Bayesian Analysis (2021) **16**, Number 4, pp. 1279–1307

Generalized Mixtures of Finite Mixtures and Telescoping Sampling*

Sylvia Frühwirth-Schnatter†, Gertraud Malsiner-Walli‡, and Bettina Grün§

Abstract. Within a Bayesian framework, a comprehensive investigation of mixtures of finite mixtures (MFMs), i.e., finite mixtures with a prior on the number of components, is performed. This model class has applications in model-based clustering as well as for semi-parametric density estimation and requires suitable prior specifications and inference methods to exploit its full potential. We contribute by considering a generalized class of MFMs where the hyperparameter γ_K of a symmetric Dirichlet prior on the weight distribution depends on the number of components. We show that this model class may be regarded as a Bayesian non-parametric mixture outside the class of Gibbs-type priors. We emphasize the distinction between the number of components K of a mixture and the number of clusters K_+, i.e., the number of filled components given the data. In the MFM model, K_+ is a random variable and its prior depends on the prior on K and on the hyperparameter γ_K. We employ a flexible prior distribution for the number of components K and derive the corresponding prior on the number of clusters K_+ for generalized MFMs. For posterior inference we propose the novel telescoping sampler which allows Bayesian inference for mixtures with arbitrary component distributions without resorting to reversible jump Markov chain Monte Carlo (MCMC) methods. The telescoping sampler explicitly samples the number of components, but otherwise requires only the usual MCMC steps of a finite mixture model. The ease of its application using different component distributions is demonstrated on several data sets.

Keywords: Bayesian mixtures, Dirichlet process mixtures, sparse finite mixtures, Pitman-Yor process mixtures, reversible jump MCMC, Gibbs-type priors.

MSC2020 subject classifications: Primary 62H30; secondary 65C40.

1 Introduction

The present paper contributes to Bayesian mixture analysis where the number of components K is unknown and a prior on K is specified. This class of mixtures of finite mixtures (MFMs) has a long tradition in Bayesian mixture modeling (Richardson and Green, 1997; Nobile, 2004; McCullagh and Yang, 2008) and has gained recent attention by Miller and Harrison (2018); Geng et al. (2019); Xie and Xu (2020), among others.

*The authors gratefully acknowledge support from the *Austrian Science Fund (FWF)*, grant P28740, and through *WU Projects*, grant IA-27001574.

†Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Wien, Austria, sylvia.fruehwirth-schnatter@wu.ac.at

‡Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Wien, Austria, gertraud.malsiner-walli@wu.ac.at

§Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Wien, Austria, bettina.gruen@wu.ac.at

 https://doi.org/10.1214/21-BA1294

Previously considered MFMs differ with respect to the prior specifications on K and the component weights. We combine the different approaches to a *generalized MFM* model specification. We base our considerations on the crucial distinction between the *number of components* K in the mixture distribution and the *number of clusters* K_+ in the data which is defined as the number of "filled" mixture components used to generate the observed data. This fundamental distinction between K and K_+ has always been prevalent in Bayesian non-parametric (BNP) mixture analysis, see, e.g., the recent work by Argiento and De Iorio (2019). In applied finite mixture analysis, however, it is still common to assume that K and K_+ are the same entity, despite earlier work by Nobile (2004), McCullagh and Yang (2008), and, more recently, Miller and Harrison (2018).

Dirichlet process mixtures (DPMs) are the most popular BNP mixture approach. Their focus naturally lies on inference on the number of clusters, with K being fixed at $+\infty$. For DPMs, the number of clusters grows as $K_+ \sim \alpha \log(N)$ as the number of observations N increases. Doubt about the usefulness of DPMs for clustering has been voiced for many years and, indeed, Miller and Harrison (2013) proved inconsistency of DPMs for the number of clusters for the simple case of univariate Gaussian mixtures with unit variances. As a two-parameter alternative to DPMs, Pitman-Yor process mixtures were introduced in the BNP literature by Pitman and Yor (1997). Malsiner-Walli et al. (2016, 2017) introduced sparse finite mixtures (SFMs) in the context of applied finite mixture analysis. As shown by De Blasi et al. (2015), both model classes are closely connected. SFMs choose a fixed, clearly overfitting value of K in the spirit of Rousseau and Mengersen (2011) and a symmetric Dirichlet prior on the weight distribution with a very small hyperparameter γ_K. Whereas K is fixed, this choice allows the number of clusters K_+ to be a random variable taking values smaller than K. However, the larger K, the smaller γ_K has to be, motivating the "dynamic" SFM introduced in Frühwirth-Schnatter and Malsiner-Walli (2019), where $\gamma_K = \alpha/K$ was chosen with α being a hyperparameter independent of K.

The class of generalized MFMs we introduce in this paper is a finite mixture model with a prior on K, where the hyperparameter γ_K may change as a function of K. We consider two special cases of this specification. The *static* MFM uses a fixed value $\gamma_K \equiv \gamma$. The *dynamic* MFM uses $\gamma_K = \alpha/K$ and can thus induce a dynamic SFM with a prior on K. This MFM specification, considered previously in McCullagh and Yang (2008), is less common in applied finite mixture analysis than the static MFM. McCullagh and Yang (2008) conjecture that the static and dynamic versions of the MFM are quite different. We shed light on this by investigating the exchangeable partition probability function (EPPF), i.e., the prior induced on the random partition of the data (Pitman, 1995) by the generalized MFM, and discuss its specific form for static and dynamic MFMs. As shown in the seminal work by Gnedin and Pitman (2006), the static MFM considered in Richardson and Green (1997) and Miller and Harrison (2018) is equivalent to a BNP mixture with a Gibbs-type prior on the random partitions. Based on the EPPF of the generalized MFM, we show that the static MFM is the only mixture within this class that induces a Gibbs-type prior. Any specification where the hyperparameter γ_K varies with K leads to a BNP mixture beyond Gibbs-type priors. We focus on the dynamic MFM where $\gamma_K = \alpha/K$ is inversely proportional to the number of components K and show that it converges to a DPM with concentration parameter α, if the prior $p(K)$ puts all mass on $+\infty$. Hence, while staying within the finite mixture

framework, the dynamic MFM can be regarded as a "natural generalization" of the celebrated Dirichlet process prior beyond the class of Gibbs-type priors.

We propose the three-parameter beta-negative-binomial distribution as a prior on the number of components K which unifies priors proposed in Richardson and Green (1997); Nobile (2004); Cerquetti (2010); Miller and Harrison (2018); Grazian et al. (2020). Building on Antoniak (1974); Nobile (2004); Gnedin and Pitman (2006), we derive the implicitly induced prior on the number of clusters K_+ for generalized MFMs.

A tremendous challenge for Bayesian mixtures with an unknown number of components is practical statistical inference. To this aim, Richardson and Green (1997) introduced reversible jump Markov chain Monte Carlo (RJMCMC) for static MFMs with univariate Gaussian components. Exploiting that static MFMs are Gibbs-type priors, Miller and Harrison (2018) introduced sampling techniques from BNP statistics to finite mixture analysis. Applying the Chinese restaurant process (CRP) sampler of Jain and Neal (2004, 2007), they sample the partitions and, in this way, the number of clusters K_+ and infer the number of components K in a post-processing step by linking the distribution of K to the distribution of K_+.

In this paper, we introduce a novel MCMC algorithm for generalized MFMs called *telescoping sampling* that updates simultaneously the number of clusters K_+ and the number of components K during sampling without resorting to RJMCMC. As opposed to the CRP sampler, telescoping sampling also works outside the class of Gibbs-type priors. Sampling K only depends on the current partition of the data and is independent of the component parameters. This makes our sampler a most generic inference tool for finite mixture models with an unknown number of components which can be applied to arbitrary mixture families. Our sampler is easily implemented, for instance, for multivariate Gaussian mixtures with an unknown number of components K, and thus provides an attractive alternative to RJMCMC which is challenging to tune in higher dimensions, see, e.g., Dellaportas and Papageorgiou (2006).

The paper is structured as follows. In Section 2, we present the generalized MFM model and derive the EPPF. Section 3 proposes the beta-negative-binomial as a prior on the number of components K and derives the prior on the number of clusters K_+ for a generalized MFM. Section 4 discusses connections between applied finite mixture analysis based on MFMs and BNP mixtures. Our novel MCMC sampler is presented in Section 5 and is benchmarked against RJMCMC and the CRP sampler in Section 6. Additionally, MFMs with various uni- and multivariate component densities are applied both to artificial and real data of varying dimension and sample size. Section 7 concludes.

2 Generalized mixtures of finite mixture models

2.1 Model formulation

Consider N observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ of a uni- or multivariate continuous or discrete-valued variable. The generalized MFM is defined in a hierarchical way:

$$K \sim p(K), \tag{2.1}$$

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

where S_i is the latent allocation variable of observation $\mathbf{y}_i$, MulNom is the multinomial distribution, and $f(\mathbf{y}_i|\boldsymbol{\theta}_k)$ is the parametric density of component k. Model (2.1) depends on a sequence $\boldsymbol{\gamma} = \{\gamma_K\}$ of positive numbers which defines for each K the hyperparameter of the symmetric Dirichlet prior $\boldsymbol{\eta}_K | K, \gamma_K \sim \mathcal{D}_K(\gamma_K)$ on the component weights $\boldsymbol{\eta}_K = (\eta_1, \ldots, \eta_K)$. The component parameters $\boldsymbol{\theta}_k$ are independent conditional on the (random) hyperparameters ϕ. In combination with the invariance of the symmetric Dirichlet prior the prior specification is therefore invariant to label-switching.

Model (2.1) contains the finite mixture model with a prior on the number of components K studied by Richardson and Green (1997) and Miller and Harrison (2018), who termed this model a mixture of finite mixtures (MFM), as that special case where $\gamma_K \equiv \gamma$. As noted by Miller and Harrison (2018), assuming the same γ for all K is a "genuine restriction" which considerably simplifies the derivation of the implied partition distribution – a crucial ingredient to their inference algorithm. McCullagh and Yang (2008) extend this "static" MFM with constant γ by specifying a "dynamic" MFM where $\gamma_K = \alpha/K$ is inversely proportional to K and depends on a hyperparameter α, i.e., $\boldsymbol{\eta}_K | K, \alpha \sim \mathcal{D}_K(\alpha/K)$.

For a given K, K_+ is defined as the number of components that generated the data, i.e., $K_+ = \sum_{k=1}^{K} I\{N_k > 0\}$, where $N_k = \#\{i : S_i = k\}$ counts the observations generated by component k. In the following we refer to K_+ as the number of clusters. Including a prior $p(K)$ leads to both K_+ and K being random a priori. As opposed to the common perception that for a finite mixture K_+ given K is deterministic and equal to K, we show in Section 3 that the sequence of hyperparameters $\boldsymbol{\gamma} = \{\gamma_K\}$ has a crucial impact on the induced prior of the data partitions and the number of clusters K_+. For a static MFM with $\gamma = 1$ (Richardson and Green, 1997; Miller and Harrison, 2018), e.g., the prior expected number of clusters, $\mathrm{E}(K_+|N, \gamma = 1)$, is indeed close to $\mathrm{E}(K)$ for many priors $p(K)$ with finite mean, even for small N. However, having γ_K decrease with increasing K induces randomness in the prior distribution of K_+ given K, allowing for a gap between K_+ and K for a wide range of α and N values.

