Jrl Syst Sci & Complexity (2008) 21: 172–183

DISCOVERY OF LATENT STRUCTURES: EXPERIENCE WITH THE COIL CHALLENGE 2000 DATA SET*

Nevin L. ZHANG · Yi WANG · Tao CHEN

Received: 13 August 2007 / Revised: 10 October 2007

Abstract The authors present a case study to demonstrate the possibility of discovering complex and interesting latent structures using hierarchical latent class (HLC) models. A similar effort was made earlier by Zhang (2002), but that study involved only small applications with 4 or 5 observed variables and no more than 2 latent variables due to the lack of efficient learning algorithms. Significant progress has been made since then on algorithmic research, and it is now possible to learn HLC models with dozens of observed variables. This allows us to demonstrate the benefits of HLC models more convincingly than before. The authors have successfully analyzed the CoIL Challenge 2000 data set using HLC models. The model obtained consists of 22 latent variables, and its structure is intuitively appealing. It is exciting to know that such a large and meaningful latent structure can be automatically inferred from data.

Key words Bayesian networks, case study, latent structure discovery, learning.

1 Introduction

Hierarchical latent class (HLC) models[1−2] are tree-structured Bayesian networks where variables at leaf nodes are observed and are hence called manifest variables, while variables at internal nodes are hidden and are hence called latent variables. All variables are assumed discrete. HLC models generalize latent class (LC) models[3] and were first identified as a potentially useful class of Bayesian networks (BNs) by Pearl[4].

HLC models can be used for latent structure discovery. Often, observed variables are correlated because they are influenced by some common hidden causes. HLC models can be seen as hypotheses about how latent causes influence observed variables and how they are correlated among themselves. Then, finding an HLC model that fits a data set amounts to finding a latent structure that explains the data well.

In general, graphical models with latent variables are of interest in many fields, including statistics[5], bioinformatics[6], and computer science[7]. When working with such a model, most researchers assume that the model structure is known. There has been relatively little work on inferring latent structures from data. One exception is the research on the linear latent variable graphs (LLVGs)[8]. The task there is to infer, from data, a two-layered Bayesian

Nevin L. ZHANG · Yi WANG · Tao CHEN
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China. Email: {lzhang,wangyi,csct}@cse.ust.hk.
*The research is supported by Hong Kong Grants Council Grants #622105 and #622307, and the National Basic Research Program of China (aka the 973 Program) under project No. 2003CB517106.

network where nodes on the upper level are latent while those on the lower level are observed. In LLVGs, all the variables are continuous, and each variable depends linearly on its parents via a regression equation. Another exception is phylogenetic tree reconstruction, which takes DNA sequences of current day species as input and infers a phylogenetic tree that exhibits the ancestral relationships among those species. The introduction of HLC models provides one addition to this small collection of tools for latent structure discovery.*

The CoIL Challenge 2000 data set[9] contains information on customers of a Dutch insurance company. The data set consists of 86 variables, around half of which are about ownerships of various insurance products. Ownership variables for different products are correlated. One who pays a high premium on one type of insurance is more likely, than those who do not, to also purchase other types of insurance in the same category. Intuitively, such correlations are due to people's (latent) attitudes toward risks. The more risk-aversion one is toward a category of risks, the more likely he is to purchase insurance products in that category. Therefore, the CoIL Challenge 2000 data set is a good testbed for latent structure discovery methods.

We have analyzed the CoIL Challenge 2000 data set using HLC models. The structure of the model obtained is given in Section 5. As the reader will see, there are 42 manifest variables and 22 latent variables, and the structure is intuitively appealing. Latent structure discovery is very difficult. It is hence exciting to know that we are able to discover such a complex and meaningful structure. Latent structures of similar sizes have been constructed before in phylogenetic tree reconstruction, but phylogenetic trees are much more restrictive than HLC models.

In addition to being a tool for latent structure discovery, HLC models can also be used for cluster analysis. In this role, they alleviate disadvantages of LC models as models for discrete data clustering. An LC model consists of one latent variable, namely the class variable, and a number of observed feature variables. It assumes that the feature variables are mutually independent given the class variable. A serious problem with the use of LC models, known as local dependence, is that this assumption is often violated. If one does not deal with local dependence explicitly, one implicitly attributes it to the latent variable. In practice, this results in spurious clusters, and degenerates the accuracy of classification[10]. In Section 6, we will show that HLC models do produce more meaningful clusters for the CoIL Challenge 2000 data set than LC models.

HLC models can also be used simply for probabilistic modeling. They possess two nice properties for this purpose. First, they have low inferential complexity due to their tree structures. Second, they can model complex dependencies among the observed variables. In Section 7, the reader will see the implications of the second property on prediction and classification accuracy in the context of the CoIL Challenge 2000 data.

