Adaptive Clustering and Beyond

Ingo Steinwart
joint work with P. Thomann and N. Schmid

ISA, Stuttgart

Leuven, January 28th, 2016

Informal Description of the Goal

Some Descriptions of Clustering

- Hartigan (1975)
  "Clustering is the grouping of similar objects."

Informal Description of the Goal

Some Descriptions of Clustering

- Hartigan (1975)
  "Clustering is the grouping of similar objects."
- Wikipedia (2011)
  "Clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense."

Informal Description of the Goal

Some Descriptions of Clustering

- Hartigan (1975)
  "Clustering is the grouping of similar objects."
- Wikipedia (2011)
  "Clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense."
- Kaufman and Rousseeuw (1990)
  "Cluster analysis is the art of finding groups in data"

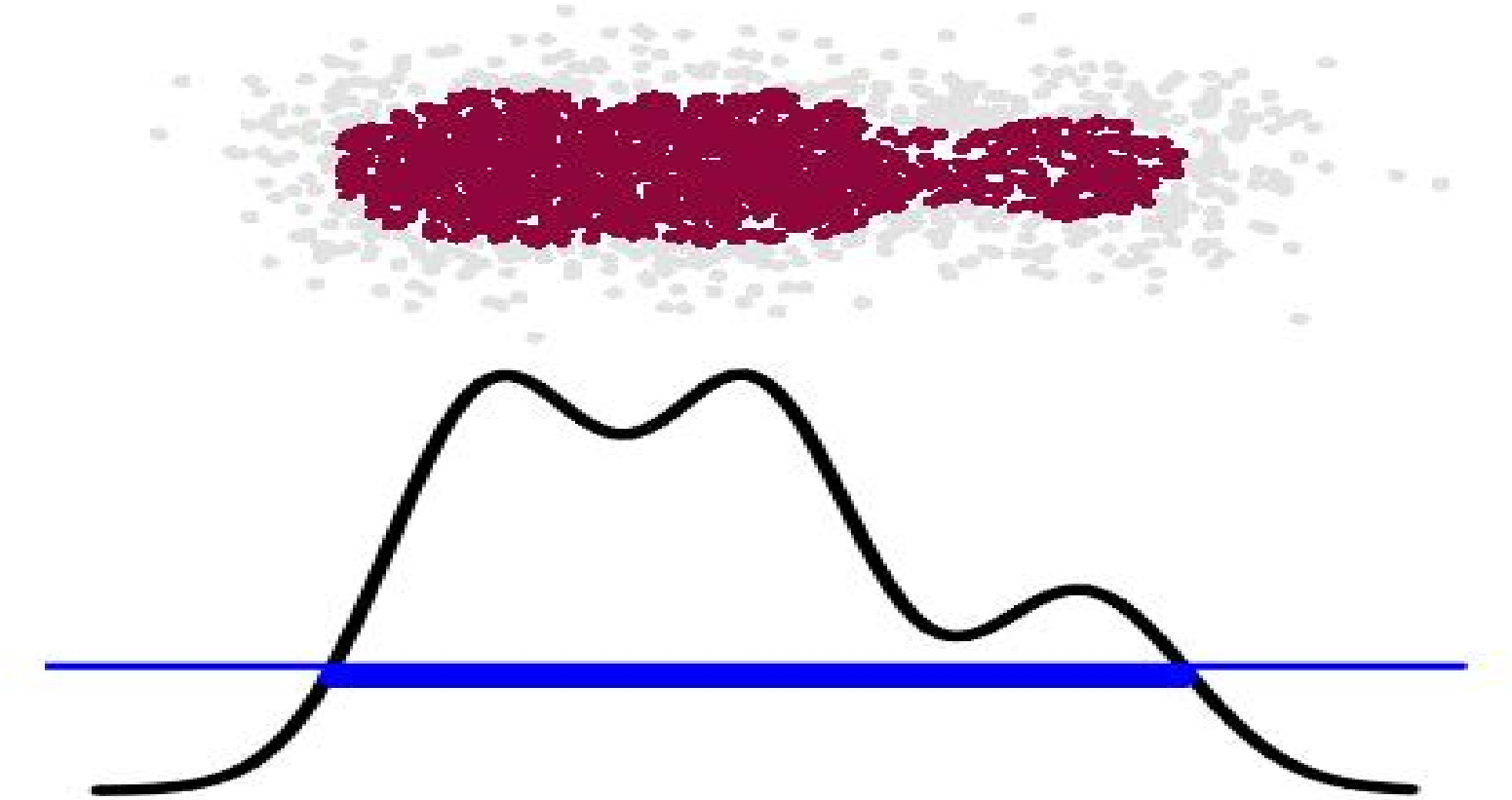
One Description of Clusters

Hartigan, 1975

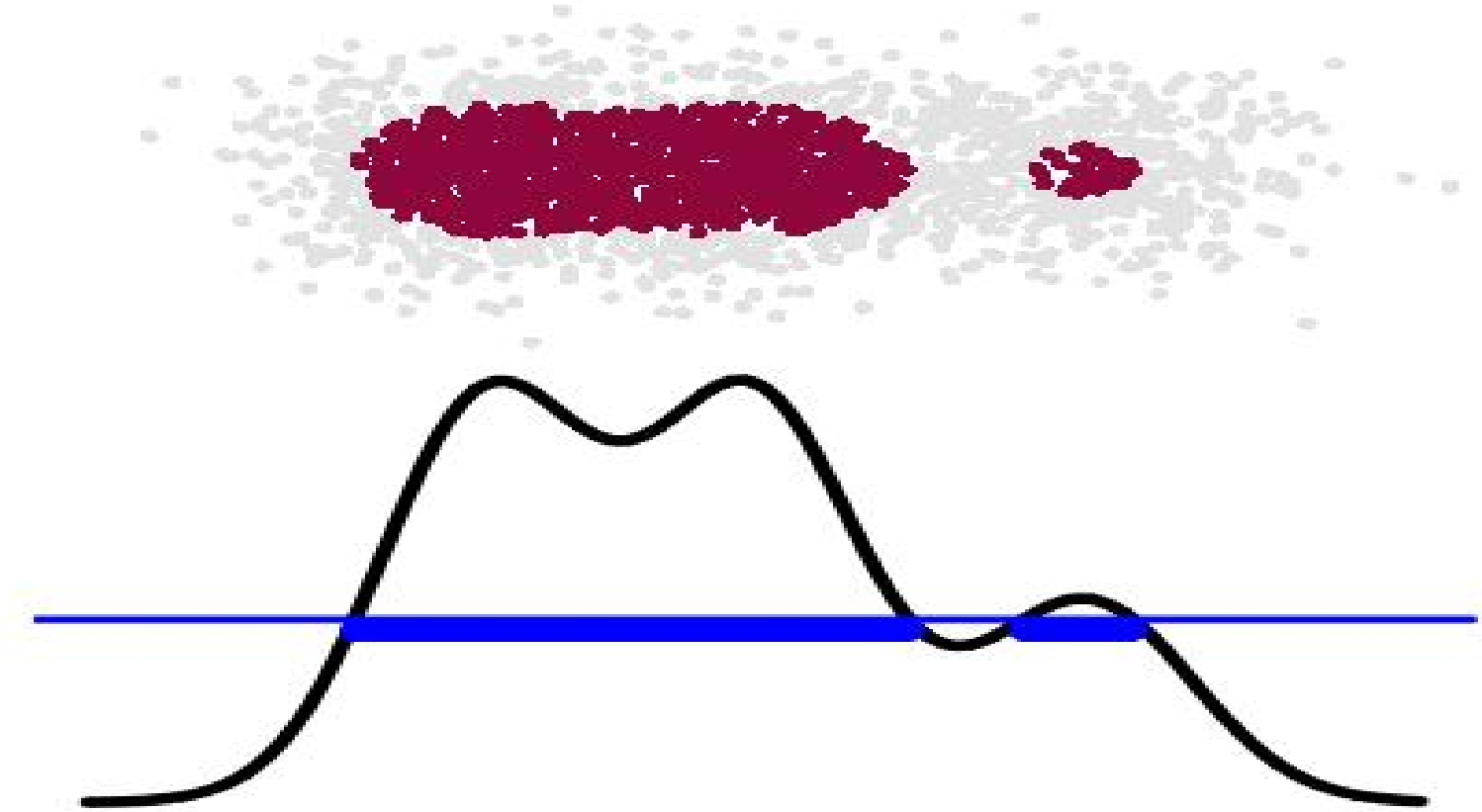
- Data $D = (x_1, \ldots, x_n) \in X^n$ are independent samples from some unknown distribution P on $X \subset \mathbb{R}^d$ that has some Lebesgue-density $h : X \to [0, \infty)$.
- For some user-specified $\rho \geq 0$, the clusters are the topologically connected components of the density level set

