Ann Inst Stat Math (2012) 64:687–714
DOI 10.1007/s10463-011-0341-x

Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds

Abhishek Bhattacharya · David B. Dunson

Received: 22 July 2010 / Revised: 19 April 2011 / Published online: 18 November 2011

Abstract This article considers a broad class of kernel mixture density models on compact metric spaces and manifolds. Following a Bayesian approach with a nonparametric prior on the location mixing distribution, sufficient conditions are obtained on the kernel, prior and the underlying space for strong posterior consistency at any continuous density. The prior is also allowed to depend on the sample size n and sufficient conditions are obtained for weak and strong consistency. These conditions are verified on compact Euclidean spaces using multivariate Gaussian kernels, on the hypersphere using a von Mises-Fisher kernel and on the planar shape space using complex Watson kernels.

Keywords Nonparametric Bayes · Density estimation · Posterior consistency · Sample-dependent prior · Riemannian manifold · Hypersphere · Shape space

1 Introduction

Density estimation on compact metric spaces, such as manifolds, is a fundamental problem in nonparametric inference on non-Euclidean spaces. Some applications include directional and axial data analysis, spatial modeling, shape analysis and dimensionality reduction problems in which the data lie on an unknown lower dimensional space. However, the literature on statistical theory and methods of density estimation in

A. Bhattacharya
Indian Statistical Institute, 203 B.T. Road, Kolkata, WB 700108, India
e-mail: abhishek@isical.ac.in

D. B. Dunson (✉)
Department of Statistical Science, Duke University, Durham, NC 27708, USA
e-mail: dunson@stat.duke.edu

non-Euclidean spaces is still underdeveloped. Our focus is on Bayesian nonparametric approaches.

For nonparametric Bayes density estimation on the real line $\Re$, there is a rich literature, with Dirichlet process mixtures of Gaussian kernels providing a commonly used approach (Escobar and West 1995) that leads to dense support (Lo 1984) and weak and strong posterior consistency (Ghosal et al. 1999). From the celebrated theorem of Schwartz (1965), weak posterior consistency results when the true density f_0 is in the Kullback–Leibler (KL) support of the prior, meaning that all KL neighborhoods around f_0 are assigned positive probability. In general, it is quite difficult to show KL support for new priors for a density, though Wu and Ghosal (2008) provide useful conditions for a class of kernel mixture priors, with Bhattacharya and Dunson (2010a) extending these conditions to general compact metric spaces. It is widely accepted that weak consistency is an insufficient property when the focus is on density estimation. For example, if f_0 is a density with respect to Lebesgue measure, weak consistency does not even ensure that the posterior assigns positive probability to the set of densities with respect to Lebesgue measure. Hence, it is important to provide stronger results.

Until very recently, essentially all the literature on theory of nonparametric Bayes density estimation focused on one-dimensional Euclidean spaces. An important development in multivariate Euclidean spaces is the article of Wu and Ghosal (2010) who provide sufficient conditions for strong consistency in nonparametric Bayes density estimation from Dirichlet process mixtures of multivariate Gaussian kernels. However, severe tail restrictions are imposed on the kernel covariance, which become overly restrictive when the data are very high dimensional. Also the theory developed in their paper is specialized and cannot be easily generalized to arbitrary kernel mixtures on more general spaces.

We are particularly interested in density estimation in the special case in which the compact metric space M corresponds to a Riemannian manifold. To extend kernel mixture models used in Euclidean spaces to manifolds M, the kernel needs to be carefully chosen. One approach is to introduce an invertible coordinate map between a subset of M and a Euclidean space (Hirsch 1976). Under such an approach, the density prior on M can be induced through a kernel mixture model in a Euclidean space. However, several major problems arise in using such an approach. First, it is not possible to cover the entire manifold with a single smooth coordinate chart except for very simple manifolds, so unless the data are very concentrated one may obtain poor performance. Different local charts can be patched together to form an atlas, but this may introduce artifactual discontinuities in the resulting density. Because the coordinate map is not isometric, the geometry of the manifold can be heavily distorted. As good choices of coordinate frames necessarily depend on the observations, additional uncertainty is automatically induced. Due to these and other shortcomings of coordinate-based methods, we focus on modeling approaches that are coordinate free in the sense that we build density models with respect to the invariant volume form on the manifold.

In Bhattacharya and Dunson (2010a), a density model is presented on a general compact metric space with respect to any fixed base measure using a random mixture of probability kernels. Under mild conditions on the kernel and the mixing prior, it is shown that the prior probability of any uniform neighborhood of any continuous

density f_0 is positive and if f_0 is positive everywhere, it lies in the KL support of the prior. This in turn implies posterior consistency in the weak sense. Density estimation on the planar shape space is presented as a special case. However, strong posterior consistency is not addressed. The everywhere positivity restriction on the true density cannot be easily relaxed. Also besides the location mixing distribution, the only other parameter in the model is a scalar bandwidth. This restricts the flexibility when the sample size is small.

Focusing on kernel mixture priors for densities on a compact metric space M, in this article, we provide sufficient conditions on the kernel, prior and the underlying space to ensure strong consistency. Theorem 2 and Corollary 1 provide sufficient conditions to ensure that all total variation neighborhoods around f_0 will be assigned probability converging to one as the sample size increases. The theoretical development relies on the method of sieves and exponentially consistent tests discussed in Barron (1989). However, applying this framework outside multivariate spaces is not standard and requires careful use of differential geometry. Through Theorem 1, we prove weak consistency for a bigger class of kernels than Bhattacharya and Dunson (2010a). The only requirement on the true density is that it is continuous everywhere. To illustrate the theory, we focus on density estimation on the unit hypersphere using von Mises-Fisher kernels and on the planar shape space using complex Watson kernels. In both these cases, it is shown that the kernels satisfy the sufficient conditions. The results also apply to Gaussian mixture densities on $\Re^d$ whenever the true density has compact support. In that case, a truncated and transformed Wishart prior on the covariance inverse, the transformation depending on the data dimension is shown to suffice as in Wu and Ghosal (2010). Appropriate kernel choices are presented on other manifolds such as axial spaces, Stiefel manifolds and Grassmannians which arise as generalizations of the two manifolds.

When the manifold is high-dimensional, priors satisfying conditions for strong consistency tend to put too little probability near bandwidths close to 0, which is undesirable for applications. A gamma prior on the inverse-bandwidth, for example, cannot be shown to satisfy the conditions. Hence, we extend the consistency results to cover cases with priors depending on the sample size n. Theorem 3 extends the Schwartz theorem to prove weak consistency, while Theorem 4 proves strong consistency using such priors. A gamma prior with scale decreasing with n at an appropriate rate satisfies the conditions for both weak and strong posterior consistency at an exponential rate. When using multivariate Gaussian mixtures, a truncated Wishart prior with hyper-parameters depending on n, is shown to work.

To maintain a free flow while reading, we put all the proofs together at the end in a section called Appendix.

2 Consistency theorems on compact metric spaces

2.1 Weak posterior consistency

Let (M, ρ) be a compact metric space, ρ being the distance metric, and let X be a random variable on M (from some measurable space $(\Omega, \mathcal{B}, Q)$). We assume that the

distribution of X has a density with respect to some fixed finite base measure λ on M. The natural choice for such a λ when M is a Riemannian manifold is the invariant volume form. We are interested in modeling this unknown density via a flexible model. Let $K(m;\mu,\mathcal{K})$ be a probability kernel on M with location $\mu \in M$ and other parameters $\mathcal{K} \in N$, N being a Polish space, that is, it is homeomorphic to a complete separable metric space. In the special case, we choose $N = (0,\infty)$ and then K may be called a location-scale kernel.

Given the parameters $(\mu,\mathcal{K})$, K satisfies $\int_M K(m;\mu,\mathcal{K})\lambda(\mathrm{d}m) = 1$. Then a location mixture density model for X is defined as

$$f(m;P,\mathcal{K}) = \int_M K(m;\mu,\mathcal{K})P(\mathrm{d}\mu) \tag{1}$$

with parameters P in the space $\mathcal{M}(M)$ of all probability distributions on M and $\mathcal{K} \in N$. We call P the location mixing distribution. When $N = (0,\infty)$, we view $\mathcal{K}^{-1}$ as the bandwidth of the kernel and hence call $\mathcal{K}$ the *precision* or inverse bandwidth parameter. More generally $\mathcal{K}$ comprises of many other parameters in different spaces determining the kernel shape, modality, etc., and the precision is a particular function of $\mathcal{K}$. The upcoming consistency theorems and examples will illustrate that function. Kernel mixture models are used routinely in Bayesian density estimation in Euclidean spaces, with Lennox et al. (2009) applying such an approach to bivariate angular data and Bhattacharya and Dunson (2010a,b) considering kernel mixtures on general metric spaces.

A prior Π_1 on $(P,\mathcal{K})$ induces a prior Π on the space of densities $\mathcal{D}(M)$ on M through the model (1). Given a random realization $X_1,\ldots,X_n$ of X, we can compute the posterior of f. The Schwartz theorem provides a useful tool in proving that the posterior assigns probability converging to one in arbitrarily small neighborhoods of the true density f_0 as the sample size $n \to \infty$. Let F_0 denote the probability distribution corresponding to f_0, let $\mathrm{KL}(f_0; f) = \int_M f_0(m)\log\{f_0(m)/f(m)\}\lambda(dm)$ denote the KL divergence of another density f from f_0, and let $K_\epsilon(f_0)$ denote the KL neighborhood $\{f \in \mathcal{D}(M) : \mathrm{KL}(f_0; f) < \epsilon\}$. f_0 is said to be in the KL support of Π, or Π is said to satisfy the KL condition at f_0 if $\Pi\{K_\epsilon(f_0)\} > 0$ for all $\epsilon > 0$.

