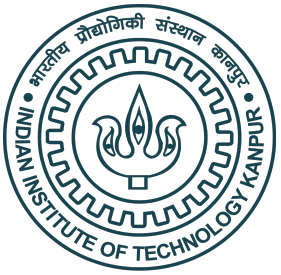
# Understanding Nonparametric Multimodal Regression via Kernel Density Estimation*

*Submitted by:*

Arkajyoti Bhattacharjee [†‡]
Rachita Mondal [§†]
Ritwik Vashishtha [¶†]
Shubha Sankar Banerjee [‖†]

*Supervised by:*

Dr. Subhra Sankar Dhar [†]




*Submitted on:*

$18^{th}$ February, 2022

**Abstract**

In this report we review non-parametric Modal Regression using Kernel Density Estimator. Instead of using conditional mean, Modal Regression uses conditional mode to summarize the relationship between the response and the explanatory variables. We describe the idea of Modal Regression and include a brief discussion regarding the superiority of Multi-modal regression over the Uni-modal case. The consistency properties of the proposed estimator and the idea of Confidence Sets have been reviewed. This report also includes an application of Prediction Sets in case of Bandwidth selection. Certain generalizations and extensions are also discussed. The report is primarily based on Chen et al. (2016).

# Contents



*This report has been prepared towards the partial fulfillment of the requirements of the course *MTH673A: Robust Statistical Methods.*

†Department of Mathematics & Statistics, Indian Institute of Kanpur, India.
‡201277, M.Sc. Statistics (Final year).
§201374, M.Sc. Statistics (Final year).
¶201389, M.Sc. Statistics (Final year).
‖201416, M.Sc. Statistics (Final year).

# 1 Introduction

A usual approach for studying the relationship between a response variable, usually denoted by $Y$, and its predictors, usually denoted by $X$, is through the conditional mean of Y given X, i.e. $\mathbb{E}(Y|X)$. An alternative approach to this problem is to replace the conditional mean with the conditional modes or local modes. Modal regression searches the conditional modes (Sager and Thisted (1982), Lee (1989)) or local modes (Einbeck and Tutz (2006), Chen et al. (2016)) of the response $Y$ given the predictor $X = x$.

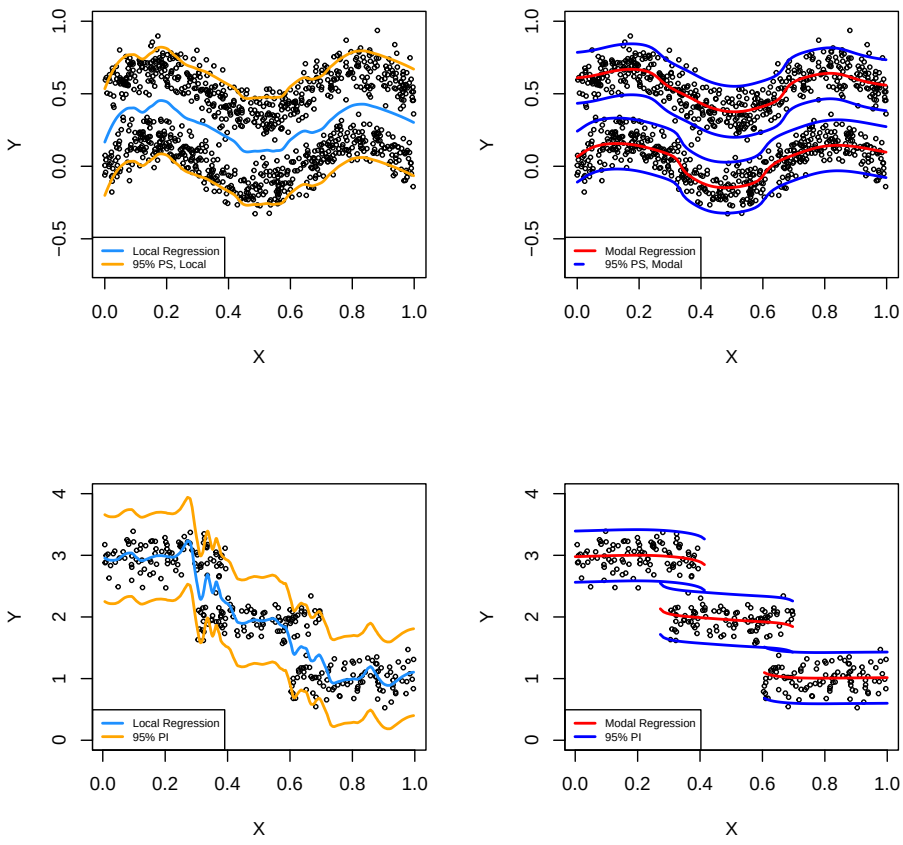



Figure 1: In the top row, we show local regression estimate and its associated 95% prediction bands alongside the modal regression and its 95% prediction bands. The bottom row does the same for a different simulated data.

There are two broad scenarios where modal regression can perform better than the conventional mean regression method in the sense that the former is better in capturing $X-Y$ relationship:
(1) When the conditional distribution of $Y$ given $X=x$ is skewed or heavy-tailed;
(2) When the conditional distribution of $Y$ given $X=x$ has multiple modes.
The conditional mean may fail to capture the inherent pattern in the data in such cases and modal regression not only provides an improvement over trend estimation but also provides narrower prediction bands as is evident in Figure 1.

Modal regression has been applied to many domains like transportation (Einbeck and Tutz (2006)), astronomy (Rojas et al. (2005)) and in various problems such as predicting Alzheimer's disease (Wang et al. (2017)), analyzing healthcare expenditure (Yao and Xiang (2016), Xiang and Yao (2022)), predicting temperature (Hyndman et al. (1996)), analyzing electricity consumption (Chaouch et al. (2017)), and studying the pattern of forest fire (Yao and Li (2014)).

The organization of the remainder of this report is as follows. In Section 2, we formally introduce modal regression and discuss its two types: uni-modal and multi-modal regression. In Section 3, we discuss about modal regression estimation based on the mean-shift algorithm (Section 3.1). In Section 4, we study the geometric properties of modal regression. In Section 5, we review the asymptotic properties of modal regression estimators. In Section 6 and Section 7, we deal with the construction of confidence sets and prediction sets. In Section 7.1, we discuss bandwidth selection for the KDE based on minimizing the prediction sets. We end the main part of the report in Section 8 with some extensions, generalizations and related work on modal regression.

## 2 Modal Regression

Consider a response variable $Y \in K \subset \mathbb{R}$ and a predictor variable $X \in D \subset \mathbb{R}^d$, where $D$ is a compact set. There is literature on broadly two types of modal regression. One, focusing on conditional (global) modes, is called *uni-modal regression* (Collomb et al. (1986), Lee (1989), Sager and Thisted (1982)). The other searches for conditional local modes and is known as *multi-modal regression* (Chen et al. (2016), Einbeck and Tutz (2006)).

More formally, let $f(z)$ denote the probability density function (PDF) of a random

variable $Z$. We define the operators:

$$\texttt{UniMode} = \arg\max_z f(z)$$

and

$$\texttt{MultiMode} = \{z : f'(z) = 0, f''(z) < 0\},$$

which return the global and local modes of the PDF of $Z$, respectively. Clearly, $f$ needs to be twice-differentiable. So, uni-modal regression searches for the function

$$m(x) = \texttt{UniMode}(Y|X = x) = \arg\max_y p(y|x), \tag{1}$$

while on the other hand, multi-modal regression searches for the function

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

Note that $M(x)$ may be a multi-valued function, i.e. it may take up multiple values at any given point $x$.

