NONPARAMETRIC BAYES INFERENCE ON MANIFOLDS WITH APPLICATIONS TO SPHERE AND PLANAR SHAPES

Abhishek Bhattacharya
Department of Statistics, Duke University
Joint work with Profs. R.Bhattacharya & D.Dunson

March 30 2010

CONTENTS I

CONTENTS II

OVERVIEW

- Present nonparametric Bayes inference results - both theoretical and applied, on general compact metric spaces, especially Riemannian manifolds.
- Focus on two particular manifolds - the unit hypersphere and landmark based shape spaces.
- Model the distributional density using random mixtures of appropriate kernels on the space.
- Prove consistency of the density estimate.
- Present algorithms to sample from the posterior distribution of the density.
- Perform np classification on manifolds and compare with standard methods.
- Perform Bayes hypothesis testing for discriminating between two or more populations.

Shapes of k-ads

- k points or landmarks are picked from an object or image in 2D or 3D called k-ad.
- In general, each observation $\mathbf{x} = (x_1, \ldots, x_k)$ consists of $k > m$ points in m-dimension (not all same).
- Shape of a k-ad is its orbit under a group G of transformations.

Different Notions of Shapes of k-ads I

- Kendall's (Direct) Similarity Shape Space Σ_m^k. Group G generated by translations, scaling and rotations.
- Planar Shape Space Σ_2^k.
- Reflection Similarity Shape Space $R\Sigma_m^k$. Remove the effects of translations, scaling and all orthogonal transformations.
- Similarity shape analysis finds applications in morphometrics - classification of biological species based on their shapes, medical diagnostics - disease detection based on change in shape of an organ due to disease or deformation, evolution studies - studying the change in shape of an organ or organism with time, age etc.

Different Notions of Shapes of k-ads II

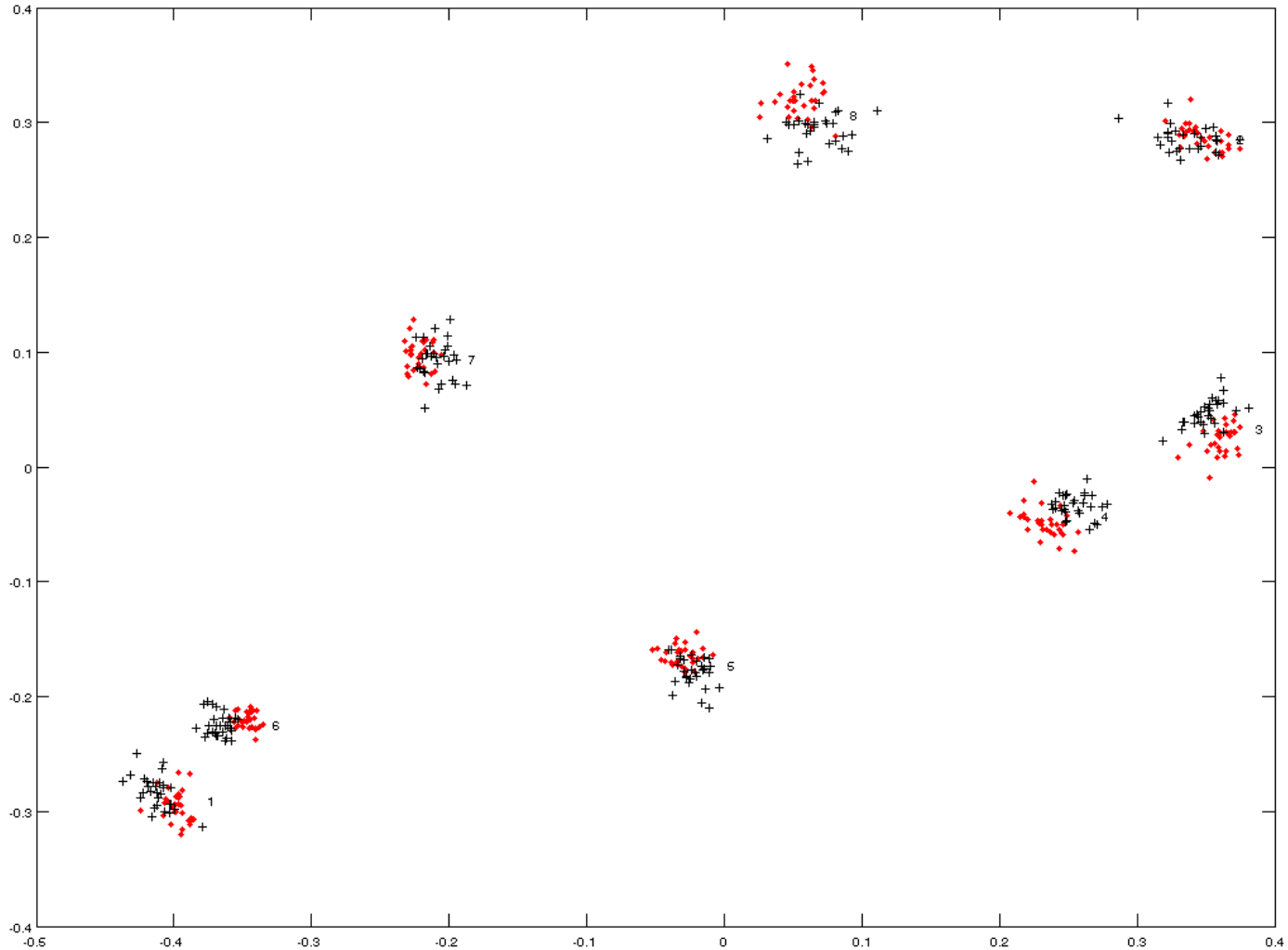
- Affine Shape Space $A\Sigma_m^k$. Group G consists of all affine transformations.
- Applications in protein matching, in scene recognition: to reconstruct a larger image from partial views in a number of ariel images of that scene.
- Projective Shape Space $P\Sigma_m^k$. Landmarks of the k-ad viewed in $\mathbb{R}P^m$-the set of all lines through origin in $\mathbb{R}^{m+1}$. The projective shape of a k-ad is invariant under all projective transformations.
- Applications in image analysis, robotics - for robots to visually recognize a scene.

Relation between different notions of shapes

- When images/photos obtained through a central projection, like a pinhole camera, projective shape analysis is useful.
- For images takes from a great distance, the rays from the object almost parallel to the camera plane. Then affine shape analysis appropriate.
- Further if the rays are perpendicular to the camera plane, similary shapes can be used.

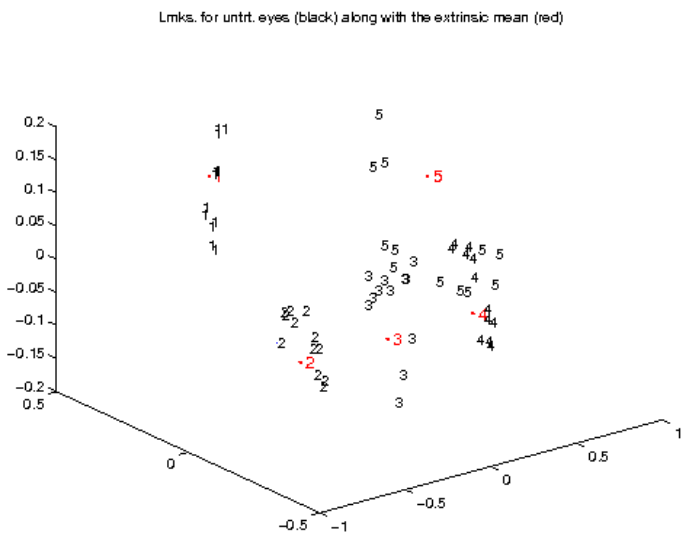
- 8 landmarks chosen on the midline plane of the 2D image of 29 male and 30 female gorilla skulls (Dryden and Mardia (1998)).
- Goal: Study the shapes of the skulls and use that to detect difference in shapes between the sexes. Finds application in morphometrics.
- Two mutually independent iid sample of planar shapes of sizes 29 and 30 on Σ_2^k, $k = 8$.

