# Informed Bayesian Finite Mixture Models via Asymmetric Dirichlet Priors

Garritt L. Page[1], Massimo Ventrucci[2], and Maria Franco-Villoria[3]

[1]Department of Statistics, Brigham Young University, Provo, USA
[2]Department of Statistical Sciences, University of Bologna, Bologna, Italy
[3]Department of Economics "Marco Biagi", University of Modena and Reggio Emilia, Italy

**Abstract**

Finite mixture models are flexible methods that are commonly used for model-based clustering. A recent focus in the model-based clustering literature is to highlight the difference between the number of components in a mixture model and the number of clusters. The number of clusters is more relevant from a practical stand point, but to date, the focus of prior distribution formulation has been on the number of components. In light of this, we develop a finite mixture methodology that permits eliciting prior information directly on the number of clusters in an intuitive way. This is done by employing an asymmetric Dirichlet distribution as a prior on the weights of a finite mixture. Further, a penalized complexity motivated prior is employed for the Dirichlet shape parameter. We illustrate the ease to which prior information can be elicited via our construction and the flexibility of the resulting induced prior on the number of clusters. We also demonstrate the utility of our approach using numerical experiments and two real world data sets.

*Keywords:* Bayesian clustering, Penalized Complexity Priors, Functional Data, Number of clusters

## 1 Introduction

Finite mixture models (FMMs) have become a popular tool in, among other things, density estimation and unsupervised learning (i.e., model-based clustering). An underlying assumption of FMMs is that each unit's measured realization comes from one of $K$ subgroups with group membership unknown *a priori*. The realizations from each of the $K$ subgroups are

then modeled with an appropriate density. This produces a procedure that is able to accommodate distributions that cannot be modeled satisfactorily with a parametric model. In its most general form a FMM can be expressed as

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K,\boldsymbol{w}) = \sum_{k=1}^{K} w_k f_k(\boldsymbol{y}_i|\boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{w} = (w_i,\ldots,w_K)$ are component weights such that $\sum_{k=1}^{K} w_k = 1$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K)$ component specific parameters, and $f_k(\cdot)$ a well defined component density. From a Bayesian perspective, the model is finished by assigning prior distributions to $\boldsymbol{w}$, $\boldsymbol{\theta}$ and possibly $K$. A key reason why FMMs like those in (1) have garnered attention is due to their extreme flexibility with regards to the shape of $f$ which seamlessly permits their use in a diverse array of applications. However, there is a cost to this flexibility as the clustering arising from FMMs can be quite delicate to model specifications (e.g., prior distributions for $\boldsymbol{\theta}$, $\boldsymbol{w}$, or $K$). Since prior decisions can have a significant impact on clustering and common noninformative priors are known to perform poorly, it would be very appealing to construct a method that connects scientifically relevant prior information to meaningful model quantities. As a result, users would be able to more easily inform and regulate the FMM.

Decisions about $K$ are particularly impactful to a FMM's model fit. As such, significant attention has been dedicated to studying it. In the Bayesian FMM literature two approaches have emerged. One is to treat $K$ as an unknown, random quantity, to which a prior distribution is assigned. Miller and Harrison (2018) have referred to this approach as a mixture of finite mixture models (MFMM). Until recently, most attempts to employ a MFMM required constructing a customized reversible-jump MCMC algorithm (RJMCMC; Richardson and Green 1997) which called for a high level of expertise. Because of this, the approach in Richardson and Green (1997) was unavailable to many practitioners (for an alternative to RJMCMC see Stephens 2000). Recently, Miller and Harrison (2018) connected MFMMs to

random partition models. This made available the computational techniques developed in the Bayesian nonparametric (BNP) literature making MFMMs more accessible.

The second approach prescribes formulating an overparametrized FMM by setting $K$ to a large value and using a prior on $\boldsymbol{w}$ that "shrinks" some of the component weights to zero. Rousseau and Mengersen (2011) have provided some theoretical justification for this approach which Malsiner-Walli et al. (2016) refer to as a sparse finite mixture model (sFMM). An alternative approach of producing a sFMM is to use a repulsive type prior on the parameter of centrality in $\boldsymbol{\theta}$. See Petralia et al. (2012), Xie and Xu (2020), Quinlan et al. (2021), Beraha et al. (2022), and Sun et al. (2022).

When $K$ is fixed at a large value and empty components are expected, it is straightforward to distinguish between the number of mixture components and the number of (data informed) clusters (which we denote as $K^+$). However, with $K$ unknown, it is less clear and until recently it was generally thought that $K = K^+$. There is now an emerging FMM literature that explicitly addresses the difference between $K$ and $K^+$ (Frühwirth-Schnatter and Malsiner-Walli 2019; Greve et al. 2021; Frühwirth-Schnatter et al. 2021; Quinlan et al. 2021; Argiento and De Iorio 2022; Alamichel et al. 2023). As a consequence, the prior distribution of $K^+$ induced by particular FMM modeling decisions has begun to garner attention.

The `R` package `fipp` (Greve, 2021) computes the implied prior on $K^+$ for three popular mixture models, namely a Dirichlet Process Model (DPM) and two versions of a MFMM that assume a symmetric Dirichlet prior on the weights with concentration parameter being fixed (static MFMM) or scaled by $K$ (dynamic MFMM). The implied prior on $K^+$ can be computed for any user-supplied prior on $K$ and the number of observations $n$. Figure 1 shows the implied prior under the DPM and static MFMM for different values of the concentration Dirichlet parameter and $n = 100$. By trial and error an expert user, having prior information on $K^+$, can tune the prior on $\boldsymbol{w}$ until the shape of the induced prior on $K^+$ resembles his/her prior belief on the number of occupied components. However, the user can only play a passive

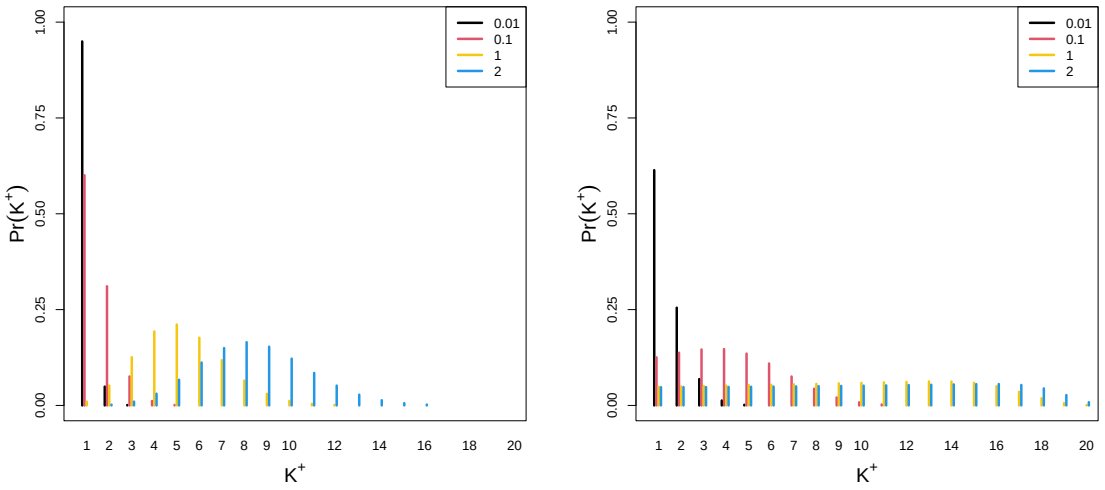


Figure 1: Right panel: Induced prior distribution of $K^+$ for DPM ($K = \infty$) using Dirichlet($\alpha$) prior on the weights. Left panel: induced prior distribution of $K^+$ for MFMM using Uniform$(1, 20)$ prior on $K$ and Dirichlet($\alpha$) prior on the weights. In both panels, different colors are associated to different values of the concentration parameter $\alpha \in \{0.01, 0.1, 1, 2\}$.

role in the sense that he/she cannot elicit the prior for $K^+$ in an intuitive way, for instance by eliciting it's mode or some other centrality parameter or by assigning a large probability mass to a specific range of values of $K^+$. Thus, although formally considering the induced prior on $K^+$ is certainly useful and improves the overall understanding of the FMM prior structure, the methods proposed don't permit users (particularly nonexperts) to "inform" the FMM ($K^+$ in particular) in a straightforward way. Our contribution is to develop an approach that does this. This is done by eliciting prior information through probabilistic statements associated with a user-supplied value of $K^+$.

Prior elicitation for $K^+$ can be challenging as it requires clearly defining what is meant by a "cluster" and to clearly define the motivation behind using a FMM (Hennig 2015). Once this has been established, our approach of eliciting prior information follows the philosophy of the penalized complexity (PC) priors outlined in Simpson et al. (2017). In particular, we first specify a base or reference model and then the role of $\boldsymbol{w}$'s prior distribution is to

"shrink" towards the reference model unless the data indicate otherwise. To do this we use an asymmetric Dirichlet distribution on the component weights. An appealing feature of this approach is that scientific questions are able to guide reference model selection (e.g., a mixture with $K^+$ equal to a user-supplied value). Practitioners can then inform the FMM *a priori* by thinking directly about $K^+$ while tuning a parameter of the asymmetric Dirichlet, that controls *a priori* the FMM's sparseness (or lack thereof).

As with any Bayesian procedure, when the data poorly inform a particular parameter, the prior can be highly influential on the resulting posterior distribution. In this setting additional analysis must be executed to understand the exact impact the prior has on the posterior. This is true for our prior construction for $K^+$. However, our method is well suited to explore the prior's impact on the posterior $K^+$ because of the coherent way in which the prior is informed. As a result, it is very straightforward to carry out a sensitivity analysis and we provide one approach of doing this.

We finish the Introduction by briefly mentioning that the random probability measures that are commonly studied in the BNP literature (Müller et al. 2015; Ghosal and van der Vaart 2017) set $K = \infty$ . This essentially side-steps the need to formally consider $K$. That said, Miller and Harrison (2013) pointed out that estimating $K^+$ using a Dirichlet Process Model (DPM) can be problematic. In fact, Cai et al. (2021) find that estimating $K^+$ consistently depends on specifying components correctly (i.e., correctly defining the meaning of a cluster). As a result, Lijoi et al. (2023) studied in more depth the finite-dimensional Bayesian clustering from a normalized random measure with independent increments (Regazzini et al. 2003) and Ascolani et al. (2023) explored conditions necessary for a DPM to consistently estimate the $K^+$. Recently, Argiento and De Iorio (2022) and Frühwirth-Schnatter et al. (2021) made very interesting connections between BNP mixtures and FMMs while Alamichel et al. (2023) studied the consistency in estimating $K^+$ (or lack there of) in a variety of mixture models.

