Repulsive Mixtures

Francesca Petralia
Department of Statistical Science
Duke University
fp12@duke.edu

Vinayak Rao
Gatsby Computational Neuroscience Unit
University College London
vrao@gatsby.ucl.ac.uk

David B. Dunson
Department of Statistical Science
Duke University
dunson@stat.duke.edu

Abstract

Discrete mixtures are used routinely in broad sweeping applications ranging from unsupervised settings to fully supervised multi-task learning. Indeed, finite mixtures and infinite mixtures, relying on Dirichlet processes and modifications, have become a standard tool. One important issue that arises in using discrete mixtures is low separation in the components; in particular, different components can be introduced that are very similar and hence redundant. Such redundancy leads to too many clusters that are too similar, degrading performance in unsupervised learning and leading to computational problems and an unnecessarily complex model in supervised settings. Redundancy can arise in the absence of a penalty on components placed close together even when a Bayesian approach is used to learn the number of components. To solve this problem, we propose a novel prior that generates components from a repulsive process, automatically penalizing redundant components. We characterize this repulsive prior theoretically and propose a Markov chain Monte Carlo sampling algorithm for posterior computation. The methods are illustrated using synthetic examples and an iris data set.

Key Words: Bayesian nonparametrics; Dirichlet process; Gaussian mixture model; Model-based clustering; Repulsive point process; Well separated mixture.

1 Introduction

Discrete mixture models characterize the density of $y \in \mathcal{Y} \subset \Re^m$ as

$$f(y) = \sum_{h=1}^{k} p_h \phi(y; \gamma_h) \tag{1}$$

where $p = (p_1, \ldots, p_k)^T$ is a vector of probabilities summing to one, and $\phi(\cdot; \gamma)$ is a kernel depending on parameters $\gamma \in \Gamma$, which may consist of location and scale parameters. In analyses of finite mixture models, a common concern is over-fitting in which *redundant* mixture components located close together are introduced. Over-fitting can have an adverse impact on predictions and degrade unsupervised learning. In particular, introducing components located close together can lead to splitting of well separated clusters into a larger number of closely overlapping clusters. Ideally, the criteria for selecting k in a frequentist analysis and the prior on k and $\{\gamma_h\}$ in a Bayesian analysis should guard against such over-fitting. However, the impact of the criteria used and prior chosen can be subtle.

Recently, [1] studied the asymptotic behavior of the posterior distribution in over-fitted Bayesian mixture models having more components than needed. They showed that a carefully chosen prior will lead to asymptotic emptying of the redundant components. However, several challenging practical issues arise. For their prior and in standard Bayesian practice, one assumes that $\gamma_h \sim P_0$ independently *a priori*. For example, if we consider a finite location-scale mixture of multivariate Gaussians, one may choose P_0 to be multivariate Gaussian-inverse Wishart. However, the behavior of the posterior can be sensitive to P_0 for finite samples, with higher variance P_0 favoring allocation to fewer clusters. In addition, drawing the component-specific parameters from a common prior tends to favor components located close together unless the variance is high.

Sensitivity to P_0 is just one of the issues. For finite samples, the weight assigned to redundant components is often substantial. This can be attributed to non- or *weak* identifiability. Each mixture component can potentially be split into multiple components having the same parameters. Even if exact equivalence is ruled out, it can be difficult to distinguish between models having different degrees of splitting of well-separated components into components located close together. This issue can lead to an unnecessarily complex model, and creates difficulties in estimating the number of components and component-specific parameters. Existing strategies, such as the incorporation of order constraints, do not adequately address this issue, since it is difficult to choose reasonable constraints in multivariate problems and even with constraints, the components can be close together.

The problem of separating components has been studied for Gaussian mixture models ([2]; [3]). Two Gaussians can be separated by placing an arbitrarily chosen lower bound on the distance between their means. Separated Gaussians have been mainly utilized to speed up convergence of the Expectation-Maximization (EM) algorithm. In choosing a minimal separation level, it is not clear how to obtain a good compromise between values that are too low to solve the problem and ones that are so large that one obtains a poor fit. To avoid such arbitrary *hard* separation thresholds, we instead propose a repulsive prior that smoothly pushes components apart.

