Statistical Inference for Shards

Yen-Chi Chen

Christopher R. Genovese   Larry Wasserman

Department of Statistics
Carnegie Mellon University

May 27, 2015

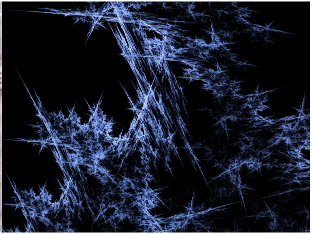
Outline

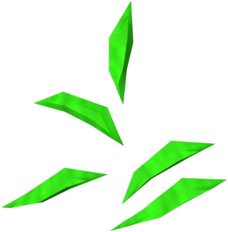
- Introduction to Shards
- Density Level Set
- Density Ridges
- Modal Regression
- Summary

Outline

- Introduction to Shards
- Density Level Set
- Density Ridges
- Modal Regression
- Summary

What are Shards?

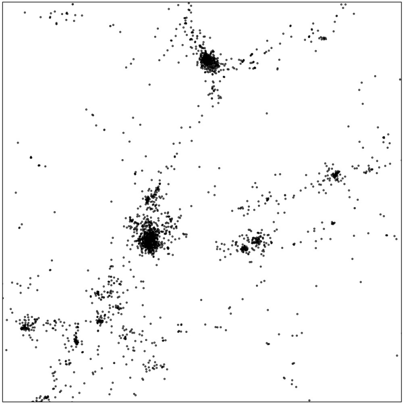




Source: odysseyseaglass.com, nsudino, the RuneScape Wiki

What are Shards?

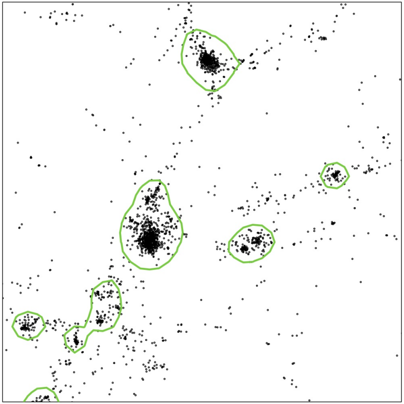
- Shards: small regions with high density.



Data: Massive Blackhole-2 Simulation

What are Shards?

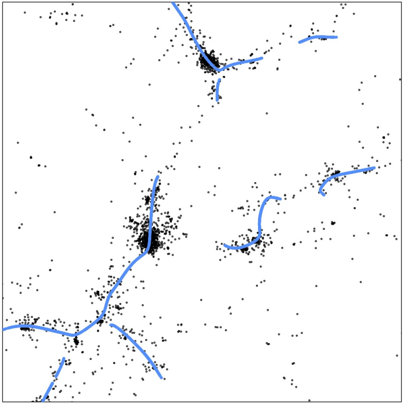
- Shards: small regions with high density.



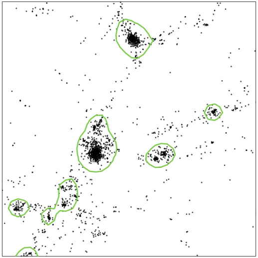
Data: Massive Blackhole-2 Simulation

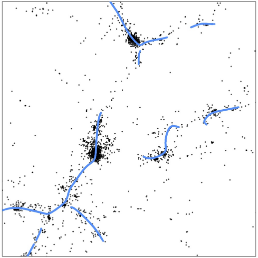
What are Shards?

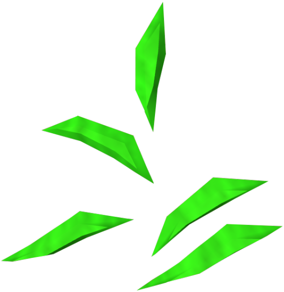
- Shards: small regions with high density.



Data: Massive Blackhole-2 Simulation







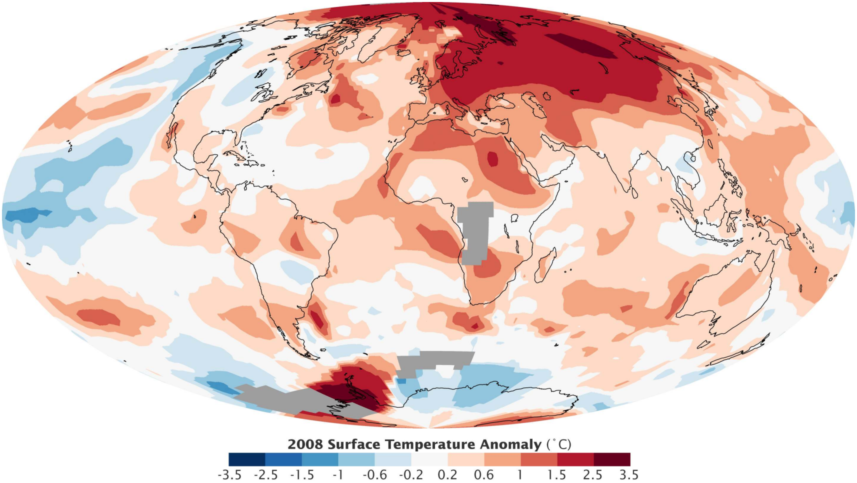
- Shards are sets, whose parameters space has infinite dimensions.
- Making inference for sets is very tough.
- There are many estimation methods but very few of them mentioned statistical inference.

- Shards are sets, whose parameters space has infinite dimensions.
- Making inference for sets is very tough.
- There are many estimation methods but very few of them mentioned statistical inference.
- $\rightarrow$ In this talk, we will see how one can make inference for sets.

Outline

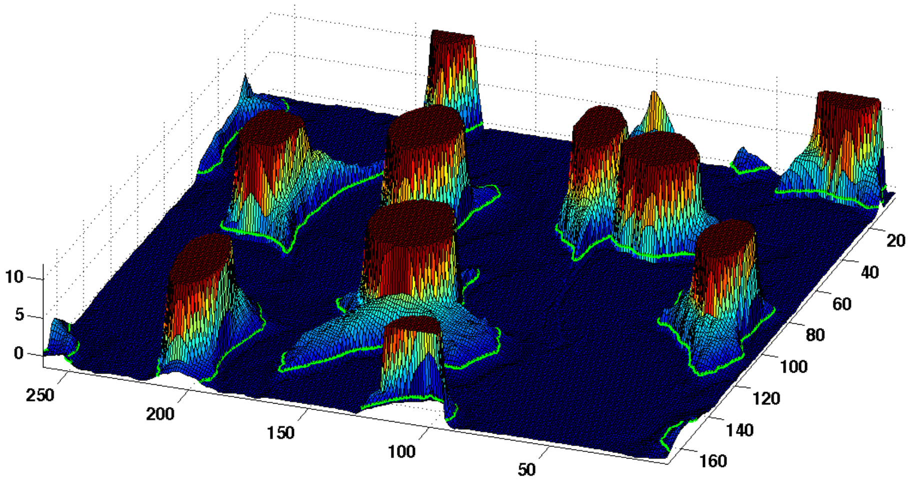
- Introduction to Shards
- Density Level Set
- Density Ridges
- Modal Regression
- Summary

Example: Climate Data



Source: NASA-GISS

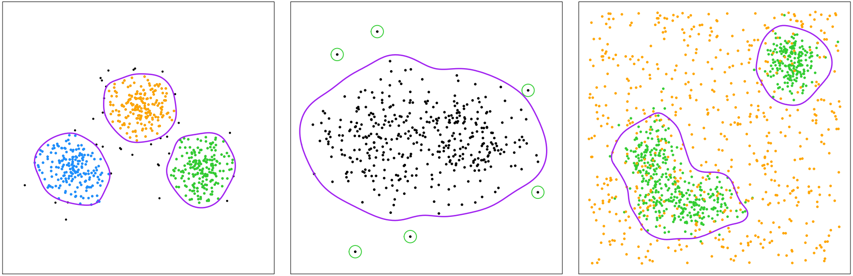
Example: Neuro Image



Source: http://neuroncyto.bii.a-star.edu.sg/

Density Level Set

- Density Level Set: The collection of points where the density is exactly at certain level.
- Applications: clustering, anomaly detection, classification, two-sample comparison



Formal Definition for Density Level Set

Let p be the probability density function.

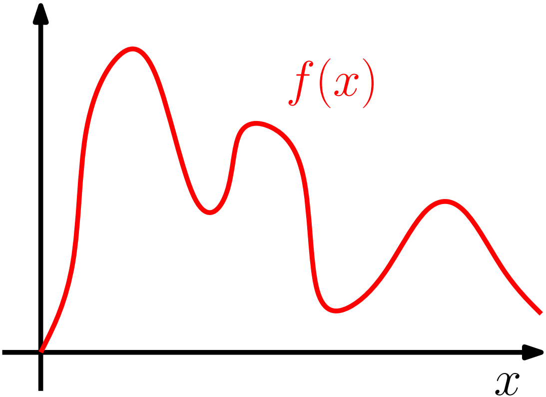
Formal Definition for Density Level Set

Let p be the probability density function.

