

Nonparametric Bayesian Density Estimation on Manifolds with Applications to Planar Shapes

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

ABSTRACT. Statistical analysis on landmark-based shape spaces has diverse applications in morphometrics, medical diagnostics, machine vision, robotics and other areas. These shape spaces are non-Euclidean quotient manifolds, often the quotient of the unit sphere under a group of transformations. To conduct nonparametric inferences, one may define notions of center and spread of a probability distribution on an arbitrary manifold and work with their estimates. There has been a significant amount of work done in this direction. However, it is useful to consider full likelihood-based methods, which allow nonparametric estimation of the probability density. This article proposes a class of mixture models constructed using suitable kernels on a general compact non-Euclidean manifold and then on the planar shape space in particular. Following a Bayesian approach with a nonparametric prior on the mixing distribution, conditions are obtained under which the Kullback-Leibler property holds, implying large support and weak posterior consistency. Gibbs sampling methods are developed for posterior computation, and the methods are applied to problems in density estimation on shape space and classification with shape-based predictors.

1. Introduction

In recent years, there has been considerable interest in the statistics literature in the analysis of data having support on a non-Euclidean manifold M . Our focus is on nonparametric approaches, which avoid modeling assumptions about the distribution of the data over M . Although we are particularly motivated by landmark-based analyses of planar shapes, we develop nonparametric Bayes theory and methods also for general manifolds.

There is a rich literature on frequentist methods of inference on manifolds, which avoid a complete likelihood specification in conducting nonparametric estimation and testing based on manifold data. Refer, for example to Bhattacharya and Bhattacharya [1] and the references cited therein. Such methods are based on estimates of center and spread, which are appropriate for manifolds. However, other aspects of the distribution other than center and spread may be important. In addition, Bayesian likelihood-based methods have the advantage of providing a full

Key words and phrases. Non-Euclidean manifold; Planar shape space; Nonparametric Bayes; Dirichlet process mixture; KL property; Posterior consistency; Discriminant analysis.

probabilistic characterization of uncertainty, which is valid even in small samples.

There is a very rich literature on nonparametric Bayes density estimation in Euclidean spaces, with the most commonly used method based on kernel mixture models of the form

$$(1.1) \quad f(y; P) = \int K(y; \theta) P(d\theta),$$

where K is a kernel and P is a mixture distribution. For example, for univariate density estimation with $y \in \mathbb{R}$, the kernel is commonly chosen as

$$K(y; \theta) = (2\pi\sigma^2)^{-1/2} \exp\left\{\frac{-1}{2\sigma^2}(y - \mu)^2\right\},$$

with $\theta = (\mu, \sigma)$, leading to a mixture of Gaussians. In allowing the mixture distribution to be unknown through a prior distribution with large support, one obtains a highly-flexible specification. A common choice of prior for P is the Dirichlet process (DP) (see Ferguson [6], [7]), resulting in a DP mixture (DPM) of Gaussians (Escobar and West [5]). Lo [14] showed that DPM location-scale mixtures of Gaussians have dense support on the space of densities with respect to Lebesgue measure, while Ghosal et al. [8] proved posterior consistency.

Our focus is on developing Bayesian methods for nonparametric density estimation on non-Euclidean manifolds M using a specification similar to (1.1). The manifold of special interest is the planar shape space Σ_2^k - the space of similarity shapes of configurations of k landmarks in 2D.

Frequentist methods for nonparametric density estimation on non-Euclidean manifolds have been developed in Pelletier [20]. In that paper, an appropriate kernel is presented on a compact manifold which generalizes the commonly used location-scale kernel on Euclidean spaces. It is used to build a kernel density estimate (KDE) which uses the sample points as the locations and a fixed known band-width. It is proved that the KDE is L^2 consistent for a sufficiently small band-width.

We use that kernel to build mixture density models on general manifolds. For landmark-based shape analyses, we focus on mixtures of Complex Watson (CW) distributions. The CW distribution was proposed in Watson [25],[26] as a convenient parametric distribution for data on spheres, and later in Dryden and Mardia [4] for planar shape data. Kume and Walker [12] recently proposed an MCMC method for posterior computation in CW parametric models.

To do Bayesian inference, as in Euclidean spaces, the kernel must be carefully chosen, so that the induced prior will have large support, meaning that the prior assigns positive probability to arbitrarily small neighborhoods around any density f_0 . Such a support condition is important in allowing the posterior to concentrate around the true density increasingly as the sample size grows. From the theorem of Schwartz [21], prior positivity of Kullback-Leibler (KL) neighborhoods around the true density f_0 implies that the posterior probability of any weak neighborhood of f_0 converges to one as $n \rightarrow \infty$. Showing that a proposed prior has KL support is important in providing a proof of concept that the prior is sufficiently flexible. Unfortunately, showing KL support tends to be quite difficult for new priors even in Euclidean spaces, though Wu and Ghosal [29] provide useful sufficient conditions.

In this paper, we extend those results to general manifolds and in particular to the planar shape space using the CW kernel.

In addition to large support, nonparametric Bayes procedures must be computationally tractable and lead to interpretable results in order to be useful in practice. The enormous success of DPM models is largely due to the availability of efficient and easy to implement computational algorithms, such as the Polya urn Gibbs sampler (Bush and MacEachern [3]), the block Gibbs sampler (Ishwaran and James [9]) and the exact block Gibbs sampler (Papaspilopoulos [18]). DP priors are characterized by a precision parameter α and a base probability measure P_0 , with computational efficiency improved when P_0 is conjugate to the likelihood. We develop efficient methods for simulating from the posterior distributions of our mixture models using DP priors.

Lennox et al. [13] proposed a DPM of bivariate von Mises distributions for protein conformation angles, modifying the finite mixture model of Mardia, Taylor and Subramaniam [16]. Posterior computation relies on the auxiliary Gibbs sampler of Neal [17], with efficiency improved through conditionally-conjugate updating. Their approach is specific to angular data and they do not present results on support of the prior. It is potentially the case that there are certain angular distributions that cannot be accurately characterized as mixtures of von Mises distributions.

This article is organized as follows. In Section 2, we develop kernel mixture density models on general compact Riemannian manifolds. Through Theorems 2.2 and 2.4, we provide mild sufficient conditions on the true density and the prior on the mixing distribution so that the induced priors satisfy the KL property. These conditions are trivially satisfied by many standard priors such as DP. We present an algorithm based on the exact block Gibbs sampler to simulate from the posterior distribution of the density. These results are then applied to the unit sphere in Section 3.

Section 4 provides a brief overview of the geometry of Σ_2^k . In Section 5, we present some important parametric distributions on this space, discuss their properties and show how to sample from them. These distributions come into much use in the later sections to build mixture density models on Σ_2^k with large support and for posterior computations. In Section 6, we carry out nonparametric density estimation on Σ_2^k using mixtures of CW kernels. We prove that the KL property holds for the induced priors under mild assumptions on the mixing priors and the true density in Theorems 6.2 and 6.3. We adapt the methods from Section 2.3 for posterior computations using a DPM of CW kernels. We present a choice for base measure P_0 and prior band-width distribution using which we get posterior conditional conjugacy and the computational efficiency is highly enhanced.

We present some applications of the methods developed in Sections 7 and 8. In Section 8, we numerically compare the performance of our density estimate with other estimates such as KDE and parametric model based estimate. To do so, we simulate data from a known distribution, estimate the distribution by each method and estimate the divergence of the density estimate from the true density. It turns out that the Bayes estimate performs much better than the other two. Finally in Section 8, we perform classification of real-world data via nonparametric Bayes

discriminant analysis. In this example, there are samples of male and female gorilla skull images. We estimate the shape density for each group and then estimate the conditional probability of a skull being female given its shape, using which we classify it as male or female.

The proofs of our major results are presented at the end in an Appendix section.

2. Nonparametric density estimation on general manifold

Let (M, g) be a compact Riemannian manifold of dimension d , g being the Riemannian metric tensor. Let d_g be the geodesic distance under g . Then (M, d_g) is a complete metric space. Let r_* denote the injectivity radius of M . Since M is compact, $0 < r_* < \infty$. For $p \in M$, let $T_p M$ be the tangent space of M at p which is isomorphic to \mathbb{R}^d . Then the exponential map at p , $\exp_p : T_p M \rightarrow M$ provides the normal coordinates at p . If we denote by $B_p(0, r_*)$ a ball of radius r_* centered at the origin in $T_p M$, then \exp_p is a diffeomorphism from $B_p(0, r_*)$ into M . This ball is contained in a normal neighborhood of p . For an Euclidean space, $r_* = \infty$ and the entire space can be covered by one coordinate patch.

For $p, m \in M$, let $G_p(m)$ be the volume density function on M . If m belongs to a normal neighborhood of p , then $G_p(m)$ is the density of the pull back of the volume measure on M to $T_p M$ with respect to the Lebesgue measure on $T_p M$ via the inverse exponential map \exp_p^{-1} . If x denotes the normal coordinates for m , then

$$G_p(m) = \det(A(x))^{1/2} \text{ where } A(x)_{ij} = g\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)(x), \quad 1 \leq i, j \leq d,$$

and $G_p(p) = 1$. In a normal neighborhood, G is strictly positive and $G_p(m) = G_m(p)$ (see Willmore [28]). This volume density function can be extended as a non-negative continuous function to the whole of M using Jacobi fields (see [28]). Note that on an Euclidean space G is identically equal to 1.

2.1. Mixture density models on M .

Consider the kernel

$$(2.1) \quad K(m; \mu, \sigma) = \mathcal{X}\left(\frac{d_g(m, \mu)}{\sigma}\right) \sigma^{-d} G_\mu^{-1}(m)$$

with variable $m \in M$ and parameters $(\mu, \sigma) \in M \times \mathbb{R}^+$. Here $\mathcal{X} : [0, \infty) \rightarrow [0, \infty)$ is a continuous function satisfying $\int_{\mathbb{R}^d} \mathcal{X}(\|x\|) dx = 1$. We also assume that there exists a constant $A > 0$ such that $\sigma \leq Ar_*$ and \mathcal{X} is zero outside $[0, \frac{1}{A})$. Then $K(\cdot; \mu, \sigma)$ is a well defined function because $G_\mu(\cdot)$ is strictly positive on the geodesic ball

$$B(\mu, r_*) = \{m \in M : d_g(\mu, m) < r_*\}$$

and due the support restriction on \mathcal{X} , K is defined to be zero outside this ball. Proposition 2.1 proves that (2.1) defines a valid probability density on M . From now on, for convenience, we will write K for $K(\cdot; \mu, \sigma)$ wherever the parameters (μ, σ) remain fixed.

PROPOSITION 2.1. *For any fixed $\mu \in M$ and $\sigma \in (0, Ar_*]$, K defines a probability density on M with respect to the volume measure.*

PROOF. We need to show that

$$(2.2) \quad \int_M K(m; \mu, \sigma) V(dm) = 1 \quad \forall \mu \in M, \quad 0 < \sigma \leq Ar_*,$$

$V(dm)$ being the volume form of M . Since K is zero outside $B(\mu, r_*)$, the integral in (2.2) can be written as $\int_{B(\mu, r_*)} K(m; \mu, \sigma) V(dm)$. Since $B(\mu, r_*)$ lies in a normal neighborhood of μ , using normal coordinates into $T_\mu M$, the integral can be written as $\int_{\|x\| < r_*} K(m; \mu, \sigma) G_\mu(m) dx$. Here x denotes the normal coordinates for m (i.e. $x = \exp_\mu^{-1}(m)$) and then $V(dm)$ becomes $G_\mu(m) dx$. Using the fact that $d_g(\mu, m) = \|x\|$, (2.2) becomes

$$(2.3) \quad \int_{\|x\| < r_*} \mathcal{X}\left(\frac{\|x\|}{\sigma}\right) \sigma^{-d} dx = \int_{\|y\| < \frac{r_*}{\sigma}} \mathcal{X}(\|y\|) dy$$

$$(2.4) \quad = \int_{\|y\| < \frac{1}{A}} \mathcal{X}(\|y\|) dy = 1$$

Equation (2.4) follows from (2.3) because $\frac{r_*}{\sigma} \geq \frac{1}{A}$ and \mathcal{X} is zero on $[\frac{1}{A}, \infty)$. This completes the proof. \square

Using the kernel K , we can define a location mixture probability density on M as

$$(2.5) \quad f(m; P, \sigma) = \int_M K(m; \mu, \sigma) P(d\mu)$$

or a location-scale mixture density

$$(2.6) \quad g(m; Q) = \int_{M \times \mathbb{R}^+} K(m; \mu, \sigma) Q(d\mu d\sigma).$$

While defining f , the parameter P is a probability distribution on M and σ is the band-width parameter which lies in $(0, Ar_*]$. For defining g , the parameter Q is a probability distribution on $M \times (0, Ar_*]$. From Proposition 2.1, it is easy to show that for parameters fixed, f and g are valid densities on M . They generalize the commonly used location and location-scale mixture models on Euclidean spaces to more complex manifolds.

