Nonparametric Bayes Inference on Manifolds with Applications

Abhishek Bhattacharya
Indian Statistical Institute

Based on the book *Nonparametric Statistics On Manifolds With Applications To Shape Spaces*

Sept.12,2011

- Non-Euclidean data arise in many applications like directional and axial data analysis, analysis of shapes of objects, images, hand writting etc.

- Non-Euclidean data arise in many applications like directional and axial data analysis, analysis of shapes of objects, images, hand writting etc.
- Requires building robust inference tools on manifolds like spheres, real and complex projective spaces, stiefel manifolds, Grassmanians, and many others.
- Frequentist np inference methods on general manifolds based on measures of centers and spreads considered in *Frechet (1948), A.Bhattacharya & R.Bhattacharya (2008, 2009, 2011), R.Bhattacharya & Patrangenaru (2002, 2003, 2005)* and others.

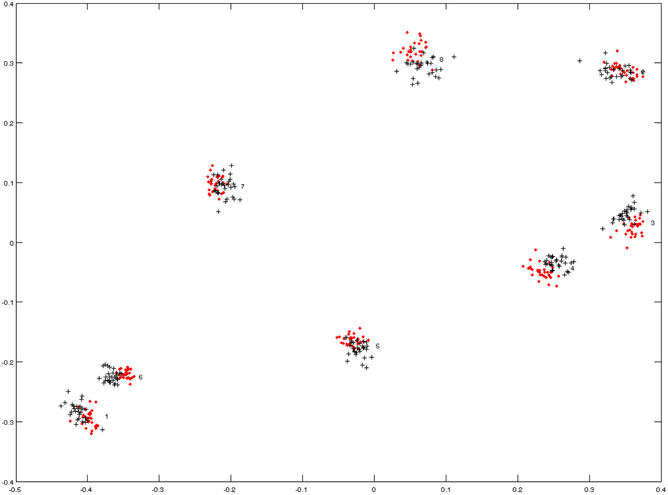
- Non-Euclidean data arise in many applications like directional and axial data analysis, analysis of shapes of objects, images, hand writting etc.
- Requires building robust inference tools on manifolds like spheres, real and complex projective spaces, stiefel manifolds, Grassmanians, and many others.
- Frequentist np inference methods on general manifolds based on measures of centers and spreads considered in *Frechet (1948), A.Bhattacharya & R.Bhattacharya (2008, 2009, 2011), R.Bhattacharya & Patrangenaru (2002, 2003, 2005)* and others.
- Density estimation plays a vital role for inference like in regression, classification and hypothesis testing.

- Most the np Bayes density estimation confined to real data.
- *Wu & Ghosal (2010)* proves strong consistency in np density estimation from Dirichlet process mixtures of multivariate Gaussian kernels on $\mathbb{R}^d$. Severe tail restrictions are imposed on the kernel covariance, which become overly restrictive with high dimensional data. Also the theory is specialized and cannot be generalized to arbitrary kernel mixtures on general spaces.
- *Bhattacharya & Dunson (2010)* considers general kernel mixture densities on compact metric spaces and proves consistency of the estimate in weak sense.
- Choice of kernels extended and weak and strong posterior consistency established under weaker assumptions in *Bhattacharya & Dunson (2011a)*.

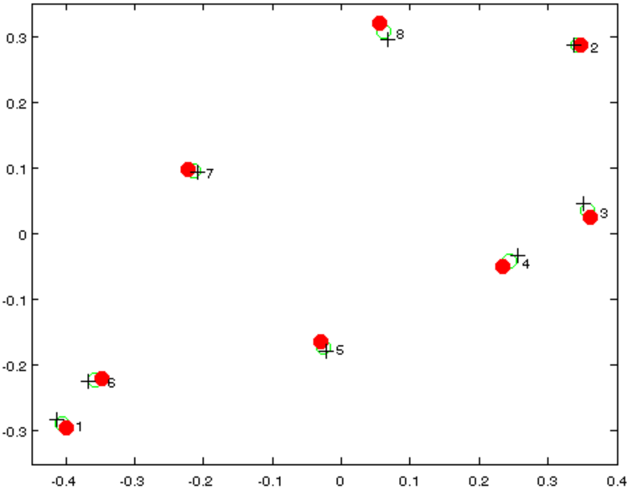
- Density estimation utilized for np regression, classification and hypothesis testing on manifolds in *Bhattacharya & Dunson (2011b)*, with even functional data in *Dunson & Bhattacharya (2010)*.