Proposition 1 (Schwartz theorem) *If (1) f_0 is in the KL support of Π, and (2) $U \subset \mathcal{D}(M)$ is such that there exists a uniformly exponentially consistent sequence of test functions for testing H_0: $f = f_0$ versus H_1: $f \in U^c$, then $\Pi(U|X_1,\ldots,X_n) \to 1$ as $n \to \infty$ a.s. F_0^∞.*

The posterior probability of U^c can be expressed as

$$\Pi(U^c|X_1,\ldots,X_n) = \frac{\int_{U^c}\prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)}\Pi(\mathrm{d}f)}{\int\prod_1^n \frac{f(X_i)}{f_0(X_i)}\Pi(\mathrm{d}f)} \tag{2}$$

Condition (1), known as the KL condition, ensures that for any $\beta > 0$,

$$\liminf_{n\to\infty} \exp(n\beta) \int \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} \Pi(\mathrm{d}f) = \infty \text{ a.s.} \tag{3}$$

while condition (2) implies that

$$\lim_{n\to\infty} \exp(n\beta_0) \int_{U^c} \prod_{i=1}^{n} \frac{f(X_i)}{f_0(X_i)} \Pi(\mathrm{d}f) = 0 \text{ a.s.}$$

for some $\beta_0 > 0$ (depending on U) and therefore

$$\lim_{n\to\infty} \exp(n\beta_0/2)\Pi(U^c|X_1, \ldots, X_n) = 0 \text{ a.s.}$$

Hence Proposition 1 provides conditions for posterior consistency at an exponential rate. Theorem 1 derives sufficient conditions on the kernel and the prior so that f_0 is in the uniform support and hence KL support of Π. They are

A1 The kernel K is continuous in its arguments.
For any continuous function $f : M \to \Re$ [written as $f \in C(M)$], for any $\epsilon > 0$, there exists a compact subset N_ϵ of N with non-empty interior, such that
A2 $\sup_{m\in M, \mathcal{K}\in N_\epsilon} \left| f(m) - \int_M K(m; \mu, \mathcal{K}) f(\mu)\lambda(\mathrm{d}\mu) \right| < \epsilon$.
A3 For any $\epsilon > 0$, the set $\{F_0\} \times N_\epsilon^o$ intersects with the (weak) support of Π_1. Here A^o denotes the interior of a set A.
A4 f_0 is a continuous density.

Theorem 1 *Under assumptions **A1–A4**, for any $\epsilon > 0$,*

$$\Pi\left\{ f \in \mathcal{D}(M) : \sup_{m\in M} |f(m) - f_0(m)| < \epsilon \right\} > 0,$$

which implies that f_0 is in the KL support of Π.

As a corollary, we obtain the KL property for the location-scale kernel, derived in Bhattacharya and Dunson (2010a). However, unlike in there, we need not assume f_0 to be positive everywhere.

When U is a weakly open neighborhood of f_0, condition (2) in Proposition 1 is always satisfied. Hence under assumptions **A1**–**A4**, weak posterior consistency at an exponential rate follows. Assumptions **A1** and **A2** impose some mild conditions on the kernel choice, which are easily satisfied by several parametric families. In particular, **A2** implies that as a probability distribution on M, $K(; \mu, \mathcal{K})$ can be made arbitrarily close in weak sense to the degenerate point mass at μ, uniformly in μ, for appropriate $\mathcal{K}$ choice, thereby justifying the name 'location' for μ. When the compact neighborhood N_ϵ can be represented as the inverse image under some function ψ ($\psi : N \to \Re^+$) of some neighborhood around infinity, then $\psi(\mathcal{K})$ can be viewed as the precision parameter. We will provide examples of kernels on some non-Euclidean

manifolds arising in shape and directional data analysis which satisfy **A1** and **A2**. A common choice for Π_1 satisfying **A3** can be a full support product prior such as a Dirichlet process $\mathrm{DP}(w_0 P_0)$ prior on P satisfying $\mathrm{supp}(P_0) = M$ and an independent everywhere positive density prior on $\mathcal{K}$.

2.2 Strong consistency

When U is a total variation neighborhood of f_0, LeCam (1973); Barron (1989) show that condition (2) of Proposition 1 will not be satisfied in most cases. In Barron (1989) (also see Ghosal et al. 1999), a sieve method is considered to obtain sufficient conditions for the numerator in (2) to decay at an exponential rate and hence get strong posterior consistency at an exponential rate. This is stated in Proposition 2. In its statement, for $\mathcal{F} \subseteq \mathcal{D}(M)$ and $\epsilon > 0$, the L_1-metric entropy $N(\epsilon, \mathcal{F})$ is defined as the logarithm of the minimum number of ϵ-sized (or smaller) L_1 subsets needed to cover $\mathcal{F}$.

Proposition 2 *If there exists a $\mathcal{D}_n \subseteq \mathcal{D}(M)$ such that (1) for n sufficiently large, $\Pi(\mathcal{D}_n^c) < exp(-n\beta)$ for some $\beta > 0$, and (2) $N(\epsilon, \mathcal{D}_n)/n \to 0$ as $n \to \infty$ for any $\epsilon > 0$, then for any total variation neighborhood U of f_0, there exists a $\beta_0 > 0$ (depending on U) such that* $\limsup_{n\to\infty} \exp(n\beta_0) \int_{U^c} \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi(\mathrm{d}f) = 0$ *a.s. F_0^∞. Hence if f_0 is in the KL support of Π, the posterior probability of any total variation neighborhood of f_0 converges to 1 almost surely.*

Theorem 2, which is the main theorem of this paper, describes a $\mathcal{D}_n$ which satisfies condition (2). We assume that there exists a continuous function $\phi : N \to [0, \infty)$ for which the following assumptions hold:

A5 There exists positive constants κ_1, a_1, A_1 such that for all $\kappa \geq \kappa_1$, $\mu, \nu \in M$,

$$\sup_{m \in M, \mathcal{K} \in \phi^{-1}[0,\kappa]} \left| K(m; \mu, \mathcal{K}) - K(m; \nu, \mathcal{K}) \right| \leq A_1 \kappa^{a_1} \rho(\mu, \nu).$$

A6 There exists positive constants a_2, A_2 such that for all $\mathcal{K}_1, \mathcal{K}_2 \in \phi^{-1}[0, \kappa]$, $\kappa \geq \kappa_1$,

$$\sup_{m, \mu \in M} \left| K(m; \mu, \mathcal{K}_1) - K(m; \mu, \mathcal{K}_2) \right| \leq A_2 \kappa^{a_2} \rho_2(\mathcal{K}_1, \mathcal{K}_2),$$

ρ_2 metrizing the topology of N.

A7 For any $\kappa \geq \kappa_1$, the subset $\phi^{-1}[0, \kappa]$ is compact and given $\epsilon > 0$, the minimum number of ϵ (or smaller) radius balls covering it (known as the *ϵ-covering number*) can be bounded by $(\kappa \epsilon^{-1})^{b_2}$ for appropriate positive constant b_2 (independent of κ and ϵ).

A8 There exists $a_3, A_3 > 0$ such that the ϵ-covering number of M is bounded by $A_3 \epsilon^{-a_3}$ for any $\epsilon > 0$.

For two positive sequences $\{a_n\}$ and $\{b_n\}$, $\{a_n\}$ is said to be 'little-o' of $\{b_n\}$, written as $a_n = o(b_n)$, if the sequence $\{a_n/b_n\}$ converges to 0 as $n \to \infty$.

Theorem 2 *For a positive sequence* $\{\kappa_n\}$ *diverging to* ∞*, define*

$$\mathcal{D}_n = \{f(P,\mathcal{K}) : P \in \mathcal{M}(M),\ \mathcal{K} \in \phi^{-1}[0,\kappa_n]\}.$$

Under assumptions ***A5–A8****, given any* $\epsilon > 0$*, for n sufficiently large,* $N(\epsilon,\mathcal{D}_n) \leq C(\epsilon)\kappa_n^{a_1a_3}$ *for some* $C(\epsilon) > 0$*. Hence,* $N(\epsilon,\mathcal{D}_n)$ *is* $o(n)$ *whenever* κ_n *is* $o\left(n^{(a_1a_3)^{-1}}\right)$.

As a corollary, we derive conditions on the prior Π_1 on $(P,\mathcal{K})$ under which strong posterior consistency at an exponential rate follows:

Corollary 1 *Under the hypothesis of Theorems 1 and 2 and* **A9** $\Pi_1(\mathcal{M}(M)\times\phi^{-1}(n^a,\infty)) < \exp(-n\beta)$ for some $a < (a_1a_3)^{-1}$ and $\beta > 0$*, the posterior probability of any total variation neighborhood of* f_0 *converges to 1 a.s.* F_0^∞.

Theorem 1 ensures that f_0 is in the KL support of Π. Theorem 2 and assumption **A9** ensure that $\mathcal{D}_n$ satisfies conditions (1) and (2) of Proposition 2. Hence from the Proposition, the proof follows.

When we use a location-scale kernel, that is, when $N = (0,\infty)$, choose a prior $\Pi_1 = \Pi_{11}\otimes\pi_1$ having full support, set ϕ to be the identity map, then a choice for π_1 for which Assumption **A9** is satisfied is a Weibull density $Weib(\mathcal{K};\alpha,\beta) \propto \mathcal{K}^{\alpha-1}\exp(-\beta\mathcal{K}^\alpha)$ whenever the shape parameter α exceeds a_1a_3 (a_1, a_3 as in Assumptions **A5** and **A8**).

Remark 1 A gamma prior on $\mathcal{K}$ satisfies the requirements for weak consistency but not **A9** (unless $a_1a_3 < 1$). However, that does not prove that it is not eligible for strong consistency because Corollary 1 provides only sufficient conditions. In Section 2.3, we prove it to be eligible as long as the hyperparameters are allowed to depend on sample size n in a suitable way.

When the underlying space is non-compact such as $\Re^d$, Corollary 1 applies to any true density f_0 supported on a compact set, say M. Then the kernel can be chosen to have non-compact support, such as Gaussian, but to apply Theorem 2, we need to restrict the prior on the location mixing distribution to have support in $\mathcal{M}(M)$. We can weaken assumptions **A5** and **A6** to

A5′ $\sup_{\mathcal{K}\in\phi^{-1}[0,\kappa]} \|K(\mu,\mathcal{K}) - K(\nu,\mathcal{K})\| \leq A_1\kappa^{a_1}\rho(\mu,\nu)$ and
A6′ $\sup_{\mu\in M} \|K(\mu,\mathcal{K}_1) - K(\mu,\mathcal{K}_2)\| \leq A_2\kappa^{a_2}\rho_2(\mathcal{K}_1,\mathcal{K}_2)\ \forall\mathcal{K}_1,\mathcal{K}_2 \in \phi^{-1}[0,\kappa]$,
respectively. Here $\|f-g\|$ denotes the L_1 distance between densities f and g. The proof of Theorem 2 can be easily modified to show consistency under the new assumptions and is left to the reader.

The multivariate Gaussian kernel can be represented as

$$K(m,\mu,\mathcal{K}) = (2\pi)^{-d/2}\det(\mathcal{K})^{1/2}\exp\big(-1/2(m-\mu)'\mathcal{K}(m-\mu)\big)\,m,$$
$$\mu \in \Re^d, \mathcal{K} \in M^+(d),$$

$M^+(d)$ being the space of all $d\times d$ positive matrices. Hence $\mathcal{K}^{-1}$ is the kernel covariance. It satisfies **A5′** and **A6′** as shown in Proposition 3. Here by $\lambda_1(\mathcal{K}),\ldots,\lambda_d(\mathcal{K})$, we denote the eigenvalues of $\mathcal{K}$ in increasing order.

