2019 NCTS & Sinica Summer Course: Nonparametric Statistics and Geometric Estimation

Yen-Chi Chen

Department of Statistics
University of Washington

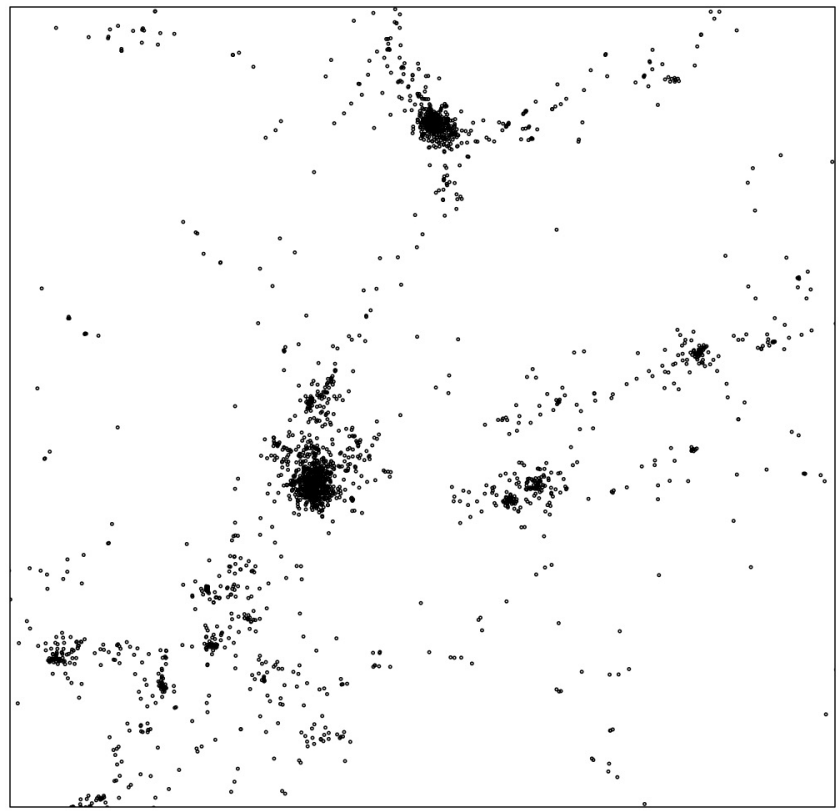
Summer 2019

Introduction

Introduction to Geometric Estimation

- Geometric estimation studies the problem of estimating a geometric feature of a function (of interest).
- Often this function is the underlying probability density function (PDF) that generates our data.
- Other the functions of interest in statistics: the regression function, the difference between two densities/regression functions, conditional probability of an event.

What are Geometric Features? - An Astronomy Example

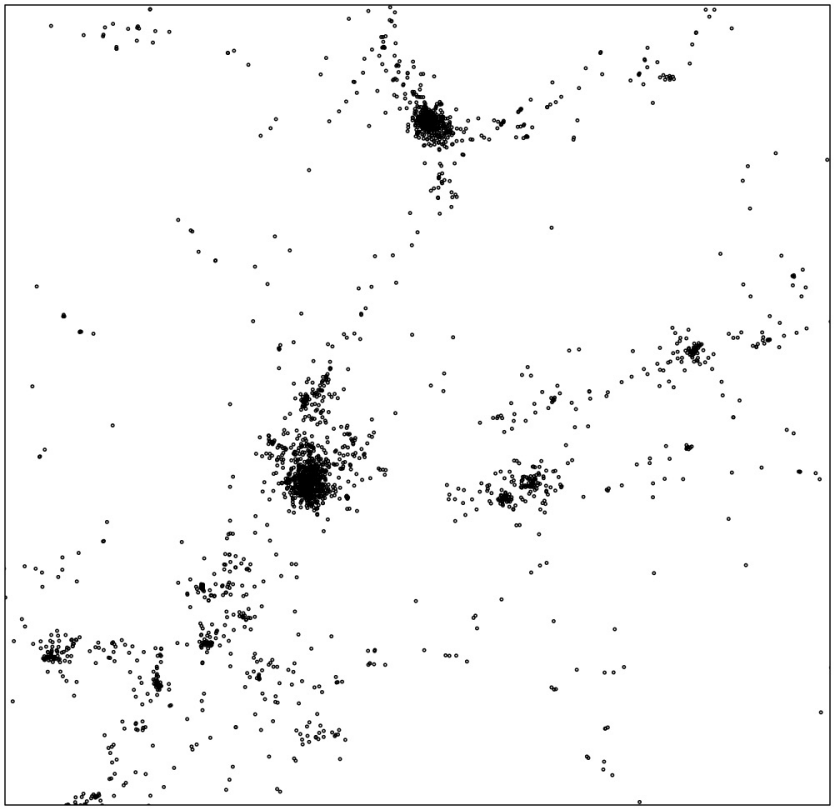


What are Geometric Features? - An Astronomy Example

The data can be viewed as

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

p is a probability density function.



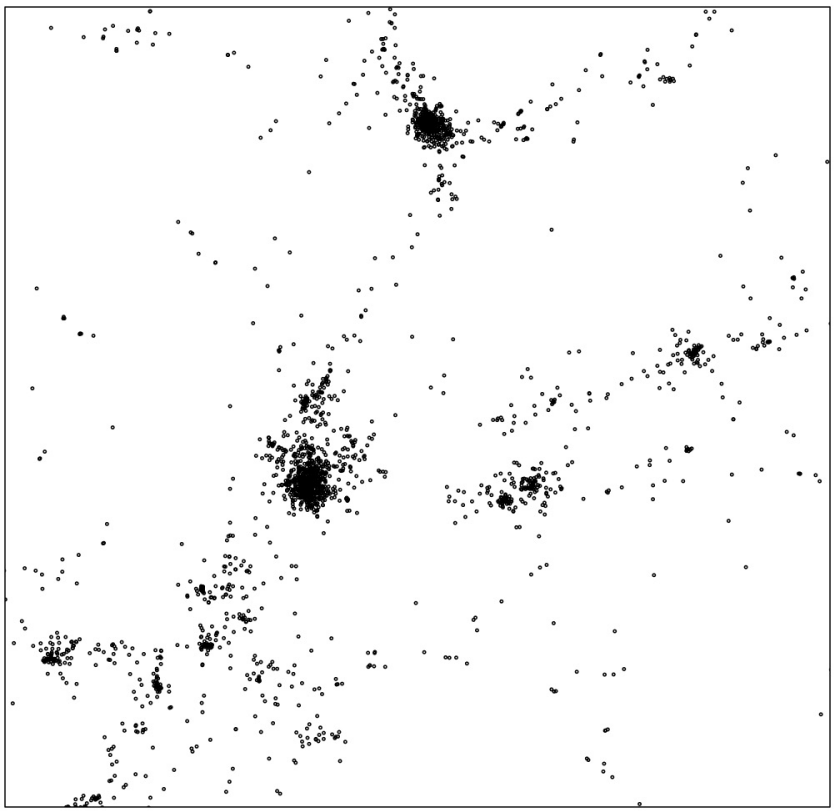
What are Geometric Features? - An Astronomy Example

The data can be viewed as

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

p is a probability density function.

Scientists are interested in *geometric features* of p.



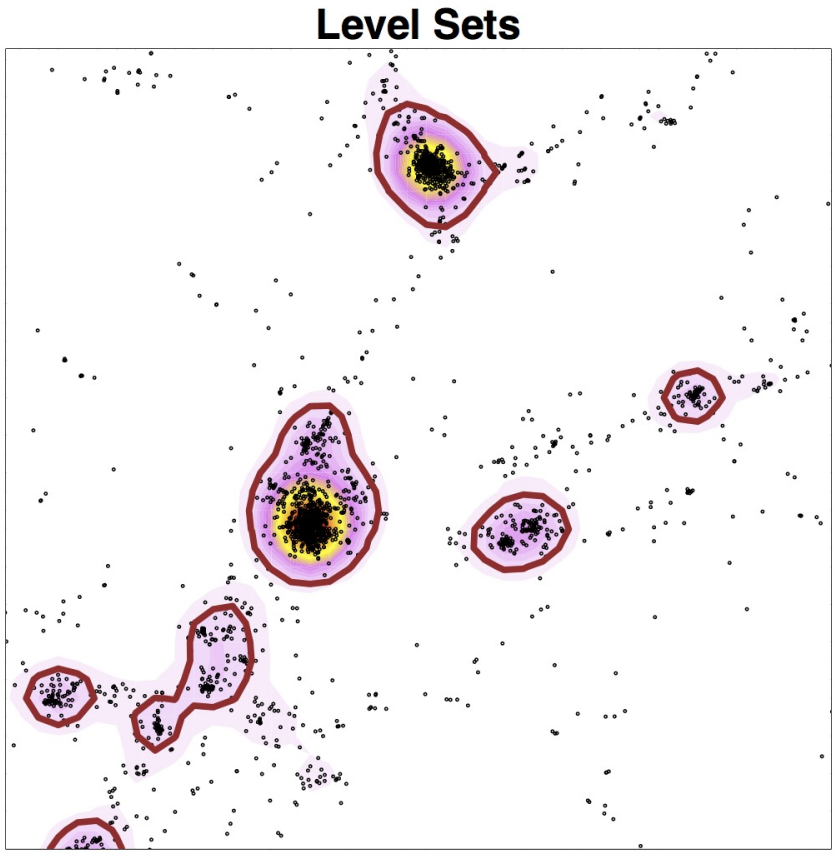
What are Geometric Features? - An Astronomy Example

The data can be viewed as

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

p is a probability density function.

Scientists are interested in *geometric features* of p.



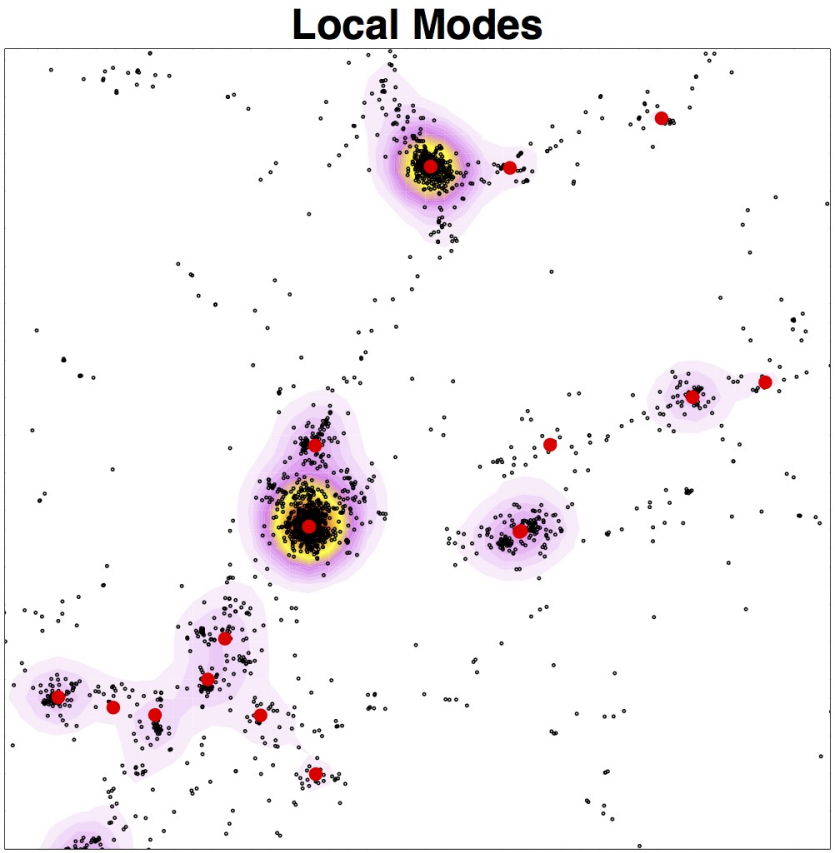
What are Geometric Features? - An Astronomy Example

The data can be viewed as

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

p is a probability density function.

Scientists are interested in *geometric features* of p.



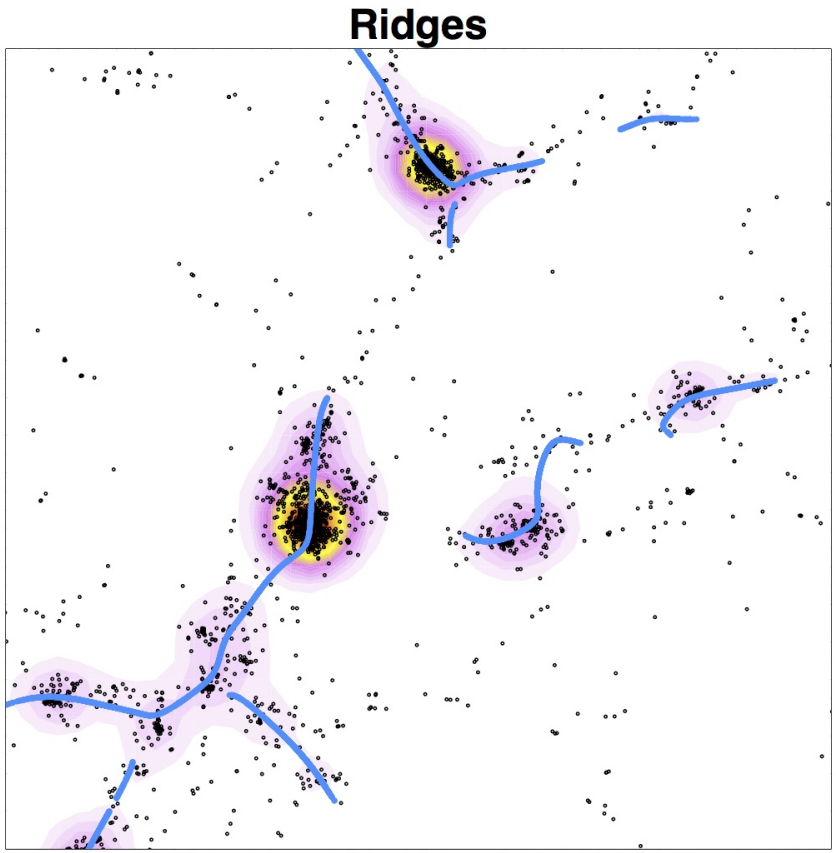
What are Geometric Features? - An Astronomy Example

The data can be viewed as

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

p is a probability density function.

Scientists are interested in *geometric features* of p.



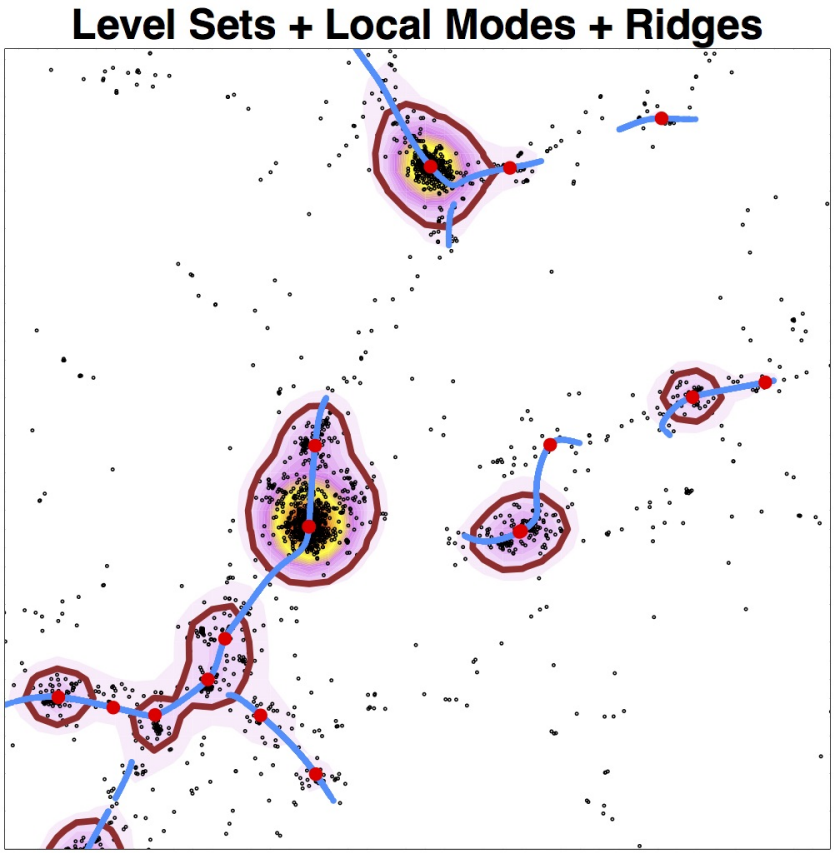
What are Geometric Features? - An Astronomy Example

The data can be viewed as

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

p is a probability density function.

Scientists are interested in *geometric features* of p.



Nonparametric Density Estimation

Density Estimation: Introduction

- A statistical model views the data as random variables $X_1, \cdots, X_n$ from an unknown distribution function $P(x)$ with a PDF $p(x)$.

Density Estimation: Introduction

- A statistical model views the data as random variables $X_1, \cdots, X_n$ from an unknown distribution function $P(x)$ with a PDF $p(x)$.
- In most cases, we do not know the PDF $p(x)$ but we want to reconstruct it from the data.
- The goal of density estimation is to *estimate* $p(x)$ *using* $X_1, \cdots, X_n$.
- In other words, the parameter of interest is the PDF $p(x)$.

Nonparametric Approach: Introduction

- A common approach to estimate $p(x)$ is to assume a parametric model such as a Gaussian and recover the parameters of the model by fitting to the data.
- However, this idea is often either too restrictive to capture the intricate structure of the PDF or computationally infeasible (in the case of mixture models).

Nonparametric Approach: Introduction

- A common approach to estimate $p(x)$ is to assume a parametric model such as a Gaussian and recover the parameters of the model by fitting to the data.
- However, this idea is often either too restrictive to capture the intricate structure of the PDF or computationally infeasible (in the case of mixture models).
- An alternative approach is to estimate the PDF nonparametrically.
- Namely, we directly estimate the PDF without assuming a parametric form of the PDF.

Kernel Density Estimation - 1

- In this lecture, we will focus on one particular nonparametric estimator–the kernel density estimator (KDE).

Kernel Density Estimation - 1

- In this lecture, we will focus on one particular nonparametric estimator–the kernel density estimator (KDE).
- The KDE estimates the PDF using the following form:

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

where $K(x)$ is a function called the kernel function and $h > 0$ is a quantity called smoothing bandwidth that controls the amount of smoothing.

Kernel Density Estimation - 1

- In this lecture, we will focus on one particular nonparametric estimator–the kernel density estimator (KDE).
- The KDE estimates the PDF using the following form:

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

  where $K(x)$ is a function called the kernel function and $h > 0$ is a quantity called smoothing bandwidth that controls the amount of smoothing.
- Common choice of $K(x)$ includes the Gaussian $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and the uniform $K(x) = \frac{1}{2} I(-1 \leq x \leq 1)$.

