NONPARAMETRIC MODAL REGRESSION

By Yen-Chi Chen, Christopher R. Genovese,
Ryan J. Tibshirani, Larry Wasserman

Carnegie Mellon University
December 8, 2014

Modal regression estimates the local modes of the distribution of Y given $X = x$, instead of the mean, as in the usual regression sense, and can hence reveal important structure missed by usual regression methods. We study a simple nonparametric method for modal regression, based on a kernel density estimate (KDE) of the joint distribution of Y and X. We derive asymptotic error bounds for this method, and propose techniques for constructing confidence sets and prediction sets. The latter is used to select the smoothing bandwidth of the underlying KDE. The idea behind modal regression is connected to many others, such as mixture regression and density ridge estimation, and we discuss these ties as well.

1. Introduction. Modal regression (Sager and Thisted, 1982; Lee, 1989; Yao et al., 2012; Yao and Li, 2013) is an alternate approach to the usual regression methods for exploring the relationship between a response variable Y and a predictor variable X. Unlike conventional regression, which is based on the conditional mean of Y given $X = x$, modal regression estimates conditional modes of Y given $X = x$.

Why would we ever use modal regression in favor a conventional regression method? The answer, at a high-level, is that conditional modes can reveal structure that is missed by the conditional mean. Figure 1 gives a definitive illustration of this point: we can see that, for the data examples in question, the conditional mean both fails to capture the major trends present in the response, and produces unecessarily wide prediction bands. Modal regression is an improvement in both of these regards (better trend estimation, and narrower prediction bands). In this paper, we rigorously study and develop its properties.

Modal regression has been used in transportation (Einbeck and Tutz, 2006), astronomy (Rojas, 2005), meteorology (Hyndman et al., 1996) and economics (Huang and Yao, 2012; Huang et al., 2013). Formally, the conditional (or local) mode set at x is defined as

$$M(x) = \left\{ y : \frac{\partial}{\partial y} p(y|x) = 0,\ \frac{\partial^2}{\partial y^2} p(y|x) < 0 \right\}, \tag{1}$$

where $p(y|x) = p(x, y)/p(x)$ is the conditional density of Y given $X = x$. As a simplification, the set $M(x)$ can be expressed in terms of the joint density alone:

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

MSC 2010 subject classifications: Primary 62G08; secondary 62G20, 62G05

Keywords and phrases: nonparametric regression, modes, mixture model, confidence set, prediction set, bootstrap

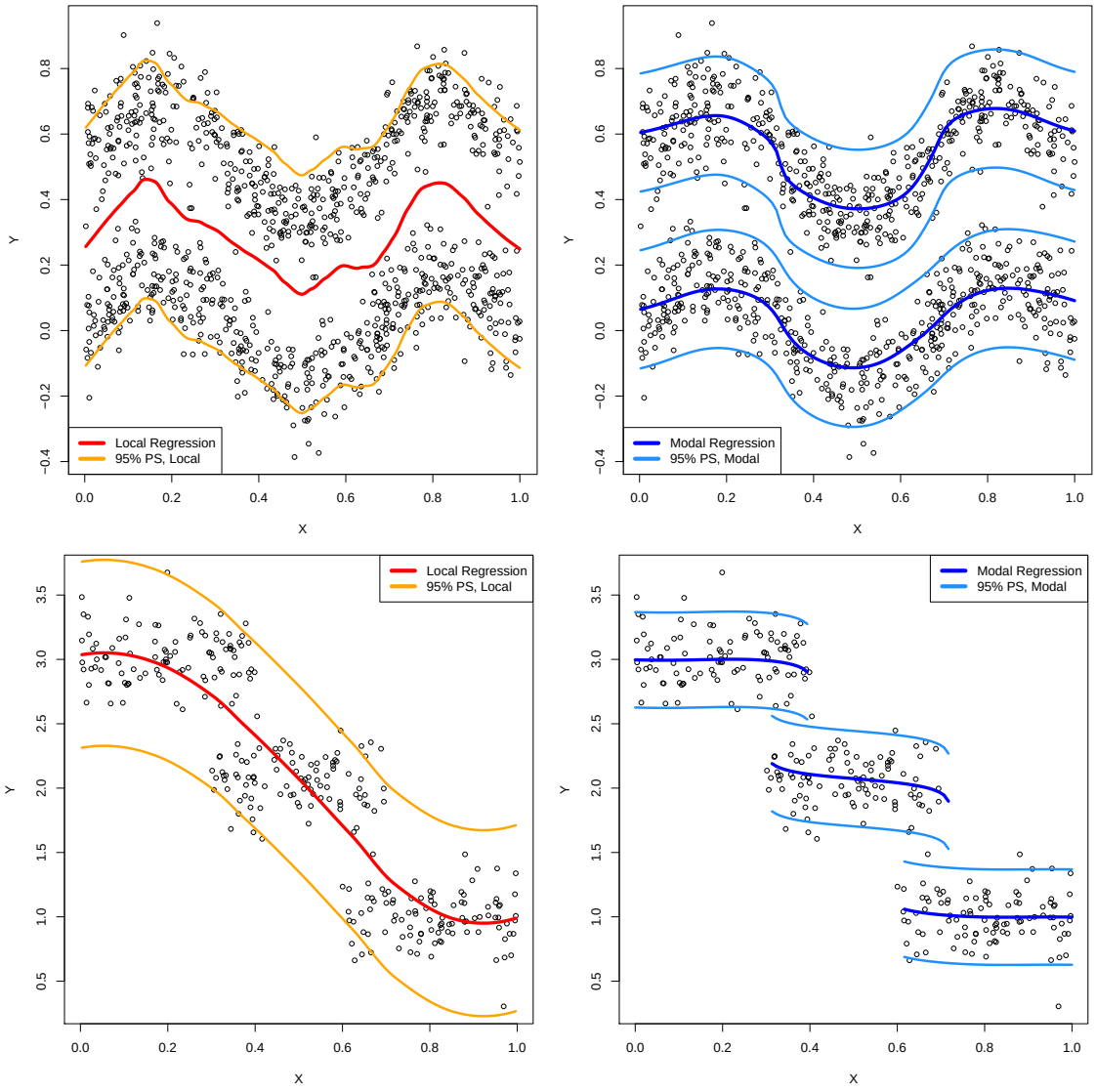


FIG 1. *Examples of modal regression versus a common nonparametric regression estimator, local linear regression. In the top row, we show local regression estimate and its associated 95% prediction bands alongside the modal regression and its 95% prediction bands. The bottom row does the same for a different data example. The local regression method fails to capture the structure, and produces prediction bands that are too wide.*

At each x, the local mode set $M(x)$ may consist of several points, and so $M(x)$ is in general a multi-valued function. Under appropriate conditions, as we will show, these modes change smoothly as x changes. Thus, local modes behave like a collection of surfaces which we call *modal manifolds*.

We focus on a nonparametric estimate of the conditional mode set, derived from a kernel density estimator (KDE):

$$\widehat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y}\widehat{p}_n(x,y) = 0,\ \frac{\partial^2}{\partial y^2}\widehat{p}_n(x,y) < 0 \right\}, \tag{3}$$

where $\widehat{p}_n(x,y)$ is the joint KDE of X, Y. Scott (1992) proposed this plug-in estimator for modal regression, and Einbeck and Tutz (2006) proposed a fast algorithm. We extend the work of these authors by giving a thorough treatment and analysis of nonparametric modal regression. In particular, our contributions are as follows.

1. We study the geometric properties of modal regression.
2. We prove consistency of the nonparametric modal regression estimator, and furthermore derive explicit convergence rates, with respect to various error metrics.
3. We propose a method for constructing confidence sets, using the bootstrap, and prove that it has proper asymptotic coverage.
4. We propose a method for constructing prediction sets, based on plug-in methods, and prove that the population prediction sets from this method can be smaller than those based on the regression function.
5. We propose a rule for selecting the smoothing bandwidth of the KDE (used to form the modal set $\widehat{M}(x)$) based on minimizing the size of prediction sets.
6. We draw enlightening comparisons to mixture regression (which suggests a clustering method using modal regression) and to density ridge estimation.

We begin by reviewing basic properties of modal regression and recalling previous work, in Section 2. Sections 3 through 8 then follow roughly the topics described in items 1–6 above. In Section 9 we end with some discussion. Simple R code for the modal regression and some simulation data sets used in this paper can be found at `http://www.stat.cmu.edu/~yenchic/ModalRegression.zip`.

2. Review of modal regression. Consider a response variable $Y \in \mathbb{K} \subseteq \mathbb{R}$ and covariate or predictor variable $X \in D \subseteq \mathbb{R}^d$, where D is a compact set. A classic take on modal regression (Sager and Thisted, 1982; Lee, 1989; Yao and Li, 2013) is to assume a linear model

$$\mathsf{Mode}(Y|X = x) = \beta_0 + \beta^T x,$$

where $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ are unknown coefficients, and $\mathsf{Mode}(Y|X = x)$ denotes the (singular) mode of Y given $X = x$. This is the same as the assuming that the set $M(x)$ in (1) is a singleton for each x, and that $M(x)$ depends linearly on x. Nonparametric modal regression is of course more flexible, because it allows $M(x)$ to be multi-valued, and also it models the components of $M(x)$ as smooth functions of x (not necessarily linear). As another nonlinear generalization of the above model, Yao et al. (2012) propose an interesting local polynomial smoothing method for mode estimation; however, they focus on $\mathsf{Mode}(Y|X = x)$ rather than $M(x)$, the collection of all conditional modes.

The estimated local mode set $\widehat{M}_n(x)$ in (3) from Scott (1992) relies on an estimated joint density function $\widehat{p}_n(x,y)$, most commonly computed using a KDE. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be the observed data samples. Then the KDE of the joint density $p(x,y)$ is

$$\widehat{p}_n(x,y) = \frac{1}{nh^{d+1}} \sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{h}\right) K\left(\frac{y-Y_i}{h}\right). \tag{4}$$

Here K is a symmetric, smooth kernel function, such as the Gaussian kernel (i.e., $K(u) = e^{-u^2/2}/\sqrt{2\pi}$), and $h>0$ is a parameter called the smoothing bandwidth. For simplicity, we have used the same kernel function K and bandwidth h for the covariates and the response, but this is not necessary. For brevity, we will write the estimated modal set as

$$\widehat{M}_n(x) = \Big\{y : \widehat{p}_{y,n}(x,y) = 0,\ \widehat{p}_{yy,n}(x,y) < 0\Big\}, \tag{5}$$

where the subscript notation denotes partial derivatives, as in $f_y = \partial f(x,y)/\partial y$ and $f_{yy} = \partial^2 f(x,y)/\partial y^2$.

In general, calculating $\widehat{M}_n(x)$ can be challenging, but for special kernels K, Einbeck and Tutz (2006) proposed a simple and efficient algorithm for computing local mode estimates. Their approach is based on the mean-shift algorithm (Cheng, 1995; Comaniciu and Meer, 2002). Mean-shift algorithms can be derived for any KDEs with radially symmetric kernels (Comaniciu and Meer, 2002), but for simplicity we assume here that K is Gaussian. The "partial" mean-shift algorithm of Einbeck and Tutz (2006), to estimate conditional modes, is described in Algorithm 1.

Algorithm 1 Partial mean-shift

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

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

$$y \longleftarrow \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|x-X_i\|}{h}\right) K\left(\frac{y-Y_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{h}\right) K\left(\frac{y-Y_i}{h}\right)}. \tag{6}$$

Output: The set $\mathcal{M}^\infty$, containing the points (x,y^∞), where x is a predictor value as fixed in $\mathcal{M}$, and y^∞ is the corresponding limit of the mean-shift iterations (6).

A straightforward calculation shows that the mean-shift update (6) is indeed a gradient ascent update on the function $f(y) = \widehat{p}_n(x,y)$ (for fixed x), with an implicit choice of step size. Because this function f is generically nonconcave, 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., 2013).

3. Geometric properties. In this section, we study the geometric properties of modal regression. Recall that $M(x)$ is a set of points at each input x. We define the *modal manifold collection* as the union of these sets over all inputs x,

$$\mathcal{S} = \{(x,y) : x \in D,\ y \in M(x)\}. \tag{7}$$

By the implicit function theorem, the set $\mathcal{S}$ has dimension d; see Figure 2 for an illustration with $d = 1$ (univariate x).

