Effective Dimensions of Partially Observable Polytrees

Tomáš Kočka[1] and Nevin L. Zhang[2]

[1] Aalborg University, Denmark
kocka@cs.auc.dk,
[2] Hong Kong University of Science and Technology, Hong Kong, China
lzhang@cs.ust.hk

Abstract. Model complexity is an important factor to consider when selecting among graphical models. When all variables are observed, the complexity of a model can be measured by its standard dimension, i.e. the number of independent parameters. When latent variables are present, however, standard dimension might no longer be appropriate. One should instead use effective dimension [5]. Zhang & Kočka [13] showed how to compute the effective dimensions of partially observable trees. In this paper we solve this problem for partially observable polytrees.

1 Introduction

Learning graphical models from data has been widely studied in recent years. Two aproaches have been developed. One approach builds models based on statistical independence tests. The other approach seaches, in a space of models, the model that maximizes some scoring function.

From the Bayesian perspective, a natural scoring function is the marginal likelihood of model given data. In [4], Cooper and Herskovitz gave a formula for computing the Bayesian score in the case of complete data. At the same time, they showed that exact computation of the score is intractable when latent variables are present. In such a case asymptotic approximations of the marginal likelihood such as Bayesian Information Criterion (BIC) [10] and Cheeseman-Stutz Criterion (CS) [3] are usually employed.

The BIC score has two parts: one evaluates the fit of the model to the data and the other penalizes the model according to its complexity. The complexity of a model is measured using the standard dimension, i.e. number of independent parameters. The standard dimension might not be correct when latent variables are present. Consider the model $O \rightarrow H$ with two variables - observed variable O and latent variable H. All the parameters in $P(H|O)$ are irrelevant as they do not influence the fit of the model to the (observed) data. Thus there is no reason to penalize the model for such parameters.

Reexamining the derivation of the BIC score, Geiger *et al.* [5] concluded that the standard dimension should be replaced by the *effective dimension*. They also showed that the effective dimension of a model is the rank of the Jacobian matrix of the transformation between the parameters of the model and the parameters of the distribution over the observed variables.

Effective dimension is useful for several reasons. First, BIC with effective dimension was shown in [5] to be asymptotic approximation of the marginal likelihood at regular points, although it was later shown not to be at singular points [9]. Second, the BIC and CS score with standard dimension can be easily shown to be inconsistent as model seletion creteria [3]. When used with effective dimension, however, they are likely to be consistent mainly because of the close relationship between effective dimension and model inclusion (see Lemma 1). Third, effective dimension fits perfectly into the penalization scheme in the AIC score [1]. Note that AIC score has quite a different objective than the marginal likelihood. Fourth, effective dimension can be used to judge upon identifiability of a model and its parameters. This approach is used for example in mark-recovery and capture-recapture studies [2].

The straigthforward method of computing effective dimensions has an exponential complexity in the number of observed nodes. The main concern of this paper is how to compute effective dimensions efficiently. For partially observed trees, this problem was solved by [13], where a theorem was proved that allows one to decompose where the problem into the same problem for a set of latent class models. The effective dimensions of latent class models can be computed either using the tight upper bound developed in [6] or by the direct computation of the rank of the Jacobian matrix. See also [11] and [12] for interesting special cases.

In this paper we present a solution for partially observable polytrees. In Section 2 we introduce our notation, definitions, special classes of models and known results concerning effective dimensions. In Section 3 we show how to compute the effective dimension of polytrees consisting of a single latent node and its observed Markov boundary. We call such a polytree a local polytree and relate its effective dimension to some latent class model. We utilize a very special parametrization to obtain this result. Section 4 utilizes a result by Zhang & Kočka [13] to decompose polytrees using some of their observed nodes. Moreover it shows how to decompose polytrees further using their latent nodes. Thus we decompose any polytree into a set of local polytrees. We end by concluding in Section 5.

2 Basic Concepts

In this section we review basic concepts of graphs, graphical models, some special classes of graphical models and known results concerning effective dimension of models with latent variables.