We begin with a review of HLC models in Section 2 and discuss the learning of HLC models in Section 3. A description of the CoIL Challenge 2000 data set is then given in Section 4.

2 Hierarchical Latent Class Models

Figure 1 shows an example HLC model (left diagram). A latent class (LC) model is an HLC model where there is only one latent node. We usually write an HLC model as a pair $M=(m, \theta)$, where the second component θ is the collection of parameters, and the first component m consists of the model structure and the cardinalities of latent variables. Here, the cardinality

*The probabilistic models of phylogenetic trees[6] can be viewed as special HLC models where 1) each latent node has exactly two children; 2) all variables have four states, namely A, C, G, and T; and 3) the conditional probability table of each variable has only one parameter, i.e., the length of the incoming edge.

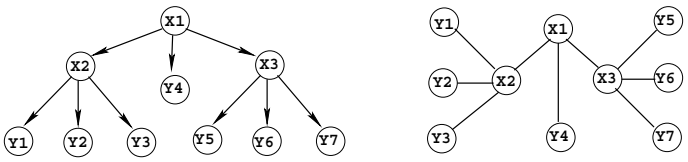


Figure 1 An example HLC model and the corresponding unrooted HLC model. The X_i's are latent variables and the Y_j's are manifest variables.

of a variable is the number of its possible values. We will sometimes refer to m also as an HLC model.

Two HLC models $M{=}(m, \theta)$ and $M'{=}(m', \theta')$ are marginally equivalent if they share the same manifest variables $Y_1, Y_2, \cdots, Y_n$ and

$$P(Y_1, Y_2, \cdots, Y_n|m, \theta) = P(Y_1, Y_2, \cdots, Y_n|m', \theta'). \tag{1}$$

An HLC model m includes another model m' if for any parameter value θ' of m', there exists a parameter value θ of m such that (m, θ) and (m', θ') are marginally equivalent. In this case, m can represent any distribution over the manifest variables that m' can represent. If m includes m' and vice versa, we say that m and m' are marginally equivalent. Marginally equivalent models are equivalent if they have the same number of independent parameters. One cannot distinguish between equivalent models using penalized likelihood scores.

Let X_1 be the root of an HLC model m. Suppose that X_2 is a child of X_1 and it is also a latent node. Define another HLC model m' by reversing the edge $X_1{\rightarrow}X_2$. Then X_2 becomes the root of m'. This operation is hence called root walking: the root has walked from X_1 to X_2. Root walking leads to equivalent models[2]. This implies that it is impossible to determine edge orientations from data. We can learn only unrooted HLC models, which are HLC models with all directions on the edges dropped. Figure 1 also shows an example unrooted HLC model. An unrooted HLC model represents a class of equivalent HLC models. Members of the class are obtained by rooting the model at various nodes. Semantically it is a Markov random field on an undirected tree. The leaf nodes are observed while the internal nodes are latent. Marginal equivalence and equivalence can be defined for unrooted HLC models in the same way as for rooted models. From now on when we speak of HLC models we always mean unrooted HLC models unless it is explicitly stated otherwise.

Let $|X|$ stand for the cardinality of a variable X. For a latent variable Z in an HLC model, enumerate its neighbors as $X_1, X_2, \cdots, X_k$. An HLC model is regular if for any latent variable Z, $|Z| \leq \prod_{i=1}^{k} |X_i| / \max_{i=1,2,\cdots,k} |X_i|$, and when Z has only two neighbors, strict inequality holds and one of the neighbors must be a latent node.

Given an irregular model m, there exists a regular model that is marginally equivalent to m and has fewer independent parameters[2]. The process of obtaining the regular model is called regularization. It is evident that if penalized likelihood scores are used for model selection, the regularized model is always preferred over m itself.

3 Learning HLC Models

Assume that there is a data set $\boldsymbol{D}$ on a given set of manifest variables. How can we induce from $\boldsymbol{D}$ an HLC model? This question can be divided into two sub-questions. First, among all the possible models which one is the best? This is the model selection problem.

Zhang[1,2] empirically examined several criteria, namely the BIC score[11], the AIC score[12], the Cheeseman-Stutz score[13], and holdout-likelihood[14]. The BIC score turns out to be the most appropriate one for the task. The BIC score of an HLC model m is given by

$$\mathrm{BIC}(m|\boldsymbol{D}) = \log P(\boldsymbol{D}|m, \theta^*) - \frac{d(m)}{2} \log N,$$

where $\boldsymbol{D}$ is the data set, θ^* is the maximum likelihood estimate of the model parameters, $d(m)$ is the number of independent parameters, and N is the sample size. Note that this definition of the BIC score is used in the machine learning community, while researchers in social sciences usually use its negation.

