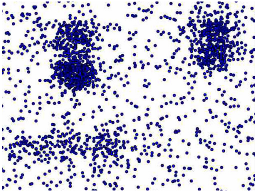
Efficient and Optimal Modal-set Estimation using kNN graphs

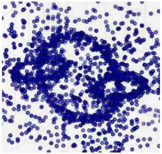
Samory Kpotufe
ORFE, Princeton University

Based on various results with Sanjoy Dasgupta, Kamalika Chaudhuri, Ulrike von Luxburg, Heinrich Jiang

Motivation:

Density-based Clustering: group points into high-density regions.





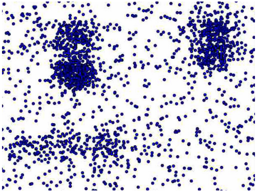
- **Flexibility:** can identify any number of rich structures in data.
- **Clear ground truth:** targets concrete mathematical objects.
- **Many applications:** medical imaging, text mining, speech, vision, ...

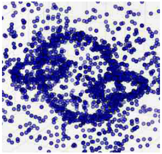
However: Heuristics work better than theoretical methods :(

We want a **practical** procedure with **theoretical** guarantees!

Motivation:

Density-based Clustering: group points into high-density regions.





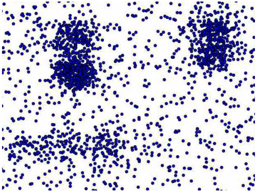
- **Flexibility:** can identify any number of rich structures in data.
- **Clear ground truth:** targets concrete mathematical objects.
- **Many applications:** medical imaging, text mining, speech, vision, ...

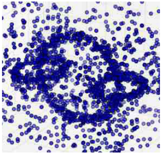
However: Heuristics work better than theoretical methods :(

We want a **practical** procedure with **theoretical** guarantees!

Motivation:

Density-based Clustering: group points into high-density regions.





- **Flexibility:** can identify any number of rich structures in data.
- **Clear ground truth:** targets concrete mathematical objects.
- **Many applications:** medical imaging, text mining, speech, vision, ...

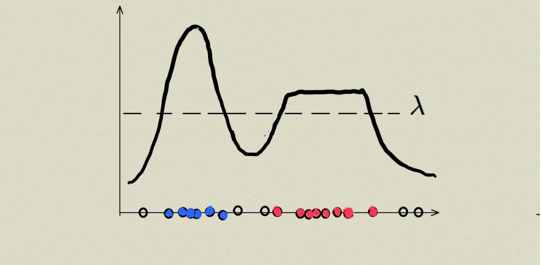
However: Heuristics work better than theoretical methods :(

We want a **practical** procedure with **theoretical** guarantees!

Formalisms of density-based clustering

Common idea: cluster data around high-density cores!

... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f

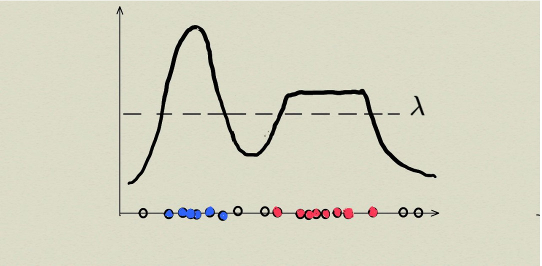


DBSCAN: cores are Connected-Components of a level set λ of f.

Problem: which level λ?

Formalisms of density-based clustering

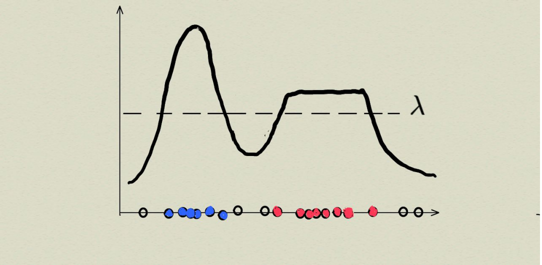
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



DBSCAN: cores are Connected-Components of a level set λ of f.
Problem: which level λ?

Formalisms of density-based clustering

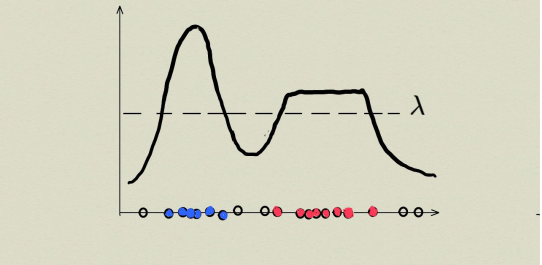
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



DBSCAN: cores are Connected-Components of a level set λ of f.
Problem: which level λ?

Formalisms of density-based clustering

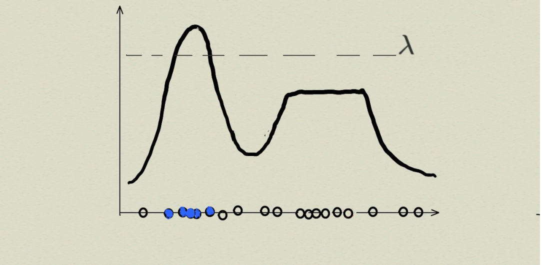
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



DBSCAN: cores are Connected-Components of a level set λ of f.
Problem: which level λ?

Formalisms of density-based clustering

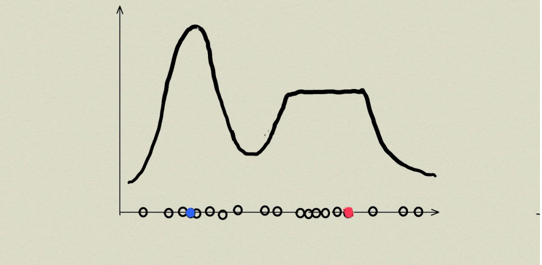
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



DBSCAN: cores are Connected-Components of a level set λ of f.
Problem: which level λ? (we can get $\neq$ results)

Formalisms of density-based clustering

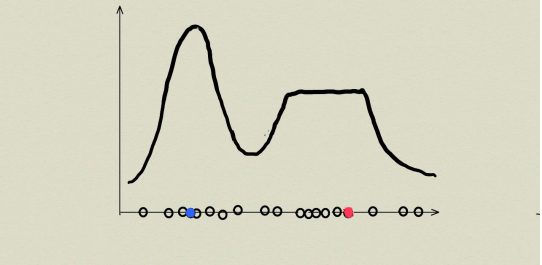
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



MEAN-SHIFT: cores are *point-modes* (maxima) of f.
Problem: unstable for general maxima ... hard to analyze.

Formalisms of density-based clustering

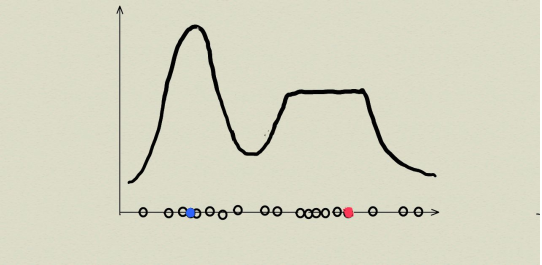
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



MEAN-SHIFT: cores are *point-modes* (maxima) of f.
Problem: unstable for general maxima ... hard to analyze.

