Extracting multiscale geometric information from high-dimensional and infinite-dimensional data with application to classification

Wolfgang Polonik

Department of Statistics, UC Davis

NIPS Dec. 8, 2017

Joint work with

Gabriel Chandler

Multiscale feature extraction in multi-dimensions

Goals:

Multiscale feature extraction in multi-dimensions

Goals:

(i) present a method for extracting geometric information of the underlying multivariate distribution;

Multiscale feature extraction in multi-dimensions

Goals:

(i) present a method for extracting geometric information of the underlying multivariate distribution;

(ii) accomplish (i) in a multiscale fashion;

Multiscale feature extraction in multi-dimensions

Goals:

(i) present a method for extracting geometric information of the underlying multivariate distribution;

(ii) accomplish (i) in a multiscale fashion;

(iii) provide a tool for the visualization of geometric aspects of a multivariate (or even infinite-dimensional) data set, and

Multiscale feature extraction in multi-dimensions

Goals:

(i) present a method for extracting geometric information of the underlying multivariate distribution;

(ii) accomplish (i) in a multiscale fashion;

(iii) provide a tool for the visualization of geometric aspects of a multivariate (or even infinite-dimensional) data set, and

(iv) show that the extracted features might be useful for inference (classification).

Multiscale feature extraction in multi-dimensions

Goals:

(i) present a method for extracting geometric information of the underlying multivariate distribution;

(ii) accomplish (i) in a multiscale fashion;

(iii) provide a tool for the visualization of geometric aspects of a multivariate (or even infinite-dimensional) data set, and

(iv) show that the extracted features might be useful for inference (classification).

Will will use notions of

- depth (Tukey depth)
- FDA

Relations and inspirations

- mass estimation (Ting et al. 2012)
- shorth plot (Einmahl et al. 2010)
- local depth (Agostinelli and Ramanazzi, 2011, Paindaveine and van Bever, 2012, and Dutta et al., 2015)

Some basic challenges in non-parametric learning

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?
  - Curse of dimensionality;
  - right 'size' of subsets; mass concentration

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?
  - Curse of dimensionality;
  - right 'size' of subsets; mass concentration
- Where to look (for features) in large dimensions?

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?
  - Curse of dimensionality;
  - right 'size' of subsets; mass concentration
- Where to look (for features) in large dimensions?
- 'kernel-trick' (non-linear methods; comput. advantage)

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?
  - Curse of dimensionality;
  - right 'size' of subsets; mass concentration
- Where to look (for features) in large dimensions?
- 'kernel-trick' (non-linear methods; comput. advantage)

  Challenge: Interpretability; how to choose kernel?

Some basic challenges in non-parametric learning

- Choosing ‘right’ scale in large dimensions?
  - Curse of dimensionality;
  - right ‘size’ of subsets; mass concentration
- Where to look (for features) in large dimensions?
- ‘kernel-trick’ (non-linear methods; comput. advantage)

  Challenge: Interpretability; how to choose kernel?
- Choose a fixed tuning parameter or consider all values?

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?
  - Curse of dimensionality;
  - right 'size' of subsets; mass concentration
- Where to look (for features) in large dimensions?
- 'kernel-trick' (non-linear methods; comput. advantage)

  Challenge: Interpretability; how to choose kernel?
- Choose a fixed tuning parameter or consider all values?
  - SiZer, mode tree, persistent homology consider all values.

Some basic challenges in non-parametric learning

- Choosing 'right' scale in large dimensions?
  - Curse of dimensionality;
  - right 'size' of subsets; mass concentration
- Where to look (for features) in large dimensions?
- 'kernel-trick' (non-linear methods; comput. advantage)

  Challenge: Interpretability; how to choose kernel?
- Choose a fixed tuning parameter or consider all values?
  - SiZer, mode tree, persistent homology consider all values.

  But: Limit cases are not meaningful.

Basic idea

IDEA:

Basic idea

IDEA:

- Construct a feature map driven by geometric consideration (in contrast to RKHS-type);

Basic idea

IDEA:

- Construct a feature map driven by geometric consideration (in contrast to RKHS-type);
- Features are real-valued functions on $[0, 1]$ that

Basic idea

IDEA:

- Construct a feature map driven by geometric consideration (in contrast to RKHS-type);
- Features are real-valued functions on $[0, 1]$ that

$$\text{thus can be plotted} \rightsquigarrow \text{visualization,}$$

Basic idea

IDEA:

- Construct a feature map driven by geometric consideration (in contrast to RKHS-type);
- Features are real-valued functions on $[0,1]$ that

thus can be plotted $\rightsquigarrow$ visualization,

contain geometric information $\rightsquigarrow$ interpretability.

Wine data

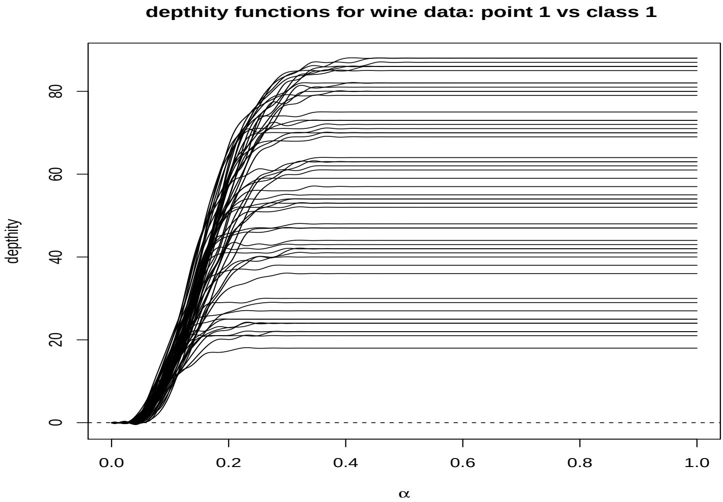
from UC Irvine Machine Learning Repository.

177 observations in 13 dimensions

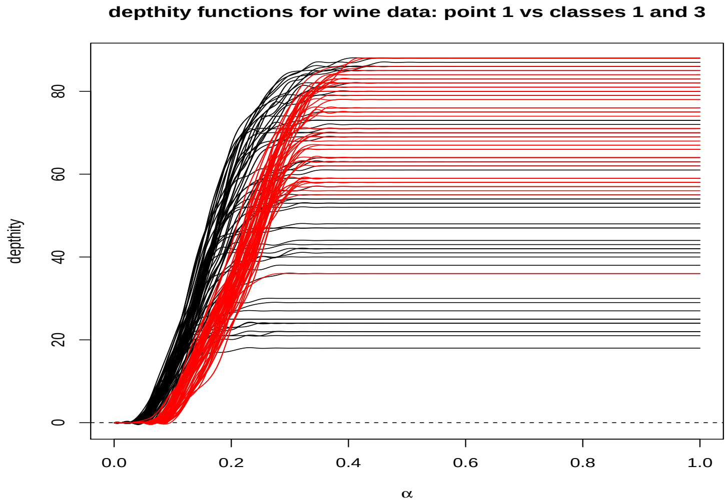
3 classes (labeled) [58 in class 1, 70 class 2, 49 class 3]

Feature functions for wine data

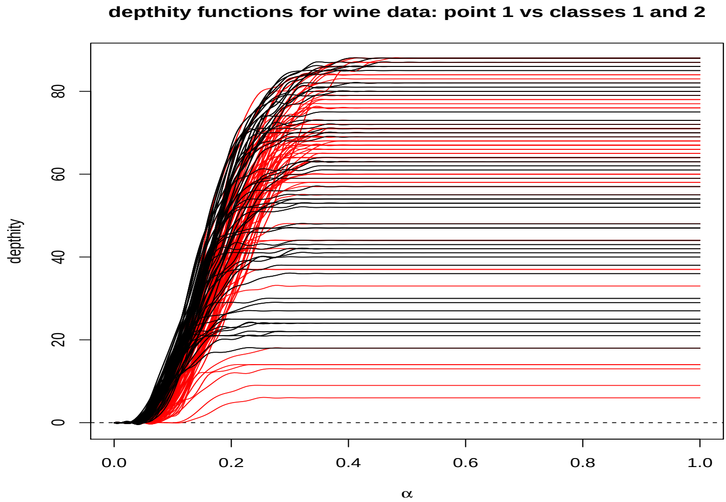
Feature functions for wine data



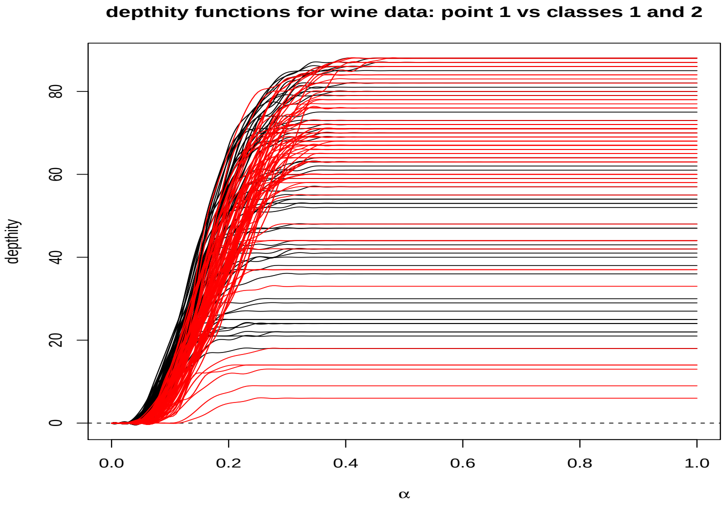
Feature functions for 13-dimensional wine data



Feature functions for 13-dimensional wine data



Feature functions for 13-dimensional wine data



Multiscale feature extraction

Multiscale feature extraction

Novel idea (inspired by Ting et al. 2012):

Multiscale feature extraction

Novel idea (inspired by Ting et al. 2012):

- Define a distribution of depths for a given point x

  $\rightsquigarrow$ corresponding quantile functions $\widehat{q}_x(\delta)$, $0 \leq \delta \leq 1$, are feature functions.

Multiscale feature extraction

Novel idea (inspired by Ting et al. 2012):

- Define a distribution of depths for a given point x

  $\rightsquigarrow$ corresponding quantile functions $\widehat{q}_x(\delta)$, $0 \leq \delta \leq 1$, are feature functions.

- Distribution of depths is constructed by randomly selecting subsets containing x, and finding depths of x within these subsets.

Multiscale feature extraction

Novel idea (inspired by Ting et al. 2012):

- Define a distribution of depths for a given point x

  $\rightsquigarrow$ corresponding quantile functions $\widehat{q}_x(\delta)$, $0 \leq \delta \leq 1$, are feature functions.

- Distribution of depths is constructed by randomly selecting subsets containing x, and finding depths of x within these subsets.

We propose to use (circular) cones as basic subsets.

Multiscale feature extraction

Multiscale feature extraction

Different scales:

Multiscale feature extraction

Different scales:

- *small quantiles* $\rightsquigarrow$ *local information (density)*

Multiscale feature extraction

Different scales:

- *small quantiles* $\rightsquigarrow$ *local information (density)*
- *large quantiles* $\rightsquigarrow$ *global information (depth)*

Multiscale feature extraction

Different scales:

- *small quantiles* $\rightsquigarrow$ *local information (density)*
- *large quantiles* $\rightsquigarrow$ *global information (depth)*
- *intermediate quantiles? (How important are they for high dimensions?)*

  $\rightsquigarrow$ *multiscale (scale given by quantile level)*

Depth quantile functions

Depth quantile functions

Intuition:

'Sit at a (data) point and look in one direction' - depth quantile function describes aspects of topographical information of what can be seen.

