ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Informed Bayesian Finite Mixture Models via Asymmetric Dirichlet Priors

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Page, G.l., Ventrucci, M., Franco-Villoria, M., Seeley, M.k. (2025). Informed Bayesian Finite Mixture Models via Asymmetric Dirichlet Priors. THE ANNALS OF APPLIED STATISTICS, 19(3 (September)), 2412-2435 [10.1214/25-AOAS2031].

Availability:
This version is available at: https://hdl.handle.net/11585/1009925 since: 2025-12-09

Published:
DOI: http://doi.org/10.1214/25-AOAS2031

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

28 September 2026

Submitted to the Annals of Applied Statistics

INFORMED BAYESIAN FINITE MIXTURE MODELS VIA ASYMMETRIC DIRICHLET PRIORS

BY GARRITT L. PAGE[1,a], MASSIMO VENTRUCCI[2,b]
MARIA FRANCO-VILLORIA[3,c] AND MATTHEW K. SEELEY[4,d]

[1]*Department of Statistics, Brigham Young University,* [a]*page@stat.byu.edu*

[2]*Department of Statistical Sciences, University of Bologna,* [b]*massimo.ventrucci@unibo.it*

[3]*Department of Economics "Marco Biagi", University of Modena and Reggio Emilia,* [c]*maria.francovilloria@unimore.it*

[4]*Department of Exercise Science, Brigham Young University,* [d]*matt_seeley@byu.edu*

There is keen interest in the field of biomechanics to identify unique strategies of negotiating specific movement tasks post lower-limb joint injury. Finite mixture models are flexible methods that are commonly used for carrying out this type of task. 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. This can make prior elicitation on the number of clusters challenging when prior information exists which is the case in the biomechanic study considered here. In light of this, we develop a finite mixture methodology that permits eliciting prior information directly on the number of clusters in a flexible and 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. In addition to applying the method to the biomechanic data, we also demonstrate utility using numerical experiments and the galaxies dataset.

1. Introduction. Biomechanics is the study of mechanical principles (e.g., forces and angles) applied to living organisms. A common focus of biomechanical studies is to discover how forces and/or angles affect musculoskeletal health, or how health influences biomechanics. For example, it is known that forces applied to knee cartilage while humans walk or run influence knee joint articular cartilage health and quality of life. Magnitude, rate, and location of knee joint forces all influence the effects that force has on knee joint articular cartilage health.

Recently researchers have learned that anterior cruciate ligament injury and the subsequent reconstructive surgery (ACLR) affect walking biomechanics, in different ways, for different ACLR patients. For example, some ACLR patients walk with a knee that is flexed less during the early part of the ground contact phase of walking, which is often associated with decreased vertical ground reaction force (vGRF) during the same walking phase. Further, these ACLR patients, who experience less vGRF during the early part of ground contact, often report less desirable outcomes in areas such as pain, stiffness, or quality of life (Dennis et al. 2024). This is one important reason why biomechanical researchers are interested in increasing basic knowledge of biomechanical patterns during movement, like walking, for

Keywords and phrases: bayesian clustering, penalized complexity priors, functional data, estimating the number of clusters.

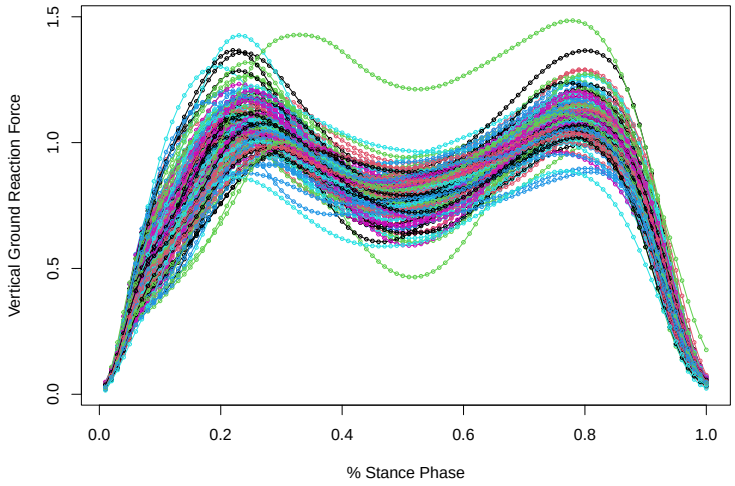

FIG 1. *Vertical ground reaction force curves for each of the 196 subjects recruited into the study. Color is provided only to help visualize individual curves*

ACLR patients. To this end, 196 ACLR patients were recruited to participate in a study that required them to walk on a treadmill. While walking vGRF was measured through the entire gait cycle. Figure 1 displays the resulting vGRF measurements for each patient. Notice that the curves share some features between subjects, but intuition dictates that each individual ACLR patient's curve is a unique vGRF pattern. Due to this, it would be extremely difficult to extract useful information without reducing the number of vGRF patterns. Thus, researchers would benefit from identifying a reduced number (i.e., less than the number of observed curves) of representative vGRF patterns for analytical considerations.

Along with biomechanical researchers, clinicians charged with rehabilitating ACLR patients are also interested in identifying a small number of distinct representative vGRF walking patterns employed by ACLR patients. However, opposed to biomechanical researchers, clinicians are more limited regarding the number of representative biomechanical patterns they can consider or effectively use. This is because current practical clinical limitations (e.g., a limited number of clinicians, or limited number of therapeutic interventions available to any one clinician) make it challenging to effectively administer treatments to patients if the number of treatments based on patient clusters is not small. For example, it is now impractical for a clinician, in the same day or week, to treat eight different patients, who each exhibit a distinct vGRF pattern, using eight distinct therapeutic approaches.

Thus, there are two perspectives for clustering the 196 ACLR patients that differ primarily on the need/ability to interpret the clustering results. Directly incorporating these types of prior information into any type of clustering model is cumbersome at best. This is particularly true of finite mixture models (FMMs), the focus of this work, where prior information is currently directly connected to the number of mixture components, not the number of clusters. Because of this, there is a need to develop an approach that allows practitioners, like the clinician and researcher in the biomechanics application, to incorporate prior information on the number of clusters in a FMM in an intuitive and flexible way (e.g., incorporating a soft upper bound on the number of clusters as informed by a clinician through prior elicitation). We seek to address this need.

FMMs have become a popular tool for model-based clustering. For treatise on FMMs see McLachlan and Peel (2005) or Frühwirth-Schnatter (2006). An underlying assumption of

FMMs is that each unit's measured realization comes from one of K possible subgroups or mixture components 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 and, as a by product, a partitioning or clustering of the n units/subjects. In its most general form a FMM can be expressed as

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

where $\boldsymbol{w} = (w_1, \ldots, w_K)$ are component weights such that $\sum_{k=1}^{K} w_k = 1$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K)$ component specific parameters, $f_k(\cdot)$ a component density, and K the number of components. Historically the number of clusters in an FMM has been connected to K. However, there is a growing literature that distinguishes between the number of mixture components K and the number of data informed clusters, which we denote as K^+ (Frühwirth-Schnatter and Malsiner-Walli 2019; Greve et al. 2022; Frühwirth-Schnatter, Malsiner-Walli and Grün 2021; Quinlan, Quintana and Page 2021; Argiento and De Iorio 2022; Alamichel et al. 2024). Note that the distinction between the number of mixture components K and the number of clusters K^+ make it so the object of interest (K^+) is not explicitly parameterized in the FMM.

It is well known that Bayesian FMMs can be quite delicate to model and prior specifications and that common vague priors perform poorly (Grazian and Robert 2018). This is particularly true for K^+ as its prior distribution is induced through the prior distribution of $\boldsymbol{w}$ and, if K is not fixed, the prior distribution of K (see Section 2 for background on how K is treated in an FMM). Because K^+ is typically of principal interest when forumulating an FMM, the prior distribution of K^+ induced by particular FMM modeling decisions has begun to garner attention in the literature. In fact, 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 mixture of FMM (MFMM) (see Miller and Harrison 2018) that assume a symmetric Dirichlet prior on the weights with shape parameter (denoted here as α) being fixed (static MFMM) or scaled by K (dynamic MFMM). Figure S1 in the supplementary material shows the implied prior computed using `fipp` for the DPM and static MFMM and different choices of the Dirichlet shape parameter. By trial and error, an expert user with prior information on K^+ like that in the biomechanic application described previously can tune the prior on $\boldsymbol{w}$ until the induced prior on K^+ resembles his/her prior belief on the number of occupied components. However, scientists can only play a passive role here, in the sense that they cannot elicit the prior for K^+ in a direct way, for instance by eliciting the mode or some other centrality parameter with an accompanying distributional shape. This is due to the fact that the prior on K^+ induced by a symmetric Dirichlet is in general quite inflexible as it is based on a single shape parameter α. As a result, it is not simple to control the probability mass assigned to a specific range of values.