Proposition 3 *The multivariate Gaussian kernel satisfies* **A5**′ *with* ϕ *being the largest eigenvalue function,* $\phi(\mathcal{K}) = \lambda_d(\mathcal{K})$ *and* $a_1 = 1/2$. *It also satisfies* **A6**′ *once we restrict* N *to be the space of all positive matrices with the least eigenvalue being bounded below by some pre-specified positive constant, say,* λ_1, *i.e.,* $N = \{\mathcal{K} \in M^+(d) : \lambda_1(\mathcal{K}) \geq \lambda_1\}$. *The space* $M^+(d)$ *(and hence* N*) satisfies* **A7** *while any compact subset* M *of* $\Re^d$ *satisfies* **A8** *from Theorem 2 with* $a_3 = d$. *Hence if*

$$\Pi_1\left(\mathcal{M}(M) \times \{\mathcal{K} \in N : \lambda_d(\mathcal{K}) > n^a\}\right) < \exp(-n\beta)$$

for some $a < 2/d$ *and* $\beta > 0$, *and if* f_0 *is in the KL support of* Π, *strong posterior consistency follows.*

Theorem 4 in Ghosal et al. (1999) provides sufficient conditions on f_0 and Π_1 for the KL condition to be satisfied while using a Gaussian mixture model in the univariate setting to model a compactly supported density. It assumes $\Pi_1 = \Pi_{11} \otimes \pi_1$ with $F_0 \in \text{supp}(\Pi_{11})$ and $\infty \in \text{supp}(\pi_1)$. The theorem can be extended to multivariate setting with the condition on Π_1 relaxed to, for any $\kappa > 0$, there exists a $\mathcal{K} \in M^+(d)$ with $\lambda_1(\mathcal{K}) \geq \kappa$ such that $(F_0, \mathcal{K}) \in \text{supp}(\Pi_1)$. Therefore, $\lambda_1(\mathcal{K})$ can be viewed as the kernel precision in this case. A full support product Π_1 on $\mathcal{M}(M) \times N$ will satisfy these requirements. Using a product prior, a choice for π_1 for which strong consistency also follows can be the so called *truncated transformed Wishart* defined as follows: Set $\mathcal{K} = \Lambda^a$ for any $a \in (0, 2/d)$ with Λ following a Wishart distribution restricted to N. Then $\mathcal{K}$ is said to follow a truncated transformed Wishart with transformation power a.

Remark 2 The truncation restriction on the space N is not undesirable, because for more precise fit, we are interested in low bandwidths and the least eigenvalue of $\mathcal{K}$ can be viewed as the inverse of the bandwidth. However, the lower the transformation power, the lower is the prior probability for high precisions which is undesirable when sample sizes are not high.

In Wu and Ghosal (2010), strong consistency is proved in the special case of Dirichlet process Gaussian mixtures used to model density f_0 having support as $\Re^d$. It requires a to be less than $1/d$ resulting in even smaller precision. In the next section, we prove that no transformation is required ($a = 1$) as long as the hyper-parameters are allowed to depend on the sample size appropriately.

2.3 Consistency with sample size-dependent priors

When the dimension of the manifold is large, as is the case in shape analysis with a large number of landmarks, the constraints on the shape parameter in the proposed Weibull prior on the inverse bandwidth become overly restrictive. In particular, for strong posterior consistency, the shape parameter needs to be very large in high-dimensional cases, implying a prior on the bandwidth that places very small probability in neighborhoods close to zero, which is undesirable in many applications. By instead allowing the prior to depend on sample size n, we can potentially obtain priors that

may have better small sample operating characteristics, while still leading to strong consistency. However, for n-dependent priors, the KL condition is no longer sufficient to ensure that (3) holds and hence the Schwartz theorem breaks down. In this section, we will modify the conditions and derive weak and strong consistency results for n-dependent priors.

As recommended in earlier sections, we let P and $\mathcal{K}$ be independent under Π_1. Then, assuming $P \sim \Pi_{11}$ is a constant prior, we focus on the case in which $\mathcal{K}$ has a sample size-dependent prior density π_n with respect to some base measure λ_1 on N, $\mathcal{K} \sim \pi_n(\mathcal{K})\lambda_1(d\mathcal{K})$. We pick λ_1 to have full support. Depending on the context, π_n will refer to both the density and distribution of $\mathcal{K}$. Denote the resulting sequence of induced priors on $\mathcal{D}(M)$ as Π_n. Theorem 3 proves weak posterior consistency under the following assumptions on the prior:

A10 The prior Π_{11} contains F_0 in its support.
A11 For any $\epsilon > 0$, for all $\mathcal{K} \in N_\epsilon$,

$$\liminf_{n\to\infty} \exp(n\epsilon)\pi_n(\mathcal{K}) = \infty.$$

Here N_ϵ is as defined in Assumption **A2**.

Theorem 3 *Under assumptions **A1** and **A2** on the kernel, **A4** on the true density f_0, and, **A10** and **A11** on the prior, the posterior probability of any weak neighborhood of f_0 converges to one a.s. F_0^∞.*

The proof is immediate from the following two lemmas:

Lemma 1 *Under assumptions **A1–A2**, **A4** and **A10–A11**, for any $\epsilon > 0$,*

$$\liminf_{n\to\infty} \exp(n\epsilon) \int \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi_n(\mathrm{d}f) = \infty \tag{4}$$

a.s. F_0^∞.

Lemma 2 *If there exists a uniformly exponentially consistent sequence of test functions for testing H_0: $f = f_0$ versus H_1: $f \in U^c$, and $\Pi_n(U^c) > 0$ for all $n > C$ with C a sufficiently large constant, then for some $\beta_0 > 0$,*

$$\lim_{n\to\infty} \exp(n\beta_0) \int_{U^c} \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi_n(\mathrm{d}f) = 0$$

a.s. F_0^∞.

The proof of Lemma 2 is related to that of Lemma 4.4.2. from Ghosh and Ramamoorthi (2003) which is stated for a constant prior Π, but with the set U^c depending on n, they call this V_n. There it is assumed that $\liminf_{n\to\infty} \Pi(V_n) > 0$, but that is not necessary as long as $\Pi(V_n) > 0$ for all large n. Lemma 1 is proved in the Appendix.

With a location-scale kernel, N being $(0, \infty)$, a gamma prior $\pi_n(\mathcal{K}) \propto \exp(-\beta_n \mathcal{K})$ $\mathcal{K}^{\alpha-1}$, $\alpha, \beta_n > 0$, denoted by $\text{Gam}(\alpha, \beta_n)$, satisfies Assumption **A11** on entire of N, as long as β_n is $o(n)$.

With a multivariate Gaussian kernel on $\Re^d$ with dispersion $\mathcal{K}^{-1}$, a Wishart prior on $\mathcal{K}$,

$$\pi_n(\mathcal{K}; \beta_n, q) = 2^{-dq/2}\Gamma_d(q/2)\beta_n^{dq/2} \exp(-\beta_n/2\text{Tr}(\mathcal{K}))\det(\mathcal{K})^{(q-d-1)/2},$$
$$q > d-1, \ \beta_n > 0,$$

denoted as $\text{Wish}(\beta_n^{-1} I_d, q)$ satisfies **A11** on entire $M^+(d)$, as long as β_n is $o(n)$. Here $\Gamma_d(.)$ denotes the *multivariate gamma function* defined as

$$\Gamma_d(q/2) = \int_{M^+(d)} \exp(-\text{Tr}(\mathcal{K}))\det(\mathcal{K})^{(q-d-1)/2} d\mathcal{K}.$$

For strong consistency, we impose the following additional condition on π_n. Let a_1 and a_3 be as in Assumptions **A5** (or **A5′**) and **A8**, respectively.

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

$$\lim_{n\to\infty} \exp(n\beta_0)\pi_n\{\phi^{-1}(n^a, \infty)\} = 0.$$

This assumption is in place of **A9** used for constant priors.

Theorem 4 *Under Assumptions **A1–A2** and **A4–A8** and **A10–A12**, the posterior probability of any total variation neighborhood of f_0 converges to 1 a.s F_0^∞.*

The proof is very similar to that of Corollary 1. This is because under assumptions **A1**–**A2**, **A4** and **A10**–**A11**, the conclusion (4) of Lemma 1 holds. The other assumptions are to show that the L_1-metric entropy of $\mathcal{D}_n$ is $o(n)$ while $\Pi_n(\mathcal{D}_n^c)$ is exponentially small, $\mathcal{D}_n$ being defined in Theorem 2. Under these assumptions, the proof of Proposition 2 goes through to prove strong consistency with sample size-dependent priors. This is also mentioned in §5 of Ghosal et al. (1999). They require $\liminf_{n\to\infty} \Pi_n(K_\epsilon(f_0)) > 0$ in place of the assumption $\Pi(K_\epsilon(f_0)) > 0$ for constant priors, but this is only to ensure that (4) holds.

Again as in Sect. 2.2, we can weaken assumptions **A5** and **A6** to **A5′** and **A6′**, respectively.

For a location-scale kernel, a $\text{Gam}(\alpha, \beta_n)$ prior on precision $\mathcal{K}$ satisfies **A12** when n^{1-a} is $o(\beta_n)$ for some $a \in (0, (a_1 a_3)^{-1})$. Hence, for example, we have weak and strong posterior consistency with $\beta_n = b_1 n/\{\log(n)\}^{b_2}$ for any $b_1, b_2 > 0$.

For a multivariate Gaussian kernel, to satisfy assumption **A6′**, we need to truncate the space N to $\{\mathcal{K} \in M^+(d) : \lambda_1(\mathcal{K}) \geq \lambda_1\}$, as proved in Proposition 3. Then we may set a truncated Wishart prior on $\mathcal{K}$, defined as

$$\pi_n(\mathcal{K}) = \frac{\exp(-\beta_n/2\text{Tr}(\mathcal{K}))\det(\mathcal{K})^{(q-d-1)/2}}{\int_{A\in N} \exp(-\beta_n/2\text{Tr}(A))\det(A)^{(q-d-1)/2} dA}, \ \mathcal{K} \in N. \tag{5}$$

Then for Assumption **A12** to be satisfied, we require n^{1-a} to be $o(\beta_n)$ for some $a \in (0, (a_1 a_3)^{-1})$. This is shown in Proposition 4. Hence we have weak and strong posterior consistency once we set $\beta_n = b_1 n/\{\log(n)\}^{b_2}$ for any $b_1, b_2 > 0$. Unlike in Sect. 2.2, we impose no transformation constraints, which is very helpful especially when sample sizes are not that high while the data dimensions are huge.

Proposition 4 *For a positive sequence $\{\beta_n\}$ diverging to infinity, Assumption **A12** is satisfied for the truncated Wishart density sequence π_n in (5) if there exists an $a \in (0, (a_1 a_3)^{-1})$ for which β_n satisfies $n^{1-a}/\beta_n \longrightarrow 0$ as $n \to \infty$.*

In the subsequent sections, we present kernel choices for density estimation on some specific non-Euclidean manifolds that arise in several applications. We illustrate how to apply Theorems 1, 2, 3 and 4 and obtain weak and strong posterior consistency.

