Bayesian Repulsive Gaussian Mixture Model

Fangzheng Xie* Yanxun Xu*†

Abstract

We develop a general class of Bayesian repulsive Gaussian mixture models that encourage well-separated clusters, aiming at reducing potentially redundant components produced by independent priors for locations (such as the Dirichlet process). The asymptotic results for the posterior distribution of the proposed models are derived, including posterior consistency and posterior contraction rate in the context of nonparametric density estimation. More importantly, we show that compared to the independent prior on the component centers, the repulsive prior introduces additional shrinkage effect on the tail probability of the posterior number of components, which serves as a measurement of the model complexity. In addition, a generalized urn model that allows a random number of components and correlated component centers is developed based on the exchangeable partition distribution, which gives rise to the corresponding blocked-collapsed Gibbs sampler for posterior inference. We evaluate the performance and demonstrate the advantages of the proposed methodology through extensive simulation studies and real data analysis.

Key Words: Blocked-Collapsed Gibbs Sampler, Density Estimation, Generalized Urn-Model, Model Complexity, Posterior Convergence

*Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, 21218
†Correspondence should be addressed to Yanxun Xu (yanxun.xu@jhu.edu)

1 Introduction

In Bayesian analysis of mixture models, independent priors on the component-specific parameters have been widely used because of their flexibility and technical convenience. A nonparametric example is the renowned Dirichlet process (DP) where the atoms in the stick-breaking representation are independent and identically distributed (i.i.d.) from a base distribution. One of the potential but non-negligible issues for such an approach is the presence of redundant components, especially when parsimony on the number of components is preferred. For example, when a mixture model is used in biomedical applications, each component of the mixture may be interpreted as clinically or biologically meaningful subpopulations (of patients, disease types, etc.). To address this challenge, in this paper we argue for a Bayesian approach for modeling repulsive mixtures as a competitive alternative, establish its posterior consistency and posterior contraction rate, and study the shrinkage effect on the posterior number of components in the presence of such a repulsion.

Mixture models have been extensively studied from both the frequentist and the Bayesian perspectives. Formally, given the parameter space Θ, a mixture model with a kernel density $\psi : \mathbb{R}^p \times \Theta \to \mathbb{R}_+$ and a mixing distribution $G \in \mathcal{M}(\Theta)$ can be represented as $\boldsymbol{y}_i \sim \int_\Theta \psi(\boldsymbol{y}, \boldsymbol{\theta}) \mathrm{d}G(\boldsymbol{\theta})$, where $\mathcal{M}(\Theta)$ is a class of probability distributions on Θ (equipped with an implicitly specified suitable σ-field). The most commonly used kernel density ψ is the normal density, which leads to the Gaussian mixture model (GMM). In particular, the GMM with a discrete (potentially infinitely supported) mixing distribution $G = \sum_k w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$ has been widely used for clustering, since an equivalent characterization is $\boldsymbol{y}_i \mid z_i \sim \mathrm{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, $\mathbb{P}(z_i = k) = w_k$, where z_i encodes the cluster membership of the corresponding observation $\boldsymbol{y}_i$. The parameters for each component $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, \cdots, K$, are referred to as the component-specific parameters. Throughout we use K to denote the (potentially infinite) number of components in a mixture model. When G is completely unknown, the GMM is referred to

as nonparametric GMM (Chen et al., 2017). Frequentists' ways of modeling mixture models require a finite and fixed K, the estimation of which could be accomplished using model selection approaches. Nonparametric Bayesian priors allow us to perform inference without *a priori* fixed and finite K. For example, the DP prior on G yields an exchangeable partition distribution on $(\boldsymbol{\theta}_{z_1}, \cdots, \boldsymbol{\theta}_{z_n})$, the inference of which indicates a distribution on the number of clusters among $(\boldsymbol{\theta}_{z_1}, \cdots, \boldsymbol{\theta}_{z_n})$. The development of Markov chain Monte Carlo sampling techniques (Ishwaran and James, 2001, 2002; Antoniak, 1974; MacEachern and Mueller, 1998; Neal, 2000; Walker, 2007) further popularized the DP mixture model in a wide array of applications, such as biomedicine, machine learning, pattern recognition, etc.

Meanwhile, the asymptotic results of the DP mixture of Gaussians as a method of nonparametric density estimation have been studied. The posterior consistency of the DP mixture of univariate Gaussians was established by Ghosal et al. (1999), and the posterior convergence rate in the context of density estimation in nonparametric GMM was studied by Ghosal and Van Der Vaart (2001). Posterior consistency in the multivariate setting (Wu and Ghosal, 2010) is harder due to the exponential growth of the L_1-entropy of sieves. Shen et al. (2013); Canale et al. (2017) derived the posterior contraction rates of general smooth densities for multivariate density estimation using the DP mixture of Gaussians.

Nevertheless, as shown in Xu et al. (2016), the DP mixture model typically produces relatively large number of clusters, some of which are typically redundant. Theoretically, Miller and Harrison (2013) showed that when the underlying data generating density is a finite mixture of Gaussians, the posterior number of clusters under the DP mixture model is not consistent for K. In other words, the posterior distribution of the number of clusters does not converge to the point mass at the underlying true K. Alternatively, finite mixture models with a prior on K, referred to as the mixture of finite mixtures (MFM) (Nobile, 1994; Miller and Harrison, 2018), was developed. The posterior inference of MFM can be carried out either by the reversible-jump Markov chain Monte Carlo (RJ-MCMC) (Green, 1995), or by

the collapsed Gibbs sampler derived via the exchangeable partition representation (Miller and Harrison, 2018). Meanwhile, the posterior asymptotics of MFM as a nonparametric density estimator, to the best of our knowledge, is restricted to the cases of univariate location-scale mixtures (Kruijer et al., 2010) and multivariate location mixtures (Shen et al., 2013), in which the priors on locations are assumed to be conditionally i.i.d. given K.

These approaches, however, assume independent prior on the component-specific parameters $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$. In the context of parametric inference, where the underlying data generating distribution is a finite mixture of Gaussians, repulsive priors (Petralia et al., 2012; Quinlan et al., 2017) and non-local priors (Fuquene et al., 2016) were developed as shrinkage methods to penalize mixture models with redundant components. In particular, theoretical properties regarding only univariate density estimations in parametric GMM (*i.e.*, assuming the ground true density is a finite mixture of Gaussians) were discussed in Petralia et al. (2012) and Quinlan et al. (2017). In addition, Xu et al. (2016) proposed repulsive mixtures via determinantal point process (DPP) with a prior on K, where the RJ-MCMC sampler for the posterior inference is potentially inefficient in high-dimensional setting.

In this paper, we propose a Bayesian repulsive Gaussian mixture (RGM) model. The main contributions of this paper are as follows. First, under certain mild regularity conditions, we establish the posterior consistency for density estimation in nonparametric GMM under the RGM prior, and obtain an "almost" parametric posterior contraction rate $(\log n)^t/\sqrt{n}$ for $t > p+1$. To the best of our knowledge, earlier work such as Ghosal and Van Der Vaart (2001), Petralia et al. (2012), and Quinlan et al. (2017), have not addressed the asymptotic analysis of repulsive mixture models for density estimation in nonparametric GMM. Ghosal and Van Der Vaart (2001) was the earliest work that discussed the posterior contraction rate for density estimation in nonparametric GMM, where the Dirichlet process (DP) prior is used. Petralia et al. (2012) and Quinlan et al. (2017) discussed the posterior contraction rate using repulsive priors, but under the parametric assumption that the mixing distribution

is finitely discrete. Second, the relationship between the posterior of K (*i.e.*, the number of components), which serves as a measurement of the model complexity, and the repulsive prior is studied as well. It turns out that compared to the independent prior on the component centers, the repulsive prior introduces additional shrinkage effect on the tail probability of the posterior of K under the nonparametric GMM assumption. Furthermore, we develop a generalized urn model for the proposed Bayesian RGM model based on the exchangeable partition distribution, generalizing the Pólya urn scheme for MFM with an independent prior on the component-specific parameters in Miller and Harrison (2018) to the case where correlation is allowed among the component centers. Such a generalized urn model not only serves as a guidance for designing a blocked-collapsed Gibbs sampler for posterior inference as an alternative to the RJ-MCMC sampler, but may also be of independent interest.

The remainder of the paper is organized as follows. In Section 2 we formulate the Bayesian repulsive Gaussian mixture model. Section 3 elaborates the theoretical properties of the posterior distribution. In particular, we establish the posterior consistency, investigate posterior contraction rate, and study the shrinkage effect on the posterior number of components in the presence of the repulsive prior. In Section 4 we develop the generalized urn model for the RGM model by integrating out the mixing weights and K, and use it as a guidance to design a blocked-collapsed Gibbs sampler for posterior inference. Section 5 demonstrates the advantages of the proposed methodology via simulation studies and real data analysis. We conclude the paper in Section 6.

2 Bayesian Repulsive Mixture Model

In this section we formulate the RGM model in a Bayesian framework. Suppose $\mathcal{S} \subset \mathbb{R}^{p\times p}$ is a collection of positive definite matrices, equipped with the Borel σ-field on $\mathcal{S}$. We consider

the Gaussian mixture model, a family of densities of the form

$$f_F(\boldsymbol{y}) = \int_{\mathbb{R}^p \times \mathcal{S}} \phi(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}F(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.1}$$

where $\phi(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \det(2\pi\boldsymbol{\Sigma})^{-1/2} \exp\left[-(1/2)(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right]$ is the density of the p-dimensional Gaussian distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and F is a distribution on $\mathbb{R}^p \times \mathcal{S}$. We shall also use the shorthand notation $\phi_{\boldsymbol{\Sigma}}(\boldsymbol{y}-\boldsymbol{\mu}) = \phi(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $f_F = \phi_{\boldsymbol{\Sigma}} * F$, where $*$ is the conventional notation for convolution of two functions. We assume that the data $(\boldsymbol{y}_i)_{i=1}^n$ are i.i.d. generated from some unknown density f_0, the estimation of which is of interest.

Denote the space of all probability distributions over $\mathbb{R}^p \times \mathcal{S}$ by $\mathcal{M}(\mathbb{R}^p \times \mathcal{S})$, and that over $\mathbb{R}^p$ by $\mathcal{M}(\mathbb{R}^p)$. We define a prior Π on f over the space of all density functions in $\mathbb{R}^p$ by the following hierarchical model:

$$\begin{aligned}
(f(\boldsymbol{y}) \mid F) &= \int_{\mathbb{R}^p \times \mathcal{S}} \phi(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}F(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\
\left(F \mid K, \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K\right) &= \sum_{k=1}^K w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \\
(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K) &\sim p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K), \\
(w_1, \cdots, w_K \mid K) \sim \mathcal{D}_K(\beta), &\qquad K \sim p_K(K), \quad K \in \mathbb{N}_+.
\end{aligned} \tag{2.2}$$

Here $p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K) > 0$ is some density function with respect to the Lebesgue measure on $(\mathbb{R}^p \times \mathcal{S})^K$, $\mathcal{D}_K(\beta)$ is the symmetric Dirichlet distribution over Δ^K with density function $p(w_1, \cdots, w_K) = \Gamma(K\beta)/\Gamma(\beta)^K \prod_{k=1}^K w_k^{\beta-1}$, where $\Delta^K = \{(w_1, \cdots, w_K)^{\mathrm{T}} : \sum_{k=1}^K w_k = 1, w_k \geq 0\}$ is the ℓ_1-simplex on $\mathbb{R}^K$. The prior on K that is supported on all positive integers is essential, as we allow the number of components to grow with the sample size in order to fit the data well.

Instead of assuming $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^K$ being i.i.d. from a "base measure", we introduce repulsion among components $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ through their centers $\boldsymbol{\mu}_k$, such that they are well

separated. We assume the density $p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K)$ is of the following form,

$$p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K \mid K) = \frac{1}{Z_K}\left[\prod_{k=1}^{K} p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_k) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_k)\right] h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K), \tag{2.3}$$

where $Z_K = \int \cdots \int_{\mathbb{R}^{p\times K}} h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)\left[\prod_{k=1}^{K} p(\boldsymbol{\mu}_k)\right] \mathrm{d}\boldsymbol{\mu}_1 \cdots \mathrm{d}\boldsymbol{\mu}_K$ is the normalizing constant, and the function $h_K : (\mathbb{R}^p)^K \to [0, 1]$ is invariant under permutation of its arguments: $h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = h_K(\boldsymbol{\mu}_{\mathfrak{T}(1)}, \cdots, \boldsymbol{\mu}_{\mathfrak{T}(K)})$ for any permutation $\mathfrak{T} : \{1, \cdots, K\} \to \{1, \cdots, K\}$. We require that h_K satisfies the following repulsive condition: $h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = 0$ if and only if $\boldsymbol{\mu}_k = \boldsymbol{\mu}_{k'}$ for some $k \neq k'$, $k, k' \in \{1, \cdots, K\}$. In this paper, we focus on the case where the repulsive property is introduced only through the mean vectors $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$, *i.e.*, we allow nonvanishing density even when distinct components share an identical covariance matrix. The case where repulsion is introduced through the covariance matrices is of independent interest and may be further explored.

We consider the following two classes of repulsive functions $h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$:

$$h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = \min_{1\leq k<k'\leq K} g(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|), \tag{2.4}$$

$$h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = \left[\prod_{1\leq k<k'\leq K} g\left(\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|\right)\right]^{1/K}, \tag{2.5}$$

for $K \geq 2$, and $h_1(\boldsymbol{\mu}_1) \equiv 1$, where $g : \mathbb{R}_+ \to [0, 1]$ is a strictly monotonically increasing function with $g(0) = 0$. Notice that the repulsive functions defined here generalize those in Petralia et al. (2012) and Quinlan et al. (2017), who fix K due to the challenges in estimating K caused by the complicated relation between Z_K and K. However, for the two repulsive functions (2.4) and (2.5), we are able to find the connection between Z_K and K in **Theorem 1**, the proof of which is deferred to the Supplementary Material. We will discuss the shrinkage behavior of the posterior distribution of K in Section 3.4.

