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.
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.
→ 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

2008 Surface Temperature Anomaly (°C)

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
Let $p$ be the probability density function.
Let $p$ be the probability density function.

- The $\lambda$-level set is

$$D = \{x : p(x) = \lambda\}.$$
Example for Level Set

\[ f(x) \]
Example for Level Set

$\lambda \quad f(x) \quad x$
Example for Level Set

\[ f(x) \]

\[ \lambda \]

\[ x \]
Example for Level Set

\[ f(x) \]

\[ \lambda \]

\[ D \]

\[ X \]
Example for Level Set
Our estimator: a plug-in from the Kernel Density Estimator (KDE).
Our estimator: a plug-in from the Kernel Density Estimator (KDE).

- The KDE $\hat{p}_n$

$$\hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right).$$
Our estimator: a plug-in from the Kernel Density Estimator (KDE).

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

- The corresponding estimators
  
  $$\hat{D}_n = \{x : \hat{p}_n(x) = \lambda\}.$$
Example: Level Set Estimator
Example: Level Set Estimator
Example: Level Set Estimator
In particular, we focus on making inference for the smoothed version of the density, denoted as $p_h$:

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

where $\otimes$ denotes the convolution.

- We define $D_h$ as the level set using $p_h$. 

---

**Smoothed Level Set**

Yen-Chi Chen (CMU-Stats)
In particular, we focus on making inference for the smoothed version of the density, denoted as $p_h$:

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

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

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

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 $D$. 
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.
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 \(L_\infty\) norm for sets.
- Consistency: \(\text{Haus}(\hat{D}_n, D_h) = o_P(1)\).
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 \( L_\infty \) norm for sets.
- Consistency: \( \text{Haus}(\hat{D}_n, D_h) = o_P(1) \).
- Useful property:

\[
A \subset B \oplus \text{Haus}(A, B), \quad B \subset A \oplus \text{Haus}(A, B),
\]

where \( A \oplus r = \{ x : d(x, A) \leq r \} \).
Hausdorff distance can be applied to construct confidence sets.

Let $F_n$ be the CDF for $\text{Haus} (\hat{D}_n, D_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile. It can be shown that $P(D_h \subset \hat{D}_n \oplus t_{1-\alpha}) \geq 1-\alpha$. This follows from the property $A \subset B \oplus \text{Haus}(A, B)$, $B \subset A \oplus \text{Haus}(A, B)$.

We need to find the distribution $F_n$. 

Yen-Chi Chen (CMU-Stats)

Inference for Shards

May 27, 2015 18 / 48
Hausdorff Distance and Confidence Sets

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

Let $F_n$ be the CDF for $\text{Haus}(\hat{D}_n, D_h)$ and $t_{1-\alpha} = F_{n}^{-1}(1 - \alpha)$ be the $1 - \alpha$ quantile.

It can be shown that

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

→ This follows from the property

$$A \subset B \oplus \text{Haus}(A, B), \quad B \subset A \oplus \text{Haus}(A, B).$$
Hausdorff distance can be applied to construct confidence sets.

Let $F_n$ be the CDF for $\text{Haus} (\hat{D}_n, D_h)$ and $t_{1-\alpha} = F_n^{-1}(1 - \alpha)$ be the $1 - \alpha$ quantile.

It can be shown that

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

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

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

→ the last approximation follows from [Chernozhukov et. al. 2014].
It can be shown that
\[ \sqrt{nh^d} \text{Haus}(\hat{D}_n, D_h) \approx \sup \{ \text{Empirical process} \} \approx \sup \{ \text{Gaussian process} \}. \]

→ 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*

\[
\sup_t \left| \mathbb{P} \left( \sqrt{nh^d} \text{Haus}(\hat{D}_n, D_h) < t \right) - \mathbb{P} \left( \sup_{f \in \mathcal{F}} |\mathbb{B}(f)| < t \right) \right| = O \left( \left( \frac{\log^7 n}{nh^d} \right)^{1/8} \right).
\]
- Good news: we have the asymptotic behavior.
- Bad news: the asymptotic behavior is complicated.
Good news: we have the asymptotic behavior.
Bad news: the asymptotic behavior is complicated.
A solution: the bootstrap.
Bootstrap sample $\rightarrow$ bootstrap level set $\hat{D}^*_n$.
Compute $\text{Haus}(\hat{D}^*_n, \hat{D}_n)$ to get a CDF estimator $\hat{F}_n$.
Choose $\hat{t}_{1-\alpha}$ be the $1 - \alpha$ quantile for $\hat{F}_n$. 
The Bootstrap Consistency

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

\[
\sqrt{nh^d} \text{Haus}(\hat{D}^*_n, \hat{D}_n) \approx \sup \{\text{Gaussian process}\} \approx \sqrt{nh^d} \text{Haus}(\hat{D}_n, D).
\]
The Bootstrap Consistency

