Indicator Selection For Latent Class Models Using Constrained Model Fitting

Brian Francis
Department of Mathematics and Statistics
Lancaster University, UK

Fuad A. A. Awwad
Department of Quantitative Analysis, KING
Saud University, Riyadh, Saudi Arabia

Abstract:

Whilst considerable attention has been paid to determining the number of classes in a latent class analysis less attention has been directed at the optimal selection of indicator variables. Indicator selection reduces redundancy and complexity, and can provide a way forward in cases where the number of indicators is large. However, determination of the optimal indicator set and the optimal number of classes is not straightforward, as the two are heavily interrelated.

This paper reports on a reformulation and extension of the Dean and Raftery algorithm. By treating subset selection as an imposition of sets of constraints on the class membership probabilities, the BIC (or any other information criterion) becomes informative both for determining the optimal subset selection and for determining the number of classes. The procedure is illustrated by a dataset on the presence or absence of psychiatric symptoms in 30 psychiatric patients.

Keywords:

Indicator selection, constrained latent class analysis, variable selection, number of classes, symptom classification.

1- Introduction:

Latent class analysis is now heavily used in medicine as a method for determining subgroup structure in a set of indicator variables. Common uses include medical diagnosis and symptom classification [8][19], investigation of response patterns in medical surveys [20], and assessment of differential need in patient groups[21].

Whilst considerable attention has been paid to determining the number of classes in a latent class analysis, less attention has been directed at the optimal selection of indicator variables, Indicator selection reduces redundancy and complexity, and can provide a way forward in cases where the number of indicators is large. However, joint determination of the optimal indicator set and the optimal number of classes is not straightforward, as the two are heavily interrelated.

Work on this topic has been carried out by Dean and Raftery [5]. Their method essentially consists of two stages. They first propose the use of the BIC to determine the number of groups based on a sequence of latent class analyses which use all of the indicator variables. Then, once the number of groups has been selected, a "headlong search" or stepwise algorithm is used to compare the BICs of latent class analyses on subsets of variables, leading to a best subset of variables. However, this two-step approach which involves first finding the number of groups and then the best subset of indicator variables may not be optimal. In addition, the headlong search algorithm is difficult to implement in practice, as it involves fitting pairs of latent class models with different numbers of indicator variables. This paper therefore proposes a modification and extension to the Dean and Raftery algorithm which simultaneously determines both the optimal number of groups and the optimal subset of indicators. In addition, a reformulation of the headlong algorithm improves its usability in standard software.

The paper proceeds as follows. Section 2 introduces the notation for the paper and the basic latent class model. Section 3 describes earlier approaches to

variable selection in latent class analysis and the Dean and Raftery headlong search algorithm. Section 4 introduces our algorithmic approach. Section 5 describes the dataset used , and Section 6 presents the results of the analysis. The paper concludes with a short discussion.

2. The Latent Class Model

The latent class model has been described by many others [11],[13],[14] and is summarized here to introduce the notation. We consider a set of J indicator variables, with each indicator representing a binary outcome on a particular characteristic or response. For individual i $(i=1...n)$, we let Y_{ij} represent the random indicator variable j which takes the value 1 or 0. We assume the existence of K latent classes. We also let $\{Y_{ij}\} = \{y_{ij}\}$ represent the full set of responses over all individuals and indicator variables. Then the basic latent class model can be written as

$$P(\{Y_{ij}\}|\{q_{ij}\}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \sum_{j=1}^{J} q_{jk}^{y_{ij}} \left(1 - q_{jk}\right)^{1-y_{ij}} \quad \ldots(1)$$

where q_{jk} is the probability of a '1' response for indicator j for an individual in class k, and the π_k represent the class sizes , with $\sum \pi_k = 1$. The parameters $\{q_{jk}\}$ and $\{\pi_k\}$ are unknown. The likelihood in the model parameters $\lambda = (K, \{q_{jk}\}, \{\pi_k\})$ is then straightforward:

$$L(\lambda) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \sum_{j=1}^{J} q_{jk}^{y_{ij}} \left(1 - q_{jk}\right)^{1-y_{ij}} \quad \ldots(2)$$

Maximization of the likelihood for fixed K can be carried out either using the EM algorithm, a Newton-Raphson approach, or hybrid schemes combining EM with Newton steps, such as that used in the Latent Gold package [9]. In general, the likelihood surface is multi-modal, and so it is essential that multiple start values are used. Typically, a large number of different random start values of the model parameters are chosen, in order to best prevent a local rather than a global maximum of the likelihood being reached.

