Bettina Grün and Gertraud Malsiner-Walli

Bayesian Latent Class Analysis with Shrinkage Priors: An Application to the Hungarian Heart Disease Data

Bayesian latent class analysis with shrinkage priors: an application to the Hungarian heart disease data

Bettina Grün[*], Gertraud Malsiner-Walli[**]

Abstract: Latent class analysis explains dependency structures in multivariate categorical data by assuming the presence of latent classes. We investigate the specification of suitable priors for the Bayesian latent class model to determine the number of classes and perform variable selection. Estimation is possible using standard tools implementing general purpose Markov chain Monte Carlo sampling techniques such as the software JAGS. However, class specific inference requires suitable post-processing in order to eliminate label switching. The proposed Bayesian specification and analysis method is applied to the Hungarian heart disease data set to determine the number of classes and identify relevant variables and results are compared to those obtained with the standard prior for the component specific parameters.

Keywords: Bayesian latent class analysis, Shrinkage prior, Variable selection.

1. Introduction

Latent class analysis (LCA) is a modeling approach for categorical data originally proposed by Lazarsfeld (1950). The observed association between the manifest categorical variables is assumed to be caused by latent classes. Conditional on class membership the categorical variables are assumed to be independent given the class specific variable distributions.

Issues in LCA are the selection of the number of classes and the identification of relevant variables. Within the frequentist framework using maximum likelihood estimation Dean and Raftery (2010) investigated the use of the BIC in combination with a headlong search algorithm to explore the model space to determine a suitable number of classes as well as subset of variables. They illustrate their approach using the Hungarian heart disease data set. Alternatively, White et al. (2016) use stochastic search methods to select the number of classes and relevant variables within the Bayesian framework.

[*] Johannes Kepler Universität Linz, bettina.gruen@jku.at

[**] Wirtschaftsuniversität Wien, gertraud.malsiner-walli@wu.ac.at

In this paper we investigate the use of sparse finite mixture models in combination with shrinkage priors. Malsiner-Walli et al. (2016) proposed the sparse finite mixture model with shrinkage priors on the means for the Gaussian finite mixture model. We extend this approach to the Bayesian latent class model. We also indicate how a general purpose Markov chain Monte Carlo (MCMC) sampler such as JAGS (Just Another Gibbs Sampler; Plummer 2003) can be used to obtain draws from the posterior and present suitable post-processing tools of the MCMC draws to eliminate label switching. This proposed model specification and analysis strategy is used to reanalyze the Hungarian heart disease data set.

2. Bayesian latent class model

Assume there are n observations $\boldsymbol{y}_i$, $i = 1, \dots, n$ given. Each observation $\boldsymbol{y}_i$ is a vector of length J, i.e., J variables are observed and each element y_{ij} contains values in $\{1, \dots, L_j\}$ implying that each variable j is a categorical variable with $L_j \geq 2$ different values.

The latent class model for observations $\boldsymbol{y}_i$, $i = 1, \dots .n$ is given by

$$f(\boldsymbol{y}_i|\boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{J} \prod_{l=1}^{L_j} \theta_{k,jl}^{\mathbb{1}(y_{ij}=l)} \right],$$

where $\boldsymbol{\pi} = (\pi_k)_{k=1,\dots,K}$, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_{k,jl})_{k=1,\dots,K;j=1,\dots,J;l=1,\dots,L_j}$, $\mathbb{1}()$ is the indicator function, and

$$\sum_{k=1}^{K} \pi_k = 1, \qquad \pi_k \geq 0, \, \forall k,$$
$$\sum_{l=1}^{L_j} \theta_{k,jl} = 1, \, \forall k, j, \qquad \theta_{k,jl} > 0, \, \forall k, j, l.$$

2.1. Prior specification

The parameter vector consists of $(\boldsymbol{\pi}, \boldsymbol{\Theta})$. In Bayesian finite mixture modeling one assumes in general that the component weights $\boldsymbol{\pi}$ and the component specific parameters $\mathbf{\Theta}$ are a-priori independent and that the component specific parameters are independently identically distributed (at least conditional on some hyperparameters). Furthermore conditionally conjugate priors are used to simplify MCMC sampling.