In contrast to the vast majority of the recent Bayesian literature on discrete mixture models, instead of drawing the component-specific parameters $\{\gamma_h\}$ independently from a common prior P_0, we propose a joint prior for $\{\gamma_1, \ldots, \gamma_k\}$ that is chosen to assign low density to γ_hs located close together. The deviation from independence is specified *a priori* by a pair of repulsion parameters. The proposed class of repulsive mixture models will only place components close together if it results in a substantial gain in model fit. As we illustrate, the prior will favor a more parsimonious representation of densities, while improving practical performance in unsupervised learning. We provide strong theoretical results on rates of posterior convergence and develop Markov chain Monte Carlo algorithms for posterior computation.

2 Bayesian repulsive mixture models

2.1 Background on Bayesian mixture modeling

Considering the finite mixture model in expression (1), a Bayesian specification is completed by choosing priors for the number of components k, the probability weights p, and the component-specific parameters $\gamma = (\gamma_1, \ldots, \gamma_k)^T$. Typically, k is assigned a Poisson or multinomial prior, p a $Dirichlet(\alpha)$ prior with $\alpha = (\alpha_1, \ldots, \alpha_k)^T$, and $\gamma_h \sim P_0$ independently, with P_0 often chosen to be conjugate to the kernel ϕ. Posterior computation can proceed via a reversible jump Markov chain Monte Carlo algorithm involving moves for adding or deleting mixture components. Unfortunately, in making a $k \to k+1$ change in model dimension, efficient moves critically depend on the choice of proposal density. [4] proposed an alternate Markov chain Monte Carlo method, which treats the parameters as a marked point process, but does not have clear computational advantages relative to reversible jump.

It has become popular to use over-fitted mixture models in which k is chosen as a conservative upper bound on the number of components under the expectation that only relatively few of the components will be occupied by subjects in the sample. From a practical perspective, the success of over-fitted mixture models has been largely due to ease in computation.

As motivated in [5], simply letting $\alpha_h = c/k$ for $h = 1, \ldots, k$ and a constant $c > 0$ leads to an approximation to a Dirichlet process mixture model for the density of y, which is obtained in the

limit as k approaches infinity. An alternative finite approximation to a Dirichlet process mixture is obtained by truncating the stick-breaking representation of [6], leading to a similarly simple Gibbs sampling algorithm [7]. These approaches are now used routinely in practice.

2.2 Repulsive densities

We seek a prior on the component parameters in (1) that automatically favors spread out components near the support of the data. Instead of generating the atoms γ_h independently from P_0, one could generate them from a repulsive process that automatically pushes the atoms apart. This idea is conceptually related to the literature on repulsive point processes [8]. In the spatial statistics literature, a variety of repulsive processes have been proposed. One such model assumes that points are clustered spatially, with the cluster centers having a Strauss density [9], that is $p(k, \gamma) \propto \beta^k \rho^{r(\gamma)}$ where k is the number of clusters, $\beta > 0$, $0 < \rho \leq 1$ and $r(\gamma)$ is the number of pairwise centers that lie within a pre-specified distance r of each other. A possibly unappealing feature is that repulsion is not directly dependent on the pairwise distances between the clusters. We propose an alternative class of priors, which smoothly push apart components based on pairwise distances.

