Statistical Inference using Geometric Features

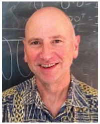
Yen-Chi Chen

Department of Statistics
University of Washington

Collaborators

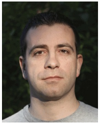
Statistics

Christopher Genovese
(CMU)

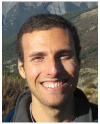


Larry Wasserman
(CMU)

Statistics



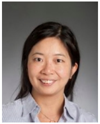
Alessandro Rinaldo
(CMU)



Ryan Tibshirani
(CMU)

Collaborators

Astronomy



Shirley Ho
(Lawrence Berkeley Lab)

Peter Freeman
(CMU)

Rachel Mandelbaum
(CMU)

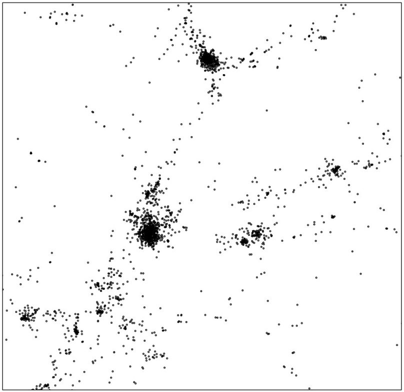
Astronomy

Andrew Connolly
(UW)

Matthew McQuinn
(UW)

Matthew Wilde
(UW)

What are Geometric Features?

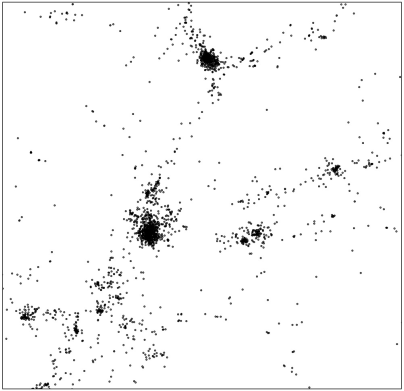


What are Geometric Features?

The data can be viewed as

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

p is a probability density function.



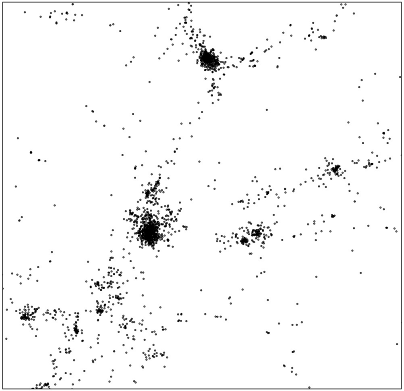
What are Geometric Features?

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.



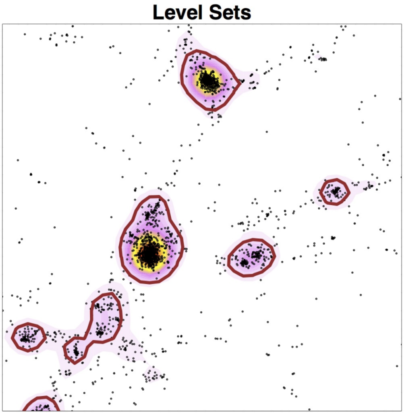
What are Geometric Features?

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.



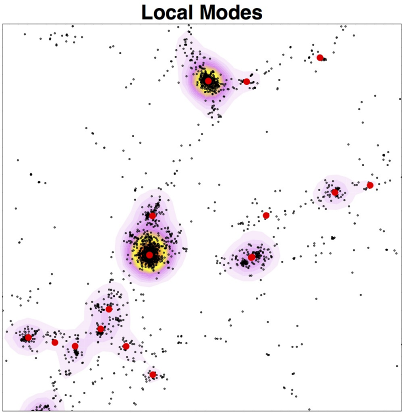
What are Geometric Features?

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.



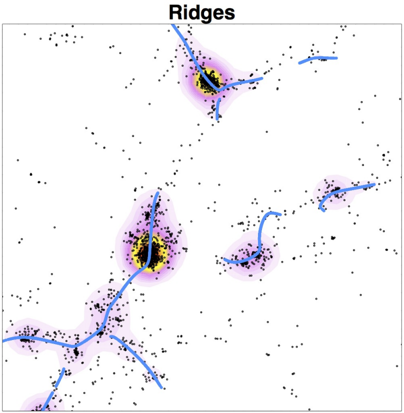
What are Geometric Features?

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.



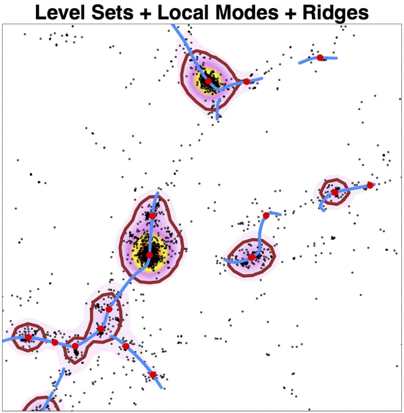
What are Geometric Features?

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.



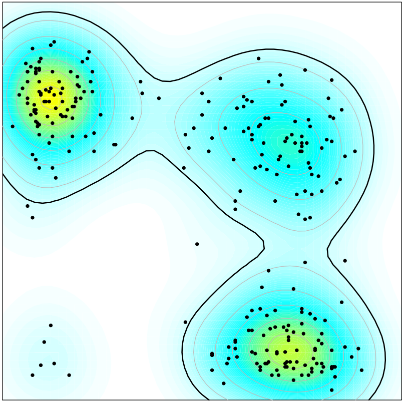
Geometric Features

Common examples:

Geometric Features

Common examples:

- Level Sets

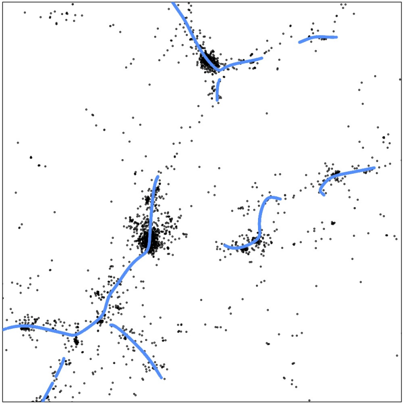


→ **Chen** et al. 'Density Level Sets: Asymptotics, Inference, and Visualization' *JASA-T&M* (2016+).

Geometric Features

Common examples:

- Level Sets
- Ridges



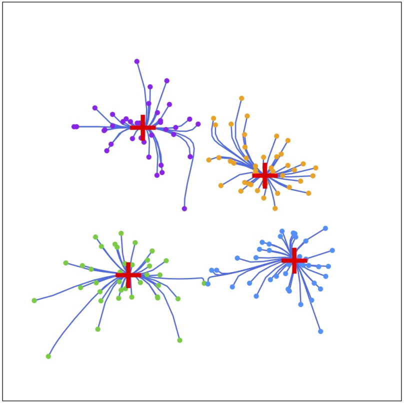
→ **Chen** et al. 'Asymptotic Theory for Density Ridges.' *The Annals of Statistics* (2015).

→ **Chen** et al. 'Optimal Ridge Detection using Coverage Risk.' *NIPS* (2015).

Geometric Features

Common examples:

- Level Sets
- Ridges
- Clusters



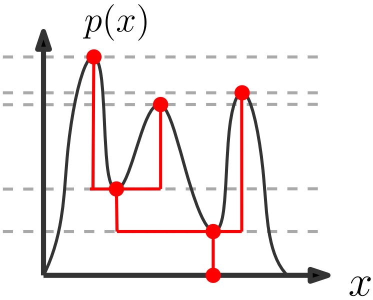
→ **Chen** et al. 'A Comprehensive Approach to Mode Clustering.' *The Electronic Journal of Statistics* (2016).

→ **Chen** et al. 'Statistical Inference Using the Morse-Smale Complex.' (2015).

Geometric Features

Common examples:

- Level Sets
- Ridges
- Clusters
- Density Trees



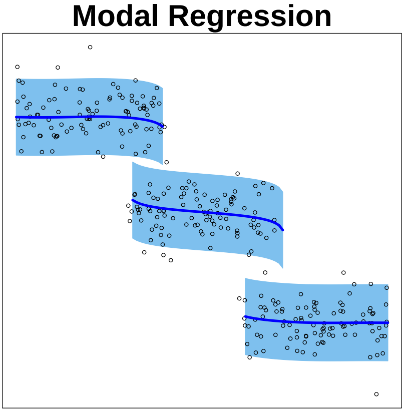
→ Kim and **Chen** et al. 'Confidence Sets for Density Trees.' *NIPS* (2016).

→ **Chen**. 'Generalized Cluster Trees and Singular Measures.' (2016)

Geometric Features

Common examples:

- Level Sets
- Ridges
- Clusters
- Density Trees
- Modal Regression



→ **Chen** et al. 'Nonparametric Modal Regression.' *The Annals of Statistics* (2016).

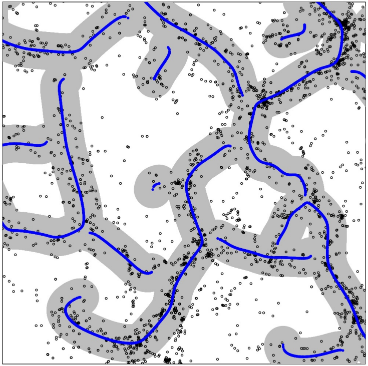
Geometric Features

Common examples:

- Level Sets
- Ridges
- Clusters
- Density Trees
- Modal Regression

Applications:

- Astronomy
- Biology
- Image Analysis



→ **Chen** et al. 'Cosmic Web Reconstruction through Density Ridges: Catalogue.' *Mon. Not. Roy. Astro. Soc.* (2016).

→ **Chen** et al. 'Cosmic Web Reconstruction through Density Ridges: Method and Algorithm.' *Mon. Not. Roy. Astro. Soc.* (2015).

Geometric Features

Common examples:

- Level Sets
- **Ridges**
- Clusters
- Density Trees
- **Modal Regression**

Applications:

- Astronomy
- Biology
- Image Analysis

→ **Chen** et al. 'Asymptotic Theory for Density Ridges.' *The Annals of Statistics* (2015).

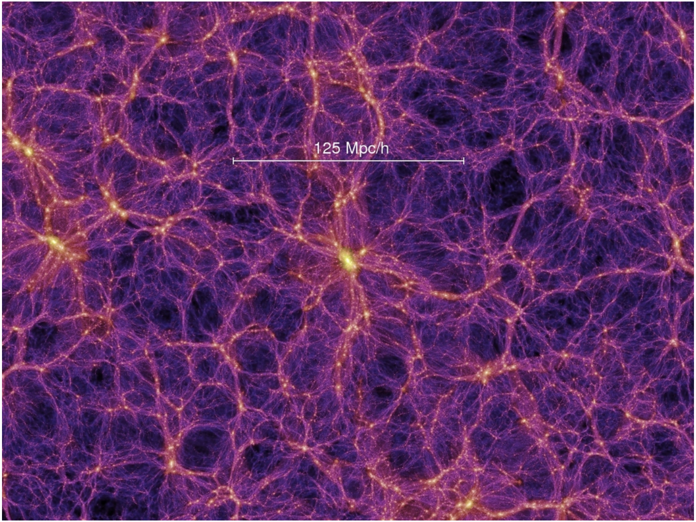
→ **Chen** et al. 'Optimal Ridge Detection using Coverage Risk.' *NIPS* (2015).

→ **Chen** et al. 'Cosmic Web Reconstruction through Density Ridges: Method and Algorithm.' *Mon. Not. Roy. Astro. Soc.* (2015).

→ **Chen** et al. 'Nonparametric Modal Regression.' *The Annals of Statistics* (2016).

