From here to infinity - bridging finite and Bayesian nonparametric mixture models in model-based clustering

Sylvia Frühwirth-Schnatter, Jan Greve, Bettina Grün, Gertraud Malsiner-Walli
(WU Vienna University of Business and Economics, Austria)

Funded by the Austrian Science Fund (FWF P28740)

2021 22nd EYSM (September 6-10) & Athens University of Economic and Business (October 7, 2021)

Common questions about finite mixture models

- Are mixtures, like tequila, inherently evil and should be avoided at all costs (Larry Wasserman on his now defunct blog *Normal Deviate*)?
- Has the number of components, K, to be known, if I want to use finite mixtures for clustering?
- If K is unknown, do I have to implement a complicated trans-dimensional MCMC sampler?
- Are finite mixtures less flexible than BNP mixtures, e.g. a Dirichlet process mixture (DPM)?

Outline of the talk

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures

Outline

Part :

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures

Finite mixture models

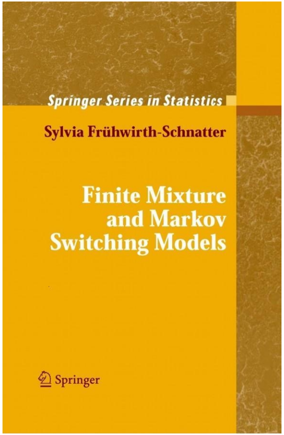
- Observations $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)$ are an iid sample from a **mixture distribution:**

$$p(\mathbf{y}_i|\vartheta) = \sum_{k=1}^{K} \eta_k f_{\mathcal{T}}(\mathbf{y}_i|\boldsymbol{\theta}_k),$$

  - K is the number of components;
  - the component densities $f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k)$ arise from the same distribution $\mathcal{T}(\boldsymbol{\theta})$;
  - $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ vary over the components;
  - $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ are the component weights, $\sum_{k=1}^{K} \eta_k = 1$, $\eta_k \geq 0$.
- Usually, group membership of the observations is unknown.
- Latent allocation variables $(S_1, \ldots, S_N)$ with $S_i \in \{1, \ldots, K\}$ are introduced to indicate the component from which each observation is drawn:

$$p(\mathbf{y}_i|S_i = k) = f_{\mathcal{T}}(\mathbf{y}_i|\boldsymbol{\theta}_k), \quad \Pr(S_i = k) = \eta_k.$$

For more details see . . .



2006

2019

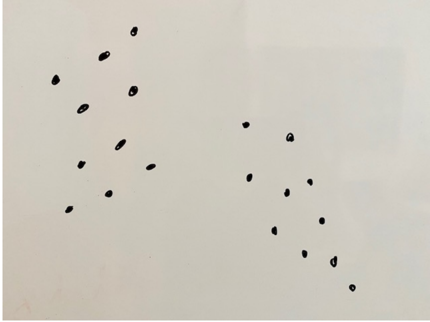
Bayesian cluster analysis

- In cluster analysis, the aim is to partition the data into groups, where within groups the observations are more "similar" than between groups.
- **Clustering** arises in a natural way in finite mixtures [Bensmail et al., 1997], recent review: [Grün, 2019]
- Each observation $\mathbf{y}_i$ has a (latent) **indicator variable** $\mathbf{S_i}$ indicating the component the observation belongs to:

$$\mathbf{y}_i | S_i \sim f_{\mathcal{T}}(\mathbf{y}_i | \boldsymbol{\theta}_{S_i}).$$

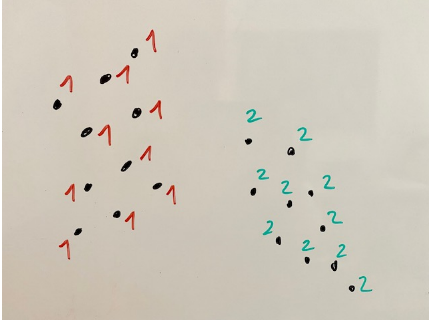
- y_i and y_j belong to the same **cluster**, iff $S_i = S_j$.

A stylized example



... with obviously two clusters

A stylized example



Fitting a mixture with two components ($K = 2$) identifies the two clusters

Induced partitions

- $(S_1, \ldots, S_N)$ define a **partition** $\mathcal{C}$ of the N data points,

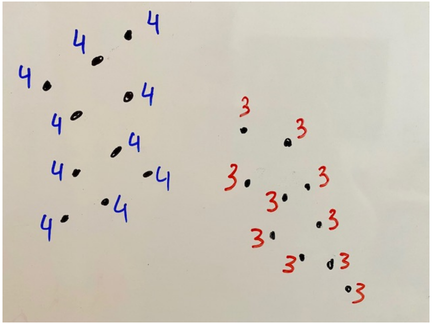
$$\mathcal{C} = \{C_1, \ldots, C_{K_+}\},$$

  which contains $K_+ = |\mathcal{C}|$ clusters [Hartigan, 1990]
- With $\mathbf{S} = (S_1, \ldots, S_N)$ being latent (random), we can look at the prior $p(\mathcal{C})$ and the posterior distribution $p(\mathcal{C}|\mathbf{y})$ [Casella et al., 2004], [Lau and Green, 2007]