The rest of the article is organized as follows. In Section 2 we provide the necessary background for FMMs. Then in Section 3 we introduce the prior construction that permits informing FMMs and provide some theoretical justifications. Section 4 details a simulation study that compares our approach to a few other FMM procedures. Section 5 describes two applications. The first is the well known galaxy dataset and the second a biomechanics functional data example. We end by providing some final comments in Section 6. All proofs and computational details are relegated to the online supplementary material along with additional details associated with the simulation study and applications detailed in Sections 4 and 5.

# 2 Background on Bayesian Finite Mixture Models

For computational purposes, the FMM in (1) is often re-expressed using latent component labels. Doing so permits describing the model hierarchically. To this end, let $z_1, \ldots, z_n$ denote $n$ component labels where $z_i = j$ implies that the $i$th unit belongs to the $j$th component. Introducing component labels in the FMM and assuming each component density belongs to the same family permits expressing the FMM in (1) as

$$\begin{aligned} y_i|z_i &\overset{ind}{\sim} f(\boldsymbol{\theta}_{z_i}), \quad i = 1, \ldots, n, \\ Pr(z_i = k|\boldsymbol{w}) &= w_k, \quad i = 1, \ldots, n. \end{aligned} \tag{2}$$

The Bayesian model is completed by assuming

$$\boldsymbol{\theta}_k \overset{iid}{\sim} \pi_\theta, \quad k = 1, \ldots, K, \tag{3}$$

$$\boldsymbol{w} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{4}$$

where $\pi_\theta$ denotes a prior distribution for $\boldsymbol{\theta}_k$ and Dirichlet($\boldsymbol{\alpha}$) denotes a Dirichlet distribution with parameter $\boldsymbol{\alpha}$. As mentioned, it is possible to assign a prior distribution to $K$ also. What we develop can be applied in that setting, but we focus on the case when $K$ is fixed to a large value (Rousseau and Mengersen 2011).

Even though $f(\cdot)$ and $\pi_\theta$ implicitly determine the type of clusters that are permitted in the FMM (e.g., spherical), their selection does not influence the implied prior on $K^+$. Althernatively, the prior on $\boldsymbol{w}$ (and/or $\boldsymbol{\alpha}$) is directly connected to the implied prior on $K^+$. However, both $f(\cdot)$ and $\pi_\theta$ and the prior on $\boldsymbol{w}$ impact the posterior distribution of $K^+$. Thus any notion of posterior consistency associated with $K^+$ must necessarily consider both the type of clusters permitted based on $f(\cdot)$ and $\pi_\theta$ *and* the *a priori* number of clusters based on the prior for $\boldsymbol{w}$. In this paper our focus is on informing the mixture through the implied prior on $K^+$ and hence focus on $\boldsymbol{w}$'s prior and assume that $f(\cdot)$ and $\pi_\theta$ are well specified.

It is common to use a symmetric Dirichlet distribution as a prior for $\boldsymbol{w}$ so that $\boldsymbol{\alpha} = \alpha\boldsymbol{j}$ where $\boldsymbol{j}$ is a $K$-dimensional vector of 1s and $\alpha > 0$. It is also quite common to fix $\alpha = 1/K$ since the resulting FMM would then approximate a Dirichlet process mixture (DPM) as $K \to \infty$ (Ishwaran and James 2001). More recently, $\alpha$ has been assigned a prior. In particular, Malsiner-Walli et al. (2016); Frühwirth-Schnatter and Malsiner-Walli (2019); Greve et al. (2021); Frühwirth-Schnatter et al. (2021) all assume $\alpha \sim Gamma(a, aK)$ so that $E(\alpha) = 1/K$. The parameter $\alpha$ regulates the sparseness of the FMM in that as $\alpha \to 0$, $K^+$ decreases.

Although a symmetric Dirichlet prior for $\boldsymbol{w}$ is quite common, it is challenging to inform the induced prior on $K^+$ so that user specified values for $K^+$ are given large prior mass. This results from the fact that the FMM is not explicitly parameterized in terms of $K^+$ and that the induced prior of $K^+$ is a function of $n$ and $K$ in addition to $\alpha$. Next, we describe a prior construction that permits introducing expert opinion with regards to $K^+$ through an asymmetric Dirichlet prior on $\boldsymbol{w}$ which we will refer to as an asymmetric FMM (aFMM).

# 3 Asymmetric Dirichlet Finite Mixture Models

We desire to develop a method in which it is straightforward to guide the implied prior on $K^+$. We do this by considering an asymmetric Dirichlet, defined below, as a prior for $\boldsymbol{w}$

**Definition 1.** *The asymmetric Dirichlet, denoted by Dirichlet($\boldsymbol{\alpha}_{1,2}$), is a Dirichlet distribution with parameters $U, \alpha_1, \alpha_2$ such that $\boldsymbol{\alpha}_{1,2} = (\alpha_1 \boldsymbol{j}_U, \alpha_2 \boldsymbol{j}_{K-U})$ where $\boldsymbol{j}_U$ and $\boldsymbol{j}_{K-U}$ are $U$ and $K-U$ dimensional vectors filled with ones.*

In Definition 1, $U$ plays a crucial role as a user-supplied value on which the induce prior of $K^+$ is "centered". An appealing property of our prior construction is that as $\alpha_1 \to \infty$ and $\alpha_2 \to 0$ prior mass concentrates on $U$ resulting in a mixture model with exactly $U$ occupied components. We show this in Proposition 1, but first build some intuition why this is the case. Note that under $\boldsymbol{w} \sim Dirichlet(\boldsymbol{\alpha}_{1,2})$

$$E(w_k) = \begin{cases} \dfrac{\alpha_1}{\alpha_1 U + \alpha_2(K-U)} & k = 1, \ldots, U \\ \dfrac{\alpha_2}{\alpha_1 U + \alpha_2(K-U)} & k = U+1, \ldots, K. \end{cases} \tag{5}$$

If $\alpha_1 \gg \alpha_2$ and $\alpha_2 \to 0$, then $E(w_k) \to 1/U$ for $k = 1, \ldots, U$ and $E(w_k) \to 0$, for $k = U+1, \ldots, K$. Thus, all prior mass is uniformly distributed over the first $U$ components, with no mass assigned to the remaining $K-U$ components. As a result, the implied prior on $K^+$ becomes a point mass at $U$ as $\alpha_1 \to \infty$ and $\alpha_2 \to 0$. We show this more carefully in the following proposition the proof of which can be found in the supplementary material.

**Proposition 1.** *Assume that $\boldsymbol{w} \sim Dirichlet(\boldsymbol{\alpha}_{1,2})$. Then as $n \to \infty$*

$$\lim_{\alpha_1 \to \infty} \lim_{\alpha_2 \to 0} Pr(K^+ = U \mid K, n, \alpha_1, \alpha_2) = 1 \tag{6}$$

In order to visualize the implied prior of $K^+$ from the aFMM, we provide Figure 2

where different values for $\alpha_1$ and $\alpha_2$ are considered and $U = 10$ which means the prior should be centered around 10 non-empty components. The plots in Figure 2 are a graphical representation of the property of the aFMM expressed in Proposition 1. Note that increasing $\alpha_1$ and decreasing $\alpha_2$ leads to an implied prior for $K^+$ highly concentrated on $U = 10$; it is sufficient fixing $\alpha_1 = U$ and $\alpha_2 = 0.001$ to get a spike on $K^+ = 10$ in this case. By moving $\alpha_1$ and $\alpha_2$ the probability mass can be moved to the left or right tails. Doing so, the probability mass can be distributed either below or above $U$, in such a way that $U$ need not be connected to the center of the distribution. In particular, by decreasing $\alpha_1$ while keeping $\alpha_2$ small (e.g. smaller than 0.001), we get more probability in the left tail hence $U = 10$ can be intended as a *soft* upper bound. Analogously, by increasing $\alpha_2$ while $\alpha_1$ being large, we obtain a prior with heavy right tail hence $U = 10$ can be intended as a *soft* lower bound.

The aFMM is constructed in such a way that the implications of the theory in Rousseau and Mengersen (2011) hold. We state this carefully in the following remark.

**Remark 1.** *The aFMM as described in* (2) - (3) *with* $\boldsymbol{w} \sim$ *Dirichlet*$(\boldsymbol{\alpha_{1,2}})$ *satisfies the assumptions in Rousseau and Mengersen (2011) so that if* $\min(\alpha_1, \alpha_2) < d/2$ *the asymmptotic vanishing weights property holds.*

The proof of Remark 1 follows from the arguments laid out in Alamichel et al. (2023). Note that Remark 1 holds only if $\alpha_1$ and $\alpha_2$ are considered fixed quantities.

## 3.1 Prior Distributions for $\alpha_1$ and $\alpha_2$

There are a number of ways to treat ($\alpha_1$, $\alpha_2$). The list includes A) Fix $\alpha_1$ and $\alpha_2$ to prespecified values, which corresponds to an asymmetric case of the static model described in Frühwirth-Schnatter et al. (2021); B) assume that $\alpha_1$ is unknown with an assigned prior distribution and fix $\alpha_2$ to a small value; C) Fix $\alpha_1$ to a prespecified value and assume $\alpha_2$ is unknown with an assigned prior distribution; and D) assume both $\alpha_1$ and $\alpha_2$ are unknown

