

# A Decision Theoretic View on Choosing Heuristics for Discovery of Graphical Models

Y. Xiang

University of Guelph, Canada

## Abstract

Discovery of graphical models is NP-hard in general, which justifies using heuristics. We consider four commonly used heuristics. We summarize the underlying assumptions and analyze their implications as to what graphical models they can and cannot discover. In particular, we consider discovery of pseudo-independent (PI) models, a subclass of probabilistic models where subsets of a set of collectively dependent variables display marginal independence. We show that some heuristics essentially exclude PI models other than the simplest from the model search space. We argue for a decision theoretic perspective for choosing heuristics and emphasize its implication to mission critical applications.

## Introduction

Graphical models such as Bayesian networks (BNs) (Pearl 1988), decomposable Markov networks (DMNs) (Xiang, Wong, & Cercone 1997), and chain graphs (Lauritzen 1996) have been applied successfully to probabilistic reasoning in intelligent systems. These models describe the state of an environment by a set of variables. Dependencies among these variables are encoded by a graph where nodes correspond to variables and links correspond to direct dependence between nodes connected. Conditional independencies (CIs) are encoded by graphical separation. Strengths of dependencies are quantified through a set of probability distributions associated with components of the graph. The graph is so structured that a joint probability distribution (JPD) over all variables can be factorized into a product of the set of probability distributions associated with the graph.

The fundamental assumption underlying the success of graphical models is that, in most practical environments, not everything is directly dependent on everything else. Under this assumption of *indirect dependency*, graphs in these models are sparse, and the models constitute concise representation of probabilistic knowledge and efficient organization for probabilistic inference.

Given the usefulness of graphical models, one way to construct them is by discovery from data, as demonstrated by pioneer work such as (Chow & Liu 1968; Rebane & Pearl 1987; Herskovits & Cooper 1990; Fung & Crawford 1990; Spirtes & Glymour 1991; Cooper & Herskovits 1992; Lam & Bacchus 1994; Heckerman, Geiger, & Chickering 1995; Xiang, Wong, & Cercone 1997). A single model or a set of complementary models may be discovered from a given data set. To discover a model, both its graph structure and the associated set of probability distributions must be determined. For the purpose of this paper, we assume, without

losing generality, the discovery of a single model and we focus on the discovery of model structure.

To this end, an environment is viewed as an unknown probabilistic model (equivalent to an unknown JPD) responsible for the generation of the data set. A space of alternative graphical models is searched to find the model judged as the best by the discovery algorithm, relative to the data, according to some criterion. The search is hampered, however, by the intractable number of alternative graphical models, and the task has been shown to be NP-hard (Chickering, Geiger, & Heckerman 1995). Use of heuristics is thus justified.

In this paper, we analyze some commonly used heuristics and summarize their underlying assumptions that are additional to the indirect dependency assumption. We then consider implications of such assumptions as to what graphical models they can and cannot discover. In particular, we consider discovery of pseudo-Independent (PI) models (Xiang 2005), a subclass of probabilistic models where subsets of a set of collectively dependent variables display marginal independence. We show that PI models other than the simplest are essentially excluded from the model search space by some heuristics. We exemplify the consequence of such exclusion and argue for a decision theoretic strategy for choosing heuristics, especially in mission critical applications.

## Background

This section *briefly* overviews terminologies on graphical models that are necessary to the remainder of the paper. Elaborations of these terminologies can be found in references such as (Pearl 1988; Spirtes, Glymour, & Scheines 1993; Lauritzen 1996; Jensen 2001; Xiang 2002; Neapolitan 2004) and those listed below.

In a graphical model, the dependence relations among environment variables are encoded as a graph  $G = (V, E)$ , where  $V$  is a set of nodes and  $E$  a set of links. In BNs,  $G$  is a directed acyclic graph (DAG) where each link is directed. In DMNs,  $G$  is a chordal graph where each link is undirected. Chain graphs have a mixture of directed and undirected links. They all admit factorized representation of JPD. For inference, BNs and chain graphs can first be converted into a DMN and algorithms assuming such a representation, e.g., (Jensen, Lauritzen, & Olesen 1990; Shafer 1996), can then be applied. The key conversion operation is *moralization* by which a DAG is converted into its *moral graph* by pairwise connecting parent nodes of each child and dropping the direction of links. A related concept is the *skeleton* of a directed graph, which is obtained by dropping the direction of links only. For our purpose, it suffices to focus on DMNs, although our results can be

generalized to other graphical models.

