Classification methods for functional data

2017 CRoNoS Spring Course on Multivariate methods with R

Enea Bongiorno

UNIVERSITÀ DEL PIEMONTE ORIENTALE

Cyprus University of Technology,
8-10/04/2017

OUTLINE

Multivariate Framework

Consider

- a random vector

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B})$$

  with positive density f

Multivariate Framework

Consider

- a random vector

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B})$$

  with positive density f
- a partition of Ω in G sub-populations: $\{\Omega_g, g = 1, \ldots, G\}$

Multivariate Framework

Consider

- a random vector

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B})$$

  with positive density f
- a partition of Ω in G sub-populations: $\{\Omega_g, g = 1, \ldots, G\}$
- the group random variable Y:

$$Y(\omega) = g \quad \text{if} \quad \omega \in \Omega_g$$

$$\mathbb{P}(Y = g) = \pi_g > 0, \quad \sum_{g=1}^{G} \pi_g = 1.$$

Multivariate Framework

Consider

- a random vector

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B})$$

  with positive density f
- a partition of Ω in G sub-populations: $\{\Omega_g, g = 1, \ldots, G\}$
- the group random variable Y:

$$Y(\omega) = g \quad \text{if} \quad \omega \in \Omega_g$$

$$\mathbb{P}(Y = g) = \pi_g > 0, \quad \sum_{g=1}^{G} \pi_g = 1.$$

- $f(x|g)$, the density function of $(X|Y = g)$

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

This is the starting point to

Model–based classification

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

This is the starting point to

Model–based classification

1. If Y is a latent/unobserved variable
2. If Y is observed

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

This is the starting point to

Model–based classification

1. If Y is a latent/unobserved variable
   $\Longrightarrow f(x)$ carries the information on the mixture
2. If Y is observed

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

This is the starting point to

Model–based classification

1. If Y is a latent/unobserved variable
   $\Longrightarrow f(x)$ carries the information on the mixture
   $\Longrightarrow$ unsupervised classification (clustering)
2. If Y is observed

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

This is the starting point to

Model–based classification

1. If Y is a latent/unobserved variable
   $\Longrightarrow f(x)$ carries the information on the mixture
   $\Longrightarrow$ unsupervised classification (clustering)
2. If Y is observed
   $\Longrightarrow$ one can compare posterior probabilities $\pi_g f(x|g)$

Multivariate Framework

Mixture Model

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

This is the starting point to

Model–based classification

1. If Y is a latent/unobserved variable
   $\Longrightarrow f(x)$ carries the information on the mixture
   $\Longrightarrow$ unsupervised classification (clustering)
2. If Y is observed
   $\Longrightarrow$ one can compare posterior probabilities $\pi_g f(x|g)$
   $\Longrightarrow$ supervised classification (discriminant) - Bayes rule

Multivariate Clustering

GOAL

- Data exploration

Multivariate Clustering

GOAL

- Data exploration
- Identify patterns in data with useful interpretation for the user

Multivariate Clustering

GOAL

- Data exploration
- Identify patterns in data with useful interpretation for the user
- Build "homogeneous" groups (clusters) of observations

Multivariate Clustering

GOAL

- Data exploration
- Identify patterns in data with useful interpretation for the user
- Build "homogeneous" groups (clusters) of observations

- Heuristic and Geometric procedures as hierarchical clustering, k-means, . . . [Ward'63, Hartigan'75]

Multivariate Clustering

GOAL

- Data exploration
- Identify patterns in data with useful interpretation for the user
- Build "homogeneous" groups (clusters) of observations

- Heuristic and Geometric procedures as hierarchical clustering, k-means, ... [Ward'63, Hartigan'75]
- Probabilistic approaches as model-based algorithm, density based clustering [Hartigan'75, Wishart'69]

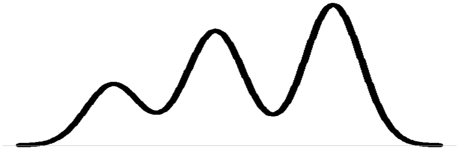
Multivariate Clustering

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

IDEA

Regions with high density identify clusters [Wishart'69]:

- Fix c, consider the connected components of $\{f > c\}$



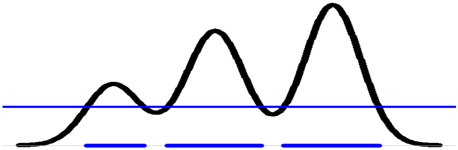
Multivariate Clustering

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

IDEA

Regions with high density identify clusters [Wishart'69]:

- Fix c, consider the connected components of $\{f > c\}$



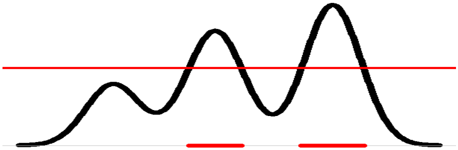
Multivariate Clustering

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

IDEA

Regions with high density identify clusters [Wishart'69]:

- Fix c, consider the connected components of $\{f > c\}$



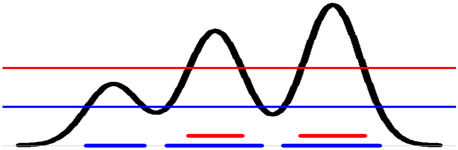
Multivariate Clustering

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

IDEA

Regions with high density identify clusters [Wishart'69]:

- Fix c, consider the connected components of $\{f > c\}$



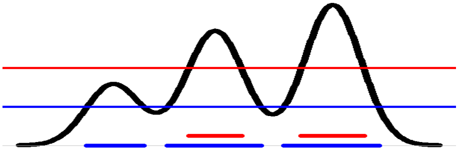
Multivariate Clustering

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

IDEA

Regions with high density identify clusters [Wishart'69]:

- Fix c, consider the connected components of $\{f > c\}$



Clusters number depends on the threshold level c

In many cases, MODES depict structural differences among data

Multivariate Discriminant Analysis

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

GOAL

To label a new incoming observation x

Multivariate Discriminant Analysis

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

GOAL

To label a new incoming observation x

- **Bayes Classification Rule**: assign x to the class $\gamma(x) \in \{1, \ldots, G\}$ with the highest posterior probability

$$\gamma(x) = \arg\max_{g=1,\ldots,G} \mathbb{P}(Y = g | X = x)$$

Multivariate Discriminant Analysis

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

GOAL

To label a new incoming observation x

- **Bayes Classification Rule**: assign x to the class $\gamma(x) \in \{1, \ldots, G\}$ with the highest posterior probability

$$\gamma(x) = \underset{g=1,\ldots,G}{\arg\max}\ \mathbb{P}(Y = g|X = x)$$

- If $f(x|g)$ were known ($f(x|g) > 0$):

$$\gamma(x) = \underset{g=1,\ldots,G}{\arg\max}\ \pi_g f(x|g).$$

OUTLINE

Functional Setting

Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \| \cdot \|)$

Functional Setting

Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \| \cdot \|)$
- $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{L}^2_{[0,1]}, \mathcal{B})$ Random Curve (RC) with

Functional Setting

Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \| \cdot \|)$
- $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}^2_{[0,1]}, \mathcal{B})$ Random Curve (RC) with
  - Mean: $\mu_X = \{\mathbb{E}[X(t)], t \in [0,1]\}$

Functional Setting

Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \| \cdot \|)$
- $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathcal{L}^2_{[0,1]}, \mathcal{B})$ Random Curve (RC) with
  - Mean: $\mu_X = \{\mathbb{E}[X(t)], t \in [0,1]\}$
  - Covariance: $\Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot \rangle (X - \mu_X)]$

Functional Setting

Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \| \cdot \|)$
- $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}^2_{[0,1]}, \mathcal{B})$ Random Curve (RC) with
  - Mean: $\mu_X = \{\mathbb{E}[X(t)], t \in [0,1]\}$
  - Covariance: $\Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot \rangle (X - \mu_X)]$
  - WLOG assume $\mu_X = 0$.

