DeBaCl: a Density-based Clustering Algorithm and its Properties

Alessandro Rinaldo
Department of Statistics
Carnegie Mellon University

joint work with Brian Kent and Fabrizio Lecci
Thanks to: Larry Wasserman, Bertrand Michel, Fred Chazal and Timothy Verstynen

November 10, 2014
Department of Statistics
Rice University

Outline

- Density-based clustering.
- The algorithm DeBaClandand some applications.
- Theoretical analysis of DeBaCl.

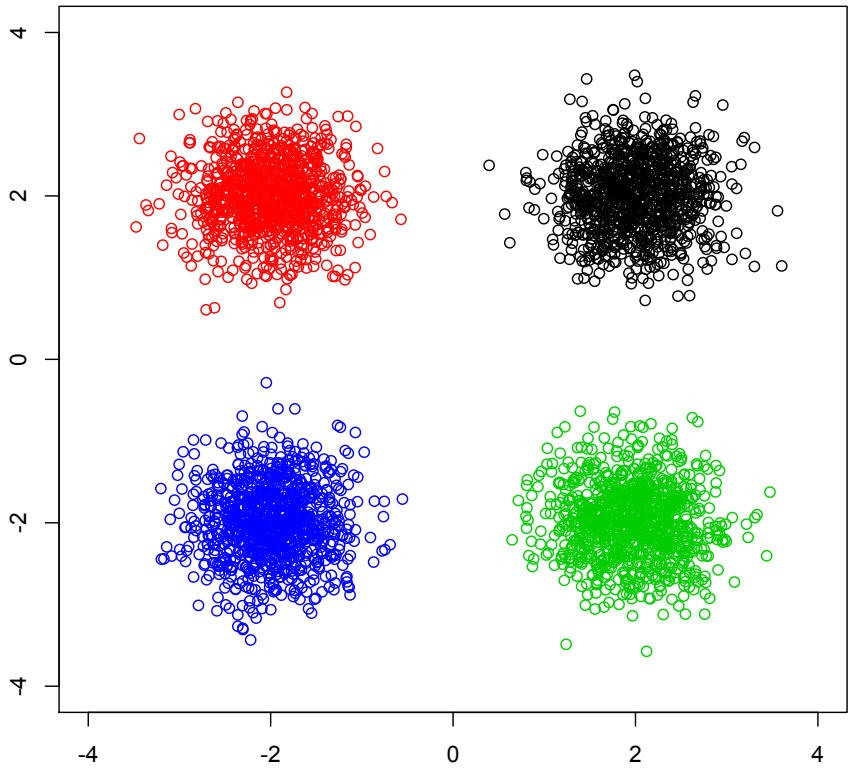
Clustering

- Classic problem in statistics, computer science, probability and many other fields. Huge literature!
- Abstract formulation: optimally organize a set of objects into groups, so that objects in the same group are maximally similar and objects in different groups are maximally dissimilar.
- Goal, scope and performance of a given clustering task is in many cases poorly or only partially defined.
- Analyses of clustering procedures often focus on the algorithmic properties, and tend to ignore the probabilistic nature of the input.

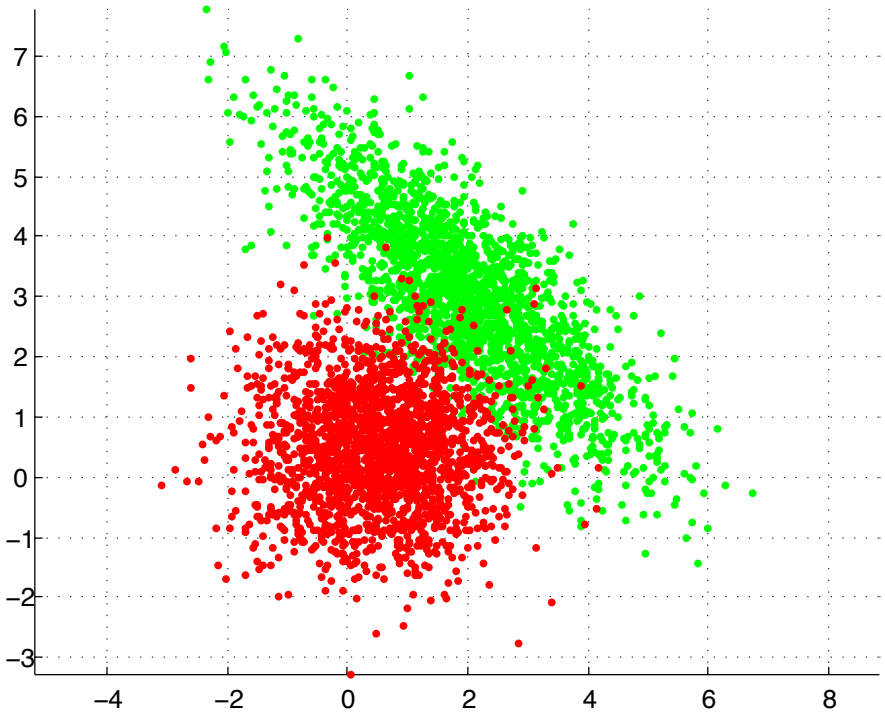
Clustering in Euclidean spaces

In much of this talk, we are interested in clustering $\mathbf{X}_n = (X_1, \ldots, X_n)$, an i.i.d. sample from a probability distribution P with support $S \subset \mathbb{R}^d$.

Clustering in Euclidean spaces

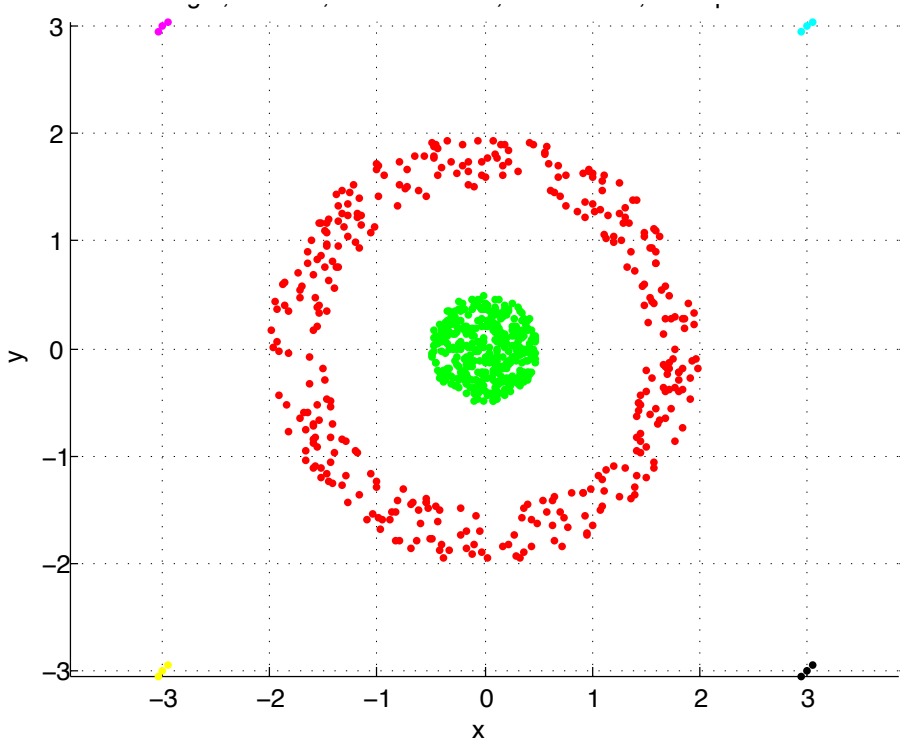


Clustering in Euclidean spaces



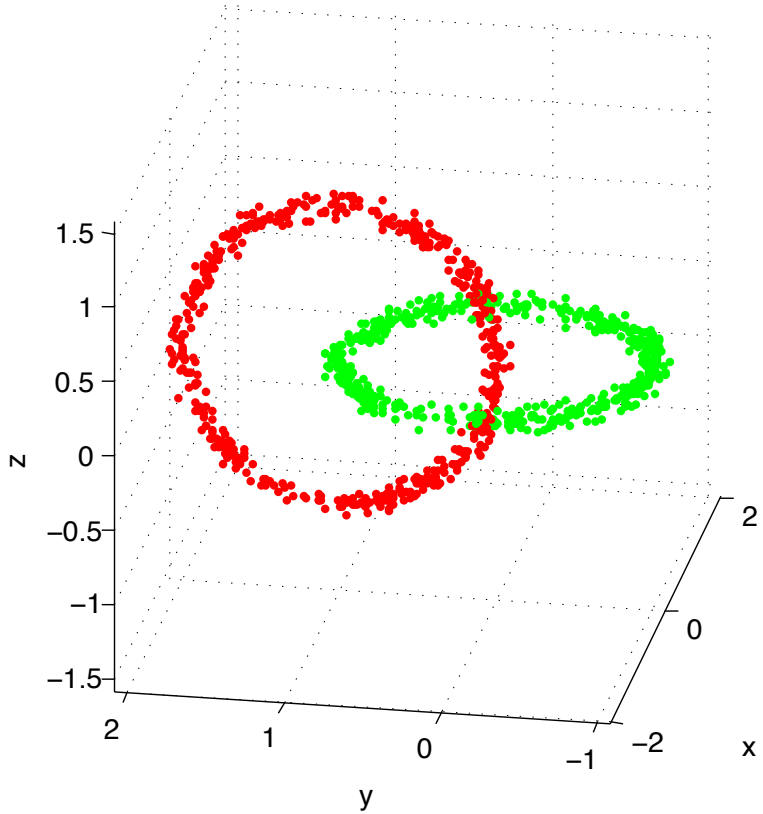
Source: Fundamental Clustering Problems Suite (FCPS).

Clustering in Euclidean spaces



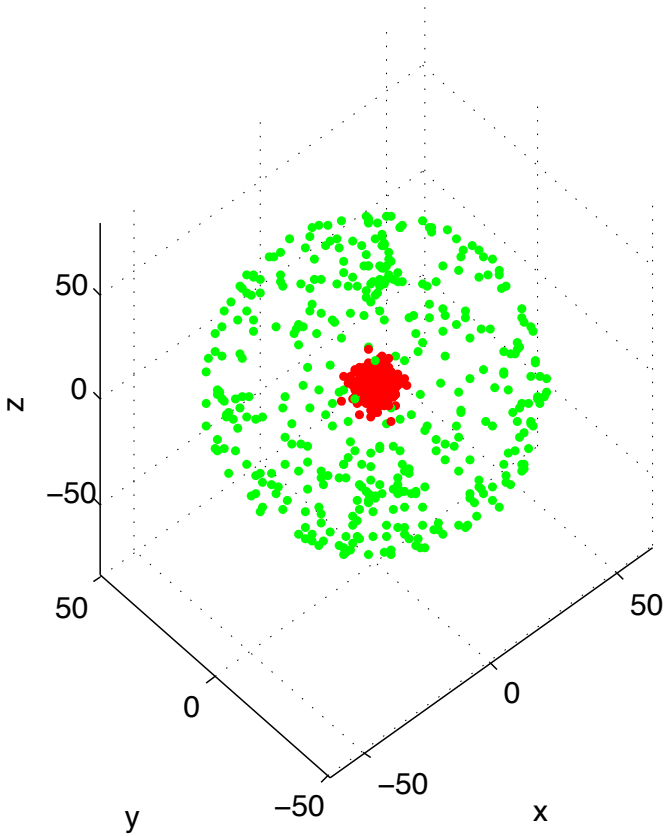
Source: Fundamental Clustering Problems Suite (FCPS).

Clustering in Euclidean spaces



Source: Fundamental Clustering Problems Suite (FCPS).

Clustering in Euclidean spaces



Source: Fundamental Clustering Problems Suite (FCPS).

Clustering in Euclidean spaces

We will try to be as agnostic as possible about P:

- P has a density with respect to k-dimensional Hausdorff measure or mixtures thereof, $k = \{1, \ldots, d\}$;
- the dimension $k = \dim(S)$ is unknown;
- the smoothness of f is unknown;
- the number of clusters is unknown;
- we are interested in both the algorithmic and statistical challenges of high dimensions.

We believe that many of our results extend to clustering of functional data.

Density-based clustering – Hartigan (1975, 1981)

- Assume P has a density f. For a threshold $\lambda > 0$, the λ-upper level set (high density region) of f is

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}.$$

Definition (λ-Clusters)

A λ-cluster of P is a maximal connected component of $L(\lambda)$.

- More interpretable twist (see Rinaldo et al., 2012).
  For $\alpha \in [0, 1]$, set $\lambda_\alpha = \sup\{\lambda : P(L(\lambda)) \geq \alpha\}$ and $L(\alpha) = L(\lambda_\alpha)$.

Definition (α-Clusters)

A α-cluster of P is maximal connected component of $L(\alpha)$. Minimal volume set of prescribed probability content.

Density-based clustering – Hartigan (1975, 1981)

- Assume P has a density f. For a threshold $\lambda > 0$, the λ-upper level set (high density region) of f is

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}.$$

Definition (λ-Clusters)

A λ-cluster of P is a maximal connected component of $L(\lambda)$.

- More interpretable twist (see Rinaldo et al., 2012).
  For $\alpha \in [0, 1]$, set $\lambda_\alpha = \sup\{\lambda : P(L(\lambda)) \geq \alpha\}$ and $L(\alpha) = L(\lambda_\alpha)$.

Definition (α-Clusters)

A α-cluster of P is maximal connected component of $L(\alpha)$. Minimal volume set of prescribed probability content.

Density-based clustering – Hartigan (1975, 1981)

- Consider all thresholds simultaneously!
- The family of all λ-clusters of P is called the cluster tree of P because it has the tree property: $A, B \in \mathcal{T}$ implies that

$$A \subset B \text{ or } B \subset A \qquad \text{or} \qquad A \cap B = \emptyset.$$

  The hierarchy of inclusions of $\mathcal{T}$ can be represented as a dendrogram, with height indexed by λ or α.

- Many subtle topological and measure-theoretical details: see Steinwart (2014).

Density-based clustering – Hartigan (1975, 1981)

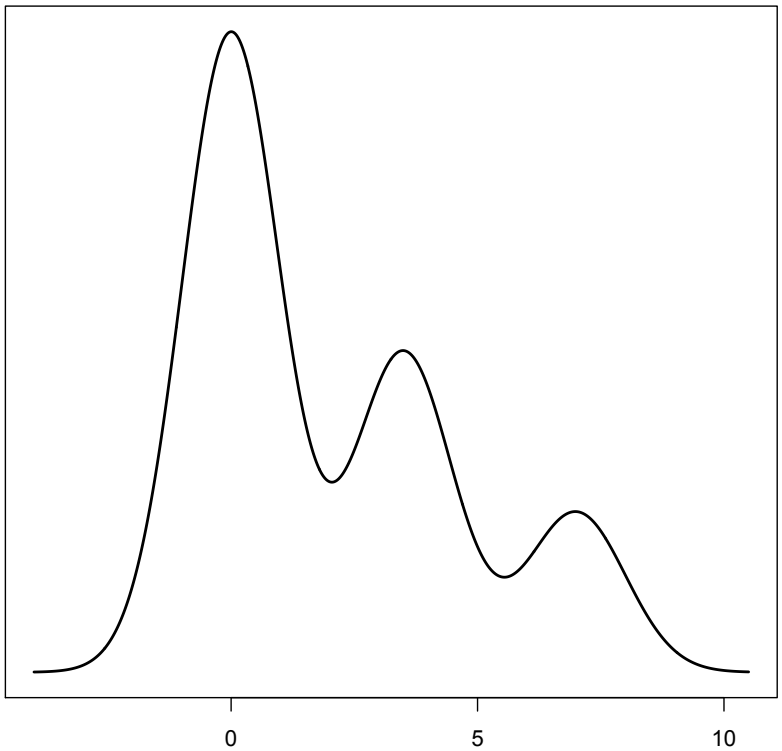
- Consider all thresholds simultaneously!
- The family of all λ-clusters of P is called the cluster tree of P because it has the tree property: $A, B \in \mathcal{T}$ implies that

$$A \subset B \text{ or } B \subset A \qquad \text{or} \qquad A \cap B = \emptyset.$$

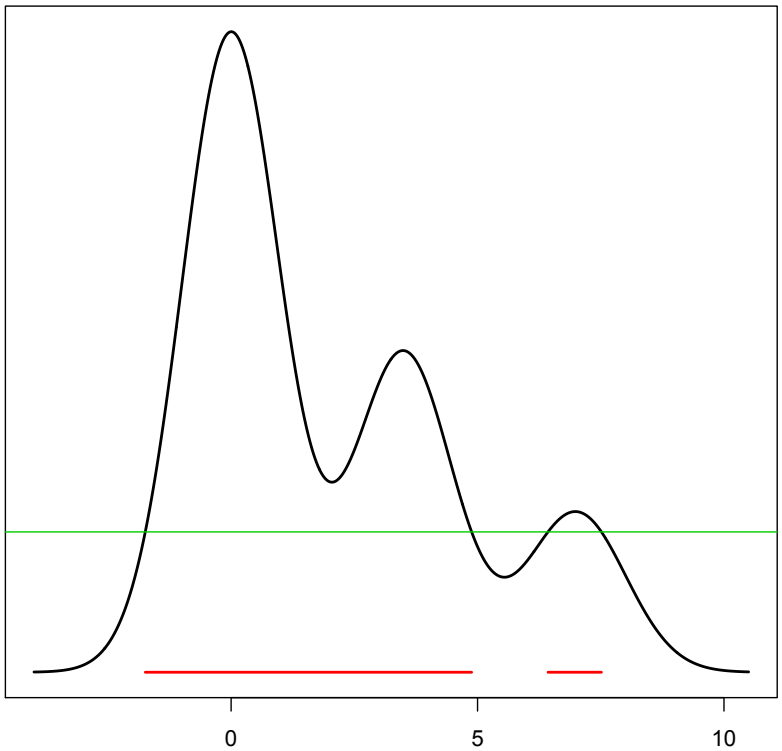
  The hierarchy of inclusions of $\mathcal{T}$ can be represented as a dendrogram, with height indexed by λ or α.

- Many subtle topological and measure-theoretical details: see Steinwart (2014).