- Interest in studying shape of an organism & using shape for classification.
- Data on 2D image of gorilla skulls and their gender. There are 29 male and 30 female gorillas, *Dryden & Mardia (1998)*.
- To make the images finite dimensional, 8 landmarks chosen on the midline plane of each skull image.
- Since different images obtained under different orientations, scale etc, it is important to be invariant to translations, rotations & scaling, i.e. analyse the image similarity shape.
- How to nonparametrically estimate a shape distribution across organisms?
- How to test for differences in shape & obtain a good method for classification based on shape?

Gorilla Skull Images: Females (red), Males (+)



Mean Shapes Plot- Female (r.), Male (+), Pooled (go)



- Let (M, ρ) be a seperable metric space with a fixed base measure λ.
- Let $K(m; \mu, \nu)$ be a probability kernel on M with location $\mu \in M$ and other parameters $\nu \in N$, N being a seperable metric space.
- Let X be a r.v. whose density we are interested in modelling.
- A location mixture density model for X is then defined as

$$f(m; P, \nu) = \int_M K(m; \mu, \nu) P(d\mu)$$

  with unknown parameters $P \in \mathcal{M}(M)$ & $\nu \in N$.
- Set prior Π_1 on (P, ν) to induce a prior Π on density space $\mathcal{D}(M)$ through the density model.

Let f_t be the unknown density of X with distribution F_t.
From now on assume M to be compact. Also assume

1. Kernel K is continuous in its arguements.
2. For any cont. function $f : M \to \Re$, for any $\epsilon > 0$, there exists a compact subset N_ϵ of N with non-empty interior, such that

$$\sup_{m \in M, \nu \in N_\epsilon} \left| f(m) - \int_M K(m; \mu, \nu) f(\mu) \lambda(d\mu) \right| < \epsilon.$$

3. Π_1 has full (weak) support.
4. f_t is continuous.

Theorem (Bhattacharya & Dunson (2010, 2011a))

Under assumptions **1-4**, *for any* $\epsilon > 0$,

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

which implies that f_t *is in the Kullback-Leibler (KL) support of* Π.

Let $\mathbf{X}_n = X_1, \dots X_n$ be iid sample from f_t.

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 $\Pi(U|\mathbf{X}_n) \to 1$ as $n \to \infty$ a.s. F_t^∞.

$$\Pi(U^c|\mathbf{X}_n) = \frac{\int_{U^c} \prod_{i=1}^n \frac{f(X_i)}{f_0(X_i)} \Pi(df)}{\int \prod_1^n \frac{f(X_i)}{f_0(X_i)} \Pi(df)}.$$

- Condition (1), known as the KL condition, ensures that the denominator isn't exponentially small while condition (2) implies that the numerator is.

- If U is a weakly open neighborhood of f_t, condition (2) always satisfied.
- Hence WPC under Assumptions 1-4.