Definition 1. *A density $h(\gamma)$ is repulsive if for any $\delta > 0$ there is a corresponding $\epsilon > 0$ such that $h(\gamma) < \delta$ for all $\gamma \in \Gamma \setminus G_\epsilon$, where $G_\epsilon = \{\gamma : d(\gamma_s, \gamma_i) > \epsilon; s = 1, \ldots, k; i < s\}$ and d is a metric.*

Depending on the specification of the metric $d(\gamma_s, \gamma_j)$, a prior satisfying definition 1 may limit overfitting or favor well separated clusters. When $d(\gamma_s, \gamma_j)$ is the distance between sub-vectors of γ_s and γ_j corresponding to only locations the proposed prior favors well separated clusters. Instead, when $d(\gamma_s, \gamma_j)$ is the distance between the sth and jth kernel, a prior satisfying definition 1 limits overfitting in density estimation. Though both cases can be implemented, in this paper we will focus exclusively on the clustering problem. As a convenient class of repulsive priors which smoothly push components apart, we propose

$$\pi(\gamma) = c_1 \left(\prod_{h=1}^{k} g_0(\gamma_h) \right) h(\gamma), \tag{2}$$

with c_1 being the normalizing constant that depends on the number of components k. The proposed prior is related to a class of point processes from the statistical physics and spatial statistics literature referred to as Gibbs processes [10]. We assume $g_0 : \Gamma \to \Re_+$ and $h : \Gamma^k \to [0, \infty)$ are continuous with respect to Lesbesgue measure, and h is bounded above by a positive constant c_2 and is repulsive according to definition 1. It follows that density π defined in (2) is also repulsive. A special hardcore repulsion is produced if the repulsion function is zero when at least one pairwise distance is smaller than a pre-specified threshold. Such a density implies choosing a minimal separation level between the atoms. As mentioned in the introduction, we avoid such arbitrary *hard* separation thresholds by considering repulsive priors that smoothly push components apart. In particular, we propose two repulsion functions defined as

$$h(\gamma) = \prod_{\{(s,j) \in A\}} g\{d(\gamma_s, \gamma_j)\} \tag{3}$$

$$h(\gamma) = \min_{\{(s,j) \in A\}} g\{d(\gamma_s, \gamma_j)\} \tag{4}$$

with $A = \{(s, j) : s = 1, \ldots, k; j < s\}$ and $g : \Re_+ \to [0, M]$ a strictly monotone differentiable function with $g(0) = 0$, $g(x) > 0$ for all $x > 0$ and $M < \infty$. It is straightforward to show that h in (3) and (4) is integrable and satisfies definition 1. The two alternative repulsion functions differ in their dependence on the relative distances between components, with all the pairwise distances playing a role in (3), while (4) only depends on the minimal separation. A flexible choice of g corresponds to

$$g\{d(\gamma_s, \gamma_j)\} = \exp\big[- \tau \{d(\gamma_s, \gamma_j)\}^{-\nu} \big], \tag{5}$$

where $\tau > 0$ is a scale parameter and ν is a positive integer controlling the rate at which g approaches zero as $d(\gamma_s, \gamma_j)$ decreases. Figure 1 shows contour plots of the prior $\pi(\gamma_1, \gamma_2)$ defined as (2) with g_0 being the standard normal density, the repulsive function defined as (3) or (4) and g defined as (5) for different values of (τ, ν). As τ and ν increase, the prior increasingly favors well separated components.

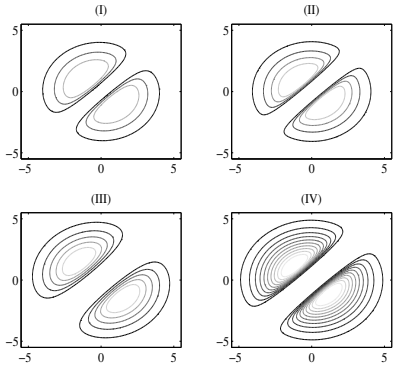

Figure 1: Contour plots of the repulsive prior $\pi(\gamma_1, \gamma_2)$ under (3), either (4) or (5) and (6) with hyperparameters (τ, ν) equal to $(I)(1,2), (II)(1,4), (III)(5,2)$ and $(IV)(5,4)$