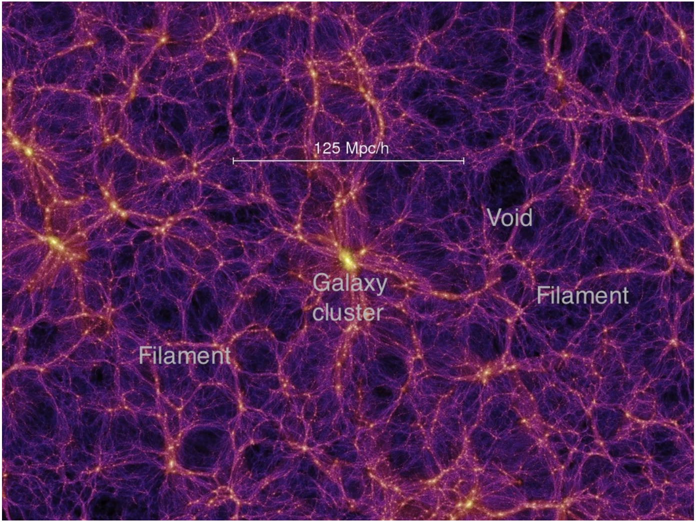
Density Ridges

Example: Cosmology



Credit: Millennium Simulation

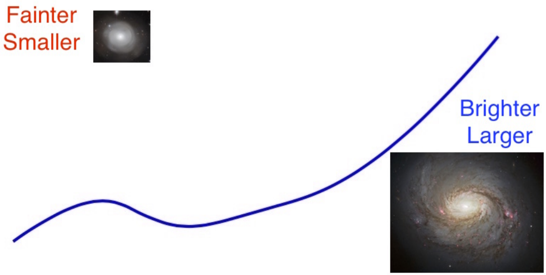
Example: Cosmology



Credit: Millennium Simulation

The Importance of Filaments

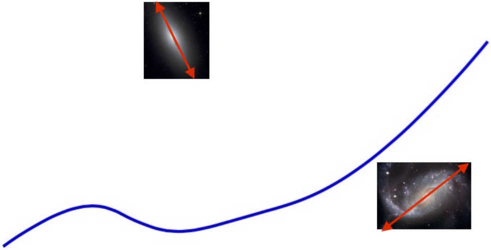
- A galaxy's brightness, size, and mass are associated with the distance to filaments.



→ **Chen** et al. 'Detecting Effects of Filaments on Galaxy Properties in Sloan Digital Sky Survey III'(Mon. Not. Roy. Astro. Soc. 2016+)

The Importance of Filaments

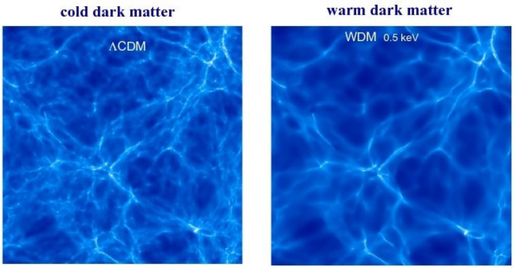
- A galaxy's brightness, size, and mass are associated with the distance to filaments.
- A galaxy's alignment is associated with filaments.



→ **Chen** et al. 'Investigating Galaxy-Filament Alignment in Hydrodynamic Simulations using Density Ridges' (Mon. Not. Roy. Astro. Soc. 2015)

The Importance of Filaments

- A galaxy's brightness, size, and mass are associated with the distance to filaments.
- A galaxy's alignment is associated with filaments.
- Filaments can be used to test cosmological theories.



- Credit: Kavli Institute for Cosmology, Cambridge

Density Ridges

We formalize the notion of filaments as *density ridges*.

Literature Review

Early work on ridges is in image analysis (Eberly 1996, Damon 1999).

Literature Review

Early work on ridges is in image analysis (Eberly 1996, Damon 1999). The concept of ridges in point clouds was introduced in Hall et al. (1992) and Hall and Peng (2001).

Literature Review

Early work on ridges is in image analysis (Eberly 1996, Damon 1999). The concept of ridges in point clouds was introduced in Hall et al. (1992) and Hall and Peng (2001).

Recent work:

- Algorithm for finding ridges: Ozertem and Erdogmus (2011).
- Consistency for ridge estimators: Genovese et al. (2014).
- Asymptotic analysis: Qiao and Polonik (2014).

Literature Review

Early work on ridges is in image analysis (Eberly 1996, Damon 1999). The concept of ridges in point clouds was introduced in Hall et al. (1992) and Hall and Peng (2001).

Recent work:

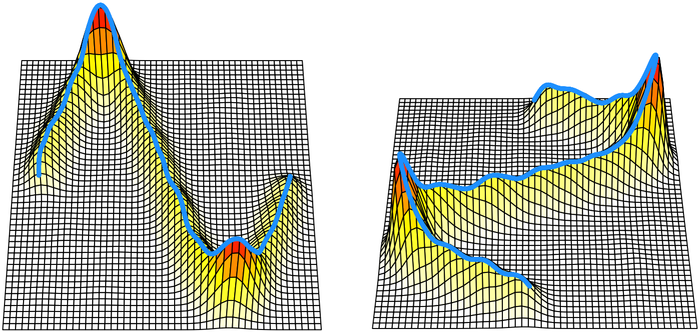
- Algorithm for finding ridges: Ozertem and Erdogmus (2011).
- Consistency for ridge estimators: Genovese et al. (2014).
- Asymptotic analysis: Qiao and Polonik (2014).

⟶In our work, we derive the asymptotic theory for ridge estimators and propose methods for constructing confidence sets.

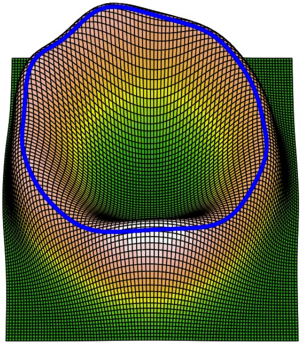
Example: Ridges in Mountains

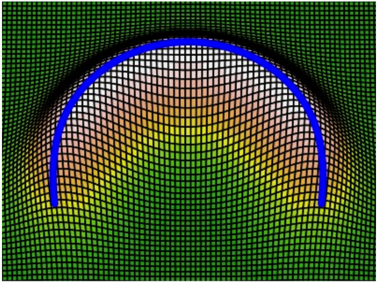
Credit: Google

Example: Ridges in Smooth Functions

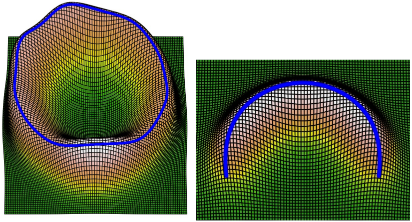


Example: Ridges in Smooth Functions



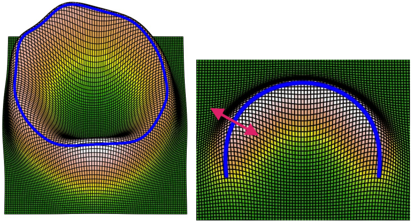


Ridges: Local Modes in Subspace



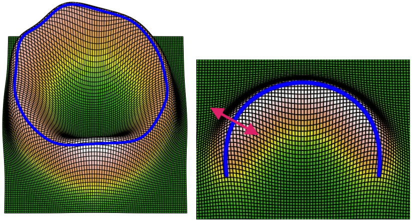
A generalized local mode in a specific ‘subspace’.

Ridges: Local Modes in Subspace

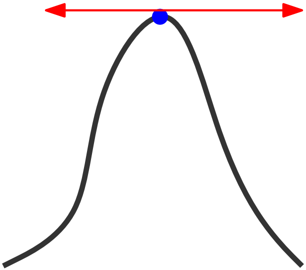


A generalized local mode in a specific 'subspace'.

Ridges: Local Modes in Subspace



A generalized local mode in a specific 'subspace'.



Formal Definition of Density Ridges

- $p : \mathbb{R}^d \mapsto \mathbb{R}$, the density function.

Formal Definition of Density Ridges

- $p : \mathbb{R}^d \mapsto \mathbb{R}$, the density function.
- $(\lambda_j(x), v_j(x))$: jth eigenvalue/vector of $H(x) = \nabla\nabla p(x)$.

Formal Definition of Density Ridges

- $p : \mathbb{R}^d \mapsto \mathbb{R}$, the 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 the 2nd eigenvector to the last eigenvector.

Formal Definition of Density Ridges

- $p : \mathbb{R}^d \mapsto \mathbb{R}$, the 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 the 2nd eigenvector to the last eigenvector.
- $V(x)V(x)^T$: a projection.

Formal Definition of Density Ridges

- $p : \mathbb{R}^d \mapsto \mathbb{R}$, the 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 the 2nd eigenvector to the last eigenvector.
- $V(x)V(x)^T$: 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

- $p : \mathbb{R}^d \mapsto \mathbb{R}$, the 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 the 2nd eigenvector to the last eigenvector.
- $V(x)V(x)^T$: a projection.
- Ridges:

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

- Local modes:

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

Dimension of Ridges

The dimension of a ridge is 1.

This is because ridges are points satisfying $V(x)V(x)^T\nabla p(x) = 0$.

$V(x)V(x)^T$ has rank $d-1$, so there are $d-1$ effective constraints.

By the Implicit Function Theorem, ridges have dimension 1.

Dimension of Ridges

The dimension of a ridge is 1.

This is because ridges are points satisfying $V(x)V(x)^T\nabla p(x) = 0$.

$V(x)V(x)^T$ has rank $d-1$, so there are $d-1$ effective constraints.

By the Implicit Function Theorem, ridges have dimension 1.

Note that there are higher dimensional ridges but in this talk, we will focus on 1 dimensional ridges.

Estimator and Algorithm

We use the plug-in estimate:

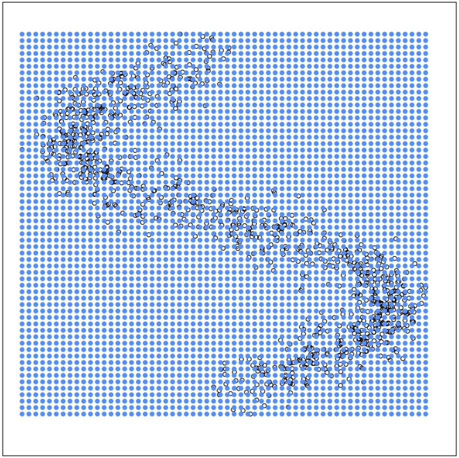
$$\widehat{R}_h = \text{Ridge}(\widehat{p}_h),$$

where $\widehat{p}_h = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$ is the kernel density estimator (KDE). h is the smoothing bandwidth, which controls the amount of smoothing.

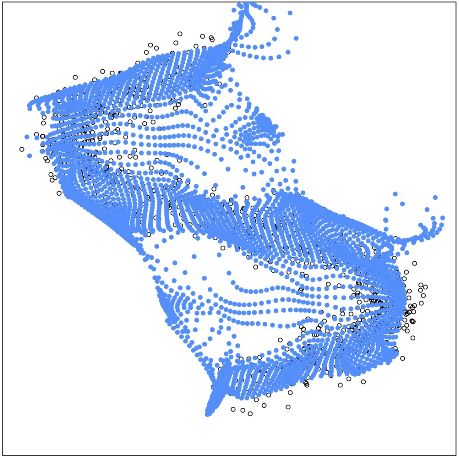
- In general, finding ridges from a given function is hard.
- The Subspace Constraint Mean Shift[1] (SCMS) algorithm allows us to find $\widehat{R}_h$, ridges of the KDE.

[1]Ozertem, Umut, and Deniz Erdogmus. "Locally defined principal curves and surfaces." JMLR (2011).