Under model (2.1), the joint distribution of the data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ has a representation as a countably infinite MFM with K components:

$$
p(\mathbf{y}) = \sum_{K=1}^{\infty} p(K) \prod_{i=1}^{N} p(\mathbf{y}_i|K), \quad p(\mathbf{y}_i|K) = \sum_{k=1}^{K} \eta_k f(\mathbf{y}_i|\boldsymbol{\theta}_k). \tag{2.2}
$$

The type of mixtures which are summed over in (2.2) vary with the prior parameter γ_K of the component weights. Using a symmetric Dirichlet prior, a priori the mean of the component weights given K is equal to a vector of dimension K with values $1/K$. However, the variance decreases with increasing hyperparameter γ_K and thus

more prior mass is assigned to balanced weight distributions. On the other hand, the variance increases and the component weights a priori become more unbalanced with decreasing values of γ_K. For a static MFM with $\gamma_K \equiv \gamma$, mixtures of a similar type are combined. For a dynamic MFM with $\gamma_K = \alpha/K$, mixtures favoring different component size distributions are combined: standard mixture models with balanced components, which emerge for small K, are mixed with SFMs for moderate K and finally, as K goes to infinity, with DPMs favoring extremely unbalanced component sizes. As will be shown in Section 2.2, the dynamic prior on the component weights increases the flexibility of the prior induced on the partitions and the number of clusters K_+ and leads outside the family of Gibbs-type priors. Moreover, a hyperprior on α, to be discussed in Section 4.3, achieves additional adaptivity of the induced prior on the partitions to the data at hand.

2.2 The EPPF and the prior distribution of cluster sizes

The MFM model (2.1) induces through the latent indicators $\mathbf{S} = (S_1, \ldots, S_N)$ a random partition $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_+}\}$ of the N data points into K_+ clusters where each cluster $\mathcal{C}_j$ contains all observations generated by the same mixture component, i.e., $S_i = S_j$ for all $\mathbf{y}_i, \mathbf{y}_j \in \mathcal{C}_j$, see Lau and Green (2007). In the tradition of Pitman (1995), we derive in Theorem 2.1 the prior partition probability function $p(\mathcal{C}|N, \boldsymbol{\gamma})$ of a generalized MFM for a given sequence $\boldsymbol{\gamma} = \{\gamma_K\}$ and discuss static MFMs with $\gamma_K \equiv \gamma$ and dynamic MFMs with $\gamma_K = \alpha/K$ as special cases. In addition, we derive the prior distribution $p(N_1, \ldots, N_{K_+}|N, \boldsymbol{\gamma})$ of the *labeled* cluster sizes $N_j = \text{card}\,(\mathcal{C}_j)$, where the K_+ clusters in $\mathcal{C}$ are arranged in some exchangeable random order and we assign label 1 to the first cluster, label 2 to the second cluster and so forth (Pitman, 2006).[1]

Theorem 2.1. *For a generalized MFM with proper prior $p(K)$ and $\boldsymbol{\eta}_K|K, \boldsymbol{\gamma} \sim \mathcal{D}_K(\gamma_K)$, the probability mass function $p(\mathcal{C}|N, \boldsymbol{\gamma})$ of the set partition $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_+}\}$ and the prior distribution $p(N_1, \ldots, N_{K_+}|N, \boldsymbol{\gamma})$ of the labeled cluster sizes are given by:*

$$p(\mathcal{C}|N, \boldsymbol{\gamma}) = \sum_{K=K_+}^{\infty} p(K) p(\mathcal{C}|N, K, \gamma_K), \tag{2.3}$$

$$p(\mathcal{C}|N, K, \gamma_K) = \frac{V_{N,K_+}^{K,\gamma_K}}{\Gamma(\gamma_K)^{K_+}} \prod_{j=1}^{K_+} \Gamma(N_j + \gamma_K), \qquad \textit{where } N_j = \text{card}\,(\mathcal{C}_j), \tag{2.4}$$

$$p(N_1, \ldots, N_{K_+}|N, \boldsymbol{\gamma}) = \frac{N!}{K_+!} \sum_{K=K_+}^{\infty} p(K) \frac{V_{N,K_+}^{K,\gamma_K}}{\Gamma(\gamma_K)^{K_+}} \prod_{j=1}^{K_+} \frac{\Gamma(N_j + \gamma_K)}{\Gamma(N_j + 1)}, \tag{2.5}$$

$$V_{N,K_+}^{K,\gamma_K} = \frac{\Gamma(\gamma_K K) K!}{\Gamma(\gamma_K K + N)(K - K_+)!}. \tag{2.6}$$

Being a symmetric function of the cluster sizes $(N_1, \ldots, N_{K_+})$, $p(\mathcal{C}|N, \boldsymbol{\gamma})$ is an EPPF (Pitman, 1995) and defines an exchangeable random partition of the N data points

[1] One such order is arrangement in order of appearance (Pitman, 1996), where the first observation $\mathbf{y}_1$ belongs to the first cluster and for each $j = 2, \ldots K_+$, the first observation not assigned to $\cup_{\ell=1}^{j-1} \mathcal{C}_\ell$ belongs to cluster $\mathcal{C}_j$. However, any other exchangeable random ordering will do.

for the class of generalized MFMs. The EPPF is instrumental for understanding the mathematical properties of the implied partitions and is a main object of interest in BNP mixtures, see, e.g., Lijoi and Prünster (2010).

An important class of BNP mixture models are mixtures relying on Gibbs-type random probability measures, or *Gibbs-type priors*, introduced in the seminal work by Gnedin and Pitman (2006). They are considered the most natural generalization of DPMs as they allow better control of the clustering behavior, see the excellent work of De Blasi et al. (2015). Under a Gibbs-type prior, the EPPF takes a specific product form and can be compared with the EPPF of a generalized MFM. Relying on Gnedin and Pitman (2006), Gnedin (2010) and De Blasi et al. (2013), among others, Miller and Harrison (2018) show that a static MFM induces a Gibbs-type prior on the partitions. Indeed, for $\gamma_K \equiv \gamma$ the EPPF in (2.3) takes the following product form:

$$p(\mathcal{C}|N,\gamma) = V_{N,K_+}^{\gamma} \prod_{j=1}^{K_+} \frac{\Gamma(N_j+\gamma)}{\Gamma(\gamma)}, \tag{2.7}$$

where $V_{N,k}^{\gamma} = \sum_{K=k}^{\infty} p(K) \frac{K!\Gamma(\gamma K)}{(K-k)!\Gamma(\gamma K+N)}$ satisfies the following recursion for $k = 1, \ldots, N-1$ (see Appendix A in the supplementary material (Frühwirth-Schnatter et al., 2021) for a proof):[2]

$$V_{N,k}^{\gamma} = (N+\gamma k) V_{N+1,k}^{\gamma} + V_{N+1,k+1}^{\gamma}. \tag{2.8}$$

However, for a generalized MFM with γ_K depending on K, we obtain a mixture model with a partition structure beyond Gibbs-type priors. For a dynamic MFM, we establish in Theorem 2.2 that the EPPF $p(\mathcal{C}|N,\alpha)$ can be expressed explicitly in relation to a DPM with precision parameter α, for which the EPPF is given by the Ewens distribution:

$$p_{\mathrm{DP}}(\mathcal{C}|N,\alpha) = \frac{\alpha^{K_+}\Gamma(\alpha)}{\Gamma(\alpha+N)} \prod_{j=1}^{K_+} \Gamma(N_j). \tag{2.9}$$

Theorem 2.2. *For a dynamic MFM with $\gamma_K = \alpha/K$, the EPPF $p(\mathcal{C}|N,\alpha)$ can be expressed as:*

$$p(\mathcal{C}|N,\alpha) = p_{\mathrm{DP}}(\mathcal{C}|N,\alpha) \times \sum_{K=K_+}^{\infty} p(K) R_{\mathbf{N},K_+}^{K,\alpha}, \tag{2.10}$$

$$R_{\mathbf{N},K_+}^{K,\alpha} = \prod_{j=1}^{K_+} \frac{\Gamma(N_j+\frac{\alpha}{K})(K-j+1)}{\Gamma(1+\frac{\alpha}{K})\Gamma(N_j)K},$$

where $p_{\mathrm{DP}}(\mathcal{C}|N,\alpha)$ is the probability mass function (pmf) of the Ewens distribution and $\mathbf{N}$ is the vector of induced cluster sizes $(N_1,\ldots,N_{K_+})$.

[2] Note that the normalization $\tilde{V}_{N,k}^{\gamma} = \gamma^k V_{N,k}^{\gamma}$ is needed to represent (2.7) as the common EPPF of a Gibbs-type prior: $p(\mathcal{C}|N,K_+=k) = \tilde{V}_{N,k}^{\gamma} \prod_{j=1}^{k} W_{N_j}$, where $W_\ell = (1+\gamma)_{(\ell-1)!} = \frac{\Gamma(\ell+\gamma)}{\Gamma(1+\gamma)}$ are the rising factorials.

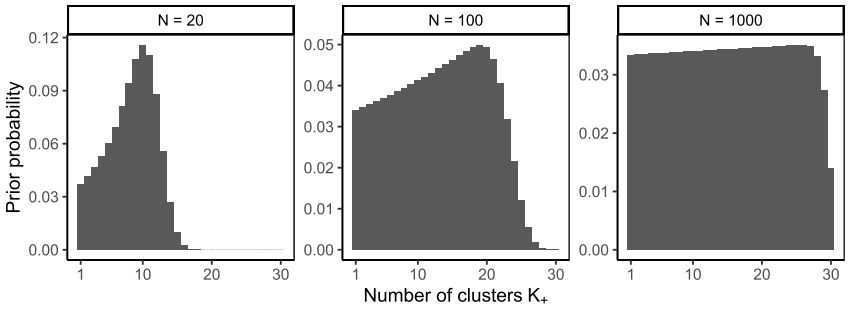

Figure 1: The implicit prior $p(K_+|N, \gamma = 1)$ on the number of clusters K_+ for the static MFM under the uniform prior $K \sim \mathcal{U}\{1, 30\}$ for various sample sizes, $N = 20, 100, 1000$.

It follows from Theorem 2.2 that dynamic MFMs can be regarded as a "natural generalization" of the celebrated Dirichlet process prior beyond the class of Gibbs-type priors. Theorems 2.1 and 2.2 (which are proven in Appendix A) are exploited further in Section 3 to derive the induced prior on the number of clusters $p(K_+|N, \boldsymbol{\gamma})$ and in Section 4 to investigate connections between applied finite mixture analysis based on MFMs and commonly used BNP mixtures in more depth.

3 The prior distributions of K and K_+

This section proposes a suitable choice for $p(K)$ and derives the implicit prior of K_+ in dependence of $p(K)$, the hyperparameters $\boldsymbol{\gamma}$ and N for a generalized MFM.

3.1 Choosing the prior on the number of components K

In their seminal paper, Richardson and Green (1997) suggest a uniform prior $K \sim \mathcal{U}\{1, K_{\max}\}$ for a static MFM with $\gamma_K \equiv 1$. However, depending on N, the prior on K_+ might be surprisingly informative and far from a uniform distribution. Figure 1 shows the implied prior $p(K_+|N, \gamma = 1)$ for a static MFM under the prior $K \sim \mathcal{U}\{1, 30\}$ for various sample sizes ($N = 20, 100, 1000$). Evidently, the prior mode depends on N and only for larger N approximately a uniform prior results.

Nobile (2004) shows that, as an alternative to the uniform prior, any proper prior $p(K)$ which satisfies $p(K) > 0$ for all $K \in \mathbb{N}$ can be adopted. While most discrete probability distributions include zero, in a mixture context the prior $p(K)$ has to exclude zero. This is often achieved by truncating the pmf at one, e.g., Nobile and Fearnside (2007) use the Poisson distribution $K \sim \mathcal{P}(1)$ restricted to $\{1, 2, \ldots, K_{\max}\}$. However, it is more convenient to work with the translated prior $K - 1 \sim p_t$, where the pmf $p(K) = p_t(K-1)$ is obtained by evaluating the translated pmf at $K-1$, as for translated priors hierarchical priors can be more easily introduced. We propose a translated prior, where $K - 1 \sim \text{BNB}(\alpha_\lambda, a_\pi, b_\pi)$ follows the beta-negative-binomial (BNB) distribution which is a hierarchical generalization of the Poisson, the geometric and the negative-

binomial distribution. The corresponding pmf is given by:

$$p(K) = p_t(K-1) = \frac{\Gamma(\alpha_\lambda + K - 1)B(\alpha_\lambda + a_\pi, K - 1 + b_\pi)}{\Gamma(\alpha_\lambda)\Gamma(K)B(a_\pi, b_\pi)}. \tag{3.1}$$

Appendix B provides the hierarchical derivation of the prior and illustrates the shapes for various parameter values. For $a_\pi > 1$, the expectation $\mathrm{E}(K) = 1 + \alpha_\lambda b_\pi/(a_\pi - 1)$ is finite. Prior (3.1) generalizes the prior derived by Cerquetti (2010) for the Gnedin-Fisher model and the prior derived by Grazian et al. (2020) from loss-based considerations which can be regarded as a $\mathrm{BNB}(1, b_\pi, a_\pi)$ prior. In their applications, Grazian et al. (2020) apply the $\mathrm{BNB}(1, 1, 1)$ prior with no finite moments.

