

---

# Dominance of Bayesian Networks and Efficient Learning of Generalized Latent Class Models

---

Yixin Chen\*

Department of Computer Science  
Washington University  
St. Louis, MO 63130  
chen@cse.wustl.edu

Dong Hua

Department of Computer Science  
George Washington University  
Washington, DC 20052  
gwuhua@gwu.edu

Fang Liu

Department of Computer Science  
University of Texas - Pan American  
Edinburg, Texas 78539  
fliu@cs.panam.edu

## Abstract

A major challenge for learning Bayesian networks is the complexity in searching the huge space of models and parameters. The computational cost is higher when the model topology is more flexible. In this paper, we propose the notion of dominance which can lead to strong pruning of the search space and significant reduction of learning complexity, and apply this notion to the Generalized Latent Class (GLC) models, a class of Bayesian networks for clustering categorical data.

GLC models can address the local dependence problem in latent class analysis by assuming a very general graph structure. However, The flexible topology of GLC leads to large increase of the learning complexity. We first propose the concept of dominance and related theoretical results which is general for all Bayesian networks. Based on dominance, we propose an efficient learning algorithm for GLC. A core technique to prune dominated models is regularization, which can eliminate dominated models, leading to significant pruning of the search space. Significantly improvements on the modeling quality and time complexity on real datasets are reported.

## 1 Introduction

Bayesian networks are graphical models for clustering and classification of data. An important class of Bayesian networks, latent class (LC) models,

is widely used to cluster categorical data (Lazarsfeld&Henry 1968; Goodman 1974b; Bartholomew and Knott 1999; Uebersax 2004).

As shown in Figure 1a, an LC model involves one latent class variable which is unobserved and denoted by the shaded node, and multiple manifest variables which are observable. The LC analysis determines both the number of states of the latent class variable and conditional probability distributions (CPDs) between the latent class variable and manifest variables.

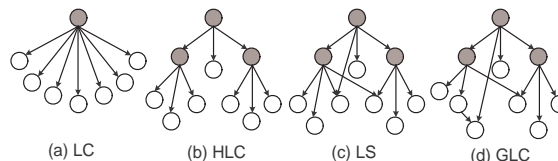


Figure 1: Illustration of LC, HLC, LS and GLC models. Shaded nodes are unobserved (latent) and solid nodes are observed (manifest).

LC models assume *local independence* which requires manifest variables be mutually independent given the latent class variable (shown in Figure 1a). But this assumption is often violated in reality, resulting in poor performance. It is well known as the *local dependence problem* (Zhang 2002). One attempt to address the local dependence problem is the *hierarchical latent class (HLC)* model (Zhang 2002, Zhang 2004). HLC requires a tree topology shown in Figure 1b. Hua *et al.* (2003) generalized HLC to a multi-layer latent structure (LS) model, which allows a manifest variable to have more than one latent parent variable. However, LS models still assume that manifest variables are mutually independent (Figure 1c), which is unrealistic for many applications such as disease class discovery (Hua *et al.* 2003).

We have experimentally observed that the flexibility

---

\*This research is supported by Microsoft Research New Faculty Fellowship and a Department of Energy ECPI grant.

of topology is a key to improve the model quality. A new model we propose to address the local dependency problem is the ***generalized latent class (GLC)*** model that allows a general graph structure for manifest variables. An example is shown in Figure 1d. With this generalization, GLC models address the local dependence problem and achieve much better model quality. For example, we applied GLC to the house building data (Hagenaars 1988) and achieved very good quality  $p=0.757$ , where  $p$  is the significance level (Read & Cressie 1988) to measure the model quality in terms of goodness-of-fit. In contrast, LC, HLC and LS models fail on this data as they all result in models with  $p = 0.000$ .

However, a ***key challenge*** we found for using GLC is that the very flexible topology of GLC models leads to a much larger search space and it becomes computationally very expensive to learn the optimal GLC models using previous learning methods. The ***main contribution*** of this paper is that we propose an efficient learning algorithm based on the novel concept of ***dominance*** which is general for all Bayesian networks.

A key theoretical property on dominance is that we can prune the dominated models during search without deteriorating the model quality. Based on this property, we propose efficient regularization operations for GLC models which lead to ***significant pruning of the search space*** by ruling out dominated models. Based on regularization, we develop an efficient learning algorithms that can converge much faster and lead to much better model quality on several real-world datasets.

## 2 Generalized Latent Class Models

In this section, we propose the generalized latent class (GLC) models that solves the local dependence problem, and discuss the quality metrics for model selection.