Component weights

For the component weights $\boldsymbol{\pi}$ a Dirichlet prior is assumed with a single parameter e_0:

$$\boldsymbol{\pi} \sim \text{Dirichlet}(e_0, \ldots, e_0).$$

Rousseau and Mengersen (2011) show that e_0 is an influential parameter if an overfitted mixture model is estimated. Based on their results Malsiner-Walli et al. (2016) propose the sparse finite mixture model where an overfitting mixture with K, the number of components, much larger than the number of latent classes is fitted together with the specification of a very small and fixed value for e_0, e.g., $e_0 = 0.0001$. Under this prior setting the posterior of an overfitting mixture asymptotically concentrates on the region of the parameter space where superfluous components have negligible component weights instead of including duplicated components.

Standard prior for the component specific parameters

In Bayesian LCA one assumes that a-priori the parameters of the variables are independent within components. This implies that for each variable j and component k the component specific parameter vector $\boldsymbol{\theta}_{k,j\cdot}$ a-priori follows a Dirichlet distribution:

$$\boldsymbol{\theta}_{k,j\cdot} \sim \text{Dirichlet}(\boldsymbol{a}_j).$$

The value for $\boldsymbol{a}_j$ is selected to regularize the likelihood which in the case of an LCA model is often multi-modal, contains spurious modes and might have modes at the boundary of the parameter space.

Shrinkage prior for the component specific parameters

To shrink irrelevant variables towards a common Dirichlet parameter a hierarchical prior is specified on $\boldsymbol{a}_j$. For this purpose the Dirichlet parameter is re-parameterized into a mean and precision parameter plus a regularizing additive constant:

$$\begin{aligned} \boldsymbol{a}_j &= \boldsymbol{a}_{0,j} + \phi_j \boldsymbol{\mu}_j, & \boldsymbol{\mu}_j &\sim \text{Dirichlet}(\boldsymbol{m}_j),\ \forall j, \\ \phi_j &= \frac{1}{\lambda_j},\ \forall j, & \lambda_j &\sim \text{Gamma}(\nu_1, \nu_2),\ \forall j. \end{aligned}$$

Following Malsiner-Walli et al. (2016) we suggest to use $\nu_1 = \nu_2 = 0.5$. Furthermore we use uniform priors for $\boldsymbol{a}_{0,j}$ and $\boldsymbol{\mu}_j$, i.e., $\boldsymbol{a}_{0,j} = 1$ and $\boldsymbol{m}_j = \boldsymbol{1}$.

2.2. MCMC estimation

Estimation of the Bayesian latent class model consists of approximating the posterior distribution of $(\boldsymbol{\pi}, \boldsymbol{\Theta})$ using MCMC methods. Diebolt and Robert (1994) suggested to use data augmentation to facilitate MCMC estimation by adding the class memberships of the observations to the sampling scheme.

Standard prior for the component specific parameters

The sampling scheme is given by:

1. Draw the class memberships S_i for all observations $i = 1, \dots, n$:

$$S_i \sim \text{Multinomial}(1, \boldsymbol{p}_i), \qquad p_{ik} \propto \pi_k \prod_{j=1}^{J} \prod_{l=1}^{L_j} \theta_{k,jl}^{\mathbb{1}(y_{ij}=l)}.$$

2. Conditional on $\boldsymbol{S} = (S_i)_{i=1,\dots,n}$ draw $\boldsymbol{\pi}$ from a Dirichlet distribution:

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}(e_0 + n_1, \dots, e_0 + n_K), \\ n_k &= \sum_{i=1}^{n} \mathbb{1}(S_i = k) \quad \forall k = 1, \dots, K. \end{aligned}$$

