Latent Class Analysis Variable Selection

Nema Dean*and Adrian E. Raftery †

July 18, 2008

Abstract

We propose a method for selecting variables in latent class analysis, which is the most common model-based clustering method for discrete data. The method assesses a variable's usefulness for clustering by comparing two models, given the clustering variables already selected. In one model the variable contributes information about cluster allocation beyond that contained in the already selected variables, and in the other model it does not. A headlong search algorithm is used to explore the model space and select clustering variables. In simulated datasets we found that the method selected the correct clustering variables, and also led to improvements in classification performance and in accuracy of the choice of the number of classes. In two real datasets, our method discovered the same group structure with fewer variables. In a dataset from the International HapMap Project consisting of 639 single nucleotide polymorphisms (SNPs) from 210 members of different groups, our method discovered the same group structure with a much smaller number of SNPs.

Keywords: Bayes factor, BIC, Categorical data, Feature Selection, Model-based clustering, Single nucleotide polymorphism (SNP).

1 Introduction

Latent class analysis is used to discover groupings in multivariate categorical data. It models the data as a finite mixture of distributions, each one corresponding to a class (or cluster or group). Because of the underlying statistical model it is possible to determine the number of classes using model selection methods. But the modeling framework does not currently address the selection of the variables to be used; typically all variables are used in the model.

Selecting variables for latent class analysis can be desirable for several reasons. It can help interpretability of the model, and it can also make it possible to fit a model with a larger number of classes than would be possible with all the variables, for identifiability reasons.

*Nema Dean, Department of Statistics, University of Glasgow, Glasgow G12 8QQ, Scotland.

†Adrian E. Raftery (corresponding author), Department of Statistics, University of Washington, Box 354320, Seattle, WA 98195-4320, USA. e-mail: raftery@u.washington.edu.

In general, removing unnecessary variables and parameters can also improve classification performance and the precision of parameter estimates.

In this paper we propose a method for selecting the variables to be used for clustering in latent class analysis. This is based on the method of Raftery and Dean (2006) for variable selection selection in model-based clustering of continuous variables. The method assesses a variables usefulness for clustering by comparing two models, given the clustering variables already selected. In one model the variable contributes information about cluster allocation beyond that contained in the already selected variables, and in the other model it does not. We then present a new search algorithm, based on Badsberg (1992), for exploring the space of possible models. The resulting method selects both the variables and the number of classes in the model.

In Section 2 we review some aspects of latent class analysis and in Section 3 we describe our variable selection methodology. In Section 4 we give results from simulated data and in Section 5 we give results for two real datasets, including one with a large number of variables and a much smaller number of data points. Issues arising with the method are discussed in Section 6.

2 Latent Class Analysis

2.1 Latent Class Analysis Model

Latent class analysis was proposed by Lazarsfeld (1950a), Lazarsfeld (1950b) and Lazarsfeld and Henry (1968) and can be viewed as a special case of model-based clustering, for multivariate discrete data. Model-based clustering assumes that each observation comes from one of a number of classes, groups or subpopulations, and models each with its own probability distribution (Wolfe 1963; McLachlan and Peel 2000; Fraley and Raftery 2002). The overall population thus follows a finite mixture model, namely

$$x \sim \sum_{g=1}^{G} \pi_g f_g(x),$$

where f_g is the density for group g, G is the number of groups, $0 < \pi_g < 1$, $\forall g$ and $\sum_{g=1}^{G} \pi_g = 1$. Often, in practice, the f_g are from the same parametric family (as is the case in latent class analysis) and we can write the overall density as:

$$x \sim \sum_{g=1}^{G} \pi_g f(x \mid \theta_g)$$

where θ_g is the set of parameters for the g^{th} group.

In latent class analysis, the variables are usually assumed to be independent given knowledge of the group an observation came from, an assumption called *local independence*. Each variable within each group is then modeled with a multinomial density. The general density of a single variable x (with categories $1, \ldots, d$) given that it is in group g is then

$$x \mid g \sim \prod_{j=1}^{d} p_{jg}^{1\{x=j\}},$$

where $1\{x = j\}$ is the indicator function equal to 1 if the observation of the variable takes value j and 0 otherwise, p_{jg} is the probability of the variable taking value j in group g, and d is the number of possible values or categories the variable can take.

Since we are assuming conditional independence, if we have k variables, their joint group density can be written as a product of their individual group densities. If we have $x = (x_1, \ldots, x_k)$, we can write the joint group density as:

$$x \mid g \sim \prod_{i=1}^{k} \prod_{j=1}^{d_i} p_{ijg}^{1\{x_i=j\}},$$

where $1\{x_i = j\}$ is the indicator function equal to 1 if the observation of the i^{th} variable takes value j and 0 otherwise, p_{ijg} is the probability of variable i taking value j in group g and d_i is the number of possible values or categories the i^{th} variable can take. The overall density is then a weighted sum of these individual product densities, namely

$$x \sim \sum_{g=1}^{G} (\pi_g \prod_{i=1}^{k} \prod_{j=1}^{d_i} p_{ijg}^{1\{x_i=j\}}),$$

where $0 < \pi_g < 1$, $\forall g$ and $\sum_{g=1}^{G} \pi_g = 1$.

The model parameters $\{p_{ijg}, \pi_g; i = 1, \ldots, k, j = 1, \ldots, d_i, g = 1, \ldots, G\}$ can be estimated from the data (for a fixed value of G) by maximum likelihood using the EM algorithm or the Newton-Rhapson algorithm or a hybrid of the two. These algorithms require starting values which are usually randomly generated. Because the algorithms are not guaranteed to find a global maximum and are usually fairly dependent on good starting values, it is routine to generate a number of random starting values and use the best solution given by one of these. In appendix B, we present an adjusted method useful for the cases where an inordinately large number of starting values is needed to get good estimates of the latent class models and $G > 2$.

Goodman (1974) discussed the issue of checking whether a latent class model with a certain number of classes was identifiable for a given number of variables. A necessary condition for identifiability when there are G classes and k variables with numbers of categories

$d = (d_1, \ldots, d_k)$ for G classes is

$$\prod_{i=1}^{k} d_i > (\sum_{i=1}^{k} d_i - k + 1) \times G,$$

This basically amounts to checking that there are enough pieces of information (or cell counts or pattern combinations) to estimate the number of parameters in the model. However, in practice, not all possible pattern combinations are observed (some or many cell counts may be zero) and so the actual information available may be less. When selecting the number of latent classes in the data, we consider only numbers of classes for which this necessary condition is satisfied.

For reviews of latent class analysis, see Clogg (1981), McCutcheon (1987), Clogg (1995) and Hagenaars and McCutcheon (2002).

2.2 Selecting the number of latent classes

Each different value of G, the number of latent classes, defines a different model for the data. A method is needed to select the number of latent classes present in the data. Since a statistical model for the data is used, model selection techniques can be applied to this question.

In order to choose the best number of classes for the data we need to choose the best model (and the related number of classes). Bayes factors (Kass and Raftery 1995) are used to compare these models.

The Bayes factor for comparing model M_i versus model M_j is equal to the ratio of the posterior odds for M_i versus M_j to the prior odds for M_i versus M_j. This reduces to the posterior odds when the prior model probabilities are equal. The general form for the Bayes factor is:

$$B_{ij} = \frac{p(Y \mid M_i)}{p(Y \mid M_j)},$$

where $p(Y \mid M_i)$ is known as the integrated likelihood of model M_i (given data Y). It is called the integrated likelihood because it is obtained by integrating over all the model parameters, namely the mixture proportions and the group variable probabilities. Unfortunately the integrated likelihood is difficult to compute (it has no closed form) and some form of approximation is needed for calculating Bayes factors in practice.

In our approximation we use the Bayesian information criterion (BIC) which is very simple to compute. The BIC is defined by

$$BIC = 2 \times \log(\text{maximized likelihood}) - (\text{no. of parameters}) \times \log(n), \qquad (1)$$

where n is the number of observations.