### 2.1 Basics of Bayesian Networks

A *Bayesian network* is a probabilistic graphical model defined by a pair  $M = (G, \theta_G)$ , where  $G = (V, E)$  is an *acyclic directed graph* (DAG).  $X \in V$  represents a random variable in the problem domain and  $V$  a set of variables. All variables are discrete.  $\theta_G$  are parameters, i.e., conditional probability distribution (CPD) for each node  $X \in V$  given its parents  $Pa(X)$ . The conditional probability  $\theta_{ijs} = P_M(X_i = j | Pa(X_i) = \mathbf{s})$ , usually written as  $P_M(X | Pa(X))$  for simplicity, represents the

probability that variable  $X_i$  assumes value  $j$  when  $Pa(X_i)$  assumes state  $\mathbf{s}$ . Hence,  $\theta_G$  is the set of all  $\theta_{ijs}$ .

Learning a Bayesian network entails determining the structure  $G$  and parameters  $\theta_G$ . Variables in  $G$  may be ***manifest variables*** that are observable and ***latent variables*** that are hidden and unobserved. Typical metrics to measure the *model quality* are ***completely*** or ***partially*** related to the joint probability distribution  $P(V|G, \theta_G)$  over all variables  $V$  using the factorization formula:

$$P_M(V|G, \theta_G) = \prod_{X \in V} P_M(X | Pa(X)).$$

### 2.2 General structure of GLC models

We propose the GLC model which accommodates local dependence by a general graph topology. A ***Generalized Latent Class (GLC) model*** (see Figure 2) is a rooted Bayesian network consisting of ***latent nodes***, ***manifest nodes***, ***latent links*** (edges between latent nodes), ***manifest links*** (edges between manifest nodes), and ***cross links*** (edges between a latent node and a manifest node), satisfying: a) latent nodes and latent links form a rooted tree; b) manifest nodes and manifest links form a DAG; and c) all cross links are from a latent node to a manifest node. The *root* latent variable is the class variable.

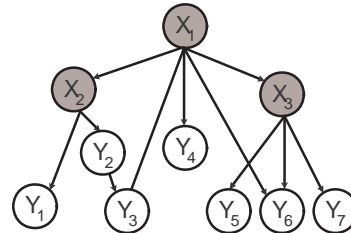


Figure 2: A Generalized Latent Class Model.

We denote the set of latent variables by  $\mathbf{x}$ , the set of manifest variables by  $\mathbf{y}$ , and the set of all variables by  $\mathbf{z}$ . The joint probability distribution over all manifest variables of a GLC model  $M$  is:

$$\begin{aligned} P_M(\mathbf{y}) &= \sum_{\mathbf{x}} \left( \prod_{Z \in \mathbf{z}} P_M(Z | Pa(Z)) \Big|_{\mathbf{x}} \right) \\ &= \sum_{\mathbf{x}} \left( P_M(\mathbf{x}) \prod_{Y \in \mathbf{y}} P_M(Y | Pa(Y)) \Big|_{\mathbf{x}} \right), \end{aligned} \quad (1)$$

where  $\sum_{\mathbf{x}}$  denotes summation over all possible states (value assignments) of  $\mathbf{x}$ . The dependency on  $\mathbf{x}$  is due to the fact that the parents of a variable

may include a node in  $\mathbf{x}$  and thus its distribution depends on the state of  $\mathbf{x}$ .

A GLC model needs to determine three components: the structure, the number of states (cardinality) of each latent variable <sup>1</sup>, and the parameters. A complete GLC model is denoted by the pair  $M = (m, \theta)$ , where  $m$  is the structure including the cardinality information and  $\theta$  denotes parameters.  $m$  is also referred to as an *uninstantiated* GLC model, and  $M$  is referred to as an *instantiated* GLC model.

### 2.3 Model quality metrics

During learning, we use the Bayesian Information Criterion (BIC) (Schwarz 1978) as the model metric score. Given data  $\mathbf{D}$ , the BIC score of an uninstantiated GLC model  $m$  is:

$$BIC(m|\mathbf{D}) = \log P(\mathbf{D}|m, \hat{\theta}) - \frac{\dim(m)}{2} \log N$$

where  $\hat{\theta}$  is the maximum likelihood estimation of model parameters and  $N$  is the number of records.

BIC is used during learning to guide the search, but typically the final quality of a model learned is measured by the significance level  $p$  (Read & Cressie 1988), a metric of the model quality in terms of goodness-of-fit. Higher value of  $p$  indicates better model fit. It is computed by the pair  $(G^2, df)$  where  $G^2$  is the likelihood-ratio  $\chi^2$  statistic with a theoretical  $\chi^2$  distribution of degree of freedom  $df$  (Uebersax 2004). We do not use  $p$  to guide search because not all candidate models can generate reasonable  $p$  value, especially when poor models are examined. But BIC can always be calculated. We have found empirically that BIC works very well to guide the learning towards GLC models with the best fit  $p$ .