In an undirected graph  $G$ , a path or cycle  $\rho$  has a *chord* if there is a link between two non-adjacent nodes in  $\rho$ .  $G$  is *chordal* if every cycle of length  $\geq 4$  has a chord. The structure of a DMN is a chordal graph. A subset  $X$  of nodes in  $G$  is *complete* if elements of  $X$  are pairwise adjacent. A maximal set of nodes that is complete is a *clique*. Two subsets  $X$  and  $Y$  of nodes in  $G$  are *separated* by a subset  $Z$  if every path from a node in  $X$  and a node in  $Y$  contains a node in  $Z$ . A node  $x$  in  $G$  is *eliminated* if nodes adjacent to it are pairwise connected before  $x$  and its incoming links are deleted. Each link added in the process is a *fill-in*. If nodes in  $G$  can be eliminated in some order such that no fill-ins are added, then  $G$  is chordal.

Let  $V$  be a set of discrete environment variables, each of which has a finite space. The space of a set  $X \subseteq V$  of variables is the Cartesian product of the spaces of variables in  $X$ . Each element in this space is a *configuration*  $\underline{x}$  of  $X$ . A *probabilistic model* (PM) over  $V$  specifies a probability value for every configuration of every subset  $X \subseteq V$ .

For any disjoint subsets  $X, Y, Z \subset V$ , subsets  $X$  and  $Y$  are *conditionally independent* given  $Z$ , if  $P(X|Y, Z) = P(X|Z)$  whenever  $P(Y, Z) > 0$ . When  $Z = \emptyset$ , subsets  $X$  and  $Y$  are *marginally independent*. If each variable  $x \in X$  is marginally independent of  $X \setminus \{x\}$ , variables in  $X$  are marginally independent. Variables in  $X$  are *collectively dependent* if, for each proper subset  $Y \subset X$ , there exists no proper subset  $Z \subset X \setminus Y$  that satisfies  $P(Y|X \setminus Y) = P(Y|Z)$ .

A graph  $G$  is an *I-map* of a PM  $M$  over  $V$  if (1) there is an one-to-one correspondence between nodes of  $G$  and variables in  $V$ , and (2) for  $X, Y, Z \subset V$ , whenever  $X$  and  $Y$  are separated in  $G$  by  $Z$ , they are conditionally independent given  $Z$  according to  $M$ . Note that  $G$  may be directed or undirected and the criterion of separation differs in each case (Xiang 2002).  $G$  is a *minimal I-map* of  $M$  if no link in  $G$  can be removed without affecting its I-mapness. If the graphical separation relations in  $G$  correspond to all conditional independence relations in  $M$  and no more, then  $G$  is a *P-map* of  $M$ . If there exists a P-map for  $M$ , then the model  $M$  is *faithful*.

A *pseudo-independent* (PI) model (Xiang 2005) is a PM where proper subsets of a set of collectively dependent variables display marginal independence.

**Definition 1 (Full PI model)** A PM over a set  $V$  ( $|V| \geq 3$ ) of variables is a full PI model if the following hold:

- ( $S_I$ ) Variables in each  $X \subset V$  are marginally independent.
- ( $S_{II}$ ) Variables in  $V$  are collectively dependent.

$S_I$  is relaxed in partial PI models into *marginally independent partition*:

**Definition 2 (Marginally independent partition)** Let  $V$  ( $|V| \geq 3$ ) be a set of variables and  $B = \{B_1, \dots, B_m\}$  ( $m \geq 2$ ) be a partition of  $V$ .  $B$  is a marginally independent partition if for every subset  $X = \{x_k | x_k \in B_k, k = 1, \dots, m\}$ , variables in  $X$  are marginally independent. Each block  $B_i$  in  $B$  is called a marginally independent block.

**Definition 3 (Partial PI model)** A PM over  $V$  ( $|V| \geq 3$ ) is a partial PI model if the following hold:

- ( $S'_I$ )  $V$  is partitioned into marginally independent blocks.
- ( $S_{II}$ ) Variables in  $V$  are collectively dependent.