Although formally considering the induced prior on K^+ is certainly useful and improves the overall understanding of the FMM prior structure, our contribution is to develop an approach that permits users to "inform" the FMM (K^+ in particular) in a direct way. For example, assigning prior weight to K^+ values that are less than or equal to some some soft upper bound (i.e., the upper bound can be exceeded). This is done by replacing the symmetric Dirichlet by what we call an asymmetric Dirichlet which is comprised of two shape parameters denoted as α_1 and α_2, plus an additional parameter denoted as U, which is the number of component weights taking value α_1. We will show in Section 3 that, under some conditions on α_1 and α_2, the parameter U can easily be connected to K^+. The first advantage

of the asymmetric Dirichlet is in terms of intuitiveness of the prior elicitation. A scientist can inform the mode of the induced prior distribution on K^+ by handling U directly. A second advantage is in terms of prior shape flexibility as, by handling α_1 and α_2, the user can control the probability mass assigned to a given range of values of K^+ more precisely than what can be done by handling α in the symmetric Dirichlet (as in the `fipp` package). Thus, this formulation is simple and yet effective in eliciting prior information through probabilistic statements associated with a user-supplied value of K^+.

In general, 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 base model unless the data indicate otherwise. An appealing feature of this approach is that scientific questions are able to guide base model selection, as for instance an FMM with a user-defined number of non-empty components K^+. Practitioners can then inform the model *a priori* by thinking directly about K^+, while tuning parameters of the asymmetric Dirichlet that control *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 of 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.

The rest of the article is organized as follows. In Section 2 we detail a functional data model used to model the vGRF curves and 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 two simulation studies; one in a univariate setting to build intuition and a second one that is more similar to the vGRF curves. In this section we also consider the galaxies data as an illustration. Section 5 describes results from the biomechanic functional data application. 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 application detailed in Sections 4 and 5.

2. Background on Functional Data Model and Bayesian Finite Mixture Models. We first provide necessary details of the functional data model we employ to model the vGRF curves in Figure 1:, the basis of which is developed in Page and Quintana (2015). Then background for FMMs is provided.

Let $y_i(t)$ denote the vGRF measurement at $t \in [0,1]$ for subjects $i = 1, \ldots, n$ ($m = 100$ measurements were taken along the timescale for each subject). We consider the following data model for each subjects measured curve

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

where $f_i(\cdot)$ denotes the ith subject's vGRF function. Note that constant variance at each time point t is assumed. With the desire to flexibly model each subject's curve, we approximate $f_i(\cdot)$ using B-splines so that $f_i(t) \approx b(t, \boldsymbol{\xi})'\boldsymbol{\beta}_i$ where $b(\cdot, \boldsymbol{\xi})$ is a B-spline basis function with knots $\boldsymbol{\xi}$ and $\boldsymbol{\beta}_i$ a p-dimensional vector of B-spline coefficients for the ith subject. Collecting each subjects m measurements into vector $\boldsymbol{y}_i = (y_i(t_1), \ldots y_i(t_m))$ permits expressing each

subject's model in the following matrix form $\boldsymbol{y}_i \sim N(\boldsymbol{B}_i\boldsymbol{\beta}_i, \sigma_i^2\boldsymbol{I})$, where $\boldsymbol{B}_i$ is a $m \times p$ matrix of B-spline basis functions created by using evenly spaced knots $\boldsymbol{\xi}$. Curve clustering is then carried out by modeling $\boldsymbol{\beta}_i$ with a multivariate Gaussian version of the FMM in (1) so that

$$\boldsymbol{\beta}_i \sim \sum_{k=1}^{K} w_k N_p(\boldsymbol{\theta}_k, \kappa_k^2\boldsymbol{I}) \tag{3}$$

and $\boldsymbol{\eta}_k = (\boldsymbol{\theta}_k, \kappa_k^2)$. Here κ_k^2 regulates the homogeneity of curve shapes in the kth mixture component. Note further that each entry of $\boldsymbol{\beta}_i$ varies independently around the corresponding entry of $\boldsymbol{\theta}_k$.

To facilitate knot selection associated with $\boldsymbol{\theta}_k$ we employ penalized B-spline (Eilers and Marx 1996; Lang and Brezger 2004) under the PC prior framework (Simpson et al. 2017; Ventrucci and Rue 2016) so that

$$Pr(\boldsymbol{\theta}_k) \propto \exp\{1/\tau_k^2 \boldsymbol{\theta}_k' \boldsymbol{S} \boldsymbol{\theta}_k\} \text{ with } \tau_k \sim Exp(\eta_\tau). \tag{4}$$

Here $\boldsymbol{S}$ is a 2nd order random walk penalty matrix and $\eta_\tau = -\log(a_\tau)/U_\tau$ where a_τ and U_τ satisfy $Pr(1/\tau_k > U_\tau) = a_\tau$ (the PC prior for the standard deviation of a Gaussian random effect, e.g. τ_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

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

We set $A = 0.001$ as the measured curves are essentially noiseless. A_0 directly impacts the posterior distribution of K^+ as it regulates the within cluster curve homogeneity. However, it has not effect on the prior of K^+. Since our focus is developing a prior on K^+ through the prior on $\boldsymbol{w}$, we simply set A_0 to a reasonable value as suggested by the simulation study in Section 4.3. More details are provided in Section 5. For τ_k^2 we set $a_\tau = 10^{-2}$ and $U_\tau = 1/0.31$ (following the rule of thumb suggested by Simpson et al. 2017). In order to avoid the challenges inherent in multivariate clustering (Chandra, Canale and Dunson 2023; Ghilotti, Beraha and Guglielmi 2023), we used a small number of interior knots (ten) in the P-spline formulation. This resulted in $\boldsymbol{\beta}_i$ being $p = 13$ dimensional which is small enough to not suffer from the curse of dimensionality (Ghilotti, Beraha and Guglielmi 2023). Since the measured vGRF curves are sufficiently smooth each subject's curve is fit well with 10 interior knots.

2.1. *Strategies for Handling K.* Decisions about K in an FMM are particularly impactful on the number of clusters. 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). 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, Frühwirth-Schnatter and Grün (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 $\boldsymbol{\eta}_k$. See Petralia, Rao and Dunson (2012), Xie and Xu (2020), Quinlan, Quintana and Page (2021), Beraha et al. (2022), and Sun, Zhang and Rao (2022).

A Bayesian nonparametric (BNP) perspective essentially side-steps the need to formally consider K by setting $K = \infty$ (Müller et al. 2015; Ghosal and van der Vaart 2017). However, Miller and Harrison (2013) pointed out that estimating K^+ using a Dirichlet Process Model (DPM) can be problematic. In fact, Cai, Campbell and Broderick (2021) find that estimating K^+ consistently depends on specifying components correctly (i.e., correctly defining the meaning of a cluster). As a result, Lijoi, Prüenster and Rigon (2023) studied in more depth the finite-dimensional Bayesian clustering from a normalized random measure with independent increments (Regazzini, Lijoi and Prüenster 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, Malsiner-Walli and Grün (2021) made very interesting connections between BNP mixtures and FMMs while Alamichel et al. (2024) studied the consistency in estimating K^+ (or lack there of) in a variety of mixture models.

2.2. *Latent Mixture Component Labels.* For computational purposes, the FMM in (3) is often re-expressed using latent component labels. Doing so permits describing the mixture model hierarchically. To this end, let $z_1, \ldots, z_n$ denote n component labels where $z_i = j$ implies that the ith unit belongs to the jth component. Then FMM in (3) can be written as

$$\begin{aligned} \boldsymbol{\beta}_i \mid z_i &\overset{ind}{\sim} N(\boldsymbol{\theta}_{z_i}, \kappa^2_{z_i}\boldsymbol{I}),\ i = 1, \ldots, n, \\ Pr(z_i = k \mid \boldsymbol{w}) &= w_k,\ i = 1, \ldots, n,\ k = 1, \ldots, K. \end{aligned} \tag{5}$$

We have discussed above the prior distributions for $\boldsymbol{\theta}_k$, $\boldsymbol{\kappa}^2_k$ which are those typical of Bayesian P-splines (Lang and Brezger 2004). For $\boldsymbol{w}$ it is common to employ a Dirichlet prior with shape parameters $\boldsymbol{\alpha}$ so that $\boldsymbol{w} \sim \text{Dirichlet}(\boldsymbol{\alpha})$. It would also be possible to assign a prior distribution to K in what we develop, but we focus on the case when K is fixed to a large value (Rousseau and Mengersen 2011).