Functional Setting

Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \|\cdot\|)$
- $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}^2_{[0,1]}, \mathcal{B})$ Random Curve (RC) with
  - Mean: $\mu_X = \{\mathbb{E}[X(t)], t \in [0,1]\}$
  - Covariance: $\Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot \rangle (X - \mu_X)]$
  - WLOG assume $\mu_X = 0$.
- G sub-populations Ω_g of Ω with Y the group–variable

Functional Setting

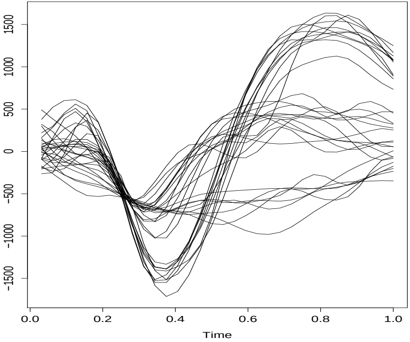
Consider

- the Hilbert space $H = (\mathcal{L}^2_{[0,1]}, \langle \cdot, \cdot \rangle, \| \cdot \|)$
- $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{L}^2_{[0,1]}, \mathcal{B})$ Random Curve (RC) with
  - Mean: $\mu_X = \{\mathbb{E}[X(t)], t \in [0,1]\}$
  - Covariance: $\Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot \rangle (X - \mu_X)]$
  - WLOG assume $\mu_X = 0$.
- G sub-populations Ω_g of Ω with Y the group–variable

[Bosq'00, Ferraty,Vieu'06, Ramsay,Silverman'05]

Example of Functional Data 1/2

Neuronal experiment on a monkey: (Voltage of neurons) vs. Time

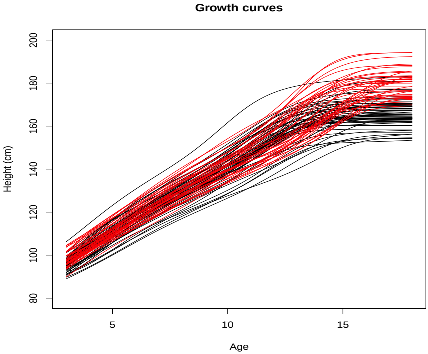


GOAL
spike sorting: distinguish different activities of neurons (clustering)

[Thanks to Andrew Schwartz motorlab in Pittsburgh]

Example of Functional Data 2/2

Berkeley growth curves: Boys (red), Girls (black)



GOAL

retrieve the gender of subjects (clustering if sex is treated as a hidden variable, discriminant otherwise)

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem
2. two stage approaches:

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem
2. two stage approaches:
   - Project data on some finite subspace (dim. reduction)

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem
2. two stage approaches:
   - Project data on some finite subspace (dim. reduction)
   - classification on basis expansion coefficients:

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem
2. two stage approaches:
   - Project data on some finite subspace (dim. reduction)
   - classification on basis expansion coefficients:
     - (non)Hierarchical techniques

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem
2. two stage approaches:
   - Project data on some finite subspace (dim. reduction)
   - classification on basis expansion coefficients:
     - (non)Hierarchical techniques
3. model based approach:

Functional Data Classification

PROBLEM

data belong to infinite dimensional space

Popular approaches ([Jacques,Preda'13] for a survey on clustering, [Ferraty,Vieu'06, James,Hastie'01] for discriminant examples):

1. non-parametric approach:
   - Consider (semi)metrics for functional data
   - Use standard clustering techniques based on dissimilarities
   - Rewrite discriminant as a regression problem
2. two stage approaches:
   - Project data on some finite subspace (dim. reduction)
   - classification on basis expansion coefficients:
     - (non)Hierarchical techniques
3. model based approach:
   - Parametric mixture model for coefficients of some basis

About the density of a random curve

GOAL: To extend density based classifications to functional data

About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure

About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.

About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

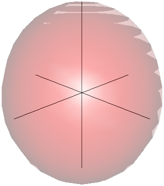
- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.
  ν should be locally finite and translation-invariant:

About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.
  ν should be locally finite and translation-invariant:
  - Let $B = \{x \in H : \|x\| < 1\}$ and $\{\xi_j\}_{j=1}^{\infty}$ an orthonormal basis of H.



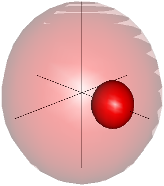
About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.

  ν should be locally finite and translation-invariant:
  - Let $B = \{x \in H : \|x\| < 1\}$ and $\{\xi_j\}_{j=1}^{\infty}$ an orthonormal basis of H.
  - Let $B_j = \{x \in H : \|x - \xi_j/2\| < 1/4\}$, with $\nu(B_j) = cost$



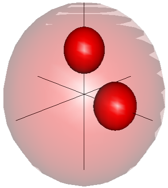
About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.

  ν should be locally finite and translation-invariant:
  - Let $B = \{x \in H : \|x\| < 1\}$ and $\{\xi_j\}_{j=1}^{\infty}$ an orthonormal basis of H.
  - Let $B_j = \{x \in H : \|x - \xi_j/2\| < 1/4\}$, with $\nu(B_j) = cost$



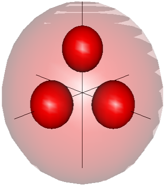
About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.

  ν should be locally finite and translation-invariant:
  - Let $B = \{x \in H : \|x\| < 1\}$ and $\{\xi_j\}_{j=1}^{\infty}$ an orthonormal basis of H.
  - Let $B_j = \{x \in H : \|x - \xi_j/2\| < 1/4\}$, with $\nu(B_j) = cost$



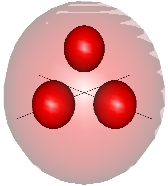
About the density of a random curve

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

- (Multivariate case) f is the Radon-Nikodym derivative w.r.t. the Lebesgue measure
- (∞–dim case) $\nexists$ a measure ν playing the role of the Leb. one.

  ν should be locally finite and translation-invariant:
  - Let $B = \{x \in H : \|x\| < 1\}$ and $\{\xi_j\}_{j=1}^{\infty}$ an orthonormal basis of H.
  - Let $B_j = \{x \in H : \|x - \xi_j/2\| < 1/4\}$, with $\nu(B_j) = cost$
  - Then $B_i \cap B_j = \emptyset$, $i \neq j$, $B \supset \bigcup B_j$ and $\nu(B) \geq \sum_j \nu(B_j) = \infty$



Small–Ball Probability (SmBP)

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

Small–Ball Probability (SmBP)

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?
"Lack" of probability density $\Longrightarrow$ IDEA: Look for a "surrogate"

Small–Ball Probability (SmBP)

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?

"Lack" of probability density $\Longrightarrow$ IDEA: Look for a "surrogate"

Small-Ball Probability (SmBP) [Ferraty,Goia,Vieu'02]

$$\varphi(x,\varepsilon) = \mathbb{P}\left(\|X - x\| < \varepsilon\right), \qquad x \in \mathcal{L}^2_{[0,1]}, \varepsilon > 0$$

Small–Ball Probability (SmBP)

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?
"Lack" of probability density $\Longrightarrow$ IDEA: Look for a "surrogate"

Small-Ball Probability (SmBP) [Ferraty,Goia,Vieu'02]

$$\varphi(x,\varepsilon) = \mathbb{P}\left(\|X - x\| < \varepsilon\right), \qquad x \in \mathcal{L}^2_{[0,1]}, \varepsilon > 0$$

- $\varphi(x,\varepsilon)$ is used as a measure of local concentration

Small–Ball Probability (SmBP)

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?
"Lack" of probability density $\Longrightarrow$ IDEA: Look for a "surrogate"

Small-Ball Probability (SmBP) [Ferraty,Goia,Vieu'02]

$$\varphi(x,\varepsilon) = \mathbb{P}\left(\|X - x\| < \varepsilon\right), \qquad x \in \mathcal{L}^2_{[0,1]}, \varepsilon > 0$$

- $\varphi(x,\varepsilon)$ is used as a measure of local concentration
- Tipical assumption:

$$\varphi(x,\varepsilon) = \Psi\left(x\right)\phi\left(\varepsilon\right) + o\left(\phi\left(\varepsilon\right)\right), \qquad \varepsilon \to 0$$

Small–Ball Probability (SmBP)

GOAL: To extend density based classifications to functional data

¿What is a probability density for random curve?
"Lack" of probability density $\Longrightarrow$ IDEA: Look for a "surrogate"

Small-Ball Probability (SmBP) [Ferraty,Goia,Vieu'02]

$$\varphi(x,\varepsilon) = \mathbb{P}\left(\|X - x\| < \varepsilon\right), \qquad x \in \mathcal{L}^2_{[0,1]}, \varepsilon > 0$$

- $\varphi(x,\varepsilon)$ is used as a measure of local concentration
- Tipical assumption:

$$\varphi(x,\varepsilon) = \Psi\left(x\right)\phi\left(\varepsilon\right) + o\left(\phi\left(\varepsilon\right)\right), \qquad \varepsilon \to 0$$

Ψ = intensity of SmBP = density "surrogate"
ϕ = Volume parameter

Some references

The limiting behaviour of the SmBP plays a key role in

- small deviations theory:
  [Li,Shao'01]
  [Lifshits'12]

Some references

The limiting behaviour of the SmBP plays a key role in

- small deviations theory:
  [Li,Shao'01]
  [Lifshits'12]
- functional regression:
  [Ferraty,Vieu'06]
  [Ferraty,Mas,Vieu'07]

Some references

The limiting behaviour of the SmBP plays a key role in

- small deviations theory:
  [Li,Shao'01]
  [Lifshits'12]
- functional regression:
  [Ferraty,Vieu'06]
  [Ferraty,Mas,Vieu'07]
- applications:

Some references

The limiting behaviour of the SmBP plays a key role in

- small deviations theory:
  [Li,Shao'01]
  [Lifshits'12]
- functional regression:
  [Ferraty,Vieu'06]
  [Ferraty,Mas,Vieu'07]
- applications:
  - to estimate the functional mode
    [Gasser, Hall, Presnell'98]
    [Delaigle,Hall'10]
    [Ferraty,Kudraszow,Vieu'12]

Some references

The limiting behaviour of the SmBP plays a key role in

- small deviations theory:
  [Li,Shao'01]
  [Lifshits'12]
- functional regression:
  [Ferraty,Vieu'06]
  [Ferraty,Mas,Vieu'07]
- applications:
  - to estimate the functional mode
    [Gasser, Hall, Presnell'98]
    [Delaigle,Hall'10]
    [Ferraty,Kudraszow,Vieu'12]
  - in classifications problems for functional data
    [Delsol,Louchet'14]
    [B.,Goia'16a]
    [Ciollaro,Genovese,Wang'16]

Some references

The limiting behaviour of the SmBP plays a key role in

- small deviations theory:
  [Li,Shao'01]
  [Lifshits'12]
- functional regression:
  [Ferraty,Vieu'06]
  [Ferraty,Mas,Vieu'07]
- applications:
  - to estimate the functional mode
    [Gasser, Hall, Presnell'98]
    [Delaigle,Hall'10]
    [Ferraty,Kudraszow,Vieu'12]
  - in classifications problems for functional data
    [Delsol,Louchet'14]
    [B.,Goia'16a]
    [Ciollaro,Genovese,Wang'16]
  - ...