Equations (1) and (2) may be further simplified noting that $p(y|x) = \frac{p(x,y)}{p(x)}$ and hence, given $x$, the global mode or local modes of $p(y|x)$ and $p(x,y)$ are equal. So, provided $p(x) > 0$, we can rewrite equations (1) and (2) in the following form:

$$m(x) = \arg\max_y p(x,y), \quad M(x) = \{y : \frac{\partial}{\partial y} p(x,y) = 0,\ \frac{\partial^2}{\partial y^2} p(x,y) < 0\}. \tag{3}$$

In other words, both the types of modal regressions may be directly defined through the joint distribution.

## 2.1 Uni-modal vs. Multi-modal Regression

There are pros and cons to both uni-modal and multi-modal regression. Uni-modal regression is an alternative to conventional regression methods for summarizing the predictor-response relationship using a single function. Multi-modal regression performs a similar task, but allows a multi-valued function. Although uni-modal regression output is easier to interpret, multi-modal regression can identify hidden structure in the predictor-response relationship in situations where the relationship is complicated or involves several distinct components. Further, the prediction intervals tend to be wider in uni-modal regression than in mutli-modal regression (see Figure 2 for example). Throughout the remainder of this report, we focus on multi-modal regression - its theory and applications.

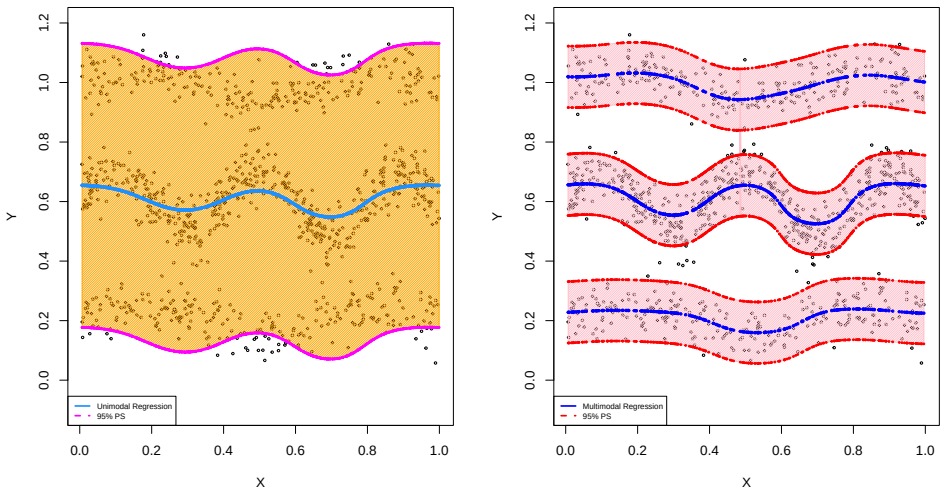



Figure 2: Uni-modal regression and multi-modal regression along with their corresponding 95% prediction sets on a simulated data with three components.

## 3 Estimation

We focus on the nonparametric estimation of the conditional mode set $M(x)$ in (3) using a plug-in estimate from the kernel density estimation (KDE) (Scott (2015), Einbeck and Tutz (2006)):

$$\hat{M}_n(x) = \{y : \frac{\partial}{\partial y}\hat{p}_n(x,y) = 0,\ \frac{\partial^2}{\partial y^2}\hat{p}_n(x,y) < 0\}, \tag{4}$$

where $\hat{p}_n(x,y)$ is the joint KDE of $X,Y$. Let $(X_1,Y_1),...,(X_n,Y_n)$, be the observed data. Then the KDE of the joint density of $p(x,y)$ is

$$\hat{p}_n(x,y) = \frac{1}{nh^{d+1}}\sum_{i=1}^{n} K\left(\frac{||x-X_i||}{h}\right) K\left(\frac{y-Y_i}{h}\right). \tag{5}$$

Here, the kernel $K$ is a smooth, symmetric function (like the Gaussian kernel[1]) and $h > 0$ is called the *bandwidth* or *smoothing parameter*.

[1] $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$

Throughout the report, we will assume the same bandwidth $h$ and kernel $K$ for both the response as well the predictor variables. For conciseness of notations, we will write the estimated modal set as

$$\hat{M}_n(x) = \{y : \hat{p}_{y,n}(x,y) = 0,\ \hat{p}_{yy,n}(x,y) < 0\}, \tag{6}$$

where $p_y(x,y) = \frac{\partial}{\partial y} p(x,y)$, $p_{yy}(x,y) = \frac{\partial^2}{\partial y^2} p(x,y)$.

## 3.1 Mean-shift Algorithm

In general, estimating (6) is not trivial. However, for special kernels, Einbeck and Tutz (2006) proposed a simple and efficient algorithm for computing local mode estimates, based on the *mean-shift algorithm* (Cheng (1995), Comaniciu and Meer (2002)).

We consider the following example to understand the mean-shift algorithm (Chen (2018)). Consider an observation of *i.i.d.* 1-d random samples $X_1, \ldots, X_n \sim p$. We take $\hat{p}$ as a KDE with Gaussian kernel $K_G(x) = \sqrt{2\pi} e^{-x^2/2}$. Gaussian kernel yields a special feature which we make use of, that is, it has a nicely behaving derivative:

$$K_G'(x) = -x \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = -x \cdot\ K_G(x)$$

The derivative of the KDE becomes:

$$\begin{aligned}
\hat{p}(x) &= \frac{d}{dx} \frac{1}{nh} \sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right) \\
&= \frac{1}{nh^3} \sum_{i=1}^{n} (X_i - x) \cdot K_G\left(\frac{X_i - x}{h}\right) \\
&= \frac{1}{nh^3} \sum_{i=1}^{n} X_i \cdot K_G\left(\frac{X_i - x}{h}\right) - x \cdot \sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)
\end{aligned}$$

Multiplying both sides of the above equation by $nh^3$ and dividing by $K_G\left(\frac{X_i - x}{h}\right)$ we get,

$$\frac{nh^3}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)} \cdot \hat{p}(x) = \frac{\sum_{i=1}^{n} X_i K_G\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)} - x$$

If we rearrange the expression, we get a known structure:

$$\underset{\text{current position}}{x} + \underbrace{\frac{nh^3}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)} \cdot \hat{p}(x)}_{\text{gradient ascent}} = \underbrace{\frac{\sum_{i=1}^{n} X_i K_G\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)}}_{\text{next location}} \tag{7}$$

Thus it takes the form of a gradient ascent algorithm. Given a point $x$, the value of $\frac{\sum_{i=1}^{n} X_i K_G\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)}$ is a shifted location by the application of gradient ascent with an amount of $\frac{nh^3}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x}{h}\right)} \cdot \hat{p}(x)$. Thus the mean-shift algorithm performs an updation of initial point $x^{(t)}$ as:

$$x^{(t)} = \frac{\sum_{i=1}^{n} X_i K_G\left(\frac{X_i - x^{(t)}}{h}\right)}{\sum_{i=1}^{n} K_G\left(\frac{X_i - x^{(t)}}{h}\right)}$$

for $t = 0, 1, 2, \ldots$. The above derivation suggests that this update moves points by gradient-ascent (7). Thus the stationary point $x^{(\infty)}$ will be one of the local modes of the KDE. Although some initial points do not converge to local modes, these points form a set with Lebesgue measure 0, so we can ignore them (Chen et al. (2017)).

Making suitable changes for our setup, for simplicity, we consider the Gaussian kernel here. We present the 'partial' mean-shift algorithm of Einbeck and Tutz (2006) in Algorithm (1).