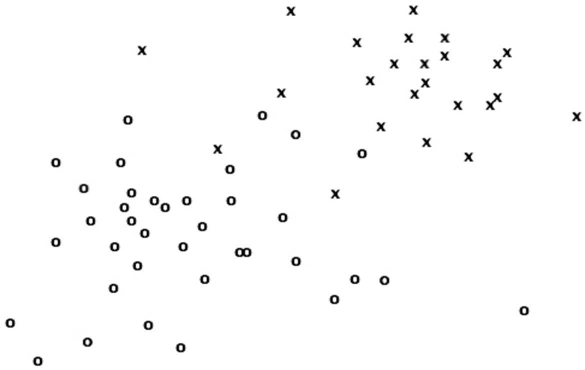
Depth quantile functions

Intuition:

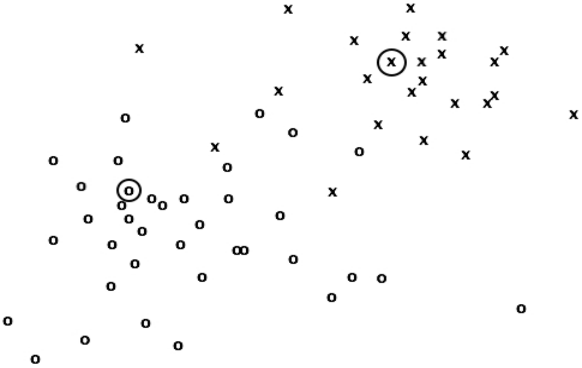
> *'Sit at a (data) point and look in one direction' - depth quantile function describes aspects of topographical information of what can be seen.*

QUESTION: How to find relevant directions?

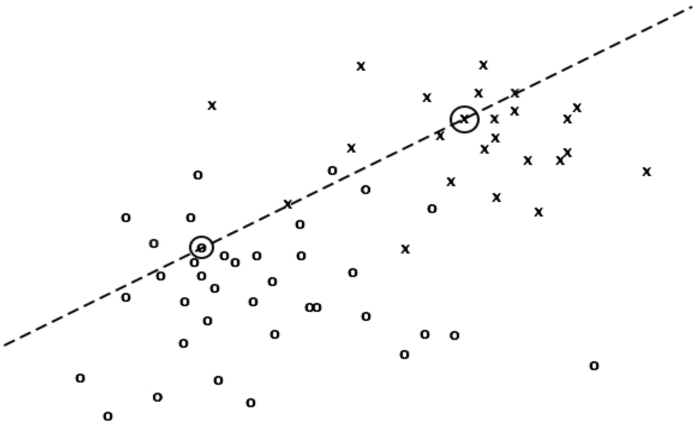
Construction of depth quantile functions



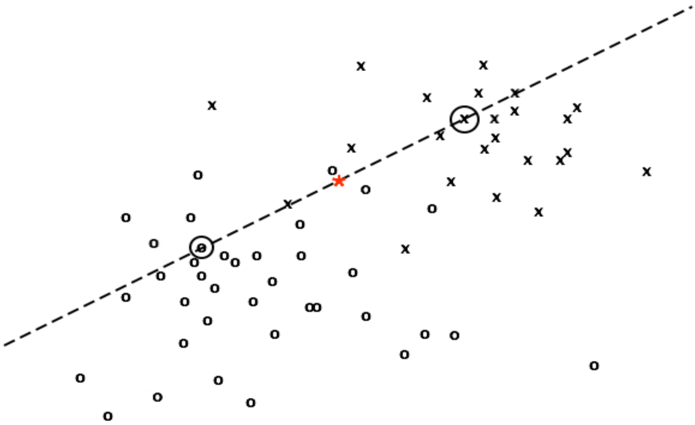
Construction of depth quantile functions



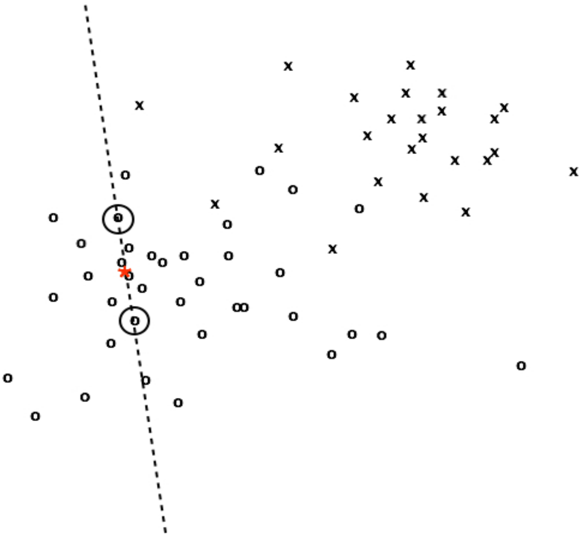
Construction of depth quantile functions



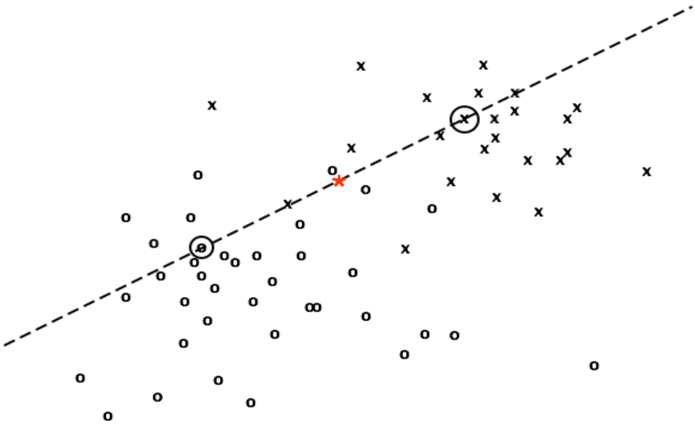
Construction of depth quantile functions



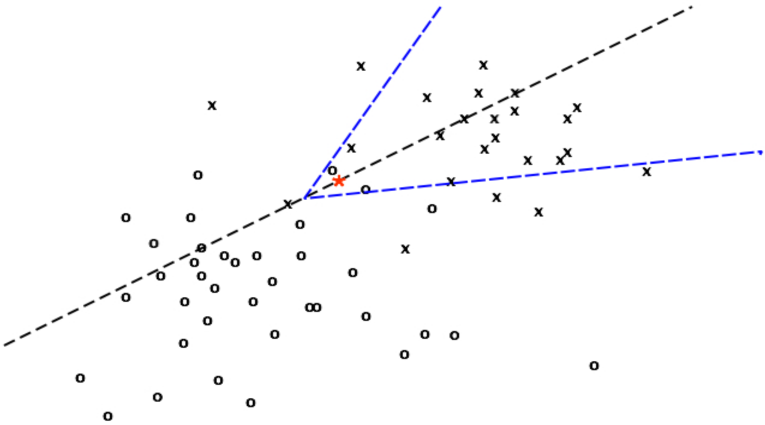
Construction of depth quantile functions



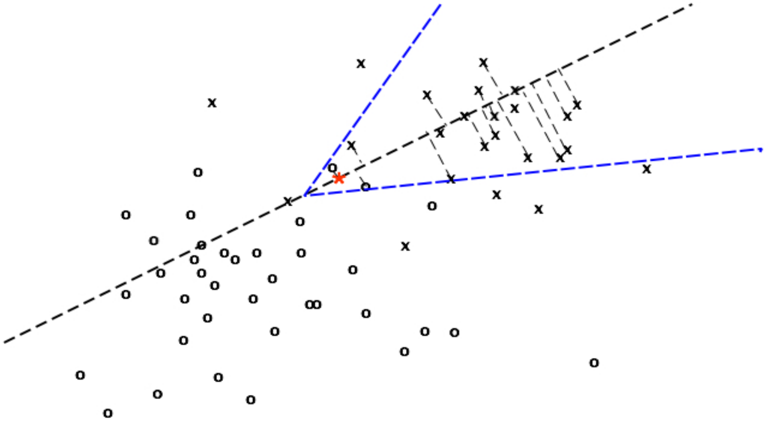
Construction of depth quantile functions



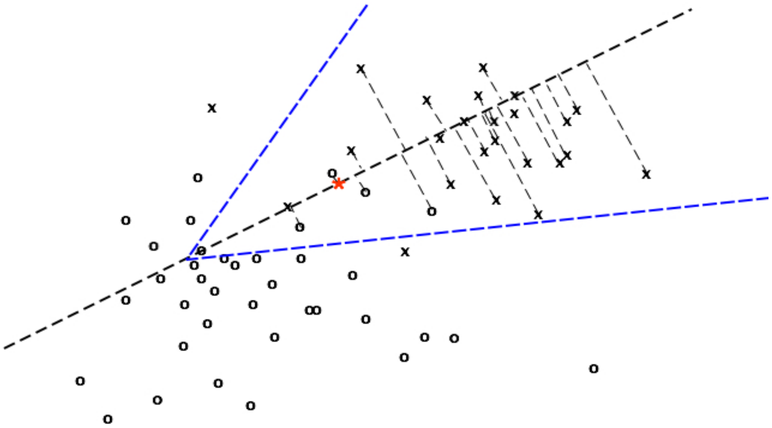
Construction of depth quantile functions



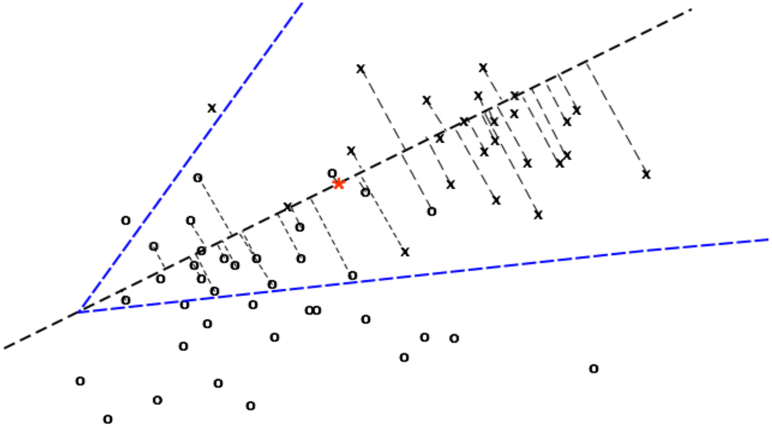
Construction of depth quantile functions



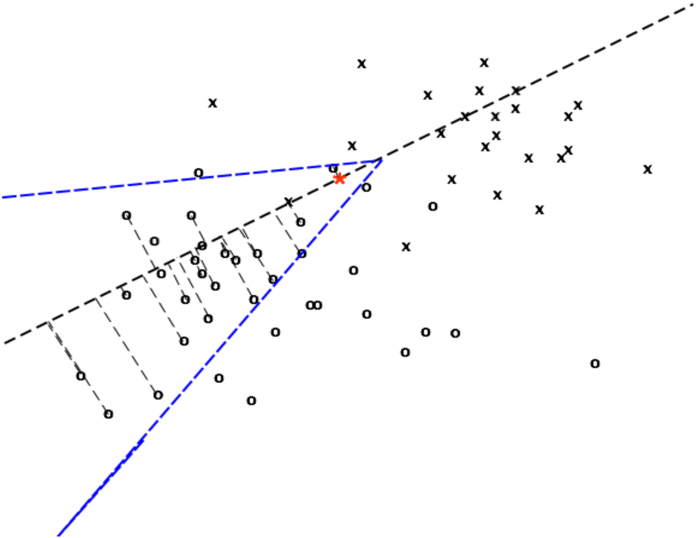
Construction of depth quantile functions