SmBP Asymptotic: fixed $d \in \mathbb{N}$

¿Under what conditions $\varphi(x, \varepsilon) \approx \Psi(x)\, \phi(\varepsilon)\,, \varepsilon \to 0$?

SmBP Asymptotic: fixed $d \in \mathbb{N}$

¿Under what conditions $\varphi(x, \varepsilon) \approx \Psi(x) \phi(\varepsilon), \varepsilon \to 0$?

- Consider:
  - $\{\xi_j\}_{j=1}^{\infty}$ **any** orthonormal basis of $\mathcal{L}^2_{[0,1]}$ and $\theta_j = \langle X, \xi_j \rangle$
  - $\{\lambda_j = Var(\theta_j)\}_{j=1}^{\infty}$ are decreasingly ordered

SmBP Asymptotic: fixed $d \in \mathbb{N}$

¿Under what conditions $\varphi(x, \varepsilon) \approx \Psi(x)\, \phi(\varepsilon), \varepsilon \to 0$?

- Consider:
  - $\{\xi_j\}_{j=1}^{\infty}$ **any** orthonormal basis of $\mathcal{L}^2_{[0,1]}$ and $\theta_j = \langle X, \xi_j \rangle$
  - $\{\lambda_j = Var(\theta_j)\}_{j=1}^{\infty}$ are decreasingly ordered
- Assume:
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ admits a smooth pdf f_d
  - $\sup_{j \geq 1}(x_j^2/\lambda_j) < \infty$, with $x_j = \langle x, \xi_j \rangle$

SmBP Asymptotic: fixed $d \in \mathbb{N}$

¿Under what conditions $\varphi(x,\varepsilon) \approx \Psi(x)\,\phi(\varepsilon), \varepsilon \to 0$?

- Consider:
  - $\{\xi_j\}_{j=1}^{\infty}$ **any** orthonormal basis of $\mathcal{L}^2_{[0,1]}$ and $\theta_j = \langle X, \xi_j \rangle$
  - $\{\lambda_j = Var(\theta_j)\}_{j=1}^{\infty}$ are decreasingly ordered
- Assume:
  - $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$ admits a smooth pdf f_d
  - $\sup_{j \geq 1}(x_j^2/\lambda_j) < \infty$, with $x_j = \langle x, \xi_j \rangle$

Proposition [B.,Goia'16b]

Fix d. As $\varepsilon \to 0$, $\quad \varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(x,\varepsilon,d)$

- $V_d(\varepsilon)$ = volume of the d–dimensional ball of radius ε.
- $\mathcal{R}(x,\varepsilon,d) = \mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]$
- $S = S(x,\varepsilon,d) = \frac{1}{\varepsilon^2}\sum_{j\geq d+1}(\theta_j - \langle x, \xi_j\rangle)^2$

SmBP Asymptotic: fixed $d \in \mathbb{N}$

Proposition [B.,Goia'16b]

Fix d. As $\varepsilon \to 0$, $\varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(x,\varepsilon,d)$

- $\mathcal{R}(x,\varepsilon,d) = \mathbb{E}\left[(1-S)^{d/2}\, \mathbb{I}_{\{S\leq 1\}}\right]$ and $S = \frac{1}{\varepsilon^2}\sum_{j\geq d+1}\left(\theta_j - \langle x,\xi_j\rangle\right)^2$
- $V_d(\varepsilon)$ = volume of the d–dimensional ball of radius ε.

Some settings lead to define an intensity $\Psi(x)$ of the SmBP:

1. $\mathcal{R}(x,\varepsilon,d)$ is independent on x,
   Ex.: when $\langle x,\xi_j\rangle = 0$ for any $j \geq d_0 + 1$
   (e.g. the Gaussian processes are included: for any $d \geq d_0$,
   $\Psi(x) = \exp\left\{-\sum_{j\leq d_0} x_j^2/(2\lambda_j)\right\}$)

SmBP Asymptotic: fixed $d \in \mathbb{N}$

Proposition [B.,Goia'16b]

Fix d. As $\varepsilon \to 0$, $\varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(x,\varepsilon,d)$

- $\mathcal{R}(x,\varepsilon,d) = \mathbb{E}\left[(1-S)^{d/2}\mathbb{I}_{\{S\leq 1\}}\right]$ and $S = \frac{1}{\varepsilon^2}\sum_{j\geq d+1}\left(\theta_j - \langle x, \xi_j\rangle\right)^2$
- $V_d(\varepsilon)$ = volume of the d–dimensional ball of radius ε.

Some settings lead to define an intensity $\Psi(x)$ of the SmBP:

1. $\mathcal{R}(x,\varepsilon,d)$ is independent on x,
   Ex.: when $\langle x, \xi_j\rangle = 0$ for any $j \geq d_0 + 1$
   (e.g. the Gaussian processes are included: for any $d \geq d_0$,
   $$\Psi(x) = \exp\left\{-\sum_{j\leq d_0} x_j^2/(2\lambda_j)\right\})$$
2. $\exists d_0$ s.t. $\forall d \geq d_0$, $\mathcal{R}(x,\varepsilon,d) = 1$,
   Ex.: X is a d_0–dimensional process, $\Psi(x) = f_{d_0}(x)$

SmBP Asymptotic: fixed $d \in \mathbb{N}$

Proposition [B.,Goia'16b]

Fix d. As $\varepsilon \to 0$, $\varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(x,\varepsilon,d)$

- $\mathcal{R}(x,\varepsilon,d) = \mathbb{E}\left[(1-S)^{d/2}\mathbb{I}_{\{S\leq 1\}}\right]$ and $S = \frac{1}{\varepsilon^2}\sum_{j\geq d+1}\left(\theta_j - \langle x, \xi_j\rangle\right)^2$
- $V_d(\varepsilon)$ = volume of the d–dimensional ball of radius ε.

Some settings lead to define an intensity $\Psi(x)$ of the SmBP:

1. $\mathcal{R}(x,\varepsilon,d)$ is independent on x,
   Ex.: when $\langle x, \xi_j\rangle = 0$ for any $j \geq d_0 + 1$
   (e.g. the Gaussian processes are included: for any $d \geq d_0$,
   $$\Psi(x) = \exp\left\{-\sum_{j\leq d_0} x_j^2/(2\lambda_j)\right\})$$
2. $\exists d_0$ s.t. $\forall d \geq d_0$, $\mathcal{R}(x,\varepsilon,d) = 1$,
   Ex.: X is a d_0–dimensional process, $\Psi(x) = f_{d_0}(x)$
3. $$\begin{cases} \mathcal{R}(x,\varepsilon,d) \to 1, \\ \varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon), \end{cases} \qquad \varepsilon \to 0,\ d(\varepsilon) \to \infty.$$

SmBP Asymptotic: let $d(\varepsilon) \to \infty$

- Consider:
  - $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ eigenelements of $\Sigma[\cdot]$
  - (Karhunen-Loève) $X(t) = \sum_{j=1}^{\infty} \theta_j \xi_j(t), \qquad t \in [0,1]$
  - $\theta_j = \langle X, \xi_j \rangle$ principal components (PCs) of X satisfying
    $\mathbb{E}[\theta_j] = 0, \qquad Var(\theta_j) = \lambda_j, \qquad \mathbb{E}[\theta_j \theta_{j'}] = 0, \qquad j \neq j'$