2.1 Graphs and Graphical Models

A *graph* G is a pair (N, E), where N is a set of nodes and E is a set of edges, i.e. subset of $N \times N$ of ordered pairs of distinct nodes. Each node $X \in N$, denoted

[3] The marginal likelihood has not been shown to be consistent for models with latent variables either.

by an upper-case letter, represents a discrete variable. We denote the number of states of a variable X by $|X|$ and a particular state of a variable X by a lower-case letter x. We often use a set of variables $R \subseteq N$ to represent a joint variable over its elements which has number of states $|R| = \prod_{X \in R} |X|$.

An *Acyclic Directed Graph* (DAG) is a graph where all edges are directed and there are no cycles. If a graph has directed edge $A \to B$, then the node A is *parent* of the node B, i.e. $A \in Pa(B)$, and B is *child* of A, i.e. $B \in Ch(A)$. The union of children and parents of a node is called *neighbours*, i.e. $Ne(A) = Pa(A) \cup Ch(A)$. The union of parents, children and parents of children of a node is called *Markov boundary*, i.e. $Mb(A) = Pa(A) \cup Ch(A) \cup_{Z \in Ch(A)} Pa(Z)$.

A *Bayesian network* is a pair (G, θ_G) where G is a DAG and θ_G are parameters. The parameters are conditional probabilities $P(X|Pa(X))$ for each node $X \in N$ given its parents $Pa(X)$. The standard number of (free, independent) parameters ds in a Bayesian network is $ds(G) = \sum_{X \in N} (|X|-1) * \prod_{Y \in Pa(X)} |Y|$.

A Bayesian network represents a joint probability distribution $P(N|G, \theta_G)$ using the *factorization* formula $P(N|G, \theta_G) = \prod_{X \in N} P(X|pa(X))$.

A node A in a DAG is d-separated by its Markov boundary $Mb(A)$ from all other nodes (see [7] for the definition of d-separation). Thus each distribution factorizing with respect to the DAG satisfies the conditional independence $N \backslash A \backslash Mb(A) \perp\!\!\!\perp A | Mb(A)$, i.e. $P(N) * P(Mb(A)) = P(N \backslash A) * P(A \cup Mb(A))$.

A *tree* is a connected undirected graph without cycles. A *directed tree* is a DAG obtained from a tree by choosing a root node and directing all edges away from this node. A tree has a unique path between any two vertices. Note that for trees Markov boundary of a node reduces to neighbours of the node.

A *polytree* is a DAG where its underlying graph (undirected graph obtained by droping the direction of arcs) is a tree. Thus it has a unique path between any two vertices as well as tree.

We say that model M_1 *includes* model M_2 if for every parametrization θ_2 of M_2 there exists a parametrization θ_1 of M_1 such that both the models M_1 and M_2 represent the same joint probability distribution over observed variables. Two model 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.

2.2 Some Special Models

A *Latent Class* (LC) model is a graphical model with one latent variable H and observed variables O. It can be represented as a directed or undirected tree where the observed nodes have exactly one neighbour. Each latent node H_i in a tree model induces, together with its neighbours $Ne(H_i)$, an LC model. It has latent variable H_i and observed variables $Ne(H_i)$.

A *Local Polytree* (LP) model is a graphical model with observed nodes O and one latent node H. The LP model is a polytree where all the observed nodes O are parents of H or children of H or parents of children of H. (In other words the nodes O form Markov boundary of the node H.)

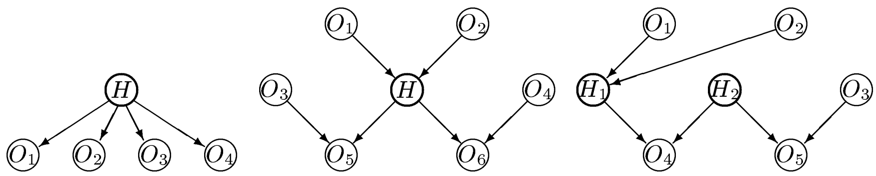

Fig. 1. Example of a) Latent Class, b) Local Polytree and c) Compact Polytree models

A *Compact Polytree* (CP) model is a graphical model with observed nodes O and latent nodes H. The CP model is a polytree where each observed node $O_i \in O$ has either no children or just one child and no parents. Each latent node H_i in a CP model induces, together with its Markov boundary $Mb(H_i)$, an LP model. It has latent variable H_i and observed variables $Mb(H_i)$.