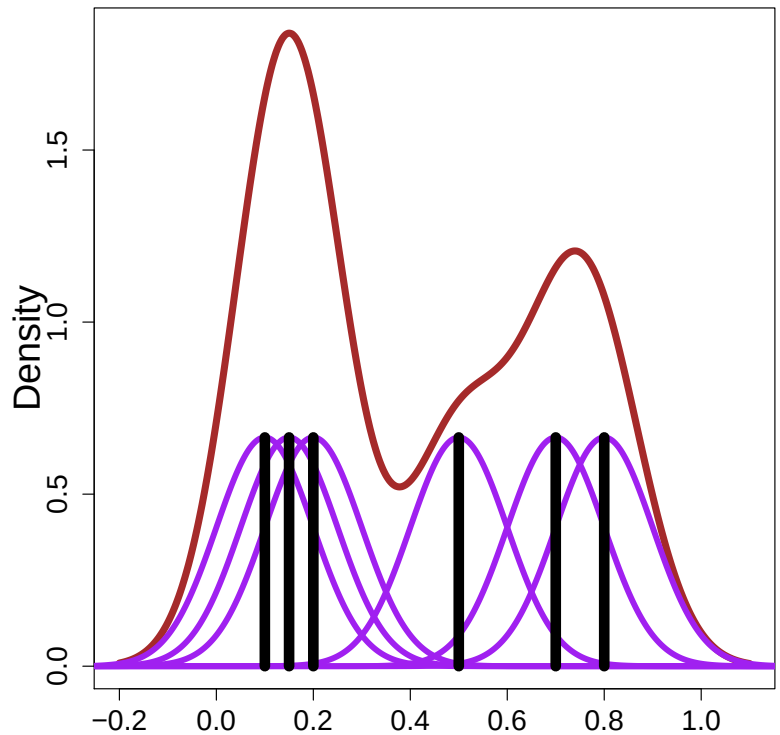
Kernel Density Estimation - 1

- In this lecture, we will focus on one particular nonparametric estimator–the kernel density estimator (KDE).
- The KDE estimates the PDF using the following form:

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

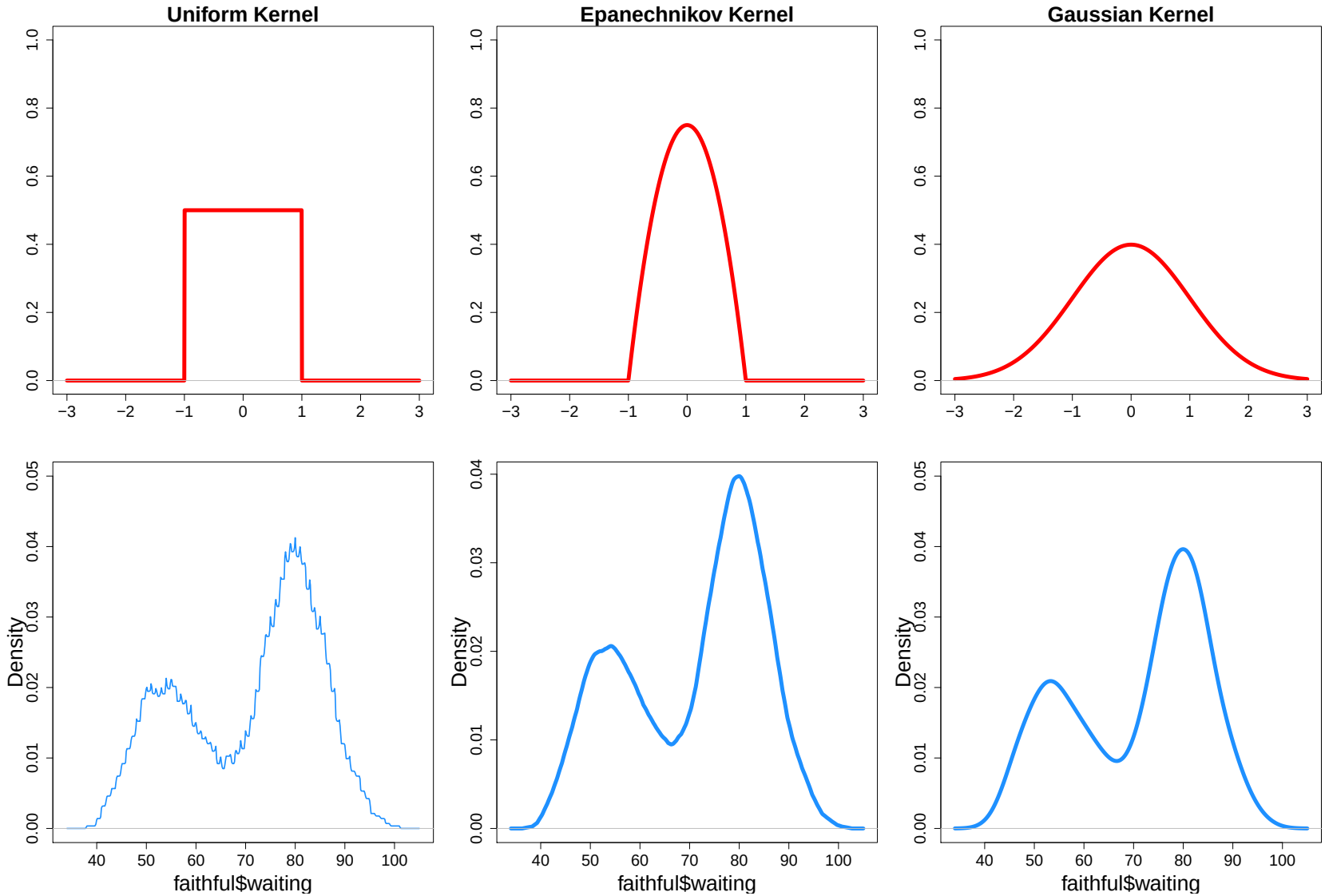
  where $K(x)$ is a function called the kernel function and $h > 0$ is a quantity called smoothing bandwidth that controls the amount of smoothing.
- Common choice of $K(x)$ includes the Gaussian $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and the uniform $K(x) = \frac{1}{2} I(-1 \leq x \leq 1)$.
- The idea of KDE is: we smooth out each data point using the kernel function into small bumps and then we sum over all bumps to obtain a density estimate.

Kernel Density Estimation - 2



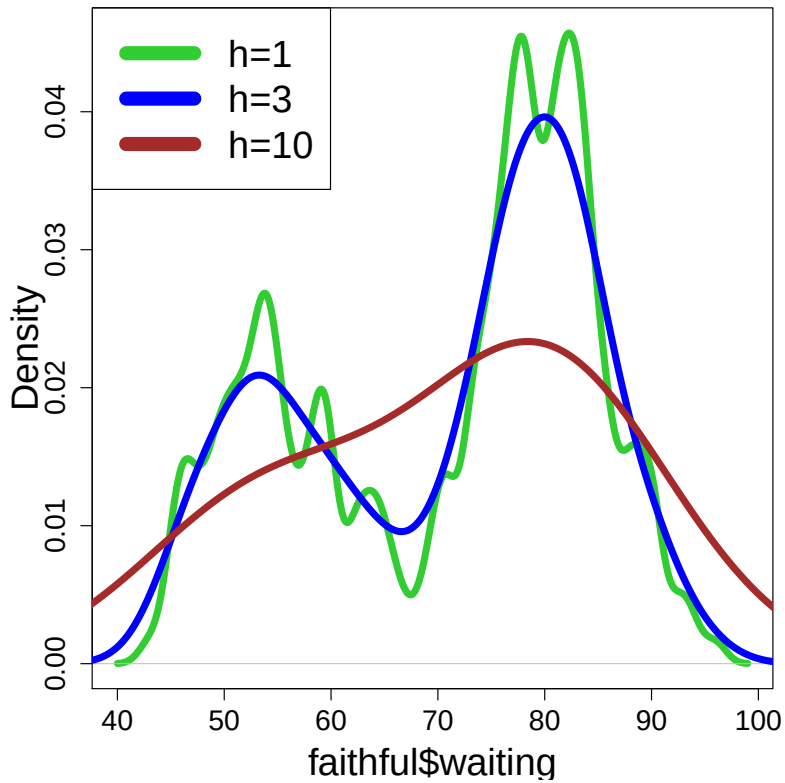
Black dots: locations of observations.
Purple bumps: the kernel function at each observation.
Brown curve: final density estimate from KDE.

Kernel Density Estimation - 3



The kernel function generally does not affect the density estimate too much.

Kernel Density Estimation - 4



The smoothing bandwidth often has a much stronger effect on the quality of estimation.

Asymptotic Theory - 1

- In statistics, there are two common quantities to measure the accuracy of estimation– bias and variance of an estimator.

Asymptotic Theory - 1

- In statistics, there are two common quantities to measure the accuracy of estimation– bias and variance of an estimator.
- When the smoothing bandwidth $h \approx 0$ and sample size n is large,

$$\textbf{bias}(\widehat{p}_h(x)) = \mathbb{E}(\widehat{p}_h(x)) - p(x) = C_{1,K} p''(x) h^2 + o(h^2),$$

and the variance has the asymptotic form:

$$\text{Var}(\widehat{p}_h(x)) = C_{2,K} \frac{p(x)}{nh^d} + o\left(\frac{1}{nh^d}\right),$$

where $C_{1,K}$ and $C_{2,K}$ are constants depending on the kernel function.

Asymptotic Theory - 1

- In statistics, there are two common quantities to measure the accuracy of estimation– bias and variance of an estimator.
- When the smoothing bandwidth $h \approx 0$ and sample size n is large,

$$\textbf{bias}(\widehat{p}_h(x)) = \mathbb{E}(\widehat{p}_h(x)) - p(x) = C_{1,K} p''(x) h^2 + o(h^2),$$

and the variance has the asymptotic form:

$$\text{Var}(\widehat{p}_h(x)) = C_{2,K} \frac{p(x)}{nh^d} + o\left(\frac{1}{nh^d}\right),$$

where $C_{1,K}$ and $C_{2,K}$ are constants depending on the kernel function.

- The mean squared error (MSE) is a common quantity of measuring the accuracy that takes both the bias and variance into consideration. For the KDE, it is

$$\begin{aligned}\text{MSE}(\widehat{p}_h(x)) &= \textbf{bias}^2(\widehat{p}_h(x)) + \text{Var}(\widehat{p}_h(x)) \\ &= C_{1,K}^2 |p''(x)|^2 h^4 + C_{2,K} \frac{p(x)}{nh^d} + o(h^4) + o\left(\frac{1}{nh^d}\right).\end{aligned}$$

Asymptotic Theory - 2

- The MSE measures the accuracy at a single point x. The overall performance is often quantified by the mean integrated squared error (MISE):

$$\begin{aligned}\mathrm{MISE}(\widehat{p}_h) &= \int \mathrm{MSE}(\widehat{p}_h(x)) \\ &= C_{1,K}^2 \left(\int |p''(x)|^2 dx \right) h^4 + \frac{C_{2,K}}{nh^d} + o(h^4) + o\left(\frac{1}{nh^d}\right).\end{aligned}$$

- This implies several interesting facts:
  1. when h is too large, we suffer from the bias.
  2. when h is too small, we suffer from the variance.
  3. the optimal choice is $h \asymp n^{-1/(d+4)}$.

L_∞ Analysis - 1

- The MISE is essentially just the L_2 distance between $\widehat{p}_h$ and p.
- We can then generalize the result to other L_p distance between the two quantities.
- Among all p, one particularly interesting case is L_∞ distance. In this case,

$$\sup_x |\widehat{p}_h(x) - p(x)| = \|\widehat{p}_h - p\|_\infty = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right).$$

- The above bound follows from the following decomposition:

$$\|\widehat{p}_h - p\|_\infty \leq \underbrace{\|\widehat{p}_h - p_h\|_\infty}_{O_P} + \underbrace{\|p_h - p\|_\infty}_{O},$$

where $p_h = \mathbb{E}(\widehat{p}_h) = p \otimes K$ is also called the smoothed density.

L_∞ Analysis - 2

$$\|\widehat{p}_h - p\|_\infty \leq \underbrace{\|\widehat{p}_h - p_h\|_\infty}_{O_P} + \underbrace{\|p_h - p\|_\infty}_{O},$$

- The fact that the bias term at rate $O(h^2)$ is from the usual analysis (Taylor expansion).
- The bound on the stochastic variation is more involved.
- In short, it follows from the Talagrand's inequality:

$$P(\|\widehat{p}_h - p_h\|_\infty > t) \leq c_0 e^{-c_1 n h^d t^2}$$

when $t > \sqrt{\frac{|\log h|}{nh^d}}$. This result is formally given in Giné and Guillou (2002).

- In fact, after rescaling, the random variable $\|\widehat{p}_h - p_h\|_\infty$ converges in distribution to an extreme value distribution (Bickel and Rosenblatt (1973)).

L_∞ Analysis - 3

- The L_∞ analysis can be generalized to derivatives of the density function.
- Gradient:

$$\sup_x \|\nabla \widehat{p}_h(x) - \nabla p(x)\|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^{d+2}}}\right).$$

- Hessian:

$$\sup_x \|\nabla\nabla \widehat{p}_h(x) - \nabla\nabla p(x)\|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^{d+4}}}\right).$$

- And higher-order derivatives can be derived accordingly.

Application of L_∞ Analysis - 1

- The L_∞ analysis implies a construction of a *simultaneous confidence band* of p.
- There are two types of confidence bands for a function p: pointwise confidence bands and simultaneous confidence bands.

Application of L_∞ Analysis - 1

- The L_∞ analysis implies a construction of a *simultaneous confidence band* of p.
- There are two types of confidence bands for a function p: pointwise confidence bands and simultaneous confidence bands.
- Pointwise CI: for any given point x and confidence level $1-\alpha$, we construct an interval $C_{1-\alpha} = [\ell_{1-\alpha}, u_{1-\alpha}]$ from the data such that

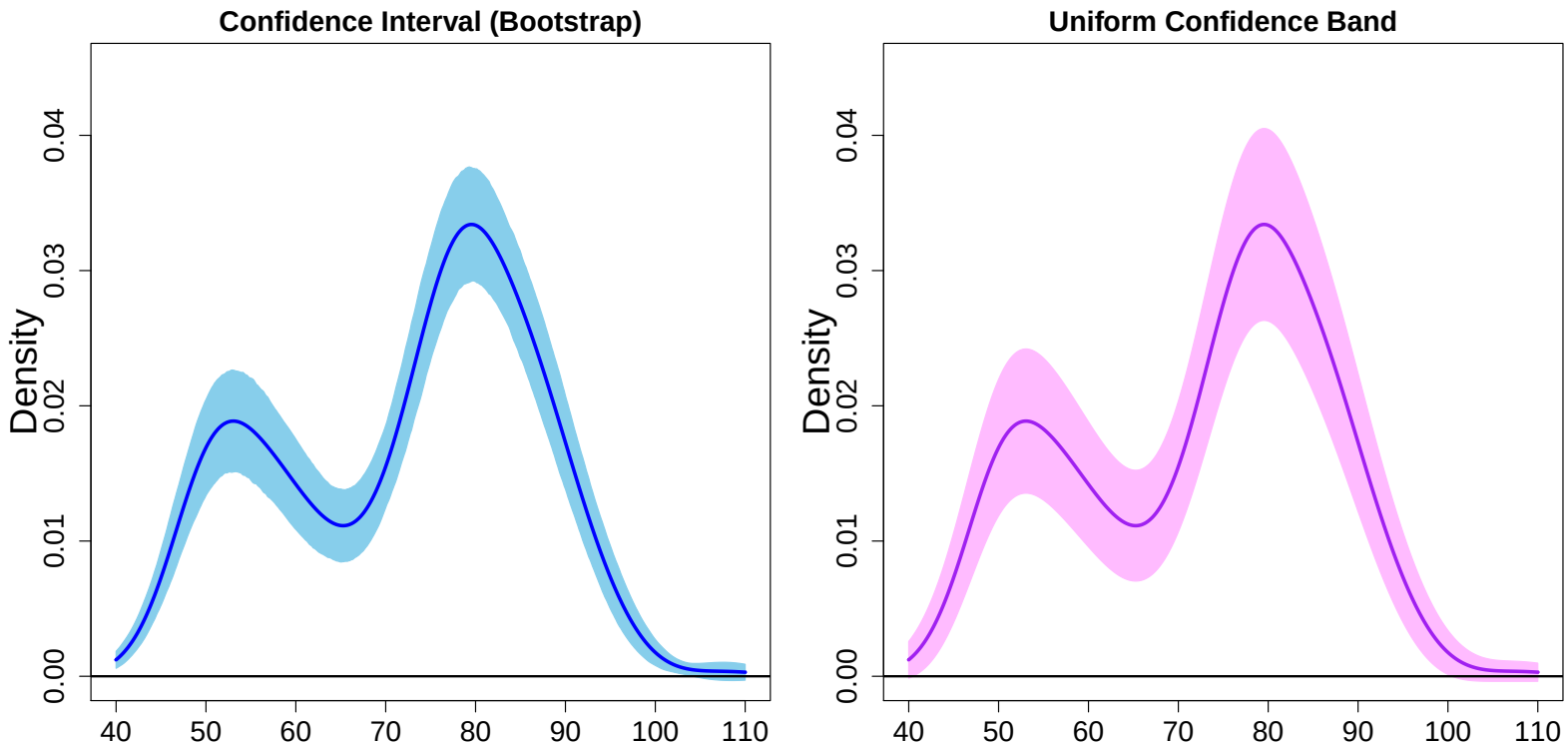
$$P(\ell_{1-\alpha} \leq p(x) \leq u_{1-\alpha}) = 1 - \alpha + o(1).$$

- Simultaneous CB (confidence band): given $1-\alpha$, we construct a band $C_{1-\alpha}(x) = [L_{1-\alpha}(x), U_{1-\alpha}(x)]$ from the data such that

$$P(L_{1-\alpha}(x) \leq p(x) \leq U_{1-\alpha}(x) \text{ for all } x) = 1 - \alpha + o(1).$$

Application of L_∞ Analysis - 2

We can construct a simultaneous confidence band by bootstrapping the L_∞ distance.



Pointwise CI (left) and simultaneous CB (right)[1].

[1] More details can be found in: **https://arxiv.org/abs/1702.07027**

Application of L_∞ Analysis - 3

- The L_∞ analysis also implies the convergence of geometric structures.
- In particular, some geometric structures converge when the derivatives converge[2].
- Convergence rate depending on $\|\widehat{p}_h - p_h\|_\infty$:
  - Level sets, cluster trees, and persistent diagrams.
- Convergence rate depending on $\|\nabla\widehat{p}_h - \nabla p_h\|_\infty$:
  - Local modes, Morse-Smale complex, and gradient system.
- Convergence rate depending on $\|\nabla\nabla\widehat{p}_h - \nabla\nabla p_h\|_\infty$:
  - Ridges.

