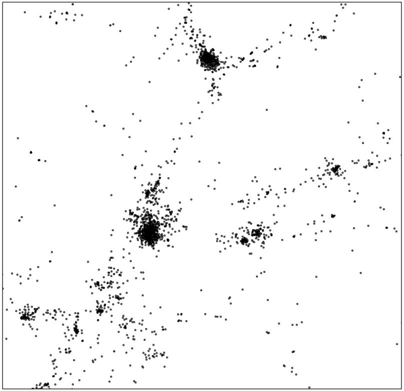
Geometric and Topological Data Analysis

Yen-Chi Chen

Department of Statistics
University of Washington

Geometric and Topological Data Analysis: Big Picture

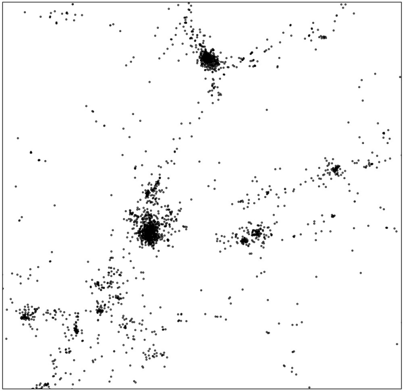


Geometric and Topological Data Analysis: Big Picture

The data can be viewed as

$$X_1, \cdots, X_n \sim p,$$

p is a probability density function.



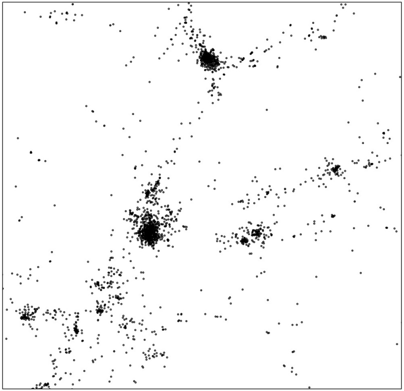
Geometric and Topological Data Analysis: Big Picture

The data can be viewed as

$$X_1, \cdots, X_n \sim p,$$

p is a probability density function.

Scientists are interested in *geometric or topological features* of p.



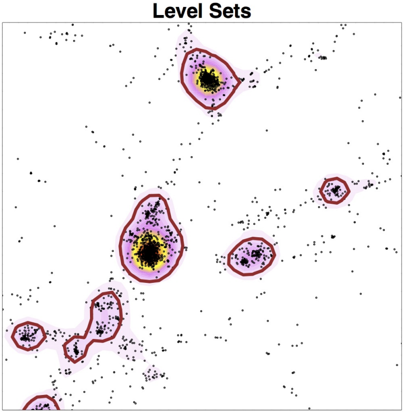
Geometric and Topological Data Analysis: Big Picture

The data can be viewed as

$$X_1, \cdots, X_n \sim p,$$

p is a probability density function.

Scientists are interested in *geometric or topological features* of p.



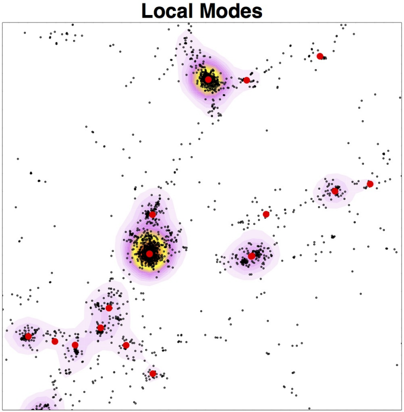
Geometric and Topological Data Analysis: Big Picture

The data can be viewed as

$$X_1, \cdots, X_n \sim p,$$

p is a probability density function.

Scientists are interested in *geometric or topological features* of p.



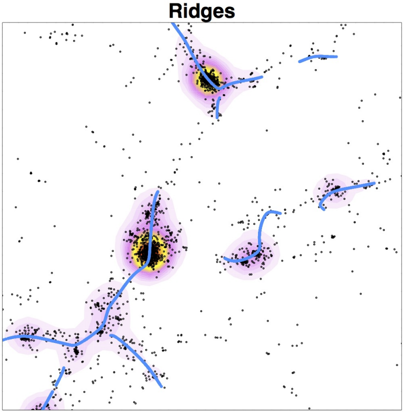
Geometric and Topological Data Analysis: Big Picture

The data can be viewed as

$$X_1, \cdots, X_n \sim p,$$

p is a probability density function.

Scientists are interested in *geometric or topological features* of p.



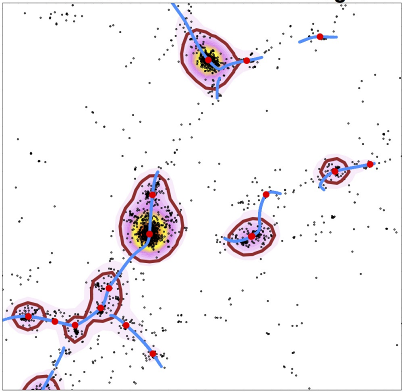
Geometric and Topological Data Analysis: Big Picture

The data can be viewed as

$$X_1, \cdots, X_n \sim p,$$

p is a probability density function.

Scientists are interested in *geometric or topological features* of p.



The Classical Approach

- In all the above examples, how we estimate the geometric/topological structures is based on plug-in estimates from the *kernel density estimator (KDE)*.

