Parameterization of Pseudo-independent Models

Y. Xiang, University of Guelph, Guelph, Canada
J. Lee, University of Waterloo, Waterloo, Canada
N. Cercone, Dalhousie University, Halifax, Canada

Abstract

Learning belief networks from data is NP-hard in general. A common search method used in heuristic learning is the single-link lookahead. It cannot learn the underlying probabilistic model when the problem domain is *pseudo-independent*. In learning these models, to explicitly trade model goodness of fit to data and model complexity, parameterization of PI models is required. In this work, we present an improved result for computing the maximum number of parameters needed to specify a full PI model. We also present results on parameterization of a subclass of partial PI models.
`Keywords`: probabilistic reasoning, knowledge discovery, data mining, machine learning, belief networks, model complexity.

Introduction

Learning belief networks from data, as an alternative or enhancement to elicitation from experts, has been an active research area in uncertain reasoning, e.g., (Lam & Bacchus 1994; Cooper & Herskovits 1992; Heckerman, Geiger, & Chickering 1995; Friedman, Murphy, & Russell 1998). As the task is NP-hard (Chickering, Geiger, & Heckerman 1995), a common search method used in heuristic learning is the single-link lookahead, where successive graphical structures adopted differ by a single link. It has been shown that a class of probabilistic models called *pseudo-independent* (PI) models cannot be learned by single-link search (Xiang, Wong, & Cercone 1996). A more sophisticated method (multi-link lookahead) is proposed in (Xiang, Wong, & Cercone 1997) and is improved in (Hu & Xiang 1997).

The method can be further improved by incorporating the model complexity (the number of parameters) explicitly in the scoring metrics of the learning algorithm (Lam & Bacchus 1994; Xiang *et al.* 2000), so that the accuracy of the model can be better traded with the complexity. This leads to the issue of how many parameters are needed to specify a PI model.

In a previous work (Xiang Oct 1997), a formula for computing the number of parameters in a full PI model was presented. However, the result was very complex and the dependency between the parameters of the PI model and the parameters of individual variables in the model is thus obscured. In this paper, we employ the concept of a hypercube to derive a much simpler and direct formula for computing the number of parameters of a full PI model. The new formula also provide good insight on the structural relation between the complexity of a full PI model and the spaces of its domain variables. We also present results on computing the number of parameters of a subclass of partial PI model.

Background

Let V be a set of n discrete variables $X_1, \ldots, X_n$ (in what follows we will focus on finite and discrete variables). Each X_i has a finite space $S_i = \{x_{i,1}, x_{i,2}, ..., x_{i,D_i}\}$ of cardinality D_i. 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 of all possible combinations of values of the variables in V. Each tuple is called a *configuration* of V, denoted by $(x_1, \ldots, x_n)$.

Let $P(X_i)$ denote the probability function over X_i and $P(x_i)$ denote the probability value $P(X_i = x_i)$. 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 $P(x_1, ..., x_n)$ denote the probability value of a configuration $(x_1, ..., x_n)$. 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 $P(x_1, ..., x_n)$ as a *joint* parameter and $P(x_i)$ as a *marginal* parameter.

For any three disjoint subsets of variables W, U and Z in V, W and U are called *conditionally independent* given Z, denoted by $I(W, Z, U)$, if $P(W|U, Z) = P(W|U)$ for all possible values in W, U and Z such that $P(U, Z) > 0$. W and U are said to be *marginally independent* if $P(W|U) = P(W)$ for all possible values W and U. Variables in a subset A are *marginally independent* if they are pairwise marginally independent. In that case, we have $P(A) = \prod_{i=1}^{|A|} P(X_i)$.

Variables in a subset A are called *generally dependent* if $P(B|A \setminus B) \neq P(B)$ for every proper subset $B \subset A$. 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 $P(B|A \setminus B) = P(B|C)$. A *pseudo-independent* (PI) model is a PDM where proper subsets of a set of collectively dependent variables display marginal independence (Xiang, Wong, & Cercone 1997).

Definition 1 (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 any proper subset of V are marginally independent.
(S_{II}) Variables in V are collectively dependent.

The condition of marginal independence is relaxed in partial PI models, which can be defined using the concept of *marginally independent partition* (Xiang *et al.* 2000).