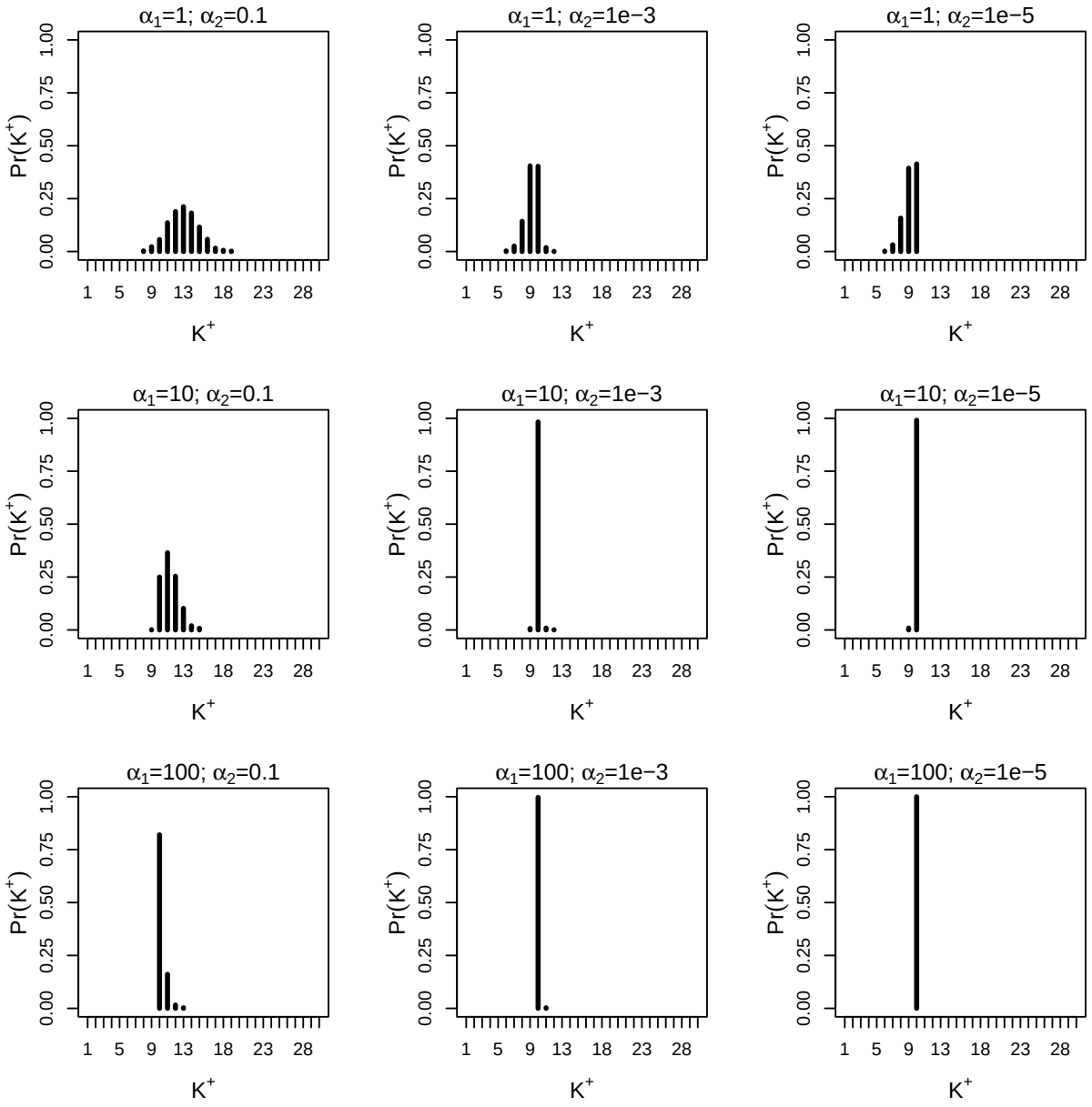


Figure 2: Induced prior distribution of $K^+$ obtained via simulations, assuming the asymmetric Dirichlet prior with $\alpha_1 = \{1, 10, 100\}$ and $\alpha_2 = \{0.1, 0.001, 0.00001\}$ and $U = 10$, $n = 100, K = 30$.

and assign both a prior distribution. Each possibility results in a specific balance between computation cost and model flexibility. In what follows, we focus primarily on the case that $\alpha_1$ is assigned a prior and $\alpha_2$ is fixed at some prespecified value. This permits us to treat $U$ as an soft upper bound for $K^+$ with $\alpha_2$ controlling the rigidity of the upper bound in the sense that as $\alpha_2 \to 0$, then $U$ becomes a hard upper bound. There are a number of prior distributions that might be considered for $\alpha_1$. For reasons we provide shortly, we focus primarily on a prior that has connections to PC priors (Simpson et al. 2017).

## 3.2 Penalized Complexity Motivated Prior

Our prior construction is guided by a key idea on which PC priors are based. Mainly, the prior is treated as a mechanism that regulates the behaviour of the model with respect to a parsimonious version of it called the *base model*. A PC prior guarantees that the base model is favored unless data support an alternative model. Typically the alternative model is assumed to be more flexible (or complex) than the base one, or an over-parameterized version of the base one. The PC prior is formally defined as an exponential distribution on a measurement scale quantifying the increased complexity of the alternative (i.e. flexible) model with respect to the base one, where *complexity* is measured by the Kullback–Leibler divergence (KLD, Kullback and Leibler (1951)). A brief review of the principles and the practical steps underlying the construction of PC priors, as originally proposed in Simpson et al. (2017), is in supplementary material.

Our aim is to construct a prior for the asymmetric Dirichlet parameters $(\alpha_1, \alpha_2)$ such that the induced prior on $K^+$ guarantees that a mixture model with a user-defined number of non-empty components, say $U \in [1, K]$, is favoured unless data support an alternative FMM. Thus, the implied prior for $K^+$ is used as a mechanism to regulate the behaviour of the FMM with respect to a *base finite mixture model* (base FMM). In general, the base FMM is a FMM with $K^+ = U$, for some $U$ selected by users according to the goal of the

analysis and/or their prior knowledge about $K^+$. A base FMM favouring $K^+ = U$ can be obtained by treating the asymmetric Dirichlet from Definition 1, Dirichlet($\boldsymbol{\alpha}_{01,02}$), as a prior on $\boldsymbol{w}$, with parameters $\alpha_{01} = \infty$, $\alpha_{02} = 0$, and $U$. Note, we will use $\boldsymbol{\alpha}_{01,02}$ to refer to the Dirichlet parameters under the base model. Because the asymmetric Dirichlet is not defined for $\alpha_1 = \infty$ and $\alpha_2 = 0$, as a *practical* base FMM we will use a "large" value for $\alpha_{01}$ and a "small" value for $\alpha_{02}$. Numerical experiments lead us to set $\alpha_{01} = U$ and $\alpha_{02} = 10^{-5}$. Thus, Dirichlet($\boldsymbol{\alpha}_{01,02}$), with $\alpha_{01} = U, \alpha_{02} = 10^{-5}$ and for a specific $U$ is our "practical base model" in general situations where we want a mixture model favouring $U$ clusters (this choice worked well for a variety of values for $n$ and $K$).

Constructing the PC prior requires quantifying how much a FMM with parameters $(\alpha_1, \alpha_2)$ *deviates* from the particular base FMM. Let $\boldsymbol{g} \sim$ Dirichlet($\boldsymbol{\alpha}_{1,2}$) be the asymmetric Dirichlet under the base FMM with parameters $\alpha_1 = \alpha_{01}$, $\alpha_2 = \alpha_{02}$ and $U$, while $\boldsymbol{p} \sim$ Dirichlet($\boldsymbol{\alpha}_{1,2}$) be the asymmetric Dirichlet under the alternative FMM with parameters $\alpha_1 > 0$, $\alpha_2 > 0$ and $U$. (Note that $\boldsymbol{g}$ and $\boldsymbol{p}$ have different values of the parameters $\alpha_1$ and $\alpha_2$ but the same $U$). The *deviation* from the alternative FMM to the base FMM is measured using the KLD between $\boldsymbol{p}$ and $\boldsymbol{g}$

$$
\begin{aligned}
KLD(\boldsymbol{p}||\boldsymbol{g}) = \log\Gamma(\alpha_1 U + \alpha_2(K-U)) - \log\Gamma(\alpha_{01}U + \alpha_{02}(K-U)) - \\
(U\log\Gamma(\alpha_1) + (K-U)\log\Gamma(\alpha_2)) + U\log\Gamma(\alpha_{01}) + (K-U)\log\Gamma(\alpha_{02}) + \\
U(\alpha_1 - \alpha_{01})[\psi(\alpha_1) - \psi(\alpha_1 U + \alpha_2(K-U))] + \\
(K-U)(\alpha_2 - \alpha_{02})[\psi(\alpha_2) - \psi(\alpha_1 U + \alpha_2(K-U))], \qquad (7)
\end{aligned}
$$

where $\Gamma$ and $\psi$ are, respectively, the gamma and digamma functions.

The function in Eq. (7) depends on the asymmetric Dirichlet parameters $(\alpha_{01}, \alpha_{02})$ under the practical base model, the asymmetric Dirichlet parameters $(\alpha_1, \alpha_2)$ under the alternative model, and the user-supplied value for $U$. Function (7) represents a suitable scale to measure

deviations from the base FMM.

For ease of interpretation (7) is transformed to a unidirectional distance measure $d(\alpha_1, \alpha_2) = d(p||g) = \sqrt{2\text{KLD}(\boldsymbol{p}||\boldsymbol{g})}$. (Note our notation for $d$ focuses on a two-dimensional function of the Dirichlet parameters $(\alpha_1, \alpha_2)$ because these are the parameters we need to assign a prior to, but $d$ also depends on the user-supplied $U$ and the choice of the practical base model $(\alpha_{01}, \alpha_{02})$). When $d = 0$, the FMM corresponds to the base FMM, i.e., a FMM favouring $K^+ = U$ non-empty components. As $d$ increases the FMM is allowed to deviate from the base FMM, with deviations occurring either as a FMM favouring $K^+ < U$ (i.e. sparser mixture) or $K^+ > U$ (i.e. less sparse mixture).

### 3.2.1 PC prior for $\alpha_1$ conditional on $\alpha_2$ being fixed

Following Simpson et al. (2017) we consider an exponential distribution on $d(\alpha_1, \alpha_2)$, with rate $\lambda > 0$, so that the mode is always at the base model $d = 0$, or $K^+ = U$, and the penalization rate is constant. However, in our case the distance $d(\alpha_1, \alpha_2)$ is a surface that varies over $\alpha_1$ and $\alpha_2$ and potentially one may consider two parameters $\lambda_1$ and $\lambda_2$ to penalize deviations along $\alpha_1, \alpha_2$ at different rates.

From Figure S1 of the supplementary material we can visually inspect the KLD in (7) and we see that it varies more sharply along $\alpha_1$ than $\alpha_2$. An exponential prior on $d(\alpha_1, \alpha_2)$ with distinct decay rates $\lambda_1$ and $\lambda_2$ will permit the user to get an induced prior on $K^+$ with mode at $U$ and, at the same time, the ability to tune the probability mass assigned to $K^+ < U$ and $K^+ > U$ independently, by careful selection of $\lambda_1$ and $\lambda_2$. This strategy is useful when users have precise information about $K^+$ and wish to center the prior for $K^+$ on $U$. From a computational point of view this strategy is quite expensive as numerically deriving the prior implies optimizing over two decay rate parameters which will slow the MCMC considerably.

