Effective Dimensions of Partially Observed Polytrees

Tao Chen
The Hong Kong University of Science and Technology

Tomáš Kočka
Prague University of Economics

Nevin L. Zhang
The Hong Kong University of Science and Technology

ABSTRACT

Model complexity is an important factor to consider when selecting among Bayesian network models. When all variables are observed, the complexity of a model can be measured by its standard dimension, i.e., the number of linearly independent network parameters. When latent variables are present, however, standard dimension is no longer appropriate and effective dimension should be used instead(Geiger et al.*, 1996). Effective dimensions of Bayesian networks are difficult to compute in general. Work has begun to develop efficient methods for calculating the effective dimensions of special networks. One such method has been developed for partially observed trees (Zhang and Kočka, 2004). In this paper, we develop a similar method for partially observed polytrees.*

Keywords: Effective dimension; Polytree models; Latent nodes; Decomposition; Regularity

Address correspondence to Tao Chen, Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. Email: csct@cs.ust.hk.

International Journal of Approximate Reasoning 1994 11:1–158

655 Avenue of the Americas, New York, NY 10010 0888-613X/94/$7.00

1. Introduction

When learning Bayesian network models from data, one needs to compare different candidate models. Usually, this is done using a scoring function. From the Bayesian perspective, a natural score is the *marginal likelihood*(Cooper and Herskovits, 1992). In some special cases, the marginal likelihood of a model can be computed using a closed-form formula(Cooper and Herskovits, 1992). In general, exact computation of marginal likelihood is intractable. In practice, an approximation of (the logarithm of) the marginal likelihood called the Bayesian Information Criterion(BIC) (Schwarz, 1978) is usually used.

The BIC score consists of two parts: one measures the fit of a model to data and the other penalizes the model according to its complexity. The complexity of a model is measured by its *standard model dimension*, i.e. the number of linearly independent model parameters.

When all variables are observed, the BIC score has been proved to be an asymptotic approximation of the marginal likelihood. Moreover it is *consistent* in the sense that, given sufficient data, the BIC score of the generative model–the model from which data were sampled–is larger than those of any other models that are not equivalent to the generative model (Haughton, 1988).

When latent variables are present, however, the BIC score is no longer an asymptotic approximation of the marginal likelihood (Geiger *et al.*, 1996). This can be, to some extent, remedied by using effective dimension to replace standard dimension. Here the *effective dimension* of a model is the rank of the Jacobian matrix of the mapping from the parameters of the model to the parameters of the marginal distribution of the observed variables. If we replace standard model dimension with effective model dimension in the BIC score, the resulting scoring function, called the *BICe score*, is an asymptotic approximation of the marginal likelihood almost everywhere except for some singular points (Geiger *et al.*, 2001, Rusakov and Geiger, 2003). There is also empirical evidence suggesting that model selection can sometimes be improved if the BICe score is used instead of the BIC score (Kočka and Zhang, 2002).

In order to use the BICe score in practice, we need to compute effective dimensions in an efficient way. This is not a trivial task. The number of rows in the Jacobian matrix increases exponentially with the number of observed variables. Hence, the construction of the Jacobian matrix and the calculation of its rank are both computationally prohibitive. Moreover they have to be done algebraically or with very high numerical precision to avoid degeneration. The necessary precision grows with the size of the matrix.

Nonetheless, fast computation of effective dimension is possible for special classes of Bayesian networks. Settimi and Smith (1998, 1999) studied trees with binary variables and latent class (LC) models with two observed variables. They have obtained a complete characterization of the effective dimensions of models in these two classes. Here, an *LC model* is the same as a Naive Bayes model except that the class variable is hidden. Zhang and Kočka (2004) have proven a theorem that decomposes, for the purpose of effective dimension calculation, a partially observed tree into a collection of LC models. The effective dimension of an LC model can be computed directly from the Jacobian matrix or be approximated using the tight upper bound provided by Kočka and Zhang (2002).

In this paper we extend the work by Zhang and Kočka (2004) to partially observed polytrees. We prove three theorems: The first theorem decomposes, for the purpose of effective dimension calculation, a partially observed polytree into what we call compact polytrees; The second theorem decomposes compact polytrees into what we call primitive polytrees; The third theorem establishes a relationship between the effective dimensions of primitive polytrees to those of some LC models. These LC models are obtained from the primitive polytrees via some simple transformation. Together, the three theorems suggest a fast method for computing the effective dimensions of partially observed polytrees.

2. Background

In this section, we quickly review the concepts and notations that will be used in subsequent sections.

We begin with some notational conventions. We will consider only random variables that have a finite number of states. Capital letters such as X and Y will denote variables and lower case letters such as x and y will denote states of variables. The domain and cardinality of a variable X will be denoted by Ω_X and $|X|$, respectively. Bold face capital letters such as $\mathbf{Y}$ denote sets of variables. $\Omega_{\mathbf{Y}}$ denotes the Cartesian product of the domains of all variables in the set $\mathbf{Y}$. Elements of $\Omega_{\mathbf{Y}}$ will be denoted by bold lower case letters such as $\mathbf{y}$ and will sometimes be referred to as states of $\mathbf{Y}$. Unless explicitly stated otherwise, the variables O, H will denote observed variable and latent variable respectively.

2.1. Graphs and Bayesian Networks

A *graph* G is a pair (N, E), where N is a set of nodes and E is a set of edges. An *Acyclic Directed Graph*(DAG) is a graph where all edges

are directed and there are no directed cycles. If two nodes X and Y are connected by a directed edge $X \rightarrow Y$, then node X is the *parent* of node Y and Y is the *child* of X. The set of all the parents of Y is denoted by $Pa(Y)$ and the set of all the children of X is denoted by $Ch(X)$. The union of a node's children and parents is called *neighbors*. We use $Ne(X)$ to denote the neighbors of node X. Then we have $Ne(X) = Pa(X) \cup Ch(X)$. The union of parents, children and parents of children of a node is called the *Markov boundary* of the node. We use $Mb(X)$ to denote the Markov boundary of node X. Then we have $Mb(X) = Pa(X) \cup Ch(X) \cup_{Z \in Ch(X)} Pa(Z)$. A node X in a DAG is *d-separated* by its Markov boundary $Mb(X)$ from all other nodes (Lauritzen 1996).

A *Bayesian network* is a pair (G, θ_G) where G is a DAG representing the *structure* of the Bayesian network and θ_G is a collection of parameters. The parameters describe the conditional probability distribution $P(X|Pa(X))$ for each variable X given its parents $Pa(X)$. A Bayesian network represents a joint probability distribution $P(\mathbf{N}|G, \theta_G)$ via the factorization formula $P(\mathbf{N}|G, \theta_G) = \prod_{X \in N} P(X|Pa(X))$. D-separation in G implies conditional independence w.r.t the joint probability P. In particular, any node A is independent of all other nodes given its Markov boundary.

A structure or *model* is *completely observed* if all its nodes are observed. Otherwise it is *partially observed*. Unobserved nodes are called *latent nodes*. A *Bayesian network model M(G)* is the set of all joint probability distributions over the observed nodes that can be represented by the Bayesian network (G, θ_G).

2.2. Tree Models

A Bayesian network model whose DAG is a rooted tree is referred to as a *tree model* or simply a *tree*. A *latent class* (LC) model is a special tree model that consists of one latent node and a number of observed nodes. A tree model is *regular* if for each latent node H the following holds $|H| \leq \frac{|Ne(H)|}{max_{Z \in Ne(H)}|Z|}$. In words, a tree model is regular if its latent nodes do not have too many states. Each irregular tree is equivalent to some regular tree, which can be obtained via a simple regularization process that reduces the cardinality of the latent nodes concerned (Zhang, 2002). Therefore one needs only to consider regular trees during model selection.

