Pruning Nearest Neighbor Cluster Trees

Samory Kpotufe
Max Planck Institute for Intelligent Systems
Tuebingen, Germany

Joint work with **Ulrike von Luxburg**

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

We'll discuss:

- An interesting notion of "clusters" (Hartigan 1982):
  Clusters are regions of high density of the data distribution μ.
- The richness of k-NN graphs G_n:
  Subgraphs of G_n encode the underlying cluster structure of μ.
- How to identify false cluster structures:
  A simple pruning procedure with strong guarantees (a first).

General motivation

More understanding of clustering

- Density yields intuitive (and clean) notion of clusters.
- Clusters take any shape $\implies$ reveals complexity of clustering?
- Popular approches (e.g. DBscan, single linkage) are density-based methods.

More understanding of k-NN graphs

These appear everywhere in various forms!

General motivation

More understanding of clustering

- Density yields intuitive (and clean) notion of clusters.
- Clusters take any shape $\implies$ reveals complexity of clustering?
- Popular approches (e.g. DBscan, single linkage) are density-based methods.

More understanding of k-NN graphs

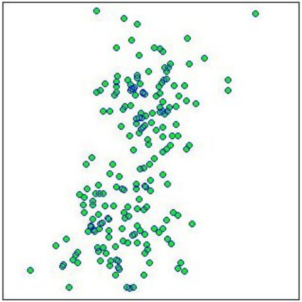
These appear everywhere in various forms!

Outline

- **Density-based clustering**
- Richness of k-NN graphs
- Guaranteed removal of false clusters

Density based clustering

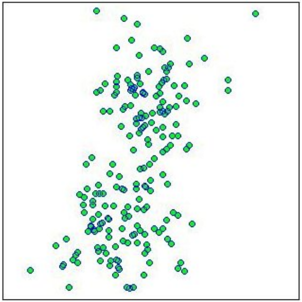
Given: data from some unknown distribution.
Goal: discover "true" high density regions.



Resolution matters!

Density based clustering

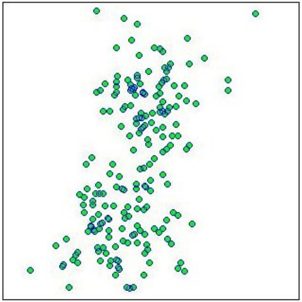
Given: data from some unknown distribution.
Goal: discover "true" high density regions.



Resolution matters!

Density based clustering

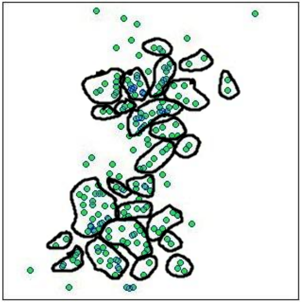
Given: data from some unknown distribution.
Goal: discover “true” high density regions.



Resolution matters!

Density based clustering

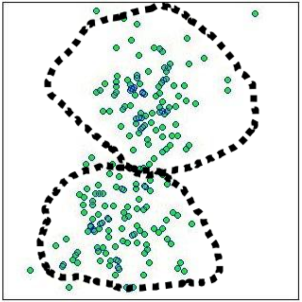
Given: data from some unknown distribution.
Goal: discover "true" high density regions.



Resolution matters!

Density based clustering

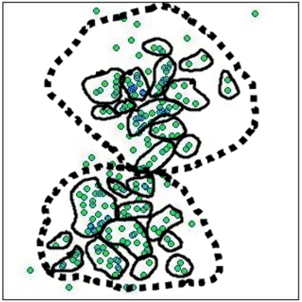
Given: data from some unknown distribution.
Goal: discover "true" high density regions.



Resolution matters!

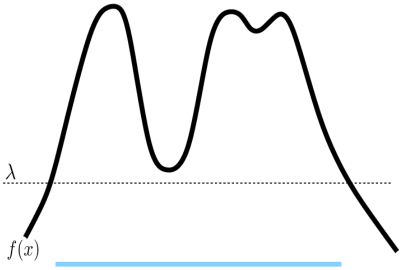
Density based clustering

Given: data from some unknown distribution.
Goal: discover "true" high density regions.



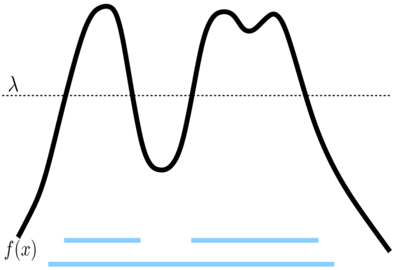
Resolution matters!

