Severity of Local Maxima for the EM Algorithm: Experiences with Hierarchical Latent Class Models

Yi Wang and Nevin L. Zhang
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong, China
{wangyi,lzhang}@cse.ust.hk

Abstract

It is common knowledge that the EM algorithm can be trapped at local maxima and consequently fails to reach global maxima. We empirically investigate the severity of this problem in the context of hierarchical latent class (HLC) models. Our experiments were run on HLC models where dependency between neighboring variables is strong. (The reason for focusing on this class of models will be made clear in the main text.) We first ran EM from randomly generated single starting points, and observed that (1) the probability of hitting global maxima is generally high, (2) it increases with the strength of dependency and sample sizes, and (3) it decreases with the amount of extreme probability values. We also observed that, at high dependence strength levels, local maxima are far apart from global ones in terms of likelihoods. Those imply that local maxima can be reliably avoided by running EM from a few starting points and hence are not a serious issue. This is confirmed by our second set of experiments.

1 Introduction

The EM algorithm (Dempster et al., 1977) is a popular method for approximating maximum likelihood estimate in the case of incomplete data. It is widely used for parameter learning in models, such as mixture models (Everitt and Hand, 1981) and latent class models (Lazarsfeld and Henry, 1968), that contain latent variables.

A well known problem associated with EM is that it can be trapped at local maxima and consequently fails to reach global maxima (Wu, 1983). One simple way to alleviate the problem is to run EM many times from randomly generated starting points, and take the highest likelihood obtained as the global maximum. One problem with this method is that it is computationally expensive when the number of starting points is large because EM converges slowly. For this reason, researchers usually adopt the *multiple restart* strategy (Chickering and Heckerman, 1997; van de Pol et al., 1998; Uebersax, 2000; Vermunt and Magidson, 2000): First run EM from multiple random starting points for a number of steps, then pick the one with the highest likelihood, and continue EM from the picked point until convergence. In addition to multiple restart EM, several more sophisticated strategies for avoiding local maxima have also been proposed (Fayyad et al., 1998; Ueda and Nakano, 1998; Ueda et al., 2000; Elidan et al., 2002; Karciauskas et al., 2004).

While there is abundant work on avoiding local maxima, we are aware of few work on the severity of the local maxima issue. In this paper, we empirically investigate the severity of local maxima for hierarchical latent class (HLC) models. Our experiments were run on HLC models where dependency between neighboring variables is strong. This class of models was chosen because we use HLC models to discover latent structures. It is our philosophical view that one cannot expect to discover latent structures reliably unless observed variables strongly depend on latent variables (Zhang, 2004; Zhang and Kocka, 2004).

In the first set of experiments, we ran EM from randomly generated single starting points, and observed that (1) the probability of hitting global maxima is generally high, (2) it increases with the strength of dependency and sample sizes, and (3) it decreases with the amount of extreme probability values. We also observed that, at high dependence strength levels, local maxima are far apart from global ones in terms of likelihoods.

Those observations have immediate practical implications. Earlier in this section, we mentioned a simple local-maxima avoidance method. We pointed out one of its problems, i.e. its high computational complexity, and said that multiple restart can alleviate this problem. There is another problem with the method: there is no guidance on how many starting points to use in order to avoid local maxima reliably. Multiple restart provides no solution for this problem.

Observations from our experiments suggest a guideline for strong dependence HLC models. As a matter of fact, they imply that local maxima can be reliably avoided by using multiple restart and a few starting points. This is confirmed by our second set of experiments.

The remainder of this paper is organized as follows. In Section 2, we review some basic concepts about HLC models and the EM algorithm. In Sections 3 and 4, we report our first and second sets of experiments respectively. We conclude this paper and point out some potential future work in Section 5.

2 Background

2.1 Hierarchical Latent Class Models

Hierarchical latent class (HLC) models (Zhang, 2004) are tree-structured Bayesian networks where the leaf nodes are observed while the internal nodes are hidden. An example HLC model is shown in Figure 1. Following the conventions in latent variable model literatures, we call the leaf nodes *manifest variables* and the internal nodes *latent variables*[1].

[1]In this paper, we do not distinguish between nodes and variables.