The most general PI models are those that embed one or more PI submodels:

**Definition 4 (Embedded PI submodel)** Let a PM be over a set  $V$  of variables. A proper subset  $X \subset V$  ( $|X| \geq 3$ ) forms an embedded PI submodel if the following hold:

- ( $S_{III}$ )  $X$  forms a partial PI model.
- ( $S_{IV}$ ) The partition  $\{B_1, \dots, B_m\}$  of  $X$  by  $S'_I$  extends into  $V$ . That is, there is a partition  $\{Y_1, \dots, Y_m\}$  of  $V$  such that  $B_i \subseteq Y_i$  ( $i = 1, \dots, m$ ), and, for each  $x \in Y_i$  and each  $y \in Y_j$  ( $i \neq j$ ),  $x$  and  $y$  are marginally independent.

For experimental discovery of PI models, see (Xiang *et al.* 2000). The result of this paper applies to all types of PI models and hence we refer to them as PI model without further distinction. To represent PI models as undirected graphical models, we distinguish between two types of links. In a minimal I-map of a non-PI model, all links are drawn as solid lines. In a minimal I-map of a PI model, for each embedded PI submodel, each pair of marginally independent variables are connected by a dotted line, signifying marginal independence as well as collective dependence. Each remaining pair of unconnected variables in the submodel are then connected by a solid line. Two nodes are adjacent if they are connected by either type of links.

## Some Discovery Heuristics

Given an environment described by a set  $V$  of variables and a data set over  $V$ , the task of discovery is to search the model space for a model judged as the best according to some criterion relative to the data. The search is hampered, however, by the prohibitive number of alternative models. To learn a BN, the number of directed acyclic graphs (DAGs) given  $V$  is 3 for  $|V| = 2$ , 25 for  $|V| = 3$ , and 29000 for  $|V| = 5$ . To learn a DMN, the number of chordal graphs given  $V$  is 2 for  $|V| = 2$ , 8 for  $|V| = 3$ , and 822 for  $|V| = 5$ . It has been shown (Chickering, Geiger, & Heckerman 1995) that, in general, discovery of graphical models from data is NP-hard. Hence, use of heuristics is justified.

Two broad types of heuristics can be identified. We refer to the first type as *limiting model space*. According to this type of heuristics, before discovery starts, the model space is pruned to contain a subset of graph structures.

The most common in this type is the *Naive Bayes heuristic*. It prunes the model space so that it contains only Naive Bayes models. The graph of a Naive Bayes model is a DAG of a single root, which we refer to as the *hypothesis*, and its set of observable child nodes, which we refer to as the *attributes*. Each attribute has a single parent, the hypothesis. For this heuristic, the DAG is given (since hypothesis is given). Hence, discovery focuses on finding the conditional probability distribution at each node, and is very efficient.

Another heuristic of this type is the *TAN heuristic*, which prunes the model space to contain only *tree augmented*

Naive Bayes (TAN) models. The DAG of a TAN model also has a single root, the hypothesis. However, attributes themselves form a tree. Each attribute has as parent the hypothesis and at most one other attribute. Figure 1 shows an example.

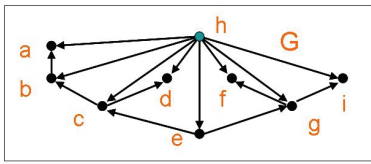


Figure 1: A TAN model where  $h$  is the hypothesis.

We refer to the second type of heuristics as *limiting search procedure*. Under this type, no explicit pruning of the model space is made before discovery. Search starts from some base model in the general model space. Based on this model, a subset of alternative models is defined by a search procedure and is evaluated. One of them, that has the best evaluation and improves over the base model, will be chosen as the new base model for the next step. The process continues until no alternative model in the subset improves over the current base model.

The above search process is fairly general and hence its behavior is mainly determined by what constitutes the candidate set given a base model. To gain efficiency, a heuristic of this type prescribes a search procedure such that from any starting base model, only a subspace of the model space will ever be visited.

The most common heuristic of this type is the *single-link lookahead*. According to this heuristic, the new base model and the current base model differ by exactly one link. An alternative is *bounded multi-link lookahead* where the two base models differ by up to  $k > 1$  links.

### Ideal Outcome from Discovery

Without limitation on what is the ideal outcome from discovery, it will be difficult, if not impossible, to compare different heuristics. For instance, how desirable the discovered graphical model is can be judged by how well it performs in a classification task. Alternatively, it can also be judged by how accurately it can estimate posterior probability. The two criteria are not equivalent. For instance, it has been reported (Rish 2001) that Naive Bayes models perform well on classification task even though they are not accurate estimators of posteriors.

