Shrinkage Priors for Sparse Latent Class Analysis

Bettina Grün

Joint work with
Gertraud Malsiner-Walli & Sylvia Frühwirth-Schnatter

WU Wien, March 6th 2020

This research is supported by the
Austrian Science Fund (FWF): P28740.

Latent class analysis

- Variables in multivariate categorical data are often associated.
- Latent class analysis assumes that this association is due to the presence of latent classes (Lazarsfeld, 1950).
- This leads to a finite mixture model where the categorical variables are assumed to be independent given latent class membership.
- The latent class model represents the standard model-based clustering approach for categorical data.
- Applications are diverse and include the social sciences, psychometrics, medicine, etc.

Latent class analysis / 2

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

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

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

$$\sum_{k=1}^{K} \eta_k = 1, \qquad \eta_k \geq 0, \forall k,$$

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

Inference and issues in latent class analysis

- Estimation:
  - Frequentist maximum likelihood estimation based on the EM algorithm (Linzer and Lewis, 2011).
  - Bayesian estimation based on data augmentation and Gibbs sampling.
- Identifiability (Goodman, 1974):
  - Only local identifiability.
  - Induced by the multivariate structure, i.e., the number of categorical variables.
- Boundary solutions:
  - Occur in a ML setting without regularization if all observations in a component have the same observed category.
- Selecting the number of classes.
- Variable selection.

Prior choices for sparse modeling

We will investigate the choice of shrinkage priors for:

- **Priors on the weights:**
  In combination with overfitting mixtures, where the likelihood is problematic.
- **Priors on the component-specific parameters:**
  Assuming the presence of cluster-irrelevant variables we investigate priors which allow to distinguish between cluster-relevant and cluster-irrelevant variables.

Prior on the weights

- Conjugate prior: Dirichlet prior

$$\boldsymbol{\eta} \sim \mathcal{D}(e_1, \ldots, e_K)$$

- The exchangeable Dirichlet prior is assumed with

$$e_k \equiv e_0, \quad k = 1, \ldots, K.$$

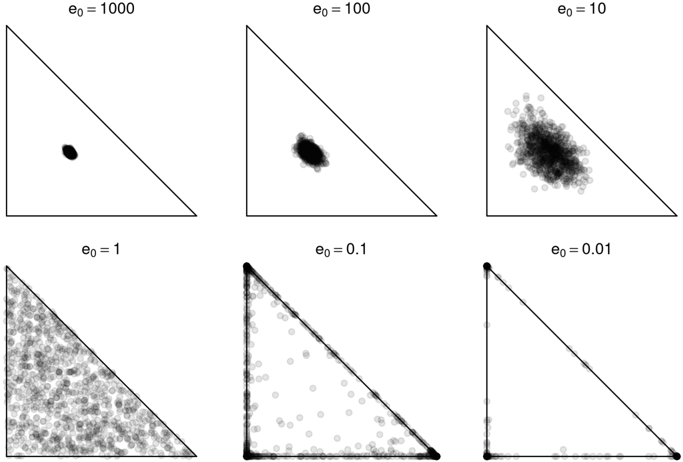
  This implies:

  - The prior expectation is

$$\mathbb{E}[\eta_k | e_0] = \frac{1}{K}$$

    regardless of the specific value of e_0.
  - The prior variance depends on the size of e_0.

Prior on the weights / 2



Dirichlet prior for overfitting mixtures

- Overfitting mixtures are mixtures where the fitted number of components K exceeds the true number of components K^{true}.
- The likelihood reflects the two possible ways of dealing with the superfluous components:
  - **Empty components:**
    - η_k is shrunken towards 0.
    - The component-specific parameters are identified only through their prior.
  - **Duplicated components:**
    - The difference of the component-specific parameters are shrunken towards 0.
    - Only the sum of the corresponding component weights is identified.
- The likelihood is multimodal, because it mixes these two unidentifiability modes.

Dirichlet prior for overfitting mixtures / 2

- Rousseau and Mengersen (2011) indicate that the value of e_0 strongly influences the asymptotic posterior density for overfitting mixtures.
- They show the following asymptotic result:
  - If $e_0 < d/2$, then asymptotically the posterior density concentrates over regions where $K - K^{\text{true}}$ groups are left empty.
  - If $e_0 > d/2$, then asymptotically the posterior density concentrates over regions with duplicated components.

  d denotes the dimension of the component-specific parameters.

Identifying the number of components

- Use overfitting mixtures with empty components (e_0 small).
  $\Rightarrow$ To obtain sparsity, e_0 very often has to be much smaller than $d/2$ in finite samples.
- Determine the number of non-empty components for each sweep m of the sampler

  $$K_0^{(m)} = K - \sum_{k=1}^{K} \boldsymbol{I}\{n_k^{(m)} = 0\}$$

  and use the most frequently visited value as estimate for K^{true}.

