Metodološki zvezki, Vol. 3, No. 1, 2006, 147-162

Latent Class Analysis Identification of Syndromes in Alzheimer's Disease: A Bayesian Approach

Cathal D. Walsh[1]

Abstract

Latent variable models have been used extensively in the social sciences. In this work a latent class analysis is used to identify syndromes within Alzheimer's disease. The fitting of the model is done in a Bayesian framework, and this is examined in detail here. In particular, the label switching problem is identified, and solutions presented. Graphical summaries of the posterior distribution are included.

1 Introduction

Latent Class Analysis (LCA) is the use of a discrete latent variable to model a situation where there are a number of categorical response variables of interest. These models have been used extensively in the social sciences to model heterogeneity of manifest responses in a multivariate sense. Many examples and guidance on practical fitting strategies may be found in Hagenaars and McCutcheon (2002). Examples of the use of LCA in the context of medical diagnosis date back to Young (1983). The methods used in this paper draw on Bayesian strategies for fitting these models, an overview of which can be found in Garrett and Zeger (2000).

2 Model

The model for LCA can be described in terms of manifest variables $\mathbf{x}$, and latent categorical variable $\mathbf{z}$. In this case, interest is on manifest variables which consist of a number of binary indicators for each individual, being presence or absence of a particular disease symptom. Let $\boldsymbol{x_j}$ be the response vector of individual j taken from a sample of J individuals. Then x_{ij} is the presence or absence of symptom

[1] Visiting Research Fellow, Mathematics and Statistics, QUT, Gardens Point, Brisbane, QLD 4001, Australia; Cathal.Walsh@tcd.ie

$i \in 1,...,I$. Let π_{ik} be the probability of a positive response on variable i for a person in class z_j=k $k \in 1,...,K$ and η_k be the probability that a randomly chosen individual is in class k. The conditional distribution of each x_{ij} is Bernouilli:

$$f(x_{ij} \mid z_j = k, \pi_{ik}) = \pi_{ik}^{x_{ij}} (1-\pi_{ik})^{(1-x_{ij})}.$$

Given the class $z_j=k$ and the j^{th} individual, independence yields:

$$f(\mathbf{x}_j \mid z_j = k) = \prod_{i=1}^{I} \pi_{ik}^{x_{ij}} (1-\pi_{ik})^{(1-x_{ij})}.$$

With K classes, the mixture becomes,

$$f(\mathbf{x}_j) = \sum_{k=1}^{K} \eta_k \prod_{i=1}^{I} \pi_{ik}^{x_{ij}} (1-\pi_{ik})^{(1-x_{ij})}.$$

The posterior probability that an individual with response x_j belongs to class k is;

$$h(z_j = k \mid \mathbf{x}_j) = \frac{\eta_k \prod_{i=1}^{I} \pi_{ik}^{x_{ij}} (1-\pi_{ik})^{(1-x_{ij})}}{f(\mathbf{x}_j)}.$$

Thus, conditioning on the unobservable class, yields a straightforward finite mixture of Bernoulli random variables. This class variable is unknown and some effort is spent on the identification of that for each individual.

3 Fitting

Fitting the latent class model involves standard techniques used to deal with missing labels. Thus, for example, in the likelihood framework the EM algorithm Dempster et al. (1977) is used.

In a Bayesian context, the missing labels are treated as parameters to be jointly estimated, and samples from the corresponding posterior can be obtained using MCMC.

3.1 EM algorithm

In the maximum likelihood setting the label information is treated as unknown, and this is completed before the parameters are estimated. Since this completion step is carried out with uncertain estimates of the parameters, it is repeated with the new estimates in an iterative fashion.

This is the most common method of fitting these models, and care must be taken to ensure that local maxima are not reached. This is done by using multiple restarts from different initial conditions.

The algorithm used to obtain point estimates of the parameters then proceeds as;

1. Choose an initial set of posterior probabilities $h(z_j = k \mid \mathbf{x}_j)$
2. Obtain a first approximation to $\hat{\eta}_k$ and $\hat{\pi}_{ik}$
3. Substitute these estimates into the expression for $h(z_j = k \mid \mathbf{x}_j)$ to get improved estimates
4. Return to stage 2 to get new approximations for $\hat{\eta}_k$ and $\hat{\pi}_{ik}$

This algorithm proceeds quickly, and there is virtually no computational overhead involved. In order to examine standard errors and goodness of fit statistics great care must be taken in the context of sparse data. Since the number of possible response patterns is large, 2^I, sparseness is a concern even where data on many hundreds have been obtained. Solutions to these problems include using bootstrap samples or lower order marginals for goodness of fit.