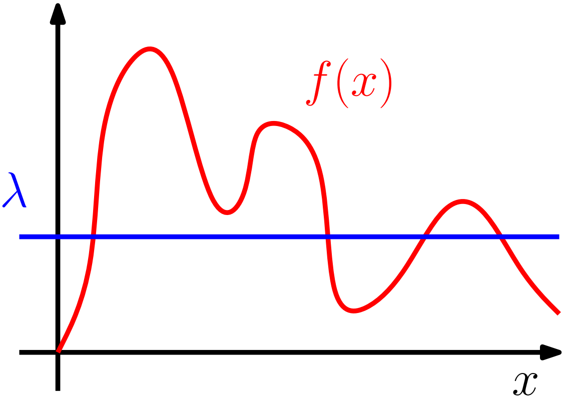
- The λ-level set is

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

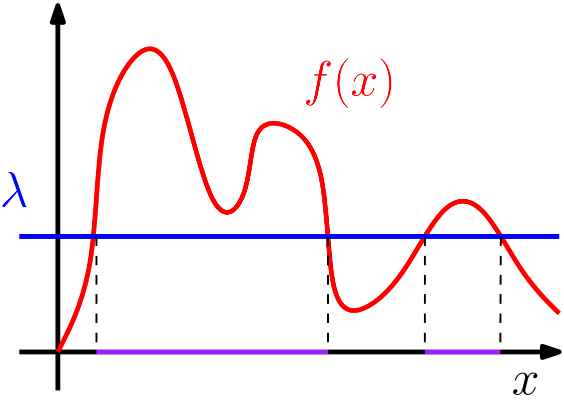
Example for Level Set



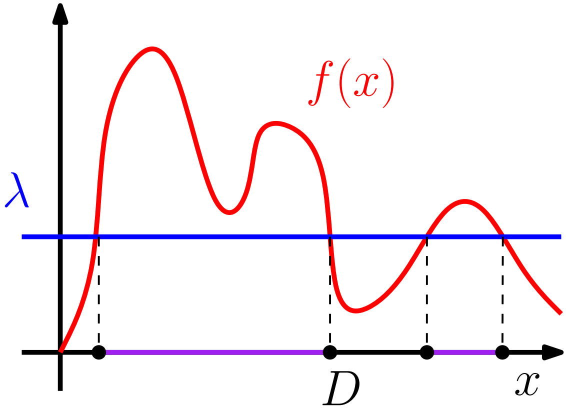
Example for Level Set



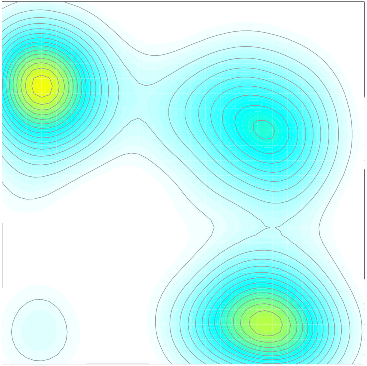
Example for Level Set

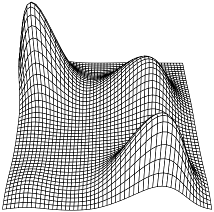


Example for Level Set

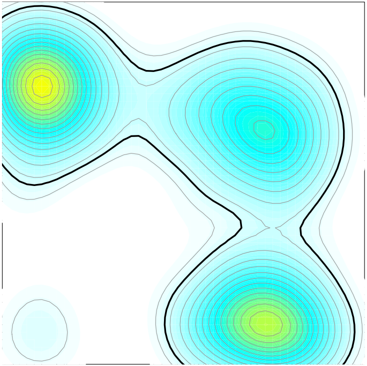


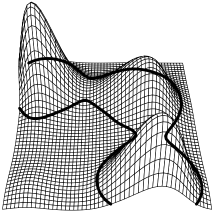
Example for Level Set





Example for Level Set





Plug-in Estimator

Our estimator: a plug-in from the Kernel Density Estimator (KDE).

Plug-in Estimator

Our estimator: a plug-in from the Kernel Density Estimator (KDE).

- The KDE $\widehat{p}_n$

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

Plug-in Estimator

Our estimator: a plug-in from the Kernel Density Estimator (KDE).

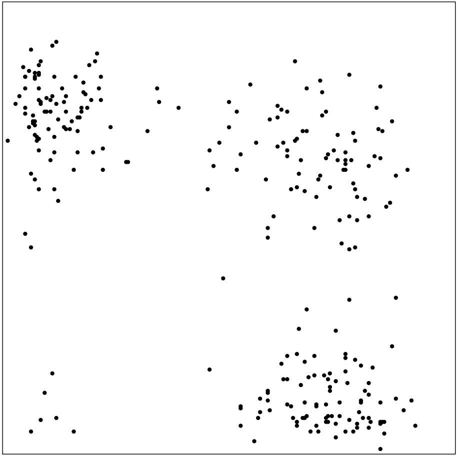
- The KDE $\widehat{p}_n$

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

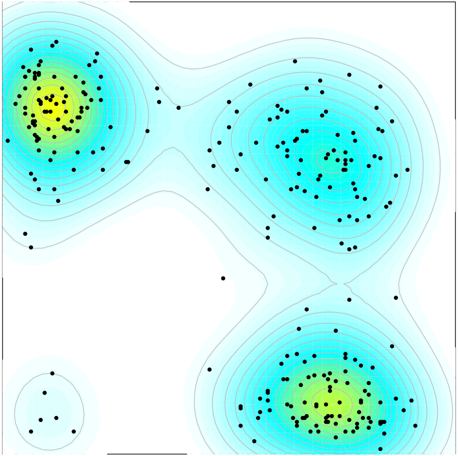
- The corresponding estimators

$$\widehat{D}_n = \{x : \widehat{p}_n(x) = \lambda\}.$$

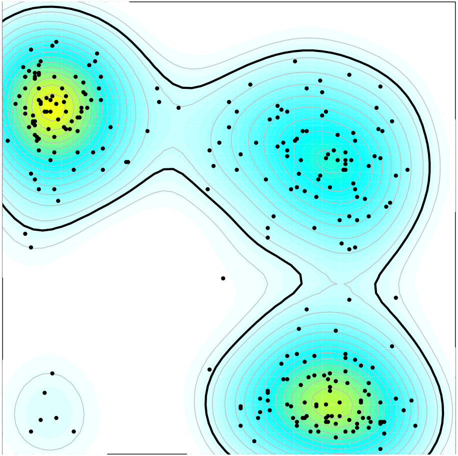
Example: Level Set Estimator



Example: Level Set Estimator



Example: Level Set Estimator



Smoothed Level Set

In particular, we focus on making inference for the smoothed version of the density, denoted as p_h:

$$p_h(x) = p \otimes K_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right), \quad K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right),$$

where $\otimes$ denotes the convolution.

- We define D_h as the level set using p_h.

Smoothed Level Set

In particular, we focus on making inference for the smoothed version of the density, denoted as p_h:

$$p_h(x) = p \otimes K_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right), \quad K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right),$$

where $\otimes$ denotes the convolution.

- We define D_h as the level set using p_h.
- The advantages for focusing on D_h:
    - Always well-defined.
    - Topologically similar.
    - Asymptotically the same.
    - Fast rate of convergence.

Smoothed Level Set

In particular, we focus on making inference for the smoothed version of the density, denoted as p_h:

$$p_h(x) = p \otimes K_h(x) = \mathbb{E}\left(\widehat{p}_n(x)\right), \quad K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right),$$

where $\otimes$ denotes the convolution.

- We define D_h as the level set using p_h.
- The advantages for focusing on D_h:
  - Always well-defined.
  - Topologically similar.
  - Asymptotically the same.
  - Fast rate of convergence.
- One can always slightly undersmooth so that inference for D_h is asymptotically valid for D.

Useful Metric: Hausdorff Distance

We introduce a useful metric–*the Hausdorff distance* for sets:

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

where $d(x, A) = \inf_{y\in A}\|x - y\|$ is the projection distance.

Useful Metric: Hausdorff Distance

We introduce a useful metric–*the Hausdorff distance* for sets:

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

where $d(x, A) = \inf_{y\in A} \|x - y\|$ is the projection distance.

- Haus is an $\mathcal{L}_\infty$ norm for sets.
- Consistency: $\text{Haus}(\widehat{D}_n, D_h) = o_{\mathbb{P}}(1)$.

Useful Metric: Hausdorff Distance

We introduce a useful metric–*the Hausdorff distance* for sets:

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

where $d(x, A) = \inf_{y\in A} \|x - y\|$ is the projection distance.

- Haus is an $\mathcal{L}_\infty$ norm for sets.
- Consistency: $\mathrm{Haus}(\widehat{D}_n, D_h) = o_{\mathbb{P}}(1)$.
- Useful property:

$$A \subset B \oplus \mathrm{Haus}(A, B), \quad B \subset A \oplus \mathrm{Haus}(A, B),$$

where $A \oplus r = \{x : d(x, A) \leq r\}$.

Hausdorff Distance and Confidence Sets

- Hausdorff distance can be applied to construct confidence sets.