Density based clustering



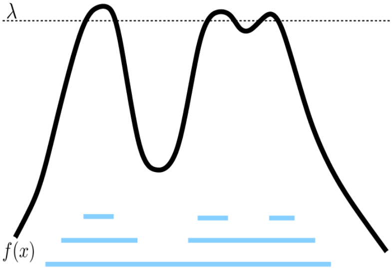
Clusters are $G(\lambda) \equiv$ CCs of $\mathcal{L}_\lambda \doteq \{x : f(x) \geq \lambda\}$.

Density based clustering



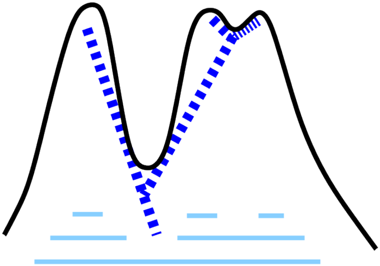
Clusters are $G(\lambda) \equiv$ CCs of $\mathcal{L}_\lambda \doteq \{x : f(x) \geq \lambda\}$.

Density based clustering



Clusters are $G(\lambda) \equiv$ CCs of $\mathcal{L}_\lambda \doteq \{x : f(x) \geq \lambda\}$.

Density based clustering



The cluster tree of f is the infinite hierarchy $\{G(\lambda)\}_{\lambda \geq 0}$.

Formal estimation problem:

Given: n i.i.d. samples $\mathbf{X} = \{x_i\}_{i\in[n]}$ from dist. with density f.

Clustering outputs: A hierarchy $\{G_n(\lambda)\}_{\lambda\geq 0}$ of subsets of $\mathbf{X}$.

We at least want consistency, i.e. for any $\lambda > 0$

$$\mathbb{P}\left(\text{Disjoint } A, A' \in G(\lambda) \text{ are in disjoint empirical clusters }\right) \to 1.$$

Formal estimation problem:

Given: n i.i.d. samples $\mathbf{X} = \{x_i\}_{i\in[n]}$ from dist. with density f.

Clustering outputs: A hierarchy $\{G_n(\lambda)\}_{\lambda\geq 0}$ of subsets of $\mathbf{X}$.

We at least want consistency, i.e. for any $\lambda > 0$

$$\mathbb{P}\left(\text{Disjoint } A, A' \in G(\lambda) \text{ are in disjoint empirical clusters }\right) \to 1.$$

Formal estimation problem:

Given: n i.i.d. samples $\mathbf{X} = \{x_i\}_{i\in[n]}$ from dist. with density f.

Clustering outputs: A hierarchy $\{G_n(\lambda)\}_{\lambda\geq 0}$ of subsets of $\mathbf{X}$.

We at least want consistency, i.e. for any $\lambda > 0$

$$\mathbb{P}\left(\text{Disjoint } A, A' \in G(\lambda) \text{ are in disjoint empirical clusters }\right) \to 1.$$

Formal estimation problem:

Given: n i.i.d. samples $\mathbf{X} = \{x_i\}_{i \in [n]}$ from dist. with density f.

Clustering outputs: A hierarchy $\{G_n(\lambda)\}_{\lambda \geq 0}$ of subsets of $\mathbf{X}$.

We at least want consistency, i.e. for any $\lambda > 0$

$$\mathbb{P}\left(\text{Disjoint } A, A' \in G(\lambda) \text{ are in disjoint empirical clusters }\right) \to 1.$$

A good procedure should satisfy:

Consistency!

Every level should be recovered for sufficiently large n.

Finite sample behavior:

- **Fast discovery of real clusters.**
- **"No false clusters !!!"**

A good procedure should satisfy:

Consistency!

Every level should be recovered for sufficiently large n.

Finite sample behavior:

- **Fast discovery of real clusters.**
- "No false clusters !!!"

A good procedure should satisfy:

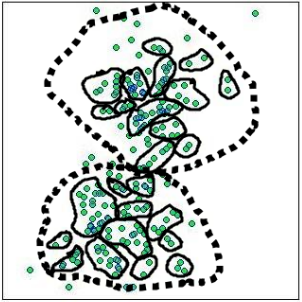
Consistency!

Every level should be recovered for sufficiently large n.

Finite sample behavior:

- **Fast discovery of real clusters.**
- **“No false clusters !!!”**

Earlier example is sampled from a **bi-modal** mixture of Gaussians!!!



My visual procedure yields false clusters at low resolution. ☹

What we'll show:

k-NN graphs guarantees

- **Finite sample:** Salient clusters recovered as subgraphs.
- **Consistency:** All clusters eventually recovered.

Generic pruning guarantees:

- **Finite sample:** No false clusters + salient clusters remain.
- **Consistency:** Pruned tree remains a consistent estimator.

What we'll show:

k-NN graphs guarantees

- **Finite sample:** Salient clusters recovered as subgraphs.
- **Consistency:** All clusters eventually recovered.

