

Nonparametric Bayesian Density Estimation on Manifolds with Applications to Planar Shapes

BY ABHISHEK BHATTACHARYA, DAVID B. DUNSON

*Department of Statistical Science, Box 90251, Duke University, Durham, North Carolina
27708-0251, U.S.A.*

ab216@stat.duke.edu dunson@duke.edu

SUMMARY

Statistical analysis on landmark-based shape spaces has diverse applications in morphometrics, medical diagnostics, machine vision and other areas. These shape spaces are non-Euclidean quotient manifolds. To conduct nonparametric inferences, one may define notions of center and spread on this manifold and work with their estimates. However, it is useful to consider full likelihood-based methods, which allow nonparametric estimation of the probability density. This article proposes a broad class of mixture models constructed using suitable kernels on a general compact metric space 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 and classification with shape-based predictors. Simulation studies show improved estimation performance relative to existing approaches.

Some key words: Dirichlet process mixture; Discriminant analysis; Kullback–Leibler property; Metric space; Nonparametric Bayes; Planar shape space; Posterior consistency; Riemannian manifold.

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 compact metric spaces and 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 & Bhattacharya (2008) 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. Pelletier (2005) develops frequentist methods for density estimation on compact Riemannian manifolds using a kernel that generalizes location-scale kernels used in Euclidean spaces. The sample points are used as the kernel locations while assuming a fixed bandwidth, and the estimator is shown to be L^2 consistent for a sufficiently small bandwidth.

Bayesian nonparametric methods have the advantage of providing a full probabilistic characterization of uncertainty, which is valid even in small samples. Nonparametric Bayes density estimation in Euclidean spaces commonly relies on kernel mixture models, with a Dirichlet pro-

cess prior (Ferguson, 1973, 1974) placed on the unknown mixture distribution and a Gaussian kernel assumed (Lo, 1984; Escobar & West, 1995). Our focus is on developing Bayesian kernel mixture models for nonparametric density estimation on compact metric spaces, with manifolds arising as a special case. The manifold of special interest is the planar shape space Σ_2^k corresponding to similarity shapes of configurations of k landmarks in two-dimensions (Kendall, 1984).

The kernel should 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 n grows. From Schwartz (1965), prior positivity of Kullback–Leibler 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 Kullback–Leibler support is important in providing a proof of concept that the prior is sufficiently flexible, but is difficult for new priors even in Euclidean spaces. We extend the sufficient conditions of Wu & Ghosal (2008) to arbitrary compact metric spaces, and apply this theory to general manifolds and planar shape spaces.

For landmark-based shape data, current Bayesian analyses focus on parametric models. For example, Kume & Walker (2006) recently proposed a method for posterior computation in complex Watson models (Watson, 1965, 1983), with Dryden & Mardia (1998) proposing the complex Watson as a convenient parametric distribution for planar shape data. Lennox et al. (2009) proposed a Dirichlet process mixture of bivariate von Mises–Fisher distributions for protein configuration angles, modifying the finite mixture model of Mardia et al. (2007). Their model arises as a special case of the framework we propose, and is not applicable to shape data. The von Mises–Fisher kernel is quite restrictive, and it is not clear whether mixtures of such kernels induce priors with large support. Lennox et al. (2009) does not present any theoretical results. However, our theory can be used to show that such a prior has full support and weak posterior consistency follows, with Bhattacharya & Dunson, unpublished, providing conditions for strong consistency. Computation in Lennox et al. (2009) relies on the auxiliary Gibbs sampler of Neal (2000). In this paper, for applications to landmark-based shape data, we focus on Dirichlet process mixtures of complex Watson distributions. We show that such priors have large support, while also developing efficient methods of posterior computation.

2. NONPARAMETRIC DENSITY ESTIMATION ON COMPACT METRIC SPACES

Let M be a compact metric space and let X be a random variable on M . We assume that the distribution of X has a density with respect to some fixed base measure λ on M and we are interested in modelling this density via a flexible model. Let $K(m; \mu, \sigma)$ be a probability kernel on M with location $\mu \in M$ and scale $\sigma \in \mathfrak{R}^+$, with $\int_M K(m; \mu, \sigma) \lambda(dm) = 1$. We can define a location mixture probability density model for X as

$$f(m; P, \sigma) = \int_M K(m; \mu, \sigma) P(d\mu) \quad (1)$$

or a location-scale mixture model

$$g(m; Q) = \int_{M \times \mathfrak{R}^+} K(m; \mu, \sigma) Q(d\mu d\sigma). \quad (2)$$

For a pre-specified kernel K , a prior on $\mathcal{D}(M)$, the space of all probability densities on M with respect to the set base measure λ , is induced through a prior $(P, \sigma) \sim \Pi_1$ in (1) and a prior $Q \sim \Pi_2$ in (2). In order to evaluate whether a particular kernel K and prior Π_1 or Π_2 induces a

97 prior for the unknown density on M that is sufficiently flexible, it would be appealing to have
 98 simple sufficient conditions to check.

99 We make the following assumptions about the kernel K :

100 ASSUMPTION 1. K is continuous on $M \times M \times (0, \sigma_0)$ for some $\sigma_0 > 0$.

101 ASSUMPTION 2. For any $\phi \in C(M)$, with $C(M)$ the space of continuous functions on M ,

102
 103
 104
$$\lim_{\sigma \rightarrow 0} \sup_{m \in M} \left| \phi(m) - \int_M K(m; \mu, \sigma) \phi(\mu) \lambda(d\mu) \right| = 0.$$

105
 106 These assumptions place minor regularity conditions on the kernel. If K is symmetric in m and
 107 μ , Assumption 2 implies that K converges weakly to the degenerate point mass at μ uniformly
 108 in μ as $\sigma \rightarrow 0$.

109 In addition, we make the following assumptions about f_0 , the true density of X , and the
 110 support of the prior Π_1 .

111
 112 ASSUMPTION 3. For any $\sigma > 0$, there exists $\tilde{\sigma} \leq \sigma$ such that $(F_0, \tilde{\sigma}) \in \text{supp}(\Pi_1)$, with F_0
 113 the probability distribution corresponding to f_0 and $\text{supp}(\Pi_1)$ denoting the weak support of Π_1 .

114 ASSUMPTION 4. The true density is continuous so that $f_0 \in C(M)$.

115
 116 THEOREM 1. Define f as in (1). Under Assumptions 1-4, given any $\epsilon > 0$,

117
 118
$$\Pi_1 \left\{ (P, \sigma) : \sup_{m \in M} |f_0(m) - f(m; P, \sigma)| < \epsilon \right\} > 0.$$

119
 120 Theorem 1 shows that the density prior induced through the location mixture model (1) assigns
 121 positive probability to arbitrarily small L^∞ neighborhoods of the true density under mild assump-
 122 tions. For a proposed prior chosen for a particular M , one can simply verify that the assumed
 123 kernel K and prior Π_1 satisfy the assumptions to show large support. We will illustrate how these
 124 assumptions are met using a complex Watson kernel on a planar shape space in §3.