The three-parameters α_λ, a_π and b_π of the $\mathrm{BNB}(\alpha_\lambda, a_\pi, b_\pi)$ prior allow simultaneous control over the expectation and the tails of $p(K)$ and the implied prior $p(K_+|N, \boldsymbol{\gamma})$ and its expectation $\mathrm{E}(K_+|N, \boldsymbol{\gamma})$. Priors $p(K)$ with finite expectation imply that $\mathrm{E}(K_+|N, \boldsymbol{\gamma})$ is finite, even for increasing N. In a clustering context, we propose to use the prior $K - 1 \sim \mathrm{BNB}(1, 4, 3)$ with $\mathrm{E}(K) = 2$. The induced prior on $p(K_+|N, \boldsymbol{\gamma})$ is investigated in more detail in Section 3.2 and differs considerably from previous choices such as the geometric or the uniform distribution. The $\mathrm{BNB}(1, 4, 3)$ prior leads to a weakly informative prior on K_+ which is concentrated on a moderate number of clusters and exhibits fat tails to ensure that also a high number of clusters may be estimated.

3.2 The induced prior on the number of clusters K_+

In applied mixture analysis, we often aim at partitions of the data with a finite, but a priori random number of clusters K_+. Since the number K of components is random a priori for a MFM, this induces K_+ to be random as well, but the induced prior $p(K_+|N)$ on K_+ does not necessarily coincide with the prior $p(K)$ for a finite number of observations N. The induced prior $p(K_+|N)$ has been derived earlier for various mixture models. For DPMs, Antoniak (1974) provides the prior of K_+ as $p_{\mathrm{DP}}(K_+|N, \alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} s_{N,K_+}$, where $s_{N,K_+} = \sum_{\mathcal{C}} \prod_{j=1}^{K_+} \Gamma(N_j)$ is the Stirling number of the first kind. Nobile (2004, Proposition 4.2) gives the prior on K_+ for a standard finite mixture, while Gnedin and Pitman (2006) derive $p(K_+|N)$ for Gibbs-type priors.

Building on this literature, we derive the prior $p(K_+|N, \boldsymbol{\gamma})$ for generalized MFMs under arbitrary priors $p(K)$. One way to obtain this prior is summing the EPPF (2.3) over all suitable partitions $\mathcal{C}$:

$$\Pr\{K_+ = k|N, \alpha\} = \sum_{K=k}^{\infty} p(K) V_{N,k}^{K,\gamma_K} (\gamma_K)^k S_{N,k}^{-1,\gamma_K}, \tag{3.2}$$

where $S_{N,k}^{-1,x} = \sum_{\mathcal{C}} \prod_{j=1}^{k} \Gamma(N_j + x)/\Gamma(1 + x)$ are the generalized Stirling numbers of the second kind. Alternatively, Theorem 3.1 derives $p(K_+|N, \boldsymbol{\gamma})$ from the prior of the *labeled* cluster sizes $p(N_1, \ldots, N_{K_+}|N, \boldsymbol{\gamma})$ given in (2.5).

Theorem 3.1. *For a generalized MFM with priors $p(K)$ and $\boldsymbol{\eta}_K|K, \boldsymbol{\gamma} \sim \mathcal{D}_K(\gamma_K)$, the prior of the number of clusters K_+ conditional on the sample size N is given for*

$k = 1, 2, \ldots, N$ *by:*

$$\Pr\{K_+ = k|N, \boldsymbol{\gamma}\} = \frac{N!}{k!} \sum_{K=k}^{\infty} p(K) \frac{V_{N,k}^{K,\gamma_K}}{\Gamma(\gamma_K)^k} C_{N,k}^{K,\gamma_K}, \tag{3.3}$$

where, for each K, $V_{N,k}^{K,\gamma_K}$ *has been defined in (2.6) and* $C_{N,k}^{K,\gamma_K}$ *is given by summation over the labeled cluster sizes* $(N_1, \ldots, N_k)$*:*

$$C_{N,k}^{K,\gamma_K} = \sum_{\substack{N_1,\ldots,N_k>0 \\ N_1+\ldots+N_k=N}} \prod_{j=1}^{k} \frac{\Gamma(N_j + \gamma_K)}{\Gamma(N_j + 1)}. \tag{3.4}$$

By matching (3.2) and (3.3), we find that $C_{N,k}^{K,\gamma_K}$ is related to the generalized Stirling numbers $S_{N,k}^{-1,\gamma_K}$ through

$$\frac{N!}{\Gamma(1+\gamma_K)^k k!} C_{N,k}^{K,\gamma_K} = S_{N,k}^{-1,\gamma_K}. \tag{3.5}$$

We found it convenient to compute $C_{N,k}^{K,\gamma_K}$ recursively through Algorithm 1. The recursion is straightforward to implement and scales well for large N, see Greve et al. (2020); Greve (2021) and Appendix A for mathematical derivations.

For a dynamic MFM with $\gamma_K = \alpha/K$, $C_{N,k}^{K,\gamma_K}$ can be written as $C_{N,k}^{K,\alpha}$ depending on K and α:

$$\Pr\{K_+ = k|N, \alpha\} = \frac{N!}{k!} \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha + N)} \sum_{K=k}^{\infty} p(K) C_{N,k}^{K,\alpha} \prod_{j=1}^{k} \frac{(K-j+1)}{K\Gamma(1+\frac{\alpha}{K})}. \tag{3.7}$$

Algorithm 1 Computing $C_{N,k}^{K,\gamma_K}$ for a generalized MFM.

1. Define the vector $\boldsymbol{c}_{K,1} \in \mathbb{R}^N$ and the $(N \times N)$ upper triangular Toeplitz matrix $\boldsymbol{W}_1$, where $w_n = \frac{\Gamma(n+\gamma_K)}{\Gamma(n+1)}$, $n = 1, \ldots, N$,

$$\boldsymbol{W}_1 = \begin{pmatrix} w_1 & \ddots & w_{N-1} & w_N \\ & w_1 & \ddots & w_{N-1} \\ & & \ddots & \ddots \\ & & & w_1 \end{pmatrix}, \qquad \boldsymbol{c}_{K,1} = \begin{pmatrix} w_N \\ w_{N-1} \\ \vdots \\ w_1 \end{pmatrix}.$$

2. For all $k \geq 2$, define the vector $\boldsymbol{c}_{K,k} \in \mathbb{R}^{N-k+1}$ as

$$\boldsymbol{c}_{K,k} = \begin{pmatrix} \mathbf{0}_{N-k+1} & \boldsymbol{W}_k \end{pmatrix} \boldsymbol{c}_{K,k-1}, \tag{3.6}$$

   where $\boldsymbol{W}_k$ is a $(N-k+1) \times (N-k+1)$ upper triangular Toeplitz matrix obtained from $\boldsymbol{W}_{k-1}$ by deleting the first row and the first column.
3. Then, for all $k \geq 1$, $C_{N,k}^{K,\gamma_K}$ is equal to the first element of the vector $\boldsymbol{c}_{K,k}$.

Putting all prior mass on $K = +\infty$, the following way to compute $p_{\mathrm{DP}}(K_+|N,\alpha)$ for a DPM emerges from (3.7),

$$p_{\mathrm{DP}}(K_+|N,\alpha) = \frac{N!}{K_+!}\frac{\alpha^{K_+}\Gamma(\alpha)}{\Gamma(\alpha+N)}C^{\infty}_{N,K_+}, \tag{3.8}$$

where C^{∞}_{N,K_+} is independent of α and obtained through recursion (3.6) with $w_n = 1/n$. For a static MFM, Theorem 3.1 simplifies to the following expression:

$$\Pr\{K_+ = k|N,\gamma\} = \frac{N!}{k!}\frac{V^{\gamma}_{N,k}}{\Gamma(\gamma)^k}C^{\gamma}_{N,k}, \tag{3.9}$$

where $V^{\gamma}_{N,k}$ is determined recursively from (2.8) and $C^{K,\gamma_K}_{N,k}$ is written as $C^{\gamma}_{N,k}$ independent of K and can be obtained in a single recursion from (3.6). Using, again, the normalization $\tilde{V}^{\gamma}_{N,k} = \gamma^k V^{\gamma}_{N,k}$, prior (3.9) is a special case of the prior given in Gnedin and Pitman (2006) for Gibbs-type priors:

$$\Pr\{K_+ = k|N,\gamma\} = \tilde{V}^{\gamma}_{N,k}B_{N,k}(W_\bullet), \tag{3.10}$$

where $B_{N,k}(W_\bullet)$ is the Bell polynomial in the W-structure $W_\bullet = \{W_\ell\}$ defined in Footnote 2.[3] Finally, putting all prior mass on $K = K_f$, (3.9) gives the result of Nobile (2004, Proposition 4.2) for a standard finite mixture:

$$\Pr\{K_+ = k|N,K = K_f,\gamma\} = \frac{N!}{k!}\frac{K_f!}{(K_f-k)!}\frac{\Gamma(\gamma K_f)}{\Gamma(\gamma K_f+N)\Gamma(\gamma)^k}C^{\gamma}_{N,k}. \tag{3.11}$$

For illustration, Figure 2 shows the impact of various priors $p(K)$ on the induced prior $p(K_+|N,\boldsymbol{\gamma})$ for static MFMs with $\gamma = 1$ (top row) and dynamic MFMs with $\alpha = 1$ (bottom row). The priors $p(K)$ in the three columns are the translated beta-negative-binomial prior $K-1 \sim \mathrm{BNB}(1,4,3)$ with $\mathrm{E}(K) = 2$ suggested in Section 3.1, the prior $K-1 \sim \mathrm{Geo}(0.1)$ with $\mathrm{E}(K) = 10$ suggested by Miller and Harrison (2018) and the uniform prior $K \sim \mathcal{U}\{1,30\}$ with $\mathrm{E}(K) = 15.5$ used by Richardson and Green (1997).

For static MFMs, $p(K_+) \approx p(K)$ for all three priors for values for K_+ and K between one and ten. In contrast, for a dynamic MFM $p(K_+)$ and $p(K)$ are only close for the BNB prior which has a small mean value. For the priors $p(K)$ with larger mean values, $p(K_+)$ considerably differs from $p(K)$ with mass being pulled towards smaller values of K_+. The corresponding posteriors of K and K_+ obtained under these priors for the famous Galaxy data are shown in Figure 6 in Section 6.2.

4 Bridging finite mixtures analysis and BNP mixtures

4.1 Connecting SFMs, MFMs and BNP mixtures

Generalized MFMs extend both Dirichlet process mixtures (DPMs) and sparse finite mixtures (SFMs). By allowing the number of components K to be finite and random,

[3] This follows from (3.5) and $S^{-1,\gamma}_{N,k} = B_{N,k}(W_\bullet)$ (Pitman, 2006, Eq. (1.20)).

