

# Nonparametric Bayes Classification and Hypothesis Testing on Manifolds

Abhishek Bhattacharya and David Dunson  
Department of Statistical Science, Duke University

**ABSTRACT.** Our first focus is prediction of a categorical response variable using features that lie on a general manifold. For example, the manifold may correspond to the surface of a hypersphere. We propose a general kernel mixture model for the joint distribution of the response and predictors, with the kernel expressed in product form and dependence induced through the unknown mixing measure. We provide simple sufficient conditions for large support and weak and strong posterior consistency in estimating both the joint distribution of the response and predictors and the conditional distribution of the response. Focusing on a Dirichlet process prior for the mixing measure, these conditions hold using von Mises-Fisher kernels when the manifold is the unit hypersphere. In this case, Bayesian methods are developed for efficient posterior computation using an exact block Gibbs sampler. Next we develop Bayesian nonparametric methods for testing whether there is difference in distributions between groups of observations on the manifold having unknown densities. We prove consistency of the Bayes Factor and develop efficient computational methods for its calculation. The proposed classification and testing methods are evaluated using simulation examples and applied to spherical data applications.

## 1. Introduction

Classification is one of the fundamental problems in statistics and machine learning. Let  $(X, Y)$  denote a pair of random variables with  $X \in \mathbb{X}$  the predictors and  $Y \in \mathbb{Y} = \{1, \dots, L\}$  the response. The focus in classification is on predicting  $Y$  given features  $X$ . From a model-based perspective, one can address this problem by first estimating  $p(y, x) = \Pr(Y = y | X = x)$ , for  $y = 1, \dots, L$  and all  $x \in \mathbb{X}$ , based on a training sample of  $n$  subjects. Then, under a 0-1 loss function, the optimal predictive value for  $y_{n+1}$ , the unknown response for an unlabeled  $(n+1)$ st subject, is simply the value of  $y$  that maximizes the estimate  $\hat{p}(y, x_{n+1})$  for  $y \in \{1, \dots, L\}$ . The model-based perspective has the advantage of providing measures of uncertainty in classification. However, performance will be critically dependent on obtaining an accurate estimate of  $p(y, x)$ .

---

*Key words and phrases.* Bayes factor; Classification; Dirichlet process mixture; Flexible prior; Hypothesis testing; Non-Euclidean manifold; Nonparametric Bayes; Posterior consistency; Spherical data.

A common strategy for addressing this problem is to use a discriminant analysis approach, which lets

$$p(y, x) = \Pr(Y = y | X = x) = \frac{\Pr(Y = y) f(x | Y = y)}{\sum_{j=1}^L \Pr(Y = j) f(x | Y = j)}, \quad y = 1, \dots, L,$$

with  $\Pr(Y = y)$  the marginal probability of having label  $y$ , and  $f(x | Y = y)$  the conditional density of the features (predictors) for subjects in class  $y$ . Then, one can simply use the proportion of subjects in class  $y$  as an estimate of  $\Pr(Y = y)$ , while applying a multivariate density estimator to learn  $f(x | Y = y)$  separately within each class. For example, [18] proposed a popular approach which estimates  $f(x | Y = y)$  using a mixture of multivariate Gaussians (refer also to [15]). [1] instead use mixtures of von Mises-Fisher distributions, but for unsupervised clustering on the unit hypersphere instead of classification. [4] extends the idea to features lying on a general manifold and uses mixtures of complex Watson distributions on the shape space.

Even when features can be assumed to have support on  $\mathbb{R}^p$ , there are two primary issues that arise. Firstly, the number of mixture components is typically unknown and can be difficult to estimate reliably, and secondly it may be difficult to accurately estimate a multivariate density specific to each class unless  $p$  is small and there are abundant training data in each class. To address these issues, a nonparametric Bayes discriminant analysis approach can be used in which the prior incorporates dependence in the unknown class-specific densities ([9]; [10]). An important challenge in implementing nonparametric Bayes methods is showing that the prior is sufficiently flexible to accurately approximate any classification function  $p(y, x)$ . A primary goal of this article is to provide methods for classification that allow the predictors to have support on a manifold, utilizing priors with full support that lead to posterior consistency for the classification function.

As noted in [17] it is routine in many applications areas, ranging from genomics to computer vision, to normalize the data prior to analysis to remove artifacts. This leads to feature vectors that lie on the surface of the unit hypersphere, though due to the lack of straightforward methods for the analysis of spherical data, Gaussian approximations in Euclidean space are typically used. [17] show that treating spherical features as Euclidean can lead to poor performance in classification if the feature vectors are not approximately spherical-homoscedastic. We will propose a class of product kernel mixture models that can be designed to have full support on the set of densities on a manifold, and that lead to strong posterior consistency. In important special cases, such as spherical data, these models also facilitate computationally convenient Gibbs sampling algorithms for posterior computation.

A closely related problem to the classification problem is testing for differences in the distribution of features across groups. In the testing setting, the nonparametric Bayes literature is surprisingly limited perhaps due to the computational challenges that arise in calculating Bayes factors. For recent articles on nonparametric testing of differences between groups, refer to [11], [24] and [19]. The former two articles considered interval null hypotheses, while the later article considered a point null for testing differences in two groups using Polya tree priors. Here, we modify the methodology developed for the classification problem to obtain an easy to implement approach for nonparametric Bayes testing of differences between

groups, with the data within each group constrained to lie on a compact metric space or Riemannian manifold, and prove consistency of this testing procedure.

Here is a very brief overview of the sections to follow. §2 describes the general modeling framework for classification on manifolds. §3 adapts the methodology to the testing problem. §4 focuses on the special case in which the features lie on the surface of a unit hyper-sphere and adapts the theory and methodology of the earlier sections to this manifold. §5 contains results from simulation studies where the developed methods of classification and testing are compared with existing ones. §6 applies the methods to spherical data applications, and §7 discusses the results. Proofs are included in an Appendix (§8).

## 2. Nonparametric Bayes Classification

**2.1. Kernel mixture model.** Let  $(\mathbb{X}, \rho)$  be a compact metric space,  $\rho$  being the distance metric and  $\mathbb{Y} = \{1, \dots, L\}$  a finite set. Consider a pair of random variables  $(X, Y)$  taking values in  $\mathbb{X} \times \mathbb{Y}$ . To induce a flexible model on the classification function  $p(y, x)$ , we propose to model the joint distribution of the  $(X, Y)$  pair. The approach of inducing a flexible model on the conditional distribution through a nonparametric model for the joint distribution was proposed by [23]. In particular, they used a DP mixture (DPM) of multivariate Gaussians for  $(X, Y)$  to induce a flexible prior on  $E(Y|X = x)$ . [27] recently generalized this approach beyond Gaussian mixtures. [7] used a joint DPM for random effects underlying a functional predictor and a response to induce a flexible model for prediction.

Our focus is on the case in which  $Y$  is an unordered categorical variable and  $X$  is constrained to have support on a compact metric space, with non-Euclidean Riemannian manifolds of particular interest. This modification is far from straightforward, as it is necessary to define a model for which large support properties can be shown, while also obtaining computational tractability. The large support property is the distinguishing feature of a nonparametric Bayes approach, as it shows that the prior can generate distributions within arbitrary small neighborhoods of the true data generating distribution. This allows uncertainty in prior knowledge, and can lead to posterior distributions that concentrate around the truth as the sample size increases.

Assume that the joint distribution of  $(X, Y)$  has a joint density with respect to some fixed base measure  $\lambda$  on  $\mathbb{X} \times \mathbb{Y}$ . Let  $\lambda = \lambda_1 \otimes \sum_{j=1}^L \delta_j$  where  $\delta$  denotes the Dirac delta measure. If  $\mathbb{X}$  is a Riemannian manifold, the natural choice for  $\lambda_1$  will be the Riemannian volume form  $V$ . The distance  $\rho$  will be chosen to maintain the topology of the manifold. Letting  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$  denote the space of all densities with respect to  $\lambda$ , we propose the following joint density model

$$(2.1) \quad f(x, y; P, \kappa) = \int_{\mathbb{X} \times S_{L-1}} \nu_y K(x; \mu, \kappa) P(d\mu d\nu), \quad (x, y) \in \mathbb{X} \times \mathbb{Y},$$

where  $\nu = (\nu_1, \dots, \nu_L)' \in S_{L-1}$  is a probability vector on the simplex  $S_{L-1} = \{\nu \in [0, 1]^L : \sum \nu_j = 1\}$ ,  $K(\cdot; \mu, \kappa)$  is a kernel located at  $\mu \in \mathbb{X}$  with precision or inverse-scale  $\kappa \in \mathfrak{R}^+$ , and  $P \in \mathcal{M}(\mathbb{X} \times S_{L-1})$  is a mixing measure, with  $\mathcal{M}(\mathbb{X} \times S_{L-1})$  denoting the space of all probability measures on  $\mathbb{X} \times S_{L-1}$ .

One can interpret this model in the following hierarchical way. Draw  $(\mu, \nu)$  from  $P$ . Given  $(\mu, \nu, \kappa)$ ,  $X$  and  $Y$  are conditionally independent with  $X$  having the

conditional density  $K(\cdot; \mu, \kappa)$  with respect to  $\lambda_1$  and

$$\Pr(Y = l | \mu, \nu, \kappa) = \nu_l, \quad 1 \leq l \leq L.$$

If  $K(\cdot; \mu, \kappa)$  is a valid probability kernel, i.e.

$$\int_{\mathbb{X}} K(x; \mu, \kappa) \lambda_1(dx) = 1, \quad \text{for all } (\mu, \kappa) \in \mathbb{X} \times \mathfrak{R}^+,$$

one can show that  $f(x, y; P, \kappa)$  is a valid probability density with respect to  $\lambda$ .

To justify model (2.1), it is necessary to show flexibility in approximating any joint density in  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$ , and hence in approximating  $p(y, x)$ . Our focus is on nonparametric Bayes methods that place a prior on the joint distribution of the measure  $P$  and the precision  $\kappa$  to induce a prior over  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$ . Flexibility of the model is quantified in terms of the size of the support of this prior. In particular, our goal is to choose a specification that leads to full  $L^\infty$  and Kullback Leibler (KL) support, meaning that the prior assigns positive probability in arbitrarily small neighborhoods of any density in  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$ . This property will not necessarily hold for arbitrarily chosen kernels, and one of our primary theoretical contributions is to provide sufficient conditions under which KL support and posterior consistency hold. This is not just of theoretical interest, as it is important to verify that the model is sufficiently flexible to approximate any classification function, with the accuracy of the estimate improving as the amount of training data grows. This is not automatic for nonparametric models in which there is often concern about over-fitting.

REMARK 2.1. In the joint model (2.1), one may also mix across the precision parameter and make the model more flexible. Posterior computation with such a model is a straight-forward extension of this one which is illustrated in §2.3.

**2.2. Support of the prior and consistency.** Assume that the joint distribution of  $(X, Y)$  has a density  $f_t(x, y) = g_t(x)p_t(y, x)$  with respect to  $\lambda$ , where  $g_t$  is the true marginal density of  $X$  and  $p_t(y, x)$  is the true  $\Pr(Y = y | X = x)$ . For Bayesian inference, we choose a prior  $\Pi_1$  on  $(P, \kappa)$  in (2.1), with one possible choice corresponding to  $DP(w_0 P_0) \otimes \text{Gam}(a, b)$ , with  $DP(w_0 P_0)$  denoting a Dirichlet process prior with precision  $w_0$  and base probability measure  $P_0 \in \mathcal{M}(\mathbb{X} \times S_{L-1})$  and  $\text{Gam}$  denoting the gamma distribution. The prior  $\Pi_1$  induces a corresponding prior  $\Pi$  on the space  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$  through (2.1). Under minor assumptions on  $\Pi_1$  and hence  $\Pi$ , the theorem below shows that the prior probability of any uniform neighborhood of a continuous true density is positive.

