Understanding Nonparametric Multimodal Regression via Kernel Density Estimation

A. Bhattacharjee[*] R. Mondal[*] R. Vasishtha[*] S. S. Banerjee[*]

[*]Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

February 20, 2022

Contents

1. Introduction
   - Modal Regression
2. Estimation
   - Mean-shift Algorithm
3. Geometry
   - Modal Manifolds
   - Derivative of Modal Manifold Collection
4. Consistency
5. Confidence Sets
6. Prediction Sets
   - Bandwidth Selection
7. References

Motivation

- Why modal regression?
- Conventional regression methods may fail when:
  - conditional distribution is heavy-tailed;
  - conditional distribution is multi-modal.
- Why nonparametric modal regression?
- Taking a nonparametric model allows for more flexibility unlike a (restrictive) parametric model: $\texttt{Mode}(Y|X = x) = \beta_0 + \beta^T x$ (Sager and Thisted (1982)).

Motivation

- Why modal regression?
- Conventional regression methods may fail when:
  - conditional distribution is heavy-tailed;
  - conditional distribution is multi-modal.
- Why nonparametric modal regression?
- Taking a nonparametric model allows for more flexibility unlike a (restrictive) parametric model: $\mathtt{Mode}(Y|X=x) = \beta_0 + \beta^T x$ (Sager and Thisted (1982)).

Motivation

- Why modal regression?
- Conventional regression methods may fail when:
  - conditional distribution is heavy-tailed;
  - conditional distribution is multi-modal.
- Why nonparametric modal regression?
- Taking a nonparametric model allows for more flexibility unlike a (restrictive) parametric model: $\texttt{Mode}(Y|X = x) = \beta_0 + \beta^T x$ (Sager and Thisted (1982)).

Motivating Examples

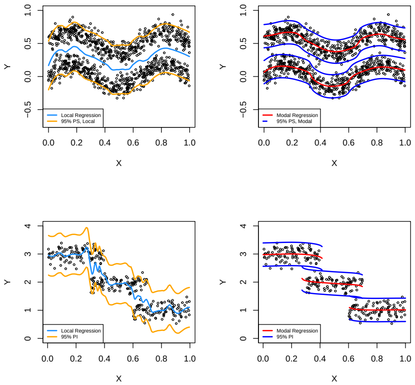

Figure: We show local regression estimate and its associated 95% prediction bands alongside the modal regression and its 95% prediction bands for two different simulated data.

Definitions

- We define operators:
  $\texttt{UniMode} = \arg\max_z f(z), \ \texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\}.$

Definition (Uni-modal function)

$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x).$

Definition (Multi-modal function)

$M(x) = \texttt{MultiMode}(Y|X = x) = \{y : \frac{\partial}{\partial y} p(y|x) = 0, \ \frac{\partial^2}{\partial y^2} p(y|x) < 0\}.$

- Equivalently, we can write,

$$m(x) = \arg\max_y p(x, y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x, y) = 0, \ \frac{\partial^2}{\partial y^2} p(x, y) < 0\}. \tag{1}$$

- We will focus on multi-modal regression (Chen et al. (2016)). Why?

Definitions

- We define operators:
$\texttt{UniMode} = \arg\max_z f(z), \ \texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\}.$

Definition (Uni-modal function)

$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x).$

Definition (Multi-modal function)

$M(x) = \texttt{MultiMode}(Y|X = x) = \{y : \frac{\partial}{\partial y} p(y|x) = 0, \ \frac{\partial^2}{\partial y^2} p(y|x) < 0\}.$

- Equivalently, we can write,

$$m(x) = \arg\max_y p(x, y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x, y) = 0, \ \frac{\partial^2}{\partial y^2} p(x, y) < 0\}. \tag{1}$$

- We will focus on multi-modal regression (Chen et al. (2016)). Why?

Definitions

- We define operators:
  $\texttt{UniMode} = \arg\max_z f(z)$, $\texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\}$.

Definition (Uni-modal function)

$$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x).$$

Definition (Multi-modal function)

$$M(x) = \texttt{MultiMode}(Y|X = x) = \{y : \tfrac{\partial}{\partial y} p(y|x) = 0, \ \tfrac{\partial^2}{\partial y^2} p(y|x) < 0\}.$$

- Equivalently, we can write,

$$m(x) = \arg\max_y p(x, y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x, y) = 0, \ \frac{\partial^2}{\partial y^2} p(x, y) < 0\}. \tag{1}$$

- We will focus on multi-modal regression (Chen et al. (2016)). Why?

Definitions

- We define operators:
  $\texttt{UniMode} = \arg\max_z f(z)$, $\texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\}$.

Definition (Uni-modal function)

$$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x).$$

Definition (Multi-modal function)

$$M(x) = \texttt{MultiMode}(Y|X = x) = \{y : \tfrac{\partial}{\partial y} p(y|x) = 0,\ \tfrac{\partial^2}{\partial y^2} p(y|x) < 0\}.$$

- Equivalently, we can write,

$$m(x) = \arg\max_y p(x,y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x,y) = 0,\ \frac{\partial^2}{\partial y^2} p(x,y) < 0\}. \tag{1}$$

- We will focus on multi-modal regression (Chen et al. (2016)). Why?

Definitions

- We define operators:
  $\texttt{UniMode} = \arg\max_z f(z)$, $\texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\}$.

Definition (Uni-modal function)

$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x)$.