2.2. KL condition. In this section and subsequent ones, we represent the space of probabilities on a space X as $\mathcal{M}(X)$. Suppose we have an iid sample X_1, \dots, X_n from some unknown distribution F_0 on M . We assume that F_0 is absolutely continuous with respect to the volume measure on M and let f_0 be its volume-density. To find a nonparametric Bayes estimate for f_0 , we approximate it by a mixture density as in (2.5) or (2.6). Then we set a prior for the parameters which induces corresponding priors on the space of densities on M via the mixture models. Using the sample and this prior, we compute the posterior distribution of the underlying probability distribution generating the sample. For model (2.5), let Π_1 be a prior for P and choose an independent prior π_1 for σ , i.e. $(P, \sigma) \sim \Pi_1 \otimes \pi_1$. For model (2.6), denote by Π_2 a prior for Q . For these models to be good approximations for the true density f_0 , we need to show that the induced priors give positive probability to arbitrarily small neighborhoods around any density f_0 on M . From the Schwartz theorem, it follows that prior positivity of the Kullback-Leibler (KL) neighborhoods around f_0 implies that the posterior probability of any weak neighborhood of f_0 converges to one almost surely as $n \rightarrow \infty$. Such a prior is said to satisfy the KL condition. We give a formal definition of KL neighborhood and KL condition in Definition 2.1.

DEFINITION 2.1. The KL neighborhood of a density f_0 of size $\epsilon > 0$ is defined as

$$KL(f_0, \epsilon) = \left\{ f : \int_M f_0(m) \log \left(\frac{f_0(m)}{f(m)} \right) V(dm) < \epsilon \right\}.$$

A prior Ψ on the space of probability densities on M (w.r.t. the volume measure) is said to satisfy the KL condition at f_0 if for any $\epsilon > 0$,

$$\Psi\{KL(f_0, \epsilon)\} > 0.$$

Corollary 2.3 provides conditions on f_0 and the prior $\Pi_1 \otimes \pi_1$ for the parameters (P, σ) corresponding to the location mixture density f in (2.5) under which the induced prior satisfies the KL condition at f_0 . Theorem 2.2 provides similar conditions on the prior Π_2 for the mixing measure Q in the location-scale mixture density g in (2.6). In fact as Theorem 2.2 shows, for the location mixture density model we can even prove that arbitrarily small L^∞ neighborhoods around f_0 get positive probability under the prior induced by $\Pi_1 \otimes \pi_1$. This implies the KL condition at f_0 as shown in corollary 2.3 and also positive prior probability for L^1 neighborhoods around f_0 . The proofs of Theorems 2.2 and 2.4 are given in the Appendix section 9.1.

THEOREM 2.2. *Let f_0 be a continuous density on M and F_0 be the corresponding probability distribution. Let $f(m; P, \sigma)$ be a density as in (2.5). Assume that the prior $\Pi_1 \otimes \pi_1$ for (P, σ) contains $(F_0, 0)$ in its support. Also assume that there exists a positive constant $r_1 < r_*$ such that $\pi_1\{(0, Ar_1]\} = 1$. Then for any $\epsilon > 0$*

$$(2.7) \quad (\Pi_1 \otimes \pi_1)\{(P, \sigma) : \sup_{m \in M} |f_0(m) - f(m; P, \sigma)| < \epsilon\} > 0.$$

COROLLARY 2.3. *Let f_0 be a strictly positive continuous density on M . Under the assumptions of Theorem 2.2, the prior induced by $\Pi_1 \otimes \pi_1$ satisfies the KL condition at f_0 .*

PROOF. From now on for simplicity we shall use $f(m)$ for $f(m; P, \sigma)$ whenever it is understood. Since M is compact, $f_0(m) > 0$ for all $m \in M$ implies that $\inf_{m \in M} f_0(m) = c_0 > 0$. For $\delta > 0$ define

$$\mathcal{W}_\delta = \{(P, \sigma) : \sup_{m \in M} |f_0(m) - f(m; P, \sigma)| < \delta\}.$$

Then if $(P, \sigma) \in \mathcal{W}_\delta$,

$$\inf_{m \in M} f(m) \geq \inf_{m \in M} f_0(m) - \delta \geq \frac{c_0}{2}$$

if we choose $\delta \leq \frac{c_0}{2}$. Then for any given $\epsilon > 0$,

$$\int_M f_0(m) \log \left(\frac{f_0(m)}{f(m)} \right) V(dm) \leq \sup_{m \in M} \left| \frac{f_0(m)}{f(m)} - 1 \right| \leq \frac{2\delta}{c_0} < \epsilon$$

if we choose $\delta < \frac{c_0 \epsilon}{2}$. Hence for δ sufficiently small, $f(\cdot; P, \sigma) \in KL(f_0, \epsilon)$ whenever $(P, \sigma) \in \mathcal{W}_\delta$. From Theorem 2.2 it follows that $(\Pi_1 \otimes \pi_1)(\mathcal{W}_\delta) > 0$ for any $\delta > 0$ and therefore

$$(\Pi_1 \otimes \pi_1)\{(P, \sigma) : f(\cdot; P, \sigma) \in KL(f_0, \epsilon)\} > 0.$$

□

THEOREM 2.4. *Let f_0 be a strictly positive continuous density on M and let F_0 denote its corresponding distribution. Let $g(m; Q)$ be a density as in (2.6). Let Π_2 be a prior on Q such that $\Pi_2\{\mathcal{M}(M \times (0, Ar_1])\} = 1$ and $F_0 \otimes \delta_0$ is in the support of Π_2 . Then the prior on the space of densities on M induced by Π_2 satisfies the KL condition at f_0 .*

REMARK 2.1. From Proposition 2.1 it follows that f and g are valid probability densities if $0 < \sigma \leq Ar_*$. However, to show the KL condition for the mixture priors in Corollary 2.3 and Theorem 2.4, we have added the stronger restriction that $0 < \sigma \leq Ar_1$. This restriction on the prior to smaller bandwidth σ is intuitively reasonable because smaller bandwidth is expected to give a finer approximation to the unknown density.

The conditions on the priors in Theorems 2.2 and 2.4 correspond to the size of the support of the priors, which are trivially satisfied by many standard nonparametric priors. For example, for model (2.5) we can choose Π_1 to be a Dirichlet process prior $DP(\omega_0 P_0)$ with $\text{supp}(P_0) = M$ and π_1 to have a density on $(0, Ar_1]$ that is strictly positive in some neighborhood of zero. For (2.6), we can instead choose the prior Π_2 for the mixing measure Q to correspond to a Dirichlet process with base $P_0 \otimes \pi_1$. These choices are convenient computationally, as we will illustrate in Section 2.3.

2.3. Posterior computation. For simplicity in describing an approach for posterior computation, we focus on the location mixture specified in (2.5) with a Dirichlet process prior for the mixing measure P . In Dirichlet process mixture models, there are two common strategies for posterior computation, with the first relying on a marginal approach that integrates out the mixing measure (MacEachern [15], West et al. [27]) and the second relying on a conditional approach (Ishwaran and James [9]). Conditional algorithms typically rely on the stick-breaking representation of Sethuraman [22], which lets $P = \sum_{j=1}^{\infty} w_j \delta_{\mu_j^s}$, with $\mu_j^s \sim P_0$, $w_j = V_j \prod_{h < j} (1 - V_h)$, and $V_j \sim \text{Be}(1, \omega_0)$, where the atoms $\{\mu_j^s\}$ and stick-breaking random variables $\{V_j\}$ are mutually independent *a priori*. A difficulty that arises is that the random measure P is expressed in terms of infinitely many unknowns, so that exact posterior computation seems impossible. Ishwaran and James [9] address this problem through a truncation approximation to P . Recently, retrospective sampling (Papaspiliopoulos and Roberts [19]) and slice sampling (Walker [24]) algorithms have been developed that avoid the need for truncation. This is possible because only a finite number of components are occupied by subjects in the sample, with the prior and posterior distributions being identical for the infinitely-many components having higher indices.

Here, we follow the exact block Gibbs sampler proposed by Papaspiliopoulos [18] and Yau et al. [30]. Let $X_i \sim K(\cdot; \mu_i, \sigma)$, for $i = 1, \dots, n$, with $\mu_i \sim P$, $P \sim DP(\omega_0 P_0)$ and $\sigma \sim \pi_1$. We introduce uniformly distributed slice sampling latent variables, $u = \{u_i\}_{i=1}^n$ and let S_i denote the mixture component for subject i , with $\mu_i = \mu_{S_i}^s$. The complete data likelihood is then

$$\prod_{i=1}^n K(X_i; \mu_{S_i}^s, \sigma) 1(u_i < w_{S_i}),$$

and we sequentially sample through the following steps.

- (1) Update S_i , for $i = 1, \dots, n$, by sampling from the multinomial conditional posterior distribution with $\Pr(S_i = j) \propto K(X_i; \mu_j^s, \sigma)$ for $j \in A_i$, where $A_i = \{j : 1 \leq j \leq l, w_j > u_i\}$ and l is the smallest index satisfying $1 - u_{(1)} < \sum_{j=1}^l w_j$. In implementing this step, draw $V_j \sim \text{Be}(1, \omega_0)$ and $\mu_j^s \sim P_0$ for $j > S_{(n)}$.
- (2) Update the atoms μ_j^s , $j = 1, \dots, S_{(n)}$, by sampling μ_j^s from the conditional posterior, which is proportional to $P_0(d\mu_j^s) \prod_{i: S_i=j} K(X_i; \mu_j^s, \sigma)$. This is equivalent to sampling from the prior for components that are unoccupied.
- (3) Update the bandwidth parameter σ by sampling from the conditional posterior, which is proportional to $\pi_1(d\sigma) \prod_{i=1}^n K(X_i; \mu_{S_i}^s, \sigma)$.
- (4) Update the stick-breaking random variables V_j , for $j = 1, \dots, S_{(n)}$, from their conditional posterior distributions given the cluster allocation but marginalizing out the slice sampling latent variables $\{u_i\}_{i=1}^n$. In particular,

$$V_j \sim \text{Be}\left(1 + \sum_i 1(S_i = j), \omega_0 + \sum_i 1(S_i > j)\right).$$

- (5) Update the slice sampling latent variables from their conditional posterior by letting $u_i \sim \text{Unif}(0, w_{S_i})$, for $i = 1, \dots, n$.

These steps are repeated a large number of iterations, with a burn-in discarded to allow convergence. In our experience, the algorithm is quite efficient, with rapid convergence and no evidence of slow mixing in cases we have considered. Due to label switching issues (Stephens [23]), we recommend assessing convergence and mixing by examining trace plots and applying standard diagnostics for the density $f(m; P, \sigma)$ evaluated at a dense grid of m values. A draw from the posterior for f or the predictive density can be calculated using

$$(2.8) \quad f(m; P, \sigma) = \sum_{j=1}^{S_{(n)}} w_j K(m; \mu_j^s, \sigma) + \left(1 - \sum_{j=1}^{S_{(n)}} w_j\right) \int K(m; \mu^s, \sigma) dP_0(\mu^s),$$

with σ and w_j , μ_j^s , $j = 1, \dots, S_{(n)}$ an MCMC draw from the joint posterior of the bandwidth and the weights and atoms for each of the components up to the maximum occupied. A Bayes estimate of f can then be obtained by averaging these draws across many samples. When P_0 is chosen to correspond to the uniform distribution over the manifold, the integral $\int K(m; \mu^s, \sigma) dP_0(\mu^s) = 1/\text{Vol}(M)$. In this case, computing the predictive density in (2.8) becomes relatively simple. However, in many cases, the uniform distribution may be overly diffuse, having a low probability of generating clusters close to the data. This can lead to a very large penalty on adding new clusters as data are added, and hence underestimation of the number of clusters. We recommend instead choosing a non-conjugate P_0 , which assigns high probability to cluster means located close to the data values, with such a P_0 chosen based on prior knowledge, past data or empirical Bayes. We can accommodate non-conjugate cases by using Metropolis-Hastings sampling in steps 2 and 3 and analytically approximating the integral in (2.8). One way to do so is to replace the integral by $K(m; \mu^*, \sigma)$, μ^* being a draw from P_0 . Alternatively, it tends to be the case unless the data set is small that $1 - \sum_{j \leq S_{(n)}} w_j \approx 0$, so that we can accurately approximate the predictive density discarding the final term in (2.8).

In the next section, we explicitly compute the kernel K on the unit sphere.