Components versus data clusters

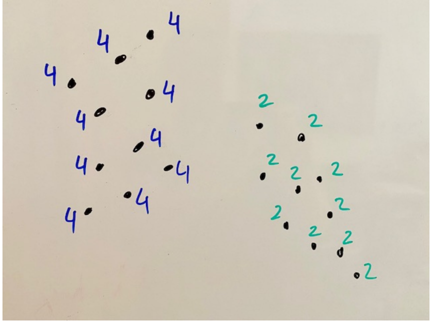
- In mixture analysis it is important to distinguish between:
  - K: the number of components in the mixture distribution.
  - K_+: the number of clusters in the data set
- In a finite sample the number of components K_+ used to generate the data (i.e. number of filled components) might be lower than K.

A stylized example



Fitting a mixture with **five** components ($K = 5$): only components 3 and 4 are used for clustering, the components 1, 2, and 5 remain "empty"

A stylized example



Fitting a mixture with **five** components ($K = 5$): only components 2 and 4 are used for clustering, the components 1, 3, and 5 remain “empty”

$K_+ = K$?

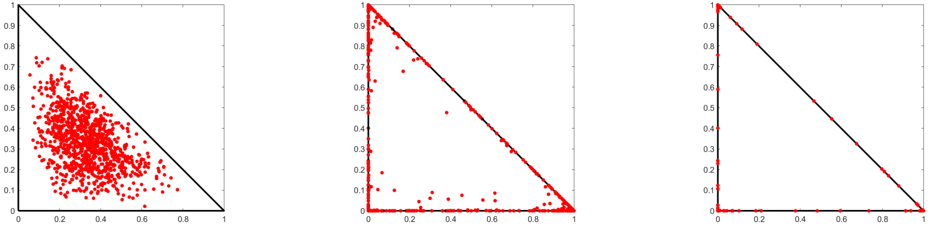
- Let N_k is the number of observations allocated to component k, $k = 1, \ldots, K$.
- Apriori, the occupation numbers are random:
  $(N_1, \ldots, N_K) \sim \mathrm{MulNom}\,(N; \eta_1, \ldots, \eta_K)$.
- Depending on the weights $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ and N, multinomial sampling may lead to partitions with **empty groups with $N_k = 0$.**
- In this case, fewer than K mixture components were used to cluster the data, i.e. the resulting partition $\mathcal{C} = \{C_1, \ldots, C_{K_+}\}$ contains $\mathbf{K_+ < K}$ clusters:

$$K_+ = K - \sum_{k=1}^{K} I\{N_k = 0\}.$$

  where K_+ is the number of **nonempty components**.
- K_+ is a **random variable** and can take a priori values $\mathbf{K_+ < K}$ with probability depending on $\boldsymbol{\eta}, N, K$.

The importance of the Dirichlet prior

- Consider a finite mixtures with K fixed
- Assume a symmetric Dirichlet prior $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K) \sim \mathcal{D}_K(\gamma)$ on the weight distribution
- The hyperparameter γ exercises strong influence on prior of the weight distribution, e.g. for $K = 3$:

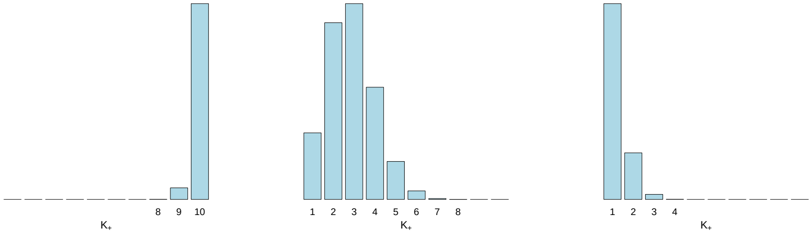


left: $\gamma = 4$, middle: $\gamma = 0.05$, right: $\gamma = 0.005$

Example: sparse finite mixtures

- (Static) Sparse finite mixtures choose a very small values of γ [Malsiner Walli et al., 2016], [Malsiner Walli et al., 2017] (overfitting mixture in the sense of [Rousseau and Mengersen, 2011])
- e.g. $K = 10$, $N = 100$:

$\gamma = 4$ $\gamma = 0.05$ $\gamma = 0.005$



K is fix; K_+ is random with an implicit prior $p(K_+|\gamma, N, K)$ concentrating on $K_+ < K$;

A fully Bayesian perspective

- In mixture analysis it is important to distinguish between:
  - K: the number of components in the mixture distribution.
  - K_+: the number of clusters in the data set
- Both K and K_+ are usually unknown and have to be estimated from the data.
- From a Bayesian perspective, the most natural approach is to treat them as unknown parameters and put priors on them:
  - Prior on K is explicitly defined.
  - Prior on K_+ is implicitly defined through priors on K and the weights and depends on N.