The BIC score is a large sample approximation of the marginal likelihood $P(\boldsymbol{D}|m)$ derived in a setting where all variables are observed. Geiger et al.[15] have re-done the derivation for latent variable models and arrived at another scoring function called the BICe score. The BICe score is the same as the BIC score except that the standard dimension $d(m)$ is replaced by the effective dimension of the model. Theoretically, BICe is advantageous over BIC. Why BIC is still used for scoring HLC models? There are three reasons. First, despite recent decomposition results[16], effective model dimensions remain difficult to compute. Second, there is no substantial empirical evidence showing that BICe is advantageous over BIC in practice. Third, our experiences with about one dozen data sets suggest that one can find good models with BIC.

The second sub-question is how to find the model with the highest BIC score in the space of all possible models. Three search-based algorithms have been proposed for this task, namely Double Hill-Climbing (DHC)[1,2], Single Hill-Climbing (SHC)[17], and Heuristic Single Hill-Climbing (HSHC)[17]. All those algorithms aim at finding the model with the highest BIC score.

In the following, we distinguish between HLC models and HLC model structures. In an HLC model, cardinalities of latent variables are specified. In an HLC model structure, they are not. DHC is the first search-based algorithm for learning HLC models. It searches in the space of HLC model structures. It starts with the structure with only one latent node. At each step, it first generates a number of candidate model structures by modifying the current structure using three search operators, namely node introduction, node deletion, and node relocation. It then optimizes the cardinalities of the latent variables in each of the candidate structures, resulting in candidate models. Finally, it evaluates the candidate models and uses the structure of the best one to seed the next search step. The search terminates when the model score stops increasing. To optimize the cardinalities of the latent variables in a model structure, the algorithm employs another hill-climbing routine. That is why it is called double hill-climbing.

SHC is the second search-based algorithm for learning HLC models. It searches in the space of HLC models. It has five search operators, namely node introduction (NI), node deletion (ND), node relocation (NR), state introduction (SI), and state deletion (SD). Although sharing the same names as the operators of DHC, the first three operators of SHC are different from the DHC operators. Their outputs are HLC models rather than HLC model structures. This means that cardinalities of latent variables were considered when designing those operators, while this was not the case with DHC. SHC does not have a separate routine to optimize the cardinalities of latent variables. It optimizes them together with the model structure. This is why SHC has the SI and SD operators.

SHC starts with the simplest HLC model and searches in two phases. In Phase 1, it hill climbs with the NI, NR, and SI operators. When model score ceases to improve, it moves to Phase 2 and continues search with the other two operators, ND and SD. If the model score is improved in Phase 2, the process repeats itself. Otherwise the algorithm terminates.

SHC is more efficient than DHC. However, it is still computationally very expensive, mainly because it needs to evaluate and hence runs the expectation-maximization (EM) algorithm on each of the candidate models. HSHC alleviates the situation by incorporating the technique of structural EM[18]. The idea is to complete the data using the current model and evaluate the candidate models using the completed data and hence avoiding EM. The main technical issue here is that the candidate models produced by NI, SI, and SD involve some different latent variables than the current model. Consequently, it is not straightforward to evaluate the candidate models using the completed data. The problem was solved using several heuristics.

4 The CoIL Challenge 2000 Data Set

The training set of the CoIL Challenge 2000 data set consists of 5,822 customer records. Each record consists of 86 attributes, containing socio-demographic information (Attributes 1–43) and insurance product ownerships (Attributes 44–86). The socio-demographic data are derived from zip codes. In previous analysis, these variables were found more or less useless. In our analysis, we included only three of them, namely Attributes 4 (average age), 5 (customer main type), and 43 (purchasing power class). All the product ownership attributes were included in the analysis.

The data was preprocessed as follows: First, similar attribute values were merged so that there are at least 30 records for each value. In the resulting data set, there are fewer than 10 records where Attributes 50, 60, 71, and 81 take "nonzero" values. Those attributes were excluded from further analysis. The final data set consists of 42 attributes, each with 2 to 9 values.

We analyzed the data using a Java implementation of the HSHC algorithm. HSHC has one algorithmic parameter K. We tried four values for K, namely 1, 5, 10, and 20. The experiments were run on a Pentium 4 PC with a clock rate of 2.26 GHz. The running times and the BIC scores of the resulting models are shown in Table 1. The best model was found in the case of $K = 10$. We denote the model by M^*. The structure of the model is shown in Figure 2[†].

Table 1

K	1	5	10	20
Time (hrs)	51	99	121	169
BIC	$-52,522$	$-51,625$	$-51,465$	$-51,592$

5 Latent Structure Discovery

Did HSHC discover interesting latent structures? The answer is positive. We will show this by examining different aspects of the model M^*. First of all, the data set contains two variables for each type of insurance. For bicycle insurance, for instance, there are "contribution to bicycle insurance policies (v_{62})" and "number of bicycle insurance policies (v_{83})". HSHC introduced a latent variable for each such pair. The latent variable introduced for v_{62} and v_{83} is h_{11}, which can be interpreted as "attitude toward bicycle risks". Similarly, h_{10} can be interpreted as "attitude toward motorcycle risks", h_9 as "attitude toward moped risks", and so on.