Definition (Multi-modal function)

$M(x) = \texttt{MultiMode}(Y|X = x) = \{y : \frac{\partial}{\partial y} p(y|x) = 0,\ \frac{\partial^2}{\partial y^2} p(y|x) < 0\}$.

- Equivalently, we can write,

$$m(x) = \arg\max_y p(x, y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x, y) = 0,\ \frac{\partial^2}{\partial y^2} p(x, y) < 0\}. \tag{1}$$

- We will focus on multi-modal regression (Chen et al. (2016)). Why?

Definitions

- We define operators:
  $\texttt{UniMode} = \arg\max_z f(z), \;\; \texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\}.$

Definition (Uni-modal function)

$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x).$

Definition (Multi-modal function)

$M(x) = \texttt{MultiMode}(Y|X = x) = \{y : \frac{\partial}{\partial y} p(y|x) = 0, \; \frac{\partial^2}{\partial y^2} p(y|x) < 0\}.$

- Equivalently, we can write,

$$m(x) = \arg\max_y p(x, y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x, y) = 0, \; \frac{\partial^2}{\partial y^2} p(x, y) < 0\}. \tag{1}$$

- We will focus on multi-modal regression (Chen et al. (2016)). Why?

Uni-modal vs. Multi-modal Regression

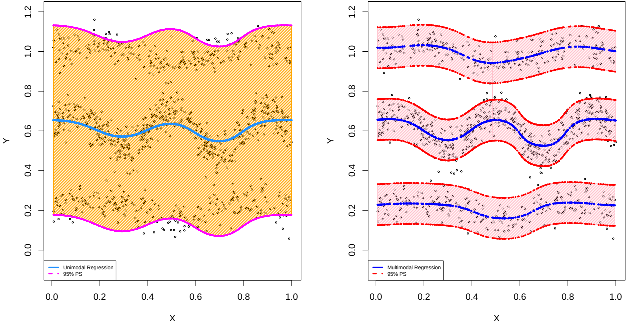

Figure: Uni-modal regression and multi-modal regression along with their corresponding 95% prediction sets on a simulated data with three components.

Modal Regression Estimators

- Our estimator is plug-in from the KDE:

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

where

$$\hat{p}_n(x,y) = \frac{1}{nh^{d+1}} \sum_{i=1}^{n} K\left(\frac{||x - X_i||}{h}\right) K\left(\frac{y - Y_i}{h}\right). \quad (3)$$

- To compute $\hat{M}_n(x)$ from the data, we use the *mean-shift algorithm* (Einbeck and Tutz (2006)).

Modal Regression Estimators

- Our estimator is plug-in from the KDE:

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

where

$$\hat{p}_n(x,y) = \frac{1}{nh^{d+1}}\sum_{i=1}^{n} K\left(\frac{||x - X_i||}{h}\right) K\left(\frac{y - Y_i}{h}\right). \quad (3)$$

- To compute $\hat{M}_n(x)$ from the data, we use the *mean-shift algorithm* (Einbeck and Tutz (2006)).

The Mean-shift Algorithm

Input: Data samples $\mathscr{D} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, bandwidth h.
(The kernel K is assumed to be Gaussian.)

1. Initialize mesh points $\mathscr{M} \subset R^{d+1}$ (a common choice is $\mathscr{M} = \mathscr{D}$, the data samples).
2. For each $(x, y) \in \mathscr{M}$, fix x, and update y using the following iterations until convergence:

$$y \longleftarrow \frac{\sum_{i=1}^{n} Y_i K\left(\frac{||x-X_i||}{h}\right) K\left(\frac{y-Y_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{||x-X_i||}{h}\right) K\left(\frac{y-Y_i}{h}\right)} \tag{4}$$

Output: The set $\mathscr{M}^{\infty}$, containing the points (x, y^{∞}), where x is a predictor value as fixed in $\mathscr{M}$, and y^{∞} is the corresponding limit of the mean-shift iterations .

Algorithm 1: Partial mean-shift algorithm

Modal Manifolds Collection: Definitions

- We define a *modal manifold collection* over all inputs x as:

$$\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\}$$

- We assume $\mathbb{S}$ can be factorized as:

$$\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\} = \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_K, \tag{5}$$

where each $\mathbb{S}_j$, $j = 1, 2, \ldots, K$ is a connected manifold defined as follows:

$$\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\} \tag{6}$$

for some function $m_j(x)$ and open set A_j.
- As a convention, $m_j(x) = \phi$ if $x \notin A_j$.
- This effectively allows us to write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}.$$

Modal Manifolds Collection: Definitions

- We define a *modal manifold collection* over all inputs x as:

$$\mathbb{S} = \{(x,y) : x \in D, y \in M(x)\}$$

- We assume $\mathbb{S}$ can be factorized as:

$$\mathbb{S} = \{(x,y) : x \in D, y \in M(x)\} = \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_K, \tag{5}$$

where each $\mathbb{S}_j$, $j = 1,2,\ldots,K$ is a connected manifold defined as follows:

$$\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\} \tag{6}$$

for some function $m_j(x)$ and open set A_j.