THEOREM 2.1. *Under the assumptions*

**A1:**  $K$  is continuous in its arguments,

**A2:** For any continuous function  $\phi$  from  $\mathbb{X}$  to  $\mathfrak{R}$ ,

$$\limsup_{\kappa \rightarrow \infty} \sup_{x \in \mathbb{X}} \left| \phi(x) - \int_{\mathbb{X}} K(x; \mu, \kappa) \phi(\mu) \lambda_1(d\mu) \right| = 0,$$

**A3:** For any  $\kappa > 0$ , there exists  $\tilde{\kappa} \geq \kappa$  such that  $(P_t, \tilde{\kappa}) \in \text{supp}(\Pi_1)$  where  $P_t \in \mathcal{M}(\mathbb{X} \times S_{L-1})$  is defined as

$$P_t(d\mu d\nu) = \sum_{j \in \mathbb{Y}} f_t(\mu, j) V(d\mu) \delta_{e_j}(d\nu),$$

with  $e_j \in \mathbb{R}^L$  denoting a zero vector with a single one in position  $j$ ,

**A4:**  $f_t(\cdot, j)$  is continuous for all  $j \in \mathbb{Y}$ ,

given any  $\epsilon > 0$ ,

$$\Pi\left(\left\{f \in \mathcal{D}(\mathbb{X} \times \mathbb{Y}) : \sup_{x \in \mathbb{X}, y \in \mathbb{Y}} |f(x, y) - f_t(x, y)| < \epsilon\right\}\right) > 0.$$

Assumption **A4** restricts the true conditional density of  $X$  given  $Y = j$  to be continuous for all  $j$ . Assumptions **A1** and **A2** place minor regularity condition on the kernel  $K$ . If  $K(x; \mu, \kappa)$  is symmetric in  $x$  and  $\mu$ , as will be the case in most examples, **A2** implies that  $K(\cdot; \mu, \kappa)$  converges to  $\delta_\mu$  in the weak sense uniformly in  $\mu$  as  $\kappa \rightarrow \infty$ . This justifies the names ‘location’ and ‘precision’ for the parameters. Assumption **A3** provides a minimal condition on the support of the prior for  $(P, \kappa)$ . These assumptions provide general sufficient conditions for the induced prior  $\Pi$  on the joint density of  $(X, Y)$  to have full  $L^\infty$  support.

Although full uniform support is an appealing property, much of the theoretical work on asymptotic properties of nonparametric Bayes estimators relies on KL support. The following corollary shows that KL support follows from **A1-A4** and the additional assumption that the true density is everywhere positive. The proof is very much on the same lines of Corollary 1, [4]. The KL divergence of a density  $f$  from  $f_t$  is defined as  $KL(f_t; f) = \int_{\mathbb{X} \times \mathbb{Y}} f_t \log \frac{f_t}{f} \lambda(dx dy)$ . Given  $\epsilon > 0$ ,  $K_\epsilon(f_t) = \{f : KL(f_t; f) < \epsilon\}$  will denote an  $\epsilon$ -sized KL neighborhood of  $f_t$ . The prior  $\Pi$  is said to satisfy the KL condition at  $f_t$ , or  $f_t$  is said to be in its KL support, if  $\Pi\{K_\epsilon(f_t)\} > 0$  for any  $\epsilon > 0$ .

**COROLLARY 2.2.** *Under assumptions **A1-A4** and*

**A5:**  $f_t(x, y) > 0$  for all  $x, y$ ,

$f_t$  is in the KL support of  $\Pi$ .

Suppose we have an independent and identically distributed (iid) sample  $(\mathbf{x}_n, \mathbf{y}_n) \equiv (x_i, y_i)_{i=1}^n$  from  $f_t$ . Since  $f_t$  is unobserved, we take the likelihood function to be  $\prod_{i=1}^n f(x_i, y_i; P, \kappa)$ . Using the prior  $\Pi$  on  $f$  and the observed sample, we find the posterior distribution of  $f$ , denoted by  $\Pi(\cdot | \mathbf{x}_n, \mathbf{y}_n)$ . Using the Schwartz theorem ([25]), Corollary 2.2 implies weak posterior consistency. This in turn implies that for any measurable subset  $A$  of  $\mathbb{X}$ , with  $\lambda_1(A) > 0$ ,  $\lambda_1(\partial A) = 0$ , and  $y \in \mathbb{Y}$ , the posterior conditional probability of  $Y$  being  $y$  given  $X$  in  $A$  converges to the true conditional probability almost surely. Here  $\partial A$  denotes the boundary of  $A$ .

To give more flexibility to the classification function expression, we may replace the location-scale kernel by some broader family of parametric distributions on  $\mathbb{X}$  such as  $K(\cdot; \mu, \kappa, \Theta)$  with  $\Theta$  denoting the other kernel parameters. Similarly when performing posterior computations, we may set hyperpriors on the parameters of the prior. Then the conclusions of Results 2.1 and 2.2 hold and hence weak consistency follows as long as the assumptions are verified given the hyperparameters over a set of positive prior probability. This is immediate and is verified in Lemma 1, [32].

Under stronger assumptions on the kernel and the prior, we prove strong posterior consistency for the joint model. We will illustrate how these conditions are met for a vMF mixture model for hyperspherical data through Proposition 4.1.

**THEOREM 2.3.** *Under assumptions **A1-A5** and*

**A6:** *There exist positive constants  $\mathcal{K}_1, a_1, A_1$  such that for all  $\mathcal{K} \geq \mathcal{K}_1, \mu, \nu \in \mathbb{X}$ ,*

$$\sup_{m \in M, \kappa \in [0, \mathcal{K}]} |K(m; \mu, \kappa) - K(m; \nu, \kappa)| \leq A_1 \mathcal{K}^{a_1} \rho(\mu, \nu).$$

**A7:** *There exists positive constants  $a_2, A_2$  such that for all  $\kappa_1, \kappa_2 \in [0, \mathcal{K}]$ ,  $\mathcal{K} \geq \mathcal{K}_1$ ,*

$$\sup_{m, \mu \in M} |K(m; \mu, \kappa_1) - K(m; \mu, \kappa_2)| \leq A_2 \mathcal{K}^{a_2} |\kappa_1 - \kappa_2|.$$

**A8:** *There exist positive constants  $a_3, A_3, A_4$  such that given any  $\epsilon > 0$ ,  $M$  can be covered by at-most  $A_3 \epsilon^{-a_3} + A_4$  many subsets of diameter at-most  $\epsilon$ .*

**A9:**  $\Pi_1(\mathcal{M}(M) \times (n^a, \infty))$  *is exponentially small for some  $a < (a_1 a_3)^{-1}$ , the posterior probability of any total variation neighborhood of  $f_t$  converges to 1 almost surely.*

Given the training data, we can classify a new feature based on a draw from the posterior of  $p$  or better still, take  $\hat{p}$  to be its posterior mean. As a corollary to Theorem 2.3, we show that  $\hat{p}$  converges to  $p_t$  in  $L^1$  sense.

COROLLARY 2.4. (a) *Strong consistency for the posterior of  $f$  implies that*

$$(2.2) \quad \Pi\{f : \max_{y \in \mathbb{Y}} \int_{\mathbb{X}} |p(y, x) - p_t(y, x)| g_t(x) \lambda_1(dx) < \epsilon | \mathbf{x}_n, \mathbf{y}_n\}$$

*converges to 1 as  $n \rightarrow \infty$  a.s.. (b) Under assumptions **A4-A5** on  $f_t$ , this implies that*

$$\Pi\{f : \max_{y \in \mathbb{Y}} \int_{\mathbb{X}} |p(y, x) - p_t(y, x)| w(x) \lambda_1(dx) < \epsilon | \mathbf{x}_n, \mathbf{y}_n\}$$

*converges to 1 a.s. for any non-negative function  $w$  with  $\sup_x w(x) < \infty$ .*

REMARK 2.2. Part (a) of Corollary 2.4 holds even when  $\mathbb{X}$  is non-compact. It just needs strong posterior consistency for the joint model.

From part (b) of Corollary 2.4, it would seem intuitive that point-wise posterior consistency can be obtained for the predictive probability function. However, this is not immediate because the convergence rate may depend on the choice of  $w$ .

Assumption **A9** is hard to satisfy, especially when the feature space is high dimensional. This type of problem was mentioned by [33] in a different setting. Then  $a_1$  and  $a_3$  turn out to be very big, so that the prior is required to have very light tails and place small mass at high precisions. This is undesirable in applications and instead we can let  $\Pi_1$  depend on the sample size  $n$  and obtain weak and strong consistency under weaker assumptions.

THEOREM 2.5. *Let  $\Pi_1 = \Pi_{11} \otimes \pi_n$  where  $\pi_n$  is a sequence of densities on  $\mathbb{R}^+$ . Assume the following.*

**A10:** *The prior  $\Pi_{11}$  has full support.*

**A11:** *For any  $\beta > 0$ , there exists a  $\kappa_0 \geq 0$ , such that for all  $\kappa \geq \kappa_0$ ,*

$$\liminf_{n \rightarrow \infty} \exp(n\beta) \pi_n(\kappa) = \infty.$$

**A12:** *For some  $\beta_0 > 0$  and  $a < (a_1 a_3)^{-1}$ ,*

$$\lim_{n \rightarrow \infty} \exp(n\beta_0) \pi_n\{(n^a, \infty)\} = 0.$$

- (a) Under assumptions **A1-A2** on the kernel, **A10-A11** on the prior and **A4-A5** on  $f_t$ , the posterior probability of any weak neighborhood of  $f_t$  converges to one a.s.  
 (b) Under assumptions **A1-A2**, **A4-A8** and **A10-A12**, the posterior probability of any total variation neighborhood of  $f_t$  converges to 1 a.s.

The proof is very similar to that of Theorems 2.6 and 2.9 in [5] and hence is omitted.

With  $\Pi_{11} = DP(\omega_0 P_0)$  and  $\pi_n = Gam(a, b_n)$ , the conditions in Theorem 2.5 are satisfied (for example) when  $P_0$  has full support and  $b_n = b_1 n / \{\log(n)\}^{b_2}$  for any  $b_1, b_2 > 0$ . Then from Corollary 2.4, we have  $L^1$  consistency of the estimated classification function.

**2.3. Computation.** Given the training sample  $(\mathbf{x}_n, \mathbf{y}_n)$ , we classify a new subject based on the predictive probability of allocating it to category  $j$ , which is expressed as

$$(2.3) \quad \Pr(y_{n+1} = j | x_{n+1}, \mathbf{x}_n, \mathbf{y}_n), \quad j \in \mathbb{Y},$$

where  $x_{n+1}$  denotes the feature for the new subject and  $y_{n+1}$  its unknown class label. It follows from Theorem 2.1 and Corollary 2.4 that the classification rule is consistent if the kernel and prior are chosen correctly. Following the recommendation in §2.2, for the prior, we let  $P \sim DP(w_0 P_0)$  independently of  $\kappa \sim \pi$ , with  $P_0 = P_{01} \otimes P_{02}$ ,  $P_{01}$  a distribution on  $\mathbb{X}$ ,  $P_{02}$  a Dirichlet distribution  $\text{Diri}(\mathbf{a})$  ( $\mathbf{a} = (a_1, \dots, a_L)$ ) on  $S_{L-1}$ , and  $\pi$  a base distribution on  $\mathfrak{R}^+$ . Since it is not possible to get a closed form expression for the predictive probability, we need a MCMC algorithm to approximate it.