We seek a simpler solution here in the form of a conditional PC prior on $\alpha_1 \in (0, U]$,

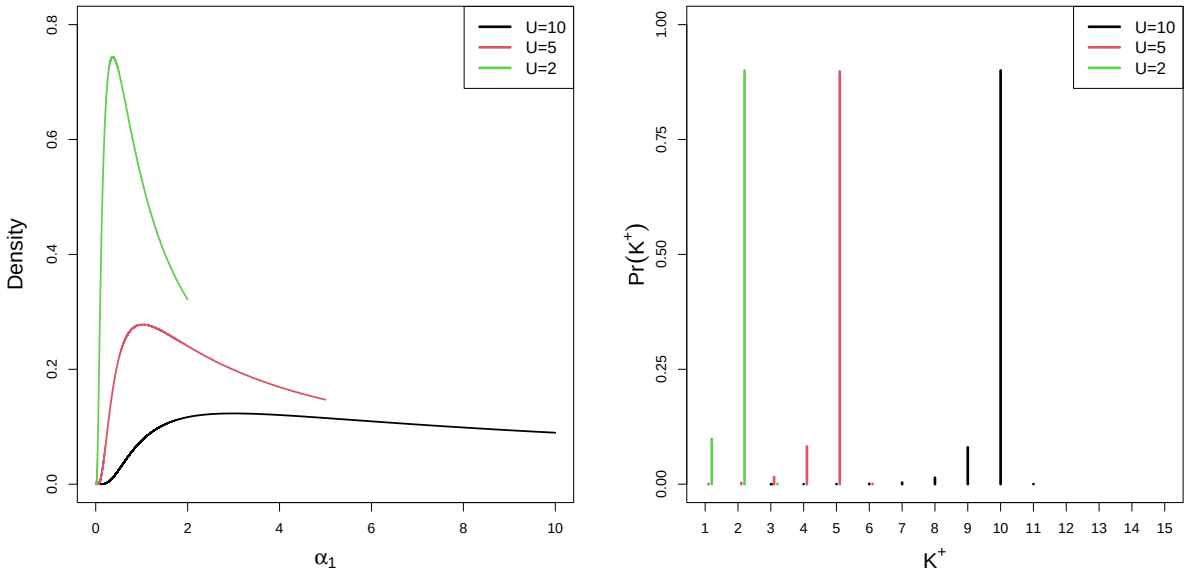


Figure 3: PC prior on $\alpha_1 \in (0, U)$ given $\alpha_2 = 1e-5$ (left) and the implied prior on $K^+$ (right), for three different choices of $U$ and tail probability $tp = 0.1$.

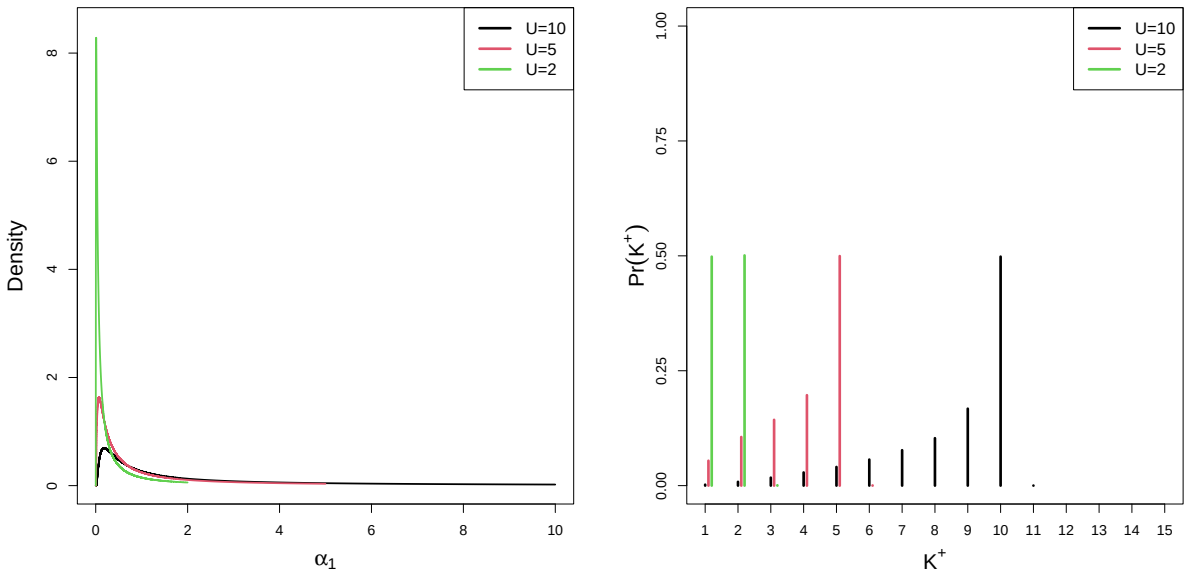


Figure 4: PC prior on $\alpha_1 \in (0, U)$ given $\alpha_2 = 1e-5$ (left) and the implied prior on $K^+$ (right), for three different choices of $U$ and tail probability $tp = 0.5$.

given $\alpha_2$ set to a small value. Doing so, the user-supplied $U$ can be intended to be an upper bound for $K^+$ which we believe works well in many applications. From our experiments $\alpha_2 = 10^{-5}$ is small enough to have $Pr(K^+ > U)$ approximately zero, with $U$ playing the role of a *lenient* upper bound; however, by increasing $\alpha_2$ the right tail probability will increase too, making $U$ a softer upper bound.

The PC prior on $\alpha_1$ conditional on $\alpha_2 = 10^{-5}$ is the (truncated) exponential prior on the distance $d(\alpha_1, \alpha_2 = 10^{-5})$, and follows by a change of variable transformation:

$$\pi(\alpha_1) = \frac{\lambda \exp\left(-\lambda d(\alpha_1, \alpha_2 = 10^{-5})\right) |d'(\alpha_1, \alpha_2 = 10^{-5})|}{1 - \exp(\lambda d(\alpha_1 = U, \alpha_2 = 10^{-5}))}, \quad \lambda > 0, \quad 0 < \alpha_1 \leq U \tag{8}$$

Details on the numerical derivation of (8) can be found in supplementary material S3. One appealing feature of (8) is that the user is only required to handle a single decay rate parameter $\lambda$, hence the scaling of the PC prior according to the user prior guess about $K^+$ greatly simplifies. To "scale the PC prior" for us means to choose $\lambda$ in Eq. (8).

Scaling the PC prior can be approached from the following situations: either the user might have information on the maximum number of clusters possibly present in the data at hand, or on the number of clusters that he/she is able to interpret. We propose computing $\lambda$ based on a user-defined probabilistic statement like

$$Pr(K^+ < U) = tp \tag{9}$$

In other words, our aim is to help the user select the $\lambda$ that corresponds to assigning a certain probability, denoted as $tp$, to the left-tail $(1, U-1)$. We use simulations to find the optimal $\lambda$ that realizes a left-tail probability equal to the user-defined $tp$; the procedure is described in supplementary material. Figures 3 and 4 display the implied prior on $K^+$ obtained by setting the left-tail probability equal to 0.1 and 0.5, respectively, and different values of the lenient upper bound $U$.

A general appealing property of the PC prior is that it accommodates the user-selected value of $K$ and the number of observations $n$ in the case study at hand in a natural way. There are two reason why this is the case. The prior in Eq. (8) derives from assuming an exponential prior on the KLD (which depends on $K$), hence it "adapts" automatically to any value of $K$ the user may choose. In addition, the simulation-based algorithm to numerically derive the prior for $K^+$ requires $n$ as an input, other than $K$, thus the number of observations in the application at hand would be automatically taken into account in the (induced) prior for $K^+$.

## 3.3 Special Cases

The aFMM has as special cases other commonly used over-parameterized FMMs. For example, if we set $U = 1$, then the asymmetric Dirichlet prior can induce sparsity in the sense that $K^+$ is much smaller than $K$. This aFMM would have shrinkage properties similar to that of sFMM as described in the following remark

**Remark 2.** *If $U = 1$ and $\alpha_2$ is fixed at a small value then as $\alpha_1 \to 0$ the $Pr(K^+ = 1 \mid K, n, \alpha_1, \alpha_2) = 1$ resulting in a sFMM.*

The proof of remark 2 follows from arguments similar to those found in the proof of Proposition 1.

Shrinkage properties similar to sFMM can also be achieved through $tp$. When $tp$ is set to a large value (i.e., close to one) then the induced prior on $K^+$ will be such that the majority of prior mass is concentrated on values (much) smaller than $U$ (when $\alpha_2$ is small). Finally, setting $U = 0$ recovers the symmetric Dirichlet prior for $\boldsymbol{w}$ with $\alpha_2$ acting as the lone concentration parameter. As a result, all FMM methods that have been developed using a symmetric Dirichlet prior can be employed.

# 4 Simulation Study

In order to illustrate the aFMM's performance in estimating $K^+$, we conduct a numerical experiment. Even though an asymmetric prior distribution on $\boldsymbol{w}$ (and as a result an informed prior for $K^+$) can be employed for any FMM, in the simulation and application that follow we focus on the case that $f_k(\cdot)$ is a Gaussian. As a result, (1) becomes

$$y_i \sim \sum_{k=1}^{K} w_k \mathrm{N}(\mu_k, \sigma_k^2) \tag{10}$$

so that $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)$. After introducing the component labels, the augmented data model becomes

$$\begin{aligned} y_i \mid z_i &\sim \mathrm{N}(\mu_{z_i}, \sigma_{z_i}^2) \\ Pr(z_i = k \mid \boldsymbol{w}) &= w_k, \end{aligned} \tag{11}$$

and we use the following prior distributions

$$\begin{aligned} \mu_k &\sim \mathrm{N}(\mu_0, \sigma_0^2) \\ \sigma_k^2 &\sim \text{Inverse-Gamma}(a_0, b_0) \\ \boldsymbol{w} &\sim \text{Dirichlet}(\boldsymbol{\alpha}_{1,2}) \\ \alpha_1 \mid \alpha_2, U &\sim PC(U, tp). \end{aligned} \tag{12}$$

Here "Inverse-Gamma" denotes an inverse Gamma distribution parameterized so that the prior mean of $\sigma_k^2$ is $b_0/(a_0 - 1)$. For hyper-prior values we set $\mu_0 = \text{mean}(\boldsymbol{y})$, $\sigma_0^2 = 10^2$ which correspond to one of the prior specifications that was employed in Grün et al. (2021). We also set $a_0 = 3$ and $b_0 = 2$ which is also very similar to one of the prior specifications in Grün et al. (2021). We set $K = 25$ in all our implementations of the aFMM. All computation is carried out using the `informed_mixture` function that can be found in the `miscPack` R-package that is available at `https://github.com/gpage2990`. Data sets are generate using

(10) as a data generating mechanism in the following two ways:

**Data Type 1**: Set $K = K^+$ for $K^+ \in \{2, 5, 10\}$ and then generate $n \in \{100, 1000\}$ observations by setting $\boldsymbol{w} = 1/K\boldsymbol{j}$, $\sigma_k = 0.5$, and $\mu_k = 3(k-1)$. In this scenario there are always exactly $K^+ \in \{2, 5, 10\}$ clusters with centers displaying little overlap. As $n$ increases from 100 to 1000 the number of observations in each component increases but is still quite uniform across the $K^+$ clusters.