Outline

Part :

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures

Generalized mixture of finite mixtures (MFM) model

- A fully Bayesian mixture model is defined in a hierarchical way:

$$
\begin{aligned}
K &\sim p(K),\\
\eta_1,\dots,\eta_K|K,\gamma_K &\sim \mathcal{D}_K(\gamma_K),\\
S_i|K,\eta_1,\dots,\eta_K &\sim \mathcal{M}(1;\eta_1,\dots,\eta_K),\ \text{independently for } i=1,\dots,N,\\
\phi &\sim p(\phi),\\
\boldsymbol{\theta}_k|\phi &\sim p(\boldsymbol{\theta}_k|\phi),\ \text{independently for } k=1,\dots,K,\\
\mathbf{y}_i|K,S_i=k,\boldsymbol{\theta}_k &\sim f_{\mathcal{T}}(\mathbf{y}_i|\boldsymbol{\theta}_k),\ \text{independently for } i=1,\dots,N.
\end{aligned}
$$

- Generic framework with no specific restrictions on
  - $f_{\mathcal{T}}(\cdot|\boldsymbol{\theta}_k)$ (parametric family),
  - observations $\mathbf{y}_i$ can be univariate or multivariate, continuous, discrete-valued, mixed-type, time series data, outcomes of a regression model, ...
  - the prior $p(K)$ (e.g., parametric pmf, $\delta_{\{K_{fix}\}}$, $\delta_{\{\infty\}}$, ...),
  - **and ...**

Static versus dynamic MFMs

- **... the sequence** $\{\gamma_K\}, K = 1, 2, \ldots$
- In a generalized MFM (SFS, Malsiner-Walli and Grün, 2021), the sequence $\{\gamma_K\}$ can either be fixed or assumed to depend on K.
- We consider two specific types of generalized MFMs:
  - Static MFMs with hyperparameter γ [Richardson and Green, 1997], [Miller and Harrison, 2018]:

$$\gamma_K \equiv \gamma.$$

  - Dynamic MFMs with hyperparameter α [McCullagh and Yang, 2008], [Guha et al., 2019]:

$$\gamma_K = \frac{\alpha}{K}.$$

The exchangeable probability partition function

- The prior on the partitions (EPPF) is well known for a static finite mixture:

$$p(\mathcal{C}|N, K, \gamma) = \binom{K}{K_+} K_+! \int p(\mathbf{S}|\boldsymbol{\eta}_K) p(\boldsymbol{\eta}_K|K, \gamma) d\boldsymbol{\eta}_K$$

$$= \frac{K!}{(K - K_+)!} \frac{\Gamma(K\gamma)}{\Gamma(K\gamma + N)} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \gamma)}{\Gamma(\gamma)}.$$

- EPPF for a generalized finite mixture with hyperparameter γ_K (K fixed):

$$p(\mathcal{C}|N, K, \gamma_K) = \frac{K!}{(K - K_+)!} \frac{\Gamma(K\gamma_K)}{\Gamma(K\gamma_K + N)} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \gamma_K)}{\Gamma(\gamma_K)}.$$

The EPPF of a generalized MFM

- EPPF for a generalized finite mixture with hyperparameter γ_K (K fixed):

$$p(\mathcal{C}|N, K, \gamma_K) = V_{N,K_+}^{K,\gamma_K} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \gamma_K)}{\Gamma(\gamma_K)},$$

$$V_{N,K_+}^{K,\gamma_K} = \frac{K!}{(K - K_+)!} \frac{\Gamma(K\gamma_K)}{\Gamma(K\gamma_K + N)}.$$

- Takes the form of a **Gibbs-type prior** [Gnedin and Pitman, 2006]
- EPPF for a generalized MFM with prior $p(K)$:

$$p(\mathcal{C}|N, \gamma_K) = \sum_{K=K_+}^{\infty} V_{N,K_+}^{K,\gamma_K} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \gamma_K)}{\Gamma(\gamma_K)} p(K).$$

- More general than a Gibbs-type prior, if γ_K depends on K.

Derivation of $p(K_+|N,\gamma)$

- Prior on K_+ obtained by summing over all partitions (challenging for large N):

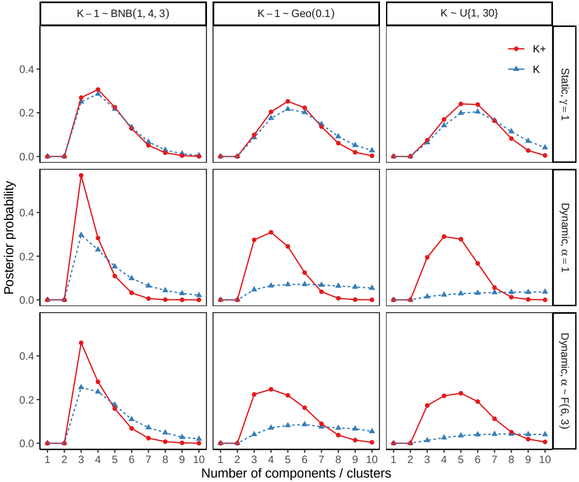
$$p(K_+ = k|N, \gamma_K) = \sum_{\mathcal{C}:K_+=k} p(\mathcal{C}|N, \gamma_K).$$