**Algorithm 1** Partial mean-shift algorithm

**Input:** Data samples $\mathscr{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, 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:

$$y \longleftarrow \frac{\sum_{i=1}^{n} Y_i K\left(\frac{||x - X_i||}{h}\right) K\left(\frac{y - Y_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{||x - X_i||}{h}\right) K\left(\frac{y - Y_i}{h}\right)} \tag{8}$$

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

As shown above, it can be shown that the mean-shift update in (8) is a gradient ascent update on the function $f(y) = p_n(x, y)$ (for fixed x), with an implicit choice of step size. Because this function $f$ is generically non-concave, we are not guaranteed that gradient ascent will actually attain a (global) maximum, but it will converge to critical points under small enough step sizes (Arias-Castro et al. (2016)).

# 4 Geometric Properties

From (2), we recall that $M(x)$ is a collection of points for each input $x$. The local modes behave like a collection of surfaces (see Figure 3) called *modal manifolds*. Going by the steps in Chen et al. (2016), we define a *modal manifold collection* as the union of these sets over all inputs $x$,

$$\mathbb{S} = \{(x,y) : x \in D, y \in M(x)\} \tag{9}$$

The dimension of the set $\mathbb{S}$ is $d$ as obtained from the implicit function theorem.

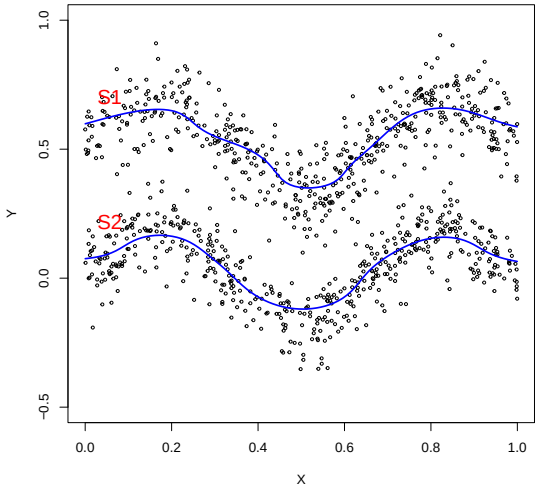



Figure 3: S1 and S2 represent modal manifolds.

We assume that the modal manifold collection $\mathbb{S}$ can be factorized as:

$$\mathbb{S} = \mathbb{S}_1 \cup \cdots \cup \mathbb{S}_K, \tag{10}$$

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

$$\mathbb{S}_j = \{(x, m_j(x)) : x \in A_j\} \tag{11}$$

for some function $m_j(x)$ and open set $A_j$. We also note that $A_1, A_2, \ldots, A_K$ form an open cover for the support $D$ of $X$. $\mathbb{S}_j$ and $m_j(x)$ are called the $j^{th}$ *modal manifold*

and the $j^{th}$ *modal function*, respectively. As a convention, $m_j(x) = \phi$ if $x \notin A_j$. This effectively allows us to write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}, \tag{12}$$

which means that for any given $x$, the values among $m_1(x), \ldots, m_K(x)$ that are not void give the local modes at that $x$.

The following lemma provides conditions under which each $m_j(x)$is differentiable and hence, in a sense, so is $M(x)$.

**Lemma 1** (Derivative of modal functions, Lemma 1, Chen et al. (2016))**.** *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 (10), (11). Then, when $x \in A_j$ ,*

$$\nabla m_j(x) = -\frac{p_{yx}(x, m_j(x))}{p_{yy}(x, m_j(x))} \tag{13}$$

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

*Proof.* We assume that $x \in A_j$. Then, by definition, we have $p_y(x, m_j(x)) = 0$. Now, taking the gradient over $x$ gives us

$$0 = \nabla_x p_y(x, m_j(x)) = p_{yx}(x, , m_j(x)) + p_{yy}(x, m_j(x)) \nabla m_j(x).$$

After rearranging the terms, we get the desired result. □

Observe that (13) is defined as long as $p_{yy}(x, m_j(x)) \neq 0$, which is trivially satisfied by the definition of local modes (3). The interpretation of the above lemma is as follows: when $p$ is smooth, the modal manifolds are smooth as well.

We have established the smoothness of each of the modal manifolds. To characterize the smoothness of $M(x)$ itself, we need to define the concept of smoothness over sets. For this purpose, we define the *Hausdorff Distance*.

**Definition 1** (Hausdroff Distance)**.** *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 define 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 $B$, $d(a,B) = \inf_{b \in B} d(a,b)$.*

An alternative and compact form of Hausdorff distance between two sets $A$ and $B$ is given by:

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

where $A \oplus r = \{x : d(x,A) \leq r\}$ with $d(x,A) = \inf_{y \in A} ||x-y||$. The distance can be viewed as a generalization of $\ell_\infty$ distance for sets. The following theorem establishes the smoothness of the modal manifold collection.

**Theorem 1** (Theorem 2, Chen et al. (2016))**.** *Assume the conditions of Lemma 1. 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*

$$\lim_{|\varepsilon| \longrightarrow 0} \frac{\texttt{Haus}(M(x), M(x+\varepsilon))}{|\varepsilon|} \leq \max_{j=1,\ldots,K} ||m_j'(x)|| \leq \frac{C}{\lambda_2} < \infty. \tag{14}$$

The proof of the above theorem follows directly from Lemma 1 and the definition of Hausdorff distance. Theorem 1 can be interpreted as a statement about Lipschitz continuity with respect to Hausdorff distance.

We now define the sample versions of the above defined population quantities. For the estimate $\hat{M}_n(x)$, we define

$$\hat{\mathbb{S}}_n = \{(x,y) : y \in \hat{M}_n(x), x \in \mathbb{R}\} = \hat{\mathbb{S}}_1 \cup \cdots \cup \hat{\mathbb{S}}_{\hat{K}}, \tag{15}$$

where each $\hat{\mathbb{S}}_j$ is a connected manifold, and $\hat{K}$ is the total number. In a similar fashion, we define $\hat{m}_j(x)$ for $j = 1,\ldots,\hat{K}$ and can, thus, write

$$\hat{M}_n(x) = \{\hat{m}_1(x), \ldots, \hat{m}_{\hat{K}}(x)\}. \tag{16}$$

In practice, it is not trivial finding the total number of manifolds $\hat{K}$ and determining the manifold memberships. In principle, although the sample manifolds $\hat{\mathbb{S}}_1, \ldots, \hat{\mathbb{S}}_{\hat{K}}$ are well defined in terms of the sample estimate $\hat{M}_n(x)$, even with a perfectly convergent mean-shift algorithm, mean-shift iterations at every input $x$ in the domain $D$ needs to be run to determine these manifold components, which is clearly not an implementable strategy. Thus, from the output of the mean-shift algorithm over a finite mesh, some type of simple post-processing technique is usually employed to determine connectivity of the outputs and hence the sample manifolds. For further discussion, the interested reader is directed to Section 7 of Chen et al. (2016).

# 5 Consistency

In this section we will focus on the convergence of the estimated modal regression set $\hat{M}_{n(x)}$ to the modal set $M_n(x)$. Now let us define the followings:

$BC^k(C)$: Collection of $k$ times continuously differentiable functions with all the partial derivatives absolutely bounded by C.

Given K, a kernel $\mathscr{K}$ is the collection of functions defined by,

$$\mathscr{K} = \left\{ v \mapsto K^{(\alpha)}\left(\frac{z-v}{h}\right) : z \in R, h > 0, \alpha = 0,1,2 \right\}$$

where $K^{(\alpha)}$ is the $\alpha^{th}$ order derivative of K.
We assume the following:

**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 $\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup ... \cup \mathbb{S}_K$, 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 $A_1, A_2, ..., A_K$ 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 $p_y(x,y) = 0$, $|p_{yy}(x,y)| > \lambda_2$.*

**Assumption K1.** *The Kernel function $K \in BC^2(C_K)$ and satisfies,*

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

*for $\alpha = 0,1,2$*

**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.*

Assumption (A1) is a smoothness condition. Fourth order derivative is required to ge the bounds of the bias of second derivatives. Assumption (A2) states that it is possible to represent the collection of all local modes as a finite collection of manifolds. (A3) ensures the sharpness of all the critical points and it also excludes

the cases that the modal manifolds bifurcate or merge. Assumption (K1) makes sure that the Kernel Density Estimator has the usual rates for its bias and variance. (K2) is assumed for the uniform bound of the kernel Density estimator.
At first we will discuss the point wise convergence of $\hat{M}_n(x)$. For this purpose the concept of Hausdroff distance will be used.

