Towards Better Scoring Metrics For Pseudo-Independent Models

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

Abstract

Learning belief networks from data is NP-hard in general. A common method used in heuristic learning is the single-link lookahead search. When the problem domain is *pseudo-independent* (PI), the method cannot discover the underlying probabilistic model. In learning these models, to explicitly trade model accuracy and model complexity, parameterization of PI models is necessary. In this work, we adopt a hypercube perspective to analyze PI models and derive 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.

1 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., [7, 3, 4] [1]. As the task is NP-hard [2], 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, where variables collectively dependent display marginal independence, cannot be learned by single-link search [16]. There are infinitely many PI models over a given set of variables. PI models found in real world data are reported in [15]. Intuitively, these methods (that fail to learn PI models correctly) update the current graph structure based on some tests for local dependence (see Section 2 for more details). The marginal independence of a PI model misleads these algorithms into ignoring the collective dependence. When an incorrectly learned model is used for making decisions, it fails to take into account of the (missed) collective dependence and results in incorrect actions. To relax this limitation, a more sophisticated method (multi-link lookahead) is proposed in [17] and is improved in [6].

[1] We apologize that references are trimmed to a minimum due to space limit.

The method can be further improved by incorporating the model complexity (the number of parameters) explicitly in the scoring metrics of the learning algorithm [7, 15], so that the accuracy of the model can be better traded with the complexity. This leads to the issue of estimation of the number of parameters needed to specify a PI model.

More generally, a practical data sample is finite and subject to sampling noise. Any attempt to model the data is an abstraction of the reality. Depending on the levels of abstraction and the bias introduced, a given data sample can potentially be abstracted into distinct models. Under this perspective, whether to abstract the data as a PI model becomes another dimension of learning in order to best balance the model complexity and accuracy of inference performed using the learned model. Accurate assessment of the complexity of PI models becomes more valuable under this perspective.

This work is an integral part of the overall research outlined above. A PI model can be *full* or *partial* as defined precisely in the next section. In a previous work [13], a formula for estimating the number of parameters in a full PI model was presented. However, the result was very complex in form, obscuring the dependency between parameters of the PI model and parameters of individual variables. This in turn makes its extension to partial PI models difficult. In this paper, we employ the perspective of a hypercube to derive a much simpler and direct formula for estimating the number of parameters of a full PI model. The new formula provides good insight on the structural relation between the complexity of a full PI model and the spaces of its domain variables. We also extend the result to estimate the number of parameters of a subclass of partial PI models.

2 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. 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 $(x_1, \ldots, x_n)$.

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 $P(x_1, \ldots, x_n)$ denote the probability value of a configuration $(x_1, \ldots, x_n)$. 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 $P(x_1, \ldots, x_n)$ as a *joint* parameter and $P(x_i)$ 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 $W \cup U$ and $U \cup Z$ separately. Conditional independence is the key property explored through belief networks [9, 8, 14].

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

Proof: By the product rule of probability, we have

$$P(A) = P(X_1|X_2,\ldots,X_n)P(X_2|X_3,\ldots,X_n)...P(X_{n-1}|X_n).$$

By the decomposition property [9] of conditional independence, namely, that

$$P(A|Z,U,W) = P(A|Z) \textit{ implies } P(A|Z,U) = P(A|Z) \textit{ and } P(A|Z,W) = P(A|Z),$$

and the assumption that X_i is marginally independent of $A \setminus \{X_i\}$, the tail in each $P(X_i|X_{i+1},\ldots,X_n)$ can be omitted.

□

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

A *pseudo-independent* (PI) model is a PDM where proper subsets of a set of collectively dependent variables display marginal independence [17].

V	$P(V)$	V	$P(V)$	V	$P(V)$	V	$P(V)$
$(0,0,0,0)$	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

Table 1: A full PI model where $V = (X_1, X_2, X_3, X_4)$.

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 any 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 $V = (X_1, X_2, X_3, X_4)$ and the 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.

The condition (S_I) of marginal independence is relaxed in partial PI models, which is defined through *marginally independent partition* [15] 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 in B is called a* `marginally independent block`.

Intuitively, a marginally independent partition of a set V of variables groups variables in V into m blocks. If one forms a subset A of V by taking one element from each block, then variables in A are always marginally independent.