Hausdorff Distance and Confidence Sets

- Hausdorff distance can be applied to construct confidence sets.
- Let F_n be the CDF for $\mathrm{Haus}(\widehat{D}_n, D_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

Hausdorff Distance and Confidence Sets

- Hausdorff distance can be applied to construct confidence sets.
- Let F_n be the CDF for $\text{Haus}(\widehat{D}_n, D_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.
- It can be shown that

$$\mathbb{P}\left(D_h \subset \widehat{D}_n \oplus t_{1-\alpha}\right) \geq 1-\alpha.$$

$\rightarrow$ This follows from the property

$$A \subset B \oplus \text{Haus}(A,B), \quad B \subset A \oplus \text{Haus}(A,B).$$

Hausdorff Distance and Confidence Sets

- Hausdorff distance can be applied to construct confidence sets.
- Let F_n be the CDF for $\text{Haus}(\widehat{D}_n, D_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.
- It can be shown that

$$\mathbb{P}\left(D_h \subset \widehat{D}_n \oplus t_{1-\alpha}\right) \geq 1-\alpha.$$

$\rightarrow$ This follows from the property

$$A \subset B \oplus \text{Haus}(A,B), \quad B \subset A \oplus \text{Haus}(A,B).$$

- We need to find the distribution F_n.

Asymptotic Theory

It can be shown that

$$\sqrt{nh^d}\text{Haus}(\widehat{D}_n, D_h) \approx \sup\{\text{Empirical process}\} \approx \sup\{\text{Gaussian process}\}.$$

$\rightarrow$ the last approximation follows from [Chernozhukov et. al. 2014].

Asymptotic Theory

It can be shown that

$$\sqrt{nh^d}\mathrm{Haus}(\widehat{D}_n, D_h) \approx \sup\left\{\text{Empirical process}\right\} \approx \sup\left\{\text{Gaussian process}\right\}.$$

$\rightarrow$ the last approximation follows from [Chernozhukov et. al. 2014].

Theorem

Under regularity condition, there exists a tight Gaussian process $\mathbb{B}$ defined on a certain function space $\mathcal{F}$ such that

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^d}\mathrm{Haus}(\widehat{D}_n, D_h) < t\right) - \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbb{B}(f)| < t\right)\right| = O\left(\left(\frac{\log^7 n}{nh^d}\right)^{1/8}\right).$$

The Bootstrap

- Good news: we have the asymptotic behavior.
- Bad news: the asymptotic behavior is complicated.

The Bootstrap

- Good news: we have the asymptotic behavior.
- Bad news: the asymptotic behavior is complicated.
- A solution: the bootstrap.

The Bootstrap Consistency

- Bootstrap sample $\Longrightarrow$ bootstrap level set $\widehat{D}_n^*$.
- Compute Haus$(\widehat{D}_n^*, \widehat{D}_n)$ to get a CDF estimator $\widehat{F}_n$.
- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.

The Bootstrap Consistency

- Bootstrap sample $\Longrightarrow$ bootstrap level set $\widehat{D}_n^*$.
- Compute Haus$(\widehat{D}_n^*, \widehat{D}_n)$ to get a CDF estimator $\widehat{F}_n$.
- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.
- It can be shown that

$$\sqrt{nh^d}\text{Haus}(\widehat{D}_n^*, \widehat{D}_n) \approx \sup\{\text{Gaussian process}\} \approx \sqrt{nh^d}\text{Haus}(\widehat{D}_n, D).$$

The Bootstrap Consistency

- Bootstrap sample $\Longrightarrow$ bootstrap level set $\widehat{D}_n^*$.
- Compute Haus$(\widehat{D}_n^*, \widehat{D}_n)$ to get a CDF estimator $\widehat{F}_n$.
- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.
- It can be shown that

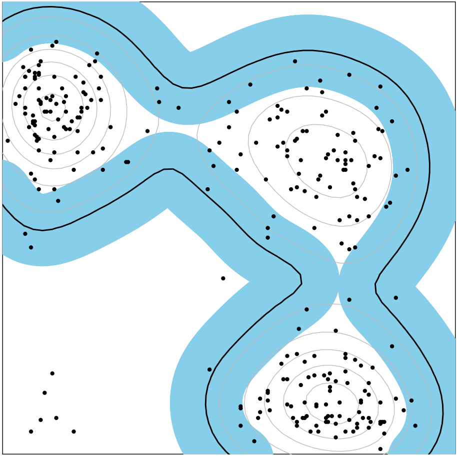
$$\sqrt{nh^d}\text{Haus}(\widehat{D}_n^*, \widehat{D}_n) \approx \sup\{\text{Gaussian process}\} \approx \sqrt{nh^d}\text{Haus}(\widehat{D}_n, D).$$

Theorem

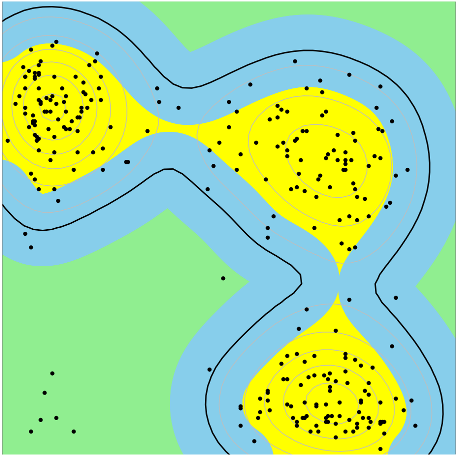
Under regularity condition,

$$\mathbb{P}\left(D_h \subset \widehat{D}_n \oplus \widehat{t}_{1-\alpha}\right) = 1-\alpha + O\left(\left(\frac{\log^7 n}{nh^d}\right)^{1/8}\right).$$

Example: Confidence Sets

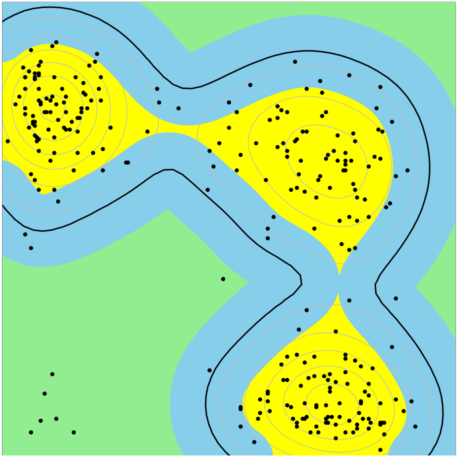


Example: Confidence Sets



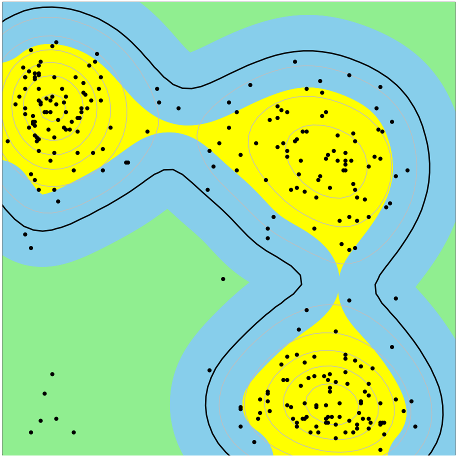
Properties for the Confidence Sets

1. Blue: confidence sets for D_h



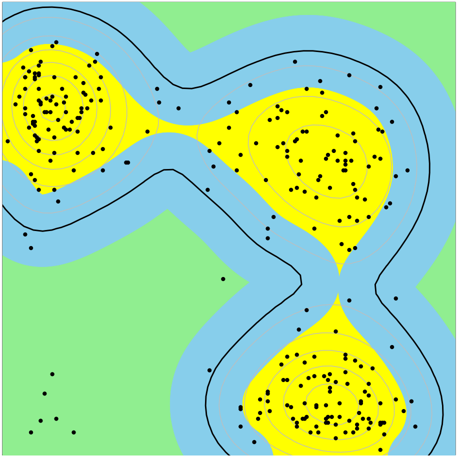
Properties for the Confidence Sets

1. Blue: confidence sets for D_h
2. Yellow: every point above λ



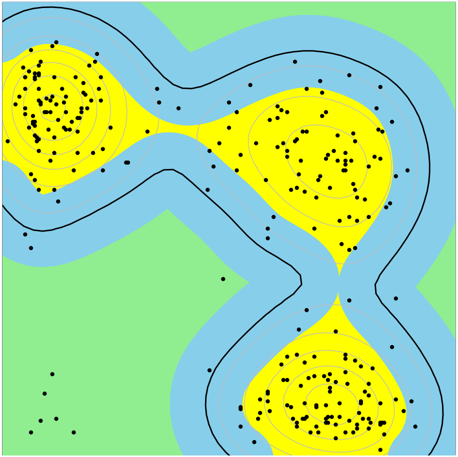
Properties for the Confidence Sets

1. Blue: confidence sets for D_h
2. Yellow: every point above λ
3. Green: every point below λ



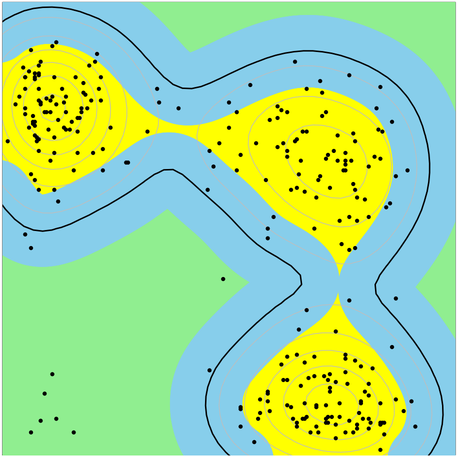
Properties for the Confidence Sets

1. Blue: confidence sets for D_h
2. Yellow: every point above λ
3. Green: every point below λ
4. Yellow+Blue: confidence sets for upper level set