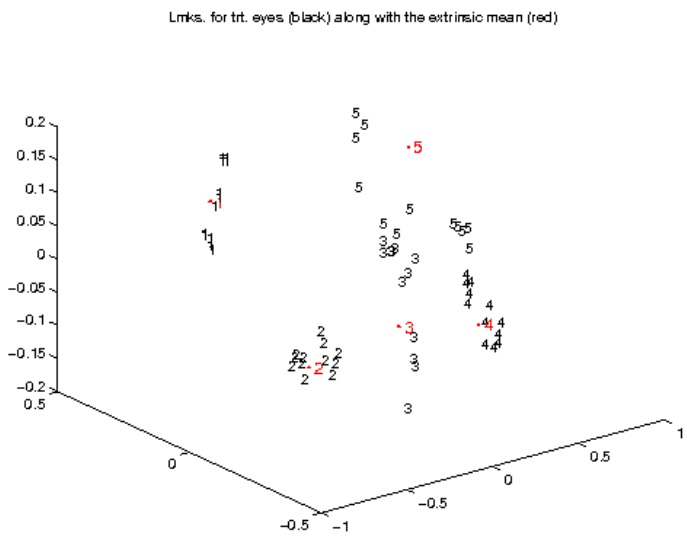
Preshapes of k-ads (k=8) from skulls of 30 females (red) & 29 male (+) gorillas



- 5 landmarks on the glaucoma induced eye and the normal eye of 12 rhesus monkeys. (BP(2005))
- Paired sample of size 12 on $R\Sigma_3^k$, $k = 5$.
- Goal: Test for any significant difference between the shapes of the glaucoma induced and normal eyes by comparing the ex. means.

Glucoma detection plots: sample k-ads





5 LMKS. FROM NORMAL & INFECTED EYES OF 12 MONKEYS ALONG WITH THE SAMPLE EX. MEANS

Fréchet Mean & Variation

- Let (M, ρ) be a metric space and Q be a prob. distn. on M.
- Fréchet mean of Q is the unique minimizer (if exists) of

$$F(p) = \int_M \rho^2(p, x) Q(dx), \ p \in M$$

- The minimum value of F is called the Fréchet variation or spread of Q.

Intrinsic & Extrinsic Mean

- Let (M, g) be a d-dimensional Riemannian manifold with metric tensor g.
- Use geodesic distance d_g in Fréchet function to define the intrinsic mean and variation.
- Let $J\colon M \to \mathbb{R}^D$ $(D >> d)$ be an embedding.
- Use the extrinsic distance d_E induced by this embedding to define the extrinsic mean and variation.

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

- Riemannian manifold of dim. d.
- Geodesic distance d_g = Great circle distance $\in [0, \pi]$.
- Intrinsic Mean of prob. Q exists if its support in a geodesic ball of radius $\frac{\pi}{2}$.

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

- Riemannian manifold of dim. d.
- Geodesic distance d_g = Great circle distance $\in [0, \pi]$.
- Intrinsic Mean of prob. Q exists if its support in a geodesic ball of radius $\frac{\pi}{2}$.
- Embedding J = Inclusion map, Extrinsic distance d_E = Chord distance.
- $\tilde{\mu} = \int_{S^d} xQ(dx) \in \mathbb{R}^{d+1}$.
- Extrinsic mean $\mu_E = \frac{\tilde{\mu}}{\|\tilde{\mu}\|}$ exists iff $\tilde{\mu} \neq 0$.

- Landmarks extracted from a 2D image.

- Landmarks extracted from a 2D image.
- k-ad x represented by a complex k-vector ($\in \mathcal{C}^k$).

- Landmarks extracted from a 2D image.
- k-ad x represented by a complex k-vector ($\in \mathcal{C}^k$).
- Its shape is its orbit under translation, scaling and rotation.

- Landmarks extracted from a 2D image.
- k-ad x represented by a complex k-vector ($\in \mathcal{C}^k$).
- Its shape is its orbit under translation, scaling and rotation.
- To remove translation, bring its cetroid to the origin: $x_c = x - \bar{x}$. x_c lies in a $(k-1)$ dim. subspace of $\mathcal{C}^k$.

- Landmarks extracted from a 2D image.
- k-ad x represented by a complex k-vector ($\in \mathcal{C}^k$).
- Its shape is its orbit under translation, scaling and rotation.
- To remove translation, bring its cetroid to the origin: $x_c = x - \bar{x}$. x_c lies in a $(k-1)$ dim. subspace of $\mathcal{C}^k$.
- Normalize x_c to have norm 1: $z = x_c/\|x_c\|$. This z is called the preshape of x. It contains shape info. + rotation. Lies on the complex unit sphere $\mathcal{C}S^{k-2}$ in $\mathcal{C}^{k-1}$.

- Landmarks extracted from a 2D image.
- k-ad x represented by a complex k-vector ($\in \mathcal{C}^k$).
- Its shape is its orbit under translation, scaling and rotation.
- To remove translation, bring its cetroid to the origin: $x_c = x - \bar{x}$. x_c lies in a $(k-1)$ dim. subspace of $\mathcal{C}^k$.
- Normalize x_c to have norm 1: $z = x_c/\|x_c\|$. This z is called the preshape of x. It contains shape info. + rotation. Lies on the complex unit sphere $\mathcal{C}S^{k-2}$ in $\mathcal{C}^{k-1}$.
- Shape of x or z is its orbit under all 2D rotations. Since rotation by angle θ equiv. to multiplication by the unit complex number $e^{i\theta}$, the shape can be represented as

$$\sigma(x) = [z] = \{e^{i\theta}z : \ -\pi < \theta \leq \pi\}.$$

- Landmarks extracted from a 2D image.
- k-ad x represented by a complex k-vector ($\in \mathcal{C}^k$).
- Its shape is its orbit under translation, scaling and rotation.
- To remove translation, bring its cetroid to the origin: $x_c = x - \bar{x}$. x_c lies in a $(k-1)$ dim. subspace of $\mathcal{C}^k$.
- Normalize x_c to have norm 1: $z = x_c/\|x_c\|$. This z is called the preshape of x. It contains shape info. + rotation. Lies on the complex unit sphere $\mathcal{C}S^{k-2}$ in $\mathcal{C}^{k-1}$.
- Shape of x or z is its orbit under all 2D rotations. Since rotation by angle θ equiv. to multiplication by the unit complex number $e^{i\theta}$, the shape can be represented as

$$\sigma(x) = [z] = \{e^{i\theta}z : \ -\pi < \theta \leq \pi\}.$$

- Σ_2^k is the collection of all such shapes.