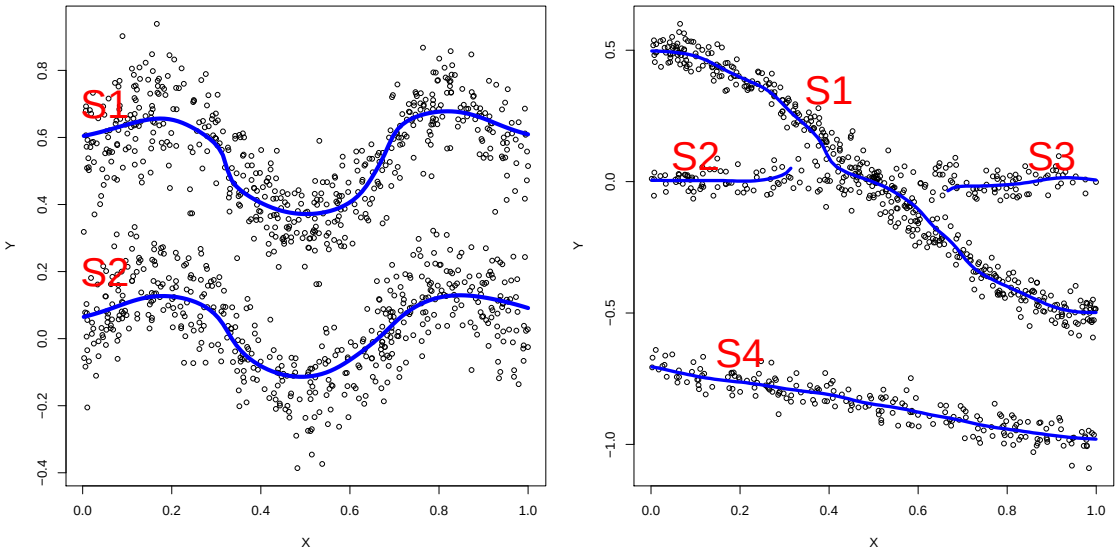


FIG 2. *Examples of modal manifolds.*

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

$$\mathcal{S} = S_1 \cup \ldots \cup S_K, \tag{8}$$

where each S_j is a connected manifold that admits a parametrization

$$S_j = \{(x, m_j(x)) : x \in A_j\}, \tag{9}$$

for some function $m_j(x)$ and open set A_j. Note that $A_1, \ldots, A_K$ form an open cover for the support D of X. We call S_j the *jth modal manifold*, and $m_j(x)$ the *jth modal function*. By convention we let $m_j(x) = \emptyset$ if $x \notin A_j$, and therefore we may write

$$M(x) = \{m_1(x), \ldots, m_K(x)\}. \tag{10}$$

That is, at any x, the values among $m_1(x), \ldots, m_K(x)$ that are nonempty give local modes.

Under weak assumptions, each $m_j(x)$ is differentiable, and so is the modal set $M(x)$, in a sense. We discuss this next.

LEMMA 1 (Derivative of modal functions). *Assume that p is twice differentiable, and let $\mathcal{S} = \{(x, y) : x \in D,\ y \in M(x)\}$ be the modal manifold collection. Assume that $\mathcal{S}$ factorizes according to (7), (8). Then, when $x \in A_j$,*

$$\nabla m_j(x) = -\frac{p_{yx}(x, m_j(x))}{p_{yy}(x, m_j(x))}, \tag{11}$$

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

PROOF. Since we assume that $x \in A_j$, we have $p_y(x, m_j(x)) = 0$ by definition. Taking a gradient over x yields

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

After rearrangement,

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

□

Lemma 1 links the geometry of the joint density function to the smoothness of the modal functions (and modal manifolds). The formula (11) is well-defined as long as $p_{yy}(x, m_j(x))$ is nonzero, which is guaranteed by the definition of local modes. Thus, when p is smooth, each modal manifold is also smooth.

To characterize smoothness of $M(x)$ itself, we require a notion for smoothness over sets. For this, we recall the *Hausdorff distance* between two sets A, B, defined as

$$\mathsf{Haus}(A, B) = \inf\{r : A \subseteq B \oplus r,\ B \subseteq A \oplus r\}, \tag{12}$$

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

THEOREM 2 (Smoothness of the modal manifold collection). *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*

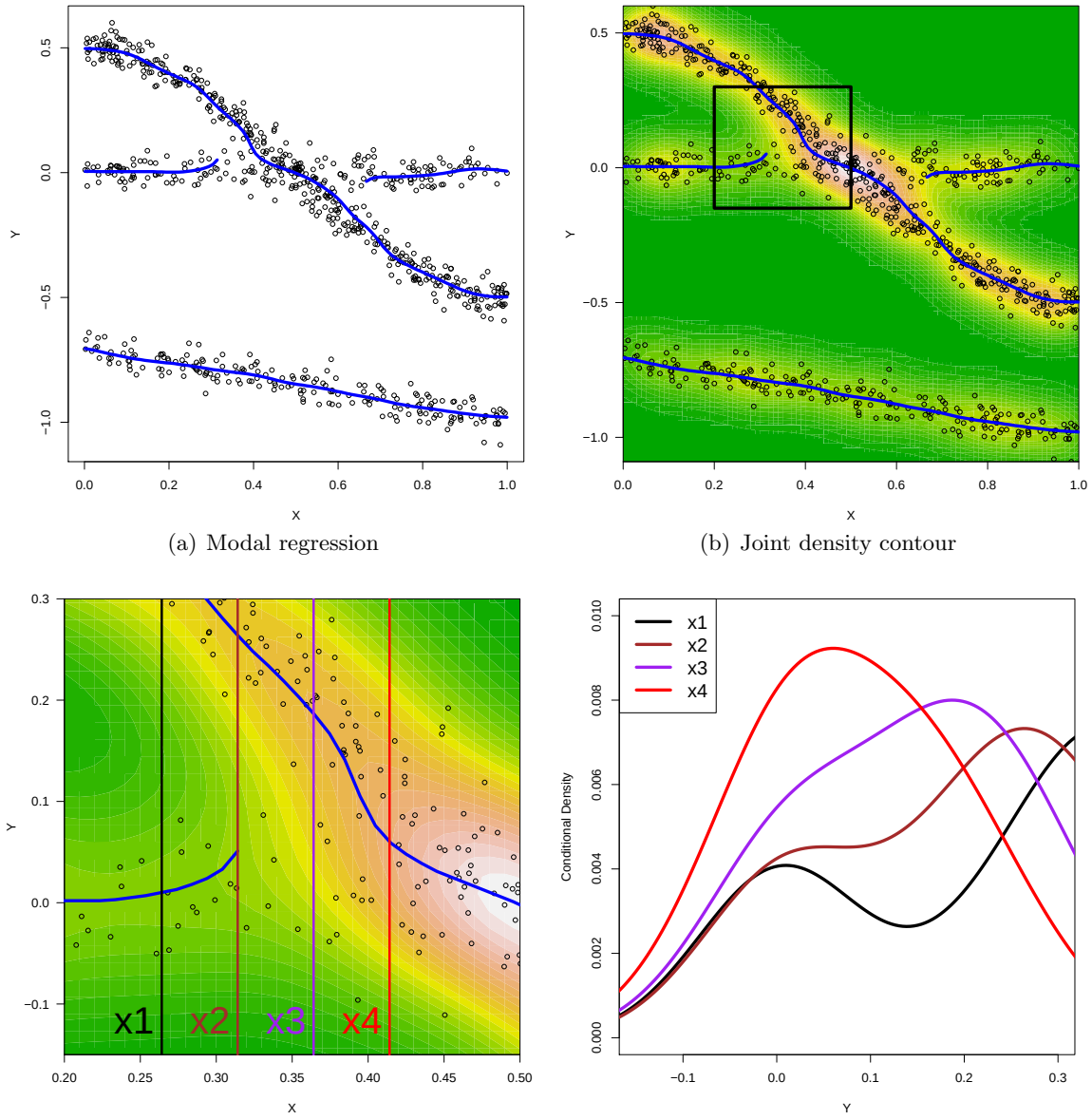
$$\lim_{|\epsilon| \to 0} \frac{\mathsf{Haus}(M(x), M(x + \epsilon))}{|\epsilon|} \leq \max_{j=1,\ldots,K} \|m_j'(x)\| \leq \frac{C}{\lambda_2} < \infty. \tag{13}$$

The proof of this result follows directly from Lemma 1 and the definition of Hausdorff distance, so we omit it. Since $M(x)$ is a multi-valued function, classic notions of smoothness cannot be applied, and Theorem 2 describes its smoothness in terms of Hausdorff distance. This distance can be thought of as a generalized ℓ_∞ distance for sets, and Theorem 2 can be interpreted as a statement about Lipschitz continuity with respect to Hausdorff distance.

Modal manifolds can merge or bifurcate as x varies. Interestingly, though, the merges or bifurcations do not necessarily occur at points of contact between two modal manifolds. See Figure 3 for an example with $d = 1$. Shown is a modal curve (manifold with $d = 1$), starting at $x = 0$ and stopping at about $x = 0.35$, which leaves a gap between itself and the neighboring modal curve. We take a closer look at the joint density contours, in panel (c), and inspect the conditional density along four slices $X = x_1, \ldots, x_4$, in panel (d). We see that after $X = x_2$, the conditional density becomes unimodal and the first (left) mode disappears, as it slides into a saddle point.

Lastly, the population quantities defined above all have sample analogs. For the estimate $\widehat{M}_n(x)$, we define

$$\widehat{\mathcal{S}}_n = \left\{(x, y) : y \in \widehat{M}_n(x), x \in \mathbb{R}\right\} = \widehat{S}_1 \cup \ldots \cup \widehat{S}_{\widehat{K}}, \tag{14}$$



(a) Modal regression

(b) Joint density contour

(c) Zoomed-in density contour

(d) Conditional density given $X = x_i$

FIG 3. *A look at bifurcation. As X moves x_1 to x_4, we can see that a local mode disappears after $X = x_2$.*

where each $\widehat{S}_j$ is a connected manifold, and $\widehat{K}$ is the total number. We also define $\widehat{m}_j(x)$ in a similar way for $j = 1, \ldots, \widehat{K}$. Thus, we can write

$$\widehat{M}_n(x) = \left\{\widehat{m}_1(x), \ldots, \widehat{m}_{\widehat{K}}(x)\right\}. \tag{15}$$

A subtle point: in practice, determining the manifold memberships and the total number of manifolds $\widehat{K}$ is not entirely trivial. In principle, the sample manifolds $\widehat{S}_1, \ldots, \widehat{S}_{\widehat{K}}$ are well-defined in terms of the sample estimate $\widehat{M}_n(x)$; but even with a perfectly convergent mean-shift algorithm, we would need to run mean-shift iterations at every input x in the domain D to determine these manifold components. Clearly this is not an implementable strategy. Thus from the output of the mean-shift algorithm over a finite mesh, we usually employ some type of simple post-processing technique to determine connectivity of the outputs and hence the sample manifolds. This is discussed further in Section 7.

4. Asymptotic error analysis. In this section we present asymptotic results about the convergence of the estimated modal regression set $\widehat{M}_n(x)$ to the underlying modal set $M(x)$. Let $\mathbf{BC}^k(C)$ denote the collection of k times continuously differentiable functions with all partial derivatives bounded in absolute value by C. (The domain of these functions should be clear from the context.) Given a kernel function $K : \mathbb{R} \to \mathbb{R}$, denote the collection of functions

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

where $K^{(\alpha)}$ denotes the α-th order derivative of K.

Our assumptions are as follows.

(A1) The joint density $p \in \mathbf{BC}^4(C_p)$ for some $C_p > 0$.

(A2) The collection of modal manifolds $\mathcal{S}$ can be factorized into $\mathcal{S} = S_1 \cup \ldots \cup S_K$, where each S_j is a connected curve that admits a parametrization $S_j = \{(x, m_j(x)) : x \in A_j\}$ for some $m_j(x)$, and $A_1, \ldots, A_K$ form an open cover for the support D of X.

(A3) There exists $\lambda_2 > 0$ such that for any $(x, y) \in D \times \mathbb{K}$ with $p_y(x, y) = 0$, $|p_{yy}(x, y)| > \lambda_2$.

(K1) The kernel function $K \in \mathbf{BC}^2(C_K)$ and satifies

$$\int_{\mathbb{R}} (K^{(\alpha)})^2(z)\, dz < \infty, \quad \int_{\mathbb{R}} z^2 K^{(\alpha)}(z)\, dz < \infty,$$

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

(K2) The collection $\mathcal{K}$ is a VC-type class, i.e. there exists $A, v > 0$ such that for $0 < \epsilon < 1$,

$$\sup_Q N\big(\mathcal{K}, L_2(Q), C_K \epsilon\big) \le \left(\frac{A}{\epsilon}\right)^v,$$

where $N(T, d, \epsilon)$ is the ϵ-covering number for a semi-metric space (T, d) and Q is any probability measure.

The assumption (A1) is an ordinary smoothness condition; we need fourth derivatives since we need to bound the bias of second derivatives. The assumption (A2) is to make sure the collection of all local modes can be represented as finite collection of manifolds. (A3) is a sharpness requirement for all critical points (local modes and minimums) and assumes no

saddle point; similar conditions appear in Romano (1988); Chen et al. (2014b) for estimating density modes. Assumption (K1) is assumed for the kernel density estimator to have the usual rates for its bias and variance. (K2) is for the uniform bounds on the kernel density estimator; this condition can be found in Gine and Guillou (2002); Einmahl and Mason (2005); Chen et al. (2014c). We study three types of error metrics for regression modes: pointwise, uniform, and mean integrated squared errors. We defer all proofs to Appendix A.