- As a convention, $m_j(x) = \phi$ if $x \notin A_j$.
- This effectively allows us to write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}.$$

Modal Manifolds Collection: Definitions

- We define a *modal manifold collection* over all inputs x as:

$$\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\}$$

- We assume $\mathbb{S}$ can be factorized as:

$$\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\} = \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_K, \quad (5)$$

  where each $\mathbb{S}_j$, $j = 1, 2, \ldots, K$ is a connected manifold defined as follows:

$$\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\} \quad (6)$$

  for some function $m_j(x)$ and open set A_j.
- As a convention, $m_j(x) = \phi$ if $x \notin A_j$.
- This effectively allows us to write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}.$$

Modal Manifolds Collection: Definitions

- We define a *modal manifold collection* over all inputs x as:

$$\mathbb{S} = \{(x,y) : x \in D, y \in M(x)\}$$

- We assume $\mathbb{S}$ can be factorized as:

$$\mathbb{S} = \{(x,y) : x \in D, y \in M(x)\} = \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_K, \tag{5}$$

  where each $\mathbb{S}_j$, $j = 1,2,\ldots,K$ is a connected manifold defined as follows:

$$\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\} \tag{6}$$

  for some function $m_j(x)$ and open set A_j.
- As a convention, $m_j(x) = \phi$ if $x \notin A_j$.
- This effectively allows us to write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}.$$

Modal Manifolds Collection: Definitions

- We define a *modal manifold collection* over all inputs x as:

$$\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\}$$

- We assume $\mathbb{S}$ can be factorized as:

$$\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\} = \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_K, \tag{5}$$

  where each $\mathbb{S}_j$, $j = 1, 2, \ldots, K$ is a connected manifold defined as follows:

$$\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\} \tag{6}$$

  for some function $m_j(x)$ and open set A_j.
- As a convention, $m_j(x) = \phi$ if $x \notin A_j$.
- This effectively allows us to write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}.$$

Modal Manifold Collection: An example

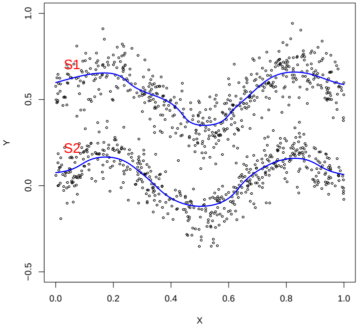

Figure: S1 and S2 represent modal manifolds.

Derivative of Modal Functions

Lemma (Derivative of modal functions)

Assume that p is twice differentiable, and let $\mathbb{S} = \{(x, y) : x \in D, y \in M(x)\}$ be the modal manifold collection. Assume that $\mathbb{S}$ factorizes according to (5), (6). Then, when $x \in A_j$,

$$\nabla m_j(x) = -\frac{p_{yx}(x, m_j(x))}{p_{yy}(x, m_j(x))} \tag{7}$$

where $p_{yx} = \nabla_x \frac{\partial}{\partial y} p(x, y)$ is the gradient over x of $p_y(x, y)$.

- **Interpretation:** When p is smooth, each modal manifold is also smooth.

Hausdorff Distance

- To characterize smoothness of $M(x)$, we require a notion of distance over sets: **Hausdorff Distance**.

Definition

Let us consider a metric space (M, d) and suppose X and Y be two non-empty subsets of the metric space. Then the Hausdroff distance between X and Y is defined by,

$$d_H(X, Y) = \max\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y)\}$$

where $d(a, B)$ is the distance from a point a to the set B, $d(a, B) = \inf_{b \in B} d(a, b)$.

- Equivalently, we can define the Hausdorff distance as:

$$\texttt{Haus}(A, B) = \inf\{r : A \subseteq B \oplus r, B \subseteq A \oplus r\},$$

where $A \oplus r = \{x : d(x, A) \leq r\}$ with $d(x, A) = \inf_{y \in A} ||x - y||$.

Hausdorff Distance

- To characterize smoothness of $M(x)$, we require a notion of distance over sets: **Hausdorff Distance**.

Definition

Let us consider a metric space (M, d) and suppose X and Y be two non-empty subsets of the metric space. Then the Hausdroff distance between X and Y is defined by,

$$d_H(X, Y) = \max\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y)\}$$

where $d(a, B)$ is the distance from a point a to the set B, $d(a, B) = \inf_{b \in B} d(a, b)$.

- Equivalently, we can define the Hausdorff distance as:

$$\mathtt{Haus}(A, B) = \inf\{r : A \subseteq B \oplus r, B \subseteq A \oplus r\},$$

where $A \oplus r = \{x : d(x, A) \leq r\}$ with $d(x, A) = \inf_{y \in A} ||x - y||$.

Hausdorff Distance

- To characterize smoothness of $M(x)$, we require a notion of distance over sets: **Hausdorff Distance**.

Definition

Let us consider a metric space (M, d) and suppose X and Y be two non-empty subsets of the metric space. Then the Hausdroff distance between X and Y is defined by,

$$d_H(X, Y) = \max\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y)\}$$

where $d(a, B)$ is the distance from a point a to the set B, $d(a, B) = \inf_{b \in B} d(a, b)$.

- Equivalently, we can define the Hausdorff distance as:

$$\mathtt{Haus}(A, B) = \inf\{r : A \subseteq B \oplus r, B \subseteq A \oplus r\},$$

where $A \oplus r = \{x : d(x, A) \leq r\}$ with $d(x, A) = \inf_{y \in A} ||x - y||$.

Hausdorff Distance

- To characterize smoothness of $M(x)$, we require a notion of distance over sets: **Hausdorff Distance**.