Let us denote the Hausdroff distance between $\hat{M}_n(x)$ and $M_n(x)$ as,

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

Also, we define the following quantities:

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

**Theorem 2** (Point wise Error Rate)**.** *Assuming (A1-3) and (K1-2) we define the stochastic process $A_n(x)$ as,*

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

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

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

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

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

*Proof.* Let $x \in D$ be a fixed point. Let $y_j \in M(x)$ be a local mode and $\hat{y}_j$ be the estimator of $y_j$.

Also, by assumption (A3) it can be shown that for sufficiently small $\|\hat{p}_n - p\|_{\infty,2}^*$, for every $x \in D$ it is possible to have a unique closest estimator $\hat{y}_j$ corresponding to $y_j$. Then we have,

$$\begin{aligned}
p_y(x, y_j) &= 0 \\
\hat{p}_{y,n}(x, \hat{y}_j) &= 0
\end{aligned}$$

Using the above facts and Taylor's theorem we have,

$$\begin{aligned}\hat{p}_{y,n}(x,y_j) &= \hat{p}_{y,n}(x,y_j) - \hat{p}_{y,n}(x,\hat{y}_j) \\ &= (y_j - \hat{y}_j)\hat{p}_{yy,n}(x,y_j^*)\end{aligned}$$

where $y_j^*$ is a point between $y_j$ and $\hat{y}_j$. Now dividing both sides by $\hat{p}_{yy,n}(x,y_j^*)$ and using the fact that,

$$|\hat{p}_{yy,n}(x,y_j^*)^{-1} - p_{yy}(x,y_j)^{-1}\ | = O_p(\|\hat{p}_n - p\|_{\infty,2}^*) \tag{17}$$

we get,

$$\begin{aligned}\hat{y}_j - y_j &= -\hat{p}_{yy,n}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j) \\ &= -p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j) + O_p(\|\hat{p}_n - p\|_{\infty,2}^* \cdot \hat{p}_{y,n}(x,y_j))\end{aligned} \tag{18}$$

Note that, (17) is valid as by (A3) $p_{yy}$ and $\hat{p}_{yy}$ both are bounded away from 0 when x and y are sufficiently close to $\mathbb{S}$ (the modal manifold collection). Hence, by (A1) and (K1) the inverses are also bounded above.
From (18) we can write,

$$|\hat{y}_j - y_j\ | - |\ p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j)\ | = O_p(\|\hat{p}_n - p\|_{\infty,2}^*.\ |\hat{p}_{y,n}(x,y_j)\ |) \tag{19}$$

Now, $\Delta_n(x) = \max_j\ |\hat{y}_j - y_j\ |$. So, taking *max* over all the local modes in (19) we get,

$$\Delta_n(x) - \max_j\ |p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j)\ | = O_p(\|\hat{p}_n - p\|_{\infty,2}^*.\max_j\ |\hat{p}_{y,n}(x,y_j)\ |)$$

Thus,

$$\max_j\{\ |\hat{p}_{y,n}(x,y_j)\ |\}^{-1}\ |\Delta_n(x) - \max_j\ |p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j)\ |\ | = O_p(\|\hat{p}_n - p\|_{\infty,2}^*) \tag{20}$$

So, $\Delta_n(x)$ can be approximated by $\max_j\ |p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j)\ |$.
Thus equation (20) implies that,

$$\frac{1}{\Delta_n(x)}\ |\Delta_n(x) - \max_j\ |p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j)\ |\ | = O_p(\|\hat{p}_n - p\|_{\infty,2}^*) \tag{21}$$

When $\Delta_n(x) > 0$, LHS of (21) = $A_n(x)$. Again, as the RHS of (21) does not depend upon x, taking *sup* to both the sides we get,

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

Now,

$$\begin{aligned} |\hat{p}_{y,n}(x,y_j)| &= |\hat{p}_{y,n}(x,y_j) - p_y(x,y_j)| \\ &\leq |\hat{p}_{y,n}(x,y_j) - E(\hat{p}_{y,n}(x,y_j))| + |E(\hat{p}_{y,n}(x,y_j)) - p_y(x,y_j)| \\ &= O_p\left(\sqrt{\frac{1}{nh^{d+3}}}\right) + O(h^2) \end{aligned} \tag{22}$$

Here (22) comes from the usual bias-variance trade-off of Kernel Density Estimator.
From (A1-3) it can be obtained that $|p_{yy}(x,y_j)^{-1}|$ is bounded from above and below. Hence, $\max_j |\hat{p}_{y,n}(x,y_j)|$ has the same rate as that of $\max_j |p_{yy}(x,y_j^*)^{-1}\hat{p}_{y,n}(x,y_j)|$ i.e. $\Delta_n(x)$. So, we have,

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

This shows that 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 $\hat{p}_{y,n}(x,z)$ after scaling. □

Now we state the theorem that shows the Uniform convergence of $\hat{M}_n$ and we define the uniform error rate as,

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

**Theorem 3** (Uniform Error rate, Theorem 4, Chen et al. (2016))**.** *Assume (A1-3) and (K1-2), then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$ *we have,*

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

Next we will consider the *Mean Integrated Squared Error* (MISE), which is a non-random quantity and is defined as,

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

**Theorem 4** (MISE rate,Theorem 5, Chen et al. (2016))**.** *Assuming (A1-3) and (K1-2), as* $\frac{nh^{d+5}}{logn} \to \infty$ *and* $h \to 0$*,*

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

*Proof.* From Theorem 2 we can show that,

$$\Delta_n(x) = O(h^2) + O_p\left(\sqrt{\frac{1}{nh^{d+3}}}\right) \tag{23}$$

Taking square and and expectation of equation (23) we get,

$$\begin{aligned} E(\Delta_n^2(x)) &= O(h^2) + O_p\left(\sqrt{\frac{1}{nh^{d+3}}}\right) \\ &= \texttt{Bias}^2(x) + \texttt{Variance}(x) \end{aligned}$$

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. □

Now, if we try to estimate the regression modes of smooth joint density $\tilde{p}(x,y) = E(\hat{p}_n(x,y))$, then it is possible to have a faster convergence rates. Let us denote the smoothed regression modes at $x \in D$ as $\tilde{M}(x) = E(\hat{M}_n(x))$. Let us also define the followings:

$$\begin{aligned} \tilde{\Delta}_n(x) &= \texttt{Haus}(\hat{M}_n(x), \tilde{M}(x)) \\ \tilde{\Delta}_n &= \sup_{x \in D} \tilde{\Delta}_n(x) \\ M\tilde{I}SE(\hat{M}_n) &= E\left(\int_{x \in D} \tilde{\Delta}_n^2(x) dx\right) \end{aligned}$$

**Corollary 4.1** (Error rates for smoothed conditional mode)**.** *Assume (A1-3) and (K1-2). Then as* $\frac{nh^{d+5}}{logn} \to \infty$ *and* $h \to 0$*,*

$$\sqrt{nh^{d+3}} \sup_{x \in D} |\tilde{\Delta}_n - \max_{z \in \tilde{M}(x)} \{\tilde{p}_{yy}^{-1}(x,z)\hat{p}_{y,n}(x,z)\}| = O_p(\varepsilon_{n,2})$$

$$\tilde{\Delta}_n(x) = O_p\left(\sqrt{\frac{1}{nh^{d+3}}}\right)$$

$$\tilde{\Delta}_n = O_p\left(\sqrt{\frac{logn}{nh^{d+3}}}\right)$$

$$M\tilde{I}SE(\hat{M}_n) = O\left(\sqrt{\frac{1}{nh^{d+3}}}\right)$$

*where,*

$$\varepsilon_{n,2} = \sup_{x,y} |\hat{p}_{yy,n}(x,y) - \tilde{p}_{yy}(x,y)| \quad = \sup_{x,y} |\hat{p}_{yy,n}(x,y) - E(\hat{p}_{yy,n}(x,y))|$$

# 6 Confidence Sets

Now that we have established the consistency of the of nonparametric modal regression estimator, we next focus on constructing *confidence sets* which can be used to obtain a band of values from around the points of the modal manifold which can contain the true population local modes up to a desired level of significance.