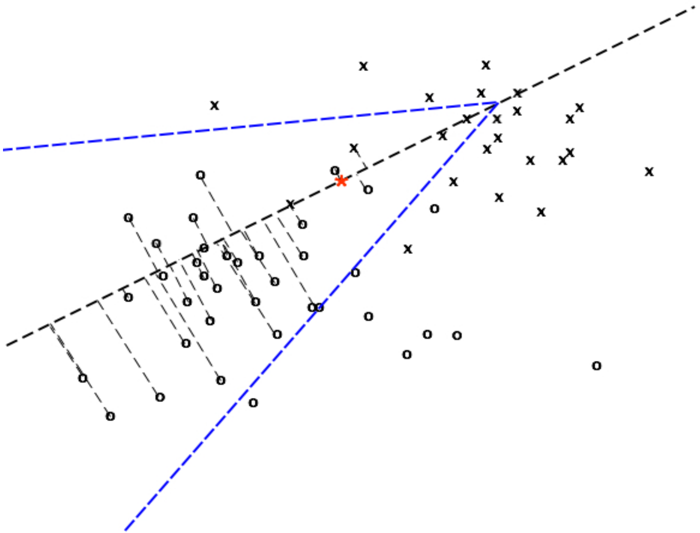
Construction of depth quantile functions



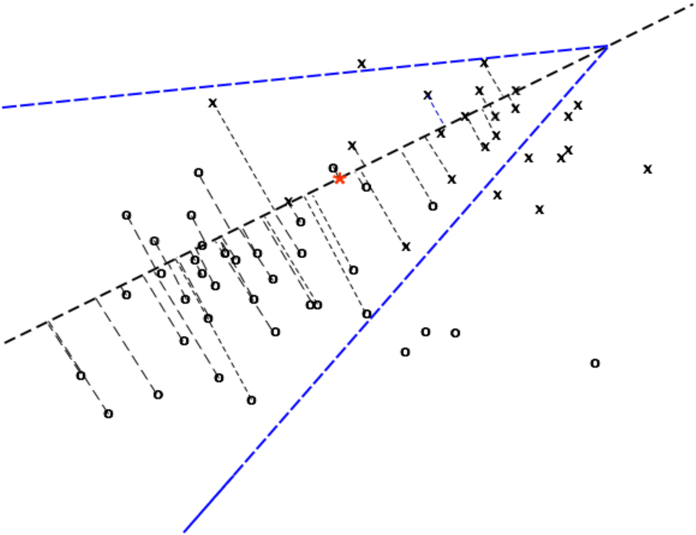
Construction of depth quantile functions



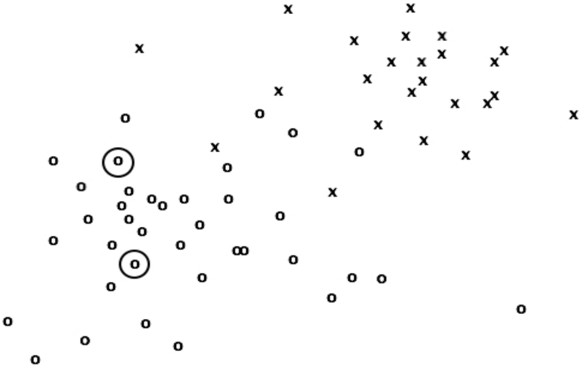
Construction of depth quantile functions



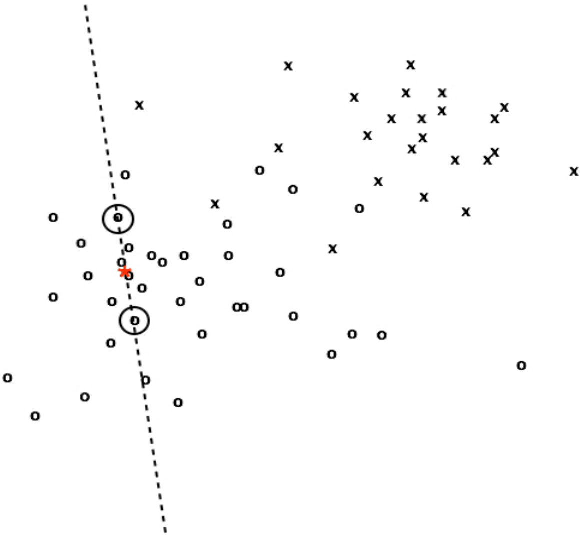
Construction of depth quantile functions



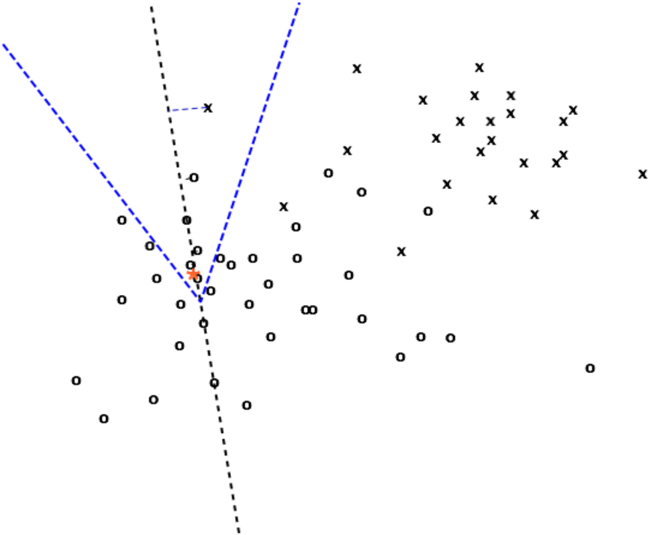
Construction of depth quantile functions



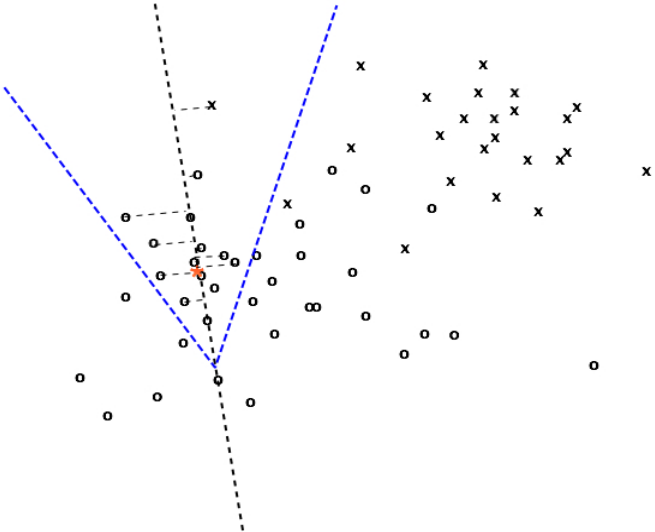
Construction of depth quantile functions



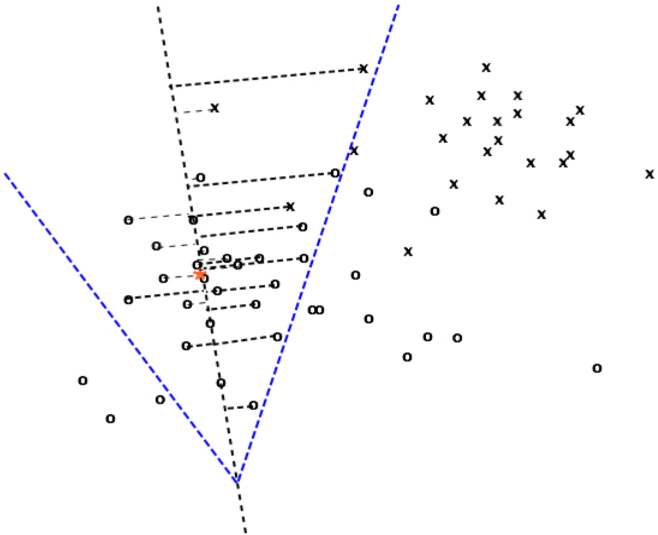
Construction of depth quantile functions



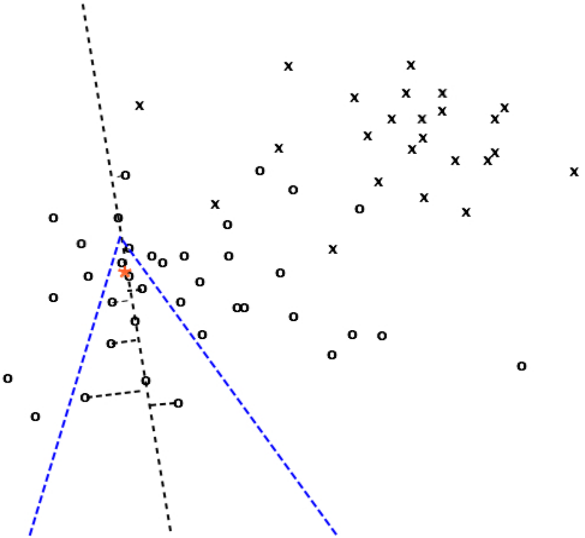
Construction of depth quantile functions



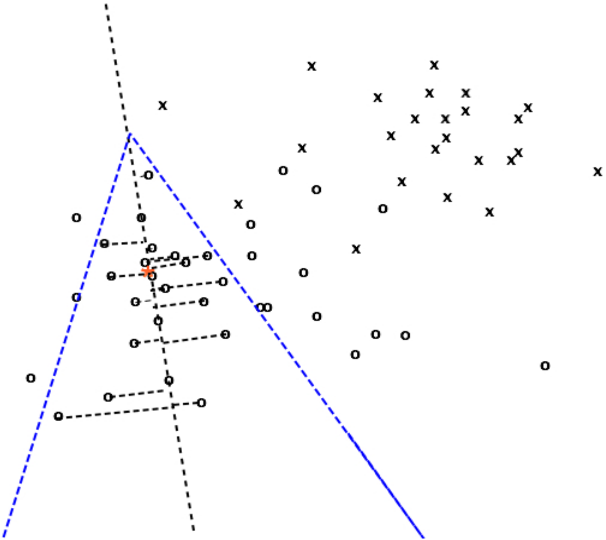
Construction of depth quantile functions



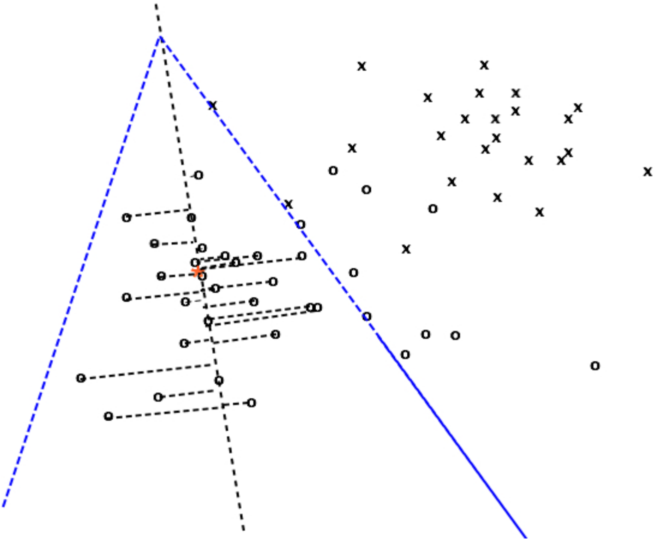
Construction of depth quantile functions



Construction of depth quantile functions



Construction of depth quantile functions



Population versions

Population versions

Given a line $\ell \in \mathbb{R}^d$ and $s, x \in \ell$, let $C_x(s)$ denote a cone with

- opening angle α fixed
- tip in $s \in \ell$
- $x \in C_x(s)$

Population versions

Given a line $\ell \in \mathbb{R}^d$ and $s, x \in \ell$, let $C_x(s)$ denote a cone with

- opening angle α fixed
- tip in $s \in \ell$
- $x \in C_x(s)$

tip s moves in both directions away from x, such that $x \in C_x(s)$.