Twice the logarithm of the Bayes factor is approximately equal to the difference between the BIC values for the two models being compared. We choose the number of latent classes by recognizing that each different number of classes defines a model, which can then be compared to others using BIC. Keribin (1998) showed BIC to be consistent for the choice of the number of components in a mixture model under certain conditions, when all variables are relevant to the grouping. A rule of thumb for differences in BIC values is that a difference of less than 2 is viewed as barely worth mentioning, while a difference greater than 10 is seen as constituting strong evidence (Kass and Raftery 1995).

3 Variable Selection in Latent Class Analysis

3.1 Variable Selection Method

At any stage in the procedure we can partition the collection of variables into three sets: $Y^{(clust)}, Y^{(?)}$ and $Y^{(other)}$, where:

- $Y^{(clust)}$ is the set of variables already selected as useful for clustering,
- $Y^{(?)}$ is the variable(s) being considered for inclusion into/exclusion from Y^{clust},
- $Y^{(other)}$ is the set of all other variables.

Given this partition and the (unknown) clustering memberships z we can recast the question of the usefulness of $Y^{(?)}$ for clustering as a model selection question. The question becomes one of choosing between two different models, M_1 which assumes that $Y^{(?)}$ is not useful for clustering, and M_2 which assumes that it is.

The two models are specified as follows:

$$\begin{aligned} M_1 : p(Y|\mathbf{z}) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)}|\mathbf{z}) \\ &= p(Y^{(other)}|Y^{(?)}, Y^{(clust)})p(Y^{(?)})p(Y^{(clust)}|\mathbf{z}) \qquad (2) \\ M_2 : p(Y|\mathbf{z}) &= p(Y^{(clust)}, Y^{(?)}, Y^{(other)}|\mathbf{z}) \\ &= p(Y^{(other)}|Y^{(?)}, Y^{(clust)})p(Y^{(?)}, Y^{(clust)}|\mathbf{z}), \\ &= p(Y^{(other)}|Y^{(?)}, Y^{(clust)})p(Y^{(?)}|\mathbf{z})p(Y^{(clust)}|\mathbf{z}), \end{aligned}$$

where $\mathbf{z}$ is the (unobserved) set of cluster memberships. Model M_1 specifies that, given $Y^{(clust)}$, $Y^{(?)}$ is independent of the cluster memberships (defined by the unobserved variables $\mathbf{z}$), that is, $Y^{(?)}$ gives no further information about the clustering. Model M_2 implies that $Y^{(?)}$ does provide information about clustering membership, beyond that given just by $Y^{(clust)}$.

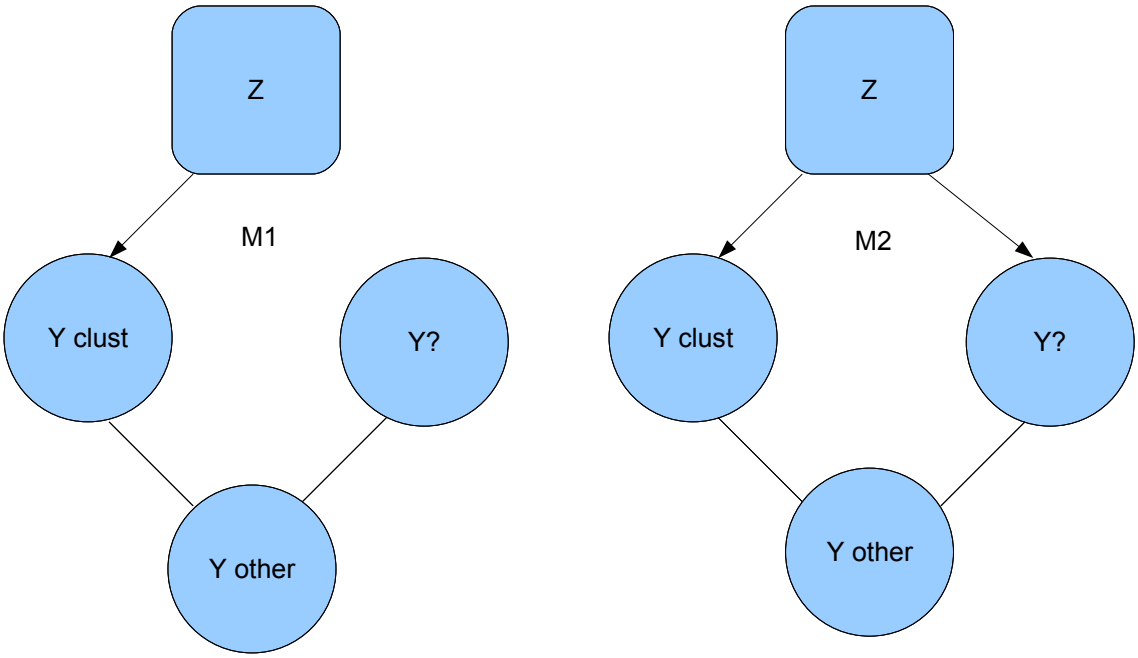

Figure 1: Graphical Representation of Models M_1 and M_2 for Latent Class Variable Selection. In model M_1, the candidate set of additional clustering variables, $Y^{(?)}$, is independent of the cluster memberships, $\mathbf{z}$, given the variables $Y^{(clust)}$ already in the model. In model M_2, this is not the case. In both models, the set of other variables considered, $Y^{(other)}$, is conditionally independent of cluster membership given $Y^{(clust)}$ and $Y^{(?)}$, but may be associated with $Y^{(clust)}$ and $Y^{(?)}$.

The difference between the assumptions underlying the two models is illustrated in Figure 1, where arrows indicate dependency.

We assume that the remaining variables $Y^{(other)}$ are conditionally independent of the clustering given $Y^{(clust)}$ and $Y^{(?)}$ and belong to the same parametric family in both models.

This basically follows the approach used in Raftery and Dean (2006) for model-based clustering with continuous data and Gaussian clusters. One difference is that conditional independence of the variables was not assumed there, so that instead of $p(Y^{(?)})$ in model M_1 we had $p(Y^{(?)} \mid Y^{(clust)})$. This assumed conditional independence instead of full independence, i.e. the assumption in model M_1 previously was that given the information in $Y^{(clust)}$, $Y^{(?)}$ had no *additional* clustering information. Note, that unlike Figure 1 in Raftery and Dean (2006) there are no lines between the subsets of variables $Y^{(clust)}$ and $Y^{(?)}$ in our Figure 1, due to the conditional independence assumption.

Models M_1 and M_2 are compared via an approximation to the Bayes factor which allows the high-dimensional $p(Y^{(other)}|Y^{(clust)}, Y^{(?)})$ to cancel from the ratio. The Bayes factor, B_{12}, for M_1 against M_2 based on the data Y is given by

$$B_{12} = p(Y|M_1)/p(Y|M_2),$$

where $p(Y|M_k)$ is the integrated likelihood of model M_k $(k = 1, 2)$, namely

$$p(Y|M_k) = \int p(Y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k. \tag{3}$$

In (3), θ_k is the vector-valued parameter of model M_k, and $p(\theta_k|M_k)$ is its prior distribution (Kass and Raftery 1995).