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

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.*

V	$P(V)$	V	$P(V)$	V	$P(V)$	V	$P(V)$	V	$P(V)$	V	$P(V)$
$(0,0,0)$	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

Table 2: A partial PI model where $V = (X_1, X_2, X_3)$.

Table 2 shows the JPD of a partial PI model over two ternary variables and one binary variable, where $V = (X_1, X_2, X_3)$ and the marginal parameters are

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

$$P(x_{2,0}) = 0.3,\ P(x_{2,1}) = 0.4,\ P(x_{2,2}) = 0.3,\ P(x_{3,0}) = 0.4,\ P(x_{3,1}) = 0.6.$$

The marginally independent partition is $\{\{X_1\}, \{X_2, X_3\}\}$. Variable X_1 is marginally independent of each variable in the other subset, e.g.,

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

However, the variables in the subset $\{X_2, X_3\}$ 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 also be a subset of V, where the remaining variables in V display conventional dependence. In such case, the subset is called an *embedded* PI submodel. A PDM can contain one or more embedded PI submodels. PDMs with embedded PI submodels are the most general type of PI models. In this work, we focus on only full or partial PI models.

Learning belief networks from data is NP-hard [2]. A common heuristic method used is a greedy 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 is shown [16] that when the underlying PDM of the given data is PI, the graph structure returned

by such algorithms misrepresents the actual dependence of the PDM. Intuitively, these algorithms update the current graph structure based on some tests for local dependence (see below for justification). The marginal independence of a PI model misleads these algorithms into ignoring the collective dependence. The primary goal of our ongoing research (to which this work is an integral part) is to develop a newer generation of algorithms that overcome this limitation.

All 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 implicitlyly 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 [5, 3, 7, 4, 12]. A Bayesian metric can often be constructed in a way that is equivalent to a description length metric, or at least approximately equal. See [1, 11] for detailed discussion. Based on the minimum description length principle, Lam and Bacchus [7] showed that the data encoding length is a monotonically increasing function of the Kullback-Leibler cross entropy between the distribution defined by a Bayesian network (BN) model and the true distribution. It has also been shown [17] that the cross entropy of a decomposable Markov network (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 [10]. Therefore, the maximization of the posterior probability of a network model given a database [3, 4], the minimization of description length [7], the minimization of cross entropy between a network model and the true model [7], the minimization of entropy of a network model [5, 12], and conditional independence tests are all closely related.

Before closing this section, we define the maximum marginally independent partition to be used later for parameterization of partial PI models:

Definition 5 (Maximum partition) *Let $B = \{B_1, \ldots, B_m\}$ 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 $|B| < |B'|$.*

Given a marginally independent partition, one can always obtain another marginally independent partition by merging. For instance, given a partition $B' = \{B_1, B_2, B_3, \ldots, B_m\}$, another partition can be defined as $B = \{B_1 \cup B_2, B_3, \ldots, B_m\}$, where $|B| < |B'|$. Therefore, a maximum marginally independent partition is the 'finest' partition that retains the property of marginal independence.

3 Why Parameterizing PI models?

In learning graphical models of PDMs, one needs to balance the accuracy of the learned model and efficiency of future inference performed based on the learned model (the model complexity). 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 the model.

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 [9] or a decomposable Markov network [17]. 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 [17, 6], 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 was 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 leads a learning algorithm to over-penalize a potential PI model and to produce incorrectly biased learning outcome. The contribution of this work is the new results leading to a theoretical foundation for correct parameterization of PI submodels.

In a previous work [13], the following result for computing the number of parameters in a full PI model was given:

Theorem 6 *[13] 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* i*th 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,* $Y_j = \{X_{j_1}, \ldots, X_{j_i}\}$ *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 complexity of the PI model and the space cardinality of each individual variable is thus obscured. In the following, we derive a much simpler and

direct formula through a new perspective. 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 present results on parameterization of a subclass of partial PI models.

4 Parameterization of Full PI Models

Consider a general PDM $\mathcal{M}$ over a set of n variables $V = \{X_1, \ldots, X_n\}$. The JPD of $\mathcal{M}$ consists of a total of $\prod_{i=1}^{n} D_i$ parameters. To facilitate visualization and analysis, we use a graphical representation of 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 $P(X_1 = x_{1,1}, X_2 = x_{2,3}, X_3 = x_{3,2})$.