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

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

Theorem

Under regularity condition,

$$\mathbb{P} \left( D_h \subset \hat{D}_n \oplus \hat{t}_{1-\alpha}\right) = 1 - \alpha + O \left( \left( \frac{\log^7 n}{nh^d} \right)^{1/8} \right).$$
Example: Confidence Sets
Example: Confidence Sets
Properties for the Confidence Sets

1. **Blue**: confidence sets for $D_h$
1. **Blue**: confidence sets for $D_h$

2. **Yellow**: every point above $\lambda$
Properties for the Confidence Sets

1. **Blue**: confidence sets for \( D_h \)
2. **Yellow**: every point above \( \lambda \)
3. **Green**: every point below \( \lambda \)
Properties for the Confidence Sets

1. **Blue**: confidence sets for $D_h$
2. **Yellow**: every point above $\lambda$
3. **Green**: every point below $\lambda$
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 $\lambda$
3. **Green**: every point below $\lambda$
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.
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 Smooth Functions
Example: Ridges in Smooth Functions
A generalized local mode in a specific ‘subspace’.
A generalized local mode in a specific ‘subspace’.
A generalized local mode in a specific ‘subspace’.
Formal Definition of Density Ridges

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

$\lambda_j(x), v_j(x))$: $j$th eigenvalue/vector of $H(x) = \nabla \nabla p(x)$.

$V(x) = [v_2(x), \ldots, v_d(x)]$: matrix of 2nd to last eigenvectors.

$V(x)V(x)^T$: a projection.

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

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

Yen-Chi Chen (CMU-Stats)
Formal Definition of Density Ridges

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

- $p(x)$: a density function.
- $(\lambda_j(x), v_j(x))$: $j$th eigenvalue/vector of $H(x) = \nabla \nabla p(x)$.
- $V(x) = [v_2(x), \ldots, v_d(x)]$: 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.
- \((\lambda_j(x), v_j(x))\): \(j\)th 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:

\[
\text{Mode}(p) = \{x : \nabla p(x) = 0, \lambda_1(x) < 0\}.
\]
We use the plug-in estimate:

\[ \hat{R}_n = \text{Ridge}(\hat{\rho}_n), \]

where \( \hat{\rho}_n \) is the KDE.
We use the plug-in estimate:

\[ \hat{R}_n = \text{Ridge}(\hat{\rho}_n), \]

where \( \hat{\rho}_n \) is the KDE.

In general, finding ridges from a given function is hard.
We use the plug-in estimate:

$$\hat{R}_n = \text{Ridge}(\hat{\rho}_n),$$

where $\hat{\rho}_n$ 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 $\hat{R}_n$, 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
Can we derive asymptotic theory and make statistical inference for density ridges?
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.
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,

\[ \sqrt{nh^{d+2}} \text{Haus}(\hat{R}_n, R_h) \approx \sup_{f \in \mathcal{F}} |\mathbb{B}(f)| \text{ for certain function space } \mathcal{F}. \]

\[ \hat{R}_n \oplus t_{1-\alpha} \text{ is an asymptotic valid confidence set for } R_h. \]

**Note:** \( R_h = \text{Ridge}(p_h) \) 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
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
Our estimator is the plug-in from the KDE:

\[ \hat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y} \hat{\rho}_n(x, y) = 0, \frac{\partial^2}{\partial y^2} \hat{\rho}(x, y) < 0 \right\}. \]
Our estimator is the plug-in from the KDE:
\[
\hat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y} \hat{\rho}_n(x, y) = 0, \frac{\partial^2}{\partial y^2} \hat{\rho}(x, y) < 0 \right\}.
\]

Finding conditional local modes is hard in general.
Our estimator is the plug-in from the KDE:

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

Finding conditional local modes is hard in general.

Partial mean shift: a simple algorithm for computing $\hat{M}_n(x)$, the plug-in estimator of the KDE, from the data (Einbeck et. al. 2006).
Example for Modal Regression
Example for Modal Regression
Let $M_h$ be the modal manifolds for $p_h$.
Define a uniform metric $\Delta_n = \sup_x \text{Haus}(\hat{M}_n(x), M_h(x))$. 

Theorem
Under regularity condition, $\sqrt{n}h_d + 3\Delta_n \approx \sup_{f \in F} |B(f)|$ for certain function space $F$. 

The set $\{ (x, y) : y \in \hat{M}_n(x) \oplus \hat{t}_1 - \alpha, x \in K \}$ is an asymptotic valid confidence set for $M_h$. 

Yen-Chi Chen (CMU-Stats) 
Inference for Shards 
May 27, 2015
Let $M_h$ be the modal manifolds for $p_h$.
Define a uniform metric $\Delta_n \equiv \sup_x \text{Haus}(\hat{M}_n(x), M_h(x))$.