Generic pruning guarantees:

- **Finite sample:** No false clusters + salient clusters remain.
- **Consistency:** Pruned tree remains a consistent estimator.

What we'll show:

k-NN graphs guarantees

- **Finite sample:** Salient clusters recovered as subgraphs.
- **Consistency:** All clusters eventually recovered.

Generic pruning guarantees:

- **Finite sample:** No false clusters + salient clusters remain.
- **Consistency:** Pruned tree remains a consistent estimator.

What we'll show:

k-NN graphs guarantees

- **Finite sample:** Salient clusters recovered as subgraphs.
- **Consistency:** All clusters eventually recovered.

Generic pruning guarantees:

- **Finite sample:** No false clusters + salient clusters remain.
- **Consistency:** Pruned tree remains a consistent estimator.

What we'll show:

k-NN graphs guarantees

- **Finite sample:** Salient clusters recovered as subgraphs.
- **Consistency:** All clusters eventually recovered.

Generic pruning guarantees:

- **Finite sample:** No false clusters + salient clusters remain.
- **Consistency:** Pruned tree remains a consistent estimator.

What was known:

People you might look up:

Wasserman, Tsybakov, Wishart, Rinaldo, Nugent, Stueltze, Rigollet, Wong, Lane, Dasgupta, Chauduri, Maeir, von Luxburg, Steinwart ...

What was known:

People you might look up:

Wasserman, Tsybakov, Wishart, Rinaldo, Nugent, Stueltze, Rigollet, Wong, Lane, Dasgupta, Chauduri, Maeir, von Luxburg, Steinwart ...

What was known:

Consistency

- $(f_n \to f) \implies$ (cluster tree of $f_n \to$ cluster tree of f). ☺
  No known practical estimators. ☹
- Various practical estimators of a **single** level set.
  Can these be extended to all levels at once?
- **Recent:** First consistent practical estimator (Ch. and Das).
  A generalization of single linkage (by Wishart) ☺

What was known:

Consistency

- $(f_n \to f) \implies$ (cluster tree of $f_n \to$ cluster tree of f). ☺
  No known practical estimators. ☹
- Various practical estimators of a **single** level set.
  Can these be extended to all levels at once?
- **Recent:** First consistent practical estimator (Ch. and Das).
  A generalization of single linkage (by Wishart) ☺

What was known:

Consistency

- $(f_n \to f) \implies$ (cluster tree of $f_n \to$ cluster tree of f).
  No known practical estimators.
- Various practical estimators of a **single** level set.
  Can these be extended to all levels at once?
- **Recent:** First consistent practical estimator (Ch. and Das).
  A generalization of single linkage (by Wishart)

What was known:

Consistency

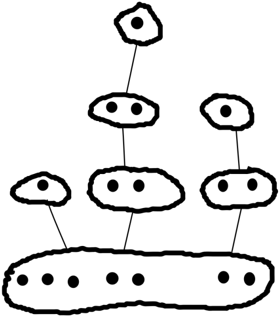
- $(f_n \to f) \implies$ (cluster tree of $f_n \to$ cluster tree of f). ☺
  No known practical estimators. ☹
- Various practical estimators of a **single** level set.
  Can these be extended to all levels at once?
- **Recent:** First consistent practical estimator (Ch. and Das).
  A generalization of single linkage (by Wishart) ☺

What was known:

Consistency

- $(f_n \to f) \implies$ (cluster tree of $f_n \to$ cluster tree of f). ☺
  No known practical estimators. ☹
- Various practical estimators of a **single** level set.
  Can these be extended to all levels at once?
- **Recent:** First consistent practical estimator (Ch. and Das).
  A generalization of single linkage (by Wishart) ☺

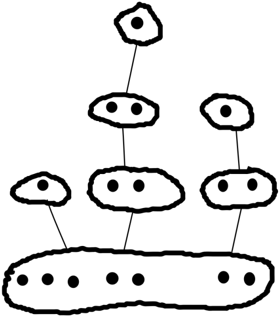
What was known:



Empirical tree contains good clusters ... but which? ☹

We need pruning guarantees!

What was known:



Empirical tree contains good clusters ... but which? ☹

We need pruning guarantees!

What was known:

Pruning

Consisted of removing **small** clusters!

Problem: Not all false clusters are "small"!!

What was known:

Pruning

Consisted of removing **small** clusters!

Problem: Not all false clusters are "small"!!

What was known:

Pruning

Consisted of removing **small** clusters!

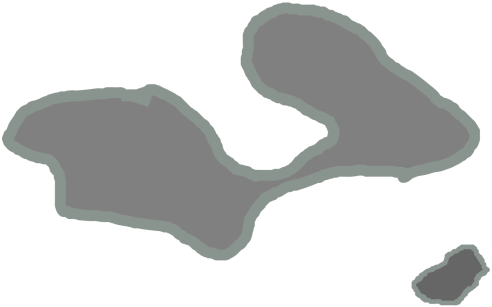
Problem: Not all false clusters are "small"!!