Formalisms of density-based clustering

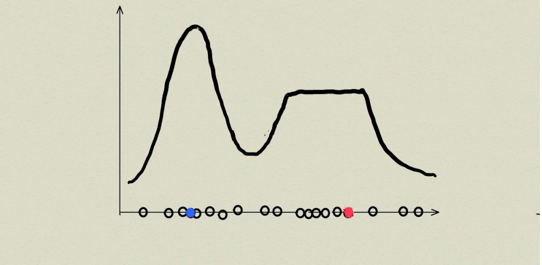
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f



MEAN-SHIFT: cores are *point-modes* (maxima) of f.
Problem: unstable for general maxima ... hard to analyze.

Formalisms of density-based clustering

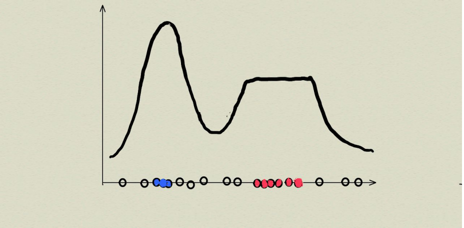
Common idea: cluster data around high-density cores!
... Suppose the data $\{X_i\}_1^n \sim_{\text{i.i.d.}}$ some density f

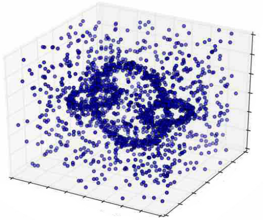


MEAN-SHIFT: cores are *point-modes* (maxima) of f.
Problem: unstable for general maxima ... hard to analyze.

Our goal:

Practical and Optimal estimator of general maxima of density f.





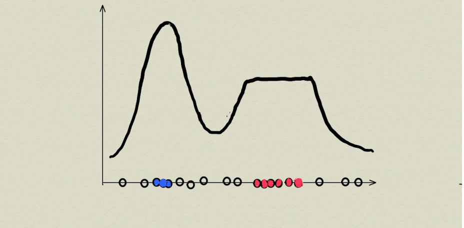
Difficulty: unknown location, dimension, shape

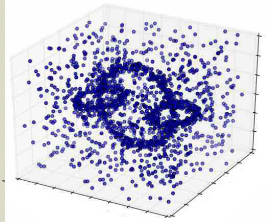
Side benefits: applies beyond clustering (e.g. manifold denoising).

Practical subroutines: *we'll traverse a k-NN graph over $\{X_i\}$*

Our goal:

Practical and Optimal estimator of general maxima of density f.





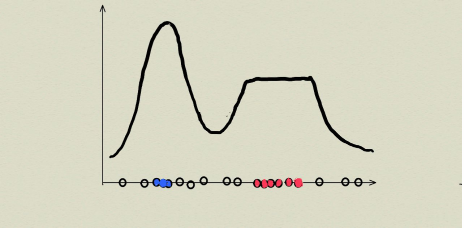
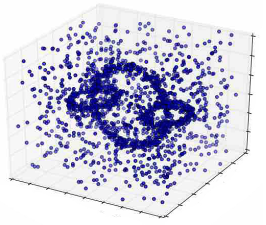
Difficulty: unknown location, dimension, shape

Side benefits: applies beyond clustering (e.g. manifold denoising).

Practical subroutines: *we'll traverse a k-NN graph over $\{X_i\}$*

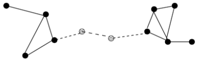
Our goal:

Practical and Optimal estimator of general maxima of density f.

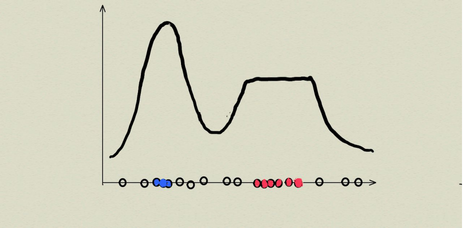
Difficulty: unknown location, dimension, shape
Side benefits: applies beyond clustering (e.g. manifold denoising).

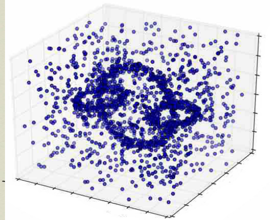
Practical subroutines: *we'll traverse a k-NN graph over $\{X_i\}$*



Our goal:

Practical and Optimal estimator of general maxima of density f.





Difficulty: unknown location, dimension, shape

Side benefits: applies beyond clustering (e.g. manifold denoising).

Practical subroutines: *we'll traverse a k-NN graph over $\{X_i\}$*

Outline:

- How k-**NN graphs** relate to f. (groundwork)
  (with Chaudhuri, Dasgupta, von Luxburg, 2011, 2014)
- Estimating all **modes** of f. (first intuition ...)
  (with S. Dasgupta, 2014)
- Estimating all **modal sets** of f. (final intuition $\square$)
  (with H. Jiang, 2017)

Outline:

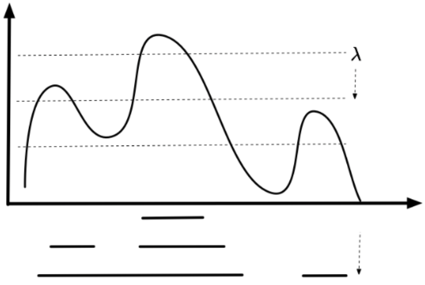
- How k-**NN graphs** relate to f. (groundwork)
  (with Chaudhuri, Dasgupta, von Luxburg, 2011, 2014)
- Estimating all **modes** of f. (first intuition ...)
  (with S. Dasgupta, 2014)
- Estimating all **modal sets** of f. (final intuition $\square$)
  (with H. Jiang, 2017)

Outline:

- How k-**NN graphs** relate to f. (groundwork)
  (with Chaudhuri, Dasgupta, von Luxburg, 2011, 2014)
- Estimating all **modes** of f. (first intuition ...)
  (with S. Dasgupta, 2014)
- Estimating all **modal sets** of f. (final intuition $\square$)
  (with H. Jiang, 2017)

How k-**NN graphs** *relate to* f.

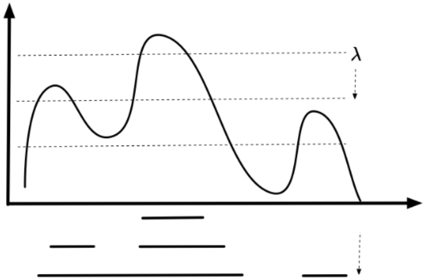
Characterize f by its *cluster-tree* [Hartigan 81]:



- [Cha., Das. 10]: first consistent estimator (extends single-linkage).
- [Kpo., vLux. 11]: Simple k-**NN subgraphs** + stronger consistency.