3 Application to unit hypersphere

Let M be the unit sphere S^d embedded in $\Re^{d+1}$. It is a compact Riemannian manifold of dimension d and a compact metric space under the chord distance $\rho(u, v) = \|u - v\|_2$, $\|.\|_2$ denoting the L^2-norm. Spherical data on S^2 arise in the context of directional data analysis. Most of the shape spaces are quotients of high-dimensional spheres. Hence it is important to develop consistent inference procedures on this space, and very few results exist in the context of Bayesian nonparametrics.

To define a probability density model as in (1) with respect to the volume form V, we need a suitable kernel which satisfies the assumptions in Sect. 2. One of the most commonly used probability densities on this space is the von Mises-Fisher (vMF) density which is given by

$$\text{vMF}(m; \mu, \mathcal{K}) = c^{-1}(\mathcal{K}) \exp(\mathcal{K} m^T \mu), \ m, \mu \in S^d, \ \mathcal{K} \in [0, \infty), \tag{6}$$

with c being the normalizing constant which can be derived to be

$$\frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_{-1}^{1} \exp(\mathcal{K}t)(1 - t^2)^{d/2-1} \mathrm{d}t. \tag{7}$$

The vMF density on S^1 was first derived in von Mises (1918) and the density in case of S^2 was given by Fisher (1953). Watson and Williams (1953) generalized this distribution to S^d and examined many of its properties. It can be shown that the parameter μ is the extrinsic mean (as defined in Bhattacharya and Patrangenaru 2003) and hence can be interpreted as the distribution location. The parameter $\mathcal{K}$ is a measure of concentration, with $\mathcal{K} = 0$ corresponding to the uniform distribution having constant density equal to $1/\int_{S^d} V(dm)$. As $\mathcal{K}$ diverges to ∞, the vMF distribution converges to a point mass at μ in an L^1 sense uniformly. This is proved in Theorem 5.

Theorem 5 *The vMF kernel satisfies Assumptions* **A1** *and* **A2**.

Hence from Theorem 1, weak posterior consistency follows using the location mixture density model (1) with a Dirichlet Process prior on P and an independent gamma

prior on $\mathcal{K}$. In the $d = 2$ special case, Lennox et al. (2009) proposed a closely related model but did not consider theoretical properties. Theorem 6 verifies the assumptions for strong consistency.

Theorem 6 *With $\phi(\mathcal{K}) = \mathcal{K}$, the vMF kernel on S^d satisfies assumption* **A5** *with $a_1 = d/2 + 1$ and* **A6** *with $a_2 = d/2$. The compact metric-space (S^d, ρ) satisfies assumption* **A8** *with $a_3 = d$.*

As a result a Weibull prior on $\mathcal{K}$ with shape parameter exceeding $(d + d^2/2)^{-1}$ satisfies the condition of Corollary 1 and strong posterior consistency follows. The proofs of Theorems 5 and 6 use the following lemma which establishes certain properties of the normalizing constant:

Lemma 3 *Define $\tilde{c}(\mathcal{K}) = \exp(-\mathcal{K})c(\mathcal{K})$, $\mathcal{K} \geq 0$. Then $\tilde{c}$ is decreasing and for $\mathcal{K} \geq 1$,*

$$\tilde{c}(\mathcal{K}) \geq C\mathcal{K}^{-d/2}$$

for some appropriate positive constant C.

When d is large, as is often the case for spherical data, a more appropriate prior on $\mathcal{K}$ for which weak and strong consistencies hold can be Gam(α, β_n) as mentioned at the end of Sect. 2.3.

It is easy to check that the vMF density is the conditional distribution, given $\|X\| = 1$, of a Gaussian random vector X on $\Re^{d+1}$ with mean μ and dispersion matrix $\mathcal{K}^{-1}I_{d+1}$. A more general family of distribution on S^d may be obtained as the conditional distribution, given $\|X\| = 1$, of a Normal X on $\Re^{d+1}$ with mean μ and dispersion matrix $\mathcal{K}^{-1}$, $\mathcal{K}$ in the space $M^+(d + 1)$ of $(d + 1) \times (d + 1)$ positive matrices. Then we obtain the *Fisher–Bingham* family of kernels. It can be interesting to show that the resulting kernel mixture satisfies the assumptions of Theorems 1, 2, 3 and 4 and obtain posterior consistency. We postpone that to later works.

A generalization of the sphere is the Stiefel manifold $V_{d+1,k}$, the space all k dimensional orthonormal frames in $\Re^{d+1}$. One can easily extend the vMF kernel to the so-called Fisher kernel on this manifold and carry out density estimation. Again, proving that consistency holds is postponed for future works.

Another important manifold arising in axial data analysis is RP^d, the space of all rays in $\Re^{d+1}$. This manifold can be obtained as the quotient of S^d after identifying antipodal points p and $-p$ as identical. In the next section, we illustrate density estimation on its complex analog, the complex projective space. It is easy and simpler to obtain analogous results on the real version.

4 Planar shape space

4.1 Background

Let M be the planar shape space Σ_2^k which is defined as follows: Consider a set of k landmark locations, $k > 2$, on a 2D image, not all points being the same. We refer to such a set as a k-ad. The similarity shape of this k-ad is what remains after removing

the Euclidean rigid body motions of translation, rotation and scaling. We use the following shape representation first proposed by Kendall (1984): Denote the k-ad by a complex k-vector z in $\mathcal{C}^k$. To remove the effect of translation from z, let $z_c = z - \overline{z}$, with $\overline{z} = (\sum_{j=1}^k z_j)/k$ being the centroid. The centered k-ad z_c lies in a $k-1$ dimensional complex subspace, and hence we can use $k-1$ complex coordinates. The effect of scaling is then removed by normalizing the coordinates of z_c to obtain a point w on the complex unit sphere $\mathcal{CS}^{k-2}$ in $\mathcal{C}^{k-1}$. Since w contains the shape information of z along with rotation, it is called the preshape of z. The similarity shape of z is the orbit of w under all rotations in 2D which is

$$[w] = \{e^{i\theta} w : \theta \in (-\pi, \pi]\}.$$

This represents a shape as the set of all intersection points of a unique complex line passing through the origin with $\mathcal{CS}^{k-2}$ and the planar shape space Σ_2^k is then the set of all such shapes. Hence Σ_2^k can be identified with the space of all complex lines passing through the origin in $\mathcal{C}^{k-1}$ which is the complex projective space and is a compact Riemannian manifold of dimension $2k-4$. The Σ_2^k can be embedded into the space of all order $k-1$ complex Hermitian matrices via the embedding $J([w]) = ww^*$, $*$ denoting the complex conjugate transpose. This embedding induces a distance on Σ_2^k called the extrinsic distance which generates the manifold topology and is given by

$$d_E([u],[v]) = \|J([u]) - J([v])\| = \sqrt{2(1-|u^*v|^2)} \quad ([u],[v] \in \Sigma_2^k).$$

For more details, see Bhattacharya and Dunson (2010a) and the references cited therein.

4.2 Density model

We define a location-mixture density on Σ_2^k as in (1) with respect to the Riemannian volume form V and the kernel being a complex Watson density. This complex Watson density was used in Dryden and Mardia (1998) for parametric density modeling and is given by

$$\text{CW}(m; \mu, \mathcal{K}) = c^{-1}(\mathcal{K}) \exp\{\mathcal{K}(|z^*v|^2 - 1)\} \quad (m = [z], \mu = [v]) \tag{8}$$

$$\text{with } c(\mathcal{K}) = \pi^{k-2}\mathcal{K}^{2-k}\left(1 - \exp(-\mathcal{K}) \sum_{r=0}^{k-3} \frac{\mathcal{K}^r}{r!}\right). \tag{9}$$

It is shown in Bhattacharya and Dunson (2010a) that the complex Watson kernel satisfies assumptions **A1** and **A2** in Sect. 2. Using a Dirichlet Process prior on the location mixing distribution and an independent gamma prior on the precision parameter, Theorem 1 implies that the density model (1) has full support in the space of all positive continuous densities on Σ_2^k in uniform and KL sense and hence the posterior is weakly consistent.

Theorem 7 verifies that the complex Watson kernel also satisfies the regularity conditions in **A5** and **A6**.

Theorem 7 *The complex Watson kernel* $\text{CW}(m; \mu, \mathcal{K})$ *on the compact metric space* Σ_2^k *endowed with the extrinsic distance* d_E *satisfies assumption* **A5** *with* $a_1 = k - 1$ *and* **A6** *with* $a_2 = 3k - 8$.

The proof uses Lemma 4 which verifies certain properties of the normalizing constant.

Lemma 4 *Let* $c(\mathcal{K})$ *be the normalizing constant for* $\text{CW}(\mu, \mathcal{K})$ *as defined in* (9). *Then* c *is decreasing on* $[0, \infty)$ *with*

$$\lim_{\kappa \to 0} c(\mathcal{K}) = \frac{\pi^{k-2}}{(k-2)!} \text{ and } \lim_{\mathcal{K} \to \infty} c(\mathcal{K}) = 0.$$

If we define

$$\tilde{c}(\mathcal{K}) = 1 - \exp(-\mathcal{K}) \sum_{r=0}^{k-3} \frac{\mathcal{K}^r}{r!},$$

it follows that $\tilde{c}$ *is increasing with*

$$\lim_{\mathcal{K} \to 0} \tilde{c}(\mathcal{K}) = 0, \lim_{\mathcal{K} \to \infty} \tilde{c}(\mathcal{K}) = 1 \text{ and}$$
$$\tilde{c}(\mathcal{K}) \geq (k-2)!^{-1} \exp(-\mathcal{K}) \mathcal{K}^{k-2}.$$

The proof follows from direct computations.

Theorem 8 verifies that assumption **A8** holds on Σ_2^k.

Theorem 8 *The compact metric space* (Σ_2^k, d_E) *satisfies assumption* ***A8*** *with* $a_3 = 2k - 3$.

As a result, Corollary 1 implies that strong posterior consistency holds with $\Pi_1 = (DP)(\omega_0 P_0) \otimes \pi_1$, for Weibull π_1 with shape parameter exceeding $(2k - 3)(k - 1)$. Alternatively, one may use a gamma prior on $\mathcal{K}$ with inverse-scale increasing with n at a suitable rate and we have consistency using Theorems 3 and 4.

The complex Watson kernel is a special case of the *complex Bingham* kernel which has density proportional to $\exp(z^* A z)$ with respect to the volume form. This kernel has location corresponding to the shape of a eigen-vector corresponding to the largest eigenvalue of A. Since it has more parameters, we expect better fit in smaller samples. We will prove that weak and strong posterior consistency holds while using this kernel in a later work.