Once estimation is complete, the posterior probability p_{ik} of individual i belonging to class k can be obtained:

$$p_{ik} = \frac{\pi_k \sum_{j=1}^{J} q_{jk}^{y_{ij}} (1-q_{jk})^{1-y_{ij}}}{\sum_{k=1}^{K} \pi_k \sum_{j=1}^{J} q_{jk}^{y_{ij}} (1-q_{jk})^{1-y_{ij}}} \quad \ldots(3)$$

In order to determine the number of classes K, a common procedure is to use the BIC

$$BIC(K) = -2 \log L(K) + p \log(n) \quad \ldots(4)$$

where p is the number of parameters being estimated in the model, and choosing the value of K which produces the lowest value of BIC.

3. Previous Approaches to Variable Selection in Latent Class Models.

Work on the general problem of variable selection in model-based clustering (sometimes known as subspace clustering) is becoming of increasing interest. Early work by Friedman and Meulman [9] on attribute selection in cluster analysis was followed by an important paper by Raftery and Dean [18], who, in the context of model-based clustering suggested a greedy search algorithm that simultaneously chose the number of classes K, the optimal subset of variables and the clustering model, based on the BIC criterion, and illustrated it in the context of a Gaussian mixture model. Other forms of penalised likelihood have been proposed; Xie et al [24] proposed a L_1 norm penalisation, and Wang and Zhu [25] suggested that penalisation should be based on grouping parameters together within a variable. More recently Guo et al [10] recommended that a pairwise penalty on pairs of cluster centres should be used. Nearly all of this work has been focused on the Gaussian mixture model.

Specific work on indicator selection in latent class analysis however, is sparse. The major work is by Dean and Raftery [5], and is an extension of their 2006 paper referred to above. Very recently, an online paper by Bartolucci et al [2] has addressed a latent class analysis problem on the quality of life of elderly nursing home patients with large N and J. These two papers form the motivation for our work. Dean and Raftery [5] proposed the following procedure for variable selection.

Stage 1. Determination of the initial number of classes.

Dean and Raftery state that the number of classes chosen should be "the largest number of classes that can be identified" from a latent class model with all variables. In practice, this is taken to be the number of classes K that minimises the BIC over a range of different values of K.

When fitting a latent class model, Dean and Raftery, in common with others writing on latent class models, suggest taking a number of starting values to avoid local maxima of the likelihood.

Stage 2. Identification of a small initial set of indicator variables.

First, we determine the smallest number of variables J^* that allow a latent class model with K classes to be identified. This is achieved by using Goodman's formula for identifiability of a latent class model [11]. For

binary indicator variables, this states that the inequality below must be satisfied

$$2^J > (2 * J - J + 1) \times K \quad \ldots(5)$$

Once J^* has been determined, the indicator variables are ranked into order according to the variance of the q_{jk} across the K classes. The J^* variables with the highest variance are taken to be the initial set. (An alternative scheme is also proposed in cases where the data cannot identify a latent class model for $K > 1$). The initial set of variables is denoted by Y(clust); all other indicator variables are in the set Y(other).

Stage 3. The iterative headlong search algorithm

Generally, at any iteration in the headlong algorithm , a stepwise procedure is performed, allowing both additions to and subtractions from Y(clust). Firstly, each of the variables in Y(other) are examined singly for potential inclusion into Y(clust). We call this candidate variable Y(?). Two quantities are constructed to test whether the new variable is important – the first is the BIC for the K-class latent class analysis for the indicator variables including Y(?) which we denote by BIC_K[Y(clust)+Y(?)], and the second is the sum of two BIC values- the BIC for the K-class latent class analysis without Y(?), and the BIC for the single class latent class analysis for Y(?) - we denote this second quantity as BIC_K[Y(clust)] + BIC_1[Y(?)]. The difference between these two quantities is then calculated, and if it is negative, the inclusion step stops and Y(?) is placed in the set Y(clust).

A exclusion step is then carried out. Each of the variables in Y(clust) is examined to see if they can be removed from the K-class latent class analysis. Again we label the candidate variable to be excluded by Y(?). The two quantities calculated now become BIC_K[Y(clust)] – the BIC for the K-class latent class analysis including Y(?) - and BIC_K[Y(clust)-Y(?)] + BIC_1[Y(?)] – the sum of the BICs for the K-class latent class analysis without Y(?), and the BIC for the single –class latent class analysis for Y(?). The difference is then examined and if the difference is positive, then the variable Y(?) is placed in the set Y(other) and the exclusion step stops

Thus at each iteration of stage 3, the set Y(clust) can both be augmented by one variable and reduced by removal of a second variable. The exclusion step is omitted at any iteration if the number of variables in Y(clust) is equal to J* (in other words, the latent class model could not be estimated with fewer variables). The algorithm (and thus the procedure) stops when there is no change in the membership of Y(clust) between successive iterations.

The Dean and Raftery algorithm can however be criticised on a number of criteria. Firstly, there is no attempt to estimate K as part of the algorithm; as the size and membership of Y(clust) may be different for different values of K.

