Nonparametric Bayes Modeling on Manifolds

Abhishek Bhattacharya
Department of Statistics, Duke University
Joint work with Prof. D.Dunson

April 17 2010

Contents I

Density Estimation via Kernel Mixtures

- Standard method for Bayes density estimation relies on

$$f(y; P) = \int K(y; \mu, \kappa) P(d\mu d\kappa), \quad y \in \Re,$$

$K(\cdot)$= kernel, P= mixture distribution

Density Estimation via Kernel Mixtures

- Standard method for Bayes density estimation relies on

$$f(y;P) = \int K(y;\mu,\kappa)P(d\mu d\kappa), \quad y \in \Re,$$

  $K(\cdot)$= kernel, P= mixture distribution
- Prior $f \sim \Pi$ is induced through $P \sim \Pi_1$

Density Estimation via Kernel Mixtures

- Standard method for Bayes density estimation relies on

$$f(y;P) = \int K(y;\mu,\kappa)P(d\mu d\kappa), \quad y \in \Re,$$

  $K(\cdot)$= kernel, P= mixture distribution
- Prior $f \sim \Pi$ is induced through $P \sim \Pi_1$
- Dirichlet process (DP, Ferguson, 73, 74) standard choice
- Leads to DP mixture (Lo, 84; Escobar & West, 95)

Large Support

- In Bayesian inference “nonparametric” implies large support

Large Support

- In Bayesian inference "nonparametric" implies large support
- $\mathcal{F}$ = set of densities wrt Lesbesgue measure on $\Re$

Large Support

- In Bayesian inference "nonparametric" implies large support
- $\mathcal{F}$ = set of densities wrt Lesbesgue measure on $\Re$
- Ideally prior $f \sim \Pi$ assigns positive probability to arbitrarily small neighborhoods of any $f_0 \in \mathcal{F}$

Large Support

- In Bayesian inference "nonparametric" implies large support
- $\mathcal{F}$ = set of densities wrt Lesbesgue measure on $\Re$
- Ideally prior $f \sim \Pi$ assigns positive probability to arbitrarily small neighborhoods of any $f_0 \in \mathcal{F}$
- Allows uncertainty in our prior beliefs & posterior consistency under some conditions

Large Support

- In Bayesian inference "nonparametric" implies large support
- $\mathcal{F}$ = set of densities wrt Lesbesgue measure on $\Re$
- Ideally prior $f \sim \Pi$ assigns positive probability to arbitrarily small neighborhoods of any $f_0 \in \mathcal{F}$
- Allows uncertainty in our prior beliefs & posterior consistency under some conditions
- If Π assigns positive probability to all KL neighborhoods of the true f_0, weak posterior consistency results

Some Tools & Results in Euclidean Spaces

- When data have support in $\Re^p$ or some subset, tools available for posterior computation & inferences

Some Tools & Results in Euclidean Spaces

- When data have support in $\Re^p$ or some subset, tools available for posterior computation & inferences
- Literature on conditions for large support & consistency

Some Tools & Results in Euclidean Spaces

- When data have support in $\Re^p$ or some subset, tools available for posterior computation & inferences
- Literature on conditions for large support & consistency
- Rich methods literature but little theory in multivariate Euclidean spaces

Some Tools & Results in Euclidean Spaces

- When data have support in $\Re^p$ or some subset, tools available for posterior computation & inferences
- Literature on conditions for large support & consistency
- Rich methods literature but little theory in multivariate Euclidean spaces
- Very little done in non-Euclidean spaces

Kernel Mixtures on Compact Metric Spaces

- M = compact metric space, with X a random variable on M

Kernel Mixtures on Compact Metric Spaces

- M = compact metric space, with X a random variable on M
- Assume that the distribution of X has a density wrt base measure λ on M

Kernel Mixtures on Compact Metric Spaces

- M = compact metric space, with X a random variable on M
- Assume that the distribution of X has a density wrt base measure λ on M
- Let $K(m; \mu, \kappa)$ denote a probability kernel on M with location $\mu \in M$ and inverse-scale/precision $\kappa \in \Re^+$

Kernel Mixtures on Compact Metric Spaces

- M = compact metric space, with X a random variable on M
- Assume that the distribution of X has a density wrt base measure λ on M
- Let $K(m; \mu, \kappa)$ denote a probability kernel on M with location $\mu \in M$ and inverse-scale/precision $\kappa \in \Re^+$
- We focus on the kernel mixture model with

$$f(m; P, \kappa) = \int_M K(m; \mu, \kappa) P(d\mu), \quad (P, \kappa) \sim \Pi_1$$

Is Your Model "Good"??

- For a particular choice of M (e.g, a compact Riemannian manifold, such as the hypersphere), we can choose a kernel K and prior Π_1

Is Your Model "Good"??