*How k-***NN graphs** *relate to f.*

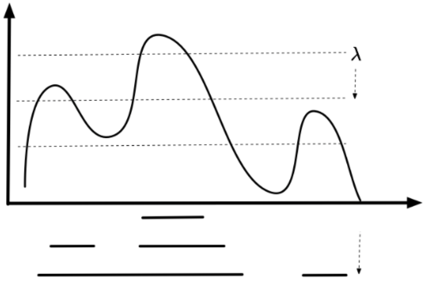
Characterize f by its *cluster-tree* [Hartigan 81]:



- [Cha., Das. 10]: first consistent estimator (extends single-linkage).
- [Kpo., vLux. 11]: Simple k-**NN subgraphs** + stronger consistency.

How k-**NN graphs** *relate to* f.

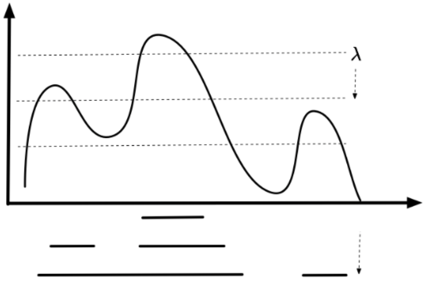
Characterize f by its *cluster-tree* [Hartigan 81]:



- [Cha., Das. 10]: first consistent estimator (extends single-linkage).
- [Kpo., vLux. 11]: Simple k-**NN subgraphs** + stronger consistency.

How k-**NN graphs** *relate to* f.

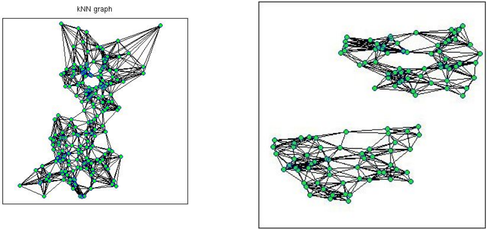
Characterize f by its *cluster-tree* [Hartigan 81]:



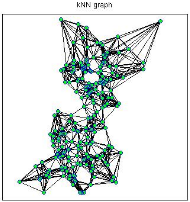
- [Cha., Das. 10]: first consistent estimator (extends single-linkage).
- [Kpo., vLux. 11]: Simple k-**NN subgraphs** $+$ stronger consistency.

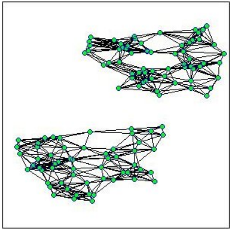
Intuition: recursively remove X_i's with farthest k'th NN ...

Graph breaks up into subgraphs corresponding to level sets of f.

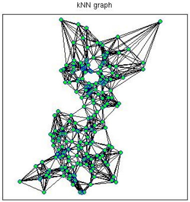


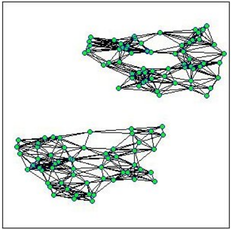
Intuition: recursively remove X_i's with farthest k'th NN ...
Graph breaks up into subgraphs corresponding to level sets of f.

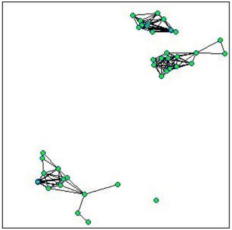




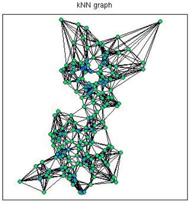
Intuition: recursively remove X_i's with farthest k'th NN ...
Graph breaks up into subgraphs corresponding to level sets of f.

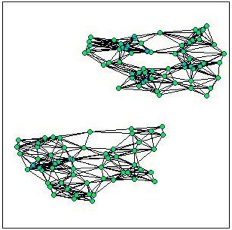


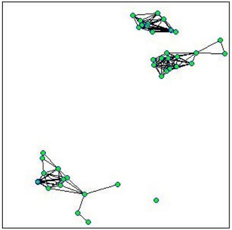




Intuition: recursively remove X_i's with farthest k'th NN ...
Graph breaks up into subgraphs corresponding to level sets of f.

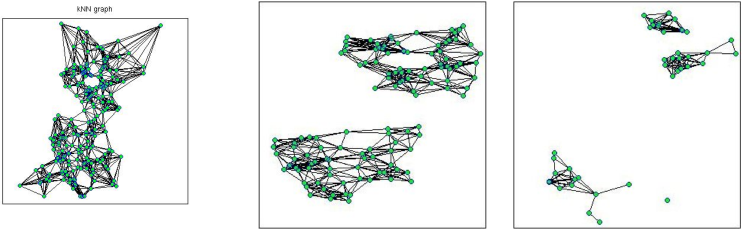






Theo. Let f uniformly cont. $+$ mild conditions on $k = \Omega(\log n)$.
Let $\mathcal{C}_1, \mathcal{C}_2$ be disjoint CCs of some $\{f \geq \lambda\}$. W.p $\to 1$,
$\mathcal{C}_1 \cap X^n$ and $\mathcal{C}_2 \cap X^n$ are in disjoint CCs of a k-NN subgraph.

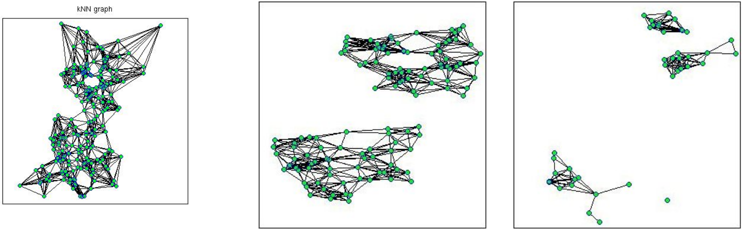
Intuition: recursively remove X_i's with farthest k'th NN ...
Graph breaks up into subgraphs corresponding to level sets of f.



Practical problem: spurious CCs due to data variability.

Intuition: recursively remove X_i's with farthest k'th NN ...

Graph breaks up into subgraphs corresponding to level sets of f.



Practical problem: spurious CCs due to data variability.

Reconnect using careful lookups to lower subgraphs.

Consistency of *Reconnect* shown in [Cha., Das., Kpo., vLux. 14]

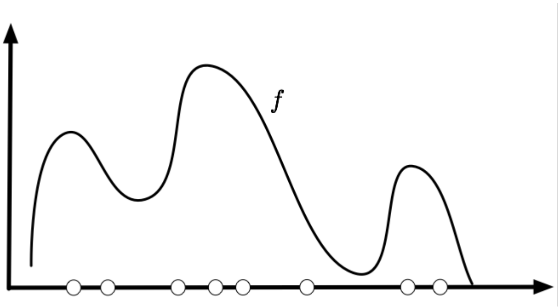
Many new refinements by various authors ...