Secondly, the algorithm is not easy to implement in practice using standard software- the criterion of inclusion or exclusion at each iteration of stage 3 is based on differences between two quantities rather than on an absolute criterion. Thirdly, the algorithm does not choose the best variable to include or exclude at each iteration, but instead chooses the *first* variable that gives an improvement in fit in each step

In the next section, we therefore propose a new algorithmic approach using constrained model fitting which addresses each of these concerns. We build on recent work by the paper by Bartolucci et al [2] who have made two important modifications to the Dean and Raftery algorithm. Firstly, they suggest that, for large numbers of indicators, starting the algorithm with different numbers of indictor variables may be worthwhile. Secondly, they point out that the absolute value of the sum of the BIC values rather than the difference between BICs is a better criterion and which can be used to select both indicators and the number of classes.

4. The Proposed Algorithm.

The starting point of our approach is to recognize that exclusion of an indicator variable from a latent class analysis has exactly the same meaning as placing a constraint on some of the q_{jk}. More exactly, the sum of the two BIC measures used in the headlong search algorithm – the BIC for the latent class model with Y(?) omitted together with the BIC for the single class latent class analysis for Y(?) alone - is exactly equal to the absolute BIC measure for a latent class analysis for Y(?) included {Y(clust), Y(?)} but with equality constraints placed on the q_{jk} relating to the indicator Y(?).

To see this, we suppose that variable Y(?) is a candidate variable for removal at an exclusion step, and the index for this variable is j_{out}. The equality constraint we propose requires that

$$q_{j_{out}k} = q_{j_{out}} . \quad \forall\, k \qquad \ldots(6)$$

Without loss of generality, we suppose that Y(clust) consists of the set of all J indicator variables, and that the candidate variable is J, the last variable of the set. This simplifies the mathematics and allows us to avoid set notation. The likelihood for the model for {Y(clust),Y(?)} with equality constraints on q_{jk} is

$$L(\lambda) = \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\left[\sum_{j=1}^{J-1} q_{jk}^{y_{ij}}\ (1-q_{jk})^{1-y_{ij}} + q_{J\cdot}^{y_{ij}}\,(1-q_{J\cdot})^{1-y_{ij}}\right]$$

$$= \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\left[\sum_{j=1}^{J-1} q_{jk}^{y_{ij}}\ (1-q_{jk})^{1-y_{ij}}\right] + \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\left[q_{J\cdot}^{y_{ij}}\,(1-q_{J\cdot})^{1-y_{ij}}\right]$$

$$= \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k\left[\sum_{j=1}^{J-1} q_{jk}^{y_{ij}}\ (1-q_{jk})^{1-y_{ij}}\right] + \prod_{i=1}^{n} q_{J\cdot}^{y_{ij}}\,(1-q_{J\cdot})^{1-y_{ij}}.$$

...(7)

This is the sum of the likelihood of a K-class latent class model on Y(clust) without Y(?) and the likelihood for a single class latent class model for Y(?) alone (ie a Bernoulli likelihood). This means that we no longer need to fit sequences of latent class models with different sets of variables in Y(clust); instead we can fit latent class models to the full set of J variables, but with constraints on the q_{Jk} which relate to subsets of the J variables. Moreover, the absolute value of BIC under different constraint can be used to determine the best model.

Moreover, the absolute value of BIC can be used to simultaneously determine the best choice of model for different values of the number of classes K – we can optimise over both K and membership of Y(clust).

Unlike Dean and Raftery, we take a backward search approach as this avoids the selection of an initial starting set of variables.

We therefore propose a new procedure:

Stage 1.

Fit a one-class latent class analysis to all J variables, and calculate the BIC. Call this BIC_1.

Stage 2.

Increase the number of latent classes by 1.

Stage 3.

For a K-class latent class analysis, we define three sets of indicator variables – Y(clust) – the set of variables included in the K-class analysis, Y(?) - the variable under consideration, and Y(out) - the set of variables omitted from the K-class analysis. At the start, Y(out) is null, and Y(clust) consists of all J indicator variables. We first fit a full J-variable latent class analysis and calculate the BIC –call this $BIC_K(J)$. We then perform

a series of backward elimination steps, testing each of the J variables in turn by setting Y(?)=Y(j) for j=1...J, setting the appropriate equality constraints on the q_{jk}, fitting the constrained latent class model and recording the resulting value of BIC. We set the lowest value of BIC at this step to be $BIC_K(J-1)$ and record the index of the variable omitted which we call j_{out}. We then add $Y(j_{out})$ to Y(out) and remove $Y(j_{out})$ from Y(clust).