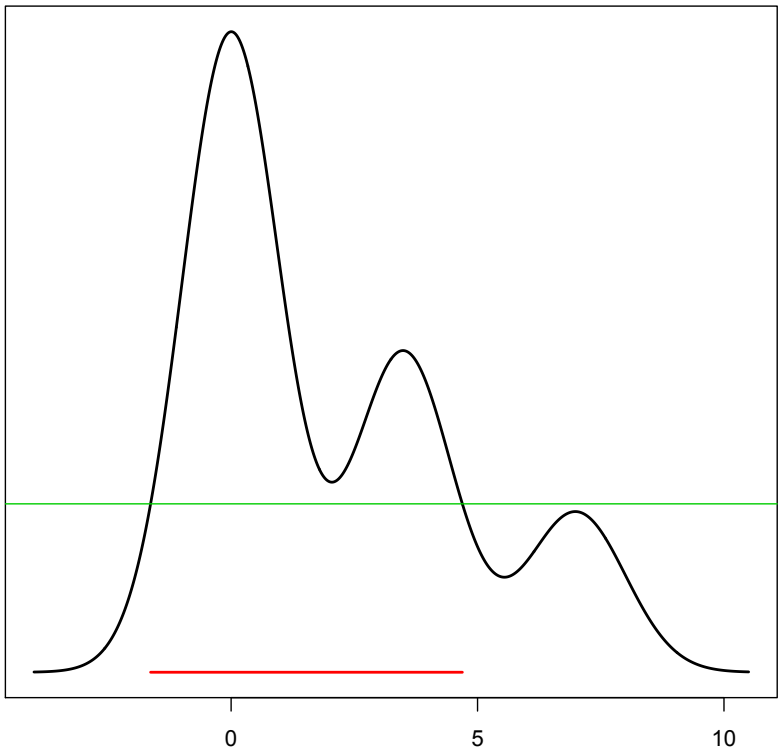
Density-based clustering in action



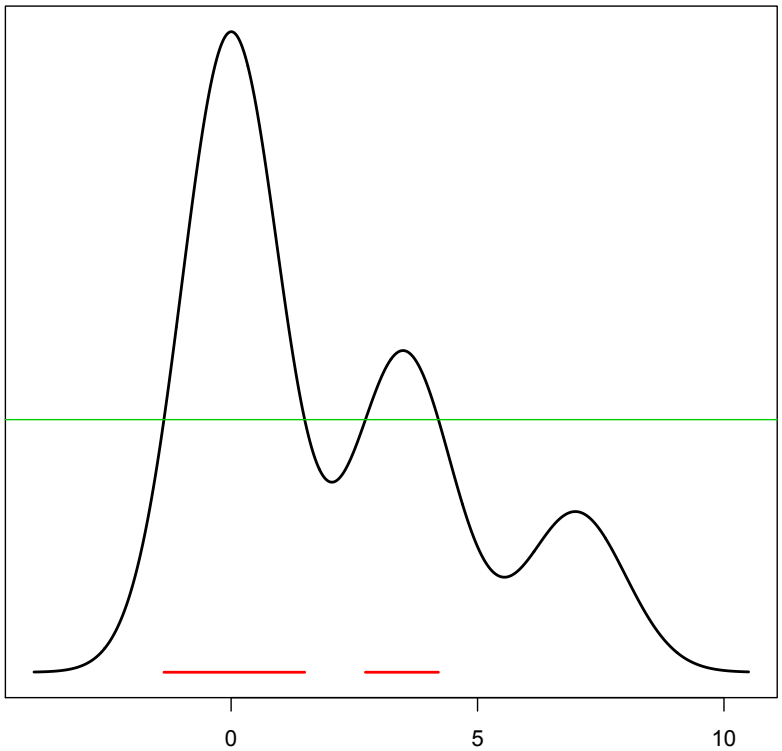
Density-based clustering in action



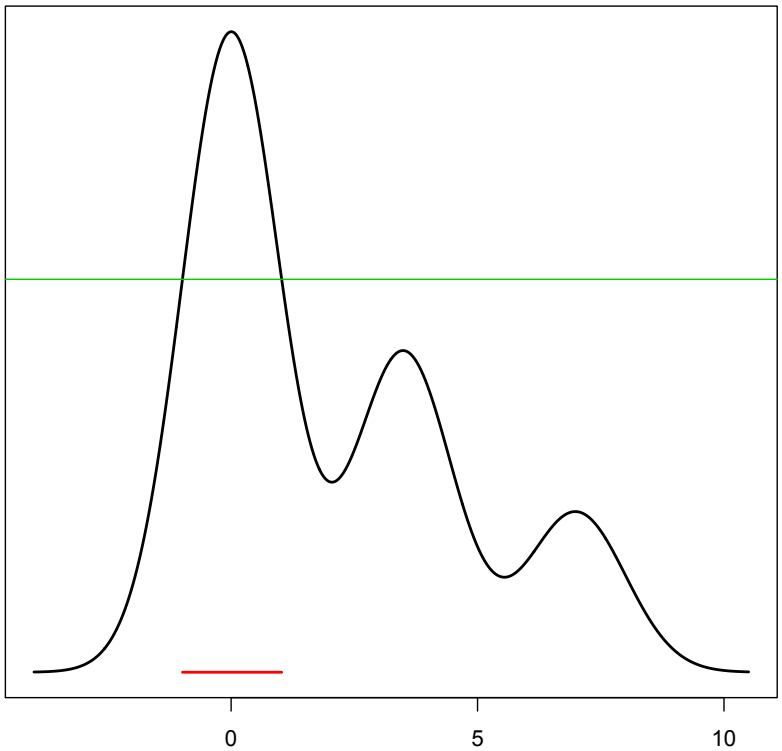
Density-based clustering in action



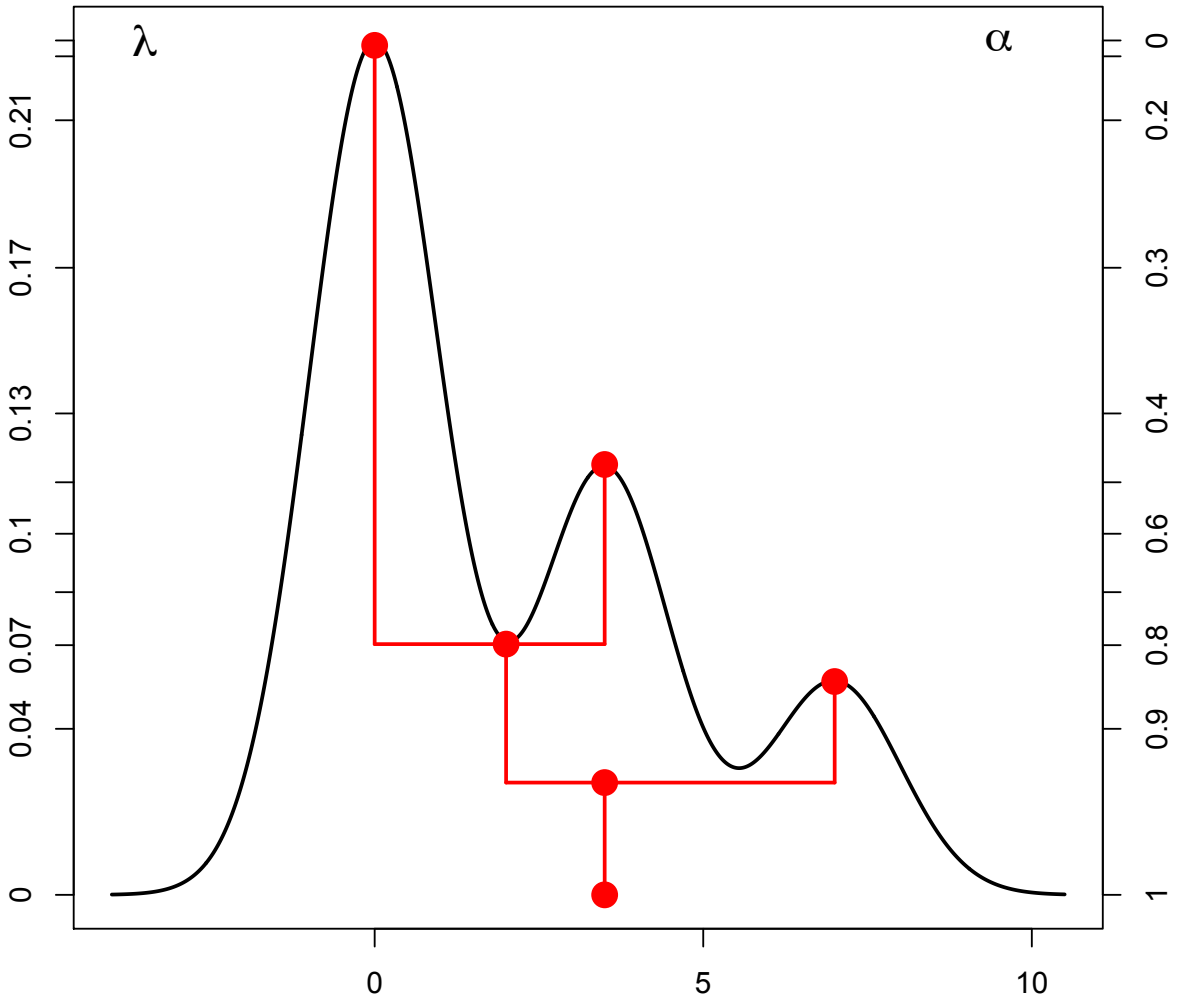
Density-based clustering in action



Density-based clustering in action

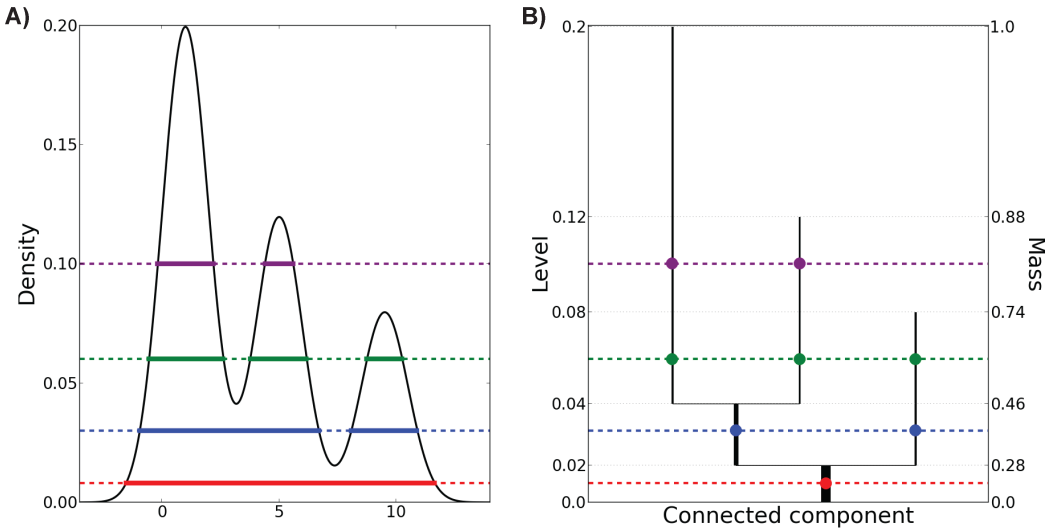


Density-based clustering in action



Cluster Trees

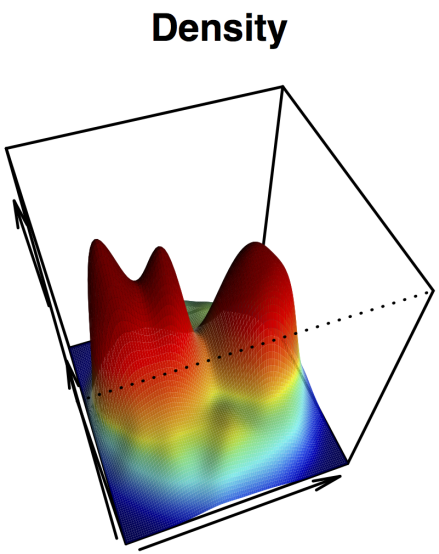
- The cluster tree $\mathcal{T}$ captures all the clustering properties of P simultaneously.
- The cluster tree is an algebraic structure for visualizing and encoding P. It is largely decoupled from the geometry and dimension of P.
- There is no need to choose the number of clusters.

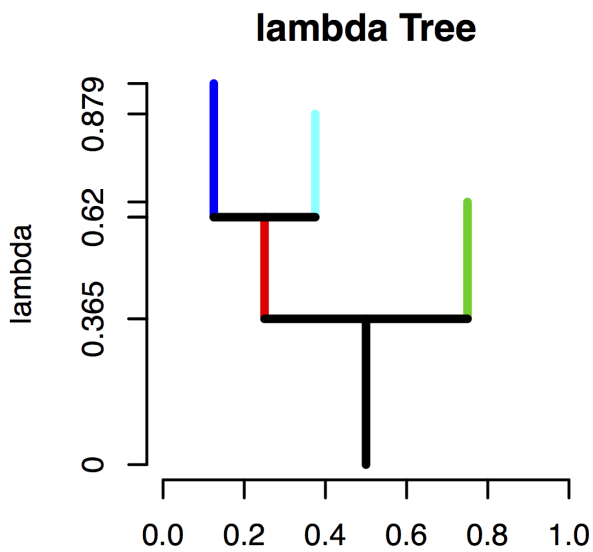


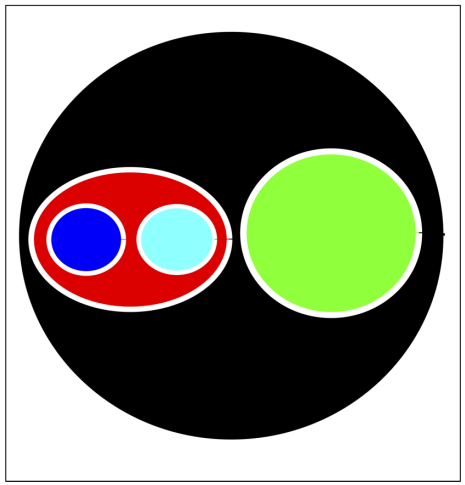
Trees, branches and leaves

Partition Property

The leaves and branches of the tree partition $S = \text{supp}(P)$.







Issue I: Density-based clustering is statistically hard

To “estimate $\mathcal{T}$ consistently" using a density estimator $\widehat{f}$, we need sup norm consistency, i.e. $\sup_x |f(x) - \widehat{f}(x)| = o_P(1)$.

- The minimax rate (attained by KDEs with vanishing bandwidth) for this problem over Hölder classes of densities is

$$\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}},$$

  where β is the smoothness parameter.
- This typically requires a sample size exponential in d. Consistent estimation of $\mathcal{T}$ is unfeasible in high-dimensions.

Issue I: Density-based clustering is statistically hard

To "estimate $\mathcal{T}$ consistently" using a density estimator $\widehat{f}$, we need sup norm consistency, i.e. $\sup_x |f(x) - \widehat{f}(x)| = o_P(1)$.

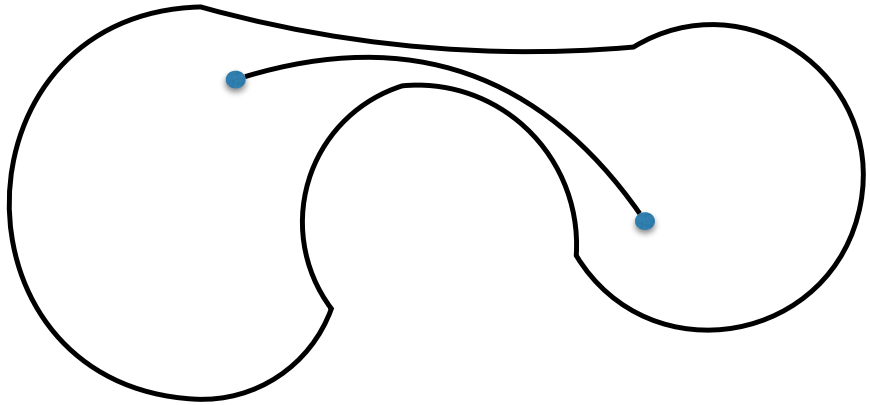
- The minimax rate (attained by KDEs with vanishing bandwidth) for this problem over Hölder classes of densities is

$$\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}},$$

  where β is the smoothness parameter.
- This typically requires a sample size exponential in d. Consistent estimation of $\mathcal{T}$ is unfeasible in high-dimensions.

Issue II: density-based clustering is algorithmically hard

- Even assuming f known, deciding whether x and y are in the same λ-cluster of f requires finding a path $\ell \subset S$ between x and y such that $f(z) \geq \lambda$ for all $z \in \ell$.



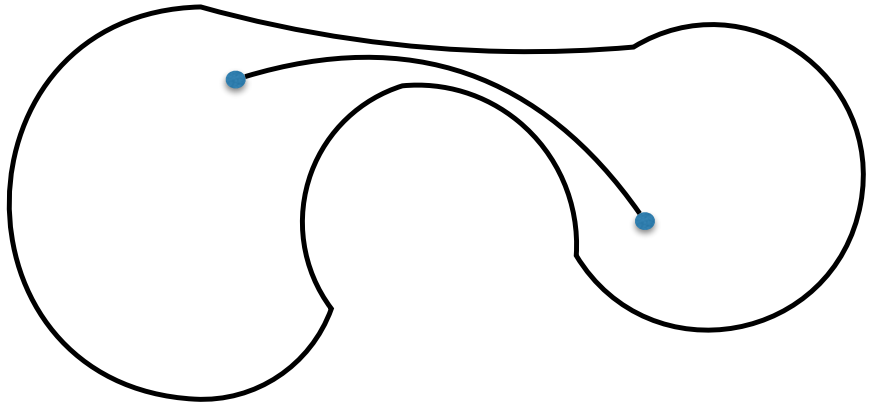
- This computation is prohibitively difficult even in moderate dimensions. Building $\mathcal{T}$ is unfeasible in high-dimensions.

Fix

Use (slightly) incorrect connectivity based on $\mathbf{X}_n$.

Issue II: density-based clustering is algorithmically hard

- Even assuming f known, deciding whether x and y are in the same λ-cluster of f requires finding a path $\ell \subset S$ between x and y such that $f(z) \geq \lambda$ for all $z \in \ell$.



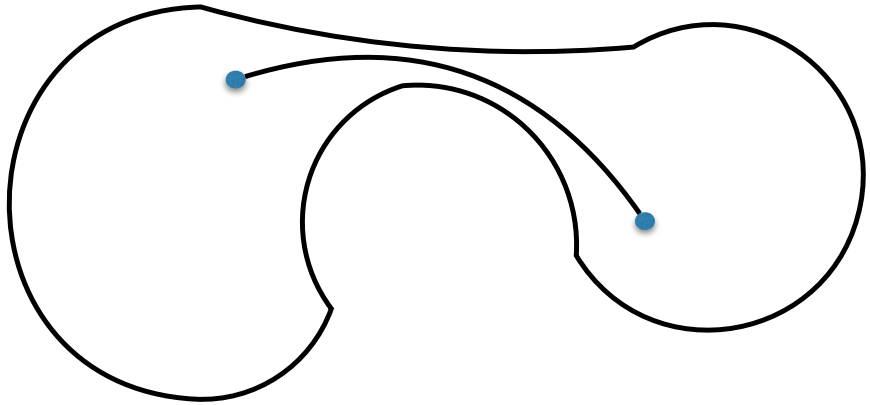
- This computation is prohibitively difficult even in moderate dimensions. Building $\mathcal{T}$ is unfeasible in high-dimensions.

Fix

Use (slightly) incorrect connectivity based on $\mathbf{X}_n$.

Issue II: density-based clustering is algorithmically hard

- Even assuming f known, deciding whether x and y are in the same λ-cluster of f requires finding a path $\ell \subset S$ between x and y such that $f(z) \geq \lambda$ for all $z \in \ell$.



- This computation is prohibitively difficult even in moderate dimensions. Building $\mathcal{T}$ is unfeasible in high-dimensions.

Fix

Use (slightly) incorrect connectivity based on $\mathbf{X}_n$.

How to deal with curse of dimensionality in density-based clustering

So you are all about the bias...big trouble!

Statistical hardness is an unavoidable bias issue. So let's ignore it!

- Suppose P has Lebesgue density f, assumed Hölder smooth with parameter β. Let $\widehat{f}_h$ be a KDE with bandwitdh h

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\Big(\frac{\|x - X_i\|}{h}\Big), \quad x \in \mathbb{R}^d.$$

  For eaxh $h > 0$, $\widehat{f}_h$ is an unbiased estimator of the density

$$f_h(x) = \frac{1}{h^d}\int_{\mathbb{R}^d} f(y) K\left(\frac{\|y - x\|}{h}\right) dx, \quad x \in \mathbb{R}^d.$$

- f_h is much easier to estimate than f!!

How to deal with curse of dimensionality in density-based clustering

So you are all about the bias...big trouble!

Statistical hardness is an unavoidable bias issue. So let's ignore it!

- Suppose P has Lebesgue density f, assumed Hölder smooth with parameter β. Let $\widehat{f}_h$ be a KDE with bandwitdh h

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{\|x - X_i\|}{h}\right), \quad x \in \mathbb{R}^d.$$

  For eaxh $h > 0$, $\widehat{f}_h$ is an unbiased estimator of the density

$$f_h(x) = \frac{1}{h^d}\int_{\mathbb{R}^d} f(y) K\left(\frac{\|y - x\|}{h}\right) dx, \quad x \in \mathbb{R}^d.$$

- f_h is much easier to estimate than f!!

How to deal with course of dimensionality

- By the Hölder assumption and Gine' and Guillon (2002):

$$\sup_x |f(x) - \widehat{f}_h(x)| \leq \underbrace{\sup_x |f(x) - f_h(x)|}_{\textit{bias}} + \underbrace{\sup_x |f_h(x) - \widehat{f}_h(x)|}_{\text{random fluctuations}}$$

$$= O(h^\beta) + O_P\left(\sqrt{\frac{1}{nh^d}}\right)$$

Ignoring the bias and for fixed h f_h can be well estimated with the nearly parameric, dimension independent rate:

$$\sup_x |f_h(x) - \widehat{f}_h(x)| = O\left(\sqrt{\frac{\log n}{n}}\right),$$

with high probability. The dimension is in the constants!

How to deal with course of dimensionality

- By the Hölder assumption and Gine' and Guillon (2002):

$$\sup_x |f(x) - \widehat{f}_h(x)| \leq \underbrace{\sup_x |f(x) - f_h(x)|}_{\textit{bias}} + \underbrace{\sup_x |f_h(x) - \widehat{f}_h(x)|}_{\text{random fluctuations}}$$

$$= O(h^\beta) + O_P\left(\sqrt{\frac{1}{nh^d}}\right)$$

Ignoring the bias and for fixed h f_h can be well estimated with the nearly parameric, dimension independent rate:

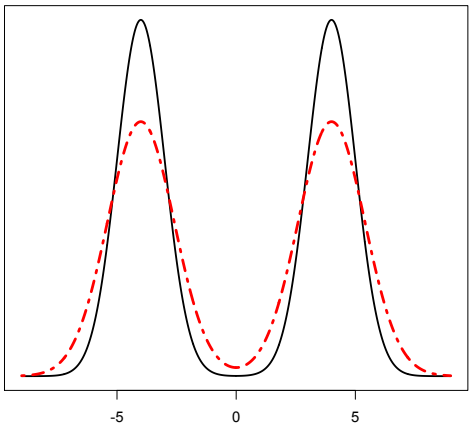
$$\sup_x |f_h(x) - \widehat{f}_h(x)| = O\left(\sqrt{\frac{\log n}{n}}\right),$$

with high probability. The dimension is in the constants!

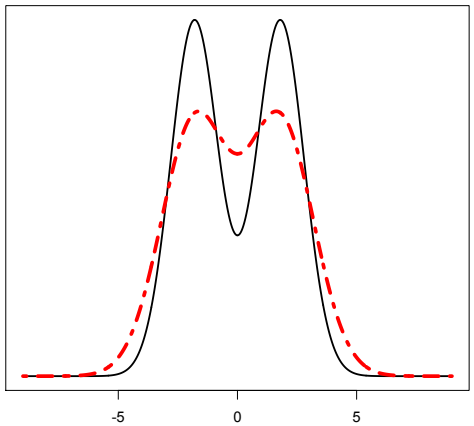
More on ignoring the bias

- One may measure the difficulty of a clustering problem depending on whether density clustering based on biased density estimation can be successful.

Easy



Hard

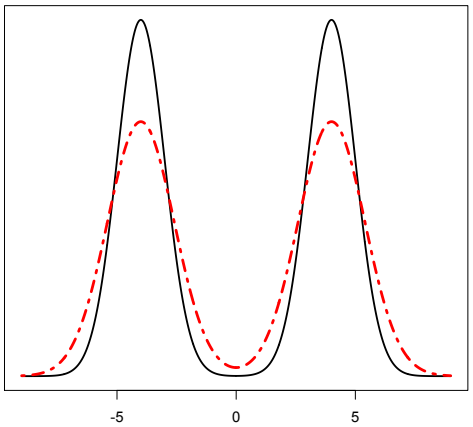


- Another major advantage of allowing for bias is that it extends the applicability of density-based clustering to singular P.
See, e.g., Rinaldo and Wasserman (2010).

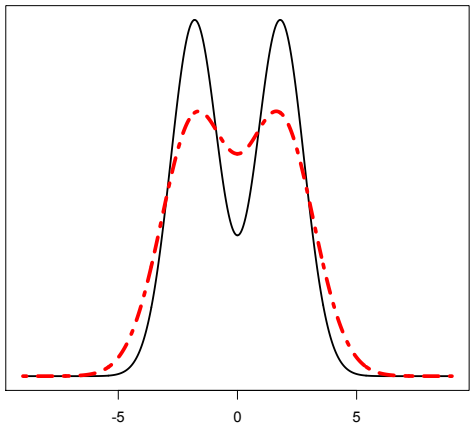
More on ignoring the bias

- One may measure the difficulty of a clustering problem depending on whether density clustering based on biased density estimation can be successful.

Easy



Hard



- Another major advantage of allowing for bias is that it extends the applicability of density-based clustering to singular P.
See, e.g., Rinaldo and Wasserman (2010).

- Very large amount of literature.
  - Cluster tree estimation: Koltchinksii (2000), Stuetzle and Nugent (2010). More recently: Chaudhuri, Dasgupta, Kptufe and von Luxburg (2013), Balakrishnan et al. (2013) and Steinwary (2014).
  - Support estimation: Korostelev and Tsybakov (1993), Mammen and Tsybakov (1995), Cuevas and Fraiman (1997), Biau, Cadre and Pellettier (2008).
  - Level set estimation for fixed λ: Polonik (1995), Tsybakov (1997), Walther (1997), Scott and Nowak (2006), Cuevas, González-Menteiga and Rodríguez-Casal (2006), Singh, Scott and Nowak (2009), Rigollet and Vert (2010).
- Some algorithms: `DBSCAN`, `OPTICS`, `denpro`.

Meet DeBaCl

- DeBaCl is a simple algorithm for density-based clustering.
  We credit Kpotufe and von Luxburg (2011).
- It is based on the k-nn density estimator.

- Implementations:
  - pyton module DeBaCl by Brian Kent (update coming soon)
    https://github.com/CoAxLab/DeBaCl
  - R package TDA by Fabrizio Lecci et al.
    http://cran.r-project.org/web/packages/TDA/index.html

Meet DeBaCl

- DeBaCl is a simple algorithm for density-based clustering.
  We credit Kpotufe and von Luxburg (2011).
- It is based on the k-nn density estimator.

- Implementations:
  - pyton module DeBaCl by Brian Kent (update coming soon)
    https://github.com/CoAxLab/DeBaCl
  - R package TDA by Fabrizio Lecci et al.
    http://cran.r-project.org/web/packages/TDA/index.html

Meet DeBaCl

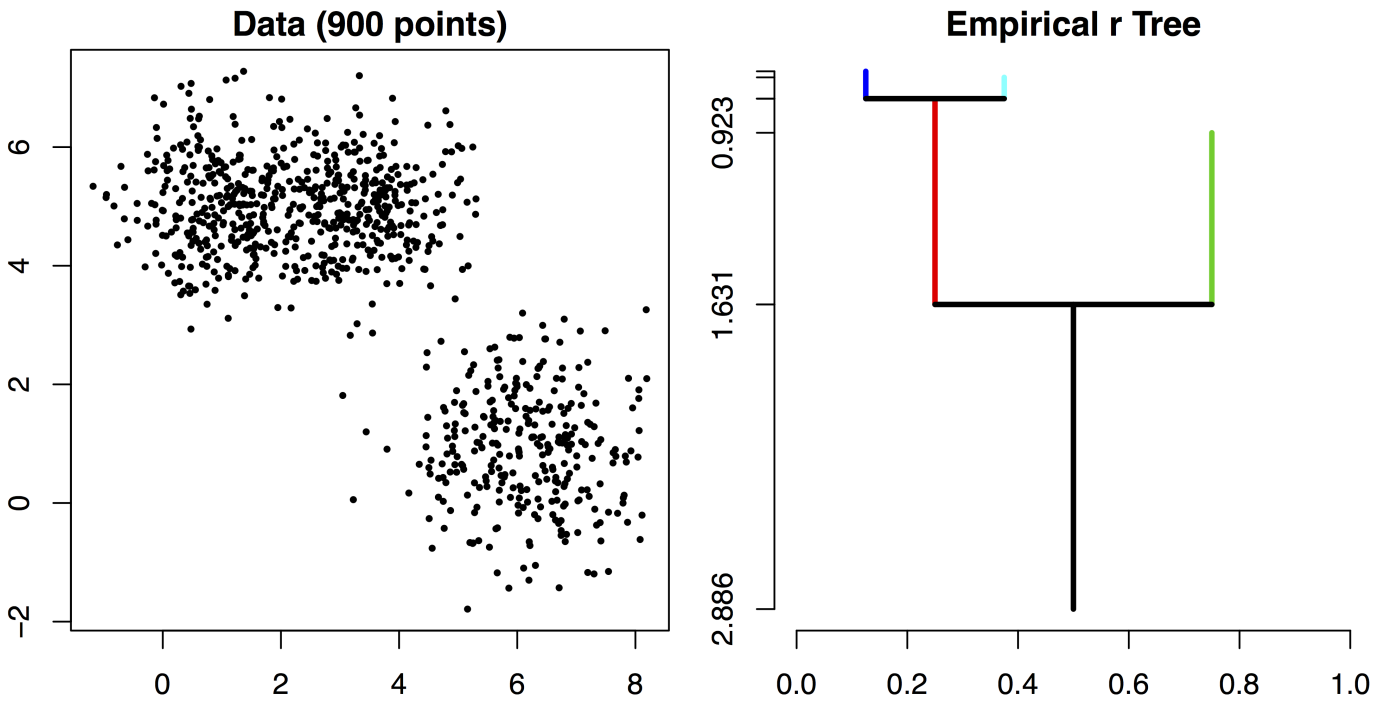
For fixed $p \in (0,1)$ and each $i = 1, \ldots, n$, set $\widehat{r}_i$ be the distance of X_i from its k-th nearest neighbors in $\mathbf{X}_n$, with $k = \lceil np \rceil$.

