Discovery of Pseudo-Independent Models from Data

Yang Xiang, University of Guelph, Canada

June 4, 2004

INTRODUCTION

Graphical models such as Bayesian networks (BNs) and decomposable Markov networks (DMNs) have been widely applied to probabilistic reasoning in intelligent systems. Figure 1 illustrates a BN and a DMN on a trivial uncertain domain: Virus can damage computer files and so can a power glitch. Power glitch also causes VCR to reset. The BN in (a) has four nodes, corresponding to four binary variables (taking values from $\{true, false\}$. The graph structure encodes a set of dependence and independence assumptions, e.g., that f is directly dependent on v and p but is independent of r once the value of p is known. Each node is associated with a conditional probability distribution conditioned on its parent nodes, e.g., $P(f|v,p)$. The joint probability distribution is the product $P(v,p,f,r) = P(f|v,p)P(r|p)P(v)P(p)$. The DMN in (b) has two groups of nodes that are maximally

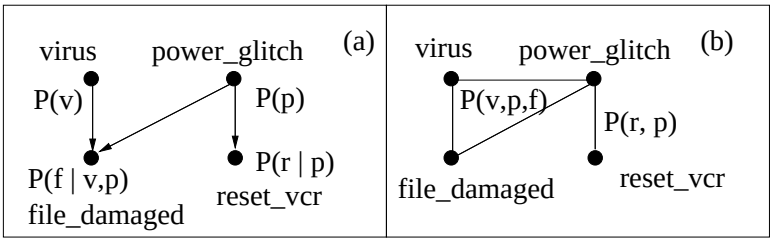

Figure 1: (a) A trivial example BN. (b) A corresponding DMN.

pairwise connected, called *cliques*. Each clique is associated with a probability distribution, e.g., clique $\{v,p,f\}$ is assigned $P(v,p,f)$. The joint probability distribution is $P(v,p,f,r) = P(v,p,f)P(r,p)/P(p)$, where $P(p)$ can be derived from one of the clique distributions. The networks can be used to reason about, say, whether there are virus in the computer system after observations on f and r are made.

Construction of such networks by elicitation from domain experts can be very time-consuming. Automatic discovery [Nea04] by exhaustively testing all possible network structures is intractable. Hence, heuristic search must be used. This chapter examines a class of graphical models that cannot be discovered using the common heuristics.

BACKGROUND

Let V be a set of n discrete variables $x_1, \ldots, x_n$ (in what follows we will focus on finite, discrete variables). Each variable x_i has a finite space $S_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,D_i}\}$ of cardinality D_i. When there is no confusion, we write $x_{i,j}$ as x_{ij} for simplicity. The space of a set V of variables is defined by the Cartesian product of the spaces of all variables in V, that is, $S_V = S_1 \times ... \times S_n$ (or $\prod_i S_i$). Thus, S_V contains the tuples made of all possible combinations of values of the variables in V. Each tuple is called a *configuration* of V, denoted by $\mathbf{v} = (x_1, \ldots, x_n)$.

Let $P(x_i)$ denote the probability function over x_i and $P(x_{ij})$ denote the probability value $P(x_i = x_{ij})$. A *probabilistic domain model* (PDM) $\mathcal{M}$ over V defines the probability values of every configuration for every subset $A \subseteq V$. Let $P(V)$ or $P(x_1, \ldots, x_n)$ denote the *joint probability distribution* (JPD) function over $x_1, \ldots, x_n$ and $P(x_{1j_1}, \ldots, x_{nj_n})$ denote the probability value of a configuration $(x_{1j_1}, \ldots, x_{nj_n})$. We refer to the function $P(A)$ over $A \subset V$ as the *marginal distribution* over A and $P(x_i)$ as the *marginal distribution* of x_i. We refer to $P(x_{1j_1}, \ldots, x_{nj_n})$ as a *joint* parameter and $P(x_{ij})$ as a *marginal* parameter of the corresponding PDM over V.