Using the stick-breaking representation of [26] and introducing cluster allocation indices  $S = (S_1, \dots, S_n)$  ( $S_i \in \{1, \dots, \infty\}$ ), the generative model (2.1) can be expressed in hierarchical form as

$$(2.4) \quad \begin{aligned} x_i &\sim K(\mu_{S_i}, \kappa), & y_i &\sim \text{Multi}(1, \dots, L; \nu_{S_i}), \\ S_i &\sim \sum_{j=1}^{\infty} w_j \delta_{\theta_j}, & \theta_j &= (\mu_j, \nu_j), \end{aligned}$$

where  $w_j = V_j \prod_{h < j} (1 - V_h)$  is the probability that subject  $i$  is allocated to cluster  $S_i = j$ ,  $\theta_j$  is the vector of parameters specific to cluster  $j$ , and  $V_j \sim \text{Beta}(1, w_0)$ ,  $\mu_j \sim P_{01}$  and  $\nu_j \sim \text{Diri}(\mathbf{a})$  are mutually independent for  $j = 1, \dots, \infty$ .

We apply the exact block Gibbs sampler ([34]) for posterior computation, with this approach adapting the blocked Gibbs sampler [20] to avoid truncation. The joint posterior density of  $V = \{V_j\}_{j=1}^{\infty}$ ,  $\theta = \{\theta_j\}_{j=1}^{\infty} = (\mu, \nu)$ ,  $S$  and  $\kappa$  given the training data is proportional to

$$\left\{ \prod_{i=1}^n K(x_i; \mu_{S_i}, \kappa) \nu_{S_i y_i} w_{S_i} \right\} \left\{ \prod_{j=1}^{\infty} \text{Beta}(V_j; 1, w_0) P_{01}(d\mu_j) \text{Diri}(\nu_j; \mathbf{a}) \right\} \pi(\kappa).$$

To avoid the need for posterior computation for infinitely-many unknowns, we introduce slice sampling latent variables  $u = \{u_i\}_{i=1}^n$  drawn iid from  $\text{Unif}(0, 1)$  such that the augmented posterior density is proportional to

$$(2.5) \quad \begin{aligned} \pi(u, V, \theta, S, \kappa | \mathbf{x}_n, \mathbf{y}_n) &\propto \left\{ \prod_{i=1}^n K(x_i; \mu_{S_i}, \kappa) \nu_{S_i y_i} I(u_i < w_{S_i}) \right\} \times \\ &\left\{ \prod_{j=1}^{\infty} \text{Beta}(V_j; 1, w_0) P_{01}(d\mu_j) \text{Diri}(\nu_j; \mathbf{a}) \right\} \pi(\kappa). \end{aligned}$$

Letting  $S_{max} = \max\{S_i\}$ , the conditional posterior distribution of  $\{V_j, \theta_j, j > S_{max}\}$  is the same as the prior, and we can use this to bypass the need for updating infinitely-many unknowns in the Gibbs sampler. After choosing initial values, the sampler iterates through the following steps.

- (1) Update  $S_i$ , for  $i = 1, \dots, n$ , given  $(u, V, \theta, \kappa, \mathbf{x}_n, \mathbf{y}_n)$  by sampling from the multinomial distribution with

$$\Pr(S_i = j) \propto K(x_i; \mu_j, \kappa) \nu_j y_i \text{ for } j \in A_i = \{j : 1 \leq j \leq J, w_j > u_i\},$$

with  $J$  being the smallest index satisfying  $1 - \min(u) < \sum_{j=1}^J w_j$ . In implementing this step, draw  $V_j \sim \text{Beta}(1, w_0)$  and  $\theta_j \sim P_0$  for  $j > S_{max}$  as needed.

- (2) Update the scale parameter  $\kappa$  by sampling from the full conditional posterior which is proportional to

$$\pi(\kappa) \prod_{i=1}^n K(x_i; \mu_{S_i}, \kappa).$$

If direct sampling is not possible, rejection sampling or Metropolis-Hastings (MH) sampling can be used.

- (3) Update the atoms  $\theta_j$ ,  $j = 1, \dots, S_{max}$  from the full conditional posterior distribution, which is equivalent to independently sampling from

$$\begin{aligned} \pi(\mu_j | -) &\propto P_{01}(d\mu_j) \prod_{i:S_i=j} K(x_i; \mu_j, \kappa) \\ (\nu_j | -) &\stackrel{D}{=} \text{Diri}\left(a_1 + \sum_{i:S_i=j} I(y_i = 1), \dots, a_L + \sum_{i:S_i=j} I(y_i = L)\right). \end{aligned}$$

If  $P_{01}$  is not conjugate, then rejection or MH sampling can be used to update  $\mu_j$ .

- (4) Update the stick-breaking random variables  $V_j$ , for  $j = 1, \dots, S_{max}$ , from their conditional posterior distributions given the cluster allocation  $S$  but marginalizing out the slice sampling latent variables  $u$ . In particular,

$$V_j \sim \text{Beta}\left(1 + \sum_i I(S_i = j), w_0 + \sum_i I(S_i > j)\right).$$

- (5) Update the slice sampling latent variables from their conditional posterior by letting

$$u_i \sim \text{Unif}(0, w_{S_i}), \quad i = 1, \dots, n.$$

These steps are repeated a large number of iterations, with a burn-in discarded to allow convergence. Given a draw from the posterior, the predictive probability of allocating a new observation to category  $l$ ,  $l \leq L$ , as defined through (2.3) is proportional to

$$(2.6) \quad \sum_{j=1}^{S_{max}} w_j \nu_{jl} K(x_{n+1}; \mu_j, \kappa) + w_{S_{max}+1} \int_{\mathbb{X} \times S_{L-1}} \nu_l K(x_{n+1}; \mu, \kappa) P_0(d\mu d\nu)$$

where  $w_{S_{max}+1} = 1 - \sum_{j=1}^{S_{max}} w_j$ . We can average these conditional predictive probabilities across the MCMC iterations after burn-in to estimate predictive probabilities. For moderate to large numbers of training samples  $n$ ,  $\sum_{j=1}^{S_{max}} w_j \approx 1$  with

high probability, so that an accurate approximation can be obtained by setting the final term equal to zero and hence bypassing need to calculate the integral.

### 3. Nonparametric Bayes Testing

**3.1. Hypotheses and Bayes factor.** A related problem to classification is testing of differences between groups. In particular, instead of wanting to predict the class label  $y_{n+1}$  for a new subject based on training data  $(\mathbf{x}_n, \mathbf{y}_n)$ , the goal is to test whether the distribution of the features differs across the classes. Although our methods can allow testing of pairwise differences between groups, we focus for simplicity in exposition on the case in which the null hypothesis corresponds to homogeneity across the groups. Formally, the alternative hypothesis  $H_1$  corresponds to any joint density in  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$  excluding densities of the form

$$(3.1) \quad H_0 : f(x, y) = f(x)f(y)$$

for all  $(x, y)$  outside of a  $\lambda$ -null set. Note that model (2.1) will in general assign zero probability to  $H_0$ , and hence is an appropriate model for the joint density under  $H_1$ .

As a model for the joint density under the null hypothesis  $H_0$  in (3.1), we replace  $P(d\mu d\nu)$  in (2.1) with  $P_1(d\mu)P_2(d\nu)$  so that the joint density becomes

$$(3.2) \quad f(x, y; P_1, P_2, \kappa) = g(x; P_1, \kappa)p(y; P_2) \text{ where}$$

$$(3.3) \quad g(x; P_1, \kappa) = \int_{\mathbb{X}} K(x; \mu, \kappa)P_1(d\mu), \quad p(y; P_2) = \int_{S_{L-1}} \nu_y P_2(d\nu).$$

We set priors  $\Pi_1$  and  $\Pi_0$  for the parameters in the models under  $H_1$  and  $H_0$ , respectively. The Bayes factor in favor of  $H_1$  over  $H_0$  is then the ratio of the marginal likelihoods under  $H_1$  and  $H_0$ ,

$$BF(H_1 : H_0) = \frac{\int_{\mathcal{M}(\mathbb{X} \times S_{L-1}) \times \mathbb{R}^+} \prod_{i=1}^n f(x_i, y_i; P, \kappa) \Pi_1(dP d\kappa)}{\int_{\mathcal{M}(\mathbb{X}) \times \mathcal{M}(S_{L-1}) \times \mathbb{R}^+} \prod_{i=1}^n g(x_i; P_1, \kappa) p(y_i; P_2) \Pi_0(dP_1 dP_2 d\kappa)}$$

The priors should be suitably constructed so that we get consistency of the Bayes factor and computation is straightforward and efficient. [2] propose an approach for calculating Bayes factors for comparing Dirichlet process mixture (DPM) models, but their algorithm is quite involved to implement and is limited to DPM models with a single DP prior on an unknown mixture distribution. Simple conditions for consistency of Bayes factors for testing a point null versus a nonparametric alternative have been provided by [8], but there has been limited work on consistency of Bayes tests in more complex cases, such as we are faced with here. An important exception is [16].

The prior  $\Pi_1$  on  $(P, \kappa)$  under  $H_1$  can be constructed as in §2. To choose a prior  $\Pi_0$  for  $(P_1, P_2, \kappa)$  under  $H_0$ , we take  $(P_1, \kappa)$  to be independent of  $P_2$  so that the marginal likelihood becomes a product of the  $X$  and  $Y$  marginals if  $H_0$  is true. Dependence in the priors for the mixing measures would induce dependence between the  $X$  and  $Y$  densities, and it is important to maintain independence under  $H_0$ .

Expression (3.3) suggests that under  $H_0$  the density of  $Y$  depends on  $P_2$  only through the  $L$ -dimensional vector

$$p = (p(1; P_2), p(2; P_2), \dots, p(L; P_2))' \in S_{L-1}.$$

Hence, it is sufficient to choose a prior for  $p$ , such as  $\text{Diri}(\mathbf{b})$  with  $\mathbf{b} = (b_1, \dots, b_L)'$ , instead of specifying a full prior for  $P_2$ . To independently choose a prior for  $(P_1, \kappa)$ ,

we recommend the marginal induced from the prior  $\Pi_1$  on  $(P, \kappa)$  under  $H_1$ . Under this choice, the marginal likelihood under  $H_0$  is

$$(3.4) \quad \int_{\mathcal{M}(\mathbb{X} \times S_{L-1}) \times \mathfrak{R}^+} \prod_{i=1}^n g(x_i; P_1, \kappa) \Pi_1(dP d\kappa) \int_{S_{L-1}} \prod_{j=1}^L p_j^{\sum_{i=1}^n I(y_i=j)} \text{Diri}(dp; \mathbf{b}) \\ = \frac{D(\mathbf{b}_n)}{D(\mathbf{b})} \int_{\mathcal{M}(\mathbb{X} \times S_{L-1}) \times \mathfrak{R}^+} \prod_{i=1}^n g(x_i; P_1, \kappa) \Pi_1(dP d\kappa),$$

with  $\mathbf{b}_n$  being the  $L$ -dimensional vector with  $j^{\text{th}}$  coordinate  $b_j + \sum_{i=1}^n I(y_i = j)$ ,  $1 \leq j \leq L$ ,  $D$  being the normalizing constant for Dirichlet distribution given by  $D(\mathbf{a}) = \frac{\prod_{j=1}^L \Gamma(a_j)}{\Gamma(\sum_{j=1}^L a_j)}$  and  $\Gamma$  denoting the gamma function. The marginal likelihood under  $H_1$  is

$$(3.5) \quad \int \prod_{i=1}^n f(x_i, y_i; P, \kappa) \Pi_1(dP d\kappa).$$

The Bayes factor in favor of  $H_1$  against  $H_0$  is the ratio of the marginal (3.5) over (3.4).