Thus we have a hierarchy (LC,) LP, CP, polytree with latent variables. This hierarchy of classes of models is the key idea behind this paper. For examples of LC, LP and CP models see Figure 1.

We say that a tree with observed variables O and latent variables H is *regular* if for each latent node $H_i \in H$ holds $|H_i| \leq \frac{|Ne(H_i)|}{max_{Z \in Ne(H_i)}|Z|}$. Each irregular tree is equivalent to some regular tree, which can be obtained by a simple regularization process of reducing cardinality of concerned latent nodes [6]. Thus by computing the effective dimension for all regular trees one solves the problem for all trees.

2.3 Effective Dimension

A graphical model G transforms its parameters θ_G into a probability distribution $P(O)$ over all observed variables O (marginal of $P(N)$). We denote by $J_O(\theta_G) = [J_{jk}] = [\frac{\partial p(o_j)}{\partial \theta_k}]$ the Jacobian matrix of this transformation. Rows of $J_O(\theta_G)$ correspond to states in the observed space O of the model G and columns to the parameters θ_G. Geiger *et al.* [5] showed that the *effective dimension* $de(G)$ of a model G is the rank of $J_O(\theta_G)$. The rank in general is a function of θ_G but was shown to be constant almost everywhere. We use $J_O(\theta)\,; \theta \subseteq \theta_G$ to denote a matrix which has only a subset of columns in $J_O(\theta_G)$.

This suggests the following numerical approach to computing de: generate random θ, compute the Jacobian and its rank with sufficient numerical precision. We used this algorithm implemented in Matlab by Rusakov [8] to study the effective dimensions of some polytrees empirically.

Effective dimension of a model M is thus defined as the dimension of the subspace spanned by the joint probability $P(O)$ of observed variables. This together with the definition of inclusion between models, which is defined as the subspace spanned by $P(O)$ of one model being subset of the subspace spanned by $P(O)$ of the other model, leads to the following lemma.

Lemma 1. *Let M_1 and M_2 be two graphical models having the same set of observed variables. If M_1 includes M_2 then $de(M_1) \geq de(M_2)$.*

The *rank* of a matrix is a number of (row or column) vectors in a basis of the matrix. A *basis* is a set of linearly independent vectors such that all other vectors can be expressed as a linear combination of the vectors in the basis. Because the rank is constant almost everywhere we use the terms basis and linear independence in this paper with the meaning almost everywhere too. Note that for any set of independent vectors there is always a basis which includes this set.

Settimi & Smith [11] solved the effective dimension of latent class models with two observed nodes. Kočka & Zhang [6] derived a tight upper bound on the effective dimension of any latent class model.

The following theorem takes advantage of the above solution(s) for LC models and solves the problem of effective dimension of a tree with latent variables.

Theorem 1. *[13] Let M be a regular tree model with observed variables O and latent variables H. Let M_i be the local LC model induced by each latent node $H_i \in H$ in M. Then the difference between the number of standard and effective parameters in M is equal to the sum of the differences over all the M_i models, i.e. $ds(M) - de(M) = \sum_{H_i \in H} ds(M_i) - de(M_i)$.*

Moreover the following stright-forward theorem proves to be essential in the next two sections.

Theorem 2. *[13] Let M be a graphical model over observed variables O and latent variables H. Let $S \in O$ be a subset of the observed nodes such that there exist two nonempty sets of variables V_1 and V_2 where $V_1 \cap S = \emptyset$, $V_2 \cap S = \emptyset$, $V_1 \cup V_2 \cup S = O \cup H$ and $V_1 \perp\!\!\!\perp V_2 | S$ is true for any distribution encoded by the model M. Let M_0, M_1 and M_2 be the submodels induced in M by the sets S, $V_1 \cup S$ and $V_2 \cup S$. Then $de(M) = de(M_1) + de(M_2) - ds(M_0)$.*

3 Effective Dimension of Local Polytrees

In this section we prove the following theorem describing how to compute the effective dimension of any local polytree model M.

