NONPARAMETRIC MODAL REGRESSION

BY YEN-CHI CHEN[1], CHRISTOPHER R. GENOVESE[2],
RYAN J. TIBSHIRANI[3] AND LARRY WASSERMAN[4]

Carnegie Mellon University

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 [Lee (1989), Sager and Thisted (1982), Yao and Li (2014), Yao, Lindsay and Li (2012)] 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 unnecessarily 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, Bashtannyk and Grunwald (1996)] and economics [Huang, Li and Wang (2013), Huang and Yao (2012)].
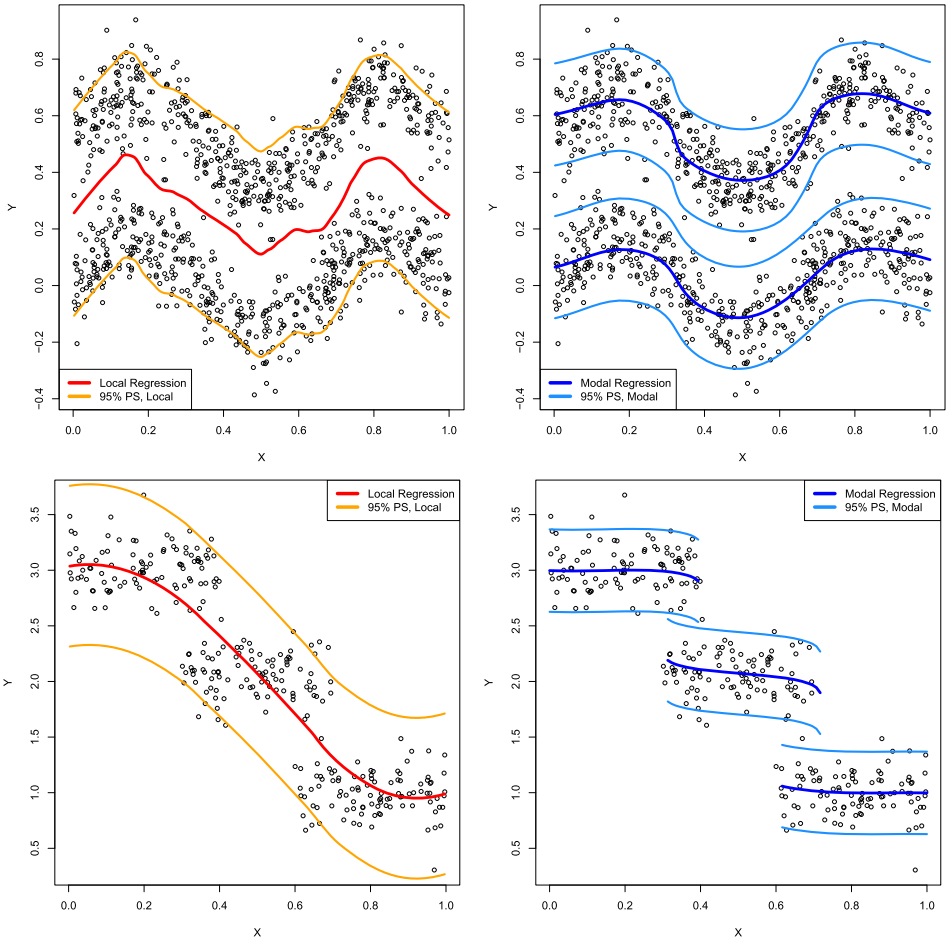
FIG. 1. *Examples of modal regression versus a common nonparametric regression estimator, local linear regression. In the top row, we show local regression estimate and its associated* 95% *prediction bands alongside the modal regression and its* 95% *prediction bands. The bottom row does the same for a different data example. The local regression method fails to capture the structure, and produces prediction bands that are too wide.*

Formally, the conditional (or local) mode set at x is defined as

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

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:

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

At each x, the local mode set $M(x)$ may consist of several points, and so $M(x)$ is in general a multivalued 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):

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

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 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 [Lee (1989), Sager and Thisted (1982), Yao and Li (2014)] 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 (global) mode of Y given $X = x$. Nonparametric modal regression is more flexible, because it allows $M(x)$ to be multivalued, 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, Lindsay and Li (2012) propose an interesting local polynomial smoothing method for mode estimation; however, they focus on the global mode $\mathsf{Mode}(Y|X=x)$ 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

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

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 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

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

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, 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)]. A related approach can be found in Yao (2013), where the authors consider a mode hunting algorithm based on the EM algorithm. For a discussion of how the mean-shift and EM algorithms relate, see Carreira-Perpiñán (2007). In general, 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.