**3.2. Consistency of the Bayes factor.** Let  $\Pi$  be the prior induced on the space of all densities  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$  through  $\Pi_1$ . For any density  $f(x, y)$ , let  $g(x) = \sum_j f(x, j)$  denote the marginal density of  $X$  while  $p(y) = \int_{\mathbb{X}} f(x, y) \lambda_1(dx)$  denotes the marginal probability vector of  $Y$ . Let  $f_t, g_t$  and  $p_t$  be the corresponding values for the true distribution of  $(X, Y)$ . The Bayes factor in favor of the alternative, as obtained in the last section, can be expressed as

$$(3.6) \quad BF = \frac{D(\mathbf{b})}{D(\mathbf{b}_n)} \frac{\int \prod_i f(x_i, y_i) \Pi(df)}{\int \prod_i g(x_i) \Pi(df)}.$$

Theorem 3.1 proves consistency of the Bayes factor at an exponential rate if the alternative hypothesis of dependence holds.

**THEOREM 3.1.** *If  $X$  and  $Y$  are not independent under the true density  $f_t$  and if the prior  $\Pi$  satisfies the KL condition at  $f_t$ , then there exists a  $\beta_0 > 0$  for which  $\liminf_{n \rightarrow \infty} \exp(-n\beta_0) BF = \infty$  a.s.  $f_t^\infty$ .*

**3.3. Computation.** We introduce a latent variable  $z = I(H_1 \text{ is true})$  which takes value 1 if  $H_1$  is true and 0 if  $H_0$  is true. Assuming equal prior probabilities for  $H_0$  and  $H_1$ , the conditional likelihood of  $(\mathbf{x}_n, \mathbf{y}_n)$  given  $z$  is

$$\Pi(\mathbf{x}_n, \mathbf{y}_n | z = 0) = \frac{D(\mathbf{b}_n)}{D(\mathbf{b})} \int \prod_{i=1}^n g(x_i; P_1, \kappa) \Pi_1(dP d\kappa) \text{ and} \\ \Pi(\mathbf{x}_n, \mathbf{y}_n | z = 1) = \int \prod_{i=1}^n f(x_i, y_i; P, \kappa) \Pi_1(dP d\kappa).$$

In addition, the Bayes factor can be expressed as

$$(3.7) \quad BF = \frac{\Pr(z = 1 | \mathbf{x}_n, \mathbf{y}_n)}{\Pr(z = 0 | \mathbf{x}_n, \mathbf{y}_n)}.$$

Next introduce latent parameters  $\mu, \nu, V, S, \kappa$  as in §2.3 such that

$$(3.8) \quad \Pi(\mathbf{x}_n, \mathbf{y}_n, \mu, V, S, \kappa, z = 0) = \frac{D(\mathbf{b}_n)}{D(\mathbf{b})} \pi(\kappa) \prod_{i=1}^n \{w_{S_i} K(x_i; \mu_{S_i}, \kappa)\} \times \prod_{j=1}^{\infty} \{\text{Be}(V_j; 1, w_0) P_{01}(d\mu_j)\},$$

$$(3.9) \quad \Pi(\mathbf{x}_n, \mathbf{y}_n, \mu, \nu, V, S, \kappa, z = 1) = \pi(\kappa) \prod_{i=1}^n \{w_{S_i} \nu_{S_i y_i} K(x_i; \mu_{S_i}, \kappa)\} \times \prod_{j=1}^{\infty} \{\text{Be}(V_j; 1, w_0) P_0(d\mu_j d\nu_j)\}.$$

Marginalize out  $\nu$  from equation (3.9) to get

$$(3.10) \quad \Pi(\mathbf{x}_n, \mathbf{y}_n, \mu, V, S, \kappa, z = 1) = \pi(\kappa) \prod_{j=1}^{\infty} \frac{D(\mathbf{a} + \tilde{a}_j(S))}{D(\mathbf{a})} \times \prod_{i=1}^n \{w_{S_i} K(x_i; \mu_{S_i}, \kappa)\} \prod_{j=1}^{\infty} \{\text{Be}(V_j; 1, w_0) P_{01}(d\mu_j)\},$$

with  $\tilde{a}_j(S)$ ,  $1 \leq j < \infty$  being  $L$ -dimensional vectors with  $l^{\text{th}}$  coordinate  $\sum_{i: S_i=j} I(y_i = l)$ ,  $l \in \mathbb{Y}$ . Integrate out  $z$  by adding equations (3.8) and (3.10) and the joint distribution of  $(\mu, V, S, \kappa)$  given the data becomes

$$(3.11) \quad \Pi(\mu, V, S, \kappa | \mathbf{x}_n, \mathbf{y}_n) \propto \{C_0 + C_1(S)\} \pi(\kappa) \prod_{i=1}^n \{w_{S_i} K(x_i; \mu_{S_i}, \kappa)\} \times \prod_{j=1}^{\infty} \{\text{Be}(V_j; 1, w_0) P_{01}(d\mu_j)\}$$

$$\text{with } C_0 = \frac{D(\mathbf{b}_n)}{D(\mathbf{b})} \text{ and } C_1(S) = \prod_{j=1}^{\infty} \frac{D(\mathbf{a} + \tilde{a}_j(S))}{D(\mathbf{a})}.$$

To estimate the Bayes factor, first make repeated draws from the posterior in (3.11). For each draw, compute the posterior probability distribution of  $z$  from equations (3.8) and (3.10) and take their average after discarding a suitable burn-in. The averages estimate the posterior distribution of  $z$  given the data, from which we can get an estimate for  $BF$  from (3.7). The sampling steps are accomplished as follows

- (1) Update the cluster labels  $S$  given  $(\mu, V, \kappa)$  and the data from their joint posterior which is proportional to

$$(3.12) \quad \{C_0 + C_1(S)\} \pi(\kappa) \prod_{i=1}^n \{w_{S_i} K(x_i; \mu_{S_i}, \kappa)\}.$$

Introduce slice sampling latent variables  $u$  as in §2.3 and replace  $w_{S_i}$  by  $I(u_i < w_{S_i})$  to make the total number of possible states finite. However unlike in §2.3, the  $S_i$ s are no more conditionally independent. We propose to use a Metropolis-Hastings block update step in which a candidate for  $(S_1, \dots, S_n)$ , or some subset of this vector if  $n$  is large, is sampled independently from multinomials with  $\Pr(S_i = j) \propto K(x_i; \mu_j, \kappa)$ , for  $j \in A_i$

where  $A_i = \{j : 1 \leq j \leq J, w_j > u_i\}$  and  $J$  is the smallest index satisfying  $1 - \min(u) < \sum_{j=1}^J w_j$ . In implementing this step, draw  $V_j \sim \text{Be}(1, w_0)$  and  $\mu_j \sim P_{01}$  for  $j > S_{max}$  as needed. The acceptance probability is simply the ratio of  $C_0 + C_1(S)$  calculated for the candidate value and the current value of  $S$ .

- (2) Update  $\kappa, \{\mu_j\}_{j=1}^{S_{max}}, \{V_j\}_{j=1}^{S_{max}}, \{u_i\}_{i=1}^n$  as in Steps (2) - (5) of the algorithm in §2.3.
- (3) Compute the full conditional posterior distribution of  $z$  which is given by

$$\Pr(z|\mu, S, \mathbf{x}_n, \mathbf{y}_n) \propto \begin{cases} \frac{D(\mathbf{b}_n)}{D(\mathbf{b})} & \text{if } z = 0, \\ \prod_{j=1}^{S_{max}} \frac{D(\mathbf{a} + \tilde{a}_j(S))}{D(\mathbf{a})} & \text{if } z = 1. \end{cases}$$

#### 4. Application to the unit sphere $S^d$

**4.1. vMF Kernel Mixture Models.** For classification using predictors  $X$  lying on the hypersphere

$$\mathbb{X} = S^d = \left\{ x \in \mathfrak{R}^{d+1} : \|x\|^2 \equiv \sum_{j=1}^{d+1} x_j^2 = 1 \right\},$$

we recommend using a von Mises-Fisher (vMF) kernel in the mixture model (2.1) to induce a prior over  $\mathcal{D}(S^d \times \mathbb{Y})$ . Although the other distributions on  $S^d$  described in [22] could be used, the vMF kernel provides a relatively simple and computationally tractable choice. As shown in Proposition 4.1, this kernel also satisfies the assumptions in §2.2 for building a flexible joint density model and for posterior consistency. For a proof, see [5].

The vMF distribution has the density ([29], [14], [30])

$$(4.1) \quad \text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu) \quad (x, \mu \in S^d, \kappa \in [0, \infty))$$

with respect to the invariant volume form on  $S^d$ , where

$$c(\kappa) = \frac{2\pi^{d/2}}{\Gamma(\frac{d}{2})} \int_{-1}^1 \exp(\kappa t) (1 - t^2)^{d/2-1} dt$$

is its normalizing constant. This distribution has a unique extrinsic mean (as defined in [6]) of  $\mu$ , thereby  $\mu$  can be interpreted as the kernel location. The parameter  $\kappa$  is a measure of concentration with  $\kappa = 0$  corresponding to the uniform distribution while as  $\kappa$  diverges to  $\infty$ , it converges weakly to  $\delta_\mu$  uniformly in  $\mu$ . Sampling from vMF is straightforward using results in [28] and [31].

**PROPOSITION 4.1.** (a) *The vMF kernel  $K$  as defined in (4.1) satisfies assumptions A1 and A2. It satisfies A6 with  $a_1 = d/2 + 1$  and A7 with  $a_2 = d/2$ . The compact metric-space  $S^d$  endowed with chord distance satisfies A8 with  $a_3 = d$ .*

In the sequel, we will apply the general methods for classification and testing developed earlier in the paper to hyperspherical features.

**4.2. MCMC Details.** When features lie on  $S^d$  and we choose vMF kernels and priors as in §2.3, simplifications result in the MCMC steps for updating  $\kappa$  and  $\mu$ . Letting  $P_{01} = \text{vMF}(\mu_0, \kappa_0)$ , we obtain conjugacy for the full conditional of the kernel locations,

$$(\mu_j | -) \stackrel{\mathcal{D}}{=} \text{vMF}(\bar{\mu}_j, \kappa_j)$$

with  $\bar{\mu}_j = v_j/\|v_j\|$ ,  $\kappa_j = \|v_j\|$ ,  $v_j = \kappa_0\mu_0 + \kappa X_j$  and  $X_j = \sum_i x_i I(S_i = j)$ . As a default, we can set  $\mu_0$  equal to the  $\mathbf{x}_n$  sample extrinsic mean or we can choose  $\kappa_0 = 0$  to induce a uniform  $P_{01}$ . The full posterior of  $\kappa$  is

$$(4.2) \quad \pi(\kappa)c^{-n}(\kappa)\kappa^{-nd/2}e^{n\kappa}\kappa^{nd/2}\exp\left\{-\kappa\left(n - \sum_i x'_i\mu_{S_i}\right)\right\}.$$

If we set

$$(4.3) \quad \pi(\kappa) \propto c^n(\kappa)\kappa^{a+\frac{nd}{2}-1}e^{-\kappa(n+b)}, \quad a, b > 0,$$

the posterior simplifies to

$$(4.4) \quad \text{Gam}\left(a + \frac{nd}{2}, b + n - \sum_i x'_i\mu_{S_i}\right).$$

We can make the MCMC more efficient by marginalizing out  $\mu$  while updating  $\kappa$ . In particular

$$\pi(\kappa|S) \propto c^{-n}(\kappa)\pi(\kappa) \prod_{j:ms(j)>0} c(\|\kappa X_j + \kappa_0\mu_0\|)$$

with  $ms(j) = \sum_i I(S_i = j)$ . This is easy to compute in the case  $d = 2$ ,  $\kappa_0 = 0$  and  $\pi \sim \text{Gam}(a, b)$ . Then it simplifies to

$$\begin{aligned} \pi(\kappa|S) &\propto \text{Gam}\left(\kappa; n - \sum_j I(ms(j) > 0) + a, n - \sum_j \|X_j\| + b\right) \\ &\times \{1 - \exp(-2\kappa)\}^{-n} \prod_{j:ms(j)>0} \{1 - \exp(-2\kappa\|X_j\|)\}. \end{aligned}$$

For this choice, we suggest a Metropolis-Hastings proposal that corresponds to the gamma component. This leads to a high acceptance probability when  $\kappa$  is high, and has good performance in general cases we have considered.