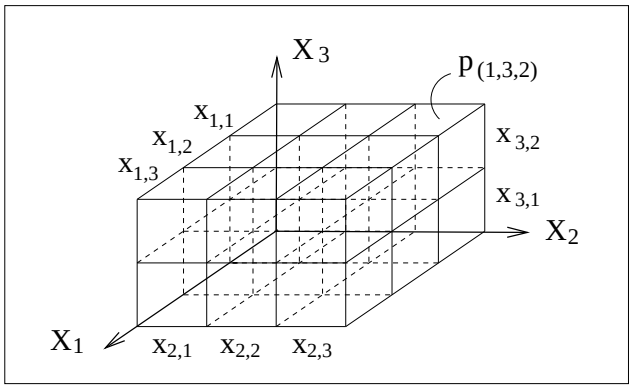

Figure 1: A 3-dimensional ($3 \times 3 \times 2$) JPD hypercube.

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 $\mathcal{M}$ requires ω_m parameters:

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

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

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

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 models, 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 7 (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 7, 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. Note that the summation index k runs from 1 to D_i except that it skips $k = r$.

Corollary 8 (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 adopt the following approach: 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 search for each cell in the hypercube of the PDM that is derivable from these 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 formalize 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 by rule of negation.

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 at the front of Figure 1

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

form a hyperplane at $X_1 = x_{1,3}$. By Corollary 8, 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

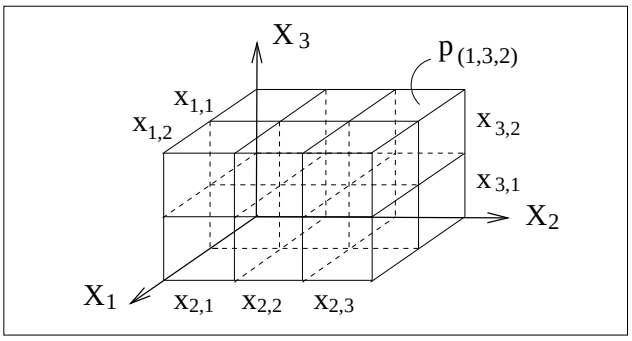

Figure 2: The 3-dimensional ($3 \times 3 \times 2$) JPD hypercube with 6 cells at $X_1 = x_{1,3}$ eliminated.

further consideration. The remaining 12 cells are shown in Figure 2.

Using the same idea, four of the remaining 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, as shown in Figure 3 (a). The remaining 4 cells at the

hyperplane at $X_3 = x_{3,2}$ can be derived. After eliminating them, only the 4 cells in Figure 3 (b) 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

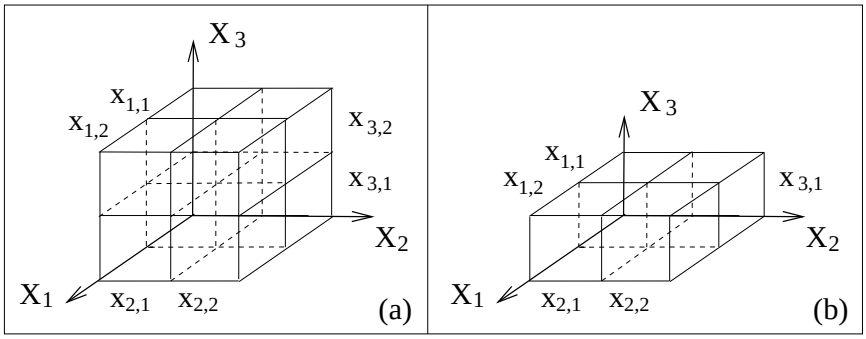

Figure 3: (a) The 3-dimensional ($3 \times 3 \times 2$) JPD hypercube with cells at $X_2 = x_{2,3}$ eliminated. (b) The cells at the hyperplane at $X_3 = x_{3,2}$ are eliminated.

needed to specify such a full PI model is 9, with 5 marginal parameters and 4 joint parameters. Note that it would take 17 parameters to specify the JPD of a general PDM over three variables of the same space cardinalities.

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

Theorem 9 (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*

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

Before proving the theorem, it can be seen that this result is significantly simpler than Theorem 6. It shows that the maximum number of parameters needed to specify a full PI model consists of two terms. One term is the cardinality of the joint space of a general PDM over the same set of variables, except the space of each variable is reduced by one element. The other term is the number of marginal parameters.