[2] A tutorial on this topic is in: **https://arxiv.org/abs/1704.03924**

Geometric Estimation

Level Sets - 1

- Given a level $\lambda > 0$, the density level set is

$$L_\lambda = \{x : p(x) = \lambda\}.$$

- A natural estimator of L_λ is the plug-in using a KDE[3]:

$$\widehat{L}_\lambda = \{x : \widehat{p}_h(x) = \lambda\}.$$

- Note that sometime in the literature, the set of interest is the upper level set:

$$S_\lambda = \{x : p(x) \geq \lambda\}.$$

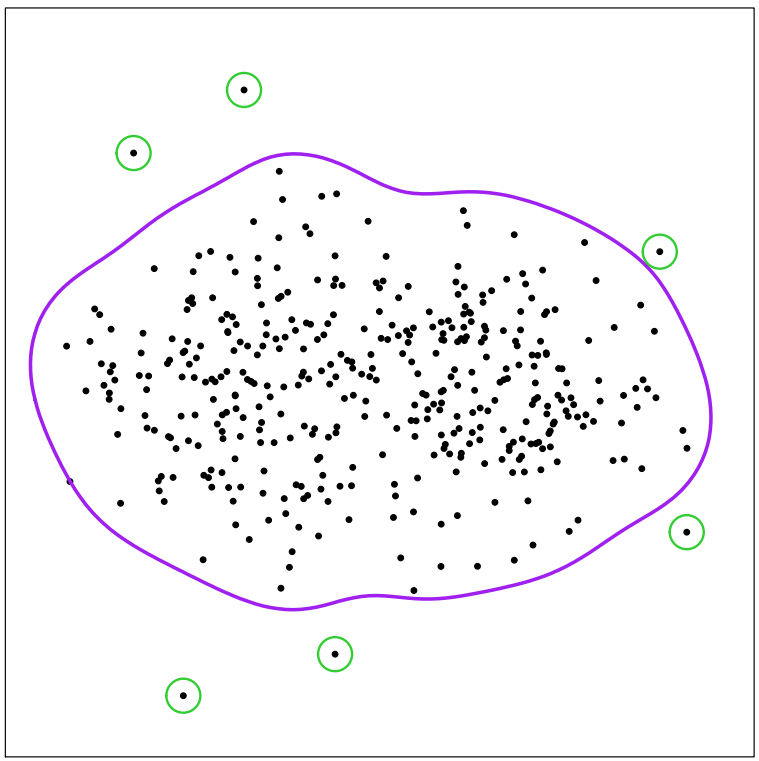
  Under smoothness conditions, the boundary of the upper level set is the level set L_λ.

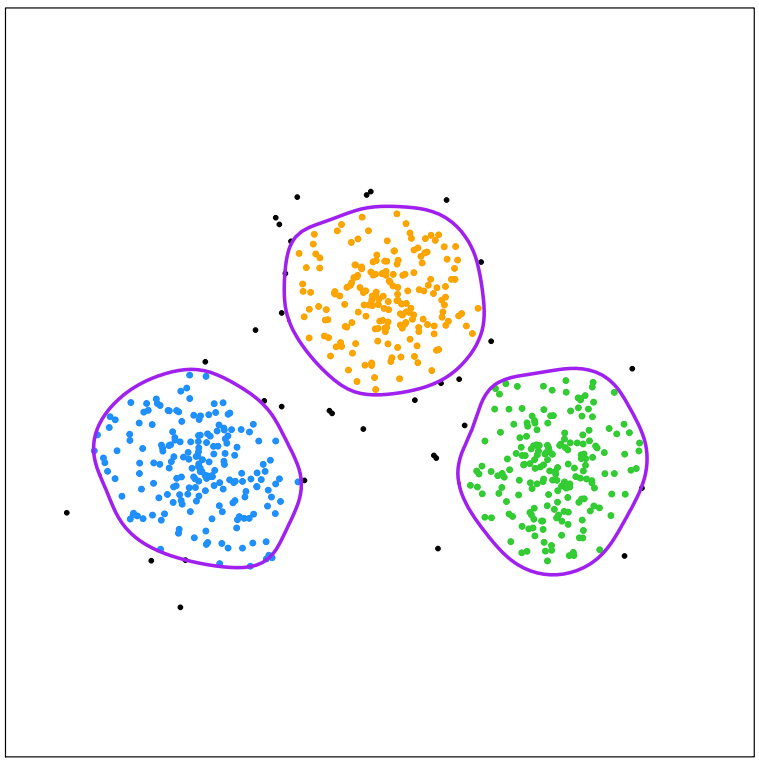
- Level set is a particularly interesting example so we take a deeper look at this problem.

[3] Materials on this topic can be found in:
https://arxiv.org/abs/1504.05438 and the reference therein.

Level Sets - 2

- Level set has two common applications:
  1. Anomaly detection–observations in the low density area (outside the level set) are going to be classified as anomaly.
  2. Clustering–observations inside the level set (high density area) are going to be clustered together.





Level Sets - 3

- There has been a tremendous amount of literature on the convergence of level set.
- Often the convergence is expressed in terms of the *Hausdorff distance*

$$\text{Haus}(A, B) = \max\{\sup_{x \in A} d(x, B), \sup_{x \in B} d(x, A)\},$$

  where $d(x, A) = \inf_{y \in A} \|x - y\|$ is the distance from a point x to a set A.
- The Hausdorff distance can be viewed as an L_∞ distance for sets.

Level Sets - 4

- A common assumption to ensure the convergence of Hausdorff distance is the gradient bound:

$$\inf_{x \in L_\lambda} \|\nabla p(x)\| \geq g_0 > 0,$$

  for some constant g_0.
- Under this assumption (and some other common assumptions),

$$\text{Haus}(\widehat{L}_\lambda, L_\lambda) = O_P\left(\|\widehat{p}_h - p\|_\infty\right).$$

Level Sets - 5

- If we further assume that $p(x)$ has bounded second derivative everywhere, then the level set L_λ is smooth in the sense that the *reach* is positive.
- The reach of a set is the longest distance away from a set that still has a unique projection back to the set.
- If L_λ has a reach r_0, then for any point x with $d(x, L_\lambda) < r_0$, x has a unique projection back to L_λ.
- The positive reach properties of level sets imply that $\widehat{L}_\lambda$ and L_λ are (asymptotically) *normal compatible*, meaning that there is a unique projection from every point in L_λ to $\widehat{L}_\lambda$ and vice versa.

Level Sets - 6

- The normal compatibility implies that we can decompose the Hausdorff distance as

$$\text{Haus}(\widehat{L}_\lambda, L_\lambda) = \sup_{x \in L_\lambda} d(x, \widehat{L}_\lambda).$$

- An more interesting fact is that for any $x \in L_\lambda$,

$$d(x, \widehat{L}_\lambda) = \frac{1}{\|\nabla p(x)\|} |\widehat{p}_h(x) - p(x)| + \text{smaller order terms.}$$

- This implies that asymptotically,

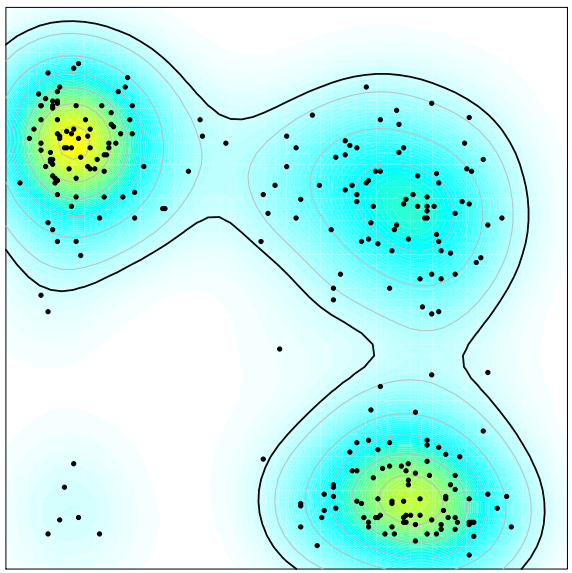
$$\text{Haus}(\widehat{L}_\lambda, L_\lambda) = \sup_{x \in L_\lambda} \frac{1}{\|\nabla p(x)\|} |\widehat{p}_h(x) - p(x)|,$$

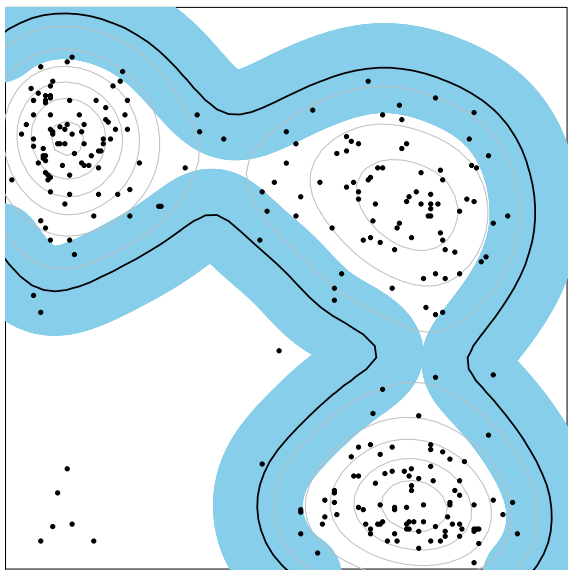
  which is the supremum of a stochastic process defined over the manifold.

Level Sets - 7

$$\text{Haus}(\widehat{L}_\lambda, L_\lambda) = \sup_{x \in L_\lambda} \frac{1}{\|\nabla p(x)\|} |\widehat{p}_h(x) - p(x)|.$$

- This shows that the Hausdorff distance follows an extreme value distribution after rescaling.
- Also, it implies that we can use the bootstrap to construct a confidence set of L_λ.





Local Modes - 1

- Another interesting geometric structure is the local modes of the PDF:

$$M = \{x : \nabla p(x) = 0, \lambda_1(x) < 0\},$$

where $\lambda_1(x)$ is the largest eigenvalue of the Hessian matrix $\nabla\nabla p(x)$[4].

- Similar to the level set problem, a simple estimator of M is the plug-in from the KDE

$$\widehat{M}_h = \{x : \nabla \widehat{p}_h(x) = 0, \widehat{\lambda}_1(x) < 0\}.$$

[4] A tutorial on this topic is in: **https://arxiv.org/abs/1406.1780** and **https://arxiv.org/abs/1408.1381**.

Local Modes - 2

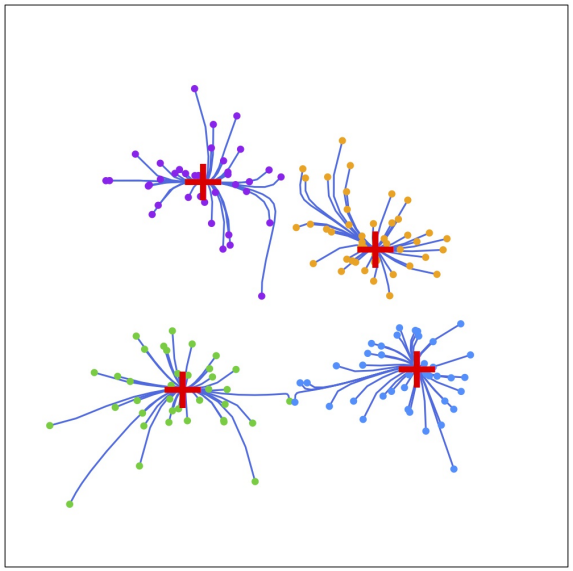
- Before talking about the applications of local modes, we first discuss some of its properties.
- If the density function p is a Morse function, i.e., all critical points of p are non-degenerated, then
  1. $\text{Haus}(\widehat{M}_h, M) = O_P\left(\sup_x \|\nabla \widehat{p}_h(x) - \nabla p(x)\|_{\max}\right)$,
  2. with a probability approaching to 1, there exists a one-to-one correspondence between elements in $\widehat{M}_h$ and elements in M.
- A common assumption to replace the Morse condition is that there exists a lower bound $\underline{\lambda} > 0$ such that

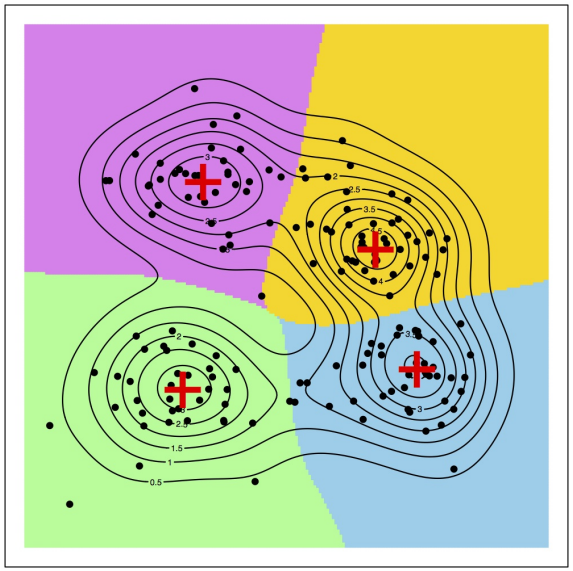
$$\min_{x \in M} |\lambda_1(x)| \geq \underline{\lambda} > 0.$$

- In fact, one can obtain a faster convergence rate of $\text{Haus}(\widehat{M}_h, M)$ without the $\log n$ term in the variance.

Local Modes - 3

- Local modes can be used to perform a cluster analysis.
- This is known as the mode clustering method (mean-shift clustering).





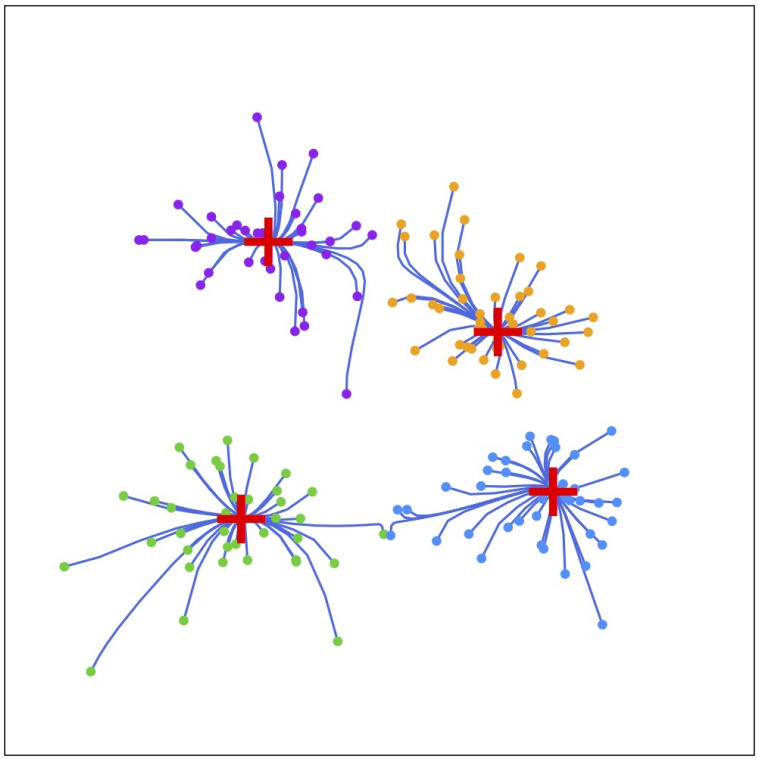
Local Modes - 4

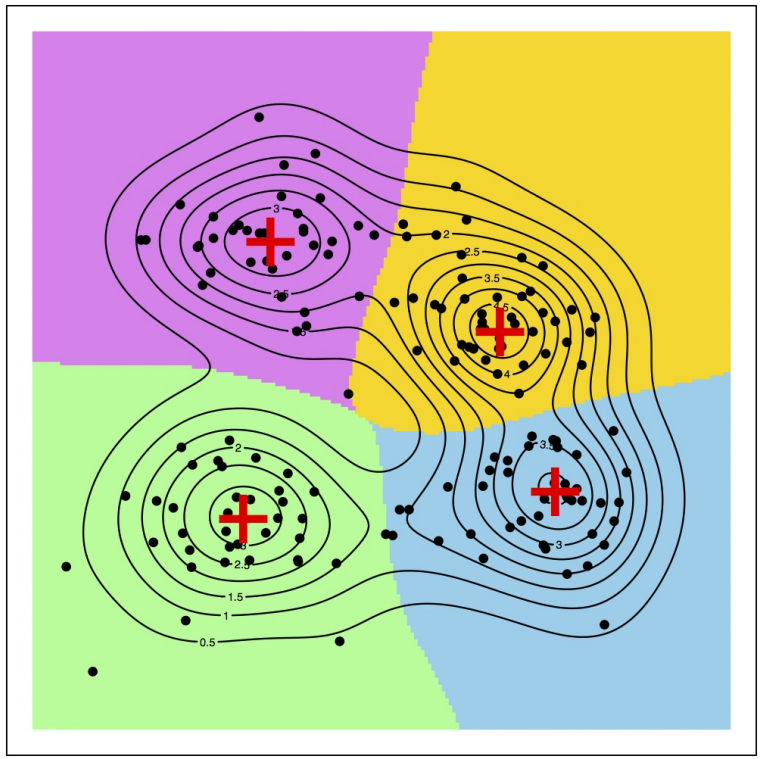
- The clusters are defined through a gradient system of p (or in the sample case, $\widehat{p}_h$).
- For a given point x, we define a gradient flow π_x such that

$$\pi_x(0) = x, \quad \pi_x'(t) = \nabla p(\pi_x(t)).$$

- The destination $\pi_x(\infty) = \lim_{t\to\infty} \pi_x(t) \in M$ when p is a Morse function for almost every x except a set of point with Lebesgue measure 0.
- Thus, we can use the destination $\pi_x(\infty)$ of each x to cluster data points. Namely, points with the same destination will be clustered together.

Local Modes - 5





Local Modes - 6

- Numerically, we use the *mean-shift* algorithm to do the gradient ascent.
- Let x_0 be the initial point.
- We iterate

$$x_{t+1} = \frac{\sum_{i=1}^{n} X_i K\left(\frac{X_i - x_t}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x_t}{h}\right)}$$

  until convergence.
- Note that this works for Gaussian kernel. Some other kernel functions also work after modifications.

Local Modes - 7

- For each $m \in M$, let $D(m) = \{x : \pi_x(\infty) = m\}$ be the basin of attraction with respect to m.
- The set $\mathcal{D} = \{D(m) : m \in M\}$ forms a partition of the entire support of p (except for a set of Lebesgue measure 0).
- Similarly, we may define the sample version of it $\{\widehat{D}(m) : m \in \widehat{M}_h\}$, where $\widehat{\mathcal{D}} = \widehat{D}(m) = \{x : \widehat{\pi}_x(\infty) = m\}$ is the basin of attraction using the gradient system of $\widehat{p}_h$.
- The sample version partition $\widehat{\mathcal{D}}$ converges to the 'population' partition $\mathcal{D}$ with proper assumptions[5].