Definition 2 (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 $X \in B_i$ and $Y \in B_j$ ($i \neq j$) imply that X and Y are marginally independent. Each block B_i in B is called a* `marginally independent block`.

In a partial PI model, it is not necessary that every proper subset is marginally independent.

Definition 3 (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.

To facilitate the parameterization of partial PI models, we define the maximum marginally independent partition as follows:

Definition 4 (Maximum partition) *Let $B = \{B_1, ..., B_m\}$ be a marginally independent partition of a partial PI model over V. B is a* maximum marginally independent partition *if there exists no marginally independent partition B' over V such that $|B| < |B'|$.*

Why parameterizing PI models?

In learning graphical models of PDMs, one needs to balance the goodness of fit of the learned model to the data and efficiency of future inference computation using the learned model. A common technique is to score each alternative model by a combination of a score of its goodness of fit to data and a score of its model complexity. The model complexity is usually measured by the number of parameters needed to fully specify it.

Variables in a full or partial PI model are collectively dependent. They are special cases of PI models. The most general type of PI models are *embedded* PI models, where the domain includes several clusters of variables each of which forms a full PI submodel or a partial PI submodel. The remaining domain variables are 'normal' variables. The normal variables are dependent on each other and on variables in the PI submodels as in a Bayesian network or a decomposable Markov network. In order to parameterize such a model, one needs to parameterize the embedded PI submodels and combine the result with the parameterization of the normal variables.

In the previous work (Xiang, Wong, & Cercone 1997; Hu & Xiang 1997), the collective dependence of a PI submodel has led to an over-parameterization. For instance, if a PI submodel contains m variables each of k possible values, its complexity is measured as $k^m - 1$. As we will show in this paper, the actual maximum number of parameters needed to specify the PI submodel can be significantly smaller than $k^m - 1$. This over-parameterization of PI submodels will lead the learning algorithm to penalize a potential PI model as both a data-fitting and a concise learned model. The contribution of this work is to present new results leading to a theoretical foundation for correct parameterization of PI submodels.

In a previous work (Xiang Oct 1997), the following theorem for computing the number of parameters in a full PI model was presented.

Theorem 5 *(Xiang Oct 1997) The total number of parameters of a full PI model is $W = W_1 + W_2$. The number W_1 is the count of marginal parameters (marginals), $W_1 = \sum_{i=1}^{n}(D_i - 1)$, where n is the total number of variables and $D_i \geq 2$ is the number of values that the ith variable can take. The number W_2 is the count of joint probability parameters (joints),*

$$W_2 = 1 + \sum_{i=1}^{n} \sum_{j=1}^{C(n,i)} \prod_{k=1, X_{j_k} \in Y_j}^{i} (D_{j_k} - 2),$$

where j ranges from 1 to the total number of combinations taking i variables out of n each time, $Y_j = \{X_{j_1}, ..., X_{j_i}\}$ denotes one combination of i variables, and D_{j_k} is the size of space for X_{j_k}.

This result is very complex (in particular, W_2). The dependency between the parameters of the PI model and the parameters of individual variables in the model is thus obscured. In the following, we derive a much simpler and direct formula through a perspective different from the previous work. The new formula also provides good insight on the structural relation between the complexity of a full PI model and the marginal parameters of its variables. In addition, we also present results on parameterization of a subclass of partial PI models.

Parameterization of full PI models

Consider a *general* PDM $\mathcal{M}$ over a set of n variables $X_1, \ldots, X_n$. The JPD of $\mathcal{M}$ consists of a total of $\prod_{i=1}^{n} D_i$ parameters. We use a graphical model to represent these parameters, called a *JPD hypercube* or simply a *hypercube*.

Given $\mathcal{M}$, its hypercube is constructed in a n-dimensional space with the axes $X_1, \ldots, X_n$. The length of the hypercube along X_i is D_i. The segment of axis X_i from $j-1$ to j, where $j = 1, 2, \ldots, D_i$, is labeled by $x_{i,j}$, the j'th value of X_i. We refer to this segment as $X_i = x_{i,j}$. The hypercube has exactly $\prod_{i=1}^{n} D_i$ cells, one for each joint parameter. The *cell* located at $X_1 = x_{1,j}, X_2 = x_{2,k}, \ldots, X_n = x_{n,m}$ is labeled by the parameter $P(X_1 = x_{1,j}, X_2 = x_{2,k}, \ldots, X_n = x_{n,m})$, or for simplicity, $p_{(j,k,\ldots,m)}$. Figure 1 shows the hypercube for a PDM with three variables, where X_1 and X_2 are ternary and X_3 is binary. The cell labeled by $p_{(1,3,2)}$ represents the probability $P(X_1 = x_{1,1}, X_2 = x_{2,3}, X_3 = x_{3,2})$.