Population versions

Given a line $\ell \in \mathbb{R}^d$ and $s, x \in \ell$, let $C_x(s)$ denote a cone with

- opening angle α fixed
- tip in $s \in \ell$
- $x \in C_x(s)$

tip s moves in both directions away from x, such that $x \in C_x(s)$.

Split cone into two parts at x:

- $A_x(s)$ subcone of $C_x(s)$ with x the midpoint of its base,
- $B_x(s) = C_x(s) \setminus A_x(s)$ ('frustum')

Population versions

Given a line $\ell \in \mathbb{R}^d$ and $s, x \in \ell$, let $C_x(s)$ denote a cone with

- opening angle α fixed
- tip in $s \in \ell$
- $x \in C_x(s)$

tip s moves in both directions away from x, such that $x \in C_x(s)$.

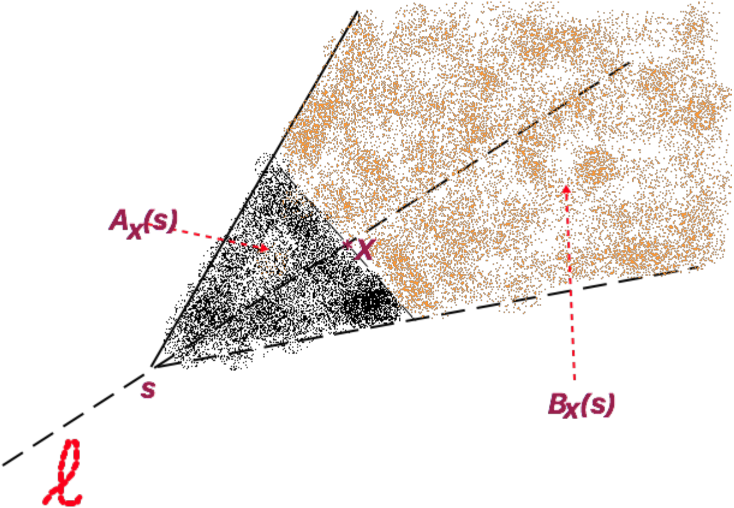
Split cone into two parts at x:

- $A_x(s)$ subcone of $C_x(s)$ with x the midpoint of its base,
- $B_x(s) = C_x(s) \setminus A_x(s)$ ('frustum')

Let

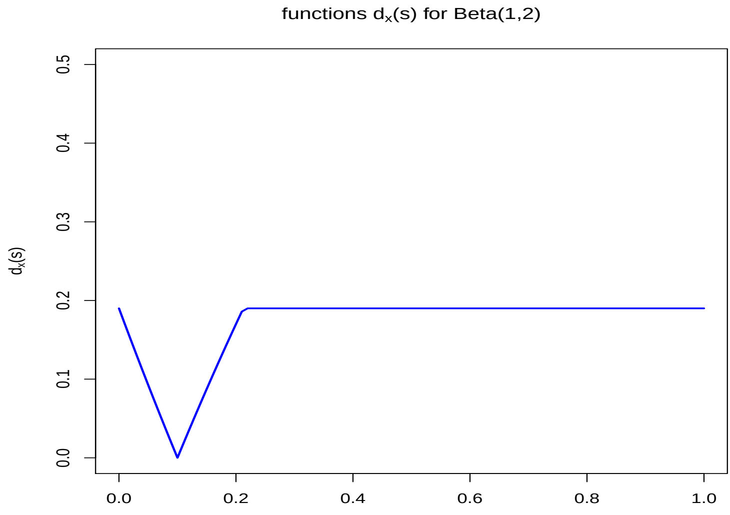
$$d_{x,\ell}(s) = \min\left\{F(A_x(s)), F(B_x(s))\right\}.$$

Population versions



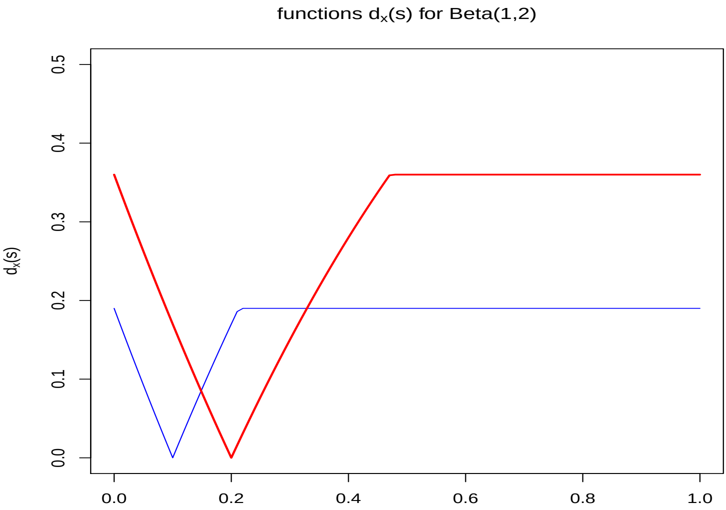
How do these depth functions look like as a function of s?

Examples



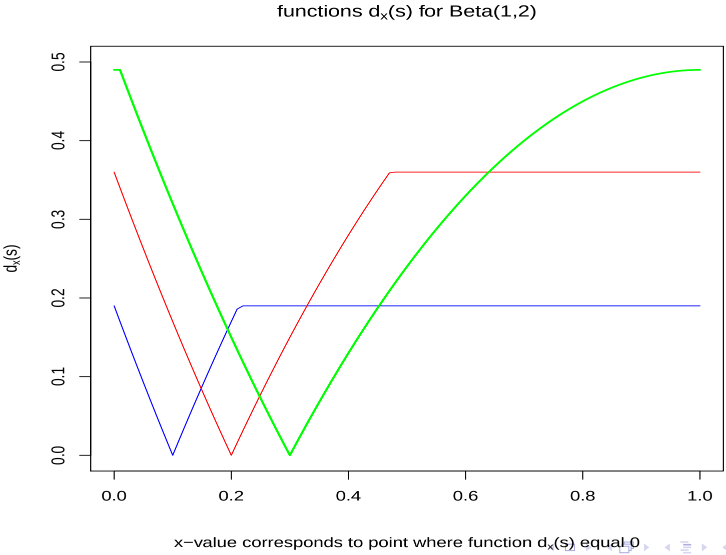
x−value corresponds to point where function $d_x(s)$ equal 0

Examples



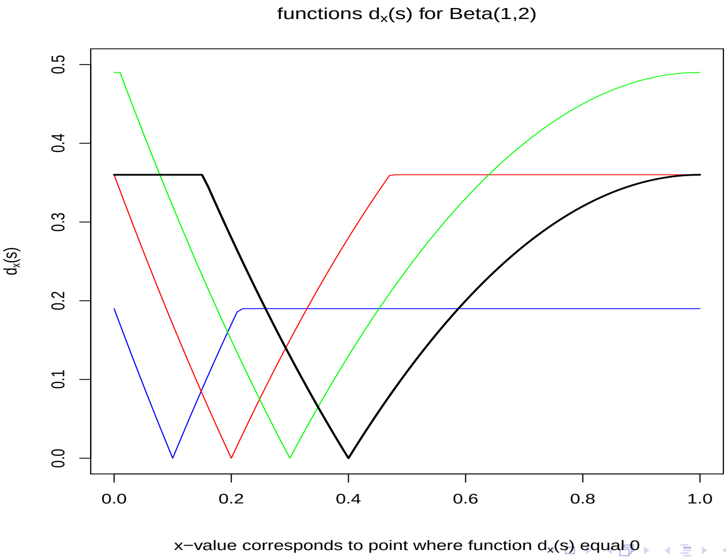
x−value corresponds to point where function $d_x(s)$ equal 0

Examples

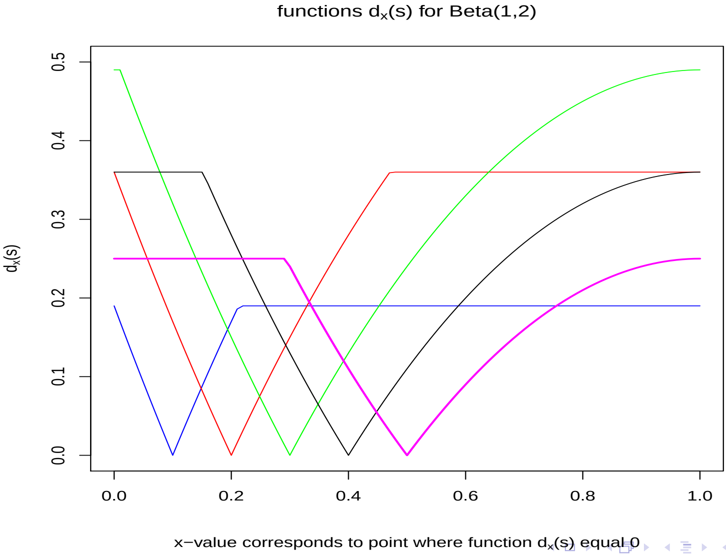


functions $d_x(s)$ for Beta(1,2)
$d_x(s)$
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
x−value corresponds to point where function $d_x(s)$ equal 0

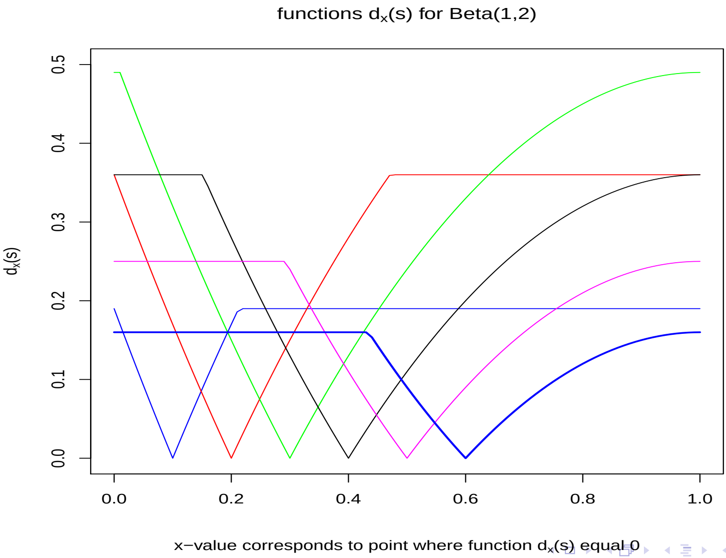
Examples


functions $d_x(s)$ for Beta(1,2)
$d_x(s)$
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
x−value corresponds to point where function $d_x(s)$ equal 0

Examples

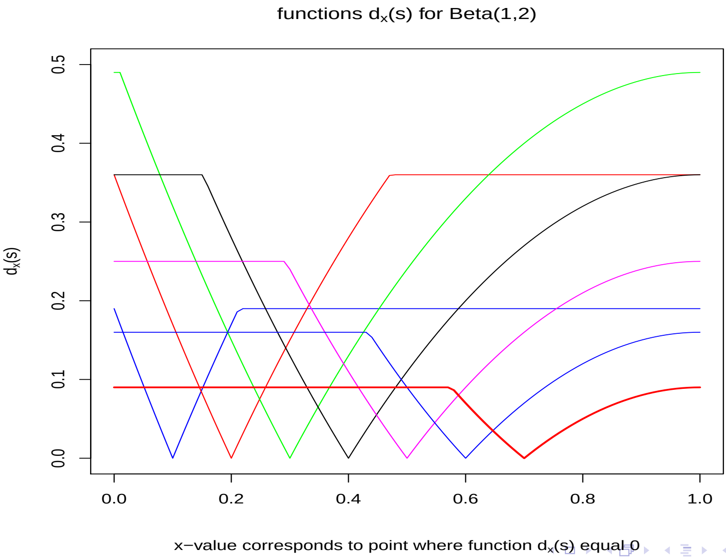

functions $d_x(s)$ for Beta(1,2)
$d_x(s)$
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
x−value corresponds to point where function $d_x(s)$ equal 0