**Data Type 2**: Use (11) - (12) to generate $n \in \{100, 1000\}$ observations by setting $K = 25$, $\alpha_1 = U$, $\alpha_2 = 10^{-3}$, $A = 1$, $\mu_0 = 0$, $\sigma_0^2 = 3$, and $U \in \{2, 5, 10\}$. In this scenario clusters may not be well separated and the number of observations in each of the $K^+$ clusters can vary greatly. As a result, this data generating scenario can be much more challenging than the first in estimating $K^+$.

Examples data sets created using the procedure just described for *Data Type 1* and *Data Type 2* are provided in Figures S2 and S3 of the supplementary material. For each data type 100 datasets are generated and to each we fit an aFMM for $U \in \{2, 5, 10\}$ under the following prior specifications

**Gam**: Fix $\alpha_2 = 10^{-5}$ and use $\alpha_1 \sim Gamma(a, b)$ where $a = 10$ and $b = (10U)^{-1}$,

**PC(0.1)**: Fix $\alpha_2 = 10^{-5}$ and use $\alpha_1 \mid \alpha_2 \sim PC(U, tp = 0.1)$, which means that the user prior statement is $Pr(K^+ < U) = 0.1$.

**PC(0.9)**: Fix $\alpha_2 = 10^{-5}$ and use $\alpha_1 \mid \alpha_2 \sim PC(U, tp = 0.9)$, which means that the user prior statement is $Pr(K^+ < U) = 0.9$.

Each of these prior specifications are found on the $x$-axis in Figures 5 and 6. In addition to fitting the aforementioned aFMM models to each dataset, for context, we also fit the following methods:

**sFMM**: sparse FMM as described in Malsiner-Walli et al. (2016) such that $\alpha \sim Gamma(10, 10K)$,

**FMM**: A FFM fit using RJMCMC as employed in the `mixAK` `R`-package (Komárek and Komárková 2014),

**DPM**: A Dirichlet Process Mixture model with dispersion parameter set at 1,

**NormIFPP**: The normalized independent finite point process FMM described in Argiento and De Iorio (2022) and fit using the `AntMAN` `R`-package (Ong et al. 2021).

Hyper-prior values for all methods listed were selected to match as much as possible those used in the aFMM procedures. For the NormIFPP method we employed values that were suggested in Argiento and De Iorio (2022). The simulation was executed using GNU parallel (Tange 2022).

To compare each methods ability to estimate $K^+$ we recorded the bias associated with the posterior mode of $K^+$ and two other metrics that evaluate the accuracy of the entire posterior distribution of $K^+$. The first is a posterior probability weighted sum of squares associated with $K^+$ as defined below

$$\text{pwss}(K^+) \stackrel{\text{def.}}{=} \sum_{k=1}^{K} (k - K^+_{\text{true}})^2 Pr(K^+ = k \mid \boldsymbol{y}), \tag{13}$$

where $K^+_{\text{true}}$ denotes the value of $K^+$ used to generate the data. This metric takes into account both the spread and location of the posterior distribution of $K^+$ relative to $K^+_{\text{true}}$ with smaller values indicating a more precise estimate of $K^+$. The second metric evaluates the accuracy of the co-clustering probabilities for each observation and is defined as follows

$$\text{ccprob\_error} \stackrel{\text{def.}}{=} \sum_{j=1}^{n} \sum_{\ell<j} (I[j \sim \ell] - Pr(z_j = z_\ell \mid \boldsymbol{y}))^2,$$

where $I[j \sim \ell] = 1$ if unit $j$ and $\ell$ belong to the same cluster and zero otherwise. Small

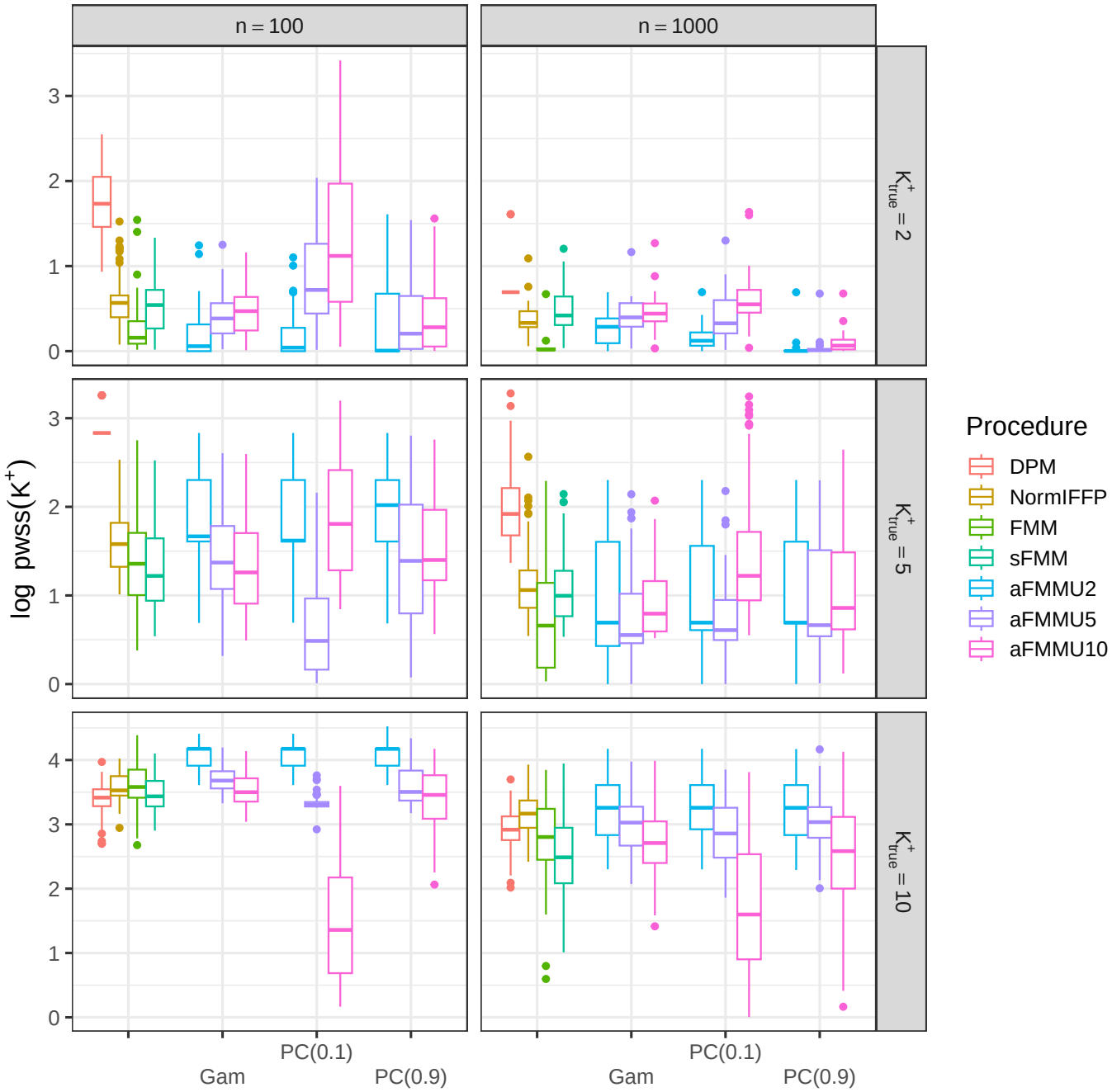


Figure 5: $\log(\text{pwss}(K^+) + 1)$ for *Data Type 2*. Each row corresponds to results associated with the value of $K^+$ used to generate data.

values of `ccprob_error` indicate more accurate estimation of the underlying partition and as a result, the number of clusters. In Figures 5 - 6 we report results for `ccprob_error` and pwss($K^+$) under *Data Type 2*. Results associated with with bias and from *Data Type 1* are provided in Figures S4-S7 of the supplementary material.

From Figure 5 note that when $U = K^+_{\text{true}}$ the aFMM performs the best regardless of prior on $\alpha_1$ except for FMM that in some scenarios (i.e., for some choice of $tp$) performs better than aFMM even when $U = K^+_{\text{true}}$. This is unsurprising. When $U \neq K^+_{\text{true}}$, aFMM continues to perform competitively (at least one aFMM procedure is the best or second best) in all scenarios while the competing methods do well in some scenarios but poorly in others. Note further how the "sparse" aFMM ($PC(0.9)$) tends to perform well when $K^+_{\text{true}} < U$. As expected, setting $U$ to a value that is far from $K^+_{\text{true}}$ tends to result in poor performance for the aFMM. Trends for *Data Type 1* are similar (see Figure S5 of the supplementary material).

Regarding Figure 6, first note that the `FMM` procedure is not included. This is due to the fact that it is not possible to compute the co-clustering probabilities based on output provided by the `mixAK` package. Now, it appears that all methods save the DPM perform similarly when $K^+_{\text{true}} = 2$ regardless of sample size. For $K^+_{\text{true}} = 5$ it appears that the aFMM performs best when $n = 100$ so long as $U > 2$ and for $n = 1000$ the aFMM performs similarly to sFMM while outperforming DPM and NormalIFPP regardless of the prior on $\alpha_1$. However, when $K^+_{\text{true}} = 10$ it appears that aFMM performs best when $U > 2$ and one of the two PC priors are employed. The upshot of the simulation study is that the estimation $K^+$ under the aFMM performs well if $U$ is not far from the truth for small $n$ and performs very competitively when $n$ is large regardless of $U$.