3. Conditional on $\boldsymbol{S} = (S_i)_{i=1,\dots,n}$ draw $\boldsymbol{\theta}_{k,j\cdot}$ from a Dirichlet distribution:

$$\begin{aligned}\boldsymbol{\theta}_{k,j\cdot} &\sim \text{Dirichlet}(a_{j1} + n_{k,j1}, \dots, a_{jL_j} + n_{k,jL_j}),\\ n_{k,jl} &= \sum_{i=1}^{n} \mathbb{1}(S_i = k)\mathbb{1}(y_{ij} = l) \quad \forall k, j, l.\end{aligned}$$

In each MCMC iteration the class memberships $\boldsymbol{S}$ induce a partition of the observations into K_+ classes, i.e., the number of non-empty components for this draw. In the overfitting mixture setting with K much larger than the number of classes and e_0 very small $K_+ \ll K$ and the posterior distribution of K_+ can be used to estimate the number of classes. Malsiner-Walli et al. (2016) proposed to use the mode as suitable point estimate.

Shrinkage prior for the component specific parameters

An additional sampling step is required to sample the hyperparameter values:

4. Conditional on $\boldsymbol{\Theta}$, sample $\boldsymbol{\mu}_j$ and λ_j for all j.

Model specification in BUGS and estimation using JAGS

The BUGS (Bayesian inference Using Gibbs Sampling; Lunn et al. 2009) model description language allows the specification of a Bayesian model based on a directed acyclic graph which contains the data as well as all parameters as nodes and where the edges are implied by the hierarchical specification of the Bayesian model.

For a Bayesian finite mixture models which is estimated using data augmentation the model specification not only includes the data $\boldsymbol{y}$ and the parameters $(\boldsymbol{\pi}, \boldsymbol{\Theta})$ but also the class memberships $\boldsymbol{S}$. The BUGS model specification for the model including the shrinkage prior is given in Figure 1. Note that for the standard prior the parameter `a[j, 1:L[j]]` is fixed and the four lines of code defining the relationships for `a`, `mu`, `phi` and `lambda` are dropped.

The model is estimated within R using package **rjags**. Only a list containing the data in an array `Y`, the dimensions `n`, `J`, `L` and the parameters needs

```
model {
  for (i in 1:n) {
    for (j in 1:J) {
      Y[i, 1:L[j], j] ~ dmulti(theta[S[i], j, 1:L[j]], 1)
    }
    S[i] ~ dcat(pi[1:K])
  }
  for (j in 1:J) {
    for (k in 1:K) {
      theta[k, j, 1:L[j]] ~ ddirch(a[j, 1:L[j]])
    }
    a[j, 1:L[j]] <- a0[1:L[j]] + phi[j] * mu[j, 1:L[j]]
    mu[j, 1:L[j]] ~ ddirch(m[1:L[j]])
    phi[j] <- 1 / lambda[j]
    lambda[j] ~ dgamma(nu1, nu2)
  }
  pi[1:K] ~ ddirch(e0[1:K]);
}
```

Figure 1. BUGS model specification for the sparse latent class model with shrinkage priors.

to be specified. Note that `Y` needs to be given as an array of dimension `n` $\times \max(L_j) \times$ `J` containing zeros and ones to indicate the observed values. `n` corresponds to the number of observations, `J` to the number of variables and `L` is a vector containing the number of categories for each variable. In addition the parameters specified are the number of components `K` and a vector `e` of length `K` containing e_0. Furthermore, for the standard prior `a` is a vector of ones of length $\max(L_j)$, whereas for the shrinkage prior, `m` and `a0` are two vectors of ones of length $\max(L_j)$, and `nu1` and `nu2`, the parameters of the Gamma prior on the shrinkage parameter λ, are both set equal to 0.5.

Then the model is defined using `jags.model()` and samples are drawn using `jags.samples()` while monitoring the parameters of interest using the argument `variable.names`.