[5] Materials on this topic can be found in:
https://arxiv.org/abs/1506.08826

Local Modes - 8

- Let $\mathcal{B} = \{\partial D(m) : m \in M\}$ be the collection of boundaries of the basins and $\widehat{\mathcal{B}} = \{\partial \widehat{D}(m) : m \in M\}$ be the sample version of it.
- The convergence of $\widehat{\mathcal{D}}$ toward $\mathcal{D}$ can be characterized by the convergence of $\widehat{\mathcal{B}}$ to $\mathcal{B}$.
- If for every $x \in M$, $p(x)$ are convex with respect to the 'normal space' of $\mathcal{B}$ at x, then

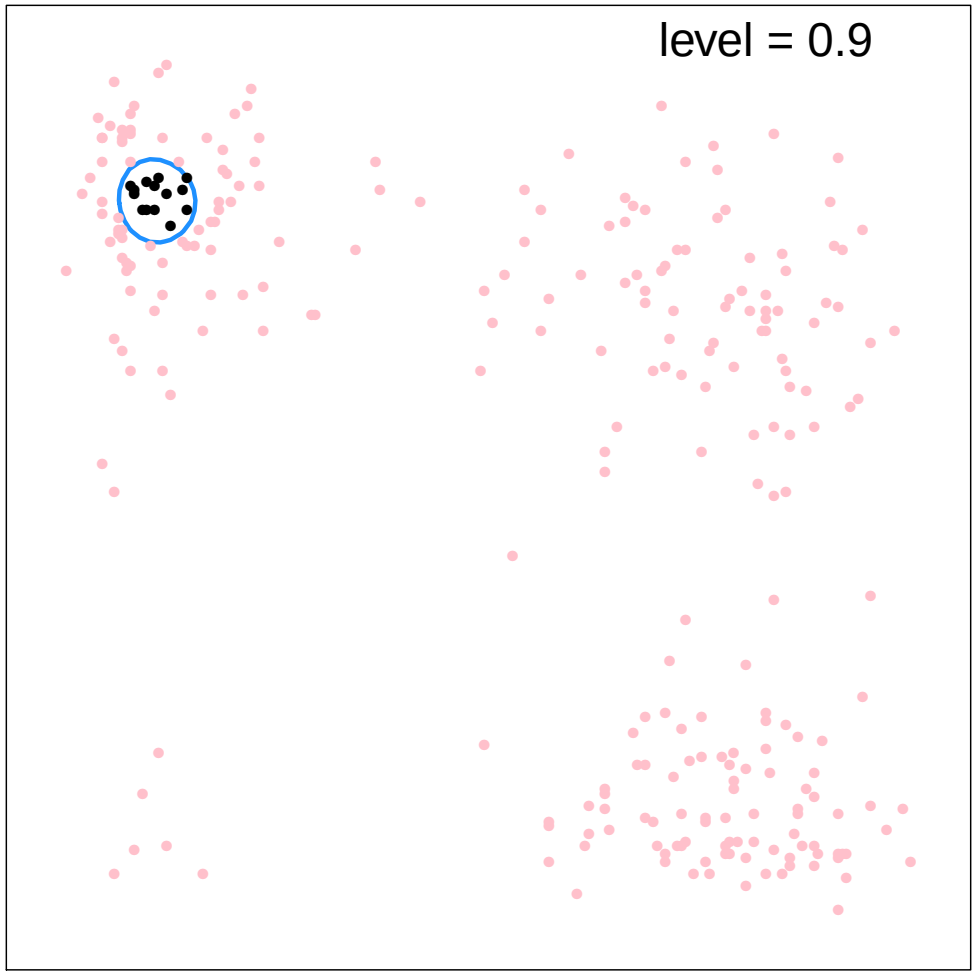
$$\text{Haus}(\widehat{\mathcal{B}}, \mathcal{B}) = O_P\left(\sup_x \|\nabla \widehat{p}_h(x) - \nabla p(x)\|_{\max}\right).$$

Cluster Tree - 1

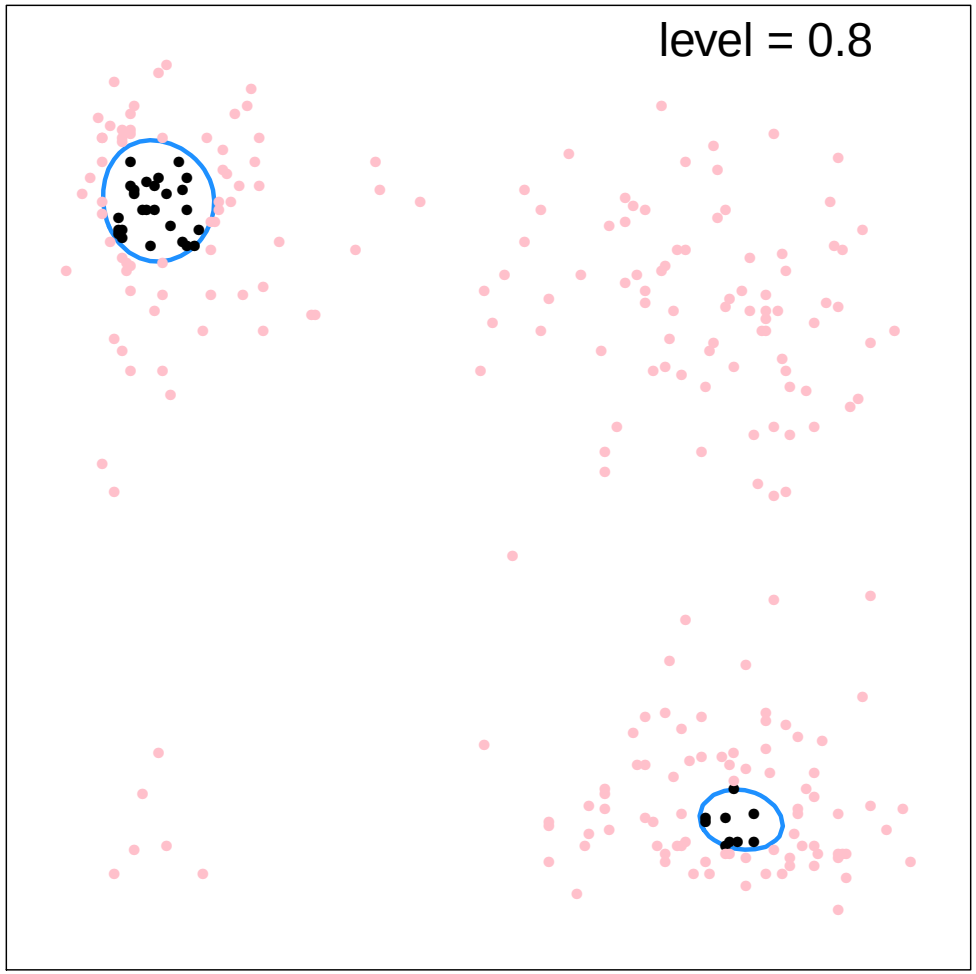
- There is an elegant idea combining both level sets and modes: cluster tree[6].
- The cluster tree considers the collection of clusters formed by the upper level sets and keeps track of their relationships.
- When applying to a density function, a cluster tree is also called a density tree.