A general backward elimination step will seek to decrease the BIC. Starting from the BIC value for the remaining J_{clust} variables in Y(clust) - $BIC_K(J_{clust})$ - we search for the next variable $Y(j_{out})$ which gives the largest decrease in BIC – call this value $BIC_K(J_{clust}-1)$. Fitting a latent class model at this general step would involve constraining suitable sets of parameters q_{jk} as follows:

$$q_{jk} = q_{j\cdot} \quad \forall\, k \text{ and for each } j \text{ where } Y(j) \in Y(\text{out}) \qquad \ldots(8)$$

This variable $Y(j_{out})$ is then placed in the set Y(out) and is removed from Y(clust).

The backward elimination steps continue either until J_{clust} is too small to allow the latent class model to be estimated – i.e.

$$2^{J_{clust}} \leq (2 * J_{clust} - J_{clust} + 1) \times K \qquad \ldots(9)$$

or until $BIC_K(J_{clust}-1) > BIC_K(J_{clust})$ - i.e. there is no decrease in BIC in eliminating a variable.

Stage 4.

Repeat stages 2 and 3 until $BIC_{K+1}(j) > BIC_K(j)$ for all j, or, alternatively, stop at a pre-determined value of K.

The BIC trajectories $BIC_K(j)$ can then be plotted against j for each value of K, and the global minimum value of BIC over both K and the variable selection procedure can be found.

Note that an amended version of the above algorithm would replace the iterative backward elimination at stage 3 by a *backward stepwise* algorithm, which would also allow variables in the set Y(out) to be tested for inclusion in Y(clust), as well as allowing variables in Y(clust) to be tested for exclusion. In this case, the BIC trajectories would not be univalued functions of j but could be multivalued functions (with potentially more than one value of BIC for some values of j). This could give a zigzag appearance of the trajectory if inclusion steps are included.

5.The Example Dataset.

We illustrate the ideas in this paper by reanalysing a dataset on the symptoms of 30 psychiatric patients. These patients were examined by an

experienced psychiatrist and the presence or absence of 23 psychiatric symptoms was recorded. The data is presented in Table 1.

Table 1 – Symptom Data for 30 Psychiatric Patients

Symptom number	Symptom name	Patient (1 to 30) (x= presence, · = absence)
1	Disorientation	..............................x..
2	obsession/compulsion	...x..........................
3	memory impairment	..................x........x..
4	lack of emotion	............x..............x
5	antisocial impulses or acts	....x............xx..........
6	speech disorganization	......................x...x.x
7	overt anger	.....x........x........x....
8	grandiosity	..........x..x.....x..x.....
9	drug abuse	x...x.........x.........x...
10	alcohol abuse	.....x...........xx.....x.x..
11	retardation	................x..xx....x.x
12	belligerence/negativism	...........x.....xxx.....x...
13	somatic concerns	..x......xx.....x........x.xx
14	suspicion/ideas of persecution	............xxx.......xx...xx
15	hallucinations/delusions	............xxx.......xx...xx
16	agitation/excitement	.....x......x.xx.......xxx..x.
17	suicide	.x....xxx..xx...xx...xx.....xx
18	anxiety	.xxx..xxxxxx...x.x..xxx.xxx..x
19	social isolation	x......xx.xxxx..xx.xxxx.x.xxxx
20	inappropriate affect or behaviour	...xx.x..x....xxxx.xxxxxxxxxxx
21	depression	xxx...xxxxxxxx..xx.xxxx..xx.xx
22	leisure time impairment	..xxxxxxxxxxxx.xxxxxxxxxxx.x
23	daily routine impairment	..xxxxxxxxxxxx.xxxxxxxxxxxxx

The original data was first described by [22] and listed in [3]. Van Mechelen and De Boeck [22] report that the purpose of the study was to determine the taxonomic structure of an individual psychiatric diagnostician. They also report that the patients were 16 men and 14 women, with ages ranging from 18 to 77 years, although individual age and gender information is not provided in the dataset.
We can immediately notice that there are nearly the same number of patients (N=30) to symptoms (J=23). This will present problems for standard latent class analysis, with the potential for multiple maxima of the likelihood surface.

5.1 Previous Analyses of the Symptoms Data.

The original paper by van Mechelen and de Boeck [22] suggested that latent class analysis can only handle a relatively small number of symptoms, and they instead approached the classification problem through a HICLAS analysis [4] – a form of hierarchical probabilistic clustering. Their paper found four patient classes, essentially defined by a suicide group, a social isolation group (both of which may be depressive or anxious), a delusional group and a substance abuse group. 11 symptoms were placed in an undefined class of symptoms and were not felt to contribute to the analysis.