Let us now consider the integrated likelihood of model M_1,
$p(Y|M_1) = p(Y^{(clust)}, Y^{(?)}, Y^{(other)}|M_1)$. From (2), the model M_1 is specified by three probability distributions: the latent class model that specifies $p(Y^{(clust)}|\theta_1, M_1)$, and the distributions $p(Y^{(?)}|\theta_1, M_1)$ and $p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, \theta_1, M_1)$. We denote the parameter vectors that specify these three probability distributions by θ_{11}, θ_{12}, and θ_{13}, and we assume that their prior distributions are independent. Then the integrated likelihood itself factors as follows:

$$p(Y|M_1) = p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, M_1)\, p(Y^{(?)}|M_1)\, p(Y^{(clust)}|M_1), \tag{4}$$

where

$$p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, M_1) = \int p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, \theta_{13}, M_1)\, p(\theta_{13}|M_1)d\theta_{13}.$$

Similar results hold for $p(Y^{(?)}|M_1)$ and $p(Y^{(clust)}|M_1)$. Similarly, we obtain

$$p(Y|M_2) = p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, M_2)\, p(Y^{(?)}, Y^{(clust)}|M_2), \tag{5}$$

where $p(Y^{(?)}, Y^{(clust)}|M_2)$ is the integrated likelihood for the latent class model for $(Y^{(?)}, Y^{(clust)})$.

The prior distribution of the parameter, θ_{13}, is assumed to be the same under M_1 as under M_2. It follows that

$$p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, M_2) = p(Y^{(other)}|Y^{(?)}, Y^{(clust)}, M_1).$$

We thus have

$$B_{12} = \frac{p(Y^{(?)}|M_1)p(Y^{(clust)}|M_1)}{p(Y^{(?)}, Y^{(clust)}|M_2)}, \tag{6}$$

which has been greatly simplified by the cancellation of the factors involving the potentially high-dimensional $Y^{(other)}$. The integrated likelihoods in (6) are still hard to evaluate analytically though, and so we approximate them using the BIC approximation of (1).

3.2 Headlong Search Algorithm

Given these models we need to find a method for creating partitions of the variables at each step. Initially we need enough variables to start $Y^{(clust)}$ so that a latent class model for $G > 1$ can be identified. If a latent class model on the set of all variables is identifiable for $G > 1$, we choose the largest number of classes that can be identified, and we then estimate the model. For each category of each variable, we then calculate the variance of its probability across groups. For each variable, we add up these variances and rank the variables according to this sum. The rationale is that variables with high values of this sum have high between-group variation in probability, and hence may be more useful for clustering.

Given this ranking we choose the top k^* variables, where k^* is the smallest number of variables that allow a latent class model with $G > 1$ to be identified. This is our starting $Y^{(clust)}$. The other variables can be left in their ordering based on variability for future order of introduction in the headlong algorithm.

If the above strategy is not possible, we instead proceed as follows. We calculate the minimum number of variables needed for identification of a latent class model with $G > 1$. We then select a number of random subsets each with this number of variables. Then for the initial $Y^{(clust)}$ we choose the variable set that gives the greatest overall average variance of categories' probabilities across the groups (given the best latent class model identified). If the minimum number of variables is small enough, we enumerate all possible subsets to choose the best initial $Y^{(clust)}$, instead of sampling.

Once we have an initial set of clustering variables $Y^{(clust)}$, we can proceed with the inclusion and exclusion steps of the headlong algorithm.

First we must define the constants *upper* and *lower*. The constant *upper* is the quantity above which the difference in BIC for models M_2 and M_1 will result in a variable being

included in $Y^{(clust)}$ and below which the difference in BIC for models M_2 and M_1 will result in a variable being excluded from $Y^{(clust)}$. The constant *lower* is the quantity below which the difference in BIC for models M_2 and M_1 will result in a variable being removed from consideration for the rest of the procedure. A natural value for *upper* is 0, by which we mean that any positive difference in BIC for models M_2 and M_1 is taken as evidence of a variable's usefulness for clustering and any negative difference is taken as evidence of a variable's lack of usefulness. A difference of *lower* is taken to indicate that a variable is unlikely to ever be useful as a clustering variable and is no longer even checked. In general a large negative number such as -100 (which by our rule of thumb would constitute strong evidence against) makes a sensible value for *lower*.

- *Inclusion Step*: Propose each variable in $Y^{(other)}$ singly in turn for $Y^{(?)}$. Calculate the difference in BIC for models M_2 and M_1 given the current $Y^{(clust)}$.
  If the variable's BIC difference is:
  - between *upper* and *lower*, do not include in $Y^{(clust)}$ and return variable to the end of the list of variables in $Y^{(other)}$;
  - below *lower*, do not include in $Y^{(clust)}$ and remove variable from $Y^{(other)}$;
  - above *upper*, include variable in $Y^{(clust)}$ and stop inclusion step.

  If we reach the end of the list of variables in $Y^{(other)}$, the inclusion step is stopped.

- *Exclusion Step*: Propose each variable in $Y^{(clust)}$ singly in turn for $Y^{(?)}$ (with the remaining variables in $Y^{(clust)}$ not including current $Y^{(?)}$ now defined as $Y^{(clust)}$ in M_1 and M_2). Calculate the difference in BIC for models M_2 and M_1. If the variable's BIC difference is:
  - between *upper* and *lower*, exclude the variable from (the original) $Y^{(clust)}$ and return variable to the end of the list of variables in $Y^{(other)}$ and stop exclusion step;
  - below *lower*, exclude the variable from (the original) $Y^{(clust)}$ and from $Y^{(other)}$ and stop exclusion step;
  - above *upper*, do not exclude the variable from (the original) $Y^{(clust)}$.

  If we reach the end of the list of variables in $Y^{(clust)}$ the exclusion step is stopped.

If $Y^{(clust)}$ remains the same after consecutive inclusion and exclusion steps the headlong algorithm stops because it has converged.

Table 1: Model parameters used to generate binary data example

Mixture proportions		
	Class 1	Class 2
	0.6	0.4
Variable	Prob. of success in class 1	Prob. of success in class 2
1	0.6	0.2
2	0.8	0.5
3	0.7	0.4
4	0.6	0.9
5	0.5	0.5
6	0.4	0.4
7	0.3	0.3
8	0.2	0.2
9	0.9	0.9
10	0.6	0.6
11	0.7	0.7
12	0.8	0.8
13	0.1	0.1

4 Simulated Data Results

4.1 Binary Simulated Data Example

Five hundred points were simulated from a two-class model satisfying the local independence assumption. There were four variables separating the classes (variables 1–4) and nine noise variables, i.e. variables that have the same probabilities in each class (variables 5–13). The actual model parameters are shown in Table 1.

When we estimated the latent class model based on all thirteen variables, BIC selected a two-class model. Since we simulated the data and hence know the actual membership of each point, we can compare the correct classification with that produced by the model estimated using all the variables. The number of observations incorrectly classified by this model was 123. The number of observations that would be incorrectly classified by using the model with the actual (true) parameter values is 110. The estimated parameters from the model with all variables are given in Table 2.

The variables ordered according the variability of their estimated probabilities (in decreasing order) are: 1, 3, 2, 4, 11, 7, 5, 6, 13, 9, 8, 10, 12. As expected, the first four

Table 2: Estimated parameters for the model involving all variables for the binary data example

Mixture proportions		
	Class 1	Class 2
	0.56	0.44
Variable	Prob. of success in class 1	Prob. of success in class 2
1	0.60	0.19
2	0.85	0.56
3	0.71	0.35
4	0.61	0.86
5	0.57	0.44
6	0.37	0.45
7	0.35	0.21
8	0.16	0.19
9	0.89	0.93
10	0.59	0.62
11	0.82	0.64
12	0.80	0.80
13	0.06	0.13

variables are the clustering variables. We note that the difference between the true probabilities across groups is 0.4 for variable 1 and 0.3 for variables 2 to 4. Since variable 1 therefore gives better separation of the classes, we would expect it to be first in the list. The number of variables needed in order to estimate a latent class model with at least 2 classes is 3. So the starting clustering variables are {1, 3, 2}. The individual step results for the variable selection procedure starting with this set are given in Table 3.

Table 3: Results for each step of the variable selection procedure for the binary data example. Note that the third and fourth row list the variables with the highest and lowest BIC difference respectively (i.e. all others were examined as well).