3. Application to the unit sphere S^d

Consider the unit sphere in \mathbb{R}^{d+1} , namely,

$$S^d = \{m \in \mathbb{R}^{d+1} : \|m\| = 1\}.$$

Statistical analysis on the sphere finds lots of application in directional data analysis. Also since most of the shape spaces are quotients of the sphere, it is important to understand its geometry and how to do inference on it.

The sphere S^d is a compact Riemannian manifold of dimension d and injectivity radius of π . For two points $m_1, m_2 \in S^d$, the geodesic distance between them is given by

$$d_g(m_1, m_2) = \arccos(m_1' m_2)$$

which lies between 0 and π . The tangent space at $m \in S^d$ is

$$T_m S^d = \{v \in \mathbb{R}^{d+1} : v' m = 0\}.$$

It is endowed with the metric tensor from \mathbb{R}^{d+1} , i.e. $g(v_1, v_2) \equiv \langle v_1, v_2 \rangle = v_1' v_2$. The exponential map takes the form

$$\exp_m : T_m S^d \rightarrow S^d, \quad \exp_m(v) = \cos(\|v\|)m + \frac{\sin(\|v\|)}{\|v\|}v.$$

It is a diffeomorphism from $B(0, \pi)$ onto $S^d \setminus \{-m\}$. Proposition 3.1 computes the volume-density function on the sphere.

PROPOSITION 3.1. For $p, m \in S^d$, $d > 1$,

$$G_m(p) = \left(\frac{\sin(d_g(m, p))}{d_g(m, p)} \right)^{d-1}.$$

On S^1 , $G_m(\cdot) \equiv 1$.

PROOF. Let $p \in S^d \setminus \{-m\}$. For a choice of orthonormal basis $\{v_1, \dots, v_d\}$ for $T_m S^d$, define

$$\begin{aligned} \phi : B(0, \pi) &\equiv \{x \in \mathbb{R}^d : \|x\| < \pi\} \rightarrow S^d \setminus \{-m\}, \\ \phi(x) &= \exp_m(x^i v_i) = \cos(\|x\|)m + \frac{\sin(\|x\|)}{\|x\|}x^i v_i. \end{aligned}$$

Then $x = \phi^{-1}(p)$ gives the normal coordinates for p . Let $D\phi(x)$ denote the derivative of ϕ at x , $D\phi(x) : \mathbb{R}^d \rightarrow T_p S^d$. Then $G_m(p) = \{\det(g(x))\}^{1/2}$ where $g(x) = ((D\phi(x)(e_i), D\phi(x)(e_j)))_{1 \leq i, j \leq d}$ and $\{e_1, \dots, e_d\}$ denotes the canonical basis for \mathbb{R}^d . Denote by V the matrix $[v_1, \dots, v_d]$. Then it is easy to show that $D\phi(x)$ is the $(d+1) \times d$ matrix,

$$D\phi(x) = -\frac{\sin(\|x\|)}{\|x\|}m x' + \frac{\sin(\|x\|)}{\|x\|}V + \left(\frac{\cos(\|x\|)}{\|x\|^2} - \frac{\sin(\|x\|)}{\|x\|^3} \right) V x x'$$

so that

$$((g(x)))_{i,j} = \frac{\sin^2 \|x\|}{\|x\|^2} \delta_{ij} + \left(1 - \frac{\sin^2 \|x\|}{\|x\|^2} \right) \frac{x^i x^j}{\|x\|^2}$$

and hence

$$g(x) = \frac{\sin^2 \|x\|}{\|x\|^2} I_d + \left(1 - \frac{\sin^2 \|x\|}{\|x\|^2} \right) \frac{x x'}{\|x\|^2}$$

which has eigen-values 1 with multiplicity 1 and $\frac{\sin^2 \|x\|}{\|x\|^2}$ with multiplicity $d - 1$. Since $\det(g(x))$ is the product of its eigen-values, we get

$$G_m(p) = \begin{cases} \left(\frac{\sin(\|x\|)}{\|x\|}\right)^{d-1} & \text{if } d > 1 \\ 1 & \text{if } d = 1. \end{cases}$$

Since $\|x\| = d_g(m, p)$, we get the desired expression for $G_m(p)$. \square

To get a kernel on the sphere as in (2.1), we choose the function \mathcal{X} such that $\mathcal{X}(\|x\|)$ defines a density on \mathbb{R}^d with compact support. Now we can build mixture density models on S^d as in Section 2.1. To sample from the posterior distribution of the density as in Section 2.3, we need to write the conditional posteriors using a suitable coordinate system on S^d . A natural choice is using normal coordinates into the tangent space of some fixed point such as the estimated center of the distribution. For different notions of centers on the sphere and their properties, see [2] and [1].

In the next section, we describe our main manifold of interest, namely the planar shape space of k -ads, and carry out density estimation on it in the subsequent sections.

4. The planar shape space Σ_2^k

Consider a set of k points, $k > 2$, on the 2D plane, not all points being the same. We refer to such a set as a k -ad or a set of k landmarks. The similarity shape of this k -ad is what remains after we remove the effects of the Euclidean rigid body motions of translation and rotation and scaling. For convenience we denote a k -ad by a complex k -vector $z = (z_1, z_2, \dots, z_k)'$, i.e., we will represent k -ads on a complex plane. To remove the effect of translation from z , one subtracts

$$\bar{z} = \frac{1}{k} \sum_{j=1}^k z_j$$

from z to bring its centroid to the origin. This centered k -ad z_c lies on the complex $(k-1)$ -dimensional subspace H^{k-1} of \mathbb{C}^k consisting of all vectors orthogonal to the vector $\mathbf{1}_k$ of all ones. Using an orthonormal basis for H^{k-1} , we compute coordinates $z_H \in \mathbb{C}^{k-1}$ for z_c , that is

$$z_c = \sum_{j=1}^{k-1} z_H^j H_j = H z_H$$

with $H = [H_1, \dots, H_{k-1}]$, columns of which form an orthonormal basis for H^{k-1} . The effect of scaling is removed by dividing z_H by its total norm $\|z_H\| = \sqrt{\sum_{j=1}^{k-1} |z_H^j|^2}$ (which is $\|z_c\|$). The normalized k -ad w lies on the complex unit sphere

$$\mathbb{C}S^{k-2} = \{w \in \mathbb{C}^{k-1} : \|w\| = 1\}$$

which can be identified with the real sphere S^{2k-3} . Since w contains the shape information of z along with rotation, it is called the preshape of z . The space of all preshapes forms the preshape sphere S_2^k which is $\mathbb{C}S^{k-2}$ or S^{2k-3} . The similarity shape of z is then the orbit of w under all rotations in 2D. Since a rotation by an

angle θ of a landmark (x, y) can be achieved by multiplying its complex version $x + iy$ by $e^{i\theta}$, the shape of z (or w) is the set (or orbit)

$$[w] = \{e^{i\theta}w : \theta \in [0, 2\pi)\}.$$

The space of all such orbits constitutes the planar shape space Σ_2^k which is the quotient of the preshape sphere under all one dimensional rotations, that is

$$\Sigma_2^k = S_2^k / [0, 2\pi) = \{[w] : w \in S_2^k\}.$$

Since any shape or orbit is the set of all intersection points of a unique line passing through the origin in \mathbb{C}^{k-1} with $\mathbb{C}S^{k-2}$, the planar shape space can be identified with the complex projective space $\mathbb{C}P^{k-2}$ which is the space of all complex lines passing through the origin in \mathbb{C}^{k-1} . With this identification, Σ_2^k is a compact Riemannian manifold of dimension $2k - 4$. It has an injective radius r_* of $\frac{\pi}{2}$. The geodesic distance between two shapes $[u], [v]$ ($u, v \in S_2^k$) is given by

$$d_g([u], [v]) = \arccos(|u^*v|)$$

where $*$ denotes the complex conjugate transpose. For $m = [u] \in \Sigma_2^k$, the tangent space $T_m \Sigma_2^k$ can be identified with the complex $(k - 2)$ -dimensional subspace

$$V_u = \{v \in \mathbb{C}^{k-1} : u^*v = 0\}.$$

For a choice of orthonormal basis $\{v_1, \dots, v_{k-2}, iv_1, \dots, iv_{k-2}\}$ for V_u (over \mathbb{R}), the normal coordinates for a shape $m_1 = [u_1]$ into $T_m \Sigma_2^k$ is given by

$$\begin{aligned} z &= (x_1, \dots, x_{k-2}, y_1, \dots, y_{k-2})', \\ x_j + iy_j &= \frac{r}{\sin(r)} e^{i\theta} v_j^* u_1, \quad j = 1, \dots, k-2, \\ r &= d_g(m, m_1) = \|z\|, \quad e^{i\theta} = \frac{u_1^* u}{|u_1^* u|}. \end{aligned}$$

Σ_2^k can be embedded into the space $S(k-1, \mathbb{C})$ of all $(k-1) \times (k-1)$ complex Hermitian matrices via the Veronese-Whitney embedding which is given by

$$J: \Sigma_2^k \rightarrow S(k-1, \mathbb{C}), \quad J([u]) = uu^*.$$

Here $S(k-1, \mathbb{C})$ is viewed as a linear subspace of $\mathbb{C}^{(k-1)^2}$ of real dimension $(k-1)^2$. The extrinsic distance between two shapes $[u], [v]$ is the one induced from this embedding, namely,

$$d_E([u], [v]) = \|J([u]) - J([v])\| = \sqrt{2(1 - |u^*v|^2)}.$$

4.1. Center and spread. Let Q be a probability distribution on Σ_2^k . The center of Q can be measured by its extrinsic or intrinsic means while its extrinsic or intrinsic variations define notions of spread of Q .

The extrinsic mean of Q is defined as the minimizer of the loss function

$$(4.1) \quad F(p) = \int_{\Sigma_2^k} d_E^2(m, p) Q(dm), \quad p \in \Sigma_2^k$$

provided F has a unique minimizer. The minimum value of F is called the extrinsic variation of Q . Given a sample X_1, \dots, X_n from Q , the extrinsic mean and variation of the empirical distribution $\frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ are called the sample analogues. Let $Q^J = Q \circ J^{-1}$ denote the push forward of Q in to $S(k-1, \mathbb{C})$ using the Veronese-Whitney

embedding J . Then Q^J has a compact support in $S(k-1, \mathbb{C})$ and hence a well defined Euclidean mean

$$\mu^J = \int_{S(k-1, \mathbb{C})} x Q^J(dx).$$

Since μ^J is the average of positive semi definite (p.s.d.) trace 1 matrices, it is also p.s.d. with trace equal to 1. Proposition 4.1 identifies the extrinsic parameters of Q as functions of μ^J . For a proof, see Bhattacharya and Bhattacharya [1] and Bhattacharya and Patrangenaru [2].

PROPOSITION 4.1. *Let λ_{k-1} denote the largest eigen-value of μ^J and let U_{k-1} be a corresponding unit norm eigen vector. (a) Q has a unique extrinsic mean iff λ_{k-1} is a eigen-value with multiplicity 1 and then the mean is given by $[U_{k-1}]$. (b) The extrinsic variation of Q equals $2(1 - \lambda_{k-1})$.*

In defining the loss function F in (4.1), if we replace d_E by the geodesic distance d_g , its minimizer defines the intrinsic mean of Q , provided it has a unique minimizer. For more details on the properties of the extrinsic and intrinsic parameters and their estimates, see [1] and [2] and the references cited therein.

5. Parametric models on the planar shape space

In this section we present some well known probability distributions on Σ_2^k and study their properties. These models will come into much use in the later sections for nonparametric density estimation.

5.1. Uniform distribution. Let $V(dm)$ and $V_1(dz)$ denote the volume-forms on the shape space Σ_2^k and the preshape sphere S_2^k respectively. The uniform measure on Σ_2^k is then given by the constant density V^{-1} where $V = \int_{\Sigma_2^k} V(dm)$ denotes the volume of Σ_2^k . Kent [10] proposed a useful coordinate chart. For $z = (z_1, \dots, z_{k-1})'$ on S_2^k , write $z_j = \sqrt{r_j} e^{i\theta_j}$, $j = 1, 2, \dots, k-1$ with $r = (r_1, \dots, r_{k-2})'$ on the unit simplex

$$S_{k-2} = \{r \in [0, 1]^{k-2} : \sum_{j=1}^{k-2} r_j \leq 1\},$$

$r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$ and $\theta_j \in (-\pi, \pi)$, $j = 1, 2, \dots, k-1$. Then $(r_1, \dots, r_{k-2}, \theta_1, \dots, \theta_{k-1})$ form the coordinates of z , we will call that Kent's preshape coordinates. Since the shape of z can be obtained by rotating it around a fixed axis, we may set $\theta_{k-1} = 0$ and use the coordinates

$$(r_1, \dots, r_{k-2}, \theta_1, \dots, \theta_{k-2})'$$

for $[z]$. These coordinates are derived in Dryden and Mardia [4], we will call them shape coordinates. The advantage of using these coordinate systems on S_2^k and Σ_2^k is that we get simple expressions for the volume forms. It can be shown that (see Kent [10])

$$\begin{aligned} V_1(dz) &= 2^{2-k} dr_1 \dots dr_{k-2} d\theta_1 \dots d\theta_{k-1}, \\ V(d[z]) &= 2^{2-k} dr_1 \dots dr_{k-2} d\theta_1 \dots d\theta_{k-2}. \end{aligned}$$

In other words, in terms of these shape coordinates, the uniform distribution on Σ_2^k remains uniform on $S_{k-2} \times (-\pi, \pi)^{k-2}$. This helps us simulate from this distribution

and also from the other models stated below. We shall also need this expression for the volume form in proving the KL condition in Section 6.