125 To show full Kullback–Leibler support for the prior, we require an additional assumption:

126 ASSUMPTION 5. The true density is everywhere positive so that $f_0(m) > 0$ for all $m \in M$.

127
 128 COROLLARY 1. Under Assumptions 1–5, the prior on $\mathcal{D}(M)$ induced by Π_1 through (1) as-
 129 signs positive probability to any Kullback–Leibler neighborhood around f_0 .

130 ASSUMPTION 6. For any $\sigma > 0$, there exists $\tilde{\sigma} \in (0, \sigma]$ such that $F_0 \otimes \delta_{\tilde{\sigma}} \in \text{supp}(\Pi_2)$.

131
 132 THEOREM 2. Let g be a density as in (2). Under Assumptions 1–2 and 4–6, the prior on $\mathcal{D}(M)$
 133 induced by Π_2 assigns positive probability to any Kullback–Leibler neighborhood around f_0 .

134 The assumptions on the priors in Theorems 1 and 2 are trivially satisfied by standard non-
 135 parametric priors. For example, for model (1) we can choose Π_1 to be $\Pi_{11} \otimes \pi_1$, with Π_{11} a
 136 Dirichlet process prior $\text{DP}(\omega_0 P_0)$ with $\text{supp}(P_0) = M$ and π_1 having a density that is strictly
 137 positive in some neighborhood of zero. For model (2), we can instead choose the prior Π_2 for the
 138 mixing measure Q to correspond to a Dirichlet process with base $P_0 \otimes \pi_1$. Under these priors,
 139 models (1) and (2) are Dirichlet process mixture models and standard algorithms can be applied
 140 for posterior computation.

141 A special case of a compact metric space is a compact Riemannian manifold with the distance
 142 metric being the geodesic distance induced by the Riemannian metric tensor. The natural choice
 143 of base measure for modelling densities is then the Riemannian volume form. For background
 144 in differential geometry, the reader is referred to Willmore (1993). Pelletier (2005) introduced

145 a geodesic distance based kernel and performed frequentist density estimation on compact Rie-
 146 mannian manifolds. Under mild restrictions on the form of this kernel, it can be shown that it
 147 satisfies the assumptions of Theorem 1. The details and proofs are omitted since we have found
 148 alternative kernels to also satisfy these assumptions for manifolds corresponding to the unit hy-
 149 persphere and the planar shape space, while having computational advantages over the Pelletier
 150 (2005) class of kernels. For the unit hypersphere, von Mises–Fisher kernels can be used (Bhat-
 151 tacharya & Dunson, unpublished), but here we focus on the landmark-based planar shape space
 152 $M = \Sigma_2^k$.

155 3. THE PLANAR SHAPE SPACE Σ_2^k

156 3.1. Geometry

157 Consider a set of k points, $k > 2$, on the two-dimensional plane, not all points being the same.
 158 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
 159 remains after we remove the effects of the Euclidean rigid body motions of translation, rotation
 160 and scaling. For convenience we denote a k -ad by a complex k -vector $z = (z_1, \dots, z_k)^T$ in
 161 \mathbb{C}^k . To remove the effect of translation from z let $z_c = z - \bar{z}$, with $\bar{z} = (\sum_{j=1}^k z_j)/k$ being the
 162 centroid. The centered k -ad z_c lies in a $k - 1$ dimensional complex subspace, and hence we can
 163 use $k - 1$ complex coordinates. The effect of scaling is removed by normalizing the coordinates
 164 of z_c to obtain a point w on the complex unit sphere CS^{k-2} in \mathbb{C}^{k-1} . Since w contains the shape
 165 information of z along with rotation, it is called the preshape of z .

166 The similarity shape of z is the orbit of w under all two-dimensional rotations. Since a rotation
 167 by an angle θ of a landmark (x, y) can be achieved by multiplying its complex version $x + iy$
 168 by $\exp(i\theta)$, the shape of z is the set or orbit $[w] = \{\exp(i\theta)w : \theta \in (-\pi, \pi]\}$. The space of all
 169 such orbits constitutes the planar shape space Σ_2^k . Any shape can be represented as the set of
 170 intersection points of a unique complex line passing through the origin with CS^{k-2} . With this
 171 identification proposed by Kendall (1984), Σ_2^k is a compact Riemannian manifold of dimension
 172 $2k - 4$. It can be embedded into the space of all complex Hermitian matrices via the embedding
 173 $J([w]) = ww^*$, with $*$ denoting the complex conjugate transpose. The extrinsic distance between
 174 two shapes $[u]$ and $[v]$ is the one induced from this embedding, namely, $d_E([u], [v]) = \|J([u]) -$
 175 $J([v])\| = \{2(1 - |u^*v|^2)\}^{1/2}$. This distance is equivalent to the geodesic distance $d_g([u], [v]) =$
 176 $\arccos(|u^*v|)$.

177 Let Q be a probability distribution on Σ_2^k . The extrinsic mean of Q is defined as the minimizer
 178 of the loss function $F(p) = \int_{\Sigma_2^k} d_E^2(m, p)Q(dm)$, $p \in \Sigma_2^k$, provided F has a unique minimizer.
 179 The minimum value of F is called the extrinsic variation of Q . Let $\tilde{\mu} = \int_{\Sigma_2^k} J(m)Q(dm)$, λ be
 180 its largest eigenvalue and U be a corresponding unit norm eigenvector. Then it can be shown
 181 that the extrinsic variation equals $2(1 - \lambda)$ and the extrinsic mean is given by $[U]$ provided λ
 182 has multiplicity 1. Given a random sample from Q , one can define the sample extrinsic mean
 183 and variation analogously. For more details see Bhattacharya & Patrangenaru (2003) and Bhat-
 184 tacharya & Bhattacharya (2008).

186 3.2. Uniform distribution

187 Let $V(dm)$ and $V_1(dz)$ denote the volume forms on the shape space Σ_2^k and the preshape
 188 sphere CS^{k-2} respectively. The uniform distribution on Σ_2^k has constant density $1/\int_{\Sigma_2^k} V(dm)$.
 189 Kent (1994) constructs a useful coordinate chart on Σ_2^k as follows. For $z = (z_1, \dots, z_{k-1})^T \in$
 190 CS^{k-2} , write $z_j = r_j^{1/2} \exp(i\theta_j)$ ($j = 1, \dots, k - 1$) with $r = (r_1, \dots, r_{k-2})^T$ lying on the $(k -$
 191 $1)$ -unit simplex S_{k-2} , and $\theta_j \in (-\pi, \pi)$ ($j = 1, \dots, k - 1$). Then $(r_1, \dots, r_{k-2}, \theta_1, \dots, \theta_{k-1})$

193 form the Kent preshape coordinates of z . Since the shape of z can be obtained by ro-
 194 tating it around a fixed axis, we may set $\theta_{k-1} = 0$ and use the Kent shape coordinates
 195 $(r_1, \dots, r_{k-2}, \theta_1, \dots, \theta_{k-2})^T$ for $[z]$ as in Dryden & Mardia (1998). These Kent (1994) coor-
 196 dinate systems have the advantage of leading to simple expressions for the volume forms,

$$197 \quad V_1(dz) = 2^{2-k} dr_1 \cdots dr_{k-2} d\theta_1 \cdots d\theta_{k-1}, \quad V(d[z]) = 2^{2-k} dr_1 \cdots dr_{k-2} d\theta_1 \cdots d\theta_{k-2}.$$

199 This implies that, in terms of these shape coordinates, the uniform distribution on Σ_2^k remains
 200 uniform on $S_{k-2} \times (-\pi, \pi)^{k-2}$. This property simplifies simulations and proofs.