Prior on the component-specific parameters

- A-priori the parameters of the variables are independent within components.
- 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 and avoid modes at the boundary of the parameter space.
- Galindo Garre and Vermunt (2006) consider the following priors for Bayesian MAP estimation to regularize ML estimation:
  - Jeffreys prior.
  - Normal prior on the logit scale.
  - Dirichlet prior for the probabilities.

Identifying cluster-irrelevant variables

- Inclusion of cluster-irrelevant variables can:
  - Mask the cluster structure.
  - Reduce the accuracy of the parameter estimates.
- Proposed approaches:
  - Variable selection using step-wise procedures or stochastic model search for ML and Bayesian estimation as well as Gaussian mixture models and latent class models (Dean and Raftery, 2010; Tadesse, Sha, and Vanucci, 2005; White and Murphy, 2016).
  - Shrinking of component means towards a common mean in the Gaussian mixture case (Yau and Holmes, 2011; Frühwirth-Schnatter, 2011).

Shrinkage priors

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

$$\boldsymbol{a}_j = \boldsymbol{a}_{0,j} + \phi_j \boldsymbol{\mu}_j.$$

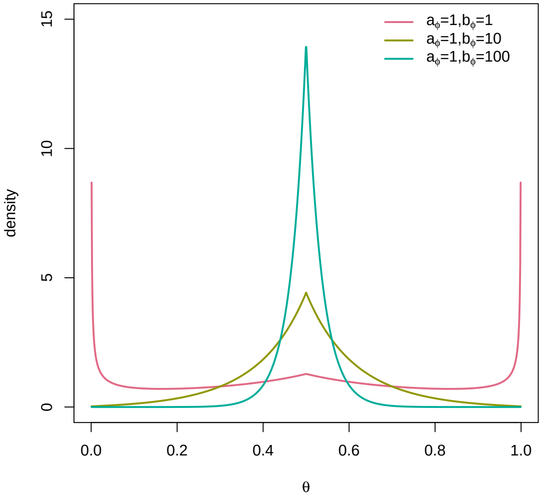
- ϕ_j represents the shrinkage factor for variable j.
- Using $\lambda_j = 1/\phi_j$ one can impose as prior

$$\lambda_j \sim \text{Gamma}(a_\phi, b_\phi), \; \forall j.$$

- $\boldsymbol{\mu}_j$ is the common mean of all components.

$$\boldsymbol{\mu}_j \sim \text{Dirichlet}(\boldsymbol{m}_j), \; \forall j.$$

Shrinkage priors / 2



Model estimation

- Since the 1990s the use of MCMC made Bayesian estimation of finite mixture models feasible.
- Like the EM algorithm (Dempster, Laird, and Rubin, 1977), practical Bayesian estimation is based on considering the class allocations as missing data and adding them in the estimation process.
  $\Rightarrow$ Data augmentation and Gibbs sampling makes sampling from the posterior density surprisingly simple (Diebolt and Robert, 1994).
- The priors assumed allow for a straightforward MCMC implementation.

Model identification

- The likelihood is invariant with respect to a permutation of the components.
- The use of symmetric priors implies that this invariance also holds for the posterior.
- Component-specific inference is impossible based on the MCMC output due to **label switching** (Redner and Walker, 1984).
- Several strategies have been proposed to determine an identified model (for an overview see Jasra, Holmes, and Stephens, 2005).

Model identification / 2

- We suggest to cluster the component-specific parameters of the MCMC draws in the point process representation, e.g., using *k*-means:
  - The point process representation is label-invariant.
  - If component-specific parameters from the same MCMC draw are assigned to the same *k*-means cluster, no unique relabeling is possible.
  - We discard the draws where no unique relabeling is achieved and use the proportion of discarded draws as a quality measure how well the fitted mixture model can be used as a clustering tool.

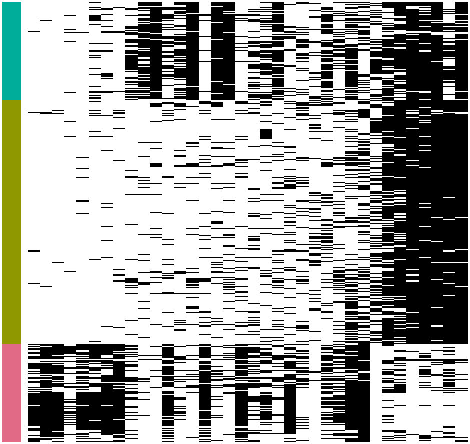
Modeling strategy

- Use a large value for K and a small e_0 in order to allow for automatic selection of a suitable number of clusters using the most frequent number of non-empty clusters during MCMC sampling.
- Use a gamma prior on the inverse precision of the component-specific parameters with $a_\phi = 1$ and b_ϕ large.
  - If component-specific parameters are pulled together with a shrinkage prior, the choice made for $\boldsymbol{\mu}_j$ is crucial.
  - Add a regularization $\boldsymbol{a}_{0,j}$ to avoid boundary solutions if precision is small, i.e., the variable is cluster-relevant.