5.2. Complex Bingham distribution. The Complex Bingham distribution on Σ_2^k has the following density with respect to the volume form:

$$f(m; A) = c^{-1}(A) \exp(z^* A z).$$

Here $z \in S_2^k$ is some preshape of $m \in \Sigma_2^k$ and the parameter $A \in S(k-1, \mathbb{C})$, $c(A)$ being the normalizing constant. It was proposed in Kent [10]. We will denote this distribution by $CB(A)$ or just CB . Note that $CB(A) = CB(A + \alpha I)$ for any $\alpha \in \mathbb{R}$, so that w.l.o.g. we may assume A to be p.s.d. with smallest eigen-value equal to 0.

5.3. Complex Watson distribution. A special case when A has complex rank equal to 1 is the Complex Watson distribution which has the density

$$f(m; \mu, \sigma) = c^{-1}(\sigma) \exp(|z^* \nu|^2 / \sigma)$$

with parameters $\mu \in \Sigma_2^k$ and $\sigma > 0$, $c(\sigma)$ being the normalizing constant. z and ν are some preshapes of m and μ respectively. We shall represent this distribution as $CW(\mu, \sigma)$ or just CW . Note that

$$CW(\mu, \sigma) = CB(J(\mu)/\sigma),$$

J being the Veronese-Whitney embedding mentioned in Section 4.

5.4. Properties of CB and CW distributions. In case of $CB(A)$, write $A = U\Lambda U^*$ with

$$U = [U_1, \dots, U_{k-1}] \in SU(k-1), \\ \Lambda = \text{diag}(\lambda_1, \dots, \lambda_{k-1}), \quad 0 = \lambda_1 \leq \dots \leq \lambda_{k-1},$$

where $SU(k-1)$ is the space of all $(k-1) \times (k-1)$ special unitary matrices ($UU^* = I$, $\det(U) = 1$). This representation is called a singular value decomposition (s.v.d.) for A . Make a change of variable $[z] \rightarrow [z_1]$, $z_1 = U^* z$. This transformation does not change the volume form on the shape space. Then use Kent's shape coordinates (r, θ) for $[z_1]$: $r = (r_1, \dots, r_{k-2}) \in S_{k-2}$, $\theta = (\theta_1, \dots, \theta_{k-2}) \in (-\pi, \pi)^{k-2}$. Then the CB distribution can be written as

$$(5.1) \quad f([z]; A)V(d[z]) = c^{-1}(A)2^{2-k} \exp\left(\sum_{j=1}^{k-1} \lambda_j r_j\right) dr_1 \dots dr_{k-2} d\theta_1 \dots d\theta_{k-2}$$

with $r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$. Hence under the CB distribution, r has the density proportional to $\exp(\sum_{j=1}^{k-1} \lambda_j r_j)$ on S_{k-2} , $\theta_1, \dots, \theta_{k-2}$ are iid $\text{Unif}(-\pi, \pi)$ and r and θ are independent.

For the $CW(\mu, \sigma)$ distribution, since $\lambda_1 = \dots = \lambda_{k-2} = 0$, $\lambda_{k-1} = \sigma^{-1}$, therefore the distribution of r can be written as $f(r) \propto \exp(\sigma^{-1} r_{k-1})$ which implies that r_{k-1} has the marginal distribution

$$g(r_{k-1}) = c_{k-1}^{-1}(\sigma) e^{r_{k-1}/\sigma} (1 - r_{k-1})^{k-3}, \quad r_{k-1} \in (0, 1) \text{ where} \\ c_{k-1}(\sigma) = \sigma^{k-2} e^{1/\sigma} \Gamma(k-2; \sigma^{-1}) \text{ and} \\ \Gamma(m; a) = \int_0^a e^{-t} t^{m-1} dt = (m-1)! e^{-a} \left[e^a - \sum_{r=0}^{m-1} \frac{a^r}{r!} \right]$$

denotes the partial gamma function. Conditioned on r_{k-1} , $r = (r_1, \dots, r_{k-2})$ has a uniform distribution on the set

$$\{r_j \geq 0, j = 1, \dots, k-2, \sum_1^{k-2} r_j = 1 - r_{k-1}\}.$$

Normalizing constants. Expression (5.1) suggests that for the $CB(A)$ distribution, $c(A)$ depends on A only through its eigen values and hence is equal to $c(\Lambda)$. For the CW distribution, $c(\sigma)$ can be derived to be

$$\begin{aligned} c(\sigma) &= 2^{2-k} (2\pi)^{k-2} \int_{S_{k-2}} e^{r_{k-1}/\sigma} dr_2 \dots dr_{k-1} \\ &= 2^{2-k} (2\pi)^{k-2} (k-3)!^{-1} \int_0^1 e^{r_{k-1}/\sigma} (1-r_{k-1})^{k-3} dr_{k-1} \\ &= (\pi\sigma)^{(k-2)} e^{1/\sigma} (k-3)!^{-1} \Gamma(k-2; \sigma^{-1}) \\ &= (\pi\sigma)^{(k-2)} \left[e^{1/\sigma} - \sum_{r=0}^{k-3} \frac{\sigma^{-r}}{r!} \right]. \end{aligned}$$

In [4], the CB & CW distributions are viewed as distributions on the preshape sphere and hence the normalizing constant is derived to be $2\pi c(\sigma)$.

Extrinsic mean and variation. Let $X_1 \sim CB(A)$. Then the extrinsic mean for the CB distribution can be expressed as the shape of a unit eigen-vector corresponding to the largest eigen-value of $\mu^J = E[J(X_1)]$. Let z be one of the preshapes of X_1 , $z_1 = U^*z$ and (r, θ) be the shape coordinates for $[z_1]$. Then

$$\mu^J = E[zz^*] = UE[z_1z_1^*]U^*.$$

Take $\theta_{k-1} = 0$. Then since

$$(z_1z_1^*)_{ij} = \sqrt{r_i r_j} e^{i(\theta_i - \theta_j)}, \quad 1 \leq i, j \leq k-1,$$

therefore

$$E(z_1z_1^*)_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ E(r_i) & \text{if } i = j. \end{cases}$$

Hence

$$\mu^J = U \text{diag}(E(r_1), \dots, E(r_{k-1})) U^*$$

and the extrinsic mean $\mu_E = [U_{j_0}]$ where $E(r_{j_0}) = \max_{1 \leq j \leq k-1} E(r_j)$ provided there is a unique such j_0 . The extrinsic variation is $2(1 - E(r_{j_0}))$.

For the CW(μ, σ) distribution,

$$\begin{aligned} E(r_1) &= \dots = E(r_{k-2}) = \frac{1-E(r_{k-1})}{k-2} \text{ and} \\ E(r_{k-1}) &= c_{k-1}^{-1}(\sigma) \int_0^1 e^{x/\sigma} x(1-x)^{k-3} dx \\ &= 1 - \sigma \frac{\Gamma(k-1; \sigma^{-1})}{\Gamma(k-2; \sigma^{-1})} = 1 - (k-2)\sigma \frac{1 - e^{-1/\sigma} \sum_0^{k-2} r!^{-1} \sigma^{-r}}{1 - e^{-1/\sigma} \sum_0^{k-3} r!^{-1} \sigma^{-r}}. \end{aligned}$$

It can be shown that $E(r_{k-1}) > \frac{1}{k-1}$ and hence $E(r_{k-1}) > E(r_j)$, $j = 1, \dots, k-2$. Therefore for this distribution, the extrinsic mean is

$$\mu_E = [U_{k-1}] = \mu$$

and the extrinsic variation equals

$$V_E = 2(1 - E(r_{k-1})) = 2(k-2)\sigma \frac{1 - e^{-1/\sigma} \sum_0^{k-2} r!^{-1} \sigma^{-r}}{1 - e^{-1/\sigma} \sum_0^{k-3} r!^{-1} \sigma^{-r}}.$$

For σ small, $E(r_{k-1}) \approx 1 - \sigma(k-2)$ and then $V_E \approx 2(k-2)\sigma$.

5.5. Simulation from CW distribution. To draw a sample from $CW(\mu, \sigma)$, we may draw $\theta_1, \dots, \theta_{k-2}$ iid from $\text{Unif}(-\pi, \pi)$, set $\theta_{k-1} = 0$, and draw $r = (r_1, \dots, r_{k-2})$, $r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$ from the distribution $f(r) \propto \exp(\sigma^{-1} r_{k-1})$ on S_{k-2} . Let $z_1 = (z_1^1, \dots, z_1^{k-1})$, $z_1^j = \sqrt{r_j} e^{i\theta_j}$, get $U \in SU(k-1)$ such that $J(\mu) = U\Lambda U^*$, $\Lambda = \text{diag}(0, \dots, 0, 1)$ and set $z = Uz_1$. Then $[z]$ is a random sample from the Complex Watson distribution.

We saw in Section 5.4 that under the distribution f , r_{k-1} has the marginal distribution $g(r_{k-1}) \propto e^{r_{k-1}/\sigma} (1 - r_{k-1})^{k-3}$ on $(0, 1)$. Make the transformation $s_{k-1} = \sigma^{-1}(1 - r_{k-1})$. Then s_{k-1} follows the distribution $h(s_{k-1}) \propto e^{-s_{k-1}} s_{k-1}^{k-3}$ on $(0, \sigma^{-1})$ which is Gamma($k-2, 1$) density restricted to $(0, \sigma^{-1})$. Draw s_{k-1} by the inverse-cdf method and set $r_{k-1} = 1 - \sigma s_{k-1}$. Then draw (s_1, \dots, s_{k-2}) from the Dirichlet distribution with all parameters set equal to 1 and set $r_j = (1 - r_{k-1})s_j$, $j = 1, \dots, k-2$. This gives us a draw from f .

5.6. Simulation from CB distribution. For the CB(A) distribution, we saw in Section 5.4 that $r \sim f(r) \propto \exp(\sum_{j=1}^{k-1} \lambda_j r_j)$ on S_{k-2} . Unless we have some more information on the eigen-values λ_j as in case of $CW(\mu, \sigma)$, it is not easy to simulate exactly from f . We may instead use a full conditional Gibbs sampling method to draw r . Draw r_j for $j = 1, \dots, k-2$ from the density proportional to $\exp((\lambda_j - \lambda_{k-1})r_j)$ on $(0, 1 - r^{(j)})$ using the inverse-cdf method where $r^{(j)} = \sum_{i=1, i \neq j}^{k-2} r_i$. Then set $r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$. Draw $\theta_1, \dots, \theta_{k-1}$ and compute $[z]$ as in Section 5.5.

Under high concentrations, that is when $\lambda_{k-1} \gg \lambda_{k-2}$, a more effective approach would be to use an independent exponential approximation. That is draw r_j , $j = 1, \dots, k-2$ independently from the density proportional to $\exp((\lambda_j - \lambda_{k-1})r_j)$ on $(0, 1)$. Accept the draw if $\sum_{i=1}^{k-2} r_i \leq 1$ and then set $r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$.

6. Density estimation on Σ_2^k

Since the planar shape space is a compact Riemannian manifold, we could use the kernel defined in Section 2.1 to build mixture density models on this space. However it is not easy to get an exact expression for the kernel because of the volume density term involved. Also to simulate from the posterior distribution of the density as in Section 2.3, if we write the conditional posteriors of the atoms and bandwidth using normal coordinates, then the expressions become messy. It is not easy to sample from them due to lack of conjugacy. In this section, we present an alternative kernel and construct mixture density models using that, for which the theoretical and numerical computations are greatly simplified, as we shall see in the subsequent sections.