- Use a Dirichlet Process prior $\Pi_{11} = DP(w_0 P_0)$ on P and an independent prior Π_{12} on ν as prior Π_1 choice.
- Use the stick breaking representation for P (*Sethuraman, 1994*) to write $f(.;P,\nu)$ as $f(x) = \sum_{i=1}^{\infty} w_i K(x;\mu_i,\nu)$ where $w_j = V_j \prod_{i=1}^{j-1}(1 - V_i)$, V_j iid Beta$(1, w_0)$, and μ_j iid P_0.
- Introduce cluster labels $\mathbf{S}_n = S_1, \ldots, S_n$ iid with $S_1 = j$ w.p. w_j for $j = 1, \ldots, \infty$, such that $X_i \sim K(.;\mu_{S_i},\nu)$, $i = 1, \ldots, n$, independently given parameters $\mathbf{S}_n$, μ, $\mathbf{V}$ and ν.
- The posterior is then $\Pi(\mathbf{S}_n, \mu, \mathbf{V}, \nu | \mathbf{X}) = \Pi_{12}(d\nu)\left(\prod_{j=1}^{\infty} P_0(d\mu_j)\mathrm{Be}(V_j;1,w_0)\right)\left(\prod_i w_{S_i} K(X_i;\mu_{S_i},\nu)\right)$.

Can perform a full conditional Gibbs sampling from the posterior of $(\mathbf{S}, \mu, \mathbf{V}, \nu)$ and hence f as follows.

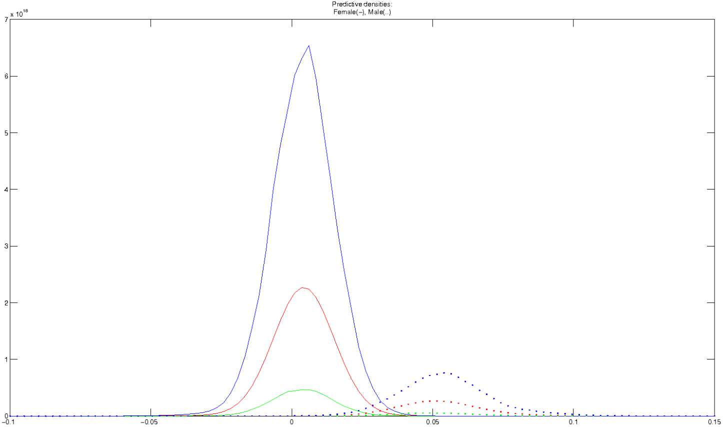
1. Update $\{S_i\}_1^n$ from the multinomial conditional posterior distribution with $\Pr(S_i = j) \propto w_j K(X_i; \mu_j, \nu)$ for $j = 1, \ldots, \infty$. To make the total number of clusters finite, either truncate to a fixed large value, or introduce latent uniform variables $u_1, \ldots, u_n$ and replace w_{S_i} by $I(u_i < w_{S_i})$, so that at any iteration the total cluster number becomes random but finite.
2. Update occupied cluster locations μ_j, $j \leq \max(S)$ from the conditional posterior $\propto P_0(d\mu_j) \prod_{i:S_i=j} K(X_i; \mu_j, \nu)$.
3. Update ν by sampling from the conditional posterior $\propto \Pi_{12}(d\nu) \prod_{i=1}^n K(X_i; \mu_{S_i}, \nu)$.
4. Update the stick-breaking random variables V_j, $j \leq \max(S)$ from $\text{Beta}(1 + \sum_i \text{I}(S_i = j), w_0 + \sum_i \text{I}(S_i > j))$.

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.

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 mean shapes.

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



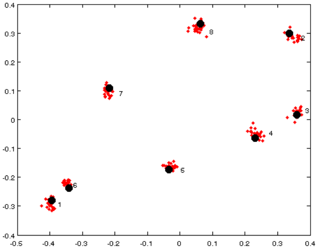
- Consider the data on gorilla skull shapes.
- From the shape density estimates for the male and female groups, can predict the gender from shape via np discriminant analysis.
- Assume the probability of being female is 0.5.
- Letting $f_1(m)$ and $f_2(m)$ denote the female and male shape densities, the conditional probability of being female given shape m is $p(m) = f_1(m)/\{f_1(m) + f_2(m)\}$.
- To estimate the posterior probability, estimate $f_i(m)$ across Markov chain iterations to obtain $\hat{p}(m)$.
- The prior Π_1 taken to be $DP(w_0 P_0) \otimes \text{Gam}(1.01, 1e-3)$ with $w_0 = 1$, $P_0 = CW(\mu_0, 1e-03)$- the complex Watson density, and μ_0 being the sample mean for the group under consideration.