In the predictive probability expression in (2.6), the integral simplifies to

$$\frac{a_j}{\sum a_i} c^{-1}(\kappa)c^{-1}(\kappa_0)c(\|\kappa X_{n+1} + \kappa_0\mu_0\|).$$

## 5. Simulation Examples

**5.1. Classification.** We draw iid samples on  $S^9 \times \mathbb{Y}$ ,  $\mathbb{Y} = \{1, 2, 3\}$  from

$$f_t(x, y) = (1/3) \sum_{l=1}^3 I(y = l) \text{vMF}(x; \mu_l, 200)$$

where  $\mu_1 = (1, 0, \dots)^T$ ,  $\mu_j = \cos(0.2)\mu_1 + \sin(0.2)v_j$ ,  $j = 2, 3$ ,  $v_2 = (0, 1, \dots)^T$  and  $v_3 = (0, 0.5, \sqrt{0.75}, 0, \dots)^T$ . Hence, the three response classes  $y \in \{1, 2, 3\}$  are equally likely and the distribution of the features within each class is a vMF on  $S^9$  with distinct location parameters. We purposely chose the separation between the kernel locations to be small, so that the classification task is challenging.

We implemented the approach described in §2.3 to perform nonparametric Bayes classification. The hyperparameters were chosen to be  $w_0 = 1$ ,  $P_0 = \text{vMF}(\mu_n, 10) \otimes \text{Diri}(1, 1, 1)$ ,  $\mu_n$  being the feature sample extrinsic mean, and  $\pi$  as in (4.3) with  $a = 1$ ,  $b = 0.1$ . Cross-validation is used to assess classification performance, with posterior computation applied to data from a training sample of size 200, and the results used to predict  $y$  given the  $x$  values for subjects in a test sample of size 100. The MCMC algorithm was run for  $5 \times 10^4$  iterations after a

$10^4$  iteration burn-in. Based on examination of trace plots for the predictive probabilities of  $y$  for representative test subjects, the proposed algorithm exhibits good rates of convergence and mixing. Note that we purposely avoid examining trace plots for component-specific parameters due to label switching. The out-of-sample misclassification rates for categories  $y = 1, 2$  and  $3$  were 18.9%, 9.7% and 12.5%, respectively, with the overall rate being 14%.

As an alternative method for flexible model-based classification, we considered a discriminant analysis approach, which models the conditional density of  $x$  given  $y$  as a finite mixture of 10-dimensional Gaussians. In the literature it is very common to treat data lying on a hypersphere as if the data had support in a Euclidean space to simplify the analysis. Using the EM algorithm to fit the finite mixture model, we encountered singularity problems when allowing more than two Gaussian components per response class. Hence, we present the results only for mixtures of one or two multivariate Gaussian components. In the one component case, we obtained class-specific misclassification rates of 27%, 12.9% and 18.8%, with the overall rate being 20%. The corresponding results for the two component mixture were 21.6%, 16.1% and 28.1% with an overall misclassification rate of 22%.

Hence, the results from a parametric Gaussian discriminant analysis and a mixture of Gaussians classifier were much worse than those for our proposed Bayesian nonparametric approach. There are several possible factors contributing to the improvement in performance. Firstly, the discriminant analysis approach requires separate fitting of different mixture models to each of the response categories. When the amount of data in each category is small, it is difficult to reliably estimate all these parameters, leading to high variance and unstable estimates. In contrast our approach of joint modeling of  $f_t$  using a DPM favors a more parsimonious representation. Secondly, inappropriately modeling the data as having support on a Euclidean space has some clear drawbacks. The size of the space over which the densities are estimated is increased from a compact subset  $S^9$  to an unbounded space  $\mathbb{R}^{10}$ . This can lead to an inflated variance and difficulties with convergence of EM and MCMC algorithms. In addition, the properties of the approach are expected to be poor even in larger samples. As Gaussian mixtures give zero probability to the embedded hypersphere, one cannot expect strong posterior consistency.

**5.2. Hypothesis Testing.** We draw an iid sample of size 100 on  $S^9 \times \mathbb{Y}$ ,  $\mathbb{Y} = \{1, 2, 3\}$ , from the distribution

$$f_t(x, y) = (1/3) \sum_{l=1}^3 I(y = l) \sum_{j=1}^3 w_{lj} \text{vMF}(x; \mu_j, 200),$$

where  $\mu_j$ ,  $j = 1, 2, 3$  are as in the earlier example and the weights  $\{w_{lj}\}$  are chosen so that  $w_{11} = 1$  and  $w_{lj} = 0.5$  for  $l = 2, 3$  and  $j = 2, 3$ . Hence, in group  $y = 1$ , the features are drawn from a single vMF density, while in groups  $y = 2$  and  $3$ , the feature distributions are equally weighted mixtures of the same two vMFs.

Letting  $f_j$  denote the conditional density of  $X$  given  $Y = j$  for  $j = 1, 2, 3$ , respectively, the global null hypothesis of no difference in the three groups is  $H_0 : f_1 = f_2 = f_3$ , while the alternative  $H_1$  is that they are not all the same. We set the hyperparameters as  $w_0 = 1$ ,  $P_0 = \text{vMF}(\mu_n, 10) \otimes \text{Diri}(\mathbf{a})$ ,  $\mu_n$  being the X-sample extrinsic mean,  $\mathbf{b} = \mathbf{a} = \hat{p} = (0.28, 0.36, 0.36)$  - the sample proportion of observations from each group, and a prior  $\pi$  on  $\kappa$  as in (4.3) with  $a = 1$  and  $b = 0.1$ .

TABLE 1. Nonparametric Bayes and frequentist test results for data simulated for three groups with the second and third groups identical.

groups	BF	p-value
(1,2,3)	$2.3 \times 10^{15}$	$2 \times 10^{-6}$
(1,2)	$2.4 \times 10^4$	$1.8 \times 10^{-4}$
(1,3)	$1.7 \times 10^6$	$1.5 \times 10^{-5}$
(2,3)	0.235	0.43

We run the proposed MCMC algorithm for calculating the Bayes factor (BF) in favor of  $H_1$  over  $H_0$  for  $6 \times 10^4$  iterations updating cluster labels  $S$  in 4 blocks of 25 each every iteration. The starting value of  $S$  is obtained by the  $k$ -means algorithm ( $k = 10$ ) applied to the  $X$  component of the sample using geodesic distance on  $S^9$  and we started with  $\kappa = 200$ . The trace plots exhibit good rate of convergence of the algorithm. After discarding a burn-in of  $4 \times 10^4$  iterations, the estimated BF was  $2.23 \times 10^{15}$ , suggesting strong evidence in the data in favor of  $H_1$ . We tried multiple starting points and different hyperparameter choices and found the conclusions to be robust, with the estimated BFs not exactly the same but within an order of magnitude. We also obtained similar estimates using substantially shorter and longer chains.

We can also use the proposed methodology for pairwise hypothesis testing of  $H_{0,l,l'} : f_l = f_{l'}$  against the alternative  $H_{1,l,l'} : f_l \neq f_{l'}$  for any two pairs,  $l, l'$ , with  $l \neq l'$ . The analysis is otherwise implemented exactly as in the global hypothesis testing case. The resulting BF in favor of  $H_{1,l,l'}$  over  $H_{0,l,l'}$  for the different possible choices of  $(l, l')$  are shown in Table 1. We obtain very large BFs in testing differences between groups 1 and 2 and 1 and 3, but a moderately small BF for testing a difference between groups 2 and 3, suggesting mild evidence that these two groups are equal. These conclusions are all consistent with the truth. We have noted a general tendency for the BF in favor of the alternative to be large when the alternative is true even in modest sample sizes, suggesting a rapid rate of convergence under the alternative in agreement with our theoretical results. When the null is true, the BF appears to converge to zero based on empirical results in our simulations, but at a slow rate.

For comparison, we also considered a frequentist nonparametric test for detecting differences in the groups based on comparing the sample extrinsic means of the  $f_l$ s. The test statistic used has an asymptotic  $\mathcal{X}_{d(L-1)}^2$  distribution where  $d = 9$  is the feature space dimension and  $L$  is the number of groups that we are comparing. This asymptotic  $\mathcal{X}^2$  test is as in §4.6 of [3], where  $L$  is taken to be 2, but it can be easily generalized to  $L > 2$ . The corresponding p-values are shown in Table 1. The conclusions are all consistent with those from the nonparametric Bayes approach.

**5.3. Testing with No Differences in Mean.** In this example, we draw iid samples on  $S^2 \times \mathbb{Y}$ ,  $\mathbb{Y} = \{1, 2\}$  from the distribution

$$f_t(x, y) = (1/2) \sum_{l=1}^2 I(y = l) \sum_{j=1}^3 w_{lj} \text{vMF}(x; \mu_j, 200),$$

TABLE 2. Nonparametric Bayes and frequentist test results for 10 simulations of 50 observations each for two groups with same population means.

BF	6.1e9	6.4e8	1.3e9	4.3e8	703.1	4.4e7	42.6	4.7e6	1.9e6	379.1
p-value	1.00	0.48	0.31	0.89	0.89	0.49	0.71	0.53	0.56	0.60

where  $w = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \end{bmatrix}$ ,  $\mu_1 = (1, 0, 0)^T$ ,  $\mu_j = \cos(0.2)\mu_1 + \sin(0.2)v_j$  ( $j = 2, 3$ ) and  $v_2 = -v_3 = (0, 1, 0)^T$ . In this case the features are drawn from two groups equally likely, one of them is a vMF, while the other is a equally weighted mixture of two different vMFs. The locations  $\mu_j$  are chosen such that both the groups have the same extrinsic mean  $\mu_1$ .

We draw 10 samples of 50 observations each from the model  $f_t$  and carry out hypothesis testing to test for association between  $X$  and  $Y$  via our method and the frequentist one. The prior, hyperparameters and the algorithm for Bayes Factor (BF) computation are as in the earlier example. In each case we get insignificant p-values, often over 0.5, but very high BFs, often exceeding  $10^6$ . The values are listed in Table 2.

The reason for the failure of the frequentist test is because it relies on comparing the group specific sample extrinsic means and in this example the difference between them is little. Our method on the other hand compares the full conditionals and hence can detect differences that are not in the means.

## 6. Applications

**6.1. Magnetization direction data.** In this example from [12], measurements of remanent magnetization in red silts and claystones were made at 4 locations. This results in samples from four group of directions on the sphere  $S^2$ , the sample sizes are 36, 39, 16 and 16. The goal is to compare the magnetization direction distributions across the groups and test for any significant difference. Figure 1 shows the 3D plot of the sample clouds. The plot suggests no major differences. To test that statistically, we calculate the Bayes factor (BF) in favor of the alternative, as in §5.2. As mixing was not quite as good as in the simulated examples, we implemented label switching moves. We updated the cluster configurations in two blocks of size 54 and 53. The estimated BF was  $\approx 1$ , suggesting no evidence in favor of the alternative hypothesis that the distribution of magnetization directions vary across locations.