[6] Materials on this topic can be found in:
https://arxiv.org/abs/1605.06416

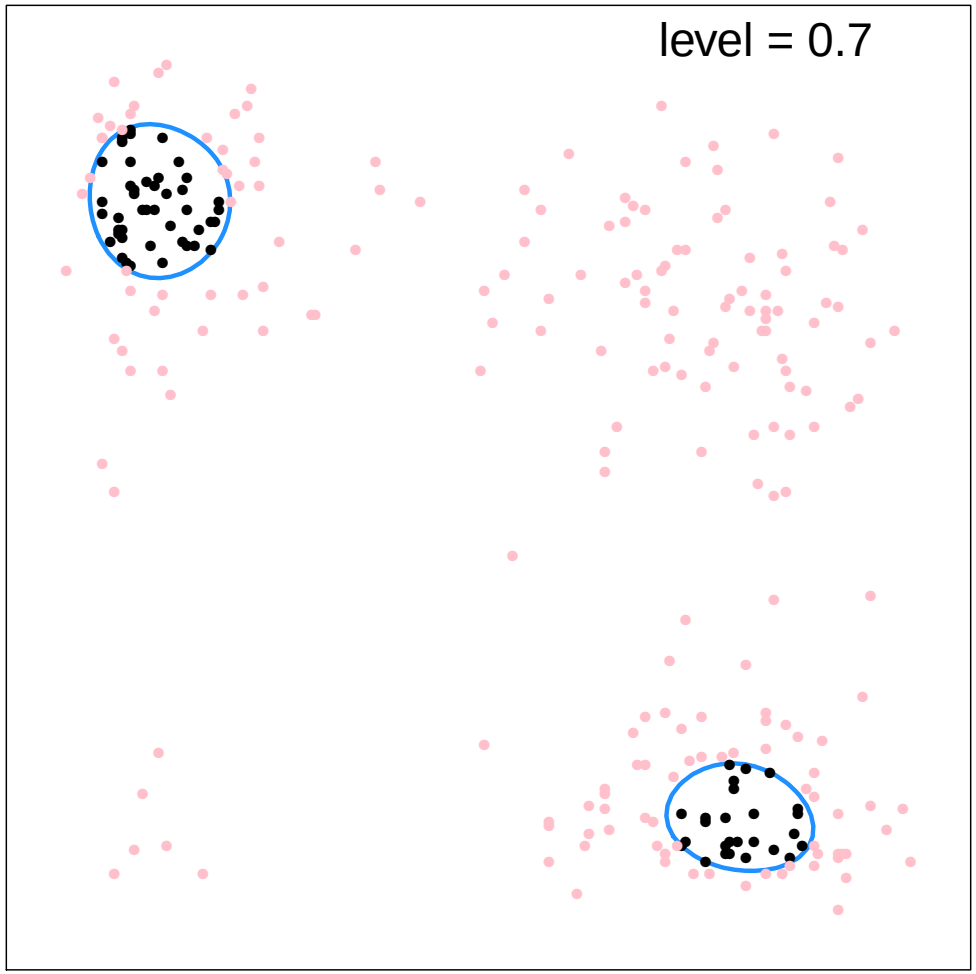
Cluster Tree - 2



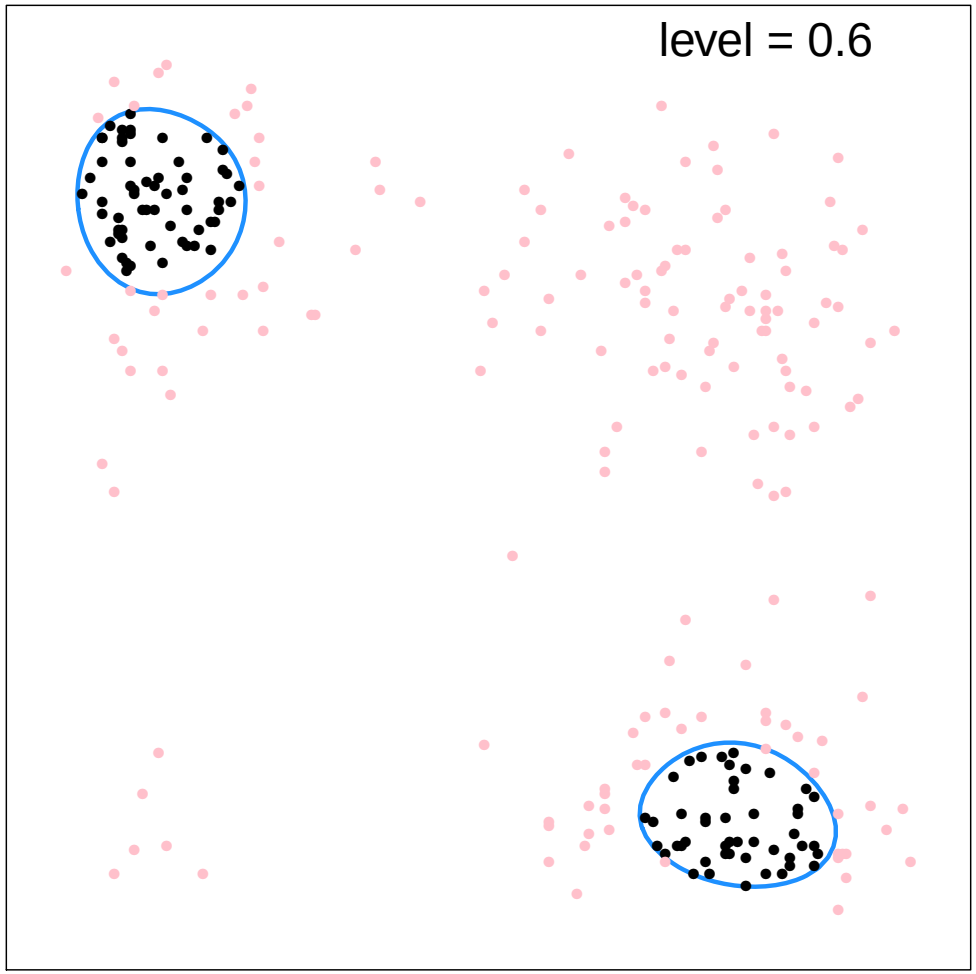
Cluster Tree - 2



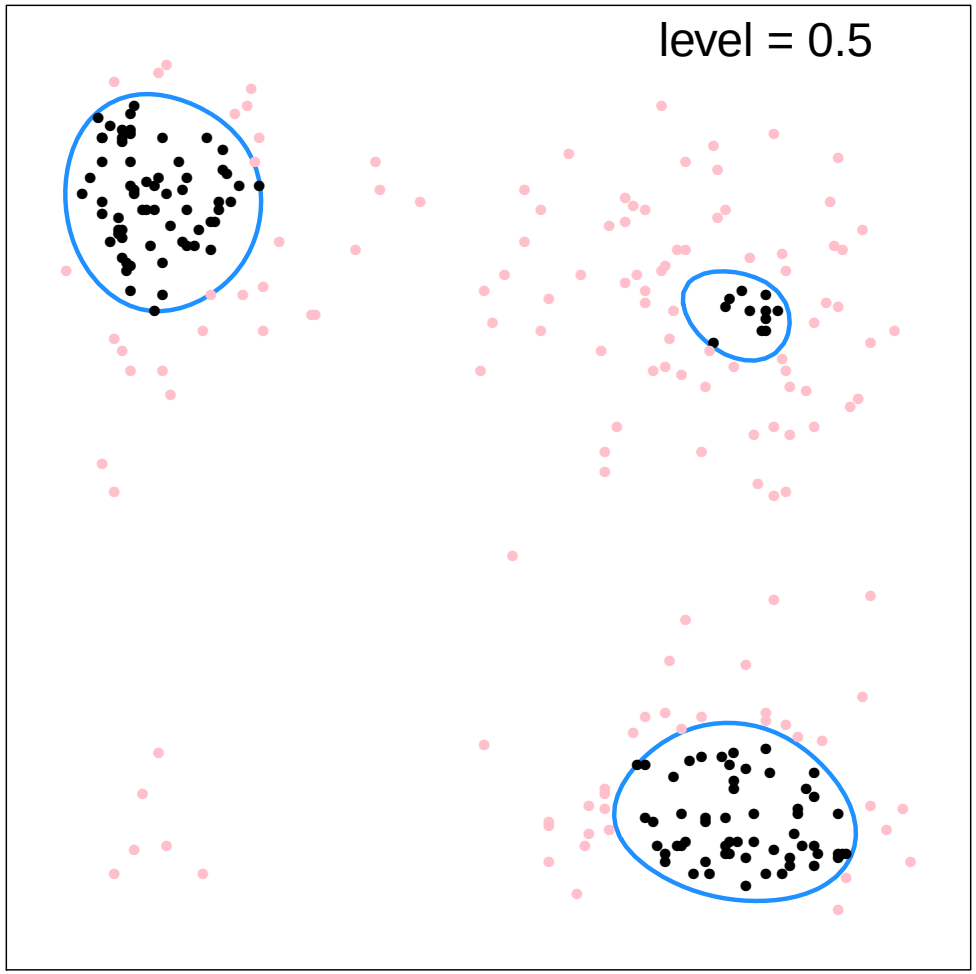
Cluster Tree - 2



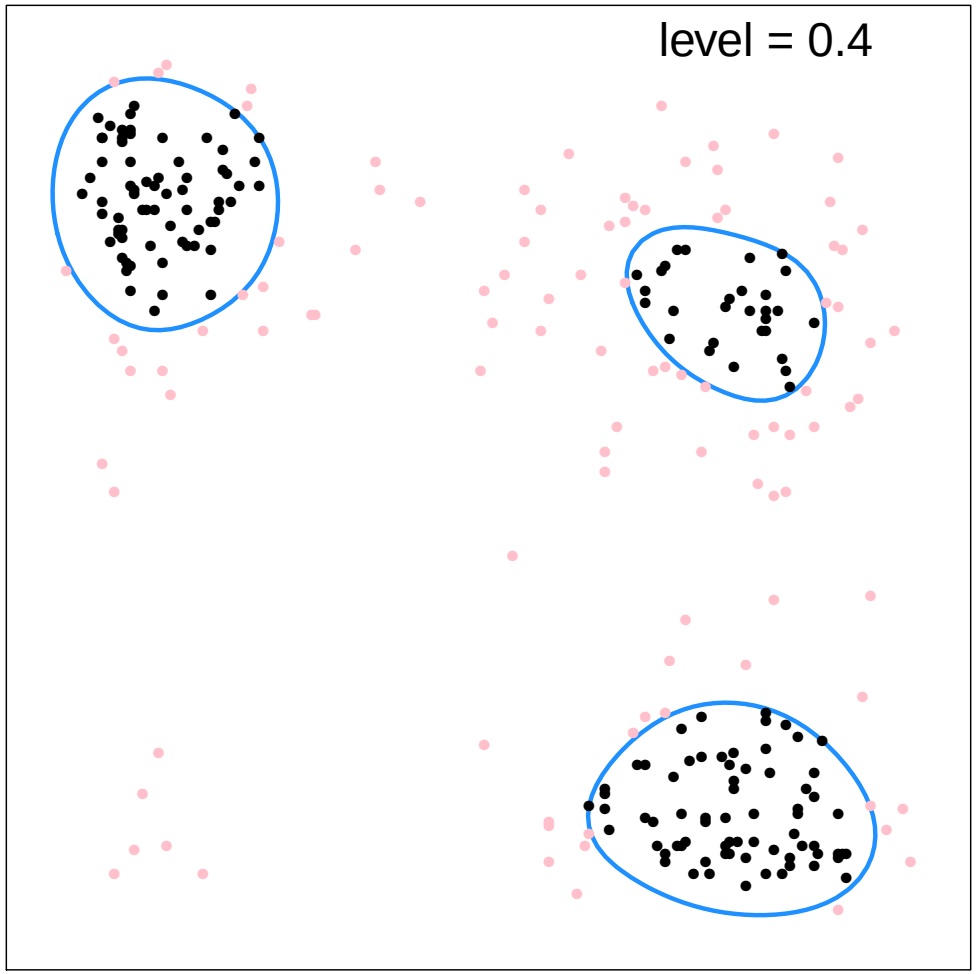
Cluster Tree - 2



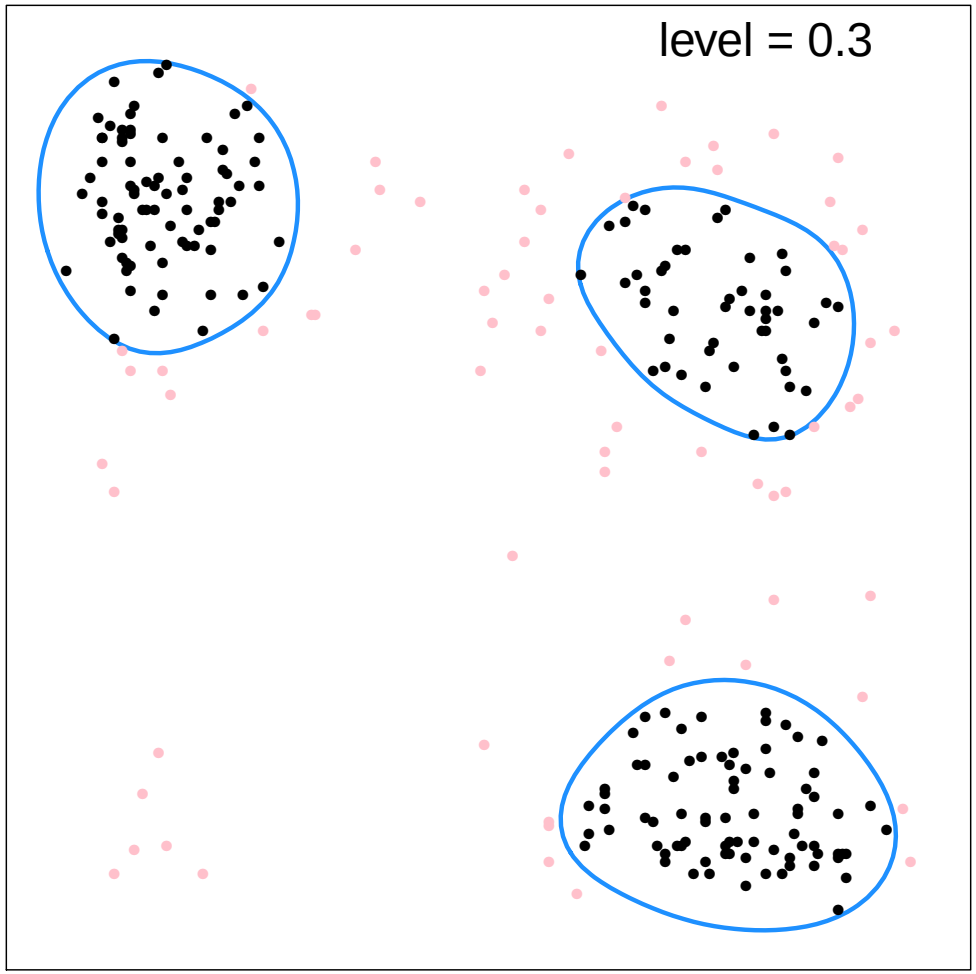
Cluster Tree - 2



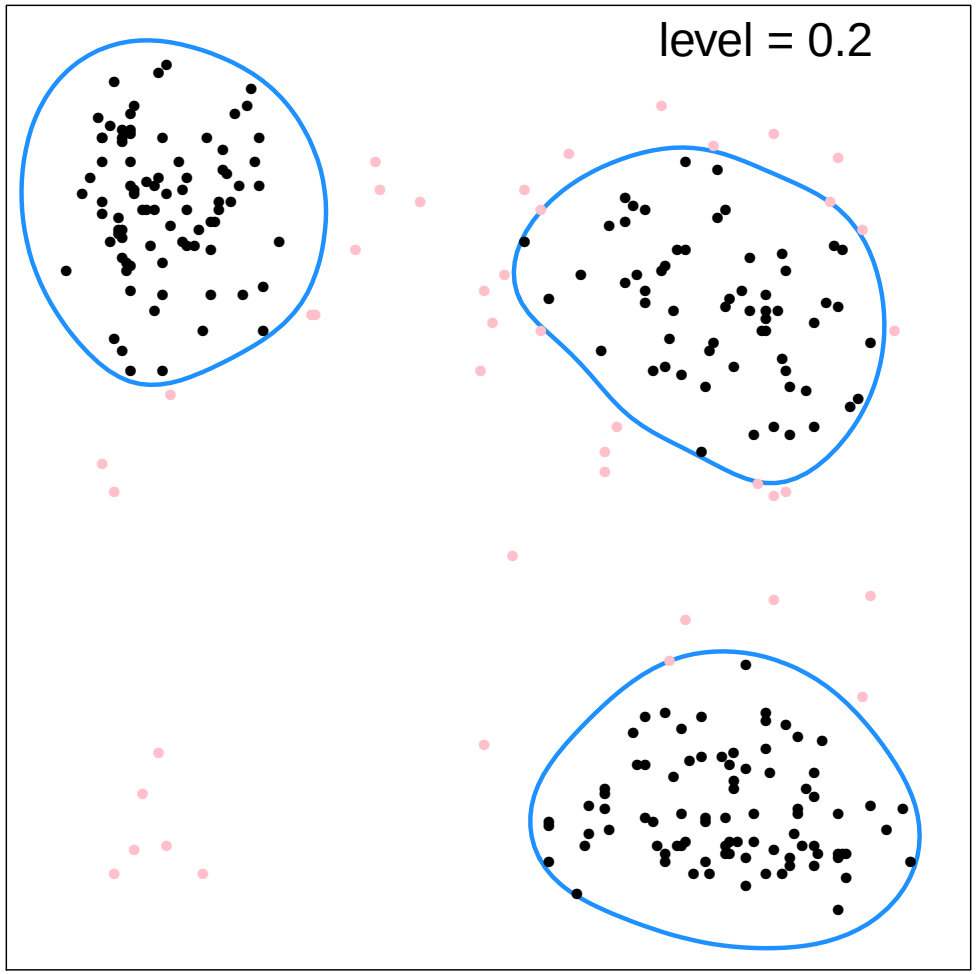
Cluster Tree - 2



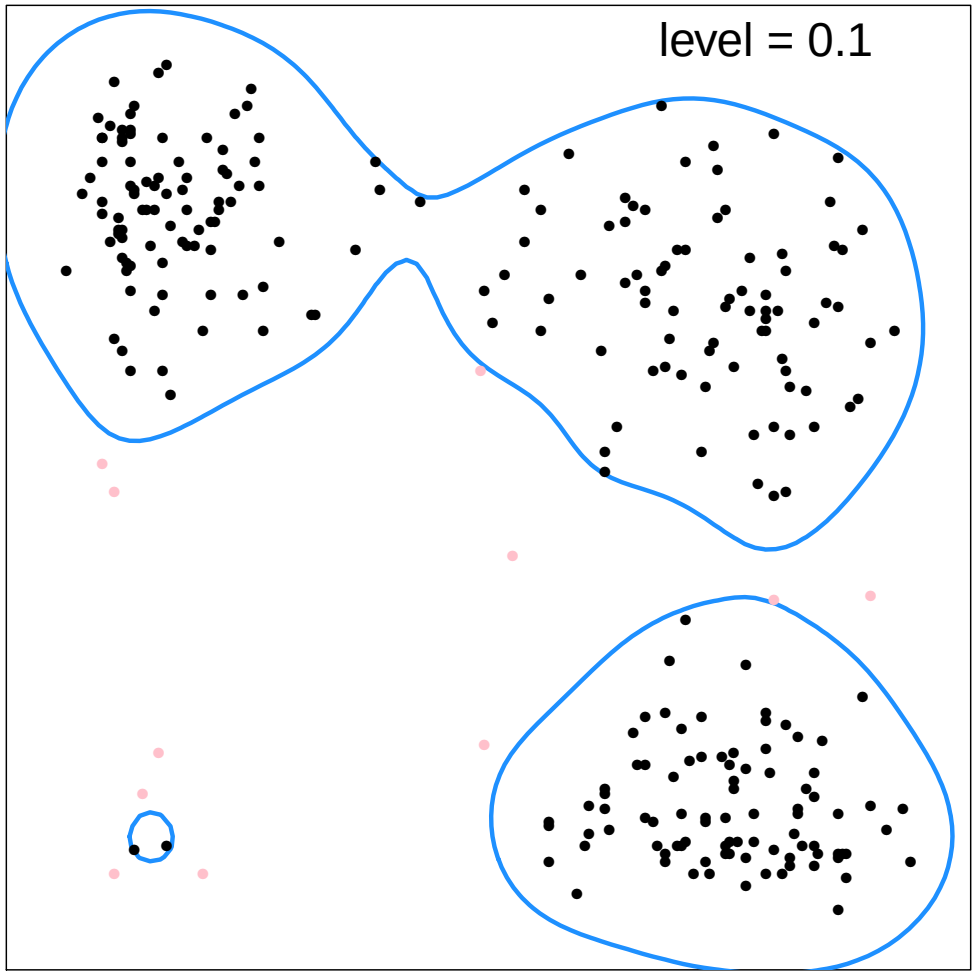
Cluster Tree - 2



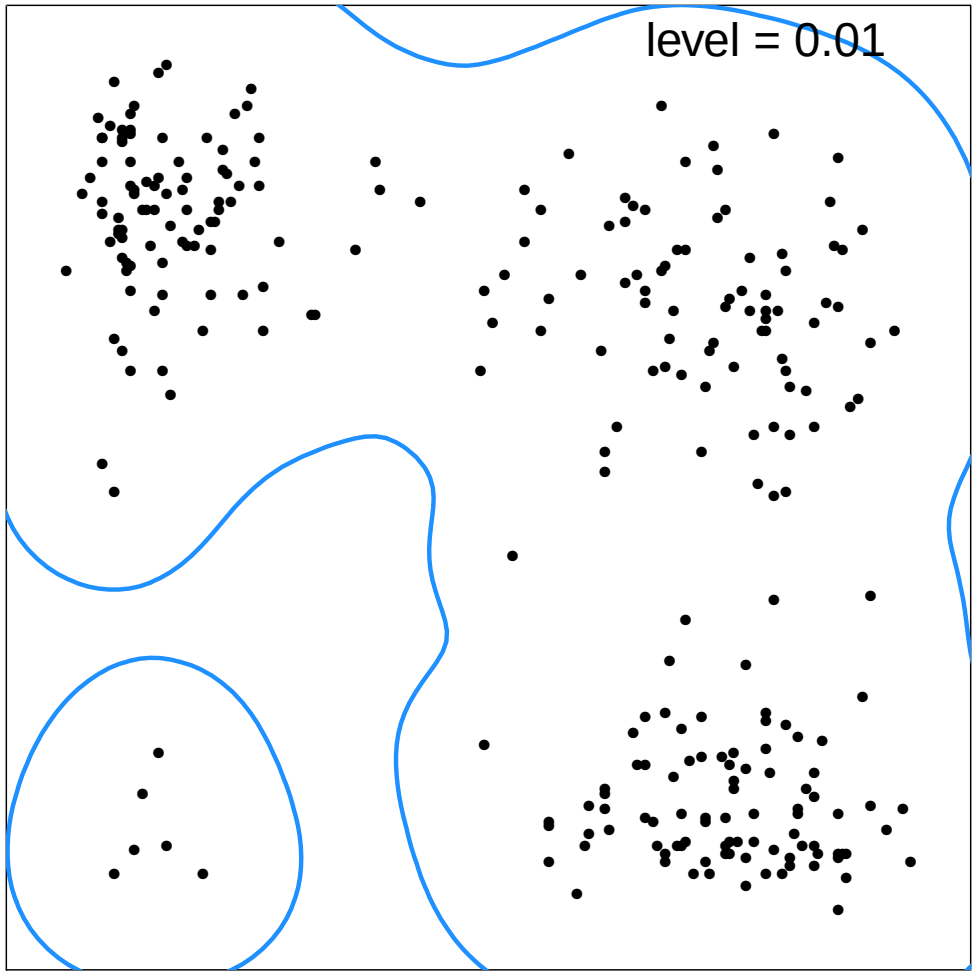
Cluster Tree - 2



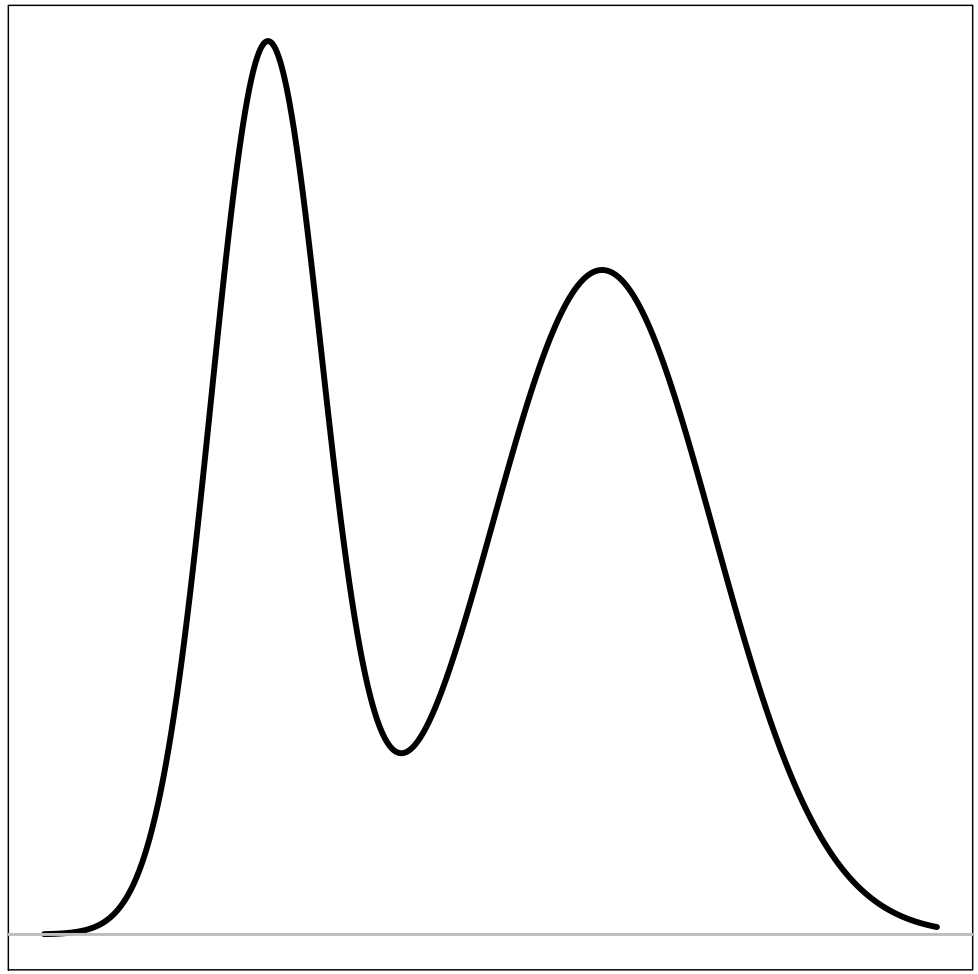
Cluster Tree - 2



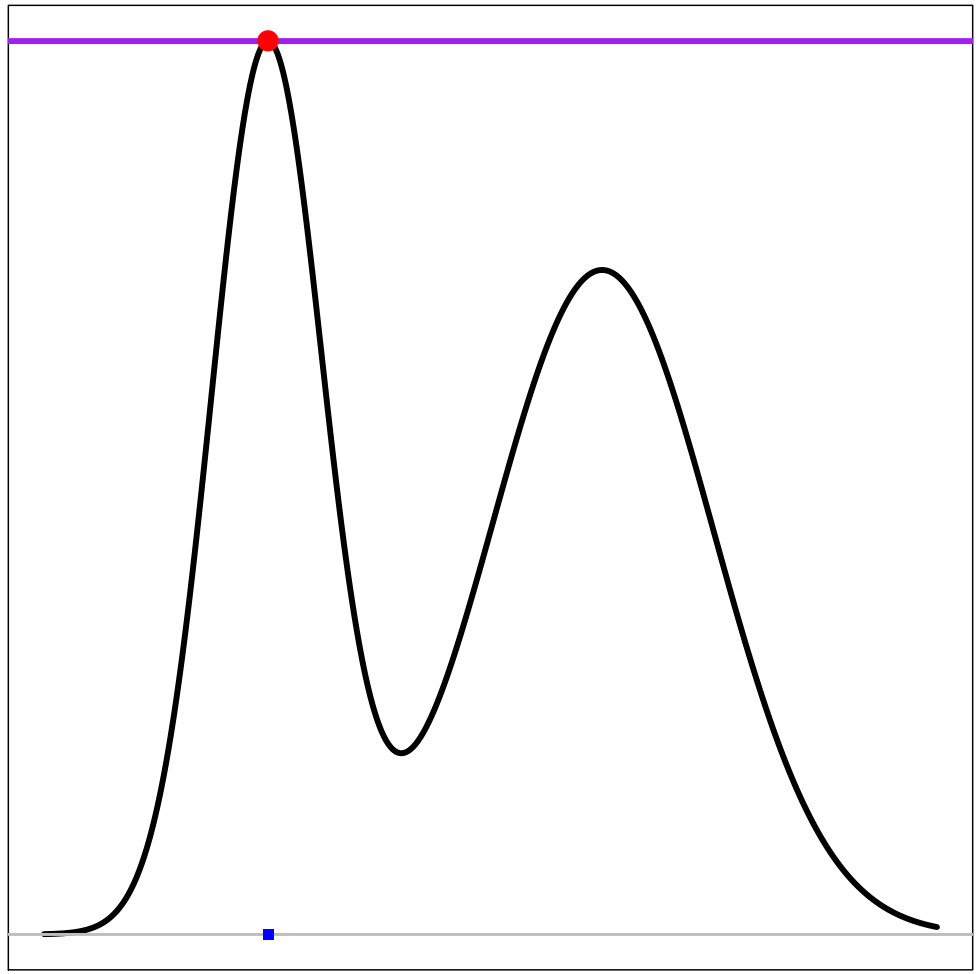
Cluster Tree - 2



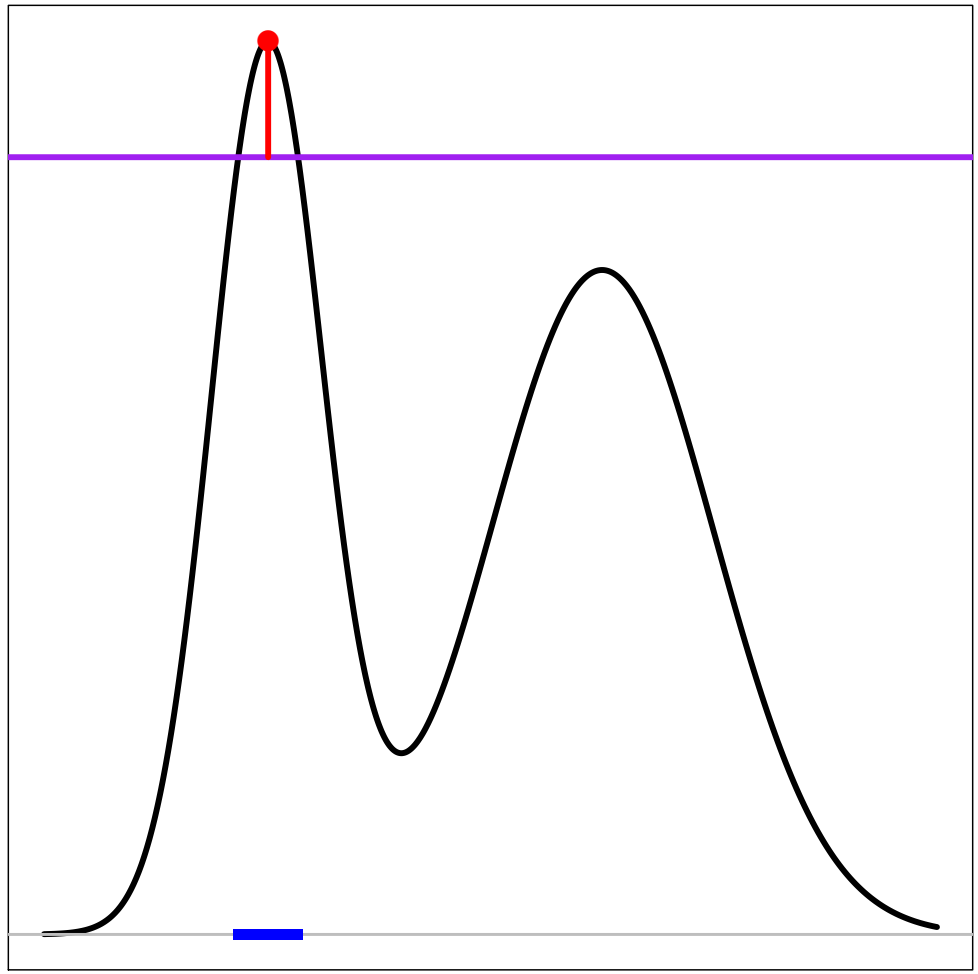
Cluster Tree - 3



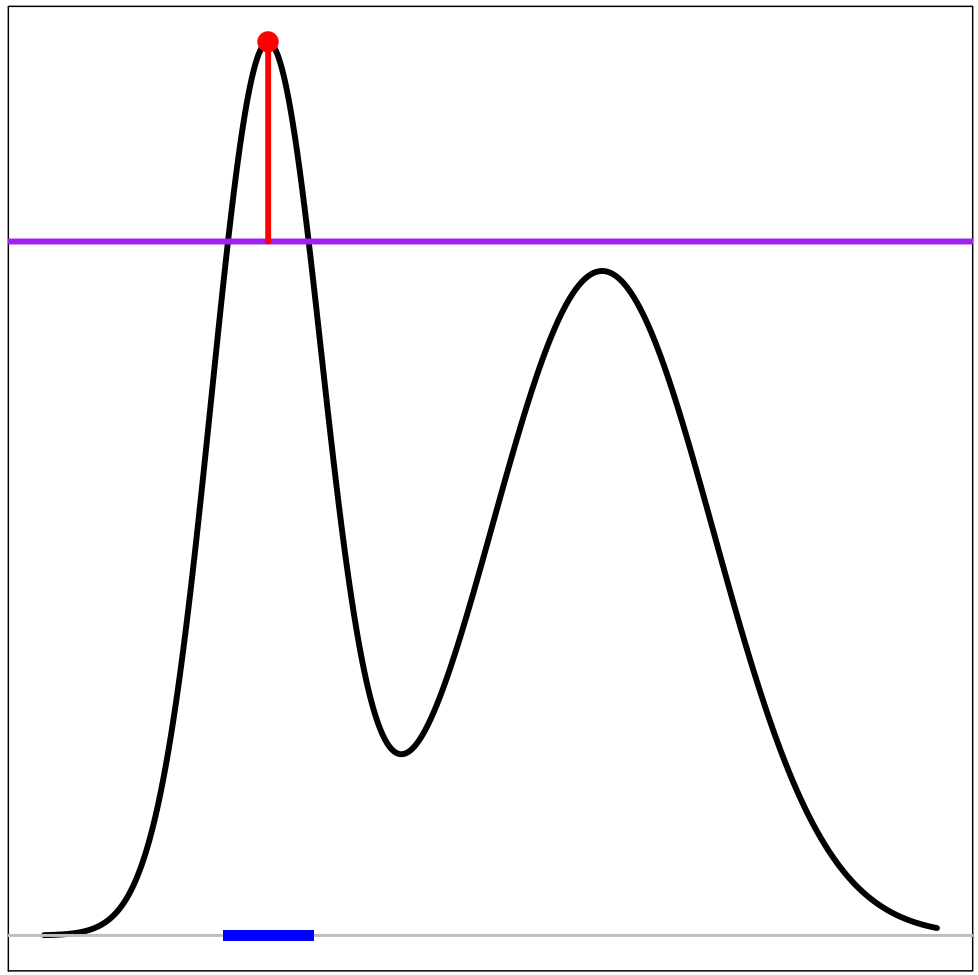
Cluster Tree - 3



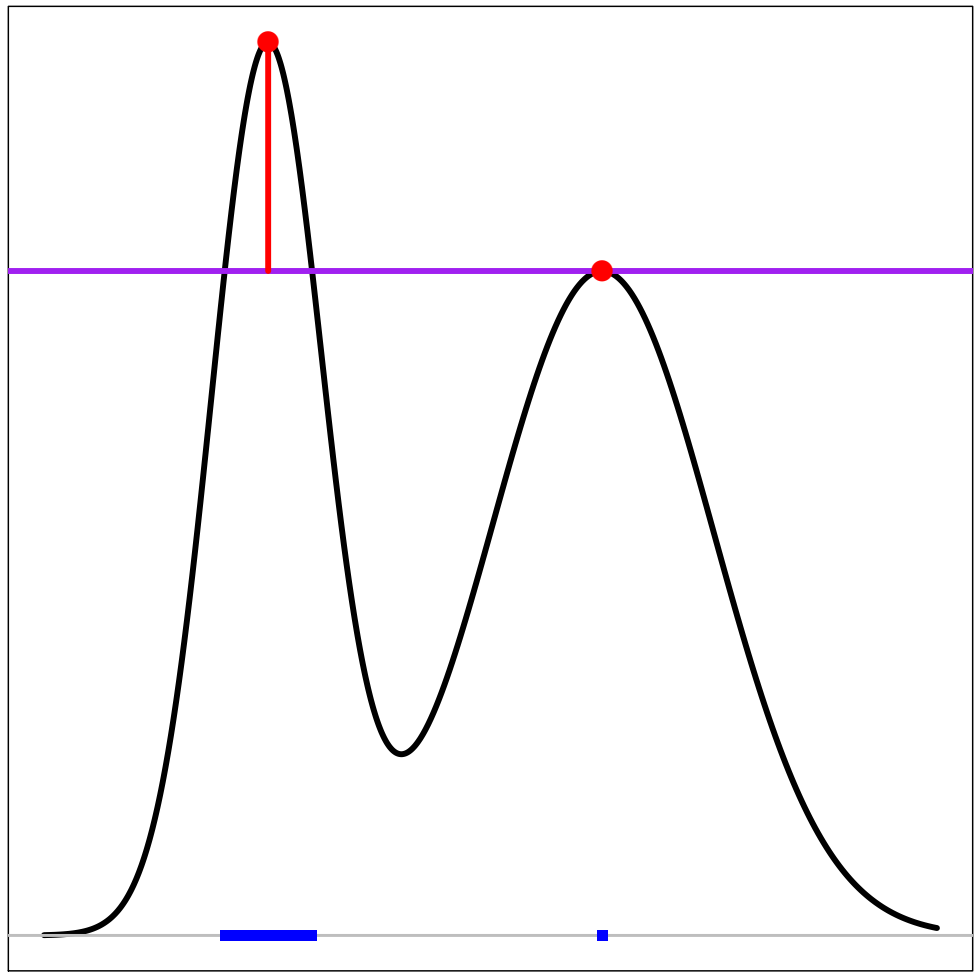
Cluster Tree - 3



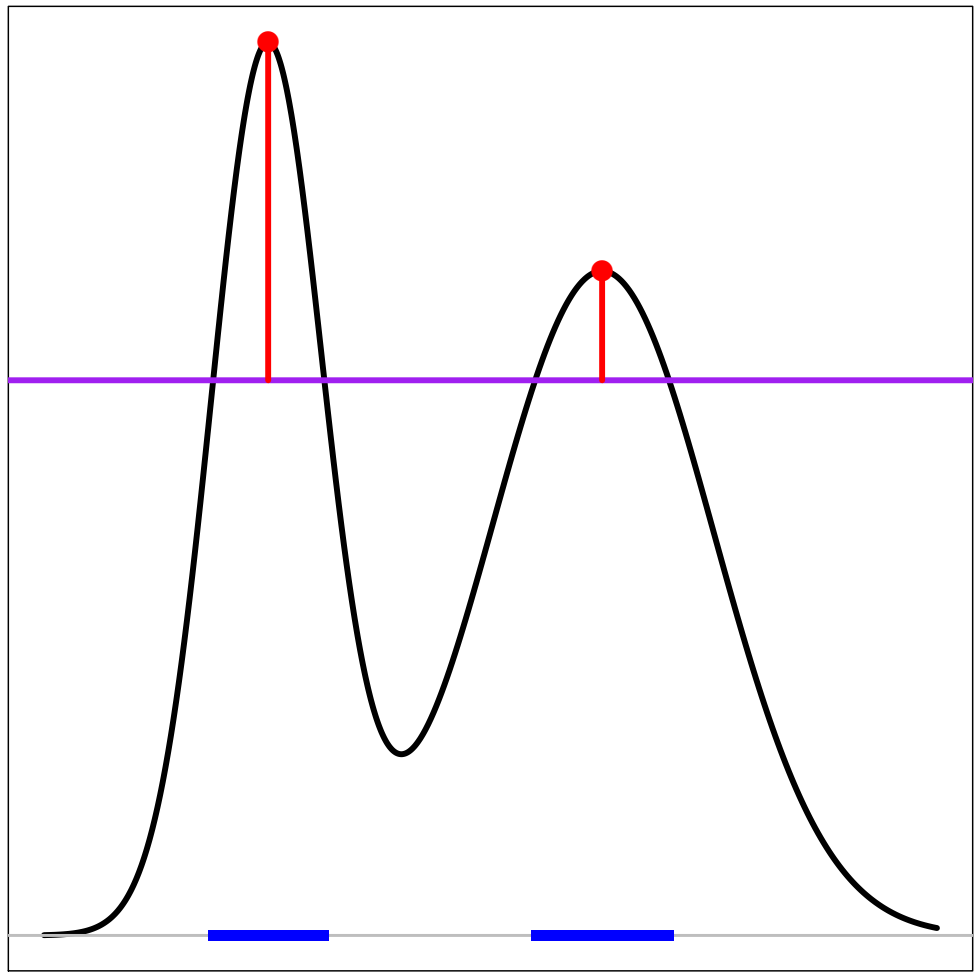
Cluster Tree - 3



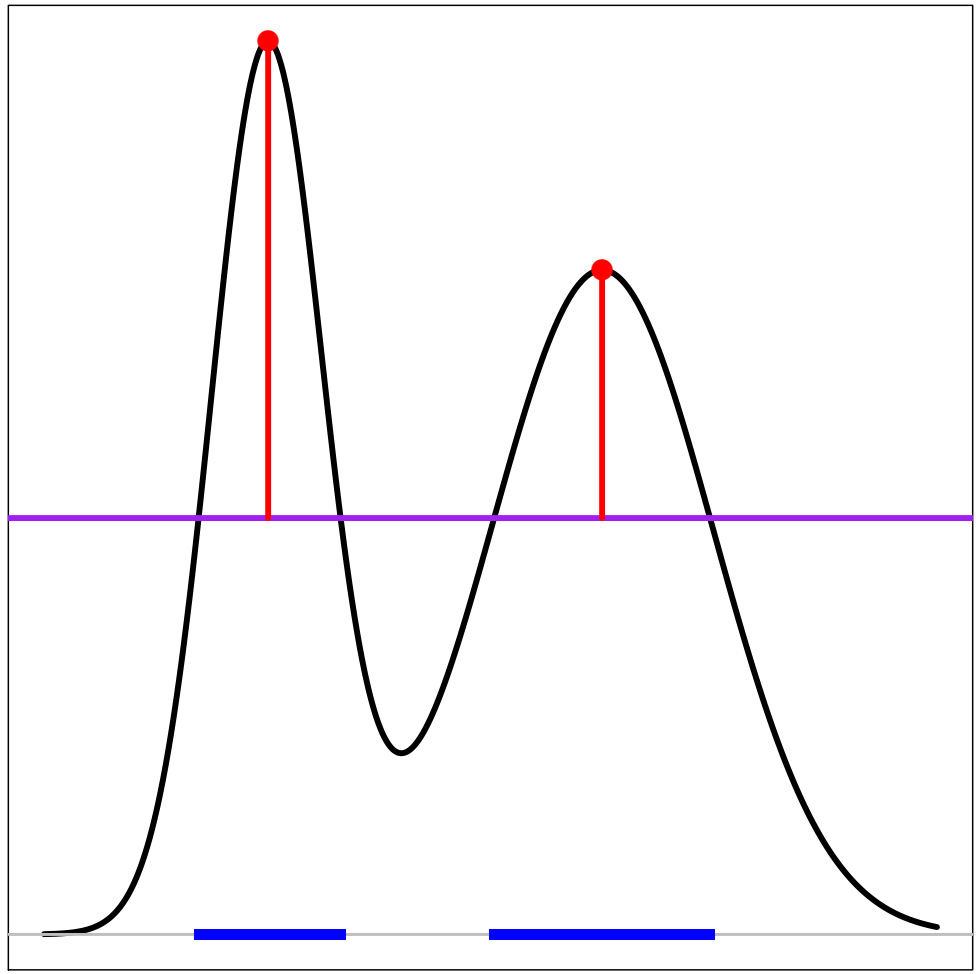
Cluster Tree - 3



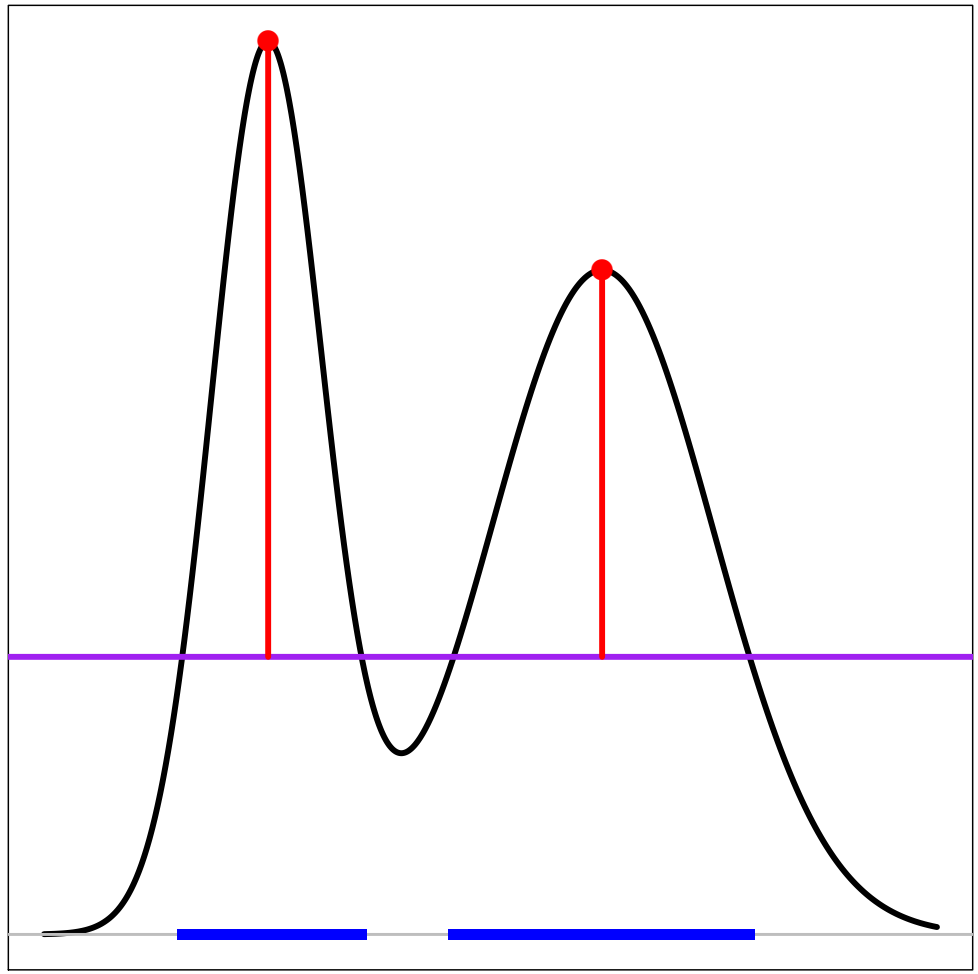
Cluster Tree - 3



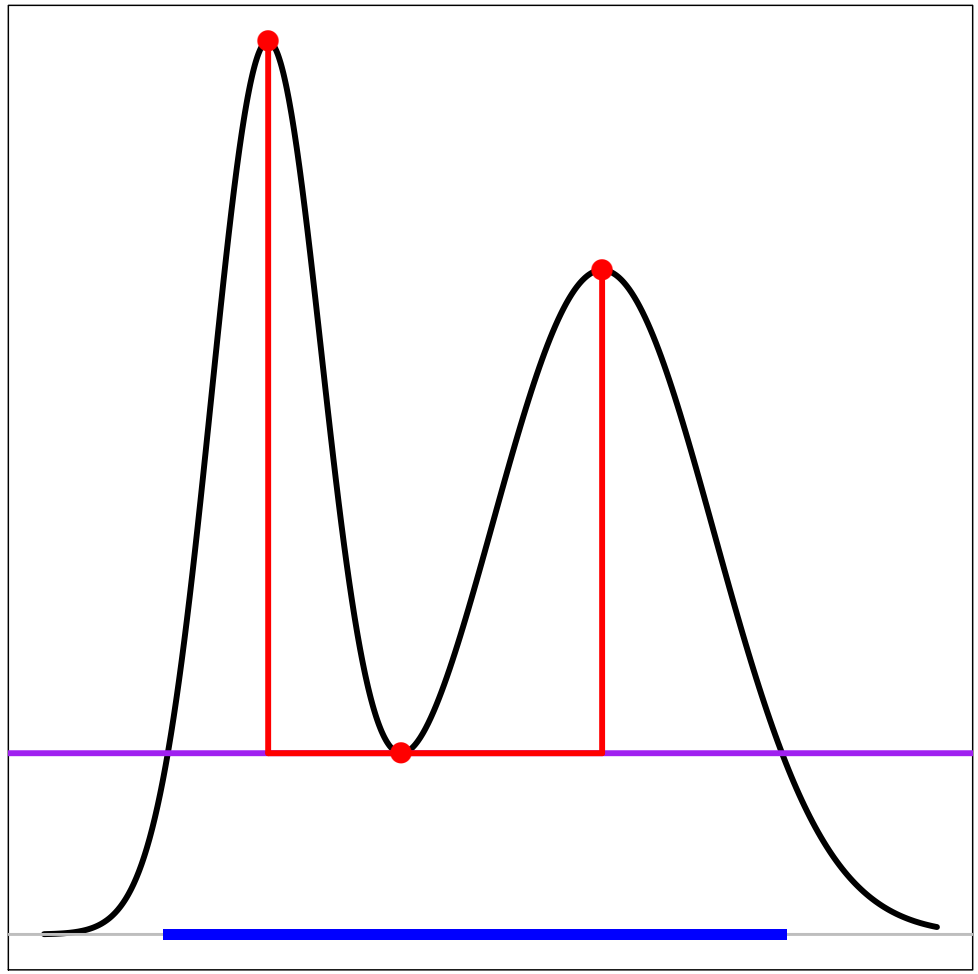
Cluster Tree - 3



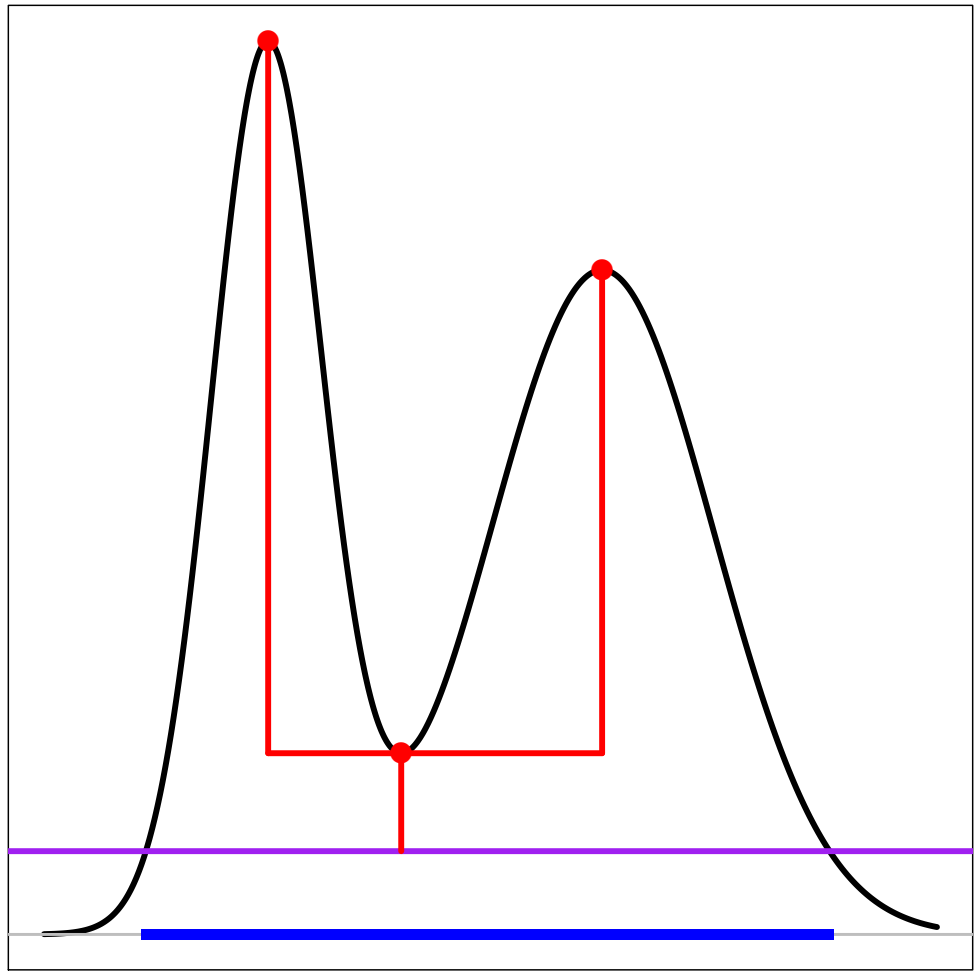
Cluster Tree - 3



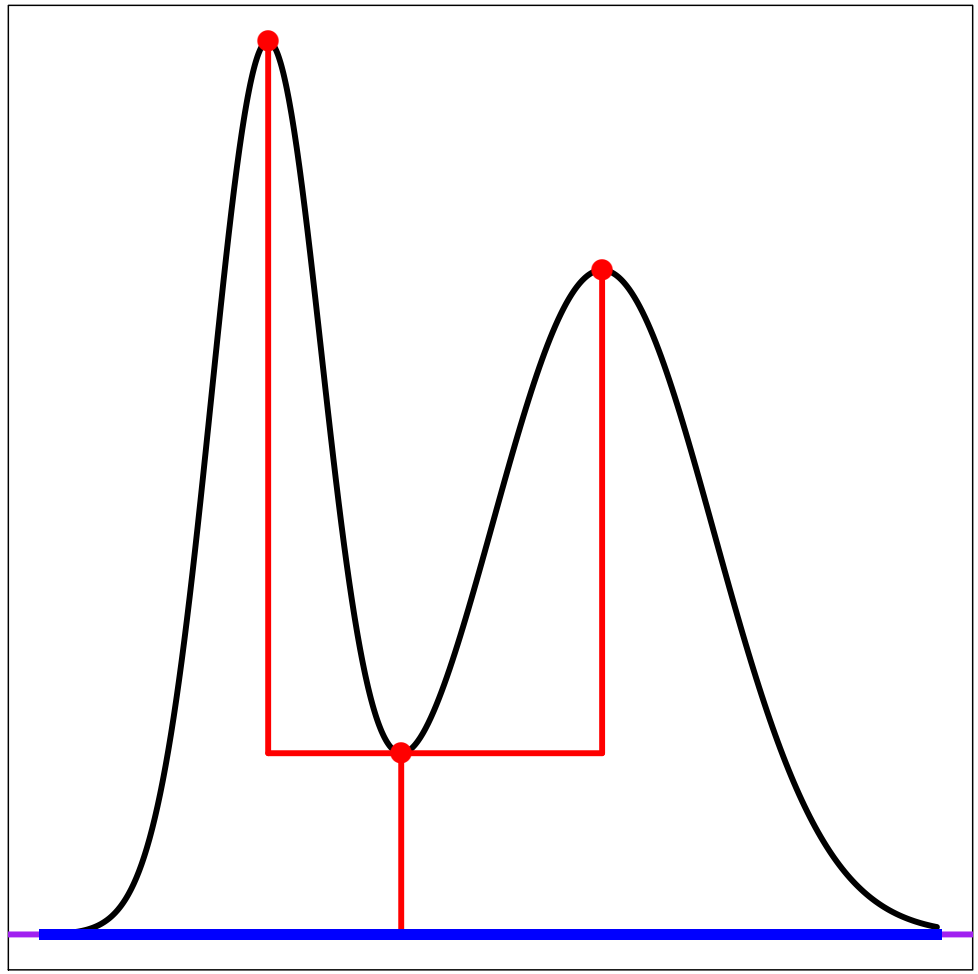
Cluster Tree - 3



Cluster Tree - 3



Cluster Tree - 3



Cluster Tree - 4

- Level sets are the basis of constructing a cluster tree.
- Local modes are associated to the creation of a new branch in a cluster tree.
- Saddle points or local minima are related to the elimination (merging) of a branch in a cluster tree.
- Cluster tree provides an elegant way to represent the shape of the PDF and can be used to visualize the data.

Cluster Tree - 5

Let T_p be the cluster tree based on the PDF and $\widehat{T}_p = T_{\widehat{p}_h}$ be the cluster tree based on the KDE.

- To measure the estimation error, a simple metric is

$$d_\infty(\widehat{T_p}, T_p) = \sup_x \|\widehat{p}_h(x) - p(x)\|,$$

so the convergence rate is

$$d_\infty(\widehat{T_p}, T_p) = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right).$$

Cluster Tree - 5

Let T_p be the cluster tree based on the PDF and $\widehat{T}_p = T_{\widehat{p}_h}$ be the cluster tree based on the KDE.

- To measure the estimation error, a simple metric is

$$d_\infty(\widehat{T_p}, T_p) = \sup_x \|\widehat{p}_h(x) - p(x)\|,$$

so the convergence rate is

$$d_\infty(\widehat{T_p}, T_p) = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right).$$