Back pain data

- Fop, Smart, and Murphy (2017) use a binary data set on low back pain to perform latent class analysis.
- The data set contains for 425 patients the information on the presence / absence of 36 binary clinical indicators.
- A classification into 3 groups is known.
- Standard clustering methods using all available variables lead to a more fine-grained clustering solution than implied by the number of known groups.
- Some of the variables might imply sub-groups and thus variable selection could help to reduce the number of clusters detected.

Back pain data / 2



Variable selection

- Fop et al. (2017) distinguish three different roles for clustering variables:
  - Relevant variables.
  - Redundant variables.
  - Irrelevant variables.
- They perform a computational expensive step-wise procedure to select a suitable model based on maximum likelihood estimation using the BIC as model selection criterion.
- Their model selection task consists of:
  - Selecting a suitable number of groups.
  - Assigning one of the three roles to the clustering variables.

Bayesian estimation

- We apply Bayesian estimation with shrinkage priors on the component weights and the component-specific parameters.
- We use the following setting for the priors:

$$
\begin{aligned}
K &= 10, & e_0 &= 0.01, \\
a_\phi &= 1, & b_\phi &= 800.
\end{aligned}
$$

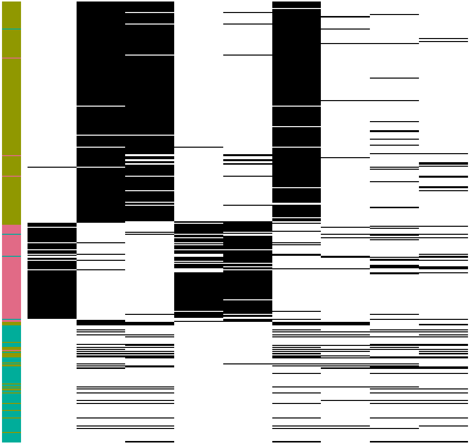
- We run MCMC sampling for 2,000 iterations burn-in and 20,000 recorded iterations.

Bayesian estimation / 2

We obtain the following results:

- 3 non-empty components occur for 85% of the MCMC draws.
- Model identification using only the variables where shrinkage factors are largest gives a non-permutation rate of 11%.
- The obtained clustering corresponds to the known classes:
  - Error rate: 8%.
  - Adjusted Rand index: 0.76.

Bayesian estimation / 3



Future work

- Investigate the impact of the parameter specification of hyper-priors and the regularization.
- In particular focus on the choice for $\boldsymbol{\mu}_j$ which is the common mean to which the parameters are shrunken.
- Use simulation studies to assess how the different roles of the clustering variables influence the performance of the Bayesian approach.
- Increase the number of variables to highlight the computational advantages of the Bayesian approach.
- Compare different prior specifications, such as also the use of the normal-gamma prior for the component-specific parameters on the probit scale.

Summary

- Shrinkage priors for Bayesian mixture models avoid overestimating heterogeneity without requiring fitting a large set of different models.
- Variable selection in particular in the context of latent class analysis is ambiguous due to the different roles which can be attributed to the variables.
- Bayesian analysis provides a flexible tool to vary how coarse or fine-grained the clustering solution obtained is depending on the amount of shrinkage imposed.

References

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

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

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

M. Fop, K. M. Smart, and T. B. Murphy. Variable selection for latent class analysis with application to low back pain diagnosis. **The Annals of Applied Statistics**, 11(4):2080–2110, 2017. doi: 10.1214/17-aoas1061.

S. Frühwirth-Schnatter. Label switching under model uncertainty. In K. Mengerson, C. Robert, and D. Titterington, editors, **Mixtures: Estimation and Application**, pages 213–239. Wiley, 2011.

F. Galindo Garre and J. K. Vermunt. Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. **Behaviormetrika**, 33(1):43–59, 2006.

References / 2

L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. **Biometrika**, 61(2):215–231, Aug. 1974.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. **Statistical Science**, 20(1):50–67, 2005.

P. F. Lazarsfeld. The logical and mathematical foundation of latent structure analysis. In **Studies in Social Psychology in World War II: Measurement and Prediction**, volume 4, pages 362–412. Princeton University Press, 1950.

D. Linzer and J. Lewis. poLCA: An R package for polytomous variable latent class analysis. **Journal of Statistical Software**, 42(10):1–29, 2011.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. **SIAM Review**, 26(2):195–239, Apr. 1984.

J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. **Journal of the Royal Statistical Society B**, 73(5):689–710, 2011.

References / 3

M. G. Tadesse, N. Sha, and M. Vanucci. Bayesian variable selection in clustering high-dimensional data. **Journal of the American Statistical Association**, 100(470):602–617, 2005.

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

C. Yau and C. Holmes. Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. **Bayesian Analysis**, 6 (2):329–352, 2011.