Variable(s) Proposed	Step Type	Clustering BIC	# of Classes	Independence BIC	Difference	Result
1, 3, 2	Inclusion	-1976.35	2	-1981.25	4.90	Accepted
4	Inclusion	-2565.37	2	-2573.62	8.25	Accepted
11	Inclusion	-3148.76	2	-3146.72	-2.04	Rejected
4	Exclusion	-2565.37	2	-2573.62	8.25	Rejected

When clustering on the four selected variables only, BIC again chose 2 classes as the best fitting model. Comparing the classification of the observations based on the estimates from this model with the correct classification we found that 110 observations had been misclassified. This seems to be optimal given that this is also the error from classifying based on the actual model parameters. The estimated parameters from the model using only selected variables are given in Table 4.

4.2 Non-Binary Simulated Data Example

One thousand points were simulated from a three-class model satisfying the local independence assumption. There are four variables that separate the classes (variables 1–4) and six noise variables that have the same probabilities in each class (variables 5–10). The actual model parameters are given in Table 5 and Table 6. Several other sets of parameters were used to simulate similar datasets where the algorithm gave results similar to this example; results are omitted.

When we estimated the latent class model based on all ten variables, BIC selected a 2-class model; recall that the actual number of classes is 3. The difference between BIC values for a 2-class and a 3-class model based on all the variables was 68. Again, since we have simulated the data and know the true membership of each point, we can compare the partition given by the true classification with that produced by the 2-class model estimated

Table 4: Estimated parameters for the model involving only the selected variables for the binary data example

Mixture proportions		
	Class 1	Class 2
	0.64	0.36
Variable	Prob. of success in class 1	Prob. of success in class 2
1	0.56	0.17
2	0.83	0.52
3	0.72	0.26
4	0.63	0.89

Table 5: Actual clustering parameters for the model with data from variables with different numbers of categories

Mixture proportions				
		Class 1	Class 2	Class 3
		0.3	0.4	0.3
Variable	Category	Prob. of category in class 1	Prob. of category in class 2	Prob. of category in class 3
Var. 1	Cat. 1	0.1	0.3	0.6
	Cat. 2	0.1	0.5	0.2
	Cat. 3	0.8	0.2	0.2
Var. 2	Cat. 1	0.5	0.1	0.7
	Cat. 2	0.5	0.9	0.3
Var. 3	Cat. 1	0.2	0.7	0.2
	Cat. 2	0.2	0.1	0.6
	Cat. 3	0.3	0.1	0.1
	Cat. 4	0.3	0.1	0.1
Var. 4	Cat .1	0.1	0.6	0.4
	Cat. 2	0.5	0.1	0.4
	Cat. 3	0.4	0.3	0.2

Table 6: Actual non-clustering parameters for the model with data from variables with different numbers of categories

Mixture proportions				
		Class 1	Class 2	Class 3
		0.3	0.4	0.3
Variable	Category	Prob. of category in class 1	Prob. of category in class 2	Prob. of category in class 3
Var. 5	Cat. 1	0.4	0.4	0.4
	Cat. 2	0.5	0.5	0.5
	Cat. 3	0.1	0.1	0.1
Var. 6	Cat. 1	0.2	0.2	0.2
	Cat. 2	0.4	0.4	0.4
	Cat. 3	0.1	0.1	0.1
	Cat. 4	0.3	0.3	0.3
Var. 7	Cat. 1	0.2	0.2	0.2
	Cat. 2	0.3	0.3	0.3
	Cat. 3	0.3	0.3	0.3
	Cat. 4	0.1	0.1	0.1
	Cat. 5	0.1	0.1	0.1
Var. 8	Cat. 1	0.2	0.2	0.2
	Cat. 2	0.8	0.8	0.8
Var. 9	Cat. 1	0.7	0.7	0.7
	Cat. 2	0.1	0.1	0.1
	Cat. 3	0.2	0.2	0.2
Var. 10	Cat. 1	0.1	0.1	0.1
	Cat. 2	0.2	0.2	0.2
	Cat. 3	0.1	0.1	0.1
	Cat. 4	0.6	0.6	0.6

Table 7: Results for each step of the variable selection procedure for the data from variables with different numbers of categories. Note that the third and fourth row list the variables with the highest and lowest BIC difference respectively (i.e. all others were examined as well).

Variable(s) Proposed	Step Type	Clustering BIC	# of Classes	Independence BIC	Difference	Result
2, 3, 1	Inclusion	-6122.65	2	-6193.37	70.72	Accepted
4	Inclusion	-8235.05	3	-8330.71	95.66	Accepted
8	Inclusion	-9261.46	3	-9248.28	-13.18	Rejected
2	Exclusion	-8235.05	3	-8322.40	87.36	Rejected

using all the variables. A cross-tabulation of the true memberships versus the estimated memberships from the 2-class model with all variables is as follows:

		Estimated classes	
		1	2
True classes	1	293	25
	2	85	324
	3	245	28

The misclassification rate from the model with the actual (true) parameters was 19.9%. If we match each true class to the best estimated class in the 2-class model with all variables we get a misclassification rate of 38.3%. If we assume that we knew the number of classes in advance to be 3 then the misclassification rate for the 3-class model with all variables is 25.7%. However this is knowledge that is not typically available in practice.

The variables ordered according the variability of their estimated probabilities in the 2-class model (in decreasing order) were: 2, 3, 1, 4, 6, 9, 7, 10, 8, 5. The first four variables are the clustering variables. The number of variables needed in order to estimate a latent class model with at least 2 classes is 3. So the starting clustering variables were {2, 3, 1}. The individual step results for the variable selection procedure starting with this set are given in Table 7.

When clustering on the four selected variables only, BIC this time chose 3 classes as the best fitting model. Comparing the partition from classifying observations based on the estimates from this model and the correct partition we found that the misclassification rate was 23.8%. The estimated parameters from the model using only selected variables are given in Table 8.

The misclassification results are summarized in Table 9. In addition to the misclassification rate, we show the Rand Index (Rand 1971) and the Adjusted Rand Index (Hubert and

Table 8: Estimated parameters for the model involving only the selected variables for the data from variables with different numbers of categories

Mixture proportions				
		Class 1	Class 2	Class 3
		0.40	0.43	0.16
Variable	Category	Prob. of category in class 1	Prob. of category in class 2	Prob. of category in class 3
Var. 1	Cat. 1	0.10	0.34	0.85
	Cat. 2	0.1	0.49	0.13
	Cat. 3	0.80	0.17	0.02
Var. 2	Cat. 10	0.49	0.12	0.82
	Cat. 2	0.51	0.88	0.18
Var. 3	Cat. 1	0.21	0.64	0.17
	Cat. 2	0.27	0.14	0.63
	Cat. 3	0.25	0.13	0.08
	Cat. 4	0.27	0.09	0.12
Var. 4	Cat .1	0.14	0.53	0.39
	Cat. 2	0.47	0.10	0.47
	Cat. 3	0.39	0.37	0.14

Arabie 1985).

Table 9: Misclassification Summary for the data from variables with different numbers of categories. (c) indicates that the number of classes was constrained to this value in advance. Recall that the minimum misclassification rate from the model based on the actual parameters is 19.9%.

Variables Included	No. of Classes selected	Misclassification Rate	Rand Index	Adjusted Rand Index
All	2	38.3%	0.65	0.30
All	3(c)	25.7%	0.72	0.40
1,2,3,4	3	23.8%	0.74	0.43

5 Real Data Examples

5.1 Hungarian Heart Disease Data

This dataset consists of five categorical variables from a larger dataset (with 10 other continuous variables) collected from the Hungarian Institute of Cardiology, Budapest by Andras Janosi, M.D. (Detrano et al. 1989; Gennari et al. 1989). The outcome of interest is diagnosis of heart disease (angiographic disease status) into two categories: $< 50\%$ diameter narrowing and $> 50\%$ diameter narrowing in any major vessel. The original paper (Detrano et al. 1989) looked at the data in a supervised learning context and achieved a 77% accuracy rate. Originally there was information about 294 subjects but 10 subjects had to be removed due to missing data. The five variables given are gender (male/female) [sex], chest pain type (typical angina/atypical angina/non-anginal pain/asymptomatic) [cp], fasting blood sugar $>$ 120 mg/dl (true/false) [fbs], resting electrocardiographic results (normal/having ST-T wave abnormality/showing probable or definite left ventricular hypertrophy by Estes' criteria) [restecg] and exercise induced angina (yes/no) [exang].

