Nonparametric Modal Regression

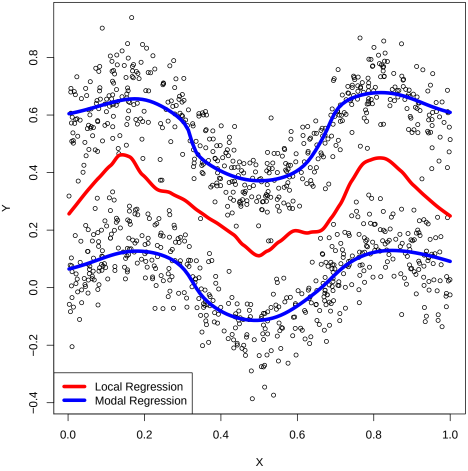
Yen-Chi Chen

Christopher R. Genovese Ryan J. Tibshirani Larry Wasserman

Department of Statistics
Carnegie Mellon University

August 4, 2015

Motivating Examples for Modal Regression



Introduction

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

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

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

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

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

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

- Equivalently,

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

Definition for Modal Regression

We assume $x \in \mathbb{K}$, a compact support.

- Regression function–the conditional **mean**:

$$m(x) = \mathbb{E}(Y|X = x) = \int yp(y|x)dy.$$

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

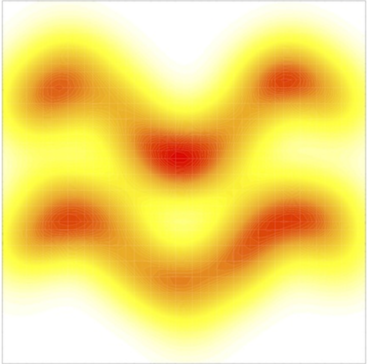
$$M(x) = \text{Mode}(Y|X = x) = \left\{ y : \frac{d}{dy} p(y|x) = 0, \frac{d^2}{dy^2} p(y|x) < 0 \right\}.$$

- Equivalently,

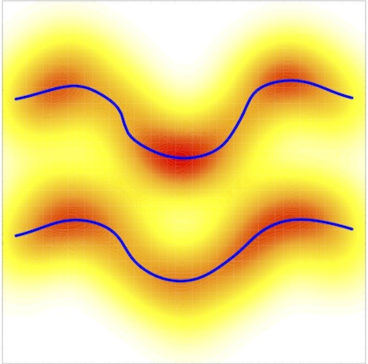
$$M(x) = \left\{ y : \frac{\partial}{\partial y} p(x, y) = 0, \frac{\partial^2}{\partial y^2} p(x, y) < 0 \right\}.$$

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

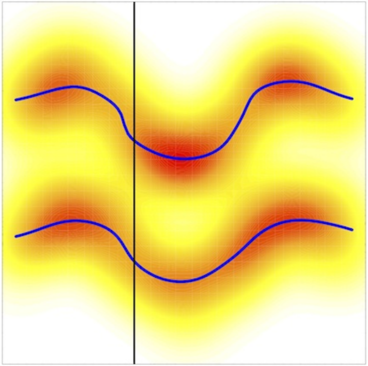
Conditional Local Modes



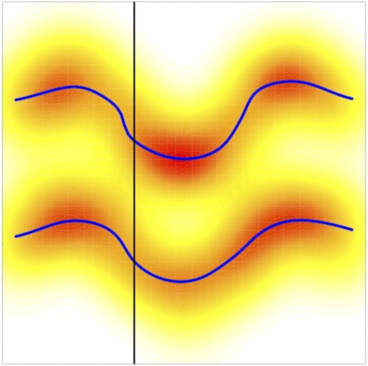
Conditional Local Modes

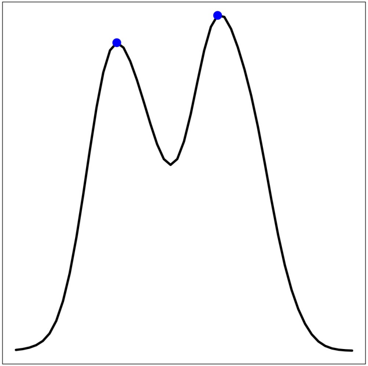


Conditional Local Modes



Conditional Local Modes





Estimator for Modal Regression

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

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

Estimator for Modal Regression

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

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

- Finding conditional local modes is hard in general.