SmBP Asymptotic: let $d(\varepsilon) \to \infty$

- Consider:
  - $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ eigenelements of $\Sigma[\cdot]$
  - (Karhunen-Loève) $\quad X(t) = \sum_{j=1}^{\infty} \theta_j \xi_j(t), \qquad t \in [0,1]$
  - $\theta_j = \langle X, \xi_j \rangle$ principal components (PCs) of X satisfying
    $\mathbb{E}[\theta_j] = 0, \qquad Var(\theta_j) = \lambda_j, \qquad \mathbb{E}[\theta_j \theta_{j'}] = 0, \qquad j \neq j'$
- Assume:
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ admits a smooth pdf f_d
  - $\sup_{j \geq 1}(x_j^2/\lambda_j) < \infty$, with $x_j = \langle x, \xi_j \rangle$
  - $\{\lambda_j\}_{j=1}^{\infty}$ decays to zero exponentially or faster
    (the spectrum of Σ is rather concentrate)

SmBP Asymptotic: let $d(\varepsilon) \to \infty$

- Consider:
  - $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ eigenelements of $\Sigma[\cdot]$
  - (Karhunen-Loève) $X(t) = \sum_{j=1}^{\infty} \theta_j \xi_j(t), \quad t \in [0,1]$
  - $\theta_j = \langle X, \xi_j \rangle$ principal components (PCs) of X satisfying
    $\mathbb{E}[\theta_j] = 0, \quad Var(\theta_j) = \lambda_j, \quad \mathbb{E}[\theta_j \theta_{j'}] = 0, \quad j \neq j'$
- Assume:
  - $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ admits a smooth pdf f_d
  - $\sup_{j \geq 1}(x_j^2/\lambda_j) < \infty$, with $x_j = \langle x, \xi_j \rangle$
  - $\{\lambda_j\}_{j=1}^{\infty}$ decays to zero exponentially or faster
    (the spectrum of Σ is rather concentrate)

Proposition [B.,Goia'16b]

As $\varepsilon \to 0$, it is possible to choose $d = d(\varepsilon) \to \infty$ so that:

$$\varphi(x, \varepsilon) \sim f_d(x_1, \ldots, x_d)\, \phi(d, \varepsilon)$$

This generalizes [Delaigle,Hall'10] where PCs are **independent**

SmBP Asymptotic: let $d(\varepsilon) \to \infty$

Proof Sketch.

STEP 1. Fix $d \in \mathbb{N}$. Taylor leads to

$$\left| \frac{\varphi(x,\varepsilon)}{f_d V_d \mathcal{R}} - 1 \right| \leq \frac{C\varepsilon^2}{\lambda_d}$$

SmBP Asymptotic: let $d(\varepsilon) \to \infty$

Proof Sketch.

STEP 1. Fix $d \in \mathbb{N}$. Taylor leads to

$$\left| \frac{\varphi(x,\varepsilon)}{f_d V_d \mathcal{R}} - 1 \right| \leq \frac{C\varepsilon^2}{\lambda_d}$$

STEP 2. Use the $\{\lambda_j\}$ decay to choose $d = d(\varepsilon)$ so that

$$d \to \infty, \qquad 1 - C\frac{d+2}{\varepsilon^2} \sum_{j \geq d+1} \lambda_j \leq \mathcal{R} \leq 1$$

That is $\mathcal{R}(x, \varepsilon, d(\varepsilon)) \to 1$ as $\varepsilon \to 0$.

SmBP Asymptotic: let $d(\varepsilon) \to \infty$

Proof Sketch.
STEP 1. Fix $d \in \mathbb{N}$. Taylor leads to

$$\left| \frac{\varphi(x,\varepsilon)}{f_d V_d \mathcal{R}} - 1 \right| \leq \frac{C\varepsilon^2}{\lambda_d}$$

STEP 2. Use the $\{\lambda_j\}$ decay to choose $d = d(\varepsilon)$ so that

$$d \to \infty, \qquad 1 - C\frac{d+2}{\varepsilon^2} \sum_{j \geq d+1} \lambda_j \leq \mathcal{R} \leq 1$$

That is $\mathcal{R}(x, \varepsilon, d(\varepsilon)) \to 1$ as $\varepsilon \to 0$.
STEP 3. Errors in 1-2 are controlled at the same time exploiting $\{\lambda_j\}$ decay (the faster they decay, the smaller the total error is)
$\Rightarrow$ The basis provided by KL decomposition is OPTIMAL

SmBP Asymptotic: comment 1/3

$\varphi(x,\varepsilon) \sim f_d(x)\phi(d,\varepsilon)$ as $\varepsilon \to 0$ and $d(\varepsilon) \to \infty$

In general, f_d still depends on ε through $d(\varepsilon)$
$\Longrightarrow$ it is not an intensity of the SmBP
BUT, some settings lead to an intensity $\Psi(x)$ of the SmBP:

- If $\{\theta_j\}$ are independent with density $\propto \exp\left\{-\left(|x_j|/\sqrt{\lambda_j}\right)^p\right\}$. Then

$$f_d(x) \to \Psi(x) = \exp\left\{-\frac{1}{2}\sum_{j=1}^{\infty}\left(|x_j|/\sqrt{\lambda_j}\right)^p\right\}, \text{ for any } x \in H$$

  is an intensity for the SmBP (Gaussian processes are included in this family).

SmBP Asymptotic: comment 2/3

$\varphi(x,\varepsilon) \sim f_d(x)\phi(d,\varepsilon)$ as $\varepsilon \to 0$ and $d(\varepsilon) \to \infty$

The form of the volumetric term $\phi(d,\varepsilon)$ depends on the eigenvalues decay rate. As example:

- Exponential decay (e.g. $\lambda_j = e^{-\beta j^\alpha},\ \beta > 0, \alpha \geq 1$):

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j < C, \quad \forall d$$

$$\phi(d,\varepsilon) = \exp\left\{\tfrac{1}{2} d \left[\log(2\pi e \varepsilon^2) - \log(d) + \delta(d,\alpha)\right]\right\}$$

- Hyper-exponential decay (e.g. $\lambda_j = e^{-\beta j^\alpha},\ \beta > 0, \alpha > 1$):

$$d\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j = o\left(1\right), \text{ as } d \to \infty$$

$\phi(d,\varepsilon) = \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2+1)}$ (Volume of a d-dim ball)

SmBP Asymptotic: comment 3/3

$\varphi(x,\varepsilon) \sim f_d(x)\phi(d,\varepsilon)$ as $\varepsilon \to 0$ and $d(\varepsilon) \to \infty$

There are situations in which $\varphi \sim f_d(x)\phi(d,\varepsilon)$ for slower (than exponential) decays. E.g. Wiener process:

$$\lambda_j \sim j^{-2}, \qquad \varphi(x,\varepsilon) \sim \exp\left\{-1/2 \int_0^1 x'(t)dt\right\} \phi(\varepsilon)$$

SmBP Mixture

- Consider the conditioned SmBP

$$\varphi(x, \varepsilon | g) = \mathbb{P}\left(\|X - x\| < \varepsilon \mid Y = g\right) \qquad g = 1, \ldots, G,$$

- Mixture

$$\varphi(x, \varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x, \varepsilon | g)$$

SmBP Mixture

- Consider the conditioned SmBP

$$\varphi(x, \varepsilon | g) = \mathbb{P}(\|X - x\| < \varepsilon \mid Y = g) \qquad g = 1, \ldots, G,$$

- Mixture $\quad + \quad$ Factorization

$$\varphi(x, \varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x, \varepsilon | g) \sim f_d(x_1, \ldots, x_d)\phi(d, \varepsilon), \quad \varepsilon \to 0$$

SmBP Mixture

- Consider the conditioned SmBP

$$\varphi(x,\varepsilon|g) = \mathbb{P}(\|X - x\| < \varepsilon \mid Y = g) \qquad g = 1,\dots,G,$$

- Mixture + Factorization

$$\varphi(x,\varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x,\varepsilon|g) \sim f_d(x_1,\dots,x_d)\phi(d,\varepsilon), \quad \varepsilon \to 0$$

Surrogate density f_d:

- carries spatial information on the mixture
- endorses "density oriented" clustering approach on f_d

SmBP Mixture

- Consider the conditioned SmBP

$$\varphi(x, \varepsilon | g) = \mathbb{P}(\|X - x\| < \varepsilon \mid Y = g) \qquad g = 1, \ldots, G,$$

- Mixture + Factorization

$$\varphi(x, \varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x, \varepsilon | g) \sim f_d(x_1, \ldots, x_d)\phi(d, \varepsilon), \quad \varepsilon \to 0$$

$$\sim \sum_{g=1}^{G} \pi_g f_{d_g}(x_1, \ldots, x_{d_g} | g)\phi(d_g, \varepsilon)$$

SmBP Mixture

- Consider the conditioned SmBP

$$\varphi(x,\varepsilon|g) = \mathbb{P}(\|X - x\| < \varepsilon \mid Y = g) \qquad g = 1,\dots,G,$$

- Mixture + Factorization

$$\varphi(x,\varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x,\varepsilon|g) \sim f_d(x_1,\dots,x_d)\phi(d,\varepsilon), \quad \varepsilon \to 0$$

$$\sim \sum_{g=1}^{G} \pi_g f_{d_g}(x_1,\dots,x_{d_g}|g)\phi(d_g,\varepsilon)$$

Under suitable assumptions on Σ, as $\varepsilon \to 0$:

$$f_d(x_1,\dots,x_d) \sim \sum_{g=1}^{G} \pi_g f_d(x_1,\dots,x_d|g), \quad \varepsilon \to 0$$

$\Longrightarrow$ discriminant analysis using $\pi_g f_d(x|g)$.