## 3 Dominance and Regularization

GLC models address the local dependence problem much more effectively than previous LC, HLC, and LS models. However, the general structure of GLC leads to a huge search space and large increase of the learning complexity. We have found that the previous learning algorithm for HLC and LS works extremely slowly for GLC and cannot converge to high-quality models.

We propose a new approach to reduce the search complexity. The approach is based on the key concepts of model **dominance** and **regularity**. First,

<sup>1</sup>In GLC models, the cardinality information of latent variables needs to be optimized and is separated from the model structure.

we show that not all the models in the search space need to be evaluated, and those dominated by others can be pruned safely. Then, we derive strong necessary conditions for one model to dominate another based on the new concept of regularity. We develop regularization operations to prune the search space and speed up learning. Finally, we develop theoretical upper bound on the search space complexity after pruning by regularization.

### 3.1 Dominance of Bayesian Networks

Based on the previous concepts of marginally equivalence and dimensionality, we define the concept of dominance that is general for Bayesian networks. Roughly speaking, a model  $M$  dominates another model  $M'$  when the two models have equivalent probability distributions but  $M$  is more compact and efficient than  $M'$ . More rigorous definitions are as follows.

**Definition 3.1** Two Bayesian networks  $M = (G, \theta_G)$  and  $M' = (G', \theta_{G'})$  induced from the same data  $\mathbf{D}$  are marginally equivalent if they represent the same probability distributions over all manifest variables, i.e.,  $P_M(V^*|G, \theta_G) = P_{M'}(V^*|G', \theta_{G'})$ , where  $V^*$  is the set of manifest variables in  $\mathbf{D}$ .

**Definition 3.2** The dimensionality, i.e. number of parameters, of a Bayesian network  $M = (G, \theta_G)$  is:

$$Dim(M) = \sum_{Z \in G} \left( (|Z| - 1) \prod_{Y \in Pa(Z)} |Y| \right)$$

where  $G$  is the set of all variables (nodes) and  $|Z|$  is the cardinality (number of states) of a node  $Z$ .  $\prod_{Y \in Pa(Z)} |Y|$  is set to 1 if  $Pa(Z) = \emptyset$ .

Intuitively, marginally equivalent models have the same modeling quality on a dataset since their probability distributions are the same. Dimensionality is a measure of model complexity and usually models with low dimensionality is favored.

**Definition 3.3** A Bayesian network  $M'$  dominates another Bayesian network  $M$  if  $M'$  and  $M$  are marginally equivalent, and  $M'$  has lower dimensionality than  $M$ , i.e.  $Dim(M') \leq Dim(M)$ .  $M$  and  $M'$  are equivalent if  $Dim(M') = Dim(M)$

If a Bayesian network  $M'$  dominates another Bayesian network  $M$ , then  $BIC(M') \geq BIC(M)$ . This is true because  $M'$  and  $M$  are the same in the first part of the BIC score formula, but  $M'$  has a lower dimensionality. Moreover, the significance level  $p$  of  $M'$  is always higher than that of  $M$ . This

is true because  $M'$  and  $M$  share the same  $\chi^2$  but  $M'$  has smaller  $df$ , which leads to larger  $p$ .

The above results show that a dominating model always has **better quality** than the dominated one. Therefore, we can discard the dominated models without sacrificing the model quality. However, to computationally detect and prune the dominated models is a challenging new problem. In the following, we develop theoretical conditions and algorithms to exploit dominance in learning GLC models.

### 3.2 Regularity and dominance condition

Regular GLC models are those that do not contain certain inefficient local structures that frequently appear during learning, including:

- **Island.** An *island* is a latent node  $X$  that has no manifest node in its descendants (see Figure 3).
- **Tie.** A *tie* is a latent node  $X$  that has a parent node and a single child node (Figure 4), satisfying:

$$\begin{cases} |X| \geq \min\{|Pa(X)|, |Ch(X)|\}, \\ \quad \text{if } X \text{ is the only parent node of } Ch(X); \\ |X| \geq |Pa(X)|, \\ \quad \text{if } Pa(X) \text{ is not a parent node of } Ch(X), \end{cases}$$

where  $Ch(X)$  denotes the child of  $X$ .

- **Star.** A *star* is the root node  $X$  which has more than two children nodes, denoted by  $Ch_1(X), \dots, Ch_k(X)$ , with no edges among these children nodes, and also satisfies

$$|X| > \frac{\prod_{i=1}^k |Ch_i(X)|}{\max_{i=1}^k \{|Ch_i(X)|\}},$$

**Definition 3.4** A GLC model is regular if none of its equivalent models (including itself) contains island, tie, or star, and is irregular otherwise.

The following is our main theorem.