First we study the pointwise case. Recall that $\widehat{p}_n$ is the KDE in (4) of the joint density based on n samples, under the kernel K, and $\widehat{M}_n(x)$ is the estimated modal regression set in (5) at a point x. Our pointwise analysis considers

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

the Hausdorff distance between $\widehat{M}_n(x)$ and $M(x)$, at a point x. For our first result, we define the quantities:

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

THEOREM 3 (Pointwise error rate). *Assume (A1-3) and (K1-2). Then when*

$$\|\widehat{p}_n - p\|_{\infty,2}^{*} = \max\left\{\|\widehat{p}_n - p\|_\infty^{(0)}, \|\widehat{p}_n - p\|_\infty^{(1)}, \|\widehat{p}_n - p\|_\infty^{(2)}\right\}$$

is sufficiently small, we have

$$\sup_{x\in D} \frac{1}{\Delta_n(x)} \left|\Delta_n(x) - \max_{z\in M(x)} \left\{|p_{yy}^{-1}(x,z)||\widehat{p}_{y,n}(x,z)|\right\}\right| = O(\|\widehat{p}_n - p\|_{\infty,2}^{*}).$$

Moreover, at any fixed $x \in D$,

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

This shows that if the curvature of joint density function along y is bounded away from 0, then the error can be approximated by the error of $\widehat{p}_{y,n}(x,z)$ after scaling. The rate of convergence follows from the fact that $\widehat{p}_{y,n}(x,z)$ is converging to 0 at the same rate (note that as z is a conditional mode, the partial derivative of the true density is 0).

For our next result, we define the uniform error

$$\Delta_n = \sup_{x\in D} \Delta_n(x) = \sup_{x\in D} \mathsf{Haus}\big(\widehat{M}_n(x), M(x)\big).$$

This is an ℓ_∞ type error for estimating regression modes (and is also closely linked to confidence sets; see Section 5).

THEOREM 4 (Uniform error rate). *Assume (A1-3) and (K1-2). Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

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

Compared to the pointwise error rate in Theorem 3, we have an additional $\sqrt{\log n}$ factor in the second term. One can view this as the price we need to pay for getting an uniform bound over all points. See Gine and Guillou (2002); Einmahl and Mason (2005) for similar findings in density estimation.

The last error metric we consider is the mean integrated squared error (MISE), defined as

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

Note that the MISE is a nonrandom quantity, unlike first two error metrics considered.

THEOREM 5 (MISE rate). *Assume (A1-3) and (K1-2). Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

$$\mathsf{MISE}\left(\widehat{M}_n\right) = O(h^2) + O\left(\sqrt{\frac{1}{nh^{d+3}}}\right).$$

If we instead focus on estimating the regression modes of the smoothed joint density $\widetilde{p}(x,y) = \mathbb{E}\left(\widehat{p}_n(x,y)\right)$, then we obtain much faster convergence rates. Let $\widetilde{M}(x) = \mathbb{E}(\widehat{M}_n(x))$ be the smoothed regression modes at $x \in D$. Analogously define

$$\begin{aligned}
\widetilde{\Delta}_n(x) &= \mathsf{Haus}\big(\widehat{M}_n(x), \widetilde{M}(x)\big) \\
\widetilde{\Delta}_n &= \sup_{x\in D} \widetilde{\Delta}_n(x) \\
\widetilde{\mathsf{MISE}}\left(\widehat{M}_n\right) &= \mathbb{E}\left(\int_{x\in D} \widetilde{\Delta}_n^2(x)\, dx\right).
\end{aligned}$$

COROLLARY 6 (Error rates for smoothed conditional modes). *Assume (A1-3) and (K1-2). Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

$$\begin{aligned}
\sqrt{nh^{d+3}} \sup_{x\in D} \left|\widetilde{\Delta}_n(x) - \max_{z\in\widetilde{M}(x)} \{\widetilde{p}_{yy}^{-1}(x,z)\widehat{p}_{y,n}(x,z)\}\right| &= O_{\mathbb{P}}(\epsilon_{n,2}) \\
\widetilde{\Delta}_n(x) &= O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^{d+3}}}\right) \\
\widetilde{\Delta}_n &= O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+3}}}\right) \\
\widetilde{\mathsf{MISE}}\left(\widehat{M}_n\right) &= O\left(\sqrt{\frac{1}{nh^{d+3}}}\right),
\end{aligned}$$

where $\epsilon_{n,2} = \sup_{x,y} |\widehat{p}_{yy,n}(x,y) - \widetilde{p}_{yy}(x,y)| = \sup_{x,y} |\widehat{p}_{yy,n}(x,y) - \mathbb{E}(\widehat{p}_{yy,n}(x,y))|$.

5. Confidence sets. In an idealized setting, we could define a confidence set at x by

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

where

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

By construction, we have $\mathbb{P}(M(x) \in \widehat{C}_n^0(x)) = 1-\alpha$. Of course, the distribution of $\Delta_n(x)$ is unknown, but we can use the bootstrap (Efron, 1979) to estimate $\delta_{n,1-\alpha}(x)$.

Given the observed data samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, we denote a bootstrap sample as $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$. Let $\widehat{M}_n^*(x)$ be the estimated regression modes based on this bootstrap sample, and

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

We repeat the bootstrap sampling B times to get $\widehat{\Delta}_{1,n}^*(x), \ldots, \widehat{\Delta}_{B,n}^*(x)$. Define $\widehat{\delta}_{n,1-\alpha}(x)$ by

$$\frac{1}{B}\sum_{j=1}^{B} I\Big(\widehat{\Delta}_{j,n}^*(x) > \widehat{\delta}_{n,1-\alpha}(x)\Big) \approx \alpha.$$

Our estimated confidence set for $M(x)$ is then given by

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

Note that this is a pointwise confidence set, at $x \in D$.

Alternatively, we can use $\Delta_n = \sup_{x\in D} \Delta_n(x)$ to build a uniform confidence set. Define $\delta_{n,1-\alpha}$ by

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

As above, we can use bootstrap sampling to form an estimate $\widehat{\delta}_{n,1-\alpha}$, based on the quantiles of the bootstrapped uniform error metric

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

Our estimated uniform confidence set is then

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

In practice, there are many possible flavors of the bootstrap that are applicable here. This includes the ordinary (nonparametric) bootstrap, the smoothed bootstrap and the residual bootstrap. See Figure 4 for an example with the ordinary bootstrap.

Theoretically, we focus on the asymptotic coverage of uniform confidence sets built with the ordinary bootstrap. We consider coverage of the smoothed regression mode set $\widetilde{M}(x)$ (to avoid issues of bias), and we employ tools developed in Chernozhukov et al. (2013a,b); Chen et al. (2014c).

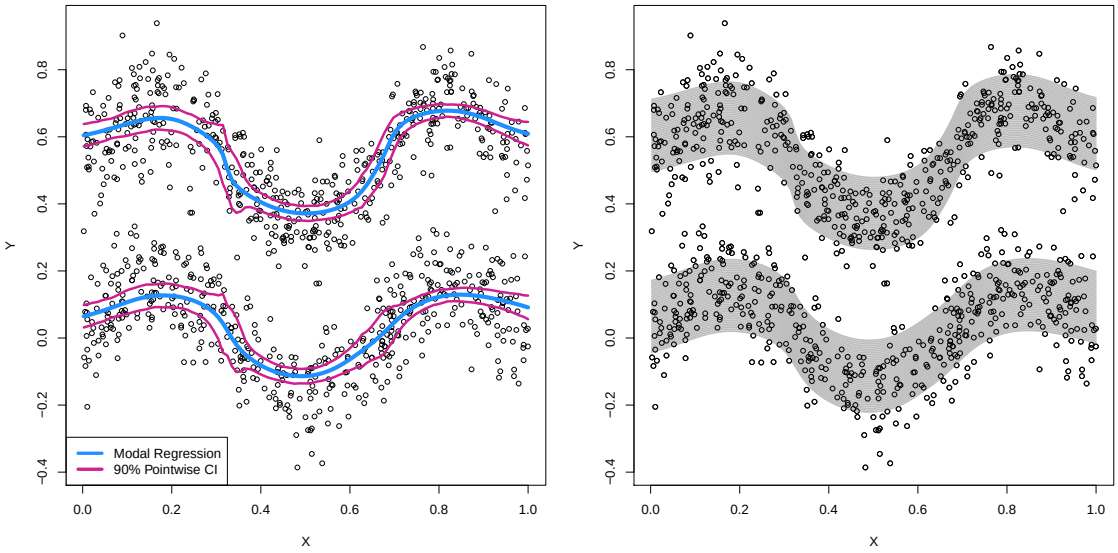


FIG 4. *An example with pointwise (left) and uniform (right) confidence sets. The significance level is* 90%.

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

$$\mathcal{F} = \left\{ (u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = \widetilde{p}_{yy}^{-1}(x,y) \times K\left(\frac{\|x-u\|}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), x \in D, y \in \widetilde{M}(x) \right\}, \tag{16}$$

and let $\mathbb{B}$ be a Gaussian process defined on $\mathcal{F}$ such that

$$\mathrm{Cov}(\mathbb{B}(f_1), \mathbb{B}(f_2)) = \mathbb{E}\left(f_1(X_i,Y_i) f_2(X_i,Y_i)\right) - \mathbb{E}\left(f_1(X_i,Y_i)\right) \mathbb{E}\left(f_2(X_i,Y_i)\right), \tag{17}$$

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

THEOREM 7 (Limiting Distribution). *Assume (A1-3) and (K1-2). Define the random variable* $\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f \in \mathcal{F}} |\mathbb{B}(f)|$. *Then as* $\frac{nh^{d+5}}{\log n} \to \infty$, $h \to 0$,

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

This theorem shows that the smoothed uniform discrepancy $\widetilde{\Delta}_n$ is distributed asymptotically as the supremum of a Gaussian process. In fact, it can be shown that the two random variables $\widetilde{\Delta}_n$ and $\mathbf{B}$ are coupled by

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

Now we turn to the limiting behavior for the bootstrap estimate. Let $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be the observed data and denote the bootstrap estimate

$$\widehat{\Delta}_n^* = \sup_{x \in D} \mathsf{Haus}\big(\widehat{M}_n^*(x), \widehat{M}_n(x)\big)$$

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

THEOREM 8 (Bootstrap Consistency). *Assume conditions (A1-3) and (K1-2). Also assume that* $\frac{nh^6}{\log n} \to \infty, h \to 0$. *Define* $\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f \in \mathcal{F}} |\mathbb{B}(f)|$. *There exists* $\mathcal{X}_n$ *such that* $\mathbb{P}(\mathcal{X}_n) \geq 1 - O(\frac{1}{n})$ *and, for all* $\mathcal{D}_n \in \mathcal{X}_n$,

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

Theorem 8 shows that the limiting distribution for the bootstrap estimate $\widehat{\Delta}_n^*$ is the same as the limiting distribution of $\widetilde{\Delta}_n$ (recall Theorem 7) with high probability. (Note that $\widehat{\Delta}_n^*$, given the data samples $\mathcal{D}_n$, is a random quantity.) Using Theorems 7 and 8, we conclude the following.

COROLLARY 9 (Uniform confidence sets). *Assume conditions (A1-3) and (K1-2). Then as* $\frac{nh^6}{\log n} \to \infty$ *and* $h \to 0$,

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

6. Prediction sets. Modal regression can be also used to construct prediction sets. Define

$$\epsilon_{1-\alpha}(x) = \inf\Big\{ \epsilon \geq 0 : \mathbb{P}\big(d(Y, M(X)) > \epsilon \,|\, X = x\big) \leq \alpha \Big\}$$
$$\epsilon_{1-\alpha} = \inf\Big\{ \epsilon \geq 0 : \mathbb{P}\big(d(Y, M(X)) > \epsilon\big) \leq \alpha \Big\}.$$

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

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

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

$$\mathbb{P}\big(Y \in \mathcal{P}_{1-\alpha}(x) \,|\, X = x\big) \geq 1 - \alpha$$
$$\mathbb{P}(Y \in \mathcal{P}_{1-\alpha}) \geq 1 - \alpha.$$

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

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

An estimated pointwise prediction set is then

$$\widehat{\mathcal{P}}_{1-\alpha}(x) = \widehat{M}_n(x) \oplus \widehat{\epsilon}_{1-\alpha}(x).$$

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

Similarly, we can define

$$\widehat{\epsilon}_{1-\alpha} = \mathsf{Quantile}\left(\left\{d\left(Y_i, \widehat{M}_n(X_i)\right) : i = 1, \ldots, n\right\}, 1-\alpha\right), \tag{18}$$

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

$$\widehat{\mathcal{P}}_{1-\alpha} = \left\{(x, y) : x \in D,\, y \in \widehat{M}_n(x) \oplus \widehat{\epsilon}_{1-\alpha}\right\}. \tag{19}$$

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.