By the rule of negation of probability, the marginal distribution of X_i can be specified by $D_i - 1$ parameters. Hence specification of marginal distributions for all n variables in

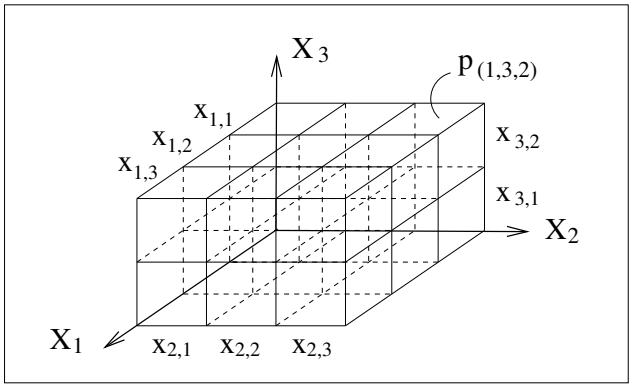

Figure 1: A 3-dimensional ($3 \times 3 \times 2$) JPD hypercube

$\mathcal{M}$ requires ω_m parameters:

$$\omega_m = \sum_{i=1}^{n} (D_i - 1). \quad (1)$$

By the rule of negation, the JPD of $\mathcal{M}$ can be specified by ω_g parameters:

$$\omega_g = \prod_{i=1}^{n} D_i - 1. \quad (2)$$

Hence, in the JPD hypercube of a general PDM, ω_g cells correspond to independent parameters (which can be freely specified) and the remaining one cell can be derived from others by the rule of negation.

By definition, a full PI model imposes constraints on the parameters of the PDM. Hence, a full PI model can be specified with fewer than ω_g parameters. In other words, more than one cell in the hypercube of a full PI model can be derived from others. From axiom S_I for full PI model, we derive the following relation between the joint parameters and marginal parameters. It says that any marginalization of the JPD is equal to the product of variable marginals.

Lemma 6 (Full PI marginal) *Let a PDM $\mathcal{M}$ be a full PI model over $V = \{X_1, \ldots, X_n\}$. Then, the following holds:*

$$\sum_{k=1}^{D_i} P(X_1, \ldots, x_{i,k}, \ldots, X_n) = P(X_1) \ldots P(X_{i-1}) P(X_{i+1}) \ldots P(X_n).$$

Proof: By marginalization, we have in general

$$\sum_{k=1}^{D_i} P(X_1, \ldots, x_{i,k}, \ldots, X_n) = P(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

By Axiom S_I, $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ are marginally independent. Therefore,

$$P(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = P(X_1) \ldots P(X_{i-1}) P(X_{i+1}) \ldots P(X_n).$$

□

From Lemma 6, the following corollary follows directly. It says that every joint parameter can be derived from marginals of $n-1$ variables plus $D_i - 1$ joint parameters.

Corollary 7 (Full PI joint) *Let a PDM $\mathcal{M}$ be a full PI model over $V = \{X_1, \ldots, X_n\}$. Then,*

$$P(X_1, \ldots, x_{i,r}, \ldots, X_n) =$$

$$P(X_1) \cdots P(X_{i-1}) P(X_{i+1}) \cdots P(X_n) - \sum_{k=1, k \neq r}^{D_i} P(X_1, \ldots, x_{i,k}, \ldots, X_n).$$

In order to determine the maximum number of independent parameters of a full PI model, we investigate in the following way: First, we specify the ω_m parameters as the marginal probability values of the n variables. Surely, these parameters are independent of each other in a general full PI model. We then investigate how many cells in the hypercube of the PDM can be derived given the ω_m marginal parameters and other cells. As soon as a cell is determined to be derivable, it is eliminated from further consideration. That is, it cannot be used to derive other cells. Once we have eliminated all derivable cells, the remaining cells and the ω_m marginal parameters constitute a maximum set of independent parameters of the full PI model. We illustrate this idea with an example before applying the idea to derive the general result.