201 202 3.3. Complex Bingham distribution

203 The complex Bingham distribution on Σ_2^k (Kent, 1994) has the following density with respect
 204 to the volume form

$$205 \quad f(m; A) = c^{-1}(A) \exp(z^* A z), \quad m = [z] \in \Sigma_2^k, \quad (3)$$

206 where A is a $(k-1) \times (k-1)$ complex Hermitian matrix and $c(A)$ is the normalizing con-
 207 stant. Denoting this density as $\text{CB}(A)$, we find that $\text{CB}(A) = \text{CB}(A + \alpha I)$ for any $\alpha \in \mathbb{R}$.
 208 Hence, without loss of generality, we may assume A to be positive semi-definite with small-
 209 est eigenvalue equal to zero. Let $A = U \Lambda U^*$ be a singular value decomposition of A with
 210 $U = [U_1, \dots, U_{k-1}] \in \text{SU}(k-1)$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_{k-1})$, and $0 = \lambda_1 \leq \dots \leq \lambda_{k-1}$, where
 211 $\text{SU}(k-1)$ is the space of all $(k-1) \times (k-1)$ special unitary matrices having $UU^* = I$ and
 212 $\det(U) = 1$. Letting $z_1 = U^* z$ and using Kent's shape coordinates (r, θ) for $[z_1]$, the complex
 213 Bingham distribution can be written as

$$214 \quad f(m; A) V(dm) = c^{-1}(A) 2^{2-k} \exp\left(\sum_{j=1}^{k-1} \lambda_j r_j\right) dr_1 \cdots dr_{k-2} d\theta_1 \cdots d\theta_{k-2} \quad (4)$$

215 with $r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$. Expression (4) suggests that r has a density $g(r) \propto \exp(\sum_{j=1}^{k-1} \lambda_j r_j)$
 216 on S_{k-2} while $\theta_1, \dots, \theta_{k-2}$ are independent and identically distributed as $\text{Unif}(-\pi, \pi)$ and r
 217 and θ are independent. This characterization is helpful in sampling from the complex Bingham.
 218 Under high concentrations, that is when $\lambda_{k-1} \gg \lambda_{k-2}$, one may use an independent Metropolis-
 219 Hasting step with an independent exponential approximation to sample from g . That is, we draw
 220 r_j , $j = 1, \dots, k-2$ independently from the density proportional to $\exp\{(\lambda_j - \lambda_{k-1})r_j\}$ on
 221 $(0, 1)$, accept the draw if $\sum_{i=1}^{k-2} r_i \leq 1$ and then set $r_{k-1} = 1 - \sum_{j=1}^{k-2} r_j$.

222 223 224 225 226 227 3.4. Complex Watson distribution

228 When A has complex rank equal to one, the complex Bingham distribution $\text{CB}(A)$ corresponds
 229 to a complex Watson distribution (Dryden & Mardia, 1998) having density

$$230 \quad f(m; \mu, \sigma) = c^{-1}(\sigma) \exp(|z^* \nu|^2 / \sigma), \quad (5)$$

231 with z and ν preshapes of m and $\mu \in \Sigma_2^k$, respectively. Here, μ is the extrinsic mean, $\sigma > 0$ is a
 232 scale parameter related to the extrinsic variation, and $c(\sigma)$ is the normalizing constant. Denoting
 233 this density as $\text{CW}(\mu, \sigma)$, $\text{CW}(\mu, \sigma)$ is equivalent to $\text{CB}(A)$ with $A = \nu \nu^* / \sigma$. As A has eigen-
 234 values $\lambda_1 = \dots = \lambda_{k-2} = 0$, $\lambda_{k-1} = \sigma^{-1}$, the distribution of r defined in §3.3 can be written as
 235 $g(r) \propto \exp(\sigma^{-1} r_{k-1})$ implying that r_{k-1} has the marginal density

$$236 \quad h(r_{k-1}) = c_{k-1}^{-1}(\sigma) \exp(r_{k-1} \sigma^{-1}) (1 - r_{k-1})^{k-3}, \quad r_{k-1} \in (0, 1),$$

237 where $c_{k-1}(\sigma) = \sigma^{k-2} \exp(\sigma^{-1}) \Gamma(k-2; \sigma^{-1})$ with $\Gamma(m, a) = \int_0^a \exp(-t) t^{m-1} dt = (m-1)!$
 238 $\exp(-a) \{ \exp(a) - \sum_{r=0}^{m-1} a^r / r! \}$ denoting the partial gamma function. Conditionally on

241 $r_{k-1}, (r_1, \dots, r_{k-2})$ has a uniform distribution on the set $\{r_j \geq 0, j = 1, \dots, k-2, \sum_{j=1}^{k-2} r_j =$
 242 $1 - r_{k-1}\}$. Transforming by letting $s = \sigma^{-1}(1 - r_{k-1})$, s has a density proportional to
 243 $\exp(-s)s^{k-3}$ on $(0, \sigma^{-1})$, which is $\text{Ga}(k-2, 1)$ restricted to $(0, \sigma^{-1})$. This characterization can
 244 be used to easily draw exact samples from a complex Watson distribution. The normalizing con-
 245 stant is $c(\sigma) = (\pi\sigma)^{(k-2)}\{\exp(\sigma^{-1}) - \sum_{r=0}^{k-3} \sigma^{-r}/r!\}$. In Dryden & Mardia (1998), $\text{CW}(\mu, \sigma)$
 246 is viewed as a distribution on the preshape sphere and the normalizing constant is instead $2\pi c(\sigma)$.
 247

248 4. DENSITY ESTIMATION ON THE PLANAR SHAPE SPACE

249 To model an unknown density on Σ_2^k , we use a mixture density as in (1) with K corresponding
 250 to the complex Watson density in expression (5).
 251

252 PROPOSITION 1. *For the complex Watson kernel, Assumptions 1 and 2 of §2 are satisfied.*
 253

254 Hence, if we choose a complex Watson kernel in (1) and choose Π_1 to satisfy Assumption 3 from
 255 Theorem 1, we induce a prior with L^∞ support on the space of continuous densities over Σ_2^k and
 256 with Kullback–Leibler support on the space of continuous and everywhere positive densities over
 257 Σ_2^k . It follows from Schwartz (1965) that this specification leads to weak posterior consistency
 258 at any continuous, everywhere positive f_0 .
 259

260 To specify a Π_1 that satisfies the assumptions and that leads to simplifications in implementing
 261 posterior computation, we follow the recommendation at the end of §2 and let $P \sim \text{DP}(\omega_0 P_0)$,
 262 with P_0 corresponding to $\text{CW}(\mu_0, \sigma_0)$, independently of $\sigma^{-1} \sim \text{Ga}(a, b)$. These priors lead to
 263 conditional conjugacy so that posterior computation can proceed via Gibbs sampling algorithms
 264 previously developed for Dirichlet process mixture models. For the location-scale mixture (2),
 265 the computations are similar and are left to the reader.
 266

267 Here, we follow the exact block Gibbs sampler proposed by Yau et al. (2010). Let $x_i \sim$
 268 $\text{CW}(\mu_i, \sigma)$, independently for $i = 1, \dots, n$, with $\mu_i \sim P$, and P, σ assigned the prior described
 269 above. We introduce uniformly distributed slice sampling latent variables, $u = \{u_i\}_{i=1}^n$ and let
 270 S_i denote the mixture component for subject i , with $\mu_i = \tilde{\mu}_{S_i}$. The complete data likelihood is
 271 then $\prod_{i=1}^n \text{CW}(X_i; \tilde{\mu}_{S_i}, \sigma) 1(u_i < w_{S_i})$, and we sequentially sample through the following steps.