- We work with the prior on the labeled cluster sizes $p(N_1, \dots, N_{K_+}|N, K, \gamma_K)$.
- Arrange the clusters in some exchangeable order (e.g. in order of appearance, see [Pitman, 1996]).
- By summing over all $N_1, \dots, N_{K_+}$ with $\sum_{j=1}^{K_+} N_j = N$, we obtain:

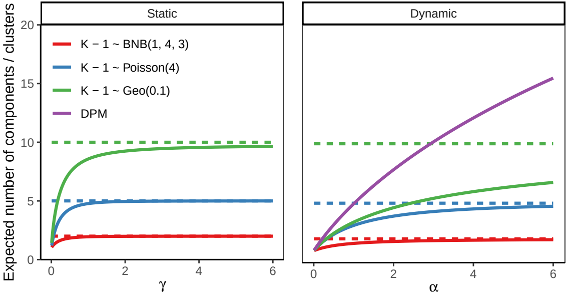
$$p(K_+ = k|N, \gamma) = \frac{N!}{k!} \sum_{K=k}^{\infty} p(K) \frac{V_{N,k}^{K,\gamma_K}}{\Gamma(\gamma_K)^k} C_{N,k}^{K,\gamma_K},$$

where $V_{N,k}^{K,\gamma_K}$ and $C_{N,k}^{K,\gamma_K}$ (possibilities to split N objects into k clusters) are determined recursively.

Examples for $p(K_+|N)$ versus $p(K)$ for $N = 82$



Example: Prior expectations $\mathrm{E}(K_+|\gamma, N)$



Prior expectations $\mathrm{E}(K_+|\gamma, N)$ for static MFMs (left) and $\mathrm{E}(K_+|\alpha, N)$ for dynamic MFMs (right) as functions of γ and α for $N = 100$ under the priors $K - 1 \sim \mathrm{BNB}(1, 4, 3)$, $K - 1 \sim \mathcal{P}(4)$, and $K - 1 \sim \mathrm{Geo}(0.1)$ in comparison to a DPM. For each prior $p(K)$, the prior expectation $\mathrm{E}(K)$ is plotted as a horizontal dashed line.

More about the implicit prior on the partitions

- [Greve et al., 2020]: "Spying on the indicators"
  - R-package **fipp**
  - functionals of the implicit prior distribution of K_+;
  - implicit prior distribution of symmetric functional of $N_1, \ldots, N_{K_+}$ such as the entropy

Outline

Part :

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures

Bayesian inference for generalized MFMs

- A trans-dimensional sampler is required.
- Traditional methods:
  - Reversible Jump MCMC (RJMCMC) [Richardson and Green, 1997]; [Robert et al., 2000]
  - Chinese Restaurant Process (CRP) based sampling schemes developed for BNP analysis [Jain and Neal, 2004], [Jain and Neal, 2007],[Miller and Harrison, 2018];
  - RJMCMC (developed for static MFMs) can be extended to dynamic MFMs.
- Telescoping Sampling:
  - suggested in SFS, Malsiner-Walli and Grün (2021) for dynamic MFMs;
  - easy to implement: generic sampler for arbitrary component densities;
  - works also for static MFM with a “gap” between K and K_+.

The Telescoping Sampler

- Exploits the exchangeable probability partition function (EPPF) of a MFM (similar to the CRP sampler)
- K is introduced as a latent variable (as in RJMCMC);
- Sample the number of components K given the partition $\mathcal{C}$:

$$
\begin{aligned}
p(K|\mathcal{C},\boldsymbol{\gamma}) &\propto p(\mathcal{C}|\gamma_K, K)p(K) \\
&\propto \frac{K!}{(K-K_+)!}\frac{\Gamma(K\gamma_K)}{\Gamma(K\gamma_K+N)}\prod_{k=1}^{K_+}\frac{\Gamma(N_k+\gamma_K)}{\Gamma(1+\gamma_K)}p(K).
\end{aligned}
$$

- Combined with any MCMC algorithm for any finite mixture with a fixed number of components K, e.g. Gibbs sampling [Diebolt and Robert, 1994]

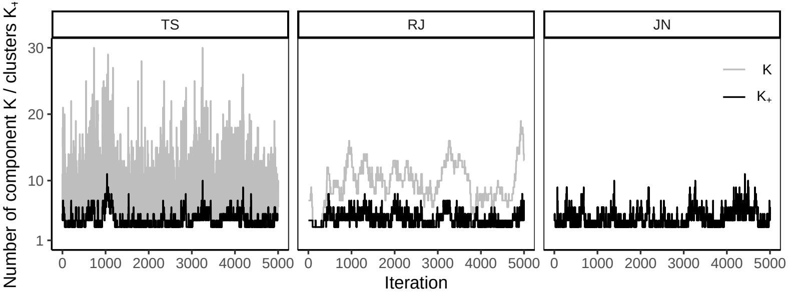
The Telescoping Sampler

- A partially marginalized sampler which switches between
  - sampling from the complete-data posterior distribution conditional on the latent allocation variables $\mathbf{S}$
  - sampling $\mathcal{C}$ from the collapsed posterior which lives in the set partition space and is marginalized over the empty components, the weight distribution and all allocations $\mathbf{S}$ inducing $\mathcal{C}$.