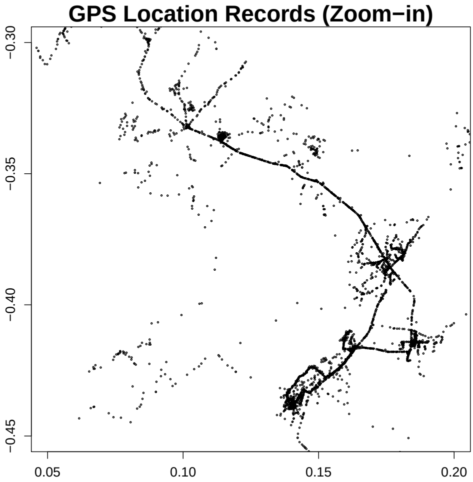
The Classical Approach

- In all the above examples, how we estimate the geometric/topological structures is based on plug-in estimates from the *kernel density estimator (KDE)*.
- Namely, we estimate the probability density function first and then convert it into an estimator of the corresponding structure.

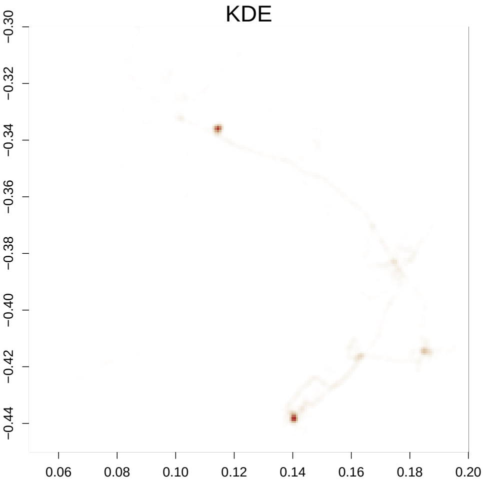
The Classical Approach

- In all the above examples, how we estimate the geometric/topological structures is based on plug-in estimates from the *kernel density estimator (KDE)*.
- Namely, we estimate the probability density function first and then convert it into an estimator of the corresponding structure.
- But this idea may fail.

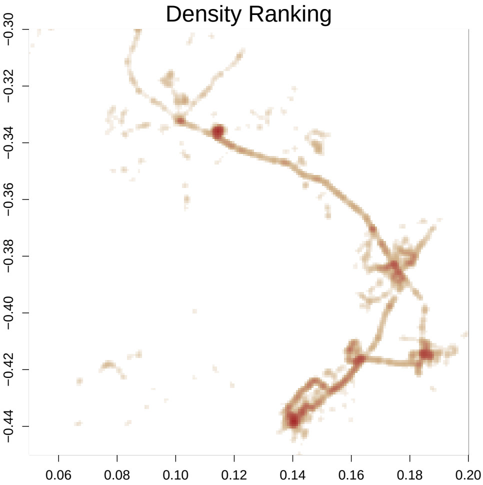
Failure of KDE in Analyzing Data



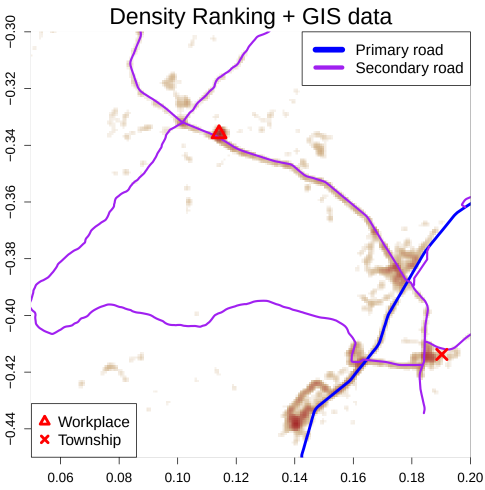
Failure of KDE in Analyzing Data



Failure of KDE in Analyzing Data



Failure of KDE in Analyzing Data



Density Ranking: Introduction

- The KDE cannot detect intricate structures inside the GPS data.
- But the density ranking works!

Density Ranking: Introduction

- The KDE cannot detect intricate structures inside the GPS data.
- But the density ranking works!
- This comes from the fact that the underlying probability density function (PDF) does not exist!
- Namely, our probability distribution function is a singular measure.

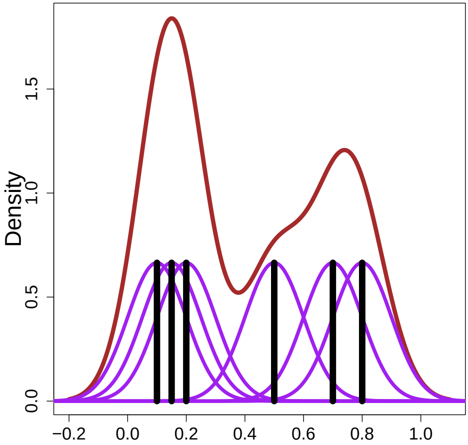
Definition of Density Ranking - 1

- Given random variables $X_1, \cdots, X_n \in \mathbb{R}^d$, the KDE is

$$\widehat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

  where $K(\cdot)$ is called the kernel function such as a Gaussian and $h > 0$ is called the smoothing bandwidth that controls the amount of smoothing.
- The KDE smoothes out the observations into small bumps and sum over all of them to obtain a PDF.

Definition of Density Ranking - 2



Definition of Density Ranking - 3

- The density ranking is a transformed quantity from the KDE.
- Instead of using the density value, we focus on the *ranking* of it.