As mentioned, a common distribution for $\boldsymbol{w}$ is the symmetric Dirichlet distribution 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, α has been assigned a prior. In particular, Malsiner-Walli, Frühwirth-Schnatter and Grün (2016); Frühwirth-Schnatter and Malsiner-Walli (2019); Greve et al. (2022); Frühwirth-Schnatter, Malsiner-Walli and Grün (2021) all assume $\alpha \sim Gamma(a, aK)$ so that $E(\alpha) = 1/K$. The parameter α 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, the induced prior on K^+ is fairly inflexible making it difficult that user specified values for K^+ are given large prior mass as is needed in the biomechanic application. 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 α. 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 3.1. The asymmetric Dirichlet, denoted by Dirichlet$(\boldsymbol{\alpha}_{1,2})$, is a Dirichlet distribution with parameters U, α_1, α_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 3.1, U plays a crucial role as a user-supplied value on which the induced 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 3.2, but first build some intuition why this is the case. Note that under $\boldsymbol{w} \sim Dirichlet(\boldsymbol{\alpha}_{1,2})$

$$E(w_k) = \begin{cases} \dfrac{\alpha_1}{\alpha_1 U + \alpha_2(K-U)} & k = 1, \ldots, U \\ \dfrac{\alpha_2}{\alpha_1 U + \alpha_2(K-U)} & k = U+1, \ldots, K. \end{cases} \tag{6}$$

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

$$\lim_{\alpha_1 \to \infty} \lim_{\alpha_2 \to 0} Pr(K^+ = U \mid K, n, \alpha_1, \alpha_2) = 1 \tag{7}$$

In order to visualize the implied prior of K^+ from the aFMM, we provide Figure 2 where different values for α_1 and α_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 3.2. Note that increasing α_1 and decreasing α_2 leads to an implied prior for K^+ highly concentrated on $U = 10$, so that U becomes the mode; it is sufficient fixing $\alpha_1 = U$ and $\alpha_2 = 0.001$ to get a spike on $K^+ = 10$ in this case. By moving α_1 and α_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 α_1 while keeping α_2 small (e.g. $\alpha_2 = 0.00001$, see right-hand panel of Figure 2), we get more probability in the left tail so that $U = 10$ can be treated as a *soft* upper bound. Analogously, by increasing α_2 while α_1 being large, we obtain a prior with heavy right tail and so $U = 10$ can be treated as a *soft* lower bound (see left-hand panel of Figure 2). Thinking about an upper/lower bound rather than the centre of the distribution may be a more intuitive way of eliciting the prior for the non-expert user and corresponds to a more informative prior than using a distribution that is centered around U. Nonetheless, a prior on K^+ more symmetric around U could be obtained by tuning α_2 accordingly.

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 3.3. The aFMM as described in (5) - (4) with $\boldsymbol{w} \sim \text{Dirichlet}(\boldsymbol{\alpha_{1,2}})$ satisfies the assumptions in Rousseau and Mengersen (2011) so that if $\min(\alpha_1, \alpha_2) < dim/2$ the asymptotic vanishing weights property holds were 'dim' is the dimension of the component densities.

The proof of Remark 3.3 follows from the arguments laid out in Alamichel et al. (2024). Note that Remark 3.3 holds only if α_1 and α_2 are considered fixed quantities.

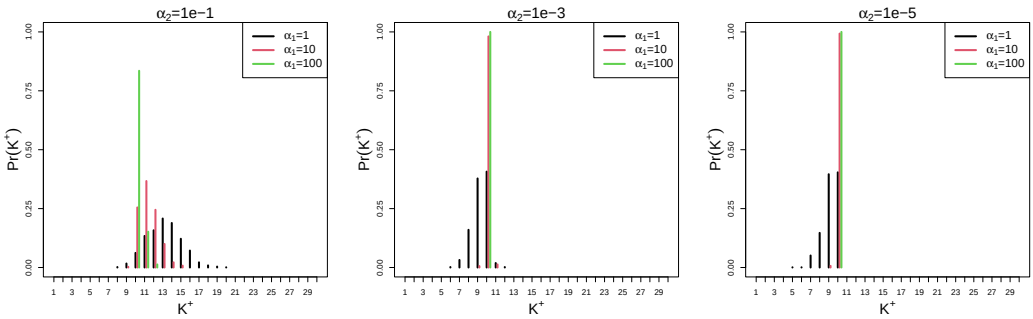

FIG 2. *Induced prior distribution of K^+ obtained via simulations, assuming the asymmetric Dirichlet prior with $\alpha_1 = \{1, 10, 100\}$ and $\alpha_2 = \{0.1, 0.001, 0.00001\}$ and $U = 10$, $n = 100, K = 30$.*

3.1. *Prior Distributions for α_1 and α_2.* There are a number of ways to treat (α_1, α_2). The list includes A) Fix α_1 and α_2 to prespecified values, which corresponds to an asymmetric case of the static model described in Frühwirth-Schnatter, Malsiner-Walli and Grün (2021); B) assume that α_1 is unknown with an assigned prior distribution and fix α_2 to a small value; C) Fix α_1 to a prespecified value and assume α_2 is unknown with an assigned prior distribution; and D) assume both α_1 and α_2 are unknown 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 α_1 is assigned a prior and α_2 is fixed at some prespecified value. This permits us to treat U as a soft upper bound for K^+ with α_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 α_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 (α_1, α_2) such that the induced prior on K^+ guarantees that an FMM with a user-defined number of non-empty components, hereafter denoted as *base finite mixture model* (base FMM), is favoured unless data support an alternative FMM. In other words, we wish to use the induced prior for K^+ as a mechanism to regulate the behaviour of the FMM with respect to a user-defined 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^+. For example, a base FMM favouring $K^+ = 10$ can be obtained by a Dirichlet$(\boldsymbol{\alpha}_{01,02})$ as a prior on $\boldsymbol{w}$ with parameters $\alpha_{01} = \infty$, $\alpha_{02} = 0$, and $U = 10$. 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 α_{01} and a "small" value for

α_{02}. Numerical experiments lead us to set $\alpha_{01} = U$ and $\alpha_{02} = 10^{-5}$, as this choice worked well for a variety of values for n and K.)

Let $\boldsymbol{g} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{01,02})$ be the asymmetric Dirichlet under the base FMM (with parameters $\alpha_{01} = U$, $\alpha_{02} = 10^{-5}$ and U), while $\boldsymbol{p} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{1,2})$ be the asymmetric Dirichlet under the alternative FMM (with parameters $\alpha_1 > 0$, $\alpha_2 > 0$ and U). The *deviation* from the alternative FMM to the base FMM is measured using the KLD between $\boldsymbol{p}$ and $\boldsymbol{g}$:

$$
\begin{aligned}
KLD(\boldsymbol{p}||\boldsymbol{g}) = &\log\Gamma(\alpha_1 U + \alpha_2(K-U)) - \log\Gamma(\alpha_{01}U + \alpha_{02}(K-U)) - \\
&(U\log\Gamma(\alpha_1) + (K-U)\log\Gamma(\alpha_2)) + U\log\Gamma(\alpha_{01}) + (K-U)\log\Gamma(\alpha_{02}) + \\
&U(\alpha_1 - \alpha_{01})[\psi(\alpha_1) - \psi(\alpha_1 U + \alpha_2(K-U))] + \\
&(K-U)(\alpha_2 - \alpha_{02})[\psi(\alpha_2) - \psi(\alpha_1 U + \alpha_2(K-U))], \qquad (8)
\end{aligned}
$$

where Γ and ψ are, respectively, the gamma and digamma functions.

For ease of interpretation (8) is transformed to a unidirectional distance measure $d(\alpha_1, \alpha_2) = \sqrt{2\text{KLD}(\boldsymbol{p}||\boldsymbol{g})}$. 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 deviates 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 α_1 conditional on α_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 α_1 and α_2 and potentially one may consider two parameters λ_1 and λ_2 to penalize deviations along α_1, α_2 at different rates.

From Figure S2 of the supplementary material we can visually inspect the KLD in (8) and we see that it varies more sharply along α_1 than α_2. This suggests a simple solution here in the form of a conditional PC prior on $\alpha_1 \in (0, U]$, given α_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.

The PC prior on α_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:

$$
\pi(\alpha_1) = \frac{\lambda \exp\left(-\lambda d(\alpha_1, \alpha_2 = 10^{-5})\right) |d'(\alpha_1, \alpha_2 = 10^{-5})|}{1 - \exp(\lambda d(\alpha_1 = U, \alpha_2 = 10^{-5}))}, \quad \lambda > 0, \quad 0 < \alpha_1 \leq U \qquad (9)
$$

One appealing feature of (9) is that the user is only required to handle a single decay rate parameter λ, 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 λ in Eq. (9).