OUTLINE

1. Density oriented classification methods: a review
2. Functional Setting
3. Estimating the surrogate Density
   - Theoretical Aspects
   - Empirical Performances
4. SmBP Clustering
5. SmBP Discriminant Analysis

Estimating the surrogate Density

Multivariate kernel density estimator:

$$\widehat{f}_{d,n}\left(\widehat{\Pi}_d x\right) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(\left\|\widehat{\Pi}_d\left(X_i - x\right)\right\|\right), \qquad \widehat{\Pi}_d x \in \mathbb{R}^d,$$

where:

- $K_H(\mathbf{u}) = \det(H)^{-1/2} K\left(H^{-1/2}\mathbf{u}\right)$,
- K kernel function,
- H symmetric semi-definite positive $d \times d$ matrix,
- $\widehat{\Pi}_d$ projection operator over the subspace spanned by the first d eigenfunctions of $\widehat{\Sigma}_n$

Estimating the surrogate Density

Multivariate kernel density estimator:

$$\widehat{f}_{d,n}\left(\widehat{\Pi}_d x\right) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(\left\|\widehat{\Pi}_d\left(X_i - x\right)\right\|\right), \qquad \widehat{\Pi}_d x \in \mathbb{R}^d,$$

where:

- $K_H(\mathbf{u}) = \det(H)^{-1/2} K\left(H^{-1/2}\mathbf{u}\right)$,
- K kernel function,
- H symmetric semi-definite positive $d \times d$ matrix,
- $\widehat{\Pi}_d$ projection operator over the subspace spanned by the first d eigenfunctions of $\widehat{\Sigma}_n$

Π_d is estimated:
¿is the rate of convergence the same as when Π_d is known?

Estimating the surrogate Density

Consider $H_n = h_n^2 I$ and suppose:

(B.1) $f_d(x) > 0$ and p times differentiable;

(B.2) $h_n \to 0 \quad \text{and} \quad nh_n^d / \log n \to \infty \quad \text{as} \quad n \to \infty$;

(B.3) K is a density, Lipschitz and bounded with cpct. support;

(B.4) $\exists s, \kappa > 0 : \quad \mathbb{E}\left[\|X - x\|^m\right] \leq m! s \kappa^{m-2}/2$.

Estimating the surrogate Density

Consider $H_n = h_n^2 I$ and suppose:

(B.1) $f_d(x) > 0$ and p times differentiable;

(B.2) $h_n \to 0 \quad \text{and} \quad nh_n^d / \log n \to \infty \quad \text{as} \quad n \to \infty$;

(B.3) K is a density, Lipschitz and bounded with cpct. support;

(B.4) $\exists s, \kappa > 0 : \quad \mathbb{E}\left[\|X - x\|^m\right] \le m! s \kappa^{m-2}/2.$

[B.,Goia'16b]

Assume (B.1)–(B.4) with $p > (3d+2)/2$ and consider the optimal bandwidth

$$c_1 n^{-\frac{1}{2p+d}} \le h_n \le c_2 n^{-\frac{1}{2p+d}} \qquad c_1, c_2 > 0.$$

Then, as $n \to \infty$ and uniformly in $\mathbb{R}^d$

$$\mathbb{E}\left[f_d(x) - \widehat{f}_n(x)\right]^2 = O\left(n^{-2p/(2p+d)}\right).$$

Finite dimensional process

$$\begin{cases} X(t) = a\sqrt{2/\pi}\sin(t) \ \text{(the random process)} \\ \left. a \sim \begin{cases} \text{pdf } f_a \\ \mathbb{E}[a] = 0, \ Var(a) = 1, \end{cases} \right\} \text{(the only random part)} \\ t \in [0, \pi] \\ \\ \end{cases}$$

Finite dimensional process

$$\begin{cases} X(t) = a\sqrt{2/\pi}\sin(t) \ \text{(the random process)} \\ a \sim \left.\begin{cases} \text{pdf } f_a \\ \mathbb{E}[a] = 0, \ Var(a) = 1, \end{cases}\right\} \text{(the only random part)} \\ t \in [0, \pi] \\ x(t) = b\sqrt{2/\pi}\sin(t) \ \text{(center of the ball)} \\ b \text{ fixed value in } \mathbb{R} \end{cases}$$

Finite dimensional process

$$\begin{cases} X(t) = a\sqrt{2/\pi}\sin(t) \text{ (the random process)} \\ a \sim \left\{ \begin{array}{l} \text{pdf } f_a \\ \mathbb{E}[a] = 0, \ Var(a) = 1, \end{array} \right\} \text{(the only random part)} \\ t \in [0, \pi] \\ x(t) = b\sqrt{2/\pi}\sin(t) \text{ (center of the ball)} \\ b \text{ fixed value in } \mathbb{R} \end{cases}$$

In this setting, it holds:

$$\varphi(x, \varepsilon) \sim f_1(x_1)\,\varepsilon\pi^{1/2}/\Gamma(1/2+1) = 2f_a(b)\,\varepsilon.$$

- $(X_1, \ldots, X_n)$ sample drawn from X
- we compare f_a with the 1–dim estimator $\widehat{f_1}$

Finite dimensional process

- Relative MSE: $\sum_b \left[\widehat{f_1}\left(\widehat{x}_1^b\right) - f_a\left(b\right)\right]^2 / \sum_b f_a^2\left(b\right)$

Finite dimensional process

- Relative MSE: $\sum_b \left[\widehat{f_1}\left(\widehat{x}_1^b\right) - f_a\left(b\right)\right]^2 / \sum_b f_a^2\left(b\right)$
- 1000 Monte Carlo replications varying the sample size n and the distribution

Finite dimensional process

- Relative MSE: $\sum_b \left[\widehat{f}_1\left(\widehat{x}_1^b\right) - f_a\left(b\right)\right]^2 / \sum_b f_a^2\left(b\right)$
- 1000 Monte Carlo replications varying the sample size n and the distribution

Finite dimensional process

- Relative MSE: $\sum_b \left[\widehat{f_1}\left(\widehat{x}_1^b\right) - f_a\left(b\right)\right]^2 / \sum_b f_a^2\left(b\right)$
- 1000 Monte Carlo replications varying the sample size n and the distribution

	$a \sim N(0,1)$, $b \in [-4,4]$		$a \sim t(5)/\sqrt{5/3}$, $b \in [-4,4]$		$a \sim \left(\chi^2(8) - 8\right)/4$, $b \in [-2,6]$	
n	Mean	Std.	Mean	Std.	Mean	Std.
50	3.235	(2.681)	5.921	(2.557)	4.081	(2.842)
100	1.860	(1.444)	4.775	(1.503)	2.401	(1.619)
200	1.091	(0.824)	4.138	(0.878)	1.422	(0.887)
500	0.546	(0.355)	3.737	(0.477)	0.753	(0.443)
1000	0.330	(0.220)	3.606	(0.327)	0.453	(0.233)

- RMSE ($\times 10^2$) decreases as the sample size increases:
  - better estimates of projections $\widehat{\theta}$ and $\widehat{x}^b$
  - better performances of the kernel estimator
- Shape of distributions matters: long tails and asymmetries worse estimates

Finite dimensional process

- 1000 Monte Carlo replications varying b with fixed sample size $n = 200$

Finite dimensional process

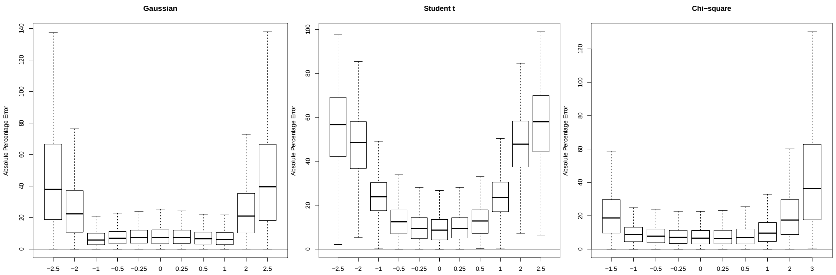
- 1000 Monte Carlo replications varying b with fixed sample size $n = 200$
- Absolute percentage errors (APE) $= \left| \widehat{f_1}\left(\widehat{x}_1^b\right) - f_a(b) \right| / f_a(b)$

Finite dimensional process

- 1000 Monte Carlo replications varying b with fixed sample size $n = 200$
- Absolute percentage errors (APE) $= \left|\widehat{f_1}\left(\widehat{x}_1^b\right) - f_a\left(b\right)\right| / f_a\left(b\right)$

Finite dimensional process

- 1000 Monte Carlo replications varying b with fixed sample size $n = 200$
- Absolute percentage errors (APE) $= \left| \widehat{f_1}\left(\widehat{x}_1^b\right) - f_a(b) \right| / f_a(b)$



Estimate worsens at the edges due to limitations of kernel density estimator in evaluate the tails.