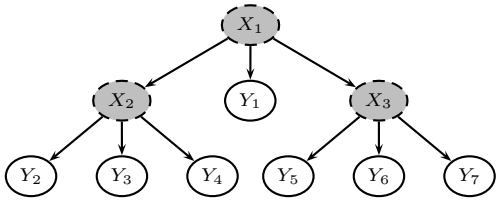


Figure 1: An example HLC model. X_1, X_2, X_3 are latent variables, and Y_1, Y_2, $\cdots$, Y_7 are manifest variables.

We usually write an HLC model as a pair $M{=}(m, \boldsymbol{\theta})$. The first component m consists of the model structure and cardinalities of the variables. The second component $\boldsymbol{\theta}$ is the collection of parameters. Two HLC models $M{=}(m, \boldsymbol{\theta})$ and $M'{=}(m', \boldsymbol{\theta}')$ are *marginally equivalent* if they share the same set of manifest variables $\mathbf{Y}$ and $P(\mathbf{Y}|m, \boldsymbol{\theta}){=}P(\mathbf{Y}|m', \boldsymbol{\theta}')$.

We denote the cardinality of a variable X by $|X|$. For a latent variable X in an HLC model, denote the set of its neighbors by $\mathbf{nb}(X)$. An HLC model is *regular* if for any latent variable X, $|X| \leq \frac{\prod_{Z \in \mathbf{nb}(X)} |Z|}{\max_{Z \in \mathbf{nb}(X)} |Z|}$, and the inequality strictly holds when X has only two neighbors, one of which being a latent variable. Zhang (2004) has shown that an irregular HLC model can always be reduced to a marginally equivalent HLC model that is regular and contains fewer independent parameters. Henceforth, our discussions are restricted to regular HLC models.

2.2 Strong Dependence HLC Models

The strength of dependency between two variables is usually measured by mutual information or correlation (Cover and Thomas, 1991). However, there is no general definition of strong dependency for HLC models yet. In this study, we use the operational definition described in the following paragraph.

Consider a probability distribution. We call the component with the largest probability mass the *major component*. We say that an HLC model is a *strong dependence model* if

- The cardinality of each node is no smaller than that of its parent.
- The major components in all conditional distributions are larger than 0.5, and

- In each conditional probability table, the major components of different rows are located in different columns.

In general, the larger the major components, the higher the dependence strength (DS) level. In the extreme case, when all major components are equal to 1, the HLC model becomes deterministic. Strong dependence HLC models defined in this way have been examined in our previous work on discovering latent structures (Zhang, 2004; Zhang and Kocka, 2004). The results show that such models can be reliably recovered from data.

2.3 EM Algorithm

Latent variables can never be observed and their values are always missing in data. This fact complicates the maximum likelihood estimation problem, since we cannot compute sufficient statistics from incomplete data records. A common method to deal with such situations is to use the *expectation-maximization (EM) algorithm* (Dempster et al., 1977). The EM algorithm starts with a randomly chosen estimation to parameters, and iteratively improves this estimation by increasing its loglikelihood. In each EM step, the task of increasing loglikelihood is delegated to the maximization of the *expected loglikelihood function*. The latter is a lower bound of the loglikelihood function. It is defined as

$$\begin{aligned} & Q(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\theta}^t) \\ = & \sum_{l=1}^{m}\sum_{\mathbf{X}_l} P(\mathbf{X}_l|\mathbf{D}_l,\boldsymbol{\theta}^t)\log P(\mathbf{X}_l,\mathbf{D}_l|\boldsymbol{\theta}), \end{aligned}$$

where $\mathcal{D}=\{\mathbf{D}_1,\mathbf{D}_2,\cdots,\mathbf{D}_m\}$ denotes the colletion of data, $\mathbf{X}_l$ denotes the set of variables whose values are missing in $\mathbf{D}_l$, and $\boldsymbol{\theta}^t$ denotes the estimation at the t-th step. The EM algorithm terminates when the increase in loglikelihood between two successive steps is smaller than a predefined stopping threshold.