**Theorem 3.5** For any irregular GLC model  $M$  on a given dataset, there exists a regular GLC model  $M'$  that dominates  $M$  on the same dataset.

### 3.3 Regularization (Proof of Theorem 3.5)

The proof of Theorem 3.5 is **constructive** and develops a regularization process used in our learning algorithm. In the following, for each of the three local structures that causes irregularity of  $M$ ,

we present a **regularization operation** that eliminates the local structure and obtains a marginally equivalent model with lower dimensionality.

- **Island Removing.** Let  $\mathbf{x}_{\text{island}}$  be the collection of variables in the island. By definition, any variable in  $\mathbf{x}_{\text{island}}$  has no descendant manifest variable. The regularization is simply eliminating  $\mathbf{x}_{\text{island}}$ . The resulting model  $M'$  is marginally equivalent to  $M$  as:

$$\begin{aligned} P_M(\mathbf{y}) &= \sum_{\mathbf{x}} P_M(\mathbf{x}) P_M(\mathbf{y}|\mathbf{x}) \\ &= \sum_{\mathbf{x}-\mathbf{x}_{\text{island}}} \sum_{\mathbf{x}_{\text{island}}} \left( P_M(\mathbf{x}_{\text{island}}) P_M(\mathbf{x}-\mathbf{x}_{\text{island}}) \right. \\ &\quad \left. P_M(\mathbf{y}|\mathbf{x}-\mathbf{x}_{\text{island}}) \right) \\ &= \sum_{\mathbf{x}-\mathbf{x}_{\text{island}}} P_{M'}(\mathbf{x}-\mathbf{x}_{\text{island}}) P_{M'}(\mathbf{y}|\mathbf{x}-\mathbf{x}_{\text{island}}) \\ &= P_{M'}(\mathbf{y}), \end{aligned}$$

and  $M'$  has lower dimensionality than  $M$ .

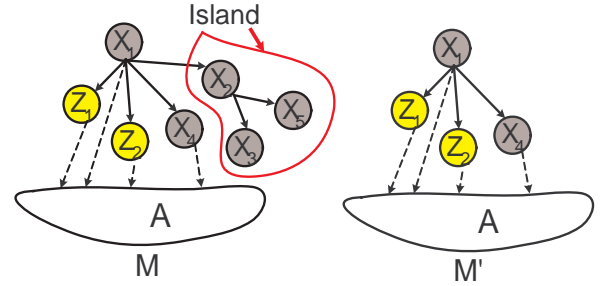


Figure 3: Illustration of island removing.

Figure 3 shows an example of island removing.

- **Tie Integration.** Figure 4 shows the two cases of a tie.

Case 1: The tie  $X_2$  in Figure 4 (a)  $M$  is the only parent node of  $Z$ . For  $M$  in Figure 4 (a), the regularization is to remove  $X_2$ , connect  $X_1$  to  $Z$ , and set  $P(Z|X_1)$  in  $M'$  as  $\sum_{X_2} P(X_2|X_1)P(Z|X_2)$  in  $M$ .

It is obvious that  $P_{M'}(\mathbf{y}) = P_M(\mathbf{y})$ . According to the definition of ties, we know  $|X_2| \geq \min\{|X_1|, |Z|\}$ . If  $|X_2| \geq |X_1|$ , we have

$$\begin{aligned} Dim(M) - Dim(M') &= |X_1|(|X_2| - 1) + |X_2|(|Z| - 1) - |X_1|(|Z| - 1) \\ &= |X_1|(|X_2| - 1) + (|Z| - 1)(|X_2| - |X_1|) \geq 0. \end{aligned}$$

If  $|X_2| \geq |Z|$ , we have

$$\begin{aligned} Dim(M) - Dim(M') &= |X_1|(|X_2| - 1) + |X_2|(|Z| - 1) - |X_1|(|Z| - 1) \\ &= |X_1|(|X_2| - |Z|) + |X_2|(|Z| - 1) \geq 0. \end{aligned}$$

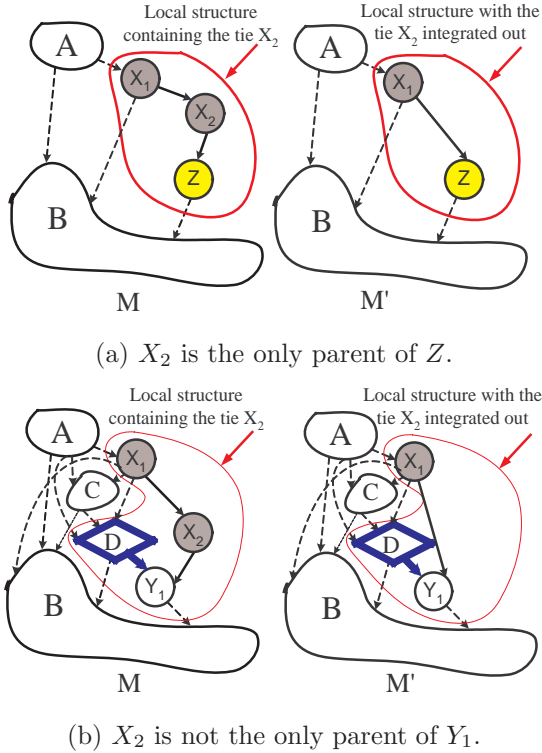


Figure 4: Illustration of tie integration.