- For a particular choice of M (e.g, a compact Riemannian manifold, such as the hypersphere), we can choose a kernel K and prior Π_1
- For example, Lennox et al. (2009) proposed a DPM of bivariate von Mises distributions for protein configuration angles

Is Your Model "Good"??

- For a particular choice of M (e.g, a compact Riemannian manifold, such as the hypersphere), we can choose a kernel K and prior Π_1
- For example, Lennox et al. (2009) proposed a DPM of bivariate von Mises distributions for protein configuration angles
- Question: Is the model flexible enough to approximate any density on M & can we at least estimate this density consistently

Is Your Model "Good"??

- For a particular choice of M (e.g, a compact Riemannian manifold, such as the hypersphere), we can choose a kernel K and prior Π_1
- For example, Lennox et al. (2009) proposed a DPM of bivariate von Mises distributions for protein configuration angles
- Question: Is the model flexible enough to approximate any density on M & can we at least estimate this density consistently
- Not at all clear for mixtures of arbitrary kernels - we want simple sufficient conditions to check

Assumptions for Large Support & Consistency

Let P_t denote the true distribution and f_t be its density. Assume

- A1 K is continuous on $M \times M \times \Re^+$.
- A2 For any cont. ϕ

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

- 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.

Theorem (Bhattacharya & Dunson, 2010a)

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

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

Corollary (BD, 2010a)

*Under assumptions **A1**-**A5**, KL condition satisfied, i.e.*

$$\Pi(\{f : KL(f_t, f) < \epsilon\}) > 0.$$

Using Schwartz theorem (*Schwartz, 1965*) weak posterior consistency follows.

Assumptions for Strong Posterior Consistency

A6 $\Pi_1(\mathcal{M}(M) \times (n^{(1/a)}, \infty)) < \exp(-n\beta)$ for some $a > a_1 a_3$ and $\beta > 0$, where

Assumptions for Strong Posterior Consistency

A6 $\Pi_1(\mathcal{M}(M) \times (n^{(1/a)}, \infty)) < \exp(-n\beta)$ for some $a > a_1 a_3$ and $\beta > 0$, where

A7 $\exists\ \mathcal{K}_1, a_1, A_1 > 0$ such that for all $\mathcal{K} \geq \mathcal{K}_1$, $\mu, \nu \in M$,

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

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}.

Strong Posterior Consistency

Theorem (Bhattacharya & Dunson, 2010b)

*Under assumptions **A1**-**A8**, strong posterior consistency holds, i.e. the posterior probability of any total variation neighborhood of f_t converges to 1 almost surely.*

If precision prior π_1 is the Weibull density: $\pi_1(\kappa) \propto \kappa^{c-1} \exp(-b\kappa^a)$ with $a > a_1 a_3$, then **A6** holds, and s.p.c. follows.

- For $M = \Re^d$, need Weibull prior on κ with $a > d^2/2$ which puts very little tail mass and that 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 $\Re^+$

A9 Π_{11} has full support

A9 Π_{11} has full support

A10 For any $\beta > 0$, there exists a $\kappa_0 \geq 0$, s.t. $\forall\ \kappa \geq \kappa_0$, $\liminf_n \exp(n\beta)\pi_n(\kappa) = \infty$

A9 Π_{11} has full support

A10 For any $\beta > 0$, there exists a $\kappa_0 \geq 0$, s.t. $\forall\ \kappa \geq \kappa_0$, $\liminf_n \exp(n\beta)\pi_n(\kappa) = \infty$

A11 For some $\beta_0 > 0$ and $a > a_1 a_3$, $\lim_n \exp(n\beta_0)\pi_n\{(n^{1/a}, \infty)\} = 0$.

A9 Π_{11} has full support

A10 For any $\beta > 0$, there exists a $\kappa_0 \geq 0$, s.t. $\forall\ \kappa \geq \kappa_0$, $\liminf_n \exp(n\beta)\pi_n(\kappa) = \infty$

A11 For some $\beta_0 > 0$ and $a > a_1 a_3$, $\lim_n \exp(n\beta_0)\pi_n\{(n^{1/a}, \infty)\} = 0$.

Theorem (BD, 2010b)

*Under assumptions **A1-A11** weak and strong p.c. hold.*

For Gamma prior $\pi_n(\kappa) \propto \kappa^{\alpha-1} \exp(-\beta_n \kappa)$, need
$n^{1-1/a} << \beta_n << n$
e.g. $\beta_n = \beta n / \log(n)$

Introduction to 2D Similarity Shapes

- k landmark points are picked from an object in 2D - similarity shape removes effects of translation, rotation & scaling.

Introduction to 2D Similarity Shapes

- k landmark points are picked from an object in 2D - similarity shape removes effects of translation, rotation & scaling.
- Denote k-ad by complex k vector $z = (z_1, \ldots, z_k)' \in C^k$ - remove translation by subtracting centroid, $z_c = z - \overline{z}$.