6.1. *Bandwidth selection.* Importantly, prediction sets can be used to select the smoothing bandwidth of the underlying KDE, as we describe here. We focus on uniform prediction sets, and we will use a subscript h throughout to denote the dependence on the smoothing bandwidth. From its definition in (19), we can see that the volume (Lebesgue measure) of the estimated uniform prediction set is

$$\mathsf{Vol}\left(\widehat{\mathcal{P}}_{1-\alpha,h}\right) = \widehat{\epsilon}_{1-\alpha,h} \int_{x \in D} \widehat{K}_h(x)\, dx,$$

where $\widehat{K}_h(x)$ is the number of estimated local modes at $X = x$, and $\widehat{\epsilon}_{1-\alpha,h}$ is as defined in (18). Roughly speaking, when h is small, $\widehat{\epsilon}_{1-\alpha,h}$ is also small, but the number of estimated manifolds is large; on the other hand, when h is large, $\widehat{\epsilon}_{1-\alpha,h}$ is large, but the number of estimated manifolds is small. This is like the bias-variance trade-off: small h corresponds to less bias ($\widehat{\epsilon}_{1-\alpha,h}$) but higher variance (number of estimated manifolds).

Our proposal is to select h by

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

Figure 5 gives an example this rule when $\alpha = 0.05$, i.e., when minimizing the size of the estimated 95% uniform prediction set. As can be seen, there is a clear trade-off in the size of the prediction set versus h in the left plot. The optimal value $h^* = 0.11$ is marked by a vertical line, and the right plot displays the corresponding modal regression estimate and uniform prediction set on the data samples.

In the same plot, we also display a local regression estimate and its corresponding 95% uniform prediction set. We can see that the prediction set from the local regression method is much larger than that from modal regression. (To even the comparison, the bandwidth for the local linear smoother was also chosen to minimize the size of the prediction set.) This illustrates a major strength of the modal regression method: because it is not constrained to modeling conditional mean structure, it can produce smaller prediction sets than the usual regression methods when the conditional mean fails to capture the main structure in the data. We investigate this claim theoretically, next.

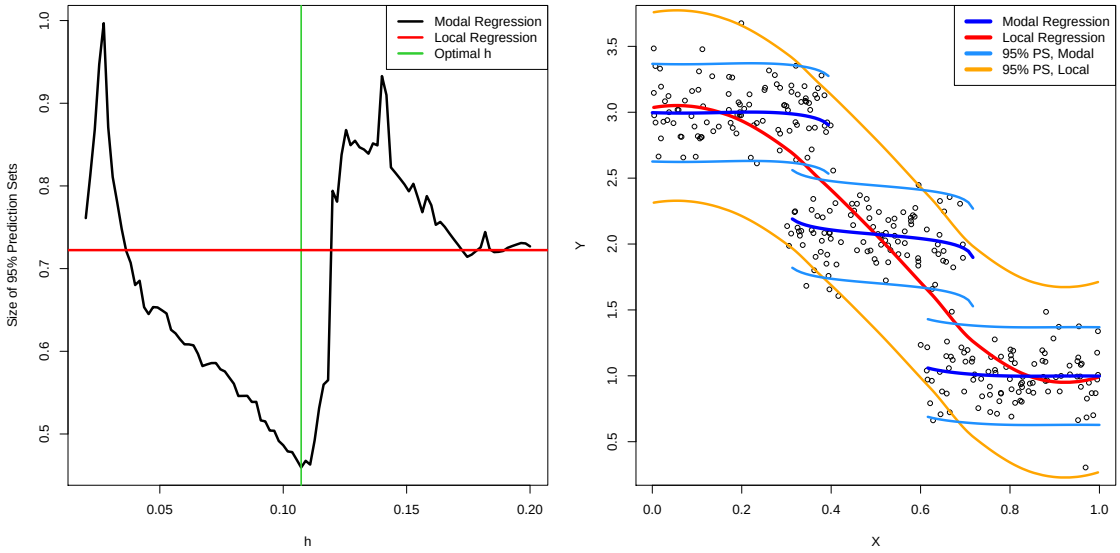


FIG 5. *An example of bandwidth selection based on the size of the prediction sets.*

6.2. *Theory on the size of prediction sets.* We will show that, at the population level, prediction sets from modal regression can be smaller than those based on the underlying regression function $\mu(x) = \mathbb{E}(Y|X = x)$. Defining

$$\eta_{1-\alpha}(x) = \inf\Big\{\eta \geq 0 : \ \mathbb{P}\big(d(Y, \mu(X)) > \eta \,|\, X = x\big) \leq \alpha\Big\}$$
$$\eta_{1-\alpha} = \inf\Big\{\eta \geq 0 : \ \mathbb{P}\big(d(Y, \mu(X)) > \eta\big) \leq \alpha\Big\},$$

pointwise and uniform prediction sets based on the regression function are

$$\mathcal{R}_{1-\alpha}(x) = \mu(x) \oplus \eta_{1-\alpha}(x) \subseteq \mathbb{R}$$
$$\mathcal{R}_{1-\alpha} = \Big\{(x, \mu(x) \oplus \eta_{1-\alpha}) : x \in D\Big\} \subseteq D \times \mathbb{R},$$

respectively.

For a pointwise prediction set $A(x)$, we write $\mathsf{length}(A(x))$ for its Lebesgue measure on $\mathbb{R}$; note that in the case of modal regression, this is somewhat of an abuse of notation because the Lebesgue measure of $A(x)$ can be a sum of interval lengths. For a uniform prediction set A, we write $\mathsf{Vol}(A)$ for its Lebesgue measure on $D \times \mathbb{R}$.

We consider the following assumption.

(GM): The conditional density satisfies

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

with $\mu_1(x) < \mu_2(x) < \ldots < \mu_{K(x)}(x)$ by convention, and $\phi(\,\cdot\,; \mu, \sigma^2)$ denoting the Gaussian density function with mean μ and variance σ^2.

The assumption that the conditional density can be written as a mixture of Gaussians is only used for the next result. It is important to note that this is an assumption made about the population density, and does not reflect modeling choices made in the sample. Indeed, recall, we are comparing prediction sets based on the modal set $M(x)$ and the regression function $\mu(x)$, both of which use true population information.

Before stating the result, we must define several quantities. Define the minimal separation between mixture centers

$$\Delta_{\min}(x) = \min\{|\mu_i(x) - \mu_j(x)| : i \neq j\}$$

and

$$\sigma^2_{\max}(x) = \max_{j=1,\ldots,K(x)} \sigma^2_j(x), \qquad \pi_{\max}(x) = \max_{j=1,\ldots,K(x)} \pi_j(x), \qquad \pi_{\min}(x) = \min_{j=1,\ldots,K(x)} \pi_j(x).$$

Also define

$$\Delta_{\min} = \inf_{x\in D} \Delta_{\min}(x), \quad \sigma^2_{\max} = \sup_{x\in D} \sigma^2_{\max}(x),$$

and

$$\pi_{\max} = \sup_{x\in D} \pi_{\max}(x), \quad \pi_{\min} = \inf_{x\in D} \pi_{\min}(x),$$

and

$$\overline{K} = \frac{\int_{x\in D} K(x)\,dx}{\int_{x\in D} dx}, \quad K_{\min} = \inf_{x\in D} K(x), \quad K_{\max} = \inf_{x\in D} K(x).$$

THEOREM 10 (Size of prediction sets). *Assume (GM). Let $\alpha < 0.1$ and assume that $\pi_1(x), \pi_{K(x)}(x) > \alpha$. If*

$$\frac{\Delta_{\min}(x)}{\sigma_{max}(x)} > \max\left\{1.1 \cdot \frac{K(x)}{K(x)-1} z_{1-\alpha/2}, \sqrt{6.4 \vee 2\log\left(4(K(x) \vee 3 - 1)\right) + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}\right\},$$

where z_α is the upper α-quantile value of a standard normal distribution and $A \vee B = \max\{A, B\}$, then

$$\mathsf{length}\big(\mathcal{P}_{1-\alpha}(x)\big) < \mathsf{length}\big(\mathcal{R}_{1-\alpha}(x)\big).$$

Moreover, if

$$\frac{\Delta_{\min}}{\sigma_{max}} > \max\left\{1.1 \cdot \left(\frac{2\overline{K}}{K_{\min}-1}\right) z_{1-\alpha/2}, \sqrt{6.4 \vee 2\log\left(4(K_{\max} \vee 3 - 1)\right) + 2\log\left(\frac{\pi_{\max}}{\pi_{\min}}\right)}\right\},$$

then

$$\mathsf{Vol}(\mathcal{P}_{1-\alpha}) < \mathsf{Vol}(\mathcal{R}_{1-\alpha}).$$

In words, the theorem shows that when the signal-to-noise ratio is sufficiently large, the modal-based prediction set is smaller than the usual regression-based prediction set.

7. Comparison to mixture regression. Mixture regression is in some ways similar to modal regression. The literature on mixture regression, also known as mixture of experts modeling, is vast; see, e.g., Jacobs et al. (1991); Jiang and Tanner (1999); Bishop (2006); Viele and Tong (2002); Khalili and Chen (2007); Hunter and Young (2012); Huang and Yao (2012); Huang et al. (2013). In mixture regression, we assume that the conditional density function takes the form

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

where each $\phi_j(y; \mu_j(x), \sigma_j^2(x))$ is a density function, parametrized by a mean $\mu_j(x)$ and variance $\sigma_j^2(x)$. The simplest and most common usage of mixture regression makes the following assumptions:

(MR1) $K(x) = K$,
(MR2) $\pi_j(x) = \pi_j$ for each j,
(MR3) $\mu_j(x) = \beta_j^T x$ for each j
(MR4) $\sigma_j^2(x) = \sigma_j^2$ for each j, and
(MR5) $\phi_j(x)$ is Gaussian for each j.

This is called linear mixture regression (Viele and Tong, 2002; Chaganty and Liang, 2013). Many authors have considered relaxing some subset of the above assumptions, but as far we can tell, no work has been proposed to effectively relax all of (MR1-5).

Modal regression is a fairly simple tool that achieves a similar goal to mixture regression models, and uses fewer assumptions. At a high level, mixture regression is inherently a model-based method, stemming from a model for the joint density $p(y|x)$; modal regression hunts directly for conditional modes, which can be estimated without a model for $p(y|x)$. Another important difference: the number of mixture components K in the mixture regression model plays a key role, and estimating K is quite difficult; in modal regression we do not need to estimate anything of this sort (e.g., we do not specify a number of modal manifolds). Instead, the flexibility of the estimated modal regression set is driven by the bandwidth parameter h of the KDE, which can be tuned by inspecting the size of prediction sets, as described in Section 6.1. Table 1 summarizes the comparison between mixture-based and mode-based methods.

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

TABLE 1
Comparison for methods based on mixtures versus modes.

Figure 6 gives a comparison between linear mixture regression and modal regression. We fit the linear mixture model using the R package mixtools, specifying $k = 3$ components, over 10,000 runs of the EM algorithm (choosing eventually the result the highest likelihood value). The modal regression estimate used a bandwidth value that minimized the volume of the corresponding prediction set, as characterized in Figure 5. The figure reveals yet another

important difference between the two methods: the estimated modal regression trends do not persist across the whole x domain, while the linear mixture model (in its default specification) carries the estimated linear trends across the entirety of the x domain. This is due to assumption (MR2), which models each component probability π_j as a constant, independent of x. As a result, the prediction set from the linear mixture model has a much larger volume than that from modal regression, since it vacuously covers the extensions of each linear fit across the whole domain. Relaxing assumption (MR2) would address this issue, but it would also make the mixture estimation more difficult.

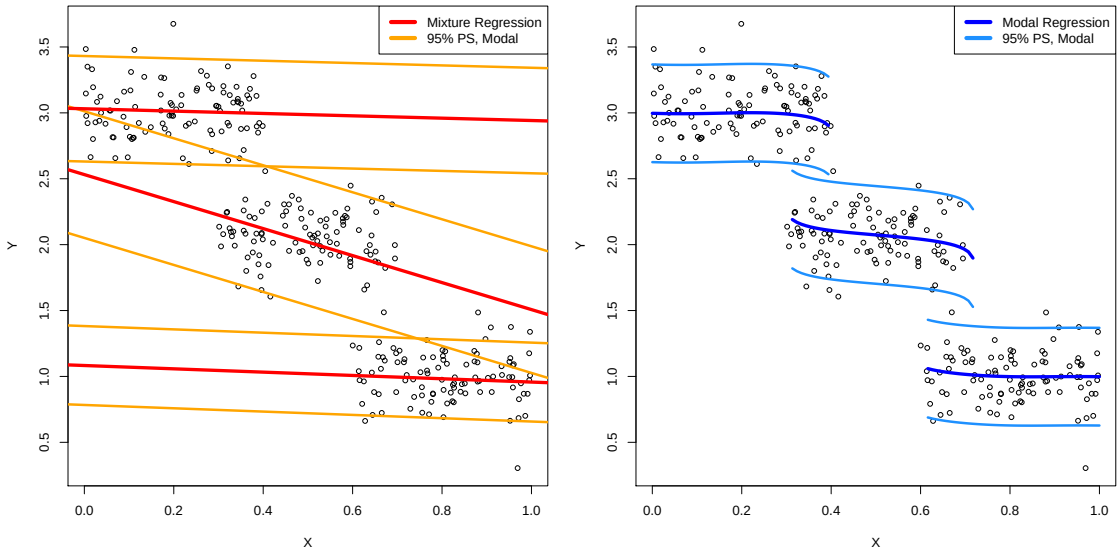


FIG 6. *A comparison between mixture regression, on the left, and modal regression, on the right.*

7.1. *Clustering with modal regression.* We now describe how modal regression can be used to conduct clustering, conditional on x. This clustering leads us define to modal proportions and modal dispersions, which are roughly analogous to the component parameters $\pi_j(x)$ and $\sigma_j^2(x)$ in mixture regression.