3.2.2. *Selection of λ.* 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. For example, in our biomechanics application, a clinician might favor three clusters to ease treatment formulation, while a biomechanicl researcher might be willing to consider a larger number of clusters but not so many as to overwhelm interpretability. We propose computing λ based on a user-defined probabilistic statement like $Pr(K^+ < U) = tp$. In other words, our aim is to help the user select the λ that corresponds to assigning a certain probability, denoted as tp, to the left-tail $(1, U-1)$. We use simulations to find the optimal λ that realizes a left-tail probability equal to the user-defined tp; the procedure is described in supplementary

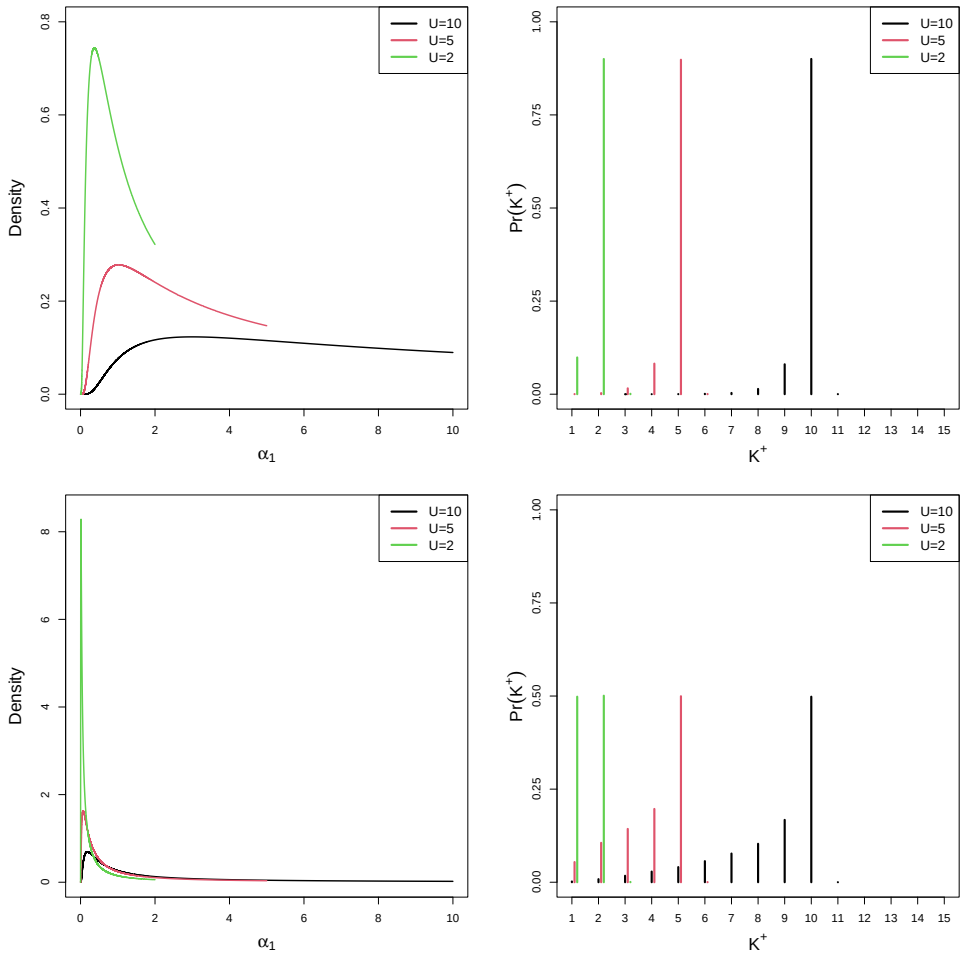

FIG 3. *PC prior on $\alpha_1|\alpha_2 = 1e-5$ (1st column panels) and the associated induced prior on K^+ (2nd column panels). The PC prior parameter λ is computed based on the user-defined statement $Pr(K^+ < U) = tp$, with $U = \{2, 5, 10\}$. Panels in the first row refer to $tp = 0.1$; panels in the second row refer to $tp = 0.5$.*

material. Figure 3 displays the prior on K^+ obtained by setting different values of the soft upper bound U and tail probability tp. Notice that as tp increases, there is more prior mass assigned to values of K^+ that are less than U which requires α_1 to be relatively small. In what follows we will use $PC(U, tp)$ to denote the PC prior just described in Section 3.2.1 where U and tp are user-supplied values satisfying $Pr(K^+ < U) = tp$.

As a final remark, an alternative PC prior approach would be to use an exponential on $d(\alpha_1, \alpha_2)$ with distinct decay rates λ_1 and λ_2. This would induce a prior on K^+ with mode at U and, at the same time, would enable the user to tune the probability mass assigned to $K^+ < U$ and $K^+ > U$ by careful selection of λ_1 and λ_2. This strategy may be useful in particular when expert users have solid prior information about K^+ and wish to control both tail probabilities, but the procedure to chose λ_1 and λ_2 would become computationally expensive. We argue that the computational burden quickly results in diminishing returns and a practical solution can be found by noting that increasing α_2 the right tail probability increases too, making U a softer upper bound. Thus, in principle one may still use the PC prior on α_1 but conditional on α_2 larger than 10^{-5} to distribute more mass above U and still enjoy the computational benefit of single λ selection.

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 3.4. If $U = 1$ and α_2 is fixed at a small value then as $\alpha_1 \to \infty$ the $Pr(K^+ = 1 \mid K, n, \alpha_1, \alpha_2) = 1$ resulting in an FMM with shrinkage properties similar to sFMM.

The proof of remark 3.4 follows from arguments similar to those found in the proof of Proposition 3.2.

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 α_2 is small); see Figure 3.2 to appreciate the impact the choice of tp has on the shape of the prior on K^+. Finally, setting $U = 0$ recovers the symmetric Dirichlet prior for $\boldsymbol{w}$ with α_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.

In practice, one might want to consider a homogeneous base model with no clusters. This can be tackled within the proposed approach by setting $U = 1$ and using the PC prior on α_1 conditional on $\alpha_2 = 10^{-5}$ in Eq. (8) with a large λ (that guarantees strong penalization for deviation from the base model); from our experiments $\lambda = 1$ is large enough.

3.4. *Computation.* Fitting an aFMM only requires a very straightforward Gibbs sampler if component densities and parameter prior distributions are conditionally conjugate. We provide more details for the Gaussian case in Section 4 of the online supplementary material. Here we only highlight how label switching is addressed. At each iteration of the MCMC algorithm we relabel components in the following way. The mixture component with the largest number of allocated units is labeled as component one, then the component with the second most allocated units is labeled as component 2, etc. Empty mixture components are labeled based on the order of their current labels.

4. Simulation Studies and Illustration. In this section we demonstrate the aFMM's performance in practice via numerical experiments. To more easily build intuition and illustrate the method we first conduct a simulation based a univariate version of (1) and also consider the well known galaxy data set. Then we conduct a simulation using a mulitvariate aFMM which is more in line with the biomechanic example in Section 5 (see equation (3)). Overall, in the univariate setting, the aFMM is competitive with regards to cluster estimation compared to existing methods in estimating K^+ and outperforms alternative methods when U is "well" specified. In the multivariate setting, U seems to have less impact, but the aFMM is at least as good at estimating the clustering as other reasonable competitors.

4.1. *Numerical Experiment Based on a Univariate aFMM.* Consider the following univariate version of an FMM in (1)

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

so that $\boldsymbol{\eta}_k = (\mu_k, \sigma_k^2)$. After introducing component labels we have

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

where the following commonly used prior distributions are employed

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

Here "Inverse-Gamma" denotes an inverse Gamma distribution parameterized so that the prior mean of σ_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, Malsiner-Walli and Frühwirth-Schnatter (2022). We also set $a_0 = 3$ and $b_0 = 2$ which is also very similar to one of the prior specifications in Grün, Malsiner-Walli and Frühwirth-Schnatter (2022). 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 generated 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 $K = 25$, $\alpha_1 = U$, $\alpha_2 = 10^{-3}$, $\mu_0 = 0$, $\sigma_0^2 = 3$, 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 S3 and S4 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 4 and 5. 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, Frühwirth-Schnatter and Grün (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) and a discrete uniform prior for $K \in \{1, \ldots, 25\}$,