To assess sensitivity to the prior specification, we repeated the analysis with different hyperparameter values of  $\mathbf{a}$ ,  $\mathbf{b}$  equal to the proportions of samples within each group and  $P_{01}$  corresponding to an uniform on the sphere. In addition, we tried different starting clusterings in the data, with a default choice obtained by implementing k-means with 10 clusters assumed. In each case, we obtain  $\text{BF} \approx 1$ , so the results were robust.

In Example 7.7 of [13], a coordinate-based parametric test was conducted to compare mean direction in these data, producing a p-value of  $1 - 1.4205 \times 10^{-5}$  based on a  $\mathcal{X}_6^2$  statistic. They also compared the mean directions for the first two groups and obtained a non-significant p-value. Repeating this two sample test using our Bayesian nonparametric method, we obtained a Bayes factor of 1.00. The

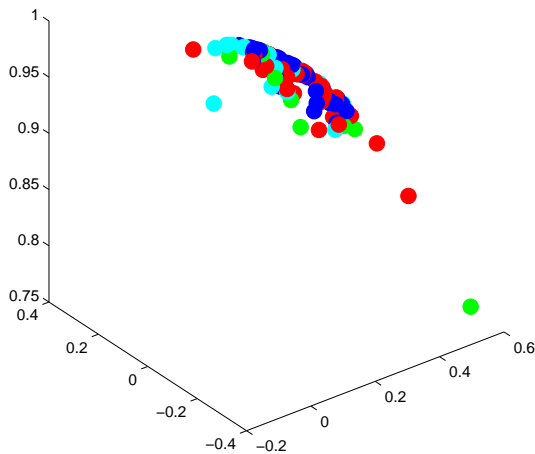


FIGURE 1. 3D coordinates of 4 groups in §6.1: 1(r), 2(b), 3(g), 4(c).

nonparametric frequentist test from §5.2 yield p-values of 0.06 and 0.38 for the two tests.

**6.2. Volcano location data.** The NOAA National Geophysical Data Center Volcano Location Database contains information on locations and characteristics of volcanoes across the globe. The locations using latitude-longitude coordinates are plotted in Figure 2. We are interested in testing if there is any association between the location and type of the volcano. We consider the most common three types which are Strato, Shield and Submarine volcanoes, with data available for 999 volcanoes of these types worldwide. Their location coordinates are shown in Figure 3. Denoting by  $X$  the volcano location which lies on  $S^2$  and by  $Y$  its type which takes values from  $\mathbb{Y} = \{1, 2, 3\}$ , we compute the Bayes factor (BF) for testing if  $X$  and  $Y$  are independent.

As should be apparent from the Figures, the volcano data are particularly challenging in terms of density estimation because the locations tend to be concentrated along fault lines. Potentially, data on distance to the closest fault could be utilized to improve performance, but we do not have access to such data. Without such information, the data present a challenging test case for the methodology in that it is clear that one may need to utilize very many vMF kernels to accurately characterize the density of volcano locations across the globe, with the use of moderate to large numbers of kernels leading to challenging mixing issues. Indeed, we did encounter a sensitivity to the starting cluster configuration in our initial analyses.

We found that one of issues that exacerbated the problem with mixing of the cluster allocation was the ordering in the weights on the stick-breaking representation utilized by the exact block Gibbs sampler. Although label switching moves can lead to some improvements, they proved to be insufficient in this case. Hence,

we modified the computational algorithm slightly to instead use the finite Dirichlet approximation to the Dirichlet process proposed in [21]. The finite Dirichlet treats the components as exchangeable so eliminates sensitivity to the indices on the starting clusters, which we obtained using  $k$ -means for 50 clusters. We used  $K = 50$  as the dimension of the finite Dirichlet and hence the upper bound on the number of occupied clusters. Another issue that lead to mixing problems was the use of a hyperprior on  $\kappa$ . In particular, when the initial clusters were not well chosen, the kernel precision would tend to drift towards smaller than optimal values and as a result too few clusters would be occupied to adequately fit the data. We did not observe such issues at all in a variety of other simulated and real data applications, but the volcano data are particularly difficult as we note above.

To address this second issue, we chose the kernel precision parameter  $\kappa$  by cross-validation. In particular, we split the sample into training and test sets, and then ran our Bayesian nonparametric analysis on the training data separately for a wide variety of  $\kappa$  values between 0 and 1,000. We chose the value that produced the highest expected posterior log likelihood in the test data, leading to  $\hat{\kappa} = 80$ . In this analysis and the subsequent analyses for estimating the BF, we chose the prior on the mixture weights to be  $\text{Diri}(w_0/K\mathbf{1}_K)$  ( $K = 50$ ). The other hyper-parameters were chosen to be  $w_0 = 1$ ,  $\mathbf{a} = \mathbf{b} = (0.71, 0.17, 0.11)$  = the sample proportion of different volcano types,  $\kappa_0 = 10$ , and  $\mu_0$  = the  $X$ -sample extrinsic mean. We collected  $5 \times 10^4$  MCMC iterations after discarding a burn-in of  $10^4$ . Using a fixed band-width considerably improved the algorithm convergence rate.

Based on the complete data set of 999 volcanoes, the resulting BF in favor of the alternative was estimated to be over  $10^{100}$ , providing conclusive evidence that the different types of volcanos have a different spatial distribution across the globe. For the same fixed  $\hat{\kappa}$  value, we reran the analysis for a variety of alternative hyperparameter values and different starting points, obtaining similar BF estimates and the same conclusion. We also repeated the analysis for a randomly selected subsample of 300 observations, obtaining  $\text{BF} = 5.4 \times 10^{11}$ . The testing is repeated for other sub-samples, each resulting in a very high BF. We also obtained a high BF in repeating the analysis with a hyperprior on  $\kappa$ .

For comparison, we perform the asymptotic  $\chi^2$  test as described in §5.2, obtaining a p-value of  $3.6 \times 10^{-7}$  which again favors  $H_1$ . The large sample sizes for the three types (713,172,114) justifies the use of asymptotic theory. However given that the volcanoes are spread all over the globe, the validity of the assumption that the three conditionals have unique extrinsic means may be questioned.

We also perform a coordinate based test by comparing the means of the latitude longitude coordinates of the three sub-samples using a  $\chi^2$  statistic. The three coordinate means are (12.6, 27.9), (21.5, 9.2), and (9.97, 21.5) (latitude, longitude). The value of the statistic is 17.07 and the asymptotic p-value equals  $1.9 \times 10^{-3}$  which is larger by orders of magnitude than its coordinate-free counterpart, but still significant. Coordinate based methods, however, can be very misleading because of the discontinuity at the boundaries. They heavily distort the geometry of the sphere which is evident from the figures.

## 7. Discussion

We have proposed a novel Bayesian approach for classification and testing relying on modeling the joint distribution of the categorical response and continuous

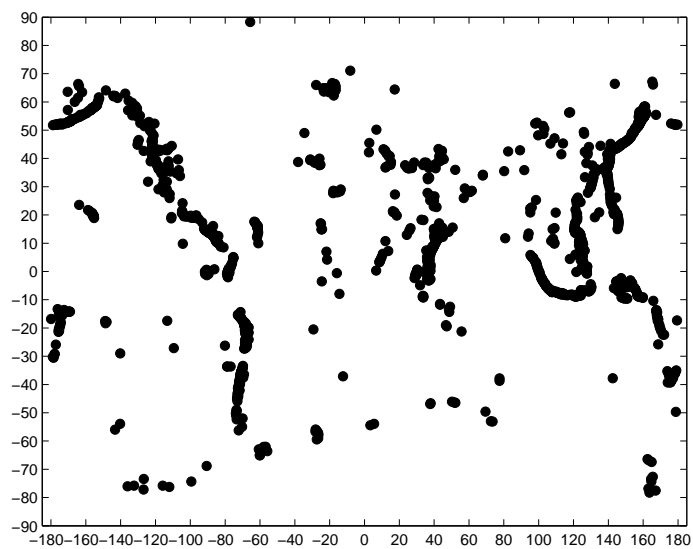


FIGURE 2. Longitude-Latitude coordinates of volcano locations in §6.2.

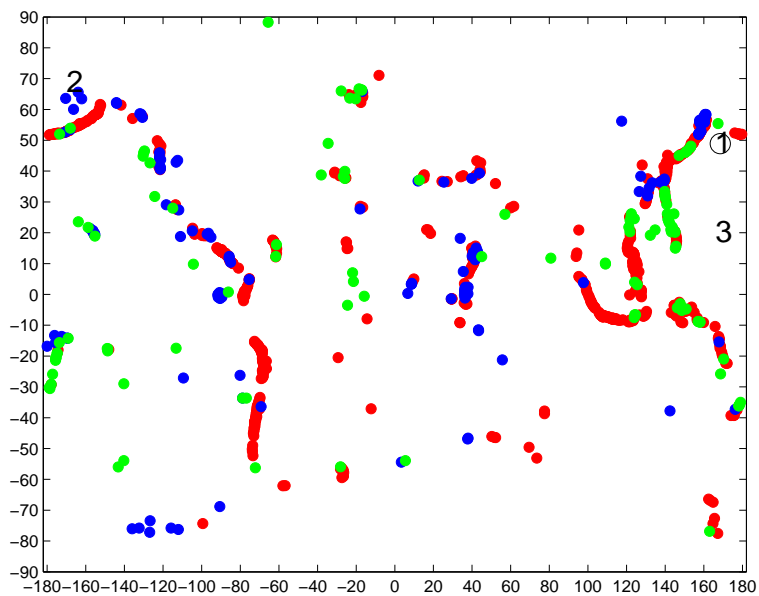


FIGURE 3. Coordinates of 3 major type volcano locations: Strato(r), Shield(b), Submarine(g). Their sample extrinsic mean locations: 1, 2, 3. Full sample extrinsic mean: o

predictors as a Dirichlet process product mixture. The product mixture likelihood includes a multinomial for the categorical response and an arbitrary kernel for the predictors, with dependence induced through the DP prior on the unknown joint mixing measure. By modifying the kernel for the predictors, one can modify the support, with multivariate Gaussian kernels for predictors in  $\mathbb{R}^p$  and von Mises Fisher kernels for predictors on the hypersphere. For other predictor spaces, one can appropriately modify the kernel.

Although our focus has been on hyperspherical predictors for concreteness, the proposed product mixture formulation is broadly applicable to classification problems for predictors in general spaces and we can easily consider predictors having a variety of supports. For example, some predictors can be in a Euclidean space and some on a hypersphere. The framework has some clear practical advantages over frequentist and nonparametric Bayes discriminant analysis approaches, which rely on separately modeling the conditional distributions of the feature (predictor) distributions specific to each response category. In particular, those approaches require substantial training data in each response category for learning of all the conditional distributions. Potentially, this can be addressed by borrowing of information via a dependent Dirichlet process as proposed in [9]. However, our approach bypasses the substantial complication of explicitly modeling differences in a collection of unknown densities.

One of our primary contributions was showing theoretical properties, including large support and posterior consistency, in modeling of the classification function.

In addition, we have added to the underdeveloped literature on nonparametric Bayes testing of differences in distributions, not only on  $\mathbb{R}^p$  but on more general manifolds. We provide a novel computational approach for estimating Bayes factors as well as prove theoretical results on Bayes factor consistency. The proposed method can be implemented in broad applications for testing differences between groups.