Consider the manifest variables on the right hand side of h_{12}. Except "social security", all the other variables are related to heavy private vehicles. HSHC concluded that they are

[†]Note that what HSHC obtains is an unrooted HLC model. The structure of the model is visually shown as a rooted tree in Figure 2 partially for readability and partially due to the discussions of the following section.

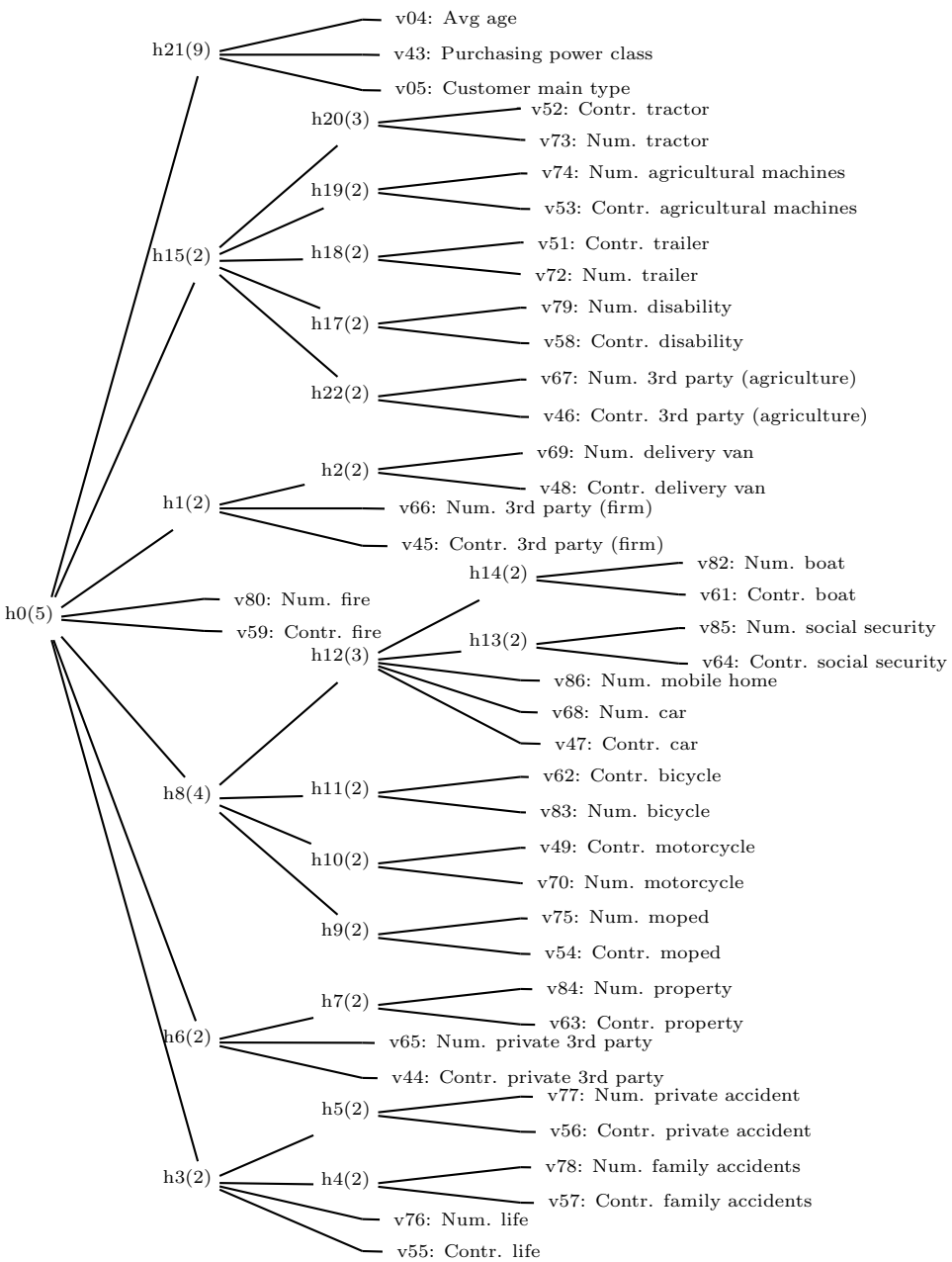


Figure 2 The structure of the best model M^* found for the CoIL data. The number next to a latent variable is the cardinality of that variable.

influenced by one common latent variable. This is clearly reasonable and h_{12} can be interpreted as "attitude toward heavy private vehicle risks". Except "social security", all the manifest variables on the right hand side of h_8 are related to private vehicles. HSHC concluded that they are influenced by one common latent variable. This is reasonable and h_8 can be interpreted as "attitude toward private vehicle risks".