What was known:

Pruning

Consisted of removing **small** clusters!

Problem: Not all false clusters are "small"!!

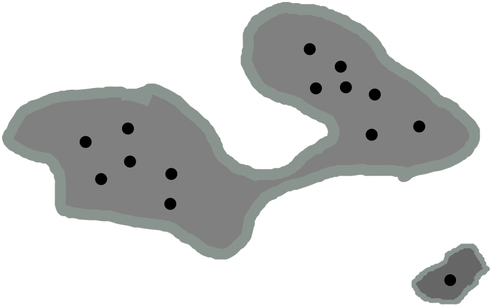


What was known:

Pruning

Consisted of removing **small** clusters!

Problem: Not all false clusters are "small"!!



Outline

- Ground-truth: Density-based clustering
- **Richness of k-NN graphs**
- Guaranteed removal of false clusters

Richness of k-NN graphs

k-**NN density estimate:** $f_n(x) \doteq k/n \cdot \mathrm{vol}(B_{k,n}(x))$.

Procedure: Remove X_i from G_n in increasing order of $f_n(X_i)$.

Level λ of the tree: $G_n(\lambda) \equiv$ subgraph with X_i s.t. $f_n(X_i) \geq \lambda$.

Richness of k-NN graphs

k-NN density estimate: $f_n(x) \doteq k/n \cdot \text{vol}(B_{k,n}(x))$.

Procedure: Remove X_i from G_n in increasing order of $f_n(X_i)$.

Level λ of the tree: $G_n(\lambda) \equiv$ subgraph with X_i s.t. $f_n(X_i) \geq \lambda$.

Richness of k-NN graphs

k-**NN density estimate:** $f_n(x) \doteq k/n \cdot \text{vol}(B_{k,n}(x))$.

Procedure: Remove X_i from G_n in increasing order of $f_n(X_i)$.

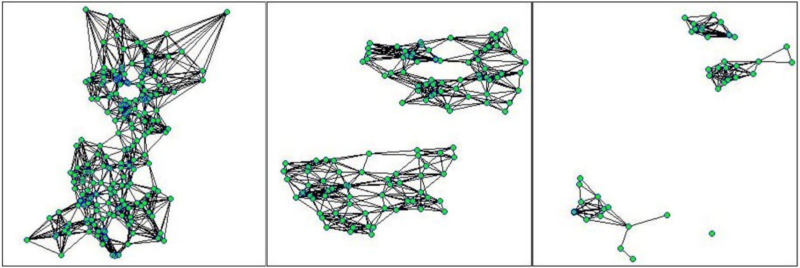
Level λ of the tree: $G_n(\lambda) \equiv$ subgraph with X_i s.t. $f_n(X_i) \geq \lambda$.

Richness of k-NN graphs

k-**NN density estimate:** $f_n(x) \doteq k/n \cdot \mathrm{vol}(B_{k,n}(x))$.

Procedure: Remove X_i from G_n in increasing order of $f_n(X_i)$.

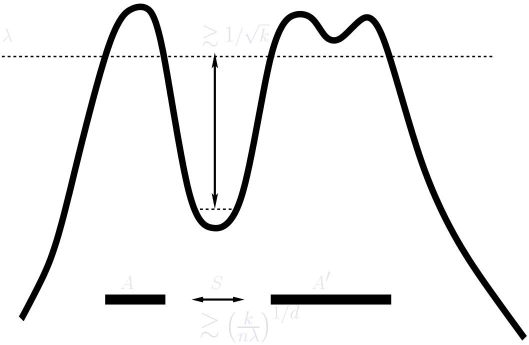
Level λ of the tree: $G_n(\lambda) \equiv$ subgraph with X_i s.t. $f_n(X_i) \geq \lambda$.



Sample from 2-modes mixture of gaussians

Theorem I:

Let $\log n \lesssim k \lesssim n^{1/O(d)}$:

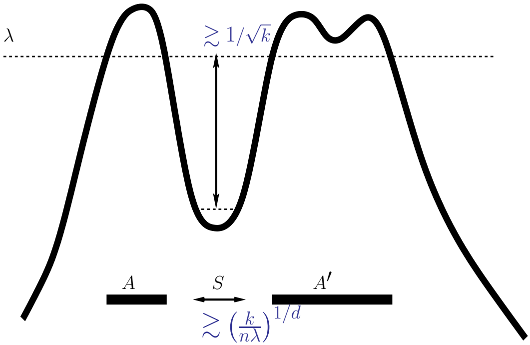


All such $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint CCs of $G_n(\lambda - O(1/\sqrt{k}))$.