A reanalysis of the data in 2003 was carried out by Berkof et al [3]. They used the data to illustrate a Bayesian latent class approach. Using three distinct forms of the Beta distribution as prior distributions for the class symptom probabilities q_{jk}, and a Dirichlet (1,1) distribution for the mixture probabilities, their analysis came to the conclusion that there was no preference between the three, four and five class solutions, being equally preferred. Additionally, for some choices of the Beta distribution prior, a two or one group solution was preferred. The Bayesian analysis is highly sensitive to the choice of prior, and makes the analysis problematic. In examining the three group solution, the authors found that Class 1 was associated with high probabilities on the symptoms agitation/excitement, suspicion/ideas of persecution, and hallucinations/delusions, and identified this as being indicative of a psychosis syndrome. Class 2 was associated with depression, anxiety, and suicide and was interpreted as an affective syndrome, while Class 3 was associated primarily with alcohol abuse.

Most recently, Aitkin et al [1] have also suggested that a three class solution is optimal, but with a four class solution also strongly supported. They again took a Bayesian latent class solution, using diffuse or reference priors for both the class symptom probabilities q_{jk} given class membership and mixture probabilities, and using posterior deviance distribution plots for inference on the number of classes.

While the consensus is towards a three class solution, we note that a frequentist latent class analysis has not been carried out, and also that, given that the number of cases is nearly equal to the number of indicators, the issue of variable selection has not been addressed.

6. Results,

We first fit a standard latent class analysis to all 23 symptom variables, over a range of values of K from 1 to 5. A large number of different random start values were chosen for each latent class analysis to ensure as far as possible that a global maximisation of the likelihood was found; we required that the smallest value of -2 log L needed to be repeated at least four times from different start values. For two and three latent classes, 100 start values were needed, for four latent classes, 1000 and for five latent classes, 10000 start values were required.

Table 2 – BIC Values for a standard latent class solUtion with no variable selection for one up to five latent classes

Number of classes K	1	2	3	4	5
BIC value	684.77	694.16	714.52	762.15	816.10
Number of starting values	1	100	100	1000	10000

Table 2 gives the resultant BIC values. A surprising result is immediately seen – based on the BIC measure, there is no evidence of a latent class structure in the data – the best model is a one- class model. Clearly, the large number of variables combined with the relatively small number of cases makes such a conclusion suspect. We therefore proceed with variable selection.

We first turn to the Dean and Raftery method. They suggest starting a variable selection procedure by determining the number of classes from examination of the BICs from a sequence of latent class analyses which use the full number of variables However, we have just seen that this method suggests a one-class solution and there is no clear recommendation on how to choose the number of classes.

The new variable selection procedure suggested in this paper allows us to proceed. For each selection step, we again adopted the same number of random starting values listed in Table 1 -100 for two and three classes and 1000 for four classes. We adopted a backward stepwise procedure allowing for both inclusion and exclusion steps.