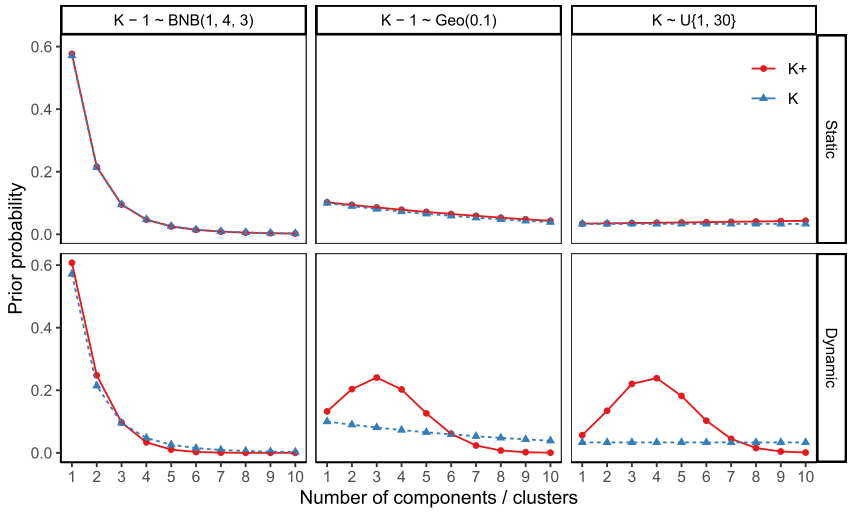

Figure 2: Priors of K (dashed blue lines, triangles) and K_+ (solid red lines, circles) under the priors $K-1 \sim \text{BNB}(1,4,3)$, $K-1 \sim \text{Geo}(0.1)$ and $K \sim \mathcal{U}\{1,30\}$ for a static MFM with $\gamma = 1$ (top) and dynamic MFM with $\alpha = 1$ (bottom), with $N = 82$.

MFMs provide notably more flexibility in the prior distribution on the partition space than DPMs and SFMs, similar to popular BNP mixtures (De Blasi et al., 2015).

SFMs result as that special case of MFMs, where $p(K) = I\{K = K_f\}$ puts all prior mass on a fixed number of components K_f. It follows from Theorem 2.2 and earlier work by Ishwaran and Zarepour (2000) that the prior distribution imposed on the partition space by a SFM lacks flexibility with increasing K_f and approaches the Ewens distribution (2.9) as $\gamma_{K_f} = \alpha/K_f$ approaches 0:

$$\lim_{K_f \to \infty} \frac{p(\mathcal{C}|\gamma_{K_f} = \alpha/K_f, K_f)}{p_{\text{DP}}(\mathcal{C}|\alpha = \gamma_{K_f} K_f)} = \lim_{K_f \to \infty} R^{K_f,\alpha}_{\mathbf{N},K_+} = 1.$$

This implies that SFMs do not easily deal with situations with many, well-balanced clusters, a behavior that is also observed for DPMs. By considering K as an additional second parameter following a prior $p(K)$, the dynamic MFM emerges as a more flexible family than a SFM with $K = K_f$ fixed. Dynamic MFMs can also be regarded as a more flexible extension of a DPM. Since $R^{K,\alpha}_{\mathbf{N},K_+}$ in Theorem 2.2 converges to 1 as K increases, putting all prior mass on $K = +\infty$ yields the Ewens distribution as limiting case. Thus, DPMs result as the limiting case of a dynamic MFM where the prior $p(K)$ increasingly concentrates all prior mass at $K = +\infty$.

Several close connections between MFMs and Pitman-Yor process mixtures (PYM) deserve to be mentioned. In Bayesian non-parametrics, mixtures based on the Pitman-Yor prior $\mathcal{PY}(\sigma,\theta)$ with $\sigma \in [0,1), \theta > -\sigma$ (Pitman and Yor, 1997) are a commonly used two-parameter alternative to DPMs which are the special case where $\sigma = 0$ and $\theta = \alpha$.

There exists a second family of PYMs, where $\sigma < 0$ and $\theta = K|\sigma|$ with $K \in \mathbb{N}$ being a natural number, see Gnedin (2010) and De Blasi et al. (2015). In the corresponding stick-breaking representation, stick $v_K = 1$ a.s. Hence, this prior yields a mixture with infinitely many components, of which only K have non-zero weights, with the symmetric Dirichlet distribution $\mathcal{D}_K(|\sigma|)$ acting as prior. Furthermore, at most K components can be populated. The EPPF of a PYM (with K known) reads:

$$p(\mathcal{C}|N,\sigma,\theta) = \frac{\Gamma(\theta)}{\Gamma(N+\theta)} \prod_{j=1}^{K_+} (\theta + \sigma(j-1)) \frac{\Gamma(N_j - \sigma)}{\Gamma(1-\sigma)}.$$

By matching EPPFs (and using $\Gamma(1-\sigma) = |\sigma|\Gamma(|\sigma|)$), it is evident that a finite mixture with K known and $\gamma_K > 0$ is equivalent to a mixture with a $\mathcal{PY}(-\gamma_K, K\gamma_K)$ prior, as proven in Gnedin and Pitman (2006). This equivalence of SFMs and PYMs provides a theoretical explanation of the empirical finding that SFMs can lead to more sensible cluster solutions than DPMs, see, e.g., Frühwirth-Schnatter and Malsiner-Walli (2019).

Even more interesting connections to BNP mixtures arise for MFMs, where K is random. As pointed out by Miller and Harrison (2014) and proven much earlier by Gnedin and Pitman (2006), for a static MFM, the dual BNP mixture is a Gibbs-type prior which arises from mixing a $\mathcal{PY}(-\gamma, K\gamma)$ prior over the concentration parameter $\theta_K = K\gamma$, while the reinforcement parameter $\sigma = -\gamma$ is fixed. The Fisher-Gnedin model studied in Gnedin (2010) is equivalent to a static MFM with $\gamma = 1$ and $K - 1 \sim \mathcal{P}(\lambda)$. The static MFM is also a special case of the class of mixtures based on normalized independent finite point processes recently introduced by Argiento and De Iorio (2019).

On the other hand, for a dynamic MFM, the prior partition distribution of the dual BNP mixture lies outside of the family of Gibbs-type priors, as it arises from mixing a $\mathcal{PY}(-\alpha/K, \alpha)$-prior over the reinforcement parameter $\sigma_K = -\alpha/K$, while the concentration parameter $\theta = \alpha$ is fixed, see also the discussion in De Blasi et al. (2015). As shown in Pitman (1996), a system of predictive distributions emerges from the EPPF, quantifying the probability that a new observation $\mathbf{y}_{N+1}$ belongs to any of the $K_+ = k$ existing clusters in $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ or creates partitions $\mathcal{C}^{\text{new}} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k, \mathcal{C}_{k+1}\}$ with a new cluster $\mathcal{C}_{k+1}$ of size $N_{k+1} = 1$. For a dynamic MFM the prior probability to introduce a new cluster for $\mathbf{y}_{N+1}$ is given by (see Appendix A for a proof):

$$\Pr\{\mathbf{y}_{N+1} \in \mathcal{C}_{k+1}|\mathbf{N}, K_+ = k, \alpha\} = \frac{\alpha}{\alpha + N} \left(1 - k \cdot \frac{\sum_{K=k}^{\infty} p(K)/K R^{K,\alpha}_{\mathbf{N},k}}{\sum_{K=k}^{\infty} p(K) R^{K,\alpha}_{\mathbf{N},k}}\right). \quad (4.1)$$

This probability (bounded by the predictive probability $\alpha/(\alpha + N)$ of a DPM) not only depends on N and the current number of clusters K_+, which characterizes Gibbs-type priors (De Blasi et al., 2013), but also on the occupation numbers $N_1, \ldots, N_{K_+}$. This confirms once more that dynamic MFMs, while staying within the finite mixture framework, are an example of a general random partition prior (De Blasi et al., 2015).

4.2 Comparing static and dynamic MFMs and DPMs

In the following we compare the induced priors on the number of clusters and the partitions for static and dynamic MFMs and DPMs in more detail and investigate the

influence of the prior on K and, respectively, the hyperparameters γ and α.

Regarding the prior on the number of clusters K_+, a fundamental question is whether a MFM allows K_+ to be different from K a priori, as for DPMs (where $K = +\infty$). To gain further understanding, we plot in Figure 3 the expectation of the induced prior $p(K_+|N, \boldsymbol{\gamma})$ as a function of γ (for static MFMs) and α (for DPMs and dynamic MFMs) for $N = 100$ under various priors $p(K)$. For both classes of MFMs, the gap between the expected number of clusters, $\mathrm{E}(K_+|N, \boldsymbol{\gamma})$, and the expected number of components, $\mathrm{E}(K)$, decreases for increasing γ or α. However, for dynamic MFMs the decrease is much slower and, even as α increases, a considerable gap remains between $\mathrm{E}(K_+|N, \boldsymbol{\gamma})$ and $\mathrm{E}(K)$. This is the effect of linking $\boldsymbol{\gamma}$ to K through $\gamma_K = \alpha/K$, thus avoiding that K_+ increases too quickly as K increases. This implies that the influence of the prior on K on the induced prior on K_+ is attenuated for an extended range of α values.

As emphasized by Green and Richardson (2001), beyond the induced prior on K_+, the *conditional* EPPF, induced for a given number of clusters $K_+ = k$,

$$p(N_1, \ldots, N_k|N, K_+ = k, \boldsymbol{\gamma}) = \frac{\Pr\{N_1, \ldots, N_k|N, \boldsymbol{\gamma}\}}{\Pr\{K_+ = k|N, \boldsymbol{\gamma}\}}, \tag{4.2}$$

is important for comparing mixture models. This prior allows a deeper understanding of the impact of choosing $\boldsymbol{\gamma}$ for MFMs on the partition distribution.

For a DPM, the conditional EPPF can be expressed using Theorem 3.1 as:

$$p_{\mathrm{DP}}(N_1, \ldots, N_k|N, K_+ = k) = \frac{1}{C^{\infty}_{N,k}} \prod_{j=1}^{k} \frac{1}{N_j}, \tag{4.3}$$

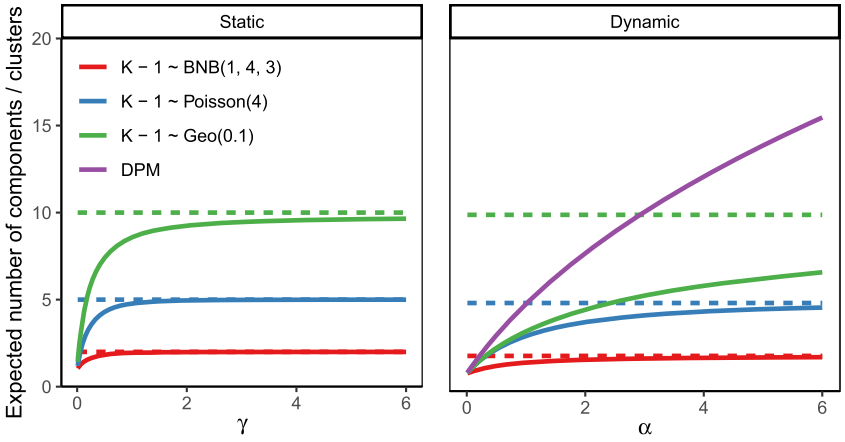

Figure 3: Prior expectations $\mathrm{E}(K_+|\gamma, N)$ for static MFMs (left) and $\mathrm{E}(K_+|\alpha, N)$ for dynamic MFMs (right) as functions of γ and α for $N = 100$ under the priors $K - 1 \sim \mathrm{BNB}(1, 4, 3)$, $K - 1 \sim \mathcal{P}(4)$, and $K - 1 \sim \mathrm{Geo}(0.1)$ in comparison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

and is known to be highly unbalanced (Antoniak, 1974), favoring partitions with some small values N_j due to the factors $1/N_j$, $j = 1, \ldots, k$ (Miller and Harrison, 2018). However, being independent of α, the conditional EPPF cannot be made more flexible for a DPM. In contrast, for a static MFM, the conditional EPPF depends on γ,[4]

$$p(N_1, \ldots, N_k | N, K_+ = k, \gamma) = \frac{1}{C_{N,k}^{\gamma}} \prod_{j=1}^{k} \frac{\Gamma(N_j + \gamma)}{\Gamma(N_j + 1)}. \tag{4.4}$$

For $\gamma = 1$, the uniform distribution over all partitions of N data points into $K_+ = k$ clusters results. Varying the hyperparameter γ introduces flexibility in the conditional EPPF for a static MFM: decreasing γ favors more unequal allocations, increasing γ favors partitions with more equal allocations. The conditional EPPF of a dynamic MFM is obtained by dividing (2.5) by (3.3):

$$p(N_1, \ldots, N_k | N, K_+ = k, \alpha) = \frac{\sum_{K=k}^{\infty} p(K) \frac{V_{N,k}^{K,\alpha}}{\Gamma(\frac{\alpha}{K})^k} \prod_{j=1}^{k} \frac{\Gamma(N_j + \frac{\alpha}{K})}{\Gamma(N_j + 1)}}{\sum_{K=k}^{\infty} p(K) \frac{V_{N,k}^{K,\alpha}}{\Gamma(\frac{\alpha}{K})^k} C_{N,k}^{K,\alpha}}. \tag{4.5}$$

This conditional EPPF depends both on α and $p(K)$, whereas the conditional EPPF of a static MFM is independent of $p(K)$. Thus, having a second parameter K, dynamic MFMs are more flexible than static MFMs regarding the conditional EPPF. Overall, in comparison to DPMs, static and dynamic MFMs induce more flexible prior structures both on the prior of the number of clusters and on the partition distribution, see Greve et al. (2020) for a detailed further investigation.