3.2 Bayesian

An alternative method of fitting these models is to use a fully Bayesian specification. This requires the model for the data, together with priors for the relevant parameters.

The model is as has been specified in Section 2. The priors can be obtained from specialists within the area of application. Alternatively, sensibly vague priors can be placed on the parameters. For example, a Dirichlet prior with equal weights on η would be considered to be ‘flat’ in the sense one would expect.

When fitting the model using MCMC a sequence of realisations of the parameters is available at each step, and derived summaries may usefully be presented in examination of model fit and interpretability. An outline of some diagnostics which can be of practical use is given in Garrett and Zeger (2000).

One of the referees emphasised that analogous methods can be used to explore the likelihood without using a fully Bayesian model. In an MLE framework, similar summaries may be constructed. The emphasis here, however, is how the Bayesian fitting proceeds.

4 Application to Alzheimer's disease

Alzheimer's disease is a degenerative condition which is affecting increasing numbers in a greying society. Largely due to improvements in health care and population based interventions, we have seen improving outcomes for those suffering from cardiac conditions and cancers. In contrast, however, no definitive cure exists for Alzheimer's Disease.

Research into this disease is highly multi-disciplinary involving psychiatrists, neuropsychologists, data managers, clinical psychologists and statisticians. The type of data that arise are complex and as identified by Kryscio and Schmitt (2000), more statisticians are needed in this area.

Of particular interest for this work, the clinical side of which is discussed in more detail in Moran et al. (2004), is the relationship between Behavioural and Psychiatric Symptoms of Dementia (BPSD) and the disease itself.

The working research hypothesis is that subclasses, or syndromes of the disease may exist. Further, it is supposed that these are the clinical phenotypes of the disease which may be related to genetic factors specific to the individual.

By identifying clinical phenotypes, genetic testing of individuals may be sampled in an efficient fashion - ensuring that the different syndromes are represented by the sample of individuals chosen. To this end, the probability of class membership will be a useful inferential summary.

4.1 Data description

The data in this case come from a memory clinic in St James's Hospital, Dublin. The Mercer's Institute houses the national memory clinic for Ireland and is the primary centre involved in the differential diagnosis of Alzheimer's disease. The sample of individuals to be examined was restricted to first visit patients with mild disease. This restriction was to ensure clinical homogeneity of the sample.

An examination of cases dealt with by the clinic revealed 240 first visits for individuals who were diagnosed as having probable disease according to the NINCDS-ADRDA criteria McKhann et al. (1984), and a Clinical Dementia Rating (CDR) Berg (1988) of 0.5 or 1.0.

This restriction to mild disease was made in order to ensure that the symptoms were related to syndrome rather than severity. This strategy ensured that a known source of heterogeneity was eliminated before the analysis began.

The Behave-AD Reisberg et al. (1996) instrument had been administered to the primary caregiver and this produces information on the prevalence of each symptom.

The Behave-AD produces scores on an ordinal scale for each of the symptoms. However, since this sample consisted of individuals in the mild stages of disease

the symptoms were each recorded as a binary variable. The symptoms of interest in this analysis were Hallucinations, Activity Disturbance, Aggression, Agitation, Diurnal Rhythm Disturbance and Affective disorder.

4.2 Data

Since the pattern of symptoms can be described by binary variables, it is convenient to write the combinations in the form $\{0,1\}^6$, so for example an individual exhibiting all symptoms would be denoted 111111, whereas an individual exhibiting none would be denoted 000000.

Thus, Table 1 summarises the data on all cases included.

Table 1: Data on prevalence of each symptom pattern.

Pattern	n	Pattern	n	Pattern	n
111111	3	011101	14	001101	2
111011	1	011011	2	001011	1
111001	1	011010	3	001010	1
110101	1	011001	9	001001	4
110011	2	011000	1	001000	2
110001	5	010111	11	000111	3
110000	2	010101	24	000110	1
101001	1	010100	3	000101	9
100101	1	010011	11	000100	3
100001	1	010010	2	000011	6
100000	1	010001	35	000010	1
011111	6	010000	20	000001	25
011110	1	001111	3	000000	18

5 Analysis

The model was fit in the maximum likelihood framework using LATCLASS Bartholomew and Knott (1999) and in the Bayesian framework using WinBUGS 1.4. Additional processing was carried out using R 1.8. R Dev Core Team (2005). All these packages ran on a 3.2GHz Pentium 4 PC under Windows XP.