We first introduce the concept of *confidence sets* under a parametric setup. We shall thereafter extend the concept under our current nonparametric setup and discuss techniques to construct them.

**Definition 2** (Confidence Sets)**.** *Let the data be distributed according to a random variable $X$, which depends on a parameter $\theta$ taken from a parameter space $\Theta$. A $1-\alpha$ confidence set, denoted by $\mathscr{C}(X)$, is a subset of the parameter space $\Theta$ that only depends on $X$ such that:*

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta(\theta \in \mathscr{C}(X)) \geq 1-\alpha$$

We now extend the concept under our setup. An ideal setting would allow us to define a confidence set at $x$ by

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

where

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

By the above construction, we have, $\mathbb{P}(M(x) \in \hat{C}_n^0(x)) = 1-\alpha$. The distribution of $\Delta_n(x)$ is however unknown and hence we use bootstrap (Efron (1979)) in order to estimate $\delta_{n,1-\alpha}$.

Given observed samples $(X_1,Y_1),\ldots,(X_n,Y_n)$, the bootstrap sample is denoted as $(X_1^*,Y_1^*),\ldots,(X_1^*,Y_1^*)$. Let $\widehat{M}_n^*(x)$ be the estimated regression modes based on bootstrap sample. The *pointwise error* at a given point $x$ based on the bootstrap sample is given by

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

Upon repeating the bootstrap sampling $B$ times to get $\hat{\Delta}^*_{1,n}(x), \ldots, \hat{\Delta}^*_{B,n}(x)$. Now we define $\hat{\delta}_{n,1-\alpha}(x)$ by

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

The confidence set thus estimated for $M(x)$ is given by

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

This confidence set is defined for a particular value of $x$ from its domain and thus it is a *pointwise confidence set*, at $x \in \mathbb{D}$.

To build a uniform convergence set, we define $\Delta_n = \sup_{x\in\mathbb{D}} \Delta_n(x)$. Further, we define $\delta_{n,1-\alpha}$ by

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

We proceed in similar lines with bootstrap sampling to form an estimate of $\delta_{n,1-\alpha}$ as $\hat{\delta}_{n,1-\alpha}$, based on quantiles of the bootstrapped uniform error metric

$$\hat{\Delta}^*_n = \sup_{x\in\mathbb{D}} \texttt{Haus}\left(\widehat{M}^*_n(x), \widehat{M}_n(x)\right)$$

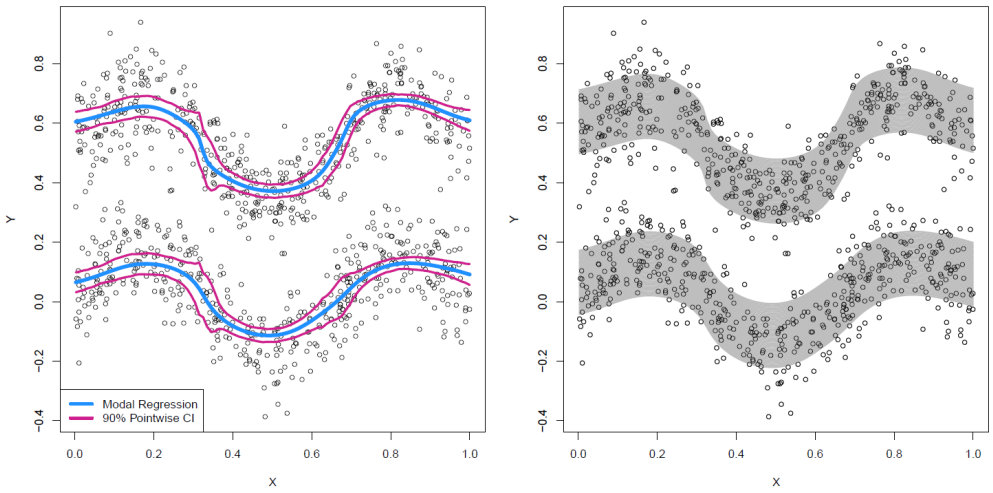



Figure 4: An example with 90% pointwise (left) and uniform (right) confidence sets. The plot has been taken from Chen et al. (2016).

See Figure 4 for an example with ordinary bootstrap.

The estimated *uniform confidence set* is

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

Our focus in this undertaking is on the theoretic asymptotic coverage of uniform confidence sets built with previously mentioned ordinary bootstrap. To avoid potential issues of the bias, we consider the coverage of smoothed regression mode set $\tilde{M}(x)$. To proceed with further calculations, we will make use of tools developed in Chernozhukov et al. (2014a); Chen et al. (2015).

Consider a function space $\mathscr{F}$ defined as

$$\mathscr{F} = \left\{(u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = \tilde{p}_{yy}^{-1}(x,y) \times K\left(\frac{||x-u||}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), x \in \mathbb{D}, y \in \tilde{M}(x)\right\} \tag{24}$$

Let $\mathbb{B}$ be a Gaussian process defined on $\mathscr{F}$ such that

$$\mathrm{Cov}(\mathbb{B}(f_1), \mathbb{B}(f_2)) = \mathbb{E}(f_1(X_i,Y_i) \cdot f_2(X_i,Y_i)) - \mathbb{E}(f_1(X_i,Y_i) \cdot \mathbb{E}(f_2(X_i,Y_i) \tag{25}$$

for all $f_1, f_2 \in \mathscr{F}$.

**Theorem 5** (Limiting Distribution, Theorem 7, Chen et al. (2016))**.** *Assume (A1-3) and (K1-2). Define a random variable* $\mathbf{B} = (h^{d+3})^{-1/2} \sup_{f \in \mathscr{F}} |\mathbb{B}(f)|$. *Then as* $\frac{nh^{d+5}}{\log n} \to \infty, h \to 0$,

$$\sup_{t \geq 0} \left| \mathbb{P}\left(\sqrt{nh^{d+3}}\tilde{\Delta}_n < t\right) - \mathbb{P}(\mathbf{B} < t) \right| = O\left(\left(\frac{\log^4 n}{nh^{d+3}}\right)^{1/8}\right)$$

*Proof.* We prove this theorem in similar lines as in Chen et al. (2015).

We consider $\mathscr{F}$ to be the functional space defined in (24). As earlier, we define $\mathbb{G}_n$ as an empirical process on $\mathscr{F}$ of the following form:

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n} (f(X_i,Y_i) - \mathbb{E}(f_k(X_i,Y_i)) \right)$$

We denote $\mathbf{G}_n = \frac{1}{\sqrt{h^{d+3}}} \sup_{f\in\mathscr{F}} |\mathbb{G}_n(f)|$ and $\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f\in\mathscr{F}} |\mathbb{B}(f)|$. Our proof will include three steps. In the first step we focus on establishing a coupling between the Hausdorff distance ($\sqrt{nh^{d+3}}\Delta_n$) and the supremum of the empirical process $\mathbf{G}_n$. The second step shows that the distribution of the maxima of the empirical process can be approximated by the maxima of a Gaussian process, i.e., a coupling between $\mathbf{G}_n$ and $\mathbf{B}$. The last step uses anti-concentration (Chernozhukov et al. (2014a)) to bound the distributions between $\sqrt{nh^{d+3}}\Delta_n$ and $\mathbf{B}$, i.e. to convert this coupling into the desired Berry-Essen bound.