For any three disjoint subsets of variables W, U and Z in V, subsets W and U are called *conditionally independent* given Z, if

$$P(W|U, Z) = P(W|Z)$$

for all possible values in W, U and Z such that $P(U, Z) > 0$. Conditional independence signifies the dependence mediated by Z. This allows the dependence among $W \cup U \cup Z$ to be modeled over subsets $W \cup Z$ and $U \cup Z$ separately. Conditional independence is the key property explored through graphical models.

Subsets W and U are said to be *marginally independent* (sometimes referred to as *unconditionally independent*) if

$$P(W|U) = P(W)$$

for all possible values W and U such that $P(U) > 0$. When two subsets of variables are marginally independent, there is no dependence between them. Hence, each subset can be modeled independently without losing information.

If each variable x_i in a subset A is marginally independent of $A \setminus \{x_i\}$, the variables in A are said to be *marginally independent*. The following proposition reveals a useful property called *factorization* when this is the case.

Proposition 1 *If each variable x_i in a subset A is marginally independent of $A \setminus \{x_i\}$, then*

$$P(A) = \prod_{x_i \in A} P(x_i).$$

Variables in a subset A are called *generally dependent* if $P(B|A \setminus B) \neq P(B)$ for every proper subset $B \subset A$. If a subset of variables is generally dependent, its proper subsets cannot be modeled independently without losing information. A generally dependent subset of variables, however, may display conditional independence within the subset. For example, consider $A = \{x_1, x_2, x_3\}$. If $P(x_1, x_2|x_3) = P(x_1, x_2)$, i.e., $\{x_1, x_2\}$ and x_3 are marginally independent, then A is *not* generally dependent. On the other hand, if

$$P(x_1, x_2|x_3) \neq P(x_1, x_2),\ P(x_2, x_3|x_1) \neq P(x_2, x_3),\ P(x_3, x_1|x_2) \neq P(x_3, x_1),$$

then A *is* generally dependent.

Variables in A are *collectively dependent* if, for each proper subset $B \subset A$, there exists no proper subset $C \subset A \setminus B$ that satisfies $P(B|A \setminus B) = P(B|C)$. Collective dependence prevents conditional independence and modeling through proper subsets of variables. Table 1

Table 1: A PDM where $\mathbf{v} = (x_1, x_2, x_3, x_4)$.

$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$
$(0,0,0,0)$	0.0586	$(0,1,0,0)$	0.0517	$(1,0,0,0)$	0.0359	$(1,1,0,0)$	0.0113
$(0,0,0,1)$	0.0884	$(0,1,0,1)$	0.0463	$(1,0,0,1)$	0.0271	$(1,1,0,1)$	0.0307
$(0,0,1,0)$	0.1304	$(0,1,1,0)$	0.0743	$(1,0,1,0)$	0.0451	$(1,1,1,0)$	0.0427
$(0,0,1,1)$	0.1426	$(0,1,1,1)$	0.1077	$(1,0,1,1)$	0.0719	$(1,1,1,1)$	0.0353

shows the JPD over a set of variables $V = \{x_1, x_2, x_3, x_4\}$. The four variables are collectively dependent, e.g.,

$$P(x_{1,1}|x_{2,0}, x_{3,1}, x_{4,0}) = 0.257$$

and

$$P(x_{1,1}|x_{2,0}, x_{3,1}) = P(x_{1,1}|x_{2,0}, x_{4,0}) = P(x_{1,1}|x_{3,0}, x_{4,0}) = 0.3.$$

MAIN THRUST OF THE CHAPTER

Pseudo-Independent (PI) Models

A *pseudo-independent* (PI) model is a PDM where proper subsets of a set of collectively dependent variables display marginal independence [XWC97]. The basic PI model is a full PI model:

Definition 2 (Full PI model) *A PDM over a set V ($|V| \geq 3$) of variables is a* `full PI model` *if the following properties (called axioms of full PI models) hold:*