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, Mason and Pelletier (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,

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

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 $\mathcal{M} = \mathcal{D}$).
2. For each $(x, y) \in \mathcal{M}$, fix x, and update y using the following iterations until convergence:

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

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).

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

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

$$(8) \qquad \mathcal{S} = S_1 \cup \cdots \cup S_K,$$

where each S_j is a connected manifold that admits a parametrization

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

for some function $m_j(x)$ and open set A_j. For instance, in Figure 2, each S_j is a connected curve. 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
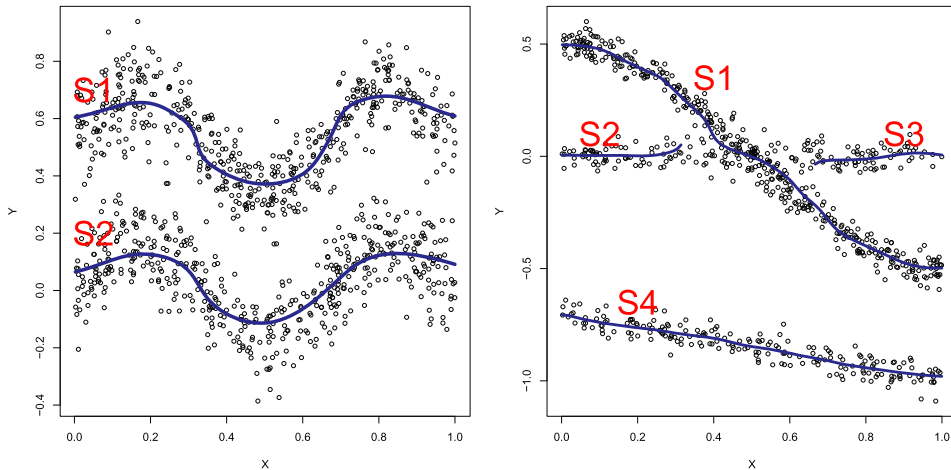

FIG. 2. *Examples of modal manifolds.*

convention, we let $m_j(x) = \varnothing$ if $x \notin A_j$ and, therefore, we may write

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

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 $x \in A_j$,*

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

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,

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

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

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

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*

$$(12) \qquad \lim_{|\varepsilon| \to 0} \frac{\mathsf{Haus}(M(x), M(x + \varepsilon))}{|\varepsilon|} \leq \max_{j=1,\ldots,K} \|m'_j(x)\| \leq \frac{C}{\lambda_2} < \infty.$$

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 multivalued 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 $x = 0.35$,
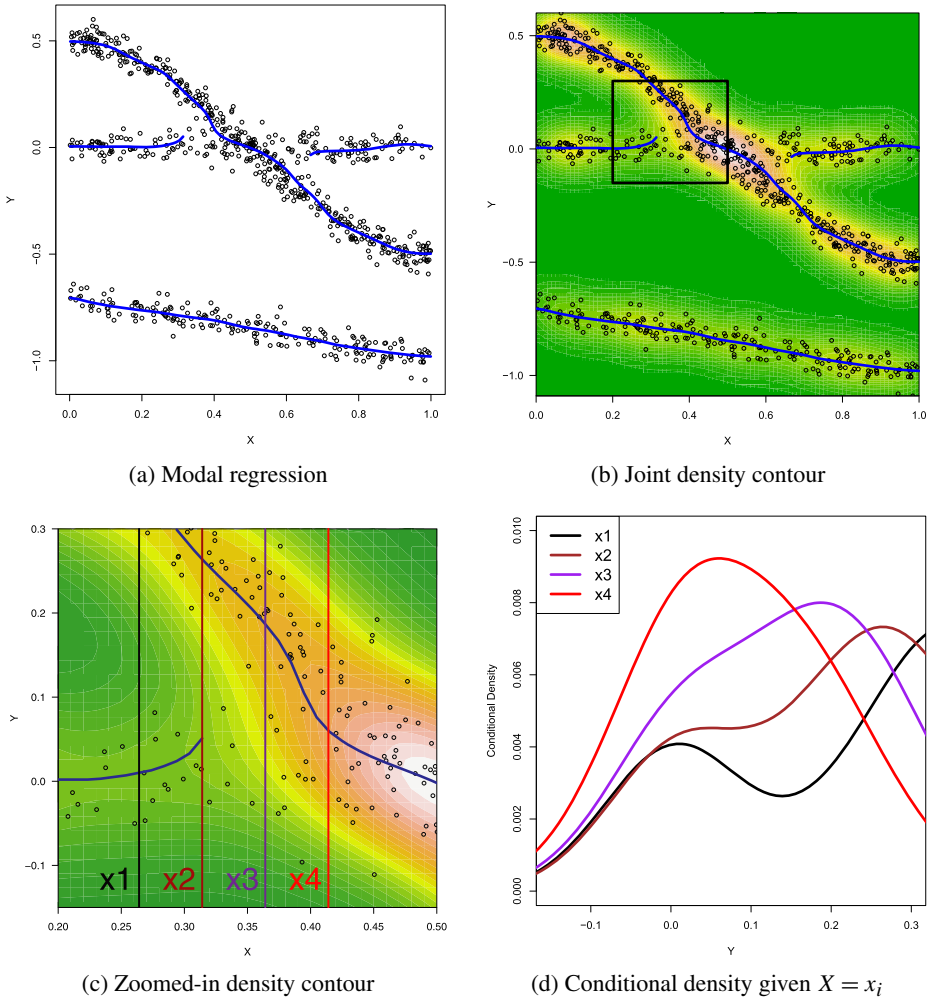