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

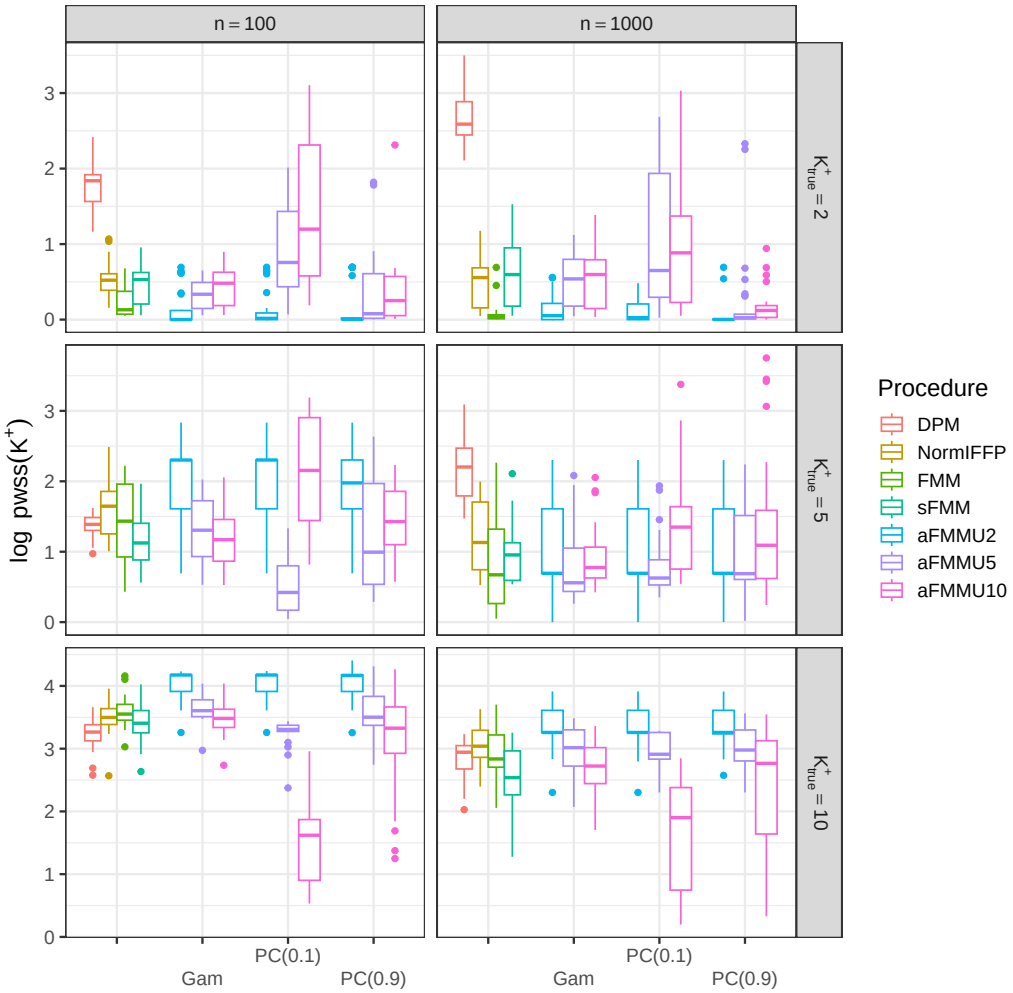

FIG 4. $\log(pwss(K^+)+1)$ *for Data Type 2. Each row corresponds to results associated with the value of* K^+ *used to generate data.*

suggested in Argiento and De Iorio (2022). The simulation was executed using GNU parallel (Tange 2022).

To compare each method's ability to estimate K^+ we recorded the bias associated with the posterior mode of K^+, the posterior mean adjust Rand index (ARI) (Hubert and Arabie 1985) between each MCMC iterate partition and the true partition, 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

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

where K^+_{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^+_{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

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

where $I[j \sim \ell] = 1$ if unit j and ℓ belong to the same cluster and zero otherwise. Small values of ccprob_error indicate more accurate estimation of the underlying partition and as a result, the number of clusters. In Figures 4 - 5 we report results for pwss(K^+) and ccprob_error under *Data Type 2*. The remaining simulation results are provided in Figures S5-S10 of the supplementary material. Similar conclusions can be drawn if we look at the bias associated with the posterior mode of K^+; results reported in Figures S7 and S8 for *Data Type 1* and *Data Type 2*, respectively.

From Figure 4 note that when $U = K^+_{\text{true}}$ the aFMM performs the best regardless of prior on α_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 (i.e.,"$PC(0.9)$") tends to perform well when $K^+_{\text{true}} < U$. As expected, setting U to a value that is far from K^+_{true} tends to result in poor performance for the aFMM. Trends for *Data Type 1* are similar (see Figure S5 of the supplementary material). Similar conclusions can be drawn from the ARI reported in Figures S9 and S10 for *Data Type 1* and *Data Type 2*, respectively.

Regarding Figure 5, 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 α_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 aFMM seems to estimate K^+ well if U is not far from the truth for small n and performs very competitively when n is large regardless of U.

4.2. *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, Malsiner-Walli and Frühwirth-Schnatter 2022). A possible reason for it's continued use is the uncertainty surrounding K^+. For example, 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, Malsiner-Walli and Frühwirth-Schnatter 2022). Due to the uncertainty associated with K^+ in these 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 S14. Notice that the posterior distribution of K^+ is influenced quite

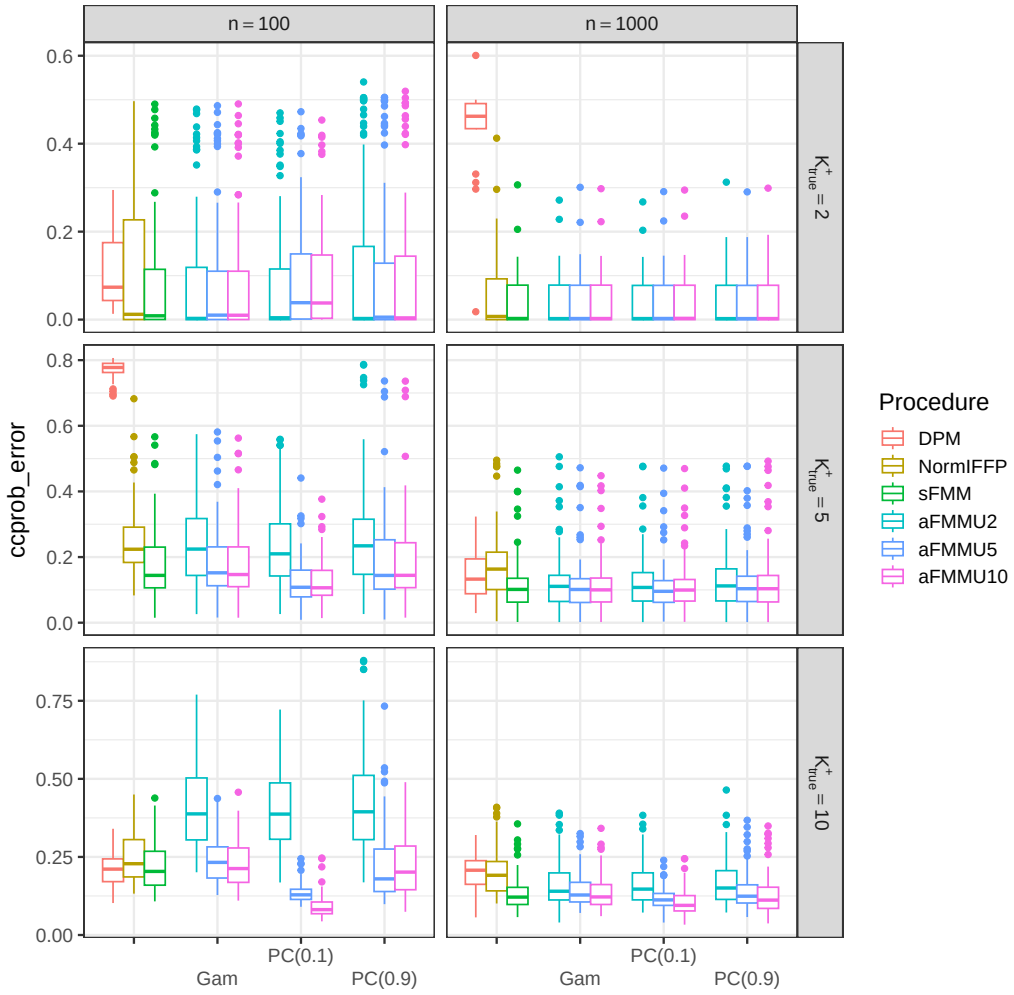

FIG 5. *Error of co-clustering probabilities for Data Type 2. Each row corresponds to results associated with the value of K^+ used to generate data.*

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 6. 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 S15 for density estimates). For $U \geq 5$ 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, Johnson and Müller 2022) results in 6 clusters for $U \geq 5$. 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 6 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 6).

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. To conclude this section, as noted in Grün, Malsiner-Walli and Frühwirth-Schnatter (2022) there is large uncertainty regarding the number of clusters in the galaxy data. An estimated K^+ value for these data can be selected depending on whether purity of cluster configuration or model fit is prioritized. Thanks to the sensitivity analysis conducted here, these two metrics can be balanced when estimating K^+.

4.3. *Numerical Experiment Based on vGRF Curves.* We now conduct a numerical experiment that is based on the vGRF curves displayed in Figure S16 of the online supplementary material. Synthetic data sets are generated using the vGRF curves in the following way. First, ignoring the functional nature of the curves, the 196 subjects were allocated to four groups using the `mclust` R-package. Then four sets of B-spline coefficients based on 30 evenly space knots were estimate using a simple linear model and the four group means. Next, subject-specific coefficients were created by adding random Gaussian noise to group-specific B-spline coefficients. Finally, subject-specific curves are created by adding Gaussian noise to the fitted values derived from the subject-specific B-spline coefficients and using 100 time points. R-code that carries out data set creation is provided as part of the online supplementary material and an example dataset created using the procedure just described is provided in Figure S11.

Using the procedure above, 100 datasets were generated and to each the functional model detailed in (2) - (5) was fit with the following priors for mixture weights:

PC(0.1): $\boldsymbol{w} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{1,2})$, 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): $\boldsymbol{w} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{1,2})$, 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 $Pr(K^+ < U) = 0.9$,

sFMM: $\boldsymbol{w} \sim \text{Dirichlet}(\alpha \boldsymbol{j})$ with $\alpha \sim Gamma(10, 10K)$ and $K = 25$ (see Malsiner-Walli, Frühwirth-Schnatter and Grün 2016),

FMM: $\boldsymbol{w} \sim \text{Dirichlet}(\alpha \boldsymbol{j})$ with $\alpha = 1/K$ and $K = 25$.