Extrinsic Distance & Mean

- Σ_2^k is a Riemannian manifold of (real) dim. $2k-4$.
- Can be embedded into space of $k \times k$ complex Hermitian matrices via $J([z]) = zz^*$.
- It induces the extrinsic distance $d_E([u],[v]) = \|J([u]) - J([v])\| = \sqrt{2(1-|u^*v|^2)}$ under which Σ_2^k is a compact m.s.

Extrinsic Distance & Mean

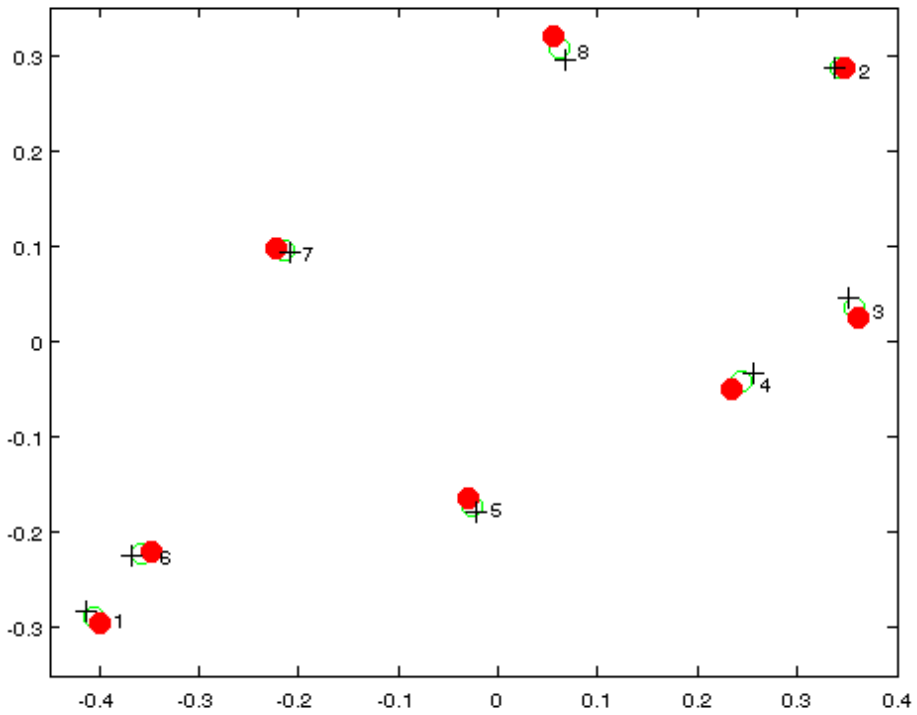
- Σ_2^k is a Riemannian manifold of (real) dim. $2k-4$.
- Can be embedded into space of $k \times k$ complex Hermitian matrices via $J([z]) = zz^*$.
- It induces the extrinsic distance $d_E([u],[v]) = \|J([u]) - J([v])\| = \sqrt{2(1-|u^*v|^2)}$ under which Σ_2^k is a compact m.s.
- Let $\tilde{\mu} =$ Mean of $Q \circ J^{-1}$. The shape of its eigenvector corresponding to the largest eigenvalue λ is the extrinsic mean of Q provided λ has multiplicity 1.

- Given a random sample from some unknown distn. Q on M, use the sample mean and variation (extrinsic or intrinsic) to identify Q.
- Construct confidence regions for the population parameters by asymptotic and bootstrap methods.
- Given two samples, distinguish between the underlying distns. by comparing the sample means or variations using two sample tests.
- Shown consistency of the sample estimates and asymptotic normality of the test statistics on various manifolds in Bhattacharya & Patrangenaru(2003,05), Bhattacharya(2008), Bhattacharya& Bhattacharya(2008,09) etc.

Gorilla Skull Application

- Compared the extrinsic and intrinsic sample means for the two sexes by two sample asymptotic and bootstrap tests.
- Asymptotic p-values $< 10^{-16}$.
- Strong evidence in favor of H_1 that the two sexes have different shape distributions.

Gorilla Skulls: Extrinsic Mean Shapes Plot

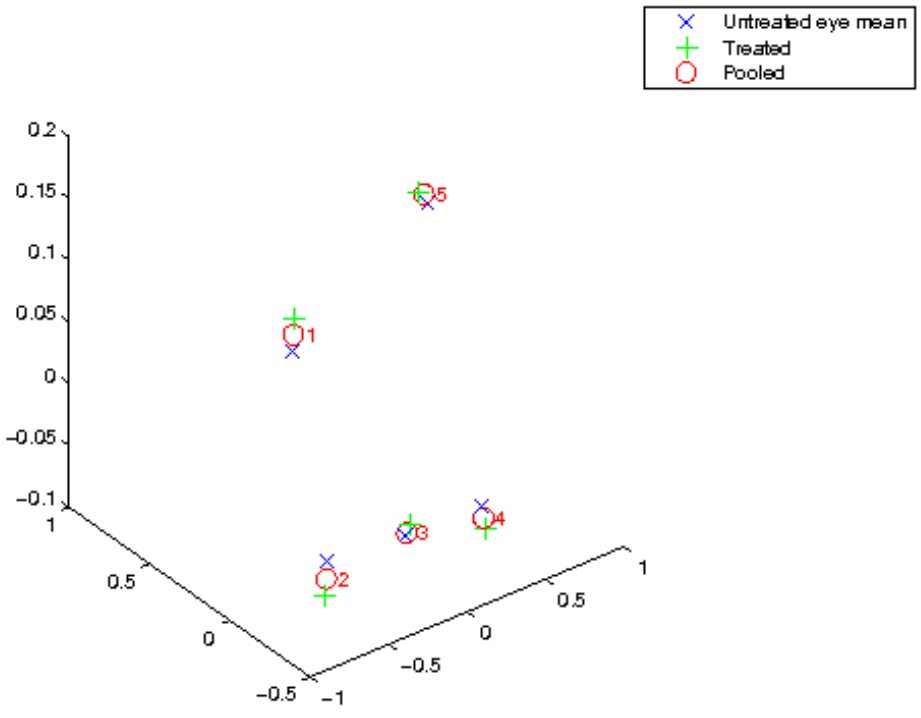


SAMPLE EX. MEANS FOR FEMALES (r.), MALES (+) ALONG WITH THE POOLED SAMPLE EX. MEAN (go)

Application to Glaucoma detection

- View the shapes in $R\Sigma_m^k$, $m = 3$, $k = 5$.
- Asymptotic p-value for comparing ex. mean shapes for two eyes = 1.38×10^{-5} (Bhattacharya 2008).
- Conclusion: Shape changes due to Glaucoma.

Glaucoma Detection Plots: Sample Extrinsic Means



EX. MEAN SHAPES FOR THE 2 EYES ALONG WITH POOLED SAMPLE EX. MEAN

Motivation

- In many applications, aspects other than location and spread may be useful.
- Then full likelihood based methods needed.
- Existing parametric models on manifolds don't fit data arising in applications very well (Bhattacharya and Bhattacharya 2008).
- Bayesian nonparametric methods have the advantage of providing a full probabilistic characterization of uncertainty, which is valid even in small samples.

Motivation

- Fundamental to full likelihood based np inference is density estimation.
- By setting suitable priors on the space of all distributions, possible to approximate the underlying model very well while avoiding over parametrization leading to efficient computations.
- Applications in np clustering, regression, classification and hypothesis testing on manifolds.