(a) Modal regression

(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 $X = x_2$.*

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.

A remark about the uniqueness of the modal manifold collection $\mathcal{S}$ in (8): this factorization is unique if the second derivative $p_{yy}(x, y)$ is uniformly bounded away from zero. This will later become one of our assumptions [assumption (A3)] in the theoretical analysis of Section 4. Note that in the left panel of Figure 2, the collection $\mathcal{S}$ is uniquely defined, while this is not the case in the right panel (at the points of intersection between curves, the density p has vanishing second derivatives with respect to y).

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

$$(13) \qquad \widehat{\mathcal{S}}_n = \{(x, y) : y \in \widehat{M}_n(x), x \in \mathbb{R}\} = \widehat{S}_1 \cup \cdots \cup \widehat{S}_{\widehat{K}},$$

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

$$(14) \qquad \widehat{M}_n(x) = \{\widehat{m}_1(x), \ldots, \widehat{m}_{\widehat{K}}(x)\}.$$

In practice, determining the manifold memberships and the total number of manifolds $\widehat{K}$ is not 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 αth 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 $\mathcal{S} = S_1 \cup \cdots \cup S_K$, 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 satisfies

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

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

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

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

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

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 it excludes the case that the modal manifolds bifurcate or merge, that is, it excludes cases such as the right panel of Figure 2. Similar conditions appear in Chen, Genovese and Wasserman (2014b), Romano (1988) 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 Chen, Genovese and Wasserman (2015), Einmahl and Mason (2005), Giné and Guillou (2002). We study three types of error metrics for regression modes: pointwise, uniform and mean integrated squared errors. We defer all proofs to the supplementary material [Chen et al. (2015)].

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

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

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

$$\|\widehat{p}_n - p\|_\infty^{(0)} = \sup_{x, y} \|\widehat{p}(x, y) - p(x, y)\|,$$

$$\|\widehat{p}_n - p\|_\infty^{(1)} = \sup_{x, y} \|\widehat{p}_y(x, y) - p_y(x, y)\|,$$

$$\|\widehat{p}_n - p\|_\infty^{(2)} = \sup_{x,y} \|\widehat{p}_{yy}(x, y) - p_{yy}(x, y)\|,$$

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

THEOREM 3 (Pointwise error rate). *Assume* (A1)–(A3) *and* (K1)–(K2). *Define a stochastic process* $A_n(x)$ *by*

$$A_n(x) = \begin{cases} \dfrac{1}{\Delta_n(x)}\Big|\Delta_n(x) - \max_{z \in M(x)}\big\{\big|p_{yy}^{-1}(x, z)\big|\big|\widehat{p}_{y,n}(x, z)\big|\big\}\Big|, \\ \qquad if\ \Delta_n(x) > 0, \\ 0, \qquad if\ \Delta_n(x) = 0. \end{cases}$$

Then, when

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

is sufficiently small, we have

$$\sup_{x \in D} A_n(x) = O_{\mathbb{P}}\big(\|\widehat{p}_n - p\|_{\infty,2}^*\big).$$

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

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

The proof is in the supplementary material [Chen et al. (2015)]. 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 $\widehat{p}_{y,n}(x, z)$ 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. We defined $A_n(x)$ as above since $\Delta_n(x) = 0$ implies $\max_{z \in M(x)}|\widehat{p}_{y,n}(x, z)| = 0$, so that the ratio would be ill-defined if $\Delta_n(x) = 0$. Also, the constraints on h in the second assertion ($\frac{nh^{d+5}}{\log n} \to \infty$ and $h \to 0$) are to ensure $\|\widehat{p}_n - p\|_{\infty,2}^* = o_{\mathbb{P}}(1)$.

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)–(A3) *and* (K1)–(K2). *Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

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

The proof is in the supplementary material [Chen et al. (2015)]. 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 Einmahl and Mason (2005), Giné and Guillou (2002) for similar findings in density estimation.