2.3 Theoretical properties

Let the true density $f_0 : \Re^m \to \Re_+$ be defined as $f_0 = \sum_{h=1}^{k_0} p_{0h}\phi(\gamma_{0h})$ with $\gamma_{0h} \in \Gamma$ and γ_{0j}s such that there exists an $\epsilon_1 > 0$ such that $\min_{\{(s,j):s<j\}} d(\gamma_{0s}, \gamma_{0j}) \geq \epsilon_1$ with d being the Euclidean distance. Let $f = \sum_{h=1}^{k} p_h\phi(\gamma_h)$ with $\gamma_h \in \Gamma$. Let $\gamma \sim \pi$ with $\gamma = (\gamma_1, \ldots, \gamma_k)^T$ and π satisfying definition 1. Let $p \sim \lambda$ with $\lambda = Dirichlet(\alpha)$ and $k \sim \mu$ with $\mu(k = k_0) > 0$. Let $\theta = (p, \gamma)$. These assumptions on f_0 and f will be referred to as condition B0. Let Π be the prior induced on $\cup_{j=1}^{\infty}\mathcal{F}_k$, where $\mathcal{F}_k$ is the space of all distributions defined as (1).

We will focus on γ being a location parameter, though the results can be extended to location-scale kernels. Let $|\cdot|_1$ denote the L_1 norm and $KL(f_0, f) = \int f_0 \log(f_0/f)$ refer to the Kullback-Leibler (K-L) divergence between f_0 and f. Density f_0 belongs to the K-L support of the prior Π if $\Pi\{f : KL(f_0, f) < \epsilon\} > 0$ for all $\epsilon > 0$. The next lemma provides sufficient conditions under which the true density is in the K-L support of the prior.

Lemma 1. *Assume condition B0 is satisfied with $m = 1$. Let D_0 be a compact set containing parameters $(\gamma_{01}, \ldots, \gamma_{0k_0})$. Suppose $\gamma \sim \pi$ with π satisfying definition 1. Let ϕ and π satisfy the following conditions:*

A1. for any $y \in \mathcal{Y}$, the map $\gamma \to \phi(y;\gamma)$ is uniformly continuous

A2. for any $y \in \mathcal{Y}$, $\phi(y;\gamma)$ is bounded above by a constant

A3. $\int f_0 \left|\log\left\{\sup_{\gamma\in D_0}\phi(\gamma)\right\} - \log\left\{\inf_{\gamma\in D_0}\phi(\gamma)\right\}\right| < \infty$

A4. π is continuous with respect to Lebesgue measure and for any vector $x \in \Gamma^k$ with $\min_{\{(s,j):s<j\}} d(x_s, x_j) \geq \upsilon$ for some $\upsilon > 0$ there is a $\delta > 0$ such that $\pi(\gamma) > 0$ for all γ satisfying $||\gamma - x||_1 < \delta$

Then f_0 is in the K-L support of the prior Π.

Lemma 2. *The repulsive density in (2) with h defined as either (3) or (4) satisfies condition A4 in lemma 1.*

The next lemma formalizes the posterior rate of concentration for univariate location mixtures of Gaussians.

Lemma 3. *Let condition B0 be satisfied, let $m = 1$ and ϕ be the normal kernel depending on a location parameter γ and a scale parameter σ. Assume that condition $(i), (ii)$ and (iii) of theorem 3.1 in [11] and assumption A4 in lemma 1 are satisfied. Furthermore, assume that*

C1) the joint density π leads to exchangeable random variables and for all k the marginal density of the location parameter γ_1 satisfies $\pi_m(|\gamma_1| \geq t) \lesssim \exp\left(-q_1 t^2\right)$ for a given $q_1 > 0$

C2) there are constants $u_1, u_2, u_3 > 0$, possibly depending on f_0, such that for any $\epsilon \leq u_3$

$$\pi(||\gamma - \gamma_0||_1 \leq \epsilon) \geq u_1 \exp(-u_2 k_0 \log(1/\epsilon))$$

Then the posterior rate of convergence relative to the L_1 metric is $\epsilon_n = n^{-1/2} \log n$.