Definition of Density Ranking - 3

- The density ranking is a transformed quantity from the KDE.
- Instead of using the density value, we focus on the *ranking* of it.
- The formal definition of density ranking is

$$\begin{aligned}\widehat{\alpha}(x) &= \frac{1}{n}\sum_{i=1}^{n} I\left(\widehat{p}(x) \geq \widehat{p}(X_i)\right)\\ &= \text{ratio of observations' density below the density of point } x.\end{aligned}$$

Definition of Density Ranking - 3

- The density ranking is a transformed quantity from the KDE.
- Instead of using the density value, we focus on the *ranking* of it.
- The formal definition of density ranking is

$$\begin{aligned}\widehat{\alpha}(x) &= \frac{1}{n}\sum_{i=1}^{n} I\left(\widehat{p}(x) \geq \widehat{p}(X_i)\right)\\ &= \text{ratio of observations' density below the density of point } x.\end{aligned}$$

- Namely, $\widehat{\alpha}(x) = 0.3$ implies that the (estimated) density of point x is above the (estimated) density of 30% of all observations.

Property of Density Ranking

- For an observation $X_{\max}$ with $\widehat{\alpha}(X_{\max}) = 1$, then it means

$$\widehat{p}(X_{\max}) = \max\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

Property of Density Ranking

- For an observation $X_{\max}$ with $\widehat{\alpha}(X_{\max}) = 1$, then it means

$$\widehat{p}(X_{\max}) = \max\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

- Similarly, for an observation $X_{\min}$ with $\widehat{\alpha}(X_{\min}) = 0$,

$$\widehat{p}(X_{\min}) = \min\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

Property of Density Ranking

- For an observation $X_{\max}$ with $\widehat{\alpha}(X_{\max}) = 1$, then it means

$$\widehat{p}(X_{\max}) = \max\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

- Similarly, for an observation $X_{\min}$ with $\widehat{\alpha}(X_{\min}) = 0$,

$$\widehat{p}(X_{\min}) = \min\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

- If an observation X_ℓ satisfies $\widehat{\alpha}(X_\ell) = 0.25$, this means that the ranking of density at X_ℓ is the 25%.

Property of Density Ranking

- For an observation $X_{\max}$ with $\widehat{\alpha}(X_{\max}) = 1$, then it means

$$\widehat{p}(X_{\max}) = \max\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

- Similarly, for an observation $X_{\min}$ with $\widehat{\alpha}(X_{\min}) = 0$,

$$\widehat{p}(X_{\min}) = \min\left\{\widehat{p}(X_1), \cdots, \widehat{p}(X_n)\right\}.$$

- If an observation X_ℓ satisfies $\widehat{\alpha}(X_\ell) = 0.25$, this means that the ranking of density at X_ℓ is the 25%.
- Moreover, for any pairs of points x_1, x_2,

$$\begin{aligned}
\widehat{p}(x_1) > \widehat{p}(x_2) &\Longrightarrow \widehat{\alpha}(x_1) > \widehat{\alpha}(x_2)\\
\widehat{p}(x_1) < \widehat{p}(x_2) &\Longrightarrow \widehat{\alpha}(x_1) < \widehat{\alpha}(x_2)\\
\widehat{p}(x_1) = \widehat{p}(x_2) &\Longrightarrow \widehat{\alpha}(x_1) = \widehat{\alpha}(x_2)
\end{aligned}$$

Density Ranking as an Estimator

- Density ranking $\widehat{\alpha}(x)$ can be viewed as an estimator to a function of the underlying population distribution.
- When the distribution function has a PDF, the population version of density ranking is defined as follows.

Density Ranking as an Estimator

- Density ranking $\widehat{\alpha}(x)$ can be viewed as an estimator to a function of the underlying population distribution.
- When the distribution function has a PDF, the population version of density ranking is defined as follows.
- Assume $X_1, \cdots, X_n$ is a random sample from an unknown distribution function P with a PDF p.

Density Ranking as an Estimator

- Density ranking $\widehat{\alpha}(x)$ can be viewed as an estimator to a function of the underlying population distribution.
- When the distribution function has a PDF, the population version of density ranking is defined as follows.
- Assume $X_1, \cdots, X_n$ is a random sample from an unknown distribution function P with a PDF p.
- Then the population version of $\widehat{\alpha}(x)$ is

$$\alpha(x) = P(p(x) \geq p(X_1)).$$

Density Ranking as an Estimator

- Density ranking $\widehat{\alpha}(x)$ can be viewed as an estimator to a function of the underlying population distribution.
- When the distribution function has a PDF, the population version of density ranking is defined as follows.
- Assume $X_1, \cdots, X_n$ is a random sample from an unknown distribution function P with a PDF p.
- Then the population version of $\widehat{\alpha}(x)$ is

$$\alpha(x) = P(p(x) \geq p(X_1)).$$

- Under regularity conditions,

$$\int \left|\widehat{\alpha}(x) - \alpha(x)\right|^2 dP(x) \xrightarrow{P} 0, \quad \sup_x \left|\widehat{\alpha}(x) - \alpha(x)\right| \xrightarrow{P} 0.$$

Density Ranking in Singular Measure

- Why density ranking works in GPS data but KDE fails is probably due to the fact that density ranking is a consistent estimator *even when the density does not exist!*

Density Ranking in Singular Measure

- Why density ranking works in GPS data but KDE fails is probably due to the fact that density ranking is a consistent estimator *even when the density does not exist!*
- To generalize population density ranking to a singular measure, we introduce the concept of *geometric density*.

Density Ranking in Singular Measure

- Why density ranking works in GPS data but KDE fails is probably due to the fact that density ranking is a consistent estimator *even when the density does not exist!*
- To generalize population density ranking to a singular measure, we introduce the concept of *geometric density*.
- Let C_d be the volume of a d dimensional ball and $B(x,r) = \{y : \|x-y\| \leq r\}$.