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

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

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

THEOREM 5 (MISE rate). *Assume* (A1)–(A3) *and* (K1)–(K2). *Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

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

The proof is in the supplementary material [Chen et al. (2015)]. If we instead focus on estimating the regression modes of the smoothed joint density $\widetilde{p}(x, y) = \mathbb{E}(\widehat{p}_n(x, y))$, 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

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

COROLLARY 6 (Error rates for smoothed conditional modes). *Assume* (A1)–(A3) *and* (K1)–(K2). *Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

$$\sqrt{nh^{d+3}} \sup_{x \in D}\left|\widetilde{\Delta}_n(x) - \max_{z \in \widetilde{M}(x)}\big\{\widetilde{p}_{yy}^{-1}(x, z)\widehat{p}_{y,n}(x, z)\big\}\right| = O_{\mathbb{P}}(\varepsilon_{n,2}),$$

$$\widetilde{\Delta}_n(x) = O_{\mathbb{P}}\left(\sqrt{\frac{1}{nh^{d+3}}}\right),$$

$$\widetilde{\Delta}_n = O_{\mathbb{P}}\left(\sqrt{\frac{\log n}{nh^{d+3}}}\right),$$

$$\widetilde{\mathsf{MISE}}(\widehat{M}_n) = O\left(\frac{1}{nh^{d+3}}\right),$$

where

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

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

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

Our 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

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

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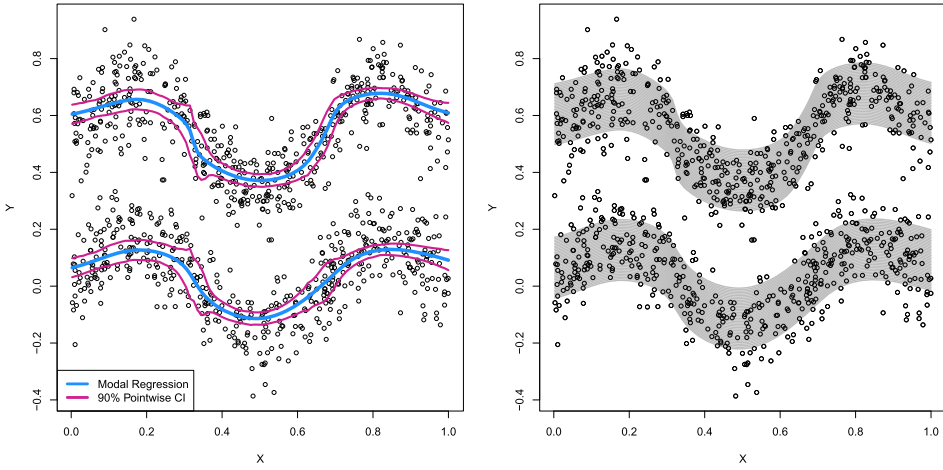
FIG. 4. *An example with pointwise (left) and uniform (right) confidence sets. The significance level is* 90%.

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 Chen, Genovese and Wasserman (2015), Chernozhukov, Chetverikov and Kato (2014a, 2014b).

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

(15)
$$\mathcal{F} = \left\{ (u, v) \mapsto f_{x,y}(u, v) : f_{x,y}(u, v) = \widetilde{p}_{yy}^{-1}(x, y) \right.$$
$$\left. \times K\left(\frac{\|x - u\|}{h}\right) K^{(1)}\left(\frac{y - v}{h}\right), x \in D, y \in \widetilde{M}(x) \right\},$$

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

(16)
$$\mathrm{Cov}\big(\mathbb{B}(f_1), \mathbb{B}(f_2)\big)$$
$$= \mathbb{E}\big(f_1(X_i, Y_i)f_2(X_i, Y_i)\big) - \mathbb{E}\big(f_1(X_i, Y_i)\big)\mathbb{E}\big(f_2(X_i, Y_i)\big),$$

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

THEOREM 7. *Assume* (A1)–(A3) *and* (K1)–(K2). *Define the random variable* $\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f \in \mathcal{F}} |\mathbb{B}(f)|$. *Then as* $\frac{nh^{d+5}}{\log n} \to \infty$, $h \to 0$,

$$\sup_{t \geq 0} \big|\mathbb{P}\big(\sqrt{nh^{d+3}}\widetilde{\Delta}_n < t\big) - \mathbb{P}(\mathbf{B} < t)\big| = O\left(\left(\frac{\log^7 n}{nh^{d+3}}\right)^{1/8}\right).$$

The proof is in the supplementary material [Chen et al. (2015)]. 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

$$\left|\sqrt{nh^{d+3}}\widetilde{\Delta}_n - \mathbf{B}\right| = O_{\mathbb{P}}\left(\left(\frac{\log^7 n}{nh^{d+3}}\right)^{1/8}\right).$$