- Another way of defining statistical convergence is based on the probability

$$P_n = P\left(\widehat{T_p} \text{ and } T_p \text{ are topological equivalent}\right).$$

Cluster Tree - 5

Let T_p be the cluster tree based on the PDF and $\widehat{T}_p = T_{\widehat{p}_h}$ be the cluster tree based on the KDE.

- To measure the estimation error, a simple metric is

$$d_\infty(\widehat{T_p}, T_p) = \sup_x \|\widehat{p}_h(x) - p(x)\|,$$

so the convergence rate is

$$d_\infty(\widehat{T_p}, T_p) = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right).$$

- Another way of defining statistical convergence is based on the probability

$$P_n = P\left(\widehat{T_p} \text{ and } T_p \text{ are topological equivalent}\right).$$

- Under smoothness conditions and $n \to \infty, h \to 0$,

$$P_n \geq 1 - e^{-nh^{d+4} \cdot C_p},$$

for some constant C_p depending on the density function p.

Cluster Tree - 6

- There are other notions of convergence/consistency of a tree estimator.

Cluster Tree - 6

- There are other notions of convergence/consistency of a tree estimator.
- Convergence in the merge distortion metric (Eldridge et al. 2015) is one example.

Cluster Tree - 6

- There are other notions of convergence/consistency of a tree estimator.
- Convergence in the merge distortion metric (Eldridge et al. 2015) is one example.
- However, it was shown in Kim et al. (2016) that this metric is equivalent to the L_∞ metric.

Cluster Tree - 6

- There are other notions of convergence/consistency of a tree estimator.
- Convergence in the merge distortion metric (Eldridge et al. 2015) is one example.
- However, it was shown in Kim et al. (2016) that this metric is equivalent to the L_∞ metric.
- Hartigan consistency (Chaudhuri and Dasgupta 2010; Balakrishnan et al. 2013) is another way to measure the consistency of a tree estimator.

Cluster Tree - 6

- There are other notions of convergence/consistency of a tree estimator.
- Convergence in the merge distortion metric (Eldridge et al. 2015) is one example.
- However, it was shown in Kim et al. (2016) that this metric is equivalent to the L_∞ metric.
- Hartigan consistency (Chaudhuri and Dasgupta 2010; Balakrishnan et al. 2013) is another way to measure the consistency of a tree estimator.
- Note: cluster tree can also be recovered by a kNN approach; see Chaudhuri and Dasgupta (2010) and Chaudhuri et al. (2014) for more details.

Persistent Diagrams - 1