MCMC scheme

1. Conditional on the parameters, update the partition $\mathcal{C}$ by sampling S_i from
$$\Pr(S_i = k|\boldsymbol{\eta}_K, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \mathbf{y}_i, K) \propto \eta_k f(\mathbf{y}_i|\boldsymbol{\theta}_k);$$
   - determine $K_+ = \sum_{k=1}^{K} I\{N_k > 0\}$, where $N_k = \#\{i|S_i = k\}$,
   - relabel the components to have the first K_+ components non-empty.
2. Conditional on $\mathcal{C}$, update parameters of filled components and hyperparameters:
   - Sample $\boldsymbol{\theta}_k$, for the components $k = 1, \dots, K_+$, from $p(\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{y}, \phi)$.
   - Sample the hyperparameter ϕ from $p(\phi|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{K_+}, K_+)$.
3. Conditional on $\mathcal{C}$, sample K from $p(K|\mathcal{C}, \boldsymbol{\gamma}) \propto p(K)p(\mathcal{C}|\gamma_K, K, N)$.
4. Conditional on $(K, \phi, \mathcal{C})$, add empty/non-filled components and update the weights:
   - If $K > K_+$: add $K - K_+$ empty components and sample $\boldsymbol{\theta}_k|\phi$ from the prior $p(\boldsymbol{\theta}_k|\phi)$, $k = K_+ + 1, \dots, K$.
   - Sample $\boldsymbol{\eta}_K|K, \gamma_K, \mathbf{S} \sim D(e_1, \dots, e_K)$, where $e_k = \gamma_K + N_k$.

Benchmarking the TS I



- Simulated data, $N = 1000$;
- Static MFM with $\gamma_K \equiv 0.1$; priors on component parameters and K as in [Richardson and Green, 1997];
- Trace plots of K (gray) and K_+ (black) for the TS, RJ and JN sampler.

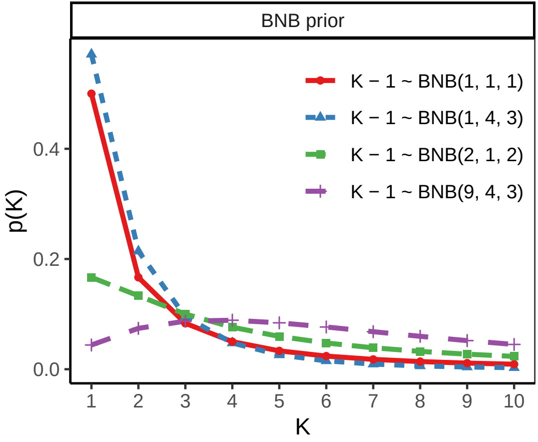
Outline

Part :

Prior choices in applied mixture analysis

- Dynamic versus static: recommend to use the dynamic version with $\gamma_K = \alpha/K$;
- Assume that α is a random hyperparameter $\alpha \sim \mathrm{F}(6,3)$ (instead of popular Gamma distribution).
- Prior on K: translated BNB distribution, $K-1 \sim \mathrm{BNB}(\alpha_\lambda, a_\pi, b_\pi)$
  - translated Poisson distribution $K-1|\lambda \sim \mathcal{P}(\lambda)$;
  - hierarchical Gamma prior $\lambda|\beta \sim \mathcal{G}(\alpha_\lambda, \beta)$ leads to the translated negative-binomial distribution, $K-1|\beta \sim \mathrm{NegBin}(\alpha_\lambda, \beta)$;
  - for $\alpha_\lambda = 1$, this reduces to the translated geometric distribution $K-1|\beta \sim \mathrm{Geo}(\pi)$ with success probability $\pi = \beta/(1+\beta)$;
  - hierarchical Beta prior on $\pi \sim \mathcal{B}(a_\pi, b_\pi)$ yields $K-1 \sim \mathrm{BNB}(\alpha_\lambda, a_\pi, b_\pi)$.
- In practice: $K-1 \sim \mathrm{BNB}(1,4,3)$

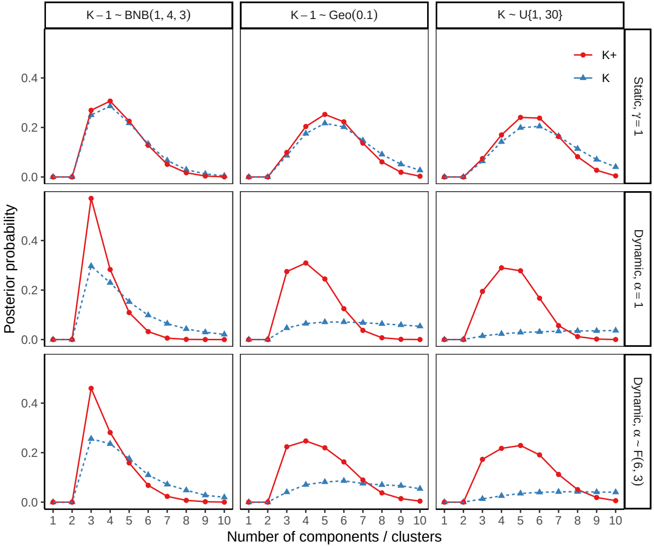
The beta-negative-binomial distribution



Galaxy data I

- [Grün et al., 2021]: How many data clusters are in the Galaxy data set?
- Fit a univariate mixture of normals with K unknown
- Prior choices are very influential for this data set [Aitkin, 2001]
- The recommended prior (dynamic MFM, $K - 1 \sim \mathrm{BNB}\,(1, 4, 3)$, $\alpha \sim \mathrm{F}\,(6, 3)$) works very well.