Theorem 3. *Let M be a local polytree with one latent variable H, its observed parents $P_i \in Pa(H) = P$, its observed childrens $T_m \in Ch(H)$ where $Pa(T_m) = \{H\}$, its observed childrens $C_j \in Ch(H)$ where $Pa(C_j) \neq \{H\}$ and their observed parents $O_{k,j} \in Pa(C_j) \backslash H = O_j$. Let M_{LC} be a latent class model with one latent variable H and observed variables $\{Y\} \cup T \cup X$ where Y is one joint variable over the cartezian product of all P_i, $T = \{T_m\}$ and $X = \{X_j\}$ where $|X_j| = 1 + (|C_j| - 1) * \prod_k |O_{k,j}|$. Then $de(M) = de(M_{LC}) + 1 + \sum_{j,k}(|O_{k,j}| - 1) + \sum_i (|P_i| - 1) - \prod_i |P_i|$.*

Proof. We prove the theorem in two steps. First we introduce a graphical model M_X with a special parametrization and show that $de(M) = de(M_X)$. Second we use the latent class model M_{LC} and show that $de(M_X) = de(M_{LC}) + 1 +$

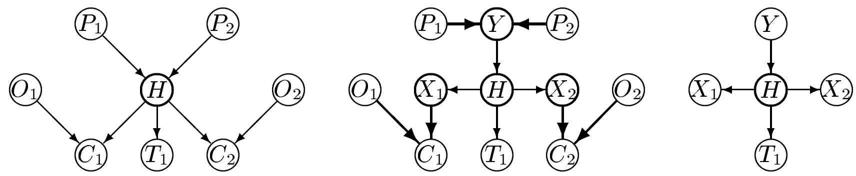

Fig. 2. a) Local Polytree model M, b) model M_X with the deterministic nodes X and Y introduced and c) Latent Class model M_{LC}

$\sum_{j,k}(|O_{k,j}|-1)+\sum_i(|P_i|-1)-\prod_i |P_i|$. We sometimes denote by O_j the cartezian product over all $O_{k,j}$.

The model M_X is obtained from the model M by introducing new latent variable Y and new latent variables X_j for each node C_j. The parameters $P(Y|P)$ are fixed in such a deterministic way that there is a one to one correspondence between the values of Y and the values of the cartezian product of all P_i. The parameters of each $P(C_j|X_j,O_j)$ are fixed in a deterministic way too. Thus these "parameters" are in fact not parameters of the model M_X. We denote each state of X_j (except one state) by a pair of numbers (c^*,o^*) where $c^* \in< 1,|C_j|-1 >$ and $o^* \in< 1,|O_j| >$. The last state of X_j is denoted by a number $c' = |C_j|$. The four states of the node X_j in the example in Table 1 are $\{(1,1),(1,2),(1,3),2\}$ in this notation. We set the $P(C_j|X_j,O_j)$ in this way: $p(C_j = c|X_j = c', O_j) = 1$ if $c = c'$; and $p(C_j = c|X_j = (c^*,o^*), O_j = o) = 1$ if $((c = c^*$ and $o = o^*)$ or $(c = c'$ and $o \neq o^*))$. All other probabilities in $P(C_j|X_j,O_j)$ are zero. An examle of such a distribution $P(C_j|X_j,O_j)$ is in Table 1. We assume that any conditional distribution $P(A|B)$ is parametrized in the standard way that the parameters are for all states of the condition B and for all but the last state of the variable A.

Table 1. Example of the deterministic distribution $P(C_j|O_j,X_j)$ in the model M_X for $|O_j| = 3$, $|C_j| = 2$ and thus $|X_j| = 4$.