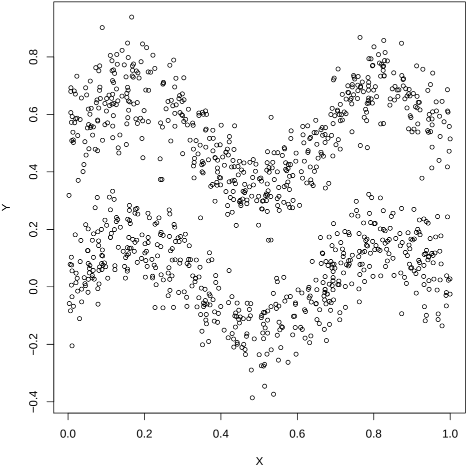
Estimator for Modal Regression

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

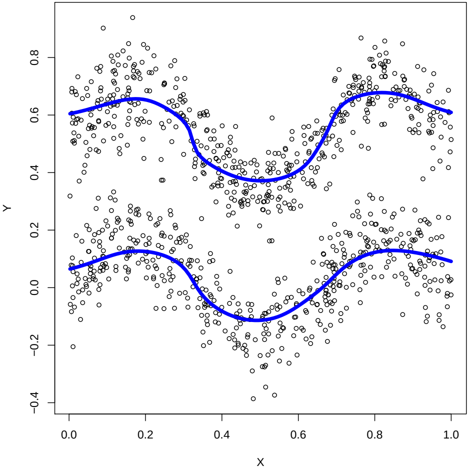
$$\widehat{M}_n(x) = \left\{ y : \frac{\partial}{\partial y}\widehat{p}_n(x,y) = 0, \frac{\partial^2}{\partial y^2}\widehat{p}_n(x,y) < 0 \right\}.$$

- Finding conditional local modes is hard in general.
- Partial mean shift: a simple algorithm for computing $\widehat{M}_n(x)$, the plug-in estimator of the KDE, from the data (Einbeck et. al. 2006).

Example for Modal Regression



Example for Modal Regression



Asymptotic Theory

Error Measurement

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

Error Measurement

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

- the *pointwise* loss

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

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

Error Measurement

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

- the *pointwise* loss

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

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

- the *uniform* loss

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

Rate of Convergence

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

Rate of Convergence

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

Theorem

Under regularity conditions,

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

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

Rate of Convergence

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

Theorem

Under regularity conditions,

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

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

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

Asymptotic Theory

To conduct statistical inferences, we ignore the bias and focus on the stochastic variation.

Asymptotic Theory

To conduct statistical inferences, we ignore the bias and focus on the stochastic variation.

Theorem

Under regularity conditions,

- $\sqrt{nh^{d+3}}\Delta_n \approx \sup\{\textit{Empirical Process}\} \approx \sup\{\textit{Gaussian process}\}$.

Asymptotic Theory

To conduct statistical inferences, we ignore the bias and focus on the stochastic variation.

Theorem

Under regularity conditions,

- $\sqrt{nh^{d+3}}\Delta_n \approx \sup\{\textit{Empirical Process}\} \approx \sup\{\textit{Gaussian process}\}$.
- $\sqrt{nh^{d+3}}\Delta_n \approx \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$ *for certain function space* $\mathcal{F}$.

Asymptotic Theory

To conduct statistical inferences, we ignore the bias and focus on the stochastic variation.

Theorem

Under regularity conditions,

- $\sqrt{nh^{d+3}}\Delta_n \approx \sup\{\textit{Empirical Process}\} \approx \sup\{\textit{Gaussian process}\}$.
- $\sqrt{nh^{d+3}}\Delta_n \approx \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$ *for certain function space* $\mathcal{F}$.

However, this is not enough for statistical inference (unknown quantities in the Gaussian Process).

Confidence Sets

We use the bootstrap to approximate Δ_n. Define another uniform metric $\widehat{\Delta}_n = \sup_x \text{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x))$.

Confidence Sets

We use the bootstrap to approximate Δ_n. Define another uniform metric $\widehat{\Delta}_n = \sup_x \text{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x))$.

Theorem

Under regularity conditions,

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

Confidence Sets

We use the bootstrap to approximate Δ_n. Define another uniform metric $\widehat{\Delta}_n = \sup_x \text{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x))$.

Theorem

Under regularity conditions,

- $\sqrt{nh^{d+3}}\widehat{\Delta}_n \approx \sup_{f\in\mathcal{F}} |\mathbb{B}(f)|$ *for certain function space* $\mathcal{F}$.
- $\sqrt{nh^{d+3}}\widehat{\Delta}_n \approx \sqrt{nh^{d+3}}\Delta_n$.

Confidence Sets

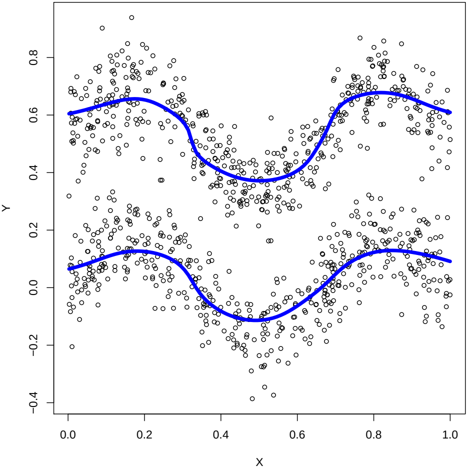
We use the bootstrap to approximate Δ_n. Define another uniform metric $\widehat{\Delta}_n = \sup_x \text{Haus}(\widehat{M}_n^*(x), \widehat{M}_n(x))$.