Theorem 1. *Suppose the repulsive function h_K is either of the form (2.4) or (2.5). If*

$$\iint_{\mathbb{R}^p\times\mathbb{R}^p} \left[\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)\right]^2 p(\boldsymbol{\mu}_1) p(\boldsymbol{\mu}_2) \mathrm{d}\boldsymbol{\mu}_1 \mathrm{d}\boldsymbol{\mu}_2 < \infty,$$

then $0 \leq -\log Z_K \leq c_1 K$ *for some constant* $c_1 > 0$.

We refer to the prior Π on $f \in \mathcal{M}(\mathbb{R}^p)$ given by (2.2), (2.3), (2.4) or (2.5) as the Bayesian *repulsive Gaussian mixture* (RGM) model, denoted by $f \sim \mathrm{RGM}_1(\beta; g, p_{\boldsymbol{\mu}}, p_{\boldsymbol{\Sigma}}, p_K)$ if h_K is of the form (2.4), or $f \sim \mathrm{RGM}_2(\beta; g, p_{\boldsymbol{\mu}}, p_{\boldsymbol{\Sigma}}, p_K)$ if h_K is of the form (2.5).

3 Theoretical Properties of the Posterior Distribution

In this section we discuss the theoretical properties of the posterior of the RGM model defined in Section 2. In particular, in the context of density estimation in nonparametric GMM, we establish the posterior consistency, discuss the posterior contraction rate, and study the shrinkage effect on the tail probability of the posterior number of components introduced by the repulsive prior. We defer the proofs of all theorems, corollaries, propositions, and lemmas to the Supplementary Material.

3.1 Preliminaries and Notations

We begin with some useful notations. Given a positive definite matrix $\boldsymbol{\Sigma}$, we use $\lambda(\boldsymbol{\Sigma})$ to denote any eigenvalue of $\boldsymbol{\Sigma}$, and $\lambda_{\max}(\boldsymbol{\Sigma})$, $\lambda_{\min}(\boldsymbol{\Sigma})$ to denote the largest and smallest eigenvalue of $\boldsymbol{\Sigma}$, respectively. Denote $\boldsymbol{I}$ the identity matrix, and $\boldsymbol{I}_p \in \mathbb{R}^{p\times p}$ the identity matrix of size $p \times p$ if specifying the matrix dimension is needed. The Kullback-Leibler (KL) divergence between two densities f and g is denoted by $\mathrm{D}(f \mid\mid g) = \int f \log(f/g)$. Denote $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^p$. We use $\|\cdot\|_1$ to denote both the L_1-norm on $L^1(\mathbb{R}^p)$ and the ℓ_1-norm on $\mathbb{R}^d$ for any $d \geq 1$. $\|\cdot\|_\infty$ is used to denote both the ℓ_∞-norm of a vector and supremum norm of a bounded function. We use $\lfloor a \rfloor$ to denote the maximum integer that does not exceed a. The notation $a \lesssim b$ is used throughout to represent $a \leq cb$ for some constant c that is universal or unimportant for the analysis. Whenever possible, we use Π to represent the prior/posterior probability measure, $\mathbb{P}_0$ and $\mathbb{E}_0$ to denote the probability

and expectation with respect to the distribution f_0, and p to denote all density functions in the model except f_0, f, and $\{f_F : F \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})\}$. For random variables, we slightly abuse the notation and do not distinguish between the random variables themselves and their realizations. We shall also use $p(x)$ or $p_x(x)$ to denote the density of the random variable x.

A weak neighborhood of f_0 is a set of densities that contains

$$V = \left\{ f \in \mathcal{M}(\mathbb{R}^p) : \left| \int \varphi_i f_0 - \int \varphi_i f \right| < \epsilon, i = 1, \cdots, I \right\}$$

for some bounded continuous functions φ_i's in $\mathbb{R}^p$ (Ghosal et al., 1999). The posterior distribution is said to be *weakly consistent* at f_0, if $\Pi(f \in U \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) \to 1$ a.s. with respect to $\mathbb{P}_0$ for any weak neighborhood U of f_0. Given a prior Π on $\mathcal{M}(\mathbb{R}^p)$, a density function $f_0 \in \mathcal{M}(\mathbb{R}^p)$ is said to be *in the KL-support* of Π, or has the *KL-property* (with respect to Π), if $\Pi(f \in \mathcal{M}(\mathbb{R}^p) : \mathrm{D}(f_0 \,||\, f) < \epsilon) > 0$ for all $\epsilon > 0$. The posterior distribution is said to be L_1*(strongly) consistent* at f_0, if for all $\epsilon > 0$, $\Pi(f \in \mathcal{M}(\mathbb{R}^p) : \|f - f_0\|_1 > \epsilon \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) \to 0$ a.s. or in $\mathbb{P}_0$-probability as $n \to \infty$. The *posterior contraction rate* is any sequence $(\epsilon_n)_{n=1}^{\infty}$ such that $\Pi(f \in \mathcal{M}(\mathbb{R}^p) : \|f - f_0\|_1 > M\epsilon_n \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) \to 0$ in $\mathbb{P}_0$-probability as $n \to \infty$ for some constant $M > 0$. Given a family of densities $\mathcal{F}$ on $\mathbb{R}^p$ with a metric d on $\mathcal{F}$, the ϵ*-covering number* of $\mathcal{F}$ with respect to d, denoted by $\mathcal{N}(\epsilon, \mathcal{F}, d)$, is defined to be the minimum number of ϵ balls of the form $\{g \in \mathcal{F} : d(f, g) < \epsilon\}$ that are needed to cover $\mathcal{F}$. The d-metric entropy is the logarithm of the covering number under the d-metric.

Above all, we assume that $f \sim \mathrm{RGM}_r(\beta; g, p_{\boldsymbol{\mu}}, p_{\boldsymbol{\Sigma}}, p_K)$, $r = 1$ or 2. In order to develop the posterior convergence theory, we need some regularity conditions, most of which are typically satisfied in practice. We group these conditions into two categories. The first set of conditions are the requirements for the sampling model.

A0 The data generating density f_0 is of the form $f_0 = \phi_{\boldsymbol{\Sigma}} * F_0$ for some $F_0 \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})$ that has a sub-Gaussian tail: $\int_{\{\|\boldsymbol{\mu}\| \geq t\}} F_0(\mathrm{d}\boldsymbol{\mu}) \leq B_1 \exp(-b_1 t^2)$ for some $B_1, b_1 > 0$.

A1 For some $\delta > 0, c_2 > 0$, we have $g(x) \geq c_2\epsilon$ whenever $x \geq \epsilon$ and $\epsilon \in (0, \delta)$.

A2 g satisfies $\iint_{\mathbb{R}^p \times \mathbb{R}^p} [\log g(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|)]^2 p(\boldsymbol{\mu}_1) p(\boldsymbol{\mu}_2) \mathrm{d}\boldsymbol{\mu}_1 \mathrm{d}\boldsymbol{\mu}_2 < \infty$.

A3 For some $\underline{\sigma}^2, \overline{\sigma}^2 \in (0, +\infty)$, we have $\underline{\sigma}^2 \leq \inf_{\mathcal{S}} \lambda(\boldsymbol{\Sigma}) \leq \sup_{\mathcal{S}} \lambda(\boldsymbol{\Sigma}) \leq \overline{\sigma}^2$.

A4 For some (non-random) unitary $\boldsymbol{U} \in \mathbb{R}^{p \times p}$, $\boldsymbol{U}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{U}$ is diagonal for all $\boldsymbol{\Sigma} \in \mathcal{S}$.

Condition A2 guarantees that $1/Z_K$ does not grow super-exponentially in K by **Theorem 1**. Conditions A0 and A3 assume that both f_0 and f are of the nonparametric GMM form, hence guaranteeing that f_0 and f are not too "spiky" such that a faster rate of convergence is obtainable. Condition A4, the simultaneous diagonalizability of all $\boldsymbol{\Sigma} \in \mathcal{S}$, appears to be of less importance, but it turns out that a structured space $\mathcal{S}$ of covariance matrices decreases the $\|\cdot\|_1$-metric entropy of the proposed sieves in Section 3.2, and hence affects the posterior contraction rate. We assume that $\boldsymbol{U}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{U} = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$ for all $\boldsymbol{\Sigma} \in \mathcal{S}$, *i.e.*, the eigenvalues of $\boldsymbol{\Sigma} \in \mathcal{S}$ are aligned according to the orthonormal eigenvectors in $\boldsymbol{U}$.

We also need some requirements for the prior distributions.

B1 $(w_1, \cdots, w_K \mid K) \sim \mathcal{D}_K(\beta)$ is weakly informative: $\beta \in (0, 1]$.

B2 $p_{\boldsymbol{\mu}}$ has a sub-Gaussian tail: $\int_{\{\|\boldsymbol{\mu}\| \geq t\}} p(\boldsymbol{\mu}) \mathrm{d}\boldsymbol{\mu} \leq B_2 \exp(-b_2 t^2)$ for some $B_2, b_2 > 0$.

B3 For all $\boldsymbol{\mu} \in \mathbb{R}^p$, $p(\boldsymbol{\mu}) \geq B_3 \exp(-b_3 \|\boldsymbol{\mu}\|^{\alpha})$ for some $\alpha \geq 2, B_3, b_3 > 0$.

B4 $p(\boldsymbol{\Sigma})$ is induced by $\prod_{j=1}^{p} p_\lambda(\lambda_j(\boldsymbol{\Sigma}))$ with $\mathrm{supp}(p_\lambda) = [\underline{\sigma}^2, \overline{\sigma}^2]$.

B5 There exists some $B_4, b_4 > 0$ such that for sufficiently large K,

$$p_K(K) \geq \exp\left(-b_4 K \log K\right), \quad \sum_{N=K}^{\infty} p_K(N) \leq \exp\left(-B_4 K \log K\right).$$

Condition B1 assumes a vague prior on $(w_1, \cdots, w_K)$. Conditions B2 and B3 are requirements for the tail behavior of the function $p_{\boldsymbol{\mu}}$ in the sense that they are neither heavier than Gaussian nor thinner than an exponential power density (Scricciolo et al., 2011). Alternatively, one may assume $p(\boldsymbol{\mu}) \propto \exp(-b_3 \|\boldsymbol{\mu}\|^{\alpha})$ for some $b_3 > 0$, as suggested by Kruijer et al. (2010). Condition B4 is adopted in Ghosal and Van Der Vaart (2001) to obtain an "almost" parametric convergence rate. We will also discuss possible extensions to the case where p_λ

has full support on $(0, +\infty)$ later in this section. Condition B5 is the requirement for the tail behavior of the prior on K. Similar assumption on the tail behavior of the prior on K is adopted in Kruijer et al. (2010) and Shen et al. (2013) for finite mixture models. As a useful example, we show that the commonly used zero-truncated Poisson prior on K satisfies condition B5.

Example. The zero-truncated Poisson prior has a density function $p_K(K) = \lambda^K/[(\mathrm{e}^\lambda - 1)K!]$, $K = 1, 2, \cdots$ with respect to the counting measure on $\mathbb{N}_+$ for some intensity parameter $\lambda > 0$. Directly compute

$$\sum_{N=K+1}^{\infty} p_K(N) = \frac{1}{\mathrm{e}^\lambda - 1}\left(\mathrm{e}^\lambda - \sum_{N=0}^{K}\frac{\lambda^N}{N!}\right) = \frac{1}{\mathrm{e}^\lambda - 1}\int_0^\lambda \frac{(\lambda - t)^K \mathrm{e}^t \mathrm{d}t}{K!} \lesssim \frac{\lambda^{K+1}}{(K+1)!},$$

where the second equality is due to Taylor's expansion. By Stirling's formula, this is further upper bounded by $[(\lambda \mathrm{e})(K+1)]^{K+1}$. Therefore, substituting $K+1$ with K, we obtain

$$\sum_{N=K}^{\infty} p_K(N) \lesssim \exp\left(K\log(\lambda \mathrm{e}) - K\log K\right) \leq \exp\left(-\frac{1}{2}K\log K\right)$$

for sufficiently large K. The constant for $\lesssim$ can be absorbed into the exponent, and hence we conclude $\sum_{N=K}^{\infty} p_K(N) \leq \exp(-B_4 K \log K)$ for some $B_4 > 0$.

For the lower bound on $p(K)$, for sufficiently large K we again use Stirling's formula,

$$p(K) = \frac{1}{\mathrm{e}^\lambda - 1}\frac{\lambda^K}{K!} \geq \exp(K\log(\lambda \mathrm{e}) - \log K - K\log K) \geq \exp(-2K\log K).$$

Hence the zero-truncated Poisson prior on K satisfies condition B5.

3.2 Posterior Consistency

Weak consistency. Using the result from Schwartz (1965), a sufficient condition for Π to be weakly consistent at f_0 is that f_0 is in the KL-support of Π. The following lemma is useful in that it provides a compactly supported F_m such that f_{F_m} can approximate f_0 arbitrarily well in the KL divergence sense.

Lemma 1. *Assume conditions A0-A4 and B1-B5 hold. For all $m \in \mathbb{N}_+$, define a sequence of distributions $(F_m)_{m=1}^{\infty}$ by $F_m(A) = c_m F_0(A \cap \mathcal{T}_m)$ for any measurable $A \subset \mathbb{R}^p \times \mathcal{S}$, where*

$$\mathcal{T}_m = \left\{ (\boldsymbol{\mu} : \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\| \leq m, \underline{\sigma}^2 + \frac{1}{m} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq \overline{\sigma}^2 - \frac{1}{m} \right\}$$

and c_m is the normalizing constant for F_m with $c_m^{-1} = F_0(\mathcal{T}_m)$. Then $\int f_0 \log(f_0/f_{F_m}) \to 0$ as $m \to \infty$.

We remark that the construction in Wu et al. (2008) is not directly applicable. The major reason is that the scale parameter of the convolving ϕ is allowed to be arbitrarily close to 0 there, whereas we impose uniform boundedness on the eigenvalues of the covariance matrices. The sequence of densities constructed in Wu et al. (2008) is $(f_m(\boldsymbol{y}))_{m=1}^{\infty} = \left(\int_{\mathbb{R}^p} \phi_{\sigma_m^2 \boldsymbol{I}}(\boldsymbol{y} - \boldsymbol{\mu}) f_0(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}\right)_{m=1}^{\infty}$, where $(\sigma_m)_{m=1}^{\infty}$ is a sequence converging to 0 at a certain rate. This construction does not apply when covariance matrices are bounded in spectrum. The construction of the sequence of densities $(f_{F_m})_{m=1}^{\infty}$ in **Lemma 1** also serves as a technical contribution to the Kullback-Leibler property of Bayesian nonparametric GMM.