$X_j =$	(1,1)			(1,2)			(1,3)			(2)		
$O_j =$	1	2	3	1	2	3	1	2	3	1	2	3
$C_j = 1$	1	0	0	0	1	0	0	0	1	0	0	0
$C_j = 2$	0	1	1	1	0	1	1	1	0	1	1	1

There is a one to one correspondence between the parameters describing $P(C_j|H,O_j)$ in M and $P(X_j|H)$ in M_X. Note that the same trivially holds for $P(H|P)$ and $P(H|Y)$. Thus it could be possible for any distribution encoded by M to be encode by M_X. It is not that simple as the parameters $P(X_j|H)$ in M_X can't give a sum higher than one if M_X is a graphical model. However we can use another model M_X^* without this condition on the parameters $P(X_j|H)$

in M_X but with the same parameters and equations used to obtain the distribution $P(O)$ over the observed variables as in M_X, i.e. M_X^* is not a graphical model but we allow only such parametrizations of it that it encodes a probability distribution over $P(O)$. Then the model M_X^* can encode any distribution $P(O)$ which can the model M and thus the model M_X^* includes the model M.

Because by marginalizing the variables X and Y out of the graphical model M_X one obtains the graphical model M, it follows that the model M includes the model M_X. The two models M_X and M_X^* have the same parameters, functional form, Jacobian matrix and thus the same effective dimension. From the facts that M_X^* includes M, M includes M_X and $de(M_X) = de(M_X^*)$ together with the Lemma 1 follows that $de(M) = de(M_X)$. Thus the first claim is proved.

Note that the nodes X and Y can be introduced in any polytree for any latent variable and all the arguments above apply to such a case too.

Back to the model M_X. The fixed deterministic distribution $P(C_j|X_j, O_j)$ has the property that as long as the marginal probability $P(O)$ is positive then for every state (c^*, o^*) of X_j there exists a state c^* of C_j and a state o^* of O_j such that $p(X_j = (c^*, o^*)|C_j = c^*, O_j = o^*) = 1$, i.e. $p(X_j = (c^*, o^*), C_j = c^*, O_j = o^*) = p(C_j = c^*, O_j = o^*)$. We have defined the distribution $P(C_j|X_j, O_j)$ including the condition $p(C_j = c^*|X_j = (c^*, o^*), O_j = o^*) = 1$, i.e. $p(C_j = c^*, X_j = (c^*, o^*), O_j = o^*) = p(X_j = (c^*, o^*), O_j = o^*)$. Moreover note that X_j and O_j are marginally independent and thus $p(X_j = (c^*, o^*), O_j = o^*) = p(X_j = (c^*, o^*)) * p(O_j = o^*)$. All these equations together imply that $p(X_j = (c^*, o^*)) = p(C_j = c^*, O_j = o^*)/p(O_j = o^*)$. We denote by B_j any set of nodes in the model M_X except the nodes X_j, C_j and O_j. Note that B_j can be for example the set of all such observed nodes. Then from knowing the distribution $P(C_j, O_j, B_j)$ one can easily compute the distribution $P(X_j, B_j)$ and $P(X_j, C_j, O_j, B_j) = P(X_j, B_j) * P(O_j) * P(C_j|X_j, O_j)$ as well. Thus the special fixed distribution $P(C_j|X_j, O_j)$ of M_X defined above causes the nodes X being in fact observed too.

Because the nodes X are observed we can apply Theorem 2 and we obtain $de(M_X) = de(M_{LC}^*) + \sum_{j,k}(|O_{j,k}| - 1)$ where M_{LC}^* is the submodel induced from the model M_X by the latent nodes H and Y and observed nodes X and T. Because of the special parametrization of $P(Y|P)$ this model is equivalent to the latent class model M_{LC} with the special requirement that the marginal distribution $P(Y)$ has to correspond to the mutually marginally independent nodes. The question is if all these $|Y| - 1 - \sum_i(|P_i| - 1)$ independent equality constraints all decrease the effective dimension of the modle M_{LC}^* compared to the model M_{LC}. Because all the parameters characterizing the observed marginal $P(Y)$ in M_{LC} are independent one can choose a basis of the Jacobian matrix of the model M_{LC} to obtain all the corresponding vectors. Thus each of the independent constrains reduces the basis by one vector. Thus we have shown that $de(M_X) = de(M_{LC}) + \sum_{j,k}(|O_{j,k}| - 1) - (\prod_i |P_i| - 1) + \sum_i(|P_i| - 1)$.