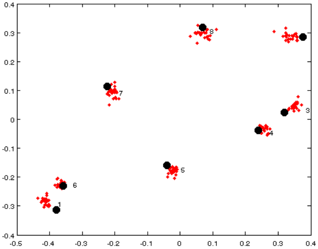
- The kernel K is also the complex Watson, resulting in conjugacy.
- To test the performance of the classifier, randomly partition the sample into training-test samples, using training samples, get the classifier and apply it to the test data.

Estimated posterior probabilities of being female for each gorilla in test sample along with a 95% credible interval (CI) for $p(m)$ for one such partition. Also shown is the distance between the sample shape m and the female($\hat{\mu}_1$), male($\hat{\mu}_2$) sample means.

gender	$\hat{p}(m)$	95% CI	$d_E(m, \hat{\mu}_1)$	$d_E(m, \hat{\mu}_2)$
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.





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

- When U is a total variation neighborhood of f_t, uniformly exponentially consistent sequence of tests seperating f_t and U^c do not exist in general. Hence Schwartz theorem does not apply.
- For $\mathcal{F} \subseteq \mathcal{D}(M)$ and $\epsilon > 0$, the log L_1-metric entropy $N(\epsilon, \mathcal{F})$ is the log of the minimum number of ϵ-sized (or smaller) L_1 subsets needed to cover $\mathcal{F}$.

Theorem (Barron(1999), Ghosal et.al.(1999))

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 $n \to \infty \ \forall \ \epsilon > 0$, then for any L_1 neighborhood U of f_t, the numerator of $\Pi(U^c|\mathbf{X}_n)$ decays exponentially fast a.s. Hence if f_t is in the KL support of Π, then $\Pi(U|\mathbf{X}_n)$ converges to 1 a.s.

Assume there exists a continuous function $\phi : N \to [0, \infty)$ for which the following assumptions hold.

1. There exists $\kappa_1, a_1, A_1 > 0$ s.t. $\forall\ \kappa \geq \kappa_1$, $\mu_1, \mu_2 \in M$,

$$\sup_{\phi(\nu)\leq\kappa} \left\|K(\mu_1, \nu) - K(\mu_2, \nu)\right\| \leq A_1 \kappa^{a_1} \rho(\mu_1, \mu_2),$$

   $\|.\|$ denoting the L_1 distance.

2. There exists $a_2, A_2 > 0$ s.t. $\forall\ \nu_1, \nu_2 \in \phi^{-1}[0, \kappa]$, $\kappa \geq \kappa_1$,

$$\sup_{\mu} \left\|K(\mu, \nu_1) - K(\mu, \nu_2)\right\| \leq A_2 \kappa^{a_2} \rho_2(\nu_1, \nu_2),$$

   ρ_2 metrizing the topology of N.

3. 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 $\epsilon > 0$.

4. 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 some $b_2 > 0$ (independent of κ and ϵ).

Theorem (Bhattacharya & Dunson (2011a))

For a +ve sequence $\{\kappa_n\}$ diverging to ∞, define

$$\mathcal{D}_n = \left\{f(P,\nu) : \phi(\nu) \leq \kappa_n\right\}.$$

Under above assumptions, given any $\epsilon > 0$, for n sufficiently large, $N(\epsilon, \mathcal{D}_n) \leq C(\epsilon)\kappa_n^{a_1 a_3}$ for some $C(\epsilon) > 0$. Hence $N(\epsilon, \mathcal{D}_n)$ is $o(n)$ whenever κ_n is $o\left(n^{(a_1 a_3)^{-1}}\right)$. SPC therefore follows under the additional assumptions for WPC and if $\Pi_1(\mathcal{M}(M) \times \phi^{-1}(n^a, \infty)) < \exp(-n\beta)$ f.s. $a < (a_1 a_3)^{-1}$, $\beta > 0$.