Galaxy data



K – 1 ~ BNB(1, 4, 3)
K – 1 ~ Geo(0.1)
K ~ U{1, 30}
K+
K
Static, γ = 1
Dynamic, α = 1
Dynamic, α ~ F(6, 3)
Posterior probability
0.0
0.2
0.4
1 2 3 4 5 6 7 8 9 10
Number of components / clusters

Changing the clustering kernel

- Thyroid data [Scrucca et al., 2016], $N = 215$
  - five-dimensional laboratory test variables
  - operation diagnosis observed (**three** potential groups)
  - multivariate mixture of Gaussian with K unknown
- Fear data [Stern et al., 1994], $N = 93$
  - three categorical features: motor activity (4 categories), fret/cry behavior (3 categories) and fear of unfamiliar events (3 categories)
  - Psychological theory suggests **two** groups
  - Latent class analysis with K unknown

Changing the clustering kernel

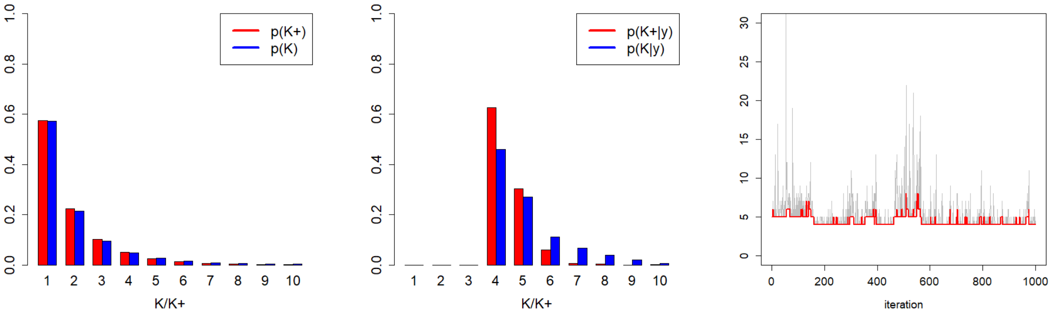
	Thyroid		Fear	
$p(K)$	$p(K_+\|\mathbf{y})$	$p(K\|\mathbf{y})$	$p(K_+\|\mathbf{y})$	$p(K\|\mathbf{y})$
$\mathcal{U}\,[1,30]$	**3** [3, 3]	3 [4, 19]	**6** [5, 9]	30 [10, 24]
$\mathrm{Geo}\,(0.1)$	**3** [3, 3]	3 [3, 7]	**4** [4, 7]	5 [5, 16]
$\mathrm{BNB}\,(1,4,3)$	**3** [3, 3]	3 [3, 4]	**2** [2, 4]	2 [2, 5]

- Posterior inference for K and K_+ for a dynamic MFM based on different priors $p(K)$ and $\alpha \sim \mathrm{F}\,(6,3)$.
- The posteriors of K_+ and K are summarized by their modes, followed by the 1st and 3rd quartiles.

Mozambique data I

- Data from the Demographics and Health Survey (DHS) for Mozambique from 2003.
- The dataset includes information on $N = 11,922$ women.
- 10 binary variables indicate which source / channel is used by women to get information on HIV (radio, TV, newspapers/magazines, posters, clinic/healthworker, church, school, community meetings, friends/relatives and working place).
- Aim: cluster women into groups according to the information sources on HIV they use.
- We apply the dynamic mixture of latent class analysis (LCA) models ($\gamma_K = 1/K$, $K - 1 \sim BNB(1, 4, 3)$, $\pi_k \sim \mathcal{B}(4, 4)$).

Mozambique data II



Prior (left) and posterior (middle) of K (blue) and K_+ (red); trace plot of TS (right)

Mozambique data III

	church/comm	TV	friends	modern
AIDSINFO-RADIO	0.51	0.98	0.76	0.95
AIDSINFO-TV	0.03	0.97	0.00	0.87
AIDSINFO-NEWS	0.03	0.23	0.00	0.99
AIDSINFO-POSTER	0.13	0.07	0.03	0.88
AIDSINFO-WKR	0.29	0.02	0.02	0.13
AIDSINFO-CHURCH	0.47	0.05	0.07	0.09
AIDSINFO-SCHOOL	0.17	0.20	0.06	0.26
AIDSINFO-COMM	0.76	0.05	0.15	0.06
AIDSINFO-FRND	0.01	0.45	0.58	0.37
AIDSINFO-WORK	0.08	0.02	0.01	0.11
Size	0.04	0.19	0.75	0.02