Introduction to 2D Similarity Shapes

- k landmark points are picked from an object in 2D - similarity shape removes effects of translation, rotation & scaling.
- Denote k-ad by complex k vector $z = (z_1, \ldots, z_k)' \in C^k$ - remove translation by subtracting centroid, $z_c = z - \overline{z}$.
- Remove scaling by normalizing coordinates of z_c to obtain point w on complex unit sphere - referred to as preshape

Introduction to 2D Similarity Shapes

- k landmark points are picked from an object in 2D - similarity shape removes effects of translation, rotation & scaling.
- Denote k-ad by complex k vector $z = (z_1, \ldots, z_k)' \in C^k$ - remove translation by subtracting centroid, $z_c = z - \overline{z}$.
- Remove scaling by normalizing coordinates of z_c to obtain point w on complex unit sphere - referred to as preshape
- Similarity shape of z is orbit of w under all rotations in 2D - the space of all such orbits is the planar shape space Σ_2^k

Example: Shapes of Gorilla Skulls

- 8 landmarks chosen on the midline plane of the 2D image of 29 male and 30 female gorilla skulls (Dryden and Mardia, 98).

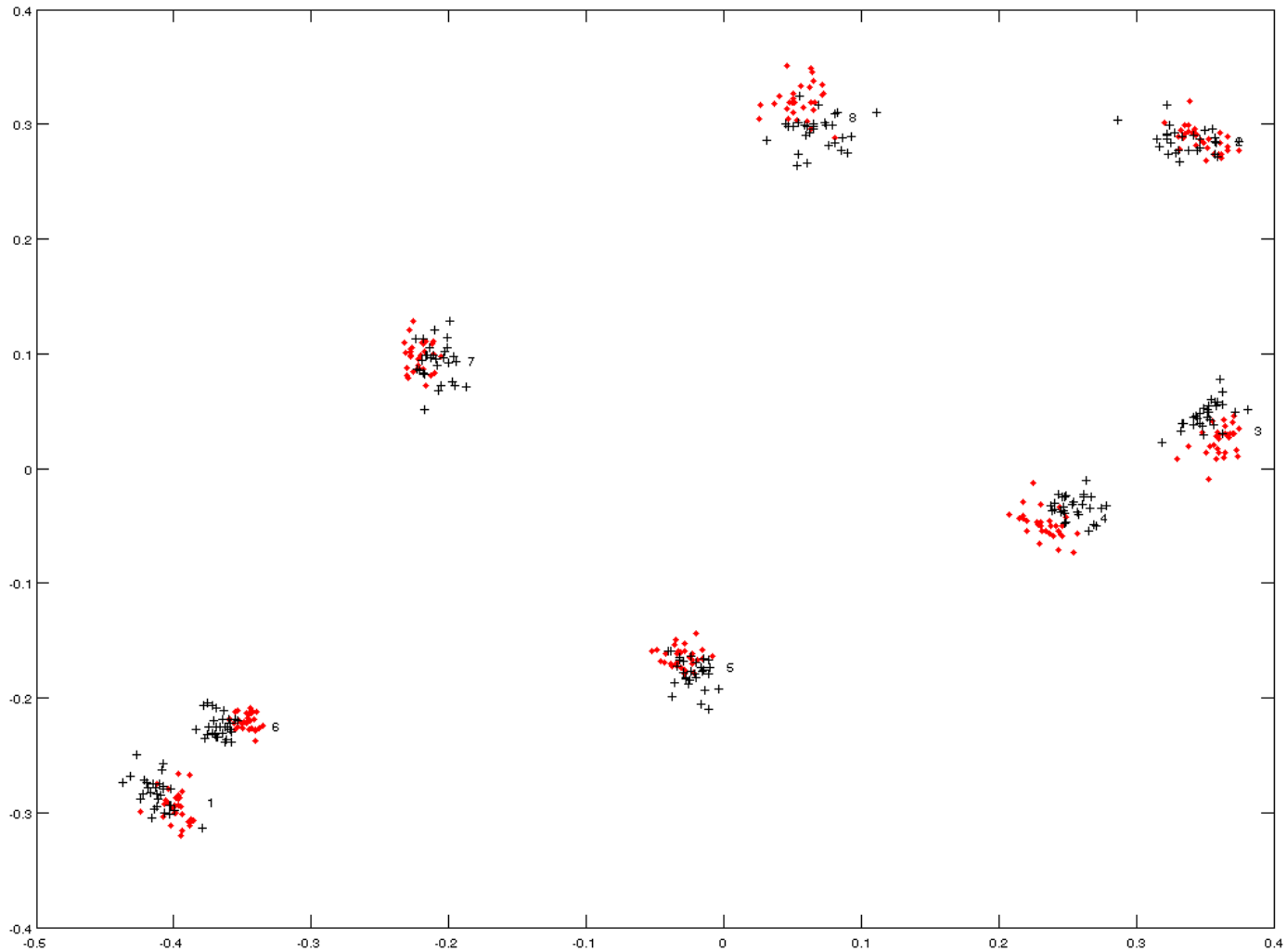
Example: Shapes of Gorilla Skulls

- 8 landmarks chosen on the midline plane of the 2D image of 29 male and 30 female gorilla skulls (Dryden and Mardia, 98).
- Goal: Study the shapes of the skulls and use that to detect difference in shapes between the sexes.