Consider the hypercube in Figure 1. For this PDM, $\omega_m = 2 + 2 + 1 = 5$. We assume that the 5 marginal parameters have been specified. Hence, the other 3 marginal probability values can be derived.

We refer to the set of cells with the identical value $X_i = x_{i,j}$ as the *hyperplane* at $X_i = x_{i,j}$. For example, the 6 cells

$$p_{(3,1,1)}, p_{(3,2,1)}, p_{(3,3,1)}, p_{(3,1,2)}, p_{(3,2,2)}, p_{(3,3,2)}$$

form a hyperplane at $X_1 = x_{1,3}$. By Corollary 7, we have

$$p_{(3,1,1)} = P(x_{2,1}) P(x_{3,1}) - (p_{(1,1,1)} + p_{(2,1,1)}).$$

That is, the cell at the front-lower-left corner can be derived by the two cells behind it and the marginal parameters. All other cells on the hyperplane at $X_1 = x_{1,3}$ can be similarly derived. Hence, we eliminate these 6 cells from further consideration. The 12 cells behind this hyperplane are left to be considered. Using the same idea, we can show that the remaining 4 cells at the hyperplane at $X_2 = x_{2,3}$ can be derived. We therefore eliminate these 4 cells from further consideration. Now only the 8 cells in the left-hand-side of this hyperplane are to be considered. The remaining 4 cells at the hyperplane at $X_3 = x_{3,2}$ can be derived. After eliminating them, only 4 cells are left:

$$p_{(1,1,1)}, p_{(2,1,1)}, p_{(1,2,1)}, p_{(2,2,1)}$$

Since no more cells can be eliminated, the maximum number of parameters needed to specify such a full PI model is 9, with 5 marginal parameters and 4 joint parameters.

Next, we present the general result on the number of parameters needed to specify a full PI model:

Theorem 8 (Full PI parameters) *Let a PDM $\mathcal{M}$ be a full PI model over $V = \{X_1, \ldots, X_n\}$. Then the maximum number of parameters needed to specify $\mathcal{M}$ is*

$$\omega_f = \prod_{i=1}^{n} (D_i - 1) + \sum_{i=1}^{n} (D_i - 1).$$

Proof: The second term $\sum_{i=1}^{n}(D_i - 1)$ corresponds to the total number of marginal parameters required to specify the marginal distributions of the n variables. We only need to show that all joint probability values can be derived given these marginal parameters plus $\prod_{i=1}^{n}(D_i - 1)$ joint probability values.

To do so, we construct a JPD hypercube for $\mathcal{M}$. Applying Corollary 7 and using the similar argument for the example in Figure 1, we can eliminate hyperplanes at $X_1 = x_{1,D_1}$, $X_2 = x_{2,D_2}$, ..., $X_n = x_{n,D_n}$ in that order such that for each variable X_i, all cells on the hyperplane at $X_i = x_{i,D_i}$ can be derived from cells outside these hyperplanes and the marginal parameters. The remaining cells form a hypercube whose length along the X_i axis is $D_i - 1$ $(i = 1, 2, ..., n)$. The total number of cells in this hypercube is $\prod_{i=1}^{n}(D_i - 1)$. □

As an example, we apply Theorem 8 to a full PI model of 10 binary variables. The number of marginal parameters is given by $\sum_{i=1}^{10}(2-1) = 10$. The number of joint parameters is obtained from $\prod_{i=1}^{10}(2-1) = 1$. Thus, the maximum number of parameters is $10+1 = 11$. This can be compared a general PDM over 10 binary variables. The number of parameters required is $\prod_{i=1}^{10} 2 - 1 = 1023$.

As another example, consider a full PI model over 10 variables. Three of them are binary, four of them are ternary, and the remaining three each has 4 possible values. The number of marginal parameters is $3 \cdot (2-1) + 4 \cdot (3-1) + 3 \cdot (4-1) = 20$. The number of joint parameters is $(2-1)^3 \cdot (3-1)^4 \cdot (4-1)^3 = 432$. Thus, the total number of parameters is $20 + 432 = 452$.