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 by

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

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

THEOREM 8 (Bootstrap consistency). *Assume conditions* (A1)–(A3) *and* (K1)–(K2). *Also assume that* $\frac{nh^{d+5}}{\log n} \to \infty, h \to 0$. *Define*

$$\mathbf{B} = \frac{1}{\sqrt{h^{d+3}}} \sup_{f \in \mathcal{F}} \left|\mathbb{B}(f)\right|.$$

There exists $\mathcal{X}_n$ *such that* $\mathbb{P}(\mathcal{X}_n) \geq 1 - O(\frac{1}{n})$ *and, for all* $\mathcal{D}_n \in \mathcal{X}_n$,

$$\sup_{t \geq 0}\left|\mathbb{P}\left(\sqrt{nh^{d+3}}\widehat{\Delta}_n^* < t | \mathcal{D}_n\right) - \mathbb{P}(\mathbf{B} < t)\right| = O_{\mathbb{P}}\left(\left(\frac{\log^7 n}{nh^{d+3}}\right)^{1/8}\right).$$

The proof is in the supplementary material [Chen et al. (2015)]. 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. Using Theorems 7 and 8, we conclude the following.

COROLLARY 9 (Uniform confidence sets). *Assume conditions* (A1)–(A3) *and* (K1)–(K2). *Then as* $\frac{nh^{d+5}}{\log n} \to \infty$ *and* $h \to 0$,

$$\mathbb{P}\left(\widetilde{M}(x) \subseteq \widehat{M}_n(x) \oplus \widehat{\delta}_{n,1-\alpha}, \forall x \in D\right) = 1 - \alpha + O\left(\left(\frac{\log^7 n}{nh^{d+3}}\right)^{1/8}\right).$$

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

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

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

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

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

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

$$\mathbb{P}(Y \in \mathcal{P}_{1-\alpha}) \geq 1 - \alpha.$$

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 $\varepsilon_{1-\alpha}(x)$ via

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

An estimated pointwise prediction set is then

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

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

$$(17) \qquad \widehat{\varepsilon}_{1-\alpha} = \mathsf{Quantile}\big(\{d\big(Y_i, \widehat{M}_n(X_i)\big) : i = 1, \ldots, n\}, 1 - \alpha\big),$$

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

$$(18) \qquad \widehat{\mathcal{P}}_{1-\alpha} = \big\{(x, y) : x \in D, y \in \widehat{M}_n(x) \oplus \widehat{\varepsilon}_{1-\alpha}\big\}.$$

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.* 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 (18), we can see that the volume (Lebesgue measure) of the estimated uniform prediction set is

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

where $\widehat{K}_h(x)$ is the number of estimated local modes at $X = x$, and $\widehat{\varepsilon}_{1-\alpha,h}$ is as defined in (17). Roughly speaking, when h is small, $\widehat{\varepsilon}_{1-\alpha,h}$ is also small, but the number of estimated manifolds is large; on the other hand, when h is large, $\widehat{\varepsilon}_{1-\alpha,h}$ is large, but the number of estimated manifolds is small. This is like the bias-variance trade-off: small h corresponds to less bias ($\widehat{\varepsilon}_{1-\alpha,h}$) but higher variance (number of estimated manifolds).

Our proposal is to select h by

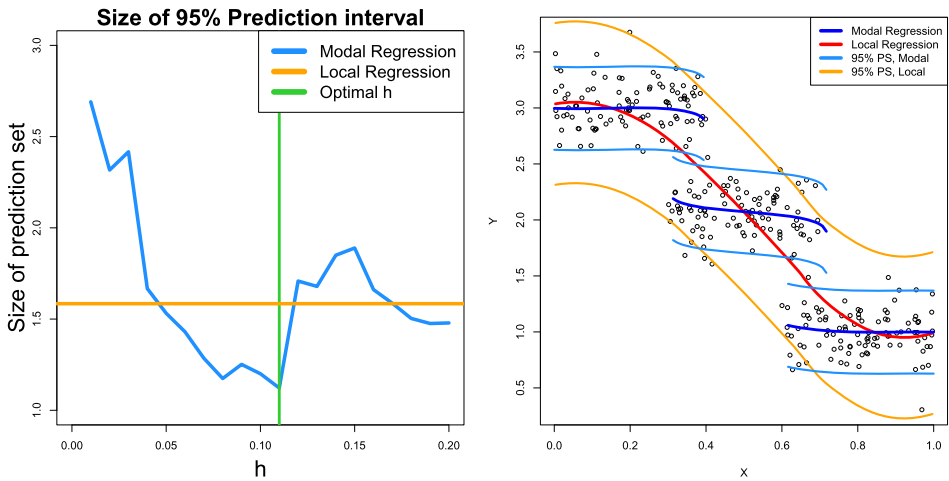$$h^* = \underset{h \geq 0}{\mathrm{argmin}}\, \mathsf{Vol}(\widehat{\mathcal{P}}_{1-\alpha,h}).$$