5.1 Label switching

A key issue which arises when sampling from the joint posterior distribution is that the label that is sampled for each individual is assigned at each step of the

sampler. Since the label is a latent marker, the assignment of the particular label is unique only up to the permutation group. An example makes this clear.

A simulation study highlights what occurs. For this simulation, two latent groups were defined. The prevalence vector was set at η=[0.2,0.8] and the symptoms within group 1 were given with prevalence π_{11} =0.1, π_{21}=0.6 and in group 2 with prevalence π_{12}=0.9, π_{22}=0.3. The total number of individuals was set at 1000.

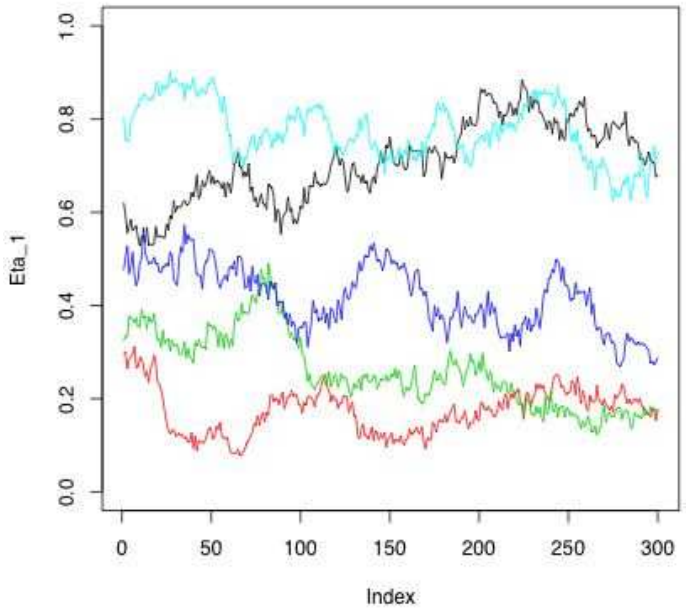

Figure 1: Chain for simulated data showing chains exploring different label space.

As is good practice, 5 chains were set running from different starting values. The output of the chain for η is shown in Figure 1. The between chain variability is influenced by the distance of two of the chains from the lower three. This is a cause for a concern when considering whether the chains have converged in distribution.

Of course, the issue here is that two of the chains have labelled individuals in one fashion whereas the other three have labeled in the other direction. By rearranging the columns of the sampled matrix, this is made clear. In particular, η_1 and η_2 have been switched for the top two chains and the resulting output is shown in Figure 2.

The arbitrary nature of the labels for latent mixtures and the difficulties caused for Bayesian inference is well known, and is discussed, for example, in Richardson and Green (1997). However, there are a number of strategies to deal with this problem some of which are discussed here.

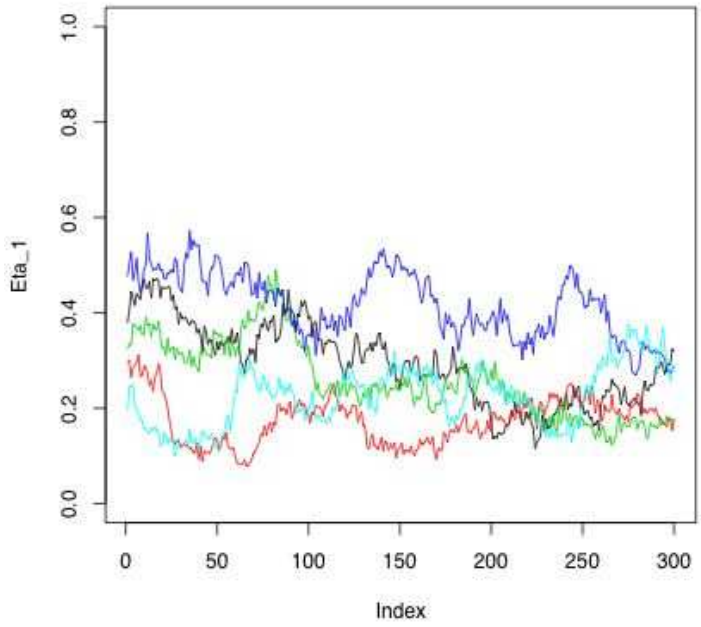

Figure 2: Transformed chains showing impact of label change.