S. Balakrishnan, L. Wasserman, A. Rinaldo, I. Steinwart, M. Belkin, Y.C. Chen, F. Chazal, J. Klëmëla, ...

Outline:

- How k-**NN graphs** relate to f.
  (with Chaudhuri, Dasgupta, von Luxburg, 2011, 2014)
- Estimating all **modes** of f.
  (with S. Dasgupta, 2014)
- Estimating all **modal sets** of f.
  (with H. Jiang, 2017)

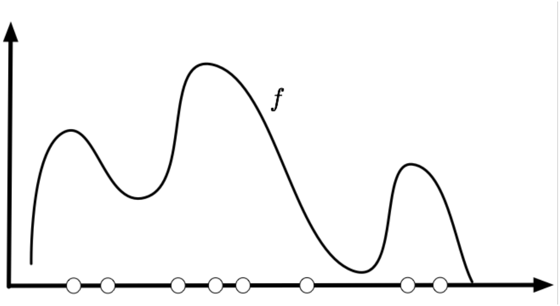
Estimating all **modes** *of f: What was known*



Rate-Optimal: single mode case ([S. Tsybakov, 90] ...).
Practical: mean-shift (hard to analyze ... see [Genovesee, ... Wasserman et.al., 13], [Arias-Castro et.al., 13] on consistency)

We derive a rate-optimal estimator based on k-NN graphs ...

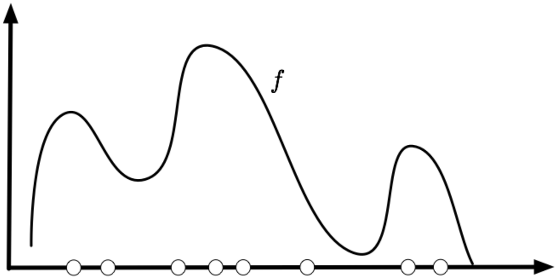
Estimating all **modes** *of f: What was known*



Rate-Optimal: single mode case ([S. Tsybakov, 90] ...).
Practical: mean-shift (hard to analyze ... see [Genovesee, ... Wasserman et.al., 13], [Arias-Castro et.al., 13] on consistency)

We derive a rate-optimal estimator based on k-NN graphs ...

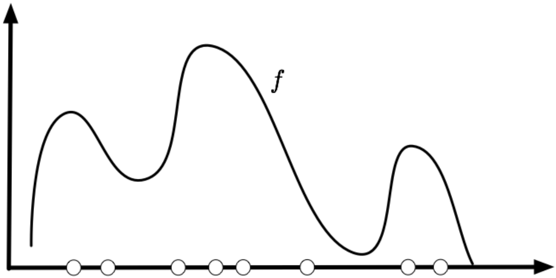
Estimating all **modes** *of f: What was known*



Rate-Optimal: single mode case ([S. Tsybakov, 90] ...).
Practical: mean-shift (hard to analyze ... see [Genovesee, ... Wasserman et.al., 13], [Arias-Castro et.al., 13] on consistency)

We derive a rate-optimal estimator based on k-NN graphs ...

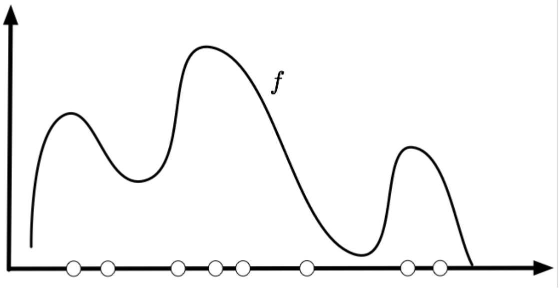
Estimating all **modes** *of f: What was known*



Rate-Optimal: single mode case ([S. Tsybakov, 90] ...).
Practical: mean-shift (hard to analyze ... see [Genovesee, ... Wasserman et.al., 13], [Arias-Castro et.al., 13] on consistency)

We derive a rate-optimal estimator based on k-NN graphs ...

Estimating all **modes** *of* f*: What was known*



Rate-Optimal: single mode case ([S. Tsybakov, 90] ...).
Practical: mean-shift (hard to analyze ... see [Genovesee, ... Wasserman et.al., 13], [Arias-Castro et.al., 13] on consistency)

We derive a rate-optimal estimator based on k-NN graphs ...

Program of construction

- k-**NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- **k-NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- **k-NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- **k-NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- **k-NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- k-**NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- **k-NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Program of construction

- **k-NN density rates:**
  asymptotic $1/\sqrt{k}$ rates (e.g. [Biau, ..., Devroye et.al., 11]).
  We show high-prob. finite sample rates.
- **Single mode:**
  Common estimator in theory: $\hat{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}(x)$.
  Practical estimator: $\tilde{x} = \arg\max_{x \in X_{1:n}} \hat{f}(x)$.
  Consistency of $\tilde{x}$ [Abraham, Biau, Cadre, 04]
  We show that $\tilde{x}$ is also minimax-optimal.
- **Multiple modes:**
  Practical procedures (e.g. meanshift) are hard to analyze.
  Our procedure recovers *just* modes at optimal rates.

Sub-Outline:

- **k-NN density rates**
- Single mode rates
- Multiple modes rates

k-NN density estimate:

Define $r_k(x) \equiv$ distance from x to its kth neighbor in $X_{1:n}$.

$$f_k(x) \triangleq \frac{k}{n \cdot \mathrm{vol}\left(B(x, r_k(x))\right)} = \frac{k}{n \cdot v_d \cdot r_k(x)^d}.$$

k-NN density estimate:

Define $r_k(x) \equiv$ distance from x to its kth neighbor in $X_{1:n}$.

$$f_k(x) \triangleq \frac{k}{n \cdot \mathsf{vol}\left(B(x, r_k(x))\right)} = \frac{k}{n \cdot v_d \cdot r_k(x)^d}.$$

Finite sample rates for f_k:

W.p $> 1 - \delta$, simult. $\forall x \in$ supp(f), $\forall \epsilon > 0$,

$$\left(1 - \frac{C_{n,\delta}}{\sqrt{k}}\right)(f(x) - \epsilon) \leq f_k(x) \leq \left(1 + \frac{C_{n,\delta}}{\sqrt{k}}\right)(f(x) + \epsilon),$$

provided **log** $\mathbf{n} \lesssim \mathbf{k} \lesssim r(\epsilon, x)^d \cdot (\mathbf{f}(\mathbf{x}) - \epsilon) \cdot \mathbf{n}$.

$r(\epsilon, x) \equiv \sup \{r : \text{ on } B(x,r), f(\cdot) \approx f(x) + \epsilon\}$.

$\therefore$ **optimal (local) rates under smoothness conditions.**
If f is α-Hölder at x:

$$|f_k(x) - f(x)| = O\left(n^{-\alpha/(2\alpha+d)}\right), \quad \text{for } k = \Theta(n^{2\alpha/(2\alpha+d)}).$$

Finite sample rates for f_k:

W.p $> 1 - \delta$, simult. $\forall x \in$ supp(f), $\forall \epsilon > 0$,

$$\left(1 - \frac{C_{n,\delta}}{\sqrt{k}}\right)(f(x) - \epsilon) \leq f_k(x) \leq \left(1 + \frac{C_{n,\delta}}{\sqrt{k}}\right)(f(x) + \epsilon),$$

provided $\textbf{log } \mathbf{n} \lesssim \mathbf{k} \lesssim r(\epsilon, x)^d \cdot (\mathbf{f}(\mathbf{x}) - \epsilon) \cdot \mathbf{n}$.

$r(\epsilon, x) \equiv \sup\{r: \text{ on } B(x, r), f(\cdot) \approx f(x) + \epsilon\}$.