Examples



functions $d_x(s)$ for Beta(1,2)
$d_x(s)$
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
x−value corresponds to point where function $d_x(s)$ equal 0

Examples


functions $d_x(s)$ for Beta(1,2)
$d_x(s)$
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
x−value corresponds to point where function $d_x(s)$ equal 0

Depth quantile functions

Choose cone tip s randomly, i.e let $S \sim G$ on ℓ. Consider the cdf of $d_{x,\ell}(S)$

$$P(d_{x,\ell}(S) \leq t) = G\big(s \in \ell : d_{x,\ell}(s) \leq t\big).$$

Depth quantile functions

Choose cone tip s randomly, i.e let $S \sim G$ on ℓ. Consider the cdf of $d_{x,\ell}(S)$

$$P(d_{x,\ell}(S) \leq t) = G\big(s \in \ell : d_{x,\ell}(s) \leq t\big).$$

and set

$$q_{x,\ell}(\delta) = \inf\big\{t : G(s : d_{x,\ell}(s)) \leq t) \geq \delta\big\}$$

Depth quantile functions

Choose cone tip s randomly, i.e let $S \sim G$ on ℓ. Consider the cdf of $d_{x,\ell}(S)$

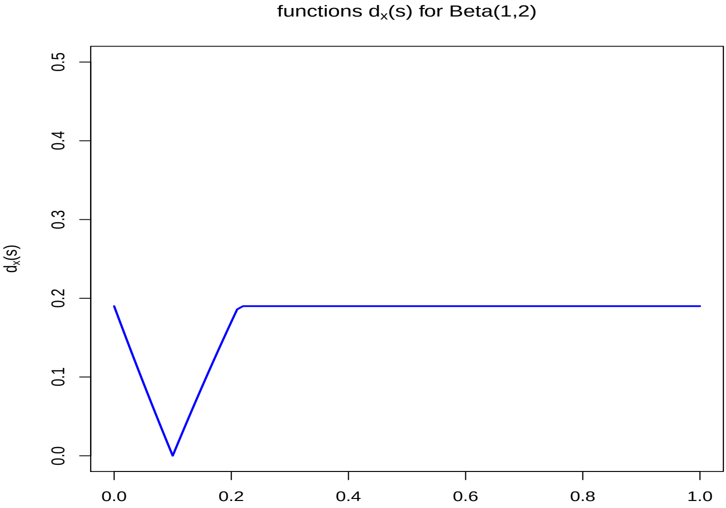
$$P(d_{x,\ell}(S) \leq t) = G\big(s \in \ell : d_{x,\ell}(s) \leq t\big).$$

and set

$$q_{x,\ell}(\delta) = \inf \big\{ t : G(s : d_{x,\ell}(s)) \leq t) \geq \delta \big\}$$

$\rightsquigarrow$ put all depth functions onto same 'scale'

Examples



x−value corresponds to point where function $d_x(s)$ equal 0

Questions

Questions

- *What information is contained in depth quantile functions?*
- *How to use for statistical inference?*

Information contained in depth quantile functions

Information contained in depth quantile functions

- $\lim_{\delta \to 1} q_{x,\ell}(\delta) =$ "Tukey depth of x among projections of data onto ℓ";

Information contained in depth quantile functions

- $\lim_{\delta \to 1} q_{x,\ell}(\delta) =$ "Tukey depth of x among projections of data onto ℓ";

- $\lim_{\delta \to 0} \frac{q_{x,\ell}(\delta)}{\alpha^d} = C \frac{f(x)}{g(x)}$, where C is known (localization)

  f, g are densities of F and G.

"multiscale"

Empirical versions

Empirical versions are obtained

- by replacing F by the empirical distribution F_n;

Empirical versions

Empirical versions are obtained

- by replacing F by the empirical distribution F_n;

- for each pair (X_i, X_j), considering the line ℓ_{ij} passing through both X_i and X_j

Empirical versions

Empirical versions are obtained

- by replacing F by the empirical distribution F_n;

- for each pair (X_i, X_j), considering the line ℓ_{ij} passing through both X_i and X_j

- letting $x = \frac{X_i+X_j}{2}$.

Resulting empirical depth quantile functions are denoted by $\widehat{q}_{ij}(\delta)$.

Empirical versions

Empirical versions are obtained

- by replacing F by the empirical distribution F_n;

- for each pair (X_i, X_j), considering the line ℓ_{ij} passing through both X_i and X_j

- letting $x = \frac{X_i + X_j}{2}$.

Resulting empirical depth quantile functions are denoted by $\widehat{q}_{ij}(\delta)$.

Computation: Embarrassingly parallelizable (if needed).

Averaged feature functions

Averaged feature functions

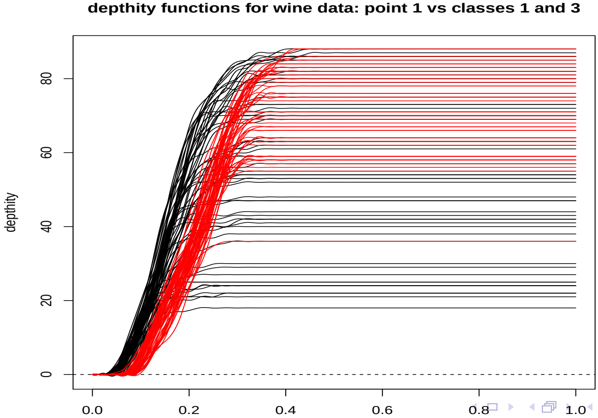
Suppose, for each pair (X_i, X_j), we have $\widehat{q}_{ij}(\delta) \rightsquigarrow \binom{n}{2}$ feature functions.

Reduce total number of functions by averaging.

Averaged feature functions

Suppose, for each pair (X_i, X_j), we have $\widehat{q}_{ij}(\delta) \rightsquigarrow \binom{n}{2}$ feature functions.

Reduce total number of functions by averaging.



Averaged feature functions

$K \geq 1$ classes

Averaged feature functions

$K \geq 1$ classes

- For each fixed X_i, average $\widehat{q}_{ij}(\delta)$ over all X_j in class k

$$\rightsquigarrow \quad \widehat{q}_i^{(k)}(\delta) = \operatorname*{ave}_{X_j \in \text{ group } k} \widehat{q}_{ij}(\delta)$$

Averaged feature functions

$K \geq 1$ classes

- For each fixed X_i, average $\widehat{q}_{ij}(\delta)$ over all X_j in class k

$$\rightsquigarrow \quad \widehat{q}_i^{(k)}(\delta) = \underset{X_j \in \text{ group } k}{\text{ave}} \widehat{q}_{ij}(\delta)$$

- For each point, we obtain K functions $(\widehat{q}_i^{(1)}(\delta), \ldots, \widehat{q}_i^{(K)}(\delta))$

Example: Iris data

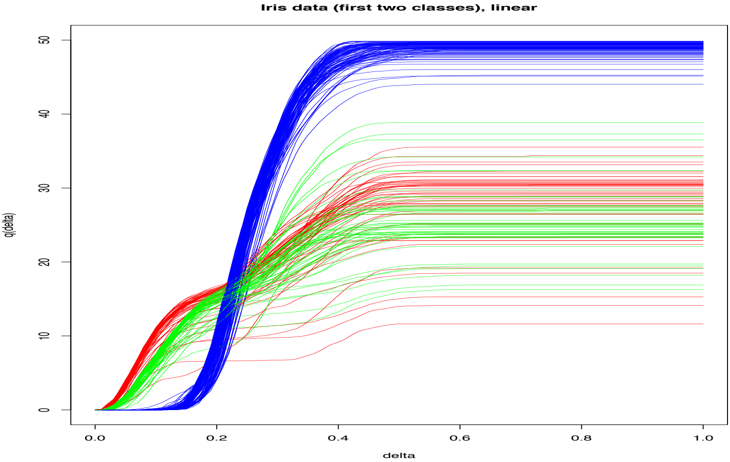
Example: Iris data

Iris data (Fisher, 1936); $d = 4$; $K = 3$, $n = 150$

Example: Iris data

Iris data (Fisher, 1936); $d = 4$; $K = 3$, $n = 150$

only used classes 1 and 2



Another example

Another example

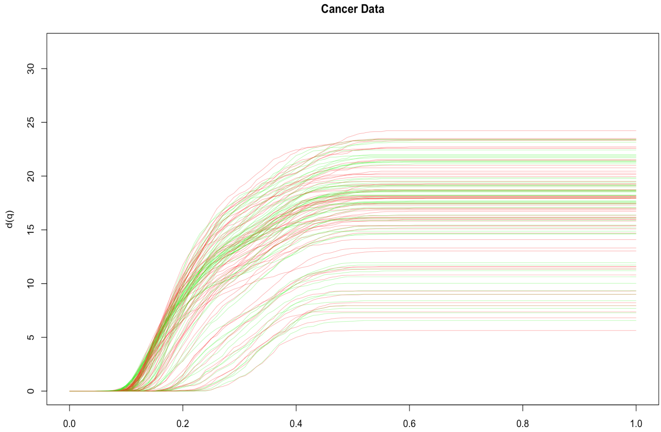
gene expression data; $d = 2000$, $n = 62$, 2 classes (normal tissue 22, tumor tissue 40) Alon et al. 1999, PNAS

(http://genomics-pubs.princeton.edu/oncology/affydata/)

Another example

gene expression data; $d = 2000$, $n = 62$, 2 classes (normal tissue 22, tumor tissue 40) Alon et al. 1999, PNAS

(http://genomics-pubs.princeton.edu/oncology/affydata/)



More information on depth quantile functions

More information on depth quantile functions

Enhance understanding of our approach by relating it to

More information on depth quantile functions

Enhance understanding of our approach by relating it to

- *multidimensional scaling*

More information on depth quantile functions

Enhance understanding of our approach by relating it to

- *multidimensional scaling*
- *Choquet capacities*

More information on depth quantile functions

Enhance understanding of our approach by relating it to

- *multidimensional scaling*
- *Choquet capacities*
- *shorth plot (one-dimensional case)*

Depth quantiles and multidimensional scaling

Depth quantiles and multidimensional scaling

Observe: Given a line $\ell \subset \mathbb{R}^d$, depth quantile functions only depend on number of points in cones (with axis of symmetry being ℓ).

Depth quantiles and multidimensional scaling

Observe: Given a line $\ell \subset \mathbb{R}^d$, depth quantile functions only depend on number of points in cones (with axis of symmetry being ℓ).

To determine whether a data point falls into a given circular cone, all we need are two one-dimensional quantities (depending on line)

Depth quantiles and multidimensional scaling

Observe: Given a line $\ell \subset \mathbb{R}^d$, depth quantile functions only depend on number of points in cones (with axis of symmetry being ℓ).

To determine whether a data point falls into a given circular cone, all we need are two one-dimensional quantities (depending on line)

- (signed) distance of projection onto line from x (Z_1^x)
- distance to line (Z_2^x)

Depth quantiles and multidimensional scaling

Observe: Given a line $\ell \subset \mathbb{R}^d$, depth quantile functions only depend on number of points in cones (with axis of symmetry being ℓ).