This surprising success of Naive Bayes has been attributed to the winner-takes-it-all nature of the classification task. That is, as long as the correct class (a possible value of the hypothesis variable) is given the highest posterior probability value by the classifier, the accuracy of the posterior does not matter. However, tasks supported by graphical models go beyond classification. For instance, in decision theoretic reasoning, the posterior must be combined with utility in order to trade between possibility and desirability. For such tasks, accuracy of the posterior matters. Models that are good at classification but poor in posterior estimation are expected to be outperformed by good posterior esti-

maters. On the other hand, models that are good at posterior estimation are expected to perform well in decision theoretic reasoning as well as in classification.

The accuracy of a graphical model in posterior estimation depends both on its graph structure and on its numerical parameters. We take the position that the ideal outcome of the discovery is an approximate minimal I-map of the unknown PM. We choose I-map because our focus here on the structural discovery. Without being concerned with the numerical parameters, the only way to judge the accuracy of a graphical model is whether it contains the correct set of conditional independence relations. We choose I-map instead of P-map because every PM has a minimal I-map, but not necessarily a P-map (Pearl 1988). Given such a criterion, we can judge a given heuristic based on its *expressiveness*, that is, whether it enables discovery of minimal I-maps of a broad range of PMs.

### Underlying Assumptions

The assumptions underlying a heuristic has a lot to do with how expressive it is. The following two propositions summarize the independence assumptions underlying Naive Bayes and TAN heuristics. Their proofs are straightforward.

**Proposition 1** *In a Naive Bayes model, every two attributes are conditionally independent given the hypothesis.*

**Proposition 2** *In a TAN model, every two non-adjacent attributes are conditionally independent given the parent of one of them and the hypothesis.*

It is unclear what is the general assumption underlying the single-link lookahead heuristic, due to the many factors that can affect the outcome of a discovery process. Known results are all based on particular algorithms that employ the heuristic and are all centered around faithfulness. Results in (Spirtes, Glymour, & Scheines 1993; Chickering & Meek 2002) are presented as sufficient condition: If a PM  $M$  is faithful, the algorithms in question can discover a minimal I-map of  $M$ . Results in (Xiang, Wong, & Cercone 1996) are presented as necessary condition (the interpretation will be clear after Theorem 4 in the next section): If  $M$  is unfaithful, the output of the algorithms in question will not be an I-map. Therefore, we will regard faithfulness as the main assumption underlying the single-link lookahead heuristic.

The bounded multi-link lookahead heuristic is more general than the single-link lookahead heuristic as it includes single-link lookahead as a search step. An analysis (Hu & Xiang 1997) shows that if a PM contains no embedded PI submodels of more than  $\mu$  dotted links, its minimal I-map can be discovered by a bounded  $\mu$ -link lookahead. Note that faithfulness is not assumed (as will be clear after Theorem 4). Instead, it has been lifted in the spirit of the much weaker assumption on indirect dependency, which underlies all the above heuristics.

### Expressiveness

Next, we consider the expressiveness of the above heuristics. The following theorem establishes that the Naive Bayes heuristic cannot discover any PI model. In the theorem, each

element in each  $\Lambda$  is a PM and is characterized by a distinct JPD.

**Theorem 1** *Let  $\Lambda$  be the set of all Naive Bayes models over a set  $V$  of variables. Let  $\Lambda'$  be the set of all PI models over  $V$ . Then  $\Lambda \cap \Lambda' = \emptyset$ .*

Proof: Let  $M \in \Lambda$  be a Naive Bayes model. Without losing generality, we assume that each attribute of  $M$  is not marginally independent of the hypothesis. Then, from Proposition 1, an undirected minimal I-map of  $M$  is a star where the center is the hypothesis and the external nodes are the attributes. That is, no three nodes are pairwise adjacent.

Let  $M' \in \Lambda'$  be a PI model. There exists three variables  $x, y, z \in V$  such that  $x, y, z$  are all contained in the same embedded PI submodel in  $M'$ . Therefore, in any minimal I-map of  $M'$ ,  $x, y, z$  must be pairwise connected, by either solid or dotted links. That is, they are pairwise adjacent.

The above implies that a minimal I-map of  $M$  is never structured as that of  $M'$ . That is,  $\Lambda \cap \Lambda' = \emptyset$ .  $\square$

The following theorem uncovers some properties of the moral graph of a TAN model. These properties are used later to establish the expressiveness of the TAN heuristic.