Lemma 3 is essentially a modification of theorem 3.1 in [11] to the proposed repulsive mixture model. Lemma 4 gives sufficient conditions for π to satisfy condition C1 and C2 in lemma 3.

Lemma 4. *Let π be defined as (2) and h be defined as either (3) or (4), then π satisfies condition C2 in lemma 3. Furthermore, if for a positive constant n_1 the function g_0 satisfies $g_0(|x| \geq t) \lesssim \exp(-n_1 t^2)$, π satisfies condition C1 in lemma 3.*

As motivated above, when the number of mixture components is chosen to be unnecessarily large, it is appealing for the posterior distribution of the weights of the extra components to be concentrated near zero. Theorem 1 formalizes the rate of concentration with increasing sample size n. One of the main assumptions required in theorem 1 is that the posterior rate of convergence relative to the L_1 metric is $\delta_n = n^{-1/2}(\log n)^q$ with $q \geq 0$. We provided the contraction rate, under the proposed prior specification and univariate Gaussian kernel, in lemma 3. However, theorem 1 is a more general statement and it applies to multivariate mixture density of any kernel.

Theorem 1. *Let assumptions $B0 - B5$ be satisfied. Let π be defined as (2) and h be defined as either (3) or (4). If $\bar{\alpha} = \max(\alpha_1, \ldots, \alpha_k) < m/2$ and for positive constants r_1, r_2, r_3 the function g satisfies $g(x) \leq r_1 x^{r_2}$ for $0 \leq x < r_3$ then*

$$\lim_{M \to \infty} \limsup_{n \to \infty} E_n^0 \left[P \left\{ \min_{\{\sigma \in S_k\}} \left(\sum_{i=k_0+1}^{k} p_{\sigma(i)} \right) > M n^{-1/2} (\log n)^{q(1+s(k_0,\alpha)/s_{r_2})} \right\} \right] = 0$$

with $s(k_0, \alpha) = k_0 - 1 + m k_0 + \bar{\alpha}(k - k_0)$, $s_{r_2} = r_2 + m/2 - \bar{\alpha}$ and S_k the set of all possible permutations of $\{1, \ldots, k\}$.

Assumptions $(B1 - B5)$ can be found in the supplementary material. Theorem 1 is a modification of theorem 1 in [1] to the proposed repulsive mixture model. Theorem 1 implies that the posterior expectation of weights of the extra components is of order $O\left\{n^{-1/2}(\log n)^{q(1+s(k_0,\alpha)/s_{r_2})}\right\}$. When g is defined as (5), parameters r_1 and r_2 can be chosen such that $r_1 = \tau$ and $r_2 = \nu$.

When the number of components is unknown, with only an upper bound known, the posterior rate of convergence is equivalent to the parametric rate $n^{-1/2}$ [12]. In this case, the rate in theorem 1 is $n^{-1/2}$ under usual priors or the repulsive prior. However, in our experience using usual priors, the sum of the extra components can be substantial in small to moderate sample sizes, and often has high variability. As we show in Section 3, for repulsive priors the sum of the extra component weights is close to zero and has small variance for small as well as large sample sizes. On the other hand, when an upper bound on the number of components is unknown, the posterior rate of concentration is $n^{-1/2}(\log n)^q$ with $q > 0$. In this case, according to theorem 1, using the proposed prior specification the logarithmic factor in theorem 1 of [1] can be improved.

2.4 Parameter calibration and posterior computation

The parameters involved in the repulsion function h are chosen such that *a priori*, with high probability, the clusters will be adequately separated. Consider the case where ϕ is a location-scale kernel with location and scale parameters (γ, Σ) and is symmetric about γ. Here, it is natural to relate the separation of two densities to the distance between their location parameters. The following definition introduces the concept of separation level between two densities.