Based on **Lemma 1**, we are able to establish the weak consistency via the KL-property.

Theorem 2. *Assume conditions A0-A4 and B1-B5 hold. Then f_0 is in the KL-support of Π, and hence $\Pi(\cdot \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ is weakly consistent at f_0.*

Strong consistency. To establish the posterior strong consistency, we utilize Theorem 1 in Canale et al. (2017), which is a standard result for proving consistency for general Bayesian nonparametric density estimation methods (see the Supplementary Material). Specializing to the RGM model, we need to construct a sequence of submodels and partitions of each of these submodels that satisfy the conditions in Theorem 1 in Canale et al. (2017). We now make these statements precise. Consider the following submodels of $\mathcal{M}(\mathbb{R}^p)$:

$$\mathcal{F}_{K_n} = \left\{ f_F : F = \sum_{k=1}^{K} w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, K \leq K_n, \boldsymbol{\mu}_k \in \mathbb{R}^p, \boldsymbol{\Sigma}_k \in \mathcal{S} \right\}$$

and the following partition of the submodel $\mathcal{F}_{K_n}$

$$\mathcal{G}_K(\boldsymbol{a}_K) = \mathcal{F}_K\left(\prod_{k=1}^{K}(a_k, a_k+1]\right), \quad \boldsymbol{a}_K = (a_1, \cdots, a_K) \in \mathbb{N}^K, \quad K = 1, \cdots, K_n,$$

where

$$\mathcal{F}_K\left(\prod_{k=1}^{K}(a_k, b_k]\right) = \left\{ f_F : F = \sum_{k=1}^{K} w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \|\boldsymbol{\mu}_k\|_\infty \in (a_k, b_k] \right\}.$$

According to Theorem 1 in Canale et al. (2017), it suffices to show the following: f_0 is in the KL-support of Π, and there exists some $b, \tilde{b} > 0$, some sequence $(K_n)_{n=1}^{\infty}$, such that $\Pi(\mathcal{F}_{K_n}^c) \lesssim \mathrm{e}^{-bn}$ for sufficiently large n, and for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{e}^{-(4-\tilde{b})n\epsilon^2} \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}\left(\epsilon, \mathcal{G}_K(\boldsymbol{a}_K), \|\cdot\|_1\right)} \sqrt{\Pi\left(\mathcal{G}_K(\boldsymbol{a}_K)\right)} = 0. \tag{3.1}$$

Lemma 2. *Let $a_k < b_k$ be non-negative integers, $k = 1, \cdots, K$. Then for sufficiently small $\delta > 0$, there exists constant $c_3 > 0$ such that*

$$\mathcal{N}\left(\delta, \mathcal{F}_K\left(\prod_{k=1}^{K}(a_k, b_k]\right), \|\cdot\|_1\right) \le \left(\frac{c_3}{\delta^{2p+1}}\right)^K \left(\prod_{k=1}^{K} b_k\right)^p.$$

Lemma 3. *Assume conditions A0-A4 and B1-B5 hold. Then we have*

$$\sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}\left(\delta, \mathcal{G}_K(\boldsymbol{a}_K), \|\cdot\|_1\right)} \sqrt{\Pi\left(\mathcal{G}_K(\boldsymbol{a}_K)\right)} \le K_n \left(\frac{M}{\delta^{p+1/2}}\right)^{K_n}.$$

for sufficiently small δ for some constant $M > 0$.

Based on **Lemma 2** and **Lemma 3**, we are able to verify (3.1) and hence establish the strong consistency.

Theorem 3. *Assume conditions A0-A4 and B1-B5 hold. Then $\Pi(\cdot \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ is strongly consistent at f_0.*

3.3 Posterior Contraction Rate

To compute the posterior contraction rate, it is sufficient to find two sequences $(\underline{\epsilon}_n)_{n=1}^{\infty}, (\bar{\epsilon}_n)_{n=1}^{\infty}$ such that

$$\Pi\left(\mathcal{F}_n^c\right) \lesssim \exp(-4n\underline{\epsilon}_n^2), \tag{3.2}$$

$$\exp(-n\bar{\epsilon}_n^2) \sum_{K=1}^{K_n} \sum_{a_1=0}^{\infty} \cdots \sum_{a_K=0}^{\infty} \sqrt{\mathcal{N}\left(\bar{\epsilon}_n, \mathcal{G}_K(\boldsymbol{a}_K), \|\cdot\|_1\right)} \sqrt{\Pi\left(\mathcal{G}_K(\boldsymbol{a}_K)\right)} \to 0, \tag{3.3}$$

$$\Pi\left(f : \int f_0 \log \frac{f_0}{f} \leq \underline{\epsilon}_n^2, \int f_0 \left(\log \frac{f_0}{f}\right)^2 \leq \underline{\epsilon}_n^2\right) \geq \exp(-n\underline{\epsilon}_n^2). \tag{3.4}$$

(See Theorem 3 in Kruijer et al., 2010, which is also provided in the Supplementary Material). For notation convenience we refer to the set of densities

$$B(f_0, \epsilon) = \left\{f : \int f_0 \log \frac{f_0}{f} \leq \epsilon_n^2, \int f_0 \left(\log \frac{f_0}{f}\right)^2 \leq \epsilon^2\right\}$$

as the KL-type ball. Equation (3.4) is also known as the prior concentration condition.

Lemma 3 not only plays a fundamental role in establishing the posterior strong consistency, but also provides an upper bound for the sum in terms of δ, which is again used to verify equation (3.3). **Proposition 1** finds the rates $(\underline{\epsilon}_n)_{n=1}^{\infty}, (\bar{\epsilon}_n)_{n=1}^{\infty}$ that satisfy (3.2) and (3.3).

Proposition 1. *Assume conditions A0-A4 and B1-B5 hold. Let* $\underline{\epsilon}_n = (\log n)^{t_0}/\sqrt{n}$, $\bar{\epsilon}_n = (\log n)^t/\sqrt{n}$ *where* t *and* t_0 *satisfy* $t > t_0 + 1/2$, $t_0 > 0$, *and* $K_n = \lfloor (p+1)^{-1}(\log n)^{2t-1} \rfloor$. *Then* (3.2) *and* (3.3) *hold.*

We are now left with finding the prior concentration rate $(\underline{\epsilon}_n)_{n=1}^{\infty}$ that satisfies (3.4). In particular, we need to bound the KL-type balls $B(f_0, \epsilon)$ by the L_1 distance. The strategy is to approximate F_0 using a finitely discrete distribution with sufficiently small number of support points. **Lemma 4** allows us to formalize this idea.

Lemma 4. *Assume conditions A0-A4 and B1-B5 hold. For some constant $\eta > 0$ and for all sufficiently small $\epsilon > 0$, there exists a discrete distribution $F^\star = \sum_{k=1}^N w_k^\star \delta_{(\boldsymbol{\mu}_k^\star, \boldsymbol{\Sigma}_k^\star)}$ supported on a subset of $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu}\|_\infty \leq 2a\}$ with $a = b_1^{-1/2} [\log(1/\epsilon)]^{1/2}$ for some constant $b_1 > 0$, $\|\boldsymbol{\mu}_k^\star - \boldsymbol{\mu}_{k'}^\star\|_\infty \geq 2\epsilon$ whenever $k \neq k'$, $N \lesssim [\log(1/\epsilon)]^{2p}$, such that*

$$\left\{ f_F : F = \sum_{k=1}^N w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} : (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \in E_k, \sum_{k=1}^N |w_k - w_k^\star| < \epsilon \right\} \subset B\left(f_0, \eta \epsilon^{\frac{1}{2}} \left(\log \frac{1}{\epsilon} \right)^{1+5p/4} \right),$$

where

$$E_k = \left\{ (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathbb{R}^p \times \mathcal{S} : \|\boldsymbol{\mu} - \boldsymbol{\mu}_k^\star\|_\infty < \frac{\epsilon}{2}, |\lambda_j(\boldsymbol{\Sigma}) - \lambda_j(\boldsymbol{\Sigma}_k^\star)| < \frac{\epsilon}{2}, j = 1, \cdots, p \right\}.$$

We are in a position to derive the posterior contraction rates for the RGM model.

Theorem 4. *Assume conditions A0-A4 and B1-B5 hold. Then the posterior distribution $\Pi(\cdot \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) \to 0$ contracts at f_0 with rate $\epsilon_n = (\log n)^t / \sqrt{n}$, $t > p + (\alpha + 2)/4$.*

It is interesting that the RGM model and some other independent-prior models (*e.g.* DP mixtures of Gaussians) yield similar posterior contraction rate. The major complication for the RGM model comes from proving that the KL-type ball is assigned with sufficiently large prior probability, since in the RGM model the repulsive function h can only be lower bounded by 0, whereas h is always unity in independent-prior models.

Remark 1. Notice that the optimal rate $(\log n)^{(p+1)+}/\sqrt{n}$ is achieved when $\alpha = 2$, where $(p+1)+$ means that any $t > p+1$ is satisfied. Namely, the posterior contraction rate is optimal when $p_{\boldsymbol{\mu}}$ has a Gaussian tail. For comparison, recall that for general location-scale Gaussian mixture problem with bounded variance, Theorem 6.2 in Ghosal and Van Der Vaart (2001) gives a contraction rate of $(\log n)^{3.5}/\sqrt{n}$ in the univariate case ($p = 1$) using the DP mixture model, in which the distribution of the location parameters is Gaussian. Analogously, in the RGM model, we may use Gaussian $p_{\boldsymbol{\mu}}$ to control the tail rate of the joint distribution of $(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K)$ as $\|\boldsymbol{\mu}_k\|$ gets large, since the repulsive function h_K is bounded. **Theorem** 4 improves the contraction rate to $(\log n)^t/\sqrt{n}$ with $t > 2$ compared to that given

by Ghosal and Van Der Vaart (2001). However, such an improvement is not due to the repulsive structure of the prior. The underlying reason is that we use Theorem 3 in Kruijer et al. (2010) to derive the posterior contraction rate, whereas Ghosal and Van Der Vaart (2001) use Theorem 2.1 in Ghosal et al. (2000), a weaker version of Theorem 3 in Kruijer et al. (2010), to derive it. In other words, this suggests that it is also possible to obtain an improved posterior contraction rate for some Bayesian GMM with independent priors on component centers using Theorem 3 in Kruijer et al. (2010).

Remark 2. The boundedness on the eigenvalues of the covariance matrices (condition A3) was originally adopted in Ghosal and Van Der Vaart (2001), which is necessary to obtain an "almost" parametric rate $(\log n)^t/\sqrt{n}$ for some $t > 0$. Walker et al. (2007) adopted the same assumption and improved the posterior contraction rate of the location mixture problem. Requiring p_λ to have full support on $(0, +\infty)$, however, is necessary in cases where the underlying true density f_0 is no longer of the form $f_0 = \phi_{\boldsymbol{\Sigma}} * F_0$ for some $F_0 \in \mathcal{M}(\mathbb{R}^p \times \mathcal{S})$. For general mixtures of finite location mixture models, the contraction rate is known to be $(\log n)^t n^{-\tilde{\beta}/(2\tilde{\beta}+d)}$ for some $t > 0$, where f_0 is in a locally $\tilde{\beta}$-Hölder class (Shen et al., 2013). It will be interesting to extend **Theorem 4** to the case where $\text{supp}(p_\lambda) = (0, +\infty)$ and explore the corresponding posterior contraction rate.

3.4 Shrinkage Effect on the Posterior of K

The behavior of the posterior of K is of great interest, since it is a measurement of the complexity of a nonparametric density estimator. If a parametric assumption on f_0 is made in the sense that $f_0 = \phi_{\boldsymbol{\Sigma}} * F_0$ for some finitely discrete $F_0 \in \mathcal{M}(\mathbb{R} \times \mathcal{S})$, then under mild regularity condition, Nobile (1994) proved that the posterior distribution $p(K \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$ converges weakly to the point mass at K_0 a.s. under the MFM model, where K_0 is the number of support points of F_0. However, when F_0 is no longer assumed to be finitely discrete, and a repulsive prior is introduced among components in MFM, there is little result

concerning the mixture complexity in the literature. This issue is addressed in **Theorem 5** in terms of the shrinkage effect on the tail probability of the posterior of K in the presence of the repulsive prior.

We first consider the case where both f_0 and the model are of location-mixture form only. We also assume that the g function is of the form $g(x) = x(g_0 + x)$ for some $g_0 \in [0, \infty)$, $x > 0$. In particular, we allow $g_0 = 0$ so that the RGM model includes the special case of the independent-prior GMM.