Therefore, we must have  $\text{Dim}(M) \geq \text{Dim}(M')$ .

Case 2: In Figure 4 (b), The tie  $X_2$  is not the only parent of the node  $Y_1$ . The diamond  $D$  represents the other parents of  $Y_1$ , enumerated as  $Z_1, \dots, Z_k$ .

For  $M$  in Figure 4 (b), the regularization is to remove  $X_2$  and set  $P(Y_1|Z_1, \dots, Z_k, X_1)$  in  $M'$  as  $\sum_{X_2} P(X_2|X_1)P(Y_1|Z_1, \dots, Z_k, X_2)$  in  $M$  to ensure  $P_{M'}(\mathbf{y}) = P_M(\mathbf{y})$ . To show  $\text{Dim}(M) \geq \text{Dim}(M')$ , since each variable has at least two states, we have

$$\begin{aligned}
& \text{Dim}(M) - \text{Dim}(M') \\
&= |X_1|(|X_2| - 1) + |X_2|(|Z_1||Z_2| \cdots |Z_k|)(|Y_1| - 1) \\
&\quad - |X_1|(|Z_1||Z_2| \cdots |Z_k|)(|Y_1| - 1) \\
&= (|Z_1||Z_2| \cdots |Z_k|)(|Y_1| - 1)(|X_2| - |X_1|) \\
&\quad + |X_1|(|X_2| - 1) > 0.
\end{aligned}$$

As a special case, if  $X_1$  is a parent node of  $Y_1$  in Figure 4 (b), no matter if  $Z_1, \dots, Z_k$  exist or not, we can always obtain a GLC model  $M'$  dominating  $M$  by removing  $X_2$  and setting  $P(Y_1|Z_1, \dots, Z_k, X_1)$  in  $M'$  as  $\sum_{X_2} P(X_2|X_1)P(Z_1, \dots, Z_k, X_2, X_1)$  in  $M$ .

• **Star Adjustment.** Finally, we show the regularization of a GLC model containing a *star*. We consider the GLC model as an undirected graph which becomes a conditional random field. Instead of us-

ing CPDs, we use potential functions. Due to the space limitation, we omit the mathematical details and outline the regularization process.

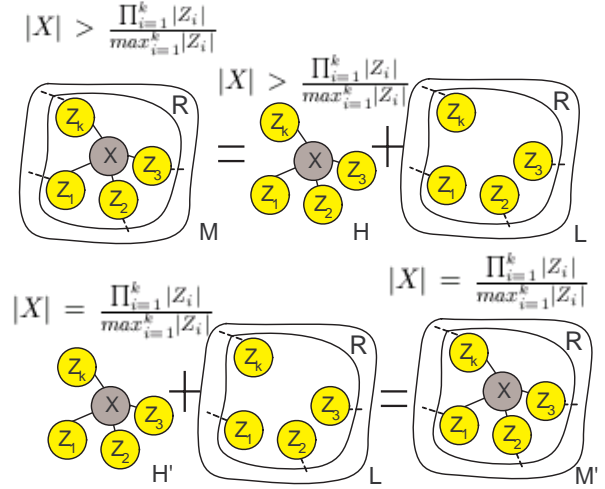


Figure 5: Illustration of star adjustment.

Figure 5 shows an example where the neighbor nodes of a star  $X$  is  $Z_1, \dots, Z_k$ . Without loss of generality, we set  $|Z_1| \leq |Z_2| \leq \dots \leq |Z_k|$ . We divide  $M$  into the local star structure  $H$  and the set of other nodes  $L$ . The product of all potential functions (including all edges and cliques) in  $M$  equals to the product of all potential functions (including all edges) in  $H$  and all potential functions (including all edges and cliques) in  $L$ . In addition,  $H$  is an undirected HLC model and satisfies  $|X| > \frac{\prod_{i=1}^k |Z_i|}{\max_{i=1}^k |Z_i|}$  (Zhang 2002). An HLC model  $H'$  can be constructed which is marginally equivalent to  $H$  but has lower dimensionality. It is achieved by setting the number of states of  $X$  as  $\prod_{i=1}^{k-1} |Z_i|$  and modifying the potential functions for each edge of  $H$  accordingly. The new GLC model  $M'$ , composed of  $H'$  and  $L$ , is marginally equivalent to  $M$  and has lower dimensionality.