Consider the Complex Watson kernel on the planar shape space as mentioned in Section 5.3. That is,

$$(6.1) \quad K(m; \mu, \sigma) = c^{-1}(\sigma) \exp\left(\frac{|x^*y|^2}{\sigma}\right)$$

where $m, \mu \in \Sigma_2^k$, x, y are some preshapes of m, μ respectively and

$$c(\sigma) = (\pi\sigma)^{(k-2)} \left[e^{1/\sigma} - \sum_{r=0}^{k-3} \frac{\sigma^{-r}}{r!} \right], \quad \sigma \in \mathbb{R}^+.$$

Then for fixed μ, σ ; $K(\cdot; \mu, \sigma)$ defines a valid probability density on Σ_2^k with respect to the volume measure. As shown in Section 5.4, it has an extrinsic mean μ_E equal to μ and extrinsic variation

$$V_E = 2(k-2)\sigma \frac{1 - e^{-1/\sigma} \sum_0^{k-2} r!^{-1} \sigma^{-r}}{1 - e^{-1/\sigma} \sum_0^{k-3} r!^{-1} \sigma^{-r}}$$

which is approximately a constant multiple of σ when σ is small. Note that K can be written as

$$c^{-1}(\sigma) \exp\left(\frac{1}{\sigma} \cos^2 d_g(m, \mu)\right).$$

Hence it is similar to the kernel in equation (2.1), except that now we need not put any constraint on the support of \mathcal{X} or σ for it to be a valid probability density.

Using this kernel, we can define a location mixture or a location-scale mixture density model on Σ_2^k as in (2.5) and (2.6) respectively. We set priors on the mixing parameters which induce corresponding priors on the space of densities. We prove that the induced priors satisfy the KL condition for both models which imply posterior consistency for the Bayes estimates of the densities. The computations are greatly simplified by using the shape coordinates described in Section 5.1 due to the fact that under this coordinate system, the uniform measure on Σ_2^k remains uniform on $S_{k-2} \times (-\pi, \pi)^{k-2}$. This observation helps us remove the constraints on the kernel parameters and prove KL property under lesser restrictions. It is proved in Theorems 6.2 and 6.3. In proving them, we will use the following lemma. The proof is given in the Appendix section 9.2.

LEMMA 6.1. *Let F_0 be an absolutely continuous probability distribution on Σ_2^k and let f_0 be its density. For a probability P on Σ_2^k and $\sigma > 0$, define*

$$f(m; P, \sigma) = \int_{\Sigma_2^k} K(m; \mu, \sigma) P(d\mu)$$

which is a valid probability density. Assume that f_0 is Holder Continuous on the metric space (Σ_2^k, d_E) , i.e. there exists constants $A, a > 0$ such that for any two points $p, q \in \Sigma_2^k$,

$$|f_0(p) - f_0(q)| \leq A d_E(p, q)^a.$$

Then

$$\sup_{m \in \Sigma_2^k} |f(m; F_0, \sigma) - f_0(m)| \longrightarrow 0$$

as $\sigma \rightarrow 0$.

Theorem 6.2 states the KL property for the prior induced on the space of densities on Σ_2^k using a location mixture density model while Theorem 6.3 states it in case of a location-scale mixture density model. The proofs follow from Lemma 6.1 just as Theorems 2.2 and 2.4 use Lemma 9.1 in their proofs, once we note that K is continuous in m, μ, σ on $\Sigma_2^k \times \Sigma_2^k \times \mathbb{R}^+$. Hence the proofs are omitted.

THEOREM 6.2. *Let f_0 be a Holder continuous density on Σ_2^k and F_0 be the corresponding probability distribution. Define*

$$(6.2) \quad f(m; P, \sigma) = \int_{\Sigma_2^k} K(m; \mu, \sigma) P(d\mu)$$

with K as in (6.1). Let Π_1 be a prior on $\mathcal{M}(\Sigma_2^k)$ which contains F_0 in its support. Let π_1 be a prior on \mathbb{R}^+ containing 0 in its support. Then for any $\epsilon > 0$,

$$(\Pi_1 \otimes \pi_1) \{ (P, \sigma) : \sup_{m \in \Sigma_2^k} |f_0(m) - f(m; P, \sigma)| < \epsilon \} > 0.$$

Further if $f_0(m) > 0 \forall m \in \Sigma_2^k$, then

$$(\Pi_1 \otimes \pi_1) \{ (P, \sigma) : f(\cdot; P, \sigma) \in KL(f_0, \epsilon) \} > 0.$$

THEOREM 6.3. *Define*

$$f(m; Q) = \int_{\Sigma_2^k \times \mathbb{R}^+} K(m; \mu, \sigma) Q(d\mu d\sigma).$$

Let Π_2 be a prior on $\mathcal{M}(M \times \mathbb{R}^+)$ containing $F_0 \otimes \delta_0$ in its support. Then if f_0 is Holder continuous and strictly positive on Σ_2^k , then

$$\Pi_2 \{ P : f(\cdot; P) \in KL(f_0, \epsilon) \} > 0.$$

6.1. Posterior computation. In this section, we describe an exact block Gibbs sampling algorithm for posterior computation in Dirichlet process location mixture of Complex Watson kernels using the mixture model in (6.2). The algorithm follows the general steps outlined in Section 2.3, with our goal being to simulate from the posterior distribution of f given an iid sample X_1, \dots, X_n from f and obtain a Bayes estimate for f . For the location-scale mixture, the computations are very similar and are left to the reader. The prior Π_1 on P is taken to be $DP(w_0 P_0)$ with $w_0 = 1$ and $P_0 = CW(\mu_0, \sigma_0)$ for some $\mu_0 \in \Sigma_2^k$, $\sigma_0 > 0$ while the prior π_1 for σ is chosen to be the Inverse Gamma distribution with some fixed hyper-parameters $a, b > 0$, i.e.,

$$\pi_1(d\sigma) \propto (\sigma^{-1})^{a+1} \exp(-b\sigma^{-1}), \quad \sigma > 0.$$

These prior choices cause posterior conjugacy as we shall see soon. In the algorithm described in Section 2.3 for sampling from the posterior distribution of the density, at any given iteration, the distinct location atoms μ_j^s are drawn from the conditional posterior

$$\begin{aligned} f(\mu_j^s) &\propto \prod_{i: S_i=j} K(X_i; \mu_j^s, \sigma) P_0(d\mu_j^s) \\ &\propto \exp\left\{ y^* \left(\frac{m_j}{\sigma} \bar{Z}_j + \frac{1}{\sigma_0} A_0 \right) y \right\} \end{aligned}$$

where y is some preshape of μ_j^s , m_j is the number of observations allocated to cluster j in the current iteration, \bar{Z}_j is the average of the embedded sample corresponding to cluster j , i.e.

$$\bar{Z}_j = \frac{1}{m_j} \sum_{i:S_i=j} J(X_i)$$

and $A_0 = J(\mu_0)$. This implies that

$$\mu_j^s | \{X_1, \dots, X_n, S_1, \dots, S_n, \sigma\} \sim \text{CB} \left(\frac{m_j}{\sigma} \bar{Z}_j + \frac{1}{\sigma_0} A_0 \right).$$

Hence the CB prior P_0 ensures conditional posterior conjugacy. We sample from this distribution by one of the methods described in Section 5.6. We draw σ from its full conditional posterior

$$\begin{aligned} g(\sigma) &\propto \prod_{i=1}^n K(X_i; \mu_i, \sigma) \pi_1(d\sigma) \\ &\propto (\sigma^{-1})^{n(k-2)+a+1} \exp \left\{ - \left(n + b - \sum_{j=1}^{S(n)} m_j y_j^* \bar{Z}_j y_j \right) \sigma^{-1} \right\} \\ &\quad \left(1 - e^{-1/\sigma} \sum_{r=0}^{k-3} \frac{1}{r!} \sigma^{-r} \right)^{-n} \end{aligned}$$

where y_j denotes some preshape for μ_j^s , $j = 1, \dots, S(n)$. For σ small, this conditional density is approximately equal to that of

$$\text{IG} \left(n(k-2) + a, b + \sum_{j=1}^{S(n)} m_j (1 - y_j^* \bar{Z}_j y_j) \right).$$

Hence we get approximate conjugacy for the conditional distribution of σ once we choose a IG prior π . Numerical studies show that this approximation is very accurate even for σ moderately small. Hence an independent Metropolis Hastings step for updating σ , with candidates generated from the IG approximation, should be highly efficient.

7. Application to simulated data

We draw an iid sample of size 200: X_1, \dots, X_n , $n = 200$, on the planar shape space Σ_2^k , $k = 4$, from the density

$$\begin{aligned} f_0 &= 0.5\text{CW}(\mu_1, \sigma_0) + 0.5\text{CW}(\mu_2, \sigma_0) \text{ with} \\ \sigma_0 &= .001, \mu_1 = [(1, 0, 0)'], \mu_2 = [(r, \sqrt{1-r^2}, 0)'] \text{ where } r = .9975. \end{aligned}$$

We try three different density estimates for f_0 , namely a nonparametric (np) Bayesian density estimate as obtained in Section 6.1, a frequentist parametric estimate and a kernel density estimate (KDE). We compare their performance by estimating the distance between the true density and the density estimate. We use two types of distances, namely the L^1 distance and the Kullback-Leibler (KL) divergence. It turns out that the np Bayes estimate performs the best in both cases.

The L^1 divergence between f_0 and another density f_1 is given by

$$d_1(f_0, f_1) = \int_{\Sigma_2^k} |f_0(m) - f_1(m)| V(dm)$$

which can be estimated consistently by

$$\hat{d}_1(f_0, f_1) = \frac{1}{n} \sum_{i=1}^n \left| 1 - \frac{f_1(X_i)}{f_0(X_i)} \right|.$$

The KL divergence between f_0 and f_1 is defined to be

$$d_2(f_0, f_1) = \int_{\Sigma_2^k} f_0(m) \log \left(\frac{f_0(m)}{f_1(m)} \right) V(dm),$$

a consistent estimator of which is given by

$$\hat{d}_2(f_0, f_1) = \frac{1}{n} \sum_{i=1}^n \log \left(\frac{f_0(X_i)}{f_1(X_i)} \right).$$

To get the Bayes estimate, we estimate f_0 using expression (2.8) averaged over a large number of iterations of the exact block Gibbs sampler described in Section 6.1 for the DP location mixture of CW kernels model. To complete a specification of the model, we let $P_0 = CW(\hat{\mu}_E, 0.1)$, with $\hat{\mu}_E$ being the sample extrinsic mean. By using the data to estimate the center of the base distribution, while choosing a moderate variance, we ensure that the prior introduces clusters close to the support of the data. This default leads to better performance than using a uniform base measure which is the limit of CW distributions as $\sigma \rightarrow \infty$. The prior π_1 for σ is set to be $IG(1, .1)$ and the DP precision parameter is fixed as $\omega_0 = 1$, which is a commonly-used default in the literature, which favors a sparse representation with few clusters.

We ran the Gibbs sampler for 100,000 iterations, with the first 15,000 discarded as a burn-in. Posterior summaries of the distances, including posterior means and credible intervals are summarized as follows:

$$\begin{aligned} \bar{d}_1 &= 0.3374, & 95\%CI &= (0.2308, 0.4538), & 99\%CI &= (0.2009, 0.4931) \\ \bar{d}_2 &= 0.0669, & 95\%CI &= (0.0234, 0.1227), & 99\%CI &= (0.0135, 0.1426) \end{aligned}$$

To get a single kernel based frequentist density estimate, we fit a $CW(\mu, \sigma)$ distribution to the data, estimating μ and σ by their MLEs $\hat{\mu}_{mle}$ and $\hat{\sigma}_{mle}$ respectively. Let \bar{Z} denote the embedded sample mean, let $\hat{\lambda}_{k-1}$ denote its largest eigen value and let \hat{U}_{k-1} be a corresponding unit eigen vector. It is shown in [4] that $\hat{\mu}_{mle} = [\hat{U}_{k-1}]$ which is the sample extrinsic mean and under high concentrations (i.e. $\hat{\lambda}_{k-1}$ close to 1) $\hat{\sigma}_{mle}$ is approximately equal to $\frac{1 - \hat{\lambda}_{k-1}}{k-2}$ which is $\frac{\hat{V}_E}{2(k-2)}$ where \hat{V}_E denotes the sample extrinsic variation. Denoting the density estimate by $\hat{f}_{mle} \equiv CW(\hat{\mu}_{mle}, \hat{\sigma}_{mle})$, the estimated distances from the true density f_0 turn out to be

$$\hat{d}_1(f_0, \hat{f}_{mle}) = 0.7182, \quad \hat{d}_2(f_0, \hat{f}_{mle}) = 0.4727.$$

Finally we use a frequentist KDE

$$\hat{f}(m) = \frac{1}{n} \sum_{j=1}^n K(m; X_j, h)$$

with K as in (6.1) and fixed band-width $h > 0$. We may take h to be equal to σ_0 or $\hat{\sigma}_{mle}$ or the Bayes mean $\bar{\sigma}$ for the posterior distribution of σ in the model $f(\cdot; P, \sigma)$. It turns out that $\hat{\sigma}_{mle} = 0.0017$ and $\bar{\sigma} = 0.0014$. The values for $\hat{d}_1(f_0, \hat{f})$ and $\hat{d}_2(f_0, \hat{f})$ for various values of h are shown in table 1. Also included are the performance of the np Bayes and single kernel estimates for a side by side comparison. It shows that the nonparametric Bayes density estimate performs much

TABLE 1. Estimated divergence from f_0 for 3 density estimates

h	KDE		np Bayes		\hat{f}_{mle}	
	\hat{d}_1	\hat{d}_2	\bar{d}_1	\bar{d}_2	\hat{d}_1	\hat{d}_2
0.001	0.8404	0.2649	0.3374	0.0669	0.7182	0.4727
0.0014	0.8473	0.4833				
0.0017	0.8691	0.6238				
0.0009	0.8548	0.20007				

better than the parametric estimate and the KDE.