To determine whether a data point falls into a given circular cone, all we need are two one-dimensional quantities (depending on line)

- (signed) distance of projection onto line from x (Z_1^x)
- distance to line (Z_2^x)

Used two-dimensional data: (Z_{1i}^x, Z_{2i}^x), $i = 1, \ldots, n$

$\rightsquigarrow$ two-dimensional depth-quantile functions with $x = 0$; exactly the same as depth quantile functions based on original high-dimensional data

$\rightsquigarrow$ spirit of multidimensional scaling

Depth quantiles and multidimensional scaling

Given data, our construction gives

$\rightsquigarrow \binom{n}{2}$ different lines

$\rightsquigarrow \binom{n}{2}$ different (Z_1, Z_2)-plots

Depth quantiles and multidimensional scaling

Given data, our construction gives

$\rightsquigarrow \binom{n}{2}$ different lines

$\rightsquigarrow \binom{n}{2}$ different (Z_1, Z_2)-plots

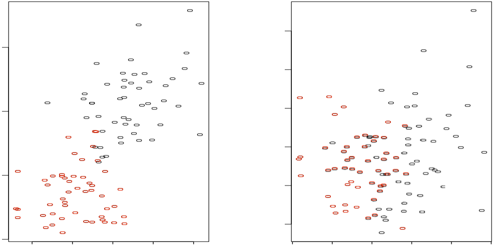


Figure 3: First two classes of iris data

different class　　　　same class

Depth quantiles and multidimensional scaling

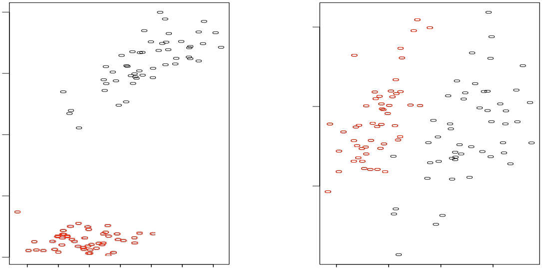
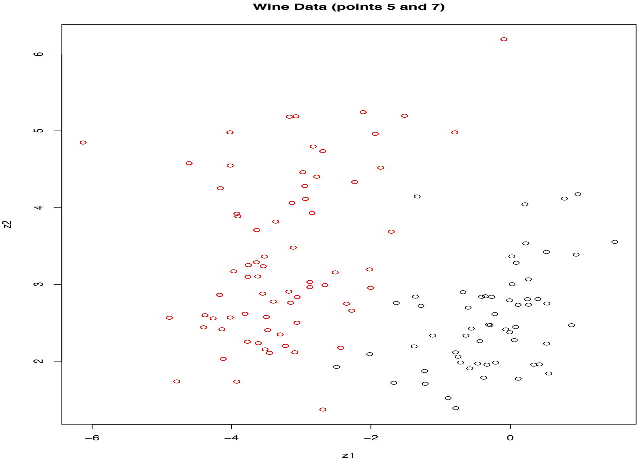


Figure 2: Second and third classes of iris data

different class　　　　　same class

Depth quantiles and multidimensional scaling



Depth quantiles in infinite dimensions

Depth quantiles in infinite dimensions

All the above only depends on dot-products

$\rightsquigarrow$ can be applied to data in Hilbert space

Depth quantiles in infinite dimensions

All the above only depends on dot-products

$\rightsquigarrow$ can be applied to data in Hilbert space

- functional data

Depth quantiles in infinite dimensions

All the above only depends on dot-products

$\rightsquigarrow$ can be applied to data in Hilbert space

- functional data
- kernelization

Depth quantiles in infinite dimensions

All the above only depends on dot-products

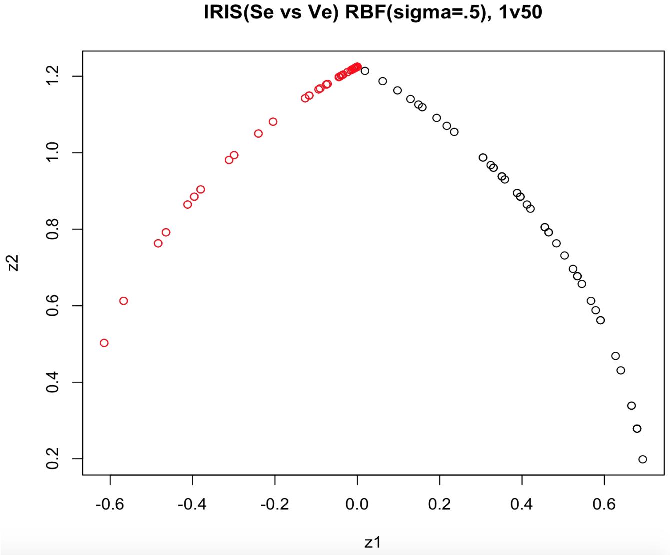
$\rightsquigarrow$ can be applied to data in Hilbert space

- functional data
- kernelization

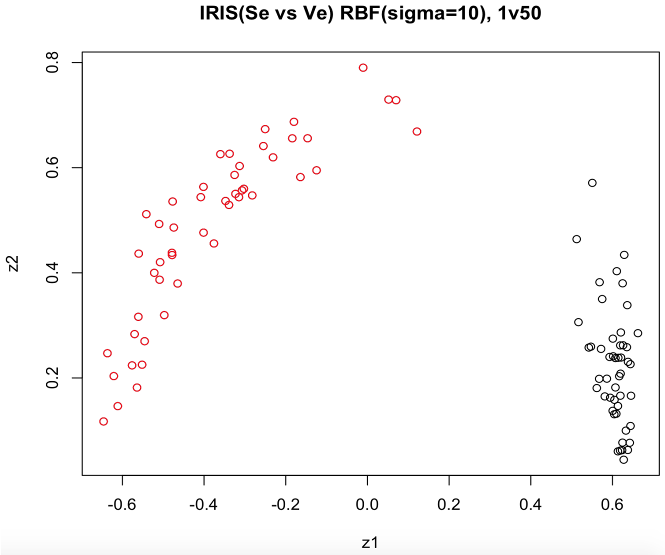
In particular:

Visualization of RKHS geometries

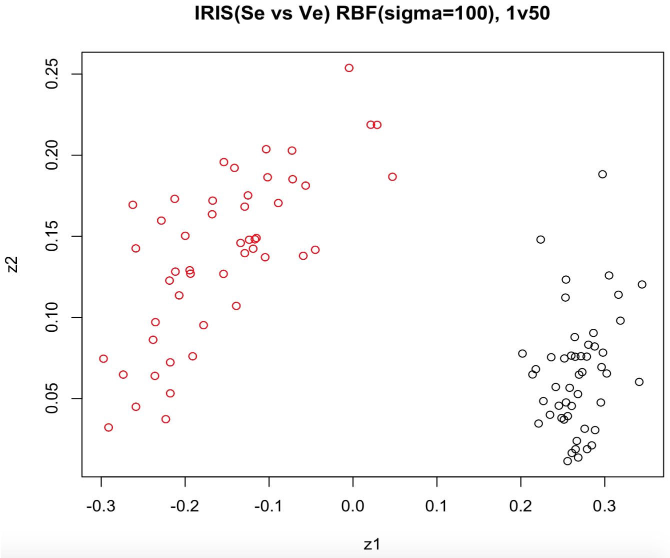
Examples: Iris data



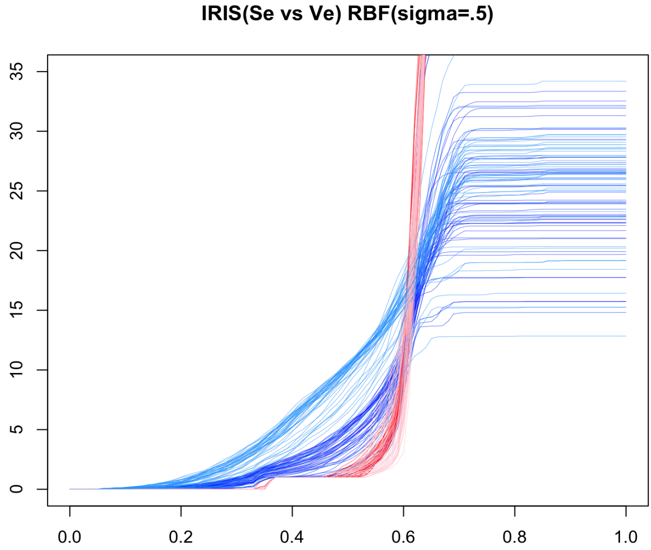
Examples: Iris data



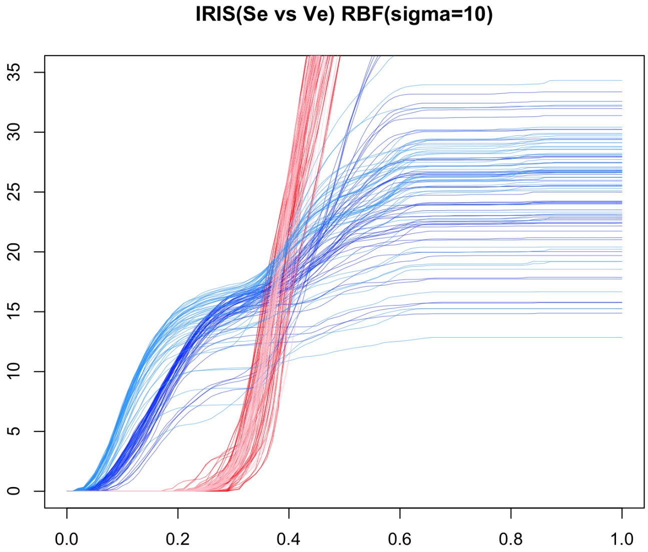
Examples: Iris data



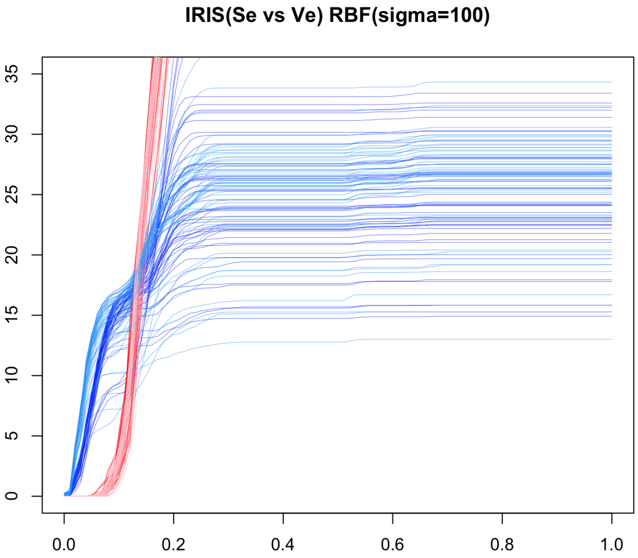
Examples: Iris data



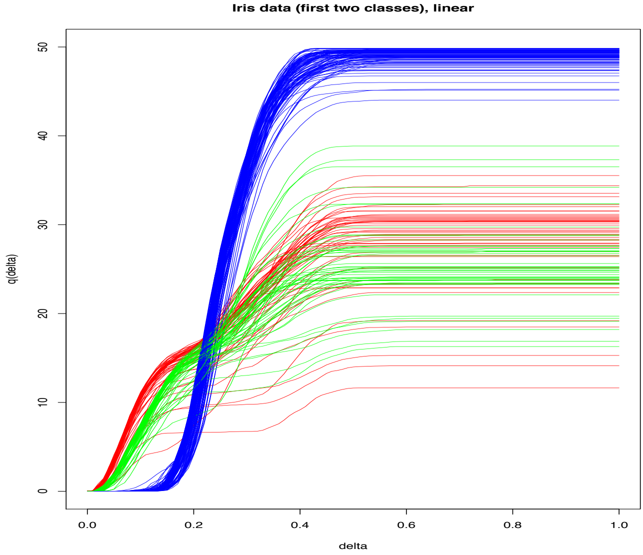
Examples: Iris data



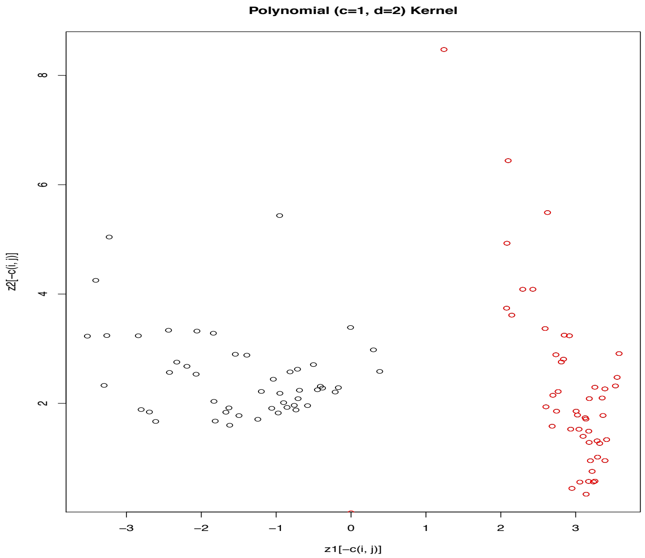
Examples: Iris data



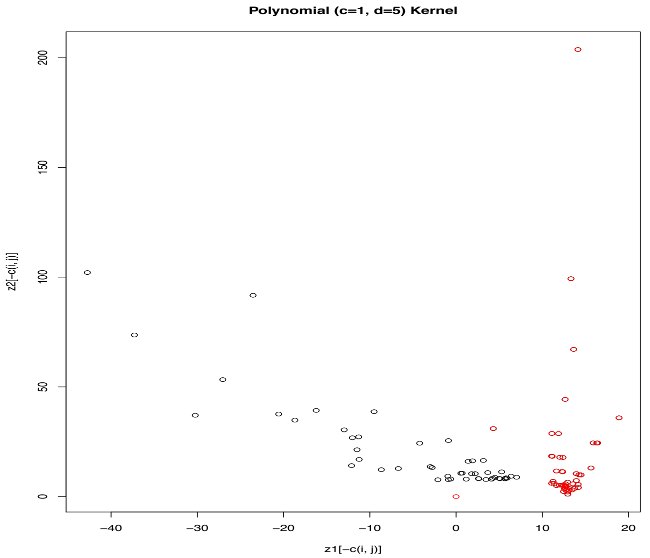
Examples: Iris data



Examples: Iris data



Examples: Iris data



Statistical inference based on averaged feature functions

Classification analysis using FDA

Statistical inference based on averaged feature functions

Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

Statistical inference based on averaged feature functions

Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

For simplicity: binary classification problem, say classes 1 and 2.

- for X^* to be classified find $\widehat{q}_i^{(1)}(\delta)$ and $\widehat{q}_i^{(2)}(\delta)$

Statistical inference based on averaged feature functions

Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

For simplicity: binary classification problem, say classes 1 and 2.

- for X^* to be classified find $\widehat{q}_i^{(1)}(\delta)$ and $\widehat{q}_i^{(2)}(\delta)$
- perform fPCA for each of these functions, keeping first p scores

Statistical inference based on averaged feature functions

Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

For simplicity: binary classification problem, say classes 1 and 2.

- for X^* to be classified find $\widehat{q}_i^{(1)}(\delta)$ and $\widehat{q}_i^{(2)}(\delta)$
- perform fPCA for each of these functions, keeping first p scores
- $\rightsquigarrow$ $2p$-dimensional vector of scores

Statistical inference based on averaged feature functions

Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

For simplicity: binary classification problem, say classes 1 and 2.

- for X^* to be classified find $\widehat{q}_i^{(1)}(\delta)$ and $\widehat{q}_i^{(2)}(\delta)$
- perform fPCA for each of these functions, keeping first p scores
- $\rightsquigarrow$ $2p$-dimensional vector of scores
- already have n such $2p$-dimensional vectors from training data (two classes)

Statistical inference based on averaged feature functions

Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

For simplicity: binary classification problem, say classes 1 and 2.

- for X^* to be classified find $\widehat{q}_i^{(1)}(\delta)$ and $\widehat{q}_i^{(2)}(\delta)$
- perform fPCA for each of these functions, keeping first p scores
- $\rightsquigarrow$ $2p$-dimensional vector of scores
- already have n such $2p$-dimensional vectors from training data (two classes)
- find classification rule (SVM; kernel SVM: etc.)

Statistical inference based on averaged feature functions

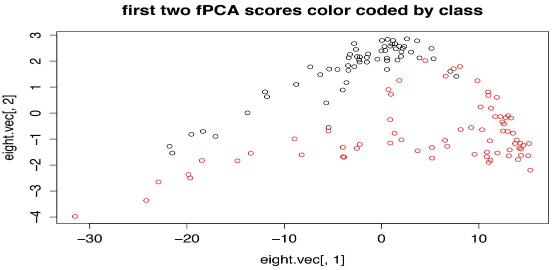
Classification analysis using FDA

Avoid defining 'features' of depth quantile functions $\rightsquigarrow$ FDA

For simplicity: binary classification problem, say classes 1 and 2.

- for X^* to be classified find $\widehat{q}_i^{(1)}(\delta)$ and $\widehat{q}_i^{(2)}(\delta)$
- perform fPCA for each of these functions, keeping first p scores
- $\rightsquigarrow$ $2p$-dimensional vector of scores
- already have n such $2p$-dimensional vectors from training data (two classes)
- find classification rule (SVM; kernel SVM: etc.)
- classify X^*

Illustration using wine data: Comparison to standard PCA



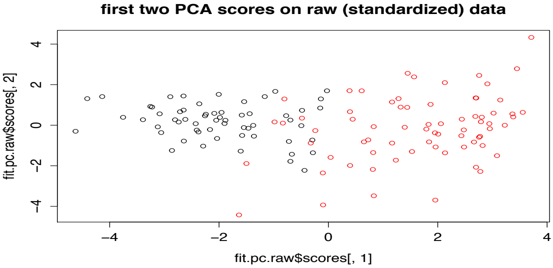


Illustration on wine data

Using leave one out procedure we obtain

misclassifications for wine data

classes	new method	1-NN
1,2	6	4
1,3	0	0
2,3	3	5

Illustration on PIMA data set

Illustration on PIMA data set

$n = 768$; $d = 8$ covariate measurements on female Pima Indians

classification in diabetes positive/negative

Illustration on PIMA data set

$n = 768$; $d = 8$ covariate measurements on female Pima Indians

classification in diabetes positive/negative

Our method is competitive with all the others tested in Dutta et al. (2015) ($\approx 25\%$ misclassification rate).

(LDA 23.37%, linear SVM 22.03%, radial SVM 24.19 %,kNN 25.73%, KDE 26.57 %, CART 27.20%, local depth based methods 25.18%)

Illustration on cancer data

gene expression data; $d = 2000, n = 62$, 2 classes (normal tissue 22, tumor tissue 40) Alon et al. 1999, PNAS

(http://genomics-pubs.princeton.edu/oncology/affydata/)

Illustration on cancer data

leave-one-out classification gives:

- opening angle $= 60$ degrees: misclassification rate 22%
- opening angle $= 85$ degrees: misclassification rate 27.4%

(LDA 35.48%, linear SVM 16.38%, radial SVM 35.48 %, kNN 22.58%, KDE 64.52 %, CART 28.77%, local depth based methods $\approx 20\%$)

Averaged feature functions and Choquet capacities

Averaged feature functions and Choquet capacities

For this, we need some more notation:

Versions with random ℓ and x

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$\widehat{d}_{ij}(s) = \min \left\{ F_n(A_{ij}(s)), F_n(B_{ij}(s)) \right\}$$

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}$$
$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}.$$

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}$$
$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}.$$

Then:

$$\widehat{q}_{ij}(\delta) = \inf\left\{t :\ G(s : \widehat{d}_{ij}(s) \leq t) \geq \delta\right\}$$

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$\widehat{d}_{ij}(s) = \min \left\{ F_n(A_{ij}(s)), F_n(B_{ij}(s)) \right\}$$
$$d_{ij}(s) = \min \left\{ F(A_{ij}(s)), F(B_{ij}(s)) \right\}.$$

Then:

$$\widehat{q}_{ij}(\delta) = \inf \left\{ t : \; G(s : \widehat{d}_{ij}(s) \leq t) \geq \delta \right\}$$
$$q_{ij}(\delta) = \inf \left\{ t : \; G(s : d_{ij}(s) \leq t) \geq \delta \right\}$$

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}$$
$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}.$$

Then:

$$\widehat{q}_{ij}(\delta) = \inf\left\{t : \; G(s : \widehat{d}_{ij}(s) \le t) \ge \delta\right\}$$
$$q_{ij}(\delta) = \inf\left\{t : \; G(s : d_{ij}(s) \le t) \ge \delta\right\}$$

$$\widehat{q}_i^{(k)}(\delta) = \mathop{\mathrm{ave}}_{X_j \in \text{ group } k} \widehat{q}_{ij}(\delta)$$
$$q_i^{(k)}(\delta) = \mathrm{E}(q_{ij}(\delta) | X_j \in \text{class } k; X_i)$$

Versions with random ℓ and x

- ℓ determined by pair (X_i, X_j)
- x is chosen as midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}$$
$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}.$$

Then:

$$\widehat{q}_{ij}(\delta) = \inf\left\{t : \ G(s : \widehat{d}_{ij}(s) \leq t) \geq \delta\right\}$$
$$q_{ij}(\delta) = \inf\left\{t : \ G(s : d_{ij}(s) \leq t) \geq \delta\right\}$$

$$\widehat{q}_i^{(k)}(\delta) = \operatorname*{ave}_{X_j \in \text{ group } k} \widehat{q}_{ij}(\delta)$$
$$q_i^{(k)}(\delta) = \mathrm{E}(q_{ij}(\delta) | X_j \in \text{class } k; X_i)$$

Note: $A_{ij}(s), B_{ij}(s)$, $d_{ij}(s), q_{ij}(\delta)$ and $q_i^{(k)}(\delta)$ are random quantities!

Averaged feature functions and Choquet capacities

Averaged feature functions and Choquet capacities

The quantity $q_i^{(k)}(\delta)$ can be expressed as:

$$q_i^{(k)}(\delta) = E_{X_j} P(Z \in \Gamma_{ij}(\delta) | X_i), \quad Z \sim F, \text{ independent of } X_i, X_j$$

where $\Gamma_{ij}(\delta)$ is a closed random set, whose distribution depends on the distributions of X_i and X_j.

Averaged feature functions and Choquet capacities

The quantity $q_i^{(k)}(\delta)$ can be expressed as:

$$q_i^{(k)}(\delta) = E_{X_j} P(Z \in \Gamma_{ij}(\delta) | X_i), \quad Z \sim F, \text{ independent of } X_i, X_j$$

where $\Gamma_{ij}(\delta)$ is a closed random set, whose distribution depends on the distributions of X_i and X_j. Let

$$\Psi_i^{(k)}(z) = P(z \in \Gamma_{ij}(\delta) | X_i), \quad X_j \in \text{group } k$$

be the hitting function of the random set $\Gamma_{ij}(\delta)$, given X_i.

Averaged feature functions and Choquet capacities

The quantity $q_i^{(k)}(\delta)$ can be expressed as:

$$q_i^{(k)}(\delta) = E_{X_j} P(Z \in \Gamma_{ij}(\delta) | X_i), \quad Z \sim F, \text{ independent of } X_i, X_j$$

where $\Gamma_{ij}(\delta)$ is a closed random set, whose distribution depends on the distributions of X_i and X_j. Let

$$\Psi_i^{(k)}(z) = P(z \in \Gamma_{ij}(\delta) | X_i), \quad X_j \in \text{group } k$$

be the hitting function of the random set $\Gamma_{ij}(\delta)$, given X_i.