Outline

Part :

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures

BNP mixtures

- Random probability measure priors like the Dirichlet process $G^{DP} \sim \mathcal{DP}(\alpha, \mathcal{G}_0)$ [Ferguson, 1973, Ferguson, 1974] lead to infinite mixtures:

$$p(\mathbf{y}) = \int f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}) G^{DP}(d\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \eta_k f_{\mathcal{T}}(\mathbf{y}|\boldsymbol{\theta}_k),$$

  where η_k are random weights such that $\sum_{k=1}^{\infty} \eta_k = 1$ almost surely.
- The **stick-breaking representation** [Sethuraman, 1994] defines the weights in terms of a sequence $v_1, v_2, v_3, \dots$ of independent random variables (the sticks):

$$\eta_1 = v_1, \quad \eta_2 = (1 - v_1)v_2, \qquad \eta_k = v_k \prod_{j=1}^{k-1} (1 - v_j).$$

It's all in the weights!

- Assume that the base measure $\mathcal{G}_0$ is the same as the prior $p(\boldsymbol{\theta}_k)$ in a finite mixture.
- The "only" difference lies in the **prior of the sticks** $v_1, v_2, v_3, \ldots$:
  - Finite mixtures: $v_k \sim \mathcal{B}\left(\gamma_K, (K-k)\gamma_K\right), k = 1, \ldots, K-1$, $v_K = 1$
  - Dirichlet process mixtures (DPM) with $\mathcal{DP}\left(\alpha, \mathcal{G}_0\right)$: $v_k \sim \mathcal{B}\left(1, \alpha\right)$
  - Pitman-Yor process mixtures with $\mathcal{PY}(\beta, \alpha)$ with reinforcement parameter $\beta \in [0, 1), \alpha > -\beta$ [Pitman and Yor, 1997]: $v_k \sim \mathcal{B}\left(1-\beta, \alpha + k\beta\right)$ (reduces to DPMs for $\beta = 0$);
  - Pitman-Yor process mixtures, where $\beta < 0$ and $\alpha = K|\beta|$ with $K \in \mathbb{N}$ [De Blasi et al., 2015]:
    - In the corresponding stick-breaking representation $v_K = 1$ a.s.
    - Yields a mixture with infinitely many components, of which only K have non-zero weights, with the symmetric Dirichlet distribution $\mathcal{D}_K\left(|\beta|\right)$ acting as prior.

Bridging generalized MFMs and PYP mixtures

- For finite mixtures with a fixed number of components $K < \infty$, a dual PYP prior with $\beta < 0$ exists which implies exactly the same EPPF [Gnedin and Pitman, 2006]:
  (a) For a static finite mixture with $\gamma > 0$, this is the PYP prior $\mathcal{PY}(-\gamma, K\gamma)$.
  (b) For a dynamic finite mixture with $\gamma_K = \alpha/K$, this is the PYP prior $\mathcal{PY}(-\alpha/K, \alpha)$.
- While being finite mixtures with a prior $p(K)$ on K, MFMs are very flexible with close connections to BNP mixture models:
  (a) A static MFM with hyperparameter γ is related to a mixture of PYMs $\mathcal{PY}(-\gamma, K\gamma)$ which are mixed over the concentration parameter $\alpha_K = K\gamma$ with prior $p(K)$, while the reinforcement parameter $\beta = -\gamma$ is kept fixed [De Blasi et al., 2013]
  (b) A dynamic MFM with hyperparameter α is related to a mixture of PYMs $\mathcal{PY}(-\alpha/K, \alpha)$ which are mixed over the reinforcement parameter $\beta_K = -\alpha/K$ with prior $p(K)$, while the concentration parameter α is kept fixed.
  Yields a model beyond the class of Gibbs-type priors. [SFS, Malsiner-Walli and Grün, 2021].

Relation of dynamic MFMs to DPMs

- (Dynamic) Sparse finite mixtures with $\gamma = \frac{\alpha}{K}$, K fixed:
  - converge to a DPM with $GDP \sim \mathcal{DP}\left(\alpha, \mathcal{G}_0\right)$ as K increases [Green and Richardson, 2001].
  - often used to approximate a DPM;
- putting a prior $p(K)$ on the "hyperparameter" K yields the dynamic MFM:

$$p(\mathcal{C}|N,\alpha) = p_{DP}(\mathcal{C}|N,\alpha) \times \sum_{K=K_+}^{\infty} p(K) R_{K_+}^{K,\alpha}, \qquad \lim_{K\to\infty} R_{K_+}^{K,\alpha} = 1.$$

  - The EPPF converges to the Ewens distribution, as the prior puts increasing mass on large values of K.
- With a proper prior $p(K)$, the dynamic MFM is a flexible natural generalization of the DPM beyond Gibbs-type priors.

A lot remains to be done . . .