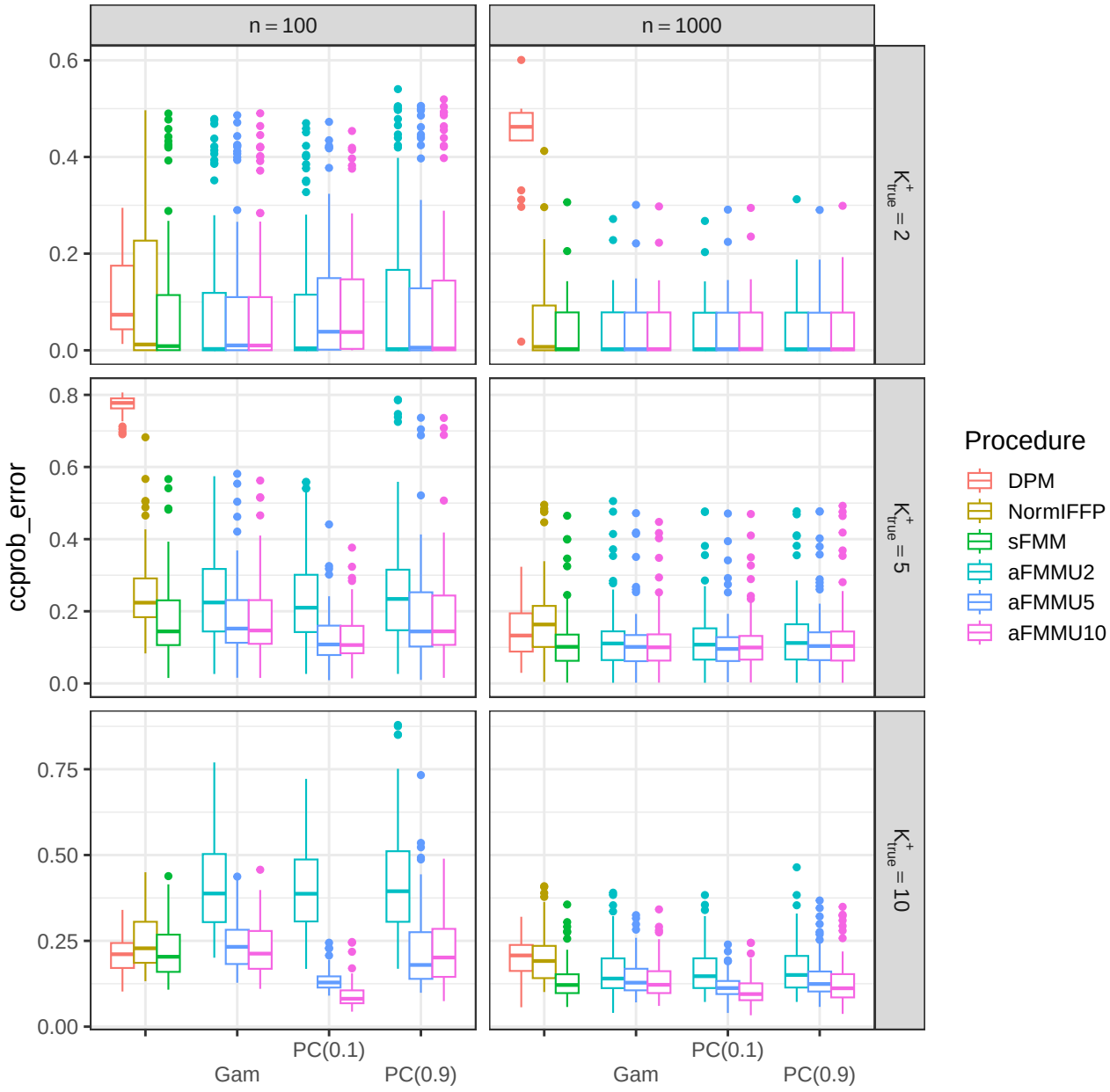


Figure 6: Error of co-clustering probabilities for *Data Type 2*. Each row corresponds to results associated with the value of $K^+$ used to generate data.

# 5 Applications

In this section we further illustrate the utility of the aFMM in two applications. The first is the well known galaxy dataset while the second is a biomechanic application where prior information is elicited from exercise scientists. We consider the galaxy data to illustrate the use of our prior construction as a principled tool to evaluate the sensitivity of the clustering configuration with regards to the induced prior on $K^+$ . The biomechanic data permits illustrating how our prior construction can be intuitively employed to accommodate prior beliefs elicited from experts that approach an analysis from different perspectives. The biomechanic data will be modelled from a functional data perspective. This requires a model that is more complex than that described in (10) - (12) which we detail in 5.2.1.

## 5.1 Galaxy Data

The well known galaxy dataset (available in the `MASS` library) contains the velocities (km/sec) of 82 galaxies. This dataset has been widely used to illustrate methods in the clustering literature (Grün et al. 2021). Aitkin (2001) argues that there are 3 clusters if equal variance components are assumed and 4 if variances are allowed to be unequal. Others claim that there are more than 4 clusters (ranging between 6 and 9 (Grün et al. 2021)). Due to the uncertainty associated with $K^+$ in the galaxy data, they are well suited to illustrate how our prior construction can be used to carry out a principled sensitivity analysis for $K^+$. This is done by fitting an aFMM for a sequence of $U$ values and then exploring the prior's impact on properties of the mixture model like model-fit and co-clustering probabilities. With this in mind, we fit the aFMM to the galaxy data for $U \in \{2, \ldots, 10\}$, $tp \in \{0.1, 0.5\}$ and $\alpha_2 = 10^{-5}$. We employ the same prior distribution specification as in Section 4. The aFMM is fit by collecting 1000 MCMC samples after discarding the first 10,000 as burn-in and thinning by 100 (i.e., 110,000 total MCMC samples).

The posterior distributions of $K^+$ for $U \in \{3, 5, 7, 10\}$ and the induced priors on $K^+$ are provided in Figure S8. Notice that the posterior distribution of $K^+$ is influenced quite heavily by $U$ for $tp = 0.1$ and also for $tp = 0.5$, but less so. In both cases $mode(K^+ \mid \boldsymbol{y}) = U$ for each value of $U$. At first glance this may seem problematic, but $U$'s impact on the $mode(K^+ \mid \boldsymbol{y})$ is not seen in the clustering configuration for $U > 4$. To see this, we provide the co-clustering probability matrices for $tp = 0.1$ in Figure 7. The rows and columns of the co-clustering matrices are ordered by velocity. Notice that for $U \leq 3$ there are three clear clusters with little movement between them. This is expected as in the galaxy data there are three groups of velocities that are well separated (see Figure S9 for density estimates). For $U \geq 6$ there appear to be six clusters, but the co-clustering probabilities among units that belong to the two big clusters decrease as $U$ increases. Thus, even though $mode(K^+ \mid \boldsymbol{y})$ based on the aFMM follows $U$ for these data, it does so not by forming clusters that don't exist but by grouping units within the two big clusters in a fairly arbitrary way. As a result, the number of clusters based on a point estimate of the cluster configuration using, for example, the `salso` `R`-package (Dahl et al. 2022) results in 6 clusters for $U \geq 6$. For each model fit we also provide the following $U$-adjusted mean squared error (MSE)

$$\text{mse} = \frac{1}{n(K-U)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

(which compares observed to fitted values taking into account the number of clusters) and the standard deviation of co-clustering probabilities averaged across units.

$$\text{sd\_ccp} = \frac{1}{n} \sum_{i=1}^{n} \text{sd}(Pr(z_i = z_{-i} \mid \boldsymbol{y})). \tag{15}$$

This metric measures the cluster "purity" as units with co-clustering probabilities that have a larger standard deviation correspond to co-clustering probabilities that are closer to either

one or zero compared to those with a smaller standard deviation.

From Figure 7 notice that for $U \leq 3$ the cluster configuration is quite "pure" but with a high mse. This is not surprising because the galaxy data exhibits three well separated groups, but $K^+ = 3$ smooths over clear data features resulting in $K^+ > 3$ exhibiting a smaller mse. On the other hand, notice that the nominal number of clusters remains at six even though $mode(K^+ \mid \boldsymbol{y})$ increases as a function of $U$. It seems that $U \in \{5, 6, 7\}$ balances best the quality of cluster configuration and model fit (see Figure 7).

Fitting the aFMM for varying values of $U$ and observing the co-clustering probability matrix for each is a principled way to study the robustness or uncertainty of the cluster configuration that is easily carried out with the aFMM.

## 5.2 Biomechanic Functional Data Application

To illustrate the portability of our prior construction among different modeling scenarios, we now employ the aFMM in a functional data example from the field of biomechanics. In this setting, a "cluster" is defined to be a collection of curves that are similar in shape and magnitude as defined by a vector of B-spline coefficients. We briefly introduce the study that produced the data we consider.

Biomechanics is the study of how mechanical principles (force and angle) are applied to living organisms. There is keen interest in learning in what way human biomechanics are connected to joint health. To this end, 196 subjects that have had reconstructive anterior cruciate ligament (ACL) surgery were recruited to participate in a study that required them to walk on a treadmill. While walking the knee angle (among other biomechanic variables) was measured through the entire gait cycle (see Figure 8). Thus, the knee angle measurements could be thought of as discretized functional realizations.

We seek to identify a subset of movement strategies that subjects adopt post ACL surgery. In this study two perspectives and motives for discovering subpopulations exist. First, from