Mode-based clustering (Cheng, 1995; Comaniciu and Meer, 2002; Chen et al., 2014a) is a nonparametric clustering method which uses local density modes to define clusters. A similar idea applies to modal regression. In words, at each point x, we find the modes of $p(y|x)$ and we cluster according to the basins of attractions of these modes. Formally, at each (x, y), we define an ascending path by

$$\gamma_{(x,y)} : \mathbb{R}^+ \to \mathbb{K} \times D, \quad \gamma_{(x,y)}(0) = (x, y), \quad \gamma'_{(x,y)}(t) = (0, p_y(x, y)).$$

That is, $\gamma_{(x,y)}$ is the gradient ascent path in the y direction (with x fixed), starting at the point y. Denote the destination of the path by $\mathsf{dest}(x, y) = \lim_{t\to\infty} \gamma_{(x,y)}(t)$. By Morse theory, $\mathsf{dest}(x, y) = m_j(x)$ for one and only one regression mode $m_j(x)$, $j = 1, \ldots, K$. Thus, we assign the cluster label j to the point (x, y).

The above was a population-level description of the clusters. In practice, we use the mean-shift algorithm (Algorithm 1) to estimate clusters and assign points according to the output

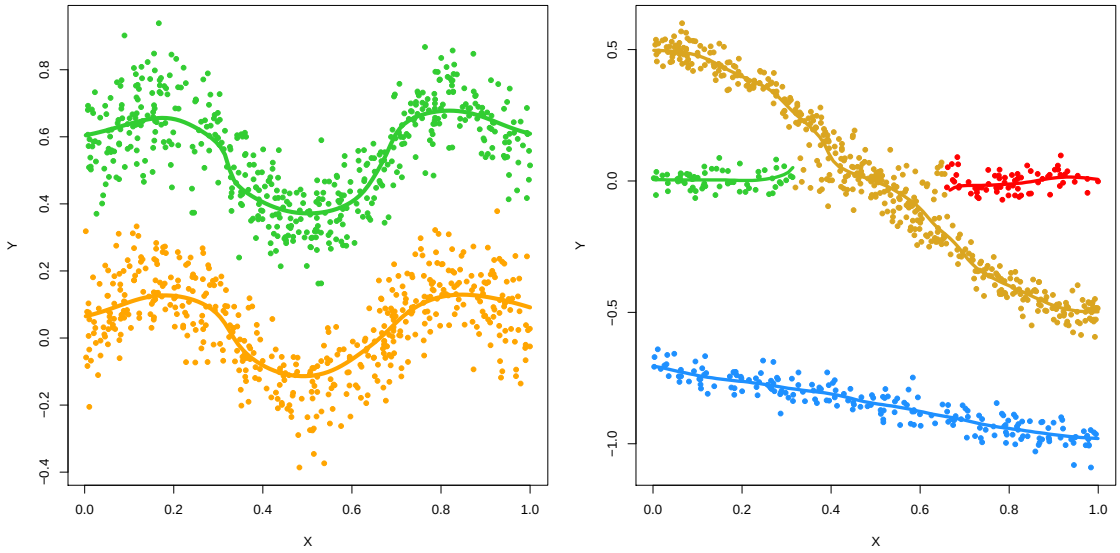


FIG 7. *Two examples of clustering based on modal regression.*

of the algorithm. That is, by iterating the mean-sfhit update (6) for each point (X_i, Y_i), with X_i fixed, we arrive at an estimated mode $\widehat{m}_j(X_i)$ for some $j = 1, \ldots, \widehat{K}$, and we hence assign (X_i, Y_i) to cluster j. An issue is that determination of the estimated modal functions $\widehat{m}_j$, $j = 1, \ldots, \widehat{K}$, or equivalently, of the modal manifolds $\widehat{S}_1, \ldots, \widehat{S}_{\widehat{K}}$, is not immediate from the data samples. These are well-defined in principle, but require running the mean-shift algorithm at each input point x. In data examples, therefore, we run mean-shift over a fine mesh (e.g., the data samples themselves) and apply hierarhical clustering to find the collection $\widehat{S}_1, \ldots, \widehat{S}_{\widehat{K}}$. It is important to note that the latter clustering task, which seeks a clustering of the outputs of the mean-shift algorithm, is trivial compared to the original task (clustering of the data samples). Some examples are shown in Figure 7.

The clustering assignments give rise to the concepts of modal proportions and modal dispersions. The modal proportion of cluster j is defined as

$$\widehat{q}_j = N_j / n,$$

where $N_j = \sum_{i=1}^n \mathbb{1}(i \in \widehat{C}_j)$ is the number of data points belonging to the jth cluster $\widehat{C}_j$. The modal disperson of cluster j is defined as

$$\widehat{\rho}_j^2 = \frac{1}{N_j} \sum_{i \in \widehat{C}_j} \big(Y_i - \widehat{m}(Y_i)\big)^2,$$

where $\widehat{m}(Y_i)$ denotes the sample destination at (X_i, Y_i) (i.e., the output of the mean-shift algorithm at (X_i, Y_i)). This is a measure of the spread of the data points around the jth estimated modal manifold.

In a mixture regression model, when we assume each density ϕ_j to be Gaussian, the local modes of $p(y|x)$ behave like the mixture centers $\mu_1(x), \ldots, \mu_K(x)$. Thus, estimating the local modes is like estimating the centers the Gaussian mixtures. The clustering based on modal

regression is like the recovery process for the mixing mechnism. Each cluster can be thought of a mixture component and hence the quantities $\widehat{q}_j, \widehat{\rho}_j^2$ are analogous to the estimates $\widehat{\pi}_j, \widehat{\sigma}_j^2$ in mixture regression (assuming (MR2) and (MR4), so that to the mixture proportions and variances do not depend on x).

8. Comparison to density ridges. Another concept related to modal regression estimation is that of density ridge estimation. Relative to mixture regression, the literature on density ridges is sparse; see Eberly (1996); Genovese et al. (2012); Chen et al. (2014c,b).

For simplicity of comparison, assume that the predictor X is univariate ($d = 1$). Let $v_1(x,y), v_2(x,y)$ be the eigenvectors corresponding to the eigenvalues $\lambda_1(x,y) \geq \lambda_2(x,y)$ of $H(x,y) = \nabla^2 p(x,y)$, the Hessian matrix of density function p at (x,y). Each point in the ridge set at x is the local mode of the the local mode of subspace spanned by $v_2(x,y)$ with $\lambda_2(x,y) < 0$. We can express this as

$$R(x) = \{y : v_2(x,y)^T \nabla p(x,y) = 0,\, v_2^T(x,y) H(x,y) v_2(x,y) < 0\}.$$

Note that we can similarly express the modal set at x as

$$M(x) = \{y : 1_Y^T \nabla p(x,y) = 0,\, 1_Y^T H(x,y) 1_Y < 0\},$$

where $1_Y^T = (0,1)$ is the unit vector in the y direction. As can be seen easily, the key difference lies in the two vectors 1_Y and $v_2(x,y)$. Every point on the density ridge is local mode with respect to a different subspace, while every point on the modal regression is the local mode with respect to the same subspace, namely, that aligned with the y-axis. The following simple lemma describes cases in which these two sets coincide.

LEMMA 11 (Equivalence of modal and ridge sets). *Assume that $d = 1$, fix any point x, and let $y \in M(x)$. Then provided that*

1. $p_x(x,y) = 0$, *or*
2. $p_{xy}(x,y) = 0$,

it also holds that $y \in R(x)$.

The lemma asserts that a conditional mode where the density is locally stationary, i.e., $p_x(x,y) = 0$, or the density is locally isotropic, i.e., $p_{xy}(x,y) = 0$, is also a density ridge. More explicitly, the first condition states that saddle points and local maximums are both local modes and ridge points, and the second condition states that when modal manifolds moving along the x-axis, they are also density ridges.

We compare modal regression, density ridges, and density modes in Figure 8. Both the estimated density ridges and modal manifolds pass through the density modes, as predicted by Lemma 11. Furthermore, at places in which the joint density is locally isotropic (i.e., spherical), the modal regression and density ridge components roughly coincide.

From a general perspective, modal regression and density ridges are looking for different types of structures; modal regression examines the conditional structure of $Y|X$, and density ridges seek out the joint structure of X, Y. Typically, density ridge estimation is less stable than modal regression estimation because in the former, both the modes and the subspace of interest (the second eigenvector $v_2(x,y)$ of the local Hessian) must be estimated.

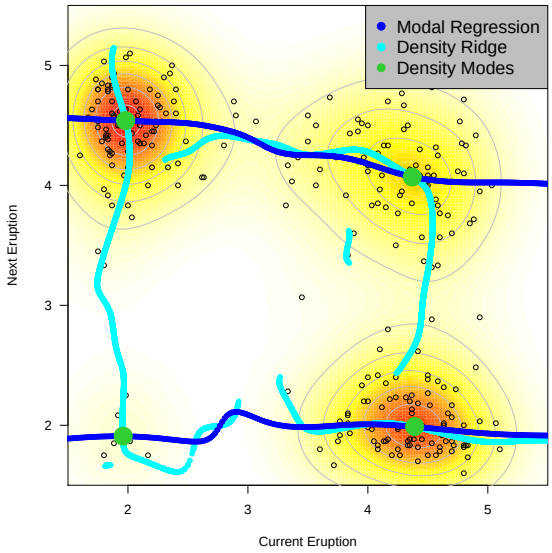


FIG 8. *A comparison between modal regression, density ridges and density modes using the old faithful data set. The background color represents the joint density (red: high density).*

9. Discussion. We have investigated 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. Finally, we compared the proposed method to the well-studied mixture of regression model, and the less well-known but also highly relevant problem of density ridge estimation. The main message is that nonparametric modal regression offers a relatively simple and useable tool to capture conditional structure missed by conventional regression methods. The advances we have developed in this paper, such those for constructing confidence sets and prediction sets, only add to its usefulness as a practical tool.

Though the discussion in this paper treated the dimension d of the predictor variable X as arbitrary, all examples used $d = 1$. We finish by giving two simple examples for $d = 2$. In the first example, the data points are normally distributed around two parabolic surfaces; in the second example, the data points come from five different components of two-dimensional structure. We apply both modal regression (in blue) and local regression (in green) to the two examples, shown in Figure 9. The estimated modal regression set identifies the appropriate structure, while local regression does not (most of the local regression surface does not lie near any of the data points at all).

Acknowledgements. YC is supported by DOE Grant DE-FOA-0000918. CG is supported in part by DOE Grant DE-FOA-0000918 and NSF Grant DMS-1208354. RT is supported by NSF Grant DMS-1309174. LW is supported by NSF Grant DMS-1208354.

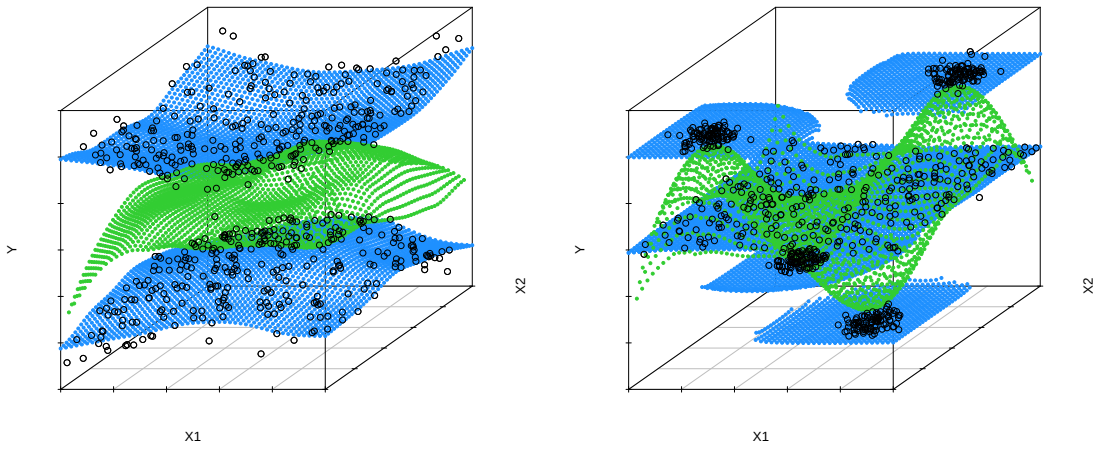


FIG 9. *Two examples for $d = 2$. Modal regression estimates are shown in blue, and local regression in green.*

APPENDIX A: PROOFS

Before we proceed, we first recall a useful theorem.