Input: $p \in (0,1)$ and $\mathbf{X_n}$

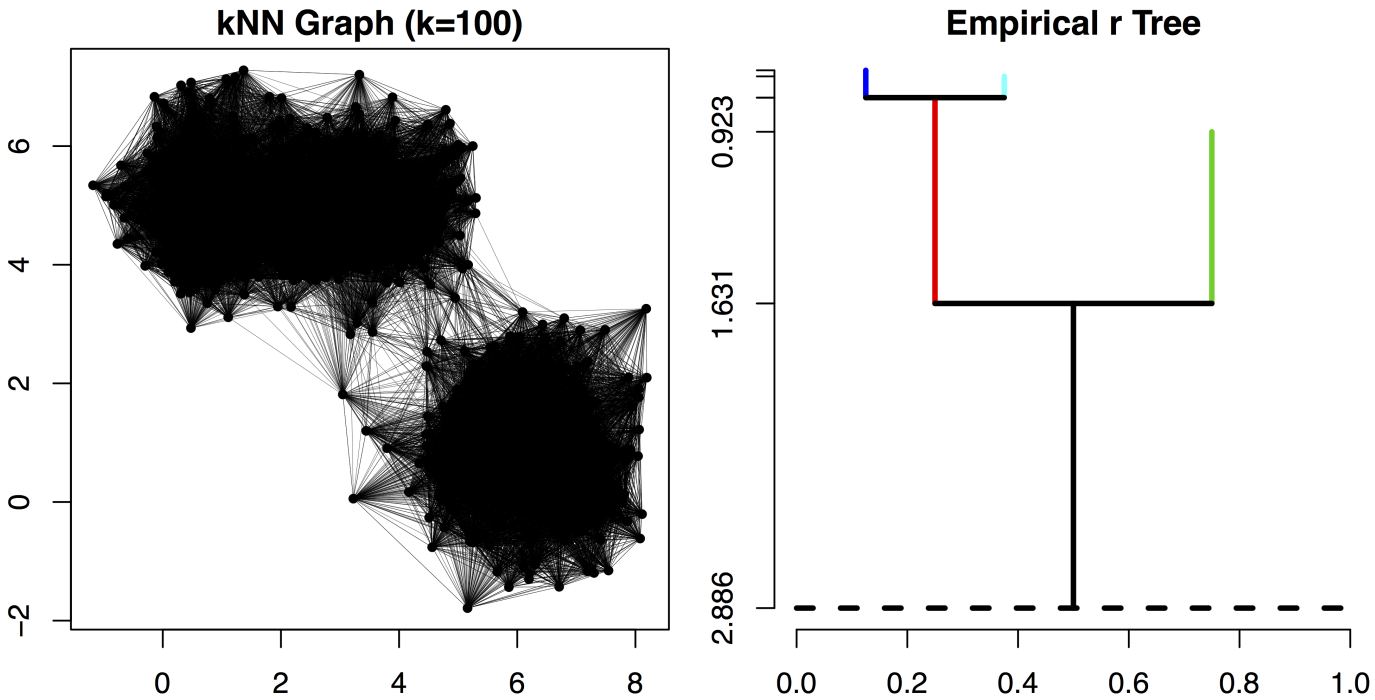
1. Construct the *knn* graph $\widehat{\mathcal{G}}_n$ with nodes $\mathbf{X}_n$ and edges (X_i, X_j)
   - (i) if $\|X_i - X_j\| \leq \max\{\widehat{r}_i, \widehat{r}_j\}$ (k-nn)
   - (ii) if $\|X_i - X_j\| \leq \max\{\widehat{r}_i, \widehat{r}_j\}$ (mutual k-nn)
2. For all $r \in R := [\min_i \widehat{r}_i, \max_i \widehat{r}_i]$
   - (i) set $\widehat{\mathcal{G}}_n(r)$ be subgraph induced by $\{X_i : \widehat{r}_i \leq r\}$.
   - (ii) compute the connected components of $\widehat{\mathcal{G}}_n(r)$.

Output $\{\widehat{\mathcal{T}}_n(r), r \in R\}$, the dendrogram of the connected components.

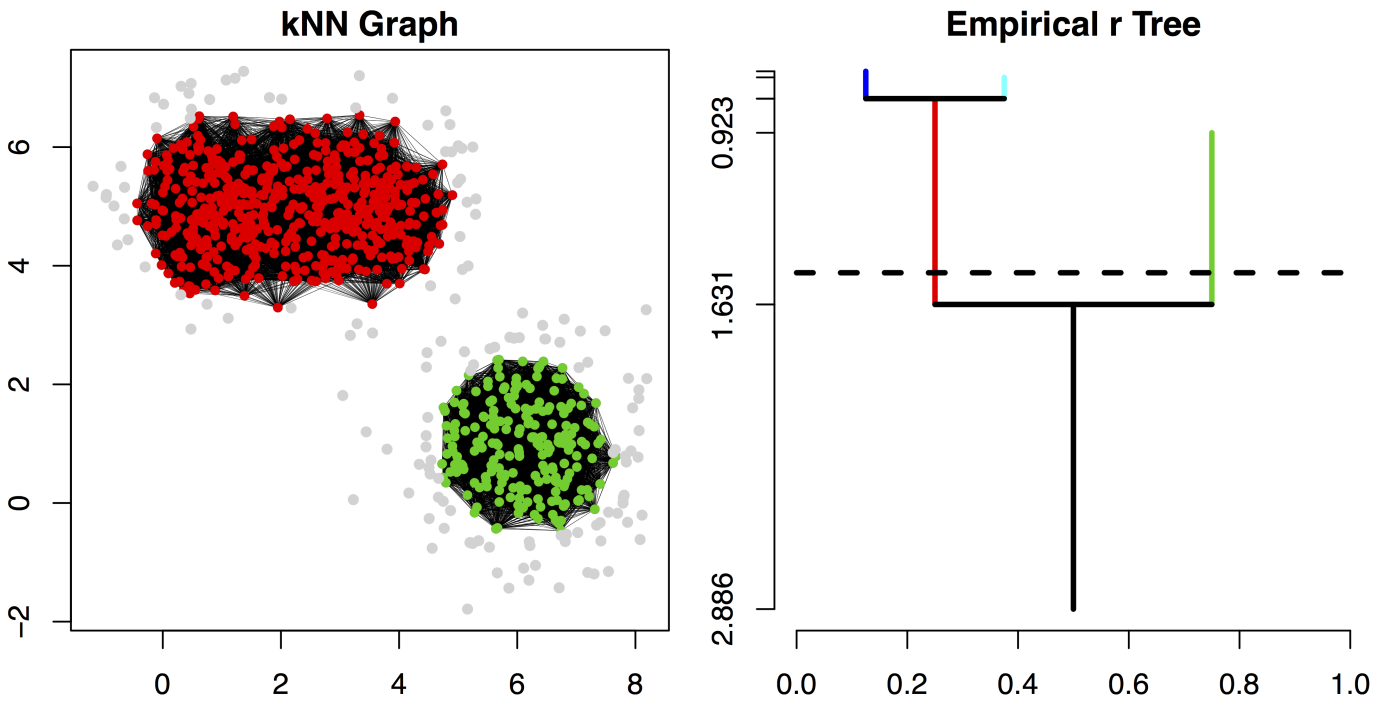
Meet DeBaCl


Data (900 points)
Empirical r Tree
0.923
1.631
2.886
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
0
2
4
6
8

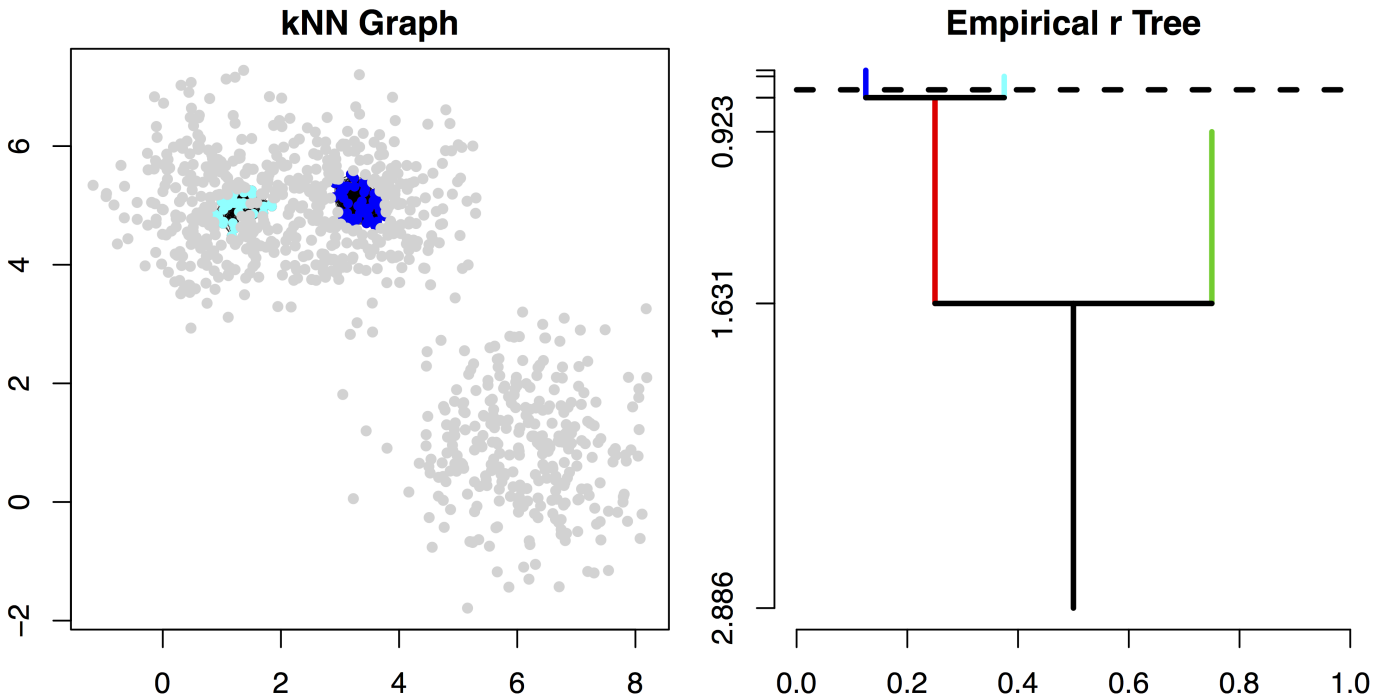
Meet DeBaCl


kNN Graph (k=100)
Empirical r Tree
0.923
1.631
2.886
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
0
2
4
6
8

Meet DeBaCl


kNN Graph
Empirical r Tree
0.923
1.631
2.886
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
0
2
4
6
8

Meet DeBaCl


kNN Graph
Empirical r Tree
0.923
1.631
2.886
0.0
0.2
0.4
0.6
0.8
1.0
−2
0
2
4
6
0
2
4
6
8

Remarks on DeBaCl

- The input to DeBaCl are the k-nn distances $\widehat{r}_1, \ldots, \widehat{r}_n$ that are used both to compute level sets and to determine connectedness.
- Computational complexity. The computation of all the k-nn's has complexity $\mathcal{O}(n \log n)$ (using k-d trees, ball-trees and cover-trees). The complexity of constructing all the connected components is nearly linear in n because it relies on a modified union-find procedure (Najman and Couprie, 2006) and never uses breadth-first search.
- DeBaCl outputs a data structure.

- Why k-nn and not KDE? A KDE version of DeBaCl is easy enough to devise but we prefer k-nn...

Remarks on DeBaCl

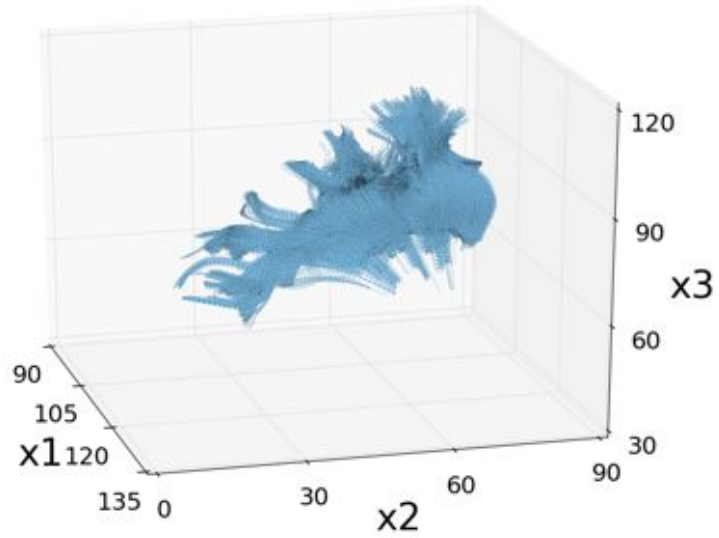
- The input to DeBaCl are the k-nn distances $\widehat{r}_1, \ldots, \widehat{r}_n$ that are used both to compute level sets and to determine connectedness.
- Computational complexity. The computation of all the k-nn's has complexity $\mathcal{O}(n \log n)$ (using k-d trees, ball-trees and cover-trees). The complexity of constructing all the connected components is nearly linear in n because it relies on a modified union-find procedure (Najman and Couprie, 2006) and never uses breadth-first search.
- DeBaCl outputs a data structure.

- Why k-nn and not KDE? A KDE version of DeBaCl is easy enough to devise but we prefer k-nn...

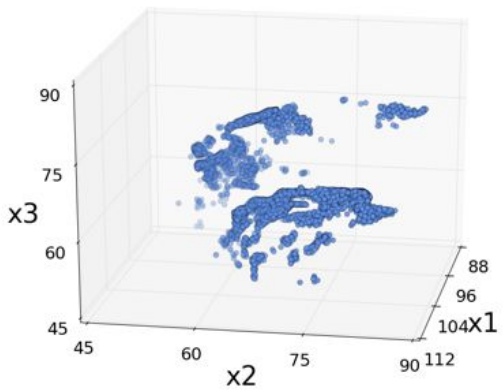
Example 1: clustering endpoints of fiber tracks in the striatum

- The fiber endpoint data is derived from in vivo difusion weighted brain imaging (DWI) collected at the Scientific Imaging and Brain Research Center at Carnegie Mellon University in 2012 for 30 neurologically healthy controls (the CMU-30 group).
- From the DWI data, deterministic fiber tractography was used to simulate smooth 1-dimensional manifolds (with boundaries) called fiber streamlines that represent tracks of strong water diffusion in the brain (Hagmann et al., 2006).
- $10,000$ fiber streamlines were mapped from the cortex into the striatum for a single subject. Only the teminal points of the streamlines were kept.
- $k = 200$
- Work done in collaboration with Timothy Verstynen:
  `http://www.psy.cmu.edu/~coaxlab/`

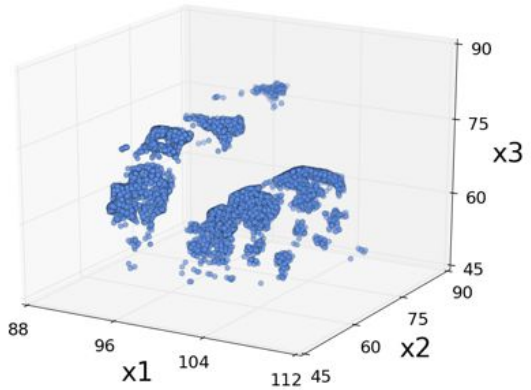
Example 1: clustering endpoints of fiber tracks in the striatum



(a)

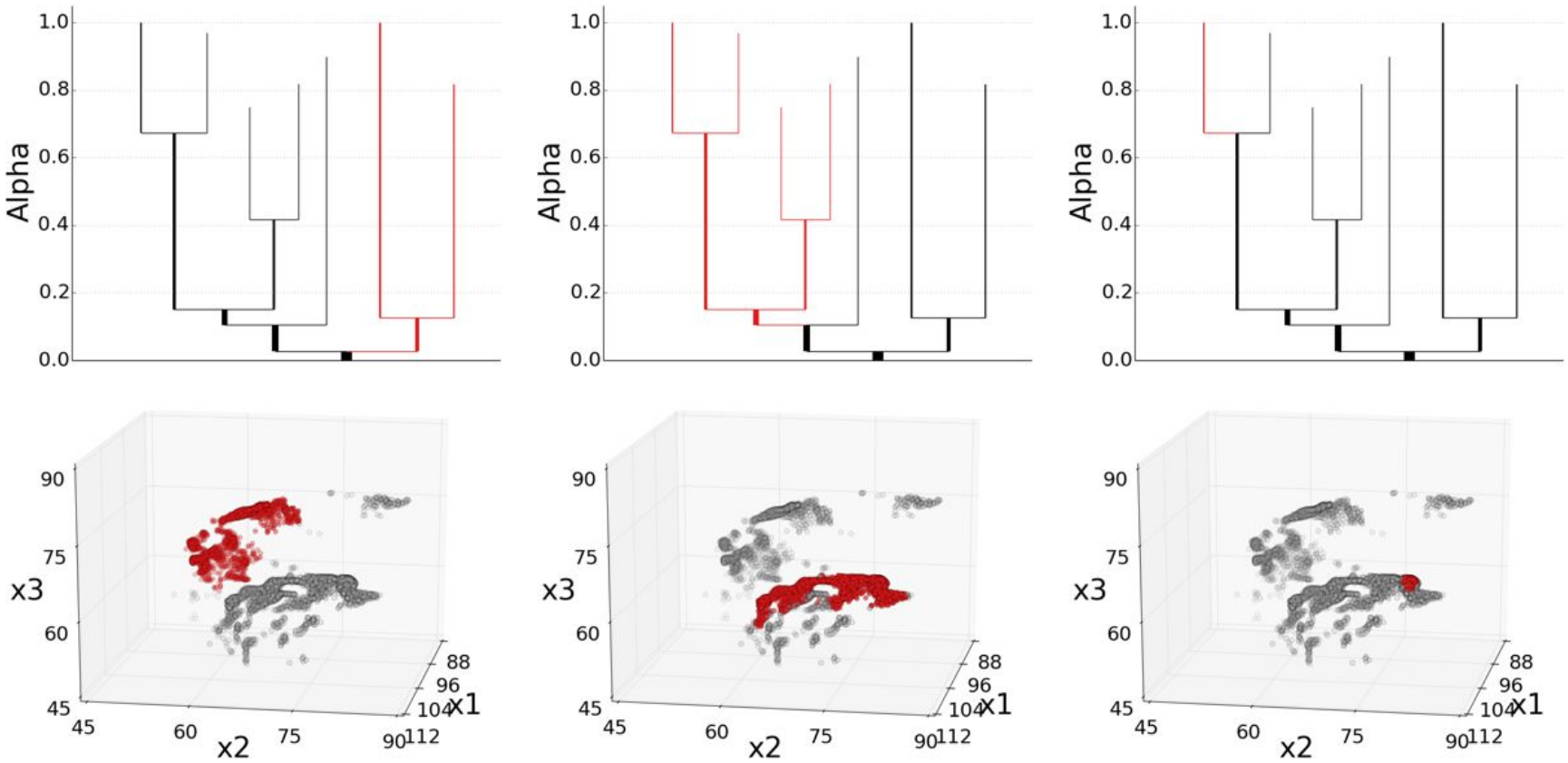


(b)

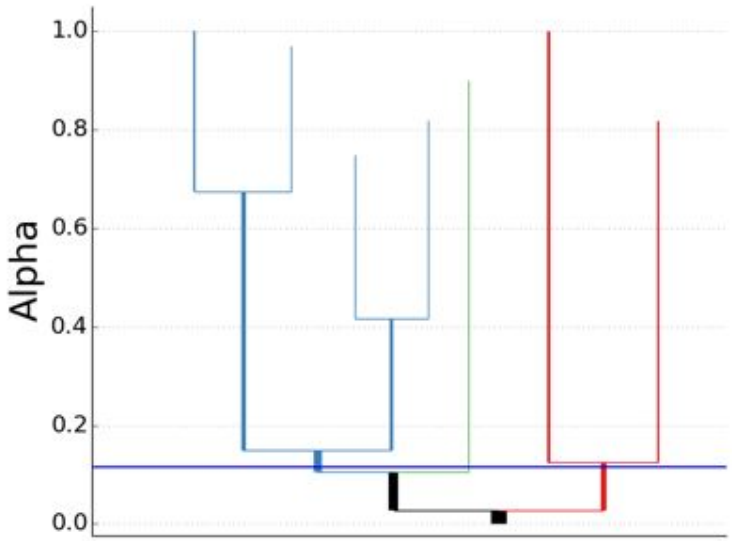


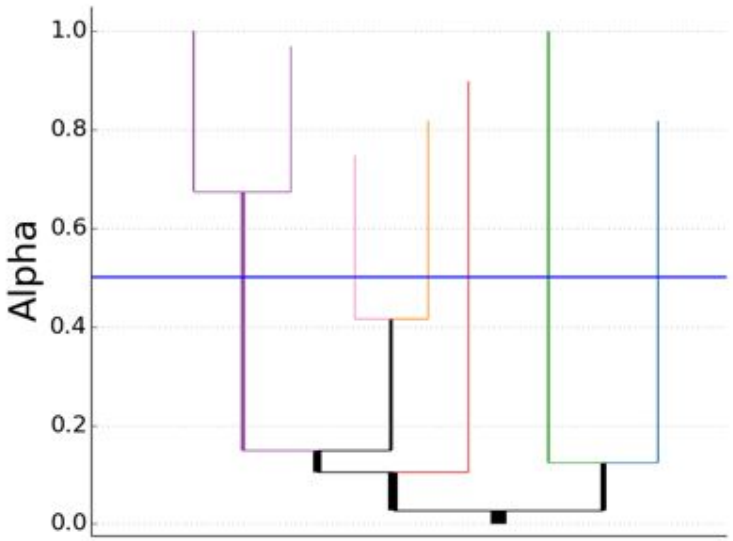
(c)

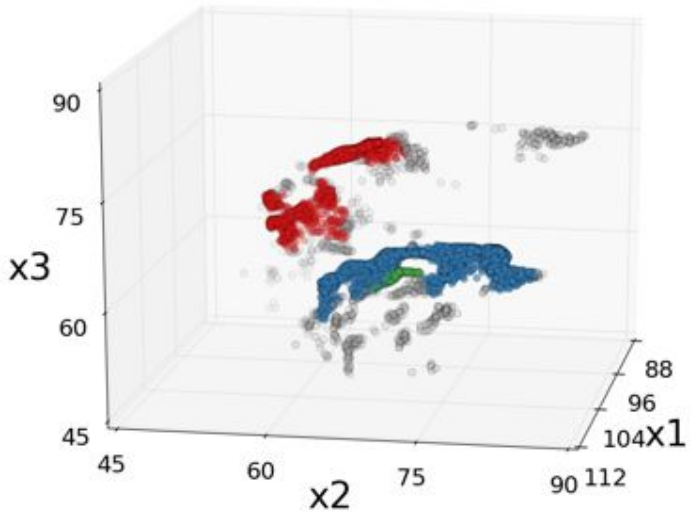
Example 1: clustering endpoints of fiber tracks in the striatum