272 *Step 1.* Update S_i , for $i = 1, \dots, n$, by sampling from the multinomial conditional poste-
 273 rior distribution with $\text{pr}(S_i = j) \propto \text{CW}(x_i; \tilde{\mu}_j, \sigma)$ for $j \in A_i$, where $A_i = \{j : 1 \leq j \leq l, w_j >$
 274 $u_i\}$ and l is the smallest index satisfying $1 - u_{(1)} < \sum_{j=1}^l w_j$ with $u_{(1)} = \min\{u_1, \dots, u_n\}$.
 275 In implementing this step, draw $V_j \sim \text{Be}(1, \omega_0)$ and $\tilde{\mu}_j \sim P_0$ for $j > S_{(n)}$, with $S_{(n)} =$
 276 $\max\{S_1, \dots, S_n\}$.
 277

278 *Step 2.* Update the kernel locations $\tilde{\mu}_j (j = 1, \dots, S_{(n)})$ by sampling from the conditional
 279 posterior
 280

$$281 \tilde{\mu}_j \sim \text{CB}\left(\frac{m_j}{\sigma} \bar{X}_j + A_0\right),$$

282 where $m_j = \sum_{i=1}^n 1(S_i = j)$, $\bar{X}_j = \sum_{i: S_i=j} z_i z_i^* / m_j$ ($x_i = [z_i]$), $A_0 = \sigma_0^{-1} \nu_0 \nu_0^*$, and $\mu_0 =$
 283 $[\nu_0]$. We use a Metropolis-Hastings step developed in §3.3 to draw $\tilde{\mu}_j$.
 284

285 *Step 3.* The full conditional posterior of σ is proportional to
 286

$$287 (\sigma^{-1})^{n(k-2)+a+1} \exp\left\{-\frac{1}{\sigma}\left(n+b-\sum_{j=1}^{S_{(n)}} m_j \nu_j^* \bar{X}_j \nu_j\right)\right\} \left\{1 - \exp(-\sigma^{-1}) \sum_{r=0}^{k-3} (r!)^{-1} \sigma^{-r}\right\}^{-n},$$

288

where $\tilde{\mu}_j = [\nu_j]$. For σ small, this conditional density is approximately equivalent to

$$\sigma^{-1} \sim \text{Ga}\left\{a + n(k - 2), b + \sum_{j=1}^{S(n)} m_j(1 - \nu_j^* \bar{X}_j \nu_j)\right\}.$$

Hence, we get approximate conjugacy for the conditional distribution of σ^{-1} under a gamma prior. Numerical studies show that this approximation is very accurate even for σ moderately small, so we recommend a Metropolis-Hastings independence step with candidates generated from the approximation.

Step 4. Update the stick-breaking random variables V_j ($j = 1, \dots, S(n)$), from their conditional posterior distributions given the cluster allocation but marginalizing out the slice sampling variables,

$$V_j \sim \text{Be}\left\{1 + m_j, \omega_0 + \sum_{i=1}^n 1(S_i > j)\right\}.$$

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

We also incorporate label-switching moves as recommended in Papaspiliopoulos & Roberts (2008). In cases we have considered, the algorithm is efficient, with rapid convergence and no evidence of slow mixing. Due to label switching issues (Stephens, 2000), 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 can be obtained using

$$f(m; P, \sigma) = \sum_{j=1}^{S(n)} w_j \text{CW}(m; \tilde{\mu}_j, \sigma) + \left(1 - \sum_{j=1}^{S(n)} w_j\right) \int \text{CW}(m; \tilde{\mu}, \sigma) \text{CW}(\tilde{\mu}; \mu_0, \sigma_0) V(d\tilde{\mu}), \tag{6}$$

with σ and $w_j, \tilde{\mu}_j$ ($j = 1, \dots, S(n)$) a Markov chain Monte Carlo 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. Since it is difficult to evaluate the integral in (6) in closed form, we replace the integral by $\text{CW}(m; \mu_1, \sigma)$, μ_1 being a draw from $\text{CW}(\mu_0, \sigma_0)$.

5. APPLICATIONS

5.1. Applications to simulated data

We draw $x_i \sim 0.5\text{CW}(\mu_1, \sigma) + 0.5\text{CW}(\mu_2, \sigma)$ independently for $i = 1, \dots, 200$, with $k = 4$, $\sigma = 0.001$, $\mu_1 = (1, 0, 0)^T$, $\mu_2 = \{r, (1 - r^2)^{1/2}, 0\}^T$ and $r = 0.9975$ so that the extrinsic distance between μ_1 and μ_2 is 0.1. We compare our Bayesian nonparametric density estimate based on Dirichlet process mixtures of complex Watson kernels to a maximum likelihood estimate under a parametric complex Watson model (Dryden & Mardia, 1998) and to the frequentist kernel density estimate using a complex Watson kernel. We generated 20 simulated data sets, with the performance evaluated based on the L^1 distance and Kullback–Leibler divergence estimated by averaging over the data points. Our Bayesian nonparametric approach was implemented as described in §4 with the Markov chain Monte Carlo algorithm run for 100,000 iterations with the first 15,000 discarded as a burn-in. The hyperparameters were chosen by setting μ_0 equal to the sample extrinsic mean and $\sigma_0 = 0.1$ in the complex Watson base measure P_0 for P , and

Table 1. *Summaries of estimated distances from the true density across the 20 simulations.*

	Bayes		MLE		KDE	
	L^1	KL	L^1	KL	L^1	KL
min	0.27	0.02	0.60	0.33	0.56	0.14
25th	0.35	0.07	0.68	0.36	0.74	0.24
50th	0.42	0.08	0.73	0.43	0.87	0.26
75th	0.48	0.16	0.83	0.46	1.20	0.27
max	0.91	0.39	0.94	0.52	2.72	0.32
mean	0.44	0.13	0.75	0.41	1.03	0.25

Bayes = our proposed approach, MLE = maximum likelihood estimate, KDE = kernel density estimate

$a = b = 0.1$ in the gamma prior for σ^{-1} . By using the data to estimate the location of the base distribution, while choosing a moderate scale, we ensure that the prior introduces clusters close to the support of the data. The Dirichlet process precision parameter is fixed as $\omega_0 = 1$, which is a commonly-used default in the literature favoring a sparse representation with few clusters.