THEOREM 12. *Assume (A1,K1-3). Then*

$$\|\widehat{p}_n - p\|^*_{\infty,k} = O(h^2) + O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+1+2k}}}\right).$$

Moreover, when n is sufficiently large and $\frac{\log n}{nh^{d+1+2k}} \to 0$,

$$\mathbb{P}\left(\|\widehat{p}_n - p\|^*_{\infty,k} > \epsilon\right) \leq (k+1)e^{Anh^{d+1+2k}}$$

for some constant $A > 0$.

The first assertion can be proved by the same method in Einmahl and Mason (2000, 2005); Gine and Guillou (2002) and the second assertion is an application of the Talagrand's inequality (Talagrand, 1996). Thus we omit the proof. Similar results for the kernel density estimator can be found in Chen et al. (2014c).

PROOF OF THEOREM 3. In this proof we will write denote elements of $M(x)$ as y_j. It can be shown that when $\|\widehat{p}_n - p\|^*_{\infty,2}$ is sufficiently small, for every x, each the local mode $y_j \in M(x)$ corresponds to an unique close estimated local mode $\widehat{y}_j$ by assumption (A3). See the proof to Theorem 4 in Chen et al. (2014b).

Part 1: Empirical Approximation. Let x be a fixed point in D. Let y_j be a local mode and $\widehat{y}_j$ be the estimator to y_j. By definition,

$$p_y(x, y_j) = 0, \qquad \widehat{p}_{y,n}(x, \widehat{y}_j) = 0.$$

By Taylor Theorem,

$$\begin{aligned}\widehat{p}_{y,n}(x,y_j) &= \widehat{p}_{y,n}(x,y_j) - \widehat{p}_{y,n}(x,\widehat{y}_j)\\ &= (y_j-\widehat{y}_j)\widehat{p}_{yy,n}(x,y_j^*),\end{aligned} \tag{20}$$

where y_j^* is a point between y_j and $\widehat{y}_j$.

Thus, after dividing $\widehat{p}_{yy,n}(x,y_j^*)$ in both sides,

$$\begin{aligned}\widehat{y}_j - y_j &= -\widehat{p}_{yy,n}(x,y_j^*)^{-1}\widehat{p}_{y,n}(x,y_j)\\ &= -p_{yy}(x,y_j)^{-1}\widehat{p}_{y,n}(x,y_j) + O(\|\widehat{p}-p\|_{\infty,2}^*)\widehat{p}_{y,n}(x,y_j).\end{aligned} \tag{21}$$

Note that we use

$$\left|\widehat{p}_{yy,n}(x,y_j^*)^{-1} - p_{yy}(x,y_j)^{-1}\right| = O(\|\widehat{p}-p\|_{\infty,2}^*).$$

This is valid since both $p_{yy}, \widehat{p}_{yy,n}$ are bounded away from 0 when x, y is closed to S by assumption (A3). Thus, the inverse is bounded above by (A1) and (K1)

Therefore, by taking absolute values we obtain

$$|\widehat{y}_j - y_j| - |p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)| = O(\|\widehat{p}-p\|_{\infty,2}^* \times |\widehat{p}_{y,n}(x,y_j)|). \tag{22}$$

Now taking max for all local modes and use the fact that $\Delta_n(x) = \max|\widehat{y}_j - y_j|$,

$$\begin{aligned}\Delta_n(x) &- \max_j \left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}\\ &= O\left(\|\widehat{p}-p\|_{\infty,2}^* \times \max_j\{|\widehat{p}_{y,n}(x,y_j)|\}\right).\end{aligned} \tag{23}$$

This implies

$$\max_j\{|\widehat{p}_{y,n}(x,y_j)|\}^{-1}\left|\Delta_n(x) - \max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}\right| = O\left(\|\widehat{p}-p\|_{\infty,2}^*\right).$$

Thus, $\Delta_n(x)$ can be approximated by $\max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)\right\}$.

Note that $|p_{yy}(x,y_j)^{-1}|$ is bounded from the above and below by assumption (A1-3). This shows that $\max_j\{|\widehat{p}_{y,n}(x,y_j)|\}$ is at the same rate as $\max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}$. Thus, equation (23) implies

$$\frac{1}{\Delta_n(x)}\left|\Delta_n(x) - \max_j\left\{|p_{yy}(x,y_j)^{-1}||\widehat{p}_{y,n}(x,y_j)|\right\}\right| = O\left(\|\widehat{p}-p\|_{\infty,2}^*\right)$$

which proves the first assertion.

Part 2: Rate of Convergence. For each j, we focus on $\widehat{p}_{y,n}(x,y_j)$ since $p_{yy}(x,y_j)^{-1}$ is bounded:

$$\begin{aligned}|\widehat{p}_{y,n}(x,y_j)| &= |\widehat{p}_{y,n}(x,y_j) - p_y(x,y_j)|\\ &\le |\widehat{p}_{y,n}(x,y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x,y_j)\right)| + |\mathbb{E}\left(\widehat{p}_{y,n}(x,y_j)\right) - p_y(x,y_j)|\\ &= O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^{d+3}}}\right) + O(h^2).\end{aligned}$$

The rate follows from the bias-variance tradeoff theory for the kernel density estimator. By repeating the above argument for each mode, the rate works for every local mode. Since there are at most $K < \infty$ local modes for fixed x, the rate is the same as we take the maximum over all local modes. Thus, we have proved the second assertion.

□

PROOF OF THEOREM 4. By Theorem 3,

$$\begin{aligned}\Delta_n(x) &= \max_j \left\{ |p_{yy}(x, y_j)^{-1}| |\widehat{p}_{y,n}(x, y_j)| \right\} + o_P(1) \\ &= \max_j \left\{ |p_{yy}(x, y_j)^{-1}| \left(|\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)| + B(x, y_j) \right) \right\} + o_P(1),\end{aligned}$$

where $B(x, y_j) = |\mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right) - p_y(x, y_j)| = O(h^2)$ is the bias and the $o_P(1)$ is from $O(\|\widehat{p} - p\|^*_{\infty,2} \Delta_n(x))$.

Since $|p_{yy}(x, y_j)^{-1}|$ is bounded, the above implies

$$\Delta_n(x) = \max_j \left\{ |p_{yy}(x, y_j)^{-1}| \, |\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)| \right\} + O(h^2) + o_P(1).$$

Note the big O term involves the bias and is independent of x. Thus, taking supremum over $x \in D$ yields

$$\Delta_n = \mathbf{Z} + O(h^2) + o_P(1), \tag{24}$$

where

$$\mathbf{Z} = \sup_{x \in D} \max_j \left\{ |p_{yy}(x, y_j)^{-1}| \, |\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)| \right\}$$

is the maximum over a stochastic process.

Now we show that $\mathbf{Z}$ is the maximum of an empirical process. Let

$$\begin{aligned}\mathcal{F}_0 = \Bigg\{ &(u, v) \mapsto f_{x,y}(u, v) : f_{x,y}(u, v) = p_{yy}^{-1}(x, y) \times \\ &K\left(\frac{\|x - u\|}{h} \right) K^{(1)}\left(\frac{y - v}{h} \right), y \in M(x), x \in \mathbb{R} \Bigg\}.\end{aligned} \tag{25}$$

be a functional space similar to the one defined in (16). We define the empirical process $\mathbb{G}_n$ to be

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n f(Z_i) - \mathbb{E}(f(Z_i)) \right) \qquad f \in \mathcal{F}_0, \tag{26}$$

where $Z_i = (X_i, Y_i)$ is the observed data.

Thus,

$$\begin{aligned}\mathbf{Z} &= \sup_{x \in D} \max_j \left\{ |p_{yy}(x, y_j)^{-1}| \, |\widehat{p}_{y,n}(x, y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x, y_j)\right)| \right\} \\ &= \frac{1}{\sqrt{n} h^{d+3}} \sup_{f \in \mathcal{F}_0} |\mathbb{G}_n(f)|.\end{aligned} \tag{27}$$

By assumption (A1) and (K1–2), $\mathcal{F}_0$ is a VC-type class with constant envelope C_K^2/λ_2. Thus, applying Theorem 2.3 in Gine and Guillou (2002) gives

$$\mathbf{Z} = \sup_{x\in D} \max_j \left\{ |p_{yy}(x,y_j)^{-1}| \, |\widehat{p}_{y,n}(x,y_j) - \mathbb{E}\left(\widehat{p}_{y,n}(x,y_j)\right)| \right\} | = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+3}}}\right).$$

Now by equation (24), we conclude the result.

□

Proof of Theorem 5. The proof of this theorem is obtained by applying to Theorem 3 so that the expected square of local error can be written as

$$\mathbb{E}\left(\Delta_n^2(x)\right) = O(h^4) + O\left(\frac{1}{nh^{d+3}}\right) = \mathsf{Bias}^2(x) + \mathsf{Variance}(x).$$

Then use the same arguments in Chacón et al. (2011); Chacón and Duong (2013) to show that integrating bias and variance over x still yield the same rate of convergence. We omit the details of this proof. □

Proof of Theorem 7. We prove this Theorem by the similar technique in the proof of Theorem 6 of Chen et al. (2014c).

Let $\mathcal{F}$ be the functional space defined in (16). We define $\mathbb{G}_n$ be an empirical process on $\mathcal{F}$ and also define $\mathbb{B}$ to be a Gaussian process on $\mathcal{F}$. Denote $\mathbf{G}_n = \frac{1}{\sqrt{h^{d+3}}} \sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)|$ and $\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$.

Our proof consists of three steps.

1. Coupling between $\sqrt{nh^{d+3}}\Delta_n$ and $\mathbf{G}_n$.
2. Coupling between $\mathbf{G}_n$ and $\mathbf{B}$.
3. Anti-concentration Chernozhukov et al. (2013a,c) to convert the coupling into the desire Berry-Esseen bound.

Step 1. Our goal is to show

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n\right| > \epsilon\right) \leq D_1 e^{-D_2 nh^{d+5}\epsilon^2} \tag{28}$$

for some constants D_1, D_2.

Recall Corollary 6,

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

Thus, there exists a constant $D_0 > 0$ such that

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

By Talagrand's inequality (Theorem A.4 in Chernozhukov et al. (2013a); see also Talagrand (1996), Massart (2000) and Gine and Guillou (2002)),

$$\begin{aligned}
&\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n\right| > \epsilon\right) \\
&\qquad \le \mathbb{P}\left(\sup_{x,y} |\widehat{p}_{yy,n}(x,y) - \mathbb{E}\left(\widehat{p}_{yy,n}(x,y)\right)| > \epsilon/D_0\right) \\
&\qquad \le D_1 e^{-D_2 n h^{d+5}\epsilon^2}
\end{aligned} \tag{29}$$

for some constants $D_1, D_2 > 0$. This gives the desire result.

Step 2. We will show

$$\mathbb{P}\left(|\mathbf{G}_n - \mathbf{B}| > A_1 \frac{b_0 \log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}\right) \le A_2\gamma. \tag{30}$$

for some constants A_1, A_2.

We first recall a useful Theorem in Chernozhukov et al. (2013a):

THEOREM 13 (Theorem 3.1 in Chernozhukov et al. (2013a)). *Let $\mathcal{G}$ be a collection of functions that is a VC-type class (see condition (K2)) with a constant envelope function b. Let σ^2 be a constant such that $\sup_{g\in\mathcal{G}} \mathbb{E}[g(X_i)^2] \le \sigma^2 \le b^2$. Let $\mathbb{B}$ be a centered, tight Gaussian process defined on $\mathcal{G}$ with covariance function*

$$\mathrm{Cov}(\mathbb{B}(g_1), \mathbb{B}(g_2)) = \mathbb{E}[g_1(X_i)g_2(X_i)] - \mathbb{E}[g_1(X_i)]\mathbb{E}[g_2(X_i)] \tag{31}$$

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

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{G}} |\mathbb{G}_n(g)| - \mathbf{B}\right| > A_1 \frac{b^{1/3}\sigma^{2/3}\log^{2/3} n}{\gamma^{1/3} n^{1/6}}\right) \le A_2\gamma \tag{32}$$

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

To apply Theorem 13, we need to verify conditions. By assumption (K3) and (A2), $\mathcal{F}$ is a VC-type class with constant envelope $b_0 = C_K^2\widetilde{\lambda}_2 < \infty$. 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 σ^2. By definition,

$$\sup_{f\in\mathcal{F}} \mathbb{E}[f(X_i)^2] \le h^{d+3} b_0^2.$$

Thus, we can pick $\sigma^2 = h^{d+3}b_0^2 \le b_0^2$ if $h \le 1$. Hence, applying Theorem 13 gives

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| - \mathbf{B}'\right| > A_1 \frac{b_0 h^{2/3} h^2 \log^{2/3} n}{\gamma^{1/3} n^{1/6}}\right) \le A_2\gamma \tag{33}$$

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

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 $\frac{1}{\sqrt{h^{d+3}}}\mathbf{B}' = \mathbf{B}$,

$$\mathbb{P}\left(|\mathbf{G}_n - \mathbf{B}| > A_1 \frac{b_0 \log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}\right) \leq A_2\gamma \tag{34}$$

which is the desire result (30).