Theorem

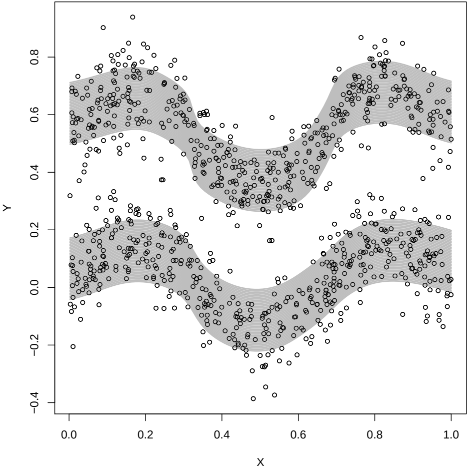
Under regularity conditions,

- $\sqrt{nh^{d+3}}\widehat{\Delta}_n \approx \sup_{f \in \mathcal{F}} |\mathbb{B}(f)|$ *for certain function space* $\mathcal{F}$.
- $\sqrt{nh^{d+3}}\widehat{\Delta}_n \approx \sqrt{nh^{d+3}}\Delta_n$.
- *The set*
$$\left\{(x, y) : y \in \widehat{M}_n(x) \oplus \widehat{t}_{1-\alpha}, x \in \mathbb{K}\right\}$$
is an asymptotic valid confidence set for M; $\widehat{t}_{1-\alpha}$ *is the upper* $1-\alpha$ *quantile of* $\widehat{\Delta}_n$.

Example for Confidence Sets



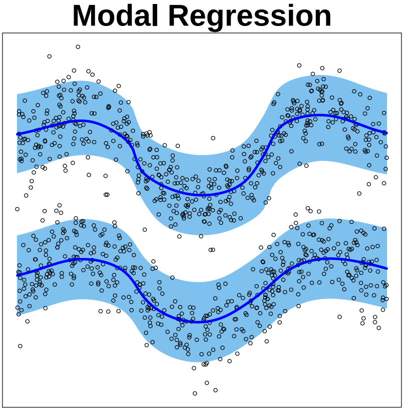
Example for Confidence Sets

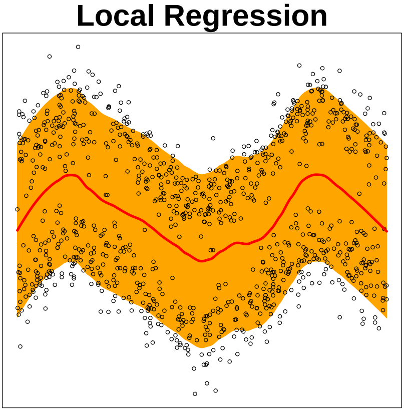


Extensions

Prediction Sets

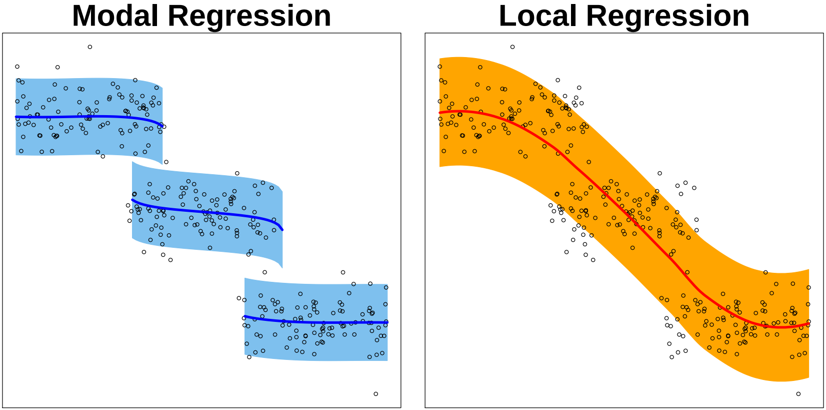
We can use modal regression to construct a compact prediction set.





Prediction Sets

We can use modal regression to construct a compact prediction set.