(S_I) *Variables in each proper subset of V are marginally independent.*

(S_{II}) *Variables in V are collectively dependent.*

Table 1 shows the JPD of a binary full PI model, where $V = \{x_1, x_2, x_3, x_4\}$. Its marginal parameters are

$$P(x_{1,0}) = 0.7,\ P(x_{2,0}) = 0.6,\ P(x_{3,0}) = 0.35,\ P(x_{4,0}) = 0.45.$$

Any subset of three variables are marginally independent, e.g.,

$$P(x_{1,1}, x_{2,0}, x_{3,1}) = P(x_{1,1})\ P(x_{2,0})\ P(x_{3,1}) = 0.117.$$

The four variables are collectively dependent as explained above.

Table 2: The color model where $\mathbf{v} = (x_1, x_2, x_3)$.

$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$
(red, red, red)	0.25	$(green, red, red)$	0
$(red, red, green)$	0	$(green, red, green)$	0.25
$(red, green, red)$	0	$(green, green, red)$	0.25
$(red, green, green)$	0.25	$(green, green, green)$	0

Table 2 is the JPD of the color model given earlier, where $V = \{x_1, x_2, x_3\}$. The marginal independence can be verified by

$$P(x_1 = red) = P(x_2 = red) = P(x_2 = red) = 0.5,$$

$$P(x_1 = red|x_2) = P(x_1 = red|x_3) = P(x_2 = red|x_3) = 0.5$$

and the collective dependence can be seen from $P(x_1 = red|x_2 = red, x_3 = red) = 1$.

By relaxing condition (S_I) on marginal independence, full PI models are generalized into partial PI models, which are defined through *marginally independent partition* [XHCH00] introduced below:

Definition 3 (Marginally independent partition) *Let V ($|V| \geq 3$) be a set of variables, and $B = \{B^1, \ldots, B^m\}$ ($m \geq 2$) be a partition of V. B is a* `marginally independent partition` *if, for every subset $A = \{x_i^k | x_i^k \in B^k, k = 1, \ldots, m\}$, variables in A are marginally independent. Each block B^i is called a* `marginally independent block`.

Intuitively, a marginally independent partition groups variables in V into m blocks. If one forms a subset A by taking one element from each block, then variables in A are marginally independent. Unlike full PI models, in a partial PI model, it is not necessary that every proper subset is marginally independent. Instead, that requirement is replaced with the marginally independent partition.

Definition 4 (Partial PI model) *A PDM over a set V ($|V| \geq 3$) of variables is* `a partial PI model` *if the following properties (called axioms of partial PI models) hold:*

(S_I') *V can be partitioned into two or more marginally independent blocks.*

(S_{II}) *Variables in V are collectively dependent.*

Table 3 shows the JPD of a partial PI model over two ternary variables and one binary variable, where $V = \{x_1, x_2, x_3\}$. Its marginal parameters are

$$P(x_{1,0}) = 0.3,\ P(x_{1,1}) = 0.2,\ P(x_{1,2}) = 0.5,$$

$$P(x_{2,0}) = 0.3,\ P(x_{2,1}) = 0.4,\ P(x_{2,2}) = 0.3,$$

$$P(x_{3,0}) = 0.4,\ P(x_{3,1}) = 0.6.$$

Table 3: A partial PI model where $\mathbf{v} = (x_1, x_2, x_3)$.

$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$
$(0,0,0)$	0.05	$(0,1,1)$	0.11	$(1,0,0)$	0.05	$(1,1,1)$	0.08	$(2,0,0)$	0.10	$(2,1,1)$	0.11
$(0,0,1)$	0.04	$(0,2,0)$	0.06	$(1,0,1)$	0.01	$(1,2,0)$	0.03	$(2,0,1)$	0.05	$(2,2,0)$	0.01
$(0,1,0)$	0.01	$(0,2,1)$	0.03	$(1,1,0)$	0	$(1,2,1)$	0.03	$(2,1,0)$	0.09	$(2,2,1)$	0.14