Theorem 5. *Suppose* $f_0(\boldsymbol{y}) = \int_{\mathbb{R}} \phi_{\boldsymbol{\Sigma}_0}(\boldsymbol{y} - \boldsymbol{\mu}) F_0(\mathrm{d}\boldsymbol{\mu})$ *for some fixed* $\boldsymbol{\Sigma}_0 \in \mathcal{S}$ *and conditions A0-A3 and B1-B3 hold with* $\beta = 1$, $p(\boldsymbol{\mu}) = \phi(\boldsymbol{\mu} \mid \mathbf{0}, \tau^2 \boldsymbol{I})$, *and* $p_{\boldsymbol{\Sigma}} = \delta_{\boldsymbol{\Sigma}_0}$. *Assume that* F_0 *is compactly supported with* $\mathrm{supp}(F_0) \subset [-B, B]^p$, *and* $\int_{\mathbb{R}^p} \boldsymbol{m} F_0(\mathrm{d}\boldsymbol{m}) = \mathbf{0}$. *Let* $p(K) = \Omega Z_K \lambda^K / K!$ *where* $\Omega = \left[\sum_{K=1}^{\infty} Z_K \lambda^K / K!\right]^{-1}$, g *be of the form* $g(x) = x/(g_0 + x)$ *for some* $g_0 \geq 0$, $x > 0$, *and* $f \sim \mathrm{RGM}_r(1, g, \phi(\boldsymbol{\mu} \mid \mathbf{0}, \tau^2 \boldsymbol{I}), \delta_{\boldsymbol{\Sigma}_0}, p(K))$, *where* $r = 1$ *or* 2. *Then when* $N \geq 3$, *we have the following result for some constant* $\zeta > 0$*:*

$$\mathbb{E}_0\left[\Pi(K \geq N \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)\right] \leq C(\lambda) \exp\left[\frac{n^2\tau^2\xi}{2}\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1})\right] \chi_r(g_0; n, N) \sum_{K=N+1}^{\infty} \frac{\lambda^K}{K!},$$

where

$$\chi_r(g_0; n, N) = \begin{cases} \dfrac{\left(1 + g_0^{2/3}\delta(\tau)\right)^{3/2} \left[2p\tau^2 + (2n\tau^4/N)\mathbb{E}_0\left(\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{m}\right)\right]^{1/2}}{g_0 + \left[2p\tau^2 + (2n\tau^4 N)\mathbb{E}_0\left(\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{m}\right)\right]^{1/2}}, & \text{if } r = 1, \\ \dfrac{\left(1 + \delta(\tau)\sqrt{g_0}\right)\left[2p\tau^2 + (2n\tau^4/N)\mathbb{E}_0\left(\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{m}\right)\right]^{1/2}}{g_0 + \left[2p\tau^2 + (2n\tau^4 N)\mathbb{E}_0\left(\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{m}\right)\right]^{1/2}}, & \text{if } r = 2. \end{cases}$$

Here $C(\lambda)$ *are some constants depending on* λ *only,* $\delta(\tau)$ *is a constant depending on* τ *only such that* $\delta(\tau) < 1$ *for sufficiently large* τ, $\mathbb{E}_0\left(\boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{m}\right) := \int_{\mathbb{R}^p} \boldsymbol{m}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{m} F_0(\mathrm{d}\boldsymbol{m})$, *and* $\chi_r(g_0; n, N)$ *is referred to as the shrinkage constant.*

As pointed out in Section 2, the normalizing constant Z_K yields complication in the posterior inference of K. In **Theorem 5** the prior density $p(K)$ of the number of components is assumed to be proportional to the Poisson density function modulus Z_K to eliminate such

effect: $p(K) \propto Z_K \lambda^K / K!$. **Theorem 5** unveils the relationship between the tail probability of the marginal posterior of K and the hyperparameter g_0 that introduces repulsion: as long as τ is moderately large so that $\delta(\tau) < 1$ (corresponding to the weakly informative prior), the upper bound for $\mathbb{E}_0 [\Pi (K > N | \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)]$ decreases as g_0 increases when g_0 is large enough. In particular, the shrinkage constant $\chi_r(g_0; n, N)$ is 1 when $g_0 = 0$ (*i.e.*, no repulsion is enforced among component centers), decreases when g_0 increases, and is smaller than 1 for large enough g_0. Namely, compared to independent priors for the component centers $\boldsymbol{\mu}_k$'s, the repulsive prior introduces additional shrinkage effect on the tail probability of the posterior of K. Another intuition for the shrinkage effect on K comes from the prior distribution p_K. Theorem 1 shows that $\exp(-c_1 K) \leq Z_K \leq 1$ for some constant $c_1 > 0$. As a consequence, the tail distribution of $p_K(K) \propto Z_K \lambda^K / K!$ is not heavier than that of the zero-truncated Poisson prior $p_K(K) \propto \lambda^K / K!$, potentially encouraging smaller K *a priori* than the latter and causing shrinkage on K *a posteriori*. In addition, it is worth mentioning that **Theorem 5** is a non-asymptotic result.

Theorem 5 also serves as a guidance for constructing a sample-size dependent RGM prior that yields a slower growth rate of K compared to the independent-prior Gaussian mixture model. Specifically, instead of using a hyperparameter g_0 that does not change with n, it is possible to choose a sample-size dependent hyperparameter $g_0(n)$ that tends to infinity and thus affects the rate of decay of $\mathbb{E}_0 [\Pi(K \geq K_n | \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)]$ for certain sequences of $(K_n)_{n=1}^{\infty}$. However, the prior concentration condition might no longer hold, potentially resulting a slower posterior contraction rate. It might be interesting to explore the trade-off between the shrinkage effect on K and the posterior contraction rate using sample-size dependent repulsive prior.

Corollary 1. *Assume the conditions in* ***Theorem 5*** *hold. If the sequence* $(K_n)_{n=1}^{\infty} \subset \mathbb{N}_+$ *satisfies* $\liminf_{n \to \infty} K_n / n > 0$, *then the tail probability of the posterior distribution of* K *satisfies* $\Pi(K \geq K_n \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) \to 0$ *in* $\mathbb{P}_0$*-probability as* $n \to \infty$.

Remark 3. In terms of K, the number of support points in the RGM model, which is a measurement of the model complexity of estimating an unknown density, **Corollary 1** says that the posterior probability that K is at least a non-negligible fraction of n (in the limit) converges to 0 in $\mathbb{P}_0$-probability as $n \to \infty$. In other words, the posterior number of components grows sub-linearly with respect to the sample size.

4 A Generalized Urn Model and Posterior Inference

For MFM with an independent prior on component-specific parameters, Miller and Harrison (2018) derived a Pólya urn scheme that resembles the renowned Blackwell-MacQueen urn process for DP (Blackwell and MacQueen, 1973) based on the random partition distribution. In this section, we generalize the urn model developed in Miller and Harrison (2018) to the proposed RGM model by representing the RGM model using the random partition distribution. The corresponding generalized urn model not only is instructive for designing a blocked-collapsed Gibbs sampler for posterior inference, but may also be of independent interest for other Bayesian mixture models with a random number of components and correlated component-specific parameters.

Let us begin with characterizing the RGM model using the latent cluster configurations. Given a random measure $F = \sum_{k=1}^{K} w_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$ with $(w_1, \cdots, w_K) \sim \mathcal{D}_K(\beta)$, we may represent the finite mixture model as follow by integrating out $(w_1, \cdots, w_K)$:

$$\begin{aligned} (\boldsymbol{y}_i \mid z_i, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, K) &\sim \mathrm{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}), \\ p(z_1, \cdots, z_n \mid K) &= \frac{\Gamma(K\beta)}{\Gamma(n + K\beta)} \prod_{k=1}^{K} \frac{\Gamma(\beta + \sum_{i=1}^{n} \mathbb{I}(z_i = k))}{\Gamma(\beta)}. \end{aligned} \tag{4.1}$$

Let $\mathcal{C}_n$ denote the partition of $\{1, \cdots, n\}$ induced by $\boldsymbol{z} = (z_1, \cdots, z_n)$ as $\mathcal{C}_n = \{E_k : |E_k| > 0\}$, where $E_k = \{i : z_i = k\}$ for $k = 1, \cdots, K$, and $|E|$ denotes the cardinality of a finite set E. For example, if one has $\boldsymbol{z} = (z_1, z_2, z_3, z_4, z_5, z_6) = (1, 3, 4, 4, 3, 1)$ with $n = 6$, then

the corresponding partition is $\mathcal{C}_6 = \{\{1,6\},\{2,5\},\{3,4\}\}$. Using the exchangeable partition distribution in Miller and Harrison (2018), we establish the generalized urn model induced by the RGM model in **Theorem 6** after marginalizing out the intractable random distribution F. The proof is provided in the Supplementary Material.

Theorem 6. *Suppose the prior* Π *on* $\mathcal{M}(\mathbb{R}^p)$ *is defined as in Section 2, and the latent class configuration variables* $\boldsymbol{z} = (z_1, \cdots, z_n)$ *is defined as in (4.1). Let* $\boldsymbol{\gamma}_i = \boldsymbol{\mu}_{z_i}, \boldsymbol{\Gamma}_i = \boldsymbol{\Sigma}_{z_i}$, $\boldsymbol{\theta}_i = (\boldsymbol{\gamma}_i, \boldsymbol{\Gamma}_i)$, $i = 1, \cdots, n$, $\mathcal{C}_{n-1}$ *be the partition on* $\{1, \cdots, n-1\}$ *induced by* $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{n-1}$, $(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_{n-1})$ *be the unique values of* $(\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{n-1})$, *and* $(\boldsymbol{\Sigma}_c^\star : c \in \mathcal{C}_{n-1})$ *be those of* $(\boldsymbol{\Gamma}_1, \cdots, \boldsymbol{\Gamma}_{n-1})$. *Let* $\ell = |\mathcal{C}_{n-1}|$ *be the number of clusters, and* K *be the number of components in* F, *where* $K \geq \ell$. *Denote* $\mathcal{C}_\varnothing \subset \mathbb{N}_+$ *the indices for the components associated with no observations with* $|\mathcal{C}_\varnothing| = K - \ell$, $((\boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star) \in \mathbb{R}^p \times \mathcal{S} : c \in \mathcal{C}_\varnothing)$ *the component-specific parameters of the components that are not associated with any observation, and* $\underline{c} = \min(c : c \in C_\varnothing)$ *provided that* $K \geq \ell + 1$. *Denote* $\Pi(\boldsymbol{\theta}_n \in \cdot \mid -)$ *the full conditional distribution of* $\boldsymbol{\theta}_n$ *with* F *marginalized out. Define the following quantities:*

$$
\begin{aligned}
V_n(\ell) &= \sum_{K=\ell}^{\infty} \frac{K(K-1)\cdots(K-\ell+1)}{(\beta K)(\beta K+1)\cdots(\beta K+n-1)} p_K(K), \\
\alpha_K &= p(K \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \times \frac{p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid K, \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\})}{p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid \mathcal{C}_n = \mathcal{C}_n \cup \{\{n\}\})} \\
&\quad \times \left[\int \phi(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n) L_K(\boldsymbol{\gamma}_n) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_n) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_n) \mathrm{d}\boldsymbol{\theta}_n\right] \left[\int L_K(\boldsymbol{\gamma}_n) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_n) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_n) \mathrm{d}\boldsymbol{\theta}_n\right]^{-1}, \\
L_K(\boldsymbol{\gamma}_{\underline{c}}^\star) &= \int \cdots \int h_K(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\varnothing) \prod_{c \in \mathcal{C}_\varnothing, c \neq \underline{c}} p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_c^\star) \mathrm{d}\boldsymbol{\gamma}_c^\star, \\
G_K(A) &\propto \int_A \phi(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n) L_K(\boldsymbol{\gamma}_n) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_n) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_n) \mathrm{d}\boldsymbol{\theta}_n,
\end{aligned}
$$

where $h_K(\boldsymbol{\gamma}_c : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\varnothing) = h_K(\boldsymbol{\gamma}_{c_1}^\star, \cdots, \boldsymbol{\gamma}_{c_K}^\star)$ *if one labels* $\mathcal{C}_{n-1} \cup \mathcal{C}_\varnothing$ *as* $\{c_1, \cdots, c_K\}$, *and*

$$
\begin{aligned}
p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid \mathcal{C}_n) &= \sum_{K=|\mathcal{C}_n|}^{\infty} p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid K, \mathcal{C}_n) p(K \mid \mathcal{C}_n), \\
p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid K, \mathcal{C}_n) &\propto \prod_{c \in \mathcal{C}_{n-1}} p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_c^\star) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_c^\star) \int \cdots \int h_K(\boldsymbol{\mu}_c^\star : c \in \mathcal{C}_{n-1} \cup \mathcal{C}_\varnothing) \prod_{c \in \mathcal{C}_\varnothing} p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_c^\star) \mathrm{d}\boldsymbol{\gamma}_c^\star.
\end{aligned}
$$

Then the following generalized urn model holds:

$$\Pi(\boldsymbol{\theta}_n \in \cdot | -) \propto \sum_{c \in \mathcal{C}_{n-1}} (|c| + \beta) p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid \mathcal{C}_n = \mathcal{C}_{n-1} \backslash \{c\} \cup \{c \cup \{n\}\}) \phi(\boldsymbol{y}_n | \boldsymbol{\theta}_c^\star) \delta_{\boldsymbol{\theta}_c^\star}(\cdot)$$
$$+ \left[\frac{V_n(\ell+1)\beta}{V_n(\ell)} \right] p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{n-1} \mid \mathcal{C}_n = \mathcal{C}_{n-1} \cup \{\{n\}\}) \sum_{K=\ell+1}^{\infty} \alpha_K G_K(\cdot). \quad (4.2)$$

For the DPP mixture model, Xu et al. (2016) developed a variation of the RJ-MCMC sampler that can be extended to the RGM model. However, the reversible-jump moves there for multi-dimensional problems could be challenging and inefficient, due to the fact that the change of K (increase or decrease) per-iteration typically is no greater than 1 in classical RJ-MCMC samplers for mixture models (Richardson and Green, 1997). In contrast, thanks to **Theorem 6**, we are able to derive a blocked-collapsed Gibbs sampler for posterior inference of the proposed RGM model.