It is common knowledge that the EM algorithm can be trapped at local maxima and consequently fails to reach global maxima (Wu, 1983). The specific result depends on the choice of the starting point. A common method to avoid local maxima is to run EM many times with randomly generated starting points, and pick the instance with the highest likelihood as the final result. The more starting points it uses, the higher the chance it can reach the global maximum. However, due to the slow convergence of EM, this method is computationally expensive. A more feasible method, called *multiple restart EM*, is to run EM with multiple random starting points, and retain the one with the highest likelihood after a specified number of initial steps. This method and its variant are commonly used to learn latent variable models in practice (Chickering and Heckerman, 1997; van de Pol et al., 1998; Uebersax, 2000; Vermunt and Magidson, 2000). Other work on escaping from poor local maxima includes (Fayyad et al., 1998; Ueda and Nakano, 1998; Ueda et al., 2000; Elidan et al., 2002; Karciauskas et al., 2004).

3 Severity of Local Maxima

Here is the strategy that we adopt to empirically investigate the severity of local maxima in strong dependence HLC models: (1) create a set of strong dependence models, (2) sample some data from each of the models, (3) learn model parameters from the data by running EM to convergence from a number of starting points, (4) graph the final loglikelihoods obtained.

The final loglikelihoods for different starting points could be different due to local maxima. Hence, an inspection of their distribution would give us a good idea about the severity of local maxima.

3.1 Experiment Setup

The structure of all models used in our experiments was the ternary tree with height equals 3, as shown in Figure 2. The cardinalities of all variables were set at 3. Parameters were randomly generated subject to the strong dependency condition. We examined 5 DS levels, labeled from 1 to 5. They correspond to restricting the major components within the following 5 intervals: $[0.5, 0.6)$, $[0.6, 0.7)$,$[0.7, 0.8)$,$[0.8, 0.9)$, and $[0.9, 1.0]$. For each DS level, 5 different parameterizations were generated. Consequently,

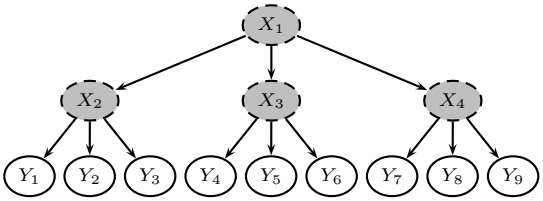


Figure 2: The structure of generative models.

we examined 25 models in total.

From each of the 25 models, four training sets of 100, 500, 1000, and 5000 records were sampled. On each training set, we ran EM independently for 100 times, each time starting from a randomly selected initial point. The stopping threshold of EM was set at 0.0001. The highest loglikelihood obtained in the 100 runs is regarded as the global maximum.

3.2 Results

The results are summarized in Figures 3 and 4. In Figure 3, there is a plot for each combination of DS level (row) and sample size (column). As mentioned earlier, 5 models were created for each DS level. The plot is for one of those 5 models. In the plot, there are three curves: solid, dashed, and dotted[2]. The dashed and dotted curves are for Section 4. The solid curve is for this section. The curve depicts a distribution function $F(x)$, where x is loglikelihood and $F(x)$ is the percent of EM runs that achieved a loglikelihood no larger than x.

While a plot in Figure 3 is about one model at a DS level, a plot in Figure 4 represents an aggregation of results about all 5 models at a DS level. Loglikelihoods for different models can be in very different ranges. To aggregate them in a meaningful way, we introduce the concept of relative likelihood shortfall of EM run.

Consider a particular model. We have run EM 100 times and hence obtained 100 loglikelihoods. The maximum is regarded as the optimum and is denoted by l^*. Suppose a particular EM run resulted in loglikelihood l. Then the *relative likelihood shortfall* of that EM run is defined as $(l-l^*)/l^*$. This value is nonnegative. The smaller it is, the higher the quality of the parameters produced by the EM run. In particular, the relative likelihood shortfall of the run that produced the global maximum l^* is 0.

For a given DS level, there are 5 models and hence 500 EM runs. We put the relative likelihood shortfalls of all those EM runs into one set and let, for any nonnegative real number x, $F(x)$ be the percent of the elements in the set that is no larger than x. We call $F(x)$ the *distribution function of (aggregated) relative likelihood shortfalls of EM runs*, or simply *distribution of EM relative likelihood shortfall*, for the DS level.