**Theorem 2** *Let  $M \in \Lambda$  be a TAN model and  $G$  be its directed minimal I-map (see Background section). Then the moral graph  $G^*$  of  $G$  has the following properties:*

1.  $G^*$  is the skeleton of  $G$ .
2.  $G^*$  is chordal.
3. All cliques of  $G^*$  have size 3.

Proof: We construct  $G^*$  from  $G$  by moralization, a common operation used to convert a BN to a DMN: For each child node in  $G$ , connect its parents pairwise and drop directions of its incoming links. To establish the first statement, we show that no link will be added during moralization: The hypothesis has no parent and hence no link is added for it. Since  $M$  is a TAN, each attribute has at most two parents and at least one. If it has one parent, no link is added for it. If it has two parents, then one of them is the hypothesis and the other is an attribute. Since the parent attribute is already connected to hypothesis, no link is added. The first statement follows.

Next, we show the second statement by constructing an order in which nodes in  $G^*$  are eliminated one by one without fill-ins. We construct the order by using both  $G$  and  $G^*$  as follows: Pick a leaf node  $x$  in  $G$ . It must have exactly two parents in  $G$  and hence adjacent to exactly these two nodes in  $G^*$ . As mentioned above, the two parents are connected in  $G^*$ . Hence,  $x$  can be eliminated from  $G^*$  without fill-ins. We eliminate  $x$  from  $G^*$  and also delete  $x$  from  $G$  (with its incoming links).

The resultant  $G$  is still a TAN structure with one less attribute, and the resultant  $G^*$  is a moral graph of the new  $G$ . Therefore, another leaf node in  $G$  can be eliminated in the same fashion. By eliminating such leaf nodes recursively from  $G$  and  $G^*$ , eventually, only one attribute is left as well as the hypothesis. Both can be eliminated without any fill-in. The second statement follows.

Since  $G^*$  is chordal and the above constructed order is a fill-in free elimination order, each clique in  $G^*$  is the adjacency of a node plus the node itself when it is eliminated. The adjacency of each node when eliminated has a size 2, except the last two nodes. The third statement now follows.  $\square$

The following theorem establishes that the TAN heuristic cannot discover PI models that contain embedded PI submodels over four or more variables.

**Theorem 3** *Let  $\Lambda$  be the set of all TAN models over a set  $V$  of variables. Let  $\Lambda'$  be the set of all PI models over  $V$  such that each PI model in  $\Lambda'$  contains at least one embedded PI submodel over 4 or more variables. Then  $\Lambda \cap \Lambda' = \emptyset$ .*

Proof: Let  $M \in \Lambda$  be a TAN model,  $G$  be a directed minimal I-map of  $M$ , and  $G^*$  be the moral graph of  $G$ . Because  $G$  is a directed minimal I-map of  $M$ ,  $G^*$  is an undirected minimal I-map (Theorem 4.8 (Xiang 2002)). By Theorem 2, the cliques of  $G^*$  all have a cardinality of 3.

Next, let  $M' \in \Lambda'$  be a PI model in  $\Lambda'$ . By assumption, there exists four variables  $w, x, y, z \in V$  such that  $w, x, y, z$  are all contained in the same embedded PI submodel in  $M'$ . Therefore, in any minimal I-map of  $M'$ ,  $w, x, y, z$  must be pairwise connected, by either solid or dotted links. That is, they are contained in a same clique and this clique has a cardinality of at least 4.

The above implies that a minimal I-map of  $M$  is never structured as that of  $M'$ . That is,  $\Lambda \cap \Lambda' = \emptyset$ .  $\square$

From Theorem 2, it can be seen that if a PI model contains only embedded PI submodels of size 3, then such PI models are not excluded by the TAN heuristic. Since a PI submodel must contain at least 3 variables, by combining Theorems 2 and 3, we conclude that the TAN heuristic cannot discover PI models other than the simple.

For the single-link lookahead heuristic, recall that its main assumption is faithfulness of the data generating model. The following theory says that PI models violate the assumption.

**Theorem 4** *A PI model is unfaithful.*

Proof: Let  $M$  be a PM over a set  $V$  of variables. If  $M$  is faithful, then it has a P-map  $G$  such that, for  $X, Y, Z \subset V$ , whenever  $X$  and  $Y$  are conditionally independent given  $Z$ , they are separated in  $G$  by  $Z$ , and whenever  $X$  and  $Y$  are not conditionally independent given  $Z$ , they are not separated in  $G$  by  $Z$ .