Infinite dimensional process

- X Stand. Brownian Motion on $[0, 1]$ ($\lambda_j \sim cost \cdot j^{-2}$).
- $x(t) = b\frac{2\sqrt{2}}{\pi}\sin\left(\frac{\pi t}{2}\right), \qquad t \in [0, 1], b \in \mathbb{R}$

$$\varphi(x, \varepsilon) \sim \exp\left\{-b^2/2\right\}\varphi(0, \varepsilon), \qquad \text{as } \varepsilon \to 0,$$

Infinite dimensional process

- X Stand. Brownian Motion on $[0,1]$ ($\lambda_j \sim cost \cdot j^{-2}$).
- $x(t) = b\frac{2\sqrt{2}}{\pi}\sin\left(\frac{\pi t}{2}\right), \qquad t \in [0,1], b \in \mathbb{R}$

$$\varphi(x,\varepsilon) \sim \exp\left\{-b^2/2\right\}\varphi(0,\varepsilon), \qquad \text{as } \varepsilon \to 0,$$

- 1000 MC replication of BM sample $(X_1, \ldots, X_n)$
- the used dimension d varies in $1, \ldots, 5$
- $b \in [-4, 4]$

Infinite dimensional process

n	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$
50	3.36 (2.51)	7.20 (3.73)	13.53 (7.34)	22.03 (12.05)	31.90 (15.87)
100	1.95 (1.20)	4.82 (2.59)	9.47 (5.54)	15.86 (8.71)	23.99 (12.27)
200	1.16 (0.72)	3.14 (1.60)	6.64 (3.77)	11.51 (6.30)	17.89 (9.48)
500	0.57 (0.33)	1.78 (0.93)	4.17 (2.36)	7.77 (4.23)	12.96 (6.88)
1000	0.35 (0.19)	1.15 (0.63)	2.82 (1.64)	5.86 (3.13)	10.09 (5.43)

- Fix d, RMSE reduces (both in mean and var.) increasing n
- Fix n, RMSE increases (both in mean and var.) with d (curse of dim.)
- d vs. n: it is possible to use large d at the cost of use large samples (read diagonally the table).

OUTLINE

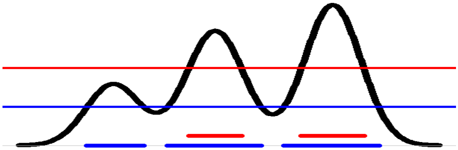
Multivariate Clustering

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

IDEA

Regions with high density identify clusters [Wishart'69]:

- Fix c, consider the connected components of $\{f > c\}$



Clusters number depends on the threshold level c

In many cases, MODES depict structural differences among data

SmBP Mixture

- Consider the conditioned SmBP

$$\varphi(x,\varepsilon|g) = \mathbb{P}\left(\|X - x\| < \varepsilon \mid Y = g\right) \qquad g = 1, \ldots, G,$$

- Mixture + Factorization

$$\varphi(x,\varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x,\varepsilon|g) \sim f_d(x_1, \ldots, x_d)\phi(d,\varepsilon), \quad \varepsilon \to 0$$

Surrogate density f_d:

- carries spatial information on the mixture
- endorses "density oriented" clustering approach on f_d

SmBP Clustering

Consider the sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn from (X, Y)

- X_i are observed
- the group variables Y_i are latent.

SmBP Clustering

Consider the sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn from (X, Y)

- X_i are observed
- the group variables Y_i are latent.

GOAL

To determine

- the range of Y (i.e. G)
- for each X_i, the membership group (that is the value of Y_i)

SmBP Clustering

Consider the sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ drawn from (X, Y)

- X_i are observed
- the group variables Y_i are latent.

GOAL

To determine

- the range of Y (i.e. G)
- for each X_i, the membership group (that is the value of Y_i)

METHOD

Regions with "locally high surrogate density (f_d)" identify groups

Algorithm

1. Estimate covariance operator and eigenelements;

Algorithm

1. Estimate covariance operator and eigenelements;
2. Fix d, compute $\widehat{f}_{d,n}$, and look for its local maxima $\widehat{m}_{d,g}$ $(g = 1, \ldots, \widehat{G})$;

Algorithm

1. Estimate covariance operator and eigenelements;
2. Fix d, compute $\widehat{f}_{d,n}$, and look for its local maxima $\widehat{m}_{d,g}$ $(g = 1, \ldots, \widehat{G})$;
3. Finding Prototypes: for each g in $\{1, \ldots, \widehat{G}\}$, the g-th "prototypes" group is formed by those X_i whose estimated PCs belong to the largest connected iso-surface of $\widehat{f}_{d,n}$ that contains only the maximum $m_{d,g}$.

Algorithm

1. Estimate covariance operator and eigenelements;
2. Fix d, compute $\widehat{f}_{d,n}$, and look for its local maxima $\widehat{m}_{d,g}$ $(g = 1, \ldots, \widehat{G})$;
3. Finding Prototypes: for each g in $\{1, \ldots, \widehat{G}\}$, the g-th "prototypes" group is formed by those X_i whose estimated PCs belong to the largest connected iso-surface of $\widehat{f}_{d,n}$ that contains only the maximum $m_{d,g}$.
4. Label the unclassified X_i with the $\widehat{K}$ prototypes groups by means of a k-NN procedure.

Comments

- Proposed method $\iff$ local adaptive threshold
  Hartigan $\iff$ global threshold
  ([Cuevas,Febreo,Fraiman'00, Cuevas,Febrero,Fraiman'01] - multivariate)

Comments

- Proposed method $\iff$ local adaptive threshold
  Hartigan $\iff$ global threshold
  ([Cuevas,Febreo,Fraiman'00, Cuevas,Febrero,Fraiman'01] - multivariate)
- Distribution-Free: no distributional assumptions on f_d
  ([Jacques,Preda'13, Jacques,Preda'14]: gaussian finite mixture)

Comments

- Proposed method $\iff$ local adaptive threshold
  Hartigan $\iff$ global threshold
  ([Cuevas,Febreo,Fraiman'00, Cuevas,Febrero,Fraiman'01] - multivariate)
- Distribution-Free: no distributional assumptions on f_d
  ([Jacques,Preda'13, Jacques,Preda'14]: gaussian finite mixture)
- Dependent Approach: PCs are just uncorrelated
  ([Jacques,Preda'13, Jacques,Preda'14]: PCs independent)

Comments

- Proposed method $\iff$ local adaptive threshold
  Hartigan $\iff$ global threshold
  ([Cuevas,Febreo,Fraiman'00, Cuevas,Febrero,Fraiman'01] - multivariate)
- Distribution-Free: no distributional assumptions on f_d
  ([Jacques,Preda'13, Jacques,Preda'14]: gaussian finite mixture)
- Dependent Approach: PCs are just uncorrelated
  ([Jacques,Preda'13, Jacques,Preda'14]: PCs independent)
- Tuning d (or ε) and/or tuning the bandwidth matrix lead to different phenomenon scales

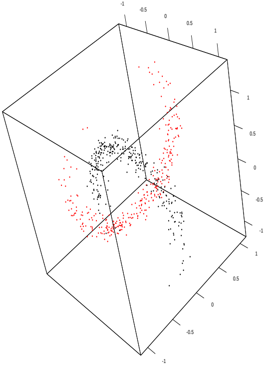
Example: Synthetic Data

$$X_i^{(g)}(t) = \sum_{l=0}^{L} \sqrt{\lambda_l}\tau_{i,l}^{(g)}\varphi_l(t),$$

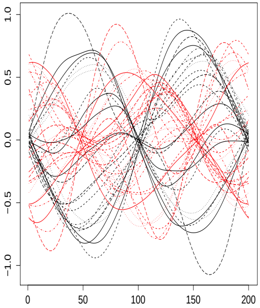
$t \in [0,1],\ i = 1,\ldots,N,\ g = 1,\ldots,G.$

- $\{\varphi_l(t)\}_l$ Fourier Basis
- $\lambda_l = 0.7 \times 3^{-l}$ $(l = 1,\ldots,L = 150)$, so that the first 3 PCs explain always more than 99%
- $G = 2$ groups and $N = 300$ curves for each group
- uncorrelated but dependent coefficients $(\tau_{i,l}^{(g)})_{l=1}^{L}$ so that the first 3 PCs looks like two bound up "bananas".