Example: Shapes of Gorilla Skulls

- 8 landmarks chosen on the midline plane of the 2D image of 29 male and 30 female gorilla skulls (Dryden and Mardia, 98).
- Goal: Study the shapes of the skulls and use that to detect difference in shapes between the sexes.
- Two mutually independent iid sample of planar shapes of sizes 29 and 30 on Σ_2^k, $k = 8$ (Dimension = 12).

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



Geometry of planar shape space

- Σ_2^k is a compact Riemannian manifold of dimension $2k - 4$

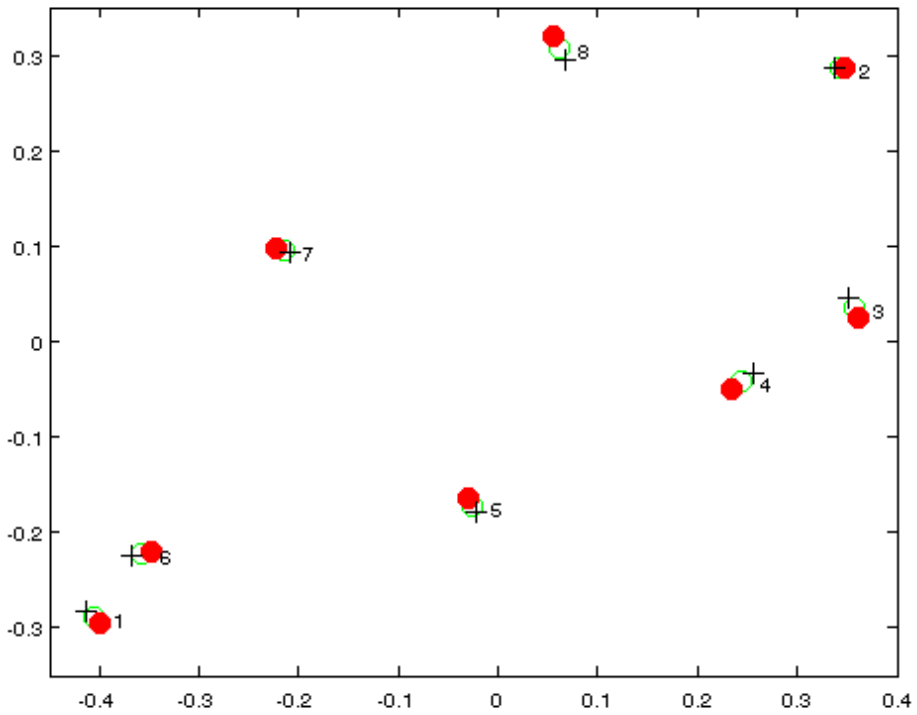
Geometry of planar shape space

- Σ_2^k is a compact Riemannian manifold of dimension $2k - 4$
- geodesic distance can be defined between two shapes, an embedding into Euclidean space induces the "extrinsic distance".

Geometry of planar shape space

- Σ_2^k is a compact Riemannian manifold of dimension $2k - 4$
- geodesic distance can be defined between two shapes, an embedding into Euclidean space induces the "extrinsic distance".
- Using an integrated distance square loss function, can define a notion of mean of a probability - intrinsic (corresponding to geodesic distance) or extrinsic (*Bhattacharya & Patrangenaru, 2003*).

Gorilla Skulls: Extrinsic Mean Shapes Plot



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

Density modelling on Σ_2^k

Density modelling on Σ_2^k

- To avoid relying on tangent space approximations, we work with the invariant volume form $V(dm)$ on $M = \Sigma_2^k$

Density modelling on Σ_2^k

- To avoid relying on tangent space approximations, we work with the invariant volume form $V(dm)$ on $M = \Sigma_2^k$
- To specify a np Bayes density model for a shape density, it remains to choose a kernel K and mixing prior Π_1.