For the presented results the call to `jags.model()` included an `inits` argument to set a specific random seed for reproducibility and an `n.adapt` argument to increase the number of iterations for adaptation to 5,000. Then `jags.samples` is called using 100,000 number of iterations with a thinning of 10.

2.3. Post-processing

The number of filled components K_+ are determined for each draw and an estimate $\hat{K}_+$ is obtained using the mode of the posterior distribution. If there is a distinct class structure in the data the MCMC sampler usually converges quickly to this number of classes and a clear mode can be identified (see Malsiner-Walli et al. 2016).

Conditional on the number of classes selected the draws are post-processed in the following way to obtain an identified model with suitable class specific parameter estimates as well as class assignments of the observations.

1. Discard all draws where $K_+ \neq \hat{K}_+$.
2. Discard all parameter draws $\boldsymbol{\theta}_{k,\cdot,\cdot}$ for empty components.
3. For each draw relabel the components to minimize the misclassification rate between the class assignments of this draw and the class assignments of the last draw.

Note that this is a very simple strategy to obtain an identified model which will only work if the data has a clear class structure. More elaborate approaches to deal with label switching have been proposed and might be required in more complicated settings to obtain good results (see Papastamoulis 2016).

3. Analyzing the Hungarian heart disease data

The Hungarian heart disease data consists of 284 patients on 5 categorical variables. For more details on the categorical variables with their levels see Table 1. Dean and Raftery (2010) analyzed this data set with LCA. They used maximum likelihood estimation in combination with the BIC to perform a joint approach for variable selection and determining the number of classes. They compared the classification results obtained with LCA to the known diagnosis of heart disease (angiographic disease status) available in the data set. The known diagnosis has two categories: "$< 50\%$" indicating less than 50%

Variable	Level	Class 1	Class 2
Chest pain type	Typical Angina	0.06 (0.02)	0.01 (0.01)
	Atypical Angina	0.57 (0.06)	0.07 (0.04)
	Non-anginal pain	0.26 (0.04)	0.08 (0.04)
	Asymptomatic	0.10 (0.07)	0.84 (0.06)
Exercise induced	No	0.95 (0.03)	0.33 (0.11)
Angina	Yes	0.05 (0.03)	0.67 (0.11)
Gender	Female	0.36 (0.04)	0.15 (0.04)
	Male	0.64 (0.04)	0.85 (0.04)
Resting	Normal	0.81 (0.03)	0.77 (0.04)
Electrocardiographic	ST-T wave	0.15 (0.03)	0.21 (0.04)
results	Estes' criteria	0.04 (0.02)	0.02 (0.01)
Fasting blood sugar	False	0.94 (0.02)	0.90 (0.03)
>120 mg/dl	True	0.06 (0.02)	0.10 (0.03)

Table 1. Posterior mean (and posterior standard deviations) of the class specific parameters for the identified 2-class sparse LCA model.

diameter narrowing and "> 50" indicating more than 50% diameter narrowing in any major vessel.

3.1. Sparse finite mixture model

An overfitting mixture model is estimated using $e_0 = 0.0001$ and $K = 10$. In addition a uniform prior is assumed for the class specific parameters, i.e., $a_{k,jl} = 1$. The posterior distribution of the number of non-empty components K_+ has a clear mode at 2 with 99.7% of the samples having 2 non-empty components. The remaining samples had 3 non-empty components (0.2%). Using the samples with 2 non-empty components to identify the model results in a posterior mean estimate for the component weight of the larger class π_1 of 0.579 with a posterior standard deviation of 0.075.