Density Ranking in Singular Measure

- Why density ranking works in GPS data but KDE fails is probably due to the fact that density ranking is a consistent estimator *even when the density does not exist!*
- To generalize population density ranking to a singular measure, we introduce the concept of *geometric density*.
- Let C_d be the volume of a d dimensional ball and $B(x,r) = \{y : \|x-y\| \leq r\}$.
- For any integer s, we define

$$\mathcal{H}_s(x) = \lim_{r\to 0} \frac{P(B(x,r))}{C_s r^s}.$$

Density Ranking in Singular Measure

- Why density ranking works in GPS data but KDE fails is probably due to the fact that density ranking is a consistent estimator *even when the density does not exist!*
- To generalize population density ranking to a singular measure, we introduce the concept of *geometric density*.
- Let C_d be the volume of a d dimensional ball and $B(x,r) = \{y : \|x-y\| \le r\}$.
- For any integer s, we define

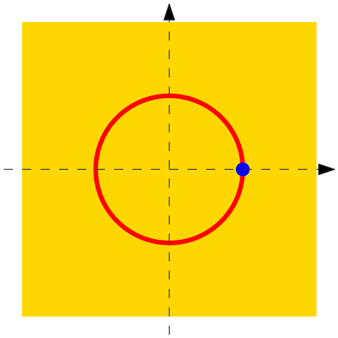
$$\mathcal{H}_s(x) = \lim_{r\to 0} \frac{P(B(x,r))}{C_s r^s}.$$

- For a point x, we then define

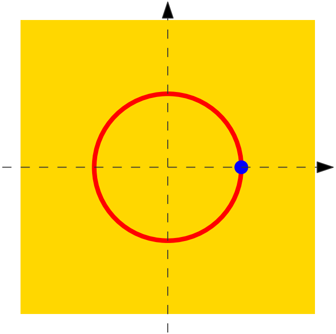
$$\tau(x) = \max\{s \le d : \mathcal{H}_s(x) < \infty\}, \quad \rho(x) = \mathcal{H}_{\tau(x)}(x).$$

Geometric Density: Example - 1

- Assume the distribution function P is a mixture of a $2D$ uniform distribution within $[-1, 1]^2$, a $1D$ uniform distribution over the ring $\{(x, y) : x^2 + y^2 = 0.5^2\}$, and a point mass at $(0.5, 0)$, then the support can be partitioned as follows:



Geometric Density: Example - 2



- Orange region: $\tau(x) = 2$.
- Red region: $\tau(x) = 1$.
- Blue region: $\tau(x) = 0$.

Geometric Density and Ranking

- The function $\tau(x)$ measures the dimension of P at point x.
- We can then use τ and ρ to compare any pairs of points and construct a ranking.

Geometric Density and Ranking

- The function $\tau(x)$ measures the dimension of P at point x.
- We can then use τ and ρ to compare any pairs of points and construct a ranking.
- For two points x_1, x_2, we define an ordering such that $x_1 \succ_{\tau,\rho} x_2$ if

$$\tau(x_1) < \tau(x_2), \qquad \text{or} \quad \tau(x_1) = \tau(x_2), \quad \rho(x_1) > \rho(x_2).$$

Geometric Density and Ranking

- The function $\tau(x)$ measures the dimension of P at point x.
- We can then use τ and ρ to compare any pairs of points and construct a ranking.
- For two points x_1, x_2, we define an ordering such that $x_1 \succ_{\tau,\rho} x_2$ if

$$\tau(x_1) < \tau(x_2), \qquad \text{or} \quad \tau(x_1) = \tau(x_2), \quad \rho(x_1) > \rho(x_2).$$

- Namely, we first compare the dimension of the two points, the lower dimensional structure wins. If they are on regions of the same dimension, we then compare the density of that dimension.

Constructing Density Ranking using Geometric Density

- Using the ordering $\succ_{\tau,\rho}$, we then define the population density ranking as

$$\alpha(x) = P(x \succeq_{\tau,\rho} X_1)$$

Constructing Density Ranking using Geometric Density

- Using the ordering $\succ_{\tau,\rho}$, we then define the population density ranking as
$$\alpha(x) = P(x \succeq_{\tau,\rho} X_1)$$
- When the PDF exists, the ordering $\succ_{\tau,\rho}$ equals to $\succ_{d,p}$ so
$$\alpha(x) = P(x \succeq_{d,p} X_1) = P(p(x) \geq p(X_1)),$$
which recovers our original definition.

Convergence under Singular Measure

- When P is a singular distribution and satisfies certain regularity conditions,

$$\int \left|\widehat{\alpha}(x) - \alpha(x)\right|^2 dP(x) \xrightarrow{P} 0$$

but no guarantee for the convergence of $\sup_x \left|\widehat{\alpha}(x) - \alpha(x)\right|$.

Convergence under Singular Measure

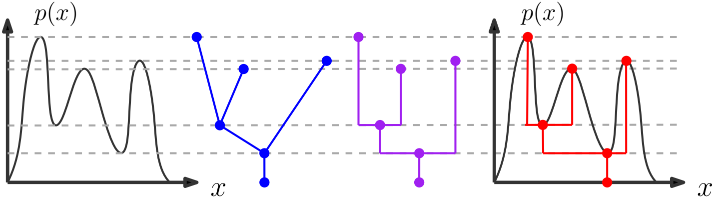
- When P is a singular distribution and satisfies certain regularity conditions,

$$\int \left|\widehat{\alpha}(x) - \alpha(x)\right|^2 dP(x) \xrightarrow{P} 0$$

  but no guarantee for the convergence of $\sup_x \left|\widehat{\alpha}(x) - \alpha(x)\right|$.
- Example of non-convergence of supreme norm: points very close to a lower dimensional structure will not converge.