2.3. Polytree Models

In a rooted tree, each node has at most one parent. In a polytree, a node may have multiple parents and there are no cycles in the underlying undirected graph. We define *leaf node* in polytrees to be the node with no children. A non-leaf node is called an *internal node*. A *polytree model* or simply a *polytree* is a Bayesian network model whose DAG is a polytree. In

a polytree M, all the nodes that are reachable(regardless of the orientation) from a node A when node X is removed are called *reachable nodes* from A when X is removed. The submodel induced by these nodes in conjunction with links among them is called *submodel at A away from X*.

In the rest of this paper, we will be concerned with partially observed polytrees. When we speak of polytrees, we always mean partially observed polytrees. Adjacent latent nodes are allowed in such models. See Figure 1 for an example.

Just as we have a concept of regularity of trees, we also have a concept of regularity for polytrees. Intuitively, a polytree is regular if its latent nodes do not have too many states. We make this concept more precise. Suppose H is a latent node in polytree model M. Denote its parents by P_i $(i = 1, ..., I)$. For each child C_j $(j = 1, ..., J)$ of H, denote the parents of C_j by $P_{k,j}$ $(k = 1, ..., K_j)$. Let

$$A_0 = \prod_{i=1}^{I} |P_i|,\ A_j = 1 + (|C_j| - 1) \prod_{k=1}^{K_j} |P_{k,j}| \quad (j = 1, ..., J), \tag{1}$$

where $\prod_{k=1}^{K_j} |P_{k,j}|$ is defined to be 1 when K_j is zero. Note that A_j is one more than the number of parameters of the conditional probability $P(C_j|Pa(C_j))$. We say a latent node H is *regular* if $|H| \leq \frac{\prod_{j=0}^{J} A_j}{max\{A_0, A_1, ..., A_J\}}$. A polytree M is *regular* if all its latent nodes are regular. Irregular polytrees can be easily transformed into equivalent regular polytrees. See Section 5.3.3 for details. From now on, when we speak of polytrees, we always mean regular (and partially observed) polytrees.

2.4. Effective Dimensions

Consider a Bayesian network model M that possibly contains latent variables. The *standard dimension ds(M)* of M is the number of linearly independent parameters in the standard parameterization of M. Suppose M consists of k variables $X_1, X_2, \ldots, X_k$. Then $ds(M)$ is given by

$$ds(M) = \sum_{i=1}^{k} |Pa(X_i)|(|X_i| - 1)$$

where $|Pa(X_i)| = 1$ if X_i has no parent.

For notational simplicity, denote the standard dimension of M by n. Let $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ be a vector of n standard model parameters of M. Each θ_l is in the form of $P(X_r = a|Pa(X_r) = b)$ where $r = 1, \ldots, k$. Here b is any specification of $Pa(X_r)$ and a is any state of X_r except for the last one. Thus the *standard parameter space* Θ is a subset of R^n that satisfies

the following conditions:

$$0 < \theta_l < 1 \qquad \text{for } l = 1, \ldots, n; \tag{2}$$

$$\sum_{a=1}^{|X_r|-1} P(X_r = a | Pa(X_r) = b) < 1 \tag{3}$$

for any variable X_r and any possible value b.

Further let $\mathbf{O}$ be the set of observed variables. Suppose $\mathbf{O}$ has $m+1$ possible states. We enumerate the first m states as $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_m$. For any $i (1 \leq i \leq m)$, $P(\mathbf{o}_i)$ is a function of the parameters $\vec{\theta}$. So we have a mapping from the the n dimensional standard parameter space (a subspace of R^n) to R^m, namely $T_{n,m} : \Theta \subseteq R^n \vdash (P(\mathbf{o}_1), P(\mathbf{o}_2), \ldots, P(\mathbf{o}_m)) \in R^m$. The Jacobian matrix of this mapping is the following $m \times n$ matrix:

$$J_M(\vec{\theta}) = [J_{ij}] = [\frac{\partial P(\mathbf{o}_i)}{\partial \theta_j}]$$

For convenience, we will often write the matrix as $J_M = [\frac{\partial P(\mathbf{O})}{\partial \theta_j}]$, with the understanding that elements of the j-th column is obtained by allowing $\mathbf{O}$ run over all its possible states except one.

For each i, $P(\mathbf{o}_i)$ is a function of $\vec{\theta}$. For most commonly used parameterizations of Bayesian networks, it is actually a polynomial function of $\vec{\theta}$. Hence we make the following assumption:

ASSUMPTION 1. *The Bayesian network M is so parameterized that the parameters for the marginal distribution of the observed variables are polynomial functions of the parameters for M.*

An obvious consequence of the assumption is that elements of J_M are also polynomial functions of $\vec{\theta}$.

For a given value of $\vec{\theta}$, J_M is a matrix of real numbers. Due to Assumption 1, the rank of this matrix is a constant d almost everywhere in the standard parameter space (Geiger *et al.*, 1996). To be more specific, the rank is d everywhere except in the set of measure zero where it is smaller than d. The constant is called the *regular rank* of J_M as well as the *effective dimension* of the Bayesian network model M. We denote it by $de(M)$. The following proposition (Geiger *et al.*, 2001) gives a geometrical interpretation of effective dimension.

PROPOSITION 1. *Suppose M is a Bayesian network model with effective dimension d. Then the space $T_{n,m}(\Theta)$ is a union of one smooth manifold with dimension d with lower dimensional smooth manifolds.*

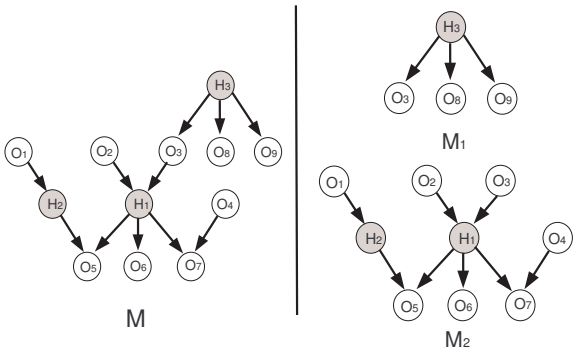

Figure 1. A polytree M and its decomposition by applying Theorem 1. Latent variables are shaded, while observed variables are not.

Points in the lower dimensional manifolds are called *singular points*. If they are removed, what remains of $T_{n,m}(\Theta)$ is a *smooth manifold*. Smooth manifolds correspond to *curved exponential families*(CEFs)(Geiger *et al.*, 2001) and the BICe score is an asymptotic approximation of the marginal likelihood for CEFs(Haughton, 1988). Therefore, BICe is an asymptotic approximation of the marginal likelihood for Bayesian network models except for some singular points.

3. Effective Dimensions of Polytrees

In this section, we present three theorems that reduce the task of computing the effective dimension of a polytree into tasks of calculating the effective dimensions of a collection of LC models. We use a running example to illustrate the theorems. The example starts from the polytree M in Figure 1. The proofs of the theorems will be given in subsequent sections. Without loss of generality, we assume that there is no latent leaf node in polytrees.[1]

3.1. Decomposition at Observed Internal Nodes

The node O_3 in M is an observed internal node. Our first theorem allows us to decompose M at O_3 into M_1 and M_2 and to obtain $de(M)$ from $de(M_1)$ and $de(M_2)$.

To present the theorem, we let M stand for a general polytree. Suppose O is an observed internal node in M. Suppose O has I children and we

[1]Since latent leaf nodes add no constraint on the set of observed variables, ignoring them does not change the effective dimension of polytrees.