Random Mixture Density Model I

- Let (M, ρ) be a compact metric space.
- Let X be a r.v. on M having a density w.r.t. some fixed base measue λ.
- Model the density f as a location mixture model

$$f(x; P, \kappa) = \int_M K(x; \mu, \kappa) P(d\mu), \ x \in M \qquad (9.1)$$

P being the mixing distn. on M and κ^{-1} scale/band-width parameter and $K(.; \mu, \kappa)$ a probability kernel on M with location μ, inverse-scale κ satisfying

$$\int_M K(x; \mu, \kappa) \lambda(dx) = 1.$$

Random Mixture Density Model II

- Set a prior Π_1 on (P, κ) which induces one, call it Π, on the space of all densities $\mathcal{D}(M)$ on M through (9.1).
- Then given an iid realisation $\{X_1, \ldots, X_n\} \equiv \mathbf{X}_n$ of X, get the posterior of f to estimate the unknown density.

- Let P_t be the true distribution of X and f_t be its density.
- Let $U \subset \mathcal{D}(M)$ be a weak open neighborhood of f_t.
- Weak consistency means $\Pi(U|\mathbf{X}_n) \longrightarrow 1$ as $n \to \infty$ almost surely (a.s.) for any such U.
- Assume

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

A1,A2 impose regularity condns. on the kernel which we'll verify for some specific kernels on some manifolds. **A3** is a condition on the prior easy to verify for many standard np priors.

Theorem (Full uniform support of prior BD(2010a))

*Under assumptions **A1**-**A4**, given any* $\epsilon > 0$,

$$\Pi\left(\left\{f : \sup_{x \in M} |f(x) - f_t(x)| < \epsilon\right\}\right) > 0$$

A1,A2 impose regularity condns. on the kernel which we'll verify for some specific kernels on some manifolds. **A3** is a condition on the prior easy to verify for many standard np priors.

Theorem (Full uniform support of prior BD(2010a))

*Under assumptions **A1**-**A4**, given any* $\epsilon > 0$,

$$\Pi\left(\left\{f : \sup_{x \in M} |f(x) - f_t(x)| < \epsilon\right\}\right) > 0$$

- Kullback-Leibler divergence of a density f from f_t defined as $\mathrm{KL}(f_t; f) = \int_M f_t(m) \log\{f_t(m)/f(m)\}\lambda(dm)$.
- Let $K_\epsilon(f_t)$ denote the neighborhood $\{f : \mathrm{KL}(f_t; f) < \epsilon\}$.

f_t is in the KL support of Π and Π satisfies the KL condn. at f_t if $\Pi\{K_\epsilon(f_t)\} > 0$ for all $\epsilon > 0$.

Theorem (KL Property of the prior BD (2010a))

*Under assumptions **A1**-**A5**, f_t is in the KL support of Π.*

f_t is in the KL support of Π and Π satisfies the KL condn. at f_t if $\Pi\{K_\epsilon(f_t)\} > 0$ for all $\epsilon > 0$.

Theorem (KL Property of the prior BD (2010a))

*Under assumptions **A1**-**A5**, f_t is in the KL support of Π.*

Then using Schwartz theorem weak posterior consistency follows.

Theorem (Schwartz 1965)

If (1) f_t 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 $H_0 : f = f_t$ versus $H_1 : f \in U^c$, then $\Pi(U|X_1, \ldots, X_n) \to 1$ as $n \to \infty$ a.s.

When U is a weak nb.hood of f_t such a sequence of tests seperating U and U^c exist.

Prior selection I

- A choice for prior Π_1 is $P \sim \mathrm{DP}(w_0 P_0)$ and independently $\kappa \sim \pi$ - a density on $\Re^+$ such as gamma.
- Denote it as DP-π prior.
- $\mathrm{DP}(w_0 P_0)$ is the Dirichlet process prior on $\mathcal{M}(M)$ with precision ω_0 and base measure P_0.
- Weak support of $\mathrm{DP}(w_0 P_0)$ is all P with $\mathrm{supp}(P) \subseteq \mathrm{supp}(P_0)$.
- Hence if $\mathrm{supp}(P_0) = M$ and π_1 has a tail near ∞, DP-π prior Π_1 satisfies **A3** & hence weak p.c. follows with appropriate choice of kernel.

Prior selection II

- If $P \sim \mathrm{DP}(\omega_0 P_0)$, then

$$(P(M_1), \ldots, P(M_k)) \sim \mathrm{Dirich}(\omega_0 P_0(M_1), \ldots, \omega_0 P_0(M_k))$$

  for any partition $M = \cup M_i$. (Fergusson 1973)
- P is discrete a.s. with the following stick breaking representation (Sethuraman 1994). $P = \sum_{j=1}^{\infty} w_j \delta_{\mu_j}$, μ_j iid P_0, $w_j = V_j \prod_{l=1}^{j}(1 - V_l)$, V_j iid $\mathrm{Be}(1, w_0)$.

- The location mixture model can be interpreted as follows.
  X_i indep. $K(\theta_i, \tau)$ $i = 1, \ldots, n$,
  θ_i iid P,
  $P \sim DP(w_0 P_0)$, $\kappa \sim \pi_1$.
- Introduce latent cluster index $S_1, \ldots, S_n$ s.t. $\theta_i = \mu_{S_i}$ and conditional likelihood given $\mathbf{S} = \{S_i\}_1^n$, $\mu = \{\mu_j\}_1^\infty$, $\mathbf{V} = \{V_j\}_1^\infty$, κ becomes $\prod_1^n K(X_i; \mu_{S_i}, \kappa)$.
- Hence $\Pi(\mathbf{S}, \mu, \mathbf{V}, \kappa | \mathbf{X}_n) \propto$
  $\pi_1(\kappa) \left(\prod_{j=1}^\infty P_0(d\mu_j) \mathrm{Be}(V_j; 1, w_0) \right) \left(\prod_i w_{S_i} K(X_i; \mu_{S_i}, \kappa) \right)$.
- Introduce latent slice sampling variables $\mathbf{u} = \{u_i\}_1^n$ and the full posterior becomes $\Pi(\mathbf{S}, \mu, \mathbf{V}, \kappa, \mathbf{u} | \mathbf{X}_n) \propto$
  $\pi_1(\kappa) \left(\prod_{j=1}^\infty P_0(d\mu_j) \mathrm{Be}(V_j; 1, w_0) \right) \left(\prod_i \mathrm{I}(u_i < w_{S_i}) K(X_i; \mu_{S_i}, \kappa) \right)$.

Can use the exact block Gibbs sampler of Yau et al. (2009) to get repeated draws from the posterior of $(\mathbf{S}, \mu, \mathbf{V}, \kappa, \mathbf{u})$ and hence f as follows.

1. Update $\{S_i\}_1^n$ by sampling from the multinomial conditional posterior distribution with $\Pr(S_i = j) \propto K(X_i; \mu_j, \kappa)$ for $j \in A_i$, $A_i = \{1 \leq j \leq l : w_j > u_i\}$ and l is the smallest index satisfying $1 - \min(u) < \sum_{j=1}^{l} w_j$. Draw $V_j \sim \text{Be}(1, w_0)$ and $\mu_j \sim P_0$ for $j > \max(S)$.
2. Update occupied cluster locations μ_j, $j \leq \max(S)$ from the conditional posterior $\propto P_0(d\mu_j) \prod_{i:S_i=j} K(X_i; \mu_j, \kappa)$.
3. Update inverse-bandwidth κ by sampling from the conditional posterior $\propto \pi_1(d\kappa) \prod_{i=1}^{n} K(X_i; \mu_{S_i}, \kappa)$.
4. Update the stick-breaking random variables V_j, $j \leq \max(S)$ from $\text{Be}(1 + \sum_i \text{I}(S_i = j), w_0 + \sum_i \text{I}(S_i > j))$.