Step 3. We first show the coupling between $\sqrt{nh^{d+3}}\Delta_n$ and $\mathbf{B}$. We pick $\epsilon = (nh^{d+5})^{-1/4}$ in (28) so that

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\Delta_n - \mathbf{G}_n\right| > (nh^{d+5})^{-1/4}\right) \leq D_1 e^{-D_2\sqrt{nh^{d+5}}}. \tag{35}$$

As n is sufficiently large, and by triangular inequality along with (30),

$$\mathbb{P}\left(\left|\sqrt{nh^{d+3}}\Delta_n - \mathbf{B}\right| > A_3 \frac{\log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}\right) \leq A_4\gamma, \tag{36}$$

for some constants $A_3, A_4 > 0$. Note that we absorb the rate $(nh^{d+5})^{-1/4}$ in (35) into $A_3\frac{\log^{2/3} n}{\gamma^{1/3}(nh^{d+3})^{1/6}}$. This is valid since $(nh^{d+5})^{-1/4}$ 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 γ converges at rate slower than $(nh^{d+5})^{-1/4}$.

Now applying the anti-concentration inequality (version of Lemma 16 in Chen et al. (2014c); see also Corollary 2.1 in Chernozhukov et al. (2013a) and Chernozhukov et al. (2013b,c)), we conclude

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

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

□

Proof of Theorem 8. The proof to this Theorem is essentially the same as proof to Theorem 7 in Chen et al. (2014c) by using the Theorem 7 of the current paper. We state the basic ideas in the following and omit the details. Note that the functional space (16) depends on the probability measure $\mathbb{P}$ and smoothing parameter h

$$\begin{aligned}\mathcal{F} = \mathcal{F}(\mathbb{P}, h) = \Bigg\{&(u,v) \mapsto f_{x,y}(u,v) : f_{x,y}(u,v) = \widetilde{p}_{yy}^{-1}(x,y) \times \\ &K\left(\frac{\|x-u\|}{h}\right) K^{(1)}\left(\frac{y-v}{h}\right), x \in D, y \in \widetilde{M}(x)\Bigg\}\end{aligned} \tag{37}$$

since the index y is defined on the smoothed local mode $\widetilde{M}(x)$ and it requires second derivatives of smooth density $\widetilde{p}(x,y)$. Both $\widetilde{M}(x)$ and $\widetilde{p}(x,y)$ are completely detemined by $\mathbb{P}$ and h. For

the bootstrap estimate, Theorem 7 implies that $\widehat{\Delta}_n^*$ can be approximated by the maximal of a certain Gaussian process

$$\sup_{f\in\mathcal{F}(\mathbb{P}_n,h)} |\mathbb{B}(f)|. \tag{38}$$

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 ($\mathcal{D}$ 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 $\widehat{M}_n(x)$ and $\widetilde{p}_{yy}(x,y)$ will be replaced by the second derivative of KDE $\widehat{p}_n(x,y)$. 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 $\mathcal{F}(\mathbb{P},h)$ and $\mathcal{F}(\mathbb{P}_n,h)$ will be asymptotically the same by Lemma 17, 19 and 20 in Chen et al. (2014c). Putting altogether, the result follows from the approximation

$$\widehat{\Delta}_n^* \approx \sup_{f\in\mathcal{F}(\mathbb{P}_n,h)} |\mathbb{B}(f)| \approx \sup_{f\in\mathcal{F}(\mathbb{P},h)} |\mathbb{B}(f)| \approx \Delta_n. \tag{39}$$

□

Before we prove Theorem 10, we first prove the following useful lemma on the Gaussian mixture and its corresponding local modes.

LEMMA 14 (Gaussian Mixture and Local Modes). *Consider a Gaussian mixture* $p(y) = \sum_{j=1}^K \pi_j \phi(y;\mu_j,\sigma_j^2)$ *with* $\mu_1 < \ldots < \mu_K$ *and* $y \in \mathbb{R}$. *Let* $W = \frac{\Delta_{\min}}{\sigma_{\max}}$ *and* $\Delta_{\min} = \min\{|\mu_j - \mu_i| : i \neq j\}$ *and* $\sigma_{\max} = \max_j \sigma_j$. *If*

$$W \geq \sqrt{2\log\left(4(K\vee 3-1)\frac{\pi_{\max}}{\pi_{\min}}\right)},$$

then

$$\max_j |\mu_j - m_j| \leq \sigma_{\max} \times 4\frac{\pi_{\max}}{\pi_{\min}}\frac{1}{W}e^{-\frac{W^2}{2}}.$$

PROOF. Given any $(\pi_j,\mu_j,\sigma_j^2 : j = 1,\ldots,K)$, we consider another mixture (but not necessarily a density)

$$h(y) = \pi_{\min}\phi(y;\mu_1,\sigma_{\max}^2) + \sum_{j=2}^K \phi(y;\mu_1 + (j-1)\Delta_{\min},\sigma_{\max}^2).$$

We assume

(MK) $h(y)$ has K distinct local modes.

Note that this implies $p(y)$ to have K distinct local modes. Later we will prove this condition. Let the ordered local modes of $h(y)$ be $m_1' < \ldots < m_K'$. It is not hard to observe that

$$|m_1' - \mu_1| \geq \max_j |m_j - \mu_j| \tag{40}$$

since all mixture components in h other than the first one are pulling the m_1' away from μ_1. One can think of $h(y)$ as the worst case scenario (the layout of mixture pulling m_j away from μ_j) of the parameters $(\pi_j, \mu_j, \sigma_j^2 : j = 1, \ldots, K)$.

Because $\mu_1 < \mu_j$, the local mode $m_1' > \mu_1$ and $h(m_1') > h(\mu_1)$. We define s_1 such that

$$h(\mu_1 + s_1) = h(\mu_1), \quad h(s) \geq h(\mu_1) \quad \forall s \in [\mu_1, \mu_1 + s_1].$$

It is easy to see that $m_1' \leq s_1 + \mu_1$ since m_1' is the smallest (in terms of location) local mode of h. Thus, if we can bound s_1, we bound the difference $|m_1' - \mu_1|$. Note that s_1 must be very small (at least smaller than $\sigma_{\max}$) otherwise we will not obtain K local modes.

Now by definition of h, we attempts to find s_1 such that

$$\begin{aligned} h(\mu_1) &= \pi_{\min}\phi(\mu_1; \mu_1, \sigma_{\max}^2) + \pi_{\max}\sum_{j=2}^{K}\phi(\mu_1; \mu_1 + (j-1)\Delta_{\min}, \sigma_{\max}^2) \\ &= \frac{1}{\sqrt{2\pi\sigma_{\max}^2}}\pi_{\min} + \frac{1}{\sqrt{2\pi\sigma_{\max}^2}}\pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}}{\sigma_{\max}})^2} \\ &= h(\mu_1 + s_1) \\ &= \frac{1}{\sqrt{2\pi\sigma_{\max}^2}}\pi_{\min}e^{-\frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2} + \frac{1}{\sqrt{2\pi\sigma_{\max}^2}}\pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min} - s_1}{\sigma_{\max}})^2}. \end{aligned} \tag{41}$$

Therefore, s_1 can be obtained by solving

$$\pi_{\min}\left(1 - e^{-\frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2}\right) = \pi_{\max}\sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}}{\sigma_{\max}})^2}(e^{\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}} - \frac{s_1^2}{2\sigma_{\max}^2}} - 1) \tag{42}$$

Note that $e^x < 1 + 2x$ if $x < 1$. Thus, when

$$\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}} < 1, \tag{43}$$

we have

$$e^{\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}} - \frac{s_1^2}{2\sigma_{\max}^2}} - 1 < 2\frac{(j-1)\Delta_{\min}}{\sigma_{\max}}\frac{s_1}{\sigma_{\max}} = 2(j-1)W\frac{s_1}{\sigma_{\max}}, \tag{44}$$

where $W = \frac{\Delta_{\min}}{\sigma_{\max}}$. Also note that

$$1 - e^{-\frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2} > \frac{1}{2}(\frac{s_1}{\sigma_{\max}})^2 \tag{45}$$

since $s_1 < \sigma_{\max}$.

Let s_2 be a small number satisfying

$$\begin{aligned}\frac{1}{2}(\frac{s_2}{\sigma_{\max}})^2 &= 2\frac{\pi_{\max}}{\pi_{\min}} W \frac{s_2}{\sigma_{\max}} \int_1^\infty x e^{-\frac{W^2}{2}x^2} dx \\ &= \frac{s_2}{\sigma_{\max}} \frac{\pi_{\max}}{\pi_{\min}} \frac{2}{W} e^{-\frac{W^2}{2}} \\ &\geq W \frac{\pi_{\max}}{\pi_{\min}} \frac{s_2}{\sigma_{\max}} \sum_{j=1}^{K} e^{-\frac{1}{2}j^2W^2} j^2 \\ &= \pi_{\max} \sum_{j=2}^{K} e^{-\frac{1}{2}(\frac{(j-1)\Delta_{\min}}{\sigma_{\max}})^2} 2(j-1) W s_2 \frac{\pi_{\max}}{\pi_{\min}}.\end{aligned} \tag{46}$$

By (42), (44) and (45), $s_2 > s_1$. The above result gives

$$s_2 = \sigma_{\max} \times \frac{\pi_{\max}}{\pi_{\min}} \frac{4}{W} e^{-W^2/2} > s_1 \geq \max_j |m_j - \mu_j| \tag{47}$$

which is the desire result.

Note that the above method requires (43), which requires

$$\frac{1}{K-1}\frac{1}{W} > \frac{s_2}{\sigma_{\max}} = \frac{\pi_{\max}}{\pi_{\min}} \frac{4}{W} e^{-W^2/2}. \tag{48}$$

This is true whenever

$$W > \sqrt{2 \log \left(4(K-1) \frac{\pi_{\max}}{\pi_{\min}} \right)} \tag{49}$$

which gives part of the condition in this Lemma.

Recall that we assume (MK) at the beginning. We prove that as W is sufficiently large, (MK) holds. It is easy to see that

$$\begin{aligned} &|\mu_i - \mu_j| > \Delta_{\min} \\ \Rightarrow\ &|m_i - m_j| > \Delta_{\min} - 2 \max_i |m_i - \mu_i|. \end{aligned} \tag{50}$$

Thus, as long as $\Delta_{\min} - 2\max_i |m_i - \mu_i| > 0$, there exists K distinct local modes for $p(y)$.

By equation (47), a sufficient condition to $\Delta_{\min} - 2\max_i |m_i - \mu_i| > 0$ is

$$\Delta_{\min} > 2\sigma_{\max} \times \frac{\pi_{\max}}{\pi_{\min}} \frac{4}{W} e^{-W^2/2} \tag{51}$$

which is equivalent to

$$W^2 e^{W^2/2} > 8 \frac{\pi_{\max}}{\pi_{\min}}.$$

As $W > 1$ (which is satisfied by (52)),

$$e^{W^2/2} > 8 \frac{\pi_{\max}}{\pi_{\min}}$$

implies (51) so that a sufficient condition for $p(y)$ having K distinct local modes is

$$W > \sqrt{2\log\left(8\frac{\pi_{\max}}{\pi_{\min}}\right)}. \tag{52}$$

Combining this condition and equation (49), we conclude the result.

□

Proof of Theorem 10. The proof consists of four steps. At first three steps, we consider the pointwise prediction sets and the last step is to extend the proof to the unifrom prediction sets. We summarize the four steps as follows:

1. We prove
$$\epsilon_{1-\alpha}(x) \leq z_{1-\alpha/2}\sigma_{\max}(x) + \max_i |u_i(x) - m_i(x)|,$$
where $m_1(x) < m_2(x) < \ldots < m_{K(x)}(x)$ are the ordered local modes.
2. We prove $\eta_{1-\alpha}(x) \geq \frac{1}{2}K(x)\Delta_{\min}(x)$.
3. We apply Lemma 14 to bound $\max_i |u_i(x) - m_i(x)|$ by $\Delta_{\min}(x)$ and use the first two steps to conclude the desire result.
4. We extend the first three steps to uniform case.