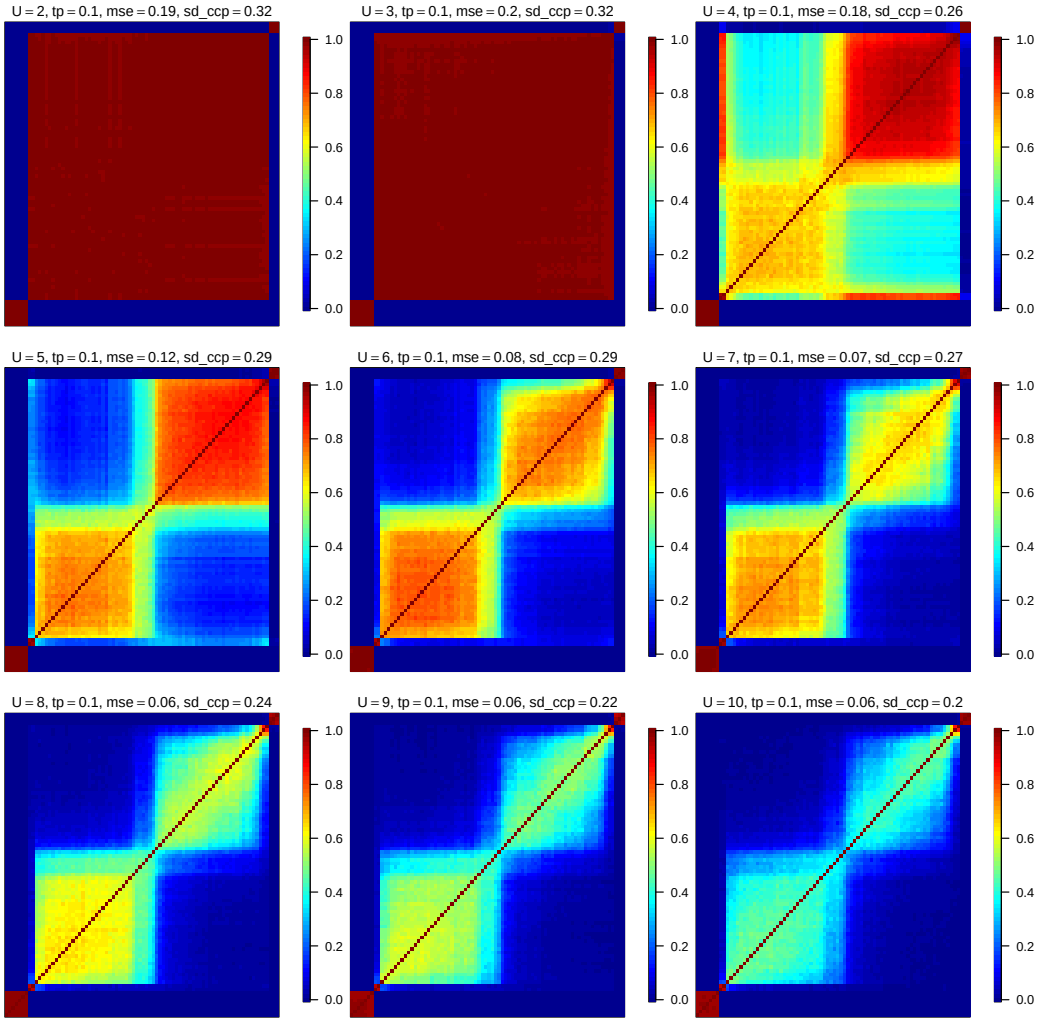

FIG 6. *Posterior co-clustering probabilities for the Galaxy data from an aFMM fit using* $U \in \{2, \ldots, 10\}$ *and* $tp = 0.1$. *In addition, sd_ccp as defined in* (15) *and mse as defined in* (14) *are provided. In middle row panels, choice of* $U \in \{5, 6, 7\}$ *balances best the quality of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

Each of these prior specifications are found on the x-axis in Figures 7 and 8 (and S12 and S13 in the online supplementary material). For both **PC(0.1)** and **PC(0.9)**, $U \in \{2, 4, 6, 8\}$ were considered (recall the the true number of clusters is four). To study how A_0 and p (B-spline coefficient vector dimension) impact the partition estimate, we considered $A_0 \in \{0.1, 0.5, 1.0\}$ and $p \in \{10, 13, 23\}$ (correspoding to 7, 10, and 20 inner knots respectively). Since this simulation scenario is based on functional data, software used to implement other competing methods in Section 4.1 are not available and as a result, they are not included in this numerical experiment.

As in Section 4.1, we considered four metrics to evaluate performance of the statistical procedures (pwss(K^+), ccprob_error, bias, and mean ARI). Results for the bias and mean ARI are provided in Figures S12 and S13 of the online supplementary material. All procedures perform similarly with regards to bias and ARI with $A_0 = 0.5$ and $p = 13$ producing the smallest bias among the aFMM procedures. This scenario also results in a slight improvement of the aFMM procedures relative to other competitors. Regarding ccprob_error,

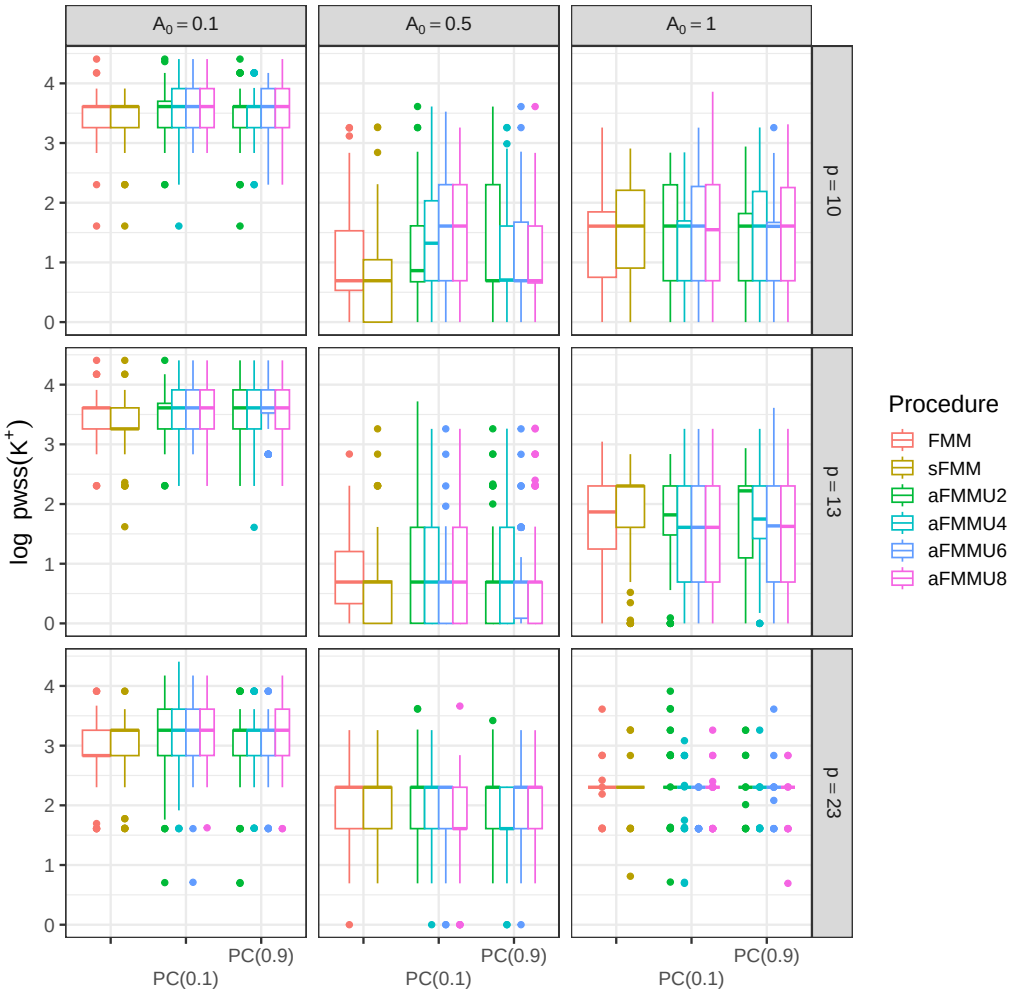

FIG 7. $\log(pwss(K^+)+1)$ *for functional data simulation.*

Figure 8 shows that the aFMM methods do at least as well as other methods for all scenarios. For pwss(K^+), the results are more mixed. Figure 7 shows that FMM and sFMM are quite competitive to the aFMM as only three scenarios result in an aFMM method that performs better than FMM and sFMM. These scenarios are $\{(A_0 = 0.5, p = 13), (A_0 = 1, p = 13), (A_0 = 0.5, p = 23)\}$. A take home message of this simulation is that overall the aFMM is very competitive to the FMM and sFMM in the functional model setting and in particular when $A_0 = 0.5$ and $p = 13$. Thus, we employ these values in Section 5.