Properties for the Confidence Sets

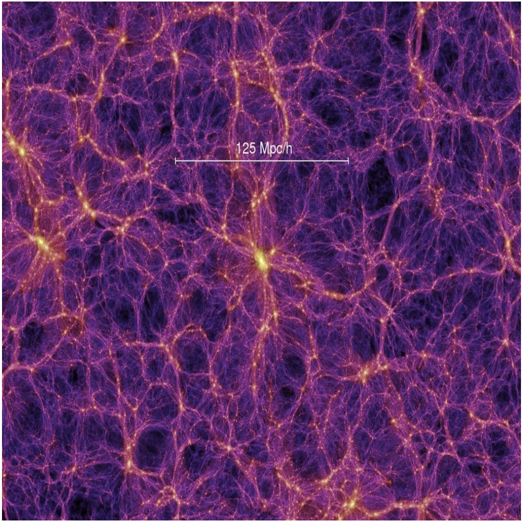
1. Blue: confidence sets for D_h
2. Yellow: every point above λ
3. Green: every point below λ
4. Yellow+Blue: confidence sets for upper level set
5. Green+Blue: confidence sets for lower level set



Outline

- Introduction to Shards
- Density Level Set
- Density Ridges
- Modal Regression
- Summary

Example: Cosmology



Credit: Millennium Simulation

Example: Neuroscience

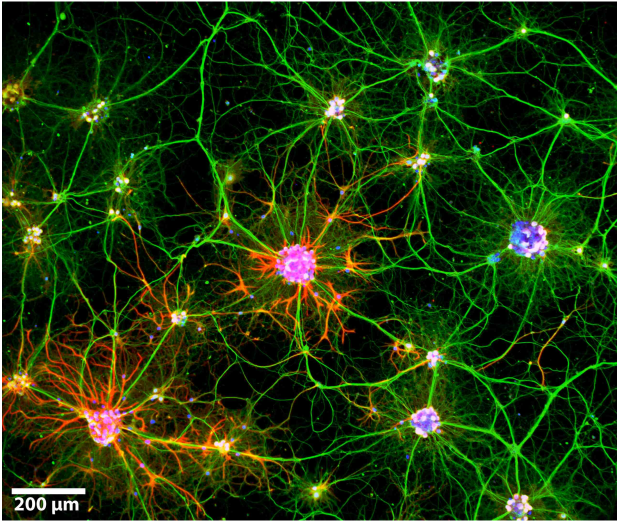


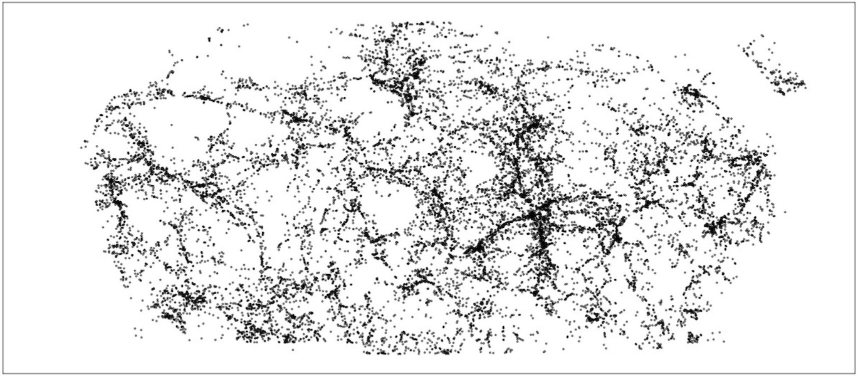
Image courtesy Eswar P. R. Iyer.

Density Ridges

- In the above examples, we see curve-like structure with high density.
- This structure can be captured by the *density ridges*.

Density Ridges

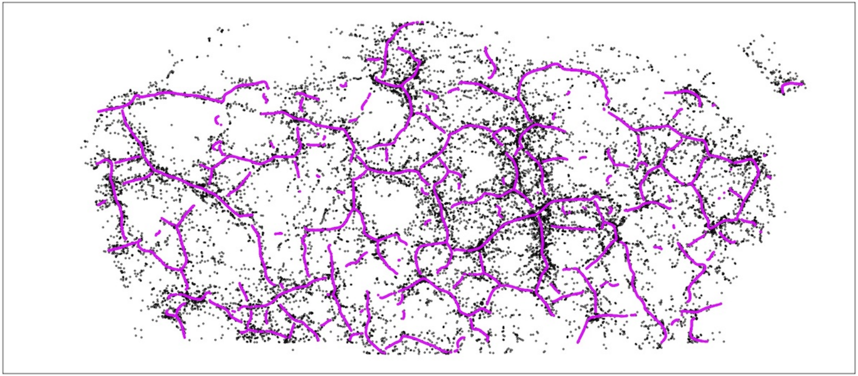
- In the above examples, we see curve-like structure with high density.
- This structure can be captured by the *density ridges*.



- Data: the Sloan Digital Sky Survey.

Density Ridges

- In the above examples, we see curve-like structure with high density.
- This structure can be captured by the *density ridges*.

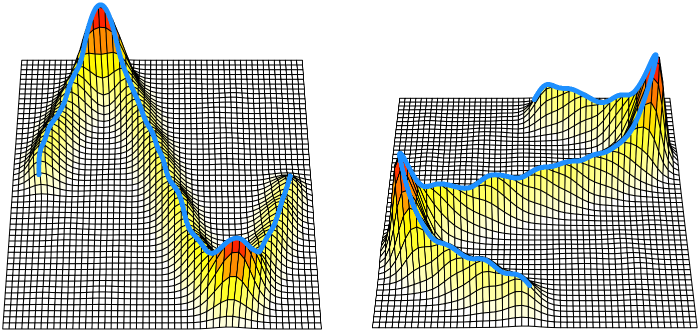


- Data: the Sloan Digital Sky Survey.

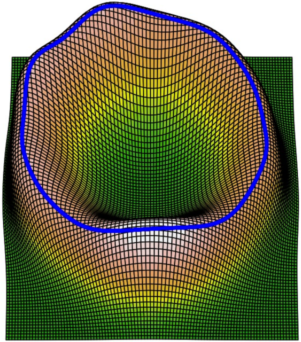
Example: Ridges in Mountains

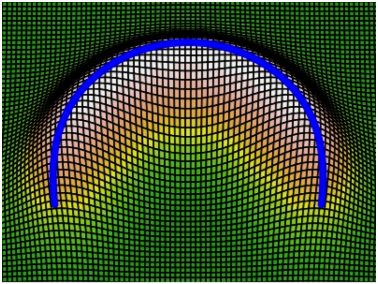
Credit: Google

Example: Ridges in Smooth Functions

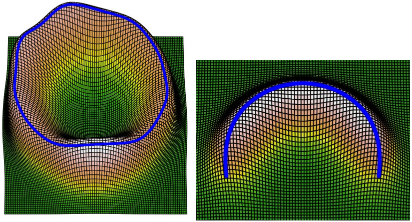


Example: Ridges in Smooth Functions



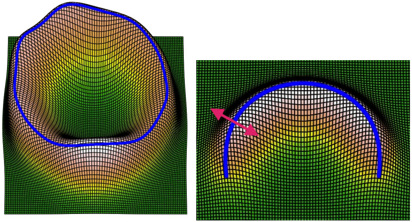


Ridges: Local Modes in Subspace



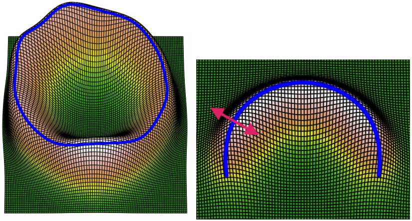
- A generalized local mode in a specific 'subspace'.

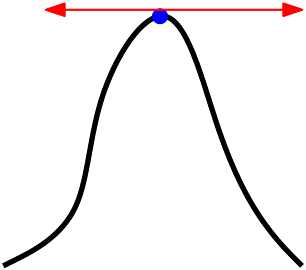
Ridges: Local Modes in Subspace



- A generalized local mode in a specific 'subspace'.

Ridges: Local Modes in Subspace





- A generalized local mode in a specific 'subspace'.

Formal Definition of Density Ridges

- $p(x)$: a density function.

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors
- $V(x)V(x)^T$: a projection

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors
- $V(x)V(x)^T$: a projection
- Ridges:

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

Formal Definition of Density Ridges

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.
- $V(x) = [v_2(x), \cdots, v_d(x)]$: matrix of 2nd to last eigenvectors
- $V(x)V(x)^T$: a projection
- Ridges:

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

- Local modes:

$$\text{Mode}(p) = \{x : \nabla p(x) = 0, \lambda_1(x) < 0\}.$$

Estimator and Algorithm

We use the plug-in estimate:

$$\widehat{R}_n = \text{Ridge}(\widehat{p}_n),$$

where $\widehat{p}_n$ is the KDE.

Estimator and Algorithm

We use the plug-in estimate:

$$\widehat{R}_n = \text{Ridge}(\widehat{p}_n),$$

where $\widehat{p}_n$ is the KDE.

- In general, finding ridges from a given function is hard.