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

These steps repeated a large number of iterations, with a burn-in discarded to allow convergence. The average of f over those draws gives the bayes estimate.

- We can also use the following density model for X.

$$f(x;Q) = \int_{M\times\Re^+} K(m;\mu,\kappa)Q(d\mu d\kappa)$$

  with parameter Q which mixes accross various locations and scales.
- Set a prior Π_2 such as DP on Q and get the posterior of f.
- That is X_i indep $K(.;\mu_i,\kappa_i)$ $i = 1,\ldots,n$,
  (μ_i,κ_i) iid Q,
  $Q \sim \Pi_2$.
- Have shown posterior consistency (BD(2010a)) using this model at any +ve cont. true density.
- Can perform posterior computations as in earlier case.

Want to show $\Pi(U|\mathbf{X}_n) \longrightarrow 1$ a.s. for any total variation/L_1 nb.hood of the true density f_t.

Theorem (Bhattacharya & Dunson 2010b)

*Under assumptions **A1-A9** strong posterior consistency holds.*

Assume

Strong Posterior consistency Assumptions

A6 $\exists\ \mathcal{K}_1, a_1, A_1 > 0$ such that for all $\mathcal{K} \geq \mathcal{K}_1$, $\mu, \nu \in M$,

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

A7 $\exists\ a_2, A_2 > 0$ such that for all $\kappa_1, \kappa_2 \in [0,\mathcal{K}]$, $\mathcal{K} \geq \mathcal{K}_1$,

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

A8 $\exists\ a_3 > 0$ s.t given any $\epsilon > 0$, M can be covered by finitely many ϵ-diameter balls of number of the order ϵ^{-a_3}.

A9 $\Pi_1(\mathcal{M}(M) \times (n^{(1/a)}, \infty)) < \exp(-n\beta)$ for some $a > a_1 a_3$ and $\beta > 0$.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

- Choose $\Pi_1 = \Pi_{11} \otimes \pi$ with a Dirichlet process Π_{11}.
- Take π to be a Weibull density: $\pi(\kappa) \propto \kappa^{c-1} \exp(-b\kappa^a)$ with $a > a_1 a_3$ and assumption **A9** holds, hence strong posterior consistency follows.
- A gamma π satisfies **A3** but not **A8** (unless $a_1 a_3 \leq 1$). However strong posterior consistency may still follow because above theorem presents only sufficient conditions.

- When dimension of the manifold is large as in shape analysis with a large number of landmarks, for strong consistency, the shape parameter in the proposed Weibull prior needs to be very large, implying a prior on the bandwidth that places very small probability near zero, which is undesirable.
- Instead allow the prior to depend on sample size n and we can obtain priors that have better small sample operating characteristics, while still leading to strong consistency.
- As before assume P and κ to be independent under Π_1. Let $P \sim \Pi_{11}$ - a constant prior while $\kappa \sim \pi_n$ - n dependent density on $\Re^+$.

A10 Π_{11} has full support.

A11 For any $\beta > 0$, there exists a $\kappa_0 \geq 0$, s.t. $\forall\ \kappa \geq \kappa_0$, $\liminf_n \exp(n\beta)\pi_n(\kappa) = \infty$.

A12 For some $\beta_0 > 0$ and $a < (a_1 a_3)^{-1}$, $\lim_n \exp(n\beta_0)\pi_n\{(n^a, \infty)\} = 0$.

Theorem (Posterior Consistency BD (2010b))

*Under assumptions **A1-A2** on the kernel, **A10-A11** on the prior and **A4-A5** on the true density, weak posterior consistency holds. Under further assumptions **A6-A7** on the kernel, **A8** on the space M and **A12** on the prior, strong p.c. also holds.*

For Gamma prior $\pi_n(\kappa) \propto \kappa^{\alpha-1}\exp(-\beta_n\kappa)$, **A11** satisfied if $\beta_n/n \to 0$. **A11** also satisfied if $n^{(a_1 a_3)^{-1}-1}\beta_n \to \infty$.
Hence weak and strong p.c. follow with $\beta_n = b_1 n/\{\log(n)\}^{b_2}$ for any $b_1, b_2 > 0$.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

Strong Posterior consistency Assumptions

A6 $\exists\ \mathcal{K}_1, a_1, A_1 > 0$ such that for all $\mathcal{K} \geq \mathcal{K}_1$, $\mu, \nu \in M$,

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

A7 $\exists\ a_2, A_2 > 0$ such that for all $\kappa_1, \kappa_2 \in [0,\mathcal{K}]$, $\mathcal{K} \geq \mathcal{K}_1$,

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

A8 $\exists\ a_3 > 0$ s.t given any $\epsilon > 0$, M can be covered by finitely many ϵ-diameter balls of number of the order ϵ^{-a_3}.

A9 $\Pi_1(\mathcal{M}(M) \times (n^{(1/a)}, \infty)) < \exp(-n\beta)$ for some $a > a_1 a_3$ and $\beta > 0$.

- To define a density model w.r.t. the volume form V, choose kernel K to be the von Mises-Fisher (vMF) density given by

$$\mathrm{vMF}(m; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa m^T \mu),$$

  with normalizing constant $c(\kappa)$.
- Its extrinsic mean is μ and hence the kernel location. κ is a measure of concentration - $\kappa = 0$ corresponds to the uniform distribution on S^d while as $\kappa \to \infty$, it converges to point mass at μ in L^1 sense.

Theorem (Consistency BD(2010b))

(1) The vMF kernel K satisfies the necessary assumptions for w.p.c.
(2) It also satisfies ***A6*** *with* $a_1 = d/2 + 1$ *and* ***A7*** *with* $a_2 = d/2$. (S^d, ρ) *satisfies* ***A8*** *with* $a_3 = d$. *As a result s.p.c follows with a DP prior on mixing distribution P and Weibull prior on* κ *with shape parameter exceeding* $d + d^2/2$.

Alternatively when d large a more preferred prior can be DP-Gam(α, β_n) with $\beta_n = n/\log(n)$.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

Weak Posterior consistency Assumptions

A1 K is continuous on $M \times M \times \Re^+$.

A2 For any cont. ϕ

$$\lim_{\kappa \to \infty} \sup_{x \in M} \left| \phi(x) - \int_M K(x; \mu, \kappa) \phi(\mu) \lambda(d\mu) \right| = 0.$$

A3 For any $\kappa_0 > 0$, $\exists\ \kappa \geq \kappa_0$ s.t. $(P_t, \kappa) \in \text{supp}(\Pi_1)$ (weak support).

A4 f_t is continuous.

A5 f_t is strictly positive.

Strong Posterior consistency Assumptions

A6 $\exists\ \mathcal{K}_1, a_1, A_1 > 0$ such that for all $\mathcal{K} \geq \mathcal{K}_1$, $\mu, \nu \in M$,

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

A7 $\exists\ a_2, A_2 > 0$ such that for all $\kappa_1, \kappa_2 \in [0,\mathcal{K}]$, $\mathcal{K} \geq \mathcal{K}_1$,

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

A8 $\exists\ a_3 > 0$ s.t given any $\epsilon > 0$, M can be covered by finitely many ϵ-diameter balls of number of the order ϵ^{-a_3}.