We follow the notation in **Theorem 6**. Let $\mathcal{C}_{-i}$ be the partition induced by $\boldsymbol{\theta}_{-i} := (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n) \backslash \{\boldsymbol{\theta}_i\}$, and $(\boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star : c \in \mathcal{C}_{-i})$ be the unique values of $\boldsymbol{\theta}_{-i}$. Notice that by exchangeability

$$\Pi(\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\} \mid \boldsymbol{y}_i, \boldsymbol{\theta}_{-i}, \mathcal{C}_{-i})$$
$$\propto \left[\frac{V_n(|\mathcal{C}_{-i}|+1)\beta}{V_n(|\mathcal{C}_{-i}|)} \right] p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}) \sum_{K=|\mathcal{C}_{-i}|+1}^{\infty} \alpha_K,$$
$$\Pi(\mathcal{C} = (\mathcal{C}_{-i} \backslash \{c\}) \cup \{c \cup \{i\}\} \mid \boldsymbol{y}_i, \boldsymbol{\theta}_{-i}, \mathcal{C}_{-i})$$
$$\propto (|c| + \beta)\, p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \backslash \{c\} \cup \{c \cup \{i\}\}) \phi(\boldsymbol{y}_i | \boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star), \quad (4.3)$$

where $c \in \mathcal{C}_{-i}$. Namely, given a partition $\mathcal{C}_{-i}$ on $\{1, \cdots, n\} \backslash \{i\}$, the left-out index i forms a new singleton cluster with probability proportional to

$$\left[\frac{V_n(|\mathcal{C}_{-i}|+1)\beta}{V_n(|\mathcal{C}_{-i}|)} \right] p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}) \sum_{K=|\mathcal{C}_{-i}|+1}^{\infty} \alpha_K,$$

and is merged into an existing cluster $c \in \mathcal{C}_{-i}$ with probability proportional to

$$(|c| + \beta)\, p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \backslash \{c\} \cup \{c \cup \{i\}\}) \phi(\boldsymbol{y}_i | \boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star).$$

Instead of directly sampling from the above categorical distribution, which involves computing the intractable α_K's, we take advantage of the integral structure of α_K and design auxiliary variables following the data augmentation technique in Neal (2000). Roughly speaking, when sampling from $p(x, y)$ via MCMC, one introduces an auxiliary variable z and samples $p(z \mid x, y)$, $p(y \mid x, z)$, and $p(x \mid z)$ alternately (collapsing). The auxiliary z is discarded after such an update.

Theorem 7. *Using above notations, define the following auxiliary distribution*

$$\widetilde{G}(A \mid \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i}) = \sum_{K=|\mathcal{C}_{-i}|+1}^{\infty} p(K \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}) \times \frac{p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid K, \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})}{p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})}$$
$$\times \left[\iint_A L_K(\boldsymbol{\gamma}_{\underline{c}}^\star) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^\star) p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^\star) \mathrm{d}\boldsymbol{\gamma}_{\underline{c}}^\star \mathrm{d}\boldsymbol{\Gamma}_{\underline{c}}^\star\right] \left[\int L_K(\boldsymbol{\gamma}_{\underline{c}}^\star) p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_{\underline{c}}^\star) \mathrm{d}\boldsymbol{\gamma}_{\underline{c}}^\star\right]^{-1},$$

where L_K is defined in ***Theorem 6****. Let $\widetilde{g}(\boldsymbol{\gamma}_{\underline{c}}^\star, \boldsymbol{\Gamma}_{\underline{c}}^\star | \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$ be the density of $\widetilde{G}(\cdot | \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$. Define the auxiliary variable $\boldsymbol{\theta}_{\underline{c}}^\star = (\boldsymbol{\gamma}_{\underline{c}}^\star, \boldsymbol{\Gamma}_{\underline{c}}^\star)$ with density*

$$p(\boldsymbol{\theta}_{\underline{c}}^\star \mid \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$$
$$\propto \left\{ \left[\frac{V_n(|\mathcal{C}_{-i}| + 1)\beta}{V_n(|\mathcal{C}_{-i}|)}\right] p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}) \phi(\boldsymbol{y}_i \mid \boldsymbol{\theta}_{\underline{c}}^\star) \right.$$
$$\left. + \sum_{c \in \mathcal{C}_{-i}} (|c| + \beta) p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \backslash \{c\} \cup \{c \cup \{i\}\}) \phi(\boldsymbol{y}_i \mid \boldsymbol{\theta}_c^\star) \right\} \widetilde{g}(\boldsymbol{\theta}_{\underline{c}}^\star \mid \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i}).$$

Given the auxiliary variable $\boldsymbol{\theta}_{\underline{c}}^\star$ and $(\boldsymbol{\theta}_{-i}, \mathcal{C}_{-i})$, suppose $\mathcal{C}$ and $\boldsymbol{\theta}_i$ are sampled as follows:

$$\mathbb{P}(\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\} \mid \boldsymbol{\theta}_{\underline{c}}^\star, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$$
$$\propto \left[\frac{V_n(|\mathcal{C}_{-i}| + 1)\beta}{V_n(|\mathcal{C}_{-i}|)}\right] p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}) \phi(\boldsymbol{y}_i \mid \boldsymbol{\theta}_{\underline{c}}^\star), \tag{4.4}$$

$$\mathbb{P}(\mathcal{C} = (\mathcal{C}_{-i} \backslash \{c\}) \cup \{c \cup \{i\}\} \mid \boldsymbol{\theta}_{\underline{c}}^\star, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$$
$$\propto (|c| + \beta)\, p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \backslash \{c\} \cup \{c \cup \{i\}\}) \phi(\boldsymbol{y}_i \mid \boldsymbol{\theta}_c^\star), \tag{4.5}$$

$$\mathbb{P}\left(\boldsymbol{\theta}_i \in A \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\}, \boldsymbol{\theta}_{\underline{c}}^\star, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i}\right) = \delta_{\boldsymbol{\theta}_{\underline{c}}^\star}(A),$$

$$\mathbb{P}\left(\boldsymbol{\theta}_i \in A \mid \mathcal{C} = (\mathcal{C}_{-i} \backslash \{c\}) \cup (\{c \cup \{i\}\}), \boldsymbol{\theta}_{\underline{c}}^\star, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i}\right) = \delta_{\boldsymbol{\theta}_c^\star}(A).$$

Then the marginal posterior $(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$ *with* $(\boldsymbol{\theta}^\star_{\underline{c}}, \mathcal{C} \mid \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$ *integrated out coincides with* (4.2)*, and the full conditional distribution of* $\boldsymbol{\theta}^\star_{\underline{c}}$ *is given by*

$$\begin{aligned}&\mathbb{P}(\boldsymbol{\theta}^\star_{\underline{c}} \in A \mid \boldsymbol{y}_i, \mathcal{C}, \boldsymbol{\theta}_{-i}, \boldsymbol{\theta}_i)\\ &\quad = \mathbb{I}(\mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})\delta_{\boldsymbol{\theta}_i}(A) + \mathbb{I}(\mathcal{C} \neq \mathcal{C}_{-i} \cup \{\{i\}\})\widetilde{G}(A \mid \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i}).\end{aligned} \tag{4.6}$$

The proof of **Theorem 7** is deferred to the Supplementary Material. Now we are in a position to introduce the blocked-collapsed Gibbs sampler for posterior inference. We remark that this Gibbs sampler can also be regarded as the generalization of the "Algorithm 8" in Neal (2000) to the case where a repulsive prior among component centers is introduced. The basic idea is to draw samples from $\mathbb{P}(\mathcal{C}_n \mid \boldsymbol{\theta}^\star_{\underline{c}}, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$, $\mathbb{P}(\boldsymbol{\theta}_i \mid \mathcal{C}_n, \boldsymbol{\theta}^\star_{\underline{c}}, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$, and $\mathbb{P}(\boldsymbol{\theta}^\star_{\underline{c}} \mid \mathcal{C}_n, \boldsymbol{\theta}_i, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$ alternately, where $\boldsymbol{\theta}^\star_{\underline{c}} = (\boldsymbol{\gamma}^\star_{\underline{c}}, \boldsymbol{\Gamma}^\star_{\underline{c}})$ is the auxiliary variable introduced in **Theorem 7**.

Algorithm. Suppose the current state of the Markov chain consists of $(\boldsymbol{\gamma}^\star_c, \boldsymbol{\Gamma}^\star_c : c \in \mathcal{C}_n)$, and a partition $\mathcal{C}_n$ on $\{1, \cdots, n\}$. We instantiate $(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n)$ using $(\boldsymbol{\gamma}^\star_c, \boldsymbol{\Gamma}^\star_c : c \in \mathcal{C}_n)$ and $\mathcal{C}_n$ by letting $\boldsymbol{\theta}_{z_i} = (\boldsymbol{\gamma}^\star_c, \boldsymbol{\Gamma}^\star_c)$ if $i \in c$. A complete iteration of the blocked-collapsed Gibbs sampler is desribed as below.

- **Step 1: For** $i = 1, \cdots, n$:
  1. **Sample auxiliary variable** $(\boldsymbol{\gamma}^\star_{\underline{c}}, \boldsymbol{\Gamma}^\star_{\underline{c}})$ **from** (4.6): If $\mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\}$, then set $(\boldsymbol{\gamma}^\star_{\underline{c}}, \boldsymbol{\Gamma}^\star_{\underline{c}}) = \boldsymbol{\theta}_i$; Otherwise sample $(\boldsymbol{\gamma}^\star_{\underline{c}}, \boldsymbol{\Gamma}^\star_{\underline{c}})$ from $\widetilde{G}(\cdot \mid \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$ as follows:
     i) Sample
     $$K \sim p(K \mid \mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\}) \times \frac{p(\boldsymbol{\theta}^\star_c : c \in \mathcal{C}_{-i} \mid K, \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})}{p(\boldsymbol{\theta}^\star_c : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})};$$
     Set $\ell = |\mathcal{C}_{-i}|$, compute $\mathcal{C}_\varnothing$ with $|\mathcal{C}_\varnothing| = K - \ell$, and set $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n) \backslash \{\boldsymbol{\theta}_i\}$.
     ii) Sample $\boldsymbol{\Gamma}^\star_{\underline{c}} \sim p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}^\star_{\underline{c}})$. Sample $(\boldsymbol{\gamma}^\star_c : c \in \mathcal{C}_\varnothing)$ by accept-reject sampling: Sample $(\boldsymbol{\gamma}^\star_c : c \in \mathcal{C}_\varnothing)$ independently from $p_{\boldsymbol{\mu}}$ and $U \sim \mathrm{Unif}(0, 1)$, independent of $(\boldsymbol{\gamma}^\star_c : c \in \mathcal{C}_\varnothing)$; If $U < h_K(\boldsymbol{\gamma}^\star_c : c \in \mathcal{C}_{-i} \cup \mathcal{C}_\varnothing)$, then accept the new

proposed samples; Otherwise resample $(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_\varnothing)$ from $p_{\boldsymbol{\mu}}$ and U until $U < h_K(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_{-i} \cup \mathcal{C}_\varnothing)$. Discard all $(\boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star : c \in \mathcal{C}_\varnothing \backslash \{\underline{c}\})$.

2. **Sample $\mathcal{C}_n$ from $p(\mathcal{C}_n|, \boldsymbol{\gamma}_{\underline{c}}^\star, \boldsymbol{\Gamma}_{\underline{c}}^\star, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$ according to (4.4) and (4.5):**

$$\Pi(\mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\} \mid -) \propto \left[\frac{V_n(|\mathcal{C}_{-i}|+1)\beta}{V_n(|\mathcal{C}_{-i}|)}\right] \phi(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_{\underline{c}}^\star, \boldsymbol{\Gamma}_{\underline{c}}^\star) \times p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})$$

$$\Pi(\mathcal{C}_n = (\mathcal{C}_{-i} \backslash \{c\}) \cup \{c \cup \{i\}\} \mid -) \propto (|c| + \beta)\, \phi(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star) \times p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \backslash \{c\} \cup \{c \cup \{i\}\}),$$

where $p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C})$ is given by **Theorem** 6.

3. **Assign $\boldsymbol{\theta}_i$ value according to $\mathbb{P}(\boldsymbol{\theta}_i \in \cdot \mid \mathcal{C}, \boldsymbol{\gamma}_{\underline{c}}^\star, \boldsymbol{\Gamma}_{\underline{c}}^\star, \boldsymbol{y}_i, \mathcal{C}_{-i}, \boldsymbol{\theta}_{-i})$.** Set $\boldsymbol{\theta}_i = (\boldsymbol{\gamma}_{\underline{c}}^\star, \boldsymbol{\Gamma}_{\underline{c}}^\star)$ if $\mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\}$, and set $\boldsymbol{\theta}_i = (\boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star)$ if $\mathcal{C}_n = (\mathcal{C}_{-i} \backslash \{c\}) \cup (\{c \cup \{i\}\})$ for some $c \in \mathcal{C}_{-i}$.

- **Step 2: Sample K from $p(K \mid \mathcal{C}_n, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n, \boldsymbol{\Gamma}_c^\star : c \in \mathcal{C}_n)$**; Set $\ell = |\mathcal{C}_n|$, and compute $\mathcal{C}_\varnothing$ such that $|\mathcal{C}_\varnothing| = K - \ell$.
- **Step 3: Sample $(\boldsymbol{\Gamma}_c^\star : c \in \mathcal{C}_n)$ from $p(\boldsymbol{\Gamma}_c^\star \mid \boldsymbol{y}_i : i \in c, \boldsymbol{\gamma}_c^\star, \mathcal{C}_n)$**: For all $c \in \mathcal{C}_n$, sample $\boldsymbol{\Gamma}_c^\star$ from

$$p(\boldsymbol{\Gamma}_c^\star \mid -) \propto p_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_c^\star) \prod_{i \in c} \phi(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_c^\star, \boldsymbol{\Sigma}_c^\star).$$

- **Step 4 (Blocking): Sample $(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_n)$ from $p(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_n \mid K, \boldsymbol{\Gamma}_c^\star, \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n, \mathcal{C}_n)$.** This can be done by accept-reject sampling: For each $c \in \mathcal{C}_n$, sample

$$p(\boldsymbol{\gamma}_c^\star \mid -) \propto p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_c^\star) \prod_{i \in c} \phi(\boldsymbol{y}_i \mid \boldsymbol{\gamma}_c^\star, \boldsymbol{\Gamma}_c^\star),$$

and for each $c \in \mathcal{C}_\varnothing$, sample $\boldsymbol{\gamma}_c^\star \sim p_{\boldsymbol{\mu}}(\boldsymbol{\gamma}_c^\star)$. Next independently sample $U \sim \text{Unif}(0,1)$; If $U < h_K(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_n \cup \mathcal{C}_\varnothing)$, then accept the new proposed samples; Otherwise resample $(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_n \cup \mathcal{C}_\varnothing)$ and U until $U < h_K(\boldsymbol{\gamma}_c^\star : c \in \mathcal{C}_n \cup \mathcal{C}_\varnothing)$.

- **Step 5:** Change the current state to $(\boldsymbol{\theta}_c^\star, c \in \mathcal{C}_n)$ and $\mathcal{C}_n$.