Estimator and Algorithm

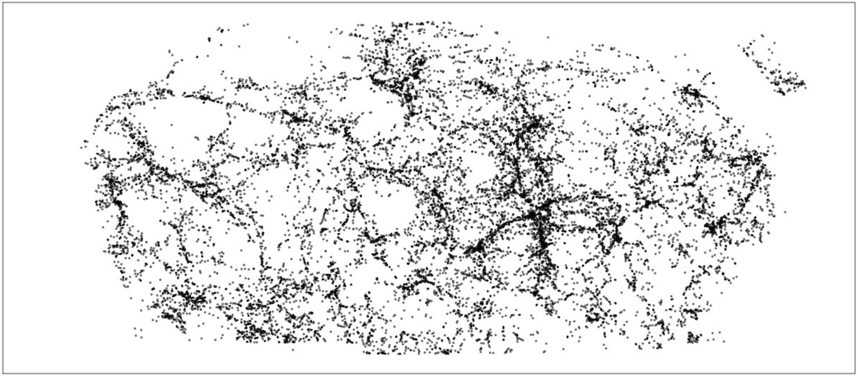
We use the plug-in estimate:

$$\widehat{R}_n = \text{Ridge}(\widehat{p}_n),$$

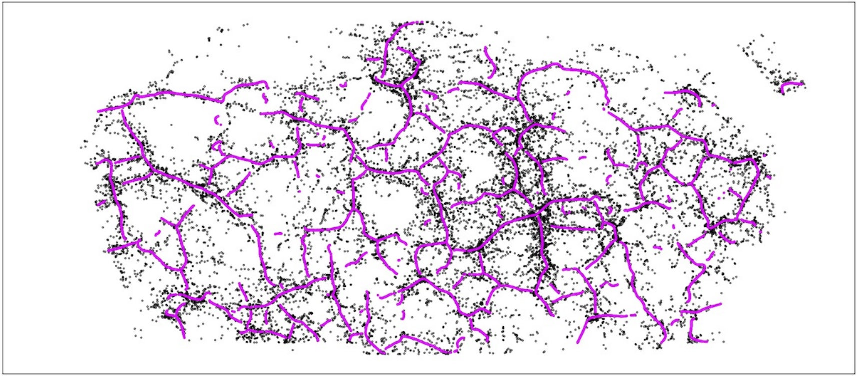
where $\widehat{p}_n$ is the KDE.

- In general, finding ridges from a given function is hard.
- The Subspace Constraint Mean Shift (SCMS; Ozertem2011) algorithm allows us to find $\widehat{R}_n$, the ridges of the KDE.

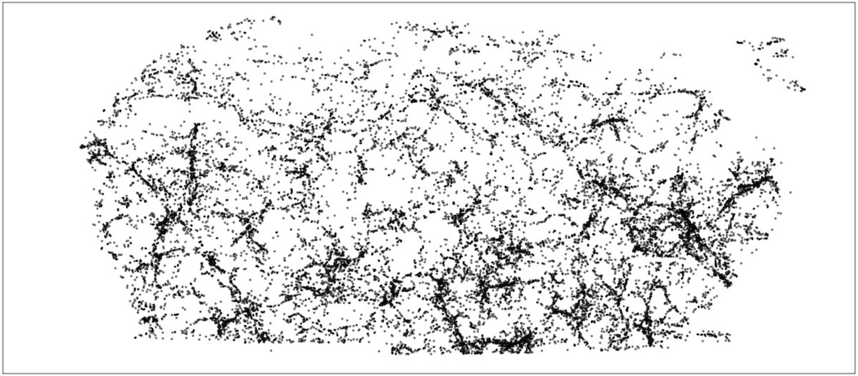
Example for Estimated Density Ridges



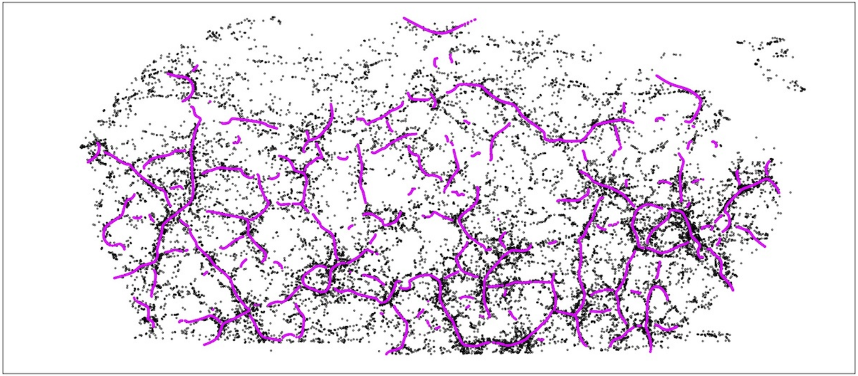
Example for Estimated Density Ridges



Example for Estimated Density Ridges



Example for Estimated Density Ridges



Asymptotic Theory and Statistical Inference

- Can we derive asymptotic theory and make statistical inference for density ridges?

Asymptotic Theory and Statistical Inference

- Can we derive asymptotic theory and make statistical inference for density ridges?
- Yes! We can make it by the similar trick to the level sets.

Asymptotic Theory and Statistical Inference

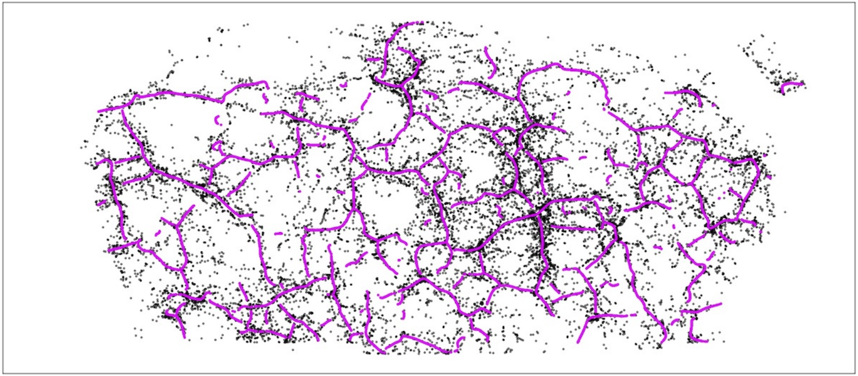
- Can we derive asymptotic theory and make statistical inference for density ridges?
- Yes! We can make it by the similar trick to the level sets.

Theorem

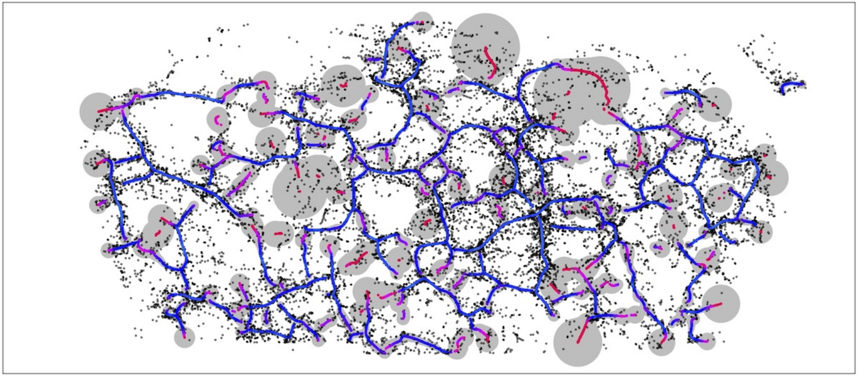
Under regularity condition,

- $\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_n, R_h) \approx \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$ *for certain function space* $\mathcal{F}$*.*
- $\widehat{R}_n \oplus \widehat{t}_{1-\alpha}$ *is an asymptotic valid confidence set for* R_h*.*
- *Note:* $R_h =$*Ridge*(p_h) *is the ridges for smoothed density* p_h*.*

Example for Confidence Sets



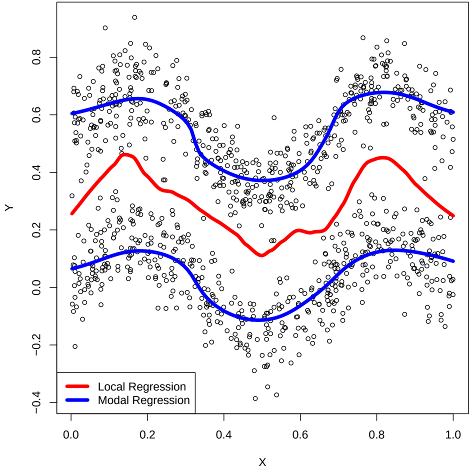
Example for Confidence Sets



Outline

- Introduction to Shards
- Density Level Set
- Density Ridges
- Modal Regression
- Summary

Motivating Examples for Modal Regression



This is a joint work with Ryan J. Tibshirani

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

- Modal function–the conditional (local) **modes**:

$$M(x) = \text{Mode}(Y|X = x) = \left\{ y : \frac{d}{dy}p(y|x) = 0, \frac{d^2}{dy^2}p(y|x) < 0 \right\}.$$

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

- Modal function–the conditional (local) **modes**:

$$M(x) = \text{Mode}(Y|X = x) = \left\{ y : \frac{d}{dy}p(y|x) = 0, \frac{d^2}{dy^2}p(y|x) < 0 \right\}.$$

- Equivalently,

$$M(x) = \left\{ y : \frac{\partial}{\partial y}p(x, y) = 0, \frac{\partial^2}{\partial y^2}p(x, y) < 0 \right\}.$$

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

- Modal function–the conditional (local) **modes**:

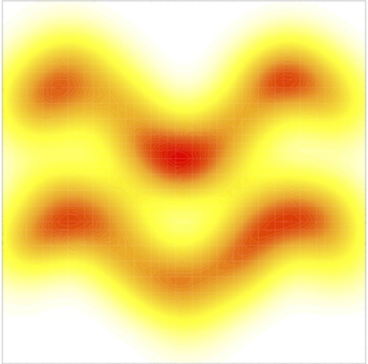
$$M(x) = \text{Mode}(Y|X = x) = \left\{ y : \frac{d}{dy} p(y|x) = 0, \frac{d^2}{dy^2} p(y|x) < 0 \right\}.$$

- Equivalently,

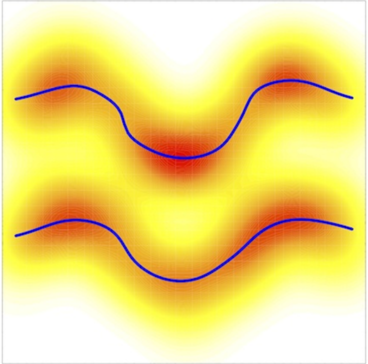
$$M(x) = \left\{ y : \frac{\partial}{\partial y} p(x, y) = 0, \frac{\partial^2}{\partial y^2} p(x, y) < 0 \right\}.$$

- $M(x)$ is a multi-value function.
- M is called modal manifolds (curves).