When BIC is used to select the number of classes in a latent class model with all of the variables, it decisively selects 2 (with a difference of at least 38 points between 2 classes and any other identifiable number of classes). When the variables are put in decreasing order of variance of estimated probabilities between classes the ordering is the following: cp, exang, sex, restecg and fbs.

Observations were classified into whichever group their estimated membership probability was greatest for. The partition estimated by this method is compared with the clinical partition below:

	<50% narrowing	>50% narrowing
Class 1	134	13
Class 2	47	90

If class 1 is matched with the <50% class and class 2 with the >50% class there is a correct classification rate of 78.9%. This gives a sensitivity of 87.4% and a specificity of 74%.

The variable selection method chooses 3 variables: cp, exang and sex. BIC selects 2 classes for the latent class model on these variables. The partition given by this model is the same as the one given by the model with all variables. The largest difference in estimated group membership probabilities between the two latent class models is 0.1. The estimated model parameters in the variables common to both latent class models and the mixing proportions differ between models by at most 0.003. Both models have the same correct classification, specificity and sensitivity rate. Thus our method identifies the fact that it is possible to reduce the number of variables from 5 to 3 with no cost in terms of clustering.

The estimated parameters for the latent class model with all variables included is given in Table 10 and the estimated parameters for the latent class model with only the selected variables included is given in Table 11.

5.2 HapMap Data

The HapMap project (The International HapMap Consortium 2003) was set up to examine patterns of DNA sequence variation across human populations. A consortium with members including the United States, United Kingdom, Canada, Nigeria, Japan and China is attempting to identify chromosomal regions where genetic variants are shared across individuals. One of the most common types of these variants is the single nucleotide polymorphism (SNP). A SNP occurs when a single nucleotide (A, T, C or G) in the genome differs across individuals. If a particular locus has either A or G then these are called the two alleles. Most SNPs have only two alleles.

This dataset is from a random selection of 3,389 SNPs on 210 individuals (out of 4 million available in the HapMap public database). Of these 801 had complete sets of measurements from all subjects and a further subset of 639 SNPs had non-zero variability. Details of the populations and numbers of subjects are given in Table 12.

There are two possible correct groupings of the data. The first one is into 3 groups: European (CEU), African (YRI) and Asian (CHB+JPT), and the second is into 4 groups: European (CEU), African (YRI), Japanese (JPT) and Chinese (CHB).

Table 10: Estimated parameters for the model involving all variables for Hungarian Heart Disease Data

Mixture proportions			
		Class 1	Class 2
		0.494	0.506
Variable	Category	Prob. of category in class 1	Prob. of category in class 2
Chest Pain Type	Typical angina	0.07	0.00
	Atypical angina	0.64	0.08
	Non-anginal pain	0.29	0.08
	Asymptomatic	0.00	0.83
Exercise Induced Angina	No	0.98	0.42
	Yes	0.02	0.58
Gender	Female	0.38	0.16
	Male	0.62	0.84
Resting Electrocardiographic Results	Normal	0.82	0.80
	Having ST-T wave abnormality	0.15	0.20
	Showing probable or definite left ventricular hypertrophy by Estes' criteria	0.03	0.01
Fasting blood sugar > 120 mg/dl	False	0.94	0.92
	True	0.06	0.08

Table 11: Estimated parameters for the model involving the selected variables for Hungarian Heart Disease Data

Mixture proportions			
		Class 1	Class 2
		0.498	0.502
Variable	Category	Prob. of category in class 1	Prob. of category in class 2
Chest Pain Type	Typical angina	0.07	0.00
	Atypical angina	0.64	0.08
	Non-anginal pain	0.28	0.08
	Asymptomatic	0.00	0.84
Exercise Induced Angina	No	0.97	0.42
	Yes	0.03	0.58
Gender	Female	0.38	0.16
	Male	0.62	0.84

Table 12: Information on the subject populations for the HapMap data

Code	Descriptions	Number of Individuals
CEU	Utah residents with ancestry from Northern and Western Europe	60
CHB	Han Chinese in Beijing, China	45
JPT	Japanese in Tokyo, Japan	45
YRI	Yoruban in Ibadan, Nigeria (West Africa)	60

Table 13: BIC values for different sets of variables and different numbers of classes for the HapMap data.

Data	2 classes	3 classes	4 classes
All variables	-142711	-141418	-146662
Selected variables	-93471	-91147	-94491

When all 639 SNPs are used to build latent class models, BIC selects the best number of classes as 3. The resulting estimated partition matches up exactly with the first 3-group partition. The variable selection procedure selects 413 SNPs as important to the clustering, reducing the number of variables by over a third. Using only the selected variables, BIC again selects a 3-class model whose estimated partition again gives perfect 3-group classification. The BIC values for models using both sets of data from 2 to 4 classes are given in Table 13. Note that comparing within rows in Table 13 is appropriate, but comparing between rows is not because different rows correspond to different datasets.

The HapMap project is also interested in the position of SNPs that differ between populations, so we can look at the distribution of all 639 SNPs across the 22 chromosomes and compare it to the distribution of the selected 413 SNPs. This is presented in Figure 2.

Although the subset of SNPs that these data come from are a random sample, it may be that some are close to each other on the same chromosome. Since genetic variants close to each other on a chromosome tend to be inherited together, this suggests that the conditional independence assumption for LCA may not hold in this case. Incorporating these dependencies may be beneficial.

6 Discussion

We have proposed a method for selecting variables in latent class analysis. In our simulated datasets the method selected the correct variables, and this also led to improved classification and more accurate selection of the number of classes. In both real data examples, the data were classified equally accurately by the smaller set of variables selected by our method as by a larger set. The HapMap data provided an example of the "$n << p$" type, and there our method reduced the number of variables (SNPs in that case) by over a third without any degradation in classification performance.

In general it appears to be a better idea to select variables before estimating the clustering model in both the discrete and continuous cases. We have seen that inclusion of noise