Definition 2. *Let f_1 and f_2 be two densities having location-scale parameters (γ_1, Σ_1) and (γ_2, Σ_2) respectively, with $\gamma_1, \gamma_2 \in \Gamma$ and $\Sigma_1, \Sigma_2 \in \Omega$. Given a metric $t(\cdot, \cdot)$, a positive constant c and a function $\omega : \Omega \times \Omega \to \Re_+$, f_1 and f_2 are c-separated if*

$$t(\gamma_1, \gamma_2) \geq c\omega(\Sigma_1, \Sigma_2)^{1/2}$$

Definition 2 is in the spirit of [2] but generalized to any symmetric location-scale kernel. A mixture of k densities is c-separated if all pairs of densities are c-separated. The parameters of the repulsion

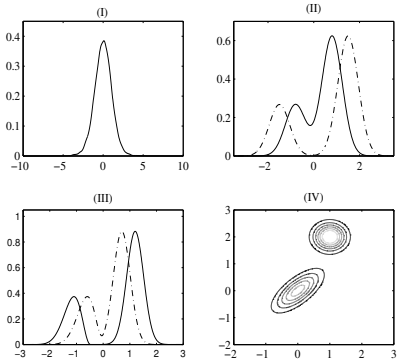

Figure 2: (I) Student's t density, (II) two-components mixture of poorly (solid) and well separated (dot-dash) Gaussian densities, referred as (IIa, IIb), (III) mixture of poorly (dot-dash) and well separated (solid) Gaussian and Pearson densities, referred as $(IIIa, IIIb)$, (IV) two-components mixture of two-dimensional non-spherical Gaussians

function, (τ, ν), will be chosen such that, for an *a priori* chosen separation level c, definition 2 is satisfied with high probability. In practice, for a given pair (τ, ν), we estimate the probability of pairwise c-separation empirically by simulating N replicates of (γ_h, Σ_h) for each component $h = 1, \ldots, k$ from the prior. The appropriate values (τ, ν) are obtained by starting with small values, and increasing until the pre-specified pairwise c-separated probability is reached. In practice, only τ will be calibrated to reach a particular probability value. This is because ν controls the rate at which the density tends to zero as two components approach but not the separation level across them. In practice we have found that $\nu = 2$ provides a good default value and we fix ν at this value in all our applications below.

A possible issue with the proposed repulsive mixture prior is that the full conditionals are nonstandard, complicating posterior computation. To address this, we propose a data augmentation scheme, introducing auxiliary slice variables to facilitate sampling [13]. This algorithm is straightforward to implement and is efficient by MCMC standards. Further details can be found in the supplementary material. It will be interesting in future work to develop fast approximations to MCMC for implementation of repulsive mixture models, such as variational methods for approximating the full posterior and optimization methods for obtaining a maximum *a posteriori* estimate. The latter approach would provide an alternative to usual maximum likelihood estimation via the EM algorithm, which provides a penalty on components located close together.

3 Synthetic examples

Synthetic toy examples were considered to assess the performance of the repulsive prior in density estimation, classification and emptying the extra components. Figure 2 plots the true densities in the various synthetic cases that we considered. For each synthetic dataset, repulsive and non-repulsive mixture models were compared considering a fixed upper bound on the number of components; extra components should be assigned small probabilities and hence effectively excluded. The auxiliary variable sampler was run for $10,000$ iterations with a burn-in of $5,000$. The chain was thinned by keeping every 10th simulated draw. To overcome the label switching problem, the samples were post-processed following the algorithm of [14]. Details on parameters involved in the true densities and choice of prior distributions can be found in the supplementary material.