- When using a location-scale kernel, i.e., $\nu \in (0, \infty)$, choose a prior $\Pi_1 = \Pi_{11} \otimes \Pi_{12}$ having full support, set ϕ to be the identity map.
- Then a choice for Π_{12} for which assumptions for SPC are satisfied is a Weibull density $\mathit{Weib}(\nu; \alpha, \beta) \propto \nu^{\alpha-1} \exp(-\beta\nu^{\alpha})$ whenever the shape parameter α exceeds $a_1 a_3$ (a_1, a_3 as in Assumptions **1** and **3**).

- Let the supp. of the data generating density f_t be M- a compact subset of $\mathbb{R}^d$.
- The kernel can have non-compact supp. like Gaussian. Then $\nu = \Sigma^{-1}$, $N = M^+(d)$ - the space of all p.d. matrices, or a subset of it.
- Assumptions for WPC are satisfied, for example, when P and ν are independent under Π_1, with the prior Π_{11} on P including all densities supported on M- $\mathcal{D}(M)$ in its support, while the prior Π_{12} on ν containing all p.d. matrices with very large eigen-values, i.e. $\forall\ \kappa > 0$, there exists a ν with $\lambda_1(\nu) > \kappa$ in its support.
- Take ϕ to be the largest eigen-value function, $\phi(\nu) = \lambda_d(\nu)$.

Theorem (Bhattacharya & Dunson (2011a))

Assumption 1 is satisfied with $a_1 = 1/2$. Ass. 2 is satisfied once ν has e-vals. bdd. below, i.e. for some $\lambda_1 > 0$, $N = \{\nu \in M^+(d) : \lambda_1(\nu) \geq \lambda_1\}$. The space $M^+(d)$ (and hence N) satisfies Ass.4 while M satisfies Ass.3 with $a_3 = d$. With independent priors Π_{11} & Π_{12} on P & ν, SPC follows once supp$(\Pi_{11}) = \mathcal{M}(M)$ & $\Pi_{12}\big(\{\nu \in N : \lambda_d(\nu) > n^a\}\big) < \exp(-n\beta)$ for some $a < 2/d$ and $\beta > 0$.

- A Wishart prior on ν,

  $$\Pi_{12}(\nu; a, b) = 2^{-db/2}\Gamma_d^{-1}(b/2)a^{db/2}\exp(-a/2\mathrm{Tr}(\nu))\det(\nu)^{(b-d-1)/2}$$

  denoted as Wish$(a^{-1}I_d, b)$ does not satisfy the theorem (unless $d = 1$).
- Here $\Gamma_d(.)$ denotes the *multivariate gamma function* defined as

  $$\Gamma_d(b/2) = \int_{M^+(d)} \exp(-\mathrm{Tr}(\nu))\det(\nu)^{(b-d-1)/2}d\nu.$$

- Instead set $\nu = \Lambda^\alpha$ for any $\alpha \in (0, 2/d)$ with Λ following a Wishart distribution restricted to N.

- k points or landmarks extracted from a 2D image.
- Represented by a complex k-vector $x(\in \mathcal{C}^k)$.
- 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: $z = x_c/\|x_c\|$ and remove scale effect. This z 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

$$m = [z] = \{e^{i\theta} z : \ -\pi < \theta \le \pi\}.$$

- Σ_2^k is the collection of all such shapes.
- Σ_2^k is a compact Riemannian manifold of (real) dim. $2k-4$.