Density modelling on Σ_2^k

- To avoid relying on tangent space approximations, we work with the invariant volume form $V(dm)$ on $M = \Sigma_2^k$
- To specify a np Bayes density model for a shape density, it remains to choose a kernel K and mixing prior Π_1.
- Simple parametric kernel corresponds to the complex Watson distribution (Dryden & Mardia, 98)

$$\mathrm{CW}(m; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa |z^* v|^2),$$

with z, v preshapes of $m, \mu \in \Sigma_2^k$, respectively, and $*$ the complex conjugate transpose

Density modelling on Σ_2^k

- To avoid relying on tangent space approximations, we work with the invariant volume form $V(dm)$ on $M = \Sigma_2^k$
- To specify a np Bayes density model for a shape density, it remains to choose a kernel K and mixing prior Π_1.
- Simple parametric kernel corresponds to the complex Watson distribution (Dryden & Mardia, 98)

$$\mathrm{CW}(m; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa |z^* v|^2),$$

with z, v preshapes of $m, \mu \in \Sigma_2^k$, respectively, and $*$ the complex conjugate transpose

- μ= extrinsic mean, κ- measure of concentration, $c(\kappa)$= norming constant

DPMs of Complex Watsons

- We use the kernel location mixture model

$$f(m; P, \kappa) = \int \mathrm{CW}(m; \mu, \kappa) P(d\mu), \quad m \in \Sigma_2^k,$$

DPMs of Complex Watsons

- We use the kernel location mixture model

$$f(m; P, \kappa) = \int \mathrm{CW}(m; \mu, \kappa) P(d\mu), \quad m \in \Sigma_2^k,$$

- We let $P \sim \mathrm{DP}(\omega_0 P_0)$, with $P_0 = \mathrm{CW}(\mu_0, \sigma_0)$ corresponding to a complex Watson base

DPMs of Complex Watsons

- We use the kernel location mixture model

$$f(m; P, \kappa) = \int \mathrm{CW}(m; \mu, \kappa) P(d\mu), \quad m \in \Sigma_2^k,$$

- We let $P \sim \mathrm{DP}(\omega_0 P_0)$, with $P_0 = \mathrm{CW}(\mu_0, \sigma_0)$ corresponding to a complex Watson base
- We let $\kappa \sim \mathrm{Ga}(a, b)$.

Theoretical Properties & Computation

- The kernel and priors can be shown to satisfy the sufficient conditions in our theorems to ensure L_∞ and KL support

Theoretical Properties & Computation

- The kernel and priors can be shown to satisfy the sufficient conditions in our theorems to ensure L_∞ and KL support
- This directly implies weak posterior consistency at all continuous, positive true densities f_0

Theoretical Properties & Computation

- The kernel and priors can be shown to satisfy the sufficient conditions in our theorems to ensure L_∞ and KL support
- This directly implies weak posterior consistency at all continuous, positive true densities f_0
- Strong consistency follows if instead of Gamma, use a Weibull prior with shape parameter exceeding $(2k-3)(k-1)$ (Bhattacharya & Dunson, 2010b).
- S.P.C. also follows with a gamma prior if the scale parameter for the gamma depends on n in an appropriate manner.

Theoretical Properties & Computation

- The kernel and priors can be shown to satisfy the sufficient conditions in our theorems to ensure L_∞ and KL support
- This directly implies weak posterior consistency at all continuous, positive true densities f_0
- Strong consistency follows if instead of Gamma, use a Weibull prior with shape parameter exceeding $(2k-3)(k-1)$ (Bhattacharya & Dunson, 2010b).
- S.P.C. also follows with a gamma prior if the scale parameter for the gamma depends on n in an appropriate manner.
- A simple exact blocked Gibbs sampler (*Papaspiliopoulos, 08*) can be used for posterior computaton

NP Bayes Classification through Joint Modeling

- Let Y be a categorical variable taking values in $\mathbb{Y} = \{1, 2, \ldots, L\}$. eg. Gorilla gender, presence/absence of disease, animal/plant species, etc

NP Bayes Classification through Joint Modeling

- Let Y be a categorical variable taking values in $\mathbb{Y} = \{1, 2, \ldots, L\}$. eg. Gorilla gender, presence/absence of disease, animal/plant species, etc
- Goal: nonparametric estimation of classification function $\Pr(Y = l \mid X = x)$ with X predictors lying on M

NP Bayes Classification through Joint Modeling

- Let Y be a categorical variable taking values in $\mathbb{Y} = \{1, 2, \ldots, L\}$. eg. Gorilla gender, presence/absence of disease, animal/plant species, etc
- Goal: nonparametric estimation of classification function $\Pr(Y = l \mid X = x)$ with X predictors lying on M
- Inspired by Müller et al. (1996)'s method of inducing a prior on $f(y|x)$ from a DPM of MVNs for the joint distribution $f(y, x)$

NP Bayes Classification through Joint Modeling

- Let Y be a categorical variable taking values in $\mathbb{Y} = \{1, 2, \ldots, L\}$. eg. Gorilla gender, presence/absence of disease, animal/plant species, etc
- Goal: nonparametric estimation of classification function $\Pr(Y = l \mid X = x)$ with X predictors lying on M
- Inspired by Müller et al. (1996)'s method of inducing a prior on $f(y|x)$ from a DPM of MVNs for the joint distribution $f(y, x)$
- Propose to nonparametrically model the joint of Y, X to induce a prior on the classification function