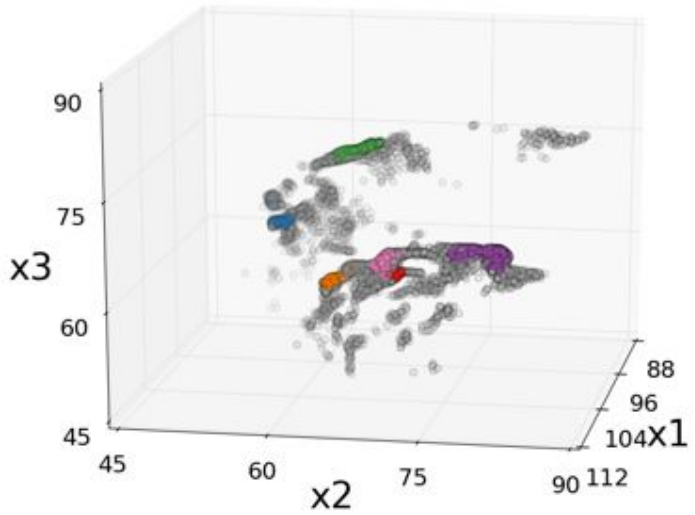


Example 1: clustering endpoints of fiber tracks in the striatum

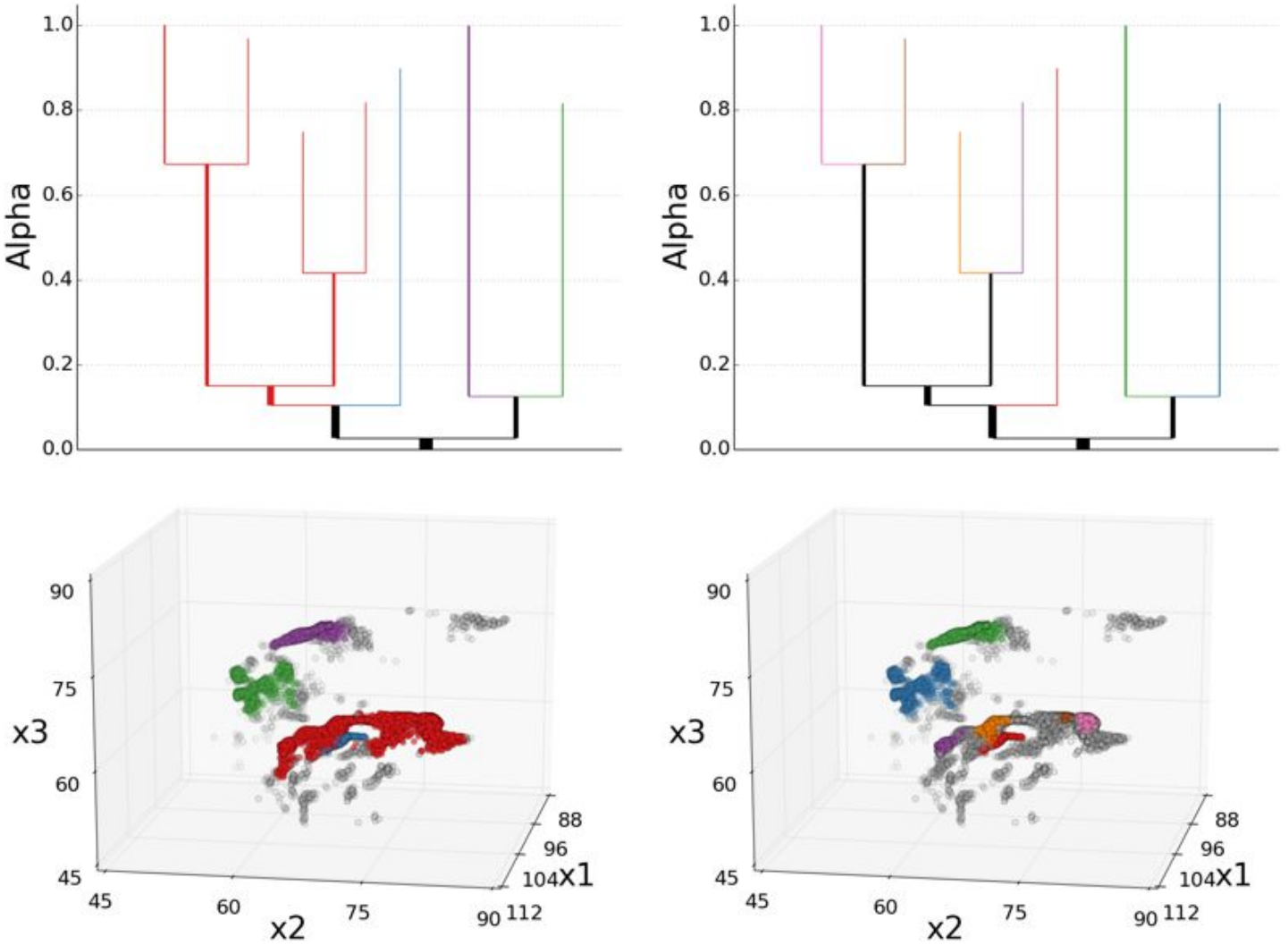








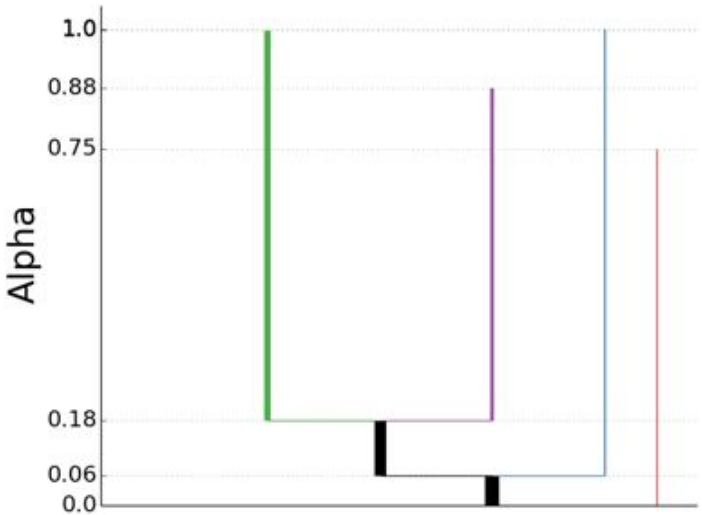
Example 1: clustering endpoints of fiber tracks in the striatum

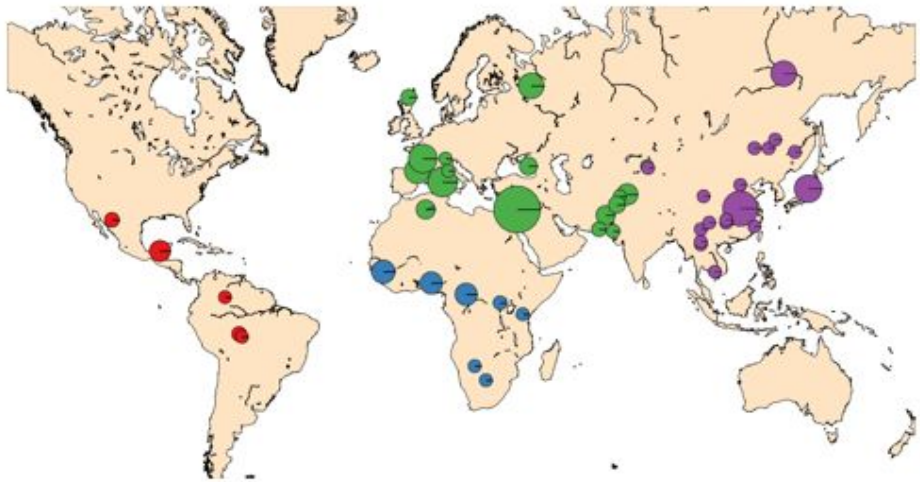


Example 2: clustering individuals into populations using SNPs

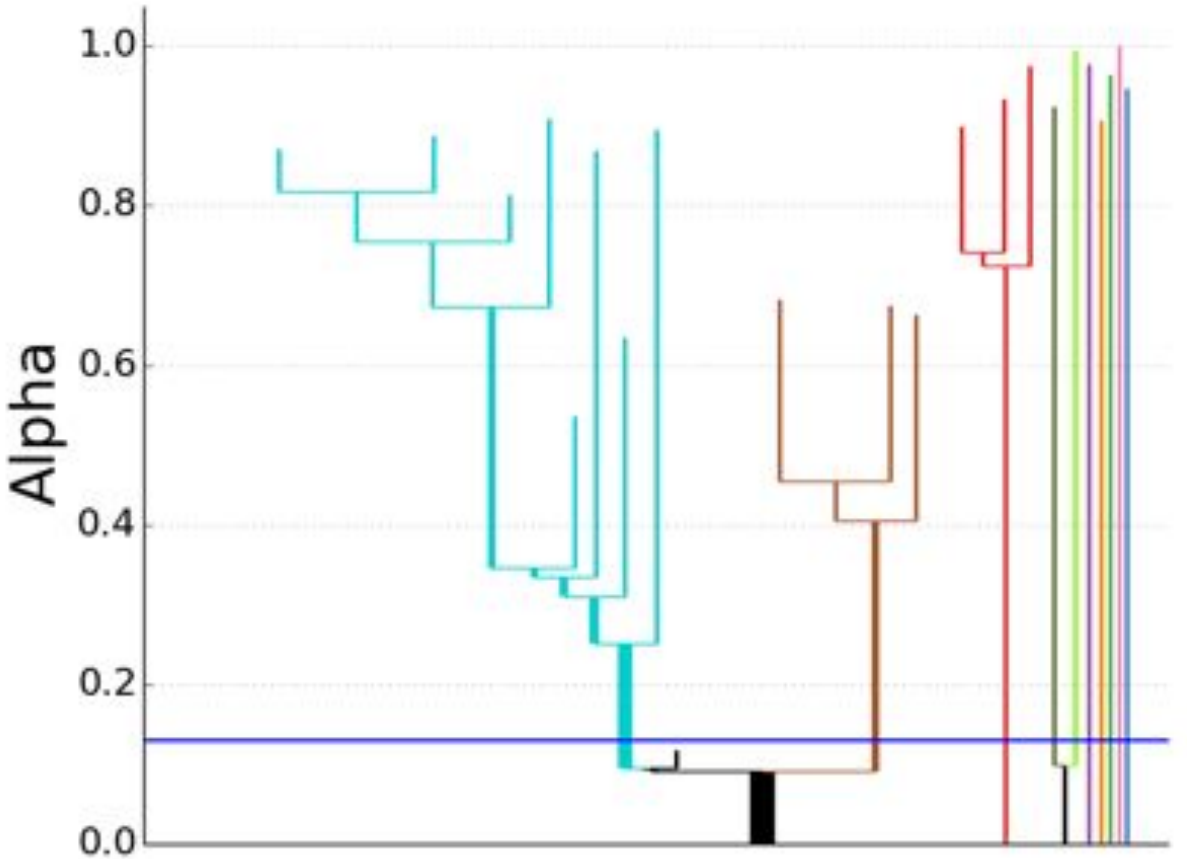
- Data from The Human Genome Diversity Project (HGDP) dataset, available at
  `http://www.hagsc.org/hgdp/files.html`.
- Cleaned-up comprised of 11,775 SNPs from 931 subjects from 53 populations from Crosset et al. (2010).
- The goal of the analysis is to identify the hierarchy of high-density clusters of individuals in the sample, ideally capturing the correct membership in populations.
- In the first level set tree $k = 40$, in the second $k = 6$.

Example 2: clustering individuals into populations using SNPs

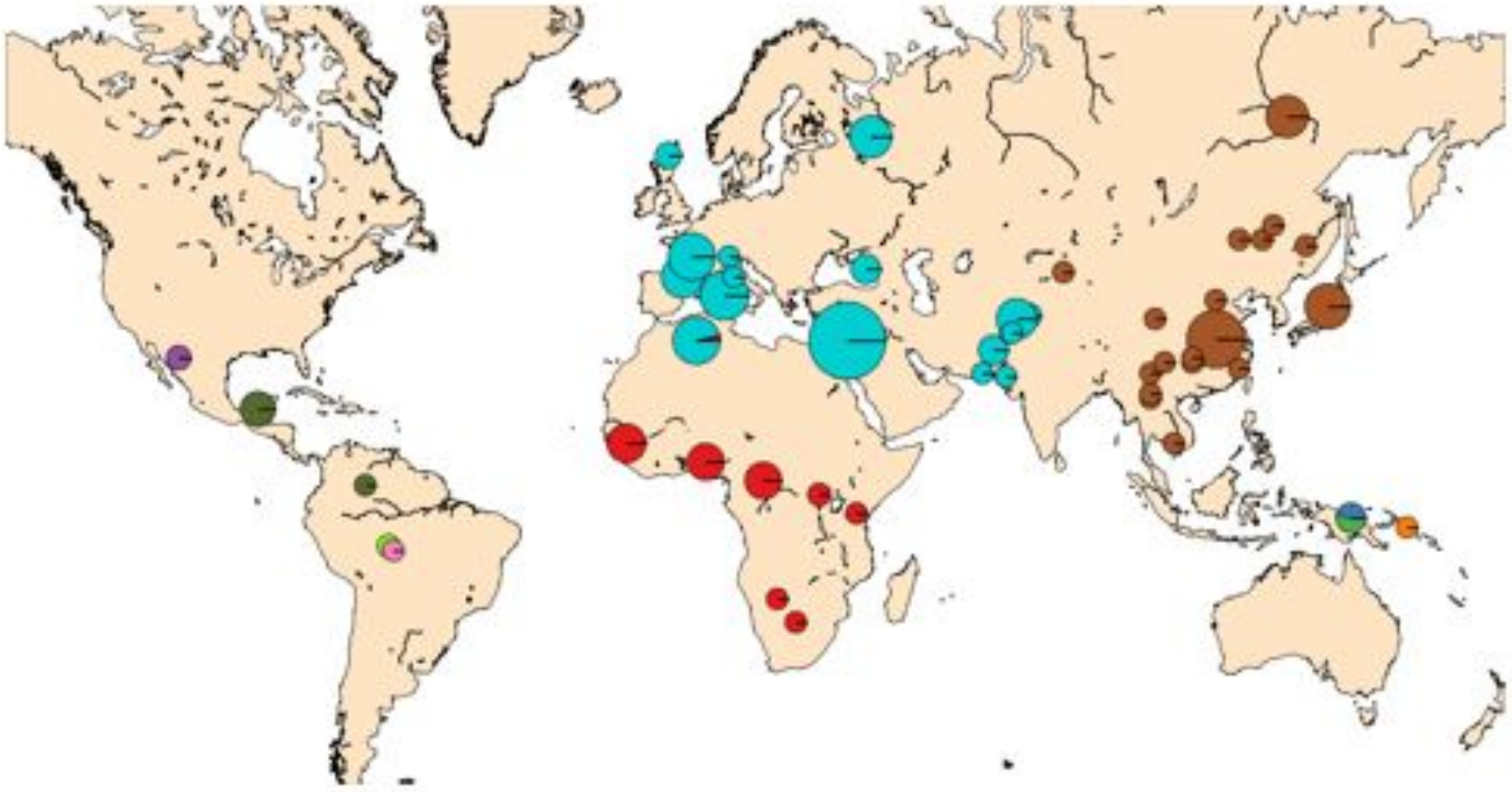




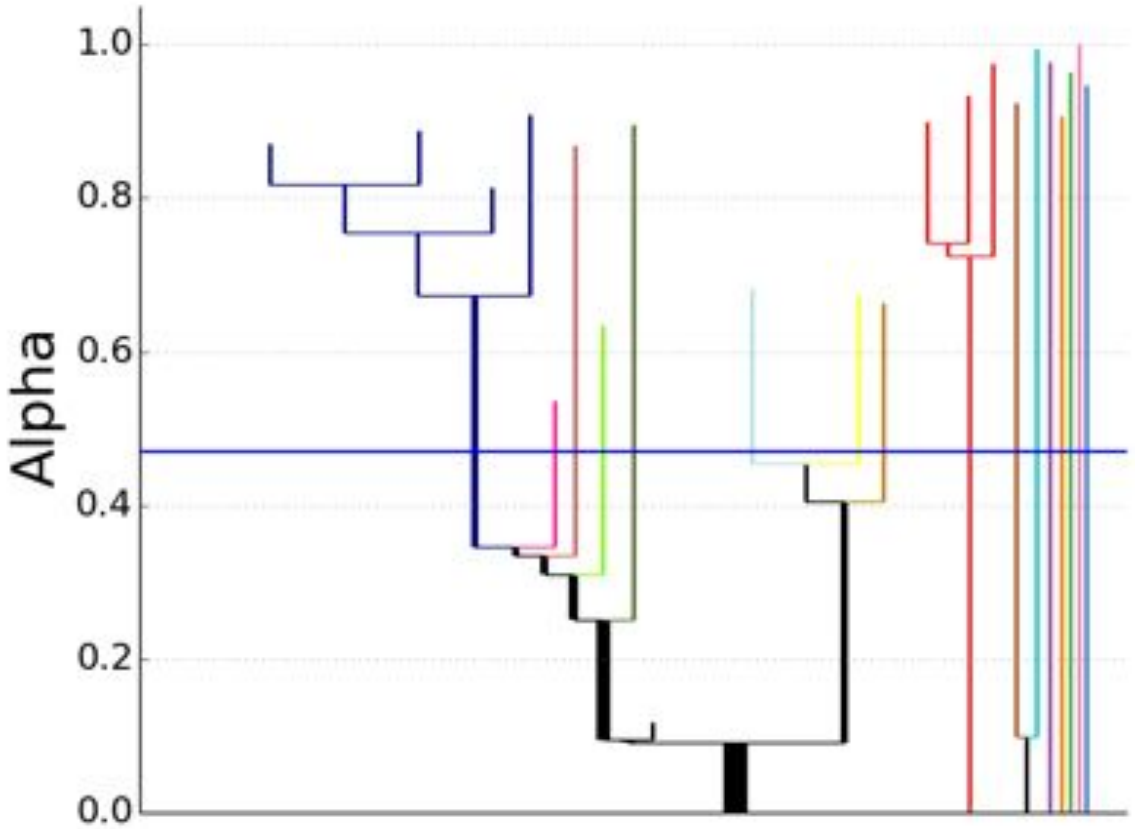
Example 2: clustering individuals into populations using SNPs



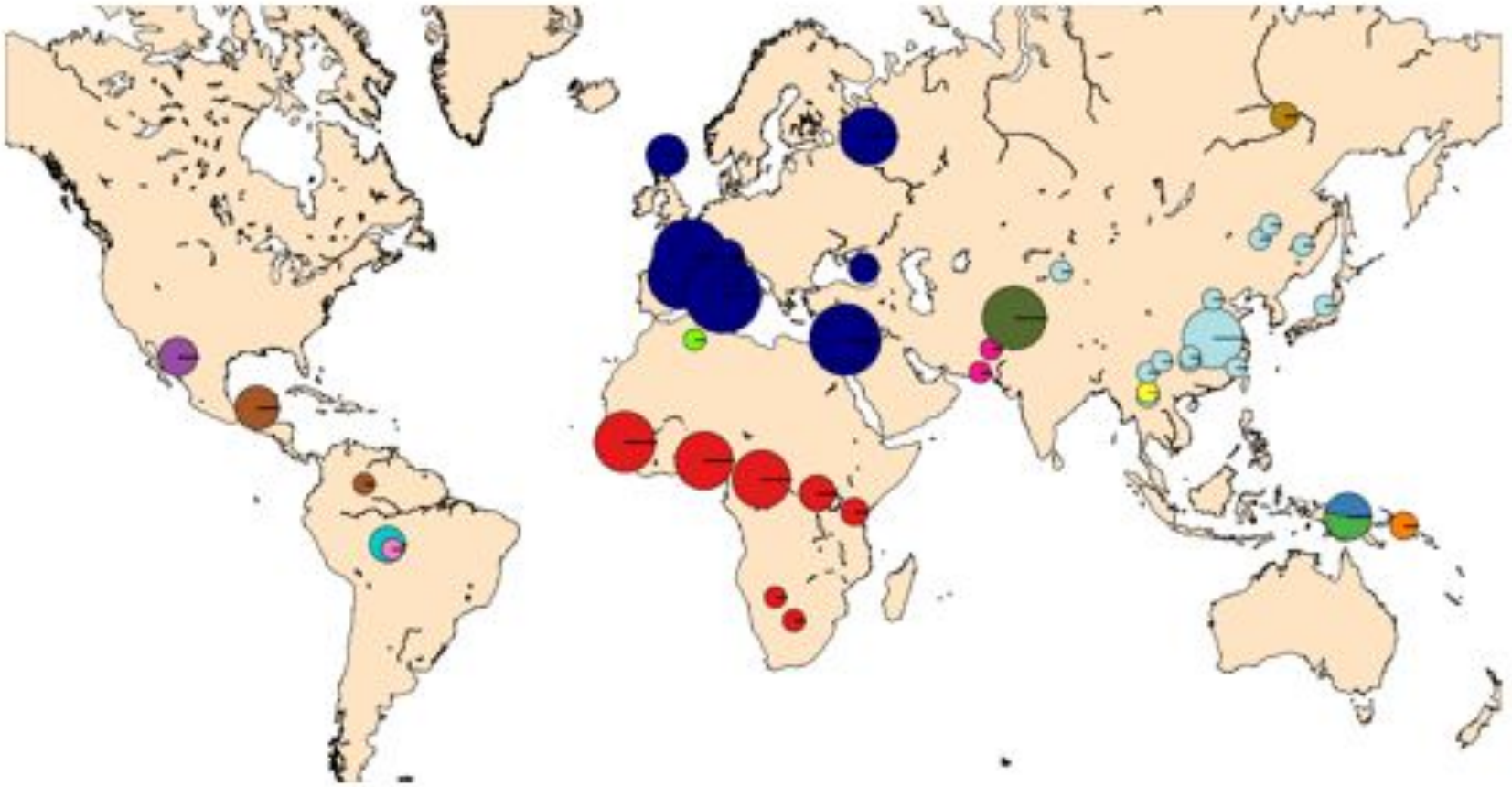
Example 2: clustering individuals into populations using SNPs



Example 2: clustering individuals into populations using SNPs



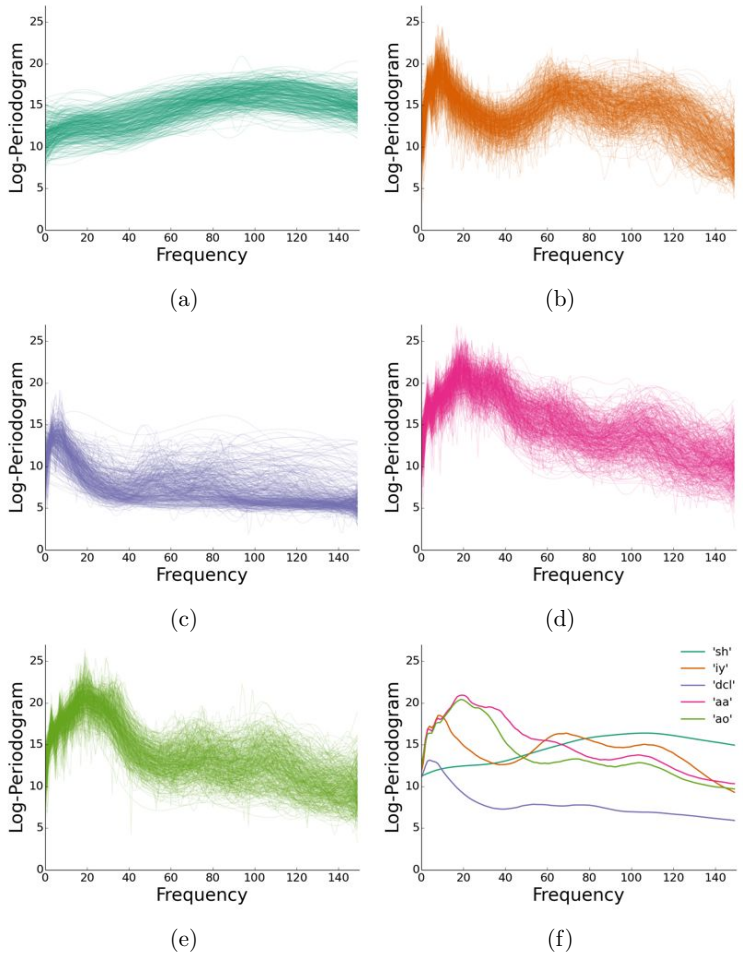
Example 2: clustering individuals into populations using SNPs



