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
Part:

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures
Observations $\mathbf{y} = (y_1, \ldots, y_N)$ are an iid sample from a mixture distribution:

$$p(\mathbf{y} | \vartheta) = \sum_{k=1}^{K} \eta_k f_T(\mathbf{y} | \theta_k),$$

- $K$ is the number of components;
- the component densities $f_T(\mathbf{y} | \theta_k)$ arise from the same distribution $\mathcal{T}(\theta)$;
- $\theta_1, \ldots, \theta_K$ vary over the components;
- $\eta = (\eta_1, \ldots, \eta_K)$ 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:

$$p(\mathbf{y}_i | S_i = k) = f_T(\mathbf{y}_i | \theta_k), \quad \Pr(S_i = k) = \eta_k.$$
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 $y_i$ has a (latent) **indicator variable** $S_i$ indicating the component the observation belongs to:

$$y_i | S_i \sim f_T(y_i | \theta_{S_i}).$$

$y_i$ 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 \textbf{partition} \(C\) of the \(N\) data points,

\[ C = \{ C_1, \ldots, C_{K_+} \}, \]

which contains \(K_+ = |C|\) clusters [Hartigan, 1990]

- With \(S = (S_1, \ldots, S_N)\) being latent (random), we can look at the prior \(p(C)\) and the posterior distribution \(p(C|y)\) [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”
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”
Let $N_k$ is the number of observations allocated to component $k$, $k = 1, \ldots, K$.

Apriori, the occupation numbers are random:

$$(N_1, \ldots, N_K) \sim \text{MulNom}(N; \eta_1, \ldots, \eta_K).$$

Depending on the weights $\eta = (\eta_1, \ldots, \eta_K)$ 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 $C = \{C_1, \ldots, C_{K+}\}$ contains $K_+ < K$ 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 $K_+ < K$ with probability depending on $\eta, N, K$. 
The importance of the Dirichlet prior

- Consider a finite mixtures with $K$ fixed
- Assume a symmetric Dirichlet prior $\eta = (\eta_1, \ldots, \eta_K) \sim D_K(\gamma)$ on the weight distribution
- The hyperparameter $\gamma$ 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 $\gamma$
  [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 $K_+ < K$;
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 $N$. 

 Finite mixtures in Bayesian cluster analysis
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
A fully Bayesian mixture model is defined in a hierarchical way:

\[ K \sim p(K), \]
\[ \eta_1, \ldots, \eta_K|K, \gamma_K \sim D_K(\gamma_K), \]
\[ S_i|K, \eta_1, \ldots, \eta_K \sim M(1; \eta_1, \ldots, \eta_K), \text{ independently for } i = 1, \ldots, N, \]
\[ \phi \sim p(\phi), \]
\[ \theta_k|\phi \sim p(\theta_k|\phi), \text{ independently for } k = 1, \ldots, K, \]
\[ y_i|K, S_i = k, \theta_k \sim f_T(y_i|\theta_k), \text{ independently for } i = 1, \ldots, N. \]

Generic framework with no specific restrictions on

- \( f_T(\cdot|\theta_k) \) (parametric family),
- observations \( y_i \) can be univariate or multivariate, continuous, discrete-valued, mixed-type, time series data, outcomes of a regression model, ...
... 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 \( \gamma \)
  [Richardson and Green, 1997], [Miller and Harrison, 2018]:

  \[
  \gamma_K \equiv \gamma.
  \]

- **Dynamic MFMs** with hyperparameter \( \alpha \)
  [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:

\[
p(C|N, K, \gamma) = \binom{K}{K_+} K_+! \int p(S|\eta_K) p(\eta_K|K, \gamma) d\eta_K
\]

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

EPPF for a generalized finite mixture with hyperparameter \(\gamma_K\) (\(K\) fixed):

\[
p(C|N, K, \gamma_K) = \frac{K!}{(K - K_+)!} \frac{\Gamma(K\gamma_K)}{\Gamma(K\gamma_K + N)} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \gamma_K)}{\Gamma(\gamma_K)}.
\]
The EPPF of a generalized MFM

▶ EPPF for a generalized finite mixture with hyperparameter $\gamma_K$ ($K$ fixed):

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

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

▶ Takes the form of a **Gibbs-type prior** [Gnedin and Pitman, 2006]

▶ EPPF for a generalized MFM with prior $p(K)$:

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

▶ More general than a Gibbs-type prior, if $\gamma_K$ depends on $K$. 

The generalized mixture of finite mixtures model
Derivation of $p(K_+|N, \gamma)$

- Prior on $K_+$ obtained by summing over all partitions (challenging for large $N$):
  \[
p(K_+ = k|N, \gamma_K) = \sum_{C: K_+ = k} p(C|N, \gamma_K).
  \]

- We work with the prior on the labeled cluster sizes $p(N_1, \ldots, N_{K_+}|N, K, \gamma_K)$.

- Arrange the clusters in some exchangeable order (e.g. in order of appearance, see [Pitman, 1996]).