8. Application to morphometrics: classification of gorilla skulls

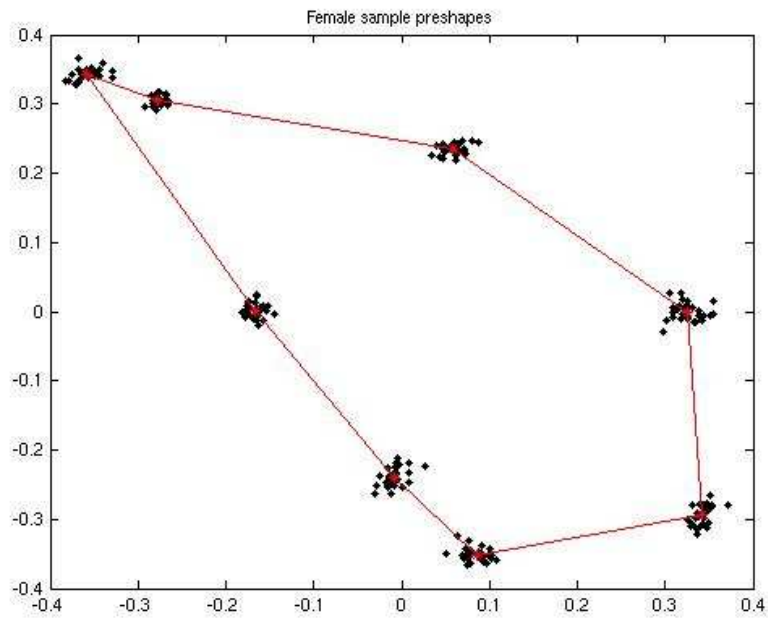
In this real life example, eight landmarks are chosen on the midline plane of 2D images of some gorilla skulls. There are 29 male and 30 female gorillas in the sample. The data can be found in Dryden and Mardia [4]. The goal is to study the shapes of the skulls and use that to build a classifier to determine the sex of a gorilla from its skull's shape. This finds application in morphometrics and other biological sciences.

Figure 1 shows the plot of the preshapes of the k-ads along with the preshapes of the sample extrinsic means for the two groups. The sample preshapes have been rotated appropriately to bring them closest to the chosen preshapes for the means. Figure 2 plots the nonparametric Bayes estimates of the shape densities for the two groups along with 95% credible regions. These estimates were obtained using the same model, prior and computational algorithm applied in Section 7 for the simulated data. The plots show the densities conditioned to the geodesic starting from the female group's mean shape and directed towards the male group's mean shape.

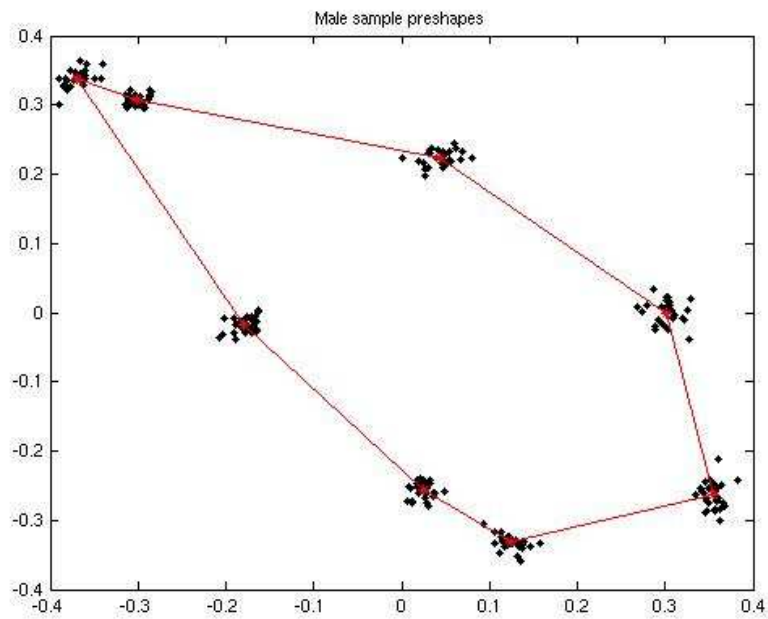
To carry out a discriminant analysis, we randomly pick 25 shapes from each sample as training data sets and the remaining 9 are used as test data. Then we estimate the shape densities independently from the test data for each sex, and find the conditional probability of being female for each of the test sample shapes. If we denote by π the prior probability of being female, by $\hat{f}_1(m)$ and $\hat{f}_2(m)$ the female and male predictive densities evaluated at a shape m , then the posterior probability of being female for the shape m given the training sample of shapes is

$$p = \frac{\pi \hat{f}_1(m)}{\pi \hat{f}_1(m) + (1 - \pi) \hat{f}_2(m)}.$$

We take $\pi = 0.5$. Table 2 presents the posterior mean \hat{p} of p along with a 95% Credible Interval for p for each of the test sample shapes. In this table the first



(a)



(b)

FIGURE 1. (a) and (b) show 8 landmarks from skulls of 30 female and 29 male gorillas respectively along with the respective sample mean shapes. * correspond to the mean shapes' landmarks.

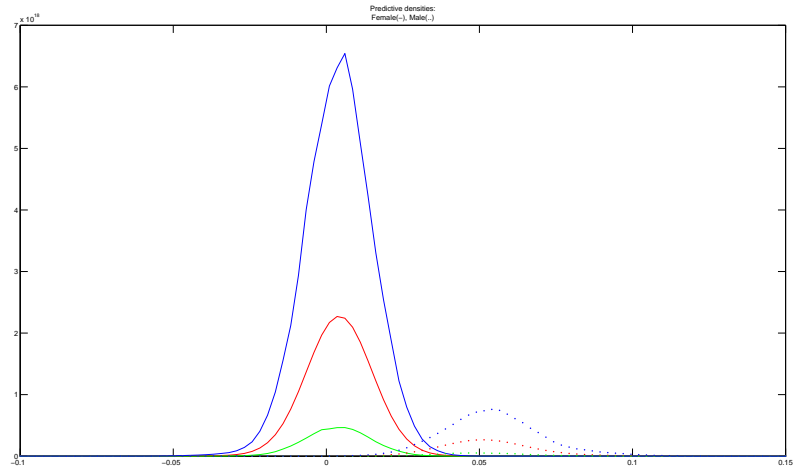


FIGURE 2. Densities for gorilla shapes

TABLE 2. Conditional prob. of being female given the shape and distances from female & male mean shapes

True Gender	\hat{p}	CI	$d(., \mu_f)$	$d(., \mu_m)$
F	1	(1,1)	.041	.1109
F	.9999	(.9992,1)	.0362	.0934
F	.16	(.008,.602)	.056	.0517
F	.9958	(.968, 1)	.0495	.0952
F	1	(1, 1)	.0755	.135
M	.0001	(0, 0)	.1672	.1033
M	.0005	(0, .003)	.087	.0417
M	.983	(.8197, 1)	.0911	.1207
M	.0003	(0, 0)	.1523	.0935

five shapes correspond to female gorillas while the last four are males. There is some uncertainty in the classification of sample 3 while sample 8 is misclassified. Figure 3 plots the preshapes of the test samples along with that of the mean shapes from the male and female groups.

We may also build a distance based classifier by comparing the distance of any given shape from the mean shapes of the female and male training sets. Columns 4 and 5 of table 2 present the extrinsic distance of each of the test sample shapes from the female and male extrinsic means respectively. Using this classifier, samples 3 and 8 are misclassified. The disadvantage of using such a classifier is that it is deterministic in nature - there is no measure for the uncertainty in classifying.

9. Appendix

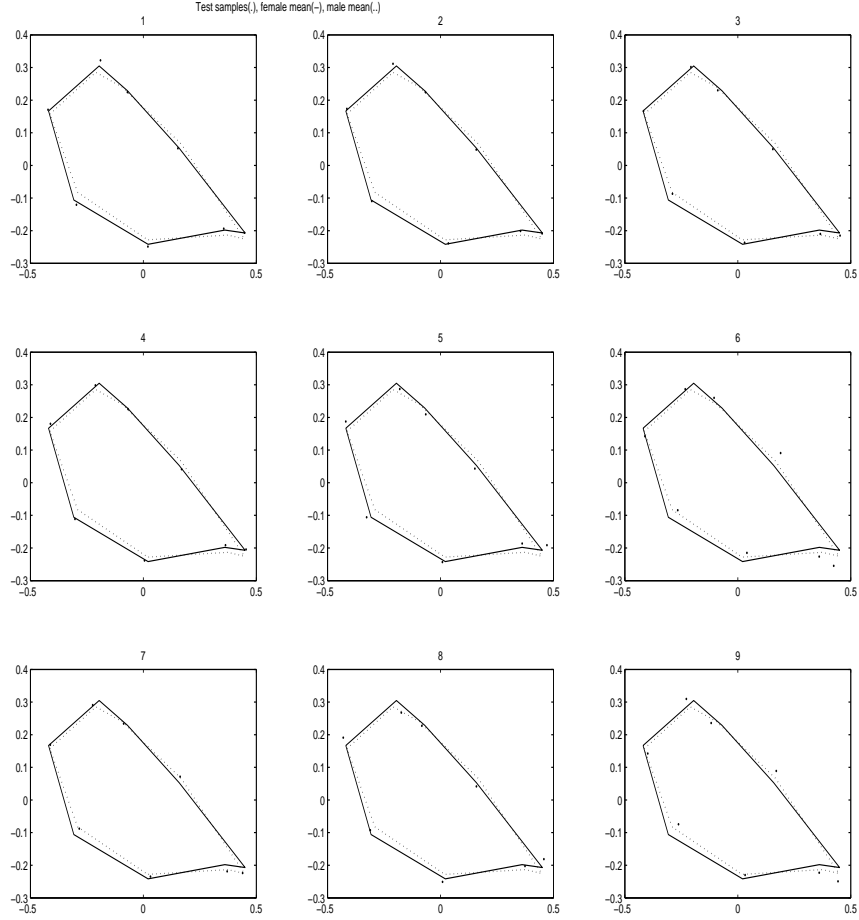


FIGURE 3. Preshapes for test samples. Sample (\cdot), Female mean ($-$), Male Mean ($\cdot\cdot$)

9.1. Proofs of Theorems 2.2 and 2.4. To prove Theorems 2.2 and 2.4, we will need the following lemmas.

LEMMA 9.1. *If f_0 is a continuous density on M , then*

(1) *for any $\epsilon > 0$, there exists a $\sigma_\epsilon \in (0, Ar_1]$ such that*

$$\sup_{m \in M} |f_0(m) - f(m; F_0, \sigma)| < \epsilon \quad \forall \sigma \leq \sigma_\epsilon.$$

(2) *Furthermore if $f_0(m) > 0 \quad \forall m \in M$, then we can choose σ_ϵ such that*

$$\int_M f_0(m) \log \left(\frac{f_0(m)}{f(m; F_0, \sigma)} \right) V(dm) < \epsilon \quad \forall \sigma \leq \sigma_\epsilon.$$

PROOF. From Proposition 2.1, it follows that

$$f_0(m) = \int_M K(\mu; m, \sigma) f_0(\mu) V(d\mu).$$

In a normal neighborhood of m , $G_\mu(m) = G_m(\mu)$ and hence K is symmetric in m and μ . Also $K(m; \cdot, \sigma)$ is zero outside $B(m, r_*)$. Therefore we can write

$$(9.1) \quad f(m; F_0, \sigma) - f_0(m) = \int_{B(m, r_*)} K(m; \mu, \sigma) \{f_0(\mu) - f_0(m)\} V(d\mu).$$

For convenience let us use $f(m)$ for $f(m; F_0, \sigma)$. Since $B(m, r_*)$ lies in a normal neighborhood of m , using normal coordinates into $T_m M$, equation (9.1) simplifies to

$$(9.2) \quad f(m) - f_0(m) = \sigma^{-d} \int_{\|y\| < r_*} \mathcal{X}\left(\frac{\|y\|}{\sigma}\right) \{f_1(y) - f_1(0)\} dy$$

$$(9.3) \quad = \int_{\|y\| \leq \frac{1}{\sigma}} \mathcal{X}(\|y\|) \{f_1(\sigma y) - f_1(0)\} dy$$

where $f_1(y) = f_0(\exp_m(y))$. Since f_0 is uniformly continuous on the compact metric-space (M, d_g) , given any $\epsilon > 0$, there exists a $\delta > 0$ such that $m_1, m_2 \in M$, $d_g(m_1, m_2) < \delta$ implies that $|f_0(m_1) - f_0(m_2)| < \epsilon$. In equation (9.3),

$$f_1(\sigma y) - f_1(0) = f_0(\exp_m(\sigma y)) - f_0(m).$$

Note that $d_g(m, \exp_m(\sigma y)) = \sigma \|y\| \leq \frac{\sigma}{A}$. Hence by choosing $\sigma < A\delta$, we can ensure that $|f_1(\sigma y) - f_1(0)| < \epsilon$ and hence $|f(m) - f_0(m)| < \epsilon$ for all m in M . This proves (1).