$\therefore$ **optimal (local) rates under smoothness conditions.**
If f is α-Hölder at x:

$$|f_k(x) - f(x)| = O\left(n^{-\alpha/(2\alpha+d)}\right), \quad \text{for } k = \Theta(n^{2\alpha/(2\alpha+d)}).$$

Finite sample rates for f_k:

W.p $> 1 - \delta$, simult. $\forall x \in$ supp(f), $\forall \epsilon > 0$,

$$\left(1 - \frac{C_{n,\delta}}{\sqrt{k}}\right)(f(x) - \epsilon) \leq f_k(x) \leq \left(1 + \frac{C_{n,\delta}}{\sqrt{k}}\right)(f(x) + \epsilon),$$

provided **log** $\mathbf{n} \lesssim \mathbf{k} \lesssim r(\epsilon, x)^d \cdot (\mathbf{f}(\mathbf{x}) - \epsilon) \cdot \mathbf{n}$.

$r(\epsilon, x) \equiv \sup\{r : \text{ on } B(x, r), f(\cdot) \approx f(x) + \epsilon\}$.

$\therefore$ **optimal (local) rates under smoothness conditions.**
If f is α-Hölder at x:

$$|f_k(x) - f(x)| = O\left(n^{-\alpha/(2\alpha+d)}\right), \quad \text{for } k = \Theta(n^{2\alpha/(2\alpha+d)}).$$

Sub-Outline:

- k-NN density rates
- **Single mode rates**
- Multiple modes rates

Most commonly studied

$\hat{x} = \arg\max_{x \in \mathbb{R}^d} f_n(x)$

Recursive estimates (One sample at a time)

[L. Devroye 79], [S. Tsybakov, 90 (optimal for Hölder classes.)]

Direct estimates (no density estimation)

$\tilde{x} = \arg\max_{x \in X_{1:n}} f_k(x) = \arg\min_{x \in X_{1:n}} r_k(x)$.
(Consistency, [Abraham, Biau, Cadre, 04])

Most commonly studied

$\hat{x} = \arg\max_{x \in \mathbb{R}^d} f_n(x)$

Recursive estimates (One sample at a time)

[L. Devroye 79], [S. Tsybakov, 90 (optimal for Hölder classes.)]

Direct estimates (no density estimation)

$\tilde{x} = \arg\max_{x \in X_{1:n}} f_k(x) = \arg\min_{x \in X_{1:n}} r_k(x)$.
(Consistency, [Abraham, Biau, Cadre, 04])

Most commonly studied

$\hat{x} = \arg\max_{x \in \mathbb{R}^d} f_n(x)$

Recursive estimates (One sample at a time)

[L. Devroye 79], [S. Tsybakov, 90 (optimal for Hölder classes.)]

Direct estimates (no density estimation)

$\tilde{x} = \arg\max_{x \in X_{1:n}} f_k(x) = \arg\min_{x \in X_{1:n}} r_k(x)$.
(Consistency, [Abraham, Biau, Cadre, 04])

A.1 (local): $x = \arg\max f(x)$, $\exists\, \nabla^2 f$ on $B(x)$, $\nabla^2 f(x) \prec 0$.
A.2 (global): level sets of f have single CC.

Theorem. Let $\tilde{x} = \arg\max_{x \in X_{1:n}} f_k(x)$. W.h.p. we have

$$\|\tilde{x} - x\| \lesssim k^{-1/4}, \quad \text{provided } \ln n \lesssim k \lesssim n^{4/(4+d)}.$$

Constants depend on $f(x)$ and $\nabla^2 f(x)$. (OPTIMAL, see Tsyb.90)

A.1 (local): $x = \arg\max f(x)$, $\exists\, \nabla^2 f$ on $B(x)$, $\nabla^2 f(x) \prec 0$.
A.2 (global): level sets of f have single CC.

Theorem. Let $\tilde{x} = \arg\max_{x \in X_{1:n}} f_k(x)$. W.h.p. we have

$$\|\tilde{x} - x\| \lesssim k^{-1/4}, \quad \text{provided } \ln n \lesssim k \lesssim n^{4/(4+d)}.$$

Constants depend on $f(x)$ and $\nabla^2 f(x)$. (OPTIMAL, see Tsyb.90)

Proof idea:

- There is a sample point at distance $\leq$ optimal rate.
- $\nabla^2 f(x) \prec 0$: $\exists$ a level set A_x:

$$c \left\| x - x' \right\|^2 \leq f(x) - f(x') \leq C \left\| x - x' \right\|^2 .$$

Proof idea:

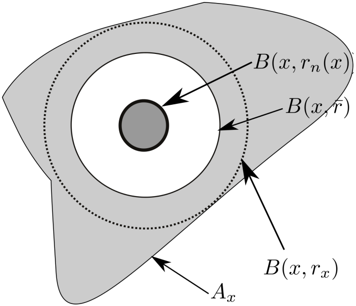
- There is a sample point at distance $\leq$ optimal rate.
- $\nabla^2 f(x) \prec 0 :$ $\exists$ a level set A_x:

$$c\,\|x - x'\|^2 \leq f(x) - f(x') \leq C\,\|x - x'\|^2.$$

Proof idea:

- There is a sample point at distance $\leq$ optimal rate.
- $\nabla^2 f(x) \prec 0$: $\exists$ a level set A_x:

$$c\|x - x'\|^2 \leq f(x) - f(x') \leq C\|x - x'\|^2.$$



Sub-Outline:

- k-NN density rates
- Single mode rates
- **Multiple modes rates**

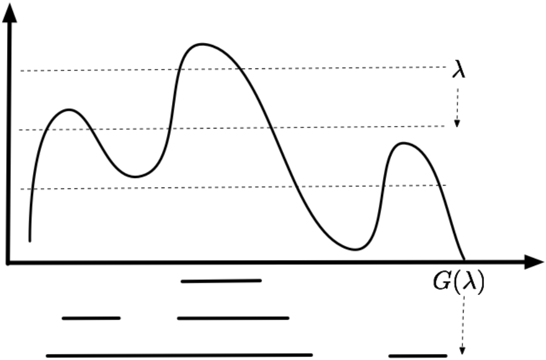
Setup:

Modes: $\mathcal{M} \equiv \{x : \exists r > 0, \forall x' \in B(x, r),\ f(x') < f(x)\}$.

A.1 (local) $\forall x \in \mathcal{M}$, $\exists\, \nabla^2 f$ on $B(x)$, $\nabla^2 f(x) \prec 0$.

A.2 (global) Any CC of any level set of f contains a mode in $\mathcal{M}$.

ALGO: As f_k goes down, pick a new mode as a new *bump* appears.



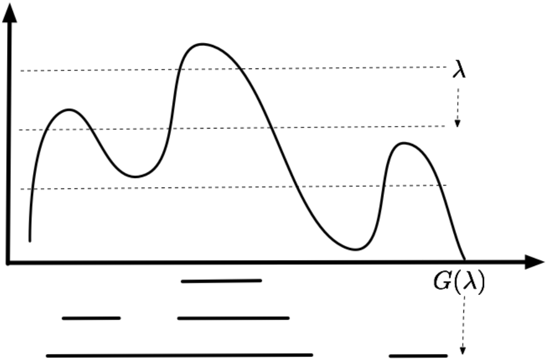
Identifying CCs of level sets:

CCs of subgraphs of a k-NN graph [Chau., Das., Kpo., v Lux., 14]

How to identify false modes in f_k?