- By summing over all $N_1, \ldots, N_{K_+}$ 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 $E(K_+ | \gamma, N)$

Prior expectations $E(K_+ | \gamma, N)$ for static MFMs (left) and $E(K_+ | \alpha, N)$ for dynamic MFMs (right) as functions of $\gamma$ and $\alpha$ for $N = 100$ under the priors $K - 1 \sim \text{BNB}(1, 4, 3)$, $K - 1 \sim \text{Poisson}(4)$, and $K - 1 \sim \text{Geo}(0.1)$ in comparison to a DPM. For each prior $p(K)$, the prior expectation $E(K)$ 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
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 MFM) can be extended to dynamic MFM.

Telescoping Sampling:
- suggested in SFS, Malsiner-Walli and Grün (2021) for dynamic MFM;
- 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 $C$:

$$p(K|C, \gamma) \propto p(C|\gamma K, K)p(K)$$

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

- Combined with any MCMC algorithm for any finite mixture with a fixed number of components $K$, e.g. Gibbs sampling [Diebolt and Robert, 1994]
A partially marginalized sampler which switches between

- sampling from the complete-data posterior distribution conditional on the latent allocation variables $S$
- sampling $C$ from the collapsed posterior which lives in the set partition space and is marginalized over the empty components, the weight distribution and all allocations $S$ inducing $C$. 
MCMC scheme

1. Conditional on the parameters, update the partition $C$ by sampling $S_i$ from
   \[ \Pr(S_i = k | \eta_K, \theta_1, \ldots, \theta_K, y_i, K) \propto \eta_k f(y_i | \theta_k); \]
   - 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 $C$, update parameters of filled components and hyperparameters:
   - Sample $\theta_k$, for the components $k = 1, \ldots, K_+$, from $p(\theta_k | S, y, \phi)$.
   - Sample the hyperparameter $\phi$ from $p(\phi | \theta_1, \ldots, \theta_{K_+}, K_+)$.

3. Conditional on $C$, sample $K$ from $p(K | C, \gamma) \propto p(K)p(C | \gamma_K, K, N)$.

4. Conditional on $(K, \phi, C)$, add empty/non-filled components and update the weights:
   - If $K > K_+$: add $K - K_+$ empty components and sample $\theta_k | \phi$ from the prior $p(\theta_k | \phi), k = K_+ + 1, \ldots, K$.
   - Sample $\eta_K | K, \gamma_K, S \sim D(e_1, \ldots, e_K)$, where $e_k = \gamma_K + N_k$. 
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 :

▶ Finite mixtures in Bayesian cluster analysis
▶ The generalized mixture of finite mixtures model
▶ Telescoping sampler
▶ Applied mixture analysis
▶ Bridging finite and BNP mixtures
Prior choices in applied mixture analysis

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

BNB prior

\[ K - 1 \sim BNB(1, 1, 1) \]
\[ K - 1 \sim BNB(1, 4, 3) \]
\[ K - 1 \sim BNB(2, 1, 2) \]
\[ K - 1 \sim BNB(9, 4, 3) \]
[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, $K - 1 \sim \text{BNB}(1, 4, 3)$, $\alpha \sim \text{F}(6, 3)$) works very well.
Galaxy data

Number of components / clusters

Posteriors of $K$ (dashed blue lines, triangles) and $K^+$ (solid red lines, circles)

Applied mixture analysis
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

<table>
<thead>
<tr>
<th>p(K)</th>
<th>Thyroid</th>
<th>Fear</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(K⁺</td>
<td>y)</td>
<td>p(K</td>
</tr>
<tr>
<td>Geo (0.1)</td>
<td>3 [3, 3] 3 [3, 7]</td>
<td>4 [4, 7] 5 [5, 16]</td>
</tr>
</tbody>
</table>

- Posterior inference for $K$ and $K_+$ for a dynamic MFM based on different priors $p(K)$ and $\alpha \sim F(6, 3)$.

- The posteriors of $K_+$ and $K$ are summarized by their modes, followed by the 1st and 3rd quartiles.
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 ($\gamma_K = 1/K$, $K − 1 \sim BNB(1, 4, 3)$, $\pi_k \sim B(4, 4)$).
Mozambique data II