Table 1 presents summaries of the results across the 20 simulated data sets. The proposed nonparametric Bayes estimator had consistently better performance across the data sets and for each choice of criterion. For the frequentist kernel density estimate, results are presented for a bandwidth of $\sigma = 0.001$. The performance was similar or worse for other choices of bandwidth, including setting σ equal to the maximum likelihood estimate under the parametric complex Watson model and the posterior mean of σ from the Bayes analysis.

5.2. Application to morphometrics: classification of gorilla skulls

The method is applied to data on the shape of 29 male and 30 female gorilla skulls, with eight landmarks chosen on the midline plane of two-dimensional images of each skull (Dryden & Mardia, 1998). The goal is to study how the shapes of the skulls vary between males and females, and build a classifier to predict gender. The shape samples lie on Σ_2^k , $k = 8$. We randomly pick 25 individuals of each gender as a training sample, with the remaining 9 used as test data. Figure 1 shows the preshapes of the sample extrinsic means for the female and male training groups. The preshape of the male mean $\hat{\mu}_2$ has been rotated appropriately so as to bring it closest to the preshape of the female mean $\hat{\mu}_1$. Most of the landmarks corresponding to the preshapes of the sample means after rotation are close for females and males, but there is a larger difference in landmarks three and eight.

Applying nonparametric discriminant analysis, we assume that the probability of being female is 0.5 and use a separate Dirichlet process mixture of complex Watson kernels for the shape density in the male and female groups. Letting $f_1(m)$ and $f_2(m)$ denote the female and male shape densities, the conditional probability of being female given shape data $[z]$ is simply $p([z]) = 1/\{1 + f_2([z])/f_1([z])\}$. To estimate the posterior probability, we average $p([z])$ across Markov chain Monte Carlo iterations to obtain $\hat{p}([z])$. The analysis was implemented as in the simulation examples, but with hyperparameters $\sigma_0 = 0.001$, $a = 1.01$ and $b = 0.001$ elicited based on our prior expectation for the gorilla example.

Table 2 presents the estimated posterior probabilities of being female for each of the gorillas in the test sample along with a 95% credible interval for $p([z])$. For most of the gorillas, there is a high posterior probability of assigning the correct gender. There is misclassification only in the

385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432

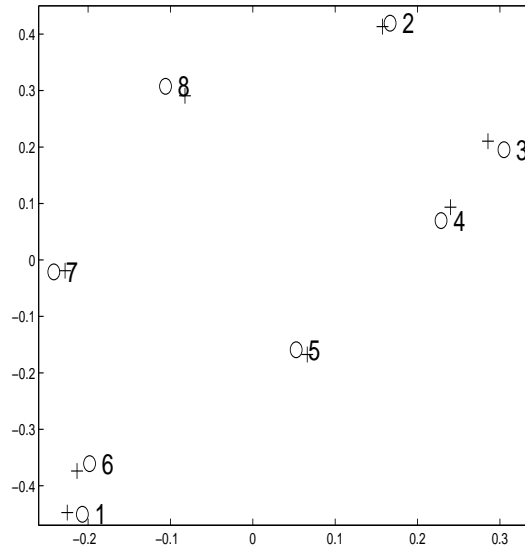


Fig. 1. Landmarks from preshapes of $\hat{\mu}_1$ (female, circle) & $\hat{\mu}_2$ (males, plus).

3rd female and 3rd male. For the 3rd female, the credible interval includes 0.5, suggesting that there is insufficient information to be confident in the classification. However, for the 3rd male, the credible interval suggests a high degree of confidence that this individual is female. Perhaps this individual is an outlier and there is something unusual about the shape of his skull, with such characteristics not represented in the training data, or alternatively he was labelled incorrectly.

In addition, we display the extrinsic distance between the shape for each gorilla and the female and male sample extrinsic means. Potentially we could define a distance-based classifier, which allocates a test subject to the group having mean shape closest to that subjects' shape. The table suggests that such a classifier will yield consistent results with our nonparametric Bayes approach. However, this distance-based classifier may be sub-optimal in not taking into account the variability within each group. In addition, the approach is deterministic and there is no measure of uncertainty in classification. Figure 2 shows the male and female training sample preshape clouds, along with the two misclassified test samples. There seems to be a substantial deviation in the coordinates of these misclassified subjects from their respective gender training groups, especially for the male gorilla, even after having rotated each training preshape separately so as to bring each closest to the plotted test sample preshapes.

It is possible that classification performance could be improved in this application by also taking into account skull size. The proposed method can be easily extended to this case by using a Dirichlet process mixture density with the kernel being the product of a complex Watson kernel for the shape component and a log-Gaussian kernel for the size. Such a model induces a prior with support on the space of densities on the manifold $\Sigma_2^k \times \mathbb{R}^+$.

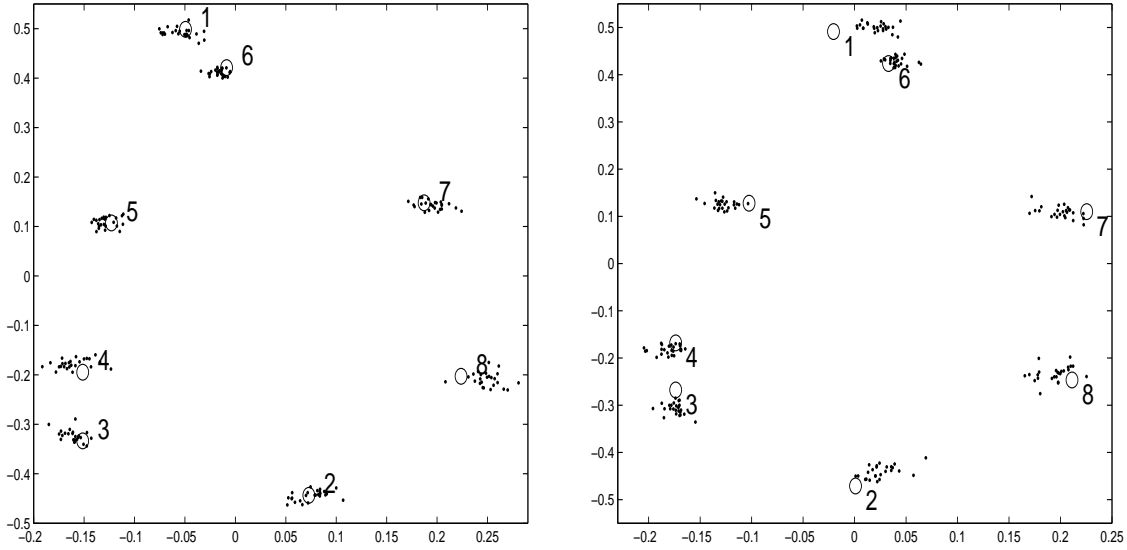


Fig. 2. Landmarks from preshapes of training (dot) & mis-classified test samples (circle) for females (left) & males (right).