Density Ranking and Cluster Tree - 1

- Cluster tree is a technique to summarize a function using a tree.
- When the PDF exists, the cluster tree of a PDF and the cluster tree of the corresponding density ranking has the same tree topology.



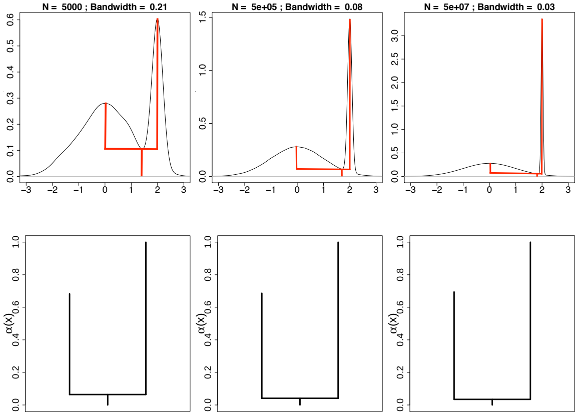
- The idea of building a cluster tree of a function f relies on matching the connecting components of level sets $\{x : f(x) \geq \lambda\}$ when we vary the level λ.

Density Ranking and Cluster Tree - 2

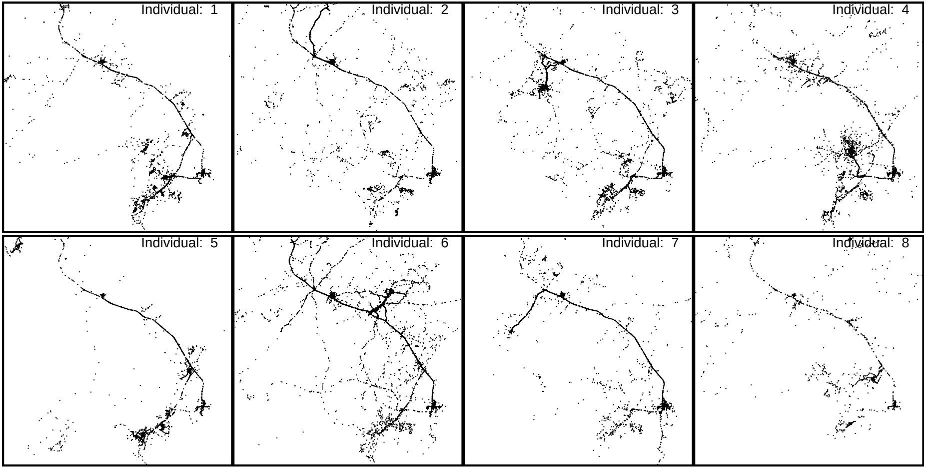
- Using the level sets of $\widehat{\alpha}(x)$ or $\alpha(x)$, we can define the cluster tree of the density ranking and the population density ranking.
- When the distribution function is singular and satisfies certain regularity conditions, the cluster tree of $\widehat{\alpha}(x)$ converges to the cluster tree of $\alpha(x)$.

Density Ranking and Cluster Tree: Example

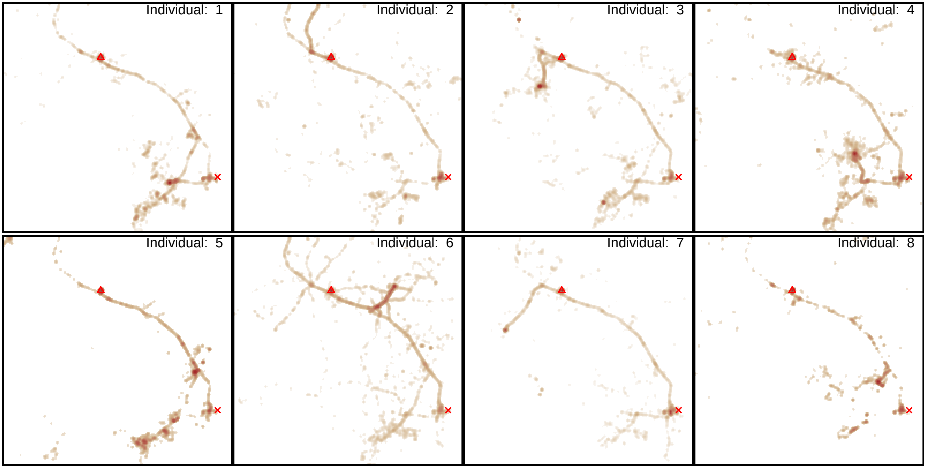
Here the population distribution function is a mixture of a $1D$ standard normal distribution and a point mass at 2. We consider three sample sizes: $n = 5 \times 10^3, 5 \times 10^5, 5 \times 10^7$.



Application of Density Ranking: GPS dataset - 1



Application of Density Ranking: GPS dataset - 2



Summarizing Multiple Density Ranking: Level Plots

- In the above example, we have multiple GPS datasets that lead to multiple density ranking.
- To compare these density rankings, a simple approach is to overlap level plots.