The solution proposed in Richardson and Green (1997) is to place a constraint on the parameter space. This could be in η space, in π space or some combination of them. For illustration, the behaviour of the sampler in π space is shown in Figure 3. It is clear from this that the chains are exploring parts of the joint space which is divided by the line of symmetry. An arbitrary ordering of the parameters, either through the prior or post-hoc will break this symmetry and force identifiability.

It is noted here that the fact that the chains do not explore the whole space (in the case of the simulated data) means that they can not have (technically) converged. Indeed, this fact is described for mixture models in Celeux et al. (2000) where the authors suggest that "almost the entirety of samplers used for mixture models has not converged."

Of course, in order for the results to be usable, what is required is samples from the posterior, modulo the permutation group. One strategy is the truncation along symmetries as suggested by placing constraints on the parameters. An alternative is to 'gather' posterior samples together as described by Stephens (2000). Here the author suggests that a loss function can be used to 'suck' the samples in the joint posterior together. Using the ideas presented, an appropriate loss function for this model is represented by a product of Dirichlet and Betas.

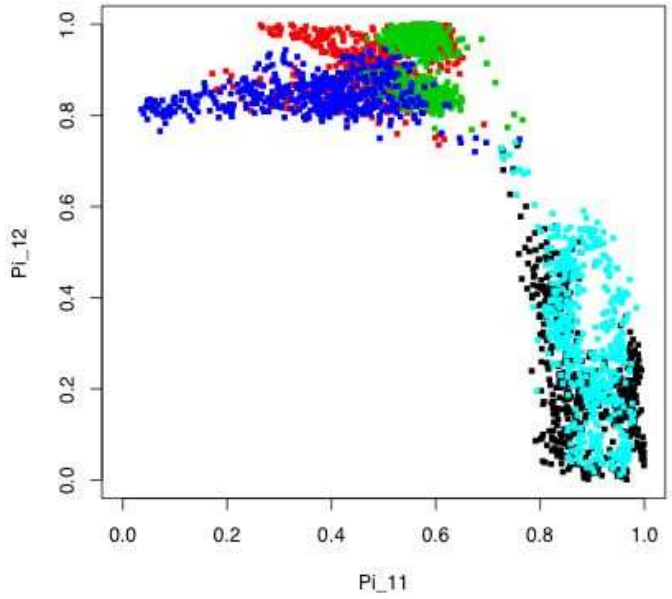

Figure 3: Transformed chains for π showing impact of label change.

Thus the algorithm for this situation is;

1. Calculated loss is based on $-\log(\cdot)$ of $\mathrm{Di}(\eta \mid \alpha)\prod_{i,k=1}^{I,K}\mathrm{Be}(\pi_{ik} \mid \theta_{ik})$.
2. Estimate α and θ from the initial set of samples.
3. Run through sampled values permuting labels to minimize the loss.
4. If changes of labels have occurred return to the start.

Offline, this algorithm has taken minutes with up to 4 classes, but of the order of an hour for 5 and many hours for 6 classes for 10,000 samples. The expense comes from the fact that the permutation group grows very quickly.

In practice, the algorithm will not change the results compared to the simple constraints for the situation given in our simulated example. This is largely due to the fact that there is a large amount of information in the data in this case. Thus the joint posterior is well defined and label switching within chains is unlikely.

However, for the case of the Alzheimer's data as recorded, a difference is observable. Due to the smaller amount of data, the information in the likelihood is less, and thus the joint posterior is flatter. This permits the sampler to commute across permutations of labels within a single chain. Indeed, a similar problem would occur with more information if the modes were closer - a feature of larger number of classes.

The Figure 4 shows the case of the sampler for η for the 3 class model.