Table 2. Posterior probability of being female for each gorilla in the test sample.

gender	$\hat{p}([z])$	95% CI	$d_E([z_i], \hat{\mu}_1)$	$d_E([z_i], \hat{\mu}_2)$
F	1.000	(1.000, 1.000)	0.041	0.111
F	1.000	(0.999, 1.000)	0.036	0.093
F	0.023	(0.021, 0.678)	0.056	0.052
F	0.998	(0.987, 1.000)	0.050	0.095
F	1.000	(1.000, 1.000)	0.076	0.135
M	0.000	(0.000, 0.000)	0.167	0.103
M	0.001	(0.000, 0.004)	0.087	0.042
M	0.992	(0.934, 1.000)	0.091	0.121
M	0.000	(0.000, 0.000)	0.152	0.094

$d_E([z_i], \hat{\mu}_i)$ = extrinsic distance from the mean shape in group i , with $i = 1$ for females and $i = 2$ for males

APPENDIX 1

Lemma 1

To prove Theorems 1 and 2, we will need the following lemma. Let $\mathcal{M}(M)$ denote the space of all probability distributions on M .

LEMMA 1. Given $\epsilon > 0$, if there exists (i) a $\sigma_\epsilon > 0$ and $P_\epsilon \in \mathcal{M}(M)$ such that

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

(ii) a set $W \subseteq \mathfrak{R}^+$ containing σ_ϵ such that

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

481 and (iii) a set $\mathcal{W} \subseteq \mathcal{M}(M)$ containing P_ϵ such that

$$482 \quad \sup_{m \in M, P \in \mathcal{W}, \sigma \in W} |f(m; P_\epsilon, \sigma) - f(m; P, \sigma)| < \frac{\epsilon}{3},$$

485 then

$$486 \quad \sup_{m \in M} |f_0(m) - f(m; P, \sigma)| < \epsilon$$

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

489 *Proof.* Follows from a direct application of the triangular inequality. \square

491 *Proof of Theorem 1*

492 *Proof.* The result follows from Lemma 1 if we can find a $W \subseteq \mathfrak{R}^+$ and $\mathcal{W} \subseteq \mathcal{M}(M)$ satisfying the
493 conditions and $\Pi_1(W \times \mathcal{W}) > 0$.

494 Condition (i) is satisfied with $P_\epsilon = F_0$ from Assumptions 2 and 4 by taking $\phi = f_0$. From Assumption
495 3, it follows that by taking σ_ϵ sufficiently small, we can ensure that $\sigma_\epsilon < \sigma_0$ and $(F_0, \sigma_\epsilon) \in \text{supp}(\Pi_1)$.

496 Next we need to find a W for which condition (ii) is satisfied. From Assumption 1, it follows that the
497 mapping from σ to K is uniformly equicontinuous on some compact neighborhood of σ_ϵ . Hence we can
498 get a compact set W containing σ_ϵ in its interior such that

$$499 \quad \sup_{(m, \mu, \sigma) \in M \times M \times W} |K(m; \mu, \sigma) - K(m; \mu, \sigma_\epsilon)| < \frac{\epsilon}{3}.$$

502 Then

$$\begin{aligned} 503 & \sup_{m \in M, \sigma \in W} |f(m; F_0, \sigma) - f(m; F_0, \sigma_\epsilon)| \\ 504 & \leq \int_M \sup_{m \in M, \sigma \in W} |K(m; \mu, \sigma) - K(m; \mu, \sigma_\epsilon)| f_0(\mu) \lambda(d\mu) \\ 505 & \leq \sup_{m, \mu \in M, \sigma \in W} |K(m; \mu, \sigma) - K(m; \mu, \sigma_\epsilon)| < \frac{\epsilon}{3}. \end{aligned}$$

507 This verifies condition (ii).

508 Lastly we need to find a \mathcal{W} for which condition (iii) is satisfied and $\Pi_1(W \times \mathcal{W}) > 0$. We claim that

$$509 \quad \mathcal{W} = \left\{ P \in \mathcal{M}(M) : \sup_{m \in M, \sigma \in W} |f(m; P, \sigma) - f(m; F_0, \sigma)| < \frac{\epsilon}{3} \right\}$$

512 contains a weakly open neighborhood of F_0 . For any $m \in M, \sigma \in W$, the mapping from μ to $K(m; \mu, \sigma)$
513 defines a continuous function on M . Hence

$$514 \quad \mathcal{W}_{m, \sigma} = \left\{ P : |f(m; P, \sigma) - f(m; F_0, \sigma)| < \frac{\epsilon}{9} \right\}$$

516 defines a weakly open neighborhood of F_0 for any (m, σ) in $M \times W$. The mapping from (m, σ) to
517 $f(m; P, \sigma)$ is a uniformly equicontinuous family of functions on $M \times W$, labeled by $P \in \mathcal{M}(M)$, be-
518 cause, for $m_1, m_2 \in M; \sigma, \tau \in W$,

$$519 \quad |f(m_1; P, \sigma) - f(m_2; P, \tau)| \leq \int_M |K(m_1; \mu, \sigma) - K(m_2; \mu, \tau)| P(d\mu)$$

522 and K is uniformly continuous on $M \times M \times W$. Therefore there exists a $\delta > 0$ such that $\rho(m_1, m_2) +$
523 $|\sigma - \tau| < \delta$ implies that

$$524 \quad \sup_{P \in \mathcal{M}(M)} |f(m_1; P, \sigma) - f(m_2; P, \tau)| < \frac{\epsilon}{9}.$$

527 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 $\mathcal{W}_1 =$
528 $\bigcap_{i=1}^N \mathcal{W}_{m_i, \sigma_i}$ 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 $|f(m; P, \sigma) - f(m; F_0, \sigma)|$

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

This proves that \mathcal{W} contains \mathcal{W}_1 and hence the claim is proved. Clearly this \mathcal{W} satisfies condition (iii). Since (F_0, σ_ϵ) is in $\text{supp}(\Pi_1)$ and in the interior of $\mathcal{W} \times W$, therefore $\Pi_1(\mathcal{W} \times W) > 0$. This completes the proof. \square

Proof of Corollary 1

Proof. Since M is compact, Assumption 5 implies that $c_0 = \inf_{m \in M} f_0(m) > 0$. For $\delta > 0$ define

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

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

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

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

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

if we choose $\delta < c_0\epsilon/2$. Hence for δ sufficiently small, $f(\cdot; P, \sigma) \in \text{KL}(f_0, \epsilon)$ whenever $(P, \sigma) \in \mathcal{W}_\delta$, with $\text{KL}(f_0, \epsilon)$ denoting an ϵ -sized Kullback–Leibler neighborhood around f_0 . From Theorem 1 it follows that $\Pi_1(\mathcal{W}_\delta) > 0$ for any $\delta > 0$ and therefore

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

Proof of Theorem 2

Proof. From the proof of Corollary 1, it follows that given any $\delta_1 > 0$, we can find a $\sigma_1 > 0$ such that with $Q_1 = F_0 \otimes \delta_{\sigma_1}$,

$$\sup_{m \in M} |f_0(m) - g(m; Q_1)| < \delta_1, \quad \int_M f_0(m) \log \left\{ \frac{f_0(m)}{g(m; Q_1)} \right\} \lambda(dm) < \delta_1. \quad (\text{A1})$$