Before proceeding with the proof, we define the following,

**Definition 3** (Reach for a set (Federer (1959)))**.** *The reach for a set A, denoted by* $\texttt{reach}(A)$*, is the largest real number r such that each* $x \in \{y : d(y,A) \leq r\}$ *has a unique projection onto A. The reach measures the smoothness of a set.*

**Step 1** Our goal is to show

$$\mathbb{P}\left(|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n| > \varepsilon\right) \leq D_1 e^{-D_2 nh^{d+5}\varepsilon^2} \tag{26}$$

for some constants $D_1,\ D_2$.

From Corollary 4.1

$$|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n| = O(\varepsilon_{n,2}) = O\left(\sup_{x,y}|\hat{p}_{yy,n}(x,y) - \mathbb{E}(\hat{p}_{yy,n}(x,y))|\right) \tag{27}$$

Thus from (27), there exists a constant $D_0 > 0$ such that

$$|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n| \leq D_0 \sup_{x,y}|\hat{p}_{yy,n}(x,y) - \mathbb{E}(\hat{p}_{yy,n}(x,y))|$$

By Talagrand's inequality (Thereom A.4 in Chernozhukov et al. (2014a); Talagrand (1996)),

$$\begin{aligned}
&\mathbb{P}\left(|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n| > \varepsilon\right) \\
&\qquad \leq \mathbb{P}\left(\sup_{x,y}|\hat{p}_{yy,n}(x,y) - \mathbb{E}(\hat{p}_{yy,n}(x,y))| > \varepsilon/D_0\right) \\
&\qquad \leq D_1 e^{D_2 nh^{d+5}\varepsilon^2}
\end{aligned} \tag{28}$$

For some constraints $D_1,\ D_2 > 0$. This gives the desired result.

Further, recalling asymptotic Hausdorff distance, $\texttt{dist}_H(A,B) = \sup_{x\in B} d(x,A)$, then,

$$|\sqrt{nh^{d+3}}\texttt{dist}_H(\widehat{M}_n, M_n) - \mathbf{G}| = O\left(||\hat{p}_n - p_n||^*_{\infty,4}\right) \tag{29}$$

This shows that the quasi-Hausdorff distance can be approximated an empirical process over the functional space $\mathscr{F}$.

When $||\hat{p}_n - p_n||^*_{\infty,5}$ is sufficiently small, the reach of $\widehat{M}_n$ is close to the reach of $M_n$, by claim 7 of Lemma 2 of Chen et al. (2015), and the Hausdorff distance is much smaller than the reach.
To proced, we require the following lemma.

**Lemma 2.** *Let $R_1$, $R_2$ be two closed, nonself-intersecting curves with positive reach. If*

$$\mathit{Haus}(R_1,R_2) < (2-\sqrt{2})\min\{\mathit{reach}(R_1), \mathit{reach}(R_2)\}$$

*then,*

$$\mathit{dist}_H(R_2,R_1) = \mathit{dist}_H(R_1,R_2) = \mathit{Haus}(R_1,R_2) \tag{30}$$

The proof can be found in Chen et al. (2015).

Following Lemma 2, the quasi-Hausdorff distance is the same as the Hausdorff distance, so that

$$|\sqrt{nh^{d+3}}\texttt{Haus}(\widehat{M}_n, M_n) - \mathbf{G}| = O(||\hat{p}_n - p_n||^*_{\infty,4}) \tag{31}$$

Equation (31) is the coupling between Hausdorff distance and the supremum of an empirical process, being the main result of step 1 (proved in another approach in (28)). It is to be noted that a sufficient condition for $||\hat{p}_n - p_n||^*_{\infty,4}$ being small is that $\frac{nh^{d+5}}{\log n} \to \infty, h \to 0$. This is the bandwidth condition that we require.

**Step 2** We will show

$$\mathbb{P}\left(|\mathbf{G}_n - \mathbf{B}| > A_1 \frac{b_0 \log^{2/3} n}{\gamma^{1/3}\left(nh^{d+3}\right)^{1/6}}\right) \leq A_2\gamma \tag{32}$$

for some constants $A_1, A_2$.

We first recall a useful Theorem in Chernozhukov et al. (2014c):

**Theorem 6** (Theorem 3.1 in Chernozhukov et al. (2014a))**.** *Let $\mathscr{G}$ be a collection of functions that is a VC-type class (see condition (K2)) with a constant envelope function b. Let $\sigma^2$ be a constant such that $\sup_{g\in\mathscr{G}}\mathbb{E}\left[g(X_i)^2\right]\leq\sigma^2\leq b^2$. Let $\mathbb{B}$ be a centered, tight Gaussian process defined on $\mathscr{G}$ with covariance function*

$$\operatorname{Cov}(\mathbb{B}(g_1),\mathbb{B}(g_2))=\mathbb{E}[g_1(X_i)g_2(X_i)]-\mathbb{E}[g_1(X_i)]\mathbb{E}[g_2(X_i)] \tag{33}$$

*where $g_1,g_2\in\mathscr{G}$. Then for any $\gamma\in(0,1)$ as $n$ is sufficiently large, there exist a random variable $\mathbf{B}\overset{d}{=}\sup_{f\in\mathscr{G}}|\mathbb{B}(g)|$ such that*

$$\mathbb{P}\left(\left|\sup_{f\in\mathscr{G}}|\mathbb{G}_n(g)|-\mathbf{B}\right|>A_1\frac{b^{1/3}\sigma^{2/3}\log^{2/3}n}{\gamma^{1/3}n^{1/6}}\right)\leq A_2\gamma \tag{34}$$

*where $A_1,A_2$ are two universal constants. Note that $A\overset{d}{=}B$ for random variables $A,B$ means that $A$ and $B$ has the same distribution*

To apply Theorem 6, we need to verify conditions. By assumption (K2) and (A3), $\mathscr{F}$ is a VC-type class with constant envelope $b_0=C_K^2\tilde{\lambda}_2<\infty$. Note that $1/\widetilde{\lambda}_2$ is the bound on the inverse second derivative of $\widetilde{p}_{yy}(x,y)$ as $y$ is closed to a local mode. Now we find $\sigma^2$. By definition,

$$\sup_{f\in\mathscr{F}}\mathbb{E}\left[f(X_i)^2\right]\leq h^{d+3}b_0^2 \tag{35}$$

Thus, we can pick $\sigma^2=h^{d+3}((b_0^2\leq b_0^2$ if $h\leq 1$. Hence, applying Theorem 6 gives

$$\mathbb{P}\left(\left|\sup_{f\in\mathscr{F}}\left|\mathbb{G}_n(f)\right|-\mathbf{B}'\right|>A_1\frac{b_0h^{2/3}h^2\log^{2/3}n}{\gamma^{1/3}n^{1/6}}\right)\leq A_2\gamma \tag{36}$$

for some constants $A_1,A_2$ and $\gamma<1$ and $\mathbf{B}'\overset{d}{=}\sup_{f\in\mathscr{F}}|\mathbb{B}(f)|$, where $\mathbb{B}$ is a Gaussian process defined on $\mathscr{F}$.

Now multiply $\sqrt{h^{-d-3}}$ in the both side of the above expression and use the definition of $\mathbf{G}_n$ and the fact that $\frac{1}{\sqrt{h^{d+3}}}\mathbf{B}'=\mathbf{B}$,

$$\mathbb{P}\left(|\mathbf{G}_n-\mathbf{B}|>A_1\frac{b_0\log^{2/3}n}{\gamma^{1/3}\left(nh^{d+3}\right)^{1/6}}\right)\leq A_2\gamma \tag{37}$$

which is the desired result (32).