To prove (2), note that $c_0 = \inf_{m \in M} f_0(m)$ is strictly positive. Given any $\delta > 0$, choose σ_δ in (1) to be such that $\sup_{m \in M} |f_0(m) - f(m)| < \delta \forall \sigma \leq \sigma_\delta$. Then $\inf_{m \in M} f(m) > c_0 - \delta > 0$ for δ sufficiently small. Hence

$$\int_M f_0(m) \log\left(\frac{f_0(m)}{f(m)}\right) V(dm) \leq \sup_{m \in M} \left| \frac{f_0(m)}{f(m)} - 1 \right| \leq \frac{\delta}{c_0 - \delta} < \epsilon$$

for δ sufficiently small. This completes the proof. \square

LEMMA 9.2. *Given $\epsilon > 0$, if there exists*

(1) *a $\sigma_\epsilon > 0$ and a $P_\epsilon \in \mathcal{M}(M)$ such that*

$$\sup_{m \in M} |f_0(m) - f(m; P_\epsilon, \sigma_\epsilon)| < \frac{\epsilon}{3},$$

(2) *a set W containing σ_ϵ and $\pi_1(W) > 0$ such that*

$$\sup_{m \in M, \sigma \in W} |f(m; P_\epsilon, \sigma_\epsilon) - f(m; P_\epsilon, \sigma)| < \frac{\epsilon}{3},$$

and

(3) *a $\mathcal{W} \subseteq \mathcal{M}(M)$ with $P_\epsilon \in \mathcal{W}$ and $\Pi_1(\mathcal{W}) > 0$ such that*

$$\sup_{m \in M, P \in \mathcal{W}, \sigma \in W} |f(m; P_\epsilon, \sigma) - f(m; P, \sigma)| < \frac{\epsilon}{3},$$

then

$$\sup_{m \in M} |f_0(m) - f(m; P, \sigma)| < \epsilon$$

for all $(P, \sigma) \in \mathcal{W} \times W$.

PROOF. Follows from a direct application of the triangular inequality. \square

PROOF OF THEOREM 2.2. The result follows from Lemma 9.2 if we can verify conditions (1), (2) and (3) because then

$$(\Pi_1 \otimes \pi_1)\{(P, \sigma) : \sup_{m \in M} |f_0(m) - f(m; P, \sigma)| < \epsilon\} \geq \Pi_1(\mathcal{W})\pi_1(W) > 0.$$

Condition (1) is verified from Lemma 9.1 (1) with $P_\epsilon = F_0$. Since $0 \in \text{supp}(\pi_1)$ and $\pi_1(\{0\}) = 0$, we can choose σ_ϵ sufficiently small so that $\sigma_\epsilon \in \text{supp}(\pi_1)$.

Next we need to find a W for which condition (2) is satisfied. First we show that $K(m; \mu, \sigma) = \sigma^{-d} \mathcal{X}\left(\frac{d_g(m, \mu)}{\sigma}\right) G_\mu^{-1}(m)$ is a continuous function of (m, μ, σ) on $M \times M \times (0, Ar_1]$ (under the product topology). We prove that as follows. Firstly note that $(m, \mu) \mapsto d_g(m, \mu)$ is continuous on $M \times M$. Since \mathcal{X} is continuous on $[0, \infty)$, therefore $(m, \mu, \sigma) \mapsto \mathcal{X}\left(\frac{d_g(m, \mu)}{\sigma}\right)$ is continuous on $M \times M \times (0, \infty)$. Also since $(m, \mu) \mapsto G_\mu(m)$ is a non-zero continuous function on

$$\{(m, \mu) \in M \times M : d_g(m, \mu) < r_*\},$$

therefore $G_\mu^{-1}(m)$ is also continuous in the above set. Therefore $K(m; \mu, \sigma)$ is continuous on

$$\{(m, \mu) \in M \times M : d_g(m, \mu) \leq r_1\} \times (0, \infty).$$

Since $K(m; \mu, \sigma) = 0$ if $d_g(m, \mu) \geq r_1$, therefore K is continuous on $M \times M \times (0, \infty)$ and hence uniformly continuous on $M \times M \times [\frac{\sigma_\epsilon}{2}, Ar_1]$ (under the L^1 metric) and bounded on this set, say by \bar{K} . This implies that $\sigma \mapsto K$ is uniformly equicontinuous on $[\frac{\sigma_\epsilon}{2}, Ar_1]$. Hence we can get a compact set $W \subseteq [\frac{\sigma_\epsilon}{2}, Ar_1]$ containing σ_ϵ in its interior such that

$$|K(m; \mu, \sigma) - K(m; \mu, \sigma_\epsilon)| < \frac{\epsilon}{3} \quad \forall (m, \mu, \sigma) \in M \times M \times W.$$

Then

$$\begin{aligned} & \sup_{m \in M, \sigma \in W} |f(m; F_0, \sigma) - f(m; F_0, \sigma_\epsilon)| \\ & \leq \int_M \sup_{m \in M, \sigma \in W} |K(m; \mu, \sigma) - K(m; \mu, \sigma_\epsilon)| f_0(\mu) V(d\mu) \\ & \leq \sup_{m, \mu \in M, \sigma \in W} |K(m; \mu, \sigma) - K(m; \mu, \sigma_\epsilon)| < \frac{\epsilon}{3}. \end{aligned}$$

Since $\sigma_\epsilon \in \text{supp}(\pi_1)$ and W contains an open neighborhood of σ_ϵ , therefore $\pi_1(W) > 0$. This verifies condition (2).

Lastly we need to find a \mathcal{W} for which condition (3) is satisfied. We claim that

$$\mathcal{W} = \{P : \sup_{m \in M, \sigma \in W} |f(m; P, \sigma) - f(m; F_0, \sigma)| < \frac{\epsilon}{3}\}$$

contains a weakly open neighborhood of F_0 . To prove this claim, note that for any $m \in M$, $\sigma \in W$, $\mu \mapsto K(m; \mu, \sigma)$ defines a bounded continuous function on M . Hence

$$\mathcal{W}_{m, \sigma} = \{P : |f(m; P, \sigma) - f(m; F_0, \sigma)| < \frac{\epsilon}{9}\}$$

defines a weakly open subset of $\mathcal{M}(M)$ for all $(m, \sigma) \in M \times W$. Now we show that $(m, \sigma) \mapsto f(m; P, \sigma)$ is a uniformly equicontinuous family of functions on $M \times W$ labeled by $P \in \mathcal{M}(M)$. That is because, for $m_1, m_2 \in M$; $\sigma, \tau \in W$,

$$|f(m_1; P, \sigma) - f(m_2; P, \tau)| \leq \int_M |K(m_1; \mu, \sigma) - K(m_2; \mu, \tau)| P(d\mu)$$

and K is uniformly continuous on $M \times M \times W$. Therefore there exists a $\delta > 0$ such that $d_g(m_1, m_2) + |\sigma - \tau| < \delta$ implies that

$$\sup_{P \in \mathcal{M}(M)} |f(m_1; P, \sigma) - f(m_2; P, \tau)| < \frac{\epsilon}{9}.$$

Cover $M \times W$ by finitely many balls of radius δ : $M \times W = \bigcup_{i=1}^N B((m_i, \sigma_i), \delta)$. Let $\mathcal{W}_1 = \bigcap_{i=1}^N \mathcal{W}_{m_i, \sigma_i}$ which is an open neighborhood of F_0 . Let $P \in \mathcal{W}_1$ and $(m, \sigma) \in M \times W$. Then there exists a (m_i, σ_i) such that $(m, \sigma) \in B((m_i, \sigma_i), \delta)$. Then

$$\begin{aligned} & |f(m; P, \sigma) - f(m; F_0, \sigma)| \\ & \leq |f(m; P, \sigma) - f(m_i; P, \sigma_i)| + |f(m_i; P, \sigma_i) - f(m_i; F_0, \sigma_i)| + |f(m_i; F_0, \sigma_i) - f(m; F_0, \sigma)| \\ & < \frac{\epsilon}{9} + \frac{\epsilon}{9} + \frac{\epsilon}{9} = \frac{\epsilon}{3}. \end{aligned}$$

This proves that \mathcal{W} contains \mathcal{W}_1 and hence the claim is proved. Since $F_0 \in \text{supp}(\Pi_1)$, therefore $\Pi_1(\mathcal{W}) > 0$. Hence condition (3) is satisfied. This completes the proof. \square

PROOF OF THEOREM 2.4. From Lemma 9.1 it follows that given any $\delta_1 > 0$, we can find a $\sigma_1 > 0$ such that with $P_1 = F_0 \otimes \delta_{\sigma_1}$,

$$(9.4) \quad \begin{aligned} & \sup_{m \in M} |f_0(m) - f(m; P_1)| < \delta_1 \text{ and} \\ & \int_M f_0(m) \log \left(\frac{f_0(m)}{f(m; P_1)} \right) V(dm) < \delta_1. \end{aligned}$$

Hence if we choose $\delta_1 \leq \frac{c_0}{2}$ where $c_0 = \inf_{m \in M} f_0(m) > 0$ then $\inf_{m \in M} f(m; P_1) \geq \frac{c_0}{2}$. Since $F_0 \otimes \delta_0 \in \text{supp}(\Pi_2)$ and $\Pi_2(\{F_0 \otimes \delta_0\}) = 0$, we can choose σ_1 sufficiently small so that $P_1 \in \text{supp}(\Pi_2)$. Get a compact set E in $(0, \infty)$ containing σ_1 in its interior. From the proof of Theorem 2.2, it follows that $K(m; \mu, \sigma)$ is continuous, hence uniformly continuous on $M \times M \times E$. For $P \in \mathcal{M}(M \times (0, Ar_1])$, define

$$f(m; P_E) = \int_{M \times E} K(m; \mu, \sigma) P(d\mu d\sigma).$$

Denote by ∂A , the boundary of any set A . Since M is a manifold, it has no boundary, hence $\partial(M \times E) = M \times \partial E$. Since $(\mu, \sigma) \mapsto K(m; \mu, \sigma)$ is uniformly equicontinuous as a family of functions labeled by $m \in M$ on $M \times E$ and $P_1\{\partial(M \times E)\} = P_1(M \times \partial E) = 0$, therefore for $\delta_2 > 0$,

$$\mathcal{W}_m(\delta_2) = \{P : |f(m; P_E) - f(m; P_1)| < \delta_2\}$$

defines a weakly open neighborhood of P_1 . Since $P_1 \in \text{supp}(\Pi_2)$, therefore $\Pi_2(\mathcal{W}_m(\delta_2)) > 0$. We also claim that if

$$\mathcal{W} = \{P : \sup_{m \in M} |f(m; P_E) - f(m; P_1)| < \delta_2\},$$

then $\Pi_2(\mathcal{W}) > 0$. To see that get $\delta_3 > 0$ such that $d_g(m_1, m_2) < \delta_3$ implies that

$$\sup_{(\mu, \sigma) \in M \times E} |K(m_1; \mu, \sigma) - K(m_2; \mu, \sigma)| < \frac{\delta_2}{3}$$

which in turn implies that

$$(9.5) \quad |f(m_1; P_E) - f(m_2; P_E)| < \frac{\delta_2}{3} \quad \forall P \in \mathcal{M}(M \times (0, Ar_1]).$$