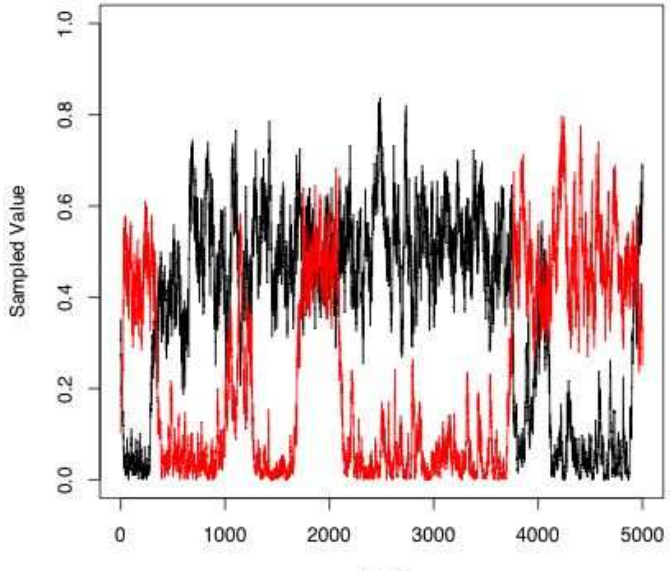

Figure 4: Chain for η showing switching.

5.2 Parameter constraints solution

Initially, constraints are placed on the η; $\eta_1 < \eta_2 < \eta_3$ and the chains are post-processed with this constraint. The result of this is to (somewhat artificially) separate out the chains. The picture in Figure 5 makes this fact clear.

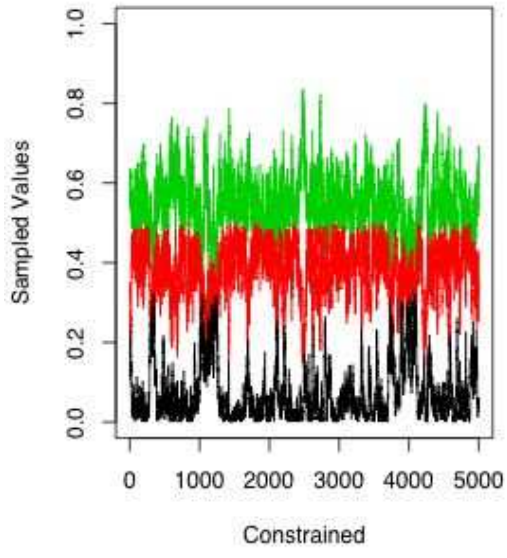

Figure 5: Chain for η showing constrained solution.

While the imposition of this constraint may seem sensible (there has to be a smallest class,) it is unreasonable to suggest that this will work in distribution. The impact is to truncate the joint posterior, which may have a strange effect on the marginal distributions.

An alternative is to use the loss function approach to group samples for parameters

5.3 Loss function processing

The strategy of post processing using the loss function described was implemented. This was a substantial computational overhead when compared with the constraint solution.

However, the advantage of the strategy is that the function jointly considers all parameters, and does not abruptly truncate any part of the joint distribution. The output is shown in Figure 6.

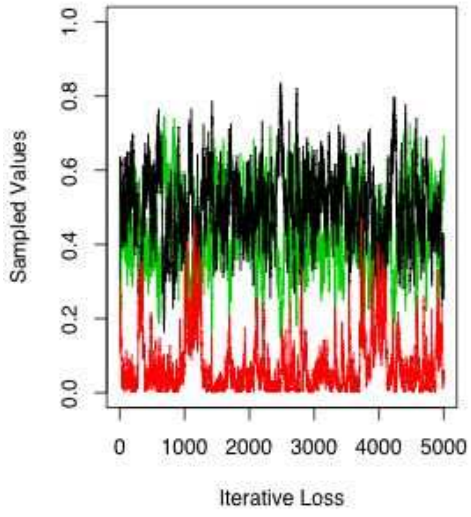

Figure 6: Chain for η showing loss switched solution.

6 Results

The results of the analysis are shown in the form of point estimates and standard errors. The maximum likelihood solution for the 3 class model is shown for comparison purposes. In the case of the Bayesian analysis, graphical summaries of the marginal posteriors are provided. Where standard errors are given in tabular format for the Bayesian summaries, these are the standard deviation of the sampled values of the parameters.

Table 2: Summaries for 3 class model. EM algorithm.

	class 0	**class 1**	**class 2**
Hallucination	0.09 (0.03)	0.01 (0.02)	0.77 (0.24)
Activity	0.57 (0.05)	0.79 (0.07)	0.98 (0.05)
Aggression	0.13 (0.04)	0.37 (0.08)	0.98 (0.06)
Agitation	0.11 (0.06)	0.82 (0.16)	0.73 (0.21)
Diurnal	0.16 (0.04)	0.33 (0.08)	0.97 (0.08)
Affective	0.63 (0.07)	0.96 (0.05)	0.99 (0.05)
η	0.58 (0.10)	0.39 (0.10)	0.03 (0.02)

It is noticeable in the summaries that the standard errors, particularly in the case of the small class, are smaller than one might expect in Table 2. This is for reasons already discussed. On the other hand, the Bayesian estimates in Table 3 give more realistic values.