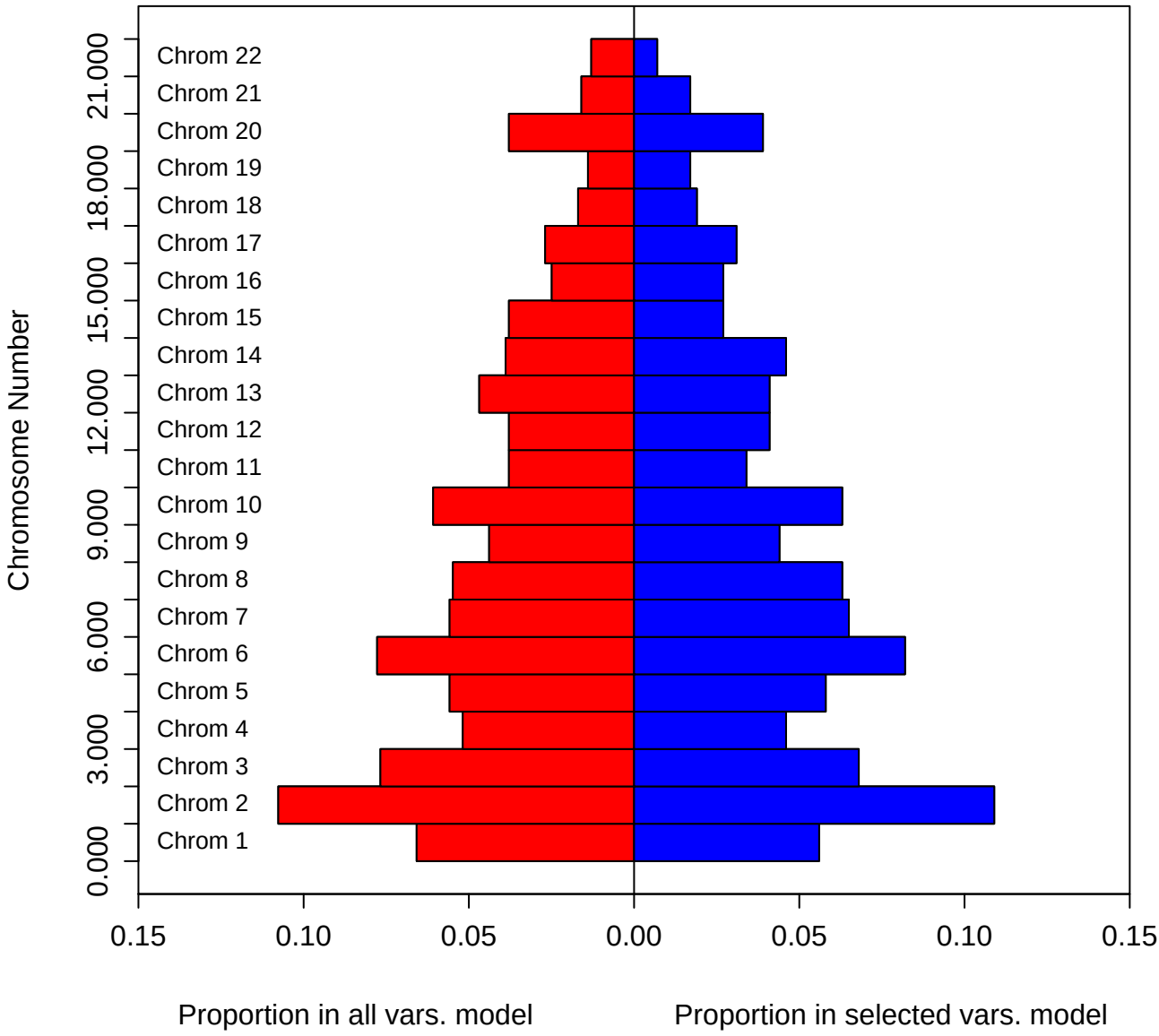

Figure 2: Distribution of SNPs for full and selected variable sets on the set of 22 chromosomes

variables can degrade the accuracy of both model estimation and choice of the number of clusters.

In terms of estimation of the model, including variables with no cluster structure can either smear out separated clusters/classes or introduce spurious classes. It is difficult without any extra knowledge to know what can happen in advance. From looking at the simulations and data sets presented here as well as others, it would appear that these problems are most likely to occur when separation between the classes is poor.

The headlong search algorithm is different from the greedy search algorithm described in Raftery and Dean (2006) in two ways:

1. The best variable (in terms of the BIC difference) is not necessarily selected in each inclusion and exclusion step in the headlong search.

2. It is possible that some variables are not looked at in any step after a certain point in the headlong algorithm (after being removed from consideration).

The headlong search is substantially faster than the greedy search and in spite of point 2 above, usually gives results comparable to or sometimes better than the results of the greedy search (perhaps due to the local nature of the search).

Galimberti and Soffritti (2006) considered the problem of finding multiple cluster structures in latent class analysis. In this problem the data are divided into subsets, each of which obeys a different latent class model. The models in the different subsets may include different variables. This is a somewhat different problem from the one we address here, but it also involves a kind of variable selection in latent class analysis.

Keribin (1998) showed that BIC was consistent for choice of the number of components in a mixture model under certain conditions, notably assuming that all variables were relevant to the clustering. Empirical evidence seems to suggest that when noise/irrelevant variables are present, BIC is less likely to select the correct number of classes. The general correctness of the BIC approximation in a specific case of binary variables with two classes in a naive Bayes network (which is equivalent to a 2-class latent class model with the local independence assumption satisfied) was looked at by Rusakov and Geiger (2005). The authors found that although the traditional BIC penalty term of # of parameters $\times \log(\#$ of observations) (or half this depending on the definition) was correct for regular points in the data space, it was not correct for singularity points (with two different types of singularity points requiring two adjusted versions of the penalty term). The first type of singularity points were those sets of parameters that could arise from a naive Bayes model with all but at most 2 links removed (type 1) and those that could arise from a model with all links removed (type 2), representing a set of mutually independent variables. Similarly in the case of redundant or

irrelevant variables being included (which is closely related to the two singularity point types) they found that the two adjusted penalty terms were correct. These issues with clustering with noise variables reinforce the arguments for variable selection in latent class analysis.

Acknowledgments

Both authors were supported by NIH grant 8 R01 EB002137-02. Raftery was also supported by NICHD grant 1 R01HD O54511, NSF grant IIS0534094 and NSF grant ATM0724721. The authors would like to thank Matthew Stephens and Paul Scheet for their preparation of the HapMap data for this paper.

References

Badsberg, J. H. (1992). Model search in contingency tables by CoCo. In Y. Dodge and J. Whittaker (Eds.), *Computational Statistics*, Volume 1, pp. 251–256.

Clogg, C. C. (1981). New developments in latent structure analysis. In D. J. Jackson and E. F. Borgatta (Eds.), *Factor Analysis and Measurement in Sociological Research*, pp. 215–246. Beverly Hills: Sage.

Clogg, C. C. (1995). Latent class models. In C. C. C. G. Arminger and M. E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pp. 311–360. New York: Plenum.

Detrano, R., A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology 64*, 304–310.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Galimberti, G. and G. Soffritti (2006). *From Data and Information Analysis to Knowledge Engineering*, Chapter , Identifying Multiple Cluster Structures Through Latent Class Models, pp. 174–181. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin: Springer.

Gennari, J. H., P. Langley, and D. Fisher (1989). Models of incremental concept formation. *Artificial Intelligence 40*, 11–61.

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

Hagenaars, J. A. and A. L. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge, U.K.: Cambridge University Press.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Keribin, C. (1998). Consistent estimate of the order of mixture models. *Comptes Rendues de l'Academie des Sciences, Série I-Mathématiques 326*, 243–248.

Lazarsfeld, P. F. (1950a). *Measurement and Prediction, Volume IV of The American Soldier: Studies in Social Psychology in World War II*, Chapter , The Logical and Mathematical Foundations of Latent Structure Analysis, pp. 362–412. Princeton University Press.

Lazarsfeld, P. F. (1950b). *Measurement and Prediction, Volume IV of The American Soldier: Studies in Social Psychology in World War II*, Chapter Some latent structures, pp. 362–412. Princeton University Press.

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

McCutcheon, A. L. (1987). *Latent Class Analysis*. Newbury Park, Calif.: Sage.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association 101*, 168–178.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 846–850.

Rusakov, D. and D. Geiger (2005). Asymptotic model selection for naive Bayesian networks. *Journal of Machine Learning Research 6*, 1–35.

The International HapMap Consortium (2003). The international hapmap project. *Nature 426*, 789–796.

Wolfe, J. H. (1963). Object cluster analysis of social areas. Master's thesis, University of California, Berkeley.

Appendix A: Headlong Search Algorithm for Variable Selection in Latent Class Analysis

Here we give a more complete description of the headlong search variable selection and clustering algorithm for the case of discrete data modeled by conditionally independent multinomially distributed groups. Note that for each latent class model fitted in this algorithm one must run a number of random starts to find the best estimate of the model (in terms of BIC). We recommend at least 5 for small to medium problems but for bigger problems hundreds may be needed to get a decent model estimate. The issue of getting good starting values without multiple generation of random starts is dealt with in appendix B.