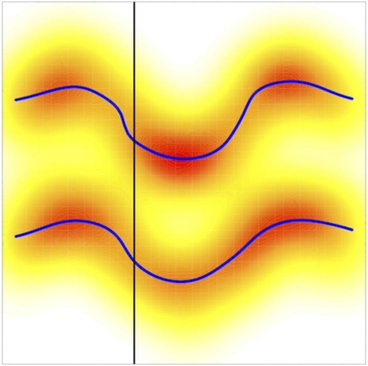
Conditional Local Modes



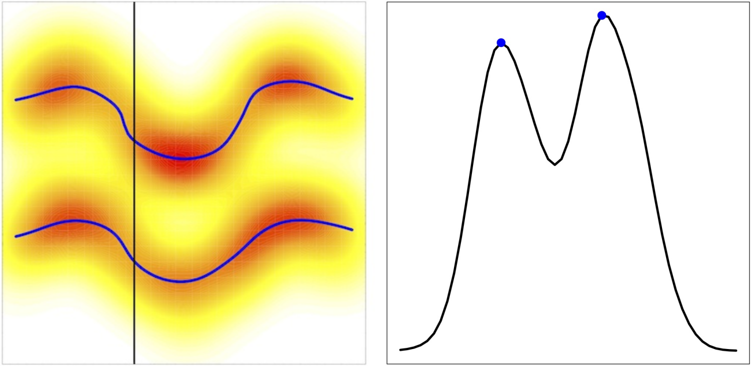
Conditional Local Modes



Conditional Local Modes



Conditional Local Modes



Estimator for Modal Regression

- Our estimator is the plug-in from the KDE:

$$\widehat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y}\widehat{p}_n(x, y) = 0, \frac{\partial^2}{\partial y^2}\widehat{p}(x, y) < 0 \right\}.$$

Estimator for Modal Regression

- Our estimator is the plug-in from the KDE:

$$\widehat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y}\widehat{p}_n(x, y) = 0, \frac{\partial^2}{\partial y^2}\widehat{p}(x, y) < 0 \right\}.$$

- Finding conditional local modes is hard in general.

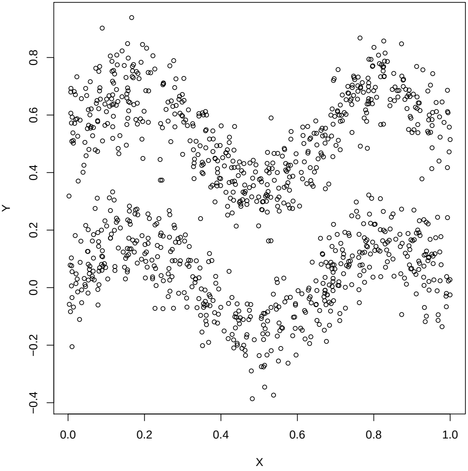
Estimator for Modal Regression

- Our estimator is the plug-in from the KDE:

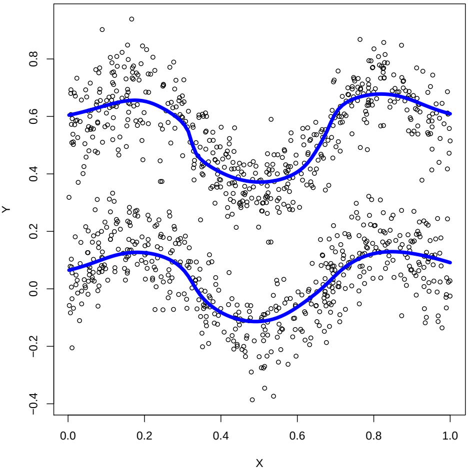
$$\widehat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y}\widehat{p}_n(x,y) = 0, \frac{\partial^2}{\partial y^2}\widehat{p}(x,y) < 0 \right\}.$$

- Finding conditional local modes is hard in general.
- Partial mean shift: a simple algorithm for computing $\widehat{M}_n(x)$, the plug-in estimator of the KDE, from the data (Einbeck et. al. 2006).

Example for Modal Regression



Example for Modal Regression



Confidence Sets

- Let M_h be the modal manifolds for p_h.
- Define a uniform metric $\Delta_n = \sup_x \mathrm{Haus}(\widehat{M}_n(x), M_h(x))$.

Confidence Sets

- Let M_h be the modal manifolds for p_h.
- Define a uniform metric $\Delta_n = \sup_x \text{Haus}(\widehat{M}_n(x), M_h(x))$.

Theorem

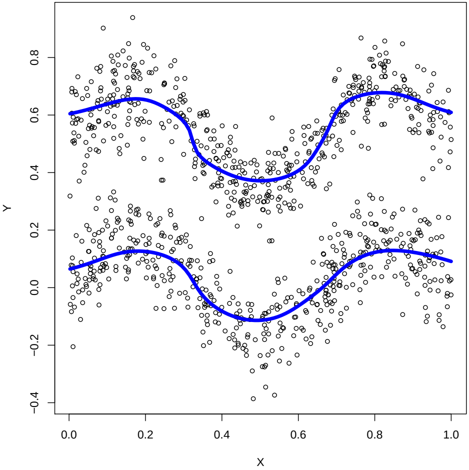
Under regularity condition,

- $\sqrt{nh^{d+3}}\Delta_n \approx \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$ *for certain function space* $\mathcal{F}$.
- *The set*

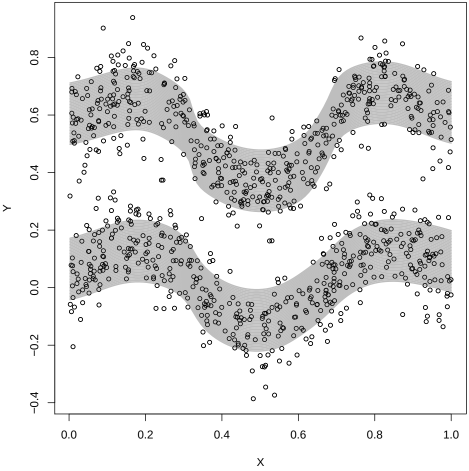
$$\left\{(x,y) : y \in \widehat{M}_n(x) \oplus \widehat{t}_{1-\alpha}, x \in \mathbb{K}\right\}$$

is an asymptotic valid confidence set for M_h.

Example for Confidence Sets

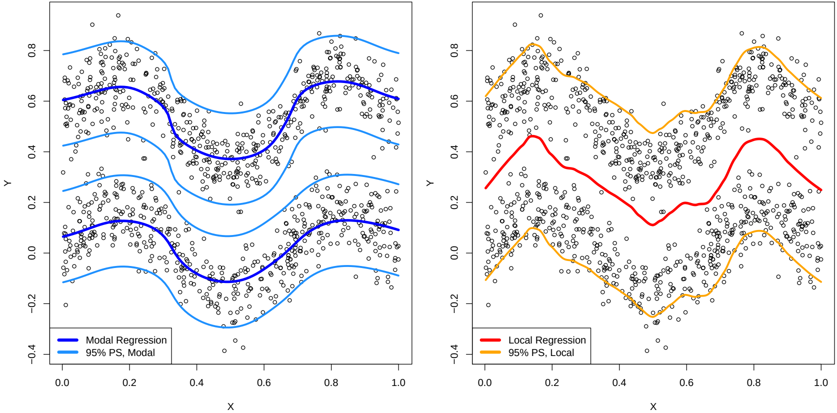


Example for Confidence Sets



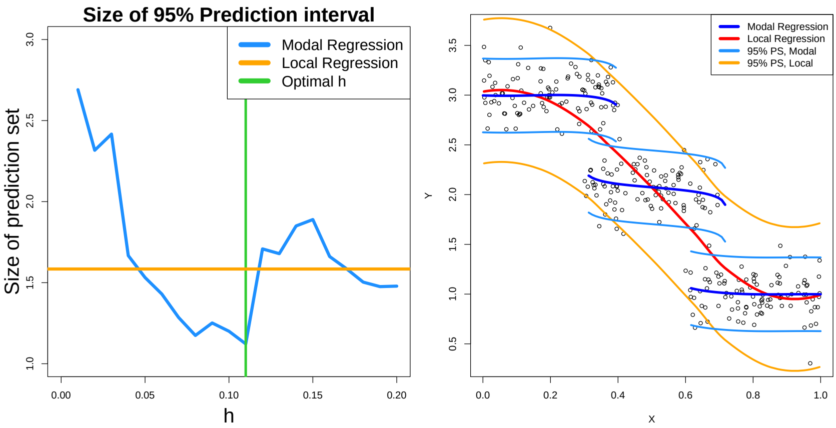
Applications for Modal Regression

- A compact prediction sets.