Table 1 shows summary statistics of the K-L divergence, the misclassification error and the sum of extra weights under repulsive and non-repulsive mixtures with six mixture components as the upper bound. Table 1 shows also the misclassification error resulting from hierarchical clustering [15]. In practice, observations drawn from the same mixture component were considered as belonging to the same category and for each dataset a similarity matrix was constructed. The misclassification error was established in terms of divergence between the true similarity matrix and the posterior similar-

ity matrix. As shown in table 1, the K-L divergences under repulsive and non-repulsive mixtures become more similar as the sample size increases. For smaller sample sizes, the results are more similar when components are very well separated. Since a repulsive prior tends to discourage overlapping mixture components, a repulsive model might not estimate the density quite as accurately when a mixture of closely overlapping components is needed. However, as the sample size increases, the fitted density approaches the true density regardless of the degree of closeness among clusters. Again, though repulsive and non-repulsive mixtures perform similarly in estimating the true density, repulsive mixtures place considerably less probability on extra components leading to more interpretable clusters. In terms of misclassification error, the repulsive model outperforms the other two approaches while, in most cases, the worst performance was obtained by the non-repulsive model. Potentially, one may favor fewer clusters, and hence possibly better separated clusters, by penalizing the introduction of new clusters more through modifying the precision in the Dirichlet prior for the weights; in the supplemental materials, we demonstrate that this cannot solve the problem.

Table 1: Mean and standard deviation of K-L divergence, misclassification error and sum of extra weights resulting from non-repulsive (N-R) and repulsive (R) mixtures with a maximum number of clusters equal to six under different synthetic data scenarios

	n=100						n=1000					
	I	IIa	IIb	$IIIa$	$IIIb$	IV	I	IIa	IIb	$IIIa$	$IIIb$	IV
K-L divergence												
N-R	0·05	0·03	0·07	0·05	0·08	0·22	0·00	0·01	0·01	0·00	0·01	0·02
	0·03	0·01	0·02	0·02	0·03	0·04	0·00	0·00	0·00	0·00	0·00	0·00
R	0·03	0·08	0·09	0·07	0·09	0·24	0·01	0·01	0·01	0·01	0·01	0·03
	0·02	0·02	0·03	0·03	0·03	0·04	0·00	0·00	0·00	0·00	0·00	0·00
Misclassification												
HCT	0·12	0·11	0·41	0·12	0·78	0·21	0·45	0·42	0·14	0·42	0·09	0·20
N-R	0·68	0·26	0·06	0·17	0·05	0·13	0·65	0·24	0·03	0·14	0·02	0·19
	0·09	0·10	0·05	0·09	0·06	0·05	0·11	0·08	0·04	0·08	0·03	0·02
R	0·06	0·09	0·00	0·05	0·00	0·09	0·05	0·08	0·00	0·03	0·00	0·18
	0·05	0·04	0·02	0·03	0·01	0·03	0·05	0·02	0·02	0·03	0·01	0·01
Sum of extra weights												
N-R	0·30	0·21	0·09	0·16	0·07	0·13	0·30	0·21	0·03	0·16	0·03	0·29
	0·10	0·11	0·07	0·09	0·07	0·07	0·11	0·11	0·04	0·10	0·03	0·03
R	0·01	0·01	0·01	0·01	0·01	0·08	0·01	0·00	0·00	0·00	0·00	0·26
	0·01	0·01	0·01	0·01	0·01	0·05	0·01	0·00	0·00	0·00	0·00	0·03

4 Real data

We assessed the clustering performance of the proposed method on a real dataset. This dataset consists of 150 observations from three different species of iris each with four measurements. This dataset was previously analyzed by [16] and [17] proposing new methods to estimate the number of clusters based on minimizing loss functions. They concluded the optimal number of clusters was two. This result did not agree with the number of species due to low separation in the data between two of the species. Such point estimates of the number of clusters do not provide a characterization of uncertainty in clustering in contrast to Bayesian approaches.

Repulsive and non-repulsive mixtures were fitted under different choices of upper bound on the number of components. Since the data contains three true biological clusters, with two of these having similar distributions of the available features, we would expect the posterior to concentrate on two or three components. Posterior means and standard deviations of the three highest weights were $(0{\cdot}30, 0{\cdot}23, 0{\cdot}13)$ and $(0{\cdot}05, 0{\cdot}04, 0{\cdot}04)$ for non-repulsive and $(0{\cdot}60, 0{\cdot}30, 0{\cdot}04)$ and $(0{\cdot}04, 0{\cdot}03, 0{\cdot}02)$ for repulsive under six components. Clearly, repulsive priors lead to a posterior more concentrated on two components, and assign low probability to more than three components.