We say that a polytree M is *reduced* if for every latent node H_i in M after adding 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. If some polytree model

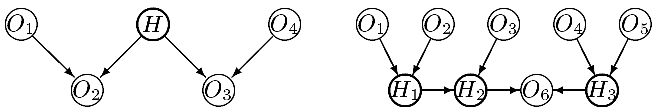

Fig. 3. Examples a) W structure in [5] with $|O_i| = 2$, b) Compact polytree model with $|O_1| = |O_2| = |O_3| = |O_4| = |O_5| = |H_3| = 2$ and $|O_6| = |H_1| = |H_2| = 3$

is not reduced we can reduce it by decreasing the cardinality of the appropriate node H_i to satisfy the regularity constraint.

Suppose having a non-reduced polytree model M and denote by M_R the model obtained by the reduction process described above. Then the two models M and M_R have the same effective dimension. We show this by showing that it holds for a single step of the reduction process decreasing the cardinality of H_i. Thus assume that only one step 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. The first part of the proof of Theorem 3 applies to any polytree because the node H_i is d-separated from all other nodes by its Markov boundary and it is exactly this boundary to which the proof applies. Thus $de(M^*) = de(M)$ 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. Again 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 too. Thus using lemma 1 they have the same effective dimension too.

We can demonstrate the use of the Theorem 3 on the W structure reported in [5]. The W structure consists from one latent node H, two binary observed childs each of them having one extra binary observed parent (see Figure 3). It was reported in [5] that this structure has $de = 9$ for $|H| = 2$, $de = 10$ for $|H| = 3$, $de = 10$ for $|H| = 4$ and $de = 11$ for $|H| = 5$, however it was later on repaired to $de = 10$ for $|H| = 5$. However there was no explanation of these results available up to now. We can apply the Theorem 3 that converts the problem into a LC problem with one latent variable H and two observed variables with three states. For these we can use the exact solution from [11]. One obtains $de = 9$ for $|H| = 2$ and $de = 10$ for $|H| \geq 3$ as all the LC models with $|H| \geq 3$ are equivalent (and irregular).

4 Decomposition of Polytrees

It is easy to realize that by applying Theorem 2 to any polytree one decomposes it into a set of compact polytrees. The sets S used for this decomposition correspond to single observed nodes having either both some parents and a child or more than one child.

In this section we show how to compute the effective dimension of any reduced compact polytree by decomposing such a polytree into a set of reduced local polytrees. It was explained in previous section that we can limit ourselves to reduced polytrees because any non-reduced polytree can be easily converted into a reduced one with the same effective dimension. The following theorem states the main result of this section, which we prove in the rest of this section.

Theorem 4. *Let M be a reduced compact polytree with observed nodes O and latent nodes H. Let M_i denote the reduced local polytree model induced in M by any latent node $H_i \in H$ and its Markov boundary $Mb(H_i)$ in M. Then $de(M) = ds(M) - \sum_{H_i \in H}(ds(M_i) - de(M_i))$.*

Proof. We prove this theorem by showing three things. First we prove a lemma characterizing how does a compact polytree having more than a single latent node look like. Second we prove a lemma describing a special parametrization of parts of reduced compact polytrees and its properties. Third we prove a lemma enabling a decomposition of any reduced compact polytree into two reduced compact polytrees each having less latent nodes than the original one. This lemma builds upon the two previous ones and it directly proves the theorem above because it ends with a set of reduced local polytrees.

Lemma 2. *Let M 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 such that two two nodes are either neighbours or both parents of an observed node O in M.*

Proof. M is a polytree thus there is a unique path between any two nodes. Choose H_2 to be such a latent node in M that the path from H_1 to H_2 in M doesn't 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 neighbours. That is possible in a compact polytree only if all its neighbours are its parents. Thus there can not be more than a single observed node in the path and the lemma is proved.