Fubini $\rightsquigarrow$ $q_i^{(k)}(\delta) = E_F \Psi_i^{(k)}(Z)$ is the (conditional) expected value of the hitting function.

Averaged feature functions and Choquet capacities

The quantity $q_i^{(k)}(\delta)$ can be expressed as:

$$q_i^{(k)}(\delta) = E_{X_j} P(Z \in \Gamma_{ij}(\delta) | X_i), \quad Z \sim F, \text{ independent of } X_i, X_j$$

where $\Gamma_{ij}(\delta)$ is a closed random set, whose distribution depends on the distributions of X_i and X_j. Let

$$\Psi_i^{(k)}(z) = P(z \in \Gamma_{ij}(\delta) | X_i), \quad X_j \in \text{group } k$$

be the hitting function of the random set $\Gamma_{ij}(\delta)$, given X_i.

Fubini $\rightsquigarrow$ $q_i^{(k)}(\delta) = E_F \Psi_i^{(k)}(Z)$ is the (conditional) expected value of the hitting function.

Our method compares expected capacity functions of the random closed sets $\Gamma_{ij}(\delta)$ (given X_i) for different distributions of the sets, determined by the distributions of X_i and X_j. This is done for each δ.

Some asymptotics

Some asymptotics

Assumptions.

(A1) F and G posses continuous, bounded densities f and g, respectively.

(A2) For every $\epsilon > 0$ and every d there exists a set $\mathcal{R}_d(\epsilon) \subset \mathbb{R}^d$ of diameter $R_d(\epsilon)$, such that

- (i) $0 \in \mathcal{R}_d(\epsilon)$
- (ii) $R_d(\epsilon) = O(d^{1/2})$, as $d \to \infty$;
- (iii) $F(\mathcal{R}_d(\epsilon)) > 1 - \epsilon$ for every d;
- (iv) there exists a constant $c > 0$, not depending on d, such that $\sup_{x \in \mathcal{R}_d(\epsilon)} f(x) \leq c$.

(A3) $\sin \alpha = 1 - O(\frac{1}{d})$

Some asymptotics

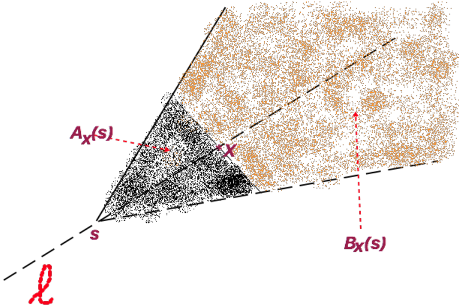
Proposition

Suppose that assumption (A1) holds, and that $X_1, \ldots, X_n, \ldots$ are iid from F. Then, for every given line $\ell \subset \mathbb{R}^d$, and every $\epsilon > 0$, there exist constants M and n_0, not varying with d, such that

$$P\left[\sup_{x,s\in\ell}\left|\sqrt{n}(\widehat{d}_{x,\ell}(s) - d_{x,\ell}(s))\right| > M\right] \leq \epsilon, \quad \textit{for } n \geq n_0.$$

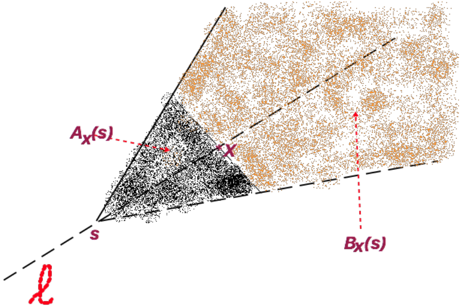
Notation

Recall:



Notation

Recall:



- Notation: $\mathcal{T}_{x,\ell}(s) = \underset{C \in \{A_x(s), B_x(s)\}}{\arg\min} \{F(C)\}.$

Some asymptotics

Theorem

Suppose that assumptions (A1) - (A3) hold. Let $S_{x,\ell}(c) = \{s \in \ell : |F(A_x(s)) - F(B_x(s))| \geq \frac{c}{\sqrt{n}}\}$*. With* $\mathcal{T}_{x,\ell}(s)$ *as above, let* $\mathbb{B}_x(s) = \min_{C \in \mathcal{T}_{x,\ell}(s)} B_F(C)$*. Then, for every* $\epsilon > 0$,

$$\lim_{c\to\infty} \limsup_{n\to\infty} P\Big[\sup_{x,s\in S_{x,\ell}(c)} \left|\sqrt{n}(\widehat{d}_{x,\ell}(s) - d_{x,\ell}(s)) - \mathbb{B}_x(s)\right| > \epsilon\Big] = 0.$$

Some asymptotics

Theorem

Suppose that assumptions (A1) - (A3) hold. Let $S_{x,\ell}(c) = \{s \in \ell : |F(A_x(s)) - F(B_x(s))| \geq \frac{c}{\sqrt{n}}\}$. *With* $\mathcal{T}_{x,\ell}(s)$ *as above, let* $\mathbb{B}_x(s) = \min_{C \in \mathcal{T}_{x,\ell}(s)} B_F(C)$. *Then, for every* $\epsilon > 0$,

$$\lim_{c\to\infty} \limsup_{n\to\infty} P\Big[\sup_{x,s \in S_{x,\ell}(c)} \left| \sqrt{n}(\widehat{d}_{x,\ell}(s) - d_{x,\ell}(s)) - \mathbb{B}_x(s) \right| > \epsilon \Big] = 0.$$

REMARK.

- This convergence is uniform in the dimension d.

Recall

Recall

- ℓ determined by pair (X_i, X_j)

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}$$

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}$$
$$\widehat{d_{ij}}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}.$$

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}$$
$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}.$$

Further:

$$q_{ij}(\delta) = \inf\left\{t : \; G(s : d_{ij}(s) \leq t) \geq \delta\right\}.$$

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}$$
$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}.$$

Further:

$$q_{ij}(\delta) = \inf\left\{t : \; G(s : d_{ij}(s) \leq t) \geq \delta\right\}.$$
$$\widehat{q}_{ij}(\delta) = \inf\left\{t : \; G(s : \widehat{d}_{ij}(s) \leq t) \geq \delta\right\}.$$

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$d_{ij}(s) = \min \left\{ F(A_{ij}(s)), F(B_{ij}(s)) \right\}$$
$$\widehat{d}_{ij}(s) = \min \left\{ F_n(A_{ij}(s)), F_n(B_{ij}(s)) \right\}.$$

Further:

$$q_{ij}(\delta) = \inf \left\{ t : \; G(s : d_{ij}(s) \leq t) \geq \delta \right\}.$$
$$\widehat{q}_{ij}(\delta) = \inf \left\{ t : \; G(s : \widehat{d}_{ij}(s) \leq t) \geq \delta \right\}.$$

$$q_i^{(k)}(\delta) = \mathrm{E}(q_{ij}(\delta) | X_j \in \text{class } k; X_i)$$

$$\widehat{q}_i^{(k)}(\delta) = \underset{X_j \in \text{ group } k}{\text{ave}} \widehat{q}_{ij}(\delta)$$

Recall

- ℓ determined by pair (X_i, X_j)
- x becomes midpoint $M_{ij} := \frac{X_i + X_j}{2}$

$$d_{ij}(s) = \min\left\{F(A_{ij}(s)), F(B_{ij}(s))\right\}$$
$$\widehat{d}_{ij}(s) = \min\left\{F_n(A_{ij}(s)), F_n(B_{ij}(s))\right\}.$$

Further:

$$q_{ij}(\delta) = \inf\left\{t:\ G(s : d_{ij}(s) \leq t) \geq \delta\right\}.$$
$$\widehat{q}_{ij}(\delta) = \inf\left\{t:\ G(s : \widehat{d}_{ij}(s) \leq t) \geq \delta\right\}.$$

$$q_i^{(k)}(\delta) = \mathrm{E}(q_{ij}(\delta) | X_j \in \mathrm{class}\ k; X_i)$$

$$\widehat{q}_i^{(k)}(\delta) = \underset{X_j \in \text{ group } k}{\mathrm{ave}} \widehat{q}_{ij}(\delta)$$

Note: $A_{ij}(s), B_{ij}(s),\ d_{ij}(s), q_{ij}(\delta)$ and $q_i^{(k)}(\delta)$ are random quantities!

Some asymptotics

Theorem

Suppose that assumption (A1) holds. With

$$D_{ij}(c) = \{\delta \in [0,1] : s_{ij}^r(\delta), s_{ij}^l(\delta) \in S_{x,\ell}(c)\},$$

we have

$$\sup_{1 \le i < j \le n} \sup_{\delta \in D_{ij}(c)} |\widehat{q}_{ij}(\delta) - q_{ij}(\delta)| = O_P\left(\sqrt{\frac{\min\{d, \log n\}}{n}}\right).$$

Some asymptotics

Theorem

Suppose that assumption (A1) holds. With

$$D_{ij}(c) = \{\delta \in [0,1] : s^r_{ij}(\delta), s^l_{ij}(\delta) \in S_{x,\ell}(c)\},$$

we have

$$\sup_{1\le i<j\le n} \sup_{\delta \in D_{ij}(c)} |\widehat{q}_{ij}(\delta) - q_{ij}(\delta)| = O_P\left(\sqrt{\frac{\min\{d, \log n\}}{n}}\right).$$

Remarks:

- Upper bound of $\left(\frac{\log n}{n}\right)^{1/2}$, independent of dimension d!

Some asymptotics

Theorem

Suppose that assumption (A1) holds. With

$$D_{ij}(c) = \{\delta \in [0,1] : s^r_{ij}(\delta), s^l_{ij}(\delta) \in S_{x,\ell}(c)\},$$

we have

$$\sup_{1\le i<j\le n} \sup_{\delta \in D_{ij}(c)} |\widehat{q}_{ij}(\delta) - q_{ij}(\delta)| = O_P\left(\sqrt{\frac{\min\{d, \log n\}}{n}}\right).$$

Remarks:

- Upper bound of $\left(\frac{\log n}{n}\right)^{1/2}$, independent of dimension d!
- If data lie in affine subspace of dimension $d^* \le d$, then d can be replaced by d^*.

Some asymptotics

Conjecture

Suppose that (A1) - (A3) hold, and assume that

$$P(X_k \notin D_{ij}(c) | X_k \text{ in class } k; X_i) = o_P(1/\sqrt{n}).$$

As $n_k \to \infty$ *(*n_k *number of obs. in class* k*), then*

$$\sup_{1 \le i \le n} \sup_{\delta \in [0,1]} \left| \widehat{q}_i^{(k)}(\delta) - q_i^{(k)}(\delta) \right| = O_P\left(\sqrt{\frac{\min\{d, \log n\}}{n}} \right).$$

Some asymptotics

Conjecture

Suppose that (A1) - (A3) hold, and assume that

$$P(X_k \notin D_{ij}(c) | X_k \text{ in class } k; X_i) = o_P(1/\sqrt{n}).$$

As $n_k \to \infty$ (n_k number of obs. in class k), then

$$\sup_{1 \leq i \leq n} \sup_{\delta \in [0,1]} \left| \widehat{q}_i^{(k)}(\delta) - q_i^{(k)}(\delta) \right| = O_P\left(\sqrt{\frac{\min\{d, \log n\}}{n}} \right).$$

Same remarks as above apply.

Open questions

- more on case of $d \to \infty$
- consider different Choquet functionals?
- investigate choice of tuning parameters
  - α (opening angle of cone)
  - G (distribution of cone tips)
- What if data lie on manifolds?
- For FDA classification: estimation of modes of variation (Petersen and Müller, 2016)
-

Illustration