Definition

Let us consider a metric space (M, d) and suppose X and Y be two non-empty subsets of the metric space. Then the Hausdroff distance between X and Y is defined by,

$$d_H(X, Y) = \max\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y)\}$$

where $d(a, B)$ is the distance from a point a to the set B, $d(a, B) = \inf_{b \in B} d(a, b)$.

- Equivalently, we can define the Hausdorff distance as:

$$\texttt{Haus}(A, B) = \inf\{r : A \subseteq B \oplus r, B \subseteq A \oplus r\},$$

where $A \oplus r = \{x : d(x, A) \leq r\}$ with $d(x, A) = \inf_{y \in A} ||x - y||$.

Derivative of Modal Manifold Collection

Theorem (Smoothness of Modal Manifold Collection)

Assume the conditions of Lemma 3. Assume furthermore all partial derivatives of p are bounded by C, and there exists $\lambda_2 > 0$ such that $p_{yy}(x,y) < -\lambda_2$ for all $y \in M(x)$ and $x \in D$. Then

$$\lim_{|\varepsilon| \longrightarrow 0} \frac{\mathtt{Haus}(M(x), M(x+\varepsilon))}{|\varepsilon|} \leq \max_{j=1,\ldots,K} ||m_j'(x)|| \leq \frac{C}{\lambda_2} < \infty. \quad (8)$$

- **Interpretation:** Can be thought of as a statement about Lipschitz continuity with respect to Hausdorff distance.

Derivative of Modal Manifold Collection

Theorem (Smoothness of Modal Manifold Collection)

Assume the conditions of Lemma 3. Assume furthermore all partial derivatives of p are bounded by C, and there exists $\lambda_2 > 0$ such that $p_{yy}(x,y) < -\lambda_2$ for all $y \in M(x)$ and $x \in D$. Then

$$\lim_{|\varepsilon|\longrightarrow 0} \frac{\mathtt{Haus}(M(x), M(x+\varepsilon))}{|\varepsilon|} \leq \max_{j=1,\dots,K} ||m_j'(x)|| \leq \frac{C}{\lambda_2} < \infty. \qquad (8)$$

- **Interpretation:** Can be thought of as a statement about Lipschitz continuity with respect to Hausdorff distance.

Error Measurements

We consider the following losses to measure the error:

- **Pointwise Error:**

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

where Haus(A,B) Hausdroff distance between the sets A and B.

- **Uniform Error:**

$$\Delta_n = \sup_{x \in D} \Delta_n(x).$$

- **Mean Integrated Squared Error (MISE):**

$$MISE(\hat{M}_n) = \mathbb{E}\left(\int_{x \in D} \Delta_n^2(x)dx\right).$$

Error Measurements

We consider the following losses to measure the error:

- **Pointwise Error:**

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

  where $\mathtt{Haus}$(A,B) Hausdroff distance between the sets A and B.

- **Uniform Error:**

$$\Delta_n = \sup_{x \in D} \Delta_n(x).$$

- **Mean Integrated Squared Error (MISE):**

$$MISE(\hat{M}_n) = \mathbb{E}\left(\int_{x \in D} \Delta_n^2(x) dx\right).$$

Error Measurements

We consider the following losses to measure the error:

- **Pointwise Error:**

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

  where $\texttt{Haus}$(A,B) Hausdroff distance between the sets A and B.

- **Uniform Error:**

$$\Delta_n = \sup_{x \in D} \Delta_n(x).$$

- **Mean Integrated Squared Error (MISE):**

$$MISE(\hat{M}_n) = \mathbb{E}\left(\int_{x \in D} \Delta_n^2(x) dx\right).$$

Assumptions on Joint Density

Assumption (A1)

The joint density $p \in BC^4(C_p)$, for some $C_p > 0$.

Assumption (A2)

The collection of modal manifolds can $\mathbb{S}$ can be factorized into $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \ldots \cup \mathbb{S}_K$, where $\mathbb{S}_j$ is a connected curve that follows a parametrization $\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\}$ for some $m_j(x)$ and $A_1, A_2, \ldots, A_K$ form an open cover for the support D of X.

Assumption (A3)

There exists $\lambda_2 > 0$ such that for any $(x, y) \in D \times K$ with $p_y(x, y) = 0$, $|p_{yy}(x, y)| > \lambda_2$.

Assumptions on Joint Density

Assumption (A1)

The joint density $p \in BC^4(C_p)$*, for some* $C_p > 0$.

Assumption (A2)

The collection of modal manifolds can $\mathbb{S}$ *can be factorized into* $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \ldots \cup \mathbb{S}_K$*, where* $\mathbb{S}_j$ *is a connected curve that follows a parametrization* $\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\}$ *for some* $m_j(x)$ *and* $A_1, A_2, \ldots, A_K$ *form an open cover for the support* D *of* X.

Assumption (A3)

There exists $\lambda_2 > 0$ *such that for any* $(x,y) \in D \times K$ *with* $p_y(x,y) = 0$, $|p_{yy}(x,y)| > \lambda_2$.

Assumptions on Joint Density

Assumption (A1)

The joint density $p \in BC^4(C_p)$*, for some* $C_p > 0$.

Assumption (A2)

The collection of modal manifolds can $\mathbb{S}$ *can be factorized into* $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \ldots \cup \mathbb{S}_K$*, where* $\mathbb{S}_j$ *is a connected curve that follows a parametrization* $\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\}$ *for some* $m_j(x)$ *and* $A_1, A_2, \ldots, A_K$ *form an open cover for the support* D *of* X.