Example: Synthetic Data



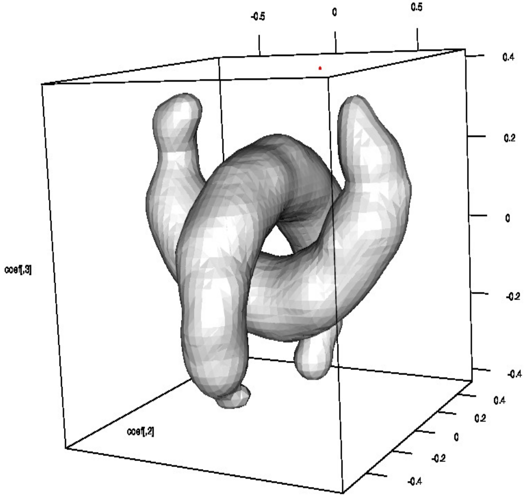
$\tau_{i,l}^{(g)}, \quad l = 1, 2, 3, \quad g = 1, 2$



sample of curves

Example: Synthetic Data

Maximum iso–surfaces (prototypes)



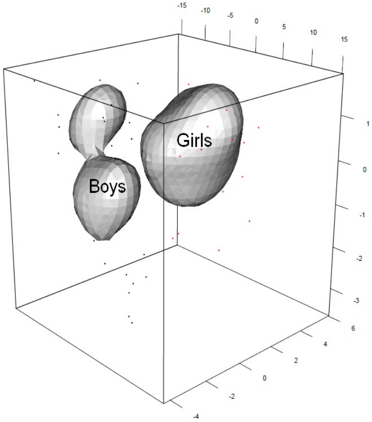
Example: Synthetic Data

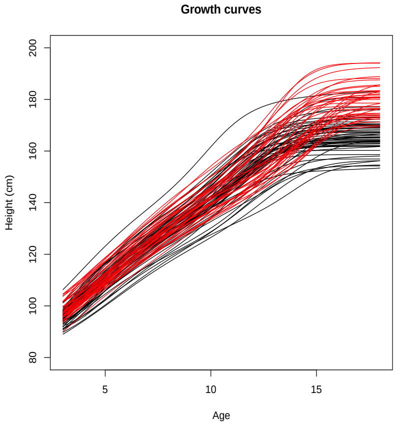
$d = 3$ (FEV criterion)
400 Monte Carlo replications

Algorithm	Miscl. Error		$\widehat{G}$		
	Mean	St. Dev.	$q_{0.5}$	$q_{0.75}$	$q_{0.9}$
SmBP	**0.088**	**0.174**	**2**	**2**	**2**
Functional K-means ($G = 2$)	0.377	0.068	–	–	–
Gaussian Mixture ($G = 2$)	0.153	0.106	–	–	–
GM BIC selection	0.666	0.034	9	10	11

Example: Growth Curves

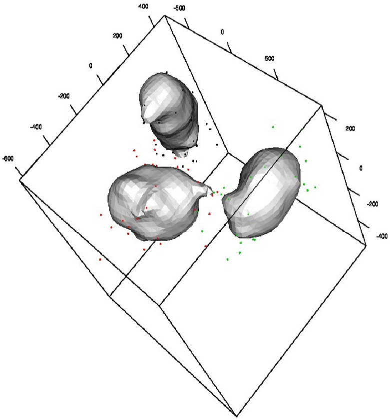
$d = 3$ (FEV) - 97.9% correct classification with respect sex

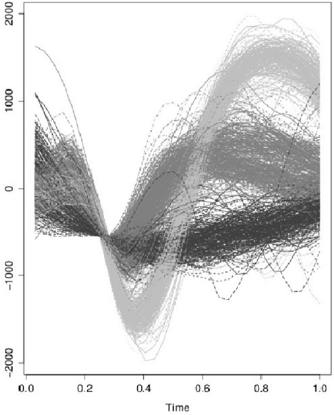




Example: Neuronal data

Iso–surfaces and Clusters





OUTLINE

Multivariate Discriminant Analysis

$$f(x) = \sum_{g=1}^{G} \pi_g f(x|g), \qquad x \in \mathbb{R}^d$$

GOAL

To label a new incoming observation x

- **Bayes Classification Rule**: assign x to the class $\gamma(x) \in \{1, \ldots, G\}$ with the highest posterior probability

$$\gamma(x) = \underset{g=1,\ldots,G}{\arg\max}\ \mathbb{P}(Y = g|X = x)$$

- If $f(x|g)$ were known ($f(x|g) > 0$):

$$\gamma(x) = \underset{g=1,\ldots,G}{\arg\max}\ \pi_g f(x|g).$$

Generalized Bayes Classification Rule

- X and Y are observed
- Consider the mixture of SmBP

$$\varphi(x,\varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x,\varepsilon|g), \qquad \varepsilon > 0.$$

Generalized Bayes Classification Rule

- X and Y are observed
- Consider the mixture of SmBP

$$\varphi(x,\varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x,\varepsilon|g), \qquad \varepsilon > 0.$$

Generalized Classification Rule: assign a new functional observation x to g-th group for which, as ε tends to 0,

$$\frac{\mathbb{P}(Y=g|\ \|X-x\|<\varepsilon)}{\mathbb{P}(Y=g'|\ \|X-x\|<\varepsilon)} > 1, \qquad \text{for any } g \in \{1,\ldots,G\},\ g' \neq g.$$

Generalized Bayes Classification Rule

- X and Y are observed
- Consider the mixture of SmBP

$$\varphi(x,\varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x,\varepsilon|g), \qquad \varepsilon > 0.$$

Generalized Classification Rule: assign a new functional observation x to g-th group for which, as ε tends to 0,

$$\frac{\mathbb{P}(Y=g|\ \|X-x\|<\varepsilon)}{\mathbb{P}(Y=g'|\ \|X-x\|<\varepsilon)} > 1, \qquad \text{for any } g \in \{1,\ldots,G\},\ g' \neq g.$$

[B.,Goia'16a]

The Generalized Classification Rule is equivalent to

$$\gamma(x,d) = \arg\max_{g=1,\ldots,G} \pi_g f_d(x|g) \qquad \text{as } d \to \infty$$

Proof Sketch

- Bayes: $\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon) = \dfrac{\pi_g \varphi(x, \varepsilon|g)}{\varphi(x, \varepsilon)}$

Proof Sketch

- Bayes: $\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon) = \dfrac{\pi_g \varphi(x, \varepsilon|g)}{\varphi(x, \varepsilon)}$
- Factorization for each group: as $\varepsilon \to 0$

$$\frac{\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon)}{\mathbb{P}(Y = g' \mid \|X - x\| < \varepsilon)} \sim \frac{\pi_g f_{d_g}(x|g)\, \phi(d_g, \varepsilon)}{\pi_{g'} f_{d_{g'}}(x|g')\, \phi(d_{g'}, \varepsilon)}$$

Proof Sketch

- Bayes: $\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon) = \dfrac{\pi_g \varphi(x, \varepsilon|g)}{\varphi(x, \varepsilon)}$
- Factorization for each group: as $\varepsilon \to 0$

$$\frac{\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon)}{\mathbb{P}(Y = g' \mid \|X - x\| < \varepsilon)} \sim \frac{\pi_g f_{d_g}(x|g)\,\phi(d_g, \varepsilon)}{\pi_{g'} f_{d_{g'}}(x|g')\,\phi(d_{g'}, \varepsilon)}$$

- $f_{d_g}(x|g) =$ joint density of the first d_g PCs of Σ_g (cov. of g-th group)

Proof Sketch

- Bayes: $\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon) = \dfrac{\pi_g \varphi(x, \varepsilon | g)}{\varphi(x, \varepsilon)}$
- Factorization for each group: as $\varepsilon \to 0$

$$\frac{\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon)}{\mathbb{P}(Y = g' \mid \|X - x\| < \varepsilon)} \sim \frac{\pi_g f_{d_g}(x|g)\,\phi(d_g, \varepsilon)}{\pi_{g'} f_{d_{g'}}(x|g')\,\phi(d_{g'}, \varepsilon)}$$

- $f_{d_g}(x|g) =$ joint density of the first d_g PCs of Σ_g (cov. of g-th group)
- Spectrum decay of Σ controls the one of Σ_g (min–max principle) $\Longrightarrow$ choose the same $d_g = d$ for any g

Proof Sketch

- Bayes: $\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon) = \dfrac{\pi_g \varphi(x, \varepsilon | g)}{\varphi(x, \varepsilon)}$
- Factorization for each group: as $\varepsilon \to 0$

$$\frac{\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon)}{\mathbb{P}(Y = g' \mid \|X - x\| < \varepsilon)} \sim \frac{\pi_g f_{d_g}(x|g)\, \phi(d_g, \varepsilon)}{\pi_{g'} f_{d_{g'}}(x|g')\, \phi(d_{g'}, \varepsilon)}$$

- $f_{d_g}(x|g) =$ joint density of the first d_g PCs of Σ_g (cov. of g-th group)
- Spectrum decay of Σ controls the one of Σ_g (min–max principle) $\implies$ choose the same $d_g = d$ for any g
- Simplification leads to Thesis.