Hence, if we choose $\delta_1 \leq c_0/2$ where $c_0 = \inf_{m \in M} f_0(m) > 0$ then $\inf_{m \in M} g(m; Q_1) \geq c_0/2$. From Assumption 6 it follows that we can choose σ_1 sufficiently small such that $\sigma_1 < \sigma_0$ and $Q_1 \in \text{supp}(\Pi_2)$. Let E denote a compact subset of $(0, \sigma_0)$ containing σ_1 in its interior. Then, being continuous in its arguments, K is uniformly continuous on $M \times M \times E$. For Q in $\mathcal{M}(M \times \mathbb{R}^+)$, define

$$g(m; Q_E) = \int_{M \times E} K(m; \mu, \sigma) Q(d\mu d\sigma).$$

For fixed $m \in M$, the integral mapping from Q to $g(m; Q_E)$ is continuous at Q_1 because

$$Q_1\{\partial(M \times E)\} = Q_1\{M \times \partial(E)\} = 0,$$

$\partial(A)$ denoting the boundary of a set A . Therefore for $\delta_2 > 0$ and $m \in M$,

$$\mathcal{W}_m(\delta_2) = \{Q : |g(m; Q_E) - g(m; Q_1)| < \delta_2\}$$

defines a weakly open neighborhood of Q_1 . We also claim that

$$\mathcal{W} = \left\{ Q : \sup_{m \in M} |g(m; Q_E) - g(m; Q_1)| < \delta_2 \right\},$$

contains an open neighborhood of Q_1 . To see this, choose a $\delta_3 > 0$ such that $\rho(m_1, m_2) < \delta_3$ 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

$$|g(m_1; Q_E) - g(m_2; Q_E)| < \frac{\delta_2}{3} \quad (\text{A2})$$

for all $Q \in \mathcal{M}(M \times \mathbb{R}^+)$. Next cover M by finitely many balls of radius δ_3 : $M = \bigcup_{i=1}^N B(m_i, \delta_3)$. Then we show that $\mathcal{W} \supseteq \bigcap_{i=1}^N \mathcal{W}_{m_i}(\delta_2/3)$. To prove that, pick Q in $\bigcap_{i=1}^N \mathcal{W}_{m_i}(\delta_2/3)$. Then for $i = 1, \dots, N$,

$$|g(m_i; Q_E) - g(m_i; Q_1)| < \delta_2. \quad (\text{A3})$$

Choosing $m \in B(m_i, \delta_3)$, (A2) implies that

$$|g(m; Q_E) - g(m_i; Q_E)| < \frac{\delta_2}{3} \quad (\text{A4})$$

for all $Q \in \mathcal{M}(M \times \mathbb{R}^+)$. From (A3) and (A4) it follows that $|g(m; Q_E) - g(m; Q_1)|$

$$\begin{aligned} &\leq |g(m; Q_E) - g(m_i; Q_E)| + |g(m_i; Q_E) - g(m_i; Q_1)| + |g(m_i; Q_1) - g(m; Q_1)| \\ &< \delta_2/3 + \delta_2/3 + \delta_2/3 = \delta_2 \end{aligned}$$

for any $m \in M$ and $Q \in \bigcap_{i=1}^N \mathcal{W}_{m_i}(\delta_2/3)$. Hence $\mathcal{W} \supseteq \bigcap_{i=1}^N \mathcal{W}_{m_i}(\delta_2/3)$ which is an open neighborhood of Q_1 . Therefore $\Pi_2(\mathcal{W}) > 0$. For $Q \in \mathcal{W}$,

$$\inf_{m \in M} g(m; Q_E) \geq \inf_{m \in M} g(m; Q_1) - \delta_2 \geq \frac{c_0}{4}$$

if $\delta_2 < \frac{c_0}{4}$. Then

$$\begin{aligned} \int_M f_0(m) \log \left\{ \frac{g(m; Q_1)}{g(m; Q)} \right\} \lambda(dm) &\leq \int_M f_0(m) \log \left\{ \frac{g(m; Q_1)}{g(m; Q_E)} \right\} \lambda(dm) \\ &\leq \sup_{m \in M} \left| \frac{g(m; Q_1)}{g(m; Q_E)} - 1 \right| \leq \frac{\delta_2}{c_0/4} < \delta_1, \end{aligned} \quad (\text{A5})$$

provided δ_2 is sufficiently small. From (A1) and (A5) we deduce that, for $Q \in \mathcal{W}$,

$$\begin{aligned} \int_M f_0(m) \log \left(\frac{f_0(m)}{g(m; Q)} \right) \lambda(dm) &= \\ \int_M f_0(m) \log \left(\frac{f_0(m)}{g(m; Q_1)} \right) \lambda(dm) &+ \int_M f_0(m) \log \left(\frac{g(m; Q_1)}{g(m; Q)} \right) \lambda(dm) < \delta_1 + \delta_1 = \epsilon \end{aligned}$$

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

$$\{g(\cdot; Q) : Q \in \mathcal{W}\} \subseteq \text{KL}(f_0, \epsilon)$$

and since $\Pi_2(\mathcal{W}) > 0$, therefore

$$\Pi_2\{Q : g(\cdot; Q) \in \text{KL}(f_0, \epsilon)\} > 0.$$

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

Proof of Proposition 1

Proof. Express K as

$$K(m; \mu, \sigma) = c^{-1}(\sigma) \exp \left\{ \frac{2 - d_E^2(m, \mu)}{2\sigma} \right\}$$

where $c(\sigma) = (\pi\sigma)^{(k-2)} \{ \exp(\sigma^{-1}) - \sum_{r=0}^{k-3} \sigma^{-r}/r! \}$ and Assumption 1 is satisfied.

As the kernel is symmetric in m and μ , for $\phi \in C(\Sigma_2^k)$,

$$I(m) \equiv \phi(m) - \int_{\Sigma_2^k} K(m; \mu, \sigma) \phi(\mu) V(d\mu) = \int_{\Sigma_2^k} \{\phi(m) - \phi(\mu)\} K(m; \mu, \sigma) V(d\mu). \quad (\text{A6})$$

Choose preshapes z and ν for m and μ , respectively, in the complex sphere CS^{k-2} , so that $m = [z]$ and $\mu = [\nu]$. Let V_1 denote the volume form on CS^{k-2} . Then for any integrable function $\phi : \Sigma_2^k \rightarrow \mathfrak{R}$,

$$\int_{\Sigma_2^k} \phi(\mu) V(d\mu) = \frac{1}{2\pi} \int_{\text{CS}^{k-2}} \phi([\nu]) V_1(d\nu).$$

Hence the integral in (A6) can be written as

$$I(m) = \frac{c^{-1}(\sigma)}{2\pi} \int_{\text{CS}^{k-2}} \{\phi([z]) - \phi([\nu])\} \exp(\sigma^{-1} \nu^* z z^* \nu) V_1(d\nu). \quad (\text{A7})$$

Consider a singular value decomposition of $z z^*$ as $z z^* = U \Lambda U^*$ where $\Lambda = \text{diag}(1, 0, \dots, 0)$ and $U = [U_1, \dots, U_{k-1}]$ with $U_1 = z$. Then $\nu^* z z^* \nu = x^* \Lambda x = |x_1|^2$ where $x = U^* \nu = (x_1, \dots, x_{k-1})^T$. Make a change of variables from ν to x in (A7). This is an orthogonal transformation, so does not change the volume form. Then (A7) becomes

$$I(m) = \frac{\exp(\sigma^{-1})}{2\pi c(\sigma)} \int_{\text{CS}^{k-2}} \{\phi([z]) - \phi([Ux])\} \exp\left(\frac{|x_1|^2 - 1}{\sigma}\right) V_1(dx). \quad (\text{A8})$$