Remove all *bumps* of height $\lesssim |f_k - f| \approx 1/\sqrt{k}$.

ALGO: As f_k goes down, pick a new mode as a new *bump* appears.



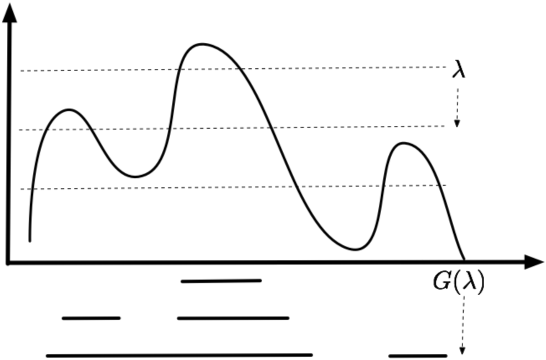
Identifying CCs of level sets:

CCs of subgraphs of a k-NN graph [Chau., Das., Kpo., v Lux., 14]

How to identify false modes in f_k?

Remove all *bumps* of height $\lesssim |f_k - f| \approx 1/\sqrt{k}$.

ALGO: As f_k goes down, pick a new mode as a new *bump* appears.



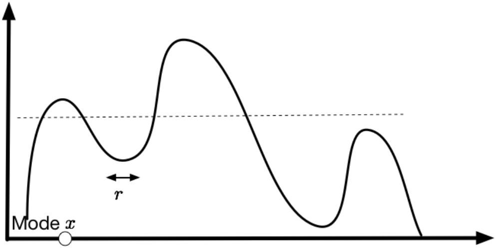
Identifying CCs of level sets:

CCs of subgraphs of a k-NN graph [Chau., Das., Kpo., v Lux., 14]

How to identify false modes in f_k?

Remove all *bumps* of height $\lesssim |f_k - f| \approx 1/\sqrt{k}$.

Identifying good modes



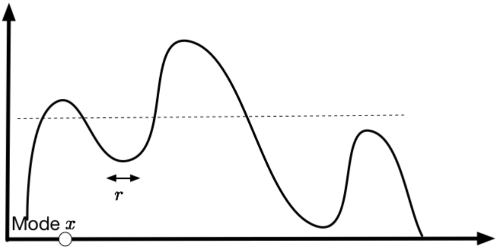
x is r-salient: separated from other modes by valley of thickness r.

Theorem. Suppose $x \in \mathcal{M}$ is r-salient. Let $n \geq N(x)$. W.h.p. $\exists \tilde{x} \in \mathcal{M}_n$ s.t.

$$\|\tilde{x} - x\| \lesssim k^{-1/4}, \quad \text{provided } \ln n / r^4 \lesssim k \lesssim n^{4/(4+d)}.$$

Constants depend on $f(x)$ and $\nabla^2 f(x)$.

Identifying good modes



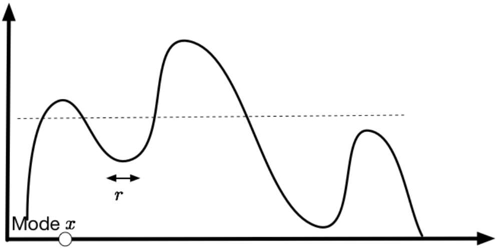
x is r-salient: separated from other modes by valley of thickness r.

Theorem. Suppose $x \in \mathcal{M}$ is r-salient. Let $n \geq N(x)$. W.h.p. $\exists \tilde{x} \in \mathcal{M}_n$ s.t.

$$\|\tilde{x} - x\| \lesssim k^{-1/4}, \quad \text{provided } \ln n / r^4 \lesssim k \lesssim n^{4/(4+d)}.$$

Constants depend on $f(x)$ and $\nabla^2 f(x)$.

Identifying good modes



x is r-salient: separated from other modes by valley of thickness r.

Theorem. Suppose $x \in \mathcal{M}$ is r-salient. Let $n \geq N(x)$. W.h.p. $\exists \tilde{x} \in \mathcal{M}_n$ s.t.

$$\|\tilde{x} - x\| \lesssim k^{-1/4}, \quad \text{provided } \ln n / r^4 \lesssim k \lesssim n^{4/(4+d)}.$$

Constants depend on $f(x)$ and $\nabla^2 f(x)$.

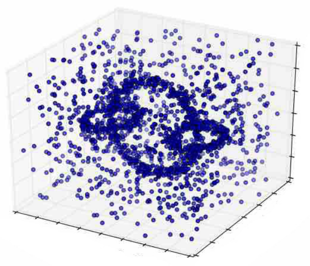
Pruning bad modes

Theorem 4. Suppose f is Lipschitz. Let $k \geq \ln n$.
All modes in $\mathcal{M}_n$ at f_k-level $\lambda > \lambda_k \approx 1/k$
can be assigned to *distinct* modes in $\mathcal{M}$ at f-level $\approx \lambda$.

Outline:

- How k-**NN graphs** relate to f.
  (with Chaudhuri, Dasgupta, von Luxburg, 2011, 2014)
- Estimating all **modes** of f.
  (with S. Dasgupta, 2014)
- Estimating all **modal sets** of f.
  (with H. Jiang, 2017)

Estimating **all modal sets** *of f.*



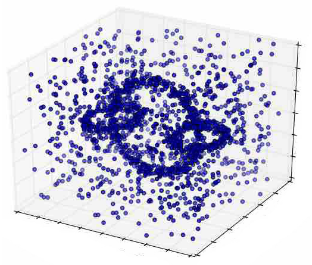
Here: **Regions** of locally high-density ...

Point-modes (0-dimensional), and more general connected sets.

Related to topological data analysis

Manifold + noise, low-dimensional ridge, ... etc

Estimating **all modal sets** *of* f.

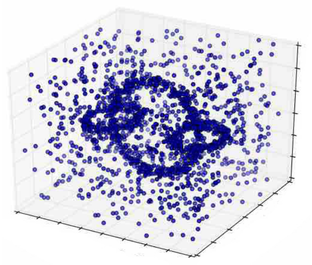


Here: **Regions** of locally high-density ...
Point-modes (0-dimensional), and more general connected sets.

Related to topological data analysis

Manifold + noise, low-dimensional ridge, ... etc

Estimating **all modal sets** *of f.*



Here: **Regions** of locally high-density ...
Point-modes (0-dimensional), and more general connected sets.

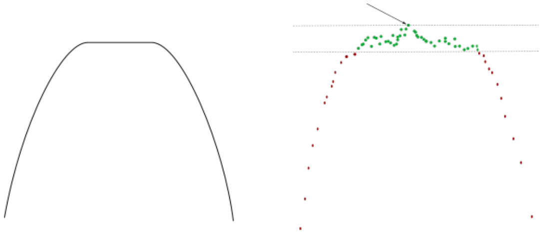
Related to topological data analysis

Manifold + noise, low-dimensional ridge, ... etc

Key changes to previous procedure:

Estimating general **modal-sets** M*:*

$M \equiv$ Region of $\mathbb{R}^d$ where f is locally maximal.
$\widehat{M} \equiv$ samples x s.t. $|\max f_k - f_k(x)| \approx \max f_k / \sqrt{k}$.