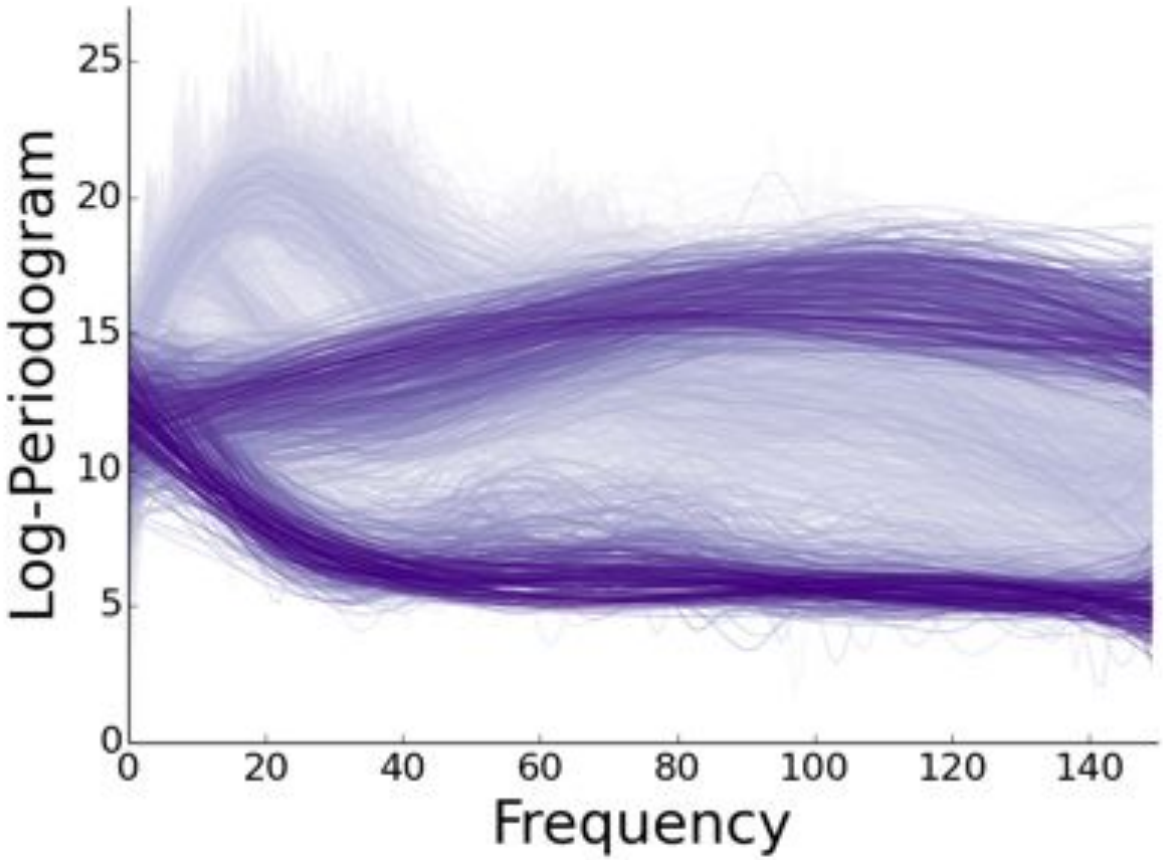
Example 3: clustering phonemes (functional data)

- The phoneme dataset, from Ferraty and Vieu (2006) contains log-periodograms of 2000 instances of digitized human speech, divided evenly between five phonemes: “sh", “dcl" (as in “dark"), “iy" (as in the vowel of “she"), “aa", and “ao". Each recording is treated as a single functional observation, which was smoothed using a cubic spline.
- Distance between function is the L_2 distace (each phoneme is observed over 150 frequencies).
- $k = 20$.

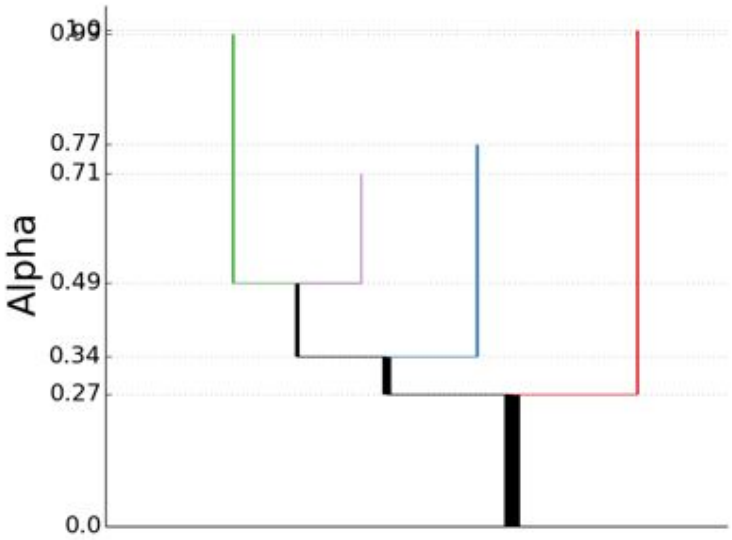
Example 3: clustering phonemes (functional data)

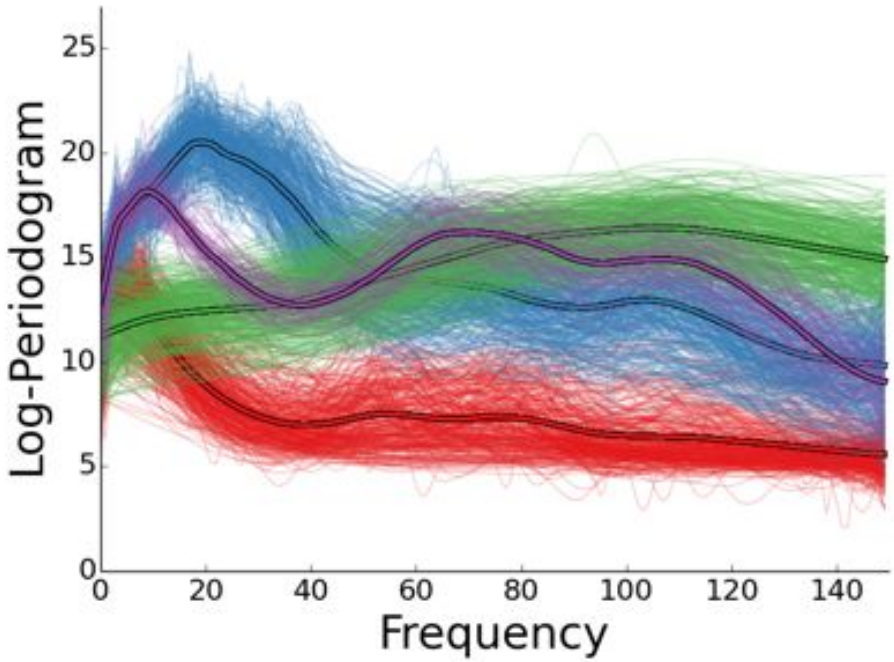


Example 3: clustering phonemes (functional data)



Example 3: clustering phonemes (functional data)

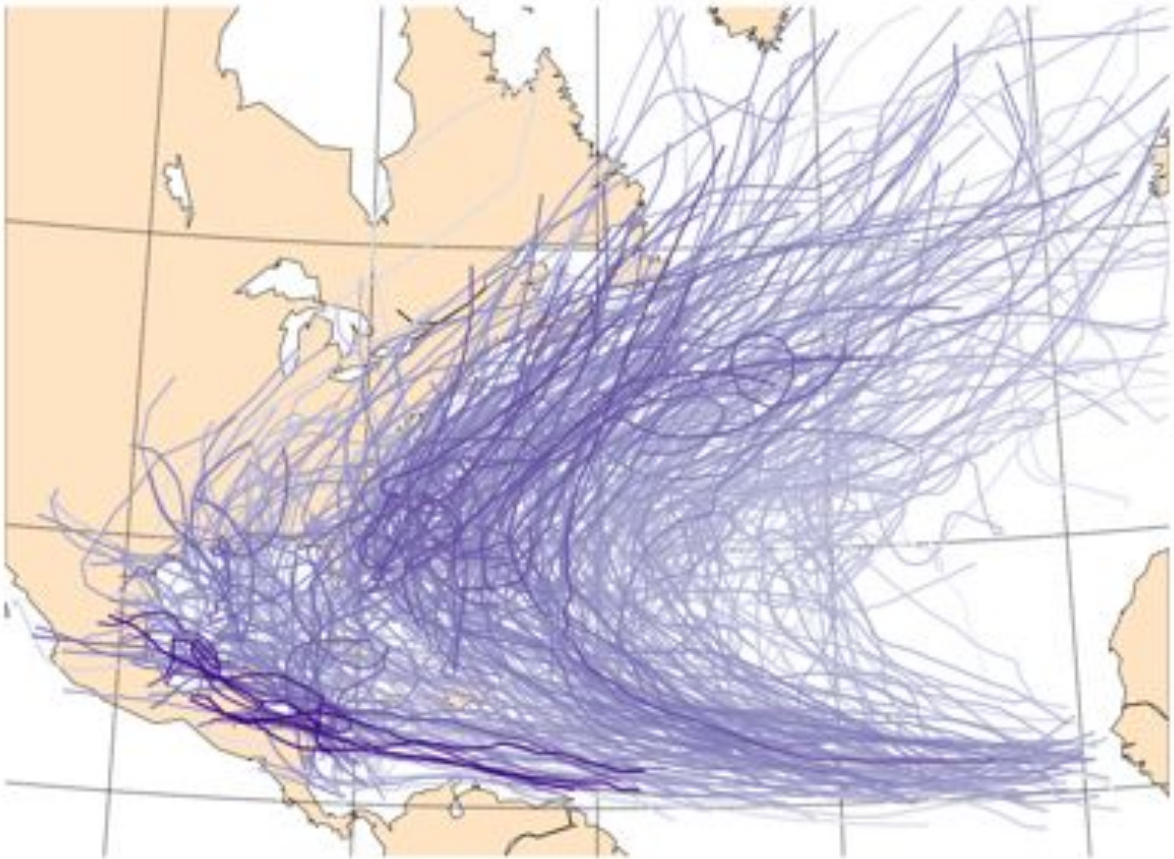




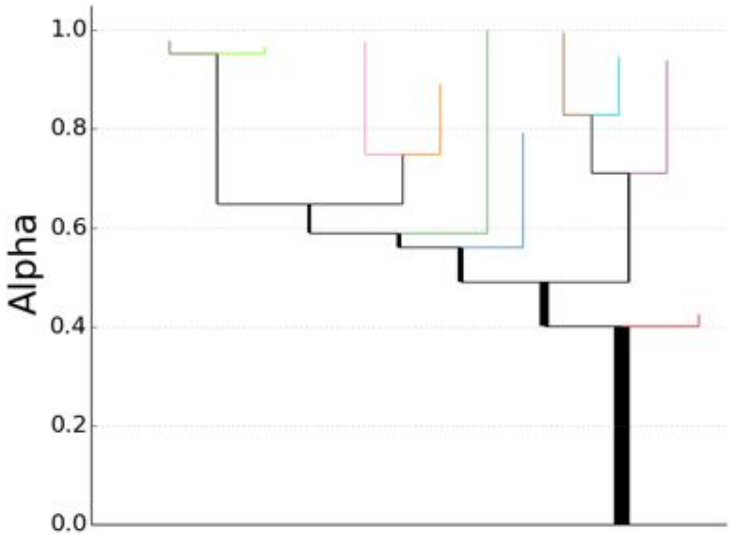
Example 4: clustering hurricane tracks (functional data)

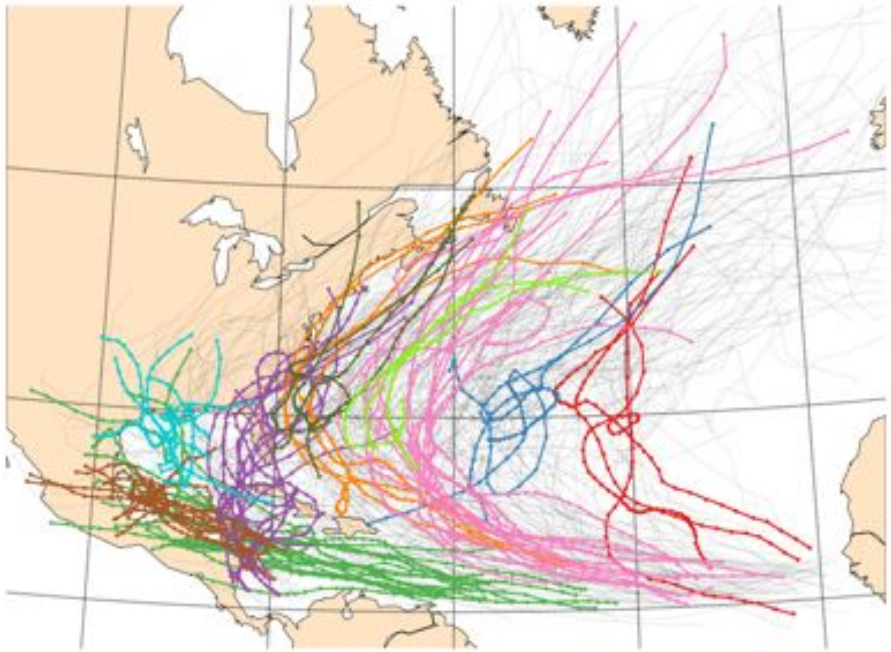
- The U.S. National Hurricane Centers's HURDAT dataset contains positional and atmospheric measurements of North Atlantic tropical cyclons from 1851 to 2012 (Landsea et al., 2013). The coordinates (in degrees latitude and longitude) for each storm are recorded at least every six hours.
- The processed dataset contained 398 hurricane tracks.
- Pairwise distances based on *max-average-min distance* (not a metric).
- $k = 6$ ($\gamma = 2$).

Example 4: clustering hurricane tracks (functional data)



Example 4: clustering hurricane tracks (functional data)

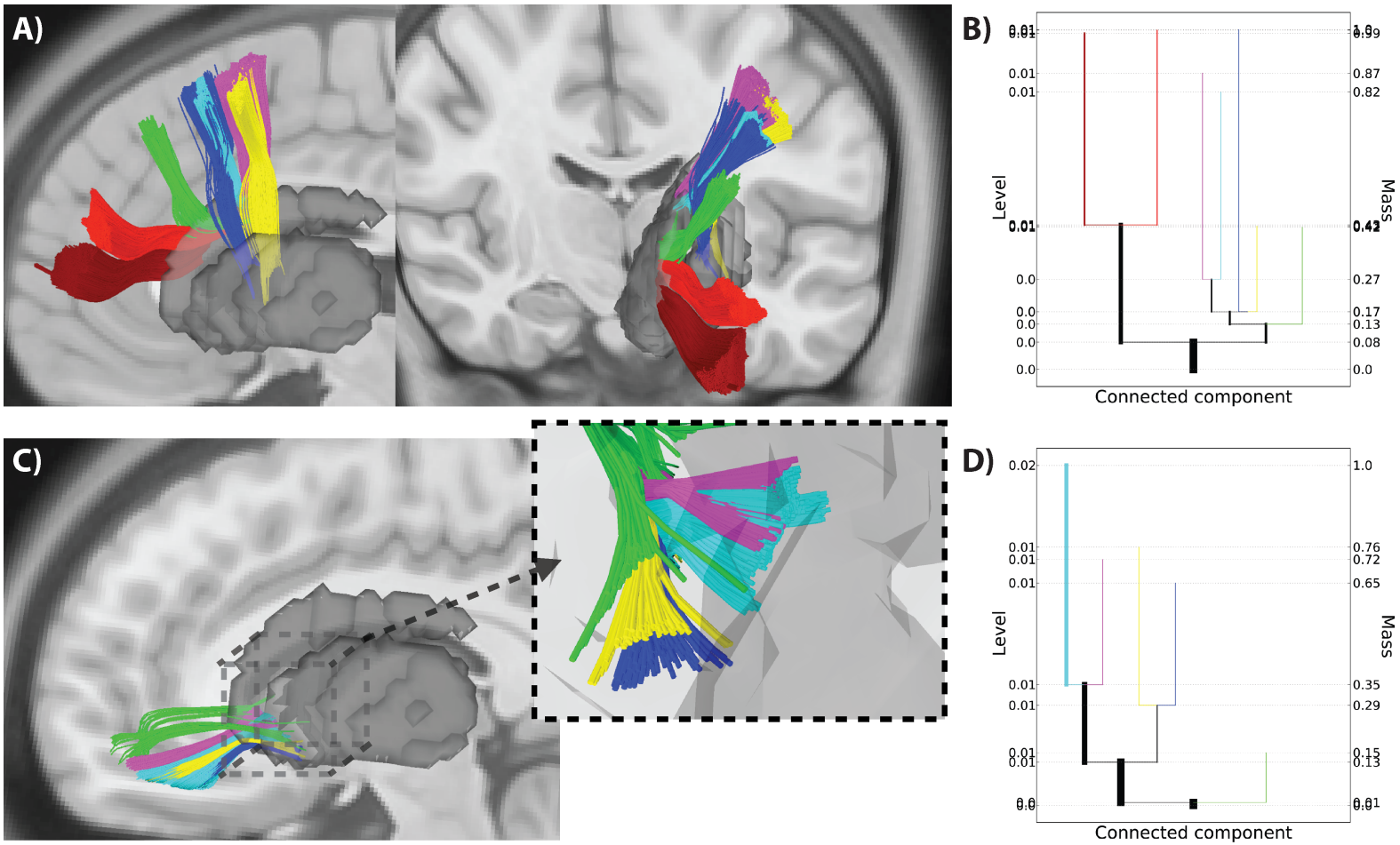




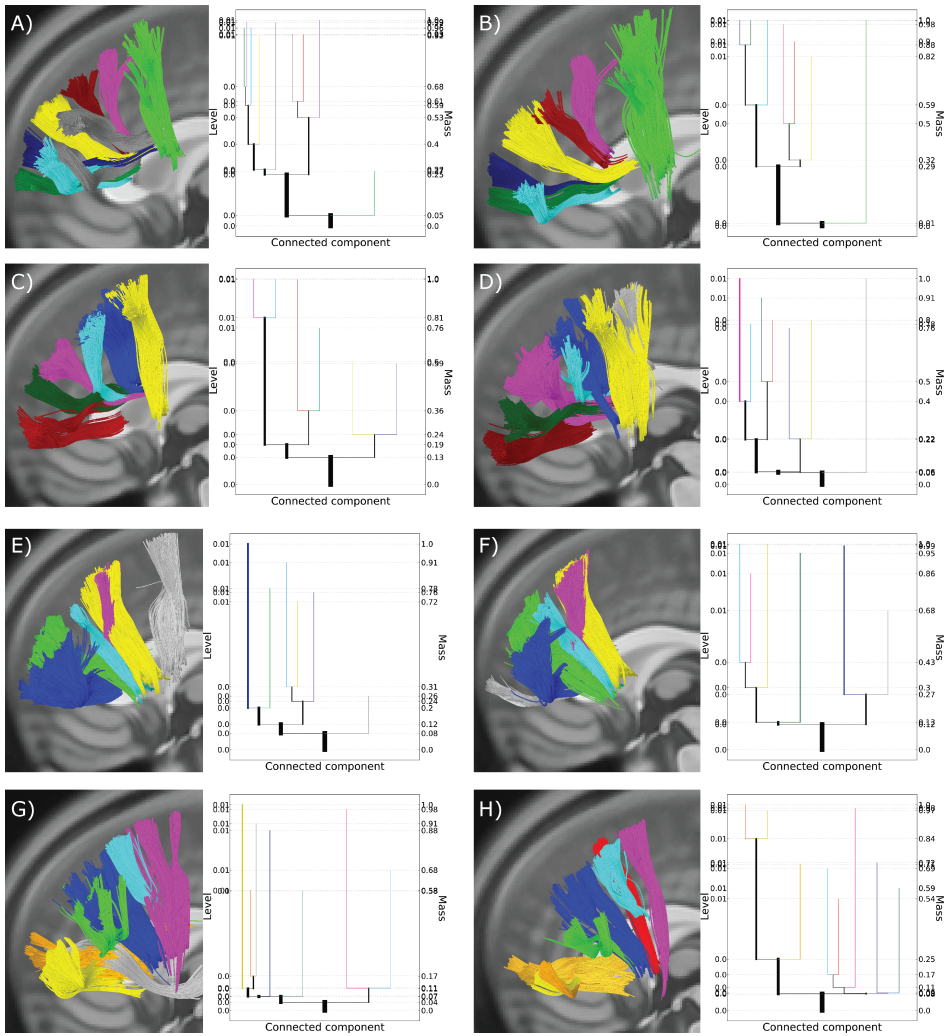
Example 5: clustering fiber tracks (functional data)

- Fiber tractography datasets obtained through DSI techniques. Focus on two corticostriatal pathways: lateral frontal (middle frontal gyrus to striatum) and orbitofrontal (gyrus rectus to striatum).
- A 30 subject template was used.
- We used `DeBaCl` to perform whole fiber tracks segmentation and looked at tracks in the lateral frontal cortex and orbitofrontal cortex. Total of 51,126 fibers.
- Pairwise distances based on *max-average-min distance* (not a metric).
- $k = \lceil 0.25 * n \rceil$
- Work done in collaboration with Timothy Verstynen:
  `http://www.psy.cmu.edu/~coaxlab/`

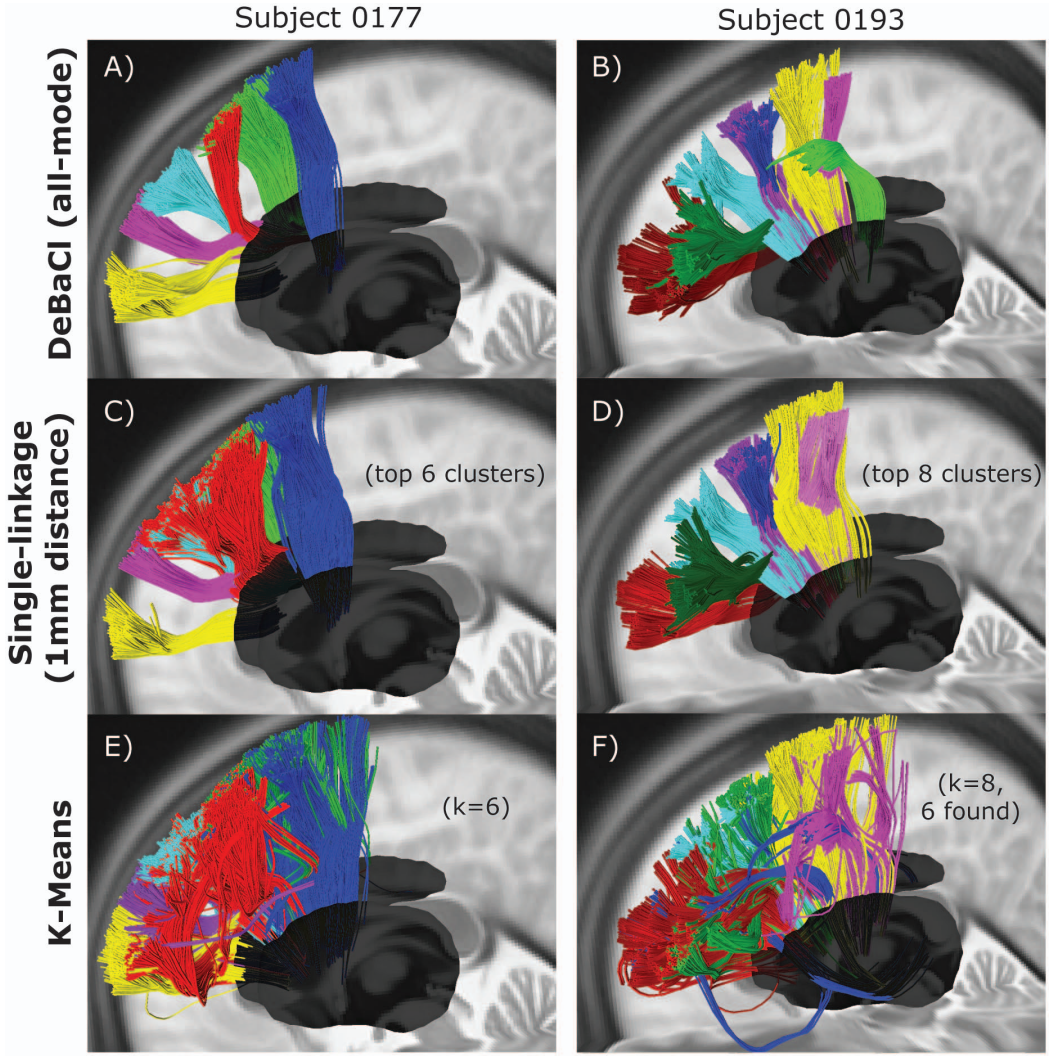
Example 5: clustering fiber tracks (functional data)



Example 5: clustering fiber tracks (functional data)



Example 5: clustering fiber tracks (functional data)



Theoretical analysis of DeBaCl: set-up

- Let P a non-atomic probability measure supported on $S \subset \mathbb{P}^d$. We allow for $\dim(S) < d$ and for S to be of mixed dimension.