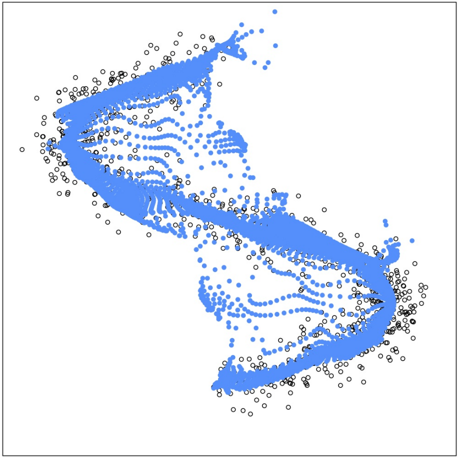
SCMS: Ridge Recovery Algorithm



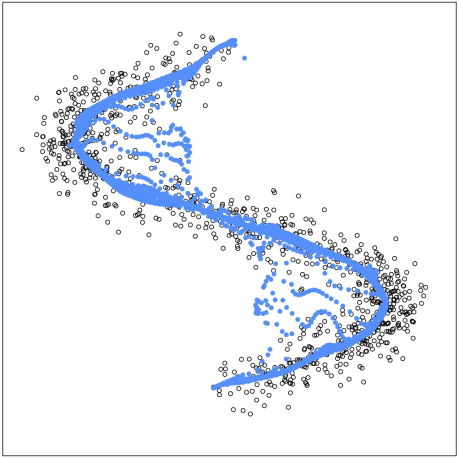
SCMS: Ridge Recovery Algorithm



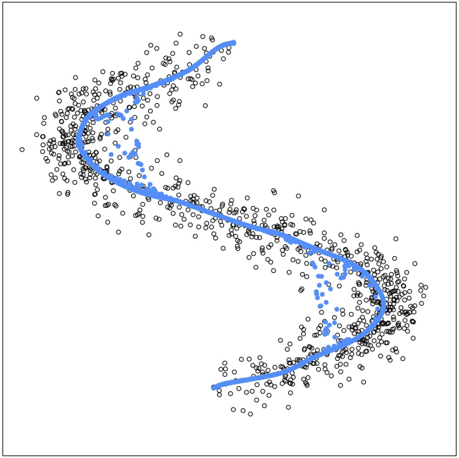
SCMS: Ridge Recovery Algorithm



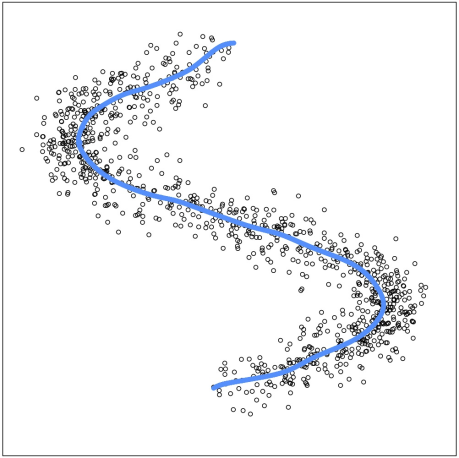
SCMS: Ridge Recovery Algorithm



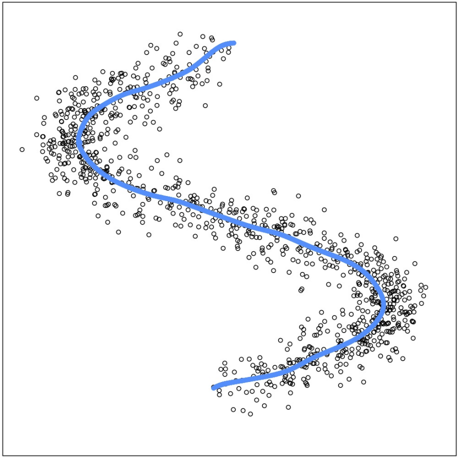
SCMS: Ridge Recovery Algorithm



SCMS: Ridge Recovery Algorithm

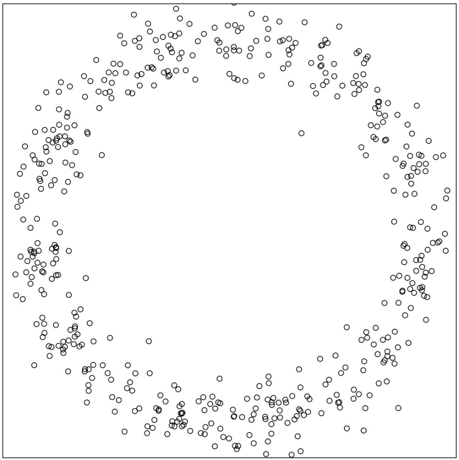


SCMS: Ridge Recovery Algorithm

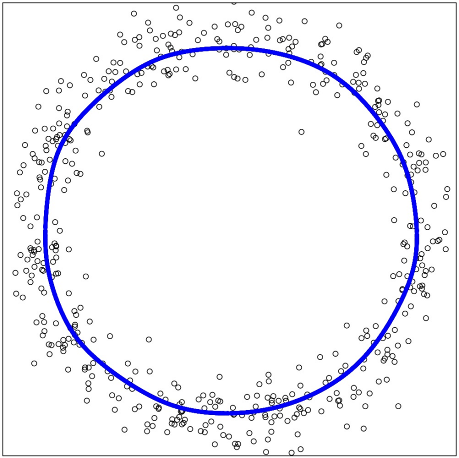


SCMS moves blue mesh points by gradient ascent and a projection.

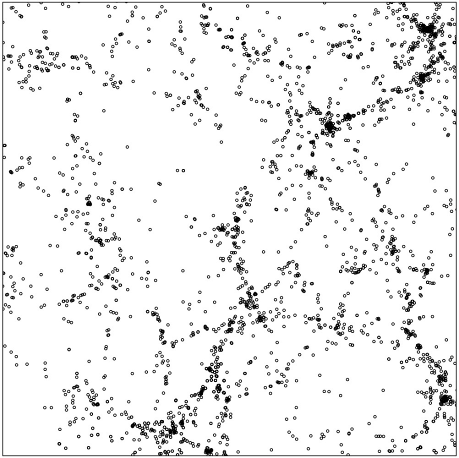
Example for Estimated Density Ridges



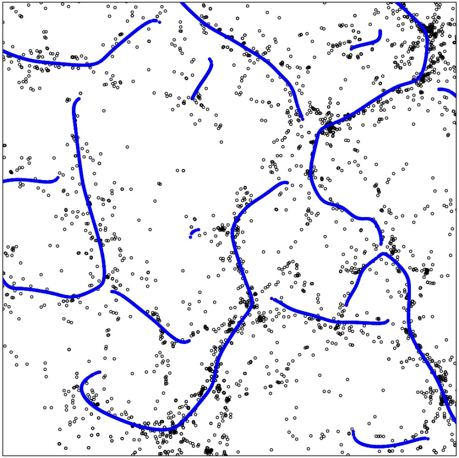
Example for Estimated Density Ridges



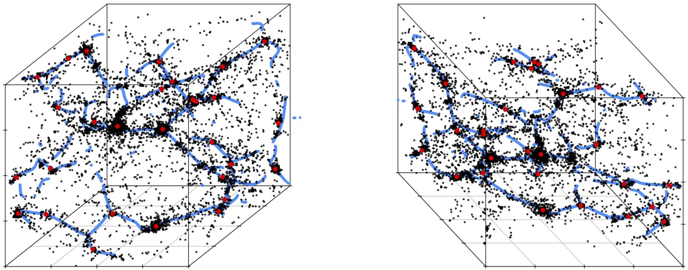
Example for Estimated Density Ridges



Example for Estimated Density Ridges



3D Example for Estimated Ridges



Blue curves: density ridges.

Red points: density local modes.

Statistical Inference: Confidence Sets

Having estimators is not enough for statistical inference.

We need confidence sets for density ridges.

Statistical Inference: Confidence Sets

Having estimators is not enough for statistical inference.

We need confidence sets for density ridges.

Namely, we want to find a set $C_{1-\alpha,n}$ from the data such that

$$\mathbb{P}\left(R \subset C_{1-\alpha,n}\right) \geq 1 - \alpha.$$

Smoothed Density Ridges

In particular, we focus on making inference for the smoothed ridges $R_h = \text{Ridge}(p_h)$.

p_h is the smoothed density function:

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

where $\otimes$ denotes the convolution.

Smoothed Density Ridges

In particular, we focus on making inference for the smoothed ridges $R_h = \text{Ridge}(p_h)$.

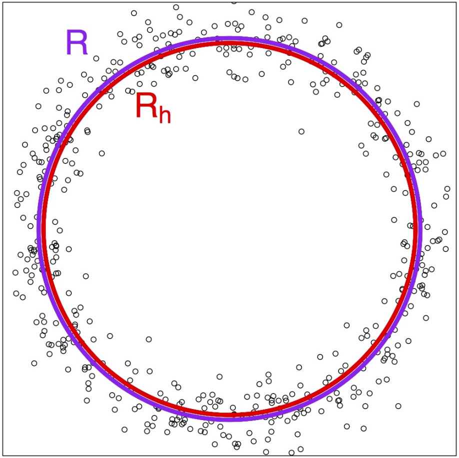
p_h is the smoothed density function:

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

where $\otimes$ denotes the convolution.

- The advantages of R_h over R:
  - Always well-defined.
  - Topologically similar.
  - We can undersmooth so that inference for R_h is also valid for R.

Ridges VS Smoothed Ridges



Useful Metric: Hausdorff Distance

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

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

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

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 from point x to a set A.

- Haus is an L_∞ metric for sets.

The $\oplus$ Operation

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

The $\oplus$ Operation

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

Then we have the following inclusion property:

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

Confidence Sets

We can use the Hausdorff distance and $\oplus$ operation to construct confidence sets.

Let F_n be the CDF for $\mathsf{Haus}(\widehat{R}_h, R_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

Confidence Sets

We can use the Hausdorff distance and $\oplus$ operation to construct confidence sets.

Let F_n be the CDF for $\mathsf{Haus}(\widehat{R}_h, R_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

- It can be shown that

$$\mathbb{P}\left(R_h \subset \widehat{R}_h \oplus t_{1-\alpha}\right) \geq 1 - \alpha.$$

$\rightarrow$ This follows from the property

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

Confidence Sets

We can use the Hausdorff distance and $\oplus$ operation to construct confidence sets.

Let F_n be the CDF for $\mathsf{Haus}(\widehat{R}_h, R_h)$ and $t_{1-\alpha} = F_n^{-1}(1-\alpha)$ be the $1-\alpha$ quantile.

- It can be shown that

$$\mathbb{P}\left(R_h \subset \widehat{R}_h \oplus t_{1-\alpha}\right) \geq 1-\alpha.$$

  → This follows from the property

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

- We need to find the distribution F_n.

Need: F_n, the CDF of $\mathsf{Haus}(\widehat{R}_h, R_h)$.

Asymptotic Theory

Need: F_n, the CDF of $\mathsf{Haus}(\widehat{R}_h, R_h)$.

Key observation:

$$\begin{aligned}\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h) &\approx \sqrt{nh^{d+2}} \sup_{x \in R_h} d(x, \widehat{R}_h) \\ &\approx \sup \{\text{Empirical process on } R_h\} \\ &\approx \sup \{\text{Gaussian process on } R_h\}.\end{aligned}$$

Asymptotic Theory

Need: F_n, the CDF of $\mathsf{Haus}(\widehat{R}_h, R_h)$.

Key observation:

$$\begin{aligned}\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h) &\approx \sqrt{nh^{d+2}} \sup_{x \in R_h} d(x, \widehat{R}_h) \\ &\approx \sup\{\text{Empirical process on } R_h\} \\ &\approx \sup\{\text{Gaussian process on } R_h\}.\end{aligned}$$

Theorem (Chen, Genovese, and Wasserman (2015))

Under regularity conditions and $\frac{\log n}{nh^{d+8}} \to 0$, *there exists a Gaussian process* $\mathbb{B}_n$ *defined on a certain function space* $\mathcal{F}$ *such that*

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h) < t\right) - \mathbb{P}\left(\sup_{f \in \mathcal{F}} |\mathbb{B}_n(f)| < t\right) \right| = O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

Asymptotic Theory

Need: F_n, the CDF of $\text{Haus}(\widehat{R}_h, R_h)$.

Key observation:

$$\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_h, R_h) \approx \sqrt{nh^{d+2}} \sup_{x \in R_h} d(x, \widehat{R}_h)$$
$$\approx \sup \{\text{Empirical process on } R_h\}$$
$$\approx \sup \{\text{Gaussian process on } R_h\}.$$

Theorem (Chen, Genovese, and Wasserman (2015))

Under regularity conditions and $\frac{\log n}{nh^{d+8}} \to 0$, *there exists a Gaussian process* $\mathbb{B}_n$ *defined on a certain function space* $\mathcal{F}$ *such that*

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+2}}\text{Haus}(\widehat{R}_h, R_h) < t\right) - \mathbb{P}\left(\sup_{f \in \mathcal{F}} |\mathbb{B}_n(f)| < t\right) \right| = O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

The Bootstrap

Theorem (Chen, Genovese, and Wasserman (2015))

Under regularity conditions and $\frac{\log n}{nh^{d+8}} \to 0$*, there exists a Gaussian process* $\mathbb{B}_n$ *defined on a certain function space* $\mathcal{F}$ *such that*

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h) < t\right) - \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\mathbb{B}_n(f)| < t\right) \right| = O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

- Good news: we have the limiting distribution.

Theorem (Chen, Genovese, and Wasserman (2015))

Under regularity conditions and $\frac{\log n}{nh^{d+8}} \to 0$, *there exists a Gaussian process* $\mathbb{B}_n$ *defined on a certain function space* $\mathcal{F}$ *such that*

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+2}} \mathsf{Haus}(\widehat{R}_h, R_h) < t \right) - \mathbb{P}\left(\sup_{f \in \mathcal{F}} |\mathbb{B}_n(f)| < t \right) \right| = O\left(\left(\frac{\log^7 n}{nh^{d+2}} \right)^{1/8} \right).$$

- Good news: we have the limiting distribution.
- Bad news: the limiting distribution involves unknown quantities.