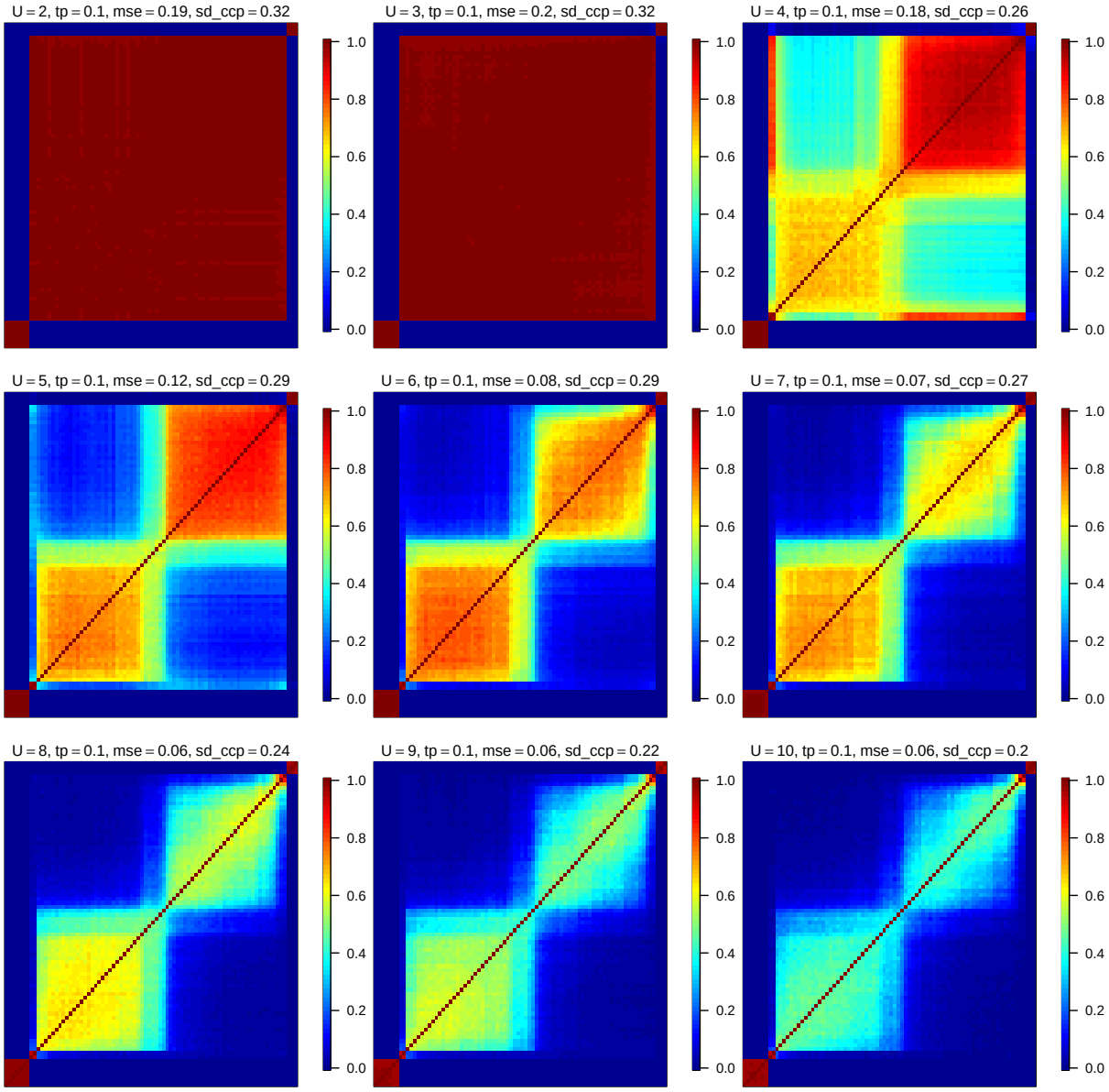


Figure 7: Posterior co-clustering probabilities for the Galaxy data from an aFMM fit using $U \in \{2, \ldots, 10\}$ and $tp = 0.1$. In addition, sd_ccp as defined in (15) and mse as defined in (14) are provided. In middle row panels, choice of $U \in \{5, 6, 7\}$ balances best the quality of cluster configuration and model fit (high purity and moderate mse).

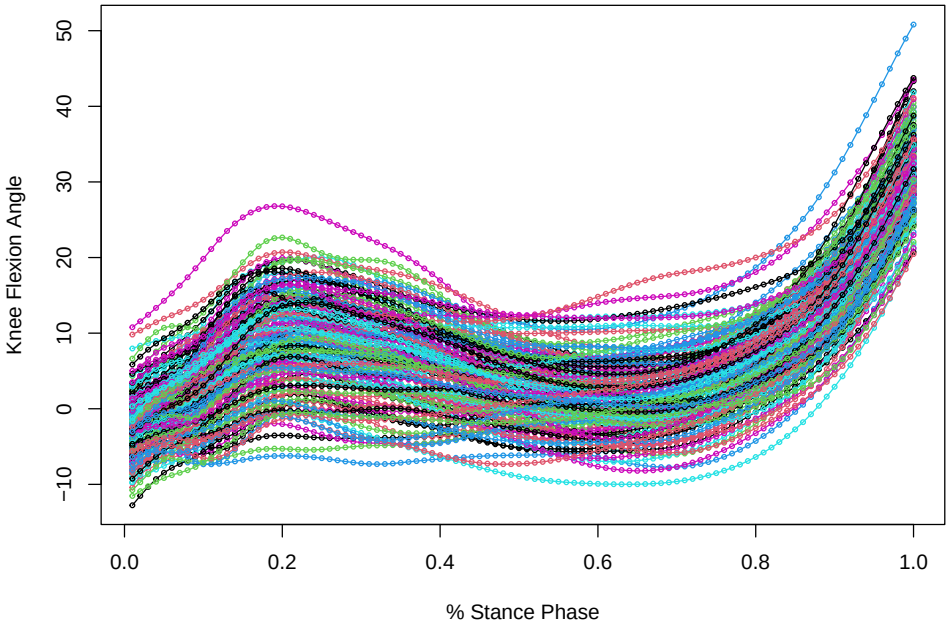


Figure 8: Knee flexion angle for each of the 196 subjects recruited into the study.

a clinician perspective, it would be very useful if a relatively few number of movement strategies are identified as this would facilitate interpretation and treatment formulation (e.g., rigid knee movement, typical knee movement, and flexible knee movement). However, an exercise scientist would not necessarily be concerned with identifying a small number of "interpretable" movement strategies, but rather understand the myriad of ways that the 196 subjects are able to accommodate the ACL surgery. So a potentially larger number of subgroups would be of interest. An appeal of the aFMM is that it can be employed to lucidly handle both situations through the specification of $U$.

### 5.2.1 Description of Functional Clustering Model

We employ a functional clustering model that is similar to that detailed in Page et al. (2020). For sake of completeness, we detail it here. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{mi})$ denote the $m$ knee angle

measurements for subject $i$. A reasonable functional data model for the data in Figure 8 is

$$y_i(t) = \beta_{0i} + f_i(t) + \epsilon_i(t) \text{ where } \epsilon_i(t) \overset{iid}{\sim} N(0, \sigma_i^2),$$

where $f_i(\cdot)$ denotes the $i$th subject's knee angle function and $\beta_{0i}$ a vertical shift. Here we assume constant variance at each time point $t \in [0, 1]$. With the desire to flexibly model each subject's curve, we approximate $f_i(\cdot)$ using B-splines which results in the following subject-specific model

$$\boldsymbol{y}_i \sim N(\beta_{0i}\boldsymbol{j} + \boldsymbol{B}_i\boldsymbol{\beta}_i, \sigma_i^2\boldsymbol{I}),$$

where $\boldsymbol{B}_i$ is a $m \times p$ matrix of B-spline basis created by using evenly spaced interior knots and $\boldsymbol{\beta}_i$ a $p$-dimensional vector of B-spline coefficients for the $i$th subject. Curve clustering is then carried out by modeling $\boldsymbol{\beta}_i$ with an aFMM

$$\begin{aligned}
\boldsymbol{\beta}_i &\sim \sum_{k=1}^{K} w_k N(\boldsymbol{\theta}_k, \kappa_k^2\boldsymbol{I}) \\
\boldsymbol{w} &\sim \text{Dirichlet}(\boldsymbol{\alpha}_{1,2}) \\
\alpha_1|\alpha_2, U &\sim PC(U, tp = 0.1).
\end{aligned}$$

Smoothing is introduced by modeling $\boldsymbol{\theta}_k$ with a penalized B-spline (Eilers and Marx 1996; Lang and Brezger 2004) under the PC prior framework (Simpson et al. 2017) such that

$$\begin{aligned}
Pr(\boldsymbol{\theta}_k) &\propto \exp\{1/\tau_k^2\boldsymbol{\theta}_k'\boldsymbol{S}\boldsymbol{\theta}_k\} \\
\tau_k &\sim Exp(\eta_\tau).
\end{aligned}$$

Here $\boldsymbol{S}$ is a 2nd order random walk penalty matrix and $\eta_\tau = -\log(a_\tau)/U_\tau$ where $a_\tau$ and $U_\tau$ satisfy $Pr(1/\tau_k > U_\tau) = a_\tau$ (the PC prior for the standard deviation of a Gaussian random effect, e.g. $\tau_k$, is the Exponential distribution). Finally, to balance borrowing-of-strength among units allocated to the same cluster and subject-specific fits, we employ the following priors on the between-subject and within-subject variance components

$$\sigma_i \sim UN(0, A)$$
$$\kappa_k \sim UN(0, A_0).$$

We set $A = 0.001$ as the measured curves are essentially noiseless and $A_0 = 0.25$ which requires clusters to be composed of similary shaped curves. For $\tau_k^2$ we set $a_\tau = 10^{-2}$ and $U_\tau = 3.22$ as suggested by Simpson et al. (2017). In order to avoid the challenges inherent in multivariate clustering (Chandra et al. 2023; Ghilotti et al. 2023), we used a small number of interior knots (seven) in the P-spline formulation. This resulted in $\boldsymbol{\beta}_i$ being $p = 10$ dimensional which is small enough to not suffer from the curse of dimensionality (Ghilotti et al. 2023). Since the measured knee angle curves are sufficiently smooth each subject's curve is fit well even with 7 interior knots. Lastly, we set $K = 25$ for all mixtures that are fit to these data.

To perform clustering from both the clinician's and exercise scientist's perspective we set $U = 3$ and $U = 10$ with $tp = 0.1$?. For additional context we also fit a sFMM and a static FMM with $\alpha = 1/K$. Each of the models were fit by collecting 1,000 MCMC iterates after discarding the first 50,000 and thinning by 100 (this required 150,000 total MCMC samples for each model).

The posterior distributions of $K^+$ under all four models turned out to be points masses at specific values. For $U = 3$, $mode(K^+|\boldsymbol{y}) = 6$ which demonstrates that $U$ is indeed a "soft" upper bound that can be exceeded when favored by the data. For $U = 10$, $mode(K^+|\boldsymbol{y}) = 9$,

while for sFMM and the static FMM $mode(K^+|\boldsymbol{y}) = 7$. To further explore the clustering results under each model, we estimated the cluster configuration based on the MCMC samples using the default settings of the `salso` function (Dahl et al. 2022). Interestingly the estimated clustering under the four models were all quite different as all the pair-wise adjusted Rand index (ARI) (Hubert and Arabie 1985) values between them were less than 0.5. To further see differences, we provide Figure S10 of the supplementary material which displays the co-clustering probabilities under each model that was fit. Clusters are labeled based on subject order. That is, subject one is always allocated to cluster one, and cluster two begins with first subject not allocated to subject one's cluster, and cluster three begins with first subject not allocated to cluster one or two, etc. Note that there are a subset of subjects that exhibit uncertainty in their cluster allocation, but for the most part the clustering is estimated with low uncertainty.

To visualize the clustering further we provide Figure 9 and Figures S11 - S14 in the supplementary material. The left column of Figure 9 displays the subject-specific curve fits with color indicating cluster memberhship and the right column displays the cluster-specific mean curves calculated cross-sectionally using all curves allocated to a particular cluster. Note that the subject-specific fits (solid lines through points) are very reasonable for the majority of subjects. Notice further that one subject was allocated to a single cluster under each model. It is clear why this is the case as the subjects curve is quite different from the others. Key differences that exist between the clusters seem to be the height of the knee angle curve early in the stance phase and also the depth of the valley and the sharpness of the drop towards the middle of the stance phase. It does appear that the desire by clinician's to have 3 clusters forced some subjects whose curves are quite different to be grouped (see Figure S11 of the supplementary material) highlighting the fact that these data highly favor more than three clusters. The aFMM for $U = 10$ generally speaking produced clusters with curves that are more homogeneous relative to the other models. Cluster seven in the aFMM

$U = 10$ model would be quite interesting to exercise scientists as it is generally agreed that a shallow valley in the knee angle curve represents "poor" biomechanics. This represents a gate that does not employ much bend at the knee. Overall, employing the aFMM provides a principled approach to consider both perspectives and the sFMM and static FMM seem to fall somewhere in between the two aFMM models with regards to clusters that exhibit curve homogeneity.