Assumption (A3)

There exists $\lambda_2 > 0$ *such that for any* $(x, y) \in D \times K$ *with* $p_y(x, y) = 0$, $|p_{yy}(x, y)| > \lambda_2$.

Assumptions on Kernel Function

Assumption (K1)

The Kernel function $K \in BC^2(C_K)$ *and satisfies for* $\alpha = 0, 1, 2$,

$$\int_R (K^{(\alpha)})^2(z)dz < \infty \qquad \int_R z^2(K^{(\alpha)})(z)dz < \infty$$

Assumption (K2)

The collection $\mathscr{K}$ *is a VC-type class, i.e. there exists* $A, v > 0$ *such that for* $0 < \varepsilon < 1$

$$\sup_Q N(\mathscr{K}, L_2(Q), C_K\varepsilon) \le \frac{A^v}{\varepsilon^v},$$

where $N(T, d, \varepsilon)$ *is the* $\varepsilon-$*covering number for the semimetric space* (T, d) *and Q is any probability measure.*

Assumptions on Kernel Function

Assumption (K1)

The Kernel function $K \in BC^2(C_K)$ *and satisfies for* $\alpha = 0, 1, 2$,

$$\int_R (K^{(\alpha)})^2(z)dz < \infty \qquad \int_R z^2(K^{(\alpha)})(z)dz < \infty$$

Assumption (K2)

The collection $\mathscr{K}$ *is a VC-type class, i.e. there exists* $A, v > 0$ *such that for* $0 < \varepsilon < 1$

$$\sup_Q N(\mathscr{K}, L_2(Q), C_{K^\varepsilon}) \leq \frac{A^v}{\varepsilon^v},$$

where $N(T, d, \varepsilon)$ *is the* $\varepsilon-$*covering number for the semimetric space* (T, d) *and* Q *is any probability measure.*

Few Notations

Before proceeding further let us define the following quantities:

$$\|\hat{p}_n - p\|_\infty^0 = \sup_{x,y} \|\hat{p}(x,y) - p(x,y)\|.$$
$$\|\hat{p}_n - p\|_\infty^1 = \sup_{x,y} \|\hat{p}_y(x,y) - p_y(x,y)\|.$$
$$\|\hat{p}_n - p\|_\infty^2 = \sup_{x,y} \|\hat{p}_{yy}(x,y) - p_{yy}(x,y)\|.$$
$$\|\hat{p}_n - p\|_{\infty,2}^* = \max\{\|\hat{p}_n - p\|_\infty^0, \|\hat{p}_n - p\|_\infty^1, \|\hat{p}_n - p\|_\infty^2\}.$$

Pointwise Rate

Theorem (Pointwise Error Rate)

Assuming (A1-3) and (K1-2) we define the stochastic process $A_n(x)$ as,

$$A_n(x) = \begin{cases} \frac{1}{\Delta_n(x)} \left|\Delta_n(x) - \max_{z \in M(x)}\{\, |p_{yy}^{-1}(x,z)|\, |\hat{p}_{y,n}(x,z)\,| \}\,\right| & \text{if } \Delta_n(x) > 0 \\ 0 & \text{if } \Delta_n(x) = 0 \end{cases}$$

Then for sufficiently small $\|\hat{p}_n - p\|^*_{\infty,2}$ we will have,

$$\sup_{x \in D}(A_n(x)) = O_p(\|\hat{p}_n - p\|^*_{\infty,2}).$$

- **Interpretation:** Under sufficient regularity conditions, $\Delta_n(x)$ can be approximated $\max_{z \in M(x)}\{\, |p_{yy}^{-1}(x,z)|\, |\hat{p}_{y,n}(x,z)\,| \}$.

Pointwise Rate

Theorem (Pointwise Error Rate)

Assuming (A1-3) and (K1-2) we define the stochastic process $A_n(x)$ as,

$$A_n(x) = \begin{cases} \frac{1}{\Delta_n(x)} \left| \Delta_n(x) - \max_{z \in M(x)} \{ |p_{yy}^{-1}(x,z)| \, |\hat{p}_{y,n}(x,z)| \} \right| & \text{if } \Delta_n(x) > 0 \\ 0 & \text{if } \Delta_n(x) = 0 \end{cases}$$

Then for sufficiently small $\|\hat{p}_n - p\|^*_{\infty,2}$ we will have,

$$\sup_{x \in D}(A_n(x)) = O_p(\|\hat{p}_n - p\|^*_{\infty,2}).$$

- **Interpretation:** Under sufficient regularity conditions, $\Delta_n(x)$ can be approximated $\max_{z \in M(x)} \{ |p_{yy}^{-1}(x,z)| \, |\hat{p}_{y,n}(x,z)| \}$.

Pointwise Rate

Theorem (Pointwise Error Rate contd.)

Moreover, at any fixed $x \in D$, when $\frac{nh^{d+5}}{\log n} \to \infty$ and $h \to 0$ we have,

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

- **Interpretation:** If the curvature of the joint density function along y is bounded away from 0, then the error can be approximated by the error of $\hat{p}_{y,n}(x, z)$.

Pointwise Rate

Theorem (Pointwise Error Rate contd.)

Moreover, at any fixed $x \in D$*, when* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$ *we have,*

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

- **Interpretation:** If the curvature of the joint density function along y is bounded away from 0, then the error can be approximated by the error of $\hat{p}_{y,n}(x,z)$.