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

Figure 5 gives an example this rule when $\alpha = 0.05$, that is, when minimizing the size of the estimated 95% uniform prediction set. Here, we actually use cross-validation to obtain the size of the prediction set; namely, we use the training set to estimate the modal manifolds and then use the validation set to estimate the width of prediction set. This helps us to avoid overfitting. 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 $h^* = 0.07$ 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.

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

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

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

pointwise and uniform prediction sets based on the regression function are

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

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

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 $\mu_1(x) < \mu_2(x) < \cdots < \mu_{K(x)}(x)$ 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

$$\sigma_{\max}^2(x) = \max_{j=1,\ldots,K(x)} \sigma_j^2(x),$$

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

Also define

$$\Delta_{\min} = \inf_{x \in D} \Delta_{\min}(x), \qquad \sigma_{\max}^2 = \sup_{x \in D} \sigma_{\max}^2(x),$$

and

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

and

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

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

$$\frac{\Delta_{\min}(x)}{\sigma_{\max}(x)} > \max\left\{1.1 \cdot \frac{K(x)}{K(x) - 1} z_{1-\alpha/2},\right.$$

$$\left.\sqrt{6.4 \vee 2\log(4(K(x) \vee 3 - 1)) + 2\log\left(\frac{\pi_{\max}(x)}{\pi_{\min}(x)}\right)}\right\},$$

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

$$\frac{\Delta_{\min}}{\sigma_{\max}} > \max\left\{1.1 \cdot \left(\frac{2\overline{K}}{K_{\min} - 1}\right) z_{1-\alpha/2},\right.$$

$$\left.\sqrt{6.4 \vee 2\log(4(K_{\max} \vee 3 - 1)) + 2\log\left(\frac{\pi_{\max}}{\pi_{\min}}\right)}\right\},$$

then

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

The proof is in the supplementary material [Chen et al. (2015)]. 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 similar to modal regression. The literature on mixture regression, also known as mixture of experts modeling, is vast; see, for example, Bishop (2006), Huang, Li and Wang (2013), Huang and Yao (2012), Hunter and Young (2012), Jacobs et al. (1991), Jiang and Tanner (1999), Khalili and Chen (2007), Viele and Tong (2002). In mixture regression, we assume that the conditional density function takes the form

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

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 [Chaganty and Liang (2013), Viele and Tong (2002)]. 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)–(MR5).

Modal regression is a fairly simple tool that achieves a similar goal to mixture regression models, and uses fewer assumptions. 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.

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

TABLE 1
Comparison for methods based on mixtures versus modes

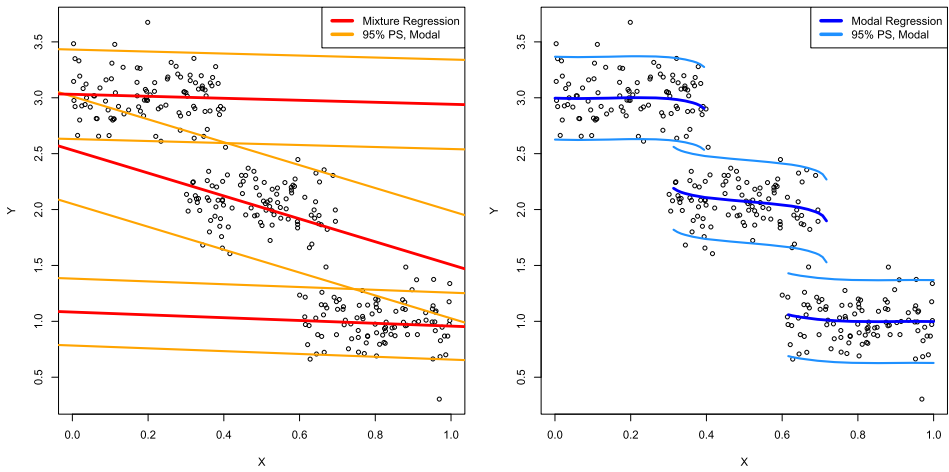|	**Mixture-based**	**Mode-based**
Density estimation	Gaussian mixture	Kernel density estimate
Clustering	K-means	Mean-shift clustering
Regression	Mixture regression	Modal regression
Algorithm	EM	Mean-shift
Complexity parameter	K (number of components)	h (smoothing bandwidth)
Type	Parametric model	Nonparametric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

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

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.