Figure 1. BIC trajectories for variable selection: one to four latent classes

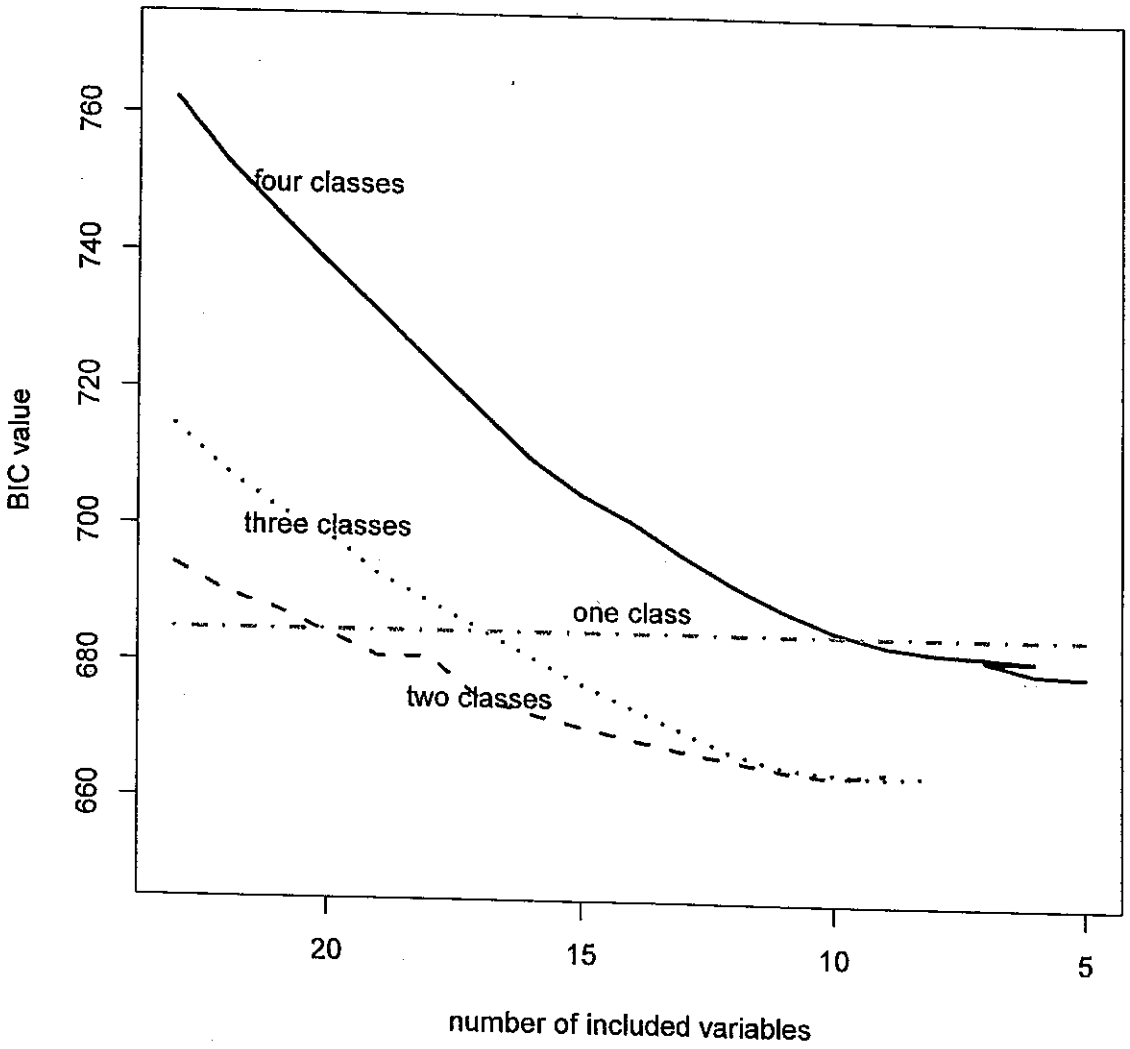

Figure 1 shows the BIC trajectory plots for one to four latent classes. The BIC values with 23 included variables are identical to those appearing in Table 1; thereafter the variable selection procedure reduces the BIC. Thus, fitting a two- class latent class model and carrying out a backward stepwise procedure gradually reduces the BIC from 694.16 to a minimum BIC of 664.01 with ten included variables; at that point, the BIC increases if further selection steps are attempted. The three-class model gives us a minimum BIC of 664.06 with nine included variables, and a four class model similarly gives a minimum BIC of 679.40 with five included variables. The global minimum of the BIC over K is found at the two class solution, but the two and three class values of the BIC are very similar.

Table 3 VARIABLE EXCLUSIONS and INCLUSIONS at each step of the selection procedure – two, three and four latent classes.

Number of latent classes	Variables included or excluded at each step of the selection procedure	Final selected variables
2	8(E); 23(E); 11(E); 12(E); 22(E); 9(E); 6(E); 15(E); 14(E); 4(E); 2(E); 1(E); 20(E)	3,5,7,10,13,16,17,18,19,21
3	9(E); 23(E); 2(E); 11(E); 20(E); 19(E), 22(E); 8(E); 1(E); 12(E); 13(E); 7(E); 6(E); 4(E)	3,5,10,14,15,16,17,18,21
4	9(E); 11(E); 23(E); 2(E); 1(E); 22(E); 12(E); 20(E); 3(E); 6(E); 8(E); 7(E); 18(E); 13(E); 5(E); 19(E); 17(E); 7(I); 21(E); 10(E)	4,7,14,15,16

Note: (E) indicates an exclusion step and (I) represents an inclusion step. In this example, all steps for two and three classes are exclusion steps; whereas there is an inclusion step for four classes.

Table 3 gives more details about the individual selections at each step. Thus, for the two class solution, the first step was an exclusion step with indicator variable 8 omitted; the next step was also an exclusion with variable 22 omitted, and so on. It is notable that even though we followed a backward *stepwise* selection, all steps for the two and three latent class analyses were exclusion steps. The four latent class analysis, however has an inclusion step, with variable 7 being reincluded at step 18 after being excluded at step 12. This is reflected in the zigzag nature of the four group trajectory in Figure 1.

Table 3 also gives the final variable selection for the two, three and four latent class models. It should be noted that the choice of variables is somewhat different for the various latent class models. With more classes, different variables are needed, and it is not simply that on set of variables is a subset of the other. Thus variables 14 and 15 are required in the three class model but not in the two class mode and contribute to defining the extra class; conversely variables 7 and 19 are needed in the two class model but do not contribute sufficient extra information for the three class model.