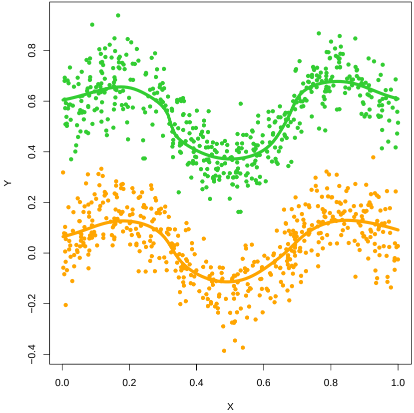
Applications for Modal Regression

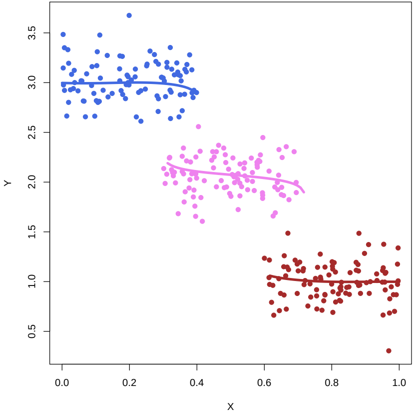
- A compact prediction sets.
- Bandwidth selection via minimizing the size of prediction sets.



Applications for Modal Regression

- A compact prediction sets.
- Bandwidth selection via minimizing the size of prediction sets.
- Regression clustering.



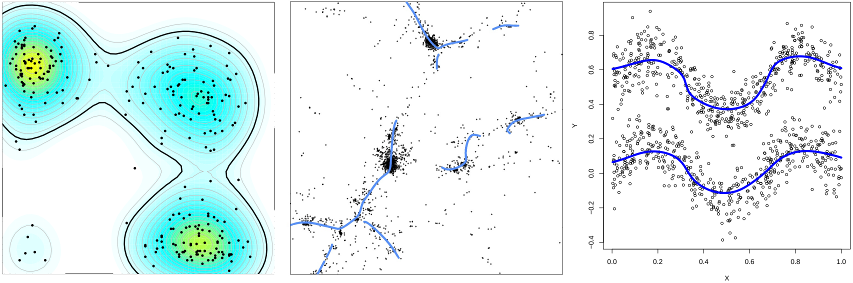


Outline

- Introduction to Shards
- Density Level Set
- Density Ridges
- Modal Regression
- Summary

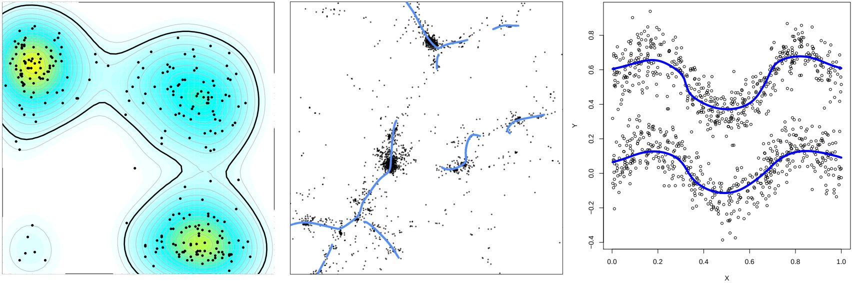
Summary

- We consider three types of Shards: level sets, ridges and conditional local modes.



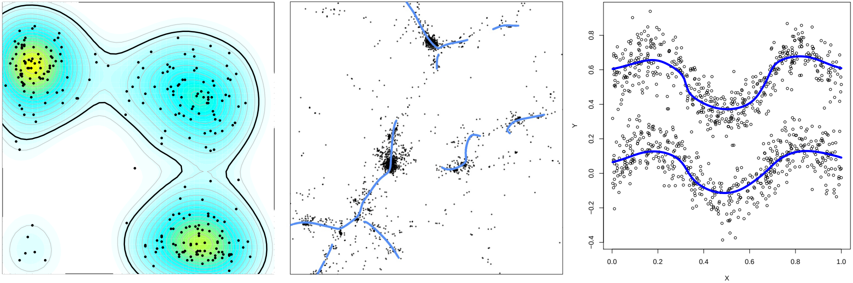
Summary

- We consider three types of Shards: level sets, ridges and conditional local modes.
- We derive asymptotic theory and propose confidence sets.



Summary

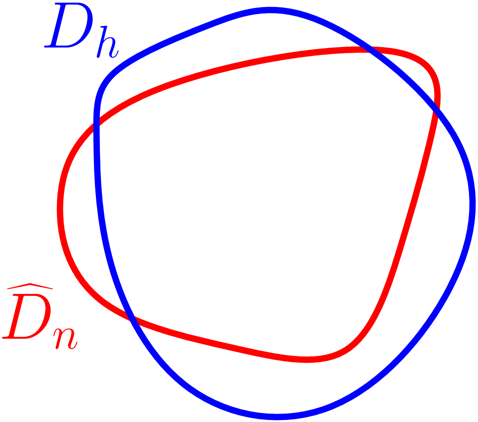
- We consider three types of Shards: level sets, ridges and conditional local modes.
- We derive asymptotic theory and propose confidence sets.
- Set estimation $\longrightarrow$ Set inference.

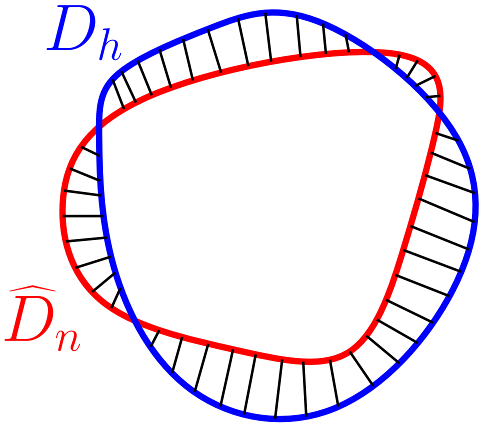


Thank you!

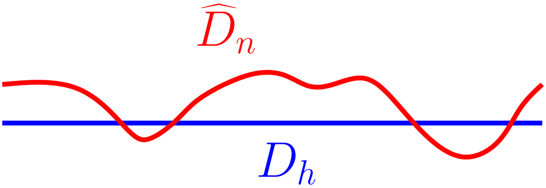
reference

1. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Density Level Sets: Asymptotics, Inference, and Visualization." Submitted to the Journal of American Statistical Association. arXiv preprint arXiv:1504.05438 (2015).
2. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Asymptotic theory for density ridges." To appear in the Annals of Statistics. arXiv preprint arXiv:1406.5663 (2014).
3. Chen, Yen-Chi, Christopher R. Genovese, Ryan J. Tibshirani, and Larry Wasserman. "Nonparametric Modal Regression." Under review of the Annals of Statistics. arXiv preprint arXiv:1412.1716 (2014).
4. Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Gaussian approximation of suprema of empirical processes." The Annals of Statistics 42, no. 4 (2014): 1564-1597.
5. Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Anti-concentration and honest, adaptive confidence bands." The Annals of Statistics 42, no. 5 (2014): 1787-1818.
6. Einbeck, Jochen, and Gerhard Tutz. "Modelling beyond regression functions: an application of multimodal regression to speedflow data." Journal of the Royal Statistical Society: Series C (Applied Statistics) 55, no. 4 (2006): 461-475.
7. Genovese, Christopher R., et al. "Nonparametric ridge estimation." The Annals of Statistics 42.4 (2014): 1511-1545.
8. Ozertem, Umut, and Deniz Erdogmus. "Locally defined principal curves and surfaces." The Journal of Machine Learning Research 12 (2011): 1249-1286.


D_h
$\widehat{D}_n$

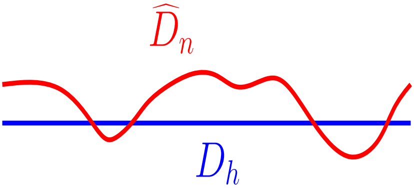

D_h
$\widehat{D}_n$

Asymptotic Theory



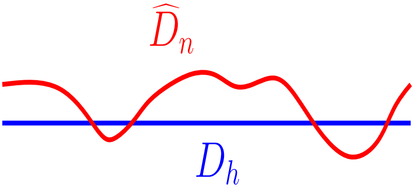
Asymptotic Theory

1. Thus, the projection distance $\approx$ a stochastic process.



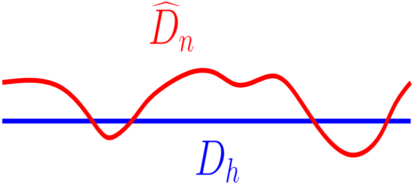
Asymptotic Theory

1. Thus, the projection distance $\approx$ a stochastic process.
2. This stochastic process $\approx$ empirical process.



Asymptotic Theory

1. Thus, the projection distance $\approx$ a stochastic process.
2. This stochastic process $\approx$ empirical process.
3. $\text{Haus}(\widehat{D}_n, D_h) =$ sup{projection distance} $\approx$ sup{Empirical process}.