Assumptions: $f(x) \leq F$ and $\forall x, x'$, $|f(x) - f(x')| \leq L \|x - x'\|^{\alpha}$.

Theorem I:

Let $\log n \lesssim k \lesssim n^{1/O(d)}$:

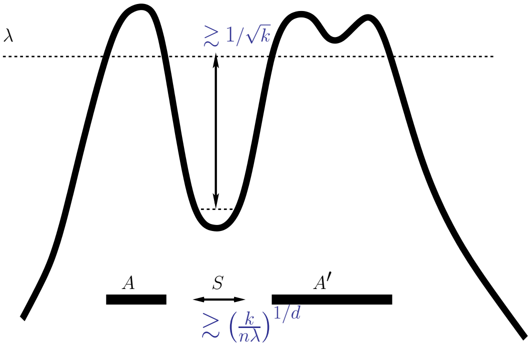


All such $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint CCs of $G_n(\lambda - O(1/\sqrt{k}))$.

Assumptions: $f(x) \leq F$ and $\forall x, x'$, $|f(x) - f(x')| \leq L \|x - x'\|^{\alpha}$.

Theorem I:

Let $\log n \lesssim k \lesssim n^{1/O(d)}$:

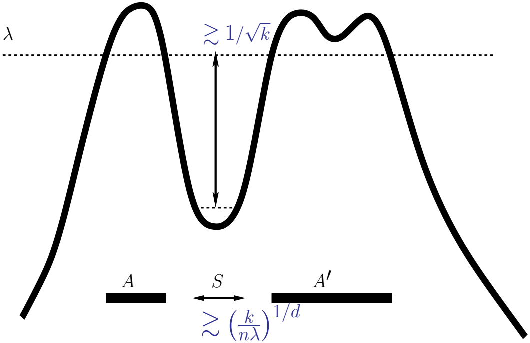


All such $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint CCs of $G_n(\lambda - O(1/\sqrt{k}))$.

Assumptions: $f(x) \leq F$ and $\forall x, x'$, $|f(x) - f(x')| \leq L \|x - x'\|^\alpha$.

Note on key quantities:

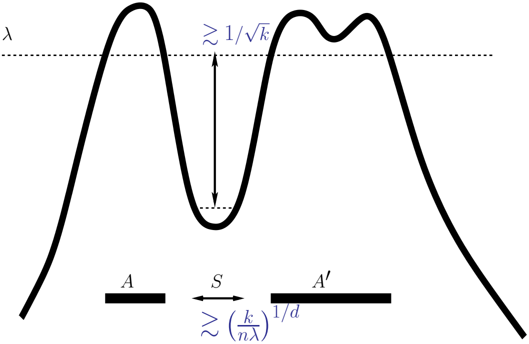
- $1/\sqrt{k} \gtrsim$ (density estimation error on samples X_i).
- $(k/n\lambda)^{1/d} \gtrsim$ (k-NN distances of X_i in $\mathcal{L}_\lambda$).



Consistency: both quantities $\to 0$, so eventually $A_n \cap A'_n = \emptyset$.

Note on key quantities:

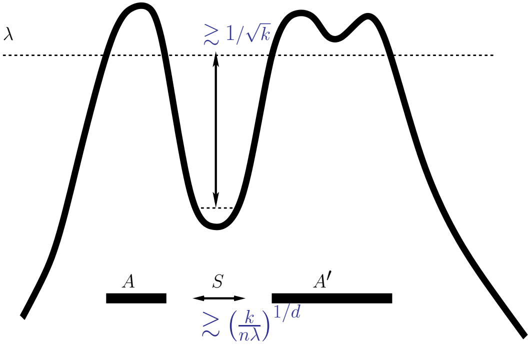
- $1/\sqrt{k} \gtrsim$ (density estimation error on samples X_i).
- $(k/n\lambda)^{1/d} \gtrsim$ (k-NN distances of X_i in $\mathcal{L}_\lambda$).



Consistency: both quantities $\to 0$, so eventually $A_n \cap A'_n = \emptyset$.

Note on key quantities:

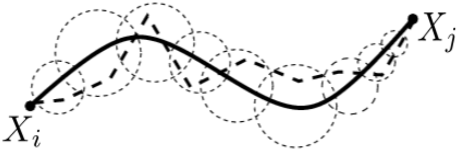
- $1/\sqrt{k} \gtrsim$ (density estimation error on samples X_i).
- $(k/n\lambda)^{1/d} \gtrsim$ (k-NN distances of X_i in $\mathcal{L}_\lambda$).



Consistency: both quantities $\to 0$, so eventually $A_n \cap A'_n = \emptyset$.

Main technicality:

Showing that $A \cap \mathbf{X}$ remains connected in $G_n(\lambda - O(1/\sqrt{k}))$.