TaBLE 4. Class symptom probabilities of class membership for the BEST two class model

Variable	Class 1 (63.0%)	Class 2 (37.0%)
3. Memory impairment	0.0	0.1800
5. Antisocial impulses or acts	0.0	0.2700
7. Overt anger	0.0	0.2700
10. Alcohol abuse	0.0	0.4500
13. Somatic concerns	0.3706	0.0
16. Agitation/excitement	0.1059	0.5400
17. Suicide	0.6353	0.0
18. Anxiety	0.7406	0.3610
19 Social Isolation	0.7882	0.2799
21. Depression	1.0000	0.1899

TaBLE 5. class symptom probabilities of class membership for the BEST tHREE class model

Variable	Class 1 (59.9%)	Class 2 (23.3%)	Class3 (16.8%)
3. Memory impairment	0.0	0.0	0.3972
5. Antisocial impulses or acts	0.0	0.0	0.5957
10. Alcohol abuse	0.0558	0.0	0.7937
14. Suspicion/ideas of persecution	0.0	1.0000	0.0
15. Hallucinations/delusions	0.0	1.0000	0.0
16. Agitation/excitement	0.1115	0.7143	0.1980
17. Suicide	0.5567	0.2857	0.0
18. Anxiety	0.8350	0.4286	0.0
21. Depression	0.9442	0.4286	0.2063

Tables 4 and 5 show the estimated class symptom probabilities of class membership q_{jk} and class proportions π_k (converted to percentages) for the selected variables contributing to the two and three class models. Table 4 identifies a large class (63%) with high probabilities q_{Jk}on variables related to anxiety and depression, and a smaller class (37%) with mid-range probabilities on anger, anti-social behaviour, and excitement; these latter variables have close to zero probabilities for q_{Jk} on the first class. The three class model is similarly described in Table 5. Class 1, representing patients with depressive and anxiety states, essentially remains the same as Class 1 for the two-class model, whereas classes 2 and 3 arise from splitting class 2 of the two-class model. The new class 2 now represents patients with paranoia and delusional symptoms, whereas the new group 3 now represents patients with alcoholic and antisocial behaviour.

Finally, we comment on the assignment of patients to groups. Both the two and three class models assign patients to their modal group (with the highest value of p_{ik}) with assignment probabilities of 0.95 or over, with excellent allocation of patients to classes in both models.

7. Discussion and Conclusions

We first comment on the data analysis of the example. We have found, after indicator variable selection, that there are either two or three classes of patients, with very similar BIC values. The three class solution with nine variables is preferred by us over the two group solution with ten variables as it uses a different selection of variables to identify a psychotic group of patients, and an alcohol abuse group, as well as the affective syndrome group identified in both the two class solution.

The results of our analysis are similar in some respects to previous analyses, and different in other respects. Other authors have suggested three or four group solutions. The three group solution of Berkhof et al [3] is close to our own three group solution in identifying groups of psychosis, affective syndrome and alcohol abuse. The four group solution of Van Mechelen and de Boeck [22] essentially splits the affective syndrome group into two, distinguishing between those with and without suicidal tendencies. An important difference in our work is that all patients are well allocated to classes with probability 0.95 or above. Berkhof et al [3] allocate only 21 of their patients to a group with probability 0.9 or above, whereas Van Mechelen and de Boeck [22] fail to allocate one case entirely. Variable selection appears to have reduced classification noise in the data by identifying the most important of the indicators.

We next consider the issue of variable selection. Van Mechelen and De Boeck [22] implicitly select 12 of the 23 variables as contributing to their analysis; however they included variables such as leisure time impairment, daily routine impairment and inappropriate affect which did not contribute to our

analysis. Berkhof et al [3], in contrast, do not consider the issue of variable selection, although in describing their three group solution in words, they use just seven of the variables (agitation/excitement, suspicion/ideas of persecution, hallucinations/delusions, depression, anxiety, suicide, and alcohol abuse.

Turning now to the methodology, we are proposing that a constrained latent class analysis framework provides a relatively straightforward procedure for guiding a latent class analysis both on the choice of variables and also the number of classes. BIC trajectory plots will help in explaining such decisions to clinicians. The procedure is particularly important where there is a large number of indicator variables.

It might be argued that fitting constrained latent class models is not strwaightforward but, in fact, the ability to constrain arbitrary subsets of the q_{jk} is available in a number of standard software packages for latent class analysis. The latent variable package M*plus* [16] fits latent class models and allows subsets of parameters including the q_{jk} to be constrained through its MODEL command [7]. The SAS add-on procedure PROC LCA [12] similarly allows the *rho* parameters (the equivalent of the q_{jk} in this paper) to be constrained in any desired way by specifying a matrix of constraint values through its RESTRICT clause. Most useful, however, for this paper is the facility for constraining estimates in Latent Gold [23]. This stand-alone package, used by us in this paper, allows the q_{jk} parameters to be constrained across latent classes by excluding an indicator variable's effect in the analysis. The package has a windows interface that allow constraints to be added or removed with a single click.