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 8 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 the hyperplane and the marginal parameters. The remaining cells form a hypercube whose length along the X_i axis is $D_i - 1$ $(i = 1, 2, \ldots, n)$. The total number of cells in this hypercube is $\prod_{i=1}^{n}(D_i - 1)$. □

As an example, we apply Theorem 9 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 with 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 $20 + 432 = 452$. The number of parameters required for a general PDM over the same set of variables is $2^3 \cdot 3^4 \cdot 4^3 - 1 = 41471$.

Clearly, a full PI model is significantly more compact than a general PDM. This compactness can be explored both for more accurate model learning and for reduced model complexity. As mentioned in Section 2, a full PI model may be present in a PDM as a submodel. The benefit from exploration of this compactness is the same.

5 Parameterization of Partial PI Models

A full PI model is a partial PI model, but the reverse is not necessarily true. Lemma 7 does not hold for a partial PI model that is not a full PI model. From axiom (S'_I) of partial PI models, we make explicit the following relation between the joint and marginal parameters:

Lemma 10 (Partial PI marginal) *Let a PDM $\mathcal{M}$ be a partial PI model over $V = \{X_1, \ldots, X_n\}$ 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:*

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

Proof: In the left hand side of the above equation, each variable in U is marginalized out from the JPD. This gives

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

The lemma follows as a direct result of the definition of marginally independent partition. □

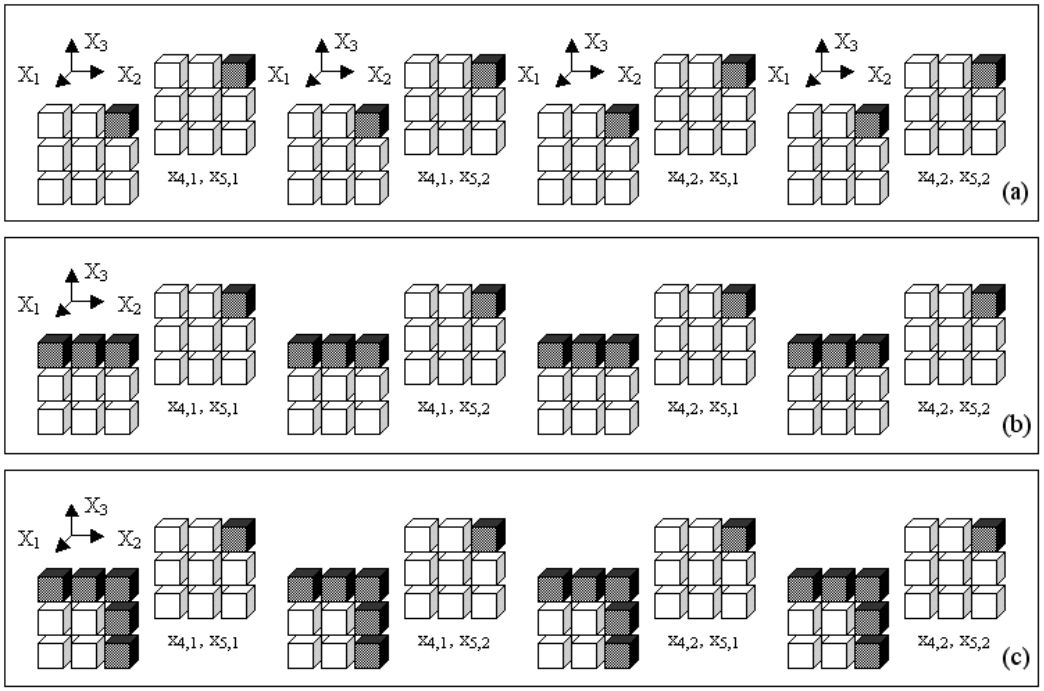

Figure 4: The joint parameters of a partial PI model.