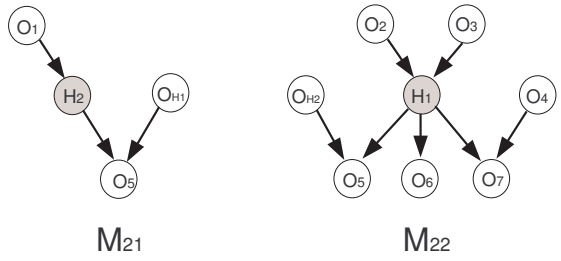

Figure 2. Decomposition of M_2 by applying Theorem 2.

denote them by $Ch(O) = \{Ch_1, Ch_2, \ldots, Ch_I\}$. For each i, let N_i be all the nodes that are reachable from Ch_i when O is removed. The nodes in $N_i \cup \{O\}$, together with the links among them, form a submodel, which we denote by M_i. If O has no parents, we say that model M *decomposes at* O into I submodels M_i ($i = 1, 2, \ldots, I$). If O has parent(s), define N_{I+1} to be all the nodes that are reachable from any parent of O when O is removed. Use M_{I+1} to denote the submodel formed by the nodes in $N_{I+1} \cup \{O\}$ together with the links among them. In this cases, we say that model M *decomposes at* O into $I+1$ submodels M_i ($i = 1, 2, \ldots, I+1$).

THEOREM 1. *Suppose that a polytree M decomposes at an observed internal node O into k submodels $M_1, M_2, \ldots, M_k$. Then*

$$de(M) = \sum_{i=1}^{k} de(M_i) - (k-1) * (|O| - 1). \tag{4}$$

In our running example, M decomposes at observed internal node O_3 into M_1 and M_2. Hence we have

$$de(M) = de(M_1) + de(M_2) - (|O_3| - 1).$$

3.2. Markov Boundary Decomposition

To continue with the running example, we notice that M_1 is an LC model. So, no further decomposition is necessary. Our second theorem allows us to further decompose M_2 into M_{21} and M_{22} and to obtain $de(M_2)$ from $de(M_{21})$ and $de(M_{22})$. See Figure 2. M_2 is an example of what we call compact polytrees. A *compact polytree* (CP) model is a polytree where each observed node has either no children or just one child and no parents. By repeatedly applying Theorem 1, one decomposes any polytree into a collection of compact polytrees.

M_{21} and M_{22} are examples of what we call *primitive polytree*. A primitive polytree (PP) is a polytree with one latent node H and a number of

observed nodes consisting of the parents of H, the children of H, and the parents of the children of H. In a polytree, each latent node and nodes in its Markov boundary form a primitive polytree.

What bridges compact polytrees and primitive polytrees is the concept of Markov boundary decomposition. Now let M be a compact polytree with observed nodes $\mathbf{O}$ and latent nodes $\mathbf{H} = \{H_1, H_2, \ldots, H_I\}$. Let M^i denote the primitive polytree formed by a latent node H_i and its Markov boundary $Mb(H_i)$. The collection of submodels $\{M^i | i = 1, 2, \ldots, I\}$ is said to be the *Markov boundary decomposition (MB-decomposition)* of M.

THEOREM 2. *Suppose that regular polytree M is a compact polytree. Let $\{M^i | i = 1, 2, \ldots, I\}$ be its MB-decomposition. Then,*

$$de(M) = ds(M) - \sum_{i=1}^{I} (ds(M^i) - de(M^i)). \tag{5}$$

Consider the compact polytree M_2 in Figure 1. We will show in Section 5.3.3 that irregular polytrees can be easily transformed into equivalent regular polytrees for the purpose of effective dimension calculation. Thus we can suppose M_2 is a regular polytree for now. It contains two latent nodes H_1 and H_2. Its MB-decomposition therefore consists of two primitive polytrees, i.e. M_{21} and M_{22} in Figure 2. Note that the observed node O_{H2} in M_{22} corresponds to H_2 in M_2 and the observed node O_{H1} in M_{21} corresponds to H_1 in M_2. By Theorem 2, we have

$$de(M_2) = ds(M_2) - [(ds(M_{21}) - de(M_{21})) + (ds(M_{22}) - de(M_{22}))].$$

3.3. Converting Primitive Polytrees into LC Models

By using Theorems 1 and 2, we have reduced the problem of computing the effective dimension of the polytree M in Figure 1 into computing $de(M_{21})$ and $de(M_{22})$. Both M_{21} and M_{22} are primitive polytrees. In this section, we present a theorem that allows us to transform them into LC models and to obtain their effective dimensions from those of LC models.

Now let M be a primitive polytree model that contains only one latent node H. Classify all the observed nodes into four categories: the parents of H, denoted by $P_i (i = 1, ..., I)$; the children of H that have only one parent(namely H), denoted by $T_r (r = 1, ..., R)$; the children of H that have more than one parent, denoted by $C_j (j = 1, ..., J)$; the observed parents of C_j for each j, denoted by $O_{k,j}$ $(k = 1, ..., K_j)$. Construct an LC model M_{LC} that has one latent variable H and observed variables Y, $T_r (r = 1, ..., R)$, and $X_j (j = 1, ..., J)$ where

$$|Y| = \prod_{i=1}^{I} |P_i|, |X_j| = 1 + (|C_j| - 1) \prod_{k=1}^{K_j} |O_{k,j}| (j = 1, ...J). \tag{6}$$

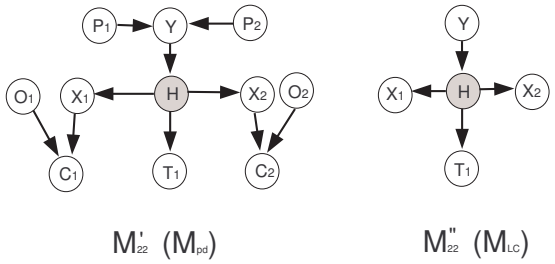

Figure 3. Transformation of primitive polytree M_{22} into LC model M''_{22}. M'_{22} is an intermediate model used in proofs.

and T_r remain the same cardinality as that in the polytree model. We call M_{LC} the *LC transformation* of primitive polytree M.

THEOREM 3. *Let M be a PP model and H be the unique latent node. Let M_{LC} be the LC transformation of M. Then we have,*

$$de(M) = de(M_{LC}) + \sum_{j,k}(|O_{j,k}| - 1) + \sum_{i}(|P_j| - 1) + 1 - \prod_{i}|P_i| \quad (7)$$

Take M_{22} as an example. Its LC transformation is M''_{22} in Figure 3. According to Equation 6, the cardinalities of observed variables in M''_{22} are:

$$\begin{aligned} |Y| &= |O_2||O_3| \\ |T_1| &= |O_6| \\ |X_1| &= 1 + (|O_5| - 1)|O_{H2}| \\ |X_2| &= 1 + (|O_7| - 1)|O_4| \end{aligned}$$

Then compute the effective dimension of M_{22} via Equation 7. We have,

$$de(M_{22}) = de(M''_{22}) + (|O_{H2}| + |O_4| - 2) + (|O_2| + |O_3| - 2) + 1 - |O_2||O_3|$$

The model M'_{22} on the left is used in proof.

4. Proof of Theorem 1

For technical convenience, we prove Theorem 1 first, following by Theorem 3 and Theorem 2. We begin with an alternative computation of effective dimension.

4.1. An Alternative Computation

Effective dimension defined in Section 2.4 can be computed in an alternative way. Note that one marginal distribution over all the observed variables can be represented using a set of conditional distributions. Suppose $(O_1, O_2, \ldots, O_k)$ is an ordered sequence of variables in $\mathbf{O}$. Mathematically we have $P(\mathbf{O}) = P(O_1) \times P(O_2|O_1) \times \ldots \times P(O_k|O_1, O_2, \ldots, O_{k-1})$. The conditional probabilities on the right side are called the *parameters of marginal factorization* of $\mathbf{O}$. Thus we define an alternative transformation of $T_{n,m}$ which is from standard parameters to the parameters of marginal factorization of $\mathbf{O}$, denoted by $\hat{T}_{n,m} : \Theta \subseteq R^n \vdash (P(O_1), P(O_2|O_1), \ldots, P(O_k|O_1, O_2, \ldots O_{k-1})) \in R^m$. The target space is m dimensional. In the space, each coordinate represents one assignment of such a conditional probability $P(O_t = o_t|O_1 = o_1, O_2 = o_2, \ldots O_{t-1} = o_{t-1})$ where $t = 1, 2, \ldots, k$ and $o_1, o_2, \ldots, o_t$ is one possible assignment of $O_1, O_2, \ldots, O_t$. Sort all the coordinates in some order and suppose that $P(O_t = o_t|O_1 = o_1, O_2 = o_2, \ldots O_{t-1} = o_{t-1})$ is the i-th coordinate. Then the Jacobian matrix of transformation $\hat{T}_{n,m}$ is :

$$\hat{J}_M(\vec{\theta}) = [J_{ij}] = [\frac{\partial P(O_t = o_t|O_1 = o_1, O_2 = o_2, \ldots O_{t-1} = o_{t-1})}{\partial \theta_j}] \quad (8)$$