## 8. Appendix

**8.1. Proof of Theorem 2.1.** Before proving the Theorem, we prove the following Lemma.

LEMMA 8.1. *Under assumptions A2 and A4,*

$$\lim_{\kappa \rightarrow \infty} \sup \{ |f(x, y; P_t, \kappa) - f_t(x, y)| : (x, y) \in \mathbb{X} \times \mathbb{Y} \} = 0.$$

PROOF. From the definition of  $P_t$ , we can write

$$f(x, y; P_t, \kappa) = \int_{\mathbb{X}} K(x; \mu, \kappa) \phi_y(\mu) \lambda_1(d\mu)$$

for  $\phi_y(\mu) = f_t(\mu, y)$ . Then from A4, it follows that  $\phi_y$  is continuous for all  $y \in \mathbb{Y}$ . Hence from A2, it follows that

$$\lim_{\kappa \rightarrow \infty} \sup_{x \in \mathbb{X}} \left| f_t(x, y) - \int_{\mathbb{X}} K(x; \mu, \kappa) f_t(\mu, y) \lambda_1(d\mu) \right| = 0$$

for any  $y \in \mathbb{Y}$ . Since  $\mathbb{Y}$  is finite, the proof is complete.  $\square$

PROOF OF THEOREM 2.1. Throughout this proof we will view  $\mathbb{X}$  as a compact metric space and  $\mathcal{M}(\mathbb{X} \times S_{L-1})$  as a topological space under the weak topology. From Lemma 8.1, it follows that there exists a  $\kappa_t \equiv \kappa_t(\epsilon) > 0$  such that

$$(8.1) \quad \sup_{x,y} |f(x,y; P_t, \kappa) - f_t(x,y)| < \frac{\epsilon}{3}$$

for all  $\kappa \geq \kappa_t$ . From assumption **A3**, it follows that by choosing  $\kappa_t$  sufficiently large, we can ensure that  $(P_t, \kappa_t) \in \text{supp}(\Pi_1)$ . From assumption **A1**, it follows that  $K$  is uniformly continuous at  $\kappa_t$ , i.e. there exists an open set  $W(\epsilon) \subseteq \mathfrak{R}^+$  containing  $\kappa_t$  s.t.

$$\sup_{x,\mu \in \mathbb{X}} |K(x; \mu, \kappa) - K(x; \mu, \kappa_t)| < \frac{\epsilon}{3} \quad \forall \kappa \in W(\epsilon).$$

This in turn implies that, for all  $\kappa \in W(\epsilon)$ ,  $P \in \mathcal{M}(\mathbb{X} \times S_{L-1})$ ,

$$(8.2) \quad \sup_{x,y} |f(x,y; P, \kappa) - f(x,y; P, \kappa_t)| < \frac{\epsilon}{3}$$

because the left expression in (8.2) is

$$\sup_{x,y} \left| \int \nu_y \{K(x; \mu, \kappa) - K(x; \mu, \kappa_t)\} P(d\mu d\nu) \right| \leq \sup_{x,\mu \in \mathbb{X}} |K(x; \mu, \kappa) - K(x; \mu, \kappa_t)|.$$

Since  $\mathbb{X}$  is compact and  $K(\cdot, \cdot, \kappa_t)$  is uniformly continuous on  $\mathbb{X} \times \mathbb{X}$ , we can cover  $\mathbb{X}$  by finitely many open sets:  $U_1, \dots, U_K$  s.t.

$$(8.3) \quad \sup_{\mu \in \mathbb{X}, x, \tilde{x} \in U_i} |K(x; \mu, \kappa_t) - K(\tilde{x}; \mu, \kappa_t)| < \frac{\epsilon}{12}$$

for each  $i \leq K$ . For fixed  $x, y, \kappa$ ,  $f(x, y; P, \kappa)$  is a continuous function of  $P$ . Hence for  $x_i \in U_i$ ,  $y = j \in \mathbb{Y}$ ,

$$\mathcal{W}_{ij}(\epsilon) = \{P \in \mathcal{M}(\mathbb{X} \times S_{L-1}) : |f(x_i, j; P, \kappa_t) - f(x_i, j; P_t, \kappa_t)| < \frac{\epsilon}{6}\},$$

$1 \leq i \leq K$ ,  $1 \leq j \leq L$ , define open neighborhoods of  $P_t$ . Let  $\mathcal{W}(\epsilon) = \bigcap_{i,j} \mathcal{W}_{ij}(\epsilon)$  which is also an open neighborhood of  $P_t$ . For a general  $x \in \mathbb{X}$ ,  $y \in \mathbb{Y}$ , find  $U_i$  containing  $x$ . Then for any  $P \in \mathcal{W}(\epsilon)$ ,

$$(8.4) \quad \begin{aligned} & |f(x,y; P, \kappa_t) - f(x,y; P_t, \kappa_t)| \leq \\ & |f(x,y; P, \kappa_t) - f(x_i, j; P, \kappa_t)| + |f(x_i, j; P, \kappa_t) - f(x_i, j; P_t, \kappa_t)| \\ & + |f(x_i, j; P_t, \kappa_t) - f(x,y; P_t, \kappa_t)|. \end{aligned}$$

Denote the three terms to the right in (8.4) as  $T_1$ ,  $T_2$  and  $T_3$ . Since  $x \in U_i$ , it follows from (8.3) that  $T_1, T_3 < \frac{\epsilon}{12}$ . Since  $P \in \mathcal{W}_{ij}(\epsilon)$ ,  $T_2 < \frac{\epsilon}{6}$  by definition of  $\mathcal{W}_{ij}(\epsilon)$ . Hence  $\sup_{x,y} |f(x,y; P, \kappa_t) - f(x,y; P_t, \kappa_t)| < \frac{\epsilon}{3}$ . Therefore

$$\mathcal{W}_2(\epsilon) \equiv \{P : \sup_{x,y} |f(x,y; P, \kappa_t) - f(x,y; P_t, \kappa_t)| < \frac{\epsilon}{3}\}$$

contains  $\mathcal{W}(\epsilon)$ . Since  $(P_t, \kappa_t) \in \text{supp}(\Pi_1)$  and  $\mathcal{W}_2(\epsilon) \times W(\epsilon)$  contains an open neighborhood of  $(P_t, \kappa_t)$ , therefore

$$\Pi_1(\mathcal{W}_2(\epsilon) \times W(\epsilon)) > 0.$$

Let  $(P, \kappa) \in \mathcal{W}_2(\epsilon) \times W(\epsilon)$ . Then for  $(x, y) \in \mathbb{X} \times \mathbb{Y}$ ,

$$(8.5) \quad \begin{aligned} & |f(x,y; P, \kappa) - f_t(x,y)| \leq \\ & |f(x,y; P, \kappa) - f(x,y; P, \kappa_t)| + |f(x,y; P, \kappa_t) - f(x,y; P_t, \kappa_t)| \\ & + |f(x,y; P_t, \kappa_t) - f_t(x,y)|. \end{aligned}$$

The first term to the right in (8.5) is  $< \frac{\epsilon}{3}$  since  $\kappa \in W(\epsilon)$ . The second one is  $< \frac{\epsilon}{3}$  because  $P \in \mathcal{W}_2(\epsilon)$ . The third one is also  $< \frac{\epsilon}{3}$  which follows from equation (8.1). Therefore

$$\Pi_1 \left( \{(P, \kappa) : \sup_{x,y} |f(x, y; P, \kappa) - f_t(x, y)| < \epsilon\} \right) > 0.$$

This completes the proof.  $\square$

**8.2. Proof of Theorem 2.3.** The proof uses Proposition 8.2 proved in [5]. Let  $M$  be a compact metric-space. Denote by  $\mathcal{D}(M)$  the space of all probability densities on  $M$  with respect to some fixed finite base measure  $\tau$ . Endow it with the total variation distance  $\|\cdot\|$ . Let  $\mathbf{z}_n = \{z_i\}_1^n$  be a iid sample from some density  $f_t$  on  $M$ . Consider a collection of mixture densities on  $M$  given by

$$(8.6) \quad f(m; P, \kappa) = \int_M K(m; \mu, \kappa) P(d\mu), \quad m \in M, \quad \kappa \in \mathbb{R}^+, \quad P \in \mathcal{M}(M)$$

with  $\int_M K(m; \mu, \kappa) \tau(dm) = 1$ . Set a prior  $\Pi_1$  on  $\mathcal{M}(M) \times \mathbb{R}^+$  which induces a prior  $\Pi$  on  $\mathcal{D}(M)$  through (8.6). For  $\mathcal{F} \subseteq \mathcal{D}(M)$  and  $\epsilon > 0$ , the  $L_1$ -metric entropy  $N(\epsilon, \mathcal{F})$  is defined as the logarithm of the minimum number of  $\epsilon$ -sized (or smaller)  $L_1$  subsets needed to cover  $\mathcal{F}$ .

PROPOSITION 8.2. *For a positive sequence  $\{\kappa_n\}$  diverging to  $\infty$ , define*

$$\mathcal{D}_n = \{f(\cdot; P, \kappa) : P \in \mathcal{M}(M), \quad \kappa \in [0, \kappa_n]\}.$$

(a) *Under assumptions A6-A8, given any  $\epsilon > 0$ , for  $n$  sufficiently large,  $N(\epsilon, \mathcal{D}_n) \leq C(\epsilon) \kappa_n^{a_1 a_3}$  for some  $C(\epsilon) > 0$ . (b) Under further assumption A9, the posterior probability of any total variation neighborhood of  $f_t$  converges to 1 a.s.  $f_t$  if  $f_t$  is in the KL support of  $\Pi$ .*

PROOF OF THEOREM 2.3. For a density  $f \in \mathcal{D}(\mathbb{X} \times \mathbb{Y})$ , let  $p(y)$  be the marginal probability of  $Y = y$  and  $g(x, y)$  be the conditional density of  $X = x$  given  $Y = y$ . Then  $f(x, y) = p(y)g(x, y)$ . For  $f_1, f_2 \in \mathcal{D}(\mathbb{X} \times \mathbb{Y})$ ,

$$(8.7) \quad \begin{aligned} \|f_1 - f_2\| &= \int |f_1(x, y) - f_2(x, y)| \lambda(dx dy) = \sum_{j=1}^L |p_1(j)g_1(x, j) - p_2(j)g_2(x, j)| \lambda_1(dx) \\ &\leq \max_j \|g_1(\cdot, j) - g_2(\cdot, j)\| + \sum_j |p_1(j) - p_2(j)|. \end{aligned}$$

Hence an  $\epsilon$  diameter ball in  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$  contains the intersection of  $L$  many  $\epsilon/2$  diameter balls from  $\mathcal{D}(\mathbb{X})$  with a  $\epsilon/2$  diameter subset of  $S_{L-1}$ . For any  $f(\cdot; P, \kappa)$  as in (2.1), its X-conditional  $g(\cdot; j)$  for  $j \in \mathbb{Y}$  can be expressed as

$$g(x, j) = \frac{\int_{\mathbb{X} \times S_{L-1}} \nu_j K(x; \mu, \kappa) P(d\mu d\nu)}{\int_{\mathbb{X} \times S_{L-1}} \nu_j P(d\mu d\nu)} = \int_{\mathbb{X}} K(x; \mu, \kappa) P_j(d\mu)$$

$$\text{with } P_j(d\mu) = \frac{\int_{S_{L-1}} \nu_j P(d\mu d\nu)}{\int_{\mathbb{X} \times S_{L-1}} \nu_j P(d\mu d\nu)}.$$

Hence  $g(\cdot, j)$  is as in (8.6) with  $M = \mathbb{X}$ . Define

$$\mathcal{D}_n = \{f(\cdot; P, \kappa) : P \in \mathcal{M}(\mathbb{X} \times \mathbb{Y}), \quad \kappa \in [0, \kappa_n]\}. \text{ Then}$$

$$(8.8) \quad \mathcal{D}_n = \{f \in \mathcal{D}(\mathbb{X} \times \mathbb{Y}) : g(\cdot; j) \in \tilde{\mathcal{D}}_n \forall j \in \mathbb{Y}\}$$

where  $\tilde{\mathcal{D}}_n = \{g(\cdot; P, \kappa) : P \in \mathcal{M}(\mathbb{X}), \kappa \in [0, \kappa_n]\}$ .