All the manifest variables on the right hand side of h_{15}, except "disability", are agriculture-related; while the manifest variables on the right hand side of h_1 are firm-related. It is therefore reasonable for HSHC to conclude that those two groups of variables are respectively influenced by two latent variables h_1 and h_{15}, which can be interpreted as "attitude toward firm risks" and "attitude toward agriculture risks" respectively.

It is interesting to note that, although delivery vans and tractors are vehicles, HSHC did not conclude that they are influenced by h_8. HSHC reached the correct conclusion that the decisions to buy insurance for tractors, for delivery vans, or for other private vehicles are influenced by different latent factors.

The manifest variables on the right hand side of h_3 intuitively belong to the same category; those on the right hand side of h_6 are also closely related to each other. It is therefore reasonable for HSHC to conclude that those two groups of variables are respectively influenced by latent variables h_3 and h_6.

The three socio-demographic variables (v_{04}, v_{05}, and v_{43}) are connected to latent variable h_{21}. Hence h_{21} can be viewed as a venue for summarizing information contained in those three variables. Latent variable h_0 can be interpreted as "general attitude toward risks". Under this interpretation, the links between h_0 and its neighbors are all intuitively reasonable: One's general attitude toward risks should be related to one's socio-demographic status (h_{21}), and should influence one's attitudes toward specific risks (h_8, h_1, h_{15}, etc).

There are also aspects of model M^* that do not match our intuition well. For example, since there is a latent variable (h_{12}) for heavy private vehicles on the right hand side of h_8, we would naturally expect a latent variable for light private vehicles. However, there is no such variable. On the right hand of h_3, we would expect a latent variable specifically for life insurance. Again, there is no such variable. The placement of the latent variables (h_{13} and h_{17}) about social security and disability is also questionable. With an eye on improvements, we have considered a number of alterations to M^*. However, none resulted in models better than M^* in terms of BIC score.

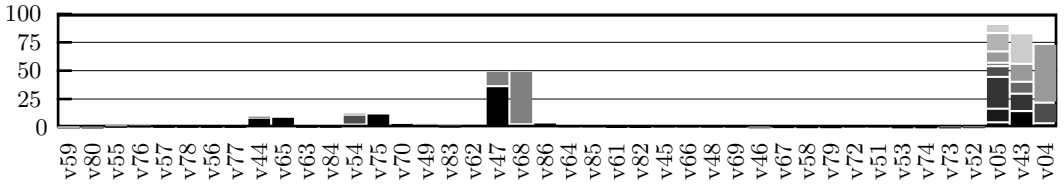
Some mismatches are partially due to the limitations of HLC models. Disability is a concern in both agriculture and firms. We would naturally expect h_{17} (attitude toward disability risks) to be connected to both h_1 (attitude toward firm risks) and h_{15} (attitude toward agriculture risks). However, that would create a cycle, which is not allowed in HLC models. Hence, there is a need to study generalizations of HLC models in the future.

6 Cluster Analysis

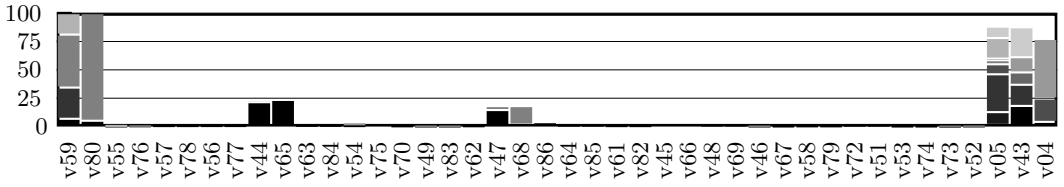
Each latent variable in an HLC model corresponds to one way to cluster data. Therefore, when learning an HLC model, one is actually performing multidimensional clustering. In contrast, latent class analysis results in one single clustering.

The latent variable h_0 in model M^* has 5 states and it is interpreted as "general attitude toward risks". This means that HSHC has identified 5 clusters in the CoIL Challenge 2000 data according to customer's general attitude toward risks. The class-conditional probability distributions of those clusters are shown as bar diagrams in Figure 3.