In addition, graphical summaries, such as Figure 7 for η and Figure 8 for the π. These are easy to present to clinicians and they can get a feeling for the substantial uncertainty that exists about the estimates of the parameters in the small classes.

Table 3: Summaries for 3 class model. Bayesian Analysis.

	class 0	**class 1**	**class 2**
Hallucination	0.07 (0.03)	0.08 (0.04)	0.22 (0.24)
Activity	0.53 (0.06)	0.79 (0.07)	0.70 (0.26)
Aggression	0.09 (0.05)	0.36 (0.08)	0.55 (0.33)
Agitation	0.13 (0.06)	0.62 (0.11)	0.41 (0.30)
Diurnal	0.11 (0.05)	0.36 (0.07)	0.55 (0.30)
Affective	0.58 (0.08)	0.95 (0.04)	0.70 (0.28)
η	0.50 (0.06)	0.43 (0.05)	0.05 (0.05)

In addition to the summary of results presented here, other graphical tools outlined in Garrett (2000) have been used. In deciding on the number of classes, a posterior predictive frequency check, Figure 9, was used. This compares the observed number of individuals in a class with the posterior predicted number in each class. The observed number is marked as a digit on a plot, with the posterior median and interquartile range shown by solid and dotted lines respectively.

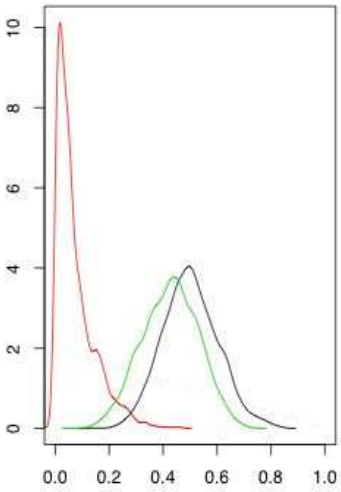

Figure 7: Marginal posterior estimates for η.