The Bootstrap

Theorem (Chen, Genovese, and Wasserman (2015))

Under regularity conditions and $\frac{\log n}{nh^{d+8}} \to 0$, *there exists a Gaussian process* $\mathbb{B}_n$ *defined on a certain function space* $\mathcal{F}$ *such that*

$$\sup_t \left| \mathbb{P}\left(\sqrt{nh^{d+2}}\,\mathsf{Haus}(\widehat{R}_h, R_h) < t\right) - \mathbb{P}\left(\sup_{f \in \mathcal{F}} |\mathbb{B}_n(f)| < t\right)\right| = O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

- Good news: we have the limiting distribution.
- Bad news: the limiting distribution involves unknown quantities.

⟶ A solution: the bootstrap.

Bootstrap Confidence Set

- Bootstrap sample $\Longrightarrow$ bootstrap ridges $\widehat{R}_h^*$.
- Repeat B times, we obtain B bootstrap ridges $\widehat{R}_h^{*(1)}, \cdots, \widehat{R}_h^{*(B)}$.
- Compute the CDF estimator $\widehat{F}_n$ by

$$\widehat{F}_n(t) = \frac{1}{B} \sum_{\ell=1}^{B} I\left(\mathsf{Haus}(\widehat{R}_h^{*(\ell)}, \widehat{R}_h) < t\right)$$

- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.
- The confidence set is

$$C_{1-\alpha,n} = \widehat{R}_h \oplus \widehat{t}_{1-\alpha}$$

Bootstrap Consistency

We proved that

$$\begin{aligned}\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h^*, \widehat{R}_h) &\approx \sup\{\text{Gaussian process on } \widehat{R}_h\}\\ &\approx \sup\{\text{Gaussian process on } R_h\}\\ &\approx \sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h).\end{aligned}$$

Bootstrap Consistency

We proved that

$$\begin{aligned}\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h^*, \widehat{R}_h) &\approx \sup\{\text{Gaussian process on } \widehat{R}_h\}\\ &\approx \sup\{\text{Gaussian process on } R_h\}\\ &\approx \sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h).\end{aligned}$$

This implies that $\widehat{t}_{1-\alpha}/t_{1-\alpha} \stackrel{P}{\to} 1$.

Bootstrap Consistency

We proved that

$$\begin{aligned}\sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h^*, \widehat{R}_h) &\approx \sup\{\text{Gaussian process on } \widehat{R}_h\} \\ &\approx \sup\{\text{Gaussian process on } R_h\} \\ &\approx \sqrt{nh^{d+2}}\mathsf{Haus}(\widehat{R}_h, R_h).\end{aligned}$$

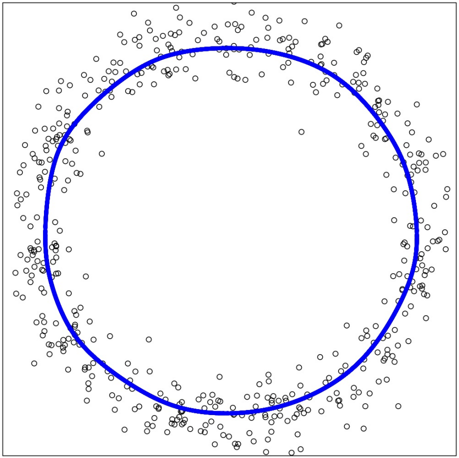
This implies that $\widehat{t}_{1-\alpha}/t_{1-\alpha} \xrightarrow{P} 1$.

Theorem (Chen, Genovese, and Wasserman (2015))

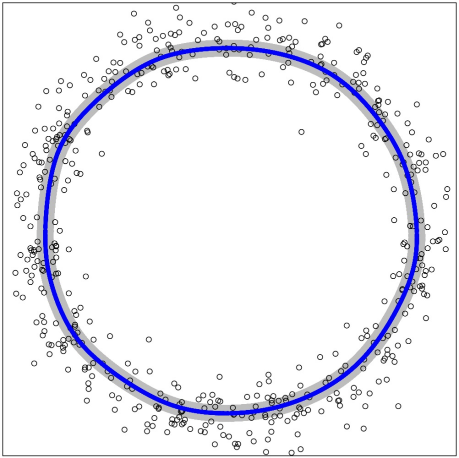
Under regularity conditions and $\frac{\log n}{nh^{d+8}} \to 0$,

$$\mathbb{P}\left(R_h \subset \widehat{R}_h \oplus \widehat{t}_{1-\alpha}\right) = 1 - \alpha + O\left(\left(\frac{\log^7 n}{nh^{d+2}}\right)^{1/8}\right).$$

Example of Confidence Sets

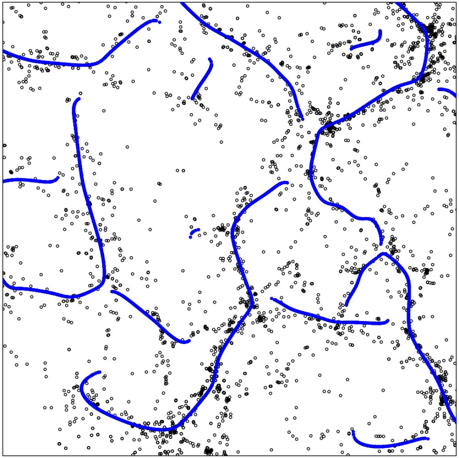


Example of Confidence Sets

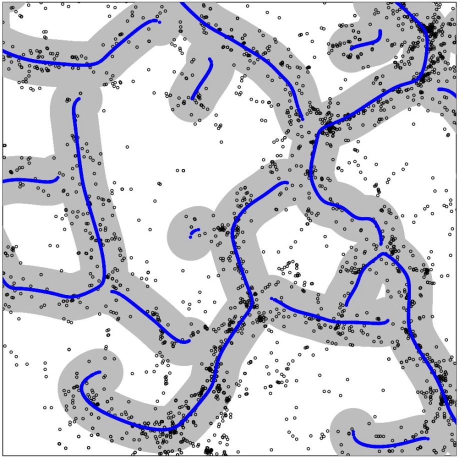


We have checked the coverage by simulation.

Example of Confidence Sets

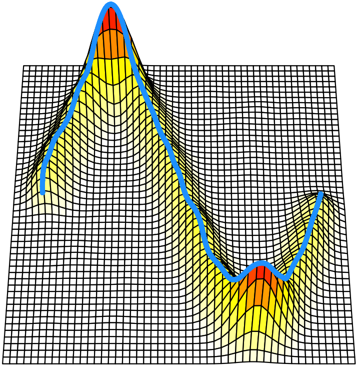


Example of Confidence Sets



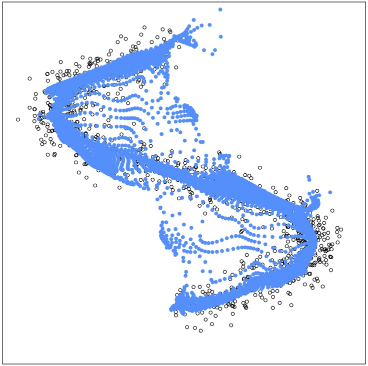
Summary for Density Ridges

- Ridges of the density function.



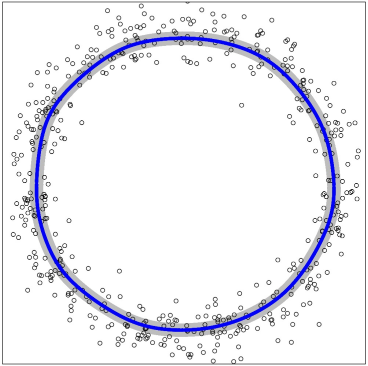
Summary for Density Ridges

- Ridges of the density function.
- An algorithm for the estimator.



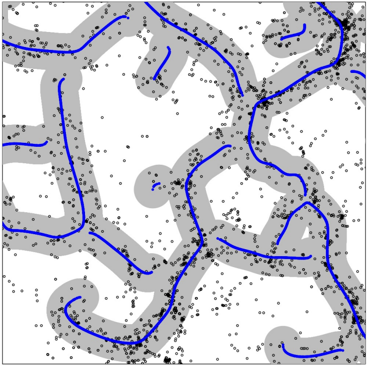
Summary for Density Ridges

- Ridges of the density function.
- An algorithm for the estimator.
- Confidence sets.



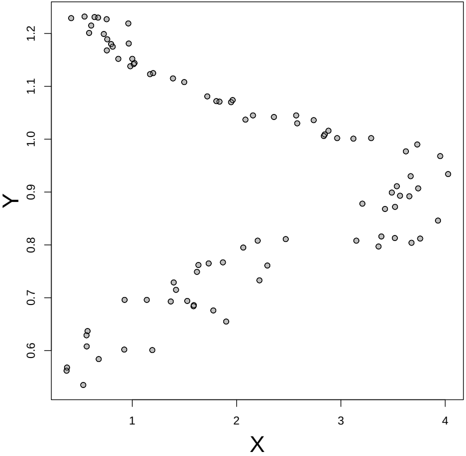
Summary for Density Ridges

- Ridges of the density function.
- An algorithm for the estimator.
- Confidence sets.
- Applications in Astronomy.

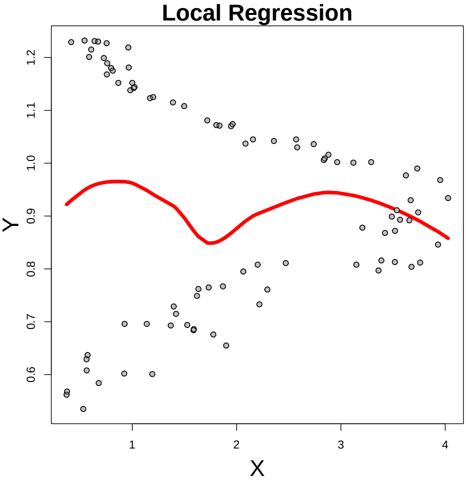


Modal Regression

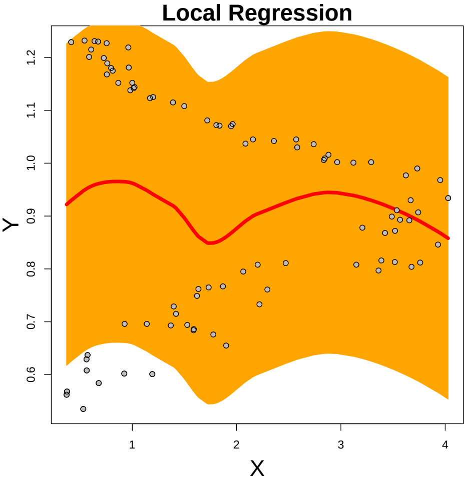
A Motivating Example for Modal Regression



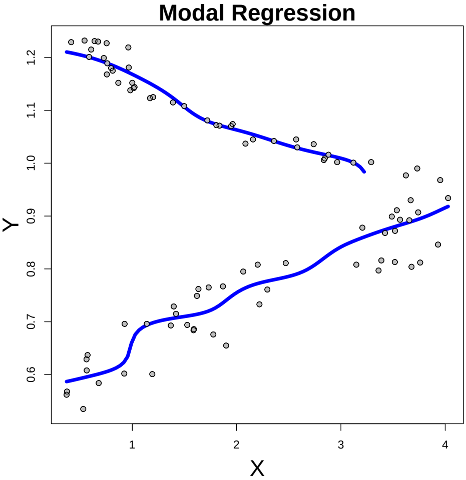
A Motivating Example for Modal Regression



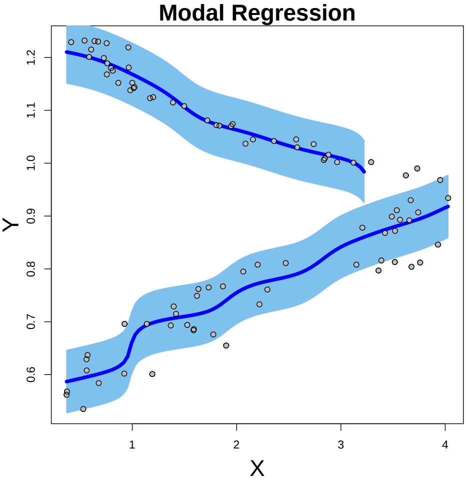
A Motivating Example for Modal Regression



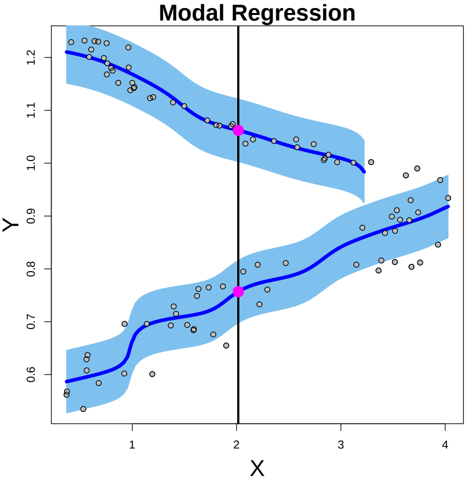
A Motivating Example for Modal Regression



A Motivating Example for Modal Regression



A Motivating Example for Modal Regression



Modal regression first appeared in Sager and Thisted (1982), Lee (1989), and Scott (1992).