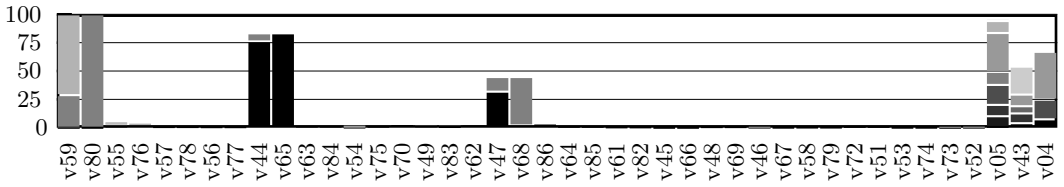
In the bar diagrams, each bar is labeled with a manifest variable. The bar depicts the



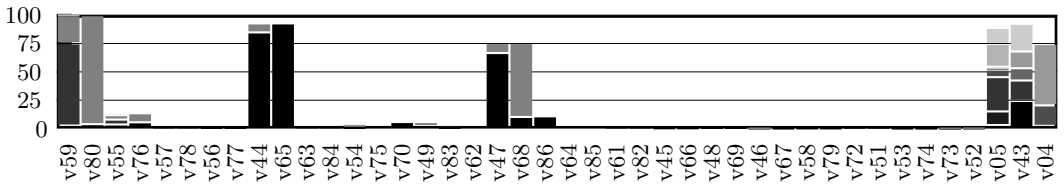
(a) $h_0 = s_0$



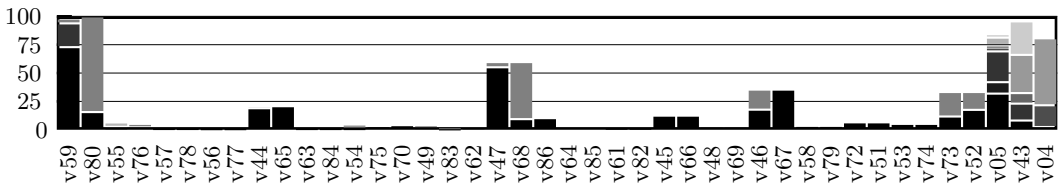
(b) $h_0 = s_1$



(c) $h_0 = s_2$



(d) $h_0 = s_3$



(e) $h_0 = s_4$

Figure 3 The class-conditional probability distributions of the 5 clusters pertaining to h_0. The sizes of the clusters are 0.46, 0.14, 0.13, 0.22, and 0.05, respectively.

distribution of that variable (in a cluster). Different segments in the bar correspond to different values of the variable. The darker the color is, the "higher" the value is. White color indicates the "lowest value" or "no". The white segment is always on top and indistinguishable from the background. Heights of the segments represent probabilities of the corresponding values.

The clusters pertaining to h_0 are meaningful. We see that $h_0{=}s_0$ is the only cluster with no fire insurance (v_{59}, v_{80}), while $h_0{=}s_4$ is the only cluster with non-zero probability of owning agriculture-related insurance (v_{46}, v_{67}, v_{52}, etc). Cluster $h_0{=}s_1$ stands out as the cluster with the lowest probability of owning car insurance (v_{47}, v_{68}). Clusters $h_0{=}s_2$ and $h_0{=}s_3$ have much higher probability of owning third-party private insurance (v_{44}, v_{65}) than the other clusters. Between these two clusters, the purchasing power (v_{43}) of cluster $h_0{=}s_2$ is significantly lower than that of cluster $h_0{=}s_3$ and, probably as a consequence, the former cluster also has much lower probability of owning car insurance (v_{47}, v_{68}) than the latter.

In contrast, latent class analysis resulted in 10 clusters. Overall, those clusters are less meaningful than the clusters pertaining to h_0. For example, there are 3, instead of 1, clusters with no fire insurance; and there are 5, instead of 2, clusters with high probability of owing third-party private insurance.

Figure 4 shows the clusters pertaining to h_8 (attitude toward private vehicle risks) and h_{12} (attitude toward heavy private vehicle risks). Those clusters are also meaningful. Among the h_8 clusters, $h_8{=}s_2$ is the only one with moped insurance (v_{54}, v_{75}). The other three h_8 clusters have different probabilities of owning car insurance (v_{47}, v_{68}). Among the h_{12} clusters, $h_{12}{=}s_0$ has no insurance on heavy private vehicles, while the other two clusters have. Cluster $h_{12}{=}s_2$ also has some probability of owning mobile home (v_{86}) insurance.

7 Probabilistic Modeling

We have so far mentioned two probabilistic models for the CoIL Challenge 2000 data, namely the HLC model M^* and the latent class model produced during latent class analysis. In this section, we will denote M^* as M_{HLC} and the latent class model as M_{LC}. For the sake of comparison, we have also used the greedy equivalence search algorithm[19] to obtain a Bayesian network model that does not contain latent variables. This model will be denoted as M_{GES}. The structure of M_{GES} is shown in Figure 5. In general, we refer to Bayesian networks that do not contain latent variables as observed BN models.

The structure of M_{HLC} is clearly more meaningful than those of M_{LC} and M_{GES}. The structure of M_{LC} is too simplistic to be informative. The relationships encoded in M_{GES} are not as interpretable as those encoded in M_{HLC}.