We define *subpolytree* at a node A away from nodes B in a polytree M with nodes N as the subgraph of M induced by all nodes $C \in N$ such that the path from A to C doesn't contain any node from the set B.

Lemma 3. *Let M be a reduced compact polytree model having nodes $N = H \cup O$, where H are latent nodes and O are observed. Let M_U be a subpolytree of M at a latent node $H_i \in H$ away from a node $C \in Ch(H_i)$ consisting from nodes U. Let M_P be a subpolytree of M at a latent node $H_i \in H$ away from the nodes $Pa(H_i)$ consisting of nodes P. Then the submodel M_U can be parametrized in such a way that $P(O)$ determines $P(O, H_i)$ and $P(H_i)$ can be chosen a positive distribution. Moreover the submodel M_P can be parametrized in such a way that $P(O)$ determines $P(O, H_i)$ and $P(H_i|(O \setminus P))$ can be any distribution.*

Proof. We present a sketch of the proof only. It is done by induction over the number of latent nodes in the model M. First for a single latent node. We can introduce the X and Y nodes from the proof of Theorem 3. Because M is reduced the induced latent class model is regular and M_P can be parametrized

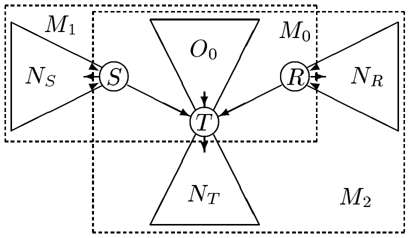

Fig. 4. Compact polytree model M and its induced submodels

to encode a bijection between the states of H_i and the cartezian product of all X nodes while for M_U one can encode a similar bijection to all but one X nodes and the states of Y which are restricted to distributions satisfying the marginal independence among $Pa(H_i)$. We have already seen that the rest of the polytree can be parametrized to make the X and Y nodes in fact 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 parametrization needed for the model M and thus prove the first induction hypothesis. Then induction step uses again 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 can be latent nodes now. For $Pa(H_i)$ we use the induction hypothesis of subpolytrees away from the node H_i, for $Ch(H_i)$ we use the subpolytrees away from their parents and for $Pa(Ch(H_i))\backslash H_i$ we use the subpolytree 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 and this is possible by the induction hypothesis. That finishes the induction step and thus the whole proof.

Lemma 4. *Let M be a reduced compact polytree model having nodes $N = H \cup O$, where H are latent nodes and O are observed. Then there is a latent node $S \in H$, its child $T \in Ch(S)$, observed parents of the child $O_0 \in O \cap Pa(T)$ and other latent parents of the child $R \in H \cap (Pa(T)\backslash\{S\})$ in M where $H \cap (\{T\} \cup R) \neq \emptyset$. The nodes S, T, R and O_0 induce in M a submodel M_0 with all nodes observed. The subpolytree of M at the node S away from T consists from nodes N_S. The nodes $N_S \cup S$, T, R and O_0 induce in M a submodel M_1 with the nodes T and R observed. The nodes $(N\backslash N_S) \cup \{S\}$ induce in M a submodel M_2 with the node S observed. For the effective dimensions of these models holds $de(M) = de(M_1) + de(M_2) - ds(M_0)$.*

Proof. We present a sketch of the proof only due to page limit. From Lemma 2 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 can contain latent nodes R too. The same proof applies to the second case, it is just simpler because the node T is observed. Moreover for simplicity we consider only a single node R, all $R_i \in R$ can be dealt in the same way.

The situation is depicted in Figure 4. We denote by J the Jacobian matrix of the polytree model M and similarly 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 subpolytrees at T, R and S except of the θ_t, θ_r and θ_s.

The columns of J_1 corresponding to the parameters $\theta_{o,t,r}$ are independent because the variables are either observed or can be observed and encode any distribution if a special parametrization from Lemma 3 is used. Thus there is a basis B_1 of J_1 which contains these and as many as possible columns corresponding to θ_s. Similarly we denote by B_2 the basis which contains all the columns $\theta_{o,t,s}$ and as many as possible θ_r. 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 parametrization of $\theta_{ss,rr,tt}$ using Lemma 3 and that leads to a dependence in B_0 what contradics the fact of B_0 being basis. Thus B is the 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 $B_0 = \theta_{o,t,r,s}$ follows $|B| = |B_1| + |B_2| - |B_0|$.