Parameterization of partial PI models

A full PI model is a partial PI model, but the reverse is not necessarily true. Lemma 6 does not hold for a partial PI model that is not a full PI model. From axiom (S_I') of partial PI models, we derive the following relation between the joint and marginal parameters:

Lemma 9 (Partial PI marginal) *Let a PDM $\mathcal{M}$ be a partial PI model over $V = \{X_1, ..., X_n\}$ with a marginally independent partition $B = \{B_1, \ldots, B_m\}$. Let $W = \{X_{i_k} | X_{i_k} \in B_k\}$ be a subset of V with one variable from each block of B and $U = V \setminus W$. Then, the following holds:*

$$\sum_{X_j \in U} P(X_1, ..., X_n) = P(X_{i_1})...P(X_{i_m}).$$

Proof: In the lemma, each variable in U is marginalized out from the JPD. This gives $\sum_{X_j \in U} P(X_1, ..., X_n) = P(X_{i_1}, ..., X_{i_m})$. Since each pair of X_{i_k} and X_{i_l} are from different marginally independent blocks, X_{i_k} and X_{i_l} are marginally independent. Hence, $P(X_{i_1}, ..., X_{i_m}) = P(X_{i_1})...P(X_{i_m})$. The lemma follows. □

We now consider a partial PI model over five variables X_1, X_2, X_3, X_4 and X_5, where X_1, X_4 and X_5 are binary and X_2 and X_3 are ternary. The marginally independent partition is $B = \{\{X_1, X_2, X_3\}, \{X_4\}, \{X_5\}\}$. Since a 5-dimensional space cannot be illustrated with a 3-D drawing,

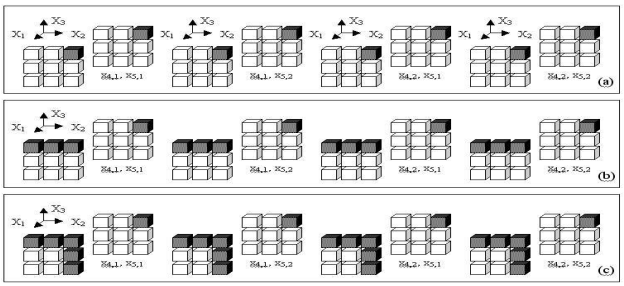

Figure 2: The joint parameters of a partial PI model.

we illustrate the corresponding hypercube using four hypercubes as shown in Figure 2 (a). All cells in each hypercube have the identical values for X_4 and X_5 as labeled beside the cube but their values on X_1, X_2 and X_3 are different. For instance, the hyperplan at the back of the first (left-most) hypercube consists of 9 cells. The cell at the bottom-left corner is the joint $P(X_1 = x_{1,1}, X_2 = x_{2,1}, X_3 = x_{3,1}, X_4 = x_{4,1}, X_5 = x_{5,1}\}$ and that at the top-left corner is $P(X_1 = x_{1,1}, X_2 = x_{2,1}, X_3 = x_{3,3}, X_4 = x_{4,1}, X_5 = x_{5,1}\}$

Apply Lemma 9 with $U = \{X_2, X_3\}$, we obtain the following equations, where each joint parameter is represented by a numerical string. For example,

$$P(X_1 = x_{1,1}, X_2 = x_{2,3}, X_3 = x_{3,2}, X_4 = x_{4,1}, X_5 = x_{5,2}\}$$

is written as (13212). The location of a digit in the string signifies the corresponding variable and the value of the digit signifies the value of the variable.

$$\begin{aligned}
&(11111) + (11211) + (11311) + (12111) + (12211) + (12311) + \\
&(13111) + (13211) + (13311) = P(x_{1,1})P(x_{4,1})P(x_{5,1}) \\
&(11112) + (11212) + (11312) + (12112) + (12212) + (12312) + \\
&(13112) + (13212) + (13312) = P(x_{1,1})P(x_{4,1})P(x_{5,2}) \\
&(11121) + (11221) + (11321) + (12121) + (12221) + (12321) + \\
&(13121) + (13221) + (13321) = P(x_{1,1})P(x_{4,2})P(x_{5,1}) \\
&(11122) + (11222) + (11322) + (12122) + (12222) + (12322) + \\
&(13122) + (13222) + (13322) = P(x_{1,1})P(x_{4,2})P(x_{5,2}) \\
&(21111) + (21211) + (21311) + (22111) + (22211) + (22311) + \\
&(23111) + (23211) + (23311) = P(x_{1,2})P(x_{4,1})P(x_{5,1}) \\
&(21112) + (21212) + (21312) + (22112) + (22212) + (22312) + \\
&(23112) + (23212) + (23312) = P(x_{1,2})P(x_{4,1})P(x_{5,2}) \\
&(21121) + (21221) + (21321) + (22121) + (22221) + (22321) + \\
&(23121) + (23221) + (23321) = P(x_{1,2})P(x_{4,2})P(x_{5,1}) \\
&(21122) + (21222) + (21322) + (22122) + (22222) + (22322) + \\
&(23122) + (23222) + (23322) = P(x_{1,2})P(x_{4,2})P(x_{5,2})
\end{aligned}$$