The first 5 plots in Figure 4 depict the distributions of EM relative likelihood shortfall for the 5 DS levels. There are four curves in each plot, each corresponding to a sample size[3].

3.2.1 Probability of Hitting Global Maxima

The most interesting question is how often EM hits global maxima. To answer this question, we first look at the solid curves in Figure 3. Most of them are stair-shaped. In each curve, the x-position of the right most stair is the global maximum, and the height of that stair is the frequency of hitting the global maximum.

We see that the frequency of hitting global maxima was generally high for high DS levels. In particular, for DS level 3 or above, EM reached global maxima more than half of the time. For sample size 500 or above, the frequency was even greater than 0.7. On the other hand, the frequency was low for DS level 1, especially when the sample size was small.

The first 5 plots in Figure 4 tell the same story. In those plots, the global maxima are represented by x=0. The heights of the curves at x=0 are the frequencies of hitting global maxima. We see that for DS level 3 or above, the frequency of hitting global maxima is larger than 0.5, except that for DS level 3 and sample size 100. And once again, the frequency was low for DS level 1.

[2]Note that in some plot (e.g., that for DS level 4 and sample size 5000) the curves overlap and are indistinguishable.

[3]Note that in Figure 4 (b) and (c) some curves are close to the y-axes and are hardly distinguishable.