A9 $\Pi_1(\mathcal{M}(M) \times (n^{(1/a)}, \infty)) < \exp(-n\beta)$ for some $a > a_1 a_3$ and $\beta > 0$.

- Let K be the complex-Watson density $\mathrm{CW}([z];[\mu],\kappa) = c^{-1}(\kappa)\exp(\kappa|z^*\mu|^2)$, $z,\mu \in CS^{k-1}$, $\kappa > 0$.
- It has extrinsic mean $[\mu]$ and hence that is the kernel location.
- κ is a measure of concentration - as $\kappa \to 0$, CW converges to the uniform density and as $\kappa \to \infty$, CW converges weakly to $\delta_{[\mu]}$ uniformly in μ.
- The kernel satisfies **A1-A2** to get weak p.c. at any true positive density using a DP-Gamma prior (BD 2010a).

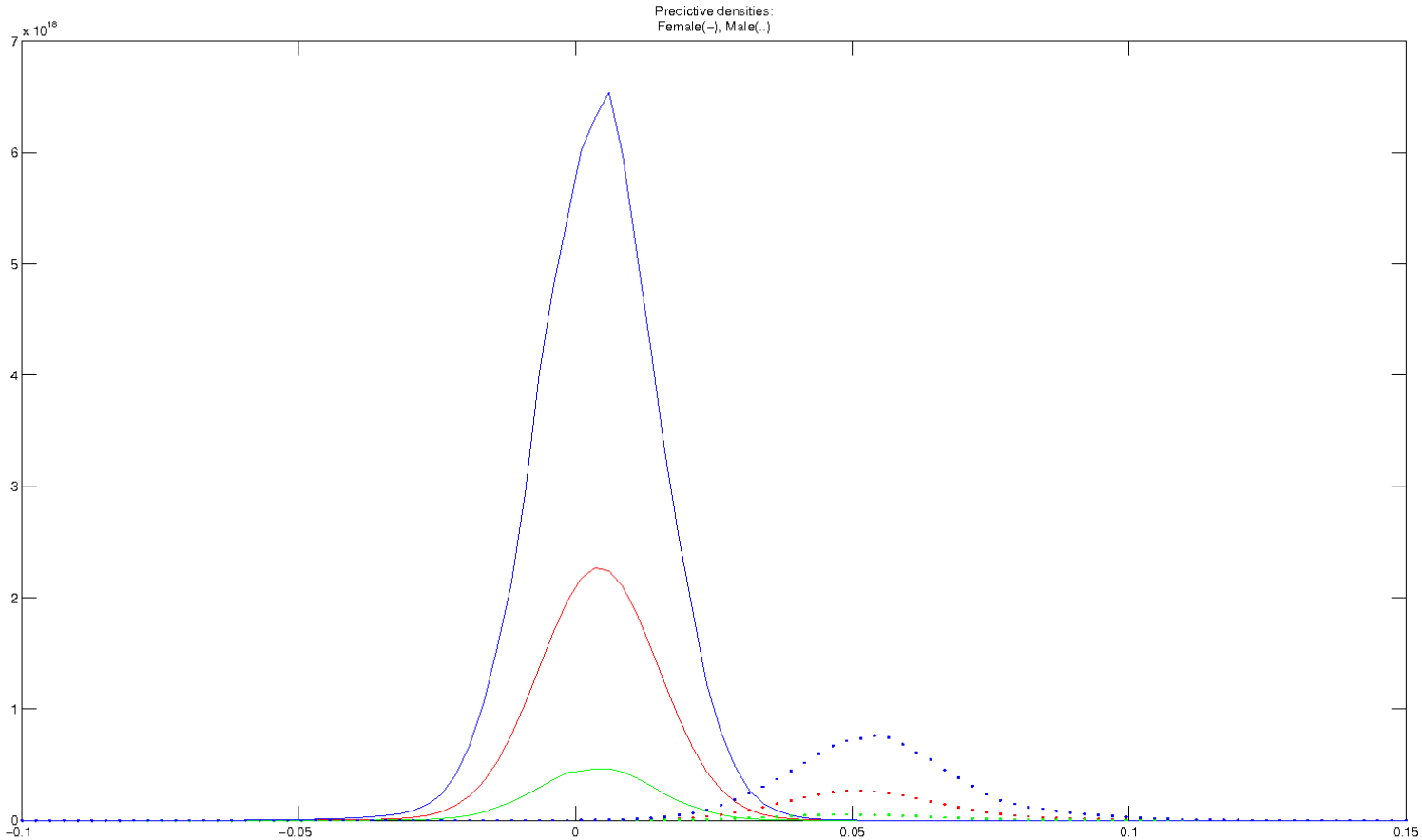
Strong Posterior Consistency

- The complex Watson kernel also satisfies **A6** with $a_1 = k - 1$ and **A7** with $a_2 = 3k - 8$.
- The compact m.s. (Σ_2^k, d_E) satisfies **A8** with $a_3 = 2k - 3$.
- Hence strong p.c. holds with a DP-Weibull prior, for Weibull with shape parameter exceeding $(2k - 3)(k - 1)$ (BD 2010b).
- When k is high, i.e. many lmks., instead use a DP-Gamma(α, β_n) prior with β_n increasing with n at a suitable rate.

Gorilla Skull shape density plot

Here is a 1D slice of the density estimates for the male and female gorrilas. Densities evaluated along the geodesic starting at the female towards the male sample ex. mean.

Female: solid, Male: dotted, Posterior mean densities: red, 95% C.R.: blue/green



- From the shape density estimates for the 2 groups, can predict the gender from the shape via np discriminant analysis.
- Assume the probability of being female is 0.5 and use a separate DP mixture of cW 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 $p([z]) = 1/\{1 + f_2([z])/f_1([z])\}$.
- To estimate the posterior probability, average $p([z])$ across Markov chain Monte Carlo iterations to obtain $\hat{p}([z])$.
- The prior Π_1 taken to be $DP(P_0) \otimes \text{Gam}(1.01, 1e-3)$ with $P_0 = CW(\mu_0, 1e-03)$ and μ_0 being the sample ex. mean for the group under consideration.