5. Biomechanic Functional Data Application. Certain characteristics of the vGRF curve are associated with clinical outcomes for ACLR patients (Pietrosimone et al. 2019). For example, vGRF load rate during early ground contact is associated with knee joint articular cartilage thickness (Andriacchi et al. 2004). Amplitudes of the vGRF at the first peak (impact) and 50% of the ground contact phase (unloading) is associated with patient-reported outcomes of knee joint function such as quality of life, pain, and/or ability to return to sport (Pietrosimone et al. 2019). Additionally, the difference between the vGRF amplitude at peak impact and unloading is associated with joint biochemical markers believed to influence long-term cartilage health and well-being (Evans-Pickett et al. 2021). Thus, finding clusters based

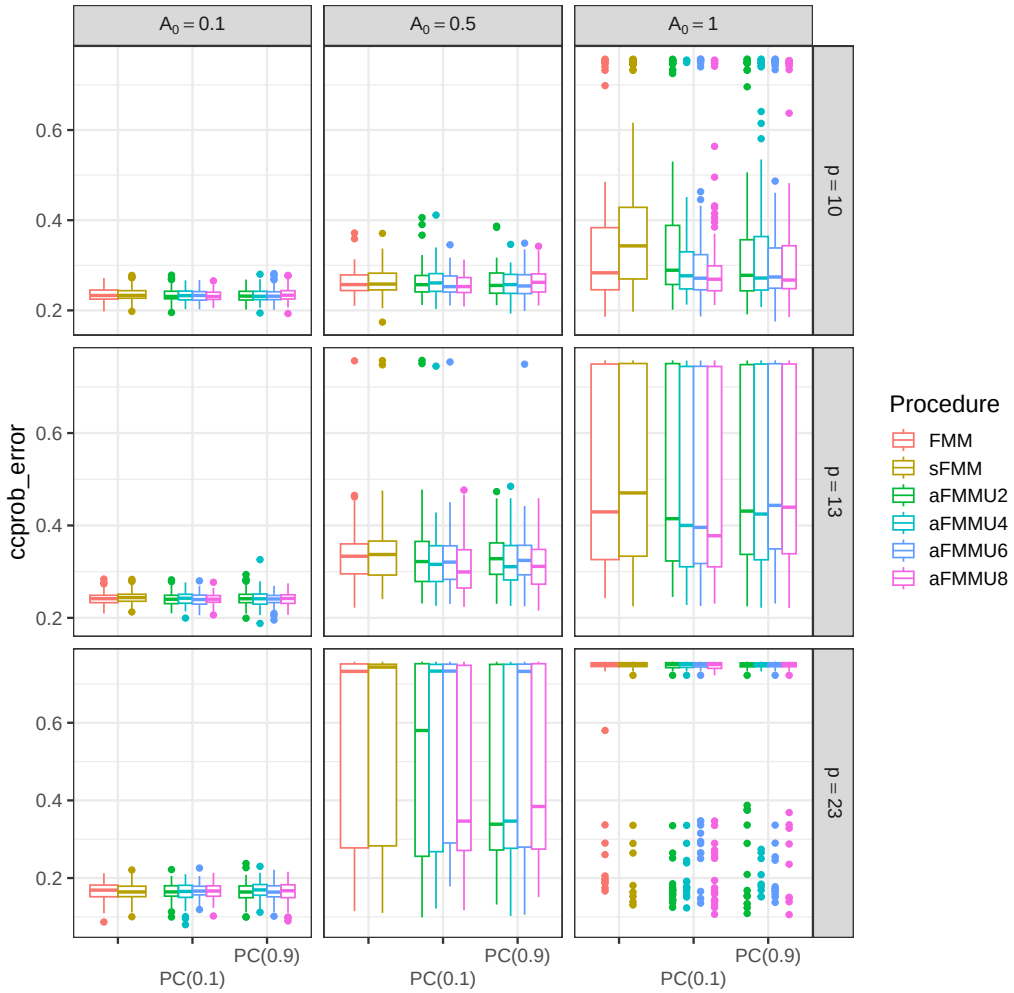

FIG 8. *Error of co-clustering probabilities for functional data simulation.*

upon individual vGRF curves is an important goal for biomechanical researchers and clinicians who ultimately wish to reveal new insights on which features of the vGRF patterns are most important to ACLR patient health and well-being.

As mentioned in Section 1, clinicians and biomechancial researchers may differ in their *a priori* desire/need of summarizing or interpreting information from clusters. For practical purposes, clinicians desire a smaller number of subpopulations in order to reduce the number of possible treatments and simplify clincial practice. Whereas, biomechanical researchers are more willing and able to summarize and interpret information from a larger number of clusters, due to fewer restraints associated with clinical practice. This necessitates that U be smaller for clinicians compared to biomechancial researchers. Thus, for a clinician, $U = 3$ would be a reasonable soft upper bound as three clusters would provide a natural number of treatments that could be effectively employed in a clinical setting due to technological and facility constraints. On the other hand, after discussions with bioemechanical researchers, we arrived at $U = 15$ as a reasonably soft upper bound on the number of clusters that could be interpreted for a researcher. Admittedly, there is some arbitrariness regarding the precise number of “interpretable" clusters. However, as the number of clusters approaches 10, interpretability quickly begins to decay which led us to select $U = 15$ as the maximum number

of clusters that can be reasonably considered. For the tail probability we use $tp = 0.1$. That said, we also fit the aFMM functional model using a variety of other tp values all resulting in very similar partition estimates. The aFMM model for $U = 3$ and $U = 15$ 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).

In contrast to the large amount of uncertainty associated with K^+ in the galaxy dataset, there is much less in the biomechanic application. In fact, for both $U = 3$ and $U = 15$ the posterior distributions of K^+ turned out to be points masses at $mode(K^+ \mid \boldsymbol{y}) = 5$. To visualize the clustering we provide Figure 9 referring to the aFMM with $U = 15$ (for $U = 3$ results are very similar and so are not reported here). The left column of Figure 9 displays the subject-specific curve fits with color indicating cluster membership and the right column displays the cluster-specific mean curves calculated cross-sectionally using all curves allocated to a particular cluster. (We estimated the partition for both aFMM configurations using the default settings of the `salso` function (Dahl, Johnson and Müller 2022) in R.) Note that the subject-specific fits (solid lines through points) are very reasonable for the majority of subjects.

To further explore the similarity between the two clustering configurations, we computed the adjusted rand index (ARI) (Hubert and Arabie 1985) between the partition estimates for both models. This turned out to be 0.98 which was unsurprising given the similarity between the two fits. That said, even though the posterior distribution of K^+ is the same between the two model fits, there are differences between the cluster configurations as can be seen through the co-clustering probabilities that are provided in Figure 10. The left plot is associated with aFMM fit with $U = 3$, the middle for $U = 15$ and the right is the difference between the two. The axis correspond to patients ordered based on their group membership. Notice that the clustering configurations, generally speaking are similar save the third cluster. For $U = 3$, there is more uncertainty on if this cluster should merge with the fourth cluster, while for $U = 15$, these two clusters are more clearly separated based on co-clustering probabilities between subjects in both clusters. Overall the co-clustering probabilities highlight the certainty that accompanies the clustering configuration; most co-clustering probabilities associated with subjects outside an assigned group are small and within group are close to 1.

In conclusion, the clustering configuration of vGRF curves is stable even if the prior on K^+ is varied drastically. The take-home message is that the prior is overwhelmed by the data and U has minimal impact in this application. This is a very informative result both from the clinician and biomechanical researcher's point of view and is only available due to the principled way prior information associated K^+ is included via the aFMM. Clinicians, who in this setting are typically capable of handling no more than 3 distinct treatments, will find it reassuring to know that patients are grouped in no more than 5 clusters. With this information clinicians can be sure that they do not lose much by focusing on only 3 patterns. Analogously, the biomechanical researchers will find it useful to know that they do not need to think of more than 5 distinct patterns. This knowledge along with cluster-specific features of the mean curves in Figure 9 will help in subsequent analysis as results are used to determine what features of the vGRF curve are associated to patient-reported outcomes post-ACLR. We would like to emphasize that the sensitivity analysis proposed here can be performed by simply moving U from small to large and that this approach is only available with our novel construction of the prior for K^+ based on the asymmetric Dirichlet.