Write $x_j = r_j^{1/2} \exp(i\theta_j)$, $j = 1, \dots, k-1$, with $r = (r_1, \dots, r_{k-1})^T \in S_{k-2}$ and $\theta = (\theta_1, \dots, \theta_{k-1})^T \in (-\pi, \pi)^{k-1}$, so that $V_1(dx) = 2^{2-k} dr_1 \cdots dr_{k-2} d\theta_1 \cdots d\theta_{k-1}$. Hence (A8) can be written as $I(m) =$

$$2^{1-k} \pi^{-1} \exp(\sigma^{-1}) c^{-1}(\sigma) \int_{S_{k-2} \times [0, 2\pi)^{k-1}} \left\{ \phi([z]) - \phi([y(r, \theta, z)]) \right\} \exp\left(\frac{r_1 - 1}{\sigma}\right) dr d\theta, \quad (\text{A9})$$

with $y \equiv y(r, \theta, z) = \sum_{j=1}^{k-1} r_j^{1/2} \exp(i\theta_j) U_j$. Then $d_E^2([y], [z]) = 2(1 - r_1)$. For $d \in \mathfrak{R}^+$, define

$$\psi(d) = \sup\{|\phi(m_1) - \phi(m_2)| : m_1, m_2 \in \Sigma_2^k, d_E^2(m_1, m_2) \leq d\}.$$

Then the absolute value of $\phi([z]) - \phi([y(r, \theta, z)])$ in (A9) is at most $\psi(2(1 - r_1))$, so that

$$\begin{aligned} \sup_{m \in \Sigma_2^k} |I(m)| &\leq \pi^{k-2} \exp(\sigma^{-1}) c^{-1}(\sigma) \int_{S_{k-2}} \psi(2(1 - r_1)) \exp\left(\frac{r_1 - 1}{\sigma}\right) dr_1 \cdots dr_{k-2} \\ &= \pi^{k-2} (k-3)!^{-1} \exp(\sigma^{-1}) c^{-1}(\sigma) \int_0^1 \psi(2(1 - r_1)) \exp\left(\frac{r_1 - 1}{\sigma}\right) (1 - r_1)^{k-3} dr_1. \end{aligned} \quad (\text{A10})$$

Make a change of variable $s = \sigma^{-1}(1 - r_1)$ to rewrite (A10) as

$$\begin{aligned} \sup_{m \in \Sigma_2^k} |I(m)| &\leq \pi^{k-2} (k-3)!^{-1} \sigma^{k-2} \exp(\sigma^{-1}) c^{-1}(\sigma) \int_0^{\sigma^{-1}} \psi(2\sigma s) \exp(-s) s^{k-3} ds \\ &\leq C_k c_1^{-1}(\sigma) \int_0^\infty \psi(2\sigma s) \exp(-s) s^{k-3} ds, \end{aligned} \quad (\text{A11})$$

where $c_1(\sigma) = 1 - \exp(-\sigma^{-1}) \sum_{r=0}^{k-3} \sigma^{-r} / r!$ and C_k is some constant depending on k . Since ϕ is uniformly continuous on the compact metric space (Σ_2^k, d_E) , ψ is bounded and $\lim_{d \rightarrow 0} \psi(d) = 0$. Also it is easy to check that $\lim_{\sigma \rightarrow 0} c_1(\sigma) = 1$. Since $\exp(-s) s^{k-3}$ is integrable on $(0, \infty)$, using the Lebesgue

673 Dominated Convergence Theorem on the integral in (A11), we conclude that

$$674 \lim_{\sigma \rightarrow 0} \sup_{m \in \Sigma_2^k} |I(m)| = 0.$$

675
676 Hence Assumption 2 is also satisfied. This completes the proof. \square

677 678 679 ACKNOWLEDGEMENT

680 This research was partially supported by a grant from the National Institute of Environmental Health
681 Sciences (NIEHS) of the National Institutes of Health (NIH).

682 683 684 REFERENCES

- 685 BHATTACHARYA, A. & BHATTACHARYA, R. N. (2008). Nonparametric statistics on manifolds with applications
686 to shape spaces. *Pushing the Limits of Contemporary Statistics: Contributions in honor of J.K. Ghosh. IMS*
687 *Collections* **3**, 282–301.
- 688 BHATTACHARYA, R. N. & PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means
689 on manifolds. *Ann. Statist.* **31**, 1–29.
- 690 DRYDEN, I. L. & MARDIA, K. V. (1998). *Statistical Shape Analysis*. Wiley N.Y.
- 691 ESCOBAR, M. D. & WEST, M. (1995). Bayesian density-estimation and inference using mixtures. *J. Am. Statist.*
692 *Assoc.* **90**, 577–588.
- 693 FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.
- 694 FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–629.
- 695 KENDALL, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. of the London*
696 *Math. Soc.* **16**, 81–121.
- 697 KENT, J. (1994). The complex Bingham distribution and shape analysis. *J. Roy. Statist. Soc. Ser. B* **56**, 285–299.
- 698 KUME, A. & WALKER, S. G. (2006). Sampling from compositional and directional distributions. *Statist. Comput.*
699 **16**, 261–65.
- 700 LENNOX, K. P., DAHL, D. B. & VANNUCCI, M. A. (2009). Density estimation for protein conformation angles
701 using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Am. Statist. Assoc.* **104**, 586–596.
- 702 LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. 1. density estimates. *Ann. Statist.* **12**, 351–357.
- 703 MARDIA, K. V., TAYLOR, C. C. & SUBRAMANIAM, G. K. (2007). Protein bioinformatics and mixtures of bivariate
704 von Mises distributions for angular data. *Biometrics* **63**, 505–12.
- 705 NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Statist.*
706 **9**, 249–65.
- 707 PAPASPILIOPOULOS & ROBERTS (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process
708 hierarchical models. *Biometrika* **95**, 169–86.
- 709 PELLETIER, B. (2005). Kernel density estimation on Riemannian manifolds. *Statist. Prob. Let.* **73**, 297–304.
- 710 SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4**, 10–26.
- 711 STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Statist. Soc. B* **62**, 795–809.
- 712 WATSON, G. (1965). Equatorial distributions on a sphere. *Biometrika* **52**, 193–201.
- 713 WATSON, G. S. (1983). *Statistics on spheres. University of Arkansas Lecture Notes in the Mathematical Sciences*,
714 vol. 6. New York: Wiley.
- 715 WILLMORE, T. (1993). *Riemannian Geometry*. Oxford: Oxford Uni. Press.
- 716 WU, Y. & GHOSAL, S. (2008). Kullback-Leibler property of kernel mixture priors in Bayesian density estimation.
717 *Elec J. Statist.* **2**, 298–331.
- 718 YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G. & HOLMES, C. (2010). Nonparametric hidden Markov models
719 with application to the analysis of copy-number-variation in mammalian genomes. *J. R. Statist. Soc. B*, under
720 revision.