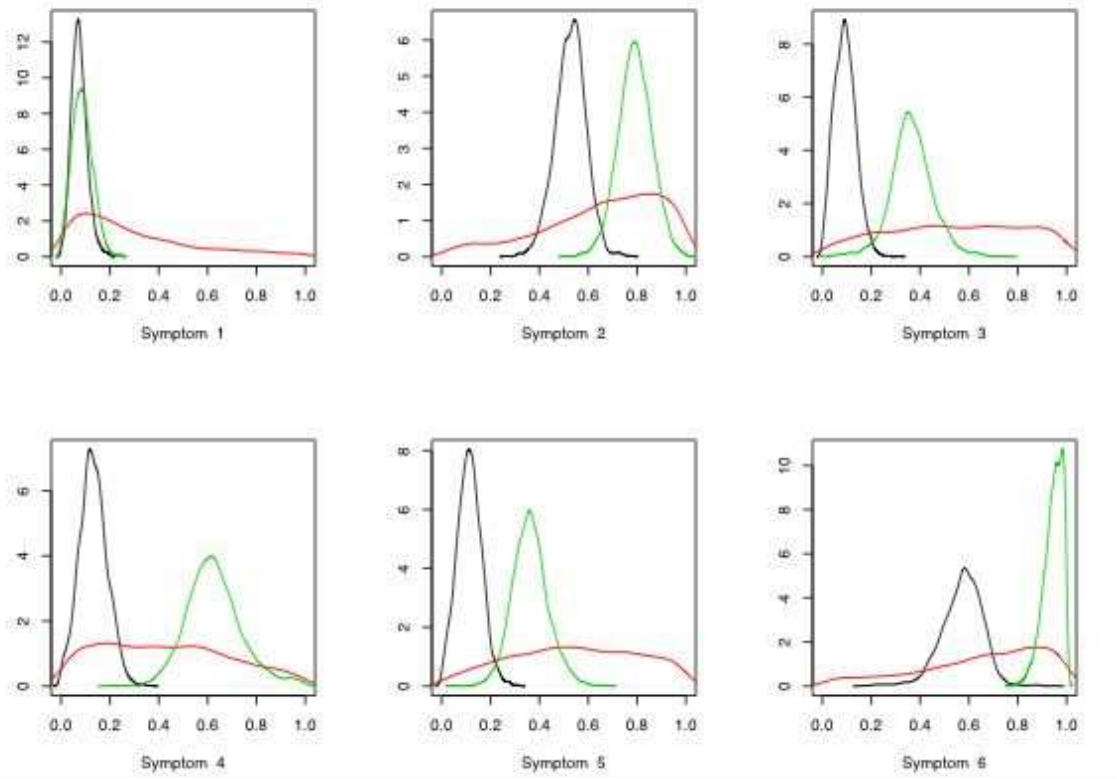

Figure 8: Marginal posterior estimates for π.

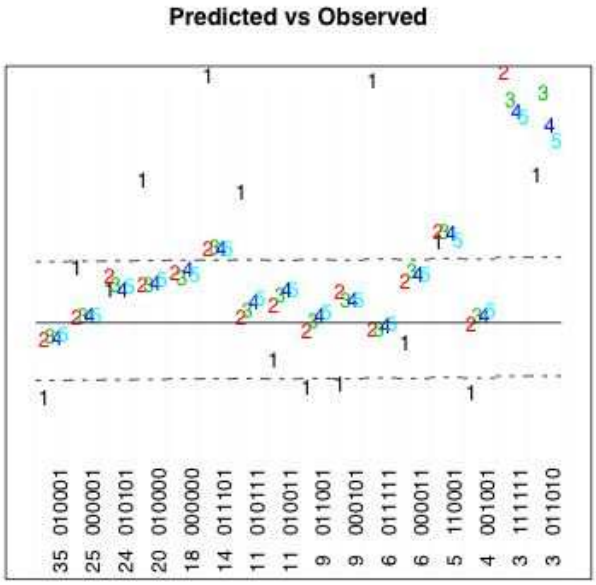

Figure 9: PFC for increasing numbers of classes. Annotations show the pattern of symptoms and number of individuals in that group.

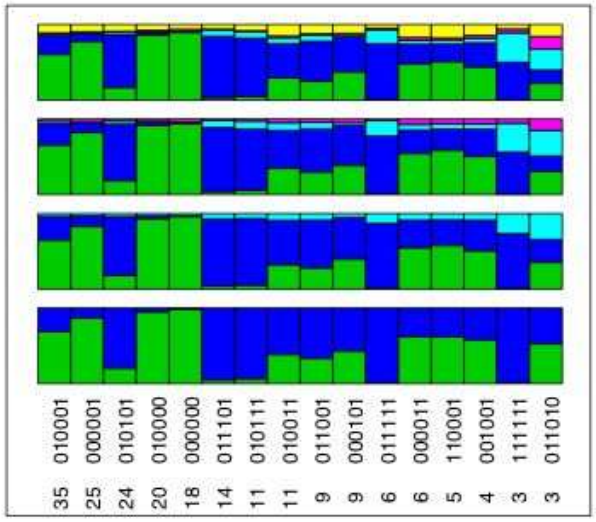

Figure 10: Probability of class membership.

In addition we summarise the probability of class membership given a particular pattern of the manifest symptoms. This is readily obtained from the posterior samples.

Models with different numbers of classes have been fitted and the resulting posterior probability of membership has been calculated in each case. These are presented in stacked fashion for different number of classes. The pattern of symptoms is shown along the x-axis ordered by the number of individuals in each class. Thus the estimated model is most heavily estimated by patterns to the left.

The bars show the probability of membership of each class, and the height of each bar thus adds to one (so no scale on the left hand side is required.)

This graphical summary is shown in Figure 10.

7 Discussion

From a clinical perspective, the latent groupings include a low symptom prevalence group, a higher symptom prevalence group and a group that includes those exhibiting hallucinations. The evidence in favour of the 3 class model is that there is a slightly better fit between observed and predicted. A 2 class model is not clinically interesting, because it shows only a high and low symptom group. The movement beyond 3 classes is not justified by an improvement in fit.

This work consists of a deeper examination of the methodology we used during the exploratory phase of Moran et al. (2004). The Bayesian analysis comes at a substantial computational cost and is hard to justify in this instance. Indeed, this was a point made by the referees in this case. However, the work emphasizes a number of important general lessons.