Additional flexibility is achieved by adjusting the hyperparameters γ and α to suit the data. In Section 4.3, a hyperprior on α is suggested, to achieve adaptivity of the induced prior on the partition to the data at hand. Also a static MFM can be combined with a prior on γ, rather than choosing a fixed value such as $\gamma = 1$.

4.3 Choosing the prior on α for dynamic MFMs

For a dynamic MFM the parameter α plays a crucial role for the prior distribution induced on the number of clusters and the partitions. On the one hand, the prior should have positive mass close to zero to allow a priori for a single cluster solution which corresponds to homogeneity. At the same time, fat tails should allow a priori larger values of K_+ and partitions with balanced cluster sizes.

The DPM literature would suggest a Gamma distribution $\alpha \sim \mathcal{G}(a, b)$ (e.g., Escobar and West, 1995; Jara et al., 2007). If $a = b \ll 1$, the expectation of α is one, while the variance is large, leading to a vague prior on α. For DPMs this induces a very informative prior on the number of clusters which is concentrated on 1 and $+\infty$ (see Dorazio, 2009; Murugiah and Sweeting, 2012). For dynamic MFMs, such a prior would – given its

[4]Note that Miller and Harrison (2018) report an approximate formula for the conditional EPPF of a static MFM, while our result is exact.

mode at zero – strongly favor homogeneity, and fail for data with balanced cluster sizes. Instead, we propose to use the F-distribution $\alpha \sim \mathcal{F}(\nu_l, \nu_r)$. The two parameters allow to control the behavior of the prior close to zero and in the tail independently. Choosing ν_r small gives fat tails. For a finite mean value, given by $\nu_r/(\nu_r - 2)$, but no higher moments, we specify $2 < \nu_r \leq 3$. Choosing a small value for ν_l allows independent control over the prior probability of homogeneity. Since the mode is given by $(\nu_l - 2)\nu_r/(\nu_l(\nu_r + 2))$, choosing $\nu_l > 2$ avoids a spike at 0. In our empirical analysis, we use $\alpha \sim \mathcal{F}(6, 3)$.

5 Inference algorithm: Telescoping sampling

A novel sampling method called *telescoping sampling* is introduced for a Bayesian analysis of finite mixtures with an unknown number of components which is related to, but also fundamentally different from RJMCMC (Richardson and Green, 1997) and the CRP sampler (Jain and Neal, 2004, 2007) applied in Miller and Harrison (2018).

Similar to Jain and Neal (2004, 2007), the telescoping sampler is a trans-dimensional Gibbs sampler which exploits the EPPF of a MFM given in (2.3). However, we do not work with the marginal EPPF $p(\mathcal{C}|N, \boldsymbol{\gamma})$, as Miller and Harrison (2018) do, but use a second level of data augmentation where we introduce the unknown number of components K, in addition to the partition $\mathcal{C}$, as a latent variable. This allows to apply the telescoping sampler outside the framework of Gibbs-type priors. We explicitly include K in the sampling scheme as in Richardson and Green (1997). However, rather than using RJMCMC, K is sampled conditional on $\mathcal{C}$ from the conditional posterior $p(K|\mathcal{C}, \gamma_K) \propto p(\mathcal{C}|N, K, \gamma_K)p(K)$ which is obtained by combining the *conditional* EPPF $p(\mathcal{C}|N, K, \gamma_K)$ provided in (2.4) with the prior $p(K)$:

$$p(K|\mathcal{C}, \gamma_K) \propto p(K)\frac{K!}{(K - K_+)!}\frac{\Gamma(\gamma_K K)}{\Gamma(N + \gamma_K K)\Gamma(\gamma_K)^{K_+}}\prod_{j:N_j>0}\Gamma(N_j + \gamma_K), \tag{5.1}$$

for $K = K_+, K_+ + 1, \ldots$, where K_+ is the number of clusters in $\mathcal{C}$.

While Miller and Harrison (2018) use (5.1) for static MFMs to infer K in a post-processing step, the telescoping (TS) sampler integrates (5.1) into a trans-dimensional Gibbs sampler for generalized MFMs and samples K and the partitions $\mathcal{C}$ (including K_+) in different blocks. Since $K \geq K_+$ by definition, the number of empty components $K - K_+$ varies over the iterations of the sampler, taking zero or a larger value. The difference between K and K_+ behaves similar to a telescope which can also be stretched or pulled together; hence the name of the sampler. Full details of the TS sampler are provided for dynamic MFMs in Algorithm 2. The TS sampler can also be applied with minor modifications to static MFMs (see Algorithm 3 in Appendix C). In both cases, the hyperparameter $\omega = \alpha$ or, respectively, $\omega = \gamma$ is assumed to be unknown.

Very conveniently, due to the conditional independence between the parameters $\boldsymbol{\theta}_k$ in the (non-empty) components and the number of components K, given the partition $\mathcal{C}$, K is sampled from the conditional posterior $p(K|\mathcal{C}, \gamma_K)$ given in (5.1) without any reference to the specific component distribution. Hence, the TS sampler is straightforward to implement and very generic, since the conditional posterior $p(K|\mathcal{C}, \gamma_K)$ does

Algorithm 2 Telescoping sampling for a dynamic MFM.

1. Update the partition $\mathcal{C}$ by sampling from $p(\mathbf{S}|\boldsymbol{\eta}_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathbf{y})$:
   (a) Sample S_i, for $i = 1, \ldots, N$, from $\Pr\{S_i = k|\boldsymbol{\eta}_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathbf{y}_i, K\} \propto \eta_k f(\mathbf{y}_i|\boldsymbol{\theta}_k)$, $k = 1, \ldots, K$.
   (b) Determine $N_k = \#\{i|S_i = k\}$ for $k = 1, \ldots, K$, the number $K_+ = \sum_{k=1}^{K} I\{N_k > 0\}$ of non-empty components and relabel such that the first K_+ components are non-empty.
2. Conditional on $\mathcal{C}$, update the parameters of the (non-empty) components:
   (a) For the (filled) components $k = 1, \ldots, K_+$, sample $\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{y}, \phi$ from
   $$p(\boldsymbol{\theta}_k|\mathbf{S}, \mathbf{y}, \phi) \propto p(\boldsymbol{\theta}_k|\phi) \prod_{i:S_i=k} f(\mathbf{y}_i|\boldsymbol{\theta}_k).$$
   (b) Sample the hyperparameter ϕ (if any) conditional on K_+ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_+}$ from
   $$p(\phi|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_+}, K_+) \propto p(\phi) \prod_{k=1}^{K_+} p(\boldsymbol{\theta}_k|\phi). \tag{5.2}$$
3. Conditional on $\mathcal{C}$, draw new values of K and α:
   (a) Sample K from
   $$p(K|\mathcal{C}, \alpha) \propto p(K) \frac{\alpha^{K_+} K!}{K^{K_+}(K - K_+)!} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(1 + \frac{\alpha}{K})}, \quad K = K_+, K_+ + 1, \ldots. \tag{5.3}$$
   (b) Use a random walk Metropolis-Hastings step with proposal $\log(\alpha^{\text{new}}) \sim \mathcal{N}\left(\log(\alpha), s_\alpha^2\right)$ to sample $\alpha|\mathcal{C}, K$ from
   $$p(\alpha|\mathcal{C}, K) \propto p(\alpha) \frac{\alpha^{K_+}\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \frac{\alpha}{K})}{\Gamma(1 + \frac{\alpha}{K})}.$$
4. Conditional on $K, \mathbf{S}, \alpha$ and ϕ, add $K - K_+$ empty components and update $\boldsymbol{\eta}_K$:
   (a) If $K > K_+$, then add $K - K_+$ empty components (i.e., $N_k = 0$ for $k = K_+ + 1, \ldots, K$) and sample $\boldsymbol{\theta}_k|\phi$ from the prior $p(\boldsymbol{\theta}_k|\phi)$ for $k = K_+ + 1, \ldots, K$.
   (b) Sample $\boldsymbol{\eta}_K|K, \alpha, \mathbf{S} \sim \mathcal{D}(e_1, \ldots, e_K)$, where $e_k = \alpha/K + N_k$.

not depend on the component parameters. This makes our sampler a most generic, easily implemented algorithm for finite mixture models with simultaneous inference on the unknown number of components and the unknown number of clusters for a wide range of component models. This greatly simplifies the application of MFMs in new application contexts allowing for arbitrary component distributions and extensions with

hierarchical priors. In contrast, the challenge to design good moves for RJMCMC is legendary. But also for CRP samplers (which are confined to static MFMs), the creation of new clusters requires knowledge of the marginal likelihood which depends on the chosen mixture family and might be difficult to work out for more complex mixtures.

More specifically, the TS sampler is a partially marginalized sampler, moving back and forth between sampling from the mixture posterior distribution $p(K, \mathbf{S}, \boldsymbol{\eta}_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \phi, \omega|\mathbf{y})$, which lives in the augmented parameter space of the mixture distribution, and sampling from the collapsed posterior $p(K, \mathcal{C}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_+}, \phi, \omega|\mathbf{y})$, which lives in the set partition space and is marginalized with respect to the parameters of the empty components, the weight distribution $\boldsymbol{\eta}_K$ and all allocations $\mathbf{S}$ that induce the same set partition $\mathcal{C}$. The full mixture posterior $p(K, \mathbf{S}, \boldsymbol{\eta}_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \phi, \omega|\mathbf{y})$ is proportional to

$$\prod_{k:N_k>0} p(\mathbf{y}^{[k]}|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\phi) \prod_{k:N_k=0} p(\boldsymbol{\theta}_k|\phi) \prod_{k=1}^{K} \eta_k^{N_k+\gamma_K-1} \frac{\Gamma(K\gamma_K)}{\Gamma(\gamma_K)^K} p(\phi)p(K)p(\omega), \quad (5.4)$$

where $\mathbf{y}^{[k]}$ are the $N_k > 0$ observations in cluster $\mathcal{C}_k$ of the partition $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_+}\}$ implied by $\mathbf{S}$ (after reordering such that the K_+ non-empty clusters appear first). The posterior (5.4) lends itself to the conditional sampling Step 1 of the TS sampler which is a standard step for finite mixtures with K known. The TS sampler is related to conditional samplers for infinite mixtures insofar, as all indicators $\mathbf{S}$ are sampled jointly due to the conditional independence of $\mathbf{S}$ given $\boldsymbol{\eta}_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K, \mathbf{y}$. As opposed to this, the CRP sampler applied in Miller and Harrison (2018) is a single-move sampler updating the allocation of each observation one-at-a-time.