**Theorem**

Under regularity condition,
\[ \sqrt{nh^{d+3}} \Delta_n \approx \sup_{f \in F} |\mathbb{B}(f)| \text{ for certain function space } F. \]

The set
\[ \{(x, y) : y \in \hat{M}_n(x) \oplus \hat{t}_{1-\alpha}, x \in \mathbb{K}\} \]

is an asymptotic valid confidence set for $M_h$. 
Example for Confidence 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.

Size of 95% Prediction interval

Size of prediction set

Modal Regression
Local Regression
95% PS, Modal
95% PS, Local
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
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.
We consider three types of Shards: level sets, ridges and conditional local modes.

We derive asymptotic theory and propose confidence sets.

Set estimation $\rightarrow$ Set inference.
Thank you!


$D_h$

$\hat{D}_n$
Asymptotic Theory

\[ D_h \]

\[ \hat{D}_n \]
\[ \hat{D}_n \]

\[ D_h \]
Thus, the projection distance \(\approx\) a stochastic process.
1. Thus, the projection distance $\approx$ a stochastic process.

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

This stochastic process $\approx$ empirical process.

$\text{Haus}(\hat{D}_n, D_h) = \sup\{\text{projection distance}\} \approx \sup\{\text{Empirical process}\}$. 
To measure the errors, we apply a local Hausdorff distance

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

This is like a pointiwise distance.
To measure the errors, we apply a *local* Hausdorff distance

\[
\Delta_n(x) = \text{Haus}(\hat{M}_n(x), M(x)).
\]

This is like a pointwise distance.

Generalized to $L_\infty$-type error:

\[
\Delta_n = \sup_x \Delta_n(x) = \sup_x \text{Haus}(\hat{M}_n(x), M(x)).
\]
The pointwise errors and $L_\infty$-type errors obey the common nonparametric rate:
The pointwise errors and $L_\infty$-type errors obey the common nonparametric rate:

\[ \Delta_n(x) = O(h^2) + O_p \left( \sqrt{\frac{1}{nh^{d+3}}} \right) \]

\[ \Delta_n = O(h^2) + O_p \left( \sqrt{\frac{\log n}{nh^{d+3}}} \right). \]
The pointwise errors and $L_\infty$-type errors obey the common nonparametric rate:

Theorem

Under regularity condition,

\[
\Delta_n(x) = O(h^2) + O_p \left( \sqrt{\frac{1}{nh^{d+3}}} \right)
\]

\[
\Delta_n = O(h^2) + O_p \left( \sqrt{\frac{\log n}{nh^{d+3}}} \right).
\]

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

$$\mathbb{P} ((X, Y) \in \mathcal{P}_{1-\alpha}) \geq 1 - \alpha.$$
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 $\hat{r}_{1-\alpha}$ such that

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

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

We can choose $\hat{r}_{1-\alpha}$ by cross-validation.
Example: Prediction Sets

Inference for Shards

- Modal Regression
- 95% PS, Modal

- Local Regression
- 95% PS, Local
Example: Prediction Sets

- Modal Regression
- Local Regression
- 95% PS, Modal
- 95% PS, Local
We can choose smoothing parameter $h$ via minimizing the size of prediction set.
We can choose smoothing parameter $h$ via minimizing the size of prediction set.

Namely, we choose

$$h^* = \arg\min_{h>0} \text{Vol} \left( \hat{P}_{1-\alpha} \right).$$
Example: Bandwidth Selection

Size of 95% Prediction interval

Size of prediction set

Modal Regression
Local Regression

Modal Regression
Local Regression
95% PS, Modal
95% PS, Local

Yen-Chi Chen (CMU-Stats)
Inference for Shards
May 27, 2015
Clustering—Exploring Hidden Structure
### Mixture Inference versus Modal Inference

<table>
<thead>
<tr>
<th></th>
<th>Mixture-based</th>
<th>Mode-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density estimation</td>
<td>Gaussian mixture</td>
<td>Kernel density estimate</td>
</tr>
<tr>
<td>Clustering</td>
<td>$K$-means</td>
<td>Mean-shift clustering</td>
</tr>
<tr>
<td>Regression</td>
<td>Mixture regression</td>
<td><strong>Modal regression</strong></td>
</tr>
<tr>
<td>Algorithm</td>
<td>EM</td>
<td>Mean-shift</td>
</tr>
<tr>
<td>Complexity parameter</td>
<td>$K$ (number of components)</td>
<td>$h$ (smoothing bandwidth)</td>
</tr>
<tr>
<td>Type</td>
<td>Parametric model</td>
<td>Nonparametric model</td>
</tr>
</tbody>
</table>

**Table:** Comparison for methods based on mixtures versus modes.
Modal Regression VS Density Ridges
A general mixture model:

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

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) \( \mathcal{K}(x) = \mathcal{K} \),
- (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 \( j \).