The joint posterior for sparse tables is moderately flat over large regions. In this situation, asymptotic results have to be used with great care. This is highlighted by for example Formann (2003).

When fitting a Bayesian mixture model, the problem of label switching can occur. Indeed, from a technical perspective if it does not occur then the samplers have not explored the full posterior. This is a point highlighted by Celeux et al. (2000) among others.

Resolving the difficulties caused by switching within chains can be done by imposing ordering constraints on the parameters. One way of thinking of these constraints is as the imposition of a prior structure on a model. This strategy is a popular one and has been implemented for mixtures by for example Richardson and Green (1997). This method essentially breaks the symmetry of the joint posterior distributions, but may result in artificial summaries for situations where a substantial posterior mass lies on the line of symmetry.

An alternative method of removing the impact of switching is to specify a loss function which ensures that the summaries are ‘well behaved’ (as defined by the loss.) This idea follows the logic of Stephens (2000) who examines mixtures of Normals. One such loss function was implemented here.

As a method of solving the switching problem, the loss function approach is expensive, but we posit is more realistic in high dimensional spaces than constraints unless the latter are chosen with great care.

One of the referees pointed out that there may be instances where the additional overhead may be worthwhile. For example, if there are known constraints on the parameters, or where there is substantial prior information, a Bayesian method of fitting may be considered.

8 Conclusions

The fitting of latent mixture models is now commonplace, as may be seen in work such as Hagenaars and McCutcheon (2002). The availability of software tools to do the necessary calculations means that it is quite quick and easy to fit such models. Routine diagnostics are produced and a particular model may be chosen.

This work demonstrates how these models are fitted within a Bayesian paradigm, the problems that may be encountered, and gives explicit guidance on their solution.

Acknowledgments

The author would like to thank the referees for their helpful comments, which has added to the clarity of this work. Many thanks to Murray Aitkin, Jim Smith and Kerrie Mengersen for comments on earlier drafts of this work. I hope some of them have been taken on board.

The author is grateful to Maria Moran and Aoibhinn Lynch for discussions regarding clinical aspects of this study. The data were collected in the memory clinic in St James's. Some funding for this work has been provided by the Dean of Research at Trinity College Dublin, which is being carried out in partial fulfilment of the promotional requirements of the University of Dublin.

References

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

[2] Berg, L. (1988): Clinical dementia rating (CDR). *Psychopharmacological Bulletin,* **24**, 637–639.

[3] Celeux, G., Hurn, M.A, and Robert, C.P. (2000): Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association,* **95**, 957-970.

[4] Dempster, A., Laird, N., and Rubin, D. (1977): Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society Series B,* **39,** 1–38.

[5] Formann, A.K. (2003): Latent class model diagnostics - a review and some proposals. *Computational Statistics and Data Analysis*, **41**,549–559.

[6] Garrett, E.S. and Zeger, S.L. (2000): Latent class model diagnosis. *Biometrics,* **56,** 1055-1067.

[7] Hagenaars, J.A. and McCutcheon, A.L. (2002): *Applied Latent Class Analysis.* Cambridge, UK: Cambridge University Press,.

[8] Kryscio, R.J. and Schmitt, F.A. (2000): Foreword to special issue on Statistics in Alzheimer's disease. *Statistics in Medicine*, **19,** 1389–1391.

[9] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E.M. (1984): Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology*, **34,** 939–944.

[10] Moran, M., Walsh, C., Lynch, A., Coen, R.F., Coakley, D., and Lawlor, B.A. (2004): Syndromes of behavioural and psychological symptoms in mild Alzheimer's disease. *International Journal of Geriatric Psychiatry*, **19,** 359–364.

[11] R Development Core Team. (2005): *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0.

[12] Reisberg, B., Auer, S.R., and Monterio, I.M. (1996): Behavioural pathology in Alzheimer's disease (BEHAVE-AD) rating scale. *International Psychogeriatrics*, **8**, 301–308.

[13] Richardson, S. and Green, P.J. (1997): On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society Series B*, **59**, 731–792.

[14] Stephens, M. (2000): Dealing with label switching in mixture models. *Journal of the Royal Statistical Society Series B*, **62**, 795–809.

[15] Young, M.A. (1983): Evaluating diagnostic criteria: a latent class paradigm. *Journal of Psychiatric Research*, **17**, 285–296.