The marginally independent partition is $\{\{x_1\}, \{x_2, x_3\}\}$. Variable x_1 is marginally independent of each variable in the other block, e.g.,

$$P(x_{1,1}, x_{2,0}) = P(x_{1,1})\ P(x_{2,0}) = 0.06.$$

However, variables within block $\{x_2, x_3\}$ are dependent, e.g.,

$$P(x_{2,0}, x_{3,1}) = 0.1 \neq P(x_{2,0})\ P(x_{3,1}) = 0.18.$$

The three variables are collectively dependent, e.g.,

$$P(x_{1,1}|x_{2,0}, x_{3,1}) = 0.1 \ \text{ and } \ P(x_{1,1}|x_{2,0}) = P(x_{1,1}|x_{3,1}) = 0.2.$$

Similarly,

$$P(x_{2,1}|x_{1,0},\ x_{3,1}) = 0.61,\ P(x_{2,1}|x_{1,0}) = 0.4,\ P(x_{2,1}|x_{3,1}) = 0.5.$$

Variables that form either a full or a partial PI model may be a proper subset of V, where the remaining variables display normal dependence and independence relations. In such a case, the subset is called an *embedded* PI submodel. A PDM can contain one or more embedded PI submodels.

Definition 5 (Embedded PI submodel) *Let a PDM be over a set V of generally dependent variables. A proper subset $V' \subset V$ ($|V'| \geq 3$) of variables forms an* `embedded` *PI submodel if the following properties (axioms of embedded PI models) hold:*

(S_{III}) *V' forms a partial PI model.*

(S_{IV}) *The marginal independent partition $\{B^1, \ldots, B^m\}$ of V' extends into V. That is, there is a partition $\{A^1, \ldots, A^m\}$ of V such that $B^i \subseteq A^i$, $(i = 1, .., m)$, and for each $x \in A_i$ and each $y \in A_j$ $(i \neq j)$, x and y are marginally independent.*

Definition 5 requires that variables in V are generally dependent. It eliminates the possibility that a proper subset is marginally independent of the rest of V.

Table 4 shows the jpd of a PDM with an embedded PI model over variables x_1, x_2 and x_3, where the marginals are

$$P(x_{1,0}) = 0.3,\ P(x_{2,0}) = 0.6,\ P(x_{3,0}) = 0.3,\ P(x_{4,0}) = 0.34,\ P(x_{5,0}) = 0.59.$$

The marginally independent partition of the embedded PI model is

$$\{B^1 = \{x_1\}, B^2 = \{x_2, x_3\}\}.$$

Table 4: A PDM containing an embedded PI model where $\mathbf{v} = (x_1, x_2, x_3, x_4, x_5)$.

$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$	$\mathbf{v}$	$P(\mathbf{v})$
$(0,0,0,0,0)$	0	$(0,1,0,0,0)$	.0018	$(1,0,0,0,0)$	.0080	$(1,1,0,0,0)$	.0004
$(0,0,0,0,1)$	0	$(0,1,0,0,1)$	.0162	$(1,0,0,0,1)$	.0720	$(1,1,0,0,1)$	.0036
$(0,0,0,1,0)$	0	$(0,1,0,1,0)$	.0072	$(1,0,0,1,0)$	.0120	$(1,1,0,1,0)$	.0006
$(0,0,0,1,1)$	0	$(0,1,0,1,1)$	.0648	$(1,0,0,1,1)$	.1080	$(1,1,0,1,1)$	.0054
$(0,0,1,0,0)$	.0288	$(0,1,1,0,0)$	.0048	$(1,0,1,0,0)$	.0704	$(1,1,1,0,0)$	.0864
$(0,0,1,0,1)$	.0072	$(0,1,1,0,1)$	.0012	$(1,0,1,0,1)$	.0176	$(1,1,1,0,1)$	.0216
$(0,0,1,1,0)$	.1152	$(0,1,1,1,0)$	.0192	$(1,0,1,1,0)$	.1056	$(1,1,1,1,0)$	.1296
$(0,0,1,1,1)$	.0288	$(0,1,1,1,1)$	.0048	$(1,0,1,1,1)$	.0264	$(1,1,1,1,1)$	.0324