Literature Review

Modal regression first appeared in Sager and Thisted (1982), Lee (1989), and Scott (1992).

Then it was used in meteorology (Hyndman et al., 1996), astronomy (Rojas, 2005), and transportation (Einbeck and Tutz, 2006).

Literature Review

Modal regression first appeared in Sager and Thisted (1982), Lee (1989), and Scott (1992).

Then it was used in meteorology (Hyndman et al., 1996), astronomy (Rojas, 2005), and transportation (Einbeck and Tutz, 2006).

Recently, some generalizations are proposed in Yao and Lindsay (2009), Yao et al. (2012), and Yao and Li (2014).

Literature Review

Modal regression first appeared in Sager and Thisted (1982), Lee (1989), and Scott (1992).

Then it was used in meteorology (Hyndman et al., 1996), astronomy (Rojas, 2005), and transportation (Einbeck and Tutz, 2006).

Recently, some generalizations are proposed in Yao and Lindsay (2009), Yao et al. (2012), and Yao and Li (2014).

In most of the above work, they consider the mode of the conditional density function.

$\longrightarrow$ In our work, we consider the multiple local modes of the conditional density function.

Definition for Modal Regression

We assume $x \in \mathbb{K} \subset \mathbb{R}^d$, where $\mathbb{K}$ is a compact set.

- 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 $x \in \mathbb{K} \subset \mathbb{R}^d$, where $\mathbb{K}$ is a compact set.

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

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

- $M(x)$ is a multi-valued function.

Definition for Modal Regression

We assume $x \in \mathbb{K} \subset \mathbb{R}^d$, where $\mathbb{K}$ is a compact set.

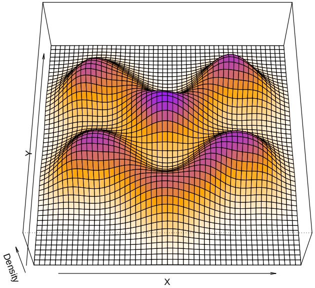
- Modal function–the conditional (local) **modes**:

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

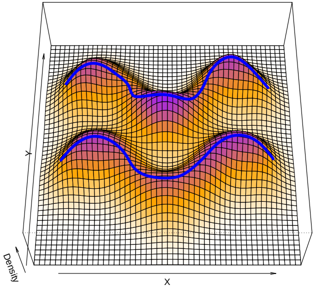
- $M(x)$ is a multi-valued function.
- An equivalent expression:

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

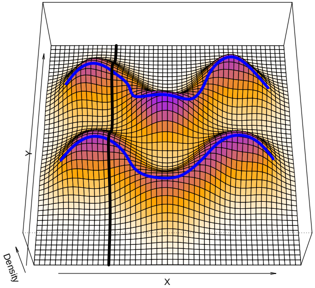
Conditional Local Modes



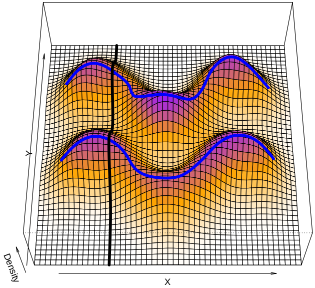
Conditional Local Modes

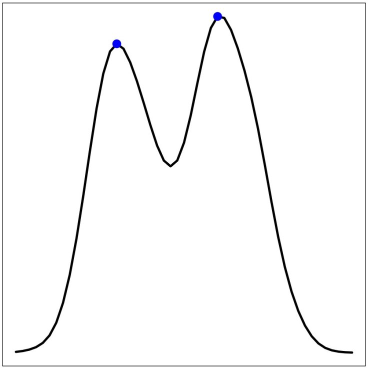


Conditional Local Modes



Conditional Local Modes





Estimator for Modal Regression

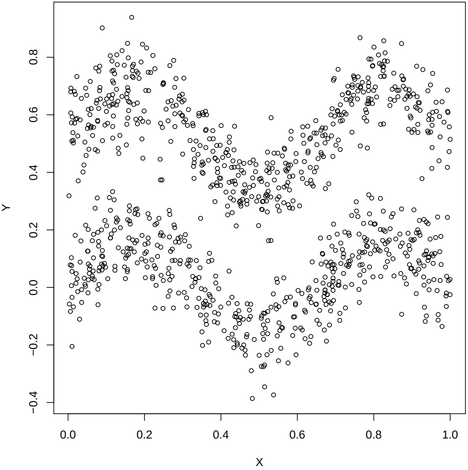
- Our estimator is the plug-in from the KDE:

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

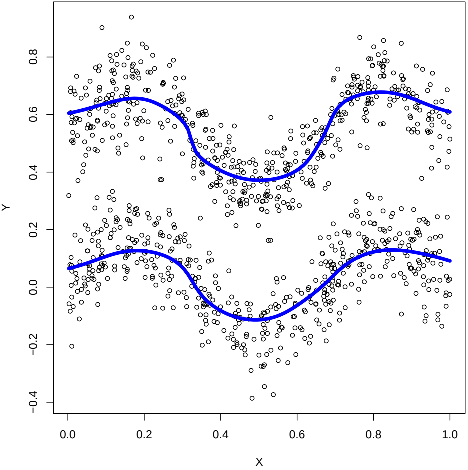
- Partial mean shift[2]: a simple algorithm for computing $\widehat{M}_n(x)$, the plug-in estimator of the KDE, from the data.

[2]Einbeck, Jochen, and Gerhard Tutz. "Modelling beyond regression functions: an application of multimodal regression to speed–flow data." JRSSC (2006)

Example for Modal Regression



Example for Modal Regression



Losses of Modal regression

To measure the errors, we consider the following two losses:

Losses of Modal regression

To measure the errors, we consider the following two losses:

- the *pointwise* loss

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

  where $\mathsf{Haus}(A, B)$ is the Hausdorff distance.

Losses of Modal regression

To measure the errors, we consider the following two losses:

- the *pointwise* loss

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

  where $\mathsf{Haus}(A, B)$ is the Hausdorff distance.
- the *uniform* loss

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

Illustration for Losses

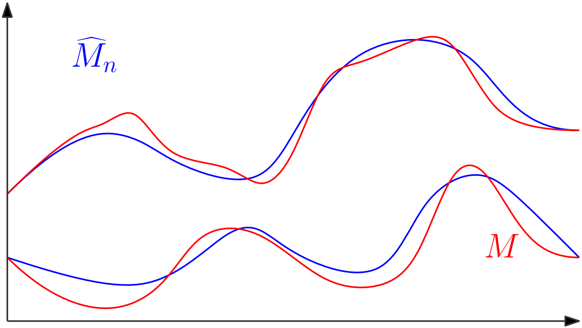


Illustration for Losses

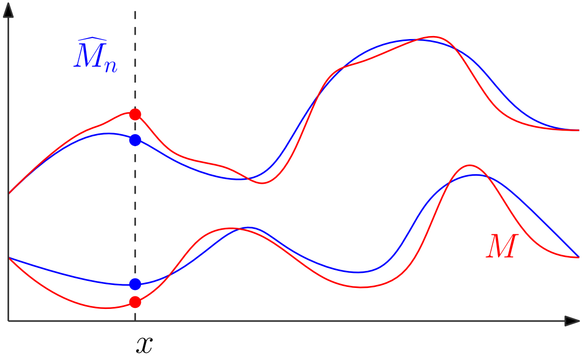


Illustration for Losses

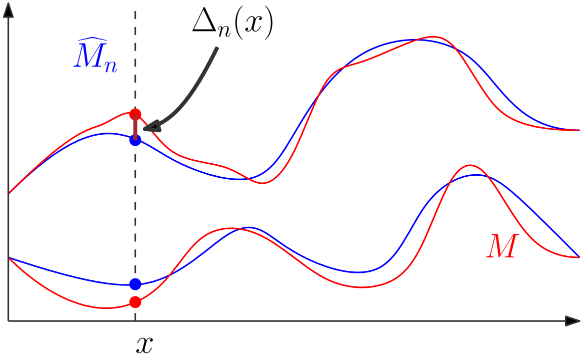
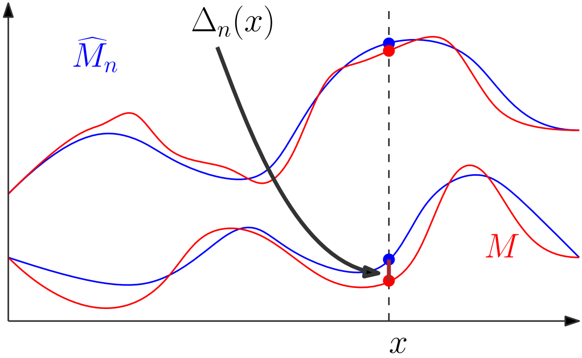


Illustration for Losses



Rate of Convergence

Both the pointwise and the uniform losses obey the common nonparametric rate:

Theorem (Chen, Genovese, and Wasserman (2016))

Under regularity conditions and $\frac{\log n}{nh^{d+3}} \to 0$,

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

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

Risk = Bias + $\sqrt{\text{Variance}}$.

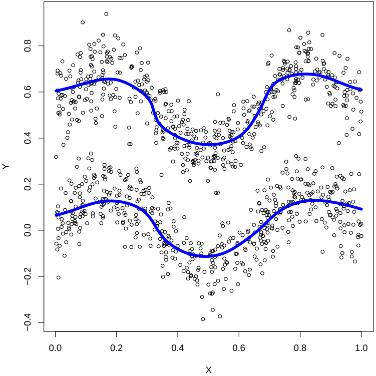
$d + 3 = d + 1 + 2 = \dim(X) + \dim(Y) + \text{gradient}.$

Confidence Sets

We can construct confidence sets using the uniform loss and the bootstrap.

Reason: the uniform loss Δ_n is an L_∞ metric for modal regression.

Bootstrap consistency follows in a similar way as density ridges.

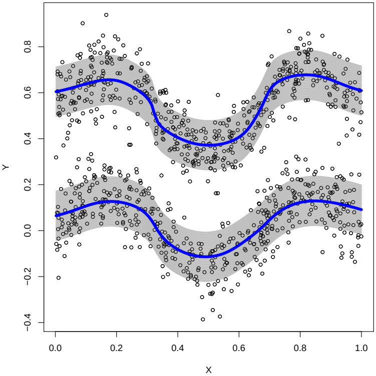


Confidence Sets

We can construct confidence sets using the uniform loss and the bootstrap.

Reason: the uniform loss Δ_n is an L_∞ metric for modal regression.

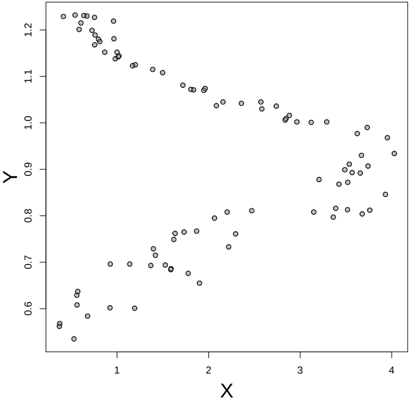
Bootstrap consistency follows in a similar way as density ridges.



Prediction Sets

We can use modal regression to construct a prediction set.