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. Assume that 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, we illustrate the corresponding hypercube using four hypercubes as shown in Figure 4 (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 hyperplane at the back of the first (left-most) hypercube consists of 9 cells. The cell at the bottom-left corner is the joint parameter

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

and that at the top-left corner is

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

Applying Lemma 10 with $U = \{X_2, X_3\}$, i.e., performing

$$\sum_{X_2,X_3} P(X_1, \ldots, X_5) = P(X_1)P(X_4)P(X_5),$$

we obtain Eqns (3) through (10) below, 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.

$$\begin{aligned}
& (11111) + (11211) + (11311) + (12111) + (12211) + (12311) + (13111) + (13211) + (13311) \\
= \;& P(x_{1,1})P(x_{4,1})P(x_{5,1}) && (3) \\
& (11112) + (11212) + (11312) + (12112) + (12212) + (12312) + (13112) + (13212) + (13312) \\
= \;& P(x_{1,1})P(x_{4,1})P(x_{5,2}) && (4) \\
& (11121) + (11221) + (11321) + (12121) + (12221) + (12321) + (13121) + (13221) + (13321) \\
= \;& P(x_{1,1})P(x_{4,2})P(x_{5,1}) && (5) \\
& (11122) + (11222) + (11322) + (12122) + (12222) + (12322) + (13122) + (13222) + (13322) \\
= \;& P(x_{1,1})P(x_{4,2})P(x_{5,2}) && (6) \\
& (21111) + (21211) + (21311) + (22111) + (22211) + (22311) + (23111) + (23211) + (23311) \\
= \;& P(x_{1,2})P(x_{4,1})P(x_{5,1}) && (7) \\
& (21112) + (21212) + (21312) + (22112) + (22212) + (22312) + (23112) + (23212) + (23312) \\
= \;& P(x_{1,2})P(x_{4,1})P(x_{5,2}) && (8) \\
& (21121) + (21221) + (21321) + (22121) + (22221) + (22321) + (23121) + (23221) + (23321) \\
= \;& P(x_{1,2})P(x_{4,2})P(x_{5,1}) && (9) \\
& (21122) + (21222) + (21322) + (22122) + (22222) + (22322) + (23122) + (23222) + (23322) \\
= \;& P(x_{1,2})P(x_{4,2})P(x_{5,2}). && (10)
\end{aligned}$$

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

$$S_1 = \{(13311), (13312), (13321), (13322), (23311), (23312), (23321), (23322)\}.$$

The corresponding cells are shaded in Figure 4 (a). For instance, from Eqn (3), we conclude that the shaded cell (13311) in the top-right corner of the hyperplane 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 10 with $U = \{X_1, X_3\}$, i.e., perform

$$\sum_{X_1,X_3} P(X_1, \ldots, X_5) = P(X_2)P(X_4)P(X_5),$$

and obtain Eqns (11) through (22) below,

$$
\begin{aligned}
(11111)+(11211)+(11311)+(21111)+(21211)+(21311) &= P(x_{2,1})P(x_{4,1})P(x_{5,1}) && (11)\\
(11112)+(11212)+(11312)+(21112)+(21212)+(21312) &= P(x_{2,1})P(x_{4,1})P(x_{5,2}) && (12)\\
(11121)+(11221)+(11321)+(21121)+(21221)+(21321) &= P(x_{2,1})P(x_{4,2})P(x_{5,1}) && (13)\\
(11122)+(11222)+(11322)+(21122)+(21222)+(21322) &= P(x_{2,1})P(x_{4,2})P(x_{5,2}) && (14)\\
(12111)+(12211)+(12311)+(22111)+(22211)+(22311) &= P(x_{2,2})P(x_{4,1})P(x_{5,1}) && (15)\\
(12112)+(12212)+(12312)+(22112)+(22212)+(22312) &= P(x_{2,2})P(x_{4,1})P(x_{5,2}) && (16)\\
(12121)+(12221)+(12321)+(22121)+(22221)+(22321) &= P(x_{2,2})P(x_{4,2})P(x_{5,1}) && (17)\\
(12122)+(12222)+(12322)+(22122)+(22222)+(22322) &= P(x_{2,2})P(x_{4,2})P(x_{5,2}) && (18)\\
(13111)+(13211)+(13311)+(23111)+(23211)+(23311) &= P(x_{2,3})P(x_{4,1})P(x_{5,1}) && (19)\\
(13112)+(13212)+(13312)+(23112)+(23212)+(23312) &= P(x_{2,3})P(x_{4,1})P(x_{5,2}) && (20)\\
(13121)+(13221)+(13321)+(23121)+(23221)+(23321) &= P(x_{2,3})P(x_{4,2})P(x_{5,1}) && (21)\\
(13122)+(13222)+(13322)+(23122)+(23222)+(23322) &= P(x_{2,3})P(x_{4,2})P(x_{5,2}). && (22)
\end{aligned}
$$

From Eqns (11) through (18), the following set of joint parameters can be derived from others:

$$S_2 = \{(21311), (21312), (21321), (21322), (22311), (22312), (22321), (22322)\}.$$

They correspond to the additional shaded cells in Figure 4 (b).