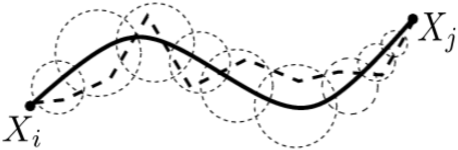
Cover high density path with balls $\{B_t\}$

- B_t's have to be large so they contain points.
- B_t's have to be small so points are connected.

So let B_t have mass about k/n!

Main technicality:

Showing that $A \cap \mathbf{X}$ remains connected in $G_n(\lambda - O(1/\sqrt{k}))$.



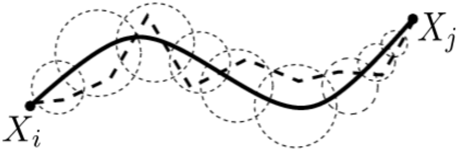
Cover high density path with balls $\{B_t\}$

- B_t's have to be large so they contain points.
- B_t's have to be small so points are connected.

So let B_t have mass about k/n!

Main technicality:

Showing that $A \cap \mathbf{X}$ remains connected in $G_n(\lambda - O(1/\sqrt{k}))$.



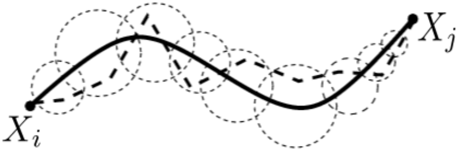
Cover high density path with balls $\{B_t\}$

- B_t's have to be large so they contain points.
- B_t's have to be small so points are connected.

So let B_t have mass about k/n!

Main technicality:

Showing that $A \cap \mathbf{X}$ remains connected in $G_n(\lambda - O(1/\sqrt{k}))$.



Cover high density path with balls $\{B_t\}$

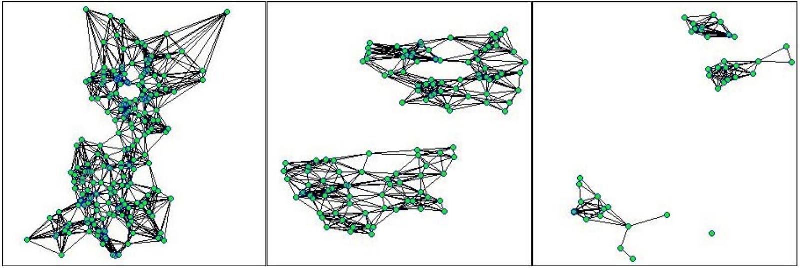
- B_t's have to be large so they contain points.
- B_t's have to be small so points are connected.

So let B_t have mass about k/n!

Outline

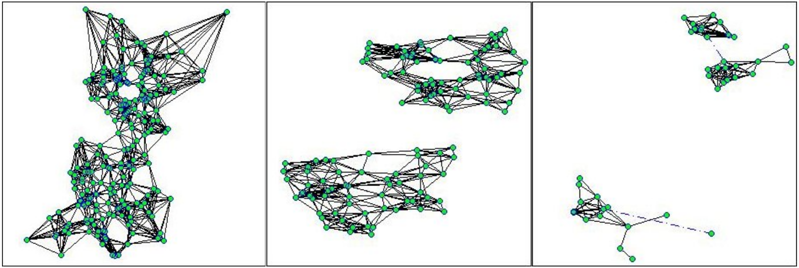
- Ground-truth: Density-based clustering
- Richness of k-NN graphs
- **Guaranteed removal of false clusters**

Guaranteed removal of false clusters



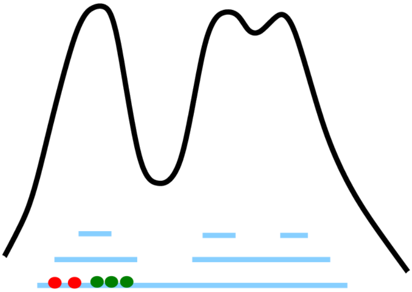
Sample from 2-modes mixture of gaussians

Guaranteed removal of false clusters



Sample from 2-modes mixture of gaussians

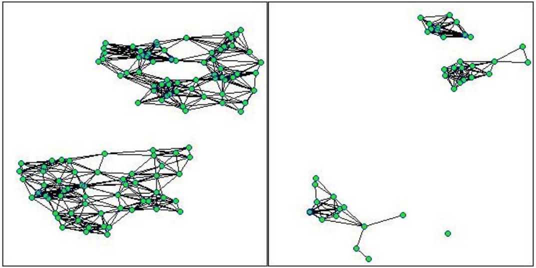
What are false clusters?



Intuitively:

A_n and A'_n in $\mathbf{X}$ should be in one (empirical) cluster if they are in the same (true) cluster at every level containing $A_n \cup A'_n$.