When the landmarks are obtained from a 3D object, it is more appropriate to carry out an affine shape analysis, that is, identify two k-configurations as identical if they are related by an affine transformation. One can identify the resulting shape space with the Grassmannian manifold—the space of all 3D subspaces of $\Re^{k-1}$, a result of Sparr (1992). The Grassmannian is an extension of the real projective space and hence one

may consider (real) Bingham kernels and construct kernel mixture density models on this space.

5 Summary and future work

We consider kernel mixture density models on general compact manifolds and obtain sufficient conditions on the kernel, priors and the space for the density estimate to be strongly consistent. Thereby we extend the existing literature on strong posterior consistency on $\Re^d$ using Gaussian kernels to more general non-Euclidean manifolds. The conditions are verified for specific kernels on two important manifolds, namely the hypersphere and the planar shape space. It is discussed how to extend the kernel choice on these manifolds and construct their counterparts on other manifolds arising as generalizations. The multivariate Gaussian mixture model with an appropriate truncated and transformed Wishart prior on the within cluster covariance inverse is also shown to satisfy the consistency conditions when used to model a compactly supported density on $\Re^d$. We also allow the prior to depend on the sample size and obtain sufficient conditions for weak and strong consistency, while expecting better small sample operating characteristics. As a result a truncated Wishart prior on the covariance inverse of a multivariate Gaussian kernel is shown to satisfy the requirements for strong consistency.

In later works we plan to prove the results for other kernels on additional manifolds arising in applications. We also plan to extend the results to cover densities with non-compact support, in particular $\Re^d$. Since most of the non-Euclidean manifolds arising in applications are compact, that is not a high priority.

6 Appendix

6.1 Proof of Theorem 1

The proof runs on the lines of that of Theorem 1 in Bhattacharya and Dunson (2010a).

Proof First of all we show that the set

$$\left\{ P \in \mathcal{M}(M) : \sup_{m \in M, \mathcal{K} \in N_\epsilon} |f(m; P, \mathcal{K}) - f(m; F_0, \mathcal{K})| < \epsilon \right\} \tag{10}$$

contains a weakly open neighborhood of F_0, F_0 being the distribution corresponding to f_0. The kernel K being continuous from assumption **A1**, for any $(m, \mathcal{K}) \in M \times N_\epsilon$,

$$\mathcal{W}_{m,\mathcal{K}} = \left\{ P : |f(m; P, \mathcal{K}) - f(m; F_0, \mathcal{K})| < \epsilon/3 \right\}$$

defines an open neighborhood of F_0. The mapping from $(m, \mathcal{K})$ to $f(m; P, \mathcal{K})$ is a uniformly equicontinuous family of functions on $M \times N_\epsilon$, labeled by $P \in \mathcal{M}(M)$, because, for $m_1, m_2 \in M; \mathcal{K}_1, \mathcal{K}_2 \in N_\epsilon$,

$$|f(m_1; P, \mathcal{K}_1) - f(m_2; P, \mathcal{K}_2)| \leq \int_M |K(m_1; \mu, \mathcal{K}_1) - K(m_2; \mu, \mathcal{K}_2)| P(\mathrm{d}\mu)$$

and K is uniformly continuous on $M \times M \times N_\epsilon$. Therefore, there exists a $\delta > 0$ such that $\rho_{12}((m_1, \mathcal{K}_1), (m_2, \mathcal{K}_2)) < \delta$ implies that

$$\sup_P |f(m_1; P, \mathcal{K}_1) - f(m_2; P, \mathcal{K}_2)| < \epsilon/3.$$

Here ρ_{12} denotes any distance on $M \times N$ inducing the product topology. Cover $M \times N_\epsilon$ by finitely many balls of radius δ: $M \times N_\epsilon = \bigcup_{i=1}^J B\{(m_i, \mathcal{K}_i), \delta\}$. Let $\mathcal{W}_\epsilon = \bigcap_{i=1}^J \mathcal{W}_{m_i, \mathcal{K}_i}$ which is an open neighborhood of F_0. Let $P \in \mathcal{W}_\epsilon$ and $(m, \mathcal{K}) \in M \times N_\epsilon$. Then there exists a $(m_i, \mathcal{K}_i)$ such that $(m, \mathcal{K}) \in B\{(m_i, \mathcal{K}_i), \delta\}$. Then

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

This proves that the set in (10) contains $\mathcal{W}_\epsilon$, an open neighborhood of F_0.

$$\begin{aligned} &\text{For } P \in \mathcal{W}_\epsilon \quad \text{and} \quad \mathcal{K} \in N_\epsilon, \quad \sup_m |f_0(m) - f(m; P, \mathcal{K})| \\ &\leq \sup_m |f_0(m) - f(m; F_0, \mathcal{K})| + \sup_m |f(m; F_0, \mathcal{K}) - f(m; P, \mathcal{K})| < 2\epsilon \end{aligned}$$

because of assumptions **A2** and **A4**. Hence

$$\Pi\left\{f : \sup_{m \in M} |f(m) - f_0(m)| < 2\epsilon\right\} \geq \Pi_1(\mathcal{W}_\epsilon \times N_\epsilon) > 0$$

because of assumption **A3** and the fact that $\mathrm{int}(\mathcal{W}_\epsilon \times N_\epsilon)$ interects with $\mathrm{supp}(\Pi_1)$. This implies the KL property when f_0 is strictly positive (and hence bounded below), as shown in Corollary 1 of Bhattacharya and Dunson (2010a).

In case f_0 is not bounded below, we use Lemma 4 in Wu and Ghosal (2008) to get a continuous everywhere positive density f_1 (depending on f_0 and ϵ) for which $\Pi(K_\epsilon(f_1)) \leq \Pi(K_{2\epsilon+\sqrt{\epsilon}}(f_0))$. From what we have proved above, $\Pi(K_\epsilon(f_1)) > 0$ and as a result the KL condition follows for f_0. □

6.2 Proof of Theorem 2

In this proof and the subsequent ones, we shall use a general symbol C for any constant not depending on n (but possibly on ϵ).

Proof Given $\delta_1 > 0$ ($\equiv \delta_1(\epsilon, n)$), cover M by N_1 ($\equiv N_1(\delta_1)$) many disjoint subsets of diameter at most δ_1: $M = \cup_{i=1}^{N_1} E_i$. Assumption **A8** implies that for δ_1 sufficiently

small, $N_1 \leq C\delta_1^{-a_3}$. Pick $\mu_i \in E_i, i = 1, \ldots, N_1$ and define for a probability P,

$$P_n = \sum_{i=1}^{N_1} P(E_i)\delta_{\mu_i}, \;\; P_n(\mathbf{E}) = (P(E_1), \ldots, P(E_{N_1}))^T. \tag{11}$$

Denoting the L_1-norm as $\|.\|$, for any $\mathcal{K}$ with $\phi(\mathcal{K}) \leq \kappa_n$,

$$\begin{aligned} \|f(P, \mathcal{K}) - f(P_n, \mathcal{K})\| &\leq \sum_{i=1}^{N_1} \int_{E_i} \|K(\mu, \mathcal{K}) - K(\mu_i, \mathcal{K})\| P(\mathrm{d}\mu) \\ &\leq C \sum_i \int_{E_i} \sup_{m \in M} |K(m; \mu, \mathcal{K}) - K(m; \mu_i, \mathcal{K})| P(\mathrm{d}\mu) \qquad (12) \\ &\leq C\kappa_n^{a_1}\delta_1. \qquad (13) \end{aligned}$$

The inequality in (13) follows from (12) using Assumption **A5**.

For $\mathcal{K}, \tilde{\mathcal{K}}$ in $\phi^{-1}[0, \kappa_n]$, $P \in \mathcal{M}(M)$,

$$\begin{aligned} \|f(P, \mathcal{K}) - f(P, \tilde{\mathcal{K}})\| &\leq C \sup_{m,\mu \in M} |K(m; \mu, \mathcal{K}) - K(m; \mu, \tilde{\mathcal{K}})| \\ &\leq C\kappa_n^{a_2}\rho_2(\mathcal{K}, \tilde{\mathcal{K}}), \qquad (14) \end{aligned}$$

the inequality in (14) following from Assumption **A6**. Given $\delta_2 > 0$ ($\equiv \delta_2(\epsilon, n)$), cover $\phi^{-1}[0, \kappa_n]$ by finitely many subsets of diameter at most δ_2, the number of such subsets required being at most $C(\kappa_n\delta_2^{-1})^{b_2}$, from Assumption **A7**. Call the collection of these subsets $W(\delta_2, n)$.

Letting $S_d = \{x \in [0, 1]^d : \sum x_i \leq 1\}$, S_d is compact under the L^1-metric ($\|x\|_{L^1} = \sum |x_i|, x \in \Re^d$), and hence given any $\delta_3 > 0$ ($\equiv \delta_3(\epsilon)$), can be covered by finitely many subsets of the cube $[0, 1]^d$ each of diameter at most δ_3. In particular, cover S_d with cubes of side length δ_3/d lying partially or totally in S_d. Then an upper bound on the number $N_2 \equiv N_2(\delta_3, d)$ of such cubes can be shown to be $\frac{\lambda(S_d(1+\delta_3))}{(\delta_3/d)^d}$, λ denoting the Lebesgue measure on $\Re^d$ and $S_d(r) = \{x \in [0, \infty)^d : \sum x_i \leq r\}$. Since $\lambda(S_d(r)) = r^d/d!$; hence

$$N_2(\delta_3, d) \leq \frac{d^d}{d!}\left(\frac{1 + \delta_3}{\delta_3}\right)^d.$$

Let $\mathcal{W}(\delta_3, d)$ denote the partition of S_d as constructed above.