Eqns (19) through (22) contain the joint parameters

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

in the set S_1. Each of them needs to be derived from others. These cells have already been shaded. Hence, no additional parameters can be derived using these equations.

Finally, we apply Lemma 10 with $U = \{X_1, X_2\}$, i.e., perform

$$\sum_{X_1, X_2} P(X_1, \ldots, X_5) = P(X_3)P(X_4)P(X_5),$$

and obtain Eqns (23) through (34)

$$
\begin{aligned}
(11111)+(12111)+(13111)+(21111)+(22111)+(23111) &= P(x_{3,1})P(x_{4,1})P(x_{5,1}) && (23)\\
(11112)+(12112)+(13112)+(21112)+(22112)+(23112) &= P(x_{3,1})P(x_{4,1})P(x_{5,2}) && (24)\\
(11121)+(12121)+(13121)+(21121)+(22121)+(23121) &= P(x_{3,1})P(x_{4,2})P(x_{5,1}) && (25)\\
(11122)+(12122)+(13122)+(21122)+(22122)+(23122) &= P(x_{3,1})P(x_{4,2})P(x_{5,2}) && (26)\\
(11211)+(12211)+(13211)+(21211)+(22211)+(23211) &= P(x_{3,2})P(x_{4,1})P(x_{5,1}) && (27)\\
(11212)+(12212)+(13212)+(21212)+(22212)+(23212) &= P(x_{3,2})P(x_{4,1})P(x_{5,2}) && (28)\\
(11221)+(12221)+(13221)+(21221)+(22221)+(23221) &= P(x_{3,2})P(x_{4,2})P(x_{5,1}) && (29)\\
(11222)+(12222)+(13222)+(21222)+(22222)+(23222) &= P(x_{3,2})P(x_{4,2})P(x_{5,2}) && (30)\\
(11311)+(12311)+(13311)+(21311)+(22311)+(23311) &= P(x_{3,3})P(x_{4,1})P(x_{5,1}) && (31)\\
(11312)+(12312)+(13312)+(21312)+(22312)+(23312) &= P(x_{3,3})P(x_{4,1})P(x_{5,2}) && (32)\\
(11321)+(12321)+(13321)+(21321)+(22321)+(23321) &= P(x_{3,3})P(x_{4,2})P(x_{5,1}) && (33)\\
(11322)+(12322)+(13322)+(21322)+(22322)+(23322) &= P(x_{3,3})P(x_{4,2})P(x_{5,2}). && (34)
\end{aligned}
$$

From Eqns (23) through (30), the following set of joint parameters can be derived from others:

$$S_3 = \{(23111), (23112), (23121), (23122), (23211), (23212), (23221), (23222)\}.$$

They correspond to the additional shaded cells in Figure 4 (c). Eqns (31) through (34) contain the joint parameters

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

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

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

$$(D_1 * D_2 * D3 - 1 - (D_1 - 1) - (D_2 - 1) - (D_3 - 1)) * D4 * D_5$$

$$= (2 * 3 * 3 - 1 - 2 - 2 - 1) * 2 * 2 = 48.$$

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

Theorem 11 (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 $m \geq 2$ variables $X_1, X_2, \ldots, X_m$ and each other block is a singleton. Then the maximum number of parameters needed to specify $\mathcal{M}$ is*

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

Before proving the theorem, we give a brief explanation about the result. There are a total of $h+m-1$ variables in $\mathcal{M}$, indexed as $1, 2, \ldots, h+m-1$. The first m of them form the block B_1. The value ω_p is the sum of two terms: The first term $\sum_{i=1}^{h+m-1}(D_i - 1)$ is the number of parameters needed to specify marginal distributions for all variables. The second term is the number of joint parameters and is obtained as the product of two factors:

The first factor is determined by variables in B_1. It can be grouped as the difference of two terms:

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

The first term $[(\prod_{i=1}^{m} D_i) - 1]$ is the number of joint parameters needed to specify the JPD over the block B_1 if itself is a general PDM. The second term $[\sum_{i=1}^{m}(D_i - 1)]$ is the number of parameters needed to specify marginal distributions for variables in B_1.