Given an irregular GLC model  $M$ , we can iteratively apply the regularization operations to eliminate all the irregular local structures and obtain a regular model  $M'$ . Since  $M'$  is marginally equivalent to  $M$  and  $\text{Dim}(M') \leq \text{Dim}(M)$ ,  $M'$  dominates  $M$ . Theorem 3.5 is proved.

### 3.4 Space complexity after regularization

We derive an important theoretical upper bound on the number of latent nodes in a regular uninstantiated GLC model given a set of manifest variables.

This theorem shows the significant pruning of search space due to dominance and regularization.

**Theorem 3.6** For a regular GLC model, let  $n$  be the number of manifest variables, and  $k$  be the maximum number of latent nodes in the parents of any manifest variable, the number of latent nodes in the model is less than  $3kn$ .

**Proof:** Consider a regular GLC model  $M'$ . Since there is no island, any latent node has at least two neighbors. Let a *singly connected* latent node be one with exactly two neighbor nodes, we have that two singly connected latent nodes  $X_1$  and  $X_2$  can not be linked. Otherwise, among the equivalent models of  $M'$ , there must exist one containing a tie, because either  $|X_1| \geq |X_2|$  or  $|X_1| \leq |X_2|$  holds and any of them can be the parent node of another in an equivalent model of  $M'$ .

We construct a new graph  $S$  by removing the manifest part of  $M$  and adding a new manifest node to the empty end of each cross link.  $S$  is a tree with  $kn$  manifest leaf nodes. A singly connected node in the original GLC model is still singly connected in  $S$ , and the non-singly connected node in the original GLC model is still non-singly connected in  $S$ . We consider two cases:

Case 1:  $S$  contains no singly connected latent nodes. Let  $m$  be the total number of latent nodes. Since  $S$  is a tree, we have  $kn + (m - 1)$  edges in total. On the other hand,  $S$  contains no singly connected latent nodes. Any latent node connects at least three edges. We have  $(kn + 3m)/2 \leq kn + (m - 1)$ , which leads to  $m \leq kn - 2 < kn$ .

Case 2:  $S$  contains singly connected latent nodes. Let  $I = m + kn$  be the total number of nodes in  $S$ . If we root  $S$  at any latent node, any singly connected latent node will have a child different from the child of any other singly connected latent node. Hence, if we eliminate all singly connected latent nodes, we obtain a new tree structure  $S^*$  with no singly connected latent nodes.  $S^*$  contains at least  $I/2$  nodes in total. We have  $I/2 - kn < kn$ . Consequently,  $m < 3kn$ . ■

Note that, without regularization, the number of latent nodes can be  $O(c^{kn})$ , an exponential of the number of manifest nodes.<sup>2</sup> Theorem 3.6 shows that pruning by dominance can reduce the original *exponential* search space to a *linear* space in terms of  $n$ . The saving is enormous.

<sup>2</sup>We have proved this conclusion using the concept of ‘model identifiability’ (Goodman 1974). The proof is out of the scope of this paper.

## 4 Regularity-Based Learning

We develop a learning algorithm similar to the Two-Phase learning of HLC models (Zhang and Kočka 2004) and the greedy equivalence search (GES) (Meek 1997), but integrated with regularization. The greedy local search algorithm is shown below.

---

### Algorithm 1 Learn a GLC model from dataset $\mathbf{D}$

---

```

1: function  $m = \text{GLC\_LEARNING}(\mathbf{D})$ 
2:   Let  $m'$  be an LC model      ▷ starting point
3:    $\Sigma \leftarrow \phi$           ▷  $\Sigma$  is the expanding pool
4:   repeat
5:      $m \leftarrow m'$ 
6:      $\Sigma \leftarrow \text{RaisingComplexity}(m, k)$ 
7:      $\Sigma' \leftarrow \text{Regularization}(\Sigma, m)$ 
8:      $m' \leftarrow \text{GreedyDescending}(\Sigma', m)$ 
9:   until  $m' = m$ 
10:  return  $m$ 
11: end function

```

---

In **RaisingComplexity**, we exhaustively generate all neighboring GLC models obtainable by adding/removing a node, adding/removing a link, or increasing/decreasing the cardinality of a node by 1. We set  $k$ , the upper bound on the number of latent parents a manifest node can have, to be the number of manifest nodes  $n$ , since typically the number of latent parents is far less than  $n$ .

In **Regularization**, the regularization operations are used to prune the search space and prevent the search from diverging. Irregular models will be regularized using *island removing*, *tie integration*, and *star adjustment*.