- Fix a number $p \in (0,1)$. Define the function $r_p \colon \mathbb{R}^d \to \mathbb{R}_+$ given by

$$x \mapsto r_p(x) = \inf\left\{r > 0 \colon P\left(B(x,r)\right) \geq p\right\}.$$

Thus $r_p(x)$ is the p-th quantile of the univariate variable $\|X - x\|$, $X \sim P$.

DeBaCl estimates r_p and its lower level sets at the sample points $\mathbf{X}_n$.

Theoretical analysis of DeBaCl: set-up

- Let P a non-atomic probability measure supported on $S \subset \mathbb{P}^d$. We allow for $\dim(S) < d$ and for S to be of mixed dimension.

- Fix a number $p \in (0,1)$. Define the function $r_p : \mathbb{R}^d \to \mathbb{R}_+$ given by

$$x \mapsto r_p(x) = \inf\left\{r > 0 : P\left(B(x,r)\right) \geq p\right\}.$$

  Thus $r_p(x)$ is the p-th quantile of the univariate variable $\|X - x\|$, $X \sim P$.

DeBaCl estimates r_p and its lower level sets at the sample points $\mathbf{X}_n$.

Theoretical analysis of DeBaCl: set-up

- Let P a non-atomic probability measure supported on $S \subset \mathbb{P}^d$. We allow for $\dim(S) < d$ and for S to be of mixed dimension.

- Fix a number $p \in (0,1)$. Define the function $r_p : \mathbb{R}^d \to \mathbb{R}_+$ given by

$$x \mapsto r_p(x) = \inf\Big\{r > 0 : P\left(B(x,r)\right) \geq p\Big\}.$$

  Thus $r_p(x)$ is the p-th quantile of the univariate variable $\|X - x\|$, $X \sim P$.

DeBaCl estimates r_p and its lower level sets at the sample points $\mathbf{X}_n$.

- Let K be the uniform kernel $K(x) = \begin{cases} 1 & \|x\| \leq 1 \\ 0 & \text{otherwise} \end{cases}$ and consider the biased, Lebesgue density

$$f_p(x) = \frac{1}{v_d r_p^d(x)} \int_{\mathbb{R}^d} K\left(\frac{\|x-y\|}{r_p(x)}\right) dP(y) = \frac{p}{v_d r_p^d(x)} \propto \frac{1}{r_p^d(x)}, \quad x \in \mathbb{R}^d,$$

with v_d the volume of unit ball in $\mathbb{R}^d$.

- The empirical equivalent of $f_p(x)$ is the k-nn density estimator, where $k = \lceil pn \rceil$:

$$\widehat{f}_p(x) = \frac{k}{n} \frac{1}{v_d \widehat{r}_p^d(x)}, \quad x \in \mathbb{R}^d$$

with $\widehat{r}_p(x)$ the distance from x to its k-th nearest neighborhood in $\mathbf{X}_n$.

- Let K be the uniform kernel $K(x) = \begin{cases} 1 & \|x\| \leq 1 \\ 0 & \text{otherwise} \end{cases}$ and consider the biased, Lebesgue density

$$f_p(x) = \frac{1}{v_d r_p^d(x)} \int_{\mathbb{R}^d} K\left(\frac{\|x-y\|}{r_p(x)}\right) dP(y) = \frac{p}{v_d r_p^d(x)} \propto \frac{1}{r_p^d(x)}, \quad x \in \mathbb{R}^d,$$

with v_d the volume of unit ball in $\mathbb{R}^d$.

- The empirical equivalent of $f_p(x)$ is the k-nn density estimator, where $k = \lceil pn \rceil$:

$$\widehat{f}_p(x) = \frac{k}{n} \frac{1}{v_d \widehat{r}_p^d(x)}, \quad x \in \mathbb{R}^d$$

with $\widehat{r}_p(x)$ the distance from x to its k-th nearest neighborhood in $\mathbf{X}_n$.

f_p is biased

If P has a continuous Lebesgue density f, then, a.e.,

$$\lim_{p \to 0} \frac{p}{v_d r_p^d(x)} = f(x),$$

and convergence holds uniformly over compacts. If P has a continuous density f w.r.t. $\mathcal{H}_k$, then, a.e.,

$$\lim_{p \to 0} \frac{p}{v_{\mathbf{k}} r_{x,p}^{\mathbf{k}}} = f(x),$$

Density-based clustering with no densities

The upper level sets of $f_p(\cdot)$ are the lower level sets of $r_p(\cdot)$.

- For $r > 0$, let $L(r) = \{x \in \mathbb{R}^d : r_p(x) \leq r\} \cap S$.
  Define the r-clusters of P as the maximal (path) connected component of $L(r)$. The resulting tree is algorithmically hard.

- Algorithmic connectedness. Two points in $L(r)$ are algorithmically connected if there exists $\{x = z_0, z_1, \ldots, z_m = y\} \subset L(r)$ such that $z_i \in B(z_{i+1}, r_{z_i+1})$.

Algorithmic Tree

The algorithmic tree $\mathcal{T}$ is the dendrogram of the r-clusters of P with path connectedness replaced by algorithmic connectedness.

Density-based clustering with no densities

The upper level sets of $f_p(\cdot)$ are the lower level sets of $r_p(\cdot)$.

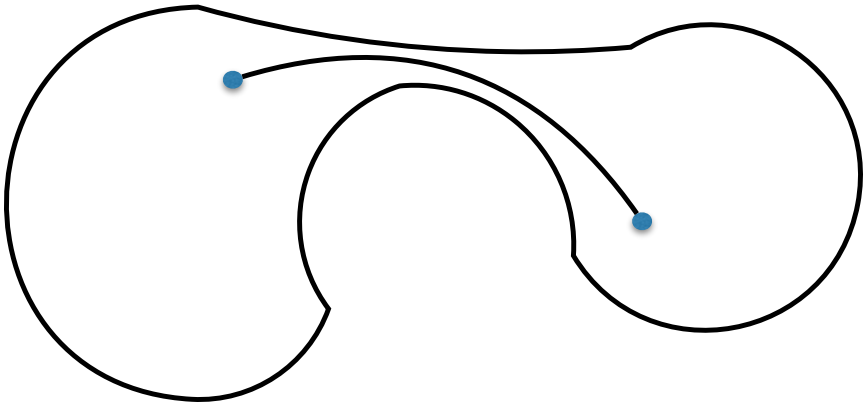
- For $r > 0$, let $L(r) = \{x \in \mathbb{R}^d : r_p(x) \leq r\} \cap S$.
  Define the r-clusters of P as the maximal (path) connected component of $L(r)$. The resulting tree is algorithmically hard.

- Algorithmic connectedness. Two points in $L(r)$ are algorithmically connected if there exists $\{x = z_0, z_1, \ldots, z_m = y\} \subset L(r)$ such that $z_i \in B(z_{i+1}, r_{z_i+1})$.

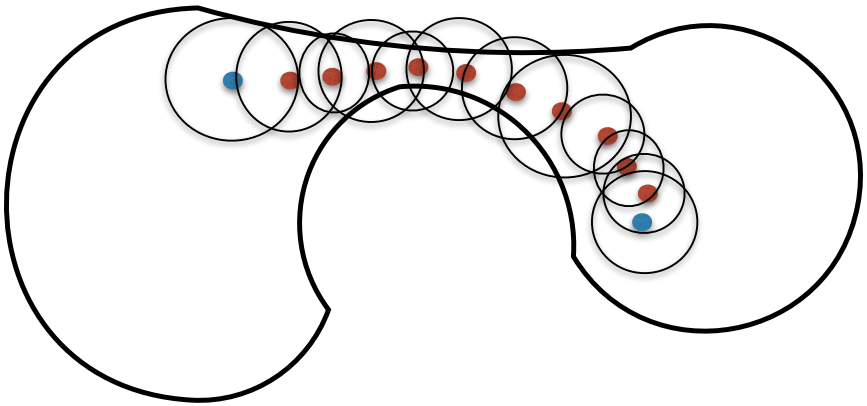
Algorithmic Tree

The algorithmic tree $\mathcal{T}$ is the dendrogram of the r-clusters of P with path connectedness replaced by algorithmic connectedness.

Density-based clustering with no densities

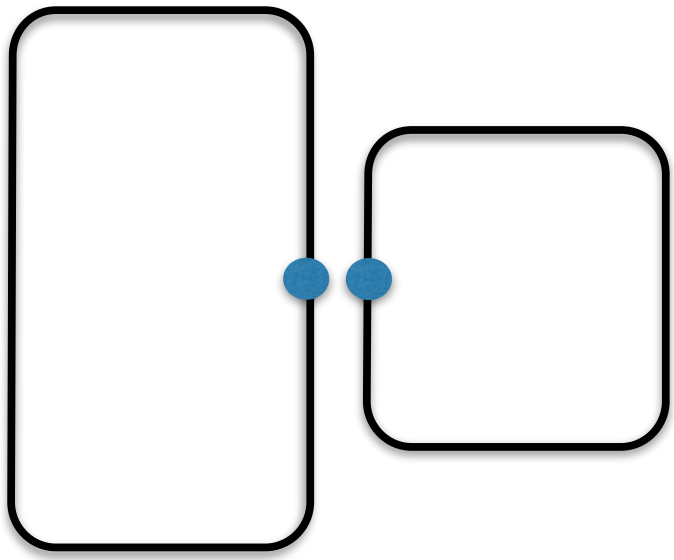


Density-based clustering with no densities



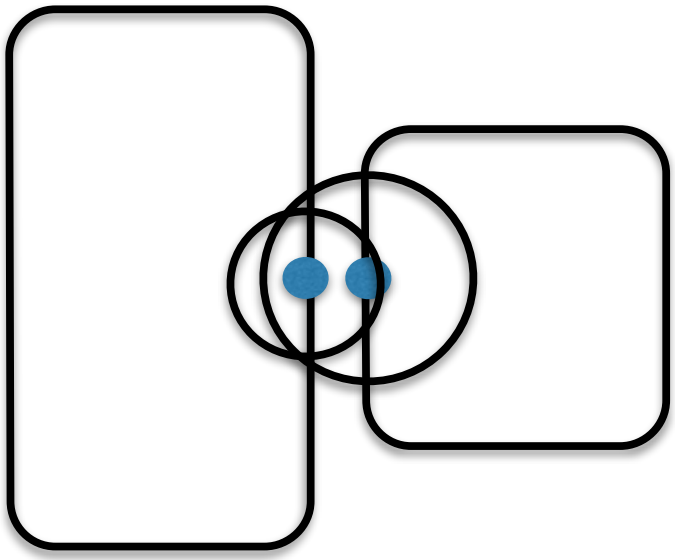
Density-based clustering with no densities

- Connectedness implies algorithmic connectedness but not the other way around.



Density-based clustering with no densities

- Connectedness implies algorithmic connectedness but not the other way around.



Density-based clustering with no densities

Empirical Tree

The empirical tree $\widehat{\mathcal{T}}_n$ is the dendrogram produced by `DeBaCl` (i.e. based on the estimated distances $\widehat{r}_1, \ldots, \widehat{r}_n$).

Oracle Tree

The oracle tree $\mathcal{T}_n$ is the dedrogram of nested subsets of $\mathbf{X}_n$ obtained by running `DeBaCl` if an oracle gave us the true values of $r_p(X_1), \ldots, r_p(X_n)$.

They are both data dependent!

A tale of many trees

- The “density" tree (high density clusters of f)
  - my not be defined (if, e.g., S is of mixed dimension)
  - is statistically and algorithmically hard in d.
- The algorithmic r-tree (tree of r-clusters with path connectedness replaced by algorithmic connectedness): our target.

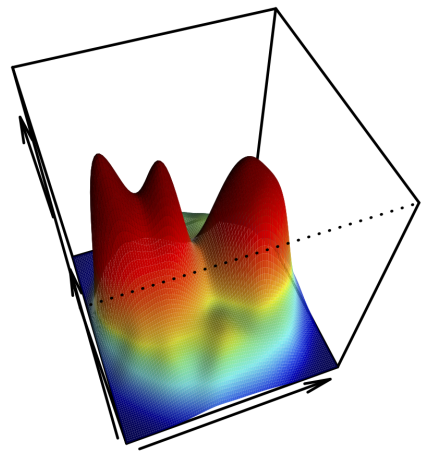
- The oracle tree: the output of DeBaCl using the true r_{X_i}'s.
- The empirical tree: the output of DeBaCl, using the estimated distances $\widehat{r}_{X_1}, \ldots, \widehat{r}_{X_n}$.

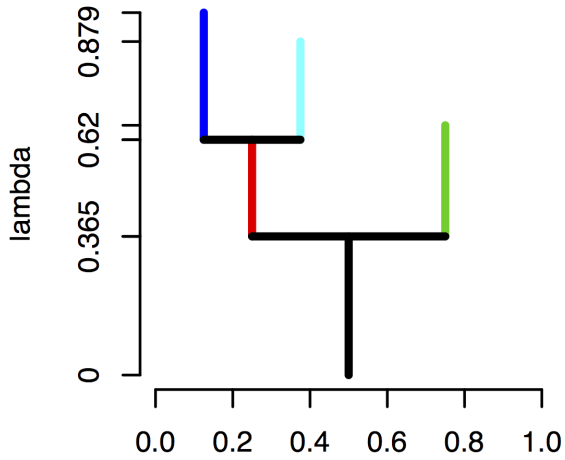
A tale of many trees

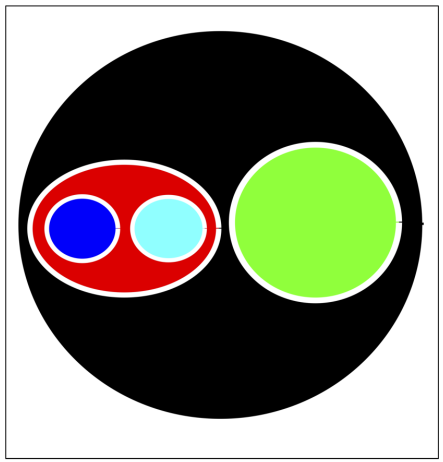
- The “density" tree (high density clusters of f)
  - my not be defined (if, e.g., S is of mixed dimension)
  - is statistically and algorithmically hard in d.
- The algorithmic r-tree (tree of r-clusters with path connectedness replaced by algorithmic connectedness): our target.

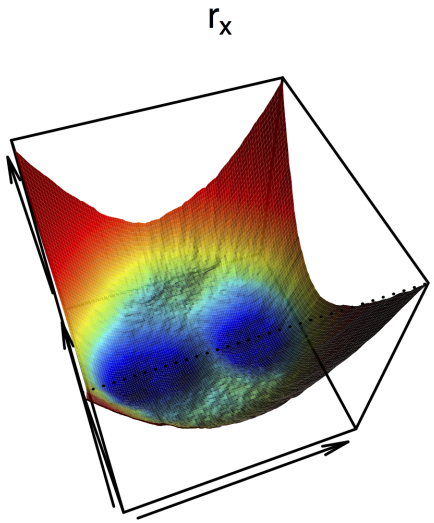
- The oracle tree: the output of DeBaCl using the true r_{X_i}'s.
- The empirical tree: the output of DeBaCl, using the estimated distances $\widehat{r}_{X_1}, \ldots, \widehat{r}_{X_n}$.