Joint Kernel Mixture Model

- Model the joint distribution of (X, Y) via

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

Joint Kernel Mixture Model

- Model the joint distribution of (X, Y) via

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

- $\nu = (\nu_1, \ldots, \nu_L)' \in S_{L-1}$ is a probability vector on the simplex $S_{L-1} = \{\nu \in [0, 1]^L : \sum \nu_l = 1\}$

Joint Kernel Mixture Model

- Model the joint distribution of (X, Y) via

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

- $\nu = (\nu_1, \ldots, \nu_L)' \in S_{L-1}$ is a probability vector on the simplex $S_{L-1} = \{\nu \in [0, 1]^L : \sum \nu_l = 1\}$
- $K(\cdot; \mu, \kappa)$ is a kernel located at $\mu \in M$ with precision $\kappa \in \Re^+$

Joint Kernel Mixture Model

- Model the joint distribution of (X, Y) via

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

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

Prior Selection

- Set a prior Π_1 on (P, κ) such as $\text{DP}(\omega_0 P_0) \otimes \pi_1$ with P_0 and π_1 a dist. on $M \times S_{c-1}$ and $\mathbb{R}^+$ respectively

Prior Selection

- Set a prior Π_1 on (P, κ) such as $\mathrm{DP}(\omega_0 P_0) \otimes \pi_1$ with P_0 and π_1 a dist. on $M \times S_{c-1}$ and $\mathbb{R}^+$ respectively
- This induces a prior Π on the joint density f of (X, Y) & hence on $\Pr(Y = j \mid X = x)$

Prior Selection

- Set a prior Π_1 on (P, κ) such as $\text{DP}(\omega_0 P_0) \otimes \pi_1$ with P_0 and π_1 a dist. on $M \times S_{c-1}$ and $\mathbb{R}^+$ respectively
- This induces a prior Π on the joint density f of (X, Y) & hence on $\Pr(Y = j \mid X = x)$
- We can use this for classification of new subjects based on features x_{n+1}

Prior Selection

- Set a prior Π_1 on (P, κ) such as $\text{DP}(\omega_0 P_0) \otimes \pi_1$ with P_0 and π_1 a dist. on $M \times S_{c-1}$ and $\mathbb{R}^+$ respectively
- This induces a prior Π on the joint density f of (X, Y) & hence on $\Pr(Y = j \mid X = x)$
- We can use this for classification of new subjects based on features x_{n+1}
- Under similar conditions to those discussed above, we obtain L_∞, KL support & L_1 consistency for the classification function (Bhattacharya & Dunson, 2010c).

- Let M be the unit sphere $S^d = \{x \in \Re^{d+1} : \sum_j x_j^2 = 1\}$.

- Let M be the unit sphere $S^d = \{x \in \Re^{d+1} : \sum_j x_j^2 = 1\}$.
- Take kernel K to be vMF density (*von Mises, 1918*)

$$\text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu)$$

- μ is the extrinsic mean, κ a measure of concentration

- Let M be the unit sphere $S^d = \{x \in \Re^{d+1} : \sum_j x_j^2 = 1\}$.
- Take kernel K to be vMF density (*von Mises, 1918*)

$$\text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu)$$

- μ is the extrinsic mean, κ a measure of concentration
- Let $\Pi_1 = \text{DP}(\omega_0 P_0) \otimes \pi_n$. Assume under P_0, μ and ν independent vMF and Dirichlet distributed and

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

- Let M be the unit sphere $S^d = \{x \in \Re^{d+1} : \sum_j x_j^2 = 1\}$.
- Take kernel K to be vMF density (*von Mises, 1918*)

$$\text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu)$$

- μ is the extrinsic mean, κ a measure of concentration
- Let $\Pi_1 = \text{DP}(\omega_0 P_0) \otimes \pi_n$. Assume under P_0, μ and ν independent vMF and Dirichlet distributed and

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

- This satisfies large support & consistency conditions (BD, 2010c)

- Let M be the unit sphere $S^d = \{x \in \Re^{d+1} : \sum_j x_j^2 = 1\}$.
- Take kernel K to be vMF density (*von Mises, 1918*)

$$\text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu)$$

- μ is the extrinsic mean, κ a measure of concentration
- Let $\Pi_1 = \text{DP}(\omega_0 P_0) \otimes \pi_n$. Assume under P_0, μ and ν independent vMF and Dirichlet distributed and

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

- This satisfies large support & consistency conditions (BD, 2010c)
- A simple exact block Gibbs sampler can be implemented with conjugate sampling steps