The method can be extended in various ways. Although this paper focuses on dichotomous indicator variables, the method can be easily extended to other forms of data – polytomous data, count data or continuous data, or mixtures of these different types. Latent class models involving other types of variable are more often referred to as mixture models, and similar procedures based on the BIC can be used. Another issue is the choice of BIC as the criterion to compare and assess models. In general the consensus is that BIC is the preferred measure to compare models in latent class analysis [17]. However, other authors have suggested that other forms of information criteria may be preferred. Thus Dias [6] suggests that AIC3 may be a better choice than BIC for latent class models; whereas Lin and Dayton [15] have suggested the use of a corrected AIC (CAIC) or Schwartz SIC statistic. It is straightforward to use these alternative criteria in our procedure.

Finally, other extensions need to be considered. One natural direction of development is to extend this procedure to models with covariates, although this would require optimization over the number of classes, choice of indicators and choice of covariates.

Research Sponsors.

We are grateful to the Australian Research Council (grant number DP120102902) for funding work on latent lcass models, and to King Saud University for travel support

References:

[1] Aitkin M, Duy V, Francis B. A new Bayesian approach for determining the number of components in a finite mixture. Submitted to *Computational Statistics and Data Analysis* 2015. Under revision.

[2] Bartolucci F., Montinari GE, Pandolfi S. Item selection by latent class methods. 2014. Online at arXiv:1407.3912v1

[3] Berkhof J, van Mechelen I, Gelman A. A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica* 2003; 13: 423-442.

[4] De Boeck P, Rosenberg, S. Hierarchical classes: Model and data analysis. *Psychometrika* 1988; 53: 361-381.

[5] Dean N, Raftery AE. Latent class analysis variable selection. Annals of the Institute of Statistical Mathematics 2010; 62; 11-35.

[6] Dias J. Latent class analysis and model selection. In Spiliopoulou M, Kruse R, Borgelt C, Nürnberger A, Gaul W. (Eds.), From data and information analysis to knowledge engineering. Springer: New York, 2006; 95-102.

[7] Finch WH, Bronk KC. Conducting Confirmatory Latent Class Analysis Using M*plus*. *Structural Equation Modeling* 2011; 18: 132-151.

[8] Formann A, Kohlmann T. Latent class analysis in medical research. *Statistical Methods in Medical Research* 1996; 5(2): 179-211.

[9] Friedman JH. Meulman JJ. Clustering objects on subsets of attributes. *Journal of Royal Statistical Society Series B* 2004 ; 66(4): 815-849.

[10] Guo J, Levian E, Michailidis G, Zhu J. Pairwise Variable Selection for High-Dimensional Model-Based Clustering. *Biometrics* 2010; 66: 793-804.

[11] Goodman L. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; 61: 215-231.

[12] Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS Procedure for Latent Class Analysis. Structural Equation Modeling 2007; 14(4): 671–694

[13] Lazarsfeld PF. The logical and mathematical foundations of latent structure analysis. In S. A. Stouffer et al (Eds.), Measurement and prediction, the American soldier: studies in social psychology in World War II (Vol. IV, Chap. 10). Princeton, NJ: Princeton University Press, 1950; 362–412.

[14] Lazarsfeld PF, Henry NW. Latent structure analysis. Houghton Mifflin: Boston, 1968.

[15] Lin TH, Dayton CM. Model Selection Information Criteria for Non-Nested Latent Class Models, Journal of Educational and Behavioral Statistics 1997: 22(3) 249-264.

[16] Muthén LK, Muthén BO. Mplus User's Guide. Sixth Edition. Muthén & Muthén: Los Angeles, CA, 2010.

[17] Nylund KL, Asparouhov T, Muthén BO. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. Structural Equation Modeling 2007;14(4): 535–569.

[18] Raftery AE, Dean N. Variable selection for model-based clustering. *Journal of the American Statistical Association* 2006; 101: 168-178.

[19] Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Statistics in Medicine* 1986; 5(1): 21-27.

[20] Soothill K, Francis B, Awwad F, Morris SM, Thomas C, Mcillmurrary M. Grouping cancer patients by psychosocial needs. *Journal of Psychosocial Oncology* 2004; 22(2): 89-109.

[21] Stewart J, Sapey B, Humphreys L, Francis B, Donaldson G. Older People and Dissatisfaction with Wheelchair Services. Scandinavian Journal of *Disability Research* 2008; 10(1): 17-28.

[22] Van Mechelen I, de Boeck P. Implicit Taxonomy in psychiatric Diagnosis: a Case Study. *Journal of Social and Clinical psychology* 1989; 8(3): 276-287

[23] Vermunt JK, Magidson J. Technical guide for Latent GOLD 5.0: Basic, advanced and syntax. Statistical Innovations Inc.: Belmont, Massachusetts, 2013.

[24] Xie B, Pan W, Shen X. Variable Selection in Penalized Model-Based Clustering Via Regularization on Grouped Parameters. *Biometrics* 2008; 64(3): 921-930.

[25] Wang S, Zhu J. Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. *Biometrics* 2008; 64: 440-448.