It is also an $m \times n$ matrix. Moreover the relations between the parameters of marginal distribution and the parameters of marginal factorization are as follows,

$$P(\mathbf{O}) = P(O_1) \times P(O_2|O_1) \times \ldots \times P(O_k|O_1, O_2, \ldots, O_{k-1}) \quad (9)$$

$$P(O_i|O_1, O_2, \ldots, O_{i-1}) = \frac{P(O_1, O_2, \ldots, O_i)}{P(O_1, O_2, \ldots, O_{i-1})} = \frac{\sum_{\{O_{i+1}, \ldots, O_k\}} P(\mathbf{O})}{\sum_{\{O_i, \ldots, O_k\}} P(\mathbf{O})} \quad (10)$$

In conjunction with Assumption 1, Equation (10) implies that elements of $\hat{J}_M$ are rational fractions of $\vec{\theta}$. Similar with the analysis on $J(M)$ in Section 2.4, the rank of $\hat{J}_M$ is also a constant d' everywhere except in the set of measure zero where it is smaller than d'. The constant d' is called the regular rank of $\hat{J}_M$. Thus we have,

PROPOSITION 2. *Matrices J_M and $\hat{J}_M$ have the same regular rank.*

Proof: Equation (9,10) show that J_M can be obtained from $\hat{J}_M$ through elementary row operations except in a set of measure zero and vice versa. Therefore J_M and $\hat{J}_M$ have the same rank except in the set of measure zero. That is the regular rank of J_M or $\hat{J}_M$. Q.E.D

Proposition 2 states the fact that effective dimension also can be computed from $\hat{J}_M$.

4.2. Proof of Theorem 1

Theorem 1 shows how to decompose a polytree model at a single observed node. In this section we first prove a lemma which generalizes this theorem then fill in the gap between the lemma and Theorem 1.

LEMMA 1. *M be a Bayesian network model over observed variables $\mathbf{O}$ and latent variables $\mathbf{H}$. Suppose a single observed node $S \in \mathbf{O}$ and two nonempty sets of variables $\mathbf{V_1}$, $\mathbf{V_2}$ form a partition of all variables $\mathbf{O} \bigcup \mathbf{H}$. $\mathbf{V_1} \perp \mathbf{V_2} \mid S$ is true for any distribution encoded by M. The submodels induced in M by the sets $\{S\}$, $\mathbf{V_1} \bigcup \{S\}$, $\mathbf{V_2} \bigcup \{S\}$ are denoted by M_0, M_1, M_2 respectively. Then $de(M) = de(M_1) + de(M_2) - (|S| - 1)$.*

Proof: For the two nonempty sets, assume $\mathbf{V_1} = \mathbf{V_1^O} \bigcup \mathbf{V_1^H}$, $\mathbf{V_2} = \mathbf{V_2^O} \bigcup \mathbf{V_2^H}$ where $\mathbf{V_i^0}$, $\mathbf{V_i^H}$ are respectively the set of observed nodes and the set of hidden nodes in $\mathbf{V_i} (i = 1, 2)$.

The condition $\mathbf{V_1} \perp \mathbf{V_2} \mid S$ implies that all the parents of S must be in either $\mathbf{V_1}$ or $\mathbf{V_2}$ exclusively. Without loss of generality, suppose all are from $\mathbf{V_1}$. Also there is no parent-child relation between variables from $\mathbf{V_1}$ and $\mathbf{V_2}$. Thus the marginal distribution over all observed variables can be written as follows:

$$
\begin{aligned}
P(O) &= P(\mathbf{V_1^O}, \mathbf{V_2^O}, S) = \sum_{\mathbf{V_1^H}, \mathbf{V_2^H}} P(\mathbf{V_1}, \mathbf{V_2}, S) \\
&= \sum_{\mathbf{V_1^H}, \mathbf{V_2^H}} \prod_{\omega \in \mathbf{O} \cup \mathbf{H}} P(\omega \mid pa(\omega)) \\
&= \sum_{\mathbf{V_1^H} \mathbf{V_2^H}} \{ [\prod_{\omega \in \mathbf{V_1} \cup \{S\}} P(\omega \mid pa(\omega))] [\prod_{\omega \in \mathbf{V_2}} P(\omega \mid pa(\omega))] \} \\
&= \{ \sum_{\mathbf{V_1^H}} [\prod_{\omega \in \mathbf{V_1} \cup \{S\}} P(\omega \mid pa(\omega))] \} \{ \sum_{\mathbf{V_2^H}} [\prod_{\omega \in \mathbf{V_2}} P(\omega \mid pa(\omega))] \} \quad (11)
\end{aligned}
$$

Thus one can write the standard parameters $\vec{\theta}$ as $\{\vec{\theta_1}, \vec{\theta_2}\}$ where $\vec{\theta_1}$ are parameters appearing in the first summation of Equation (11) and $\vec{\theta_2}$ are parameters appearing in the second term. Moreover $\vec{\theta_i}$ can be viewed as parameters of model $M_i (i = 1, 2)$.

We compute the effective dimension in the alternative way, i.e. compute the regular rank of Jacobian matrix $\hat{J}_M$ which is defined in Section 4.1. In order to define the mapping $\hat{T}_{n,m}$, an ordered sequence of all the observed variables should be specified beforehand. A special kind of ordered sequence which is called *topological sort* is adopted here. In a topological sort of $\mathbf{O}$, for any pair of variables O_i and O_j, if O_i is an ancestor of O_j, then O_i must precede O_j in the ordering. We also require that the single observed

node S is preceded by $\mathbf{V_{1O}}$ and precedes $\mathbf{V_{2O}}$. Consequently this sequence can be represented as $\left(V_{1O}^{(1)}, V_{1O}^{(2)}, \ldots, V_{1O}^{(k_1)}, S, V_{2O}^{(1)}, V_{2O}^{(2)}, \ldots, V_{2O}^{(k_2)}\right)$ where $V_{1O}^{(i)}(i = 1, \ldots, k_1)$ are from $\mathbf{V_{1O}}$ and $V_{2O}^{(j)}(i = 1, \ldots, k_2)$ are from $\mathbf{V_{2O}}$. Under the particular ordering, the target coordinate of $\hat{T}_{n,m}$ is simply the probability of each variable given all the preceding variables. Naturally, these are divided into two parts:

1. **Part 1:**

   $$\left(P(V_{1O}^{(1)}), P(V_{1O}^{(2)}|V_{1O}^{(1)}), \ldots, P(V_{1O}^{(k_1)}|V_{1O}^{(1)}, V_{1O}^{(2)}, \ldots, V_{1O}^{(k_1-1)}), P(S|\mathbf{V_{1O}})\right)$$

2. **Part 2:**

   $$\left(P(V_{2O}^{(1)}|\mathbf{V_{1O}}, S), P(V_{2O}^{(2)}|\mathbf{V_{1O}}, S, V_{2O}^{(1)}), \ldots, P(V_{2O}^{(k_2)}|\mathbf{V_{1O}}, S, V_{2O}^{(1)}, \ldots, V_{2O}^{(k_2-1)})\right)$$

   Since $\mathbf{V_1} \perp \mathbf{V_2} \mid S$, this part can be represented as

   $$\left(P(V_{2O}^{(1)}|, S), P(V_{2O}^{(2)}|S, V_{2O}^{(1)}), \ldots, P(V_{2O}^{(k_2)}|S, V_{2O}^{(1)}, \ldots, V_{2O}^{(k_2-1)})\right)$$