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

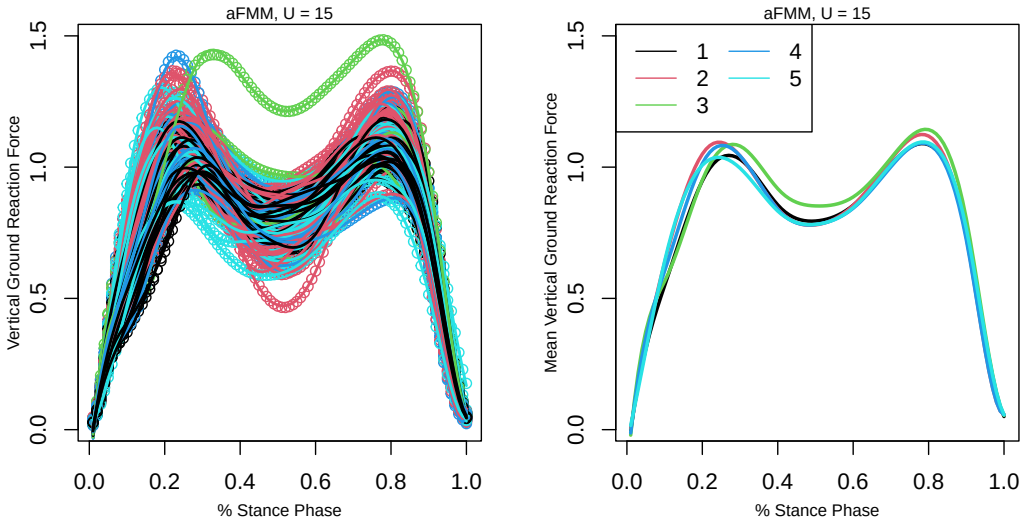

FIG 9. *aFMM functional model fits for $U = 15$. The left plot displays the curve fits for all 196 subjects with color indicating cluster (estimated using default settings of the* `salso` *function). The right plot corresponds to the cluster means which were calculated by finding cross-sectional mean among all curves in a cluster.*

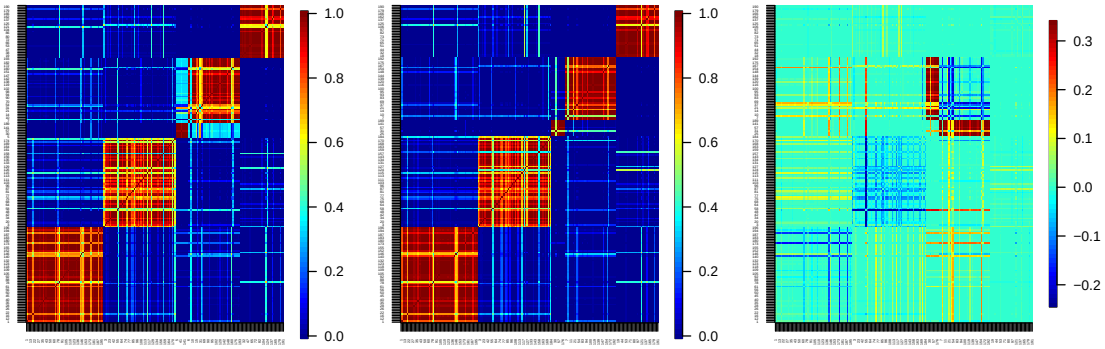

FIG 10. *Co-clustering probabilities for $U = 3$ (left),$U = 15$ (middle) and difference between the two (right).*

of the Dirichlet that permits eliciting prior information through intuitive statements that can be asked of the user.

An appealing property of our implied prior on K^+ is that it accommodates the user-selected value of K and the number of observations n in a natural way. There are two reasons why this is the case. The PC prior in Eq. (9) is derived by assuming an exponential 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, in addition to K, thus the number of observations would be automatically taken into account in the prior for K^+. We have experimented building the implied prior on K^+ starting from different choices of K (e.g. $10, 20, 30, 50$) and it resulted having no impact.

As with any new procedure, it is important for big picture purposes to outline the steps necessary to carry out a full analysis. We envision the workflow of the proposed approach to consist of the following steps.

1. Provide a clear definition of a cluster, which in our application implies the definition of meaningful prior quantities such between-subjects and within-subject parameters A and A_0.
2. Inform the aFMM through the choice of U and tp following the approach described in Section 3; this step requires the user to make use of prior information related to K^+ (e.g., clinician versus biomechanic information). Details on the derivation of the PC prior are in supplementary material Section 4.2.
3. Implement the model via MCMC and analyze MCMC convergence.
4. Sample from the posterior distribution of K^+ via MCMC and assume the posterior mode or median as an estimator for the number of clusters. Additionally, the entire cluster configuration can be explored via co-clustering probabilities or partition estimate using the `salso` `R`-package. Details on MCMC are in supplementary material Section 4.1.
5. Evaluate uncertainty on the clustering and in particular on the estimated K^+. Our methodology permits a principled study of the uncertainty associated with the clustering configuration. In this paper we have explored two ways to measure uncertainty on the estimated K^+. The first is through co-clustering probabilities with those that are more "pure" indicating a more certain clustering. The second is through a sensitivity analysis. The proposed asymmetric Dirichlet leads to naturally study stability of the clustering configuration and purity (i.e., Eq. (14) and (15), respectively) by measuring the impact of moving the parameter U from small to large. Essentially, a user can check sensitivity due to modifying their prior guess on K^+ from small (clinician perspective) to large (biomechanic perspective). If results exhibit uncertainty, then the data are not that informative regarding the number of clusters.

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

Acknowledgments. We would like to thank 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.

SUPPLEMENTARY MATERIAL

The supplementary material contains the proof of Proposition 1, further background and details concerning PC priors, computational details including the numerical evaluation of the PC prior for $\alpha_1|\alpha_2$, and additional results from the simulations and applications. All computer codes employed in the applications and simulation are also provided. These rely on the `miscPack` `R`-package that is available at `https://github.com/gpage2990`.

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. (2024). Bayesian mixture models (in)consistency for the number of clusters. *Scandinavian Journal of Statistics* **0** 0–0.

ANDRIACCHI, T. P., MÜNDERMANN, A., SMITH, R. L., ALEXANDER, E. J., DYRBY, C. O. and KOO, S. (2004). A Framework for the in Vivo Pathomechanics of Osteoarthritis at the Knee. *Annals of Biomedical Engineering* **32** 447-457.

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* **31** 422–435.

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* (M. MEILA and T. ZHANG, eds.). *Proceedings of Machine Learning Research* **139** 1158–1169. PMLR.

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.

DENNIS, H., EVANS-PICKETT, A., PAGE, G. L., PIETROSIMONE, B. G., BLACKBURN, J. T. and SEELEY, M. K. (2024). Cluster Analysis of Walking Biomechanics Post Anterior Cruciate Ligament Reconstruction Technical Report, Brigham Young University.

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

EVANS-PICKETT, A., LONGOBARDI, L., SPANG, J., CREIGHTON, R., KAMATH, G., DAVIS-WILSON, H., LOESER, R., BLACKBURN, T. and PIETROSIMONE, B. (2021). Synovial Fluid Concentrations of Matrix Metalloproteinase-3 and Interluekin-6 Following Anterior Cruciate Ligament Injury Associate with Gait Biomechanics 6 months Following Reconstruction. *Osteoarthritis and Cartilage* **29** 1006-1019.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite mixture and Markov switching models*. *Springer Series in Statistics*. Springer, New York.

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.

GRAZIAN, C. and ROBERT, C. P. (2018). Jeffreys priors for mixture estimation: Properties and alternatives. *Computational Statistics & Data Analysis* **121** 149-163.

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. (2022). Spying on the Prior of the Number of Data Clusters and the Partition Distribution in Bayesian cluster Analysis. *Australian & New Zealand Journal of Statistics* **64** 205-229.

GRÜN, B., MALSINER-WALLI, G. and FRÜHWIRTH-SCHNATTER, S. (2022). How many data clusters are in the Galaxy data set? Bayesian cluster analysis in action. *Advances in Data Analysis and Classification* **16** 325–349.

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* **119** 929–941.

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.

MCLACHLAN, G. and PEEL, D. (2005). *Finite Mixture Models*. John Wiley & Sons, Inc.

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

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.

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. and QUINTANA, F. A. (2015). Predictions Based on the Clustering of Heterogenous Functions via Shape and Subject-Specific Covariates. *Bayesian Analysis* **10** 379-410.

PETRALIA, F., RAO, V. and DUNSON, D. B. (2012). Repulsive Mixtures. In *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 1889-1897. Curran Associates, Inc.

PIETROSIMONE, B., SEELEY, M. K., JOHNSTON, C., PFEIFFER, S. J., SPANG, J. T. and BLACKBURN, J. T. (2019). Walking Ground Reaction Force Post-ACL Reconstruction: Analysis of Time and Symptoms. *Medicine & Science in Sports & Exercise* **51** 246-254.

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.

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.

SUN, H., ZHANG, B. and RAO, V. (2022). Bayesian Repulsive Mixture Modeling with Matérn Point Processes. arXiv:2210.04140. https://doi.org/10.48550/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. https://doi.org/10.5281/zenodo.6891516

VENTRUCCI, M. and RUE, H. (2016). Penalized complexity priors for degrees of freedom in Bayesian P-splines. *Statistical Modelling* **16** 429-453.

XIE, F. and XU, Y. (2020). Bayesian Repulsive Gaussian Mixture Model. *Journal of the American Statistical Association* **115** 187-203.