Remark 4. Technically speaking, the aforementioned blocked-collapsed Gibbs sampler is only approximate for posterior inference. It is worth noting that in theory, only **Step 1** is necessary to create a Markov chain with the stationary distribution being the full posterior distribution. The only approximation step occurring in **Step 1** is

$$K \sim p(K \mid \mathcal{C}_n = \mathcal{C}_{-i} \cup \{\{i\}\}) \times \frac{p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid K, \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})}{p(\boldsymbol{\theta}_c^\star : c \in \mathcal{C}_{-i} \mid \mathcal{C} = \mathcal{C}_{-i} \cup \{\{i\}\})},$$

which, in turn, depends on approximately drawing from $p(K \mid \mathcal{C}_n)$ as follows: By equation (3.7) in Miller and Harrison (2018), $p(K \mid \mathcal{C})/p(K = |\mathcal{C}| \mid \mathcal{C}) \approx 0$ for $K \geq |\mathcal{C}| + 2$, since it is expected that $K << n$ in practice. Thus the following approximate sampling scheme is applicable:

$$p(K = |\mathcal{C}| \mid \mathcal{C}) \propto |\mathcal{C}| + n + 1, \quad p(K = |\mathcal{C}| + 1 \mid \mathcal{C}) \propto |\mathcal{C}| + 1.$$

Nevertheless, such an urn-model-based sampler could potentially cause slow convergence (Neal, 2000). Although the resampling steps (**Step 2** through **Step 5**) involve approximation (see the Supplementary Material), introducing these steps can improve the performance of posterior inference in practice, as illustrated in Section 4.

Remark 5. The proposed sampler can be easiliy extended to the case where a non-Gaussian mixture model is used, provided that we use priors $p_{\boldsymbol{\mu}}, p_{\boldsymbol{\Sigma}}$ in (2.3) that are conjugate to the non-Gaussian kernel density. In such cases, it is also possible to extend the blocked-collapsed Gibbs sampler either by a method of "no-gaps" proposed by MacEachern and Mueller (1998) or a Metropolis-within-Gibbs sampler (Neal, 2000).

Remark 6. The above blocked-collapsed Gibbs sampler can be extended to other Bayesian mixture models with correlated priors. In particular, when the normalizing constant Z_K is in closed-form (*e.g.*, the non-local priors proposed in Fuquene et al., 2016), the numerical computation of intractable integrals involved are no longer required, greatly facilitating the use of the proposed blocked-collapsed Gibbs sampler in practice.

5 Numerical Examples

We evaluate the performance of the RGM model and the blocked-collapsed Gibbs sampler proposed in Section 4 through extensive simulation studies and real data analysis. Subsections 5.1 and 5.2 aim to illustrate the advantages of the RGM model concerning accurate density estimation, identification of correct number of components, and shrinkage effect on the model complexity. Subsection 5.3 demonstrates the performance of the proposed blocked-collapsed Gibbs sampler compared to those of the DP mixture model and the DPP mixture model (Xu et al., 2016). In Subsection 5.4 we apply the RGM model to analyze the Old Faithful geyser eruption data (Silverman, 1986). We assume $\beta = 1$, indicating a uniform prior on $(w_1, \cdots, w_K \mid K)$. We assign a zero-truncated Poisson prior on K with intensity $\lambda = 1$ (*i.e.*, $p(K) = 1/[(\mathrm{e}-1)K!]$, $K = 1, 2, \cdots$) for all numerical examples except the location-mixture problem in Section 5.2. The repulsive function is chosen to be $g(x) = x/(g_0 + x)$ for some $g_0 > 0$, and without loss of generality, we let h_K to be of the form (2.4). Lastly, we assume $p(\boldsymbol{\mu}) = \phi(\boldsymbol{\mu} \mid 0, \tau^2 \boldsymbol{I}_p)$ and a truncated inverse Gamma prior on $\lambda(\boldsymbol{\Sigma})$, $p(\lambda) \propto \mathbb{I}(\underline{\sigma}^2 \leq \lambda \leq \overline{\sigma}^2)\lambda^{-a_0-1}\exp(-b_0/\lambda)$ for some $a_0, b_0 > 0$.

We give the convergence diagnostics via trace plots and autocorrelation plots in the Supplementary Material. To compare the performance of the proposed models with the competitors (*e.g.* the DP mixture (DPM) model and the DPP mixture model), we follow the ideas in Pettit (1990) and compute the *logarithm of the conditional predictive ordinate* (log-CPO) of different models using the post-burn-in samples as follows:

$$\text{log-CPO} = -\sum_{i=1}^{n} \log \left[\frac{1}{n_{\text{mc}}} \sum_{i_{\text{it}}=1}^{n_{\text{mc}}} p(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_{\text{mc}}^{i_{\text{it}}}) \right],$$

where n_{mc} is the number of the post-burn-in MCMC samples, i_{it} indices the post-burn-in iterations, and $\boldsymbol{\Theta}_{\text{mc}}^{i_{\text{it}}}$ represents the post-burn-in samples of all parameters generated by the MCMC at the i_{it}th iteration.

5.1 Fitting Multi-modal Density: Finite Gaussian Mixtures

In this subsection, to demonstrate multi-modal density fitting, we fit a finite mixture of Gaussians using the RGM model, and evaluate its performance regarding density estimation and identification of the number of components. The simulation setup is as follows. Suppose the simulated data $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n$, $n = 1000$, are i.i.d. generated from the bivariate density:

$$f_0(\boldsymbol{y}) = 0.4\phi(\boldsymbol{y} \mid \boldsymbol{0}, \mathrm{diag}(2,1)) + 0.3\phi(\boldsymbol{y} \mid (-6,-6)^{\mathrm{T}}, 3\boldsymbol{I}_2) + 0.3\phi(\boldsymbol{y} \mid (6,6)^{\mathrm{T}}, 2\boldsymbol{I}_2).$$

We implement the proposed blocked-collapsed Gibbs sampler with $g_0 = 10$, $\tau = 10$, $m = 2, \underline{\sigma} = 0.1, \overline{\sigma} = 10$, and a total number of 2000 iterations with the first 1000 iterations discarded as burn-in. For comparison, we consider the following DPM model,

$$(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \sim N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}), \quad (\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} \mid G) \overset{\text{i.i.d.}}{\sim} G, \quad \text{and } (G \mid \alpha, G_0) \sim \mathrm{DP}(\alpha, G_0),$$

where $G_0 = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \sim \mathrm{N}(\boldsymbol{m}_1, \boldsymbol{\Sigma}/k_0)$, $\boldsymbol{\Sigma} \sim$ Inv-Wishart$(4, \boldsymbol{\Psi}_1)$, $\alpha \sim$ Gamma$(1,1)$, $\boldsymbol{m}_1 \sim \mathrm{N}(\boldsymbol{0}, 2\boldsymbol{I}_2)$, $k_0 \sim$ Gamma$(0.5, 0.5)$, and $\boldsymbol{\Psi}_1 \sim$ Inv-Wishart$(4, 0.5\boldsymbol{I}_2)$. For the DP mixture model, we use K to denote the number of clusters throughout this section, since the number of components is always infinity.

Table 1 shows that the log-CPO of the RGM model is higher than that of the DPM model, indicating that RGM is preferred according to the data. Figures 1a and 1c show the contour plots of the posterior predictive densities under the RGM model and the DP mixture model, respectively, indicating that both methods perform well in terms of density estimation.

Table 1: Log-Conditional Predictive Ordinate (log-CPO) for Numerical Results

Model	Subsection 5.1	Subsection 5.2	Subsection 5.4
RGM model	-3593.586	-3060.687	-227.236
DPM model	-4599.204	-3483.667	-315.103
DPP mixture model			-512.6564

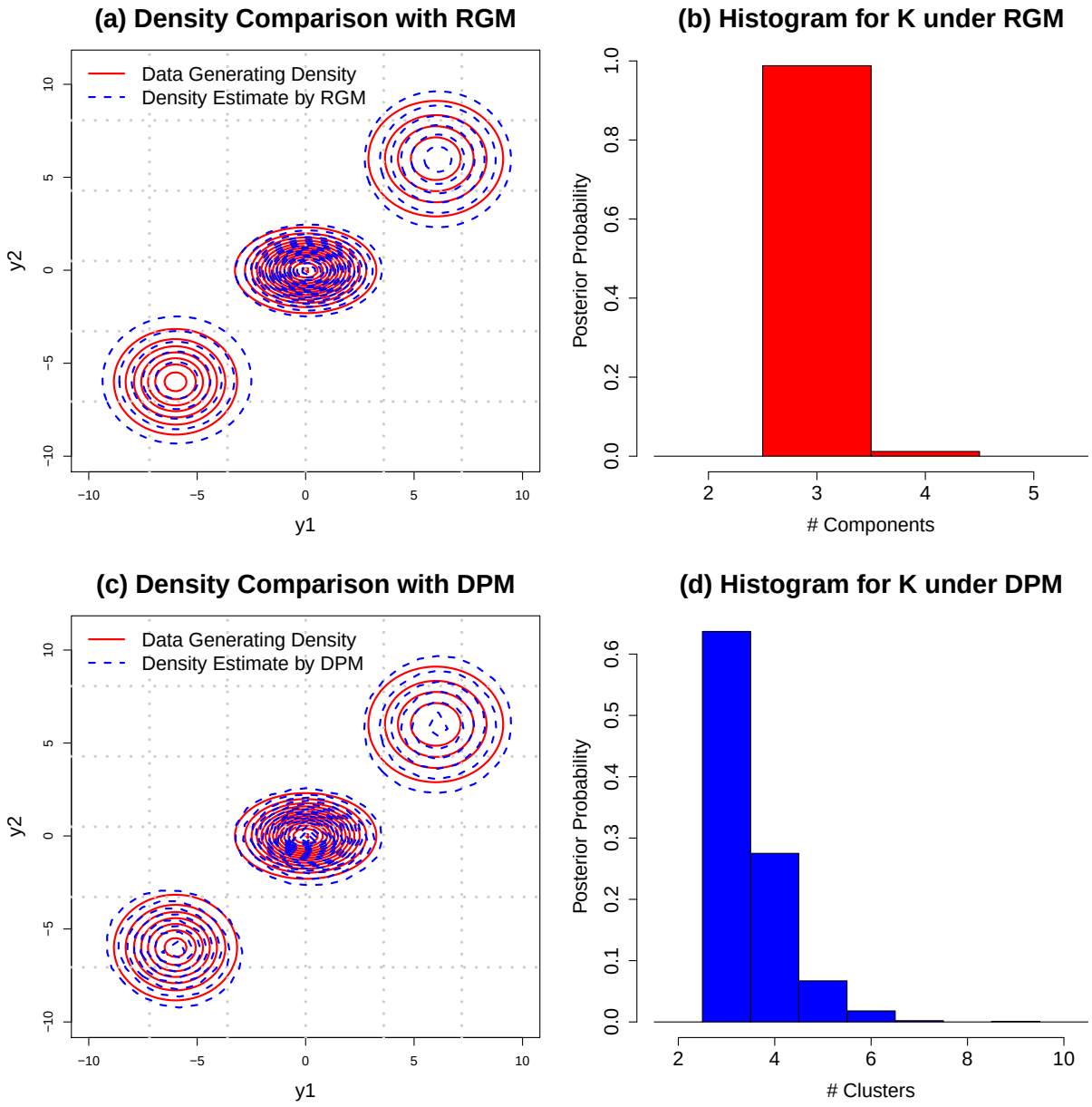

Figure 1: Fiting Multi-modal Density: Panels (a) and (c) are the contour plots for the posterior density estimation of the RGM model and the DPM model, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the RGM model and the posterior number of clusters under the DPM model, respectively, where the underlying true number of components is $K = 3$.

However, as shown in the histograms of the posterior numbers of components/clusters in Figures 1b and 1d, the posterior distribution of the number of components is highly concentrated around the underlying true K under the RGM model, whereas the DPM model assigns relatively higher posterior probability to larger K, introducing redundant clusters. This agrees with the inconsistency phenomenon of the DPM model in terms of identifying the number of components, which is reported in Miller and Harrison (2013).

5.2 Fitting Uni-modal Density: Continuous Gaussian Mixtures

Besides generating the simulated data from a finite discrete Gaussian mixture model, in this subsection we consider a continuous mixture of Gaussians,

$$f_0(y_1, y_2) = \prod_{j=1}^{2} \int_0^{\infty} \phi(y_i - \mu_i - \mu_0 \mid 0, 1) \exp(-\mu_i) \mathrm{d}\mu_i. \tag{5.1}$$

Notice that f_0 is uni-modal. The random variables y_i, $i = 1, 2$ can be i.i.d. generated as the sum of a normal random variable and an exponential random variable with intensity parameter 1, *i.e.*, $y_i = z_i + \mu_i$ where $z_i \sim \mathrm{N}(\mu_0, 1)$ and $\mu_i \sim \mathrm{Exp}(1)$, $i = 1, 2$. Then $\boldsymbol{y} = (y_1, y_2)$ is the random vector following the distribution in (5.1). The marginal distribution of y_i is referred to as the *exponentially modified Gaussian* (EMG) distribution, the density of which can be alternatively represented as $f(y) = (1/2) \exp\left(\mu_0 - y + 1/2\right) \mathrm{erfc}\left((\mu_0 + 1 - y)/\sqrt{2}\right)$, where erfc is the well-known complementary error function $\mathrm{erfc}(x) = (2/\sqrt{\pi}) \int_x^{\infty} \exp(-t^2) \mathrm{d}t$. We generate $n = 1000$ i.i.d. samples from f_0 with $\mu_0 = -4$, and implement the proposed blocked-collapsed Gibbs sampler with $g_0 = 7$, $\tau = 10$, $m = 2, \underline{\sigma} = 0.1, \overline{\sigma} = 10$, and a total number of 2000 iterations with the first 1000 iterations discarded as burn-in. For comparison, we consider the DPM model with the same setting as in Susbsection 5.1.