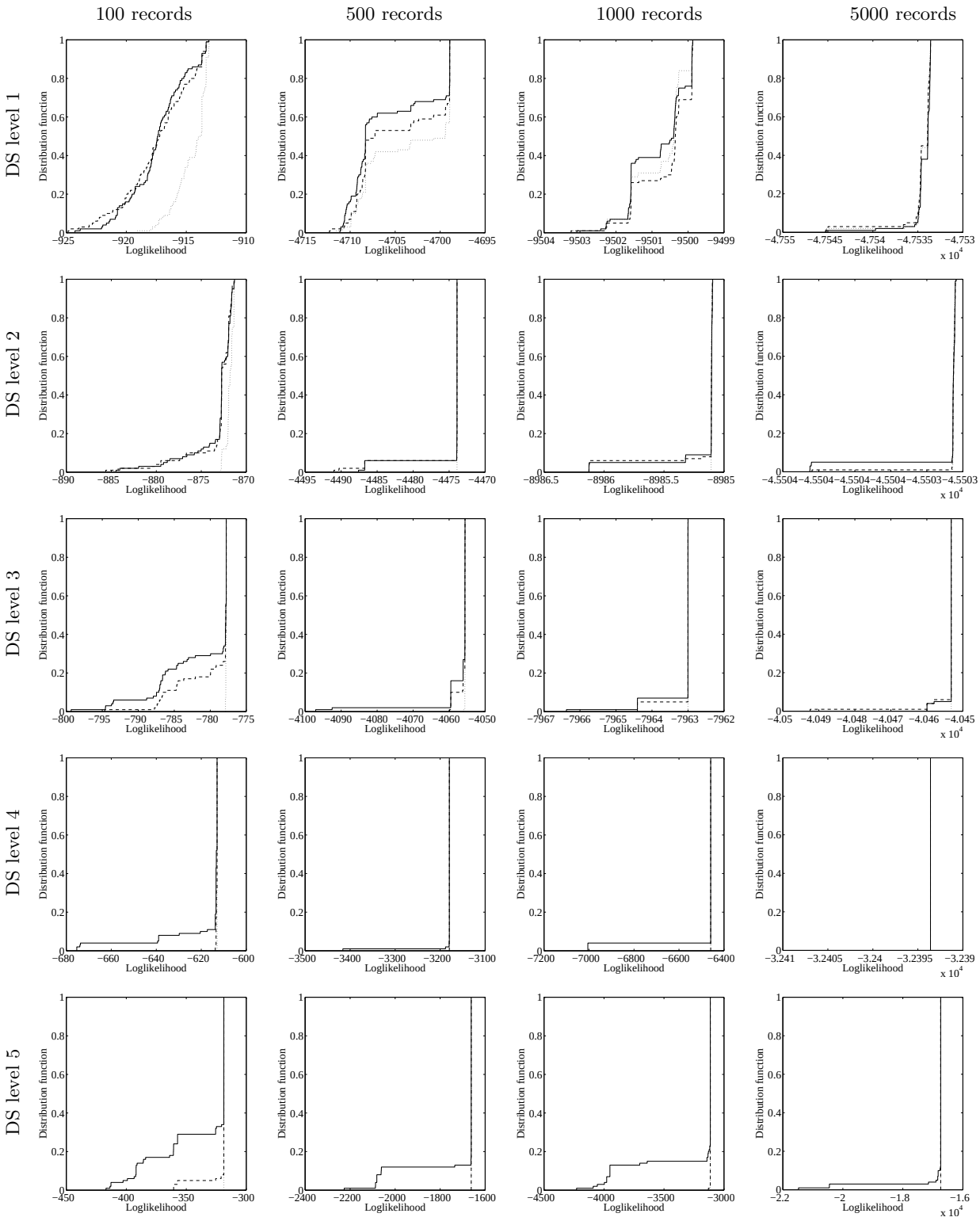
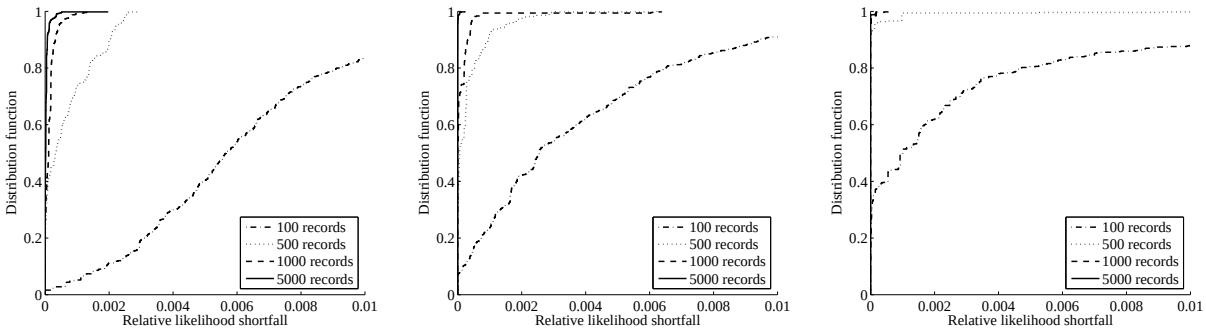
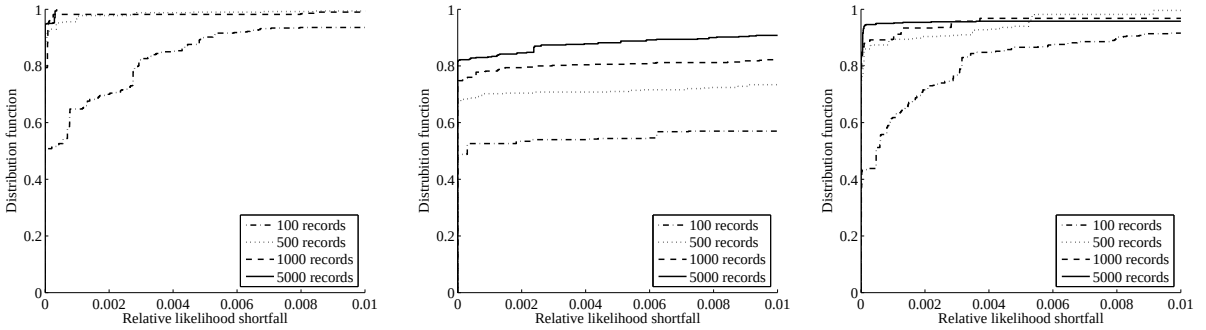


Figure 3: Distributions of loglikelihoods obtained by different EM runs. Solid curves are for EM runs with single starting points. Dashed and dotted curves are for runs of multiple restart EM with setting 4×10 and 16×50 respectively (see Section 4). Each curve depicts a distribution function $F(x)$, where x is loglikelihood and $F(x)$ is the percent of runs that achieved a loglikelihood no larger than x.