The class specific parameters for the categorical variables are given in Table 1. These results can be compared to those in Dean and Raftery (2010) who reported the maximum likelihood estimates for the parameters of a two-class latent class model. The posterior mean and the maximum likelihood estimates are similar. However, the Bayesian approach also provides uncer-

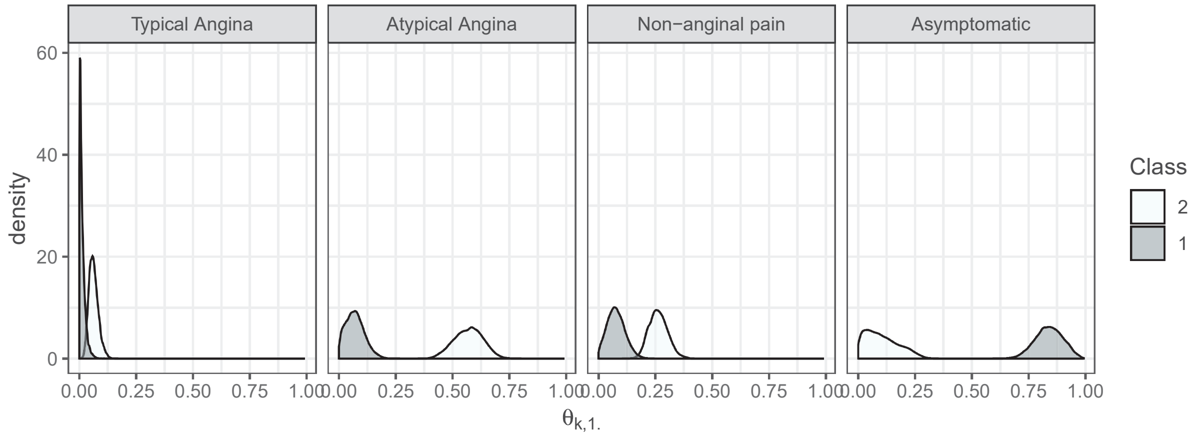

Figure 2. Posterior distribution of the class specific parameters for the variable "Chest pain type".

tainty estimates as given by the posterior standard deviations and the full posterior distributions which are visualized in Figure 2 for the variable "Chest pain type". In particular for parameter values which are estimated to be close to the boundary the posterior is non-normal and the full posterior allows to estimate suitable credible intervals for these parameters.

Observations can also be classified to the class they are most often assigned to during MCMC sampling after model identification. This partition is compared to the clinical partition contained in the data (see Table 3 on the left). The congruence between these two partitions is very high and results are similar to those reported in Dean and Raftery (2010).

3.2. Sparse finite mixture model with shrinkage prior

An overfitting mixture model is estimated using $e_0 = 0.0001$ and $K = 10$. In addition the shrinkage prior is imposed on the class specific parameters. The posterior distribution of the number of non-empty components K_+ has a clear mode at 2, with 99.9% of the samples having 2 non-empty components. The remaining samples had 3 non-empty components (0.2%). Using the samples with 2 non-empty components to identify the model results in a posterior mean estimate for the component weight of the larger class π_1 of 0.572 with a posterior standard deviation of 0.068. The class specific parameters for the variables are given in Table 2 and the congruence between the partitions in

Variable	Level	Class 1	Class 2
Chest pain type	Typical Angina	0.06 (0.02)	0.01 (0.01)
	Atypical Angina	0.57 (0.06)	0.07 (0.04)
	Non-anginal pain	0.26 (0.04)	0.08 (0.04)
	Asymptomatic	0.10 (0.07)	0.83 (0.06)
Exercise induced Angina	No	0.94 (0.03)	0.34 (0.10)
	Yes	0.06 (0.03)	0.66 (0.10)
Gender	Female	0.36 (0.04)	0.15 (0.04)
	Male	0.64 (0.04)	0.85 (0.04)
Resting Electrocardiographic results	Normal	0.81 (0.03)	0.79 (0.04)
	ST-T wave	0.16 (0.03)	0.19 (0.04)
	Estes' criteria	0.03 (0.01)	0.02 (0.01)
Fasting blood sugar >120 mg/dl	False	0.94 (0.02)	0.91 (0.03)
	True	0.06 (0.02)	0.09 (0.03)

Table 2. Posterior mean (and posterior standard deviations) of the class specific parameters for the identified 2-class sparse LCA model with shrinkage prior.