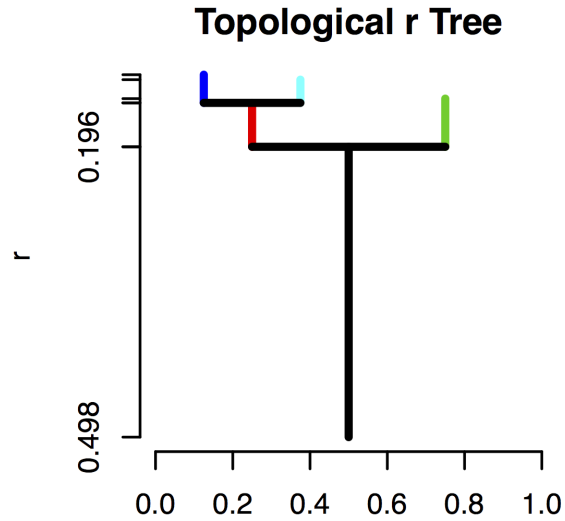
A tale of many trees

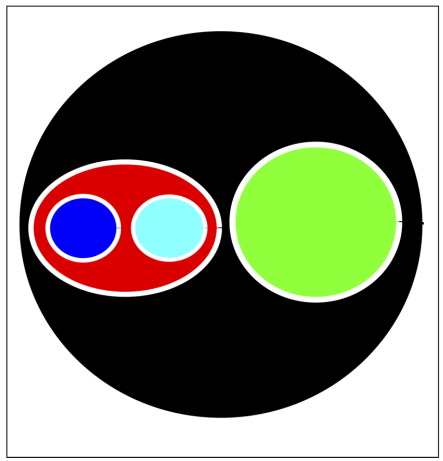




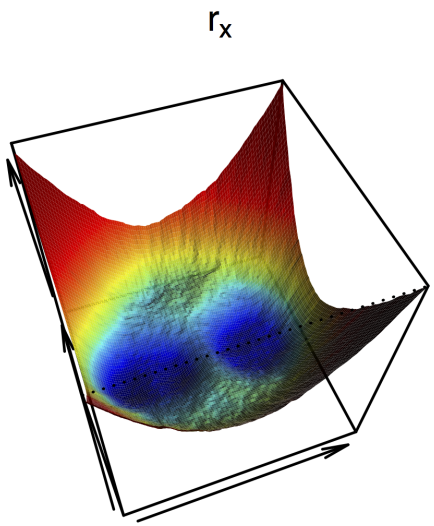


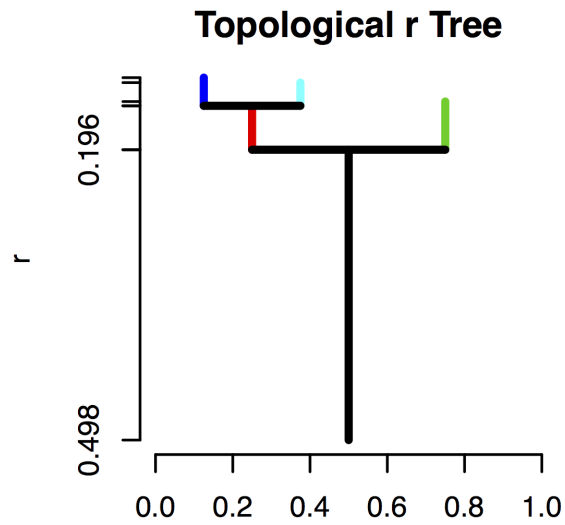


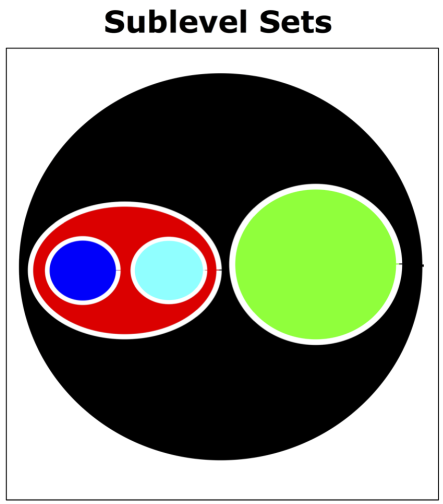


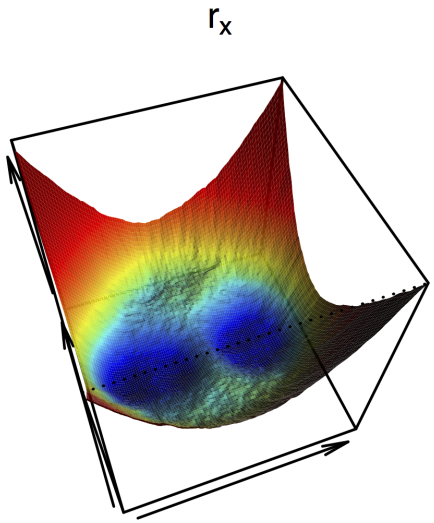


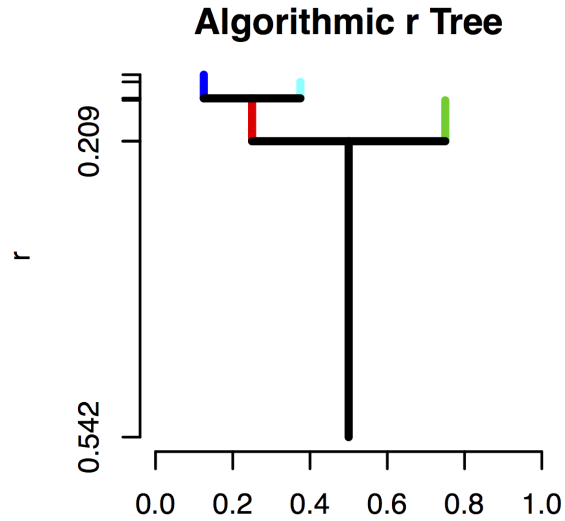
A tale of many trees

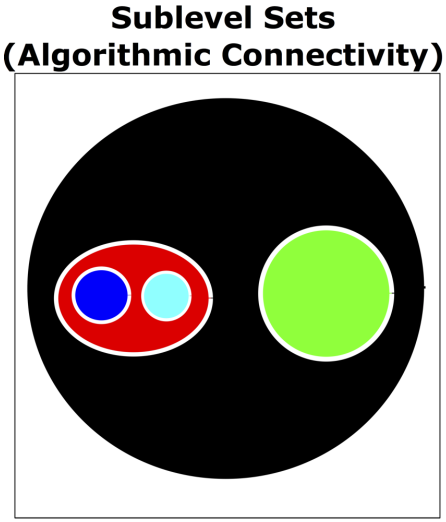




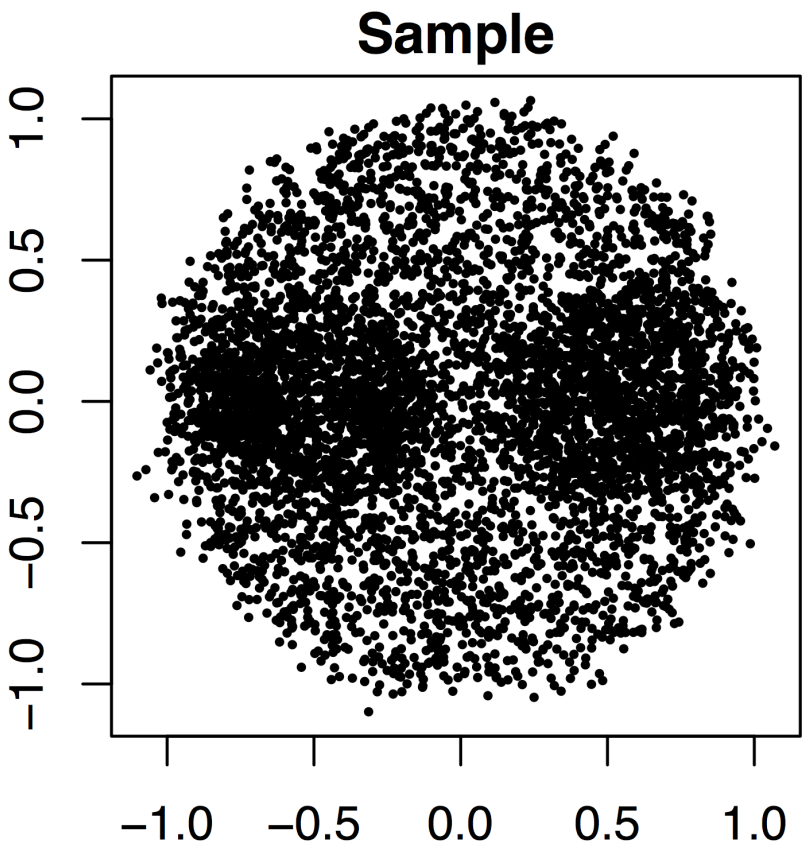




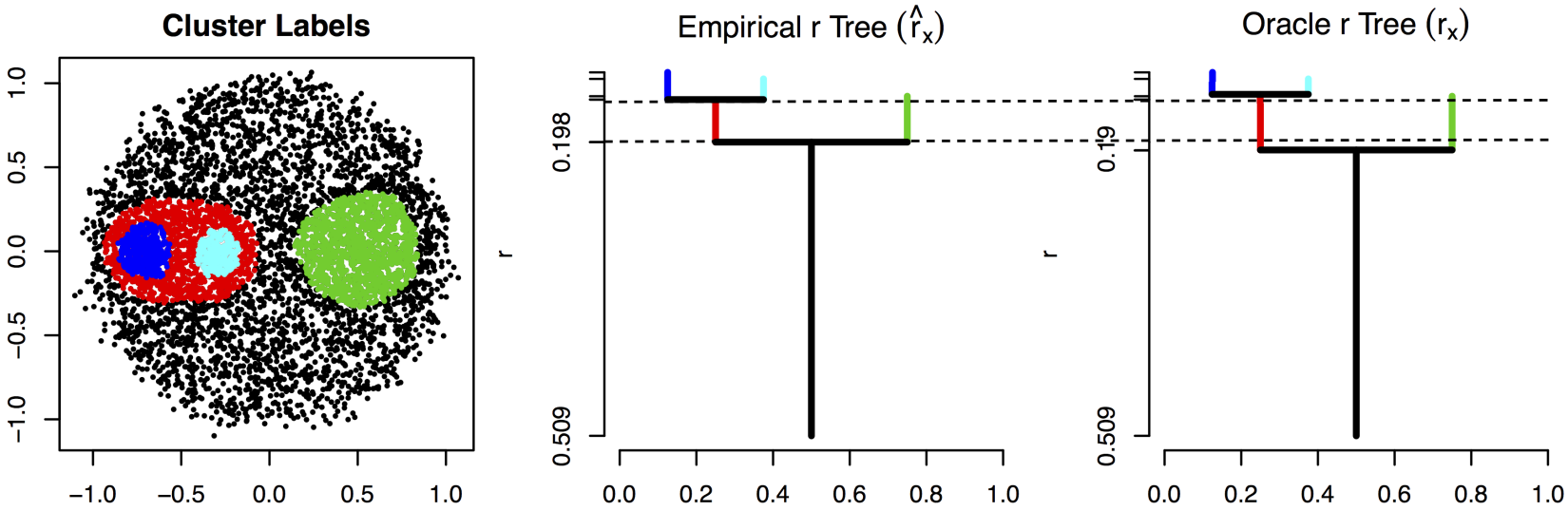




A tale of many trees



A tale of many trees



Consistency of DeBaCl

Tree Consistency

A cluster A of $\mathcal{T}$ is said to be consistency estimable if, with high probability, $\widehat{A}_n := A \cap \mathbf{X}_n$ is a cluster of $\widehat{\mathcal{T}}_n$. Consistency of the tree follows from uniform consistency over clusters.

To prove consistency we will show that

- $\widehat{\mathcal{T}}_n$ the (empirical tree) and $\mathcal{T}_n$ (the oracle tree) yield nearly the same hierarchy of clusters over $\mathbf{X}_n$, with high probability;
- $\mathcal{T}_n$ (the oracle tree) consistently estimates most of $\mathcal{T}$ (the algorithmic tree).

Consistency of DeBaCl

Tree Consistency

A cluster A of $\mathcal{T}$ is said to be consistency estimable if, with high probability, $\widehat{A}_n := A \cap \mathbf{X}_n$ is a cluster of $\widehat{\mathcal{T}}_n$. Consistency of the tree follows from uniform consistency over clusters.

To prove consistency we will show that

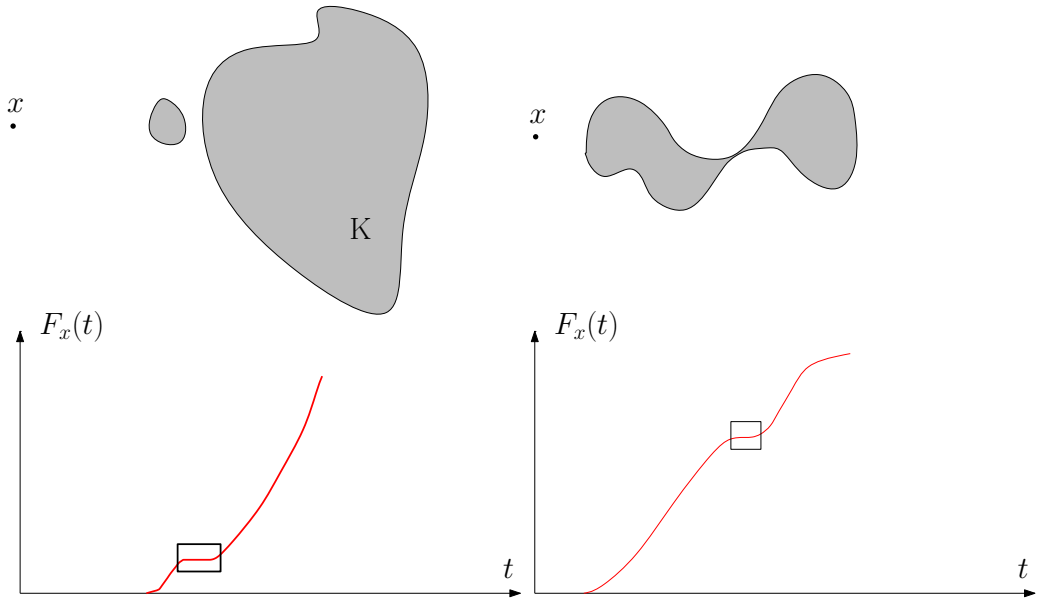
- $\widehat{\mathcal{T}}_n$ the (empirical tree) and $\mathcal{T}_n$ (the oracle tree) yield nearly the same hierarchy of clusters over $\mathbf{X}_n$, with high probability;
- $\mathcal{T}_n$ (the oracle tree) consistently estimates most of $\mathcal{T}$ (the algorithmic tree).

The empirical tree $\widehat{\mathcal{T}}_n$ is close to the oracle tree $\mathcal{T}_n$

Write r_i and $\widehat{r}_i$ for r_{X_i} and $\widehat{r}_{X_i}$.

Regularity Assumption

Let F_x be the cdf of $\|X - x\|$, $X \sim P$. For some $c > 0$ and P-almost all x, $F'_x(r_p(x)) \geq c$. The constant c depends on $\dim(P)$.



Courtesy of Betrand Michel

The empirical tree $\widehat{\mathcal{T}}_n$ is close to the oracle tree $\mathcal{T}_n$

Lemma

There exists a constant C, depending on $\dim(P)$*, such that, with probability at least* $1/n$*,*

$$\max_i \left|\widehat{r}_i - r_i\right| \leq C\sqrt{\frac{\log n}{n}} := \epsilon_n.$$

- For a uniform distribution in full dimension $C \propto \frac{1}{p^d}$.

For fixed $r > 0$, consider the (random) sublevel sets:

$$L_n(r) = \{X_i \in \{1, \ldots, n\} : r_i \leq r\}, \quad \text{(oracle)} \tag{1}$$

$$\widehat{L}_n(r) = \{X_i \in \{1, \ldots, n\} : \widehat{r}_i \leq r\}, \quad \text{(empirical)}. \tag{2}$$

Corollary (Level set consistency)

Uniformly over all $r > \epsilon_n$, with probability at least $1 - 1/n$,

$$L_n(r - \epsilon_n) \subset \widehat{L}_n(r) \subset L_n(r + \epsilon_n),$$

and

$$\widehat{L}_n(r - \epsilon_n) \subset L_n(r) \subset \widehat{L}_n(r + \epsilon_n).$$

If the c.d.f. of $r_p(X)$ with $X \sim P$ is locally Lipschizt at r, with high probability the proportion of misclustered nodes at level r is $O\left(\sqrt{\frac{\log n}{n}}\right)$.

Stability (yields correct connectivity)

The interval of heights $I = (a, b)$ is *stable* for $\mathcal{T}_n$ if the dendrogram does not change when the k-nn distances involved are perturbed by ϵ_n. Sufficient conditions are that the clusters are at least $2\epsilon_n$- apart and $\mathbf{X}_n$ is "sufficiently dense".

Corollary (Cluster consistency)

If I is a stable interval, then for each $a + \epsilon_n \leq r \leq b - \epsilon_n$, the number of clusters in $L_r(n)$ and $\widehat{L}_n(r)$ is the same and constant and each cluster A of $L_n(r)$ is such that $B \subset A \subset B'$, for some B cluster of $\widehat{L}_n(r - \epsilon_n)$ and B' of $\widehat{L}_n(r + \epsilon_n)$.

Stability is a local property and does not hold around at values r of near branching points.

Stability (yields correct connectivity)

The interval of heights $I = (a, b)$ is *stable* for $\mathcal{T}_n$ if the dendrogram does not change when the k-nn distances involved are perturbed by ϵ_n. Sufficient conditions are that the clusters are at least $2\epsilon_n$- apart and $\mathbf{X}_n$ is "sufficiently dense".

Corollary (Cluster consistency)

If I is a stable interval, then for each $a + \epsilon_n \leq r \leq b - \epsilon_n$, the number of clusters in $L_r(n)$ and $\widehat{L}_n(r)$ is the same and constant and each cluster A of $L_n(r)$ is such that $B \subset A \subset B'$, for some B cluster of $\widehat{L}_n(r - \epsilon_n)$ and B' of $\widehat{L}_n(r + \epsilon_n)$.

Stability is a local property and does not hold around at values r of near branching points.

Stability (yields correct connectivity)

The interval of heights $I = (a, b)$ is *stable* for $\mathcal{T}_n$ if the dendrogram does not change when the k-nn distances involved are perturbed by ϵ_n. Sufficient conditions are that the clusters are at least $2\epsilon_n$- apart and $\mathbf{X}_n$ is "sufficiently dense".

Corollary (Cluster consistency)

If I is a stable interval, then for each $a + \epsilon_n \leq r \leq b - \epsilon_n$, the number of clusters in $L_r(n)$ and $\widehat{L}_n(r)$ is the same and constant and each cluster A of $L_n(r)$ is such that $B \subset A \subset B'$, for some B cluster of $\widehat{L}_n(r - \epsilon_n)$ and B' of $\widehat{L}_n(r + \epsilon_n)$.

Stability is a local property and does not hold around at values r of near branching points.

The oracle tree $\mathcal{T}_n$ is close to the algorithmic tree $\mathcal{T}$

- We need clusters to be “well-shaped".

Thickness

A cluster is (γ, c)-thick, where $\gamma > 0$ and $c \geq 0$ if, for any $x \in A$, there exists a $y \in B(x, \gamma) \cap A$ such that

$$B(y, \gamma/(2 + c)) \cap \text{supp}(P) \subset B(x, \gamma) \cap A.$$

- Thickness provides controls how narrow a cluster A can get *compared to clusters at contiguous higher levels in the tree*. It is satisfied for well-behaved manifold if γ is small compared to the reach.

The oracle tree $\mathcal{T}_n$ is close to the algorithmic tree $\mathcal{T}$

- We need clusters to be “well-shaped".

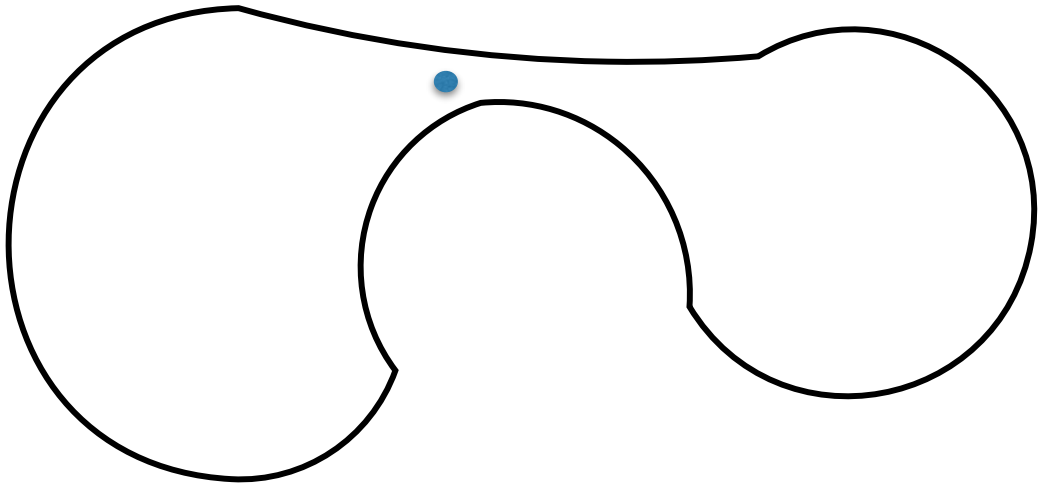
Thickness

A cluster is (γ, c)-thick, where $\gamma > 0$ and $c \geq 0$ if, for any $x \in A$, there exists a $y \in B(x, \gamma) \cap A$ such that

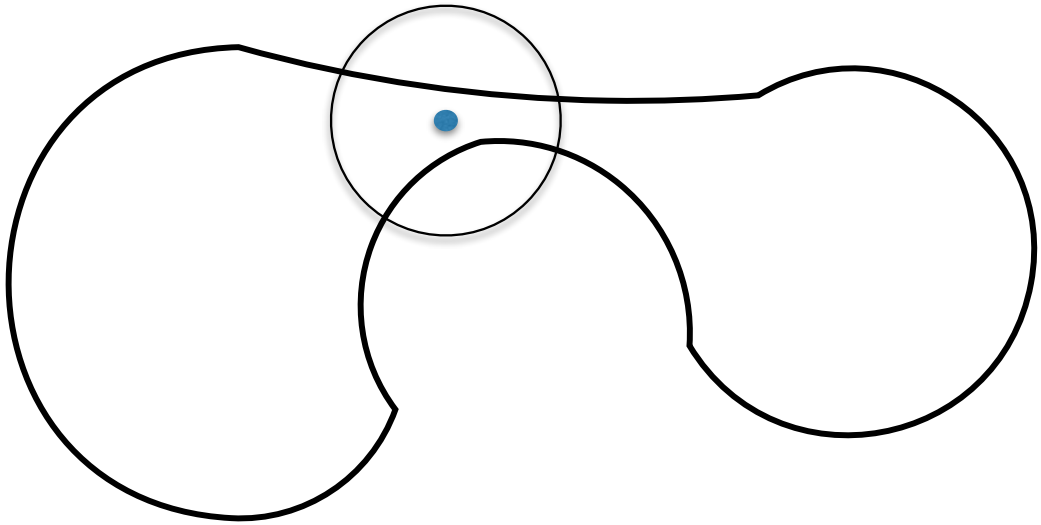
$$B(y, \gamma/(2 + c)) \cap \text{supp}(P) \subset B(x, \gamma) \cap A.$$

- Thickness provides controls how narrow a cluster A can get *compared to clusters at contiguous higher levels in the tree*. It is satisfied for well-behaved manifold if γ is small compared to the reach.

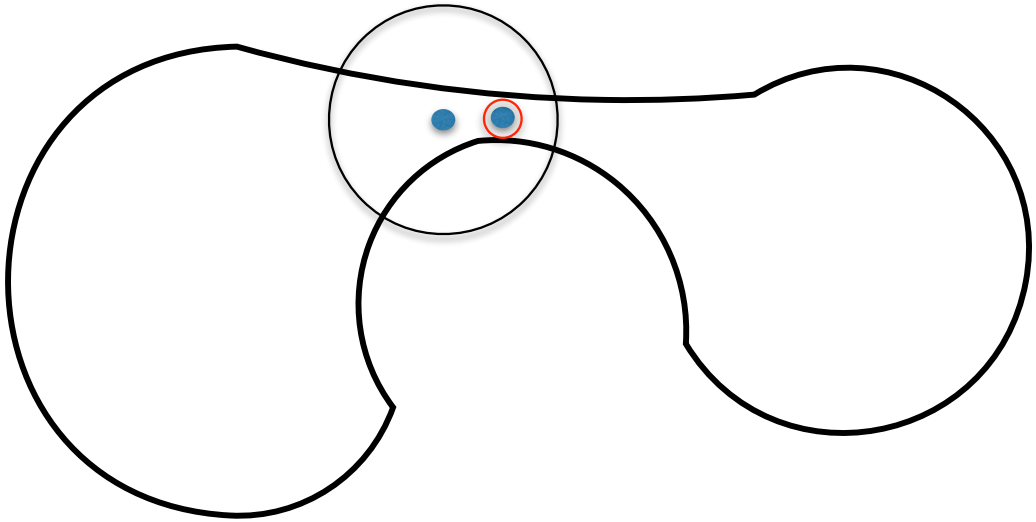
The oracle tree $\mathcal{T}_n$ is close to the algorithmic tree $\mathcal{T}$



The oracle tree $\mathcal{T}_n$ is close to the algorithmic tree $\mathcal{T}$



The oracle tree $\mathcal{T}_n$ is close to the algorithmic tree $\mathcal{T}$



For a cluster A of $\mathcal{T}$, let $r_A = \inf_{x \in A} r_p(x)$.

Consistency

Assume that the sample $\mathbf{X} = (X_1, \ldots, X_n)$ is a γ_n-covering of $\mathrm{supp}(P)$. Then, if A is an r-cluster that is (γ_n, c)-thick, where $\gamma_n < r_A/(2(2+c))$, then $\mathbf{X}_{n,A}$ is a cluster of $L_n(r)$. This holds uniformly over all such clusters.

That is, when the sample is dense enough, the oracle tree will produce the same clustering as the if we knew the algorithmic tree.

When is $\mathbf{X}_n$ a dense covering?

When is $\mathbf{X}_n$ a γ_n-covering of a set $A \subset S$, with $\gamma_n \leq \frac{r_A}{2(c+2)}$?

- Assume A is "well-behaved" (standard assumption) and

$$\inf_x P(B(x, \gamma/2) \cap A) = c_A > 0.$$

Then, $\mathbf{X}_{n,A}$ is a γ covering of A with high probability if

$$c_A = \Omega\left(\frac{\dim(A) + \log n}{n}\right).$$

- If we let $p \to 0$, then $c_A \to 0$ at a rate $\Theta(\gamma^k)$, with $\gamma \to 0$. This would give dimension dependent rate.

When is $\mathbf{X}_n$ a dense covering?

When is $\mathbf{X}_n$ a γ_n-covering of a set $A \subset S$, with $\gamma_n \leq \frac{r_A}{2(c+2)}$?

- Assume A is "well-behaved" (standard assumption) and

$$\inf_x P(B(x, \gamma/2) \cap A) = c_A > 0.$$

Then, $\mathbf{X}_{n,A}$ is a γ covering of A with high probability if

$$c_A = \Omega\left(\frac{\dim(A) + \log n}{n}\right).$$

- If we let $p \to 0$, then $c_A \to 0$ at a rate $\Theta(\gamma^k)$, with $\gamma \to 0$. This would give dimension dependent rate.

When is $\mathbf{X}_n$ a dense covering?

When is $\mathbf{X}_n$ a γ_n-covering of a set $A \subset S$, with $\gamma_n \leq \frac{r_A}{2(c+2)}$?

- Assume A is "well-behaved" (standard assumption) and

$$\inf_x P(B(x, \gamma/2) \cap A) = c_A > 0.$$

Then, $\mathbf{X}_{n,A}$ is a γ covering of A with high probability if

$$c_A = \Omega\left(\frac{\dim(A) + \log n}{n}\right).$$

- If we let $p \to 0$, then $c_A \to 0$ at a rate $\Theta(\gamma^k)$, with $\gamma \to 0$. This would give dimension dependent rate.

Consistency of DeBaCl

Consistency of DeBaCl

Let A be an r-cluster of $\mathcal{T}$ such that its $r - \epsilon_n$ subcluster A' is (γ_n, c)-thick, with $\gamma_n < (r_{A'} - \epsilon_n)/(2 + c)$ and $c \geq 0$. If the sample $\mathbf{X} = (X_1, \ldots, X_n)$ is a γ_n covering of $\mathrm{supp}(P)$ then the subgraph of $\widehat{\mathcal{G}}(r - \epsilon_n)$ induced by $\mathbf{X}_{n,A'}$ is connected; if A is also $2\epsilon_n$-away from all the other r clusters, then $\mathbf{X}_{n,A'}$ is a cluster of $\widehat{\mathcal{T}}_n(r - \epsilon_n)$. This holds uniformly over such clusters of $\mathcal{T}$.

Cluster trees come in many flavors: r, λ, α...and κ

DeBaCloutputs a data structure, which can be visualized in different ways.

- The r-tree: the tree height is indexed by r. Counterintuitive: the tree grows as r gets smaller and the higher portions of the tree are shorter!
- λ-tree: the tree height is indexed by $1/r^d$, a rescaling proportional to the values of the k-nn density estimator.
  The proportions are right but it is not very interpretable.
- α-tree: for each $\alpha \in (0,1)$ set $\widehat{r}_\alpha$ to be the α-quantile of $\widehat{r}_1, \ldots, \widehat{r}_n$. The tree height is indexed by α: the α-level of the tree represents $\widehat{L}(\widehat{r}_\alpha)$, i.e. the α-fraction of "most clusterable" data points. Highly interpretable.
- κ-tree: each branch has length equal to the fraction of points comprising it. The sum of the length of all branches and leaves is 1.
  Highly interpretable.

Cluster trees come in many flavors: r, λ, α...and κ

DeBaCloutputs a data structure, which can be visualized in different ways.

- The r-tree: the tree height is indexed by r. Counterintuitive: the tree grows as r gets smaller and the higher portions of the tree are shorter!
- λ-tree: the tree height is indexed by $1/r^d$, a rescaling proportional to the values of the k-nn density estimator.
  The proportions are right but it is not very interpretable.
- α-tree: for each $\alpha \in (0,1)$ set $\widehat{r}_\alpha$ to be the α-quantile of $\widehat{r}_1, \ldots, \widehat{r}_n$. The tree height is indexed by α: the α-level of the tree represents $\widehat{L}(\widehat{r}_\alpha)$, i.e. the α-fraction of "most clusterable" data points. Highly interpretable.
- κ-tree: each branch has length equal to the fraction of points comprising it. The sum of the length of all branches and leaves is 1.
  Highly interpretable.