(a) DS level 1 (b) DS level 2 (c) DS level 3



(d) DS level 4 (e) DS level 5 (f) DS level 3 with 15% zero components

Figure 4: Distributions of EM relative likelihood shortfall for different DS levels.

3.2.2 DS Level and Local Maxima

To see how DS level influences the severity of local maxima, read each column of Figure 3 (for a given sample size) from top to bottom. Notice that, in general, the height of the right most stair increases with the DS level. It means that the frequency of hitting global maxima increases with DS level. The same conclusion can be read off from Figure 4. Compare the curves for a given sample size from the first 5 plots. In general, their values at x=0 rise up as we move from DS level 1 to DS level 5.

3.2.3 Extreme Parameter Values and Local Maxima

There are some exceptions to the regularities mentioned in the previous paragraph. We see in both figures that the frequency of hitting global maxima decreases as the DS level changes from 4 to 5. By comparing Figure 4 (c) and (d), we also find that the frequency slightly drops when the DS level changes from 3 to 4, given that the sample size is large.

We conjecture that this is because that, in generative models for DS levels 4 and 5, there are many parameter values close to 0. It has been reported in latent variable model literatures that extreme parameter values cause local maxima (Uebersax, 2000; McCutcheon, 2002).

To confirm our conjecture, we created another set of five models for DS level 3. The parameters were generated in the same way as before, except that we randomly set 15% of the non-major components in the probability distributions to 0. We then repeated the experiments for this set of models. The EM relative likelihood shortfall distributions are given in Figure 4 (f). We see that, as expected, the frequency of hitting global maxima did decrease considerably compared with Figure 4 (c).

3.2.4 Sample Size and Local Maxima

We also noticed the power of the sample size. By going through each row of Figure 3, we found

that the frequency of hitting global maxima increases with the sample size. The phenomenon is more apparent if we look at Figure 4, where curves for different sample sizes are plotted in the same picture. As the sample size goes up, the curve becomes steeper and its value at x=0 increases significantly.

3.2.5 Quality of Local Maxima

In addition to the frequency of hitting global maxima, another interesting question is how bad local maxima could be. Let us first examine the local maxima in Figure 3. They are marked by the x-positions of the inflection points on curves. For DS level 1, the local maxima are not far from the global ones. The discrepancies are less than 15. Moreover, there are a lot inflection points on the curves, namely, distinct local maximal solutions. They distributed evenly within the interval between the worst local maxima and the global ones.

For the other DS levels, things are different. We first observed that the quality of local maxima can be extremely poor in some situation. The worst case is that of DS level 5 with sample size 5000. The loglikelihood of the local maximum can be lower than that of the global one by more than 4000. The second thing we observed is that the curves contain much fewer inflection points. In other words, there are only a few distinct local maximal solutions in such cases. Moreover, those local maxima stay far apart from global ones since the steps of the staircases are fairly large. Those observations can be confirmed by studying the first 5 plots in Figure 4, where the curves and the inflection points can be interpreted similarly.

4 Performance of Multiple Restart EM

We emphasize two observations mentioned in the previous section: (1) the probability of hitting global maxima is generally high, and (2) likelihoods of local maxima are far apart from those of global maxima at high DS levels. We will see that these observations have immediate implications on the performance of multiple restart EM.

We say that a starting point is *optimal* if it converges to the global maximum. The first observation can be restated as follows: the probability for a randomly generated starting point to be optimal is generally high. Consequently, it is almost sure that there is an optimal starting point within a few randomly generated ones.

As it is well known, the EM algorithm increases the likelihood quickly in its early stage and slows down when it is converging. In other words, the likelihood should become relatively stable after a few steps. Therefore, the second observation implies that we can easily separate the optimal starting point from the others after running a few EM steps on them.

A straightforward consequence of the above inference is that multiple restart EM with a few starting points and initial steps should reliably avoid local maxima for strong dependence HLC models. To confirm this conjecture, we ran multiple restart EM independently for 100 times on each training set that was presented in Figure 3. We tested two settings for multiple restart EM: (1) 4 random starting points with 10 initial steps (in short, 4×10), and (2) 16 random starting points with 50 initial steps (in short, 16×50). As before, we plotted the distributions of loglikelihoods in Figure 3. The dashed and the dotted curves denote the results for settings of 4×10 and 16×50, respectively.