How well do the models fit the data? Before answering this question, we note that HLC models and observed BN models both have their pros and cons when it comes to represent interactions among manifest variables. The advantage of HLC models over observed BN models is that they can model high-order interactions. In M_{HLC}, latent variable h_{12} models some of the interactions among the heavy private vehicle variables; h_8 models some of the interactions among the private vehicle variables; while h_0 models some of the interactions among all manifest variables. On the other hand, observed BN models are better than HLC models in modeling details of variable interactions. In M_{GES}, the conditional probability distributions $P(v_{59}|v_{44})$ and $P(v_{67}|v_{59}, v_{44})$ contain all information about the interactions among the three variables v_{44}, v_{59}, and v_{67}.

As shown in Table 2, the logscore of M_{HLC} on training data is slightly higher than that of M_{GES}. On the other hand, M_{GES} is less complex than M_{HLC}, and its BIC score is higher than that of M_{HLC}. Here the complexity of a model is measured by the number of independent

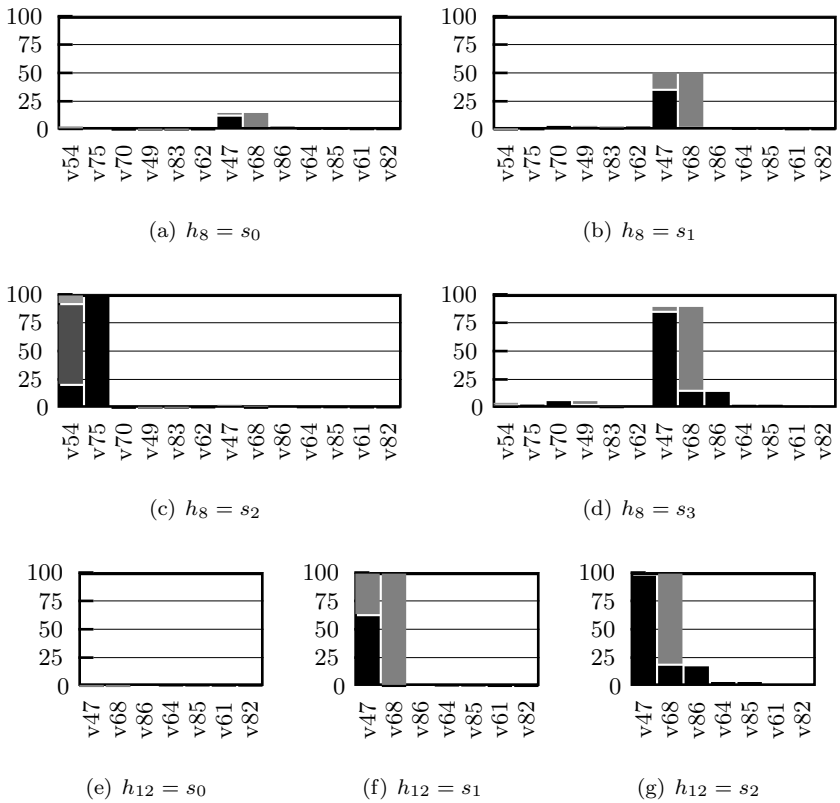


Figure 4 The clusters pertaining to h_8 and h_{12}

parameters. In CoIL Challenge 2000, there is a test set of 4,000 records. The logscore of M_{HLC} on the test data is higher than that of M_{GES} and the difference is larger than that on the training data. In other words, M_{HLC} is better than M_{GES} when it comes to predicting the test data. It is also clear that both M_{HLC} and M_{GES} significantly outperform M_{LC}.

Table 2

Model	Logscore	Complexity	BIC	Logscore (test data)
M_{LC}	−62328	739	−65532	−43248
M_{GES}	−49792	284	−51023	−34627
M_{HLC}	−49688	410	−51465	−34282

Because HLC models capture high-order variable interactions, M_{HLC} should perform better than M_{GES} in classification tasks. Out of the 4,000 customers in the CoIL Challenge 2000 test data, 238 own mobile home policies (v_{86}). The classification task is to identify a subset of 800 customers that contains as many mobile home policy owners as possible. As we can see from Table 3, M_{HLC} does perform significantly better than M_{GES}.

The classification performance of M_{HLC} ranks at Number 5 among the 43 entries to the CoIL Challenge 2000 contest, and it is not far from the performance of the best entry. This is impressive considering that no attempt was made to minimize classification error when learning M_{HLC}. In terms of model interpretability, M_{HLC} would rank Number 1 because all the other entries focus on classification accuracy rather than data modeling.