Let $d_n = N_1(\delta_1)$. For $1 \leq i \leq N_2(\delta_3, d_n)$, $1 \leq j \leq C(\kappa_n\delta_2^{-1})^{b_2}$, define

$$\mathcal{D}_{ij} = \{f(P, \mathcal{K}) : P_n(\mathbf{E}) \in \mathcal{W}_i, \; \mathcal{K} \in W_j\},$$

with $\mathcal{W}_i$ and W_j being elements of $\mathcal{W}(\delta_3, d_n)$ and $W(\delta_2, n)$, respectively. We claim that this subset of $\mathcal{D}_n$ has L^1 diameter of at most ϵ. For $f(P, \mathcal{K})$, $f(\tilde{P}, \tilde{\mathcal{K}})$ in this set,

$$\begin{aligned}&\|f(P,\mathcal{K})-f(\tilde{P},\tilde{\mathcal{K}})\|\\&\leq \|f(P,\mathcal{K})-f(P_n,\mathcal{K})\|+\|f(P_n,\mathcal{K})-f(\tilde{P}_n,\mathcal{K})\|\\&\quad+\|f(\tilde{P}_n,\mathcal{K})-f(\tilde{P},\mathcal{K})\|+\|f(\tilde{P},\mathcal{K})-f(\tilde{P},\tilde{\mathcal{K}})\|. \qquad (15)\end{aligned}$$

From inequality (13), it follows that the first and third terms in (15) are at most $C\kappa_n^{a_1}\delta_1$. The second term can be bounded by

$$\sum_{i=1}^{d_n}|P(E_i)-\tilde{P}(E_i)|<\delta_3$$

and from the inequality in (14), the fourth term is bounded by $C\kappa_n^{a_2}\delta_2$. Hence the claim holds if we choose $\delta_1 = C\kappa_n^{-a_1}$, $\delta_2 = C\kappa_n^{-a_2}$, and $\delta_3 = C$. The number of such subsets covering $\mathcal{D}_n$ is at most $CN_2(\delta_3, d_n)(\kappa_n\delta_2^{-1})^{b_2}$. From Assumption **A8**, it follows that for n sufficiently large,

$$d_n = N_1(\delta_1) \leq C\kappa_n^{a_1a_3}.$$

Using the Stirling's formula, we can bound $\log(N_2(\delta_3, d_n))$ by Cd_n. Also $\kappa_n\delta_2^{-1}$ is bounded by $C\kappa_n^{a_2+1}$, so that

$$N(\epsilon,\mathcal{D}_n)\leq C+C\log(\kappa_n)+Cd_n\leq C\kappa_n^{a_1a_3}$$

for n sufficiently large. □

6.3 Proof of Proposition 3

Proof In Lemma 5 of Wu and Ghosal (2010), it is shown that

$$\|K(\mu,\mathcal{K})-K(\nu,\mathcal{K})\|\leq C\lambda_d(\mathcal{K})^{1/2}\rho(\mu,\nu)$$

for all $\mu, \nu \in \mathbb{R}^d$ and $\mathcal{K} \in M^+(d)$. This means that **A5**′ is satisfied with $a_1 = 1/2$. Also by the geometry of $\Re^d$, **A8** is satisfied with $a_3 = d$.

To show **A7**, note that $\phi^{-1}[0,\kappa]$ is a subset of $M^+(d)$ consisting of those positive matrices whose eigenvalues are bounded by κ. We equip $M^+(d)$ with the L_2 norm distance, i.e.,

$$\rho_2(\mathcal{K}_1,\mathcal{K}_2)=\|\mathcal{K}_1-\mathcal{K}_2\|_2,\ \ \|\mathcal{K}\|_2^2=\sum_{ij}\mathcal{K}_{ij}^2=\mathrm{Trace}(\mathcal{K}^2)$$

and view it as a subset of $M(d)$—the space of all order d real matrices. Then $\phi^{-1}[0,\kappa]$ is contained in a ball of radius $\sqrt{d\kappa}$ around the zero matrix. The ϵ-covering number of a such a ball is of the order $(\sqrt{\kappa}/\epsilon)^{d^2}$. Hence **A7** is also satisfied.

Remains to check **A6**′. Since $\phi^{-1}[0,\kappa]$ is a convex subset of $M(d)$, use the Taylor's theorem to get

$$|K(x;\mu,\mathcal{K}_1) - K(x;\mu,\mathcal{K}_2)| \le \rho_2(\mathcal{K}_1,\mathcal{K}_2) \sup_{\mathcal{K}\in\phi^{-1}[0,\kappa]} \left\| \frac{\partial}{\partial \mathcal{K}} K(x;\mu,\mathcal{K}) \right\|_2$$

for all $x, \mu \in \Re^d$, and $\mathcal{K}_1, \mathcal{K}_2 \in \phi^{-1}[0,\kappa]$. This in turn implies that

$$\|K(\mu,\mathcal{K}_1) - K(\mu,\mathcal{K}_2)\| \le \rho_2(\mathcal{K}_1,\mathcal{K}_2) \int_{\Re^d} \sup_{\mathcal{K}\in\phi^{-1}[0,\kappa]} \left\| \frac{\partial}{\partial \mathcal{K}} K(x;\mu,\mathcal{K}) \right\|_2 \mathrm{d}x.$$

Some calculation will show that

$$\left\| \frac{\partial}{\partial \mathcal{K}} K(x;\mu,\mathcal{K}) \right\|_2 \le C K(x;\mu,\mathcal{K}) \left(\sqrt{\sum_1^d \lambda_j^{-2}(\mathcal{K}) + \|x-\mu\|_2^2} \right),$$

C being some constant independent of x, μ or $\mathcal{K}$. Since $\phi^{-1}[0,\kappa]$ consists of all positive matrices $\mathcal{K}$ whose eigenvalues lie in $[\lambda_1, \kappa]$, this will imply that

$$\|K(\mu,\mathcal{K}_1) - K(\mu,\mathcal{K}_2)\| \le C\kappa^{d/2} \rho_2(\mathcal{K}_1,\mathcal{K}_2)$$

which means that **A6**′ is also satisfied with $a_2 = d/2$. Here C denotes a constant independent of $\mu, \mathcal{K}_1, \mathcal{K}_2$ or κ. The rest of the proof follows from Theorem 2 and Proposition 2. □

6.4 Proof of Lemma 1

Proof Fix $\epsilon > 0$. Under assumptions **A1**, **A2** and **A4**, it follows from the proof of Theorem 1 that there exists a weakly open neighborhood $\mathcal{W}$ of F_0 (depending on ϵ) such that $K_\epsilon(f_0)$ contains $\{f(P,\mathcal{K}) : P \in \mathcal{W}, \mathcal{K} \in N_\epsilon\}$. Hence

$$\begin{aligned} \int \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi_n(\mathrm{d}f) &\ge \int_{K_\epsilon(f_0)} \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi_n(\mathrm{d}f) \\ &\ge \int_{\mathcal{W}\times N_\epsilon} \prod_1^n \frac{f(X_i;P,\mathcal{K})}{f_0(X_i)} \pi_n(\mathcal{K}) \Pi_{11}(\mathrm{d}P) \lambda_1(\mathrm{d}\mathcal{K}). \end{aligned}$$

By the law of large numbers, for any $f \in K_\epsilon(f_0)$,

$$\frac{1}{n} \sum_i \log\{(f_0/f)(X_i)\} \to \mathrm{KL}(f_0; f) < \epsilon$$

a.s. F_0^∞ as $n \to \infty$. Therefore, for any $P \in \mathcal{W}$ and $\mathcal{K} \in N_\epsilon$,

$$\liminf_n \exp(2n\epsilon) \prod_1^n \frac{f(X_i; P, \mathcal{K})}{f_0(X_i)}$$
$$= \liminf_n \exp\left[n \left[2\epsilon - (1/n) \sum_i \log\{f_0(X_i)/f(X_i; P, \mathcal{K})\} \right] \right] = \infty \text{ a.s. } F_0^\infty.$$

Also from Assumption **A11**, $\liminf_n \exp(n\epsilon)\pi_n(\mathcal{K}) = \infty \forall \mathcal{K} \in N_\epsilon$ and hence

$$\liminf_n \exp(3n\epsilon) \prod_1^n \frac{f(X_i; P, \mathcal{K})}{f_0(X_i)} \pi_n(\mathcal{K}) = \infty \text{ a.s. } F_0^\infty.$$

By Fubini–Tonelli theorem, there exists a $\Omega_0 \subset \Omega$ with probability 1 such that for any $\omega \in \Omega_0$,

$$\liminf_n \exp(3n\epsilon) \prod_1^n \frac{f(X_i(\omega); P, \mathcal{K})}{f_0(X_i(\omega))} \pi_n(\mathcal{K}) = \infty$$

for all $(P, \mathcal{K}) \in \mathcal{W} \times N_\epsilon$ outside of a $\Pi_{11}(\mathrm{d}P) \otimes \lambda_1(\mathrm{d}\mathcal{K})$ measure 0 subset. By Assumption **A10** and since λ_1 has full support and N_ϵ has a non-empty interior, $(\Pi_{11} \otimes \lambda_1)(\mathcal{W} \times N_\epsilon) > 0$. Therefore, using the Fatou's lemma, we conclude that

$$\liminf_n \exp(3n\epsilon) \int \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi_n(\mathrm{d}f)$$
$$\geq \int_{\mathcal{W} \times N_\epsilon} \liminf_n \left\{ \exp(3n\epsilon) \prod_1^n \frac{f(X_i; P, \mathcal{K})}{f_0(X_i)} \pi_n(\mathcal{K}) \right\}$$
$$\times \Pi_{11}(\mathrm{d}P)\lambda_1(\mathrm{d}\mathcal{K}) = \infty \text{ a.s. } F_0^\infty.$$

Since ϵ was arbitrary, the proof is completed. □

6.5 Proof of Proposition 4

Proof For any $a > 0$,

$$\begin{aligned} \pi_n\{\phi^{-1}(n^a, \infty)\} &= Pr(\lambda_d(\mathcal{K}) > n^a),\ \mathcal{K} \sim \pi_n \\ &= Pr(\lambda_d(X) > n^a | \lambda_1(X) \geq \lambda_1),\ X \sim \text{Wish}(\beta_n^{-1} I_d, q) \\ &\leq Pr(\lambda_d(X) > n^a)/Pr(\lambda_1(X) \geq \lambda_1) \\ &= Pr(\lambda_d(Z) > \beta_n n^a)/Pr(\lambda_1(Z) \geq \beta_n \lambda_1),\ Z \sim \text{Wish}(I_d, q), \end{aligned} \tag{16}$$

the last identity following because X equals to $\beta_n^{-1}Z$ in distribution. The numerator in (16) is less than $Pr(\mathrm{Tr}(Z) > \beta_n n^a)$. The trace of Z follows a $\mathrm{Gam}(1/2, qd/2)$ distribution which has exponentially decaying tail. Hence the numerator is less than $\exp(-C\beta_n n^a)$ for some $C > 0$ when n is sufficiently large.