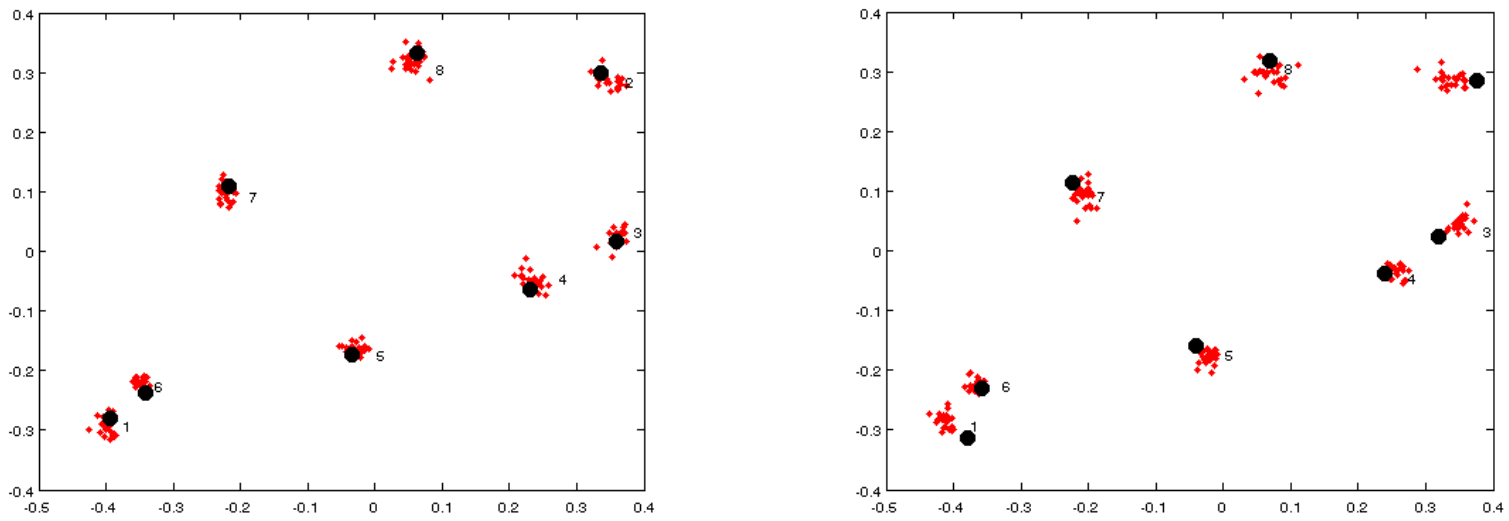
- To test the performance of the classifier, randomly partition the sample into training-test samples, using training samples, get the classifier and apply it to the test data.

This table presents the estimated posterior probabilities of being female for each gorilla in the test sample along with a 95% credible interval (CI) for $p([z])$ for one such partition. Also shown is the ex. dist. between the sample shape and the female($\hat{\mu}_1$), male($\hat{\mu}_1$) sample ex. means.

gender	$\hat{p}([z])$	95% CI	$d_E([z], \hat{\mu}_1)$	$d_E([z], \hat{\mu}_2)$
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

- There is misclassification in the 3rd female and 3rd male.
- Based on the CI, there is some uncertainity in classifying the 3rd female.
- Perhaps there is something unusual about the shapes for these individuals, which was not represented in the training data, or they were labelled incorrectly.
- Can also define a distance-based classifier, which allocates a test subject to the group having mean shape closest to that subjects' shape.
- The 2 classifiers give consistent results.
- However, such a 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.

Gorilla Skull Training and Test Samples



Training(red) & mis-classified test(black) samples corresponding to females (left) & males (right).

- Let Y be a categorical variable taking values in $\mathbb{Y} = \{1, 2, \ldots, L\}$. Eg. Gorilla gender, Presence/Absence of Glaucoma e.t.c.
- Wanto predict the conditional of Y given predictors X lying on M.
- Model the joint distbn. of (X, Y) via density

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

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

- Interpret the model in following heirchical way.
  Draw (μ, ν) from P.
  Given (μ, ν, κ), X and Y are conditionally independent with X having the conditional density $K(.; \mu, \kappa)$ w.r.t. the volume form $V(dx)$ and $\Pr(Y = j|\mu, \nu, \kappa) = \nu_j$, $1 \leq j \leq L$.

- Set a prior Π_1 on (P, κ) such as DP$(\omega_0 P_0)$-Gamma with P_0 some distn. on $M \times S_{c-1}$.
- This induces a prior Π on the joint density f of (X, Y).
- Get the posterior of f from a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ and compute the posterior predicive probability of allocating a new feature x_{n+1} to different categories and hence classify it as y_{n+1}.

Define $P_t \in \mathcal{M}(M \times S_{L-1})$ as

$$P_t(d\mu d\nu) = \sum_{j=1}^{L} f_t(\mu, j) V(d\mu) \delta_{e_j}(d\nu)$$

f_t being the true jt. density of (X, Y) and e_j is the L-dimensional vector with 1 in j^{th} postn. and 0 else-where.

Theorem (Model Flexibility BD(2010c))

*Under assumptions **A1**-**A5**, (a) Π includes f_t in its uniform and KL support. (b) Hence weak p.c. follows if the true density is +ve and cont. everywhere.*

Let $p_j(x)$ be the predictive probability function for class j, i.e. $f(x,j)/\{\sum_l f(x,l)\}$, $j \leq c$ and $p_j^t(x)$ be the true function.

Theorem (Posterior Consistency BD(2010c))

*Under assumptions **A1**-**A9**, (a) strong p.c. also follows. (b) As a result*

$$\Pi\left(\int_M |p_j(x) - p_j^t(x)| < \epsilon | \mathbf{X_n}, \mathbf{Y_n}\right) \to 1 \text{ a.s.}$$

as $n \to \infty$ *for* $1 \leq j \leq c$.

- Let M be S^d.
- Take kernel K to be vMF density.
- Let $\Pi_1 = \mathrm{DP}(\omega_0 P_0) \otimes \pi_n$. Under P_0, μ and ν independent vMF and Dirichlet distributed and

$$\pi_n(\kappa) \propto c^n(\kappa)\kappa^{a+\frac{nd}{2}-1}e^{-\kappa(n+b)} \text{ for } a, b > 0.$$

- Then we get posterior conjugacy in the Exact Blocked Gibbs sampling algorithm to get a draw from the posterior of $f(x, y; P, \kappa)$.

Application to Simulated Data

Get a (X,Y) training sample sized 200 on $S^9 \times \{1,2,3\}$ from

$$f_t(x,y) = 1/3 \sum_{j=1}^{c} I(y=j)\Big\{0.1 \sum_{l=1}^{2} \text{vMF}(x;\mu_l;\kappa) + 0.8\text{vMF}(x;\tilde{\mu}_j;\kappa)\Big\}$$