It is obvious that *Part 1* can be represented by $\vec{\theta_1}$ and *Part 2* can be represented by $\vec{\theta_2}$. Thus the Jacobian matrix $\hat{J}_M(\vec{\theta})$ is in the form of:

$$\begin{pmatrix} \frac{\partial Part\ 1}{\partial \vec{\theta_1}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial Part\ 2}{\partial \vec{\theta_2}} \end{pmatrix}$$

where $\mathbf{0}$ denotes zero matrix. Hence,

$$de(M) = Rank(\frac{\partial Part\ 1}{\partial \vec{\theta_1}}) + Rank(\frac{\partial Part\ 2}{\partial \vec{\theta_2}}) \tag{12}$$

For model M_1, we set the ordered sequence of all the observed variables in M_1 to be $(V_{1O}^{(1)}, V_{1O}^{(2)}, \ldots, V_{1O}^{(k_1)}, S)$, the Jacobian matrix of M_1 is exactly $[\frac{\partial Part\ 1}{\partial \vec{\theta_1}}]$. Hence,

$$de(M_1) = Rank(\frac{\partial Part\ 1}{\partial \vec{\theta_1}}) \tag{13}$$

Similarly for model M_2, we set the ordered sequence of all the observed variables of M_2 to be $(S, V_{2O}^{(1)}, V_{2O}^{(2)}, \ldots, V_{2O}^{(k_2)})$. Inheriting the notations of M, we notice that the parameter $\vec{\theta_1}$ is $P(S)$ and *Part 1* is also $P(S)$. Hence, the Jacobian matrix of M_2 is in the form of:

$$\begin{pmatrix} \frac{\partial Part\ 1}{\partial \vec{\theta_1}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial Part\ 2}{\partial \vec{\theta_2}} \end{pmatrix} = \begin{pmatrix} I_{|S|-1} & \mathbf{0} \\ \mathbf{0} & \frac{\partial Part\ 2}{\partial \vec{\theta_2}} \end{pmatrix}$$

where $I_{|S|-1}$ is an identity matrix with dimension $|S|-1$. Hence

$$de(M_2) = (|S|-1) + Rank(\frac{\partial Part\ 2}{\partial \vec{\theta_2}}) \tag{14}$$

Consequently the theorem is proved by combining (12),(13) and (14). Q.E.D

Proof of Theorem 1: We prove this theorem by induction on the number of children of O. First Considering the case that $Pa(O)$ is not empty. When the number is one only two models M_0 and M_1 are induced and $N_0 \perp N_1|O$. Direct use of Lemma 1 will yield the proof.

When the number of children of O is greater than one, namely $I > 1$, we first decompose the model into two parts. The model M_I is induced by Ch_I in M and the model M' is the model by deleting all the node in M_I except O. Given O and except O in both models, the nodes in M_I are independent of those in M'. According to Lemma 1, we have $de(M) = de(M_I) + de(M') - (|O|-1)$. Notice that the node O has $I-1$ children in model M'. By the induction hypothesis, we have $de(M') = \sum_{i=0}^{I-1} de(M_i) - (I-1)(|O|-1)$. Together, these two equations yield that $de(M) = \sum_{i=0}^{I} de(M_i) - I * (|O|-1)$.

In the case that O does not have parents, the proof is even more simpler. Therefore Theorem 1 has been proved. Q.E.D

5. Proof of Theorem 3

This section is devoted to the proof of Theorem 3. We begin with some properties about effective dimension.

5.1. Effective Dimensions of Inclusion Models

We say that model M_1 *includes* model M_2 if for every parameterization $\vec{\theta^2}$ of M_2 there exists a parameterization $\vec{\theta^1}$ of M_1 such that M_1 and M_2 represent the same marginal probability distribution of observed variables. Two models M_1 and M_2 are said to be *equivalent* if M_1 includes M_2 and M_2 includes M_1. Note that these definitions extend the standard ones by considering the possibility of having both latent and observed variables.

LEMMA 2. *Let M_1, M_2 be two graphical models having the same set of observed variables. If M_1 includes M_2 then $de(M_1) \geq de(M_2)$.*

Proof: As showed in Section 2.4, the space $T_{n,m}(\Theta)$ contains all the possible marginal distributions of observed variables. Suppose $T^1_{n,m}(\Theta_1)$ and

$T^2_{n,m}(\Theta_2)$ is the corresponding space of M_1 and M_2 respectively. It follows from M_1 includes M_2 that $T^2_{n,m}(\Theta_2)$ is a subset of $T^1_{n,m}(\Theta_1)$.

The lemma can be proved readily by contradiction. Denote $de(M_1)$ and $de(M_2)$ by d_1, d_2 respectively. By Proposition 1, $T^1_{n,m}(\Theta_1)$ is a d_1-dimensional smooth manifold in conjunction with a set of measure zero; $T^2_{n,m}(\Theta_2)$ is a d_2-dimensional smooth manifold in conjunction with a set of measure zero. Assume $d_2 > d_1$. As a matter of fact, a higher dimensional smooth manifold cannot be a subset of a lower dimensional one. Therefore we conclude that $T^2_{n,m}(\Theta_2)$ cannot be contained in $T^1_{n,m}(\Theta_1)$. This contradicts the fact that $T^2_{n,m}(\Theta_2)$ is a subset of $T^1_{n,m}(\Theta_1)$. Lemma 2 must be true. Q.E.D

COROLLARY 1. *Equivalent models have the same effective dimension.*

5.2. Effective Dimensions of domain extended Bayesian networks

In this part a new operation on model M, called *domain extension*, is introduced for technical convenience.

Suppose M has observed variables $\mathbf{O}$ and latent variables $\mathbf{H}$. The standard parameter space Θ of M is restricted by Condition (2) and (3) which are given in Section 2.4. In this section we relax these conditions and impose a weaker condition that only requires the marginal probability of $\mathbf{O}$ to be non-negative. This operation on Θ is defined as *domain extension*. By domain extension, the resulting domain is called the *extended standard parameter space*, denoted by Θ_{ext}. The induced new model is called *domain extended Bayesian network(DEBN)* of M and denoted by M_{ext}.

LEMMA 3. *Suppose M_{ext} is a DEBN of M, then $de(M) = de(M_{ext})$.*

Proof: By domain extension, we have $\Theta \subseteq \Theta_{ext}$. Suppose A is the positive measure subset of Θ in which the Jacobian matrix J_M has rank $de(M)$. It follows that the Jacobian matrix $J_{M_{ext}}$ also has rank $de(M)$ in the set A and A is also a positive measure set of Θ_{ext}. Therefore $de(M_{ext}) = de(M)$. Q.E.D

5.3. Proof of Theorem 3

The notations are inherited from Section 3.3 except that we use M_{PP} to replace the original PP model M. First assume each C_j $(j = 1, ..., J)$ has only one observed parent O_j for simplicity. Three intermediate models between M_{PP} and M_{LC} are introduced for technical convenience.

1. See Figure 3. M_{pd} is the *partially determined* model induced from M_{PP}. Define the model as follows.

Insert a latent variable Y between all P_i and H in model M_{PP}. For each node C_j, introduce a latent variable X_j as the parent of C_j and the child of H. The cardinality of Y and X_j are stated in Equation (6) . We will denote the probability concerning M_{pd} by adding the subscript "M_{pd}".