The second factor $\prod_{i=m+1}^{h+m-1} D_i$ is determined by variables in B_2 through B_h. It is the cardinality of the joint space of these variables. We give the proof of the theorem below. To make the proof comprehensible, we use the above example to illustrate the general ideas from time to time.

Proof: The partial PI model is over $h+m-1$ variables. Without losing generality, we assume that $D_i = k \geq 2$ for $i = 1, \ldots, h+m-1$. The PI model has a total of k^{h+m-1} joint parameters. We represent the joint parameters using a $(h+m-1)$-dimensional hypercube, which can be alternatively represented as k^{h-1} hypercubes each of m-dimension. For example, Figure 4 (a) has $2^2 = 4$ hypercubes each being 3-dimensional. We choose each hypercube such that all cells in the hypercube have identical values for $X_{m+1}, \ldots, X_{h+m-1}$. Hence each hypercube is essentially a hypercube in the hyperspace of $X_1, \ldots, X_m$, namely, variables in B_1.

The remaining proof proceeds with a number of steps. In each step, a subset of $m-1$ variables is selected from B_1 arbitrarily. There are $C(m, m-1) = m$ ways to select such a subset. Hence, m steps are needed.

In the first step, suppose that the subset $\{X_2, \ldots, X_m\}$ is chosen. Applying Lemma 10 with $U = \{X_2, \ldots, X_m\}$, we obtain

$$\sum_{X_2, \ldots, X_m} P(X_1, X_2, \ldots, X_{h+m-1}) = P(X_1)P(X_{m+1})...P(X_{h+m-1}).$$

This equation can be expanded into k^h equations since the right-hand-side has h terms and each can take k possible values. In each expanded equation, the left-hand-side has k^{m-1} terms since the summation is performed over $m-1$ variables and each can take k possible values. A total of k^{h+m-1} joint parameters appear in the left-hand-side of all the k^h equations. Note that all joint parameters of the PDM appear, with each appearing exactly once.

For the above example, when $U = \{X_2, X_3\}$, Eqns (3) through (10) (a total of $2^3 = 8$ equations) are obtained. The left-hand-side of each equation has $3^2 = 9$ terms. Each of the 72 joint parameters appears in one equation.

From the k^h equations, k^h joint parameters can be derived (one from each equation) from the other joint parameters plus the marginal parameters. If we mark these cells in hypercubes, then k cells will be marked from each m-dimensional hypercube. For the above example, when $U = \{X_2, X_3\}$, 8 cells are marked with 2 (the value of D_1) cells from each 3-dimensional hypercube (see Figure 4 (a)). Suppose that we choose to mark cells with

$$X_2 = x_{2,k}, X_3 = x_{3,k}, \ldots, X_m = x_{m,k}.$$

This is valid because each of the k^h equations contains exactly one item of the form $F = P(X_1, X_2 = x_{2,k}, \ldots, X_m = x_{m,k}, X_{m+1}, \ldots, X_{h+m-1})$.

In the second step, suppose that the subset $\{X_1, \ldots, X_{m-1}\}$ is selected. Applying Lemma 10 with $U = \{X_1, \ldots, X_{m-1}\}$, we obtain

$$\sum_{X_1, \ldots, X_{m-1}} P(X_1, \ldots, X_{h+m-1}) = P(X_m)P(X_{m+1})...P(X_{h+m-1}).$$

Using a similar argument above, we can derive (mark) cells of the form

$$F' = P(X_1 = x_{1,k}, \ldots, X_{m-1} = x_{m-1,k}, X_m, \ldots, X_{h+m-1}).$$

Since cells of the form $F'' = P(X_1 = x_{1,k}, \ldots, X_m = x_{m,k}, X_{m+1}, \ldots, X_{h+m-1})$ are consistent with both F and F', these cells have already been marked in the previous group. Note that there is exactly one such cell in each of the m-dimensional hypercubes. Hence, only $k-1$ additional cells are marked in each hypercube.

For the above example, when $U = \{X_1, X_3\}$, Eqns (11) through (22) (a total of $3 * 2 * 2 = 12$ equations) are obtained. In each of Eqns (19) through (22), one cell is contained in the set S_1. Hence, only $D_2 - 1 = 3 - 1 = 2$ additional cells are marked in each hypercube in Figure 4 (b), where D_2 substitutes the value of k.