- Cluster trees contain only the information about connected components of level sets.
- Connected components are 0th order homology group.
- One can generalize this concept to higher order homology groups.
- The creation and elimination of homology groups can be summarized using the persistent diagram.

Persistent Diagrams - 2

- Again, we may define the persistent diagram formed by using the population PDF p and the KDE $\widehat{p}_h$.
- Let $\mathrm{PD} = \mathrm{PD}(p)$ be the persistent diagram formed by level sets of p and $\widehat{\mathrm{PD}} = \mathrm{PD}(\widehat{p}_h)$ be the one formed by level sets of $\widehat{p}_h$.
- Using the fact the bottleneck distance of persistent diagrams is bounded by the L_∞ distance of the generated function[7], we conclude that

$$d_B(\widehat{\mathrm{PD}}, \mathrm{PD}) \leq \|\widehat{p}_h - p\|_\infty = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^d}}\right).$$

[7] https://link.springer.com/article/10.1007/s00454-006-1276-5.

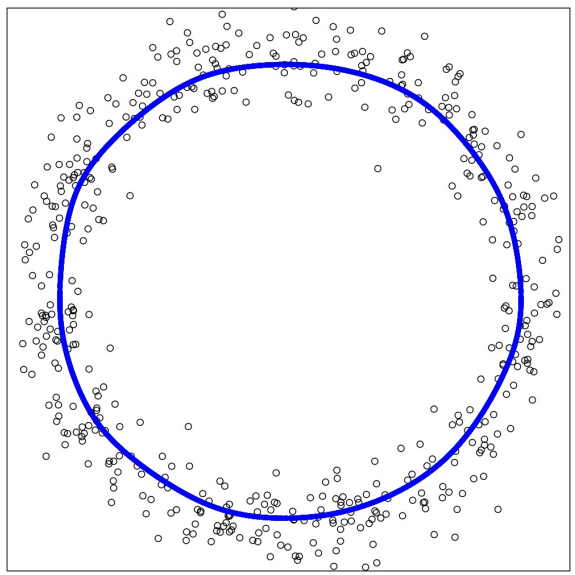
Persistent Diagrams - 3

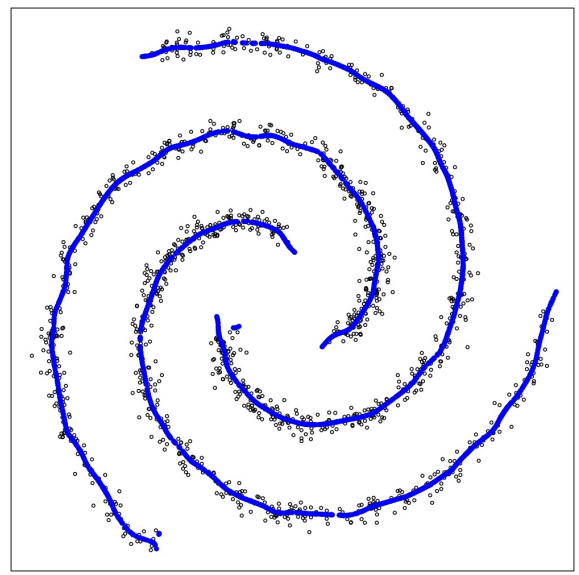
- There are other distance for persistent diagrams such as the Wasserstein distance.
- But the bottleneck distance has a nice property that it is an L_∞ type distance so we can use it to construct a confidence set[8].
- This is often done by bootstrapping the upper bound $\|\widehat{p}_h - p\|_\infty$ since it is unclear if bootstrapping the bottleneck distance will work or not.
- Also, computing the bottleneck distance is challenging.

[8] See **https://arxiv.org/abs/1303.7117** for more details..

Ridges - 1

- Ridges are another interesting geometric structure that we may want to study[9].
- They can be viewed as generalized local modes.





[9] Materials on this topic can be found in:
https://arxiv.org/abs/1406.5663 and
https://arxiv.org/abs/1212.5156

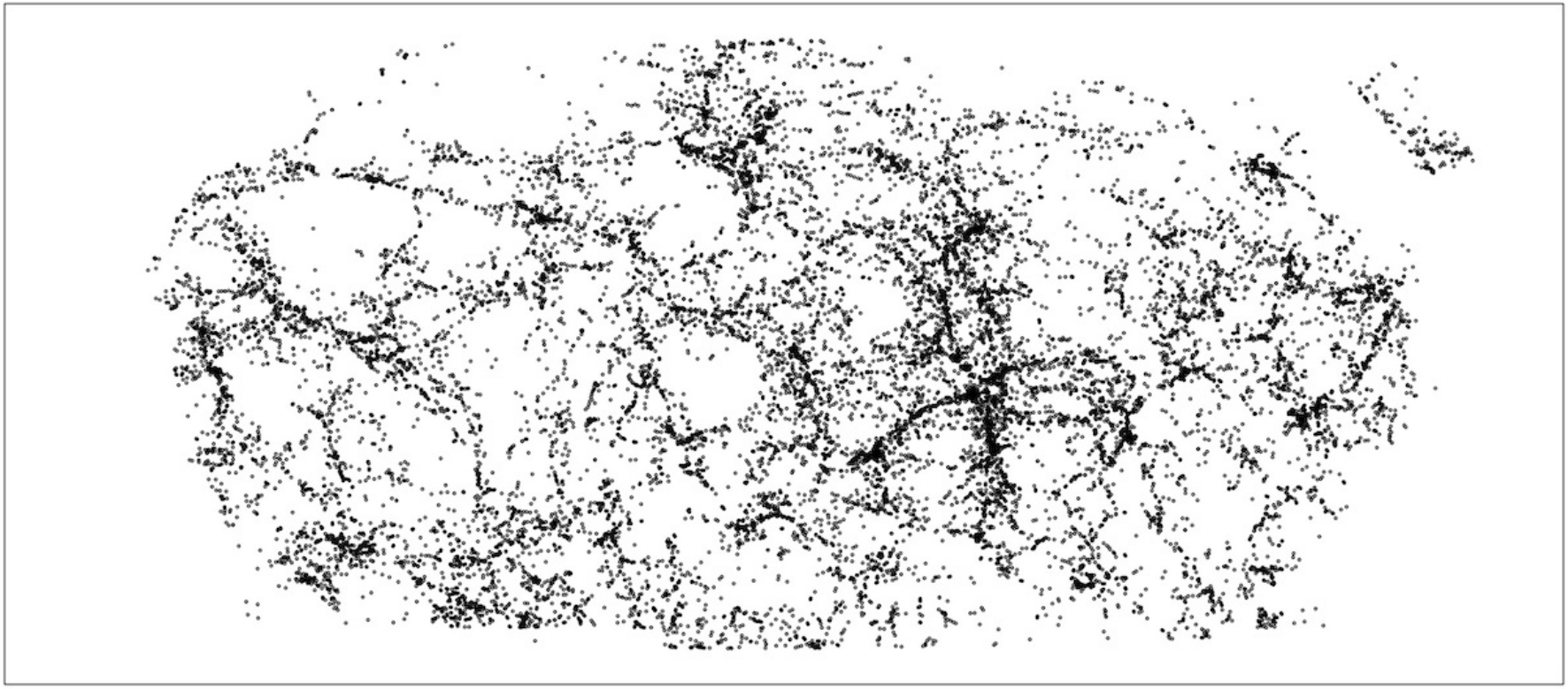
Ridges - 2

- Here is the formal definition of ridges.
- Let $v_1(x), \cdots, v_d(x)$ be the ordered eigenvectors of $\nabla\nabla p(x)$, where $v_1(x)$ corresponds to the largest eigenvalue.
- Define $V(x) = [v_2(x), \cdots, v_d(x)] \in \mathbb{R}^{d \times (d-1)}$.
- Ridges are defined as the collection:

$$R = \{x : V(x)V(x)^T \nabla p(x) = 0, \lambda_2(x) < 0\}.$$

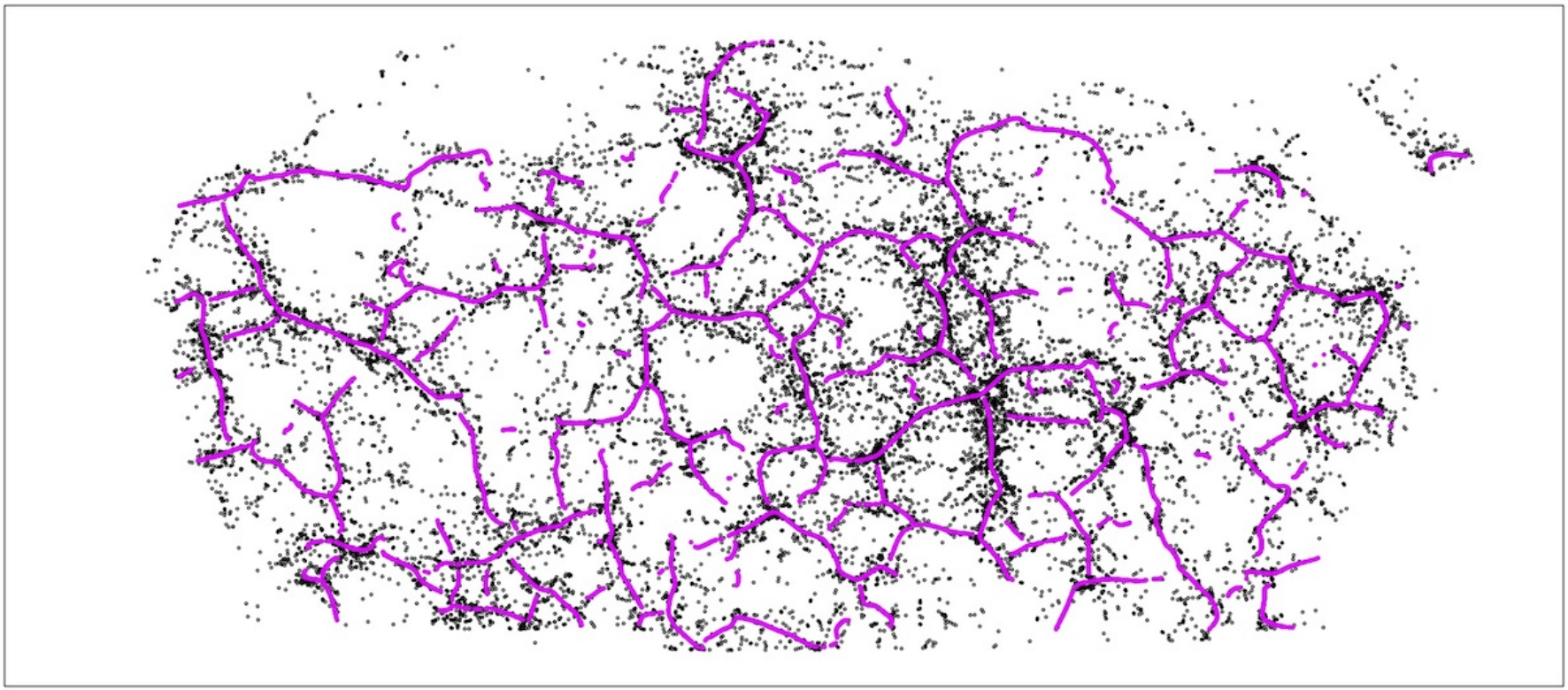
- $V(x)V(x)^T$ is the projection matrix onto the subspace spanned by $v_2(x), \cdots, v_d(x)$.
- Thus, the ridge R is the collection of *projected local modes*.