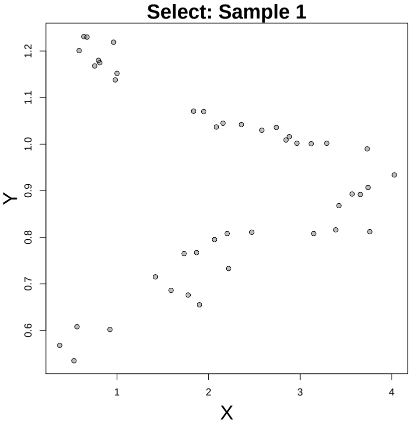
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

We can use modal regression to construct a prediction set.

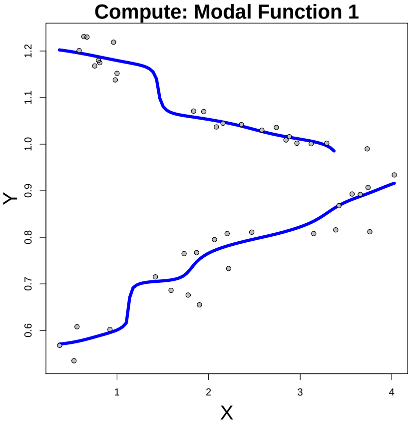
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

We can use modal regression to construct a prediction set.

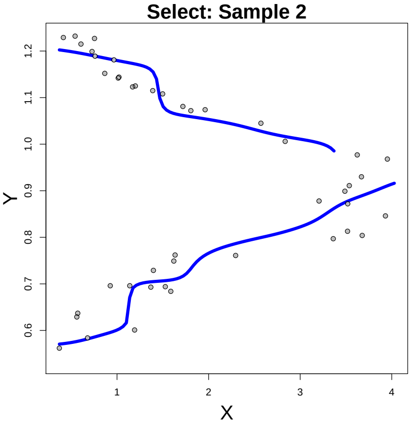
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

We can use modal regression to construct a prediction set.

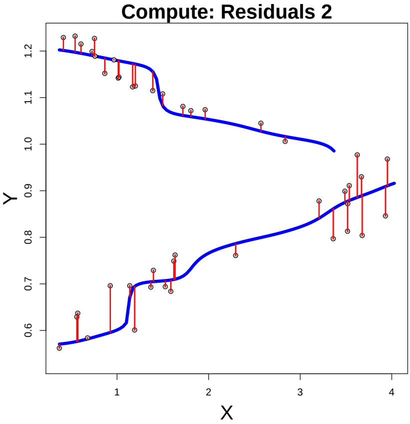
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

We can use modal regression to construct a prediction set.

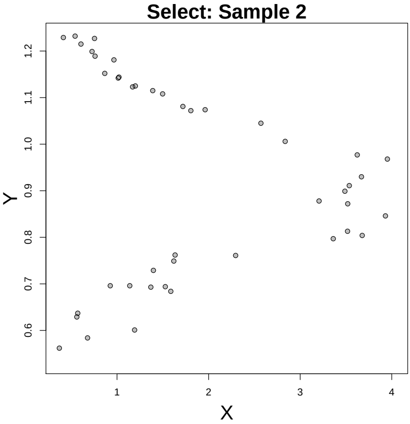
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

We can use modal regression to construct a prediction set.

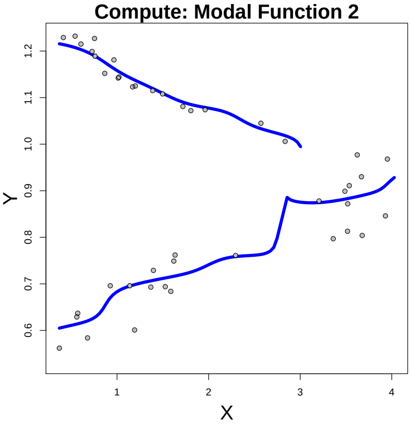
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

We can use modal regression to construct a prediction set.

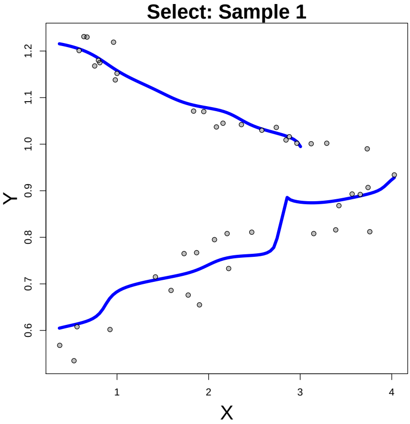
Here we use 2-fold cross validation to compute the prediction set.



Prediction Sets

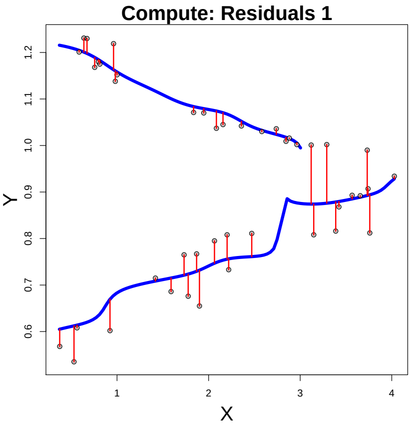
We can use modal regression to construct a prediction set.

Here we use 2-fold cross validation to compute the prediction set.



We can use modal regression to construct a prediction set.

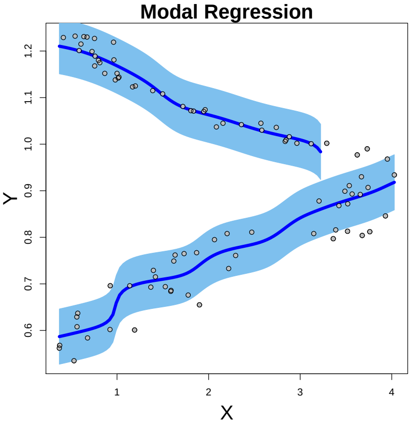
Here we use 2-fold cross validation to compute the prediction set.



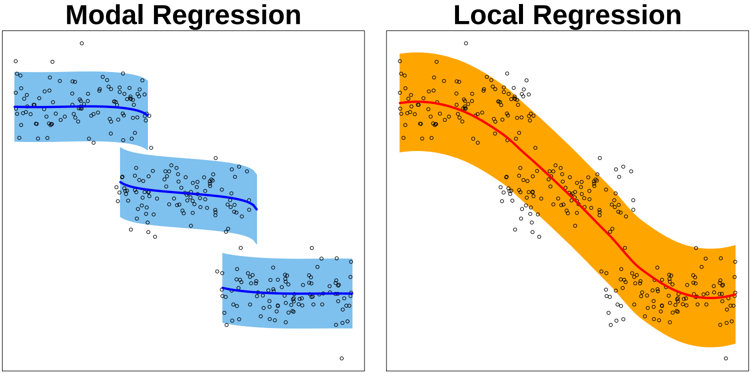
Prediction Sets

We can use modal regression to construct a prediction set.

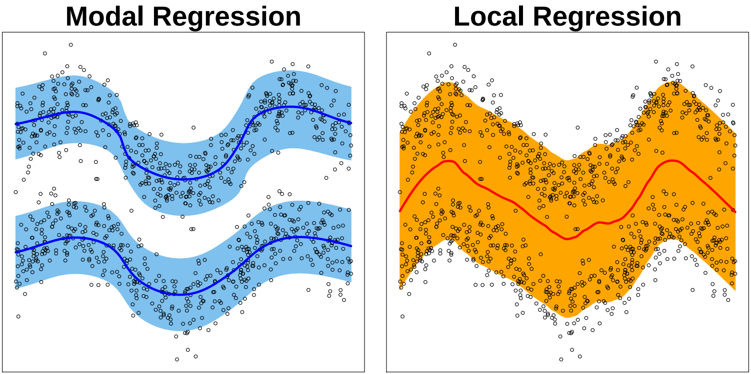
Here we use 2-fold cross validation to compute the prediction set.



Examples of Prediction Sets



Examples of Prediction Sets



Bandwidth Selection

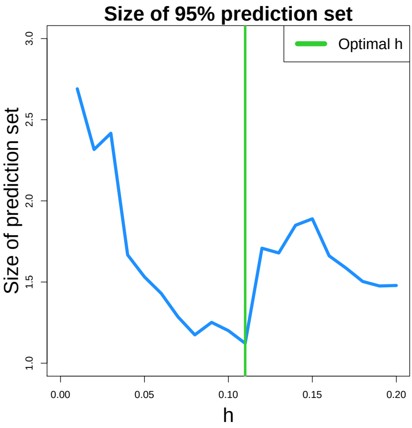
We can choose the smoothing parameter h via minimizing the size of the prediction set.

Namely, we choose

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

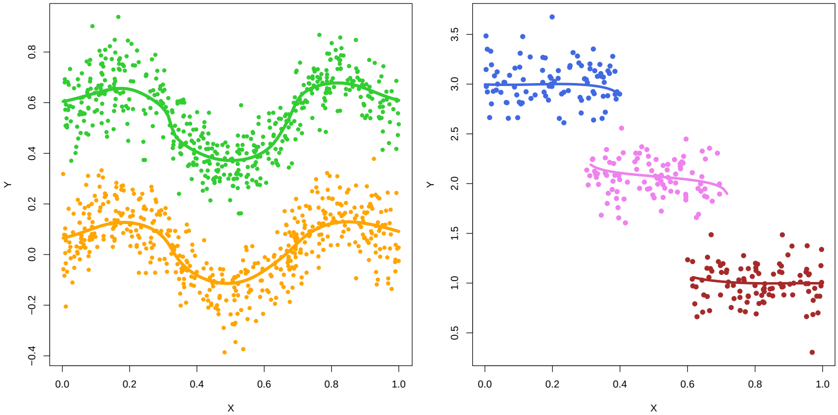
where $\widehat{\mathcal{P}}_{1-\alpha}$ is the prediction set.

Example: Bandwidth Selection



Regression Clustering

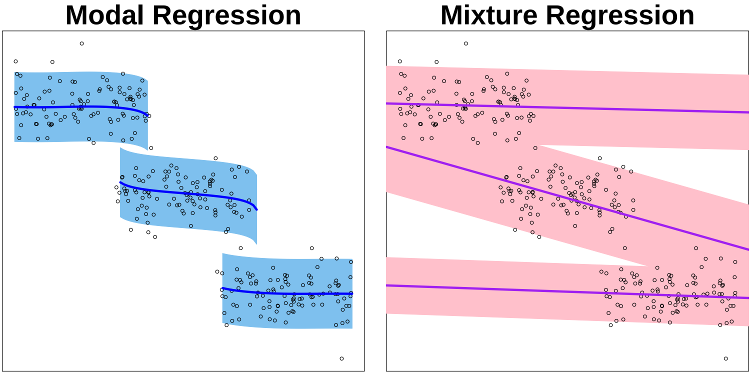
- Clustering based on the response Y.
- Clusters as functions of covariates X.



Modal Regression VS Mixture Regression

Modal regression and mixture regression are solving different problems.

Here is a case where modal regression gives a better result.



Concluding Remarks

Geometric Features

Common examples:

- Level Sets
- **Ridges**
- Clusters
- Density Trees
- **Modal Regression**

Applications:

- Astronomy
- Biology
- Image Analysis

→ **Chen** et al. 'Asymptotic Theory for Density Ridges.' *The Annals of Statistics* (2015).

→ **Chen** et al. 'Optimal Ridge Detection using Coverage Risk.' *NIPS* (2015).

→ **Chen** et al. 'Cosmic Web Reconstruction through Density Ridges: Method and Algorithm.' *Mon. Not. Roy. Astro. Soc.* (2015).

→ **Chen** et al. 'Nonparametric Modal Regression.' *The Annals of Statistics* (2016).

Geometric Features

Common examples:

- **Level Sets**
- Ridges
- **Clusters**
- **Density Trees**
- Modal Regression

Applications:

- Astronomy
- Biology
- Image Analysis

→ **Chen** et al. 'Density Level Sets: Asymptotics, Inference, and Visualization' *JASA-T&M* (2016+).

→ **Chen** et al. 'A Comprehensive Approach to Mode Clustering.' *The Electronic Journal of Statistics* (2016).

→ **Chen** et al. 'Statistical Inference Using the Morse-Smale Complex.' (2015).

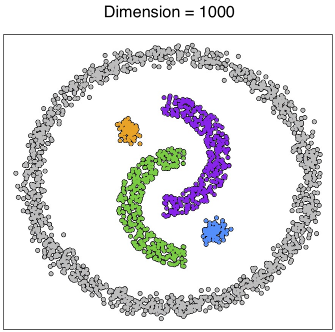
→ Kim and **Chen** et al. 'Confidence Sets for Density Trees.' *NIPS* (2016).