Assuming that we have specified all the marginal parameters, the following joint parameters (the last joint in the right-hand-side of each equation) can be derived from other joints and do not need to be specified:

$$(13311), (13312), (13321), (13322), (23311), (23312), (23321), (23322).$$

The corresponding cells are shaded in Figure 2 (a). For instance, from the first equation, we conclude that the shaded cell (13311) in the top-right corner of the hyperplan at the back of the first hypercube can be derived once we know the other cells in the same hyperplane (and the relevant marginals).

Next we apply Lemma 9 with $U = \{X_1, X_3\}$, we obtain the following equations.

$$\begin{aligned}
&(11111) + (11211) + (11311) + (21111) + (21211) + (21311) \\
&= P(x_{2,1})P(x_{4,1})P(x_{5,1}) \\
&(11112) + (11212) + (11312) + (21112) + (21212) + (21312) \\
&= P(x_{2,1})P(x_{4,1})P(x_{5,2}) \\
&(11121) + (11221) + (11321) + (21121) + (21221) + (21321) \\
&= P(x_{2,1})P(x_{4,2})P(x_{5,1})
\end{aligned}$$

$$(11122)+(11222)+(11322)+(21122)+(21222)+(21322)$$
$$=P(x_{2,1})P(x_{4,2})P(x_{5,2})$$
$$(12111)+(12211)+(12311)+(22111)+(22211)+(22311)$$
$$=P(x_{2,2})P(x_{4,1})P(x_{5,1})$$
$$(12112)+(12212)+(12312)+(22112)+(22212)+(22312)$$
$$=P(x_{2,2})P(x_{4,1})P(x_{5,2})$$
$$(12121)+(12221)+(12321)+(22121)+(22221)+(22321)$$
$$=P(x_{2,2})P(x_{4,2})P(x_{5,1})$$
$$(12122)+(12222)+(12322)+(22122)+(22222)+(22322)$$
$$=P(x_{2,2})P(x_{4,2})P(x_{5,2})$$
$$(13111)+(13211)+(13311)+(23111)+(23211)+(23311)$$
$$=P(x_{2,3})P(x_{4,1})P(x_{5,1})$$
$$(13112)+(13212)+(13312)+(23112)+(23212)+(23312)$$
$$=P(x_{2,3})P(x_{4,1})P(x_{5,2})$$
$$(13121)+(13221)+(13321)+(23121)+(23221)+(23321)$$
$$=P(x_{2,3})P(x_{4,2})P(x_{5,1})$$
$$(13122)+(13222)+(13322)+(23122)+(23222)+(23322)$$
$$=P(x_{2,3})P(x_{4,2})P(x_{5,2})$$

From the first 8 equations, the following joint parameters can be derived from other joints:

$$(21311),(21312),(21321),(21322),(22311),(22312),(22321),(22322).$$

They correspond to the additional shaded cells in Figure 2 (b). Since the last 4 equations contain the joint parameters

$$(23311),(23312),(23321),(23322)$$

in the previously to-be-derived group, each of them needs to be derived from others. These cells have already been shadded. Hence, no additional parameters can be derived using these equations.