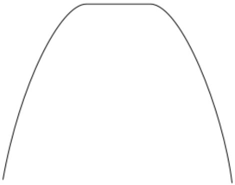
Needs local pruning! (Else some $\widehat{M}$ have wrong dimension).

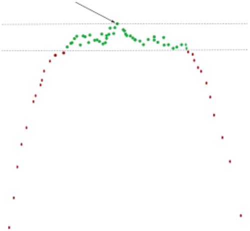
Key changes to previous procedure:

Estimating general **modal-sets** M*:*

$M \equiv$ Region of $\mathbb{R}^d$ where f is locally maximal.

$\widehat{M} \equiv$ samples x s.t. $|\max f_k - f_k(x)| \approx \max f_k/\sqrt{k}$.



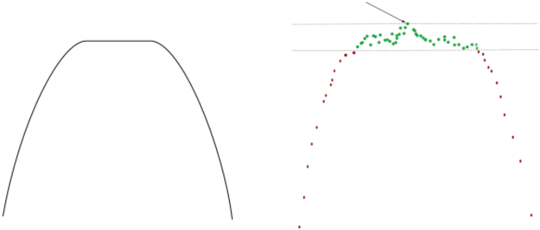


Needs local pruning! (Else some $\widehat{M}$ *have wrong dimension).*

Key changes to previous procedure:

Estimating general **modal-sets** M*:*

$M \equiv$ Region of $\mathbb{R}^d$ where f is locally maximal.
$\widehat{M} \equiv$ samples x s.t. $|\max f_k - f_k(x)| \approx \max f_k/\sqrt{k}$.

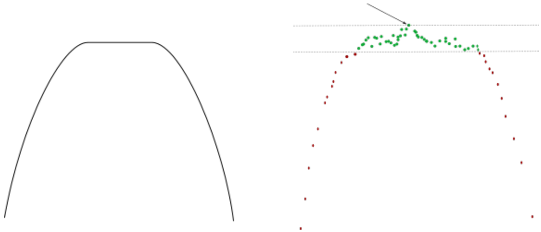


Needs local pruning! (Else some $\widehat{M}$ have wrong dimension).

Key changes to previous procedure:

Estimating general **modal-sets** M*:*

$M \equiv$ Region of $\mathbb{R}^d$ where f is locally maximal.
$\widehat{M} \equiv$ samples x s.t. $|\max f_k - f_k(x)| \approx \max f_k/\sqrt{k}$.



Needs local pruning! (Else some $\widehat{M}$ have wrong dimension).

Key change in analysis:

f might not be smooth at boundary of M

However, if f is uniformly continuous on some $B(M,r)$, then for all $x \in B(M,r)$,

$$L(d(x,M)) \leq f_M - f(x) \leq U(d(x,M))$$

for some $L(\cdot), U(\cdot)$ increasing on $[0,\infty)$, 0 at 0.

Key change in analysis:

f might not be smooth at boundary of M

However, if f is uniformly continuous on some $B(M, r)$, then for all $x \in B(M, r)$,

$$L(d(x, M)) \le f_M - f(x) \le U(d(x, M))$$

for some $L(\cdot), U(\cdot)$ increasing on $[0, \infty)$, 0 at 0.

Key change in analysis:

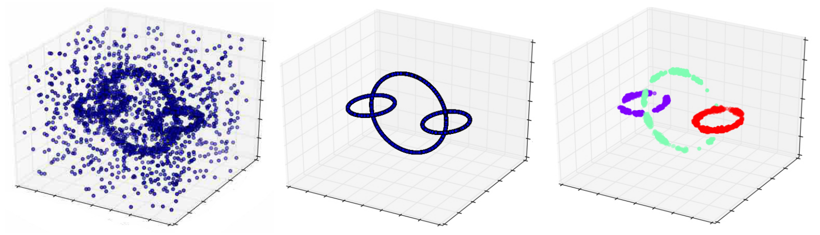
f might not be smooth at boundary of M

However, if f is uniformly continuous on some $B(M, r)$, then for all $x \in B(M, r)$,

$$L(d(x, M)) \leq f_M - f(x) \leq U(d(x, M))$$

for some $L(\cdot), U(\cdot)$ increasing on $[0, \infty)$, 0 at 0.

Consistency



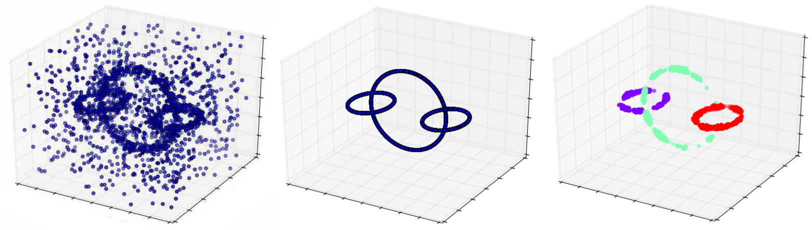
Theorem. The following holds w.h.p. Suppose M is r-salient. Let $n \geq N(M, r)$. We recover an $\widehat{M}$ s.t.

$$d_{\text{Hausdorff}}(\widehat{M}, M) \lesssim L^{-1}(f_M/\sqrt{k}),$$

$$\text{provided } \log \mathbf{n}/L^2(r) \lesssim \mathbf{k} \lesssim \left(U^{-1}(f_M/\sqrt{k})\right)^d \cdot \mathbf{n}.$$

Similar pruning guarantees with a bit more work.

Consistency



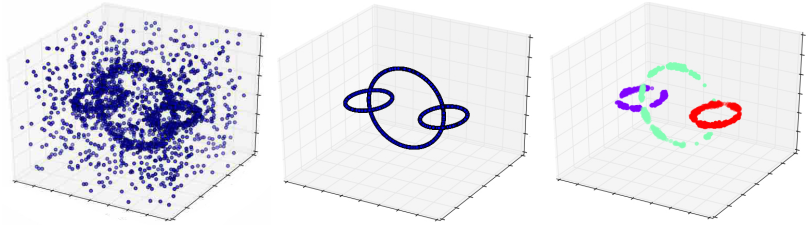
Theorem. The following holds w.h.p. Suppose M is r-salient. Let $n \geq N(M, r)$. We recover an $\widehat{M}$ s.t.

$$d_{\text{Hausdorff}}(\widehat{M}, M) \lesssim L^{-1}(f_M/\sqrt{k}),$$

$$\text{provided } \log \mathbf{n}/L^2(r) \lesssim \mathbf{k} \lesssim \left(U^{-1}(f_M/\sqrt{k})\right)^d \cdot \mathbf{n}.$$

Similar pruning guarantees with a bit more work.