Uniform Rate

Theorem (Uniform Error rate)

Assume (A1-3) and (K1-2), then as $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$ *we have,*

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

- Both the Pointwise and Uniform Error have the usual nonparametric rate, where $Rate = Bias + \sqrt{Variance}$.

Uniform Rate

Theorem (Uniform Error rate)

Assume (A1-3) and (K1-2), then as $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$ *we have,*

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

- Both the Pointwise and Uniform Error have the usual nonparametric rate, where $Rate = Bias + \sqrt{Variance}$.

MISE Rate

Theorem (MISE rate)

Assuming (A1-3) and (K1-2), as $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

$$MISE(\hat{M}_n) = O(h^4) + O\left(\frac{1}{nh^{d+3}}\right).$$

- Starting from Pointwise Error rate, Following the arguments from Chacón et al. (2011);Chacón and Duong (2013) it can be shown that the integrated bias and variance yields the same rate of convergence.

MISE Rate

Theorem (MISE rate)

Assuming (A1-3) and (K1-2), as $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

$$MISE(\hat{M}_n) = O(h^4) + O\left(\frac{1}{nh^{d+3}}\right).$$

- Starting from Pointwise Error rate, Following the arguments from Chacón et al. (2011);Chacón and Duong (2013) it can be shown that the integrated bias and variance yields the same rate of convergence.

Ideal Confidence Sets

In an ideal setting, following the estimation of $M_n(x)$, we could define confidence set at x by

$$\hat{C}_n^0(x) = \widehat{M}_n(x) \oplus \delta_{n,1-\alpha}(x)$$

where, $$\mathbb{P}(\Delta_n(x) > \delta_{n,1-\alpha}(x)) = \alpha.$$

We have, by construction, $\mathbb{P}(M(x) \in \hat{C}_n^0(x)) = 1 - \alpha$.

Since the distribution of $\Delta_n(x)$ is unknown, we estimate $\hat{\delta}_{n,1-\alpha}$ using bootstrap.

Modified setup with Bootstrap sample

Considering Bootstrap samples $(X_1^*, Y_1^*), \dots, (X_n^*, Y_n^*)$, we define error metric based on estimated regression mode $\widehat{M}_n^*(x)$:

$$\hat{\Delta}_n^*(x) = \texttt{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x)).$$

Repeating bootstrap sampling B times to get $\hat{\Delta}_{1,n}^*, \dots, \hat{\Delta}_{B,n}^*$, we get $\hat{\delta}_{n,1-\alpha}(x)$ as the solution to the equation:

$$B^{-1} \sum_{j=1}^{B} \mathbb{I}\left(\hat{\Delta}_{j,n}^*(x) > \hat{\delta}_{n,1-\alpha}\right) \approx \alpha.$$

Pointwise and Uniform confidence sets

The estimated pointwise confidence set is therefore given by

$$\hat{C}_n(x) = \widehat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha}(x), \; x \in D.$$

Further, defining $\delta_{m,1-\alpha}$ by

$$\mathbb{P}\left(M(x) \subseteq \widehat{M}_n^* \oplus \delta_{n,1-\alpha}, \; \forall x \in D\right) = 1-\alpha,$$

and estimating $\delta_{n,1-\alpha}$ based on quantiles of bootstrapped error metric

$$\hat{\Delta}_n^* = \sup_{x \in D} \mathrm{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x)).$$

Our uniform confidence set is then given by

$$\hat{C}_n = \left\{(x,y) : x \in D, y \in \widehat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha}\right\}. \qquad (9)$$

Pointwise and Uniform confidence sets

The estimated pointwise confidence set is therefore given by

$$\hat{C}_n(x) = \widehat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha}(x), \; x \in D.$$

Further, defining $\delta_{m,1-\alpha}$ by

$$\mathbb{P}\left(M(x) \subseteq \widehat{M}_n^* \oplus \delta_{n,1-\alpha}, \; \forall x \in D\right) = 1 - \alpha,$$

and estimating $\delta_{n,1-\alpha}$ based on quantiles of bootstrapped error metric

$$\hat{\Delta}_n^* = \sup_{x \in D} \mathtt{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x)).$$

Our uniform confidence set is then given by

$$\hat{C}_n = \left\{(x, y) : x \in D, y \in \widehat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha}\right\}. \tag{9}$$

Pointwise and Uniform confidence sets

The estimated pointwise confidence set is therefore given by

$$\hat{C}_n(x) = \widehat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha}(x), \; x \in D.$$

Further, defining $\delta_{m,1-\alpha}$ by

$$\mathbb{P}\left(M(x) \subseteq \widehat{M}_n^* \oplus \delta_{n,1-\alpha}, \; \forall x \in D \right) = 1 - \alpha,$$

and estimating $\delta_{n,1-\alpha}$ based on quantiles of bootstrapped error metric

$$\hat{\Delta}_n^* = \sup_{x \in D} \mathtt{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x)).$$

Our uniform confidence set is then given by

$$\hat{C}_n = \left\{ (x, y) : x \in D, y \in \widehat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha} \right\}. \quad (9)$$