Now we derive a lower bound for the probability in the denominator of (16). In Mallik (2003), the joint density of $\lambda_1(Z), \ldots, \lambda_d(Z)$ has been shown to be

$$f(x_1, \ldots, x_d) = \frac{\left(\prod_{i=1}^d x_i\right)^{q-d} \exp\left(-\sum_{i=1}^d x_i\right) \prod_{1\le i<j\le d}(x_i - x_j)^2}{\prod_{i=1}^d (d-i)!(q-i)!},$$
$$0 < x_1 < \cdots < x_d < \infty.$$

Hence $Pr(\lambda_1(Z) \ge \beta_n\lambda_1) = Pr(\lambda_j(Z) \ge \beta_n\lambda_1 \ \forall j)$

$$= C\int_{\beta_n\lambda_1\le x_1<\cdots<x_d<\infty} \left(\prod_{i=1}^d x_i\right)^{q-d} \exp\left(-\sum_{i=1}^d x_i\right) \prod_{1\le i<j\le d}(x_i - x_j)^2 \prod_{i=1}^d \mathrm{d}x_i$$

$$\ge C\int_{\beta_n\lambda_1\le x_1<\cdots<x_d<\infty} \exp\left(-2\sum_{i=1}^d x_i\right) \prod_{1\le i<j\le d}(x_i - x_j)^2 \prod_{i=1}^d \mathrm{d}x_i \tag{17}$$

for n sufficiently large. Integrate (17) by parts to get $Pr(\lambda_1(Z) \ge \beta_n\lambda_1) \ge \exp(-C\beta_n)$ for appropriate $C > 0$ when n is sufficiently large. Hence there exists a $C > 0$ such that the ratio in (16) is less than $\exp(-C\beta_n n^a)$. If we pick a as in the Proposition, for any $\beta_0 > 0$, it follows that

$$\exp(n\beta_0)\pi_n\{\phi^{-1}(n^a, \infty)\} < \exp\{-n(C\beta_n n^{a-1} - \beta_0)\}$$

which converges to zero because $\beta_n n^{a-1}$ diverges to infinity. This verifies Assumption **A12** with a as in the Proposition and β_0 being any positive constant. □

6.6 Proof of Theorem 5

Proof Denote by M the unit sphere S^d and by ρ the chord distance on it. Express the vMF kernel as

$$K(m; \mu, \mathcal{K}) = c^{-1}(\mathcal{K}) \exp[\mathcal{K}\{1 - \rho^2(m, \mu)/2\}] \quad (m, \mu \in M; \mathcal{K} \in [0, \infty)).$$

Since ρ is continuous on the product space $M \times M$ and c is continuous and non-vanishing on $[0, \infty)$, K is continuous on $M \times M \times [0, \infty)$ and assumption **A1** follows:

For a given continuous function f on M, $m \in M$, $\mathcal{K} \ge 0$, define

$$I(m, \mathcal{K}) = f(m) - \int_M K(m; \mu, \mathcal{K}) f(\mu) V(\mathrm{d}\mu)$$
$$= \int_M K(m; \mu, \mathcal{K})\{f(m) - f(\mu)\} V(\mathrm{d}\mu).$$

Then assumption **A2** follows once we show that

$$\lim_{\mathcal{K}\to\infty} (\sup_{m\in M} |I(m,\mathcal{K})|) = 0.$$

To simplify $I(m,\mathcal{K})$, make a change of coordinates $\mu \mapsto \tilde{\mu} = U(m)^T\mu$, $\tilde{\mu} \mapsto \theta \in \Theta_d \equiv (0,\pi)^{d-1} \times (0,2\pi)$ where $U(m)$ is an orthogonal matrix with first column equal to m and $\theta = (\theta_1,\ldots,\theta_d)^T$ are the spherical coordinates of $\tilde{\mu} \equiv \tilde{\mu}(\theta)$ which are given by

$$\tilde{\mu}_j = \cos\theta_j \prod_{h<j} \sin\theta_h, \;\; j = 1,\ldots,d, \quad \tilde{\mu}_{d+1} = \prod_{j=1}^{d} \sin\theta_j.$$

Using these coordinates, the volume form can be written as

$$V(\mathrm{d}\mu) = V(\mathrm{d}\tilde{\mu}) = \sin^{d-1}(\theta_1)\sin^{d-2}(\theta_2)\cdots\sin(\theta_{d-1})\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d$$

and hence $I(m,\mathcal{K})$ equals

$$\begin{aligned}
&\int_{\Theta_d} c^{-1}(\mathcal{K})\exp\{\mathcal{K}\cos(\theta_1)\}\{f(m) - f(U(m)\tilde{\mu})\}\sin^{d-1}(\theta_1)\cdots\sin(\theta_{d-1})\mathrm{d}\theta_1\cdots\mathrm{d}\theta_d \\
&\quad = c^{-1}(\mathcal{K})\int_{\Theta_{d-1}\times(-1,1)} \exp(\mathcal{K}t)\{f(m) - f(U(m)\tilde{\mu})\}(1-t^2)^{d/2-1} \\
&\qquad \times \sin^{d-2}(\theta_2)\cdots\sin(\theta_{d-1})\mathrm{d}\theta_2\cdots\mathrm{d}\theta_d\mathrm{d}t \qquad (18)
\end{aligned}$$

where $t = \cos(\theta_1)$, $\tilde{\mu} = \tilde{\mu}(\theta(t))$ and $\theta(t) = (\arccos(t),\theta_2,\ldots,\theta_d)^T$. In the integrand in (18), the distance between m and $U(m)\tilde{\mu}$ is $\sqrt{2(1-t)}$. Substitute $t = 1 - \mathcal{K}^{-1}s$ in the integral with $s \in (0,2\mathcal{K})$. Define

$$\Phi(s,\mathcal{K}) = \sup\{|f(m) - f(\tilde{m})| : \; m,\tilde{m} \in M, \; \rho(m,\tilde{m}) \le \sqrt{2\mathcal{K}^{-1}s}\}.$$

Then

$$|f(m) - f(U(m)\tilde{\mu})| \le \Phi(s,\mathcal{K}).$$

Since f is uniformly continuous on (M,ρ), Φ is bounded on $(\Re^+)^2$ and $\lim_{\mathcal{K}\to\infty}\Phi(s,\mathcal{K}) = 0$. Hence from (18), we deduce that

$$\begin{aligned}
&\sup_{m\in M}|I(m,\mathcal{K})| \\
&\quad \le c^{-1}(\mathcal{K})\mathcal{K}^{-1}\int_{\Theta_{d-1}\times(0,2\mathcal{K})} \exp(\mathcal{K}-s)\Phi(s,\mathcal{K})(\mathcal{K}^{-1}s(2-\mathcal{K}^{-1}s))^{d/2-1} \\
&\qquad \times \sin^{d-2}(\theta_2)\cdots\sin(\theta_{d-1})\mathrm{d}\theta_2\cdots\mathrm{d}\theta_d\mathrm{d}s \\
&\quad \le C\mathcal{K}^{-d/2}\tilde{c}^{-1}(\mathcal{K})\int_0^\infty \Phi(s,\mathcal{K})e^{-s}s^{d/2-1}\mathrm{d}s. \qquad (19)
\end{aligned}$$

From Lemma 3, it follows that

$$\limsup_{\mathcal{K}\to\infty} \mathcal{K}^{-d/2}\tilde{c}^{-1}(\mathcal{K}) < \infty.$$

This in turn, using the Lebesgue Dominated Convergence Theorem implies that the expression in (19) converges to 0 as $\mathcal{K} \to \infty$. This verifies assumption **A2**. □

6.7 Proof of Theorem 6

In the proof, $B_d(r)$ denotes the ball of radius r around 0 in $\Re^d$:

$$B_d(r) = \{x \in \Re^d : \|x\|_2 \le r\}$$

and B_d refers to $B_d(1)$.

Proof It is clear from (6) and (7) that the vMF kernel K is continuously differentiable on $\Re^{d+1} \times \Re^{d+1} \times [0, \infty)$. Hence

$$\sup_{m\in S^d, \mathcal{K}\in[0,\kappa]} \left|K(m;\mu,\mathcal{K}) - K(m;\nu,\mathcal{K})\right| \le \sup_{m\in S^d, x\in B_{d+1}, \mathcal{K}\in[0,\kappa]} \left\|\frac{\partial}{\partial x} K(m;x,\mathcal{K})\right\|_2 \|\mu-\nu\|_2.$$

Since

$$\frac{\partial}{\partial x} K(m;x,\mathcal{K}) = \mathcal{K}\tilde{c}^{-1}(\mathcal{K}) \exp\{-\mathcal{K}(1-m^T x)\}m,$$

its norm is bounded by $\mathcal{K}\tilde{c}^{-1}(\mathcal{K})$. Lemma 3 implies that this in turn is bounded by

$$\kappa\tilde{c}^{-1}(\kappa) \le C\kappa^{d/2+1}$$

for $\mathcal{K} \le \kappa$ and $\mathcal{K} \ge 1$. This proves assumption **A5** with $a_1 = d/2 + 1$.

To verify **A6**, given $\mathcal{K}_1, \mathcal{K}_2 \le \kappa$, use the inequality,

$$\sup_{m,\mu\in S^d} \left|K(m;\mu,\mathcal{K}_1) - K(m;\mu,\mathcal{K}_2)\right| \le \sup_{m,\mu\in S^d, \mathcal{K}\le\kappa} \left|\frac{\partial}{\partial\kappa} K(m;\mu,\mathcal{K})\right| |\mathcal{K}_1 - \mathcal{K}_2|.$$

By direct computations, one can show that

$$\frac{\partial}{\partial\mathcal{K}} K(m;\mu,\mathcal{K}) = -\frac{\partial}{\partial\mathcal{K}}\tilde{c}(\mathcal{K})\tilde{c}^{-2}(\mathcal{K}) \exp\{-\mathcal{K}(1-m^T\mu)\} - \tilde{c}^{-1}(\mathcal{K}) \exp\{-\mathcal{K}(1-m^T\mu)\}(1-m^T\mu),$$

$$\frac{\partial}{\partial \mathcal{K}}\tilde{c}(\mathcal{K}) = -C\int_{-1}^{1}\exp\{-\mathcal{K}(1-t)\}(1-t)(1-t^2)^{d/2-1}\mathrm{d}t,$$
$$\left|\frac{\partial}{\partial \mathcal{K}}\tilde{c}(\mathcal{K})\right| \leq C\tilde{c}(\mathcal{K}).$$

Therefore, using Lemma 3,

$$\left|\frac{\partial}{\partial \mathcal{K}}K(m;\mu,\mathcal{K})\right| \leq C\tilde{c}^{-1}(\mathcal{K}) \leq C\tilde{c}^{-1}(\kappa) \leq C\kappa^{d/2}$$

for any $\mathcal{K} \leq \kappa$ and $\kappa \geq 1$. Hence **A6** is verified with $a_2 = d/2$.