Step 1. By assumption (GP), the set

$$A = \bigcup_{j=1}^{K(x)} \mu_j(x) \oplus \left(z_{1-\alpha/2}\sigma_j(x)\right)$$

is a $(1-\alpha)\%$ prediction set. Let $m_1(x) < m_2(x) < \ldots < m_{K(x)}(x)$ be the ordered local modes of $p(y|x)$. Then we have

$$\begin{aligned} \mu_j(x) \oplus \left(z_{1-\alpha/2}\sigma_j(x)\right) &\subseteq m_j(x) \oplus \left(z_{1-\alpha/2}\sigma_j(x) + |\mu_j(x) - m_j(x)|\right) \\ &\subseteq m_j(x) \oplus \left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|\right). \end{aligned} \tag{53}$$

This works for all j. Note that the regression mode set $M(x) = \{m_1(x), \ldots, m_{K(x)}(x)\}$ so that

$$A \subseteq M(x) \oplus \left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|\right)$$

which implies

$$\epsilon_{1-\alpha}(x) \leq z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)| \tag{54}$$

since $\epsilon_{1-\alpha}(x)$ is the smallest size to construct a pointwise prediction set with with $1-\alpha$ prediction accuracy.

Step 2. Since we pick α such that $\alpha < \pi_1(x), \pi_{K(x)}(x)$. The prediction set from regression function must contains all the mixture centers. Thus,

$$2\eta_{1-\alpha}(x) \geq \mu_{K(x)}(x) - \mu_1(x) \geq (K(x)-1)\Delta_{min}(x).$$

Step 3. The length of prediction set $\mathcal{P}_{1-\alpha} = M(x) \oplus \epsilon_{1-\alpha}(x)$ is $2K(x)\epsilon_{1-\alpha}(x)$ and the length of prediction set $\mathcal{R}_{1-\alpha} = m(x) \oplus \eta_{1-\alpha}(x)$ is $2\eta_{1-\alpha}(x)$. Thus, we need to show

$$\eta_{1-\alpha}(x) > K(x)\epsilon_{1-\alpha}(x). \tag{55}$$

By (54) and Step 2, a sufficient condition for (55) is

$$(K(x)-1)\Delta_{min}(x) > K(x)\left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)|\right).$$

Applying Lemma 14 yields

$$\max_j |\mu_j(x) - m_j(x)| \le \sigma_{\max}(x) \times 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\frac{1}{W(x)}e^{-\frac{W(x)^2}{2}} \tag{56}$$

whenever

$$W(x) \ge \sqrt{2\log\left(4(K(x)\vee 3 - 1)\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}, \tag{57}$$

where $W(x) = \frac{\Delta_{min}(x)}{\sigma_{\max}(x)}$.

Now assume $\alpha < 0.1$, the z-score $z_{1-\alpha/2} > 1.64$. We bound $\max_j |\mu_j(x) - m_j(x)|$ by $0.1 \times z_{1-\alpha/2}\sigma_{\max}(x)$ so that we can have a reference rule only depends on z-score. Thus, we can use (56):

$$\begin{aligned} 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\frac{1}{W(x)}e^{-\frac{W(x)^2}{2}} &\le 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)}e^{-\frac{W(x)^2}{2}} \\ &\le 0.1 \times z_{1-\alpha/2} \\ &< 0.1 \times 1.64 \end{aligned} \tag{58}$$

Thus, we need

$$\begin{aligned} W(x) &> \sqrt{2\log\left(\frac{40}{1.64}\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)} \\ &= \sqrt{6.4 + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}. \end{aligned} \tag{59}$$

Hence, as $W(x) > \sqrt{6.4 + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}$, we need

$$(K(x)-1)\Delta_{min}(x) > 1.1 \times K(x)z_{1-\alpha/2}\sigma_{\max}(x).$$

This requires

$$W(x) > 1.1 \times \frac{K(x)}{K(x)-1}z_{1-\alpha/2}. \tag{60}$$

Now the conditions on $W(x) = \frac{\Delta_{\min}(x)}{\sigma_{\max}(x)}$ involve equations (57),(59) and (60). We conclude that whenever

$$W(x) = \frac{\Delta_{\min}(x)}{\sigma_{\max}(x)} > \max\left\{1.1 \times \frac{K(x)}{K(x)-1} z_{1-\alpha/2}, \sqrt{6.4 \vee 2\log\left(4(K(x) \vee 3 - 1)\right) + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}\right\},$$

the prediction set $\mathcal{P}_{1-\alpha}(x)$ is smaller than $\mathcal{R}_{1-\alpha}(x)$.

Step 4. Now we extend to the uniform case. Note that

$$\begin{aligned} \epsilon_{1-\alpha} &\leq \sup_x \epsilon_{1-\alpha}(x) \\ \eta_{1-\alpha} &\geq \inf_x \eta_{1-\alpha}(x). \end{aligned} \tag{61}$$

Therefore,

$$\begin{aligned} \epsilon_{1-\alpha} &\leq \sup_x \epsilon_{1-\alpha}(x) \\ &\leq \sup_x \left(z_{1-\alpha/2}\sigma_{\max}(x) + \max_j |\mu_j(x) - m_j(x)| \right) \\ &\leq z_{1-\alpha/2}\sigma_{\max} + \sup_x \max_j |\mu_j(x) - m_j(x)| \end{aligned} \tag{62}$$

and similarly

$$\begin{aligned} \eta_{1-\alpha} &\geq \inf_x \eta_{1-\alpha}(x) \\ &\geq \inf_x (K(x)-1)\Delta_{\min}(x) \\ &\geq (K_{\min}-1)\Delta_{\min}. \end{aligned} \tag{63}$$

Now note that the second term in the last inequality of (62) can be bounded by

$$\begin{aligned} \sup_x \max_j |\mu_j(x) - m_j(x)| &\leq \sup_x \sigma_{\max}(x) \times 4\frac{\pi_{\max}(x)}{\pi_{\min}(x)} \frac{1}{W(x)} e^{-W(x)^2/2} \\ &\leq \sigma_{\max} \times 4\frac{\pi_{\max}}{\pi_{\min}} \frac{1}{W} e^{-W^2/2}, \end{aligned} \tag{64}$$

where $W = \frac{\Delta_{\min}}{\sigma_{\max}} \leq W(x)$.

Now use the same way as (58) and combine (62) and (64),

$$\epsilon_{1-\alpha} \leq 1.1 \times z_{1-\alpha/2}\sigma_{\max} \tag{65}$$

whenever

$$W = \frac{\Delta_{\min}}{\sigma_{\max}} > \max\left\{\sqrt{6.4 \vee 2\log\left(4(K_{\max} \vee 3 - 1)\right) + 2\log\left(\frac{\pi_{\max}}{\pi_{\min}}\right)}\right\}. \tag{66}$$

The volume of prediction sets $\mathcal{P}_{1-\alpha}$ and $\mathcal{R}_{1-\alpha}$ is

$$\mathsf{Vol}(\mathcal{P}_{1-\alpha}) = 2\epsilon_{1-\alpha}\int_D K(x)dx, \qquad \mathsf{Vol}(\mathcal{R}_{1-\alpha}) = 2\eta_{1-\alpha}\int_D dx.$$

Thus, $\mathsf{Vol}(\mathcal{P}_{1-\alpha}) < \mathsf{Vol}(\mathcal{R}_{1-\alpha})$ if and only if

$$\epsilon_{1-\alpha}\overline{K} < \eta_{1-\alpha}. \tag{67}$$

Applying equation (65) and (64) to (67), we require

$$\eta_{1-\alpha} \geq (K_{\min}-1)\Delta_{\min} > \overline{K} \times 1.1 \times z_{1-\alpha/2}\sigma_{\max} \geq \epsilon_{1-\alpha} \tag{68}$$

which leads to

$$\frac{\Delta_{\min}}{\sigma_{\max}} > 1.1 \times \frac{\overline{K}}{K_{\min}-1} z_{1-\alpha/2}.$$

Combining this condition and (66), we complete the proof.

□

Proof for Lemma 11. Let the Hessian matrix of $p(x,y)$ be $H \equiv H(x,y)$. The eigenvalue of a 2×2 matrix has an explicit formula:

$$\begin{aligned}\lambda_1(x,y) &= \mathsf{tr}(H)/2 + \sqrt{\mathsf{tr}(H)^2/2 - \mathsf{det}(H)} \\ \lambda_2(x,y) &= \mathsf{tr}(H)/2 - \sqrt{\mathsf{tr}(H)^2/2 - \mathsf{det}(H)}\end{aligned} \tag{69}$$

and the corresponding eigenvectors are

$$v_1(x,y) = \begin{bmatrix}\lambda_1(x,y) - H_{22} \\ H_{21}\end{bmatrix} \quad v_2(x,y) = \begin{bmatrix}\lambda_2(x,y) - H_{22} \\ H_{21}\end{bmatrix}, \tag{70}$$

where H_{ij} is the (i,j) element of H and $\mathsf{tr}(H)$ is the trace of H and $\mathsf{det}(H)$ is the determinant of H.

Thus, $\lambda_2(x,y) < 0$ if and only if ($\mathsf{tr}(H) < 0$ or $\mathsf{det}(H) < 0$). Namely,

$$\lambda_2(x,y) < 0 \iff \left(H_{11} + H_{22} < 0 \text{ or } H_{11}H_{22} < H_{12}^2\right).$$

However, since $y \in M(x)$, $H_{22} < 0$. This implies $\lambda_2(x,y) < 0$. (since whatever the sign of H_{11} is, one of the above conditions must hold)

Thus, all we need is to show $v_2^T(x,y)\nabla p(x,y) = 0$. By the formula for eigenvectors,

$$\begin{aligned}v_2^T(x,y)\nabla p(x,y) &= (\lambda_2(x,y) - H_{22})p_x(x,y) + H_{21}p_y(x,y) \\ &= (\lambda_2(x,y) - H_{22})p_x(x,y)\end{aligned}$$

since $p_y(x,y) = 0$ for $y \in M(x)$. Therefore, $v_2^T(x,y)\nabla p(x,y) = 0$ if and only if $p_x(x,y) = 0$ or $\lambda_2(x,y) - H_{22} = 0$. The former case corresponds to the first condition and by (69), $\lambda_2(x,y) = H_{22}$ if and only if $H_{12} = 0$. This completes the proof.

□

REFERENCES

E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Unpublished Manuscript*, 2013.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

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

J. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 2011.

A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. *arXiv preprint arXiv:1306.3729*, 2013.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Enhanced mode clustering. arXiv: 1406.1780, 2014a.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Generalized mode and ridge estimation. arXiv: 1406.1803, 2014b.

Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. arXiv: 1406.5663, 2014c.

Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.

V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest adaptive confidence bands. *arXiv:1303.7152*, 2013a.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes. *arXiv:1212.6885*, 2013b.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *arXiv:1301.4807*, 2013c.

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

D. Eberly. *Ridges in Image and Data Analysis*. Springer, 1996.

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

J. Einbeck and G. Tutz. 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, 2006. ISSN 1467-9876. . URL http://dx.doi.org/10.1111/j.1467-9876.2006.00547.x.

U. Einmahl and D. M. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoretical Probability*, 2000.

U. Einmahl and D. M. Mason. Uniform in bandwidth consistency for kernel-type function estimators. *The Annals of Statistics*, 2005.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *arXiv:1212.5156v1*, 2012.

E. Gine and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *In Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 2002.

M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724, 2012. . URL http://dx.doi.org/10.1080/01621459.2012.682541.

M. Huang, R. Li, and S. Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941, 2013. . URL http://dx.doi.org/10.1080/01621459.2013.772897.

D. R. Hunter and D. S. Young. Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24(1):19–38, 2012.

R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald. Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4):315–336, 1996. . URL http://www.tandfonline.com/doi/abs/10.1080/10618600.1996.10474715.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87, Mar. 1991. ISSN 0899-7667. . URL http://dx.doi.org/10.1162/neco.1991.3.1.79.

W. Jiang and M. A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics*, 27(3):987–1011, 06 1999. . URL http://dx.doi.org/10.1214/aos/1018031265.

A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Sta-*

tistical Association, 102(479):1025–1038, 2007. . URL http://dx.doi.org/10.1198/016214507000000590.

M.-j. Lee. Mode regression. *Journal of Econometrics*, 42(3):337–349, November 1989. URL http://ideas.repec.org/a/eee/econom/v42y1989i3p337-349.html.

P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 2000.

A. Rojas. *Nonparametric Mixture Regression*. PhD thesis, Carnegie Mellon University, 2005.

J. P. Romano. On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, pages 629–647, 1988.

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

D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 1992.

M. Talagrand. Newconcentration inequalities in product spaces. *Invent. Math*, 1996.

K. Viele and B. Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12(4):315–330, 2002.

W. Yao and L. Li. A new regression model: Modal linear regression. *Scandinavian Journal of Statistics*, pages n/a–n/a, 2013. ISSN 1467-9469. . URL http://dx.doi.org/10.1111/sjos.12054.

W. Yao, B. G. Lindsay, and R. Li. Local modal regression. *Journal of Nonparametric Statistics*, 24(3):647–663, 2012. . URL http://dx.doi.org/10.1080/10485252.2012.678848.

Department of Statistics
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
E-mail: yenchic@andrew.cmu.edu