Error Measurement

- To measure the errors, we apply a *local* Hausdorff distance

$$\Delta_n(x) = \text{Haus}(\widehat{M}_n(x), M(x)).$$

This is like a pointwise distance.

Error Measurement

- To measure the errors, we apply a *local* Hausdorff distance

$$\Delta_n(x) = \text{Haus}(\widehat{M}_n(x), M(x)).$$

  This is like a pointwise distance.

- Generalized to $\mathcal{L}_\infty$-type error:

$$\Delta_n = \sup_x \Delta_n(x) = \sup_x \text{Haus}(\widehat{M}_n(x), M(x)).$$

Asymptotic Theory

The pointwise errors and $\mathcal{L}_\infty$-type errors obey the common nonparametric rate:

Asymptotic Theory

The pointwise errors and $\mathcal{L}_\infty$-type errors obey the common nonparametric rate:

Theorem

Under regularity condition,

$$\Delta_n(x) = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^{d+3}}}\right)$$

$$\Delta_n = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+3}}}\right).$$

Asymptotic Theory

The pointwise errors and $\mathcal{L}_\infty$-type errors obey the common nonparametric rate:

Theorem

Under regularity condition,

$$\Delta_n(x) = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^{d+3}}}\right)$$

$$\Delta_n = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+3}}}\right).$$

Rate = Bias + Variance.

Prediction Sets

- Goal: to construct a set $\mathcal{P}_{1-\alpha} \subset \mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbb{P}\left((X, Y) \in \mathcal{P}_{1-\alpha}\right) \geq 1 - \alpha.$$

Prediction Sets

- Goal: to construct a set $\mathcal{P}_{1-\alpha} \subset \mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbb{P}\left((X, Y) \in \mathcal{P}_{1-\alpha}\right) \geq 1 - \alpha.$$

- A simple approach–pick $\widehat{r}_{1-\alpha}$ such that

$$\widehat{\mathcal{P}}_{1-\alpha} = \left\{(x, y) : y \in \widehat{M}_n(x) \oplus \widehat{r}_{1-\alpha}, x \in \mathbb{K}\right\}.$$

Prediction Sets

- Goal: to construct a set $\mathcal{P}_{1-\alpha} \subset \mathbb{R}^d \times \mathbb{R}$ such that

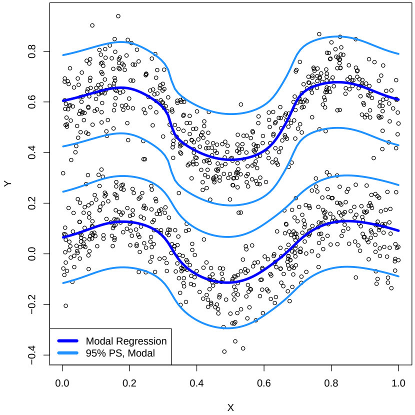
$$\mathbb{P}\left((X, Y) \in \mathcal{P}_{1-\alpha}\right) \geq 1 - \alpha.$$

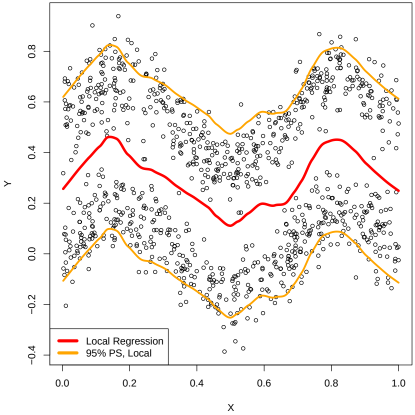
- A simple approach–pick $\widehat{r}_{1-\alpha}$ such that

$$\widehat{\mathcal{P}}_{1-\alpha} = \left\{(x, y) : y \in \widehat{M}_n(x) \oplus \widehat{r}_{1-\alpha}, x \in \mathbb{K}\right\}.$$

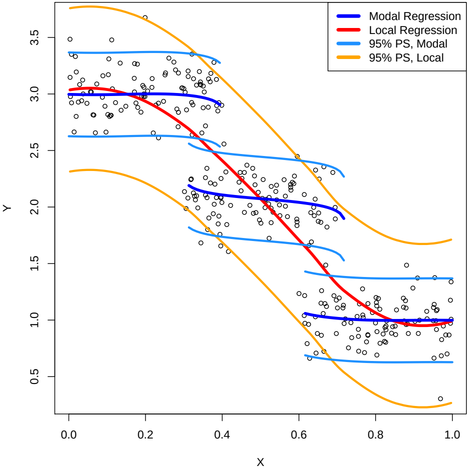
- We can choose $\widehat{r}_{1-\alpha}$ by cross-validation.

Example: Prediction Sets





Example: Prediction Sets



Bandwidth Selection

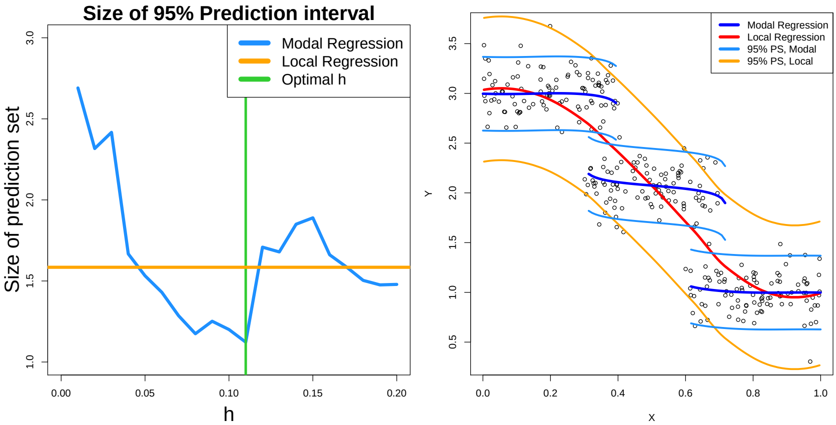
- We can choose smoothing parameter h via minimizing the size of prediction set.

Bandwidth Selection

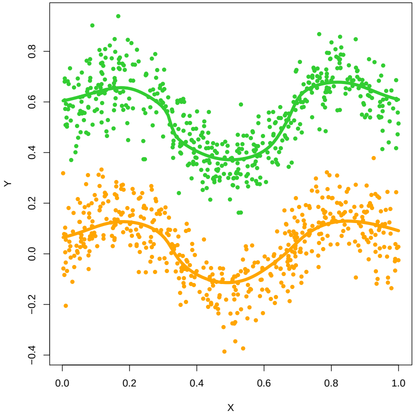
- We can choose smoothing parameter h via minimizing the size of prediction set.
- Namely, we choose

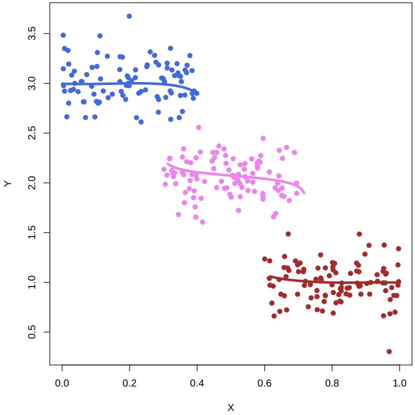
$$h^* = \underset{h>0}{argmin} \text{Vol}\left(\widehat{\mathcal{P}}_{1-\alpha}\right).$$

Example: Bandwidth Selection



Clustering–Exploring Hidden Structure



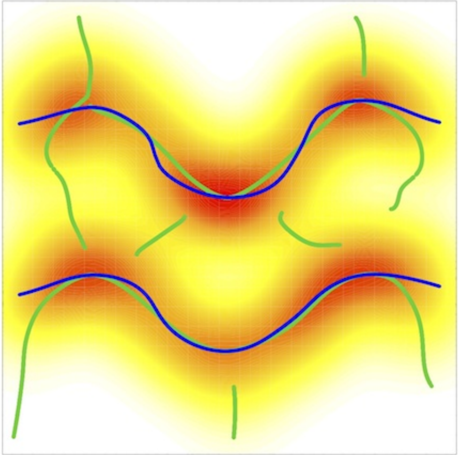


Mixture Inference versus Modal Inference

	Mixture-based	**Mode-based**
Density estimation	Gaussian mixture	Kernel density estimate
Clustering	K-means	Mean-shift clustering
Regression	Mixture regression	**Modal regression**
Algorithm	EM	Mean-shift
Complexity parameter	K (number of components)	h (smoothing bandwidth)
Type	Parametric model	Nonparametric model

Table: Comparison for methods based on mixtures versus modes.

Modal Regression VS Density Ridges



Mixture Regression

A general mixture model:

$$p(y|x) = \sum_{j=1}^{K(x)} \pi_j(x)\phi_j(y; \mu_j(x), \sigma_j^2(x)),$$

where each $\phi_j(y; \mu_j(x), \sigma_j^2(x))$ is a density function, parametrized by a mean $\mu_j(x)$ and variance $\sigma_j^2(x)$.

Common assumptions:

(MR1) $K(x) = K$,

(MR2) $\pi_j(x) = \pi_j$ for each j,

(MR3) $\mu_j(x) = \beta_j^T x$ for each j,

(MR4) $\sigma_j^2(x) = \sigma_j^2$ for each j, and

(MR5) $\phi_j(x)$ is Gaussian for each j.