**Step 3** We first show the coupling between $\sqrt{nh^{d+3}}\Delta_n$ and B. We pick $\varepsilon = \left(nh^{d+5}\right)^{-1/4}$ in (26) so that

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n\right| > \left(nh^{d+5}\right)^{-1/4}\right) \leq D_1 e^{-D_2\sqrt{nh^{d+5}}} \tag{38}$$

As $n$ is sufficiently large, and by triangular inequality along with (32),

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\Delta_n - \mathbf{B}\right| > A_3 \frac{\log^{2/3} n}{\gamma^{1/3}\left(nh^{d+3}\right)^{1/6}}\right) \leq A_4\gamma, \tag{39}$$

for some constants $A_3, A_4 > 0$. Note that we absorb the rate $\left(nh^{d+5}\right)^{-1/4}$ in (38) into $A_3 \frac{\log^{2/3} n}{\gamma^{1/3}\left(nh^{d+3}\right)^{1/6}}$. This is valid since $\left(nh^{d+5}\right)^{-1/4}$ converges faster. Also, we absorb $D_1 e^{-D_2\sqrt{nh^{d+5}}}$ into $A_4\gamma$. We allow $\gamma \to 0$ as long as $\gamma$ converges at rate slower than $\left(nh^{d+5}\right)^{-1/4}$.

Now applying the anti-concentration inequality (version of Lemma 16 in Chen et al. (2015)); see also Corollary 2.1 in Chernozhukov et al. (2014a), Chernozhukov et al. (2014c), Chernozhukov et al. (2014b)), we conclude

$$\sup_t \left|\mathbb{P}\left(\sqrt{nh^{d+3}}\Delta_n < t\right) - \mathbb{P}(\mathbf{B} < t)\right| \leq A_5\left(A_3 \frac{\log^{2/3} n}{\gamma^{1/3}\left(nh^{d+3}\right)^{1/6}} + A_4\gamma\right) \tag{40}$$

for some $A_5 > 0$. Now by taking $\gamma = \left(\frac{\log n}{nh^{d+1}}\right)^{1/8}$, we obtain the desired result. □

The Theorem 5 shows that the smoothed uniform error $\tilde{\Delta}_n$ is distributed asymptotically (coupled) as the supremum of a Gaussian Process. In other words, we can mathematically define the relationship as follow:

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

We are next interested in the limiting behaviour of the bootstrap estimate. Let $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be the observed data, the bootstrap estimate is given by

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

where $\widehat{M}_n^*(x)$ is tge bootstrap regression mode set at $x$.

**Theorem 7** (Bootstrap Consistency, Theorem 8, Chen et al. (2016))**.** *Assume conditions (A1-3) and (K1-2). Also assume that* $nh^6/\log n \to \infty$, $h \to 0$. *Define* $\mathbf{B} = (h^{d+3})^{-1/2} \sup_{f \in \mathscr{F}} |\mathbb{B}(f)|$. *There exists* $\chi_n$ *such that* $\mathbb{P}(\chi_n) \geq 1 - O\left(\frac{1}{n}\right)$, *and for all* $\mathscr{D}_n \in \chi_n$,

$$\sup_{t \geq 0} \left| \mathbb{P}\left( \sqrt{nh^{d+3}} \hat{\Delta}_n^* < t \;\middle|\; \mathscr{D}_n \right) - \mathbb{P}(\mathbf{B} < t) \right| = O\left( \left( \frac{\log^4 n}{nh^{d+3}} \right)^{1/8} \right)$$

*Proof.* The proof of this theorem runs along the same line of the previous Theorem 5 (based on Chen et al. (2015)). We state the basic ideas and omit the details.

Chen et al. (2015) proves the theorem in three steps. First it is shown that the Hausdorff distance $\texttt{Haus}(\widehat{M}_n^*, \widehat{M}_n)$ conditioned on the observed data $\mathscr{D}_n$ can be approximated by an empirical process. Second, using result of Theorem 5, we bound the difference between the distributions of $\texttt{Haus}(\widehat{M}_n^*, \widehat{M}_n)$ and a Gaussian process defined on $\widehat{M}_n$. This uses the second an third steps of Theorem 5. The last step shows that the Gaussian process defined on $\widehat{M}_n$ is asymptotically the same being defined on $M_n$).

We note that the functional space defined in (24) depends on the probability measure $\mathbb{P}$ and smoothing parameter $h$. Since $y$ is defined on the smoothed local mode $\tilde{M}(n)$ and it requires second derivatives of smooth density $\tilde{p}(x,y)$. Both $\tilde{M}(x)$ and $\tilde{p}(x,y)$ are completely defined by $\mathbb{P}$ and $h$. For the bootstrap estimate, Theorem 5 implies that $\hat{\Delta}_n^*$ can be approximated by the marginal of a certain Gaussian process,

$$\sup_{f \in \mathscr{F}(\mathbb{P}_n, h)} |\mathbb{B}(f)| \tag{41}$$

Note now the function space depends on $\mathbb{P}_n$ and $h$. This is because for the bootstrap case, we are conditioned on the data ($\mathscr{D}$ i.e. empirical measure $\mathbb{P}_n$) and sampling from $\mathbb{P}_n$. The role of $\mathbb{P}$ is completely replaced by $\mathbb{P}_n$. For the functional space, the index $y$ takes values at the 'estimated' local modes $\hat{M}_n(x)$ and $\tilde{p}_{yy}(x,y)$ will be replaced by the second derivative of KDE $\hat{p}_n(x,y)$. Both quantities now are determined by the empirical measure $\mathbb{P}_n$ and the smoothing parameter $h$.

The maximal of Gaussian processes defined on the two functional space $\mathscr{F}(\mathbb{P}, h)$ and $\mathscr{F}(\mathbb{P}_n, h)$ will be asymptotically the same by Lemma 17, 19 and 20 in Chen

et al. (2015). Putting altogether, the result follows from the approximation

$$\hat{\Delta}_n^* \approx \sup_{f \in \mathscr{F}(\mathbb{P}_n, h)} |\mathbb{B}(f)| \approx \sup_{f \in \mathscr{F}(\mathbb{P}, h)} |\mathbb{B}(f)| \approx \Delta_n. \tag{42}$$

□

Theorem 7 shows that the limiting distribution for the bootstrap estimate $\hat{\Delta}_n^*$ is the same as the limiting distribution of $\tilde{\Delta}_n$ (recall Theorem 5) with high probability. (Note that $\hat{\Delta}_n^*$, given the data samples $\mathscr{D}_n$, is a random quantity.) Using Theorems 5 and 7, we conclude the following.