Cluster trees come in many flavors: r, λ, α...and κ

DeBaCloutputs a data structure, which can be visualized in different ways.

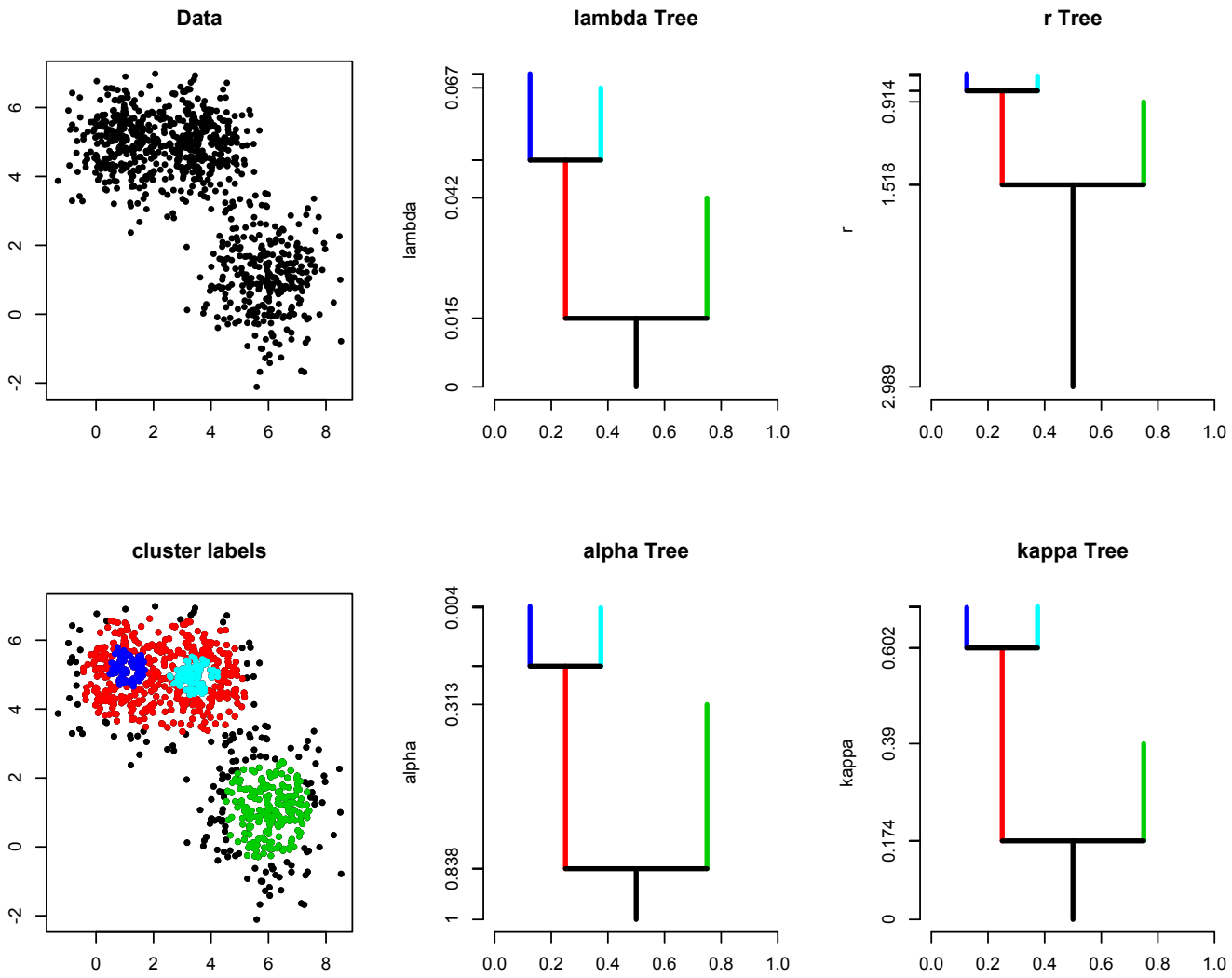
- The r-tree: the tree height is indexed by r. Counterintuitive: the tree grows as r gets smaller and the higher portions of the tree are shorter!
- λ-tree: the tree height is indexed by $1/r^d$, a rescaling proportional to the values of the k-nn density estimator.
  The proportions are right but it is not very interpretable.
- α-tree: for each $\alpha \in (0,1)$ set $\widehat{r}_\alpha$ to be the α-quantile of $\widehat{r}_1, \ldots, \widehat{r}_n$. The tree height is indexed by α: the α-level of the tree represents $\widehat{L}(\widehat{r}_\alpha)$, i.e. the α-fraction of "most clusterable" data points. Highly interpretable.
- κ-tree: each branch has length equal to the fraction of points comprising it. The sum of the length of all branches and leaves is 1.
  Highly interpretable.

Cluster trees come in many flavors: r, λ, α...and κ

DeBaCloutputs a data structure, which can be visualized in different ways.

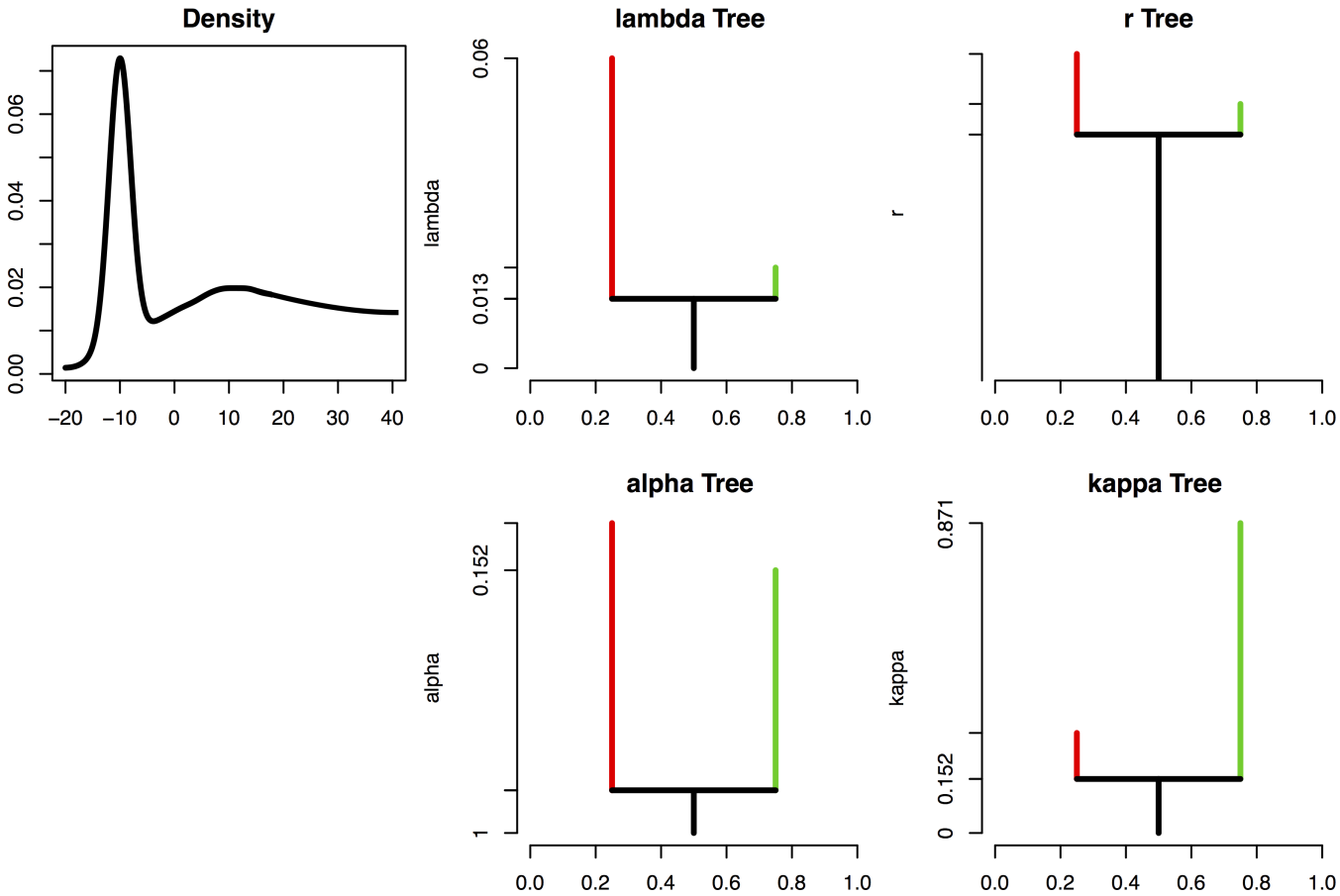
- The r-tree: the tree height is indexed by r. Counterintuitive: the tree grows as r gets smaller and the higher portions of the tree are shorter!
- λ-tree: the tree height is indexed by $1/r^d$, a rescaling proportional to the values of the k-nn density estimator.
  The proportions are right but it is not very interpretable.
- α-tree: for each $\alpha \in (0,1)$ set $\widehat{r}_\alpha$ to be the α-quantile of $\widehat{r}_1, \ldots, \widehat{r}_n$. The tree height is indexed by α: the α-level of the tree represents $\widehat{L}(\widehat{r}_\alpha)$, i.e. the α-fraction of "most clusterable" data points. Highly interpretable.
- κ-tree: each branch has length equal to the fraction of points comprising it. The sum of the length of all branches and leaves is 1.
  Highly interpretable.

Cluster trees come in many flavors: r, λ, α...and κ



Cluster trees come in many flavors: r, λ, α...and κ

- The difference can be substantial.



Weak convergence and the bootstrap

Let $\mathbb{B}$ the P-Brownian bridge indexed by the closed balls in $\mathbb{R}^d$: a centered Gaussian process with covariance function

$$(B, B') \to P(B \cap B') = P(B)P(B').$$

Denote by $\mathbb{B}(x, r)$ the value of $\mathbb{B}$ at the ball $B(x, r)$.

Recall that, for $x \in \mathbb{R}^d$, F_x is the c.d.f. of $\|X - x\|$ with $X \sim P$ and $r_p(x)$ is the p-th quantile of F_x.

Theorem (Function delta method)

Assume that, for some constant $c > 0$ and for P-almost all x,

$$F'_x(r_p(x)) \geq c.$$

Then,

$$\left\{ \sqrt{n} \left(\widehat{r}_p(x) - r_p(x) \right), x \in \mathbb{R}^d \right\} \rightsquigarrow \left\{ \frac{\mathbb{B}(x, r_p(x))}{F'_x(r_p(x))}, x \in \mathbb{R}^d \right\}$$

Weak convergence and the bootstrap

Let $\mathbb{B}$ the P-Brownian bridge indexed by the closed balls in $\mathbb{R}^d$: a centered Gaussian process with covariance function

$$(B, B') \to P(B \cap B') = P(B)P(B').$$

Denote by $\mathbb{B}(x, r)$ the value of $\mathbb{B}$ at the ball $B(x, r)$.

Recall that, for $x \in \mathbb{R}^d$, F_x is the c.d.f. of $\|X - x\|$ with $X \sim P$ and $r_p(x)$ is the p-th quantile of F_x.

Theorem (Function delta method)

Assume that, for some constant $c > 0$ and for P-almost all x,

$$F'_x(r_p(x)) \geq c.$$

Then,

$$\left\{ \sqrt{n} \left(\widehat{r}_p(x) - r_p(x) \right), x \in \mathbb{R}^d \right\} \rightsquigarrow \left\{ \frac{\mathbb{B}(x, r_p(x))}{F'_x(r_p(x))}, x \in \mathbb{R}^d \right\}$$

Weak convergence and the bootstrap

Let $\mathbb{B}$ the P-Brownian bridge indexed by the closed balls in $\mathbb{R}^d$: a centered Gaussian process with covariance function

$$(B, B') \to P(B \cap B') = P(B)P(B').$$

Denote by $\mathbb{B}(x, r)$ the value of $\mathbb{B}$ at the ball $B(x, r)$.

Recall that, for $x \in \mathbb{R}^d$, F_x is the c.d.f. of $\|X - x\|$ with $X \sim P$ and $r_p(x)$ is the p-th quantile of F_x.

Theorem (Function delta method)

Assume that, for some constant $c > 0$ and for P-almost all x,

$$F'_x(r_p(x)) \geq c.$$

Then,

$$\left\{ \sqrt{n} \left(\widehat{r}_p(x) - r_p(x) \right), x \in \mathbb{R}^d \right\} \rightsquigarrow \left\{ \frac{\mathbb{B}(x, r_p(x))}{F'_x(r_p(x))}, x \in \mathbb{R}^d \right\}$$

Weak convergence and the bootstrap

Let $\{\widehat{r}_p^*(x), x \in \mathbb{R}^d\}$ be the bootstrap version of $\widehat{r}_p$ based on the empirical distribution of $\mathbf{X}_n$.

Corollary (Boostrap validity)

Conditionally almost surely,

$$\left\{\sqrt{n}\left(\widehat{r}_p^*(x) - \widehat{r}_p(x)\right), x \in \mathbb{R}^d\right\} \rightsquigarrow \left\{\frac{\mathbb{B}(x, r_p(x))}{F_x'(r_p(x))}, x \in \mathbb{R}^d\right\}$$

Using the bootstrap, we can construct asymptotically correct confidence bands for r_p and the DeBaCltrees.

Weak convergence and the bootstrap

Let $\{\widehat{r}_p^*(x), x \in \mathbb{R}^d\}$ be the bootstrap version of $\widehat{r}_p$ based on the empirical distribution of $\mathbf{X}_n$.

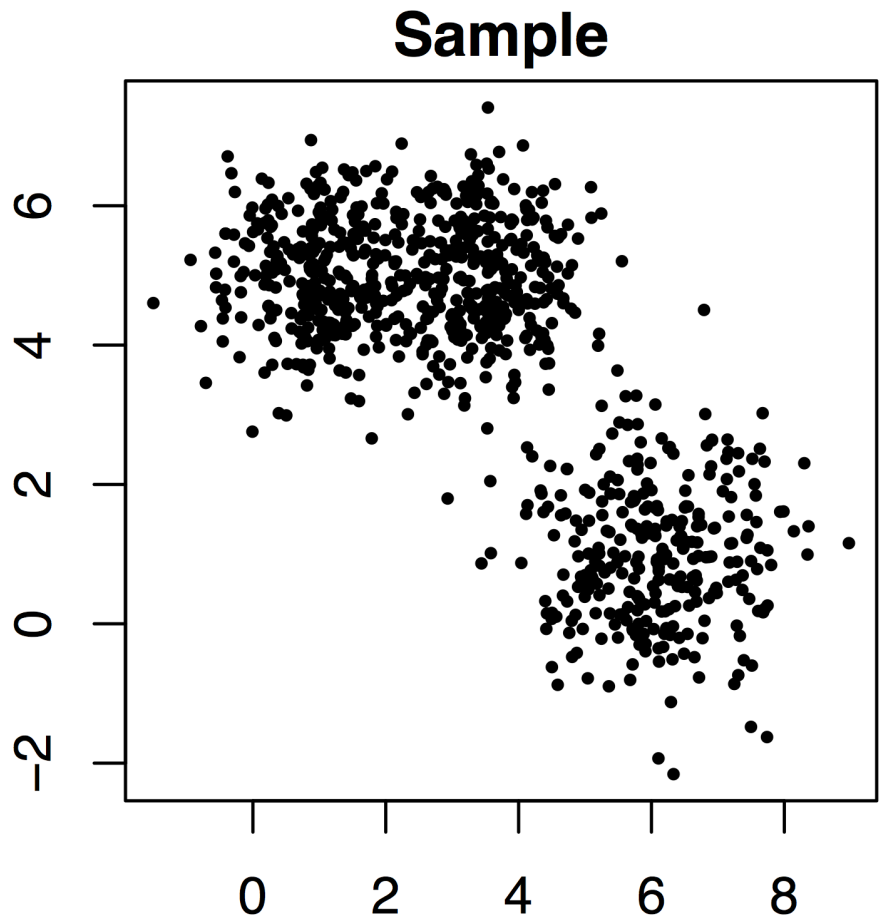
Corollary (Boostrap validity)

Conditionally almost surely,

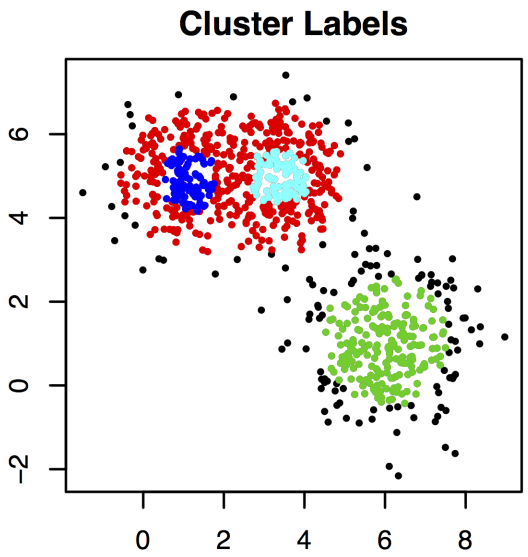
$$\left\{\sqrt{n}\left(\widehat{r}_p^*(x) - \widehat{r}_p(x)\right), x \in \mathbb{R}^d\right\} \rightsquigarrow \left\{\frac{\mathbb{B}(x, r_p(x))}{F_x'(r_p(x))}, x \in \mathbb{R}^d\right\}$$

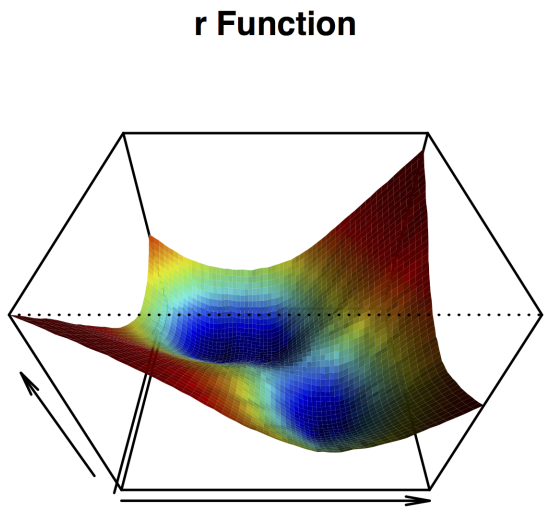
Using the bootstrap, we can construct asymptotically correct confidence bands for r_p and the DeBaCltrees.

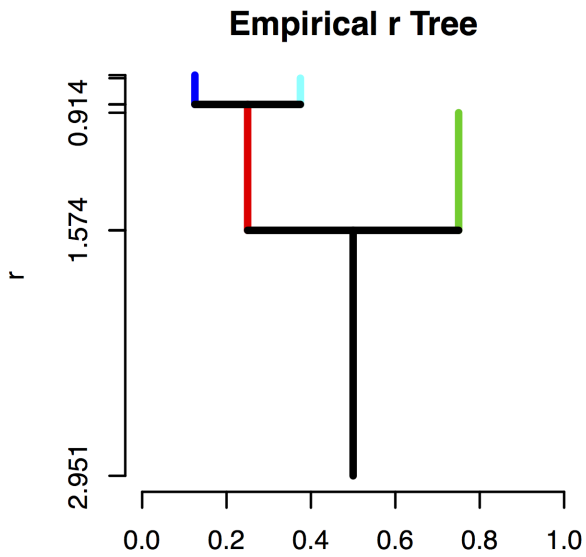
Bootstrap confidence sets example



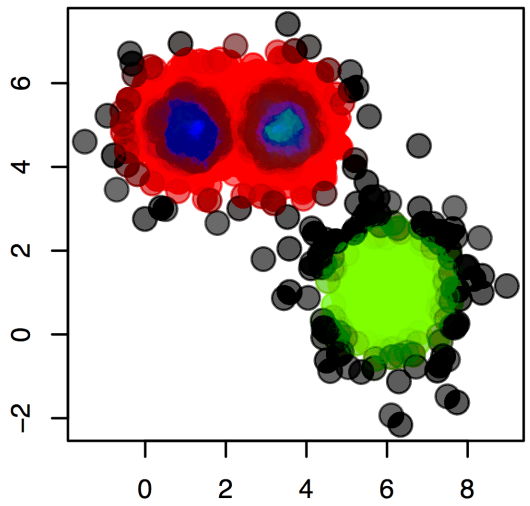
Bootstrap confidence sets example

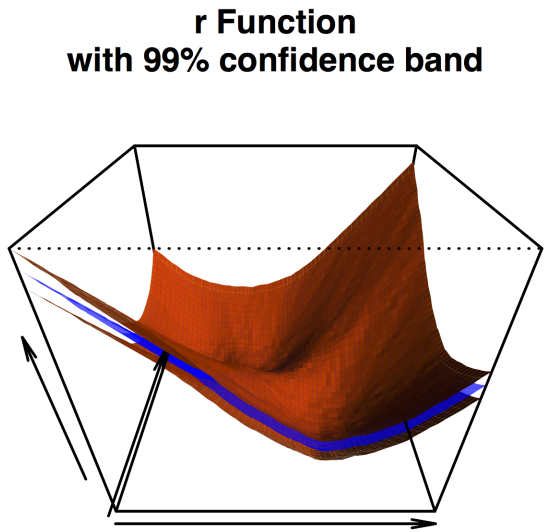


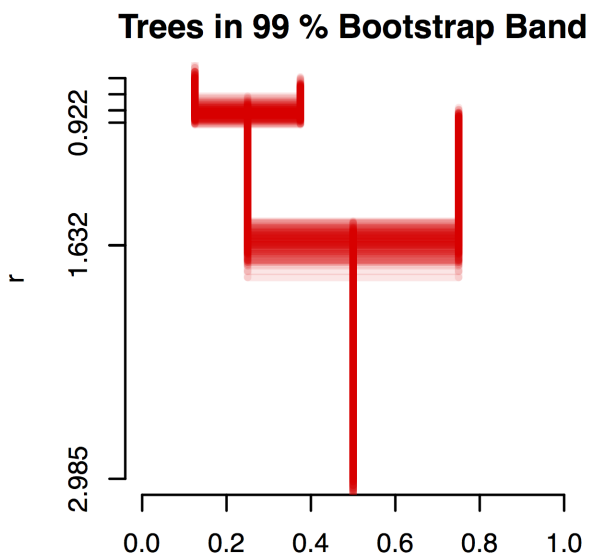




Bootstrap confidence sets example







Conclusions

- Density-based clustering is a principled paradigm for clustering.
- The cluster tree provides an interpretable and highy informative encoding of all the clustering properties of P.
- DeBaClis a simple, computationally efficient algorithm for consistently estimating the cluster tree, even in high dimensions.

Current and future work:

- Work out the theoretical properties of the α and κ tree.
- Develop and study methods for constructing confidence sets for cluster trees and, more generally, for using cluster trees for statistical inference.
- Provide guidelines on how to choose p!

Conclusions

- Density-based clustering is a principled paradigm for clustering.
- The cluster tree provides an interpretable and highy informative encoding of all the clustering properties of P.
- DeBaClis a simple, computationally efficient algorithm for consistently estimating the cluster tree, even in high dimensions.

Current and future work:

- Work out the theoretical properties of the α and κ tree.
- Develop and study methods for constructing confidence sets for cluster trees and, more generally, for using cluster trees for statistical inference.
- Provide guidelines on how to choose p!

CMU TopStat

CMU TopStat

Homepage Projects Papers Reading Group

People

Sivaraman Balakrishnan
Yen-Chi Chen
Jessi Cisewski
Brittany Fasy
Christopher Genovese
Brian Kent
Fabrizio Lecci
Alessandro Rinaldo
Aarti Singh
Isa Verdinelli
Larry Wasserman

The CMU Topological Statistics group is a research group at Carnegie Mellon University. The emphasis of our research is on statistical problems related to topological inference.
Visit the **Projects** page to see descriptions of our projects and relevant publications or preprints.