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 to define 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 [Chen, Genovese and Wasserman (2014a), Cheng (1995), Comaniciu and Meer (2002), Li, Ray and Lindsay (2007), Yao and Lindsay (2009)] 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

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

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). Similar ideas have been used by Li, Ray and Lindsay (2007), Yao and Lindsay (2009), and the former authors also discuss how the modes and clustering results merge as the bandwidth increases.

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 of the algorithm. That is, by iterating the mean-shift 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 hierarchical 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 dispersion 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.
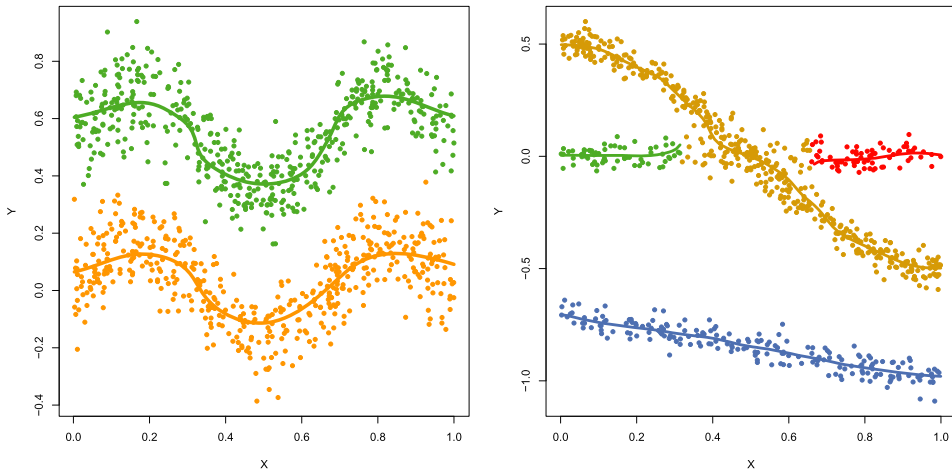

FIG. 7. *Two examples of clustering based on modal regression.*

In a mixture regression model, where each ϕ_j is assumed 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 of the Gaussian mixtures. The clustering based on modal regression is like the recovery process for the mixing mechanism. 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 Chen, Genovese and Wasserman (2014b, 2015), Eberly (1996), Genovese et al. (2014).

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 local mode of subspace spanned by $v_2(x, y)$ with $\lambda_2(x, y) < 0$. We can express this as

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

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

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

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 proof is in the supplementary material [Chen et al. (2015)]. The lemma asserts that a conditional mode where the density is locally stationary, that is, $p_x(x, y) = 0$, or the density is locally isotropic, that is, $p_{xy}(x, y) = 0$, 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.
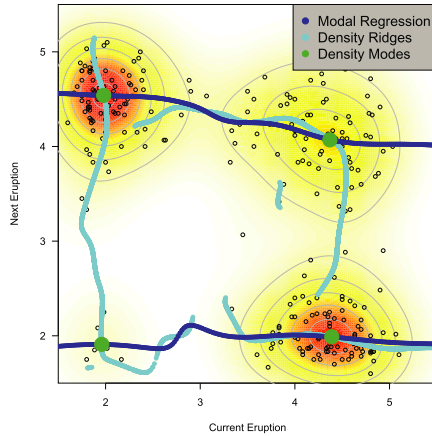
FIG. 8. *A comparison between modal regression, density ridges and density modes using the old faithful data set. The background color represents the joint density (red: high density).*

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.

9. Discussion. We have investigated a nonparametric method for modal regression estimation, based on a KDE of a joint sample of data points $(X_1, Y_1), \dots, (X_n, Y_n)$. 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, 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
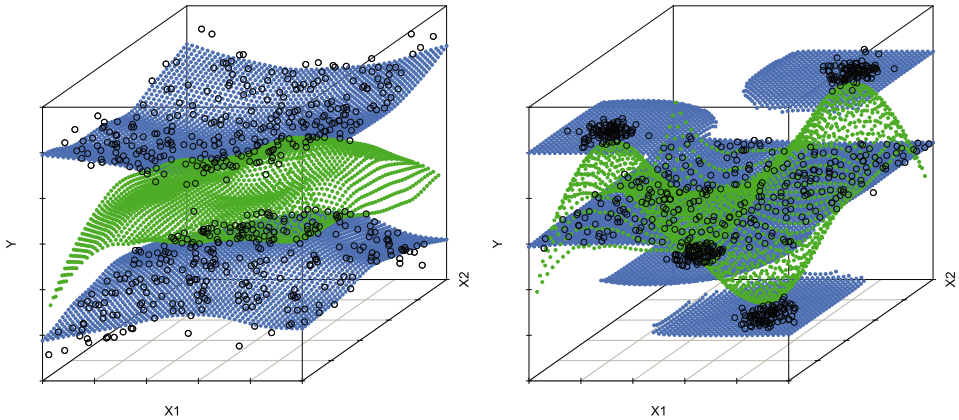
FIG. 9.   *Two examples for d = 2. Modal regression estimates are shown in blue, and local regression in green.*

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).