Suppose  $M$  is a PI model. There exists a subset  $X \subseteq V$  that forms a PI submodel<sup>1</sup> and in the submodel there exist  $x, y \in X$  such that  $x$  and  $y$  are marginally independent. If  $M$  has a P-map  $G$ , because variables in  $X$  are collectively dependent, each pair of variables in  $X$  must be directly connected in  $G$ . Because  $x$  and  $y$  are marginally independent, they must be separated in  $G$  (by  $\emptyset$ ). No graph  $G$  can satisfy both conditions simultaneously. Therefore,  $M$  has no P-map.  $\square$

Theorem 4 establishes that discovery algorithms that assume faithfulness essentially exclude PI models in its model

<sup>1</sup>In the case  $X = V$ , the submodel is  $M$  itself.

search space. It has been shown (Xiang, Wong, & Cercone 1996) that indeed several algorithms are unable to discover the minimal I-map of PI model.

We conclude this section with an example PI-model and how each of the above heuristics behaves when the data is generated by this model.

**Example 1** Patient of a chronic disease changes his/her health state (denoted by variable  $s$ ) daily between stable (denoted by value  $t$ ) and unstable (denoted by value  $u$ ). Patient suffers badly in an unstable day unless treated in the morning, at which time no indicator of the state is detectable. However, if treated at the onset of a stable day, the day is spoiled due to side effect. From historical data, patient's states in four consecutive days observe the following estimated probability distribution:

$(s_1, s_2, s_3, s_4)$	$P(\cdot)$	$(s_1, s_2, s_3, s_4)$	$P(\cdot)$
$(t, t, t, t)$	0.125	$(u, t, t, t)$	0
$(t, t, t, u)$	0	$(u, t, t, u)$	0.125
$(t, t, u, t)$	0	$(u, t, u, t)$	0.125
$(t, t, u, u)$	0.125	$(u, t, u, u)$	0
$(t, u, t, t)$	0	$(u, u, t, t)$	0.125
$(t, u, t, u)$	0.125	$(u, u, t, u)$	0
$(t, u, u, t)$	0.125	$(u, u, u, t)$	0
$(t, u, u, u)$	0	$(u, u, u, u)$	0.125

What is the best strategy for treatment?

For each of the four days, the state is uniformly distributed between  $t$  and  $u$ , that is,

$$P(s_i = t) = 0.5 \quad (i = 1, 2, 3, 4).$$

The state of each day is independent of that of the previous day, that is,

$$P(s_i = t | s_{i-1}) = 0.5 \quad (i = 2, 3, 4).$$

The state of each day is independent of that of the previous two days, that is,

$$P(s_i = t | s_{i-1}, s_{i-2}) = 0.5 \quad (i = 3, 4).$$

However, the state of the last day can be precisely predicted given the states of the previous three days, for instance,

$$P(s_4 = u | s_3 = u, s_2 = t, s_1 = t) = 1.$$

In the minimal I-map of this PM, each pair of variables are directly connected by a dotted link. We denote it by  $G_6$ .

Clearly, the I-map cannot be represented either as a Naive Bayes or as a TAN. Hence, discovery algorithms assuming Naive Bayes or TAN will not discover it. In fact, since each pair of variables are marginally independent, the Naive Bayes heuristic will return an empty graph  $G_0$  (four nodes without links). Similarly, since each subset of 3 variables are marginally independent, the TAN heuristic will also return  $G_0$ .

For the single-link lookahead heuristic, we consider an algorithm that scores a graphical structure using the K-L cross entropy and starts with  $G_0$ . It can be shown (Xiang, Wong, & Cercone 1997) that the score  $KLS(G_0)$  is

$$KLS(G_0) = \sum_{i=1}^4 H(s_i),$$

where  $H(s_i)$  is the entropy of  $s_i$ ,

$$H(s_i) = -P(s_i = t) \log P(s_i = t) - P(s_i = u) \log P(s_i = u).$$

Let  $G_1$  be an alternative structure, according to the single-link lookahead heuristic, with a single link between  $s_1$  and  $s_2$ . Its score  $KLS(G_1)$  is

$$KLS(G_1) = H(s_1, s_2) + \sum_{i=3}^4 H(s_i),$$

where  $H(s_1, s_2)$  is the entropy over the variable set  $\{s_1, s_2\}$ . Since  $s_1$  and  $s_2$  are marginally independent, we have

$$H(s_1, s_2) = H(s_1) + H(s_2)$$

and hence  $KLS(G_1) = KLS(G_0)$ . The discovery algorithm thus regards  $G_1$  as no better than  $G_0$ . Since this analysis applies to any other single link added to  $G_0$ , the final outcome of the discovery will be  $G_0$ .<sup>2</sup>