Integrating (5.4) with respect to the weight distribution $\boldsymbol{\eta}_K$, the parameters $\boldsymbol{\theta}_k$ of the empty components and all allocations $\mathbf{S}$ that induce the same partition $\mathcal{C}$ yield (after suitable relabeling) the collapsed posterior which lives in the set partition space:

$$\begin{aligned} p(K, \mathcal{C}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_+}, \phi, \omega|\mathbf{y}) &\propto \prod_{k=1}^{K_+} p(\mathbf{y}^{[k]}|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\phi) \frac{\Gamma(K\gamma_K)}{\Gamma(\gamma_K)^K} p(\phi)p(K)p(\omega) \\ &\quad \cdot \int \prod_{k:N_k=0} p(\boldsymbol{\theta}_k|\phi) d(\boldsymbol{\theta}_{K_++1}, \ldots, \boldsymbol{\theta}_K) \sum_{\mathbf{S}:\mathbf{S}\in\mathcal{C}} \int \prod_{k=1}^{K} \eta_k^{N_k+\gamma_K-1} d\boldsymbol{\eta}_K \\ &= \prod_{k=1}^{K_+} p(\mathbf{y}^{[k]}|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\phi) \frac{K!}{(K-K_+)!} \frac{\Gamma(K\gamma_K)\prod_{k=1}^{K_+}\Gamma(N_k+\gamma_K)}{\Gamma(N+K\gamma_K)\Gamma(\gamma_K)^{K_+}} p(\phi)p(K)p(\omega). \quad (5.5) \end{aligned}$$

We see in (5.5) that updating of the parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{K_+}$ and ϕ (Step 2) can be performed independently from updating K and the hyperparameter ω (Step 3). It should be noted that the conditional posterior $p(K|\mathcal{C}, \omega)$ of K given $\mathcal{C}$ that results from (5.5) is identical with (5.1), verifying the validity of Step 3(a) (or 3(a*)) in our partially marginalized sampler. In practice, Step 3(a) (or 3(a*)) is implemented by considering an upper bound $K_{\max}$ for K and sampling K from a multinomial distribution over $\{K_+, \ldots, K_{\max}\}$, with the success probabilities being proportional to the non-normalized posterior probability of K. In the following empirical analysis we use a maximum value of $K_{\max} = 100$.

The sampler returns to conditional sampling from the full mixture posterior in Step 4(b) (or 4(b*)), by sampling the parameters of the empty components conditional on ϕ and sampling the weight distribution $\boldsymbol{\eta}_K$ from the conventional Dirichlet posterior distribution. Using the stick breaking representation of a finite mixture, with the sticks following $v_k|K, \gamma_K \sim \mathcal{B}(\gamma_K, (K-k)\gamma_K)$, Step 4(b) (or 4(b*)) can be rewritten in terms of sampling the sticks from a generalized Dirichlet distribution, see, e.g., Algorithm 1 of Frühwirth-Schnatter and Malsiner-Walli (2019).

In order to learn the component parameters, a hierarchical prior structure is introduced in the Bayesian mixture model (2.1). Basically, in Step 2(b) of the TS sampler, any hierarchical prior $p(\phi)$ on the model parameters can be used. For other samplers, such as the allocation sampler (Nobile and Fearnside, 2007), the prior $p(\phi)$ has to be conditionally conjugate to easily integrate out the component parameters $\boldsymbol{\theta}_k$. A specific feature of the TS sampler is that the hyperparameters ϕ are learned in Step 2(b) only from the K_+ *filled* components and that the parameters of the $K - K_+$ empty components are sampled subsequently in Step 4(a) from the conditional prior $p(\boldsymbol{\theta}_k|\phi)$ for $k = K_+ + 1, \ldots, K$. In this way, the parameters of the filled components inform the parameters of the empty components. In our opinion, this is an elegant way to handle hierarchical priors for component parameters in a dimension changing framework.

The TS sampler allows for a varying, but conditionally finite model dimension K. Truncation, however, does not result from slice sampling (Kalli et al., 2011), a popular method for DPMs to turn the infinite mixture into a conditionally finite one. The TS sampler adds and deletes components as follows. Step 3(a) is a birth move, where new components are created, if a value $K > K_+$ is sampled. These components are empty, since we leave the filled components in partition $\mathcal{C}$ unchanged. Observations are allocated to any empty component during the subsequent sweep of the sampler in Step 1(a). Components can only disappear, if they get emptied in the allocation Step 1(a). Hence, for the TS sampler to work well, the tail probability $\sum_{K>K_+} p(K|\mathcal{C}, \omega)$ cannot be too small, as this probability controls how many empty components are added in Step 3(a) (or 3(a*)). The more $p(K|\mathcal{C}, \omega)$ is concentrated at K_+, the more likely mixing for K_+ and K will be poor for the TS sampler. This is true both for static and dynamic MFMs.

Finally, we allow the hyperparameter of the weight distribution, either α or γ, to be an unknown parameter estimated from the data under a hyperprior. α (or γ) are updated in Step 3(b) (or 3(b*)), which is the only updating step where a random walk Metropolis-Hastings step is employed.

6 Empirical demonstrations

6.1 Benchmarking the telescoping sampler

We compare the performance of the TS sampler to two other samplers previously proposed to fit a static MFM with univariate Gaussian components, namely, reversible jump MCMC (RJ; Richardson and Green, 1997) and the Jain-Neal split-merge algorithm (JN; Jain and Neal, 2004, 2007; Miller and Harrison, 2018). In contrast to the TS sampler, where in each iteration both K and K_+ are updated, the RJ sampler just

Sampler	1	2	3	4	5	6	7	8	9	10	11	≥ 12
TS	.000	.000	.070	.161	.228	.228	.159	.087	.040	.017	.006	.003
RJ	.006	.000	.070	.161	.227	.226	.158	.086	.040	.017	.006	.003
JN	.000	.000	.070	.162	.228	.228	.159	.087	.040	.017	.006	.003

Table 1: Galaxy data. Mean estimates over 100 MCMC runs of the posterior of K_+ for the telescoping (TS), the RJMCMC (RJ) and the Jain-Neal (JN) sampler.

samples K while K_+ is calculated a posteriori from the sampled allocations, and the JN sampler just samples the partitions and thus K_+, whereas the posterior of K is reconstructed in a post-processing step (see Miller and Harrison, 2018, Equation (3.7)).

For this comparison we consider the well-known Galaxy data (Roeder, 1990), which is a small data set of $N = 82$ measurements on velocities of different galaxies from six well-separated sections of the space, and fit univariate Gaussian mixtures, $y_i|S_i = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right)$, with K unknown. Priors are chosen as in Richardson and Green (1997), namely $p(K)$ is a uniform distribution $\mathcal{U}\{1, 30\}$, $\boldsymbol{\eta}_K|K \sim \mathcal{D}_K\left(\gamma_K\right)$ with $\gamma_K \equiv 1$ is uniform, whereas $\mu_k \sim \mathcal{N}\left(m, R^2\right)$, $\sigma_k^2 \sim \mathcal{G}^{-1}\left(2, C_0\right)$, and $C_0 \sim \mathcal{G}\left(0.2, 10/R^2\right)$, where m and R are the midpoint and the length of the observation interval. These priors are imposed for sake of comparison with previous results, but not motivated by modeling considerations nor selected to favor the TS sampler.

Results were obtained for the RJ sampler using the Nmix software provided by Peter Green and for the JN sampler as implemented in Miller and Harrison (2018).[5] Each sampler was run for 1,000,000 iterations without thinning after discarding the first 10,000 iterations and using 100 different initializations. Table 1 summarizes the posterior $p(K_+|\mathbf{y})$ over all 100 runs based on the means for all three samplers (see Appendix D.1 for more detailed results). The posteriors estimated by all three samplers are very similar indicating that the TS sampler provides suitable draws from this posterior distribution.

The performance of the three samplers is compared by inspecting the number of clusters K_+ as well as the number of components K obtained for the MCMC iterations, if available. For this comparison, we use a simulated data set with a data generating process similar to the Galaxy data set and draw $N = 1000$ observations from a three-component univariate Gaussian mixture (see Figure D.1 in Appendix D.1). We specify priors on the component parameters as used in Richardson and Green (1997) for the Galaxy data set and fit a static MFM with $\gamma = 0.1$. The smaller value for the Dirichlet parameter increases the gap between the prior on K and K_+ and thus improves the mixing of the TS and RJ samplers. Each sampler is run for 100,000 iterations without thinning. The first 10% iterations are omitted as burn-in.

Figure 4 shows a combined trace plot of K_+ and K (if available) for each of the three samplers using the first 5,000 iterations after omitting burn-in. In each trace plot the black line shows how the number of clusters K_+ induced by the sampled partitions varies over the iterations. For the TS and RJ samplers, in addition, the gray lines show how the number of components K vary. For the TS sampler, K is sampled given K_+,

[5]Both are included in the supplementary material to Miller and Harrison (2018).

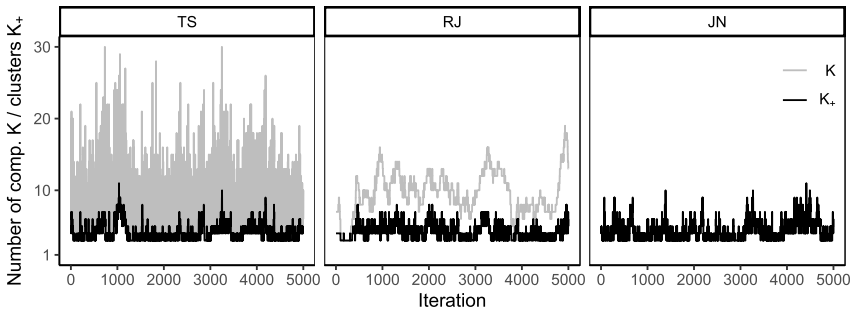

Figure 4: Simulated data, $N = 1000$, $\gamma_K \equiv 0.1$, all other priors as in Richardson and Green (1997). Trace plots of K (gray) and K_+ (black) for the TS, RJ and JN sampler.

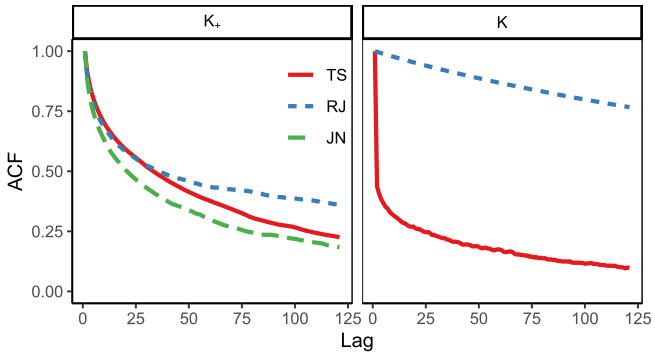

Figure 5: Simulated data, $N = 1000$, $\gamma_K \equiv 0.1$, all other priors as in Richardson and Green (1997). Auto-correlation function (ACF) for K_+ (left) and K (right) for the TS (solid red line), RJ (dashed blue line) and JN (long dashed green line) sampler.

while for the RJ sampler K changes if components are split or combined or due to a birth or death of an empty component. This difference is clearly visible in the trace plots with poorer mixing in K for the RJ relative to the TN sampler.

We assess the efficiency of the three samplers by estimating auto-correlation functions (ACFs) for the sampled K_+ and K values (if available) and visualizing them in Figure 5. Regarding K_+, the efficiency is rather comparable over the three samplers, with slight advantages for JN followed by TS and RJ being the least efficient. Comparing the ACFs for K clearly confirms that TS outperforms RJ.

The performance comparison indicates that TS is competitive with the other samplers, while providing the advantage of being easily adjusted and immediately applicable for mixtures with other component distributions or models. Note however, that an appropriate choice of γ_K has an impact on the efficiency of the sampler. While too large values of γ_K prevent that empty components are created, too small values induce many (superfluous) additional empty components.

6.2 Sensitivity to the prior choice on the number of components

In the following we use the TS sampler to investigate how the posteriors of K and K_+ vary in dependence of different prior specifications $p(K, \gamma_K)$ for the Galaxy data set. Although this data set is very popular in the clustering literature, there is no consensus on the number of clusters in the sample, see for instance Aitkin (2001), Grün et al. (2021) and the discussion in Appendix D.2.

In contrast to these previous Bayesian analyses, we keep the priors on the component parameters fixed to those as specified by Richardson and Green (1997) for all analyses. In this way, the impact of the priors on K and the component weights can be investigated without mixing these effects with those of different prior specifications on the component parameters. We consider static and dynamic MFMs with the same priors $p(K)$ and γ_K as specified in Figure 2, i.e., $K-1 \sim \text{BNB}(1,4,3)$, $K-1 \sim \text{Geo}(0.1)$, and $K \sim \mathcal{U}\{1,30\}$, and $\gamma = 1$ for the static MFM and $\alpha = 1$ for the dynamic MFM.