Few Definitions

- We consider the estimation problem of regression modes of smoothed joint density $\tilde{p}(x,y) = \mathbb{E}(\hat{p}_n(x,y))$, since we obtain faster convergence rate.
- Similarly let $\tilde{M}(x) = \mathbb{E}(\widehat{M}_n(x))$ be smoothed regression modes at $x \in D$.
- Define $\tilde{\Delta}_n(x) = \mathtt{Haus}(\widehat{M}_n(x), \tilde{M}(x))$ and $\tilde{\Delta}_n = \sup_{x \in D} \tilde{\Delta}_n(x)$.
- We consider function space

$$\mathscr{F} = \left\{(u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = \tilde{p}_{yy}^{-1}(x,y) \times K\left(\frac{||x-u||}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), x \in \mathbb{D}, y \in \tilde{M}(x)\right\}.$$

- Let $\mathbb{B}$ be a Gaussian process defined on $\mathscr{F}$ such that $\forall f_1,\ f_2 \in \mathscr{F}$

$$\mathrm{Cov}(\mathbb{B}(f_1), \mathbb{B}(f_2)) = \mathbb{E}(f_1(X_i, Y_i) \cdot f_2(X_i, Y_i)) - \mathbb{E}(f_1(X_i, Y_i)) \cdot \mathbb{E}(f_2(X_i, Y_i)).$$

Limiting Distribution

Consider an empirical process $\mathbb{G}_n$defined on $\mathscr{F}$ as

$$\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^{n} f(D_i) - \mathbb{E}(f(D_i)),\ D_i = (X_i, Y_i).$$

Theorem (Asymptotic Theory)

Under regularity conditions,

- $\sqrt{nh^{d+3}}\tilde{\Delta}_n \approx \sup_{f\in\mathscr{F}}\{|G_n(f)|\} \approx \sup_{f\in\mathscr{F}}\{\mathbb{B}(f)\}$.
- *More precisely,*

$$\left|\sqrt{nh^{d+3}}\tilde{\Delta}_n - \mathbb{B}\right| = O_{\mathbb{P}}\left(\left(\frac{\log^4 n}{nh^{d+3}}\right)^{1/8}\right).$$

Since Gaussian Process involves unknown quantities, this in itself is not sufficient to conduct statistical inferences.

Bootstrap Consistency

We use bootstrap to approximate Δ_n. We define another metric $\hat{\Delta}_n^* = \sup_{x \in D} \texttt{Haus}(\widehat{M}_n^*, \widehat{M}_n(x))$.

Theorem

Under regularity conditions,

- $\sqrt{nh^{d+3}}\hat{\Delta}_n^* \approx \sup_{f \in \mathscr{F}} |\mathbb{B}(f)|$ *for function space* $\mathscr{F}$*,*
- $\sqrt{nh^{d+3}}\hat{\Delta}_n^* \approx \sqrt{nh^{d+3}}\tilde{\Delta}_n$.

- **Interpretation** This theorem brings forth an equivalence in limiting distribution of $\hat{\Delta}_n^*$ and $\tilde{\Delta}_n$. Infact, The rate of convergence in distribution is $O\left(\left(\frac{\log^4 n}{nh^{d+3}}\right)^{1/8}\right)$.

Bootstrap Consistency

We use bootstrap to approximate Δ_n. We define another metric $\hat{\Delta}_n^* = \sup_{x \in D} \mathtt{Haus}(\widehat{M}_n^*, \widehat{M}_n(x))$.

Theorem

Under regularity conditions,

- $\sqrt{nh^{d+3}}\hat{\Delta}_n^* \approx \sup_{f \in \mathscr{F}} |\mathbb{B}(f)|$ *for function space* $\mathscr{F}$*,*
- $\sqrt{nh^{d+3}}\hat{\Delta}_n^* \approx \sqrt{nh^{d+3}}\tilde{\Delta}_n$.

- **Interpretation** This theorem brings forth an equivalence in limiting distribution of $\hat{\Delta}_n^*$ and $\tilde{\Delta}_n$. Infact, The rate of convergence in distribution is $O\left(\left(\frac{\log^4 n}{nh^{d+3}}\right)^{1/8}\right)$.

Uniform Confidence Sets

Corollary (Uniform confidence sets)

Assume (A1-3) and (K1-2). Then as $\frac{nh^6}{\log n} \to \infty$ *and* $h \to 0$,

$$\mathbb{P}\left(\tilde{M}(x) \subseteq \hat{M}_n(x) \oplus \hat{\delta}_{n,1-\alpha},\ \forall x \in D\right) = 1 - \alpha + O\left(\left(\frac{\log^4 n}{nh^{d+3}}\right)^{1/8}\right).$$

Therefore. the asymptotic valid confidence for M is given as

$$\left\{(x, y) : y \in \widehat{M}_n(x) \oplus \hat{\delta}_{1-\alpha}, x \in D\right\},$$

$\hat{\delta}_{n,1-\alpha}$ is the upper $1 - \alpha$ quantile of $\hat{\Delta}_n$.