	Standard prior		Shrinkage prior	
	<50%	>50%	<50%	>50%
Class 1	139	15	135	14
Class 2	42	88	46	89

Table 3. Estimated versus clinical partition for the identified 2-component sparse LCA model with standard or shrinkage prior.

Table 3 on the right. Overall similar results are obtained for the two different component specific priors. However, using a shrinkage prior reduces the risk of overfitting heterogeneity and thus allows to obtain more precise estimates in case irrelevant variables are identified. Figure 3 shows the posterior distributions of the shrinkage parameters λ for each variable. Small values indicate that a variable is identified as not being relevant for distinguishing between the two classes and that similar parameter values are estimated for both classes. These results confirm those by Dean and Raftery (2010) who concluded that the variables "Resting Electrocardiographic results" and "Fasting blood sugar >120 mg/dl" are irrelevant.

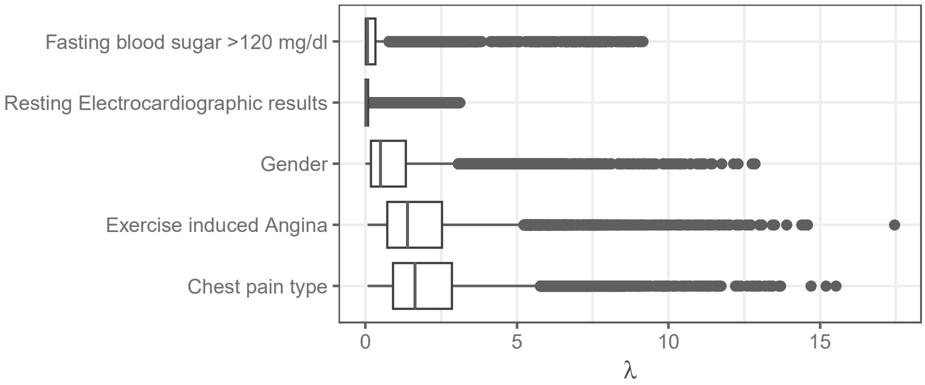

Figure 3. Box plot of the shrinkage parameter λ for each variable.

4. Conclusion

Suitable priors for Bayesian LCA are presented which regularize the likelihoods to avoid boundary solutions, induce sparse solutions with respect to the number of classes as well as shrinkage to perform implicit variable selection. Their application is demonstrated on the Hungarian heard disease data which was previously analyzed based on maximum likelihood estimation. This data set contains a clear structure with respect to the number of classes as well as the relevance of variables for clustering. Suitable priors for such a setting were proposed. Future research needs to investigate how these priors perform and need to be adapted in more challenging settings.

Acknowledgements: This research was funded by the Austrian Science Fund (FWF): P28740.

References

Dean N., Raftery A.E. (2010) Latent class analysis variable selection. *The Annals of the Institute of Statistical Mathematics*, 62, 11-35.

Diebolt J., Robert C.P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society B*, 56, 363-375.

Lazarsfeld P. (1950) The logical and mathematical foundation of latent structure analysis. In *Measurement and Prediction, the American Soldier: Studies in Social Psychology in World War II*, IV, 362-412. Princeton University Press.

Lunn D., Spiegelhalter D., Thomas A., Best N. (2009) The BUGS Project: Evolution, Critique and Future Directions. *Statistics in Medicine*, 28, 3049-3067.

Malsiner-Walli G., Frühwirth-Schnatter S., Grün B. (2016) Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26, 303-324.

Papastamoulis P. (2016) label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software*, 69, 1-24.

Plummer M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Technische Universität Wien, Vienna, Austria.

Rousseau J., Mengersen K. (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society B*, 73, 689-710.

White A., Wyse J., Murphy T.B. (2016) Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. *Statistics and Computing*, 26, 511-527.