Summarizing Multiple Density Ranking: Level Plots

- In the above example, we have multiple GPS datasets that lead to multiple density ranking.
- To compare these density rankings, a simple approach is to overlap level plots.
- For a density ranking $\widehat{\alpha}$, let

$$\widehat{A}_\gamma = \{x : \widehat{\alpha}(x) \geq 1 - \gamma\}$$

  be the (upper) level set.

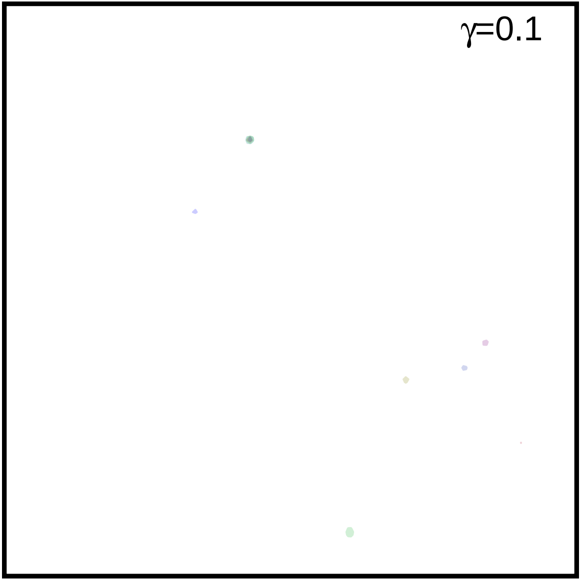
Summarizing Multiple Density Ranking: Level Plots

- In the above example, we have multiple GPS datasets that lead to multiple density ranking.
- To compare these density rankings, a simple approach is to overlap level plots.
- For a density ranking $\widehat{\alpha}$, let

$$\widehat{A}_\gamma = \{x : \widehat{\alpha}(x) \geq 1 - \gamma\}$$

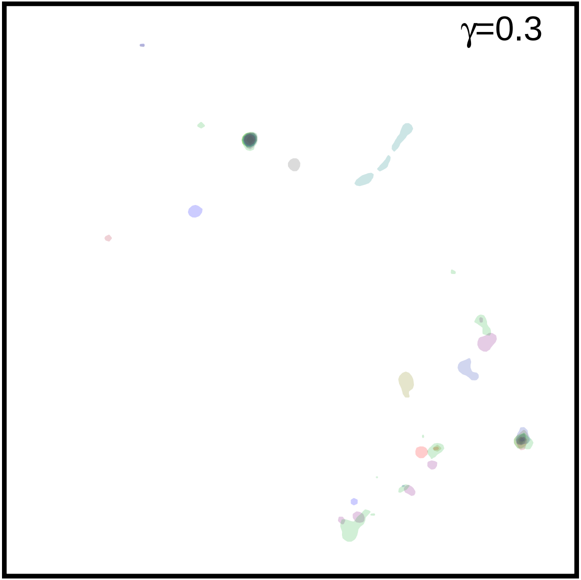
  be the (upper) level set.
- We can compare the density ranking of each individual by overlapping their level sets at each level.