Outside the PI submodel, B^1 extends to include x_4 and B^2 extends to include x_5. Each variable in one block is marginally independent of each variable in the other block, e.g.,

$$P(x_{1,1}, x_{5,0}) = P(x_{1,1})\ P(x_{5,0}) = 0.413.$$

Variables in the same block are pairwise dependent, e.g.,

$$P(x_{2,1}, x_{3,0}) = 0.1 \neq P(x_{2,1})\ P(x_{3,0}) = 0.12.$$

The three variables in the submodel are collectively dependent, e.g.,

$$P(x_{1,1}|x_{2,0}, x_{3,1}) = 0.55, P(x_{1,1}|x_{2,0}) = P(x_{1,1}|x_{3,1}) = 0.7.$$

However, x_4 is independent of other variables given x_1 and x_5 is independent of other variables given x_3, displaying the normal conditional independence relation, e.g.,

$$P(x_{5,1}|x_{2,0}, x_{3,0}, x_{4,0}) = P(x_{5,1}|x_{3,0}) = 0.9.$$

PDMs with embedded PI submodels are the most general type of PI models.

Discovery of PI Models

Given a data set over n variables, the number of possible network structures is super-exponential. To make the discovery tractable, a common heuristic method is the single-link lookahead search. Learning starts with some initial graphical structure. Successive graphical structures representing different sets of conditional independence assumptions are adopted. Each adopted structure differs from its predecessor by a single link and improves a score metric optimally.

PI models pose a challenge to such algorithms. It has been shown [XWC96] that when the underlying PDM of the given data is PI, the graph structure returned by such algorithms misrepresents the actual dependence relations of the PDM. Intuitively, these algorithms update the current graph structure based on some tests for local dependence (see the next paragraph for justification). The marginal independence of a PI model misleads these algorithms into ignoring the collective dependence.

Consider a full PI model over n binary variables $x_1, ..., x_n$ where $n \geq 3$. Each x_i ($1 \leq i < n$) takes value *red* or *green* with equal chance. Variable x_n takes *red* if even number of other variables take *red* and takes *green* otherwise. If the search starts with an empty graph, then the single-link lookahead will return an empty graph because every proper subset of variables is marginally independent. From the values of any n-1 variables, this learned model will predict the n'th variable as equally likely to be *red* or *green*. In fact, when the values of any n-1 variables are known, the value of the n'th variable can be determined with certainty! When one has a life or death decision to make, one certainly does not want to use such incorrectly learned model.

Most known algorithms use a scoring metric and a search procedure. The scoring metric evaluates the goodness-of-fit of a structure to the data, and the search procedure generates alternative structures and selects the best based on the evaluation. Although not all scoring metrics explicitly test for local dependence, they are implicitly doing so or approximately doing so: Bayesian metrics (based on posterior probability of the model given the data with variations on possible prior probability of the model), description length metrics, and entropy metrics have been used by many [HC90, Bun91, CH92, LB94, MR94, HGC95, Bou94, WX94]. A Bayesian metric can often be constructed in a way that is equivalent to a description length metric, or at least approximately equal. See [Che93, Scl94] for detailed discussion. Based on the minimum description length principle, Lam and Bacchus [LB94] showed that the data encoding length is a monotonically increasing function of the Kullback-Leibler cross entropy between the distribution defined by a BN model and the true distribution. It has also been shown [XWC97] that the cross entropy of a DMN can be expressed as the difference between the entropy of the distribution defined by the DMN and the entropy of the true distribution which is a constant given a static domain. Entropy has also been used as a means to test conditional independence in learning BNs [RP87]. Therefore, the maximization of the posterior probability of a graphical model given a database [CH92, HGC95], the minimization of description length [LB94], the minimization of cross entropy between a graphical model and the true model [LB94], the minimization of entropy of a graphical model [HC90, WX94], and conditional independence tests are all closely related. The inability of several common algorithms to discover PI models is another testimony of this close relationship.