In summary, each heuristic above has the outcome of  $G_0$ . The model says that the state of the patient is unpredictable and hence there is nothing we can do to help the patient. However, if a bounded 6-link lookahead heuristic is used, the correct minimal I-map  $G_6$  will have the score

$$KLS(G_6) = H(s_1, s_2, s_3, s_4) < KLS(G_0)$$

and  $G_6$  will be discovered. Note that even if this PI model is an embedded submodel in a much large PM, the bounded 6-link lookahead search is still sufficient to discover the corresponding minimal I-map. The model says that patient state can be predicted accurately from states of the previous three days. Hence, patient can be helped by treatment at the onset of each predicted unstable day.

## Decision Theoretic Perspective

The limited expressiveness of some heuristics analyzed above can be attributed to their underlying assumptions. Naive Bayes makes the strongest assumption, followed by TAN, followed by single-link lookahead, followed by bounded multi-link lookahead. Note that none of these assumptions (including the assumption of indirect dependency) are subject to the verification of the discovery process. Hence, the stronger the assumption made, the more likely that the discovered model is not the minimal I-map.

As expected, according to the complexity of the discovery computation, these heuristics are reversely ordered, although all of them are efficient. Therefore, the heuristics with stronger assumptions are mainly motivated by efficiency. In addition, the faithfulness has also been motivated by the much higher likelihood of faithful models over unfaithful ones. We argue that choosing discovery heuristics based on a decision theoretic strategy should be preferred over one based mainly on efficiency and prior model likelihood, as elaborated below. We focus on single-link lookahead versus bounded multi-link lookahead, which differ in whether to assume faithfulness. The argument is equally

<sup>2</sup>For analysis of behavior of a constraint-based algorithm, such as PC, see the above reference.

valid between heuristics that differ relative to other assumptions such as those presented in Propositions 1 and 2.

Let the choice be one of two alternative discovery algorithms  $A$  and  $A'$ , to be applied to a given environment, where  $A$  assumes faithfulness (more precisely, the heuristic used by  $A$  assumes faithfulness) and  $A'$  does not. Both algorithms are efficient but  $A'$  has higher computational cost than  $A$ . That is, discovery cost  $C_{disc}(A) = d$ ,  $C_{disc}(A') = d'$ , and  $d < d'$ . The unknown PM  $M$  of the environment has a small probability  $\epsilon$  to be unfaithful and a probability  $1 - \epsilon$  to be faithful. Choosing  $A$ , if  $M$  is faithful, the discovered model supports optimal actions. If  $M$  is unfaithful, the discovered model causes suboptimal actions. Choosing  $A'$ , no matter  $M$  is faithful or not, the discovered model supports optimal actions.

Let the action cost of a correct model (a minimal I-map) be  $C_{opt}$  and that of an incorrect model be  $C_{sub}$ . We assume that optimal actions has zero cost, that is,  $C_{opt} = 0$  and  $C_{sub} = \omega > 0$ . Then the expected cost of choosing  $A$  is

$$ECost(A) = C_{disc}(A) + (1 - \epsilon)C_{opt} + \epsilon C_{sub} = d + \epsilon \omega.$$

The expected cost of choosing  $A'$  is

$$ECost(A') = C_{disc}(A') + C_{opt} = d'.$$

Therefore, according to decision theory,  $A'$  is a better choice if and only if

$$\omega > (d' - d)/\epsilon.$$

Note that  $d$  and  $d'$  are preference functions of the actual discovery cost which may be measured, say, by execution time of the discovery algorithm. It can be estimated (Xiang & Lee 2006) as  $O(|V|^{2\mu} \mu \kappa c')$ , where  $\mu$  is the max number of lookahead links,  $\kappa$  is the max number of values of a variable,  $c'$  is the size of the maximum clique in the subgraph with newly modified links. For reasonably sized  $\mu$ ,  $\kappa$  and  $c'$ , the difference between single link lookahead ( $\mu = 1$ ) and bounded multi-link lookahead are no more than several hours in execution time. Since discovery is often off-line, such difference in execution time may mean a very small difference in  $d' - d$ . Hence, for mission critical applications, such as treatment strategy in the chronic patient example, the above inequation often holds, which suggests the less efficient but more open-minded choice,  $A'$ .