**Corollary 7.1** (Uniform confidence sets, Corollary 9, Chen et al. (2016)).**.** *Assume (A1-3) and (K1-2). Then as* $\frac{nh^6}{logn} \to \infty$ *and* $h \to 0$*,*

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

# 7 Prediction Sets

In this section, we discuss the application of modal regression in constructing prediction sets (Chen et al. (2016)). We define the following

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

For a point $x$ and a set $A$, recall that $d(x,A) = \inf_{y \in A} |x - y|$ . Then

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

are pointwise and uniform prediction sets, respectively, at the population level. This is because

$$\begin{aligned} \mathbb{P}(Y \in \mathscr{P}_{1-\alpha}(x) \mid X = x) &\geq 1 - \alpha \\ \mathbb{P}(Y \in \mathscr{P}_{1-\alpha}) &\geq 1 - \alpha \end{aligned}$$

Correspondingly, at the sample level, we use a KDE of the conditional density $\widehat{p}_n(y \mid x) = \widehat{p}_n(x,y)/\widehat{p}_n(x)$, and estimate $\varepsilon_{1-\alpha}(x)$ via

$$\widehat{\varepsilon}_{1-\alpha}(x) = \inf\left\{\varepsilon \geq 0 : \int_{\widehat{M}_n(x) \oplus \varepsilon} \widehat{p}_n(y \mid x) dy \geq 1 - \alpha\right\}$$

An estimated pointwise prediction set is then

$$\widehat{\mathscr{P}}_{1-\alpha}(x) = \widehat{M}_n(x) \oplus \widehat{\varepsilon}_{1-\alpha}(x)$$

This has the proper pointwise coverage with respect to samples drawn according to $\widehat{p}_n(y \mid x)$, so in an asymptotic regime in which $\widehat{p}_n(y \mid x) \to p_n(y \mid x)$, it will have the correct coverage with respect to the population distribution, as well. Similarly, we can define

$$\widehat{\varepsilon}_{1-\alpha} = \text{ Quantile } \left( \left\{ d\left(Y_i, \widehat{M}_n\left(X_i\right)\right) : i = 1, \ldots, n \right\}, 1-\alpha \right), \tag{43}$$

the $(1-\alpha)$ quantile of $d\left(Y_i, \widehat{M}_n\left(X_i\right)\right), i = 1, \ldots, n$, and then the estimated uniform prediction set is

$$\widehat{\mathscr{P}}_{1-\alpha} = \left\{ (x,y) : x \in D, y \in \widehat{M}_n(x) \oplus \widehat{\varepsilon}_{1-\alpha} \right\} \tag{44}$$

The estimated uniform prediction set has proper coverage with respect to the empirical distribution, and so certain conditions, it will have valid limiting population coverage.

## 7.1 Bandwidth Selection

In this sub-section, we will discuss the application of uniform prediction sets in selecting the smoothing parameter $h$ of the underlying KDE. Based on definition (44), the volume (Lebesgue measure) of the estimated uniform prediction set is defined as

$$\texttt{Vol}(\hat{\mathscr{P}}_{1-\alpha,h}) = \hat{\varepsilon}_{1-\alpha,h} \int_{x \in D} \hat{K}_h(x) dx,$$

where $\hat{K}_h(x)$ is the number of estimated local modes at $X = x$, and $\hat{\varepsilon}_{1-\alpha,h}$ is as defined in (43). Roughly speaking, a small $h$ corresponds to a small $\hat{\varepsilon}_{1-\alpha,h}$, but a large $\hat{K}_h(x)$ and a large $h$ corresponds to a large $\hat{\varepsilon}_{1-\alpha,h}$, but a small $\hat{K}_h(x)$. So, we select an optimal $h$, $h^*$ say, as

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

Figure 5 illustrates the above rule with $\alpha = 0.05$. There is a evidently a trade-of in the size of the prediction set versus $h$ in the plot. We also obtain the optimal $h$ for the local regression method. The figure helps in visualizing the strength of modal regression, which not being constrained to modeling conditional mean structure,

can produce smaller prediction sets than the usual regression methods when the conditional mean fails to capture the main structure in the data.

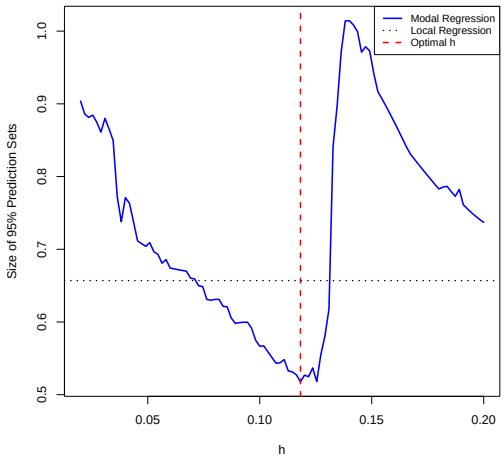



Figure 5: Bandwidth selection based on size of prediction sets.

It can be shown that, at the population level, under certain assumptions, the prediction sets from modal regression ca be smaller than those based the underlying regression function $\mu(x) = \mathbb{E}(Y|X = x)$. For a comprehensive review, we direct the reader to Section 6.2 of Chen et al. (2016).

# 8 Discussion

In this report we reviewed multi-modal regression along with some relevant topics like asymptotic theory, confidence and prediction sets and bandwidth selection. Here we outline some other relevant topics that were out of the scope of this report.

Two concepts related to modal regression are mixture regression and density ridges. A comprehensive analysis is done in Chen et al. (2016), which also describes how clustering can be used to conduct modal clustering (Section 7.1).

Chen (2018) reviews uni-modal regression and provides references to certain extensions and generalizations to measurement error case and censored response variables. Similar extensions and generalizations in the multi-modal regression are yet to be explored more vigorously.

We have reviewed confidence band construction based on a bootstrap approach. However, because this confidence band does not correct bias in KDE, it requires an undersmoothing assumption. Recently, Calonico et al. (2018) proposed a debiased approach that constructs a bootstrap nonparameteric confidence set without undersmoothing. The application of this approach to modal regression is another possible future direction (Chen (2018)).

A classical problem in nonparameteric statistics is bump hunting (Burman and Polonik (2009), Hall et al. (2004)), which detects the number of significant local modes. An interesting study of the bump hunting problem may be in the modal regression setting.

## 9 Supplementary Material

The interested reader is directed to https://github.com/ArkaB-DS/NPmodalReg which contains all the figures present here in the directory `images` and the corresponding codes to generate them in the `R` directory.

## 10 Acknowledgements

We take this opportunity to heartily thank our supervisor Prof. Subhra Sankar Dhar for his valuable feedback and constant guidance on this project.

## References

Arias-Castro, E., Mason, D., and Pelletier, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514.

Burman, P. and Polonik, W. (2009). Multivariate mode hunting: Data analytic tools with measures of significance. *Journal of Multivariate Analysis*, 100(6):1198–1218.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779.

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.

Chaouch, M., Laïb, N., and Louani, D. (2017). Rate of uniform consistency for a class of mode regression on functional stationary ergodic data. *Statistical Methods & Applications*, 26(1):19–47.

Chen, Y.-C. (2018). Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(4):e1431.

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

Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2015). Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5).

Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2017). Statistical inference using the Morse-Smale complex. *Electronic Journal of Statistics*, 11(1):1390 – 1433.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5).

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014b). Comparison and anti-concentration bounds for maxima of gaussian random vectors.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014c). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4).

Collomb, G., Härdle, W., and Hassani, S. (1986). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, 15:227–236.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

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.

Federer, H. (1959). Curvature measures. *Trans. Am. Math. Soc.*, 93:418–491.

Hall, P., Minnotte, M. C., and Zhang, C. (2004). Bump hunting with non-gaussian kernels. *The Annals of Statistics*, 32(5):2124–2141.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336.

Lee, M.-j. (1989). Mode regression. *Journal of Econometrics*, 42(3):337–349.

Rojas, A. L., Genovese, C. R., Miller, C. J., Nichol, R., and Wasserman, L. (2005). Conditional density estimation using finite mixture models with an application to astrophysics. *Cetner for Automatic Learning and Discovery, Department of Statistics, Carnegie Mellon University*.

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

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126:505–563.

Wang, X., Chen, H., Cai, W., Shen, D., and Huang, H. (2017). Regularized modal regression with applications in cognitive impairment prediction. *Advances in neural information processing systems*, 30.

Xiang, S. and Yao, W. (2022). Nonparametric statistical learning based on modal regression. *Journal of Computational and Applied Mathematics*, page 114130.

Yao, W. and Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671.

Yao, W. and Xiang, S. (2016). Nonparametric and varying coefficient modal regression. *arXiv preprint arXiv:1602.06609*.