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.
The latent class model for observations $y_i, i = 1, \ldots, n$ is given by

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

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

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

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

\[ \eta \sim D(e_1, \ldots, e_K) \]

- 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

e_0 = 1000  

\[ \text{Triangle} \]

e_0 = 100  

\[ \text{Triangle} \]

e_0 = 10  

\[ \text{Triangle} \]

e_0 = 1  

\[ \text{Triangle} \]

e_0 = 0.1  

\[ \text{Triangle} \]

e_0 = 0.01  

\[ \text{Triangle} \]
Overfitting mixtures are mixtures where the fitted number of components $K$ exceeds the true number of components $K^{\text{true}}$.

The likelihood reflects the two possible ways of dealing with the superfluous components:

- **Empty components:**
  - $\eta_k$ 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.
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 \((e_0 \text{ small})\).
  \(\Rightarrow\) 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

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

and use the most frequently visited value as estimate for \(K^{\text{true}}\).
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 $\theta_{k,j}$. a-priori follows a Dirichlet distribution:
  \[
  \theta_{k,j} \sim \text{Dirichlet}(a_j).
  \]
- The value for $a_j$ 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 $a_j$.
- Re-parameterize the Dirichlet parameter into a mean and precision parameter plus a regularizing additive constant:

$$a_j = a_{0,j} + \phi_j \mu_j.$$ 

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

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

- $\mu_j$ is the common mean of all components.

$$\mu_j \sim \text{Dirichlet}(m_j), \ \forall j.$$
Shrinkage priors

\[
\begin{align*}
\phi & = 1, b_\phi = 1 \\
\phi & = 1, b_\phi = 10 \\
\phi & = 1, b_\phi = 100
\end{align*}
\]
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.

⇒ 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).
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 $b_\phi$ large.
  - If component-specific parameters are pulled together with a shrinkage prior, the choice made for $\mu_j$ is crucial.
  - Add a regularization $a_{0,j}$ 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 computationally 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:

\[
K = 10, \quad e_0 = 0.01,
\]

\[
a_\phi = 1, \quad b_\phi = 800.
\]

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

- Investigate the impact of the parameter specification of hyper-priors and the regularization.

- In particular focus on the choice for $\mu_j$ 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.