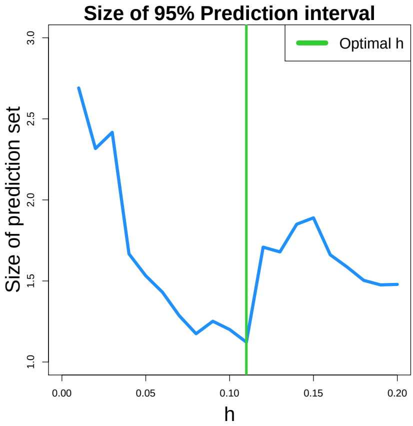


We can choose smoothing parameter h via minimizing the size of prediction set.
Namely, we choose

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

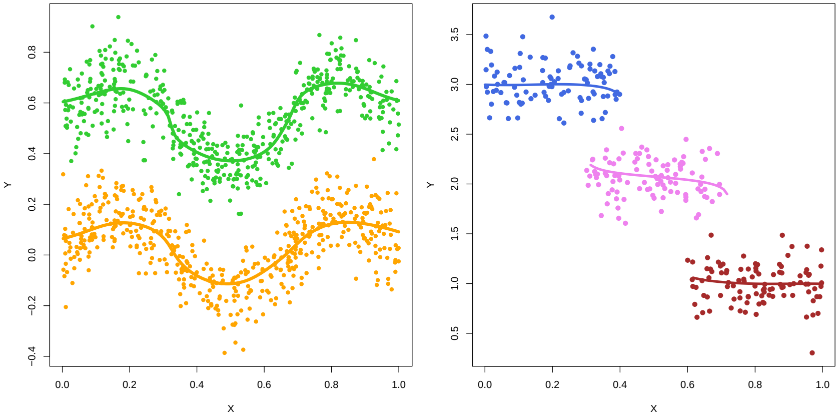
where $\widehat{\mathcal{P}}_{1-\alpha}$ is the prediction set.

Example: Bandwidth Selection



Regression Clustering

We can use modal regression to do 'clustering'–exploring the hidden structures.



Concluding Remarks

- Modal regression is very similar to mixture regression.

Concluding Remarks

- Modal regression is very similar to mixture regression.
- However, our approach is purely nonparametric–no Gaussian assumption, free from number of mixture components.

Concluding Remarks

- Modal regression is very similar to mixture regression.
- However, our approach is purely nonparametric–no Gaussian assumption, free from number of mixture components.
- Fast to compute–no need to use EM algorithm.

Thank you!

More information and R source code can be found in

- `http://www.stat.cmu.edu/~yenchic`

reference

1. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Density Level Sets: Asymptotics, Inference, and Visualization." Submitted to the Journal of American Statistical Association. arXiv preprint arXiv:1504.05438 (2015).
2. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Asymptotic theory for density ridges." To appear in the Annals of Statistics. arXiv preprint arXiv:1406.5663 (2014).
3. Chen, Yen-Chi, Christopher R. Genovese, Ryan J. Tibshirani, and Larry Wasserman. "Nonparametric Modal Regression." Under review of the Annals of Statistics. arXiv preprint arXiv:1412.1716 (2014).
4. Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Gaussian approximation of suprema of empirical processes." The Annals of Statistics 42, no. 4 (2014): 1564-1597.
5. Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. "Anti-concentration and honest, adaptive confidence bands." The Annals of Statistics 42, no. 5 (2014): 1787-1818.
6. Einbeck, Jochen, and Gerhard Tutz. "Modelling beyond regression functions: an application of multimodal regression to speedflow data." Journal of the Royal Statistical Society: Series C (Applied Statistics) 55, no. 4 (2006): 461-475.
7. Genovese, Christopher R., et al. "Nonparametric ridge estimation." The Annals of Statistics 42.4 (2014): 1511-1545.
8. Ozertem, Umut, and Deniz Erdogmus. "Locally defined principal curves and surfaces." The Journal of Machine Learning Research 12 (2011): 1249-1286.

Regularity Conditions

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

(A2) There exists $\lambda_2 > 0$ such that for any $(x, y) \in \mathbb{K} \times \mathbb{K}$ 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

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

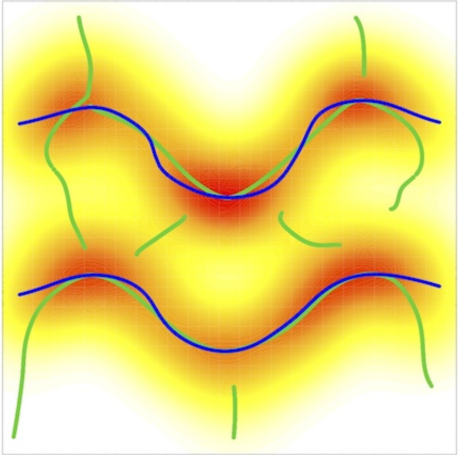
for $\alpha = 0, 1, 2$.

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

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

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

Modal Regression VS Density Ridges



A general mixture model:

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

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 $\sigma_j^2(x)$.
Common 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.

Mixture Inference versus Modal Inference

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

Table: Comparison for methods based on mixtures versus modes.

3D examples