Let the set of all states of Y to be the Cartesian product of that of P_i. Specifically speaking, the state of Y is represented by $(p_1, p_2, \ldots, p_I)$ where p_i is any possible state of P_i $(i = 1, 2, \ldots, I)$. The i-th element of Y is denoted by $Y^{(i)}$. The parameters of $P_{M_{pd}}(Y|P_1, ...P_I)$ are fixed in such a deterministic way that there is a one-to-one correspondence between the states of Y and those of all P_i:

$$P_{M_{pd}}(Y|P_1, \ldots P_I) = \begin{cases} 1 & \text{if } Y = (P_1, P_2, \ldots, P_I) \\ 0 & \text{otherwise} \end{cases}$$

Considering the parameters for each $P_{M_{pd}}(C_j|X_j, O_j)$. We denote each state of X_j(except one state) by a pair of numbers (c^*, o^*) where $c^* \in \{1, 2, \ldots, |C_j| - 1\}$ and $o^* \in \{1, 2, \ldots, |O_j|\}$. The last state of X_j is denoted by a number $c^{'} = |C_j|$. Therefore the cardinality of X_j is $|X_j| = 1 + (|C_j| - 1)|O_j|$. We set parameters $P_{M_{pd}}(C_j|X_j, O_j)$ in this way:

$$P_{M_{pd}}(C_j|X_j, O_j) = \begin{cases} 1 & \text{if } X_j = c^{'} \text{ and } C_j = c^{'} \\ 1 & \text{if } X_j = (c^*, o^*) \text{ and } C_j = c^*,\ O_j = o^* \\ 1 & \text{if } X_j = (c^*, o^*) \text{ and } C_j = c^{'},\ O_j \neq o^* \\ 0 & \text{otherwise} \end{cases}$$

As showed by the definition, the term *partially determined* indicates that the parameters in M_{pd} are determined partially .

2. The model M^*_{pd} is the same as M_{pd} except that Y and X_j are observed variables.

3. The model M_{ext_pd} is a DEBN model of M_{pd}.

We will show

Claim 1: M_{PP}, M_{pd} and M_{ext_pd} have the same effective dimension.

Claim 2: M_{pd} and M^*_{pd} have the same effective dimension.

Proof of Theorem 3: By the two claims, we conclude that $de(M^*_{pd})$ is equal to $de(M_{PP})$. Note that Y and X_j $(j = 1, \ldots, J)$ are observed nodes in model M^*_{pd}. According to Lemma 1, decomposition of M^*_{pd} at these nodes will result in an LC model M_{LC} and a collection of completely observed

models. Moreover the effective dimension of M^*_{pd} as well as M_{PP} can be computed from Equation (7). Q.E.D

To removing the assumption that each C_j has one observed parent O_j, we need only replace the parents $O_{k,j}$ $(k = 1, ..., K_j)$ of C_j using one node O_j, which is the Cartesian product of all such parents. Then we can introduce the new node X_j and set the parameters of $P_{M_{pd}}(C_j|X_j, O_j)$ in the same way.

5.3.1. *Proof of Claim 1*

LEMMA 4. *For M_{pd} and M_{PP}, we have $de(M_{pd}) \leq de(M_{PP})$.*

Proof: As indicated by Lemma 2, it suffices to show that M_{PP} includes M_{pd}. For this purpose, write the marginal probabilities of M_{PP} and M_{pd} according to the structure.

$$\begin{aligned} & P_{M_{PP}}(P_1, ...P_I, T_1, ...T_R, O_1, ...O_J, C_1, ...C_J) \\ = \; & [\prod_{i=1}^{I} P_{M_{PP}}(P_i)][\prod_{j=1}^{J} P_{M_{PP}}(O_j)] \sum_{H} \{[P_{M_{PP}}(H|P_1, \dots P_I)] \\ & [\prod_{r=1}^{R} P_{M_{PP}}(T_r|H)][\prod_{j=1}^{J} P_{M_{PP}}(C_j|H, O_j)]\} \end{aligned} \tag{15}$$

$$\begin{aligned} & P_{M_{pd}}(P_1, ...P_I, T_1, ...T_R, O_1, ...O_J, C_1, ...C_J) \\ = \; & [\prod_{i=1}^{I} P_{M_{pd}}(P_i)][\prod_{j=1}^{J} P_{M_{pd}}(O_j)] \sum_{H} \{[\sum_{Y} P_{M_{pd}}(Y|P_1, \dots P_I) P_{M_{pd}}(H|Y)] \\ & [\prod_{r=1}^{R} P_{M_{pd}}(T_r|H)][\prod_{j=1}^{J} \sum_{X_j} P_{M_{pd}}(C_j|X_j, O_j) P_{M_{pd}}(X_j|H)]\} \end{aligned} \tag{16}$$

Providing one parameterization of M_{pd}, we can set parameters of M_{PP} as follows,

$$P_{M_{PP}}(H|P_1, \dots, P_I) = \sum_{Y} P_{M_{pd}}(Y|P_1, \dots P_I) P_{M_{pd}}(H|Y)$$

$$P_{M_{PP}}(C_j|H, O_j) = \sum_{X_j} P_{M_{pd}}(C_j|X_j, O_j) P_{M_{pd}}(X_j|H) \qquad \text{for } j = 1, \dots J$$

By (15) and (16), the marginal distribution of M_{PP} is identical to that of M_{pd}. Therefore any marginal probability of observed variables of M_{pd} can be represented by M_{PP}. In other words, M_{PP} includes M_{pd}. Q.E.D

LEMMA 5. *For M_{PP} and M_{ext_pd}, we have $de(M_{PP}) \leq de(M_{ext\_pd})$.*

Proof: It suffices to verify that M_{ext_pd} includes M_{PP}. The marginal distribution of M_{ext_pd} is in the form of (16) except that the subscript is M_{ext_pd}. Suppose that we have a parameterization of M_{PP}, the problem is to find parameters of M_{ext_pd} such that

1. $\sum_{X_j} P_{M_{ext\_pd}}(X_j|H)P_{M_{ext\_pd}}(C_j|X_j, O_j) = P_{M_{PP}}(C_j|H, O_j) \quad j = 1 \; to \; J$

2. $\sum_{Y} P_{M_{ext\_pd}}(H|Y)P_{M_{ext\_pd}}(Y|P_1, \ldots, P_I) = P_{M_{PP}}(H|P_1, \ldots, P_I)$

These are two systems of equations. In the rest of the proof , we call these as "System 1" and "System 2" respectively. The remaining question is how to solve them.

Consider System 1 first. For each j, $P_{M_{ext\_pd}}(C_j|X_j, O_j)$ are fixed parameters and $P_{M_{ext\_pd}}(X_j|H)$ are variables. When $C_j = c^*$,$O_j = o^*$ and $H = h$, we have

$$\begin{aligned} & \sum_{X_j} P_{M_{ext\_pd}}(X_j|H = h)P_{M_{ext\_pd}}(C_j = c^*|X_j, O_j = o^*) \\ = \quad & P_{M_{PP}}(C_j = c^*|H = h, O_j = o^*) \end{aligned}$$

By the definition of $P_{M_{ext\_pd}}(C_j|X_j, O_j)$, the coefficient $P_{M_{ext\_pd}}(C_j = c^*|X_j, O_j = o^*)$ is one if $X_j = (c^*, o^*)$ and zero otherwise. Hence

$$P_{M_{ext\_pd}}(X_j = (c^*, o^*)|H = h) = P_{M_{PP}}(C_j = c^*|H = h, O_j = o^*) \quad (17)$$

Moreover,

$$P_{M_{ext\_pd}}(X_j = c^{'}|H = h) = 1 - \sum_{C_j \neq c^{'}, O_j} P_{M_{PP}}(C_j|H = h, O_j) \quad (18)$$

Together, Equations (17) and (18) provide the solution for System 1.

Substituting the fixed parameters of $P_{ext\_pd}(Y|P_1, P_2, \ldots, P_I)$ into System 2, we have

$$P_{M_{ext\_pd}}(H|Y) = P_{M_{PP}}(H|P_1 = Y^{(1)}, \ldots P_I = Y^{(I)}) \quad (19)$$

Equation (19) provides the solution for System 2. Q.E.D

Remark: The parameters of M_{ext_pd} might be negative by Equation (18). Nonetheless the solution still makes sense since $M_{ext_{pd}}$ is a DEBN.