- Choose G_{max}, the maximum number of clusters/classes to be considered for the data. Make sure that this number is identifiable for your data! Define constants *upper* (default 0) and *lower* (default -100), where *upper* is the quantity above which the difference in BIC for models M_2 and M_1 will result in a variable being included in $Y^{(clust)}$ and below which the difference in BIC for models M_2 and M_1 will result in a variable being excluded from $Y^{(clust)}$, and *lower* is the quantity below which the difference in BIC for models M_2 and M_1 will result in a variable being removed from consideration for the rest of the procedure.

- **First step :** One way of choosing the initial clustering variable set is by estimating a latent class model with at least 2 classes for all variables (if more classes are identifiable, estimate all identifiable class numbers and choose the model with the best number of classes via BIC). Order the variables in terms of variability of their estimated probabilities across classes. Choose the minimum top variables that allow at least a 2-class model to be identified. This is the initial $Y^{(clust)}$. We do not require that the BIC difference between clustering and a model with a single class for our $Y^{(clust)}$ to be positive at this point because we need a set of starting variables for the algorithm. These can be removed later if there are not truly clustering variables.

  Specifically we estimate the $\{p_{ijg}, i = 1, \ldots, k, j = 1, \ldots, d_i, g = 1, \ldots, G\}$ where k is the number of variables, d_i is the number of categories for the i^{th} variables and G is the number of classes. For each variable i we calculate $V(i) = \sum_{j=1}^{d_i} Var(p_{ijg})$. We order the variables in decreasing order of $V(i)$: $y^{(1)}, y^{(2)}, \ldots, y^{(k)}$ and find m the minimum number of top variables that will identify a latent class model with $G \geq 2$.

$$\begin{aligned} Y^{(clust)} &= \{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\} \\ Y^{(other)} &= \{y^{(m+1)}, \ldots, y^{(k)}\} \end{aligned}$$

If the previous method is not possible (data cannot identify latent class model for $G > 1$) then split the variables randomly into subsets with enough variables to identify a latent class model for at least 2 classes, estimate the latent models for each subset and calculate the BICs, estimate the single class (G=1) models for each subset and calculate these 1 class BICs and choose the subset with the highest difference between latent class model (G≥2) and 1 class model BICs as the initial $Y^{(clust)}$.

Specifically look at the list of numbers of categories $d = (d_1, \ldots, d_k)$ and work out the minimum number of variables m that allows a latent class model for $G \geq 2$ to be identified. Split the variables into S subsets of at least m variables in each. For each set $Y_s, s = 1, \ldots, S$ estimate:

$$BIC_{\text{diff}}(Y_s) = BIC_{\text{clust}}(Y_s) - BIC_{\text{not clust}}(Y_s)$$

where $BIC_{\text{clust}}(Y_s) = \max_{2 \leq G \leq G_{maxs}}\{BIC_G(Y_s)\}$, with $BIC_G(Y_s)$ being the BIC given in (1) for the latent class model for Y_s with G classes and G_{maxs} being the maximum number of identifiable classes for Y_s, and $BIC_{\text{not clust}}(Y_s) = BIC_1(Y_s)$.

We choose the best variable subset, Y_{s^1}, such that

$$s^1 = \arg \max_{s: Y_s \in Y} (BIC_{\text{diff}}(Y_s))$$

and create

$$\begin{aligned} Y^{(clust)} &= Y_{s^1} \\ \text{and } Y^{(other)} &= Y \backslash Y_{s^1} \end{aligned}$$

where $Y \backslash Y_{s^1}$ denotes the set of variables Y excluding the subset Y_{s^1}.

- **Second step :** Next we look at each variable in $Y^{(other)}$ singly in order as the new variable under consideration for inclusion into $Y^{(clust)}$. For each variable we look at the difference between the BIC for clustering on the set of variables including the variables selected in the first set and the new variable (maximized over number of clusters from 2 up to G_{max}) and the sum of the BIC for the clustering of the variables chosen in the first step and the BIC for the single class latent class model for the new variable. If this difference is less than *lower* the variable is removed from consideration for the rest of the procedure and we continue checking the next variable. Once the difference is greater than *upper* we stop and this variable is included in the set of clustering variables. Note that if no variable has difference greater than *upper* we include the variable with the largest difference in the set of clustering variables. We force a variable to be selected at this stage to give one final extra starting variable.

Specifically, we split $Y^{(other)}$ into its variables $y^1, \ldots, y^{D_2}$. For each j in $1, \ldots, D_2$ until $BIC_{\text{diff}}(y^j) > upper$, we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}}(y^j) - BIC_{\text{not clust}}(y^j)$$

where $BIC_{\text{clust}}(y^j) = \max_{2 \leq G \leq G_{max}j}\{BIC_G(Y^{(clust)}, y^j)\}$ with $BIC_G(Y^{(clust)}, y^j)$ being the BIC given in (1) for the latent class clustering model for the dataset including both the previously selected variables (contained in $Y^{(clust)}$) and the new variable y^j with G classes, and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}} + BIC_{\text{clust}}(Y^{(clust)})$ where BIC_{reg} is $BIC_1(y^j)$ and $BIC_{\text{clust}}(Y^{(clust)})$ is the BIC for the latent class clustering model with only the currently selected variables in $Y^{(clust)}$.

We choose the first variable, y^{j_2}, such that

$$BIC_{\text{diff}}(y^{j_2}) > upper$$

or if no such j_2 exists,

$$j_2 = \arg \max_{j: y^j \in Y^{(other)}} (BIC_{\text{diff}}(y^j))$$

and create

$$\begin{aligned} Y^{(clust)} &= Y^{(clust)} \cup y^{j_2} \\ \text{and } Y^{(other)} &= Y^{(other)} \backslash y^{j_2} \end{aligned}$$

where $Y^{(clust)} \cup y^{j_2}$ denotes the set of variables including those in $Y^{(clust)}$ and variable y^{j_2}.

- **General Step [Inclusion part] :** Each variable in $Y^{(other)}$ is proposed singly (in order), until the difference between the BIC for clustering with this variable included in the set of currently selected clustering variables (maximized over numbers of clusters from 2 up to G_{max}) and the sum of the BIC for the clustering with only the currently selected clustering variables and the BIC for the single class latent class model of the new variable, is greater than *upper*.

- The variable with BIC difference greater than upper is then included in the set of clustering variables and we stop the step. Any variable whose BIC difference is less than *lower* is removed from consideration for the rest of the procedure. If no variable has BIC difference greater than *upper* no new variable is included in the set of clustering variables

Specifically, at step t we split $Y^{(other)}$ into its variables $y^1, \ldots, y^{D_t}$. For j in $1, \ldots, D_t$ we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}}(y^j) - BIC_{\text{not clust}}(y^j) \tag{7}$$

where $BIC_{\text{clust}}(y^j) = \max_{2 \leq G \leq G_{max}} \{BIC_G(Y^{(clust)}, y^j)\}$, with $BIC_G(Y^{(clust)}, y^j)$ being the BIC given in (1) for the latent class clustering model for the dataset including both the previously selected variables (contained in $Y^{(clust)}$) and the new variable y^j with G clusters, and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}} + BIC_{\text{clust}}(Y^{(clust)})$ where BIC_{reg} is the single class latent class model for variable y^j and $BIC_{\text{clust}}(Y^{(clust)})$ is the BIC for the clustering with only the currently selected variables in $Y^{(clust)}$.

We check if $BIC_{\text{diff}}(y^j) > upper$,
if so we stop and set

$$\begin{aligned} Y^{(clust)} &= Y^{(clust)} \cup y^j \text{ if } BIC_{\text{diff}}(y^j) > 0 \\ \text{and } Y^{(other)} &= Y^{(other)} \backslash y^j \text{ if } BIC_{\text{diff}}(y^j) > 0 \end{aligned}$$

if not we increment j and re-calculate $BIC_{\text{diff}}(y^j)$ If $BIC_{\text{diff}}(y^j) < lower$ we remove it from both $Y^{(clust)}$ and $Y^{(other)}$

If no j has $BIC_{\text{diff}}(y^j) > upper$ leave $Y^{(clust)} = Y^{(clust)}$ and $Y^{(other)} = Y^{(other)}$.