Level Plots: Example

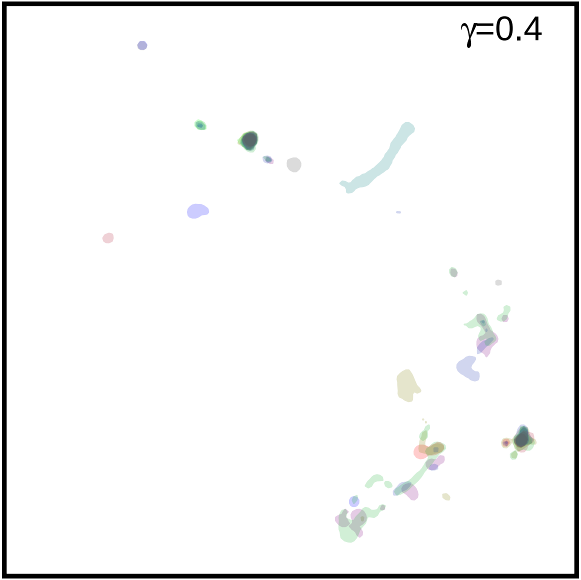


Level Plots: Example

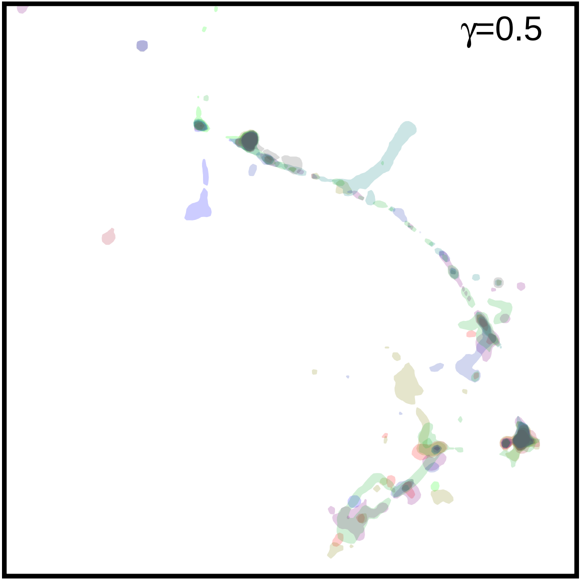
Level Plots: Example



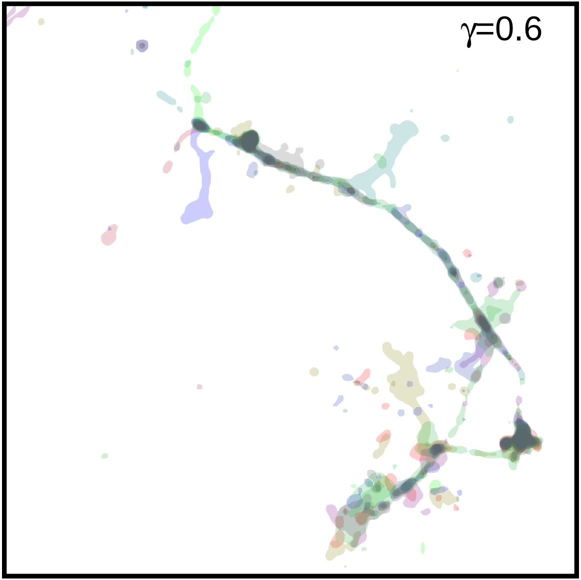
Level Plots: Example



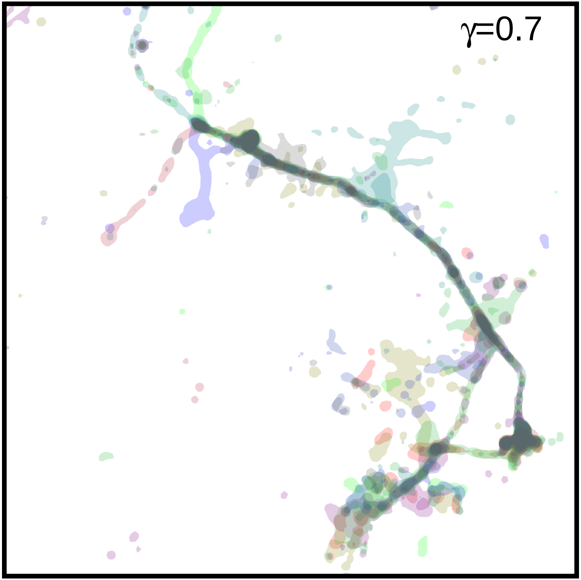
Level Plots: Example



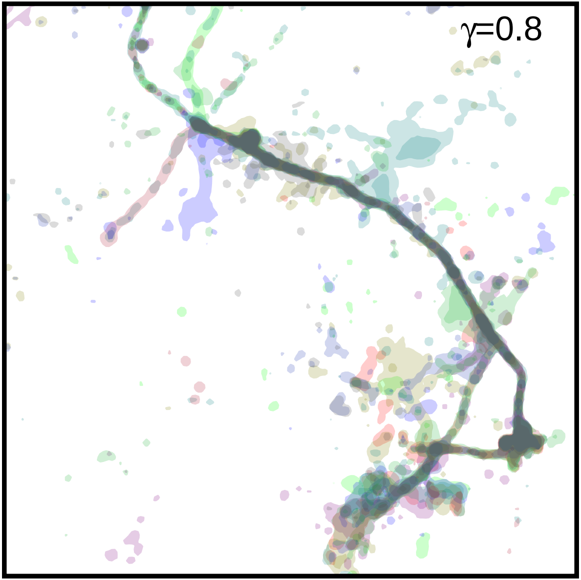
Level Plots: Example



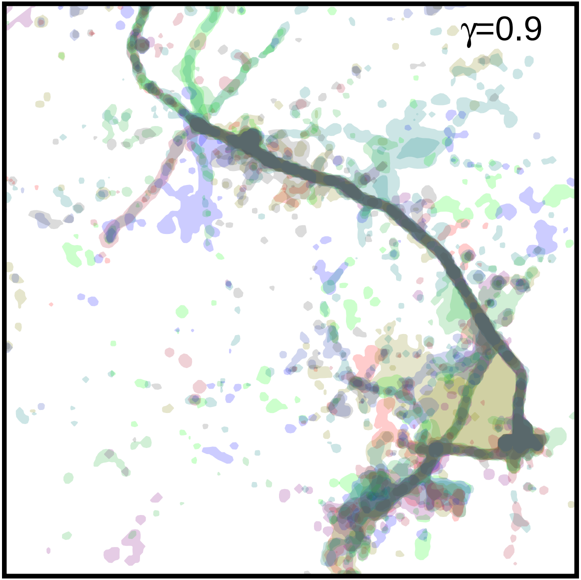
Level Plots: Example

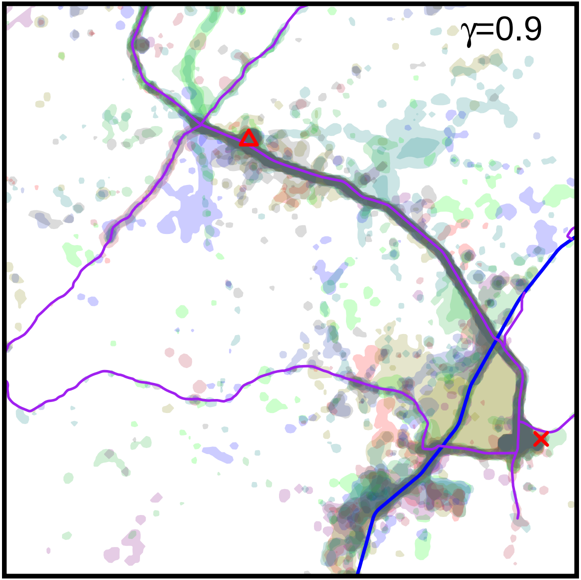


Level Plots: Example



Level Plots: Example




γ=0.9

Summary Curves of Density Ranking

- The level plot allows us to compare GPS datasets from different individuals.