We can demonstrate the use of Theorem 4 on the reduced compact polytree in Figure 3. This model has $ds = 41$. It has three latent nodes and they induce three reduced local polytree models with $ds = 23$, $ds = 28$ and $ds = 20$. These local polytree models have $de = 17$, $de = 22$ and $de = 18$. Thus the compact polytree model in Figure 3 has $de = 27$.

The effective dimension of the compact polytree model in Figure 3 can be computed directly using its numerical Jacobian matrix. This approach gives the same answer and takes 34 seconds on nowdays desktop computer. The effective dimensions of the Latent Class models needed for computing the local polytrees can be computed exactly because they have only two observed nodes. In general one would have to use the approximation proposed in [6] or again the numerical Jacobian matrix approach which takes for these three latent class models 0.46, 0.74 and 0.44 seconds. The direct approach is exponential in the number of observed nodes and thus infeasible in general.

5 Conclusion

In this paper we present two important results concerned with the computation of effective dimension of graphical models with latent variables. The first result enables us to compute the effective dimension of any local polytree, i.e. polytree having just one latent node. It converts the problem into the same problem for some latent class model. The effective dimension of latent class models was studied more and exact solutions are known for special cases and a tight upper bound for effective dimension of latent class models can be used in practice too.

The second result enables us to compute the effective dimension of any reduced polytree with latent nodes by decomposing it into the same problem for a set of local polytrees.

Our results can be summarized as a method how to compute the effective dimension for any polytree with latent nodes. Many of the results are achieved only thanks to the underlying tree property and thus are not directly applicable to the general problem for any graphical model. It is an open research question if a similar method could be used for graphs with cycles too, at least as an approximation of the effective dimension. We hope that studying effective dimensions can shed some light on the size of the class og graphical models with latent variables and inclusion among them.

References

1. Akaike H. (1974). A new look at the statistical model identification. In *IEEE Transactions on Automatic Control*, AC-19, 716-723.
2. Catchpole E.A. and Morgan B.J.T. (2001). Deficiency of parameter-redundant models. In *Biometrika*, 88, 2, 593-598.
3. Cheeseman P., Stutz J. (1995). Bayesian classification (AutoClass): Theory and results. In Fayyad U., Paitesky-Shapiro G., Smyth P., Uthurusamy R. (Eds.), *Advances in knowledge discovery and data mining*, pp. 153-180, Menlo Park, CA: AAAI Press.
4. Cooper G., Herskovits E. (1992). A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, 9, pp. 309-347.
5. Geiger D., Heckerman D. and Meek C. (1996). Asymptotic Model Selection for Directed Networks with Hidden Variables. In *Proceedings of the Twlefth Conferece on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, pp. 283-290.
6. 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)*.
7. Lauritzen S. L. (1996). Graphical models. Clarendon Press, Oxford.
8. Rusakov D. (2002). Effective dimension calculations for Bayesian networks, code in Matlab, *http : //www.cs.technion.ac.il/ ~ rusakov/archive/ bn_dimension/bn_dimension.m* .
9. Rusakov D. and Geiger D. (2002). Asymptotic Model Selection for Naive Bayesian Networks. In *Proc. of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*.
10. Schwarz G. (1978). Estimating the dimension of a model. In *Annals of Statistics*, 6, pp. 461-464.
11. Settimi, R. and Smith, J.Q. (1998). On the geometry of Bayesian graphical models with hidden variables. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, S. Francisco, CA, pp. 472-479.
12. Settimi, R. and Smith, J.Q. (1999). Geometry, moments and Bayesian networks with hidden variables. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, Florida (3-6 January 1999), Morgan Kaufmann Publishers, S. Francisco, CA.

13. Zhang, N. L. and Kočka, T. (2002). Effective Dimensions of Hierarchical Latent Class Models, submitted.