Consistency



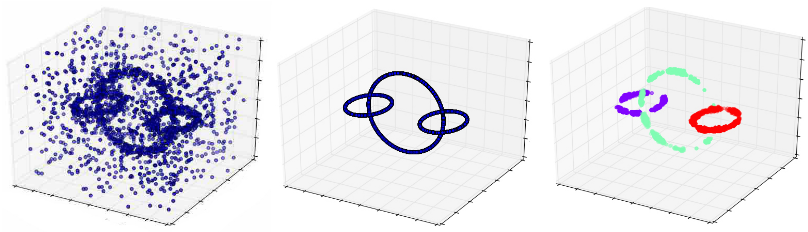
Theorem. The following holds w.h.p. Suppose M is r-salient. Let $n \geq N(M, r)$. We recover an $\widehat{M}$ s.t.

$$d_{\text{Hausdorff}}(\widehat{M}, M) \lesssim L^{-1}(f_M/\sqrt{k}),$$

$$\text{provided } \log \mathbf{n}/L^2(r) \lesssim \mathbf{k} \lesssim \left(U^{-1}(f_M/\sqrt{k})\right)^d \cdot \mathbf{n}.$$

Similar pruning guarantees with a bit more work.

Consistency



Theorem. The following holds w.h.p. Suppose M is r-salient. Let $n \geq N(M, r)$. We recover an $\widehat{M}$ s.t.

$$d_{\text{Hausdorff}}(\widehat{M}, M) \lesssim L^{-1}(f_M/\sqrt{k}),$$

provided $\log \mathbf{n}/L^2(r) \lesssim \mathbf{k} \lesssim \left(U^{-1}(f_M/\sqrt{k})\right)^d \cdot \mathbf{n}.$

Similar pruning guarantees with a bit more work.

Clustering procedure:

QuickShift $++$

- Estimate modal-sets $M_1, M_2, \ldots, M_K$;
- Assign every point x (by gradient ascent) to some M_i:
  Follow sample path $x_0 \to x_1 \to x_2 \ldots \rightsquigarrow M_i$,
  ... $x_{t+1} \equiv$ closest point to x_t s.t. $\hat{f}(x_{t+1}) > \hat{f}(x_t)$

Clustering procedure:

QuickShift $++$

- Estimate modal-sets $M_1, M_2, \ldots, M_K$;
- Assign every point x (by gradient ascent) to some M_i:
  Follow sample path $x_0 \rightarrow x_1 \rightarrow x_2 \ldots \rightsquigarrow M_i$,
  ... $x_{t+1} \equiv$ closest point to x_t s.t. $\hat{f}(x_{t+1}) > \hat{f}(x_t)$

Clustering procedure:

QuickShift $++$

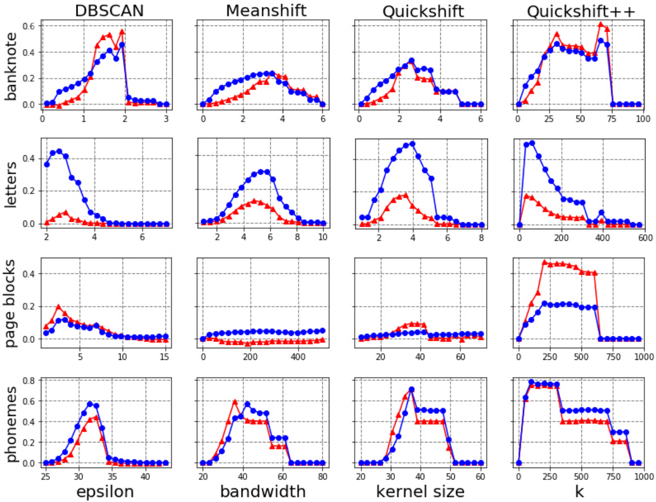
- Estimate modal-sets $M_1, M_2, \ldots, M_K$;
- Assign every point x (by gradient ascent) to some M_i:
  Follow sample path $x_0 \to x_1 \to x_2 \ldots \rightsquigarrow M_i$,
  ... $x_{t+1} \equiv$ closest point to x_t s.t. $\hat{f}(x_{t+1}) > \hat{f}(x_t)$

More experiments on UCI datasets:

Data/ Algo	DBSCAN	MnShift	QkShift	QkShift++
seeds	0.4473	**0.7319**	0.6715	**0.7261**
	0.4429	0.6769	0.6360	**0.7085**
phonemes	0.6333	0.5732	**0.7653**	**0.7663**
	0.7280	0.5396	**0.7954**	**0.8019**
banknote	0.5584	0.2434	0.3318	**0.6152**
	0.4594	0.2351	0.3397	**0.4866**
images	0.3313	0.3497	0.4077	**0.5359**
	0.5264	0.4656	0.5364	**0.6456**
letters	0.0460	0.1506	0.1335	**0.2128**
	0.2338	0.3457	**0.3706**	**0.4122**
page blocks	0.0132	0.0028	0.0925	**0.4727**
	0.0578	0.0526	0.0397	**0.2192**

Clustering scores are: Mutual information, and Rand-Index.

Sensitivity to tuning parameters.



(Blue) Mutual Information score, (Red) Rand-index score

We've started investigating **related applications ...**

- Medical imaging (detecting low-d structures).
- Image segmentation (detecting object boundaries).
- Internet of Things (outlier detection).

We've started investigating **related applications ...**

- Medical imaging (detecting low-d structures).
- Image segmentation (detecting object boundaries).
- Internet of Things (outlier detection).

Unsupervised Image Segmentation

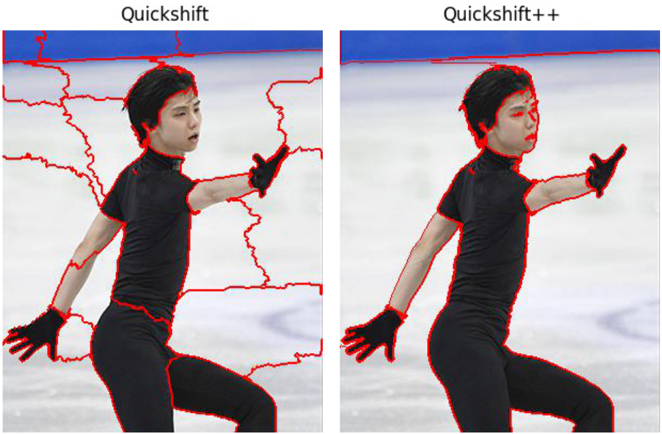


Figure: Best tradeoff between over and under segmentation.

Future questions:

- Adaptive (data-driven) choices of hyperparameters.
- High-dimensional clustering:
  Feature selection, *spectral and projection* methods.

That's all. Thanks!

Future questions:

- Adaptive (data-driven) choices of hyperparameters.
- High-dimensional clustering:
  Feature selection, *spectral and projection* methods.

That's all. Thanks!

Future questions:

- Adaptive (data-driven) choices of hyperparameters.
- High-dimensional clustering:
  Feature selection, *spectral and projection* methods.

That's all. Thanks!

Future questions:

- Adaptive (data-driven) choices of hyperparameters.
- High-dimensional clustering:
  Feature selection, *spectral and projection* methods.

That's all. Thanks!