Pruning Intuition:
key connecting points are missing!!!

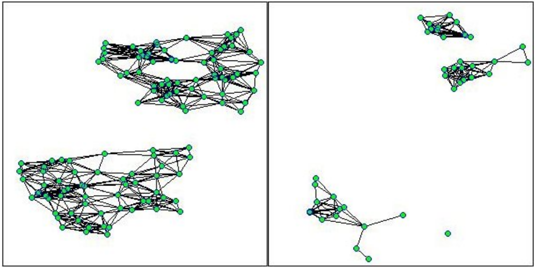


Sample from 2-modes mixture of gaussians

Pruning: Connect $G_n(0)$.
Re-connect A_n, A'_n in $G_n(\lambda_n)$ if they are connected in $G_n(\lambda_n - \tilde{\epsilon})$.
How do we set $\tilde{\epsilon}$?

Pruning Intuition:

key connecting points are missing!!!



Sample from 2-modes mixture of gaussians

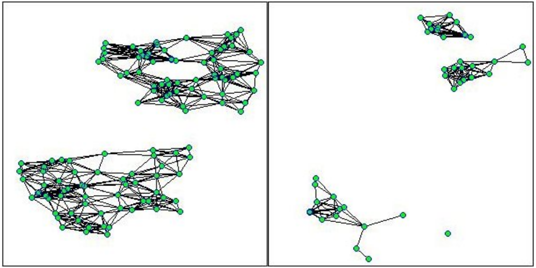
Pruning: Connect $G_n(0)$.

Re-connect A_n, A'_n in $G_n(\lambda_n)$ if they are connected in $G_n(\lambda_n - \tilde{\epsilon})$.

How do we set $\tilde{\epsilon}$?

Pruning Intuition:

key connecting points are missing!!!



Sample from 2-modes mixture of gaussians

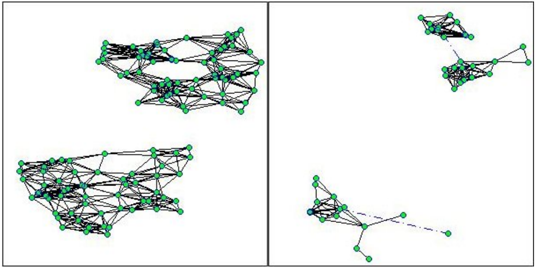
Pruning: Connect $G_n(0)$.

Re-connect A_n, A'_n in $G_n(\lambda_n)$ if they are connected in $G_n(\lambda_n - \tilde{\epsilon})$.

How do we set $\tilde{\epsilon}$?

Pruning Intuition:

key connecting points are missing!!!



Sample from 2-modes mixture of gaussians

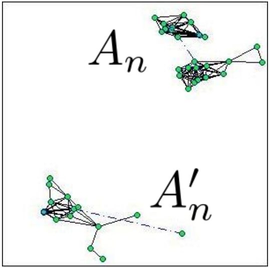
Pruning: Connect $G_n(0)$.

Re-connect A_n, A'_n in $G_n(\lambda_n)$ if they are connected in $G_n(\lambda_n - \tilde{\epsilon})$.

How do we set $\tilde{\epsilon}$?

Theorem II:

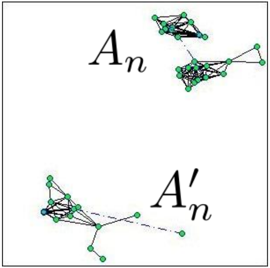
Suppose $\tilde{\epsilon} \gtrsim 1/\sqrt{k}$.



- A_n and A'_n belong to disjoint A and A' in some $G(\lambda)$.
- $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint A_n and A'_n of $G_n(\lambda - O(1/\sqrt{k}))$.
- $(\tilde{\epsilon}, k, n)$-salient modes map 1-1 to leaves of empirical tree.

Theorem II:

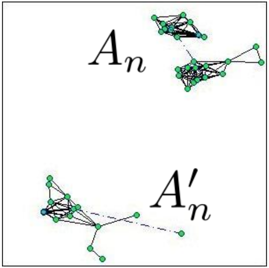
Suppose $\tilde{\epsilon} \gtrsim 1/\sqrt{k}$.

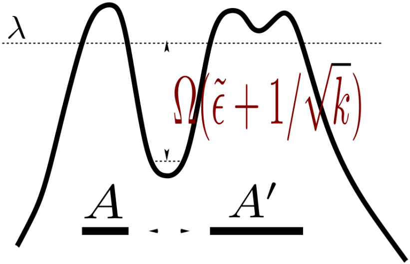


- A_n and A'_n belong to disjoint A and A' in some $G(\lambda)$.
- $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint A_n and A'_n of $G_n(\lambda - O(1/\sqrt{k}))$.
- $(\tilde{\epsilon}, k, n)$-salient modes map 1-1 to leaves of empirical tree.