# 6 Discussion

In this paper we've constructed a prior distribution for arguably the most relevant quantity in model-based clustering; the number of clusters. This was done by employing an asymmetric Dirichlet distribution as a prior on the weights of a finite mixture. Further, employing PC prior type technology, we formulated a prior distribution on the shape of the Dirichlet that permits eliciting prior information through intuitive statements that can be asked of the user.

Our methodology also permits a principled study of the uncertainty associated with the clustering configuration. The uncertainty associated with $K^+$ can be studied in two ways. The first is through co-clustering probabilities with those that are more "pure" indicating a more certain clustering. Uncertainty can also be explored by studying the stability of the clustering configuration as the value of $K^+$ is changed *a priori*. If either of these two perspectives exhibit uncertainty, then the data are not that informative regarding the number of clusters. Our prior construction leads to naturally being able to employ both perspectives.

Finally, model-based clustering procedures are, at the end of the day, exploratory approaches that permit users to discover structure in their data. Our procedure, according to our knowledge, is the first to provide users the ability of carrying out the exploratory data analysis in a principled way based on $K^+$. As a result, the influence that the prior

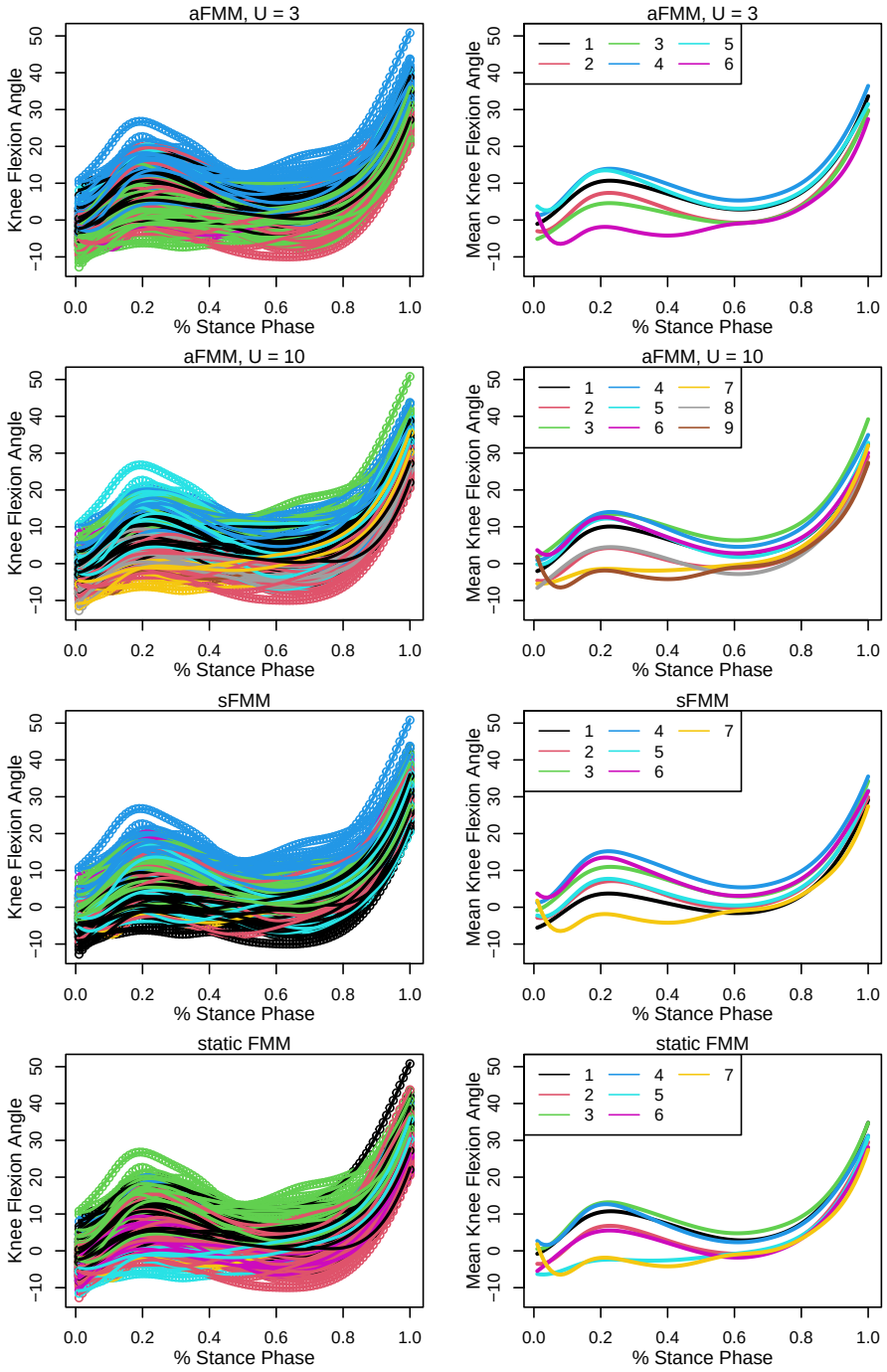


Figure 9: The left column displays the curve fits for all 196 subjects with color indicating cluster (estimated using default settings of the `salso` function). The right column correspond to the cluster means which were calculated by finding cross-sectional mean among all curves in a cluster. The first row corresponds to an aFMM with $U = 3$, the second an aFMM with $U = 10$, the third a sFMM, and the fourth a static FMM.

distribution of $K^+$ has on its posterior is something that can easily be studied.

**Acknowledgements**: We would like to thank to Christian Hennig for stimulating discussions about the definition of a cluster, José J Quinlan for discussions on details associated with the proof of Proposition 1, and the Motion Science Institute at the University of North Carolina for access to the biomechanics dataset.

# References

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

Alamichel, L., Bystrova, D., Arbel, J., and King, G. K. K. (2023), "Bayesian mixture models (in)consistency for the number of clusters," arXiv:2210.14201.

Argiento, R. and De Iorio, M. (2022), "Is Infinity that Far? A Bayesian Nonparametric Perspective of Finite Mixture Models," *Annals of Statistics*, 50, 2641–2663.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023), "Clustering consistency with Dirichlet process mixtures," *Biometrika*, 110, 551–558.

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2022), "MCMC Computations for Bayesian Mixture Models Using Repulsive Point Processes," *Journal of Computational and Graphical Statistics*, 0, 1–14.

Cai, D., Campbell, T., and Broderick, T. (2021), "Finite mixture models do not reliably learn the number of components," in *Proceedings of the 38th International Conference on Machine Learning*, eds. Meila, M. and Zhang, T., PMLR, vol. 139 of *Proceedings of Machine Learning Research*, pp. 1158–1169.

Chandra, N. K., Canale, A., and Dunson, D. B. (2023), "Escaping The Curse of Dimensionality in Bayesian Model-Based Clustering," *Journal of Machine Learning Research*, 24, 1–42.

Dahl, D. B., Johnson, D. J., and Müller, P. (2022), *salso: Search Algorithms and Loss Functions for Bayesian Clustering*, r package version 0.3.29.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89–121.

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 volume*, 13, 33–64.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021), "Generalized Mixtures of Finite Mixtures and Telescoping Sampling," *Bayesian Analysis*, 16, 1279–1307.

Ghilotti, L., Beraha, M., and Guglielmi, A. (2023), "Bayesian clustering of high-dimensional data via latent repulsive mixtures," arXiv:2303.02438.

Ghosal, S. and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Greve, J. (2021), *fipp: Induced Priors in Bayesian Mixture Models*, r package version 1.0.0.

Greve, J., Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. (2021), "Spying on the Prior of the Number of Data Clusters and the Partition Distribution in Bayesian cluster Analysis," *Australian & New Zealand Journal of Statistics*, 0.

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*.

Hennig, C. (2015), "What are the true clusters?" *Pattern Recognition Letters*, 64, 53–62.

Hubert, L. and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

Ishwaran, H. and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Komárek, A. and Komárková, L. (2014), "Capabilities of R Package mixAK for Clustering Based on Multivariate Continuous and Discrete Longitudinal Data," *Journal of Statistical Software*, 59, 1–38.

Kullback, S. and Leibler, R. A. (1951), "On information and sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Lang, S. and Brezger, A. (2004), "Bayesian P-Splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Lijoi, A., Prüenster, I., and Rigon, T. (2023), "Finite-dimensional Discrete Random Structures and Bayesian Clustering," *Journal of the American Statistical Society*, 0, 0–0.

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.

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 26*, eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., Curran Associates, Inc., vol. 26, pp. 199–206.

— (2018), "Mixture Models With a Prior on the Number of Components," *Journal of the American Statistical Association*, 113, 340–356.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, Springer.

Ong, P., Argiento, R., Bodin, B., and De Iorio, M. (2021), *AntMAN: Anthology of Mixture Analysis Tools*, r package version 1.1.0.

Page, G. L., Rodríguez-Álvarez, M. X., and Lee, D.-J. (2020), "Bayesian Hierarchical Modelling of Growth Curve Derivatives via Sequences of Quotient Differences," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69, 459–481.

Petralia, F., Rao, V., and Dunson, D. B. (2012), "Repulsive Mixtures," in *Advances in Neural Information Processing Systems 25*, eds. Pereira, F., Burges, C., Bottou, L., and Weinberger, K., Curran Associates, Inc., pp. 1889–1897.

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2021), "On a Class of Repulsive Mixture Models," *Test*, 30, 445–446.

Regazzini, E., Lijoi, A., and Prüenster, I. (2003), "Distributional Results for Means of Normalized Random Measures with Independent Increments," *Annals of Statistics*, 31, 560–585.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society: Series B*, 859, 731–792.

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

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017), "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors," *Statistical Science*, 32, 1–28.

Stephens, M. (2000), "Bayesian Analysis of Mixture Models with an Unknown Number of Components- An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40–74.

Sun, H., Zhang, B., and Rao, V. (2022), "Bayesian Repulsive Mixture Modeling with Matérn Point Processes," arXiv:2210.04140.

Tange, O. (2022), "GNU Parallel 20220722 ('Roe vs Wade')," GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

Xie, F. and Xu, Y. (2020), "Bayesian Repulsive Gaussian Mixture Model," *Journal of the American Statistical Association*, 115, 187–203.