In **GreedyDescending**, for each regular model left in  $\Sigma'$ , its parameters are optimized using the EM algorithm (Dempster *et al.* 1977). We select a best model based on the BIC improvement per unit increase in model complexity (Zhang and Kočka 2004) defined as:

$$U(m', m|\mathbf{D}) = \frac{BIC(m'|\mathbf{D}) - BIC(m|\mathbf{D})}{\dim(m') - \dim(m)},$$

where  $m$  is current model and  $m'$  is a candidate model. Models with  $BIC(m'|\mathbf{D}) \leq BIC(m|\mathbf{D})$  are removed from the pool  $\Sigma'$ . If  $\Sigma'$  is empty, we return  $m$  and quit. If there exists a model  $m'$  such that  $\dim(m') \leq \dim(m)$  in the rest of models, the model with the smallest  $\dim(m')$  is chosen for the seed. Otherwise, the model with the highest  $U(m', m|\mathbf{D})$  is chosen. We start the next iteration using the selected  $m'$  as the starting model.

**Remarks.** a) A majority of learning time is spent in optimizing model parameters using the EM algorithm. Hence, regularization greatly reduces the number of EM evaluations by eliminating all irregular models. b) The naive scheme of simply discarding irregular models does not work. The reason is that the regions of regular models are not connected by neighborhoods. There are many irregular models whose neighbors are all irregular. Therefore, if we discard all irregular models, the search usually quickly gets stuck at an irregular model. Regularization allows the search to traverse directly from an irregular model to a regular one.

## 5 Experimental Results

In this section, we perform our experiments on various real datasets. The experiments were conducted on a 1.8GHz IBM Pentium 4 laptop. The program is implemented using the Bayes Net Toolbox (Murphy 2006) on Matlab 7.0.

Since regular GLC models are always better than previous LC, HLC and LS models in terms of solution quality for all the datasets, we focus on comparing the GLC learning with and without regularization. We show that, if we do not use regularization, the learning algorithm makes extremely slow progress, and fails to converge to optimal GLC models for all tests. In contrast, if we use regularization, the learning is very efficient and converges to high-quality models.

**Dataset 1: The hannover rheumatoid arthritis data** There are 7,162 records in this data taken from the study by Wasmus *et al.* (1989) on the prevalence of rheumatoid arthritis in the adult population. Five symptoms (A-E) formed five yes/no questions to be answered. Figure 6 and Table 1 summarize the results. In Table 1, the first column indicates whether or not (‘Yes/No’) regularization is used during learning. The learning algorithm without using regularization converged in 2 hours and 35 minutes to a final model (Figure 6) with quality  $p:0.396$ . In contrast, using regularization, the algorithm converged to the optimal GLC model with  $p : 0.998$ . The best previous model is an LS model with  $p : 0.936$ .

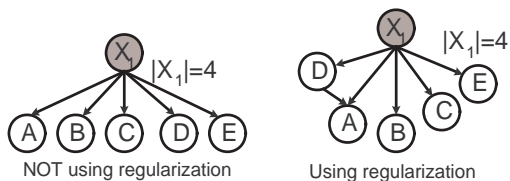


Figure 6: Models learned for dataset 1.

Regularization	$G^2$	$df$	$p$	Time
No	8.394	8	0.396	2.5hrs
Yes	0.132	4	0.998	8.5min

Table 1: Results for dataset 1.

### Dataset 2: The HIV data

This data set (Alvord *et al.* 1988) was collected from the human HIV virus test results performed on 428 subjects. There are four diagnostic tests denoted by four letters A to D. The learning algorithm without using regularization converged in 22 minutes (see Table 2) to the model in Figure 7 with  $p:0.549$ . In contrast, using regularization, the algorithm converged to the a model with  $p:0.910$  in 5 minutes. The best previous model is an HLC model with  $p : 0.549$ .

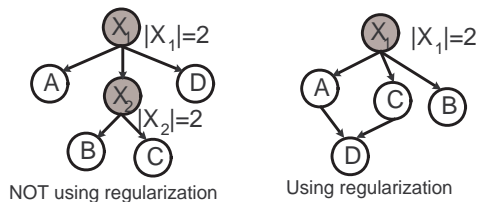


Figure 7: Models learned for dataset 2.

Regularization	$G^2$	$df$	$p$	Time
No	3.056	4	0.549	22min
Yes	1.00	4	0.910	5min

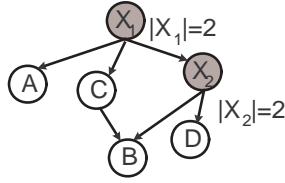
Table 2: Results for dataset 2.