Theorem II:

Suppose $\tilde{\epsilon} \gtrsim 1/\sqrt{k}$.

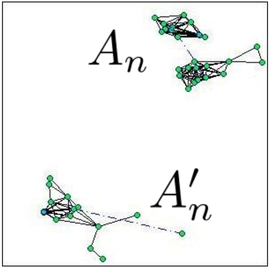


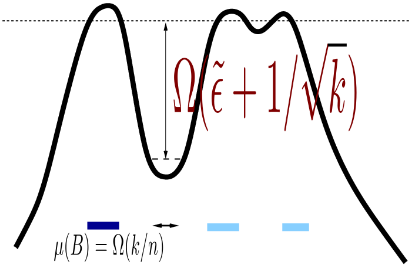


- A_n and A'_n belong to disjoint A and A' in some $G(\lambda)$.
- $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint A_n and A'_n of $G_n(\lambda - O(1/\sqrt{k}))$.
- $(\tilde{\epsilon}, k, n)$-salient modes map 1-1 to leaves of empirical tree.

Theorem II:

Suppose $\tilde{\epsilon} \gtrsim 1/\sqrt{k}$.

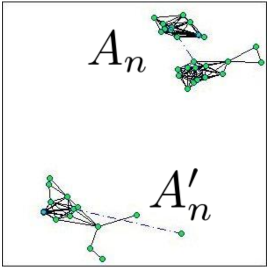


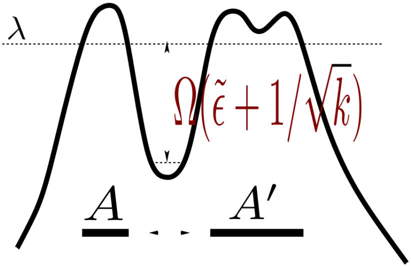


- A_n and A'_n belong to disjoint A and A' in some $G(\lambda)$.
- $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ belong to disjoint A_n and A'_n of $G_n(\lambda - O(1/\sqrt{k}))$.
- $(\tilde{\epsilon}, k, n)$-salient modes map 1-1 to leaves of empirical tree.

Consistency even after pruning:

We just require $\tilde{\epsilon} \to 0$ *as* $n \to \infty$.





Some last technical points:

[Ch. and Das. 2010] seem to be first to allow any cluster shape besides mild requirements on envelopes of clusters.

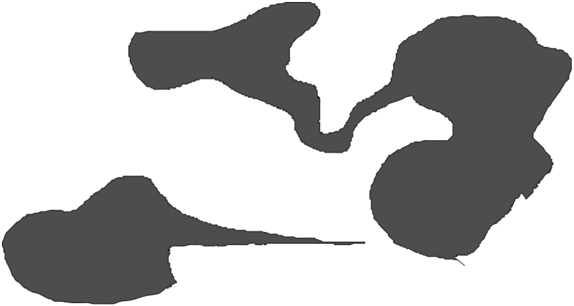
We allow any cluster shape up to smoothness of f and can explicitely relate empirical clusters to true clusters!

Some last technical points:

[Ch. and Das. 2010] seem to be first to allow any cluster shape besides mild requirements on envelopes of clusters.

We allow any cluster shape up to smoothness of f and can explicitely relate empirical clusters to true clusters!

Some last technical points:



[Ch. and Das. 2010] seem to be first to allow any cluster shape besides mild requirements on envelopes of clusters.

We allow any cluster shape up to smoothness of f and can explicitely relate empirical clusters to true clusters!

We have thus discussed:

- Density based clustering - Hartigan 1982).
- The richness of k-NN graphs G_n.
  Subgraphs of G_n consistently recover cluster tree of μ.
- Guaranteed pruning of false clusters.
  While discovering salient clusters and maintaining consistency!

We have thus discussed:

- Density based clustering - Hartigan 1982).
- The richness of k-NN graphs G_n.
  Subgraphs of G_n consistently recover cluster tree of μ.
- Guaranteed pruning of false clusters.
  While discovering salient clusters and maintaining consistency!

We have thus discussed:

- Density based clustering - Hartigan 1982).
- The richness of k-NN graphs G_n.
  Subgraphs of G_n consistently recover cluster tree of μ.
- Guaranteed pruning of false clusters.
  While discovering salient clusters and maintaining consistency!

We have thus discussed:

- Density based clustering - Hartigan 1982).
- The richness of k-NN graphs G_n.
  Subgraphs of G_n consistently recover cluster tree of μ.
- Guaranteed pruning of false clusters.
  While discovering salient clusters and maintaining consistency!

Thank you! ☺