- Let M be the unit sphere $S^d = \{x \in \Re^{d+1} : \sum_j x_j^2 = 1\}$.
- Take kernel K to be vMF density (*von Mises, 1918*)

$$\text{vMF}(x; \mu, \kappa) = c^{-1}(\kappa) \exp(\kappa x' \mu)$$

- μ is the extrinsic mean, κ a measure of concentration
- Let $\Pi_1 = \text{DP}(\omega_0 P_0) \otimes \pi_n$. Assume under P_0, μ and ν independent vMF and Dirichlet distributed and

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

- This satisfies large support & consistency conditions (BD, 2010c)
- A simple exact block Gibbs sampler can be implemented with conjugate sampling steps
- Better performance than discriminant analysis methods based on mixtures of Gaussians

- Simulated 200 iid training data on $S^9 \times \mathbb{Y}$, $\mathbb{Y} = \{1, 2, 3\}$ from

$$(X, Y) \sim f_t(x, y) = (1/3) \sum_{l=1}^{3} I(y = l) \mathrm{vMF}(x; \mu_l, 200) \text{ where}$$

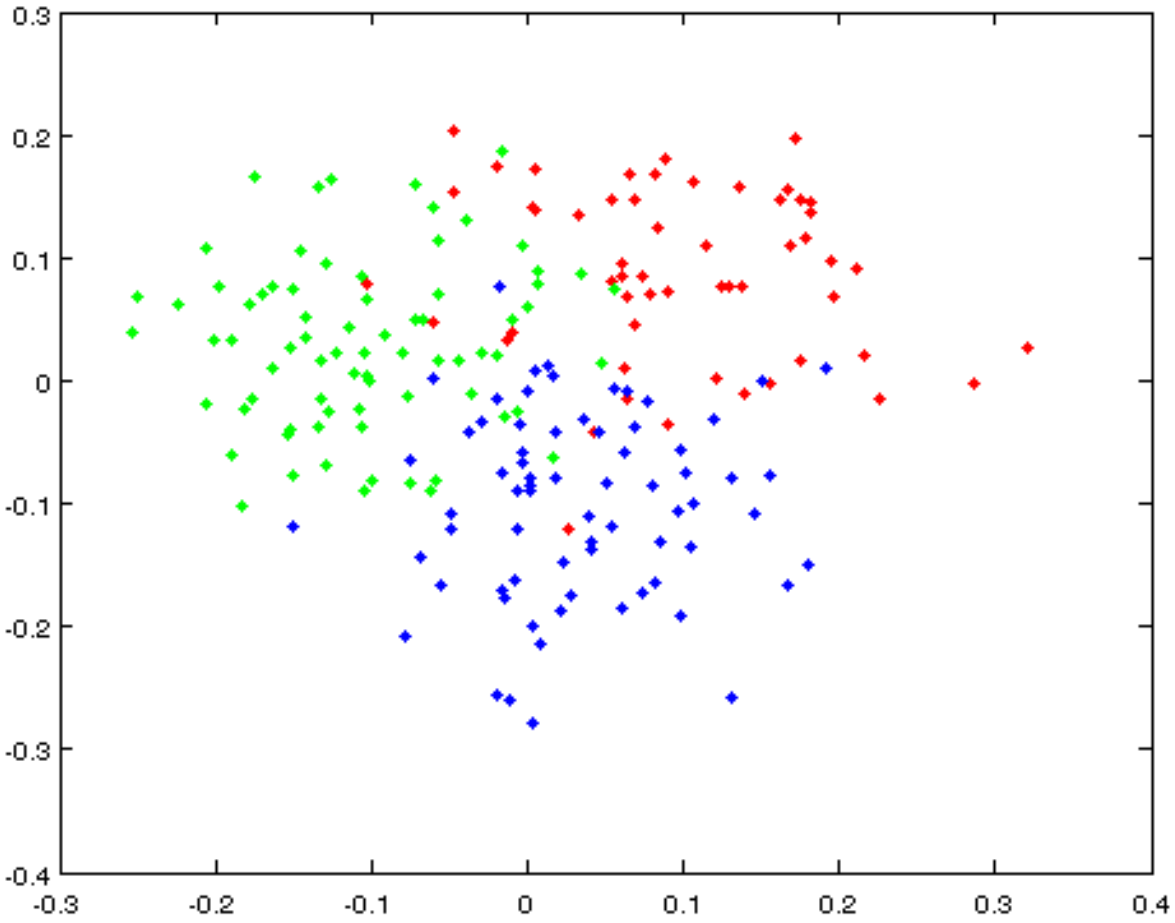
- $\mu_1 = (1, 0, \ldots)^T$, $\mu_j = \cos 0.2\mu_1 + \sin 0.2 v_j$, $j = 2, 3$, $v_2 = (0, 1, \ldots)^T$ and $v_3 = (0, 0.5, \sqrt{0.75}, 0, \ldots)^T$.