The key to correctly discover a PI model from data is to identify collective dependence. In particular, given a large problem domain which contains embedded PI submodels, the key to discovery is to identify collective dependence among variables in each submodel. This requires multi-link lookahead search, during which candidate graph structures with $k > 1$ additional links are examined before the best candidate is adopted. The multiple additional links define their endpoints as a subset of variables whose potential collective dependence is tested explicitly. Once such collective dependence is confirmed, the subset will be identified as a PI submodel. Clearly, if improperly organized, multi-link lookahead search can become intractable. Hu and Xiang [HX97] presented an algorithm, which applys single-link lookahead search and low-order (small k value) multi-link lookahead search as much as possible, and uses high-order (large k value) multi-link lookahead search only when necessary.

An experiment using data from social survey was reported in [XHCH00]. A PI model was discovered from the data on Harmful Drinking (see Table 5). The discovered DMN graphical structure is shown in Figure 2. The discovered PI model performed 10% better in prediction than the model discovered using single-link lookahead search.

Table 5: Variables in social survey data on *Harmful Drinking*

i	$Variable$	$Question$
0	$HarmSocial$	$Did\ alcohol\ harm\ friendships/social\ life?$
1	$HarmHealth$	$Did\ alcohol\ harm\ your\ physical\ health?$
2	$HrmLifOutlk$	$Did\ alcohol\ harm\ your\ outlook\ on\ life?$
3	$HarmLifMrig$	$Did\ alcohol\ harm\ your\ life\ or\ marriage?$
4	$HarmWorkSty$	$Did\ alcohol\ harm\ your\ work,\ studies,\ etc?$
5	$HarmFinance$	$Did\ alcohol\ harm\ your\ financial\ position?$
6	$NumDrivrDrink$	$How\ many\ drinks\ should\ a\ designated\ driver\ have?$
7	$NmNonDrvrDrink$	$How\ many\ drinks\ should\ non-designated\ driver\ have?$

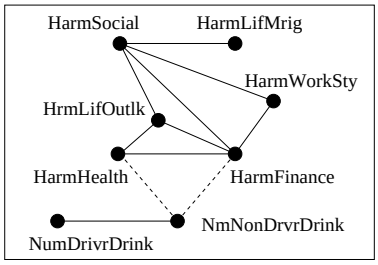

Figure 2: DMN learned from data on *Harmful drinking*.

FUTURE TRENDS

A number of issues are still open for research. A PI submodel is highly constrained by its collective dependence. Therefore, a PI submodel over k binary variables is specified by less than $2^n - 1$ probability parameters. This means that a PI submodel, though collective dependent, is simpler than a conventional complete graphical submodel. Research is needed to quantify this difference. The outcome will allow more precise scoring metrics to be devised in the next generation of discovery algorithms.

Collective dependence in PI models does not allow the conventional factorization, which is a powerful tool in both knowledge representation and probabilistic inference with graphical models. On the other hand, PI submodels are simple submodels as argued above. Research into formalisms and techniques that can explore this simplicity in both representation and inference is needed.

Causal models are stronger models than dependence models as they provides a basis for successful manipulation and control. What is the relation between PI models and its causal counterpart? How can one discover the causal structure within a PI model? Answers to these questions will make useful contributions to knowledge discovery both theoretically as well as practically.

CONCLUSION

Research in the last decade indicated that PI models exist in practice. This fact complements the theoretical analysis that for any given set of $n \geq 3$ variables, there exist infinitely many PI models, each of which is characterized by a distinct JPD. Knowledge discovery by definition is an open-minded process. The newer generation of discovery algorithms equipped with the theoretical understanding of PI models are more open-minded. They admit PI models when the data say so, thus improving the quality of knowledge discovery and allowing more accurate predictions from more accurately discovered models. The first generation of algorithms that are capable of discovering PI models demonstrates that, with a reasonable amount of extra computation (relative to single-link lookahead search), many PI models can be effectively discovered and effectively used in inference.