Finally, we apply Lemma 9 with $U=\{X_1,X_2\}$ and obtain the following equations.

$$(11111)+(12111)+(13111)+(21111)+(22111)+(23111)$$
$$=P(x_{3,1})P(x_{4,1})P(x_{5,1})$$
$$(11112)+(12112)+(13112)+(21112)+(22112)+(23112)$$
$$=P(x_{3,1})P(x_{4,1})P(x_{5,2})$$
$$(11121)+(12121)+(13121)+(21121)+(22121)+(23121)$$
$$=P(x_{3,1})P(x_{4,2})P(x_{5,1})$$
$$(11122)+(12122)+(13122)+(21122)+(22122)+(23122)$$
$$=P(x_{3,1})P(x_{4,2})P(x_{5,2})$$
$$(11211)+(12211)+(13211)+(21211)+(22211)+(23211)$$
$$=P(x_{3,2})P(x_{4,1})P(x_{5,1})$$
$$(11212)+(12212)+(13212)+(21212)+(22212)+(23212)$$
$$=P(x_{3,2})P(x_{4,1})P(x_{5,2})$$
$$(11221)+(12221)+(13221)+(21221)+(22221)+(23221)$$
$$=P(x_{3,2})P(x_{4,2})P(x_{5,1})$$
$$(11222)+(12222)+(13222)+(21222)+(22222)+(23222)$$
$$=P(x_{3,2})P(x_{4,2})P(x_{5,2})$$
$$(11311)+(12311)+(13311)+(21311)+(22311)+(23311)$$
$$=P(x_{3,3})P(x_{4,1})P(x_{5,1})$$
$$(11312)+(12312)+(13312)+(21312)+(22312)+(23312)$$
$$=P(x_{3,3})P(x_{4,1})P(x_{5,2})$$
$$(11321)+(12321)+(13321)+(21321)+(22321)+(23321)$$
$$=P(x_{3,3})P(x_{4,2})P(x_{5,1})$$
$$(11322)+(12322)+(13322)+(21322)+(22322)+(23322)$$
$$=P(x_{3,3})P(x_{4,2})P(x_{5,2})$$

From the first 8 equations, the following joint parameters can be derived from other joints:

$$(23111),(23112),(23121),(23122),(23211),(23212),(23221),(23222).$$

They correspond to the additional shaded cells in Figure 2 (c). The last 4 equations contain the joint parameters

$$(23311),(23312),(23321),(23322).$$

that have already been shadded. No additional parameters can be derived using these equations.

From the figure, there are 48 joint parameters unshaded. With the 7 marginals counted, the maximum number of parameters needed to specify this partial PI model is 55. The number of joints needed can be calculated as

$$(D_1*D_2*D3-(D_1-1)-(D_2-1)-(D_3-1)-1)*D4*D_5$$
$$=(2*3*3-1-2-2-1)*2*2=48.$$

Below we prove the general case for such partial PI models.

Theorem 10 (Partial PI parameter) *Let a PDM $\mathcal{M}$ be a partial PI model with a maximum marginally independent partition $B=\{B_1,\ldots,B_h\}$, where B_1 contains m variables $X_1,X_2,...,X_m$ and each other block is a singleton. Then the maximum number of parameters needed to specify $\mathcal{M}$ is*

$$\omega_p=[(\prod_{i=1}^{m}D_i)-(\sum_{i=1}^{m}(D_i-1))-1][\prod_{i=m+1}^{h+m-1}D_i]+[\sum_{i=1}^{h+m-1}(D_i-1)].$$

Conclusion

In this work, we present an improved parameterization of full PI models, that is simple and more insightful than the previous result. We present a parameterization of partial PI models whose maximum marginal independent partition contains all singleton blocks except one. We employ the hypercube representation for analyzing the parameterization of PI models, which provide a visually appealing tool that facilitates the task.

The hypercube representation and the parameterization of the subclass of partial PI models provide a new base for research into the parameterization of general partial PI models and ultimately general PI models (the embedded PI models). The parameterization of general PI models will provide a foundation to a new generation of algorithms for learning probabilistic graphical models with embedded PI submodels.

References

Chickering, D.; Geiger, D.; and Heckerman, D. 1995. Learning Bayesian networks: serach methods and experimental results. In *Proc. of 5th Conf. on Artificial Intelligence and Statistics*, 112–128. Ft. Lauderdale: Society for AI and Statistics.

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

Friedman, N.; Murphy, K.; and Russell, S. 1998. Learning the structure of dynamic probabilistic networks. In Cooper, G., and Moral, S., eds., *Proc. 14th Conf. on Uncertainty in Artificial Intelligence*, 139–147. Madison, Wisconsin: Morgan Kaufmann.

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

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.

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

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. Oct 1997. Towards understanding of pseudo-independent domains. In *Poster Proc. 10th Inter. Symposium on Methodologies for Intelligent Systems*.