Continuing with the process, we claim that each additional step marks $k-1$ cells. We show this from a graphical perspective. At each step, each of the k^h equations is associated with a unique hyperplane in a m-dimensional hypercube. Each such hyperplane is orthogonal to the same axis of the the corresponding hypercube. For example, Eqn (3) corresponds to the hyperplane at the back of the first hypercube of Figure 4 (a). It is orthogonal to the X_1 axis, and so are the hyperplanes corresponding to Eqns (4) through (10). Before the first step, no cell has been marked. Hence, one cell can be marked for each hyperplane as in Figure 4 (a). These cells can be chosen such that they differ only in their position along the orthogonal axis.

In the second step, a different set of equations is involved, that corresponds to a different set of hyperplanes. All these hyperplanes are orthogonal to another axis. For instance, hyperplanes corresponding to Eqns (11) through (22) are orthogonal to the X_2 axis. With the above convention for cell marking, exactly one such hyperplane per hypercube has cells marked in the first step. The remaining $k-1$ hyperplanes contain no marked cells. Hence, one cell per hyperplane can be marked, yielding a total of $k-1$ additional marked cells, as we analyzed above. These cells can be chosen such that they and cells marked in the first step are contained in the same hyperplane spanned by two orthogonal axes. In Figure 4 (b), we see that two additional cells ($k = 3$) are marked per hypercube. In each hypercube, all cells marked so far are contained in the same hyperplane spanned by axes X_1 and X_2.

For each remaining step, a different orthogonal axis is used. By selecting cells to mark from the same hyperplane spanned by orthogonal axes used so far, there always exist $k-1$ hyperplanes that are orthogonal to the current selected axis and contain no marked cells. Hence, $k-1$ cells can be marked for each remaining step.

At the end of the m'th step, $m \cdot (k-1) + 1$ joint parameters are marked in each m-dimensional hypercube. The number of cells unmarked in each hypercube is then $k^m - m(k-1) - 1$. The total number of cells unmarked is $[k^m - m(k-1) - 1] \cdot k^{h-1}$. This is the maximum number of joint parameters needed to specify the partial PI model, given also the marginal parameters. In the general case where $D_i \neq D_j$ for some i and j, the maximum number

of joint parameters needed becomes

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

□

As another example, consider a partial PI model over 10 variables. Three of them are binary, four of them are ternary, and the remaining three each has 4 possible values. Suppose that the maximum marginally independent partition consists of 8 blocks with the three binary variables in one block. The number of marginal parameters is 20. The number of joint parameters is $[2^3 - 1 - (2 - 1) \cdot 3] \cdot 3^4 \cdot 4^3 = 20736$. Thus, the total number of parameters is 20756. The number of parameters required for a general PDM over the same set of variables is 41471.

By comparing the results in the previous section, it can be seen that a partial PI model is also more compact than a general PDM but is less compact than a full PI model. As full PI models, this compactness can be explored both for more accurate model learning (due to fewer number of parameters to estimate) and for reduced model complexity.

6 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 only one multi-variable block. We employ the hypercube perspective for analyzing the parameterization of PI models, which provides a visually appealing tool that facilitates the task. The hypercube representation is equivalent to the tabular or numerical representations in that it conveys the same structural information about a joint probability distribution. Nevertheless, it allows the same structure to be examined from an alternative perspective. In fact, our presentation has been switching between these alternative perspectives (visual, tabular, and numerical). The visual perspective has been crucial in helping us to perform the analysis presented in the paper.

This work is an integral part of a longer term project which explores learning of graphical models in PI and related problem domains. The hypercube perspective 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 PDMs with embedded PI submodels. 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. In practice, a given data set can potentially be abstracted into a number of distinct models depending on the levels of abstraction and the learning bias. The new algorithms will provide a new dimension for trading model complexity and model accuracy in learning.

References

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

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

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

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

[5] 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.

[6] 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.

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

[8] S.L. Lauritzen. *Graphical Models*. Oxford, 1996.

[9] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Xiang. Towards understanding of pseudo-independent domains. In *Poster Proc. 10th Inter. Symposium on Methodologies for Intelligent Systems*, Charlotte, 1997.

[14] Y. Xiang. *Probabilistic Reasoning in Multi-Agent Systems: A Graphical Models Approach*. Cambridge University Press, 2002.

[15] 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.

[16] 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.

[17] 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.