Cover M by finitely many balls of radius δ_3 : $M = \bigcup_{i=1}^N B(m_i, \delta_3)$. Then we show that $\mathcal{W} \supseteq \bigcap_{i=1}^N \mathcal{W}_{m_i}(\frac{\delta_2}{3})$. To prove that pick $P \in \bigcap_{i=1}^N \mathcal{W}_{m_i}(\frac{\delta_2}{3})$. Then

$$|f(m_i; P_E) - f(m_i; P_1)| < \delta_2 \quad \forall i = 1, 2, \dots, N.$$

Pick $m \in M$, say $m \in B(m_i, \delta_3)$. Equation (9.5) implies that

$$|f(m; P_E) - f(m_i; P_E)| < \delta_2 \quad \forall P.$$

Hence

$$\begin{aligned} & |f(m; P_E) - f(m; P_1)| \\ & \leq |f(m; P_E) - f(m_i; P_E)| + |f(m_i; P_E) - f(m_i; P_1)| + |f(m_i; P_1) - f(m; P_1)| \\ & < \frac{\delta_2}{3} + \frac{\delta_2}{3} + \frac{\delta_2}{3} = \delta_2. \end{aligned}$$

Hence $\mathcal{W} \supseteq \bigcap_{i=1}^N \mathcal{W}_{m_i}(\frac{\delta_2}{3})$ which is a open neighborhood of P_1 . Therefore $\Pi_2(\mathcal{W}) > 0$. For $P \in \mathcal{W}$,

$$\inf_{m \in M} f(m; P_E) \geq \inf_{m \in M} f(m; P_1) - \delta_2 \geq \frac{c_1}{4}$$

if $\delta_2 < \frac{c_1}{4}$. Then

$$\begin{aligned} & \int_M f_0(m) \log \left(\frac{f(m; P_1)}{f(m; P)} \right) V(dm) < \int_M f_0(m) \log \left(\frac{f(m; P_1)}{f(m; P_E)} \right) V(dm) \\ (9.6) \quad & \leq \sup_{m \in M} \left| \frac{f(m; P_1)}{f(m; P_E)} - 1 \right| \leq \frac{\delta_2}{c_1/4} < \delta_1 \end{aligned}$$

provided δ_2 is sufficiently small. From (9.4) and (9.6) we deduce that, for $P \in \mathcal{W}$,

$$\begin{aligned} & \int_M f_0(m) \log \left(\frac{f_0(m)}{f(m; P)} \right) V(dm) = \\ & \int_M f_0(m) \log \left(\frac{f_0(m)}{f_1(m)} \right) V(dm) + \int_M f_0(m) \log \left(\frac{f_1(m)}{f(m; P)} \right) V(dm) \\ & < \delta_1 + \delta_1 = \epsilon \end{aligned}$$

if $\delta_1 = \epsilon/2$. Hence

$$\{f(\cdot; P) : P \in \mathcal{W}\} \subseteq KL(f_0, \epsilon)$$

and therefore

$$\Pi_2\{P : f(\cdot; P) \in KL(f_0, \epsilon)\} > 0.$$

Since ϵ was arbitrary, the proof is completed. \square

9.2. Proof of Lemma 6.1.

. Since K is symmetric in m and μ , therefore

$$\int K(m; \mu, \sigma) V(dm) = \int K(m; \mu, \sigma) V(d\mu)$$

Hence we can write $|f(m; F_0, \sigma) - f_0(m)|$ as

$$\begin{aligned} & \left| \int_{\Sigma_2^k} K(m; \mu, \sigma) f_0(\mu) V(d\mu) - \int_{\Sigma_2^k} K(m; \mu, \sigma) f_0(m) V(d\mu) \right| \\ (9.7) \quad & = \left| \int_{\Sigma_2^k} \{f_0(\mu) - f_0(m)\} K(m; \mu, \sigma) V(d\mu) \right|. \end{aligned}$$

Let x and y be some preshapes for m and μ respectively in $\mathbb{C}S^{k-2}$, so that $m = [x]$ and $\mu = [y]$. Let V_1 denote the volume-form on $\mathbb{C}S^{k-2}$. Then for any integrable function $\phi : \Sigma_2^k \rightarrow \mathbb{R}$,

$$\int_{\Sigma_2^k} \phi(m) V(dm) = \frac{1}{2\pi} \int_{\mathbb{C}S^{k-2}} \phi([x]) V_1(dx).$$

Hence the integral in (9.7) can be written as

$$(9.8) \quad \frac{c^{-1}(\sigma)}{2\pi} \left| \int_{\mathbb{C}S^{k-2}} \{f_0([y]) - f_0([x])\} \exp(\sigma^{-1} y^* x x^* y) V_1(dy) \right|.$$

Consider a s.v.d. of $x x^*$ as $x x^* = U \Lambda U^*$ where $\Lambda = \text{diag}(1, 0, \dots, 0)$ and $U = [U_1, \dots, U_{k-1}]$ with $U_1 = x$. Then

$$y^* x x^* y = z^* \Lambda z = |z_1|^2$$

where

$$z = U^* y = (z_1, \dots, z_{k-1})'.$$

Make a change of variable $y \mapsto z$ in (9.8). This does not change the volume form because of it being an orthogonal transformation. Then (9.8) becomes

$$(9.9) \quad \frac{c^{-1}(\sigma)}{2\pi} \left| \int_{\mathbb{C}S^{k-2}} \{f_0([Uz]) - f_0([x])\} \exp(\sigma^{-1} |z_1|^2) V_1(dz) \right|.$$

Write $z_j = \sqrt{r_j} e^{i\theta_j}$, $j = 1, \dots, k-1$, where $r = (r_1, \dots, r_{k-1})' \in S_{k-2}$ and $\theta = (\theta_1, \dots, \theta_{k-1})' \in [0, 2\pi)^{k-1}$, then

$$V_1(dz) = 2^{2-k} dr_1 \dots dr_{k-2} d\theta_1 \dots d\theta_{k-1}.$$

Hence (9.9) can be written as

$$(9.10) \quad \frac{c^{-1}(\sigma) \pi^{-1} 2^{1-k}}{2\pi} \left| \int_{S_{k-2} \times [0, 2\pi)^{k-1}} \{f_0([y(r, \theta, x)]) - f_0([x])\} \exp\left(\frac{r_1}{\sigma}\right) dr d\theta \right|$$

where

$$y \equiv y(r, \theta, x) = \sum_{j=1}^{k-1} \sqrt{r_j} e^{i\theta_j} U_j.$$

Then

$$d_E^2([y], [x]) = 2(1 - r_1).$$

By the Holder continuity of f_0 , we get that

$$|f_0([y]) - f_0([x])| \leq A(1 - r_1)^\alpha$$

for some $A, \alpha > 0$. Then from (9.10), we deduce that

$$(9.11) \quad \begin{aligned} & \sup_{m \in \Sigma_2^k} |f(m; F_0, \sigma) - f_0(m)| \\ & \leq c^{-1}(\sigma) \pi^{-1} 2^{1-k} A \int_{S_{k-2} \times [0, 2\pi)^{k-1}} (1 - r_1)^\alpha \exp\left(\frac{r_1}{\sigma}\right) dr d\theta \\ & = \frac{\pi^{k-2}}{(k-3)!} c^{-1}(\sigma) A \int_0^1 (1 - r_1)^{\alpha+k-3} \exp\left(\frac{r_1}{\sigma}\right) dr_1 \\ & = \frac{\pi^{k-2}}{(k-3)!} c^{-1}(\sigma) \sigma^{k-2+\alpha} e^{1/\sigma} A \int_0^{\sigma^{-1}} e^{-s} s^{k-3+\alpha} ds \\ & \leq g(\sigma) \int_0^\infty e^{-s} s^{k-3+\alpha} ds \end{aligned}$$

with

$$g(\sigma) = \frac{\pi^{k-2}}{(k-3)!} c^{-1}(\sigma) \sigma^{k-2+\alpha} e^{1/\sigma} A.$$

Hence (9.11) converges to zero if $g(\sigma) \rightarrow 0$ as $\sigma \rightarrow 0$. Using the expression for $c(\sigma)$, $g(\sigma)$ can be written as

$$\begin{aligned} g(\sigma) &= (k-3)!^{-1} \sigma^\alpha e^{1/\sigma} [e^{1/\sigma} - \sum_{r=0}^{k-3} r!^{-1} \sigma^{-r}]^{-1} \\ &= (k-3)!^{-1} \sigma^\alpha [1 - \sum_{r=0}^{k-3} e^{-1/\sigma} r!^{-1} \sigma^{-r}]^{-1}. \end{aligned}$$

Since $1 - \sum_{r=0}^{k-3} e^{-1/\sigma} r!^{-1} \sigma^{-r} \rightarrow 1$ and $\sigma^\alpha \rightarrow 0$ as $\sigma \rightarrow 0$, therefore $g(\sigma) \rightarrow 0$ and this completes the proof. \square

References

- [1] BHATTACHARYA, A. AND BHATTACHARYA, R. (2008). Nonparametric Statistics on Manifolds with Applications to Shape Spaces. *Pushing the Limits of Contemporary Statistics: Contributions in honor of J.K. Ghosh. IMS Collections* **3** 282-301.
- [2] BHATTACHARYA, R. N. AND PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds-I. *Ann. Statist.* **31** 1-29.
- [3] BUSH & MACEACHERN, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83** 275-285.
- [4] DRYDEN, I. L. & MARDIA, K.V. (1998). *Statistical Shape Analysis*. Wiley N.Y.
- [5] ESCOBAR, M.D. & WEST, M. (1995). Bayesian density-estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90** 577-588.
- [6] FERGUSON, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209-230.
- [7] FERGUSON, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.
- [8] GHOSAL, S., GHOSH, J.K. AND RAMAMOORTHI, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143-158.
- [9] ISHWARAN, H. & JAMES, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.* **96**, 161-73.
- [10] KENT, J.T. (1994). The complex Bingham distribution and shape analysis. *J. Roy. Statist. Soc. Ser. B* **56** no.2, 285-299.
- [11] KIM, J., CETIN, M. & WILLSKY, A.S. (2007). Nonparametric shape priors for active contour-based image segmentation. *Signal Processing* **87**, 3021-3044.
- [12] KUME, A. & WALKER, S.G. (2006). Sampling from compositional and directional distributions. *Statist. Comput.* **16**, 261-265.
- [13] LENNOX, K.P., DAHL, D.B., VANNUCCI, M. & TSAI, J.W. (2009). Density estimation for protein configuration angles using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Am. Statist. Assoc.*, to appear.
- [14] LO, A.Y. (1984). On a class of Bayesian nonparametric estimates - 1. Density estimates. *Ann. Statist.* **12**, 351-357.
- [15] MACEACHERN, S.N. (1994). Estimating Normal means with a conjugate style Dirichlet Process prior. *Commun. in Statist.-simulation and computat.* **23**, 727-741.
- [16] MARDIA, K.V., TAYLOR, C.C. & SUBRAMANIAM, G.K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* **63**, 505-512.
- [17] NEAL, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Statist.*, **9**, 249-265.
- [18] PAPASPILIOPOULOS (2008). A note on posterior sampling from Dirichlet mixture models. *Working Paper, 08-20*, Centre for Research in Statistical Methodology, Uni. Warwick, Coventry, U.K.
- [19] PAPASPILIOPOULOS & ROBERTS (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95** 169-186.
- [20] PELLETIER, B. (2005). Kernel density estimation on Riemannian manifolds. *Stat. & Prob. Letters* **73** 297-304.
- [21] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10-26.

- [22] SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639-650.
- [23] STEPHENS, M. (2000). Dealing with label switching in mixture models. *J.R. Stat. Soc. Ser. B Stat. Methodol.* **62** 795-809.
- [24] WALKER, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, **36**, 45-54.
- [25] WATSON, G.S. (1965). Equitorial distributions on a sphere *Biometrika* **52** 193-201.
- [26] WATSON, G.S. (1983). *Statistics on spheres. University of Arkansas Lecture Notes in the Mathematical Sciences*, **6**. Wiley, NY.
- [27] WEST, M., MULLER, P. & ESCOBAR, M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation.
- [28] WILLMORE, T. (1993). *Riemannian Geometry*. Oxford Uni. Press, Oxford.
- [29] WU, Y. & GHOSAL, S. (2008). Kullback-Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, **2**, 298-331.
- [30] YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G.O. & HOLMES, C. (2008). Bayesian nonparametric hidden Markov models with application to the analysis of copy-number-variation in mammalian genomes. *Working Paper*, Oxford-Man Institute, Uni. of Oxford, Oxford, U.K.

DEPARTMENT OF STATISTICAL SCIENCE, DUKE UNIVERSITY, DURHAM, NC, USA