### Dataset 3: The Coleman data

The Coleman data studies the social system of adolescents in ten high schools (Coleman 1964). It is concerned with the membership in leading crowds and attitudes towards it. Responses of schoolboys to two questions were collected on two interviews at two different time points. Four manifest variables, i.e., A-D, are formed for tests. The learning algorithm without using regularization converged in 53 minutes to a model with  $p:0.870$ . Using regularization, the algorithm converged to the same model in only 4 minutes. The result is shown in Figure 8 and Table 3. The best previous model is an LS model with  $p : 0.867$ .

### Dataset 4: The house building data

This data was used by Hagenaars (1988) on the study of people’s view about what a new government should do. Responses to two kinds of questions were recorded at two different dates. One question was about whether house building was an important problem, and the others were about how important house building was in relation to some other issues. There are four dichotomous (binary) manifest variables: A and C denote the answers to the first ques-



Using or NOT using regularization

Figure 8: Model learned for dataset 3.

Regularization	$G^2$	$df$	$p$	Time
No	0.278	2	0.870	53min
Yes	0.278	2	0.870	4min

Table 3: Results for dataset 3.

tion at two separated interview times, and B and D denote the answers to the second question at the two different times. Figure 9 and Table 4 summarize the results. The learning algorithm without using regularization converged in 40 minutes to a model with  $p:0.446$ . The algorithm using regularization converged to a model with  $p:0.757$  in 4.5 minutes. Previous HLC, LS, and LC models all fail on this data ( $p : 0.000$ ).

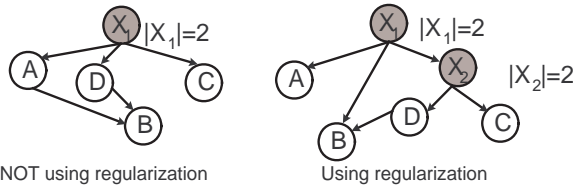


Figure 9: Models learned for dataset 4.

## 6 Conclusions

In this paper, we have proposed the concept of dominance of Bayesian networks and regularization operations that can generate dominating nodes for a generalized class of latent class models. Integrating the regularization operations in the learning algorithm leads to significant pruning of the search space. Based on the flexible topology of GLC models and the efficient learning algorithm, for several datasets, we have obtained GLC models with better quality than all the previous latent class models. We plan to study dominance and pruning for learning other classes of Bayesian networks in the future.

## References

Alvord, W. G. *et al.* (1988). A method for predicting individual HIV infection status in the absence of clinical information. *AIDS Res Hum Retroviruses*, 4(4): 295-304.

Bartholomew, D. J. and Knott, M. (1999). *Latent vari-*

Reg.	$G^2$	$df$	$p$	Time
No	3.713	4	0.446	40min
Yes	0.558	2	0.757	4.5min

Table 4: Results for dataset 4.

*able models and factor analysis*, 2nd edition. Kendall's Library of Statistics 7. London: Arnold.

Coleman, J. S. (1964). *Introduction to Mathematical Sociology*. Glencoe, Illinois: Free Press.

Dempster, A. P., *et al.* (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1-38.

Friedman, N. (1997) Learning belief networks in the presence of missing values and hidden variables. *ICML-97*, 125-133.

Geiger, D., Heckerman, D., and Meek, C. (1996) Asymptotic Model Selection for Directed Networks with Hidden Variables. *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence, Portland, Oregon, GLCA (UAI-96)*, 158-168.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, 215-231.

Hagenaars, J. A. (1988). Latent structure models with direct effects between indicators: local dependence models. *Sociological Methods and Research*, 16, 379-405.

Hua, D., *et al.* (2003). Latent Structure Models for the Analysis of Gene Expression Data, *Proceedings of the IEEE Computer Society Bioinformatics 2003*, pp. 496-499.

Lazarsfeld, P. R. and Henry, N. W. (1968). *Latent Structure Analysis*. Boston:Houghton-Mifflin.

Meek, C. (1997). Graphical models: *Selection causal and statistical models*. *Ph.D. Thesis*, Carnegie Mellon University.

Murphy, K.P. (2006) <http://bnt.sourceforge.net/>

Read, T. R. C., and Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464.

Uebersax, J. (2004). A Practical Guide to Local Dependence in Latent Class Models. [ourworld.compuserve.com/homepages/jsuebersax/](http://ourworld.compuserve.com/homepages/jsuebersax/)

Wasmus, A., *et al.* (1989). Activity and severity of rheumatoid arthritis in Hannover/FRG and in one regional referral center. *Scandinavian Journal of Rheumatology*, Suppl. 79, 33-44.

Zhang, N. L. (2002). Hierarchical latent class models for cluster analysis. *AAAI'02*, 230-237.

Zhang, N. L. (2004) Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5, 697-723.