- **General Step [Removal part] :** Each variable in $Y^{(clust)}$ is proposed singly (in order), until the difference between the BIC for clustering with this variable included in the set of currently selected clustering variables (maximized over numbers of clusters from 2 up to G_{max}) and the sum of the BIC for the clustering with only the other currently selected clustering variables (and not the variable under consideration) and the BIC for the single class latent class model of the variable under consideration, is less than *upper*.

- The variable with BIC difference less than upper is then removed from the set of clustering variables and we stop the step. If the difference is greater than *lower* we include the variable at the end of the list of variables in $Y^{(other)}$. If not we remove it entirely from consideration for the rest of the procedure. If no variable has BIC difference less than *upper* no variable is excluded from the current set of clustering variables

  In terms of equations for step $t+1$, we split $Y^{(clust)}$ into its variables $y^1, \ldots, y^{D_{t+1}}$. For each j in $1, \ldots, D_{t+1}$ we compute the approximation to the Bayes factor in (6) by

$$BIC_{\text{diff}}(y^j) = BIC_{\text{clust}} - BIC_{\text{not clust}}(y^j)$$

where $BIC_{\text{clust}} = \max_{2 \leq G \leq G_{max}} \{BIC_G(Y^{(clust)})\}$ with $BIC_{G,m}(Y^{(clust)})$ being the BIC given in (1) for the model-based clustering model for the dataset including the previously selected variables (contained in $Y^{(clust)}$) with G clusters, and $BIC_{\text{not clust}}(y^j) = BIC_{\text{reg}} + BIC_{\text{clust}}(Y^{(clust)} \backslash y^j)$ where BIC_{reg} is the single class latent class model for variable y^j and $BIC_{\text{clust}}(Y^{(clust)} \backslash y^j)$ is the BIC for the clustering with all the currently selected variables in $Y^{(clust)}$ except for y^j.

We check if $BIC_{\text{diff}}(y^j) < upper$,
if so we stop and set

$$\begin{aligned} Y^{(clust)} &= Y^{(clust)} \backslash y^j \text{ if } BIC_{\text{diff}}(y^j) < upper \\ \text{and } Y^{(other)} &= Y^{(other)} \cup y^j \text{ if } lower < BIC_{\text{diff}}(y^j) < upper \end{aligned}$$

if not we increment j and re-calculate $BIC_{\text{diff}}(y^j)$. If $BIC_{\text{diff}}(y^j) < lower$ we remove it from both $Y^{(clust)}$ and $Y^{(other)}$.

If no j has $BIC_{\text{diff}}(y^j) < upper$ leave $Y^{(clust)} = Y^{(clust)}$ and $Y^{(other)} = Y^{(other)}$.

- After the first and second steps the general step is iterated until consecutive inclusion and removal proposals are rejected. At this point the algorithm stops as any further proposals will be the same ones already rejected.

Appendix B: Smart Starting Values for Latent Class Analysis in the Headlong Search Algorithm

In the previous appendix we discussed the details of the headlong algorithm for latent class variable selection. In each step multiple latent class models for different sets of data/variables and classes are estimated. Previously we have only mentioned that starting values are generated randomly for each model several times and the best (in terms of BIC/likelihood) of the resulting estimated models is chosen as the single estimate for a particular latent class model. This means that for each different dataset and each different number of classes we are required to generate random starting values and estimate the model via EM numerous times. For datasets with reasonable numbers of variables this is not too computationally expensive but for more complex datasets it is burdensome. Also with increasing numbers of observations and/or variables and/or classes more random starts are needed to have any confidence in finding the global maximum likelihood for the model as the likelihood surface becomes more complex, with increasing numbers of local maxima.

Because of the stepwise nature of the algorithm we can use models estimated before to give good starting values for new models. By starting values here we mean the matrix z of conditional probabilities of membership in the different classes for each observation.

At the end of each step (either inclusion or exclusion) we have a set of currently selected clustering variables. At some point in the step we have estimated the latent class model for this set over a range of classes (or sometimes just one, 2 classes) and chosen the model with the number of classes that gives us the highest BIC. We can call this model $LCA_{current}$ and the number of classes in this best model for the current set of clustering variables $G_{current}$. We can also save the z matrix for this model and call it $z_{current}$.

In our next step we will be either looking at models for $Y^{(clust)}$ with a new additional variable (inclusion step) or models for $Y^{(clust)}$ leaving out one of the current clustering variables (exclusion step). It seems obvious that a reasonable starting z matrix for models involving the new dataset (which is either a sub- or super-set of the old one) and number of classes $G_{current}$ would be $z_{current}$, because the dataset will only have changed by one variable. So instead of randomly generating multiple z matrixes (or other starting parameters) to try to get the global maximum likelihood for our latent class model, we merely use what we believe to be a good set starting z matrix (which hopefully will be reasonably close to the global maximum in the new likelihood space).

However, we may still wish to have good starting values for the new dataset with different numbers of classes, $G_{current} \pm c$. But our $z_{current}$ will be an $n \times G_{current}$ matrix (where n is the number of observations) and we need $n \times (G_{current} \pm c)$ matrices. How can we sensibly create a new matrix with c more/less columns given our $z_{current}$?

We will look at the case for $+1$ and -1 separately (the analogue for general $+c$ and $-c$ should be obvious). It will be rare in practice to need more than ± 1 at each step as the number of identifiable classes will only generally increase fairly slowly with the number of variables selected.

For -1 we want to reduce the number of columns of our $z_{current}$ by 1. A sensible way to do this is to collapse the two closest classes (in terms of Euclidean distance in the parameter space). We calculate the distances between the classes' estimated parameters/probabilities from $LCA_{current}$ and select the closest two. We then simply remove the two columns corresponding to those classes from $z_{current}$ and replace them with one column equal to the sum (across rows) of the removed columns. This is our new starting z matrix for the model with $G_{current} - 1$ classes. In terms of a single observation with probability p_1 of being in the first chosen class and probability p_2 of being the second chosen class we are saying the observation has probability $p_1 + p_2$ of being in the new class created from the amalgamation of the two i.e. the observation will be in the new class if he is in either of the old classes. Note that if

we wish to, we can weight the distances with the mixing proportions, making it more likely that we would join smaller close classes.

For $-c$ we can use the resulting matrix from the process described in the previous paragraph to estimate the model for $G_{current} - 1$ classes and then reduce the resulting estimated z from this model by one column in the same fashion, continuing on in the same way until we have removed c columns.

For $+1$ we want to increase the number of columns of our $z_{current}$ by 1. An obvious way to do this is by splitting a class in two. We choose the largest class (in terms of mixing proportions). We then remove the column corresponding to that class from $z_{current}$ and call this w and estimate a two class latent class model using the data points weighted by w. Obviously we have returned to problem of needing starting values for estimating our 2-class model. However usually a small number of randomly generated starts, say 5, for this number of classes will result in an estimated model achieving the global maximum likelihood and this is usually not too computationally expensive. Once we have our 2-class model estimate of the z matrix, called z_2, we can (scalar/column) multiply this by w and add the resulting two columns to the original $z_{current}$ (less the removed column), giving us a starting z matrix for estimating the $G_{current} + 1$ class model. We can think of w as being the conditional probability of an observation being in the old selected class and then the new z_2 matrix as being the probability for an observation being in either of the two new sub-classes *given* it was in the old class.

Again for $+c$ we can use the resulting matrix from the process described in the previous paragraph to estimate the model for $G_{current} + 1$ classes and then increase the resulting estimated z from this model by one column in the same fashion, continuing on in the same way until we have added c columns.