Finally to verify **A8**, note that $S^d \subset B_{d+1} \subset [-1,1]^{d+1}$ which can be covered by finitely many cubes of side length $\epsilon/(d+1)$. Each such cube has L_2 diameter ϵ. Hence their intersections with S^d provides a finite ϵ-cover for this manifold. If $\epsilon < 1$, such a cube intersects with S^d only if it lies entirely in $B_{d+1}(1+\epsilon) \cap B_{d+1}(1-\epsilon)^c$. The number of such cubes, and hence the ϵ-cover size can be bounded by

$$C\epsilon^{-(d+1)}\{(1+\epsilon)^{d+1} - (1-\epsilon)^{d+1}\} \leq C\epsilon^{-d}$$

for some $C > 0$ not depending on ϵ. This verifies **A8** for appropriate positive constant A_3 and $a_3 = d$. □

6.8 Proof of Lemma 3

Proof Express $\tilde{c}(\mathcal{K})$ as

$$C\int_{-1}^{1}\exp\{-\mathcal{K}(1-t)\}(1-t^2)^{d/2-1}\mathrm{d}t$$

and it is clear that it is decreasing. This expression suggests that

$$\begin{aligned}
\tilde{c}(\mathcal{K}) &\geq C\int_{0}^{1}\exp\{-\mathcal{K}(1-t)\}(1-t^2)^{d/2-1}\mathrm{d}t \\
&\geq C\int_{0}^{1}\exp\{-\mathcal{K}(1-t^2)\}(1-t^2)^{d/2-1}\mathrm{d}t \\
&= C\int_{0}^{1}\exp(-\mathcal{K}u)u^{d/2-1}(1-u)^{-1/2}\mathrm{d}u \\
&\geq C\int_{0}^{1}\exp(-\mathcal{K}u)u^{d/2-1}\mathrm{d}u
\end{aligned}$$

$$= C\mathcal{K}^{-d/2}\int_0^{\mathcal{K}} \exp(-v)v^{d/2-1}\mathrm{d}v$$
$$\geq C\left\{\int_0^1 \exp(-v)v^{d/2-1}\mathrm{d}v\right\}\mathcal{K}^{-d/2}$$

if $\mathcal{K} \geq 1$. □

6.9 Proof of Theorem 7

Proof Express the complex Watson kernel as

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

Given $\mathcal{K} \geq 0$, define

$$\phi(t) = \exp\left(\frac{-\mathcal{K}}{2}t^2\right),\ t \in [0,\sqrt{2}].$$

Then $|\phi^{'}(t)| \leq \sqrt{2}\mathcal{K}$, so that

$$|\phi(t)-\phi(s)| \leq \sqrt{2}\mathcal{K}|s-t|,\ s,t \in [0,\sqrt{2}]$$

which implies that

$$\begin{aligned}|K(m;\mu,\mathcal{K}) - K(m;\nu,\mathcal{K})| &\leq c^{-1}(\mathcal{K})\sqrt{2}\mathcal{K}|d_E(m,\mu)-d_E(m,\nu)| \\ &\leq \sqrt{2}\mathcal{K}c^{-1}(\mathcal{K})d_E(\mu,\nu). \end{aligned} \tag{20}$$

For $\mathcal{K} \leq \kappa$, from Lemma 4, it follows that

$$\begin{aligned}\mathcal{K}c^{-1}(\mathcal{K}) \leq \kappa c^{-1}(\kappa) &= \pi^{2-k}\kappa^{k-1}\tilde{c}^{-1}(\kappa) \\ &\leq \pi^{2-k}\kappa^{k-1}\tilde{c}^{-1}(1)\end{aligned}$$

provided $\kappa \geq 1$. Hence for any $\kappa \geq 1$,

$$\sup_{\mathcal{K}\in[0,\kappa]} \mathcal{K}c^{-1}(\mathcal{K}) \leq C\kappa^{k-1}$$

and from inequality (20), $a_1 = k-1$ follows.

By direct computation, one can show that

$$\begin{aligned}\frac{\partial}{\partial\mathcal{K}}K(m;\mu,\mathcal{K}) = \pi^{k-2}\exp\left\{-\frac{1}{2}\mathcal{K}d_E^2(m,\mu)-\mathcal{K}\right\} \\ \times c^{-2}(\mathcal{K})\mathcal{K}^{2-k}\left[\sum_{r=k-1}^{\infty}\frac{\mathcal{K}^{r-1}}{r!}\left\{k-2-\frac{r}{2}d_E^2(m,\mu)\right\}\right].\end{aligned} \tag{21}$$

Denote by S the sum in the second line of (21). Since $d_E^2(m,\mu) \le 2$, it can be shown that

$$|k-2-\frac{r}{2}d_E^2(m,\mu)| \le \begin{cases} k-2 & \text{if } k-1 \le r \le 2k-4, \\ r-k+2 & \text{if } 2k-3 \le r, \end{cases}$$

so that

$$\begin{aligned} |S| &\le (k-2)\sum_{r=k-1}^{2k-4}\frac{\mathcal{K}^{r-1}}{r!} + \sum_{r=2k-3}^{\infty}\frac{\mathcal{K}^{r-1}}{r!}(r-k+2) \\ &= (k-2)\mathcal{K}^{k-2}\sum_{r=0}^{k-3}\frac{\mathcal{K}^r}{(r+k-1)!} + \mathcal{K}^{2k-4}\sum_{r=0}^{\infty}\frac{\mathcal{K}^r}{(r+2k-3)!}(r+k-1) \\ &\le C\mathcal{K}^{k-2}e^{\mathcal{K}} + \mathcal{K}^{2k-4}e^{\mathcal{K}}. \end{aligned}$$

Plug the above inequality in (21) to get

$$\begin{aligned} \left|\frac{\partial}{\partial\mathcal{K}}K(m;\mu,\mathcal{K})\right| &\le Cc^{-2}(\mathcal{K})\mathcal{K}^{2-k}\exp\left\{-\frac{1}{2}\mathcal{K}d_E^2(m,\mu)\right\}(C\mathcal{K}^{k-2}+\mathcal{K}^{2k-4}) \\ &\le Cc^{-2}(\mathcal{K})(C+\mathcal{K}^{k-2}). \end{aligned} \tag{22}$$

For $\mathcal{K} \le \kappa$ and $\kappa \ge 1$, using Lemma 4, we bound the expression in (22) by

$$\begin{aligned} Cc^{-2}(\kappa)(C+\kappa^{k-2}) &= C\kappa^{2k-6}\tilde{c}^{-2}(\kappa)(C+\kappa^{k-2}) \\ &\le C\kappa^{2k-6}\tilde{c}^{-2}(1)(C+\kappa^{k-2}) \le C\kappa^{3k-8} \end{aligned} \tag{23}$$

for κ sufficiently large. Since K is a continuously differentiable in $\mathcal{K}$, from (23) it follows that there exists $\kappa_1 > 0$ such that for all $\kappa \ge \kappa_1, \mathcal{K}_1, \mathcal{K}_2 \le \kappa$,

$$\begin{aligned} \sup_{m,\mu\in\Sigma_2^k}|K(m;\mu,\mathcal{K}_1)-K(m;\mu,\mathcal{K}_2)| &\le \sup_{m,\mu\in\Sigma_2^k,\mathcal{K}\in[0,\kappa]}\left|\frac{\partial}{\partial\mathcal{K}}K(m;\mu,\mathcal{K})\right||\mathcal{K}_1-\mathcal{K}_2| \\ &\le C\kappa^{3k-8}|\mathcal{K}_1-\mathcal{K}_2|. \end{aligned}$$

This proves Assumption **A6** with $a_2 = 3k-8$. □

6.10 Proof of Theorem 8

In the proof, $C_i, i = 1, 2, \ldots$ denote positive constants possibly depending on k.

Proof The preshape sphere $\mathcal{CS}^{k-2}$, as a real manifold, can be identified with the real unit sphere S^{2k-3}. Endow it with the chord distance induced by the L^2-norm

$$\|u\|_2 = \sqrt{\sum_{i=1}^{k-1} |u_i|^2} \quad (u = (u_1, \ldots, u_{k-1})^T).$$

Then from Theorem 6, it follows that given any $\delta > 0, \mathcal{CS}^{k-2}$ can be covered by finitely many subsets of diameter less than or equal to δ, the number of such subsets being bounded by $C_1\delta^{-(2k-3)} + C_2$. The extrinsic distance d_E on Σ_2^k can be bounded by the chord distance on $\mathcal{CS}^{k-2}$ as follows: For $u, v \in \mathcal{CS}^{k-2}$,

$$\begin{aligned} \|u - v\|_2^2 &= 2 - 2\mathrm{Re}(u^* v) \geq 2 - 2|u^* v| = 2(1 - |u^* v|) \\ &\geq (1 + |u^* v|)(1 - |u^* v|) = \frac{1}{2} d_E^2([u], [v]). \end{aligned}$$

Hence $d_E([u], [v]) \leq \sqrt{2}\|u - v\|_2$, so that given any $\epsilon > 0$, the shape image of a δ-cover for $\mathcal{CS}^{k-2}$ with $\delta = \epsilon/\sqrt{2}$ provides an ϵ-cover for Σ_2^k. Hence the ϵ-covering size for Σ_2^k can be bounded by $C_1\epsilon^{-(2k-3)} + C_2$. □

References

Barron, A. R. (1989). Uniformly powerful goodness of fit tests. *Annals of Statistics, 17*, 107–124.

Bhattacharya, A., Dunson, D. (2010a). Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika, 97*(4), 851–865.

Bhattacharya, A., Dunson, D. (2010b). Nonparametric Bayes classification and hypothesis testing on manifolds. Discussion Paper, Department of Statistical Science, Duke University.

Bhattacharya, R. N., Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of Statistics, 31*, 1–29.

Dryden, I. L., Mardia, K. V. (1998). *Statistical Shape Analysis*. New York: Wiley.

Escobar, M. D., West, M. (1995). Bayesian density-estimation and inference using mixtures. *Journal of the American Statistical Association, 90*, 577–588.

Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London A, 1130*, 295–305.

Ghosal, S., Ghosh, J. K., Ramamoorthi, R. V. (1999). Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics, 27*, 143–158.

Ghosh, J. K., Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. New York: Springer.

Hirsch, M. (1976). *Differential Topology*. New York: Springer.

Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society, 16*, 81–121.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics, 1*, 38–53.

Lennox, K. P., Dahl, D. B., Vannucci, M., Tsai, J. W. (2009). Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *Journal of the American Statistical Association, 104*, 586–596.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. 1. density estimates. *Annals of Statistics, 12*, 351–357.

Mallik, R. K. (2003). The pseudo-wishart distribution and its application to mimo systems. *IEEE Transactions on Information Theory, 49(10)*, 2761–2769.

Schwartz, L. (1965) On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 4*, 10–26.

Sparr, G. (1992). Depth-computations from polihedral images. *Proceedings of 2nd European Conference on Computer Vision, ECCV-2*, 378–386.

von Mises, R. V. (1918). Uber die "Ganzzahligkeit" der Atomgewicht und verwandte Fragen. *Physik Z, 19*, 490–500.

Watson, G. S., Williams, E. J. (1953). Construction of significance tests on the circle and sphere. *Biometrika, 43*, 344–352.

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

Wu, Y., Ghosal, S. (2010). The L_1-consistency of dirichlet mixtures in multivariate bayesian density estimation on bayes procedures. *Journal of Mutivariate Analysis, 101*, 2411–2419.