Ridges -3



An application of ridges in Astronomy.

Ridges -3



An application of ridges in Astronomy.

Ridges -4

- Ridges can be estimated by the KDE.
- Let $\widehat{V}_h(x)$ be the KDE version of $V(x)$ and $\widehat{\lambda}_2(x)$ be the KDE version of $\lambda_2(x)$. Then the ridge estimator is

$$\widehat{R}_h = \{x : \widehat{V}_h(x)\widehat{V}_h(x)^T \nabla \widehat{p}_h(x) = 0, \widehat{\lambda}_2(x) < 0\}.$$

- One can use the subspace constrained mean shift algorithm[10] to numerically calculate the estimator.
- The convergence rate is

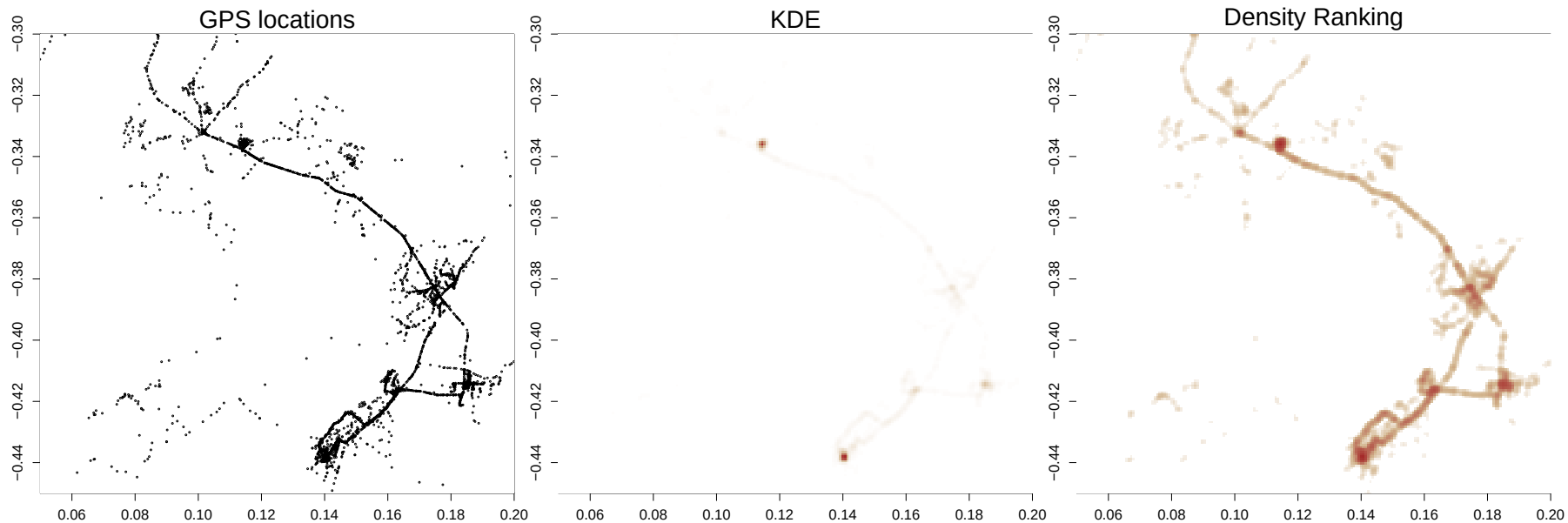
$$\text{Haus}(\widehat{R}_h, R) = O(\sup_x \|\nabla\nabla \widehat{p}_h(x) - \nabla\nabla p(x)\|_{\max}).$$

[10] See **http://www.jmlr.org/papers/v12/ozertem11a.html**.

Singular distribution

Failure of the KDE - 1

- In the previous few sections, we see that the KDE is a powerful tool.
- However, it may not work in certain situations.



Failure of the KDE - 2

- The KDE does not work because there is no underlying PDF for GPS data!
- A better model to describe a GPS data is the following distribution:

$$P_{\text{GPS}}(x) = \pi_0 P_0(x) + \pi_1 P_1(x) + \pi_2 P_2(x),$$

where $P_0(x)$ is a distribution of point mass, and $P_1(x)$ is a distribution of a $1D$ density function, and $P_2(x)$ is a distribution of a $2D$ density function, and $\pi_0 + \pi_1 + \pi_2 = 1$ with $\pi_j \geq 0$ are proportions.

- The components P_0 and P_1 make the distribution function singular so the KDE diverges.

Density Ranking - 1

- Although the KDE fails, the *density ranking* still works[11]!
- The density ranking is a density surrogate that

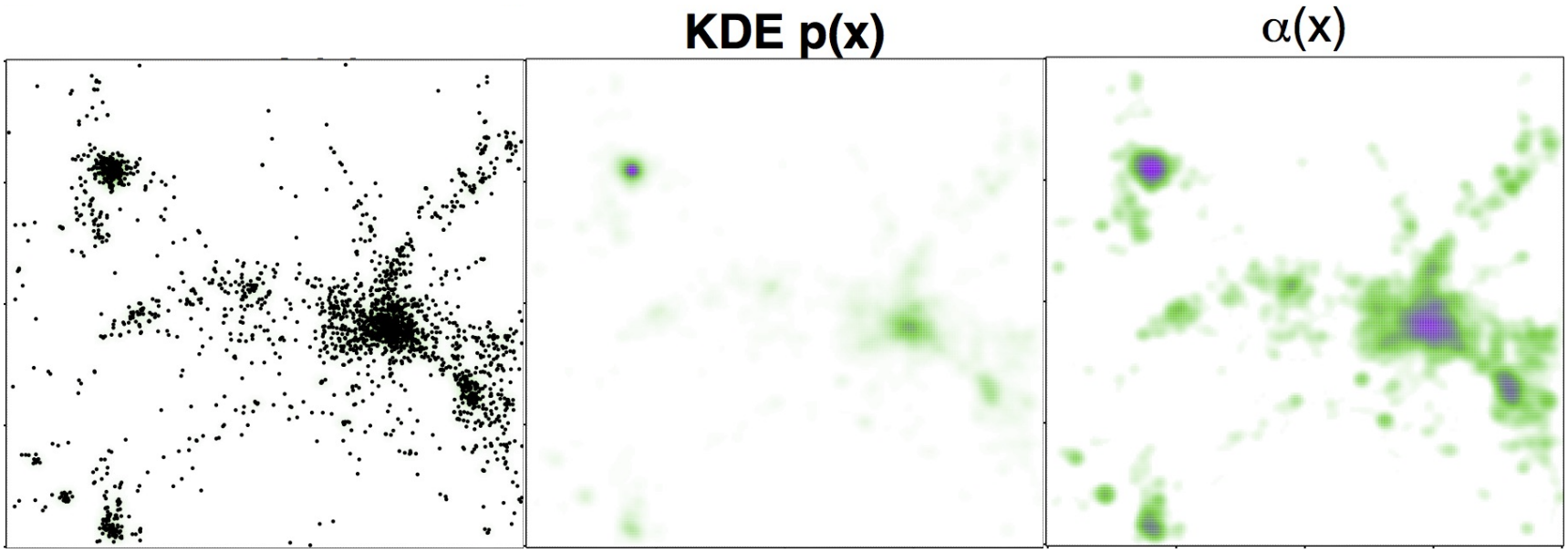
$$\widehat{\alpha}_h(x) = \frac{1}{n}\sum_{i=1}^{n} I(\widehat{p}_h(x) \geq \widehat{p}_h(X_i)).$$

- $\widehat{\alpha}_h(x)$ is the proportion of observations' density below the density of the point x.
- It preserves the ordering of $\widehat{p}_h(x)$ on each observed data points.
- $\widehat{\alpha}_h(x) \in [0, 1]$ so it will not diverge.

[11] See **https://arxiv.org/abs/1611.02762** and **https://arxiv.org/abs/1708.05017**.

Density Ranking - 2

Another example of possibly singular distribution in an Astronomy dataset.



Density Ranking - 3

- To see why the density ranking is stable in handling GPS data (or more general, singular distributions), we consider its population quantity.
- When the PDF exists, it is easy to see that the population quantity is

$$\alpha(x) = P(p(x) \geq p(X)),$$

  where $X \sim P$.
- When the distribution is singular, we need to use the concept of *Hausdorff density*.

Density Ranking - 4

- Let C_d be the volume of a d dimensional unit ball and $B(x, r) = \{y : \|x - y\| \leq r\}$.
- For any integer s, we define

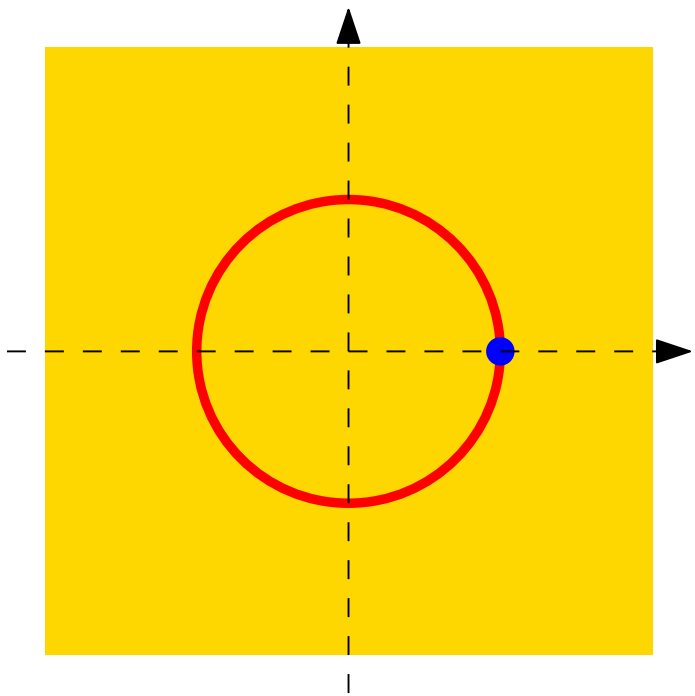
$$\mathcal{H}_s(x) = \lim_{r \to 0} \frac{P(B(x, r))}{C_s r^s}.$$

- $\mathcal{H}_s(x)$ occurs in three regimes: 0, ∞, or a number between $(0, \infty)$.
- Example of 0: $s = 1$ on a place with $2D$ density ($s <$ the structural dimension).
- Example of ∞: $s = 1$ on a point mass ($s >$ the structural dimension).
- For a point x, we then define

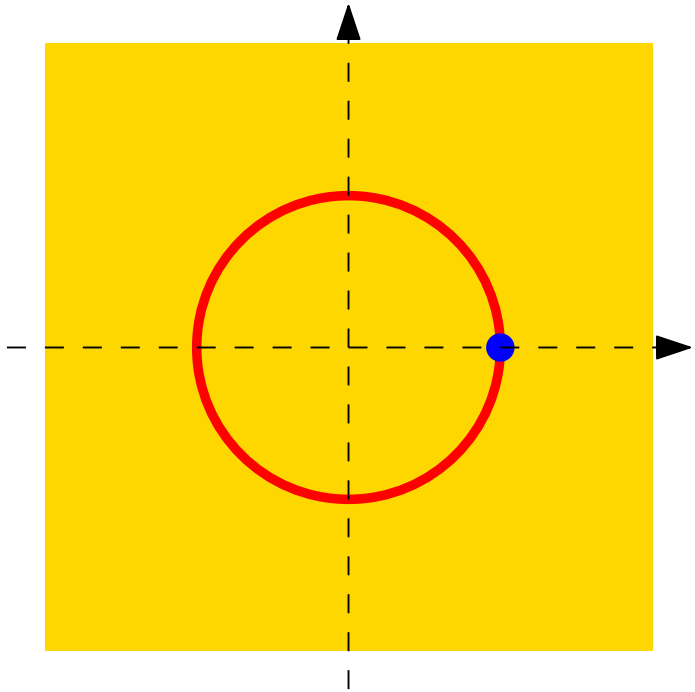
$$\tau(x) = \max\{s \leq d : \mathcal{H}_s(x) < \infty\}, \quad \rho(x) = \mathcal{H}_{\tau(x)}(x).$$

Density Ranking: 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:



Density Ranking: Example - 2



- Orange region: $\tau(x) = 2 \Leftrightarrow$ contribution of $P_2(x)$.
- Red region: $\tau(x) = 1 \Leftrightarrow$ contribution of $P_1(x)$.
- Blue region: $\tau(x) = 0 \Leftrightarrow$ contribution of $P_0(x)$.

Hausdorff Density and Ranking - 1

- The function $\tau(x)$ measures the dimension of P at point x.
- The function $\rho(x)$ describes the density of that corresponding dimension.
- We can 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.

Hausdorff Density and Ranking - 2

- 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

$$\alpha(x) = P(x \succeq_{d,p} X_1) = P(p(x) \geq p(X_1)),$$

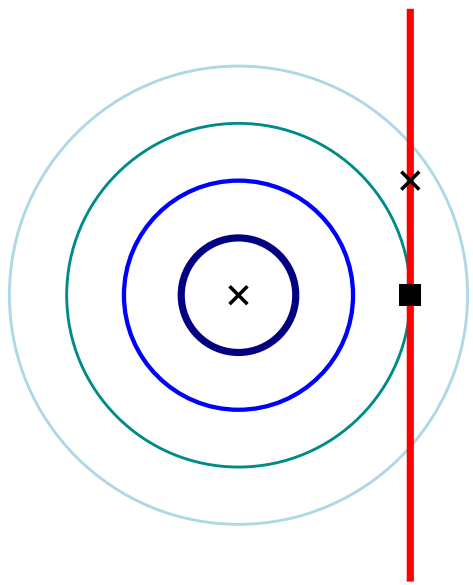
which recovers the definition in the simple case.

Dimensional Critical Points - 1

- In singular measure, there is a new type of critical points. We call them the *dimensional critical points*.
- These critical points contribute to the change of topology of level sets as the usual critical points but they cannot be defined by setting gradient to be 0.

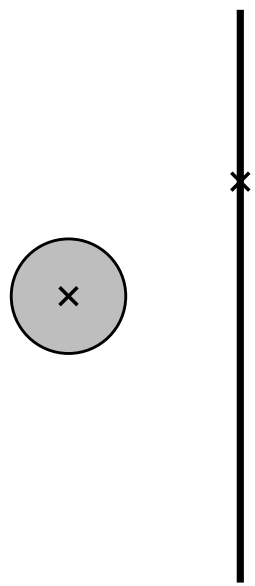
Dimensional Critical Points - 2

- The box in the following figure is a dimensional critical point.
- Note: this is a mixture of 2D distribution and a 1D distribution on the black line (maximum value occurs at the cross).



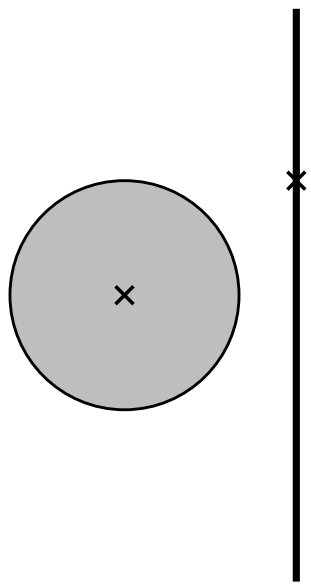
Dimensional Critical Points - 2

- The box in the following figure is a dimensional critical point.
- Note: this is a mixture of 2D distribution and a 1D distribution on the black line (maximum value occurs at the cross).



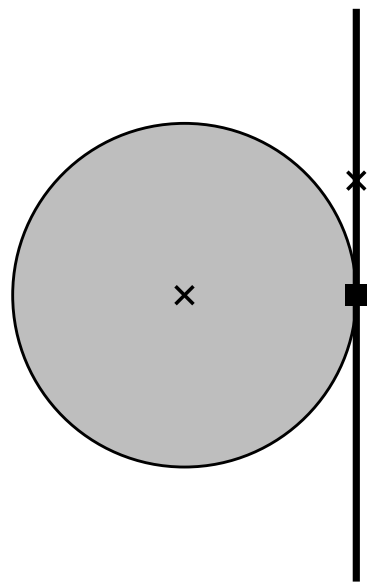
Dimensional Critical Points - 2

- The box in the following figure is a dimensional critical point.
- Note: this is a mixture of 2D distribution and a 1D distribution on the black line (maximum value occurs at the cross).



Dimensional Critical Points - 2

- The box in the following figure is a dimensional critical point.
- Note: this is a mixture of 2D distribution and a 1D distribution on the black line (maximum value occurs at the cross).



Convergence under Singular Measure: Density Ranking - 1

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

$$\int |\widehat{\alpha}_h(x) - \alpha(x)|^2 \, dP(x) = O(h) + O_P\left(\sqrt{\frac{1}{nh^d}}\right)$$

- Intuition of convergence: as $h \to 0$, the KDE

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

  diverges when x is in a lower dimensional structure ($\tau(x) < d$).
- The bias of order $O(h)$ is due to the smoothing from a nearby lower dimensional structure.
- However, the speed of diverging depends on $\tau(x)$. The smaller $\tau(x)$, the faster (actually the diverging rate is $O(h^{\tau(x)-d})$).
- Note: we do not estimate $\tau(x)$ when using $\widehat{\alpha}_h(x)$!

Convergence under Singular Measure: Density Ranking - 2

- Although we have $L_2(P)$ convergence (also we have L_2 and pointwise convergence), we do not have a uniform convergence.
- Example of non-convergence of supreme norm: consider a sequence of points on a higher dimensional space but moving toward a lower dimensional structure within distance $\frac{h}{2}$.
- Interestingly, we can still prove that some topological features (local modes, level sets, cluster trees, persistent diagrams) are converging.

Convergence under Singular Measure: Density Ranking - 3

- Although we do not have uniform convergence, many geometric structures still converge.
- The cluster tree, local modes, and 0-th order homology groups of density ranking converge to the population version.
- However, it is unclear if other quantities such as ridges or higher-order homology groups converge.

Conclusion

- In this lecture, we study the problem of estimating a geometric structure of the underlying PDF.
- We show that we can estimate the PDF, level sets, local modes, gradient systems, cluster trees, persistent diagrams, and ridges using the KDE.
- There are two alternative ways to view this problem.
- First, when the KDE converges to the true population PDF, many geometric structures also converge.
- Second, many geometric structures have an intrinsic stability with respect to the underlying function so small perturbation will not change it drastically.

Useful References

- Wasserman, Larry. **"Topological data analysis."** Annual Review of Statistics and Its Application 5 (2018): 501-532.
- Chen, Yen-Chi. **"A tutorial on kernel density estimation and recent advances."** Biostatistics & Epidemiology 1, no. 1 (2017): 161-187.
- **All of statistics: a concise course in statistical inference.** Larry Wasserman. Springer Science & Business Media, 2013.
- **All of nonparametric statistics.** Larry Wasserman. Springer, 2006.
- **Multivariate density estimation: theory, practice, and visualization.** David Scott. John Wiley & Sons, 2015.