→ **Chen**. 'Generalized Cluster Trees and Singular Measures.' (2016).

Future Work

Some future directions:

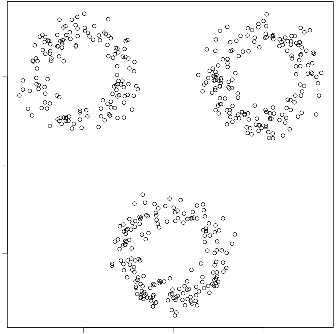
- More to do in geometric features.
- High-dimensional density clustering.
- Topological data analysis.



Future Work

Some future directions:

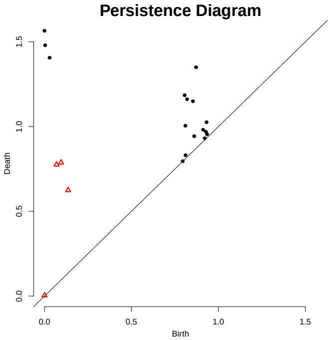
- More to do in geometric features.
- High-dimensional density clustering.
- Topological data analysis.



Future Work

Some future directions:

- More to do in geometric features.
- High-dimensional density clustering.
- Topological data analysis.



More details can be found in:

http://faculty.washington.edu/yenchic/

More details can be found in:

http://faculty.washington.edu/yenchic/

Thank you!

Backups for Density Ridges

Regularity Conditions

(K1) The kernel function K is $\mathbf{BC}^4$ and integrable.

(K2) K satisfies the VC-type class condition.

(P1) The density p is in $\mathbf{BC}^4$.

(P2) The eigengap $\lambda_1(x) - \lambda_2(x) \geq \beta_0 > 0$ for points around ridges.

(P3) The orientation of each ridge point is close to the gradient.

Regularity Conditions on Kernel Functions

(K1) The kernel K is in $\mathbf{BC}^4$ and $\|K\|^*_{\infty,4} < \infty$.

(K2) Let

$$\mathcal{K}_r = \left\{ y \mapsto K^{(\alpha)}\left(\frac{x-y}{h}\right) : x \in \mathbb{R}^d, |\alpha| = r \right\},$$

where $K^{(\alpha)}$ is the α-th derivative and let $\mathcal{K}^*_l = \bigcup_{r=0}^{l} \mathcal{K}_r$. We assume that $\mathcal{K}^*_4$ is a VC-type class. i.e. there exists constants A, v and a constant envelope b_0 such that

$$\sup_Q N(\mathcal{K}^*_4, \mathcal{L}^2(Q), b_0\epsilon) \leq \left(\frac{A}{\epsilon}\right)^v, \tag{1}$$

where $N(T, d_T, \epsilon)$ is the ϵ-covering number for an semi-metric set T with metric d_T and $\mathcal{L}^2(Q)$ is the L_2 norm with respect to the probability measure Q.

Regularity Conditions on Distributions

(P1) The density p_h is in $\mathbf{BC}^4$.

(P2) There exists constants $\beta_0, \beta_1, \beta_2, \delta_0 > 0$ such that

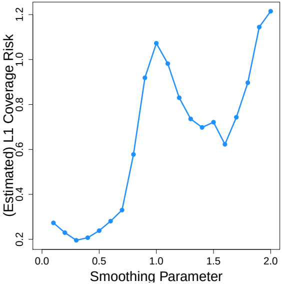
$$\begin{aligned} \lambda_2(x) &\leq -\beta_1 \\ \lambda_1(x) &\geq \beta_0 - \beta_1 \\ \|g_h(x)\| \max_{|\alpha|=3} |p_h^{(\alpha)}(x)| &\leq \beta_0(\beta_1 - \beta_2) \end{aligned} \tag{2}$$

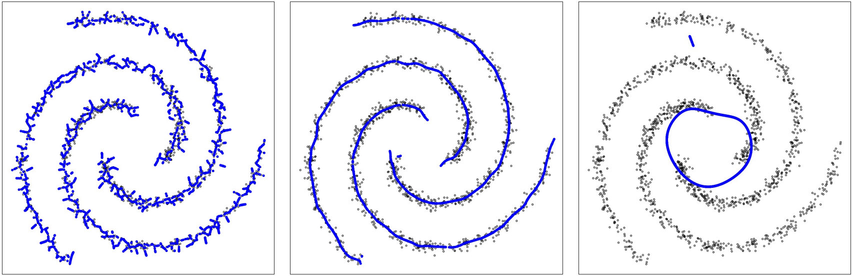
for all $x \in R_h \oplus \delta_0$.

(P3) For each $x \in R_h$, $|e(x)^T g_h(x)|^2 \geq \frac{\lambda_1(x)}{\lambda_1(x) - \lambda_2(x)}$ where $e(x)$ is the direction of R_h at point $x \in R_h$.

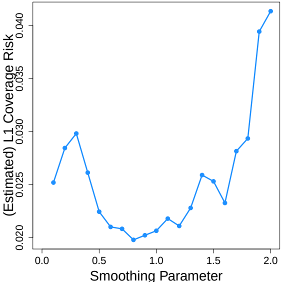
(P4) The above assumptions hold for all sufficiently small h.

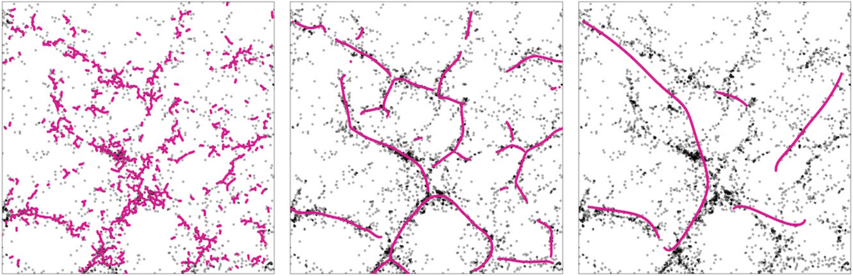
Bandwidth Selection



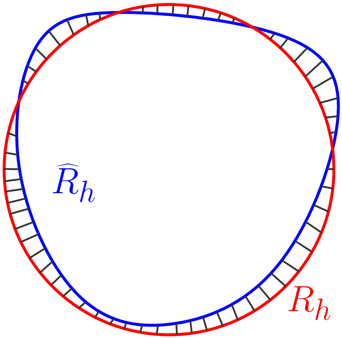


Bandwidth Selection





Bandwidth Selection

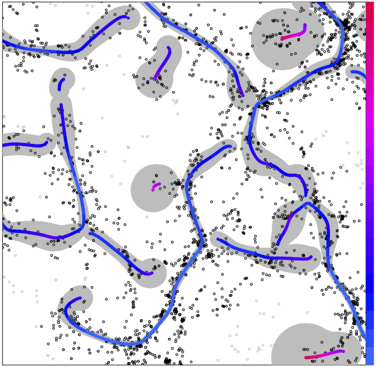


L_1 distance are like the area of the shady regions.

We estimate this distance by data splitting or the bootstrap.

Reference: **Chen** et al. 'Optimal Ridge Detection using Coverage Risk' (NIPS 2015).

Local Uncertainty and Pointwise Confidence Sets

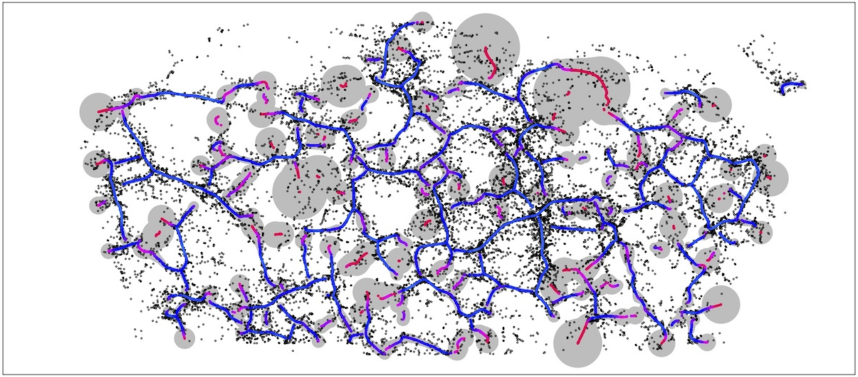


Color denotes the amount of uncertainty.

Red: unstable filaments.

Blue: stable filaments.

Local Uncertainty and Pointwise Confidence Sets



Color denotes the amount of uncertainty.

Red: unstable filaments.

Blue: stable filaments.

Why Smoothed Density? - Bias Consideration

We have the following decomposition:

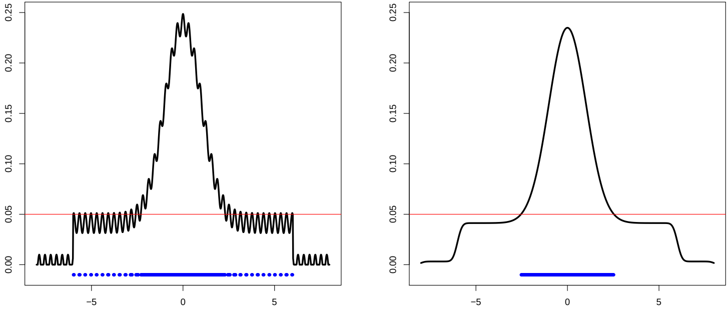
$$
\begin{aligned}
\mathsf{Haus}(\widehat{R}_h, R) &\le \mathsf{Haus}(R_h, R) + \mathsf{Haus}(\widehat{R}_h, R) \\
&= O(h^2) + O_P\left(\sqrt{\frac{\log n}{n h^{d+2}}}\right).
\end{aligned}
$$

Bias + $\sqrt{\text{Variance}}$.

Work on smoothed ridges R_h allows us to avoid the problem of bias.

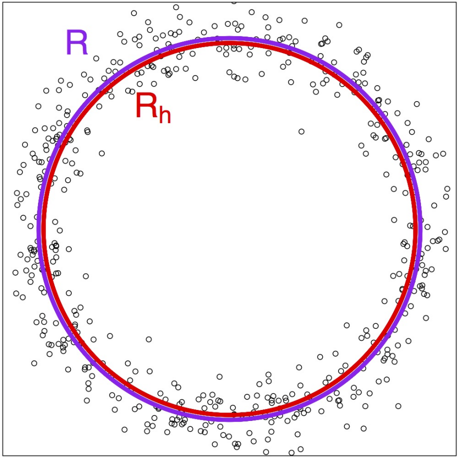
Optimal rate: $O_P\left(\left(\frac{\log n}{n}\right)^{\frac{2}{d+6}}\right)$ when we choose $h = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d+6}}\right)$.

Why Smoothed Density? - A Level Set Example



Ridges VS Smoothed Ridges

Radius of ring: $r = 1$. Smoothing bandwidth: $h = 0.25$. Gaussian noise level: $\sigma = 0.1$



General Ridges

We can generalize ridges to higher dimensions. Pick $V_r(x) = [v_{r+1}(x), \cdots, v_d(x)]$.

We define

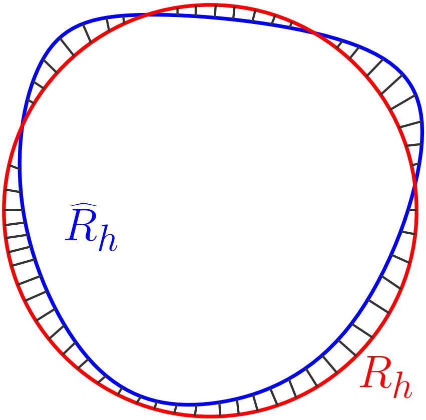
$$r\text{-Ridge}(p) = \{x : V_r(x)V_r(x)^T \nabla p(x) = 0, \lambda_{r+1}(x) < 0\}.$$

$V_r(x)$ is a $d \times (d-r)$ matrix. There are $d-r$ constraints.

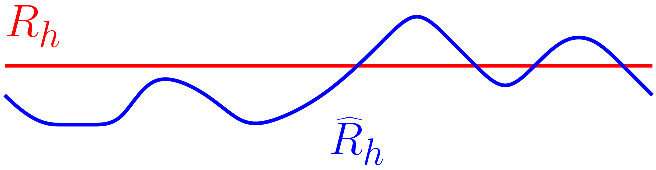
By Implicit Function Theorem, r-ridges are r-manifolds.

In Astronomy, $r = 2$ can be used to model 'Cosmic Sheets (Walls)'.

$r = 0$ coincides with the definition of local modes.