From Figure 3, we see that multiple restart EM with setting 16×50 can reliably avoid local maxima for DS level 2 or above. Actually, the dotted curves are parallel to the y-axes except that for DS level 2 and sample size 100. It means that global maxima can always be reached in such cases. For setting 4×10, similar behaviors are observed for DS level 4 or 5 and sample size 500 or above. Note that dashed and dotted curves overlap in those plots.

Nonetheless, we also notice that, for DS level 1, multiple restart EM with both settings still can not find global maxima reliably. This is consistent with our reasoning. As we have mentioned in Section 3.2.1, when we ran EM with randomly generated single starting points, the frequency of hitting global maxima is low for DS level 1. In other words, it is hard to hit an op-

timal starting point by chance. Moreover, due to the small discrepancy among local maxima (see Section 3.2.5), it demands a large number of initial steps to distinguish an optimal starting point from the others. Therefore, the effectiveness of multiple restart EM degenerates. In such situations, one can either increase the number of starting points and initial steps, or appeal to more sophisticated methods for avoiding local maxima.

5 Conclusions

We have empirically investigated the severity of local maxima for EM in the context of strong dependence HLC models. We have observed that (1) the probability of hitting global maxima is generally high, (2) it increases with the strength of dependency and sample sizes, (3) it decreases with the amount of extreme probability values, and (4) likelihoods of local maxima are far apart from those of global maxima at high dependence strength levels. We have also empirically shown that the local maxima can be reliably avoided by using multiple restart EM with a few starting points and hence are not a serious issue.

Our discussion has been restricted to a specific class of HLC models. In particular, we have defined the strong dependency in an operational way. One can devise more formal definitions and carry on similar studies for generalized strong dependence models. Another future work would be the theoretical exploration to support our experiences.

Based on our observations, we have analyzed the performance of multiple restart EM. One can exploit those observations to analyze more sophisticated strategies for avoiding local maxima. We believe that those observations can also give some inspirations to develop new methods on this direction.

Acknowledgments

Research on this work was supported by Hong Kong Grants Council Grant #622105.

References

D. M. Chickering and D. Heckerman. 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, 29:181–212.

T. M. Cover and J. A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons, Inc.

A. P. Dempster, N. M. Laird, and D. R. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

G. Elidan, M. Ninio, N. Friedman, and D. Schuurmans. 2002. Data perturbation for escaping local maxima in learning. In *AAAI-02*, pages 132–139.

B. S. Everitt and D. J. Hand. 1981. *Finite Mixture Distributions*. Chapman and Hall.

U. M. Fayyad, C. A. Reina, and P. S. Bradley. 1998. Initialization of iterative refinement clustering algorithms. In *KDD-98*, pages 194–198.

G. Karciauskas, T. Kocka, F. V. Jensen, P. Larranaga, and J. A. Lozano. 2004. Learning of latent class models by splitting and merging components. In *PGM-04*.

P. F. Lazarsfeld and N. W. Henry. 1968. *Latent Structure Analysis*. Houghton Mifflin.

A. L. McCutcheon. 2002. Basic concepts and procedures in single- and multiple-group latent class analysis. In *Applied Latent Class Analysis*, chapter 2, pages 56–85. Cambridge University Press.

J. S. Uebersax. 2000. A brief study of local maximum solutions in latent class analysis. http://ourworld.compuserve.com/homepages/jsuebersax/local.htm.

N. Ueda and R. Nakano. 1998. Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–282.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. 2000. SMEM algorithm for mixture models. *Neural Computation*, 12(9):2109–2128.

F. van de Pol, R. Langeheine, and W. de Jong, 1998. *PANMARK user manual, version 3*. Netherlands Central Bureau of Statistics.

J. K. Vermunt and J. Magidson, 2000. *Latent GOLD User's Guide*. Statistical Innovations Inc.

C. F. J. Wu. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.

N. L. Zhang and T. Kocka. 2004. Efficient learning of hierarchical latent class models. In *ICTAI-04*, pages 585–593.

N. L. Zhang. 2004. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5(6):697–723.