Figures 2a and 2c show that the RGM model and the DPM model provide similarly accurate density estimation of the underlying true density f_0. However, Figures 2b and 2d indicate that under the DPM model, the number of active components tends be larger than

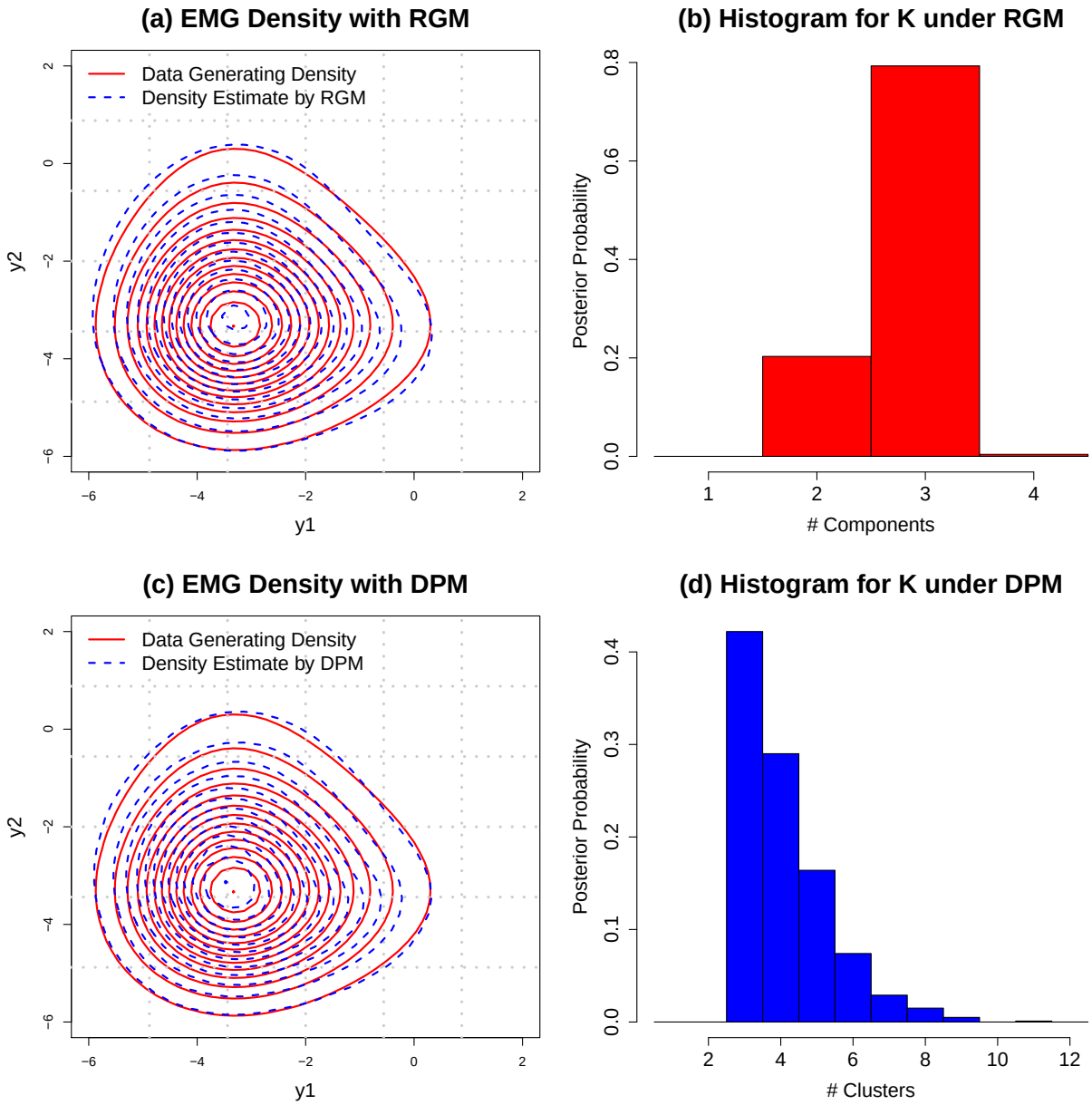

Figure 2: Fitting Uni-modal Density: Panels (a) and (c) are the contour plots for the posterior density estimation under the RGM model and the DPM model, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the RGM model and the posterior number of clusters under the DPM model, respectively.

that under the RGM model in order to fit the data well. In other words, the posterior of the RGM model provides the same level of accuracy in density estimation as the DPM model does, but with less number of components. With high posterior probability, the RGM model only utilizes 3 components to fit the density, whereas the DPM model assigns large posterior probability to utilizing 4 or more components. The log-CPO comparison in Table 1, clearly shows that the RGM model outperforms the DPM model.

To demonstrate the parsimony effect on the number K of necessary components to fit the density well, we perform comparison between the RGM and the independent-prior MFM. Suggested by **Theorem 5**, we consider location-mixture problem here. That is, the covariance matrices for all components under both RGM and MFM are fixed at $\boldsymbol{\Sigma}_k = \boldsymbol{I}_2$, $k = 1, \cdots, K$. We use the prior $p(K) \propto Z_K/K!$, $K = 1, 2, \cdots$, for the RGM, and $p(K) \propto 1/K!$, $K = 1, 2, \cdots$, for MFM. We implement the proposed blocked-collapsed Gibbs sampler with $\tau = 10$, $m = 2$, $g_0 = 7$ for location-RGM, $g_0 = 0$ for MFM, and a total number of 2000 iterations with the first 1000 iterations discarded as burn-in.

Since the data generating density is a continuous mixture of Gaussians, there is no "ground true" K. We evaluate the two methods in terms of the posterior of K and the log-CPO values. Figures 3a and 3c show that location-RGM and MFM provide similarly accurate density estimation of the underlying true density f_0 and yield similar log-CPO. Nevertheless, it can be seen from Figures 3b and 3d that the MFM model assigns larger number components than location-RGM. This phenomenon also numerically verifies **Theorem 5**: overall, compared to the independent prior ($g_0 = 0$), the posterior number of components K under the repulsive prior ($g_0 > 0$) tends to be smaller. We also observe that both location-RGM and MFM provide similar performance in terms of the density estimation, measured by the log-CPO (-3559.83 and -3575.346 under location-RGM and MFM, respectively).

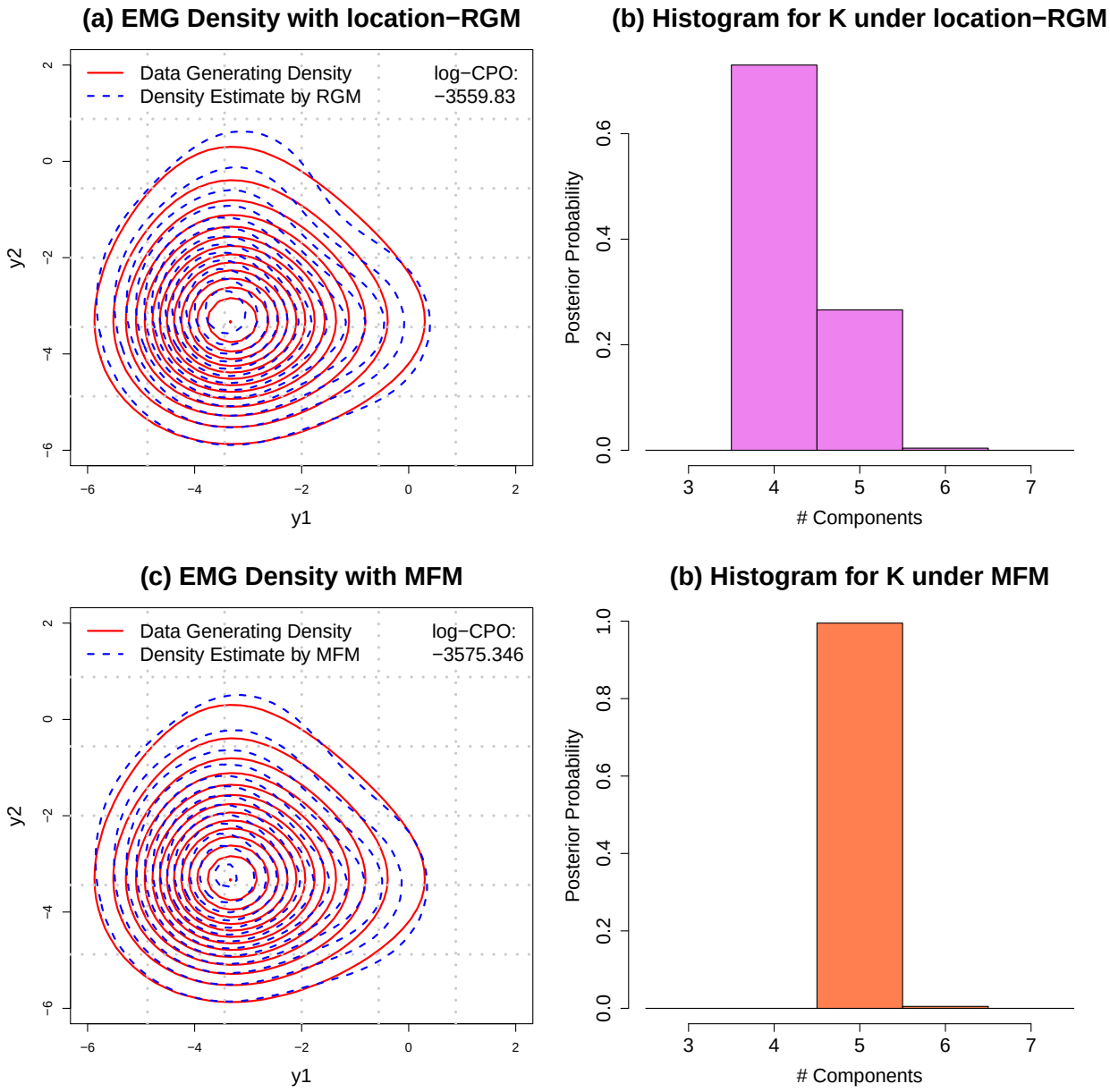

Figure 3: Fitting Uni-modal Density using Location-Mixtures only: Panels (a) and (c) are the contour plots for the posterior density estimation under location-RGM and MFM, respectively. Panels (b) and (d) are the histograms of the posterior number of components under location-RGM and MFM, respectively.

5.3 Multivariate Model-Based Clustering

Now we focus on a higher dimensional model-based clustering problem. Suppose that we generate $n = 500$ i.i.d. samples from a mixture of 3 10-dimensional Gaussians:

$$f_0(\boldsymbol{y}) = 0.4\phi(\boldsymbol{y} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.3\phi(\boldsymbol{y} \mid \boldsymbol{\mu}_2, 3\boldsymbol{I}_{10}) + 0.3\phi(\boldsymbol{y} \mid \boldsymbol{\mu}_3, 2\boldsymbol{I}_{10}),$$

where the covariance matrix for the first component is a randomly generated diagonal matrix:

$$\boldsymbol{\Sigma}_1 = \text{diag}(5.5729, 5.0110, 3.6832, 8.1931, 5.7717, 3.0267, 3.5011, 7.8291, 4.2233, 4.3885),$$

and $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = (-6, \cdots, -6)^{\mathrm{T}} \in \mathbb{R}^{10}$, $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$. In this simulation study, we focus on the model-based clustering without fixing the number K of components *a priori*. Due to the challenge of visualizing high-dimensional clustering, we only show the scatter plot of the 4th versus 8th coordinate of the simulated data in Figure 4a. These two dimensions correspond to the first two largest eigenvalues of the covariance matrix. The projection of the data onto this 2-dimensional subspace shows that the three clusters are not well-separated. We implement the proposed blocked-collapsed Gibbs sampler with $g_0 = 70$, $\tau = 10$, $m = 2, \underline{\sigma} = 0.1, \overline{\sigma} = 10$. To demonstrate the efficiency of the proposed sampler, we keep all MCMC samples and compare the efficiency of the algorithms in terms of their numbers of burn-in iterations.

For comparison, we consider the two alternative clustering models and evaluate their performance in terms of efficiency in estimating posterior number of components. The first one is the DPM model: $(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \sim N(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$, $(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} \mid G) \overset{\text{i.i.d.}}{\sim} G$, and $(G \mid \alpha, G_0) \sim \text{DP}(\alpha, G_0)$, where $G_0 = \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}/k_0)$, $\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(12, \boldsymbol{\Psi}_1)$, $\alpha = 1$, $k_0 \sim \text{Gamma}(0.005, 0.005)$, and $\boldsymbol{\Psi}_1 = 0.1\boldsymbol{I}_{10}$. The Second alternative model is the *DPP mixture model* proposed in Xu et al. (2016), who used the determinantal point process as a repulsive function: $h_K(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K) = \det\left\{[\exp\left(-0.5\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|^2/\theta^2\right)]_{K\times K}\right\}$ for $K \geq 2$, $h_K \equiv 1$ otherwise. Posterior inference of the DPP mixture model is performed using a potentially inefficient RJ-MCMC sampler.