## Conclusion

Due to the complexity, heuristics must be used to render discovery of graphical models computationally tractable. All such heuristics assume indirect dependency. Each also makes additional assumptions on the data generating PM. We have demonstrated that heuristics that make stronger assumptions tend to be more efficient, but are more likely to discover a graphical model that does not correctly encode dependence relations in the PM. We have argued for a decision-theoretic strategy in choosing the heuristic based on discovery efficiency, likelihood of discovering an incorrect model, as well as consequence in applying an incorrectly discovered model in decision making.

## References

- Chickering, D., and Meek, C. 2002. Finding optimal Bayesian networks. In Darwiche, A., and Friedman, N., eds., *Proc. 18th Conf. on Uncertainty in Artificial Intelligence*, 94–102. Morgan Kaufmann.
- Chickering, D.; Geiger, D.; and Heckerman, D. 1995. Learning Bayesian networks: search methods and experimental results. In *Proc. of 5th Conf. on Artificial Intelligence and Statistics*, 112–128. Ft. Lauderdale: Society for AI and Statistics.
- Chow, C., and Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Information Theory* (14):462–467.
- Cooper, G., and Herskovits, E. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9:309–347.
- Fung, R., and Crawford, S. 1990. Constructor: A system for the induction of probabilistic models. In *Proc. of AAAI*, 762–769. Boston, MA: MIT Press.
- Heckerman, D.; Geiger, D.; and Chickering, D. 1995. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 20:197–243.
- Herskovits, E., and Cooper, G. 1990. Kutato: an entropy-driven system for construction of probabilistic expert systems from database. In *Proc. 6th Conf. on Uncertainty in Artificial Intelligence*, 54–62.
- Hu, J., and Xiang, Y. 1997. Learning belief networks in domains with recursively embedded pseudo independent submodels. In *Proc. 13th Conf. on Uncertainty in Artificial Intelligence*, 258–265.
- Jensen, F.; Lauritzen, S.; and Olesen, K. 1990. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly* (4):269–282.
- Jensen, F. 2001. *Bayesian Networks and Decision Graphs*. Springer-Verlag, New York.
- Lam, W., and Bacchus, F. 1994. Learning Bayesian networks: an approach based on the MDL principle. *Computational Intelligence* 10(3):269–293.
- Lauritzen, S. 1996. *Graphical Models*. Oxford.
- Neapolitan, R. 2004. *Learning Bayesian Networks*. Prentice Hall.
- Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- Rebane, G., and Pearl, J. 1987. The recovery of causal ploy-trees from statistical data. In Kanal, L.; Lemmer, J.; and Levitt, T., eds., *Proc. of Workshop on Uncertainty in Artificial Intelligence*, 222–228. Seattle: Elsevier Science, Amsterdam.
- Rish, I. 2001. An empirical study of the naive Bayes classifier. In *Proc. IJCAI-01 Workshop on Empirical Methods in AI*.
- Shafer, G. 1996. *Probabilistic Expert Systems*. Society for Industrial and Applied Mathematics, Philadelphia.
- Spirtes, P., and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9(1):62–73.
- Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. Springer-Verlag.
- Xiang, Y., and Lee, J. 2006. Learning decomposable Markov networks in pseudo-independent domains with local evaluation. *Machine Learning* 65(1):199–227.
- Xiang, Y.; Hu, J.; Cercone, N.; and Hamilton, H. 2000. Learning pseudo-independent models: analytical and experimental results. In Hamilton, H., ed., *Advances in Artificial Intelligence*. Springer. 227–239.
- Xiang, Y.; Wong, S.; and Cercone, N. 1996. Critical remarks on single link search in learning belief networks. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, 564–571.
- Xiang, Y.; Wong, S.; and Cercone, N. 1997. A ‘microscopic’ study of minimum entropy search in learning decomposable Markov networks. *Machine Learning* 26(1):65–92.
- Xiang, Y. 2002. *Probabilistic Reasoning in Multi-Agent Systems: A Graphical Models Approach*. Cambridge University Press, Cambridge, UK.
- Xiang, Y. 2005. Pseudo independent models. In Wang, J., ed., *Encyclopedia of Data Warehousing and Mining*. Information Science Publishing. 935–940.