with $\mu_1 = e_1,\ \mu_2 = \exp_{e_1}(0.2e_2)$,

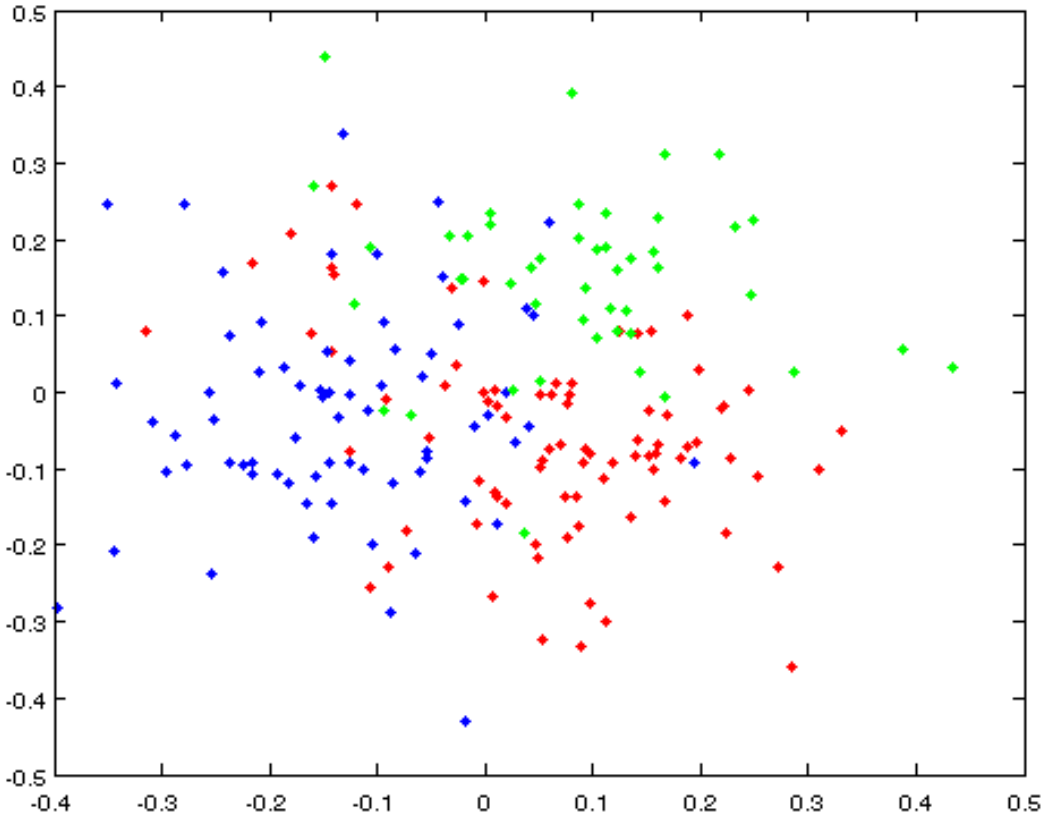
$\tilde{\mu}_1 = \exp_{e_1}(-0.2e_2), \tilde{\mu}_2 = \exp_{e_1}(0.2e_3), \tilde{\mu}_3 = \exp_{e_1}(-0.2e_3)$

and $\kappa = 100$.

Here $q = \exp_p(v)$ lies on the geodesic (great circle) starting at p in direction V and at a distance $\|V\|$

Training sample Plot

1st two P.C.s of the normal coordinates of the features. Color being the class label: r=(Y=1), b=(Y=2),g=(Y=3).



- Run 50,000 it. of the Blocked Gibb's sampler to draw from the posterior jt. density and compute the posterior predictive probability function $p_n(j|x, \text{data})$ for category $j = 1, 2, 3$ in each iteration.
- Prior $\Pi_1 = DP(P_0) \otimes \pi_n(a, b)$ with $P_0 = vMF(10\mu_0) \otimes Dirich(1, 1, 1)$, μ_0 being training sample ex. mean, $a = 1, b = 0.1$.
- Take their average after discarding first 15,000 draws and use that to predict categories of 100 test observations.
- Miss-classification rate per category = (32, 22, 37)% for each category.
- Over-all miss-classification rate = 30%.

- We can also use a frequentist parametric discriminat analysis approach.
- Fit finite mixture of Gaussians for 3 categories independently, 2 clusters per category and classify using Discriminant analysis approach.
- Miss-classification rate per category = (39, 40, 59)%
- Over-all miss-classification rate = 46%.
- Run in-to singularity problems while fitting more than 2 clusters per category.

Null and Alternative Hypothesis

- Instead of classifying a new feature, goal is to test whether the distribution of the features differs across the classes.
- Hence test for independence between X and Y from the training sample.
- Alternative hypothesis H_1 corresponds to all joint density $f(x, y)$ while null corrsponds to

$$H_0 : f(x, y) = f(x)f(y) \text{ for all } (x, y) \in M \times \mathbb{Y}.$$

- Model under H_1 is $f(x, y; P, \kappa)$. Set prior Π_1 on parameters (P, κ).

Model Selection

- As a model for the joint density under H_0, replace $P(d\mu d\nu)$ with $P_1(d\mu)P_2(d\nu)$ so that the joint density becomes $f(x, y; P_1, P_2, \kappa) = f_X(x; P_1, \kappa) f_Y(y; P_2)$ where

$$f_X(x; P_1, \kappa) = \int_M K(x; \mu, \kappa) P_1(d\mu), \; f_Y(y; P_2) = \int_{S_{c-1}} \nu_y P_2(d\nu).$$

- Set prior Π_0 on parameters (P_1, P_2, κ).
- The Bayes-factor in favor of H_1 over H_0, BF, is then the ratio of the marginal likelihoods under H_1 and H_0,

$$BF = \frac{\int \prod_{i=1}^n f(x_i, y_i; P, \kappa) \Pi_1(dP d\kappa)}{\int \prod_{i=1}^n f_X(x_i; P_1, \kappa) f_Y(y_i; P_2) \Pi_0(dP_1 dP_2 d\kappa)}$$

Bayes Factor Computation

- The *BF* expression suggests that under H_0 the density of Y depends on P_2 only through the L-dimensional prob. vector

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

  Hence under Π_0, its sufficient to choose a prior for (P_1, κ, p), instead of specifying a full prior for (P_1, κ, P_2).
- We take (P_1, κ) to follow the marginal induced from Π_1 on (P, κ) independently of $p \sim$ Dirichlet.
- Then BD(2010c) presents an algo to approximate *BF* by introducing a latent variable which is the indicator of accepting H_1.

Gorilla Skull Application

- Goal to test if the skull shape distribution differs accross gender.
- Introduce category label Y and test for independence between X and Y.
- $BF(H_1 : H_0) > 10^{16}$.
  Conclusion: Strong evidence in favor of H_1 hence shape distbn. for two sexes different.

Gorilla Skull Application

- Goal to test if the skull shape distribution differs accross gender.
- Introduce category label Y and test for independence between X and Y.
- $BF(H_1 : H_0) > 10^{16}$.
  Conclusion: Strong evidence in favor of H_1 hence shape distbn. for two sexes different.
- Next permute the labels randomly, so that we expect independence.
- $BF(H_1 : H_0) = 2.11$.
  Conclusion: Not enough evidence in favor of H_1.

References


BHATTACHARYA, A. (2008). *Sankhya* Vol.**70** Part **3** 0-43.


BHATTACHARYA, A. AND BHATTACHARYA, R. (BB) (2008a). *IMS Collections* **3** 282-301.


BHATTACHARYA, A. AND BHATTACHARYA, R. (BB) (2008b). *Proc. Amer. Math. Soc.* **136** 2957-2967.


BHATTACHARYA, A. AND DUNSON, D. (BD) (2010a). Nonparametric Bayesian Density Estimation on Manifolds with applications to Planar Shapes. *To Appear in Biometrika*.


BHATTACHARYA, A. AND DUNSON, D. (BD) (2010b). Strong consistency of nonparametric Bayes density estimation on compact metric spaces *To Appear in Annals of Statistics*.

References contd.


BHATTACHARYA, A. AND DUNSON, D. (BD) (2010c). Nonparametric Bayes Classification and Testing on Manifolds. *Working Paper*.


BHATTACHARYA, R. AND PATRANGENARU, V. (BP) (2003). *Ann. Statist.*


BHATTACHARYA, R. AND PATRANGENARU, V. (BP) (2005). *Ann. Statist.*


DRYDEN, I. L. AND MARDIA, K. V. (1998). Wiley N.Y.


KENDALL, W.S. (1990). *London Math. Soc.*


SCHWARTZ, L. (1965).*Z. Wahrsch. Verw. Gebiete* **4** 10-26.


YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G.O. & HOLMES, C. (2009). Uni. of Oxford, U.K.