Prior (left) and posterior (middle) of $K$ (blue) and $K_+$ (red); trace plot of TS (right)
<table>
<thead>
<tr>
<th></th>
<th>church/comm</th>
<th>TV</th>
<th>friends</th>
<th>modern</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDSINFO-RADIO</td>
<td>0.51</td>
<td>0.98</td>
<td>0.76</td>
<td>0.95</td>
</tr>
<tr>
<td>AIDSINFO-TV</td>
<td>0.03</td>
<td>0.97</td>
<td>0.00</td>
<td>0.87</td>
</tr>
<tr>
<td>AIDSINFO-NEWS</td>
<td>0.03</td>
<td>0.23</td>
<td>0.00</td>
<td>0.99</td>
</tr>
<tr>
<td>AIDSINFO-POSTER</td>
<td>0.13</td>
<td>0.07</td>
<td>0.03</td>
<td>0.88</td>
</tr>
<tr>
<td>AIDSINFO-WKR</td>
<td>0.29</td>
<td>0.02</td>
<td>0.02</td>
<td>0.13</td>
</tr>
<tr>
<td>AIDSINFO-CHURCH</td>
<td>0.47</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>AIDSINFO-SCHOOL</td>
<td>0.17</td>
<td>0.20</td>
<td>0.06</td>
<td>0.26</td>
</tr>
<tr>
<td>AIDSINFO-COMM</td>
<td>0.76</td>
<td>0.05</td>
<td>0.15</td>
<td>0.06</td>
</tr>
<tr>
<td>AIDSINFO-FRND</td>
<td>0.01</td>
<td>0.45</td>
<td>0.58</td>
<td>0.37</td>
</tr>
<tr>
<td>AIDSINFO-WORK</td>
<td>0.08</td>
<td>0.02</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Size</td>
<td>0.04</td>
<td>0.19</td>
<td>0.75</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Part:

- Finite mixtures in Bayesian cluster analysis
- The generalized mixture of finite mixtures model
- Telescoping sampler
- Applied mixture analysis
- Bridging finite and BNP mixtures
Random probability measure priors like the Dirichlet process $GDP \sim \text{DP}(\alpha, G_0)$ [Ferguson, 1973, Ferguson, 1974] lead to infinite mixtures:

$$p(y) = \int f_T(y|\theta) GDP(d\theta) = \sum_{k=1}^{\infty} \eta_k f_T(y|\theta_k),$$

where $\eta_k$ 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 $v_1, v_2, v_3, \ldots$ of independent random variables (the sticks):

$$\eta_1 = v_1, \quad \eta_2 = (1 - v_1)v_2, \quad \eta_k = v_k \prod_{j=1}^{k-1} (1 - v_j).$$
Assume that the base measure $G_0$ is the same as the prior $p(\theta_k)$ in a finite mixture.

The “only” difference lies in the **prior of the sticks** $\nu_1, \nu_2, \nu_3, \ldots$:

- **Finite mixtures:** $\nu_k \sim B(\gamma_K, (K - k)\gamma_K), k = 1, \ldots, K - 1, \nu_K = 1$
- **Dirichlet process mixtures (DPM)** with $\text{DP}(\alpha, G_0)$: $\nu_k \sim B(1, \alpha)$
- **Pitman-Yor process mixtures** with $\text{PY}(\beta, \alpha)$ with reinforcement parameter $\beta \in [0, 1), \alpha > -\beta$ [Pitman and Yor, 1997]: $\nu_k \sim B(1 - \beta, \alpha + k\beta)$ (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 $\nu_K = 1$ a.s.
  - Yields a mixture with infinitely many components, of which only $K$ have non-zero weights, with the symmetric Dirichlet distribution $\mathcal{D}_K(|\beta|)$ 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 $\gamma$ 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 $\alpha$ 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 $\alpha$ is kept fixed. Yields a model beyond the class of Gibbs-type priors. [SFS, Malsiner-Walli and Grün, 2021].
Relation of dynamic MFM to DPMs

- (Dynamic) Sparse finite mixtures with $\gamma = \frac{\alpha}{K}$, $K$ fixed:
  - converge to a DPM with $GDP \sim DP(\alpha, G_0)$ 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:

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

- 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 MFM$s$ under correctly specified and misspecified components [Guha et al., 2019]
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).


An asymptotic analysis of a class of discrete nonparametric priors.
Statistica Sinica, 23:1299–1321.

Estimation of finite mixture distributions through Bayesian sampling.

A Bayesian analysis of some nonparametric problems.

Prior distributions on spaces of probability measures.
*Finite Mixture and Markov Switching Models.*

*Handbook of Mixture Analysis.*
CRC Press, Boca Raton, FL.

Generalized mixtures of finite mixtures and telescoping sampling. 
*Bayesian Analysis,* page conditionally accepted.

Exchangeable Gibbs partitions and Stirling triangles. 


How many data clusters are in the Galaxy data set? Bayesian cluster analysis in action.
*Advances in Data Analysis and Classification, XX*:forthcoming.

On posterior contraction of parameters and interpretability in Bayesian mixture modeling.
*arXiv, 1901.05078*.

Partition models.

A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model.
Splitting and merging components of a nonconjugate Dirichlet process mixture model.  
*Bayesian Analysis*, 3:445–500.

Bayesian model-based clustering procedures.  

Model-based clustering based on sparse finite Gaussian mixtures.  

Identifying mixtures of mixtures using Bayesian estimation.  
How many clusters?
_Bayesian Analysis, 3:101–120._

Mixture models with a prior on the number of components.

Some developments of the Blackwell-MacQueen urn scheme.

The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator.