From Proposition 8.2(a),  $N(\epsilon, \tilde{\mathcal{D}}_n)$  is of order at-most  $\kappa_n^{a_1 a_3}$  and hence from (8.7) and (8.8),  $N(\epsilon, \mathcal{D}_n) \leq C \kappa_n^{a_1 a_3}$ ,  $C$  depending on  $\epsilon$ . Therefore from part (b) of the Proposition, under assumptions **A1-A9**, strong posterior consistency follows.  $\square$

### 8.3. Proof of Corollary 2.4.

PROOF. (a) For any  $y \in \mathbb{Y}$ ,

$$\begin{aligned} & \int_{\mathbb{X}} |p(y, x) - p_t(y, x)| g_t(x) \lambda_1(dx) = \\ & \int_{\mathbb{X}} |f_t(x, y) - f(x, y) + p(y, x)g(x) - p(y, x)g_t(x)| \lambda_1(dx) \\ & \leq \int_{\mathbb{X}} |f_t(x, y) - f(x, y)| \lambda_1(dx) + \int_{\mathbb{X}} |g_t(x) - g(x)| \lambda_1(dx) \\ & \leq 2 \sum_{j=1}^L \int_{\mathbb{X}} |f(x, j) - f_t(x, j)| \lambda_1(dx) = \|f - f_t\|_1 \end{aligned}$$

and hence any neighborhood of  $p_t$  of the form  $\{p : \max_{y \in \mathbb{Y}} \int_{\mathbb{X}} |p(y, x) - p_t(y, x)| g_t(x) \lambda_1(dx) < \epsilon\}$  contains an  $L^1$  neighborhood of  $f_t$ . Now part(a) follows from strong consistency of the posterior distribution of  $f$ .

(b) Since  $\mathbb{X}$  is compact,  $f_t$  being continuous and positive implies that  $c = \inf_{x \in \mathbb{X}} g_t(x) > 0$ . Hence

$$\int_{\mathbb{X}} |p(y, x) - p_t(y, x)| w(x) \lambda_1(dx) \leq c^{-1} \sup(w(x)) \int_{\mathbb{X}} g_t(x) |p(y, x) - p_t(y, x)| \lambda_1(dx)$$

Now the result follows from part (a).  $\square$

**8.4. Proof of Theorem 3.1.** The proof uses Lemma 8.3. This lemma is fundamental to proving weak posterior consistency using the Schwartz theorem and its proof can be found in any standard text which contains the theorem's proof.

LEMMA 8.3. (a) If  $\Pi$  includes  $f_t$  in its KL support, then

$$\liminf_{n \rightarrow \infty} \exp(n\beta) \int \prod_i \frac{f(x_i, y_i)}{f_t(x_i, y_i)} \Pi(df) = \infty$$

a.s.  $f_t^\infty$  for any  $\beta > 0$ . (b) If  $U$  is a weak open neighborhood of  $f_t$  and  $\Pi_0$  is a prior on  $\mathcal{D}(\mathbb{X} \times \mathbb{Y})$  with support in  $U^c$ , then there exists a  $\beta_0 > 0$  for which

$$\lim_{n \rightarrow \infty} \exp(n\beta_0) \int \prod_i \frac{f(x_i, y_i)}{f_t(x_i, y_i)} \Pi_0(df) = 0$$

a.s.  $f_t^\infty$ .

PROOF OF THEOREM 3.1. Express  $BF$  as

$$BF = \left\{ \prod_i p_t(y_i) \right\} \frac{D(\mathbf{b}) \int \prod_i \frac{f(x_i, y_i)}{f_t(x_i, y_i)} \Pi(df)}{D(\mathbf{b}_n) \int \prod_i \frac{g(x_i) p_t(y_i)}{f_t(x_i, y_i)} \Pi(df)} = T_1 T_2 / T_3$$

with  $T_1 = \{\prod_i p_t(y_i)\} \frac{D(\mathbf{b})}{D(\mathbf{b}_n)}$ ,  $T_2 = \int \prod_i \frac{f(x_i, y_i)}{f_t(x_i, y_i)} \Pi(df)$  and  $T_3 = \int \prod_i \frac{g(x_i) p_t(y_i)}{f_t(x_i, y_i)} \Pi(df)$ . Since  $\Pi$  satisfies the KL condition, Lemma 8.3(a) implies that  $\liminf_{n \rightarrow \infty} \exp(n\beta) T_2 = \infty$  a.s. for any  $\beta > 0$ .

Let  $U$  be the space of all dependent densities, that is

$$U^c = \{f \in \mathcal{D}(\mathbb{X} \times \mathbb{Y}) : f(x, y) = g(x)p(y) \text{ a.s. } \lambda(dxdy)\}.$$

The prior  $\Pi$  induces a prior  $\Pi_0$  on  $U^c$  via  $f \mapsto \{\sum_j f(\cdot, j)\} p_t$  and  $T_3$  can be expressed as  $\int \prod_i \frac{f(x_i, y_i)}{f_t(x_i, y_i)} \Pi_0(df)$ . It is easy to show that  $U$  is open under the weak topology and hence under  $H_1$  is a weak open neighborhood of  $f_t$ . Then using Lemma 8.3(b), it follows that  $\lim_{n \rightarrow \infty} \exp(n\beta_0) T_3 = 0$  a.s. for some  $\beta_0 > 0$ .

The proof is complete if we can show that  $\liminf_{n \rightarrow \infty} \exp(n\beta) T_1 = \infty$  a.s. for any  $\beta > 0$  or  $\log(T_1) = o(n)$  a.s. For a positive sequence  $a_n$  diverging to  $\infty$ , the Stirling's formula implies that  $\log \Gamma(a_n) = a_n \log(a_n) - a_n + o(a_n)$ . Express  $\log(T_1)$  as

$$(8.9) \quad \sum_i \log(p_t(y_i)) - \log(D(\mathbf{b}_n)) + o(n).$$

Since  $p_t(j) > 0 \forall j \leq L$ , by the SLLN,

$$(8.10) \quad \sum_i \log(p_t(y_i)) = n \sum_j p_t(j) \log(p_t(j)) + o(n) \text{ a.s.}$$

Let  $b_{nj} = b_j + \sum_i I(y_i = j)$  be the  $j$ th component of  $\mathbf{b}_n$ . Then  $\lim_{n \rightarrow \infty} b_{nj}/n = p_t(j)$ , that is  $b_{nj} = np_t(j) + o(n)$  a.s. and hence the Stirling's formula implies that

$$\begin{aligned} \log(\Gamma(b_{nj})) &= b_{nj} \log(b_{nj}) - b_{nj} + o(n) \\ &= np_t(j) \log(p_t(j)) - np_t(j) + \log(n) b_{nj} + o(n) \text{ a.s.} \end{aligned}$$

which implies

$$(8.11) \quad \begin{aligned} \log(D(\mathbf{b}_n)) &= \sum_{j=1}^L \log(\Gamma(b_{nj})) - \log \Gamma(\sum_j b_j + n) \\ &= n \sum_j p_t(j) \log(p_t(j)) + o(n) \text{ a.s.} \end{aligned}$$

From (8.9), (8.10) and (8.11),  $\log(T_1) = o(n)$  a.s. and this completes the proof.  $\square$

## References

- [1] A. Banerjee, I.S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Jour. Machine Learning Res.*, 6:1345–1382, 2005.
- [2] S. Basu and S. Chib. Marginal likelihood and Bayes factors for D-irichlet process mixture models. *Jour. of Amer. Statist. Assoc.*, 98:224–235, 2003.
- [3] A. Bhattacharya. *Nonparametric Statistics on Manifolds with Applications to Shape Space*. 2008. PhD Thesis.
- [4] A. Bhattacharya and D. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 2010. In Press.
- [5] A. Bhattacharya and D. Dunson. Strong consistency of nonparametric Bayes density estimation on compact metric spaces. *Duke Discussion Paper*, 2010.
- [6] R. N. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Ann. Statist.*, 31:1–29, 2003.
- [7] J. Bigelow and D. Dunson. Bayesian semiparametric joint models for functional predictors. *J. Am. Statist. Ass.*, 104:26–36, 2009.
- [8] S.C. Dass and J. Lee. A note on the consistency of bayes factors for testing point null versus non-parametric alternatives. *J. Statist. Plan. Infer.*, 119:143–152, 2004.

- [9] R. De la Cruz-Mesia, F.A. Quintana, and P. Müller. Semiparametric bayesian classification with longitudinal markers. *Applied Statist.*, 56:119–137, 2007.
- [10] D.B. Dunson. Multivariate kernel partition process mixtures. *Statistica Sinica*, 20:1395–1422, 2010.
- [11] D.B. Dunson and S.D. Peddada. Bayesian nonparametric inference on stochastic ordering. *Biometrika*, 95:859–874, 2008.
- [12] B.J.J. Embleton and K.L. McDonnell. Magnetostratigraphy in the Sydney Basin, SouthEastern Australia. *J. Geomag. Geoelectr.*, 32:304, 1980.
- [13] N.I. Fisher, T. Lewis, and B.J.J. Embleton. *Statistical Analysis of Spherical Data*. Cambridge Uni. Press, N.Y., 1987.
- [14] R. A. Fisher. Dispersion on a sphere. *Proc. of the Royal Soc. of London Ser. A - Math. and Phy. Sci.*, 1130:295–305, 1953.
- [15] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.
- [16] S. Ghosal, J. Lember, and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.
- [17] O.C. Hamsici and A.M. Martinez. Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8:1583–1623, 2007.
- [18] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society Series B*, 58:155–176, 1996.
- [19] C.C. Holmes, F. Caron, J.E. Griffin, and D.A. Stephens. Two-sample bayesian nonparametric hypothesis testing. Technical Report, <http://arxiv.org/abs/0910.5060>.
- [20] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.*, 96:161–73, 2001.
- [21] H. Ishwaran and M. Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963, 2002.
- [22] K.V. Mardia and P.E. Jupp. *Directional Statistics*. John Wiley & Sons, West Sussex, England, 2000.
- [23] P. Müller, A. Erkanli, and M. West. Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, 83:67–79, 1996.
- [24] M.L. Pennell and D.B. Dunson. Nonparametric Bayes testing of changes in a response distribution with an ordinal predictor. *Biometrics*, 64:413–423, 2008.
- [25] L. Schwartz. On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4:10–26, 1965.
- [26] J. Sethuraman. A constructive definition of Dirichlet priors. *Statist. Sinica*, 4:639–50, 1994.
- [27] B. Shahbaba and R. Neal. Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, 10:1829–1850, 2009.
- [28] G. Urich. Computer generation of distributions on the  $m$ -sphere. *Appl. Statist. Soc.*, B16:885–898, 1984.
- [29] R.V. von Mises. Über die “Ganzzahligkeit” der Atomgewicht und verwandte Fragen. *Physik Z*, 19:490–500, 1918.
- [30] G.S. Watson and E.J. Williams. Construction of significance tests on the circle and sphere. *Biometrika*, 43:344–52, 1953.
- [31] A.T.A. Wood. Simulation of the Von Mises Fisher distribution. *Commun. Statist.-Simula.*, 23(1):157–164, 1994.
- [32] Y. Wu and S. Ghosal. Kullback-Leibler property of kernel mixture priors in Bayesian density estimation. *Elec J. Statist.*, 2:298–331, 2008.
- [33] Y. Wu and S. Ghosal.  $L_1$  - consistency of dirichlet mixtures in multivariate bayesian density estimation. *Jour. of Multivar. Analysis*, 101:2411–2419, 2010.
- [34] C. Yau, O. Papaspiliopoulos, G.O. Roberts, and C. Holmes. Nonparametric hidden Markov models with applications in genomics. *J. R. Statist. Soc. B*, 73, 2010.