Summary Curves of Density Ranking

- The level plot allows us to compare GPS datasets from different individuals.
- However, it has two drawbacks:
  - When we have more individuals, this approach might not work (too many contours).
  - We often need to choose a level γ to show the plot but which level to be chosen is unclear.

Summary Curves of Density Ranking

- The level plot allows us to compare GPS datasets from different individuals.
- However, it has two drawbacks:
  - When we have more individuals, this approach might not work (too many contours).
  - We often need to choose a level γ to show the plot but which level to be chosen is unclear.
- Here we introduce a few curves to summarize geometric and topological features of density ranking.

Mass-Volume Curve

- Recall that $\widehat{A}_\gamma = \{x : \widehat{\alpha}(x) \geq 1 - \gamma\}$ is the level set of density ranking.

Mass-Volume Curve

- Recall that $\widehat{A}_\gamma = \{x : \widehat{\alpha}(x) \geq 1 - \gamma\}$ is the level set of density ranking.
- The mass-volume curve is a curve of

$$\left(\gamma, \mathsf{Vol}(\widehat{A}_\gamma)\right) : \gamma \in [0, 1].$$

- Namely, we are plotting the size of set $\widehat{A}_\gamma$ at various level.

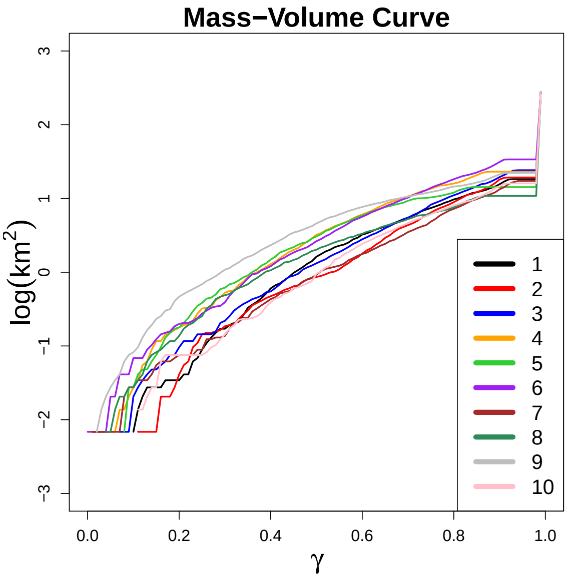
Mass-Volume Curve

- Recall that $\widehat{A}_\gamma = \{x : \widehat{\alpha}(x) \geq 1 - \gamma\}$ is the level set of density ranking.
- The mass-volume curve is a curve of

$$\left(\gamma, \mathsf{Vol}(\widehat{A}_\gamma)\right) : \gamma \in [0, 1].$$

- Namely, we are plotting the size of set $\widehat{A}_\gamma$ at various level.
- In practice, we often plot γ versus $\log \mathsf{Vol}(\widehat{\alpha})_\gamma$.

Mass-Volume Curve: Example



Betti Number Curve

- The Betti number curve is a curve quantifying topological features of the density ranking.
- It counts the number of connected components of $\widehat{A}_\gamma$ at various level γ.

Betti Number Curve

- The Betti number curve is a curve quantifying topological features of the density ranking.
- It counts the number of connected components of $\widehat{A}_\gamma$ at various level γ.
- Formally, the Betti number curve is

$$\left(\gamma, \mathsf{Betti}_0(\widehat{A}_\gamma)\right) : \gamma \in [0,1],$$

where for a set A

$$\mathsf{Betti}_0(A) = \text{ number of connected components inside } A.$$

Betti Number Curve

- The Betti number curve is a curve quantifying topological features of the density ranking.
- It counts the number of connected components of $\widehat{A}_\gamma$ at various level γ.
- Formally, the Betti number curve is

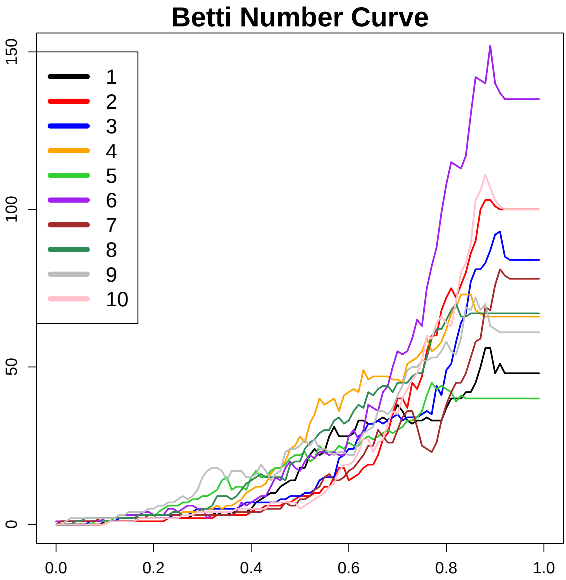
$$\left(\gamma, \mathsf{Betti}_0(\widehat{A}_\gamma)\right) : \gamma \in [0,1],$$

where for a set A

$$\mathsf{Betti}_0(A) = \text{ number of connected components inside } A.$$

- Note that the number of connected component is called the 0th order Betti number (0th order topological structure); one can generalize this idea to higher order topological structures.

Betti Number Curve: Example



Density Ranking: Open Questions

- Convergence of density ranking level sets.
- Convergence of summary curves under singular/non-singular measure.
- Other summary curves.
- Convergence of higher order topological structures.
- Connection to stratified space.