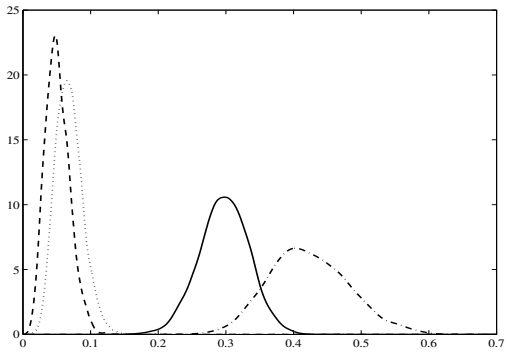

Figure 3: Posterior density of the total probability weight assigned to more than three components in the Iris data under a max of 6 or 10 components for non-repulsive (6:solid, 10:dash-dot) and repulsive (6:dash, 10:dot) mixtures.

Figure 3 shows the density of the total probability assigned to the extra components. This quantity was computed considering the number of species as the true number of clusters. According to figure 3, our repulsive prior specification leads to extra component weights very close to zero regardless of the upper bound on the number of components. The posterior uncertainty is also small. Non-repulsive mixtures assign large weight to extra components, with posterior uncertainty increasing considerably as the number of components increases.

Discussions

We have proposed a new repulsive mixture modeling framework, which should lead to substantially improved unsupervised learning (clustering) performance in general applications. A key aspect is *soft* penalization of components located close together to favor, without sharply enforcing, well separated clusters that should be more likely to correspond to the true missing labels. We have focused on Bayesian MCMC-based methods, but there are numerous interesting directions for ongoing research, including fast optimization-based approaches for learning mixture models with repulsive penalties.

Acknowledgments

This research was partially supported by grant 5R01-ES-017436-04 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH) and DARPA MSEE.

References

[1] J. Rousseau and K. Mengersen. Asymptotic Behaviour of the Posterior Distribution in Over-Fitted Models. *Journal of the Royal Statistical Society B*, 73:689–710, 2011.

[2] S. Dasgupta. Learning Mixtures of Gaussians. *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 633–644, 1999.

[3] S. Dasgupta and L. Schulman. A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians. *The Journal of Machine Learning Research*, 8:203–226, 2007.

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

[5] H. Ishwaran and M. Zarepour. Dirichlet Prior Sieves in Finite Normal Mixtures. *Statistica Sinica*, 12:941–963, 2002.

[6] J. Sethuraman. A Constructive Denition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

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

[8] M. L. Huber and R. L. Wolpert. Likelihood-Based Inference for Matern Type-III Repulsive Point Processes. *Advances in Applied Probability*, 41:958–977, 2009.

[9] A. Lawson and A. Clark. *Spatial Cluster Modeling*. Chapman & Hall CRC, London, UK, 2002.

[10] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, 2008.

[11] Catia Scricciolo. Posterior Rates of Convergence for Dirichlet Mixtures of Exponential Power Densities. *Electronic Journal of Statistics*, 5:270–308, 2011.

[12] H. Ishwaran, L. F. James, and J. Sun. Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions. *Journal of American Statistical Association*, 96:1316–1332, 2001.

[13] Paul Damien, Jon Wakefield, and Stephen Walker. Gibbs Sampling for Bayesian Non-Conjugate and Hierarchical Models by Using Auxiliary Variables. *Journal of the Royal Statistical Society B*, 61:331–344, 1999.

[14] M. Stephens. Dealing with label switching in mixture models. *Journal of the Roya; statistical society B*, 62:795–810, 2000.

[15] H. Locarek-Junge and C. Weihs. *Classification as a Tool for Research*. Springer, 2009.

[16] C. Sugar and G. James. Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association*, 98:750–763, 2003.

[17] J. Wang. Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97:893–904, 2010.