- Simulated 200 iid training data on $S^9 \times \mathbb{Y}$, $\mathbb{Y} = \{1, 2, 3\}$ from

$$(X, Y) \sim f_t(x, y) = (1/3) \sum_{l=1}^{3} I(y = l) \mathrm{vMF}(x; \mu_l, 200) \text{ where}$$

- $\mu_1 = (1, 0, \ldots)^T$, $\mu_j = \cos 0.2\mu_1 + \sin 0.2 v_j$, $j = 2, 3$, $v_2 = (0, 1, \ldots)^T$ and $v_3 = (0, 0.5, \sqrt{0.75}, 0, \ldots)^T$.
- Goal To estimate the conditional of Y given X and use that for classification

Training Sample Plot: 1st two principal tangent space coordinates. r:Y=1, b:Y=2, g:Y=3



Performance of Classifier

- To test the performance, we draw a test sample of size 100 from f_t, classify them and calculate the miss-classifications rates for each category
- They turn out to be 18.9%, 9.7% and 12.5% for categories 1, 2 and 3 respectively
- Overall percent of test data mis-labelled = 14%

Performance of Classifier

- To test the performance, we draw a test sample of size 100 from f_t, classify them and calculate the miss-classifications rates for each category
- They turn out to be 18.9%, 9.7% and 12.5% for categories 1, 2 and 3 respectively
- Overall percent of test data mis-labelled = 14%
- Corresponding values from fitting Gaussian mixtures (2 clusters) to each category:
- (21.6, 16.1, 28.1)%, Overall = 22%

Null and Alternative Hypothesis

- Instead of classifying a new feature, goal is to test whether the distribution of the features differs across the classes

Null and Alternative Hypothesis

- Instead of classifying a new feature, goal is to test whether the distribution of the features differs across the classes
- Test for independence between X and Y

Null and Alternative Hypothesis

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

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

Null and Alternative Hypothesis

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

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

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

Bayes Factor

- Under H_0 replace $P(d\mu d\nu)$ with $P_1(d\mu)P_2(d\nu)$ so that $f(x, y; P_1, P_2, \kappa) = f_X(x; P_1, \kappa) f_Y(y; P_2)$ where

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

Bayes Factor

- Under H_0 replace $P(d\mu d\nu)$ with $P_1(d\mu)P_2(d\nu)$ so that $f(x, y; P_1, P_2, \kappa) = f_X(x; P_1, \kappa)f_Y(y; P_2)$ 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).

Bayes Factor

- Under H_0 replace $P(d\mu d\nu)$ with $P_1(d\mu)P_2(d\nu)$ so that $f(x, y; P_1, P_2, \kappa) = f_X(x; P_1, \kappa)f_Y(y; P_2)$ 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

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

Bayes Factor Computation & Consistency

- BY selecting suitable DP mixture priors and
- introducing a latent variable z which is the indicator of accepting H_1, we devise a simple algorithm for computing BF (BD, 2010c)

Bayes Factor Computation & Consistency

- BY selecting suitable DP mixture priors and
- introducing a latent variable z which is the indicator of accepting H_1, we devise a simple algorithm for computing BF (BD, 2010c)
- BD(2010c) also proves consistency of the Bayes factor, $BF \to \infty$ a.s. if H_1 is true

Gorilla Skull Application

- Goal to test if the skull shape distribution differs accross gender.

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.

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 dists 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 dists for two sexes different.
- Next permute the labels randomly, so that we expect independence.

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 dists 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.

- Considered a broad class of kernel mixture models for density estimation, classification & testing in general spaces

- Considered a broad class of kernel mixture models for density estimation, classification & testing in general spaces
- Theory to verify that a prior leads to large support & consistency

- Considered a broad class of kernel mixture models for density estimation, classification & testing in general spaces
- Theory to verify that a prior leads to large support & consistency
- Applied to hyperspheres & shapes - new computational methods also developed

- Considered a broad class of kernel mixture models for density estimation, classification & testing in general spaces
- Theory to verify that a prior leads to large support & consistency
- Applied to hyperspheres & shapes - new computational methods also developed
- Ongoing: factor models for manifolds & methods for manifold learning

 BHATTACHARYA, A. AND BHATTACHARYA, R. (2008).

 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*.

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


BHATTACHARYA, R. AND PATRANGENARU, V. (BP) (2003).


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


ESCOBAR, M. D. AND WEST, M. (1995).


LO, A. Y. (1984).


LENNOX, K.P., DAHL, D.B., VANNUCCI, M. AND TSAI, J.W. (2009).


MÜLLER, P., ERKANLI, A. AND WEST, M. (1996).


SCHWARTZ, L. (1965).


YAU, C., PAPASPILIOPOULOS, O., ROBERTS, G.O. & HOLMES, C. (2009).