References

[Bou94] R.R. Bouckaert. Properties of Bayesian belief network learning algorithms. In R. Lopez de Mantaras and D. Poole, editors, *Proc. 10th Conf. on Uncertainty in Artificial Intelligence*, pages 102–109, Seattle, Washington, 1994. Morgan Kaufmann.

[Bun91] W. Buntine. Classifiers: a theoretical and empirical study. In *Proc. 1991 Inter. Joint Conf. on Artificial Intelligence*, pages 638–644, Sydney, 1991.

[CH92] G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[Che93] P. Cheeseman. Overview of model selection. In *Proc. 4th Inter. Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale, 1993. Society for AI and Statistics.

[HC90] E.H. Herskovits and G.F. Cooper. Kutato: an entropy-driven system for construction of probabilistic expert systems from database. In *Proc. 6th Conf. on Uncertainty in Artificial Intelligence*, pages 54–62, Cambridge,, 1990.

[HGC95] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[HX97] J. Hu and Y. Xiang. Learning belief networks in domains with recursively embedded pseudo independent submodels. In *Proc. 13th Conf. on Uncertainty in Artificial Intelligence*, pages 258–265, Providence, 1997.

[LB94] W. Lam and F. Bacchus. Learning Bayesian networks: an approach based on the MDL principle. *Computational Intelligence*, 10(3):269–293, 1994.

[MR94] D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Stat. Association*, 89(428):1535–1546, 1994.

[Nea04] R.E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2004.

[RP87] G. Rebane and J. Pearl. The recovery of causal ploy-trees from statistical data. In *Proc. of Workshop on Uncertainty in Artificial Intelligence*, pages 222–228, Seattle, 1987.

[Scl94] S.L. Sclove. Small-sample and large-sample statistical model selection criteria. In P. Cheeseman and R.W. Oldford, editors, *Selecting Models from Data*, pages 31–39. Springer-Verlag, 1994.

[WX94] S.K.M. Wong and Y. Xiang. Construction of a Markov network from data for probabilistic inference. In *Proc. 3rd Inter. Workshop on Rough Sets and Soft Computing*, pages 562–569, San Jose, 1994.

[XHCH00] Y. Xiang, J. Hu, N. Cercone, and H. Hamilton. Learning pseudo-independent models: analytical and experimental results. In H. Hamilton, editor, *Advances in Artificial Intelligence*, pages 227–239. Springer, 2000.

[XWC96] Y. Xiang, S.K.M. Wong, and N. Cercone. Critical remarks on single link search in learning belief networks. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence*, pages 564–571, Portland, 1996.

[XWC97] Y. Xiang, S.K.M. Wong, and N. Cercone. A 'microscopic' study of minimum entropy search in learning decomposable Markov networks. *Machine Learning*, 26(1):65–92, 1997.

TERMS AND THEIR DEFINITION

Conditional independence: Two sets X and Y of variables are conditionally independent given a third set Z, if knowledge on Z (what value Z takes) makes knowledge on Y irrelevant to guessing the value of X.

Marginal independence: Two sets X and Y of variables are marginally independent, if knowledge on Y is irrelevant to guessing the value of X.

Collective dependence: A set V of variables is collectively dependent if V cannot be split into nonempty subsets X and Y such that X and Y are marginally independent, nor can V be partitioned into nonempty subsets X, Y and Z such that X and Y are conditionally independent given Z.

Full PI model: A full PI model is a PI model where every proper subset of variables is marginally independent. Full PI models are the most basic PI models.

Partial PI model: A partial PI model is a PI model where some proper subsets of variables are not marginally independent. A partial PI model is also a full PI model, but the converse is not true. Hence, partial PI models are more general than full PI models.

Embedded PI submodel: An embedded PI submodel is a full or partial PI model over a proper subset of domain variables. The most general PI models are those over large problem domains which contain embedded PI submodels.