Proof of Claim 1: Together, Lemmas 3, 4 and 5 give:

$$de(M_{pd}) \leq de(M_{PP}) \leq de(M_{ext\_pd}) = de(M_{pd})$$

Hence,

$$de(M_{PP}) = de(M_{ext\_pd}) = de(M_{pd})$$

Q.E.D

5.3.2. *Proof of Claim 2*

Proof of Claim 2: Denote by $\mathbf{B_j}$ any set of nodes in the model M_{pd} except the nodes X_j, O_j and C_j. Note that $\mathbf{B_j}$ can be for example the set of all such observed nodes. By the definition of parameters of M_{pd}, we have

$$P_{M_{pd}}(X_j = (c^*, o^*)|C_j = c^*, O_j = o^*) = 1 \tag{20}$$
$$P_{M_{pd}}(C_j = c^*|O_j = o^*, X_j = (c^*, o^*)) = 1 \tag{21}$$
$$P_{M_{pd}}(C_j = c^{'}|O_j \neq o^*, X_j = (c^*, o^*)) = 1 \tag{22}$$

From the structure of M_{pd}, we have

$$O_j \perp \{X_j, \mathbf{B_j}\} \tag{23}$$

We will show how to compute $P_{M_{pd}}(X_j, C_j, O_j, \mathbf{B_j})$ from $P_{M_{pd}}(C_j, O_j, \mathbf{B_j})$.

$$\begin{aligned}
& P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b) \\
= & \sum_{C_j} P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j) \\
= & P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^{'}) + P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^*) \\
& + \sum_{C_j \neq c^{'}, c^*} P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j)
\end{aligned}$$

Denoted the three terms in the last equality by "Term 1", "Term 2" and "Term 3" respectively. Term 3 is zero by Equation (21) and (22). Moreover,

$$\begin{aligned}
Term1 &= \sum_{O_j} P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^{'}, O_j) \\
&= P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^{'}, O_j = o^*) \\
&\quad + P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^{'}, O_j \neq o^*) \\
&= P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^{'}, O_j \neq o^*) \text{ (By Equation 21)}
\end{aligned}$$

$$\begin{aligned}
Term2 &= \sum_{O_j} P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^*, O_j) \\
&= P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^*, O_j = o^*) \\
&\quad + P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^*, O_j \neq o^*) \\
&= P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^*, O_j = o^*) \text{ (By Equation 22)}
\end{aligned}$$

Hence,

$$
\begin{aligned}
& P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b) \\
= \; & P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^{'}, O_j \neq o^*) \\
& + P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, C_j = c^*, O_j = o^*) \\
= \; & P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b, O_j \neq o^*) \\
& + P_{M_{pd}}(\mathbf{B_j} = b, C_j = c^*, O_j = o^*) \quad \text{(By Equations 20 and 22)} \\
= \; & P_{M_{pd}}(O_j \neq o^*) \times P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b) \\
& + P_{M_{pd}}(\mathbf{B_j} = b, C_j = c^*, O_j = o^*) \qquad \text{(By Equation 23)}
\end{aligned}
$$

When $P(O_j = o^*) \neq 0$, we have

$$P_{M_{pd}}(X_j = (c^*, o^*), \mathbf{B_j} = b) = P_{M_{pd}}(\mathbf{B_j} = b, C_j = c^* | O_j = o^*)$$

Thus $P_{M_{pd}}(X_j, \mathbf{B_j})$ can be computed from $P_{M_{pd}}(C_j, O_j, \mathbf{B_j})$ almost everywhere. So can $P_{M_{pd}}(X_j, C_j, O_j, \mathbf{B_j}) = P_{M_{pd}}(X_j, \mathbf{B_j}) \times P_{M_{pd}}(O_j) \times P_{M_{pd}}(C_j | O_j, X_j)$. This result can be generalized for M_{pd} and M^*_{pd} as follows: suppose M_{pd} has observed variables $\mathbf{O}$, then the observed variables of M^*_{pd} is $\mathbf{O} \cup \{Y, X_1, X_2, \ldots, X_J\}$, and $P_{M^*_{pd}}(\mathbf{O}, Y, X_1, X_2, \ldots, X_J)$ can be computed from $P_{M_{pd}}(\mathbf{O})$. The result suggests that the Jacobian matrix of M^*_{pd} can be obtained from that of M_{pd} via elementary row operations except in a set of measure zero and vice versa. Therefore $de(M_{pd}) = de(M^*_{pd})$. Q.E.D.

5.3.3. *Regular Polytrees* We have defined the concept of regularity of polytrees in Section 2.3. The essence of regularity becomes clear now. A latent node H_i is regular means that after the addition of the X and Y nodes around H_i (as in the proof above) the latent class model induced by the Markov boundary of H_i is regular. A polytree is regular if all its latent node are regular. If some polytree model is irregular at some latent node, we can make it regular by decreasing the cardinality of the node H_i. The procedure is called *regularity reduction*. It is straightforward that an irregular polytree can be reduced to a regular one by conducting regularity reduction multiply times.

PROPOSITION 3. *Suppose having an irregular polytree model M and denote by M_R the model obtained by the regularity reduction process described above. Then the two models M and M_R have the same effective dimension.*

Proof: We show this by showing that it holds for a single step of the reduction process. Thus, assume that only one step which decreases the cardinality of H_i was needed to reduce M. Denote by M^* and M^*_R the

models obtained from M and M_R by adding the nodes X and Y. By using the technique of proof of claim 1, we have $de(M^*) = de(M)$ and $de(M_R^*) = de(M_R)$. And now in the two models M^* and M_R^* the node H_i has different cardinality but the same Markov boundary, which forms a latent class model. Using the d-separation of H_i from all other nodes given its Markov boundary and the fact that the two latent class models are equivalent, it follows that the two models M^* and M_R^* are equivalent. By Corollary 1 we have $de(M^*) = de(M_R^*)$. Therefore $de(M) = de(M_R)$. Q.E.D

6. Proof of Theorem 2

Proof of Theorem 2: We prove this theorem by showing three things. First, we prove a lemma characterizing what a compact polytree having more than a single latent node looks like. Second, we prove a lemma describing a special parameterization of parts of regular compact polytrees and its properties. Third, we prove a lemma enabling a decomposition of any regular compact polytree into two regular compact polytrees, each having less latent nodes than the original one. This lemma builds upon the two previous ones and directly proves Theorem 2 because it ends with a set of regular primitive polytrees. Q.E.D

LEMMA 6. *Let M_{CP} be a compact polytree model having more than a single latent node. For any latent node H_1 there is a latent node H_2 in M_{CP} such that H_1 and H_2 are either neighbors or both parents of an observed node O in M_{CP}.*

Proof: M_{CP} is a polytree, thus there is a unique path between any two nodes. Choose H_2 to be such a latent node in M_{CP} that the path from H_1 to H_2 in M_{CP} does not contain any other latent node. The path can thus contain only observed nodes or no node at all (except H_1 and H_2). Every observed node in the path has at least two neighbors. This is possible in a compact polytree only if all its neighbors are its parents. Thus, there can not be more than a single observed node in the path and the lemma is proved. Q.E.D

LEMMA 7. *Let M_{CP} be a regular compact polytree model having latent nodes $\mathbf{H}$ and observed nodes $\mathbf{O}$. Suppose $H_i \in \mathbf{H}$ and $C \in Ch(H_i)$. U is all the nodes that are reachable from H_i when C is removed and the induced submodel is M_U. W is all the nodes that are reachable from*

H_i when $Pa(H_i)$ are removed and the induced submodel is M_W. Then M_U can be parameterized in such a way that $P(\mathbf{O})$ determines $P(\mathbf{O}, H_i)$ and $P(H_i)$ can be chosen a positive distribution. Moreover M_W can be parameterized in such a way that $P(\mathbf{O})$ determines $P(\mathbf{O}, H_i)$ and $P(H_i|(\mathbf{O}\backslash W))$ can be any distribution.