Acknowledgement.   We thank the reviewers for useful comments.

SUPPLEMENTARY MATERIAL

Supplementary Proofs: Nonparametric modal regression (DOI: 10.1214/15-AOS1373SUPP; .pdf). This document contains all proofs to the theorems and lemmas in this paper.

REFERENCES

ARIAS-CASTRO, E., MASON, D. and PELLETIER, B. (2013). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Unpublished Manuscript.

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

CARREIRA-PERPIÑÁN, M. Á. (2007). Gaussian mean-shift is an em algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 0767–0776.

CHAGANTY, A. T. and LIANG, P. (2013). Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the* 30*th International Conference on Machine Learning* (*ICML*-13) 1040–1048. ACM, New York.

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2014a). Enhanced mode clustering. Available at arXiv:1406.1780.

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2014b). Generalized mode and ridge estimation. Available at arXiv:1406.1803.

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2015). Asymptotic theory for density ridges. *Ann. Statist.* **43** 1896–1928. MR3375871

CHEN, Y.-C., GENOVESE, C. R., TIBSHIRANI, R. J. and WASSERMAN, L. (2015). Supplement to "Nonparametric modal regression." DOI:10.1214/15-AOS1373SUPP.

CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17** 790–799.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014b). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461

COMANICIU, D. and MEER, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 603–619.

EBERLY, D. (1996). *Ridges in Image and Data Analysis*. Springer, Berlin.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

EINBECK, J. and TUTZ, G. (2006). Modelling beyond regression functions: An application of multimodal regression to speed-flow data. *J. Roy. Statist. Soc. Ser. C* **55** 461–475. MR2242274

EINMAHL, U. and MASON, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380–1403. MR2195639

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2014). Nonparametric ridge estimation. *Ann. Statist.* **42** 1511–1545. MR3262459

GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 907–921. MR1955344

HUANG, M., LI, R. and WANG, S. (2013). Nonparametric mixture of regression models. *J. Amer. Statist. Assoc.* **108** 929–941. MR3174674

HUANG, M. and YAO, W. (2012). Mixture of regression models with varying mixing proportions: A semiparametric approach. *J. Amer. Statist. Assoc.* **107** 711–724. MR2980079

HUNTER, D. R. and YOUNG, D. S. (2012). Semiparametric mixtures of regressions. *J. Nonparametr. Stat.* **24** 19–38. MR2885823

HYNDMAN, R. J., BASHTANNYK, D. M. and GRUNWALD, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.* **5** 315–336. MR1422114

JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87. ISSN 0899-7667. Available at http://dx.doi.org/10.1162/neco.1991.3.1.79.

JIANG, W. and TANNER, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Ann. Statist.* **27** 987–1011. MR1724038

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. MR2411662

LEE, M.-J. (1989). Mode regression. *J. Econometrics* **42** 337–349. MR1040748

LI, J., RAY, S. and LINDSAY, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8** 1687–1723. MR2332445

ROJAS, A. (2005). Nonparametric mixture regression. Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA.

ROMANO, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647. MR0947566

SAGER, T. W. and THISTED, R. A. (1982). Maximum likelihood estimation of isotonic modal regression. *Ann. Statist.* **10** 690–707. MR0663426

SCOTT, D. W. (1992). *Multivariate Density Estimation*: *Theory*, *Practice*, *and Visualization*. Wiley, New York. MR1191168

VIELE, K. and TONG, B. (2002). Modeling with mixtures of linear regressions. *Stat. Comput.* **12** 315–330. MR1951705

YAO, W. (2013). A note on EM algorithm for mixture models. *Statist. Probab. Lett.* **83** 519–526. MR3006984

YAO, W. and LI, L. (2014). A new regression model: Modal linear regression. *Scand. J. Stat.* **41** 656–671. MR3249422

YAO, W. and LINDSAY, B. G. (2009). Bayesian mixture labeling by highest posterior density. *J. Amer. Statist. Assoc.* **104** 758–767. MR2751453

YAO, W., LINDSAY, B. G. and LI, R. (2012). Local modal regression. *J. Nonparametr. Stat.* **24** 647–663. MR2968894

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVE.
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: yenchic@andrew.cmu.edu
          genovese@cmu.edu
          ryantibs@cmu.edu
          larry@cmu.edu