- Let K be the complex-Watson density $\mathrm{CW}([z];[\mu],\nu) = c^{-1}(\nu)\exp(\nu|z^*\mu|^2)$, $z,\mu \in CS^{k-1}$, $\nu > 0$.
- It has mean $[\mu]$.
- ν is a measure of concentration - as $\nu \to 0$, CW converges to the uniform density and as $\nu \to \infty$, CW converges weakly to $\delta_{[\mu]}$ uniformly in μ.
- The kernel satisfies conditions for W.P.C. using a DP-Gamma prior (*Bhattacharya & Dunson 2010*).

Strong Posterior Consistency I

Assume there exists a continuous function $\phi : N \to [0, \infty)$ for which the following assumptions hold.

1. There exists $\kappa_1, a_1, A_1 > 0$ s.t. $\forall\ \kappa \geq \kappa_1$, $\mu_1, \mu_2 \in M$,

$$\sup_{\phi(\nu)\leq\kappa} \left\|K(\mu_1, \nu) - K(\mu_2, \nu)\right\| \leq A_1 \kappa^{a_1} \rho(\mu_1, \mu_2),$$

   $\|.\|$ denoting the L_1 distance.

2. There exists $a_2, A_2 > 0$ s.t. $\forall\ \nu_1, \nu_2 \in \phi^{-1}[0, \kappa]$, $\kappa \geq \kappa_1$,

$$\sup_{\mu} \left\|K(\mu, \nu_1) - K(\mu, \nu_2)\right\| \leq A_2 \kappa^{a_2} \rho_2(\nu_1, \nu_2),$$

   ρ_2 metrizing the topology of N.

Strong Posterior Consistency II

3. 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 $\epsilon > 0$.
4. 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 some $b_2 > 0$ (independent of κ and ϵ).

- The complex Watson kernel Ass.1 for SPC with $a_1 = k - 1$ and Ass.2 with $a_2 = 3k - 8$.
- Σ_2^k satisfies Ass.3 with $a_3 = 2k - 3$.
- Hence SPC holds with a DP-Weibull prior, for Weibull with shape parameter exceeding $(2k - 3)(k - 1)$ (*Bhattacharya & Dunson 2011a*).

References I

- BHATTACHARYA, A. (2008). *Sankhya* **70**-A, Part **2** 223-266.
- BHATTACHARYA, A. AND BHATTACHARYA, R. (BB) (2008a). Nonparametric statistics on manifolds with application to shape spaces. *IMS Collections* **3** 282-301.
- BHATTACHARYA, A. AND BHATTACHARYA, R. (BB) (2008b). Statistical on Riemannian manifolds: asymptotic distribution and curvature. *Proc. Amer. Math. Soc.* **136** 2957-2967.
- BHATTACHARYA, A. AND BHATTACHARYA, R. (BB) (2009). Statistical on manifolds with application to shape spaces. *Perspectives in Math. Sci., I.S.I.* 41-70.

References II

- BHATTACHARYA, A. AND DUNSON, D. (BD) (2010). Nonparametric Bayesian Density Estimation on Manifolds with applications to Planar Shapes. *Biometrika* **97**-4 851-865.
- BHATTACHARYA, A. AND DUNSON, D. (BD) (2011a). Strong consistency of nonparametric Bayes density estimation on compact metric spaces with applications to specific manifolds. *Ann. of Instit. of Statist. Math.*. In Press.
- BHATTACHARYA, A. AND DUNSON, D. (BD) (2011b). Nonparametric Bayes Classification and Testing on Manifolds with Applications on Hypersphere. Submitted.

References III

BARRON, A.; SCHERVISH, J.; WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. of Stat.* **27** 2 536-561.

GHOSAL, S.; GHOSH, J. K.; RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. of Stat.* **27** 143-158.

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

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639-50.

References IV

WU, Y. AND GHOSAL, S. (2010). L_1- consistency of Dirichlet mixtures in multivariate Bayesian density estimation on Bayes procedures. *Jour. of Multiv. Analysis* **101** 2411-19.