$\widehat{R}_h$
R_h

Asymptotic Theory



Backups for Modal Regression

Regularity Conditions

(K1) The kernel function K is $\mathbf{BC}^4$ and integrable.

(K2) K satisfies the VC-type class condition.

(P1) The density p is in $\mathbf{BC}^4$.

(P2) The second derivative along y axis is bounded away from 0 for points on M.

(P3) M contains L well-separated, connected components.

Regularity Conditions on Kernel Functions

(K1) The kernel K is in $\mathbf{BC}^4$ and $\|K\|^*_{\infty,4} < \infty$.

(K2) Let

$$\mathcal{K}_r = \left\{ y \mapsto K^{(\alpha)}\left(\frac{x-y}{h}\right) : x \in \mathbb{R}^d, |\alpha| = r \right\},$$

where $K^{(\alpha)}$ is the α-th derivative and let $\mathcal{K}^*_l = \bigcup_{r=0}^{l} \mathcal{K}_r$. We assume that $\mathcal{K}^*_2$ is a VC-type class. i.e. there exists constants A, v and a constant envelope b_0 such that

$$\sup_Q N(\mathcal{K}^*_2, \mathcal{L}^2(Q), b_0\epsilon) \leq \left(\frac{A}{\epsilon}\right)^v, \tag{3}$$

where $N(T, d_T, \epsilon)$ is the ϵ-covering number for an semi-metric set T with metric d_T and $\mathcal{L}^2(Q)$ is the L_2 norm with respect to the probability measure Q.

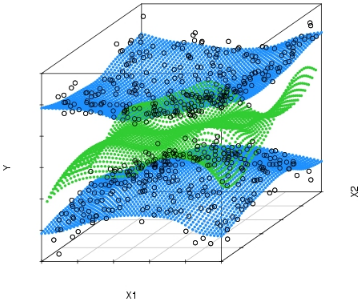
Regularity Conditions on Distributions

(P1) The density p is in $\mathbf{BC}^4$.

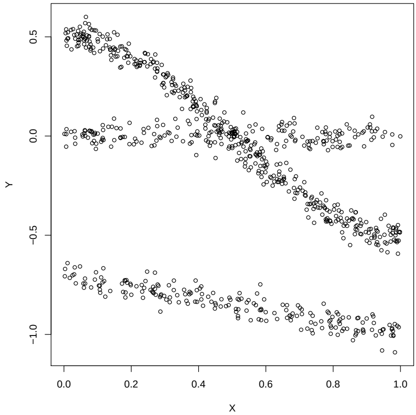
(P2) There exists constants $\lambda_0 > 0$ such that for any $(x, y) \in \mathbb{K} \times \mathbb{R}$ with $\frac{\partial}{\partial y} p(x, y) > 0$, the second derivative satisfies $\frac{\partial^2}{\partial^2 y} p(x, y) \leq -\lambda_0 < 0$.

(P3) Modal function $M = \cup_{j=1}^{L} M_j$, where each M_j is a connected component with $M_j \cap M_i = \phi$ for $i \neq j$.

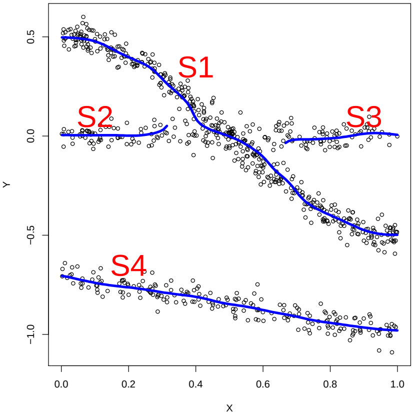
3D Modal Regression



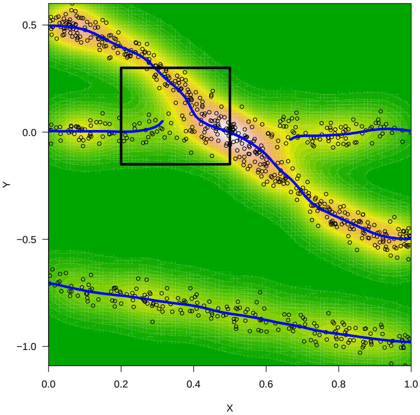
Bifurcation and Merge



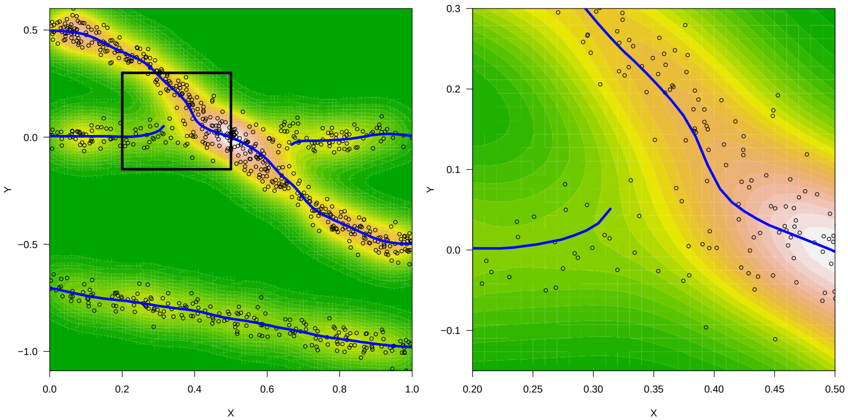
Bifurcation and Merge



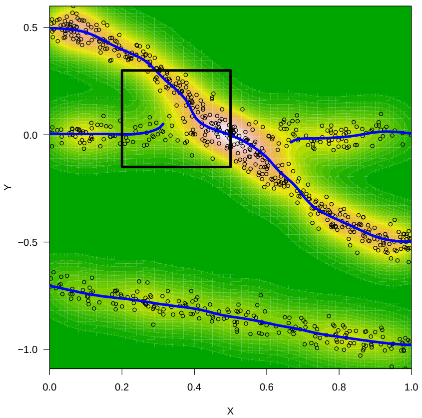
Bifurcation and Merge

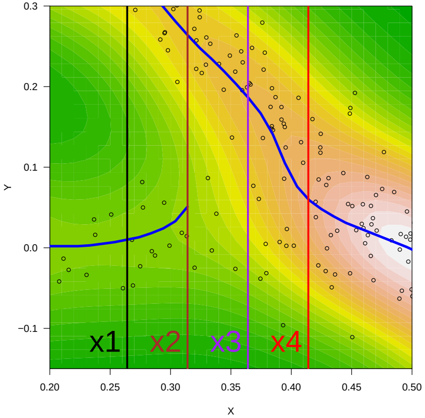


Bifurcation and Merge

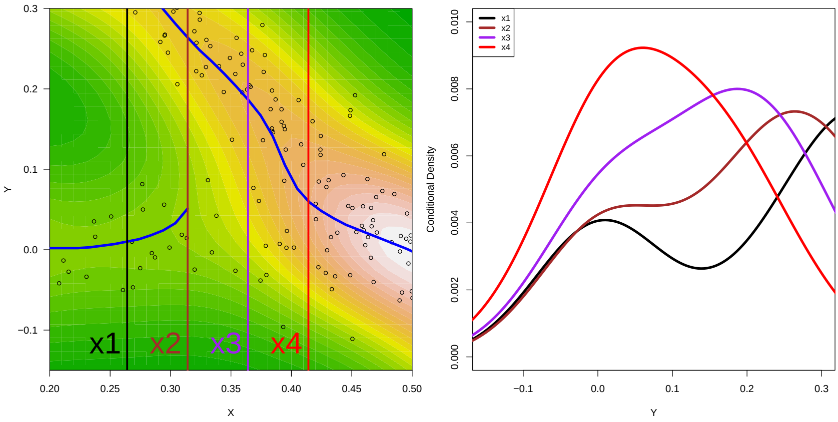


Bifurcation and Merge

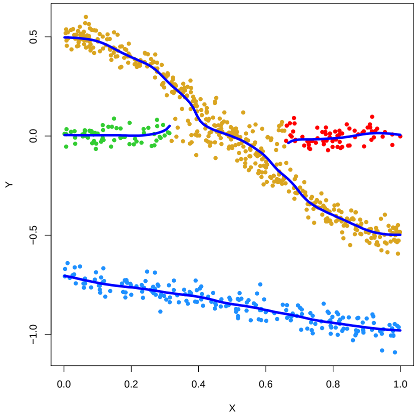




Bifurcation and Merge



Bifurcation and Merge



Comments on Mixture Regression

A general model for mixture regression:

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

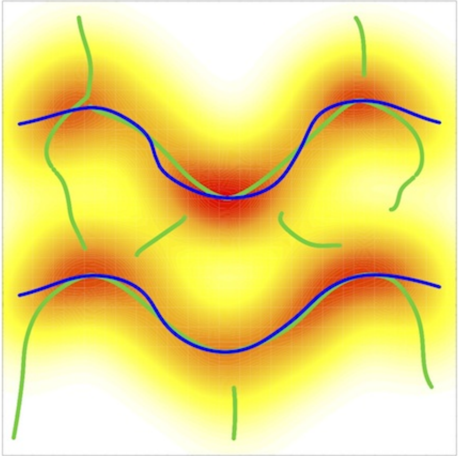
where each $\phi_j(y; \mu, \sigma^2)$ is a density function with mean μ and variance σ^2.

Common assumptions:

1. $\pi_j(x) = \pi_j$.
2. $\mu_j(x) = \beta_j^T x$.
3. $\sigma_j^2(x) = \sigma_j^2$.
4. ϕ_j is a Gaussian.

Generally, we need to use EM algorithm to estimate the parameters.

Modal Regression VS Density Ridges

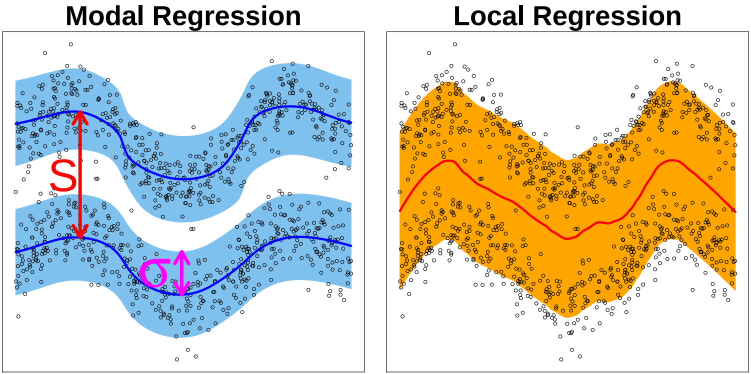


Mixture Inference versus Modal Inference

	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.

Theory for Prediction Sets



Theorem (Chen, Genovese, and Wasserman (2015))

When the signal-to-noise ratio S/σ is sufficiently large, the modal regression has a smaller prediction set than the nonparametric regression.

Confidence Sets

We can construct confidence sets using the uniform loss.

Reason: the uniform loss Δ_n is like an L_∞ metric for modal regression.

Let $t_{1-\alpha}$ be the $1-\alpha$ quantile of F_n, the CDF of Δ_n.

$\widehat{M}_n(x) \pm t_{1-\alpha}$ is a confidence set for $M(x)$ uniformly for all x.

Problem: $t_{1-\alpha}$ cannot be computed.

Solution: the bootstrap.

The Bootstrap

- Bootstrap sample $\Longrightarrow$ bootstrap modal function $\widehat{M}_n^*$.
- Repeat B times, we obtain B bootstrap modal functions $\widehat{M}_n^{*(1)}, \cdots, \widehat{M}_n^{*(B)}$.
- Compute $\widehat{\Delta}_n^{*(1)}, \cdots, \widehat{\Delta}_n^{*(B)}$ by $\widehat{\Delta}_n^{*(\ell)} = \sup_x \mathsf{Haus}(\widehat{M}_n^{*(\ell)}(x), \widehat{M}_n(x))$.
- Compute the CDF estimator $\widehat{F}_n$ by

$$\widehat{F}_n(t) = \frac{1}{B} \sum_{\ell=1}^{B} I\left(\widehat{\Delta}_n^{*(\ell)} < t\right).$$

- Choose $\widehat{t}_{1-\alpha}$ be the $1-\alpha$ quantile for $\widehat{F}_n$.
- $\widehat{M}_n(x) \pm \widehat{t}_{1-\alpha}$ is an asymptotic confidence set uniformly for all x.

Bootstrap consistency follows in the similar way as ridges.

Pointwise Confidence Sets