In Figure 6 in the top row, the posteriors of K and K_+ are reported for the static MFM with $\gamma_K \equiv 1$. The posteriors $p(K_+|\mathbf{y})$ and $p(K|\mathbf{y})$ are very similar to each other

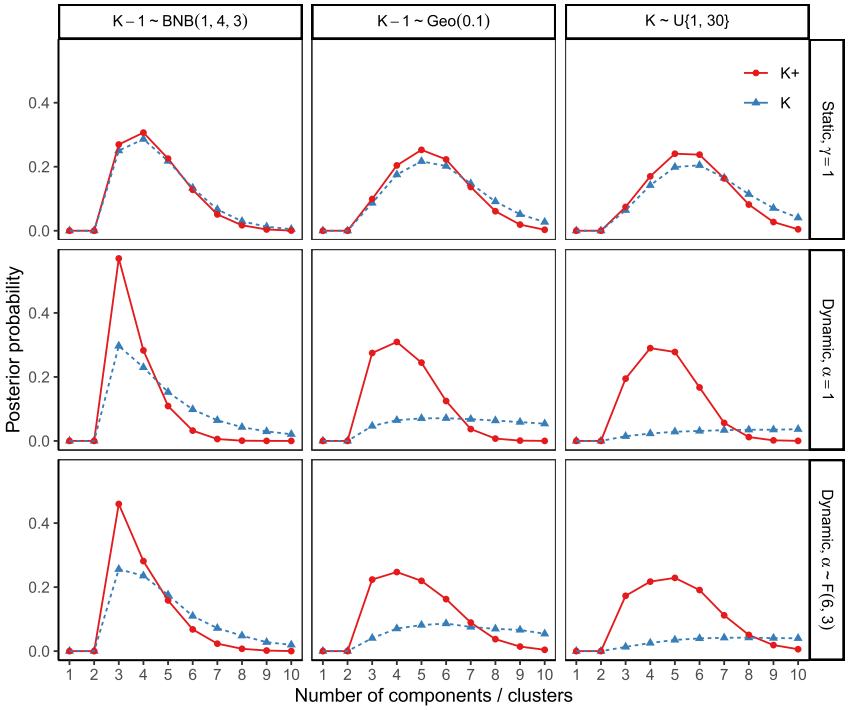

Figure 6: Galaxy data. Posteriors of K (dashed blue lines, triangles) and K_+ (solid red lines, circles) under priors $K-1 \sim \text{BNB}(1,4,3)$ (left), $K-1 \sim \text{Geo}(0.1)$ (middle) and $K \sim \mathcal{U}\{1,30\}$ (right) under a static MFM with $\gamma = 1$ (top) and dynamic MFMs with $\alpha = 1$ (middle) and $\alpha \sim \mathcal{F}(6,3)$ (bottom), for $N = 82$.

regardless of $p(K)$ specified. In contrast, for the dynamic prior $\gamma_K = 1/K$, shown in the middle row, the posteriors $p(K|\mathbf{y})$ and $p(K_+|\mathbf{y})$ differ considerably. While the posterior $p(K|\mathbf{y})$ becomes flatter compared to fixed $\gamma = 1$, most of the posterior mass of $p(K_+|\mathbf{y})$ concentrates on K_+ equal to 3, 4 or 5 which are reasonable values for the number of clusters in this data set. Comparing the posteriors of K_+ and K to the corresponding priors in Figure 2 indicates that the posteriors are strongly influenced by the prior distributions. E.g., the flat prior for K_+ induced by the uniform distribution and $\gamma_K = 1$ (plot in Figure 2 on the top right) results in a posterior of K_+ favoring large values between 4 and 7 clusters which clearly overestimates the number of clusters in this small data set. In contrast, a sparse prior on K_+ in combination with a dynamic MFM favors a sparse estimation of the number of clusters also a posteriori, see, e.g., the posterior $p(K_+|\mathbf{y})$ for the BNB $(1, 4, 3)$ prior where three clusters are estimated.

Under the hyperprior $\alpha \sim \mathcal{F}(6, 3)$, the posterior of K_+ looks rather similar to assuming that $\alpha = 1$ fixed, see Figure 6 at the bottom. However, if the shrinkage prior $\alpha \sim \mathcal{G}(1, 20)$ is specified, the posterior of K_+ becomes completely independent of both the prior and posterior of K, see Appendix D.2 where also results for other specifications on K and the weights are reported, in particular a FM, a SFM and a DPM model.

Figure 6 shows that depending on the prior on K and whether a static or dynamic MFM is specified, the posterior mode of $p(K_+|y)$ varies. This highlights the impact of the implicitly specified prior on K_+ on the posterior of K_+. This especially applies to the Galaxy data set which contains only $N = 82$ observations and has no clear cluster structure. If, in contrast, there is considerable information in the data, the posteriors of K_+ for different prior specifications $p(K)$ coincide, as can be seen in the next section when analyzing the Thyroid data set.

6.3 Changing the clustering kernel

We use the TS sampler to fit dynamic MFMs with different component distributions, i.e., the multivariate Gaussian distribution and the latent class model for multivariate categorical data. This demonstrates how easily the TS sampler can be used to fit a MFM regardless of the component distributions. For K we use the same priors $p(K)$ as in the previous section. It will turn out that a prior specification for K where $\mathrm{E}(K)$ is small and the tails are not too light, in combination with the dynamic prior $\gamma_K = \alpha/K$ on the component weights and $\alpha \sim \mathcal{F}(6, 3)$ gives good clustering results.

The final partition is obtained by identifying the models through the post-processing procedure suggested by Frühwirth-Schnatter (2006) and applied in Malsiner-Walli et al. (2016, 2017). First, the number of clusters $\hat{K}_+$ is estimated by the mode of the posterior $p(K_+|\mathbf{y})$. Then for all posterior draws where $K_+^{(m)} = \hat{K}_+$, the component parameters are clustered in the point process representation into $\hat{K}_+$ clusters using k-means clustering. A unique labeling of the draws is obtained and used to reorder all draws, including the sampled allocations. The final partition of the data is then determined by the maximum a posteriori (MAP) estimate of the relabeled cluster allocations.

	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]
Geo (0.1)	3	[3, 3]	3	[3, 7]	4	[4, 7]	5	[5, 16]
BNB $(1,4,3)$	3	[3, 3]	3	[3, 4]	2	[2, 4]	2	[2, 5]

Table 2: Thyroid and Fear data. Posterior inference for K and K_+ for a dynamic MFM based on different priors $p(K)$ and $\alpha \sim \mathcal{F}(6,3)$. The posteriors of K_+ and K are summarized by their modes, followed by the 1st and 3rd quartiles.

Multivariate Gaussian mixtures: Thyroid data

The Thyroid data are a benchmark data set for multivariate normal mixtures included in the *R* package *mclust* (Scrucca et al., 2016). It consists of five laboratory test variables which are used for clustering and a categorical variable indicating the operation diagnosis (with three potential values) for 215 patients. A dynamic MFM with multivariate normal component densities is fitted using a simplified version of the priors proposed in Malsiner-Walli et al. (2016) for the component parameters (for details see Appendix D.3). As can be seen in the left-hand column of Table 2, for all priors on K the mode of the posteriors for K_+ lies at three, even for the uniform prior. Also the posterior mode of K is three, indicating that rarely empty components were sampled. For the $K-1 \sim \text{BNB}(1,4,3)$ prior, the final partition obtained through the MAP estimate consists of three clusters with 28, 37 and 150 patients. The adjusted Rand index (ARI) of this partition with the known operation diagnosis is 0.88, which is equal to the ARI of the *mclust* solution. Overall these results suggest that, if the data are informative regarding a specific cluster structure, the clustering result is not susceptible to the prior specification of $p(K)$.

Latent class analysis: Fear data

Stern et al. (1994) consider data of 93 children in the context of infant temperamental research. For each child, three categorical features are observed, namely motor activity (M) with 4 categories, fret/cry behavior (C) with 3 categories, and fear of unfamiliar events (F) with 3 categories, see Frühwirth-Schnatter and Malsiner-Walli (2019) for the contingency table of the data. The scientific hypothesis is that two different profiles in children are present. To test this, a latent class model is fitted using a dynamic MFM with a uniform Dirichlet prior on the component parameters. Table 2 shows that the prior $K-1 \sim \text{BNB}(1,4,3)$ selects $\hat{K}_+ = 2$, confirming the theoretically expected number of clusters. The geometric prior with $\text{E}(K) = 10$ and the truncated uniform prior, however, overestimates the number of clusters with the mode of K_+ at 4 and 6, respectively. The results obtained when identifying the MCMC output from a dynamic MFM with $K-1 \sim \text{BNB}(1,4,3)$ and $\alpha \sim \mathcal{F}(6,3)$ indicate that the two classes have a rather different profile regarding the occurrence probabilities of the categories (see Appendix D.3), which coincides with the findings in Stern et al. (1994).

6.4 Investigating the telescoping sampler with artificial data

We perform a simulation study with artificial data to investigate how the TS sampler performs in dependence of sample size N, dimension r and number of clusters K_+. In addition, we vary the priors for $p(K, \gamma_K)$ considering static and dynamic MFMs and in particular include the suggested priors $K-1 \sim \text{BNB}(1,4,3)$ and $\alpha \sim \mathcal{F}(6,3)$. We sample 100 data sets from a multivariate normal mixture with eight equally sized components, varying dimension ($r = 2, 8, 12$) and increasing sample size ($N = 400, 4000, 10000$), combining higher values of the dimension r with larger sample sizes N. A detailed description of the data generating processes of the simulated data as well as the specified priors $p(K)$ and Dirichlet parameters γ and α and the inference scheme employed is given in Appendix D.4.

Results are visualized in a bubble plot in Figure 7. The area of the bubbles is proportional to the percentage of data sets with a specific number of clusters K_+ estimated as indicated on the y-axis. The results show how the influence of the prior $p(K, \gamma_K)$ decreases when the information in the sample increases. If the information is weak, i.e., for $N = 400$ and $r = 2$, the prior specifications on K and on γ_K have considerable impact on the clustering result (first column of Figure 7). The estimated number of clusters K_+ tends to be lower for the Poisson prior regardless of the prior imposed on γ_K. While the Poisson prior with $\lambda = 1$ induces the same prior mean $\text{E}(K) = 2$ as the $\text{BNB}(1,4,3)$ prior, it has also light tails. Thus, the fatter tails of the $\text{BNB}(1,4,3)$ prior allow to estimate the number of clusters in the data correctly despite its sparsity inducing properties. Regarding the prior on the Dirichlet parameter γ_K, the results of the static MFM clearly indicate that the estimated number of clusters decreases for decreasing values of γ. In the dynamic case, using $\alpha \sim \mathcal{F}(6,3)$ gives more reliable results than the other specifications for α regardless of the prior on K. In contrast, the influence of the sparsity inducing prior $\alpha \sim \mathcal{G}(1,20)$ is clearly visible across all priors on K, leading to less than eight clusters being estimated for the majority of the data sets. Overall, the results for the combination $K - 1 \sim \text{BNB}(1,4,3)$ and $\alpha \sim \mathcal{F}(6,3)$ confirm the suitability of this prior specification for determining the number of clusters in a Bayesian cluster analysis application. For $N = 4000$ the estimated number of data clusters $\hat{K}_+$ is equal to eight for nearly all data sets regardless of the prior specifications. Results are similar for $N = 10000$.

7 Concluding remarks

Being a finite mixture model where the number of components is unknown, the MFM model has a long tradition in Bayesian mixture analysis. Building on this tradition, a key aspect of our work is to explicitly distinguish between the number of components K in the mixture distribution and the number of clusters K_+ in the partition of the data, corresponding to non-empty components given the data. With this fundamental distinction in mind, we contribute to MFMs both from a methodological as well as a computational perspective.