- R-package **bmbclust**:
  - A range of parametric component densities (uni- and multivariate Gaussians, Poisson, latent class analysis)
  - Semi-parametric component densities in the spirit of [Malsiner Walli et al., 2017]
- Posterior consistency for the number of clusters for generalized MFMs under correctly specified and misspecified components [Guha et al., 2019]

Concluding Q & A about finite mixtures

- Are mixtures, like tequila, inherently evil and should be avoided at all costs?

  No, mixtures are really interesting and useful, but can be challenging.
- Has the number of components K to be known, if I want to use finite mixtures for clustering?

  No, put a prior on K and check implicit priors, e.g. $p(K_+|N)$, using the `fipp`-package.
- If K is unknown, do I have to implement a complicated trans-dimensional MCMC sampler?

  No, you can use the telescoping sampler.
- Are finite mixtures less flexible than BNP mixtures such as Dirichlet process mixtures (DPM)?

  No, dynamic MFMs are more general than DPMs and are very closely related to mixtures of Pitman-Yor process mixtures with a finite number of components (clusters).

References I

Aitkin, M. (2001).
Likelihood and Bayesian analysis of mixtures.
Statistical Modelling, 1:287–304.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997).
Inference in model-based cluster analysis.
Statistics and Computing, 7:1–10.

Casella, G., Robert, C. P., and Wells, M. T. (2004).
Mixture models, latent variables and partitioned importance sampling.
Statistical Methodology, 1:1–18.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I., and Ruggiero, M. (2015).
Are Gibbs-type priors the most natural generalization of the Dirichlet process?
IEEE Transactions on Pattern Analysis and Machine Intelligence, 37:212–229.

References II

De Blasi, P., Lijoi, A., and Prünster, I. (2013).
An asymptotic analysis of a class of discrete nonparametric priors.
Statistica Sinica, 23:1299–1321.

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

Ferguson, T. S. (1973).
A Bayesian analysis of some nonparametric problems.
The Annals of Statistics, 1:209–230.

Ferguson, T. S. (1974).
Prior distributions on spaces of probability measures.
The Annals of Statistics, 2:615–629.

References III

Frühwirth-Schnatter, S. (2006).
Finite Mixture and Markov Switching Models.
Springer, New York.

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors (2019).
Handbook of Mixture Analysis.
CRC Press, Boca Raton, FL.

Frühwirth-Schnatter, S., Walli, G. M., and Grün, B. (2021).
Generalized mixtures of finite mixtures and telescoping sampling.
Bayesian Analysis, page conditionally accepted.

Gnedin, A. and Pitman, J. (2006).
Exchangeable Gibbs partitions and Stirling triangles.
Journal of Mathematical Scienes, 138:5674–5684.

References IV

Green, P. J. and Richardson, S. (2001).
Modelling heterogeneity with and without the Dirichlet process.
Scandinavian Journal of Statistics, 28:355–375.

Greve, J., Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. (2020).
Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis.
arXiv, 2012.12337.

Grün, B. (2019).
Model-based clustering.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 8, pages 157–192. CRC Press, Boca Raton, FL.

References V

Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. (2021).
How many data clusters are in the Galaxy data set? Bayesian cluster analysis in action.
Advances in Data Analysis and Classification, XX:forthcoming.

Guha, A., Ho, N., and Nguyen, X. (2019).
On posterior contraction of parameters and interpretability in Bayesian mixture modeling.
arXiv, 1901.05078.

Hartigan, J. A. (1990).
Partition models.
Communications in Statistics, Part A – Theory and Methods, 19:2745–2756.

Jain, S. and Neal, R. M. (2004).
A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.
Journal of Computational and Graphical Statistics, 13:158–182.

References VI

Jain, S. and Neal, R. M. (2007).
Splitting and merging components of a nonconjugate Dirichlet process mixture model.
Bayesian Analysis, 3:445–500.

Lau, J. W. and Green, P. (2007).
Bayesian model-based clustering procedures.
Journal of Computational and Graphical Statistics, 16:526–558.

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

Malsiner Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017).
Identifying mixtures of mixtures using Bayesian estimation.
Journal of Computational and Graphical Statistics, 26:285–295.

References VII

McCullagh, P. and Yang, J. (2008).
How many clusters?
Bayesian Analysis, 3:101–120.

Miller, J. W. and Harrison, M. T. (2018).
Mixture models with a prior on the number of components.
Journal of the American Statistical Association, 113:340–356.

Pitman, J. (1996).
Some developements of the Blackwell-MacQueen urn scheme.
In *Statistics, Probability and Game Theory*, volume 30 of *IMS Lecture Notes - Monograph Series*, pages 245–267.

Pitman, J. and Yor, M. (1997).
The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.
Annals of Probability, 25:855–900.

References VIII

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

Robert, C. P., Rydén, T., and Titterington, D. M. (2000).
Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method.
Journal of the Royal Statistical Society, Ser. B, 62:57–75.

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

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016).
mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models.
The R Journal, 8(1):289–317.

References IX

Sethuraman, J. (1994).
A constructive definition of Dirichlet priors.
Statistica Sinica, 4:639–650.

Stern, H., Arcus, D., Kagan, J., Rubin, D. B., and Snidman, N. (1994).
Statistical choices in infant temperament research.
Behaviormetrika, 21:1–17.