Prediction Sets

- We define:

$$\varepsilon_{1-\alpha}(x) = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(Y, M(x)) > \varepsilon \mid X = x) \leq \alpha\}.$$

$$\varepsilon_{1-\alpha} = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(Y, M(X)) > \varepsilon) \leq \alpha\}.$$

Definition (Pointwise Prediction Set)

$\mathscr{P}_{1-\alpha}(x) = M(x) \oplus \varepsilon_{1-\alpha}(x) \subseteq \mathbb{R}.$

Definition (Uniform Prediction Set)

$\mathscr{P}_{1-\alpha} = \{(x, y) : x \in D,\ y \in M(x) \oplus \varepsilon_{1-\alpha}\} \subseteq D \times \mathbb{R}.$

Prediction Sets

- We define:

$$\varepsilon_{1-\alpha}(x) = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(Y, M(x)) > \varepsilon \mid X = x) \leq \alpha\}.$$

$$\varepsilon_{1-\alpha} = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(Y, M(X)) > \varepsilon) \leq \alpha\}.$$

Definition (Pointwise Prediction Set)

$\mathscr{P}_{1-\alpha}(x) = M(x) \oplus \varepsilon_{1-\alpha}(x) \subseteq \mathbb{R}.$

Definition (Uniform Prediction Set)

$\mathscr{P}_{1-\alpha} = \{(x, y) : x \in D,\ y \in M(x) \oplus \varepsilon_{1-\alpha}\} \subseteq D \times \mathbb{R}.$

Prediction Sets

- We define:

$$\varepsilon_{1-\alpha}(x) = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(Y, M(x)) > \varepsilon \mid X = x) \leq \alpha\}.$$

$$\varepsilon_{1-\alpha} = \inf\{\varepsilon \geq 0 : \mathbb{P}(d(Y, M(X)) > \varepsilon) \leq \alpha\}.$$

Definition (Pointwise Prediction Set)

$\mathscr{P}_{1-\alpha}(x) = M(x) \oplus \varepsilon_{1-\alpha}(x) \subseteq \mathbb{R}.$

Definition (Uniform Prediction Set)

$\mathscr{P}_{1-\alpha} = \{(x, y) : x \in D,\ y \in M(x) \oplus \varepsilon_{1-\alpha}\} \subseteq D \times \mathbb{R}.$

Bandwidth Selection

- We can choose the bandwidth of the KDE by minimizing the size of the prediction set.
- Choose

$$h^* = \arg\min_{h \geq 0} \texttt{Vol}(\hat{\mathscr{P}}_{1-\alpha,h}),$$

where $\hat{\mathscr{P}}_{1-\alpha,h}$ is the estimated uniform prediction set.

Bandwidth Selection

- We can choose the bandwidth of the KDE by minimizing the size of the prediction set.
- Choose

$$h^* = \arg\min_{h \geq 0} \mathtt{Vol}(\hat{\mathscr{P}}_{1-\alpha,h}),$$

where $\hat{\mathscr{P}}_{1-\alpha,h}$ is the estimated uniform prediction set.

Bandwidth Selection: Example

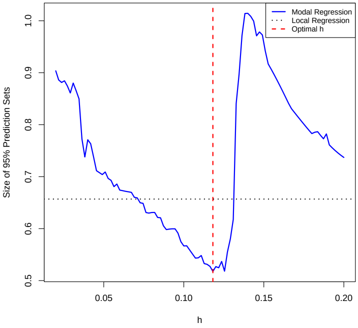

Figure: Bandwidth selection based on size of prediction sets.

Final Remarks

- We reviewed a nonparametric method for modal regression estimation, based on a KDE of a joint sample of data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We studied some of the geometry underlying the modal regression set, and described techniques for confidence set estimation, prediction set estimation, and bandwidth selection for the underlying KDE.
- The main message is that nonparametric modal regression offers a relatively simple and usable tool to capture conditional structure missed by conventional regression methods.
- For more information: Report R Codes

Thank You!

Final Remarks

- We reviewed a nonparametric method for modal regression estimation, based on a KDE of a joint sample of data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We studied some of the geometry underlying the modal regression set, and described techniques for confidence set estimation, prediction set estimation, and bandwidth selection for the underlying KDE.
- The main message is that nonparametric modal regression offers a relatively simple and usable tool to capture conditional structure missed by conventional regression methods.
- For more information: Report R Codes

Thank You!

Final Remarks

- We reviewed a nonparametric method for modal regression estimation, based on a KDE of a joint sample of data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We studied some of the geometry underlying the modal regression set, and described techniques for confidence set estimation, prediction set estimation, and bandwidth selection for the underlying KDE.
- The main message is that nonparametric modal regression offers a relatively simple and usable tool to capture conditional structure missed by conventional regression methods.
- For more information: Report R Codes

Thank You!

Final Remarks

- We reviewed a nonparametric method for modal regression estimation, based on a KDE of a joint sample of data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We studied some of the geometry underlying the modal regression set, and described techniques for confidence set estimation, prediction set estimation, and bandwidth selection for the underlying KDE.
- The main message is that nonparametric modal regression offers a relatively simple and usable tool to capture conditional structure missed by conventional regression methods.
- For more information: Report R Codes

Thank You!

Final Remarks

- We reviewed a nonparametric method for modal regression estimation, based on a KDE of a joint sample of data points $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We studied some of the geometry underlying the modal regression set, and described techniques for confidence set estimation, prediction set estimation, and bandwidth selection for the underlying KDE.
- The main message is that nonparametric modal regression offers a relatively simple and usable tool to capture conditional structure missed by conventional regression methods.
- For more information: Report R Codes

Thank You!

Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, 7:499–532.

Chacón, J. E., Duong, T., and Wand, M. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, pages 807–840.

Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric modal regression. *The Annals of Statistics*, 44(2):489–514.

Einbeck, J. and Tutz, G. (2006). Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(4):461–475.

Sager, T. W. and Thisted, R. A. (1982). Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics*, pages 690–707.