Traditionally, the hyperparameter γ of a symmetric Dirichlet prior on the component weights is a fixed value, often equal to one. In this paper, we investigate in detail a more general MFM specification which defines the hyperparameter γ_K of the symmetric

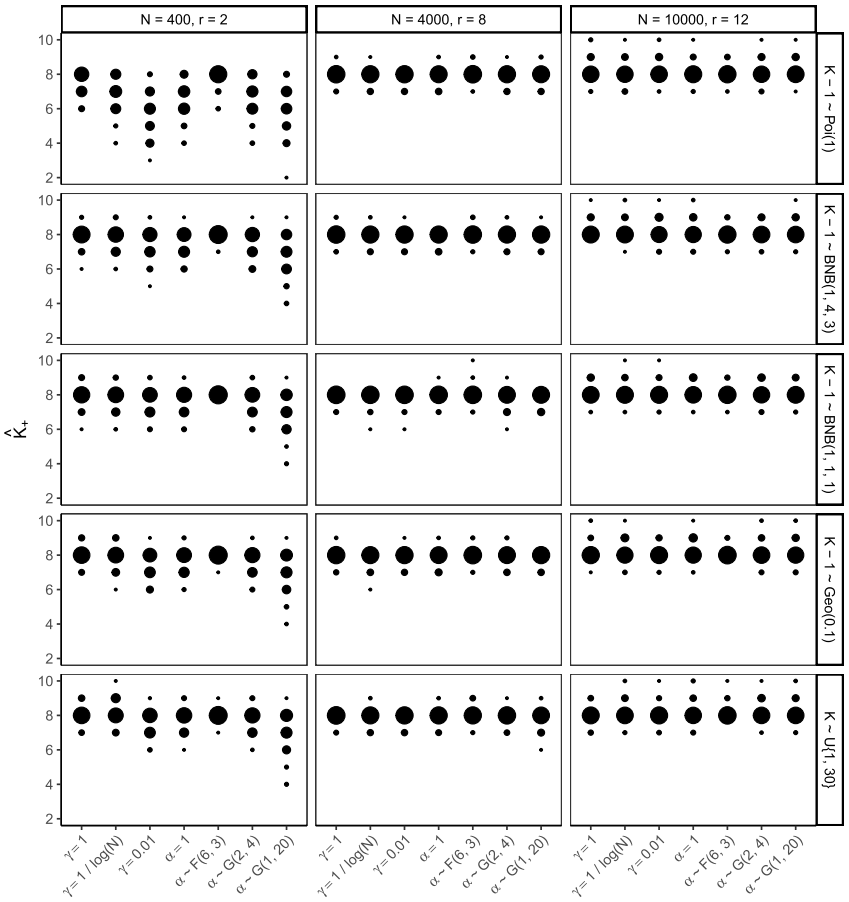

Figure 7: Simulation study. Estimated number of clusters for 100 artificial data sets drawn from mixtures of multivariate Gaussian distributions with eight components. Results based on TS sampling for varying sample size N and dimension r (columns), priors on K (rows) and Dirichlet parameter values γ_K (x-axis). The size of a bubble point shows the percentage of artificial data sets with a specific number of clusters estimated.

Dirichlet prior dynamically and dependent on K. We provide theoretical results that characterize how this specification of a dynamic symmetric Dirichlet prior on the component weights influences the induced prior on the number of clusters and the partition structure. While a static MFM with fixed γ corresponds to a Bayesian non-parametric mixture within the class of Gibbs-type priors, our dynamic version where γ_K depends on K leads to a more flexible mixture outside the class of Gibbs-type priors.

Regarding posterior inference, we introduce the novel telescoping (TS) sampler which is a trans-dimensional Gibbs sampler that simultaneously infers the posterior on the number of components K and the number of clusters K_+. As illustrated, for instance,

for multivariate Gaussian mixtures, the TS sampler can be easily implemented for any kind of component model or distribution. Based on the TS sampler, in future work many different kinds of mixture models can be easily fitted to cluster different types of data which require the use of specific component distributions and models. Future work should also investigate the potential to improve the computational efficiency of the TS sampler, e.g., by reducing the computational burden due to the empty components.

Supplementary Material

Supplementary material for: "Generalized mixtures of finite mixtures and telescoping sampling" (DOI: 10.1214/21-BA1294SUPP; .pdf).

References

Aitkin, M. (2001). "Likelihood and Bayesian Analysis of Mixtures." *Statistical Modelling*, 1: 287–304. 1299

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2: 1152–1174. MR0365969. 1281, 1286, 1292

Argiento, R. and De Iorio, M. (2019). "Is Infinity That Far? A Bayesian Nonparametric Perspective of Finite Mixture Models." *arXiv preprint* arXiv:1904.09733. 1280, 1290

Cerquetti, A. (2010). "A New Parametrization of the Gnedin-Fisher Species Sampling Model." *arXiv preprint* arXiv:1008.2285. 1281, 1286

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. 1280, 1284, 1289, 1290

De Blasi, P., Lijoi, A., and Prünster, I. (2013). "An Asymptotic Analysis of a Class of Discrete Nonparametric Priors." *Statistica Sinica*, 23: 1299–1321. MR3114715. 1284, 1290

Dellaportas, P. and Papageorgiou, I. (2006). "Multivariate Mixtures of Normals with Unknown Number of Components." *Statistics and Computing*, 16: 57–68. MR2224189. doi: https://doi.org/10.1007/s11222-006-5338-6. 1281

Dorazio, R. M. (2009). "On Selecting a Prior for the Precision Parameter of Dirichlet Process Mixture Models." *Journal of Statistical Planning and Inference*, 139: 3384–3390. MR2538090. doi: https://doi.org/10.1016/j.jspi.2009.03.009. 1292

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 1292

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer. MR2265601. 1300

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). "From Here to Infinity: Sparse Finite Versus Dirichlet Process Mixtures in Model-based Clustering." *Advances in Data Analysis and Classification*, 13: 33–64. MR3935190. doi: https://doi.org/10.1007/s11634-018-0329-y. 1280, 1290, 1296, 1301

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Supplementary Material for: "Generalized Mixtures of Finite Mixtures and Telescoping Sampling"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1294SUPP. 1284

Geng, J., Bhattacharya, A., and Pati, D. (2019). "Probabilistic Community Detection With Unknown Number of Communities." *Journal of the American Statistical Association*, 114: 893–905. MR3963189. doi: https://doi.org/10.1080/01621459.2018.1458618. 1279

Gnedin, A. (2010). "A Species Sampling Model with Finitely Many Types." *Electronic Communications in Probability*, 15: 79–88. MR2606505. doi: https://doi.org/10.1214/ECP.v15-1532. 1284, 1290

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs Partitions and Stirling Triangles." *Journal of Mathematical Scienes*, 138: 5674–5684. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 1280, 1281, 1284, 1286, 1288, 1290

Grazian, C., Villa, C., and Lisero, B. (2020). "On a Loss-based Prior for the Number of Components in Mixture Models." *Statistics & Probability Letters*, 158: 108656. MR4025678. doi: https://doi.org/10.1016/j.spl.2019.108656. 1281, 1286

Green, P. J. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28: 355–375. MR1842255. doi: https://doi.org/10.1111/1467-9469.00242. 1291

Greve, J. (2021). *fipp: Induced Priors in Bayesian Mixture Models*. R package version 1.0.0 (https://CRAN.R-project.org/package=fipp). 1287

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 preprint* arXiv:2012.12337. 1287, 1292

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*, doi: https://doi.org/10.1007/s11634-021-00461-8. 1299

Ishwaran, H. and Zarepour, M. (2000). "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-parameter Process Hierarchical Models." *Biometrika*, 87: 371–390. MR1782485. doi: https://doi.org/10.1093/biomet/87.2.371. 1289

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. MR2044876. doi: https://doi.org/10.1198/1061860043001. 1281, 1293, 1296

Jain, S. and Neal, R. M. (2007). "Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model." *Bayesian Analysis*, 3: 445–500. 1281, 1293, 1296

Jara, A., García-Zattera, M. J., and Lesaffre, E. (2007). "A Dirichlet Process Mixture Model for the Analysis of Correlated Binary Responses." *Computational Statistics & Data Analysis*, 51: 5402–5415. MR2370880. doi: https://doi.org/10.1016/j.csda.2006.09.010. 1292

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). "Slice Sampling Mixture Models." *Statistics and Computing*, 21: 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 1296

Lau, J. W. and Green, P. (2007). "Bayesian Model-based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16: 526–558. MR2351079. doi: https://doi.org/10.1198/106186007X238855. 1283

Lijoi, A. and Prünster, I. (2010). "Models Beyond the Dirichlet Process." In Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G. (eds.), *Bayesian Nonparametrics*, 80–136. Cambridge: Cambridge University Press. MR2730661. 1284

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. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 1280, 1300, 1301

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. MR3640186. doi: https://doi.org/10.1080/10618600.2016.1200472. 1280, 1300

McCullagh, P. and Yang, J. (2008). "How Many Clusters?" *Bayesian Analysis*, 3: 101–120. MR2383253. doi: https://doi.org/10.1214/08-BA304. 1279, 1280, 1282

Miller, J. W. and Harrison, M. T. (2013). "A Simple Example of Dirichlet Process Mixture Inconsistency for the Number of Components." In *Advances in Neural Information Processing Systems*, 199–206. 1280

Miller, J. W. and Harrison, M. T. (2014). "Inconsistency of the Pitman-Yor Process for the Number of Components." *Journal of Machine Learning Research*, 15: 3333–3370. MR3277163. 1290

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. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 1279, 1280, 1281, 1282, 1284, 1288, 1292, 1293, 1295, 1296, 1297

Murugiah, S. and Sweeting, T. (2012). "Selecting the Precision Parameter Prior in Dirichlet Process Mixture Models." *Journal of Statistical Planning and Inference*, 142: 1947–1959. MR2903404. doi: https://doi.org/10.1016/j.jspi.2012.02.013. 1292

Nobile, A. (2004). "On the Posterior Distribution of the Number of Components in a

Finite Mixture." *The Annals of Statistics*, 32: 2044–2073. MR2102502. doi: https://doi.org/10.1214/009053604000000788. 1279, 1280, 1281, 1285, 1286, 1288

Nobile, A. and Fearnside, A. (2007). "Bayesian Finite Mixtures With an Unknown Number of Components: The Allocation Sampler." *Statistics and Computing*, 17: 147–162. MR2380643. doi: https://doi.org/10.1007/s11222-006-9014-7. 1285, 1296

Pitman, J. (1995). "Exchangeable and Partially Exchangeable Random Partitions." *Probability Theory and Related Fields*, 102: 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 1280, 1283

Pitman, J. (1996). "Some Developments of the Blackwell-MacQueen Urn Scheme." In *Statistics, Probability and Game Theory*, volume 30 of *IMS Lecture Notes – Monograph Series*, 245–267. MR1481784. doi: https://doi.org/10.1214/lnms/1215453576. 1283, 1290

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Springer. 1283, 1288

Pitman, J. and Yor, M. (1997). "The Two-parameter Poisson-Dirichlet Distribution Derived From a Stable Subordinator." *Annals of Probability*, 25: 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 1280, 1289

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. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 1279, 1280, 1281, 1282, 1285, 1288, 1293, 1296, 1297, 1298, 1299

Roeder, K. (1990). "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in Galaxies." *Journal of the American Statistical Association*, 85: 617–624. 1297

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. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 1280

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. 1301

Stern, H., Arcus, D., Kagan, J., Rubin, D. B., and Snidman, N. (1994). "Statistical Choices in Infant Temperament Research." *Behaviormetrika*, 21: 1–17. 1301

Xie, F. and Xu, Y. (2020). "Bayesian Repulsive Gaussian Mixture Model." *Journal of the American Statistical Association*, 115: 187–203. MR4078456. doi: https://doi.org/10.1080/01621459.2018.1537918. 1279

Acknowledgments

The authors would like to thank Raffaele Argiento, Pierpaolo De Blasi, and Annalisa Cerquetti as well as an anonymous reviewer and the associate editor for valuable suggestions and feedback which helped to improve this work.