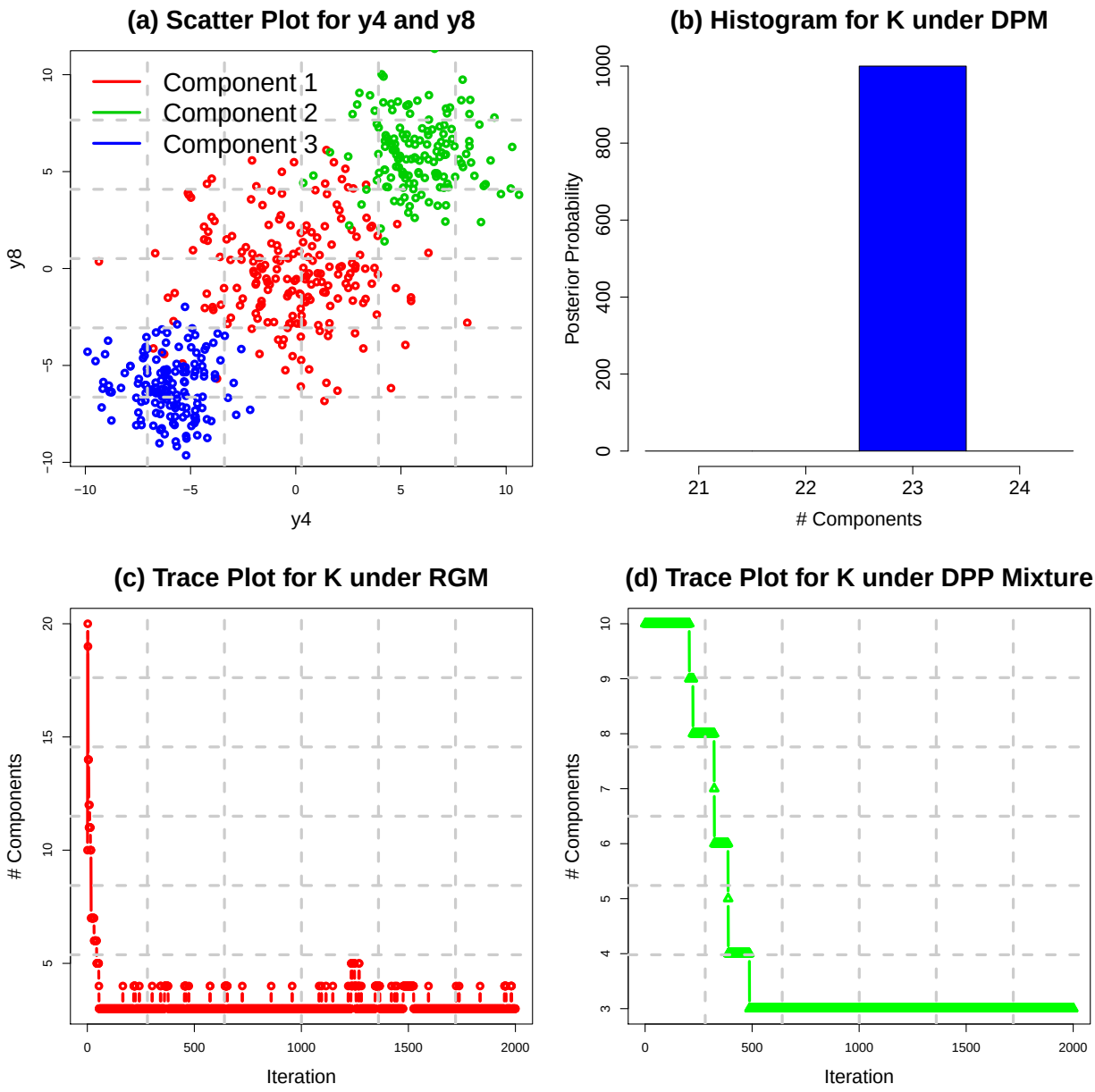

Figure 4: Multivariate Model-Based Clustering: Panel (a) is the scatter plot of the 4th-versus-8th coordinate of the simulated data; Panel (b) is the histogram of the posterior number of clusters under the DPM model; Panels (c) and (d) are the trace plots of the posterior samples of K under the RGM model, and that of the number of clusters under the DPP mixture model, respectively.

We initialize the Markov chains with $K = 10$ for all three models. By comparing the histogram and trace plots of the posterior number of components/clusters in Figures 4b,

4c, and 4d, we find the DPM model significantly over-estimates the number of components ($K = 23$ with high posterior probability) in order to fit the 10-dimensional data well; The DPP mixture inferred with RJ-MCMC, though eventually stabilizes at the correct $K = 3$, requires relatively more number of iterations to find the underlying truth (approximately 500 iterations). In contrast, the posterior number of components under the RGM model highly concentrates around the underlying true $K = 3$, and stabilizes within only 100 iterations. It can be clearly seen that the blocked-collapsed Gibbs sampler stabilizes around the true K using fewer number of iterations than the RJ-MCMC, possibly due to the reason that K is allowed to change by greater than 1 within a single iteration in the blocked-collapsed Gibbs sampler, but can only increase or decrease by at most one in the RJ-MCMC sampler for the DPP mixture model. However, it should be noting that the blocked-collapsed Gibbs sampler requires the computation of several intractable integrals within each iteration, which could potentially increase the per-iteration computation complexity compared to the standard RJ-MCMC.

We further report the performance of the model-based clustering procedure under the RGM model. Adopting the ideas in Xu et al. (2016) and Dahl (2006), we define the association matrix $S \in \{0,1\}^{n\times n}$ with (i,j)th entries being $\mathbb{I}(\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j)$, and $H \in \{0,1\}^{n\times n}$ with (i,j)th entries being $\mathbb{I}(\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$. Using the posterior samples, H can be approximated using the posterior mean of $\mathbb{I}(\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j)$ for all (i,j) pairs. We compute the mean of the absolute mis-classification matrix $(|H_{ij} - S_{ij}|)_{n\times n}$. The mis-classification error defined by $(1/n^2)\|\hat{H} - S\|_{\mathrm{F}}$ is 1.0215×10^{-5}, where $\hat{H}$ is computed using the posterior means.

5.4 Old Faithful Geyser Eruption Data

In this subsection, we consider the Old Faithful geyser eruption data that record the eruption length of the Old faithful geyser in the Yellowstone National Park with the number of observations $n = 272$ as a real world example. Following the procedure described in Qin

and Priebe (2013) and Garcia-Escudero and Gordaliza (1999), for each observed eruption duration time, we pair it with the time length of the next eruption, so that we have a bivariate data of sample size 271. The points with the "short followed by short" eruption property were identified as outliers in Garcia-Escudero and Gordaliza (1999), in which a robust trimmed mean procedure was used to reduce the error caused by these outliers. Here we apply the RGM model to analyze the bivariate dataset, and show that the outliers can actually be identified as an extra component. We also compare the proposed method with the two alternative models: the DPM model and the DPP mixture model as described in subsection 5.3.

Figure 5 shows the predictive densities and the histograms of the number of components/clusters estimated by the three models: the RGM model, the DPM model, and the DPP mixture model. The proposed RGM, not only identifies the outliers component (Figure 5a), but also provides the posterior number of components that is highly concentrated at $K = 4$ (Figure 5b). In contrast, Figure 5c shows that DPP mixture fails to identify the outliers at the bottom-left corner of the scatter plot – instead, they are merged into the existing cluster located at the bottom-right corner. The corresponding posterior number of components K, as illustrated in Figure 5d, is highly concentrated at $K = 3$, failing to detect the outlier component. In addition, notice that failure in identifying the outliers significantly affects the posterior predictive density estimate, as shown from the comparison of the level curves among Figures 5a, 5c, and 5e. The DPM model in Figure 5e, although successfully detects the outliers component, still assigns relatively larger posterior probability to redundant components(Figure 5f). Hence the proposed RGM model outperforms the other two alternatives in terms of either the robustness or the model complexity measured by the posterior of K. This conclusion is also supported by the fact that log-CPO of the RGM model is higher than those of the DPM model and the DPP mixture model (Table 1).

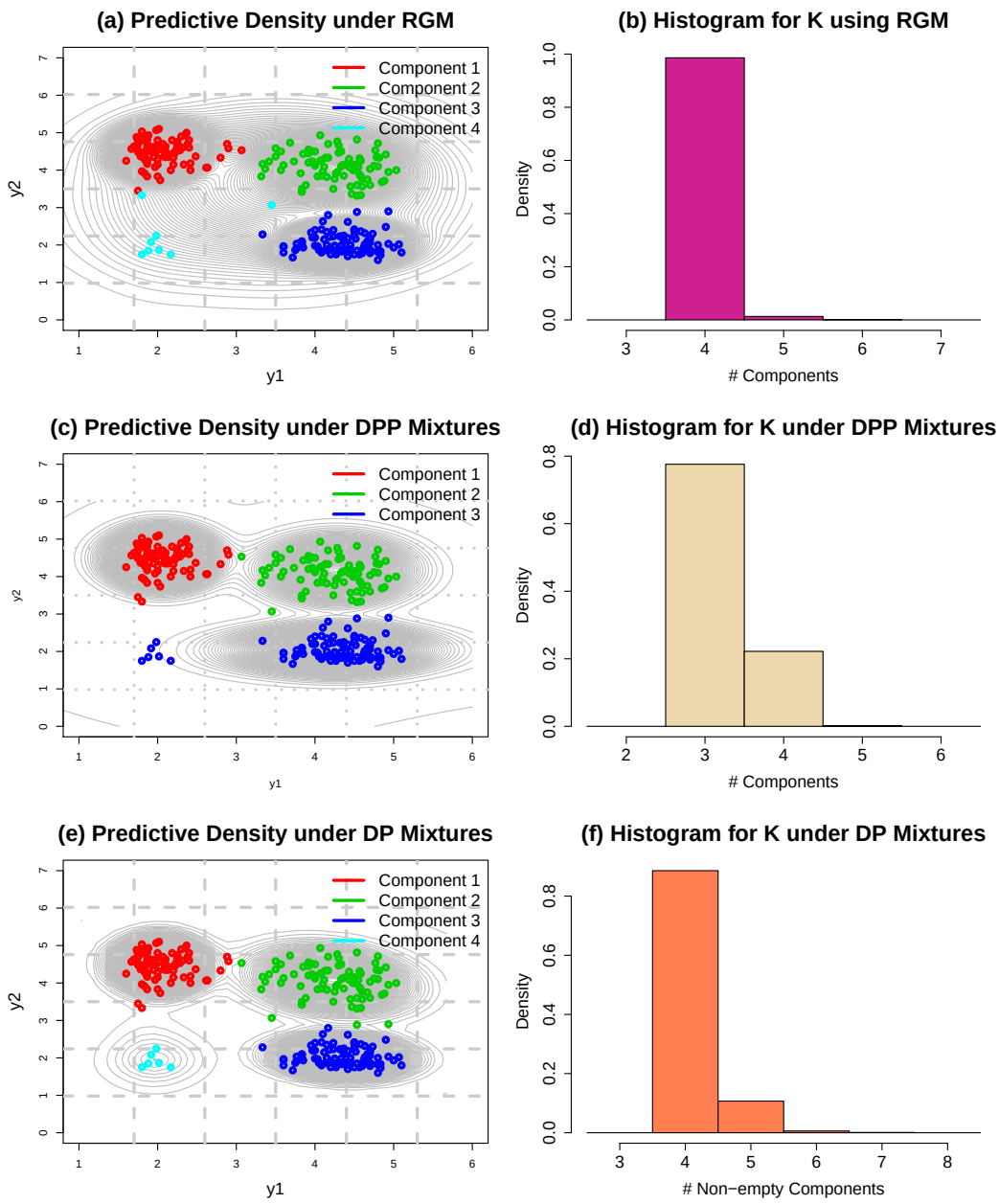

Figure 5: Old Faithful Geyser Eruption Data: Panels (a), (c), and (e) are the scatter plots of the observations with their corresponding clusters and contour plots of the posterior predictive density estimate(grey level curves) stratified by the RGM model, the DPP mixture model, and the DPM model, respectively. Panels (b), (d), and (f) are the histograms of the posterior distributions of the number of components/clusters under the RGM model, the DPP mixture model, and the DPM model, respectively.

6 Conclusion

We propose the (Bayesian) RGM model, in which the location parameters for each component are not *a priori* independent, but jointly distributed according to some symmetric repulsive distribution that encourages the separation of the locations for different components. We establish the posterior consistency and obtain an "almost" parametric posterior contraction rate($(\log n)^t/\sqrt{n}$ with $t > p+1$), generalizing the repulsive mixture model proposed by Petralia et al. (2012) and Quinlan et al. (2017) to the context of density estimation in nonparametric GMM. Furthermore, we study the shrinkage effect on the model complexity of the proposed RGM model regarding the number of necessary components needed to fit the data well.

Based on the exchangeable partition distribution, we develop a blocked-collapsed Gibbs sampler for the posterior inference. Through extensive simulation studies and real data analysis, we demonstrate that the proposed RGM model is able to detect outliers, penalize the number of components to reduce model complexity, and accurately estimate the underlying true density.

There are several potential further extensions. Beyond mixture models for density estimation, it is also interesting to extend the repulsive mixture model to the nested clustering of grouped data, and perform simultaneous clustering of individuals within each group and the group level features when the inference prefers the parsimonious model and the focus is the interpretation of the clusters as meaningful subgroups. Secondly, the posterior distribution of the number of components under the RGM model is potentially sensitive to the hyperparameters in the repulsive function h_K. Performing sensitivity analysis by imposing suitable priors on the hyperparameters is possible if an efficient updating rule for them can be integrated within the blocked-collapsed Gibbs sampler. Lastly, instead of implementing a Gibbs sampler, which is not scalable to large number of observations, one can develop an

optimization-based fast inference algorithm, which would greatly improve the computational efficiency and scalability of posterior inference.

Supplementary Material

The supplementary material contains the proofs of all technical results in Sections 2, 3, and 4, additional numerical results, and the MATLAB code for implementing the blocked-collapsed Gibbs sampler in Section 4.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.

Canale, A., De Blasi, P., et al. (2017). Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404.

Chen, J. et al. (2017). Consistency of the mle under mixture models. *Statistical Science*, 32(1):47–63.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian Inference for Gene Expression and Proteomics*, pages 201–218.

Fuquene, J., Steel, M., and Rossell, D. (2016). On choosing mixture components via non-local priors. *arXiv preprint arXiv:1604.00314*.

Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, 94(447):956–969.

Ghosal, S., Ghosh, J. K., Ramamoorthi, R., et al. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531.

Ghosal, S. and Van Der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29(5):1233–1263.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and James, L. F. (2002). Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532.

Kruijer, W., Rousseau, J., Van Der Vaart, A., et al. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257.

MacEachern, S. N. and Mueller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

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*, pages 199–206.

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(521):340–356.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Petralia, F., Rao, V., and Dunson, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897.

Pettit, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Societ: Series B (Statistical Methodology)*, 52(1):175–184.

Qin, Y. and Priebe, C. E. (2013). Maximum Lq-likelihood estimation via the expectation-maximization algorithm: a robust estimation of mixture models. *Journal of the American Statistical Association*, 108(503):914–928.

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2017). Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (Statistical Methodology)*, 59(4):731–792.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Scricciolo, C. et al. (2011). Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electronic Journal of Statistics*, 5:270–308.

Shen, W., Tokdar, S. T., and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation®*, 36(1):45–54.

Walker, S. G., Lijoi, A., Pruenster, I., et al. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35(2):738–746.

Wu, Y. and Ghosal, S. (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419.

Wu, Y., Ghosal, S., et al. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331.

Xu, Y., Mueller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964.