Proof: We present a sketch of the proof only. The proof is done by induction over the number of latent nodes in model M_{CP}. First for a single latent node. We can introduce the X and Y nodes from the proof of Theorem 3. Specifically, they are the nodes from the partially determined model, i.e. M_{pd}. Because M_{CP} is regular, the induced latent class model is regular and M_W can be parameterized to encode an injective mapping between the states of H_i and the Cartesian product of all X nodes. For M_U one can encode a similar injective mapping to all X nodes but one and the states of Y which are restricted to distributions satisfying the marginal independence among $Pa(H_i)$. We have already seen in Section 3 that the rest of the polytree can be parameterized to make the X and Y nodes de facto observed and we note that a positive distribution satisfying the marginal independence is always possible. The nodes X and Y can be marginalized out and we obtain the parameterization needed for the model M_{CP} and thus prove the first induction hypothesis. The induction step again uses a latent node H_i and the nodes X and Y around it. But the $Pa(H_i)$, $Ch(H_i)$ and $Pa(Ch(H_i))\backslash H_i$ in M_{CP} can be latent nodes now. For $Pa(H_i)$ we use the induction hypothesis of submodels away from the node H_i, for $Ch(H_i)$ we use the submodels away from their parents and for $Pa(Ch(H_i))\backslash H_i$ we use the submodel away from $Ch(H_i)$, resp. the C nodes. Note that for both $Pa(H_i)$ and $Pa(Ch(H_i))\backslash H_i$ any positive marginal distribution is sufficient, while for $Ch(H_i)$ one needs to be able to encode any distribution as needed which is possible by the induction hypothesis. This finishes the induction step and thus the whole proof. Q.E.D

LEMMA 8. *Let M_{CP} be a regular compact polytree model having nodes $\mathbf{N} = \mathbf{H} \cup \mathbf{O}$, where $\mathbf{H}$ are latent nodes and $\mathbf{O}$ are observed. Then there is a latent node $S \in \mathbf{H}$, its child $T \in Ch(S)$, observed parents of the child $O_0 \in \mathbf{O} \cap Pa(T)$ and other latent parents of the child $R \in \mathbf{H} \cap (Pa(T)\backslash\{S\})$ in M_{CP} where $\mathbf{H} \cap (\{T\} \cup R) \neq \emptyset$. The nodes S, T, R and O_0 induce in M_{CP} a submodel M_0 with all nodes observed. $\mathbf{N_S}$ is all the nodes that are reachable from S when T is removed. The nodes $\mathbf{N_S} \cup \{S\}$, T, R and O_0 induce in M_{CP} a submodel M_1 with the nodes T and R observed. The nodes $(\mathbf{N}\backslash\mathbf{N_S}) \cup \{S\}$ induce in M_{CP} a submodel M_2 with the node S observed. For the effective dimensions of these models holds $de(M_{CP}) = de(M_1) + de(M_2) - ds(M_0)$.*

Proof: We present a sketch of the proof only due to page limit. From

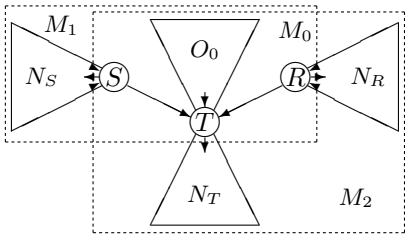

Figure 4. Compact polytree model M_{CP} and its induced sub-models

Lemma 6 follows either the existence of the latent nodes S and T or the latent nodes S and $R_i \in R$ having a common observed child T. We consider the first case only, which may contain latent nodes R, too. The same proof applies to the second case, it is just simpler because node T is observed. Moreover, for simplicity we consider only a single node R, all $R_i \in R$ can be dealt with in the same way.

The situation is depicted in Figure 4. We denote by J the Jacobian matrix of the polytree model M_{CP} and similarly use J_1 and J_2 for M_1 and M_2. Moreover, we denote by θ_O, θ_t, θ_r and θ_s the marginal parameters of O_0, T, R and S and by θ_{tt}, θ_{rr} and θ_{ss} the parameters of the sub polytrees at T, R and S except for θ_t, θ_r and θ_s.

The columns of J_2 corresponding to the parameters $\theta_{o,t,s}$ are independent because the variables are either observed or can be observed and encode any distribution if the special parameterization of θ_{tt} from Lemma 7 is used. Thus, there is a basis B_2 of J_2 which contains these and as many columns corresponding to θ_r as possible. Similarly, we denote by B_1 the basis of J_1 which contains all the columns $\theta_{o,t,r}$ and as many θ_s as possible. Obviously, B_0 contains all the columns $\theta_{o,t,r,s}$. Let $B = (B_1 \backslash B_0) \cup (B_2 \backslash B_0) \cup (B_1 \cap B_2)$.

All vectors in J depend on the vectors in B because $\theta_{ss,s}$ depend on $B_1 \backslash \theta_r$ in M_1, $\theta_{rr,r,tt,o,t}$ on $B_2 \backslash \theta_s$ in M_2 and these dependencies imply dependence in B because of the d-separations. The fact that all vectors in B are independent is proved by contradiction. If there is a dependence then it has to hold even with the special parameterization of $\theta_{ss,rr,tt}$ using Lemma 7 and this leads to a dependence in B_0 what contradicts the fact of B_0 being basis. Thus, B is a basis of J. From $B = (B_1 \backslash B_0) \cup (B_2 \backslash B_0) \cup (B_1 \cap B_2)$, $\theta_{o,t,r} \subseteq B_1$, $\theta_{o,t,s} \subseteq B_2$ and $B_0 = \theta_{o,t,r,s}$ it follows that $|B| = |B_1| + |B_2| - |B_0|$. Q.E.D

7. Conclusion

In this paper, we have proved three theorems concerning the effective dimensions of partially observed polytrees, which are a special class of

Bayesian networks. The first theorem decomposes, for the purpose of effective dimension calculation, a partially observed polytree into compact polytrees and the second theorem further decomposes the compact polytrees into primitive polytrees. The third theorem establishes a relationship between the effective dimensions of the primitive polytrees and those of some LC models obtained from them via some simple transformation. Together, the three theorems suggest a fast method for computing the effective dimensions of partially observed polytrees.

Acknowledgement

The work was partially supported by Hong Kong Research Grants Council under Grant HKUST6088/01E.

References

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

2. Geiger, D., Heckerman, D. and Meek, C. (1996). Asymptotic Model Selection for Directed Networks with Hidden Variables. In *Proc. of the 12th Conference on Uncertainty in Artificial Intelligence*, 283-290.

3. Geiger, D., Heckerman, D., King, H. and Meek, C. (2001). Stratified exponential families: Graphical models and model selection. In *Annals of Statistics*, 29, 505-529.

4. Haughton, D. (1988). On the choice of a model to fit data from an exponential family. In *Annals of Statistics*, 16, 342-555.

5. Kočka, T. and Zhang, N.L. (2002). Dimension correction for hierarchical latent class models. In *Proc. of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*.

6. Lauritzen, S. L. (1996). Graphical models. Clarendon Press, Oxford.

7. Rusakov, D. and Geiger, D (2002). Asymptotic model selection for Naive Bayesian networks. *UAI-02*.

8. Rusakov, D. and Geiger, D (2003). Automated analytic asymptotic evaluation of marginal likelihood for latent models. *UAI-03*.

9. Schwarz, G. (1978). Estimating the dimension of a model. In *Annals of Statistics*, 6, 461-464.

10. Settimi, R. and Smith, J.Q. (1998). On the geometry of Bayesian graphical models with hidden variables. In *Proc. of the 14th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, S. Francisco, CA, 472-479.

11. Settimi, R. and Smith, J.Q. (1999). Geometry, moments and Bayesian networks with hidden variables. In *Proc. of the 7th International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, Morgan Kaufmann Publishers, S. Francisco, CA.

12. Zhang, N.L. (2002). Hierarchical Latent Class models for Cluster Analysis. *AAAI-02,230-237.*

13. Zhang, N.L. and Kočka, T. (2004). Effective Dimensions of Hierarchical Latent Class Models. In *Journal of Artificial Intelligence Research*, 21, 1-17.