Estimate the Classifier 1/2

Consider

- $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ sample from (X, Y)
- $d_g = d$ for each $g = 1, \ldots, G$
- $\widehat{\Pi}_{g,d}$ = proj.oper. over the subspace spanned by the first d eigenfunctions of $\widehat{\Sigma}_g$

A <u>kernel estimator</u> for $\gamma(x, d)$ is:

$$\widehat{\gamma}_n(x, d) = \underset{g=1,\ldots,G}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{Y_i = g\}} K_{H_g}\left(\left\|\widehat{\Pi}_{g,d}(X_i - x)\right\|\right)$$

Estimate the Classifier 2/2

- Consider the Bayes probability of error

$$L^\star = \min_\gamma \mathbb{P}\left(\gamma\left(X\right) \neq Y\right)$$

and the conditional probability of error

$$L_n = \mathbb{P}\left(\widehat{\gamma}_n\left(X, d\right) \neq Y \mid \left\{\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right)\right\}\right)$$

- Assume that there exists a positive integer d^* such that $\gamma(x, d) = \gamma(x, d^*)$ for any $d \geq d^*$.

[B.,Goia'16a]

Take $H_g = h_g I$.

$$L_n \longrightarrow L^\star \qquad \text{in probability} \qquad n \to \infty$$

Comments

- Proposed method $\Longleftrightarrow$ Bayes Discrimination based on PCs

  (compare with)

  [James,Hastie'01] $\Longleftrightarrow$ projective approach based on PCs Gaussian Mixture

Comments

- Proposed method $\Longleftrightarrow$ Bayes Discrimination based on PCs
  (compare with)
  [James,Hastie'01] $\Longleftrightarrow$ projective approach based on PCs Gaussian Mixture
- Distribution-Free: no distributional assumptions on f_d

Comments

- Proposed method $\Longleftrightarrow$ Bayes Discrimination based on PCs
  (compare with)
  [James,Hastie'01] $\Longleftrightarrow$ projective approach based on PCs Gaussian Mixture
- Distribution-Free: no distributional assumptions on f_d
- Model-based approach theoretically justified (not only pure projective)

Comments

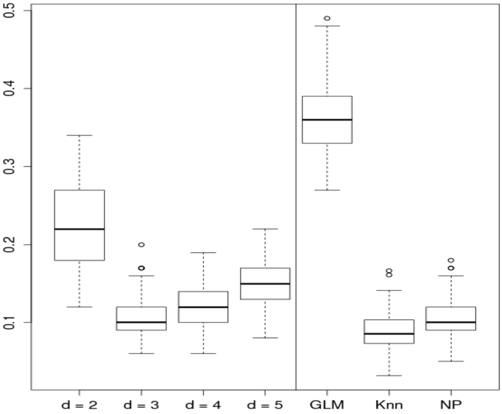
- Proposed method $\Longleftrightarrow$ Bayes Discrimination based on PCs
  (compare with)
  [James,Hastie'01] $\Longleftrightarrow$ projective approach based on PCs Gaussian Mixture
- Distribution-Free: no distributional assumptions on f_d
- Model-based approach theoretically justified (not only pure projective)
- Indeed, if eigenvalues decay of Σ is slow we can not ensure that $d_g = d \implies$ the volume terms ϕ can not be neglected (insufficiency of pure projective approach)

Comments

- Proposed method $\iff$ Bayes Discrimination based on PCs
  (compare with)
  [James,Hastie'01] $\iff$ projective approach based on PCs Gaussian Mixture
- Distribution-Free: no distributional assumptions on f_d
- Model-based approach theoretically justified (not only pure projective)
- Indeed, if eigenvalues decay of Σ is slow we can not ensure that $d_g = d \implies$ the volume terms ϕ can not be neglected (insufficiency of pure projective approach)
- Curse-of-dimensionality affects performances of the methodology

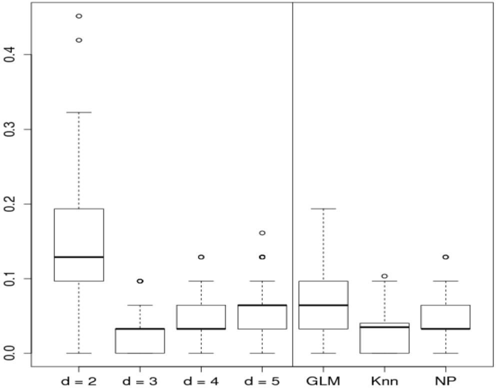
Example: Synthetic data

- Two Bananas data-set with $\pi_g = 0.5$
- Sample of size $n = 300$ split in two parts:
  training-sets ($n_{in} = 200$) $+$ test-set ($n_{out} = 100$)
- Out-of-sample misclassification error over 100 MC replications



Example: Growth curves

Out-of-sample misclassification errors over 100 MC replications (training-set is 2/3 of the sample)



Conclusions

- We introduced conditions to obtain a factorization of the SmBP

$$\text{SmBP} \sim \text{Intensity} \ \times \ \text{Volume}$$

Conclusions

- We introduced conditions to obtain a factorization of the SmBP
$$\text{SmBP} \sim \text{Intensity} \;\times\; \text{Volume}$$
- We use the intensity term as surrogate density in classification motivating theoretically the use of PCs

Conclusions

- We introduced conditions to obtain a factorization of the SmBP

$$\text{SmBP} \sim \text{Intensity} \;\times\; \text{Volume}$$

- We use the intensity term as surrogate density in classification motivating theoretically the use of PCs
- Unsupervised classification: High-Intensity Region Principle

Conclusions

- We introduced conditions to obtain a factorization of the SmBP

$$\text{SmBP} \sim \text{Intensity} \ \times \ \text{Volume}$$

- We use the intensity term as surrogate density in classification motivating theoretically the use of PCs
- Unsupervised classification: High-Intensity Region Principle
- Supervised classification: Bayes Rule Generalization

Conclusions

- We introduced conditions to obtain a factorization of the SmBP

$$\text{SmBP} \sim \text{Intensity} \times \text{Volume}$$

- We use the intensity term as surrogate density in classification motivating theoretically the use of PCs
- Unsupervised classification: High-Intensity Region Principle
- Supervised classification: Bayes Rule Generalization
- Performances are studied on simulated and real dataset

Some References

- E.B.,A.Goia:Classification methods for Hilbert data based on surrogate density. *CSDA*, 99, 204–222, (2016)
- E.B.,A.Goia:Some Insights About the Small Ball Probability Factorization for Hilbert Random Elements. *Statistica Sinica* (Preprint) doi:10.5705/ss.202016.0128
- D.Bosq: Linear Processes in Function Spaces: Theory and Applications. Lectures Notes in Statistics, 149, Springer–Verlag, Berlin (2000)
- A.Cuevas,M.Febrero,R.Fraiman. Estimating the number of clusters. *Can.J.Stat.*,28(2),367-382 (2000).
- A.Cuevas,M.Febrero,R.Fraiman. Cluster analysis: a further approach based on density estimation. *CSDA*, 36, 441–459 (2001).
- M.Ciollaro, C.R.Genovese, D.Wang: Nonparametric Clustering of Functional Data Using Pseudo-Densities (preprint)
- Delaigle,Hall:Defining probability density for a distribution of random functions.*AoS* 38(2),1171-1193 (2010)
- L.Delsol, C.Louchet (2014) Segmentation of hyperspectral images from functional kernel density estimation. in Contributions in infinite–dimensional statistics and related topics. [Eds Bongiorno et. al], Esculapio
- F.Ferraty,P.Vieu: Nonparametric functional data analysis. Theory and practice. Springer Series Stat.(2006)
- F.Ferraty, A.Goia, P.Vieu: Functional nonparametric model for time series: a fractal approach to dimension reduction. TEST. **11** 317–344 (2002)
- F.Ferraty, N.Kudraszow, P.Vieu: Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. J. Nonparametr. Stat. 24 (2), 447–464 (2012).

Some References

- F.Ferraty, A.Mas, P.Vieu: Nonparametric regression on functional data: inference and practical aspects. Aust. N. Z. J. Stat. **49**(3), 267–286 (2007)
- T.Gasser, P.Hall, B.Presnell: Nonparametric estimation of the mode of a distribution of random curves. JRSS B **60** (4), 681–691 (1998).
- J.A.Hartigan: Clustering Algorithm. Wiley, New York (1975)
- J.Jacques, C.Preda: Functional data clustering: a survey. ADAC (2014)
- J.Jacques, C.Preda: Model-based clustering for multivariate functional data. CSDA **71**, 92–106 (2014)
- G.M. James, T.J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society: Series B, **63**.3: 533–550 (2001).
- W.V.Li,Q.-M.Shao. Gaussian processes: inequalities, small ball probabilities and applications. In Stochastic processes: theory and methods,V.19 Handbook of Statist., p.533–597. North-Holland, Amsterdam, 2001.
- M.A.Lifshits: Lectures on Gaussian processes. Springer Briefs in Mathematics. Springer, Heidelberg (2012)
- J.O.Ramsay, B.W.Silverman: Functional data analysis, 2nd ed., New York: Springer (2005)
- J.H.Ward Jr.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, **58**, 236–244 (1963).
- D.Wishart (1969). Mode Analysis: A Generalization of Nearest Neighbor which Reduces Chaining Effects, in A.J. Cole (Ed.), *Numerical Taxonomy*, Academic Press, 282–311.

Thanks for your Attention