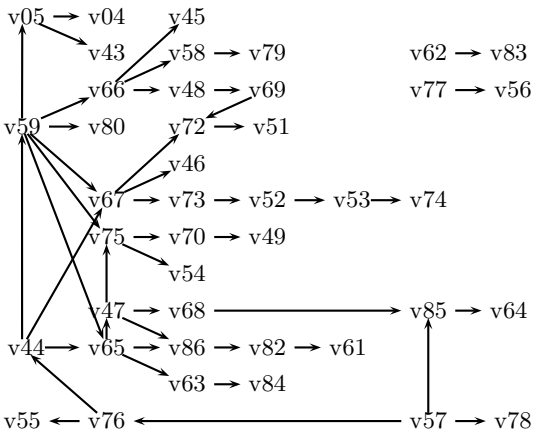


Figure 5 Bayesian network model without latent variables

Table 3

Model/Method	# of Mobile Home Policy Holders Identified	Hit Ratio
Random	42	17.6%
M_{GES}	83	34.9%
M_{LC}	105	44.1%
M_{HLC}	110	46.2%
CoIL 2000 Best	121	50.8%

In an HLC model, one can also compute probability distributions of latent variables. In M_{HLC}, one can calculate, for a given customer, the posterior distributions of h_8 (attitude toward private vehicle risks), h_1 (attitude toward firm risks), h_{15} (attitude toward agriculture risks), and so on. Collectively, those distributions can be used as a profile for the customer. Such profiling may have interesting applications.

8 Conclusions

Through the analysis of the CoIL Challenge 2000 data set, we have demonstrated that it is possible to infer complex and meaningful latent structures from data using HLC models. This indicates that HLC models are a viable tool for latent structure discovery, and calls for further study on HLC models and further explorations of their application potentials. One immediate future work is to relax the limitation identified in Section 5, namely, one manifest node cannot be connected to more than one latent node.

We have also demonstrated the usefulness of HLC models in cluster analysis and probabilistic modeling. As a tool for cluster analysis, they produce more meaningful clusters than latent class models and they allow multi-way clustering at the same time. As a tool for probabilistic modeling, they can model high-order interactions among variables and hence lead to better prediction and classification performance than observed BN models. They also facilitate unsupervised profiling.

References

[1] N. L. Zhang, Hierarchical latent class models for cluster analysis, in *Proceedings of the* 18*th National Conference on Artificial Intelligence*, AAAI Press, Menlo Park, 2002, 230–237.

[2] N. L. Zhang, Hierarchical latent class models for cluster analysis, *Journal of Machine Learning Research*, 2004, **5**(Jun): 697–723.

[3] P. F. Lazarsfeld and N. W Henry, *Latent Structure Analysis*, Houghton Mifflin, Boston, 1968.

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

[5] D. J. Bartholomew and M. Knott, *Latent Variable Models and Factor Analysis* (2nd edition), Arnold, London, 1999.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, 1998.

[7] G. Elidan, N. Lotner, N. Friedman, and D. Koller, Discovering hidden variables: a structure-based approach, in *Advances in Neural Information Processing Systems* 13, MIT Press, Cambridge, 2001, 479–485.

[8] R. Silva, R. Scheines, C. Glymour, and P. Spirtes, Learning measurement models for unobserved variables, in *Proceedings of the* 19*th Conference on Uncertainty in Artificial Intelligence*, 2003, 545–555.

[9] P. van der Putten and M. van Someren, A bias-variance analysis of a real world learning problem: the CoIL Challenge 2000, *Machine Learning*, 2004, **57**(1–2): 177–195.

[10] J. K. Vermunt and J. Magidson, Latent class cluster analysis, in *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, 2002, 89–106.

[11] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics*, 1978, **6**(2): 461–464.

[12] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 1974, **19**(6): 716–723.

[13] P. Cheeseman and J. Stutz, Bayesian classification (AutoClass): theory and results, in *Advances in Knowledge Discovery and Data Mining* (ed. by U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy), AAAI Press, Menlo Park, 1996.

[14] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Springer, New York, 1999.

[15] D. Geiger, D. Heckerman, and C.Meek, Asymptotic model selection for directed networks with hidden variables, in *Proceedings of the* 12*th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Fransisco, 1996, 283–290.

[16] T. Kočka and N. L. Zhang, Dimension correction for hierarchical latent class models, in *Proceedings of the* 18*th Conference on Uncertainty in Artificial Intelligence* (ed. by A. Darwiche and N. Friedman), Morgan Kaufmann Publishers, San Fransisco, 2002, 267–274.

[17] N. L. Zhang and T. Kočka, Efficient learning of hierarchical latent class models, in *Proceedings of the* 16*th IEEE International Conference on Tools with Artificial Intelligence*, IEEE Computer Society, Los Alamitos, CA, 2004, 585–593.

[18] N. Friedman, Learning belief networks in the presence of missing values and hidden variables, in *Proceedings of the* 14*th International Conference on Machine Learning*, Morgan Kaufmann Publishers, San Fransisco, 1997, 125–133.

[19] D. M. Chickering, Learning equivalence classes of Bayesian-network structures, *Journal of Machine Learning Research*, 2002, **2**(Feb): 445–498.