$$M_\rho := \{h \geq \rho\}.$$

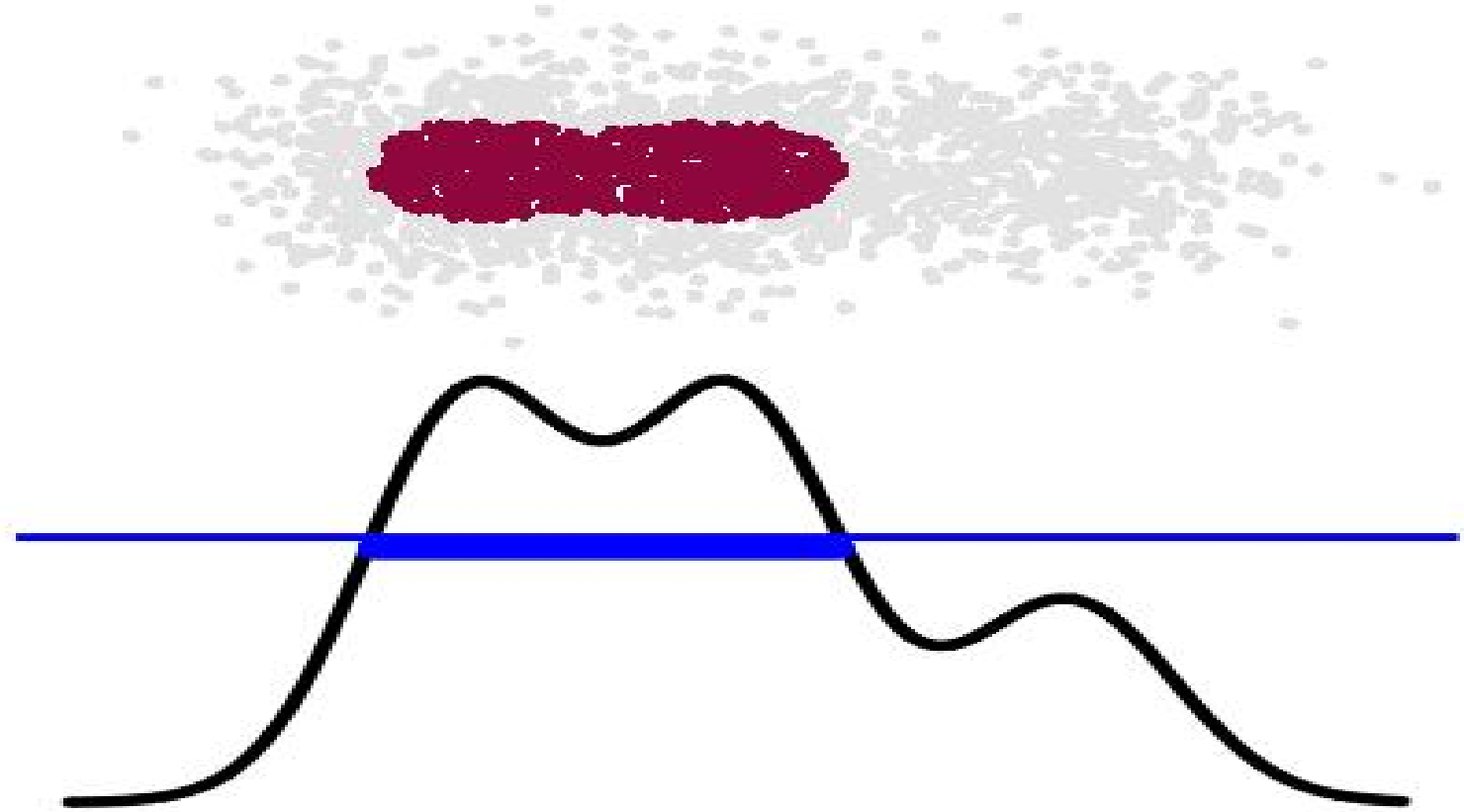
Example: Density Clustering



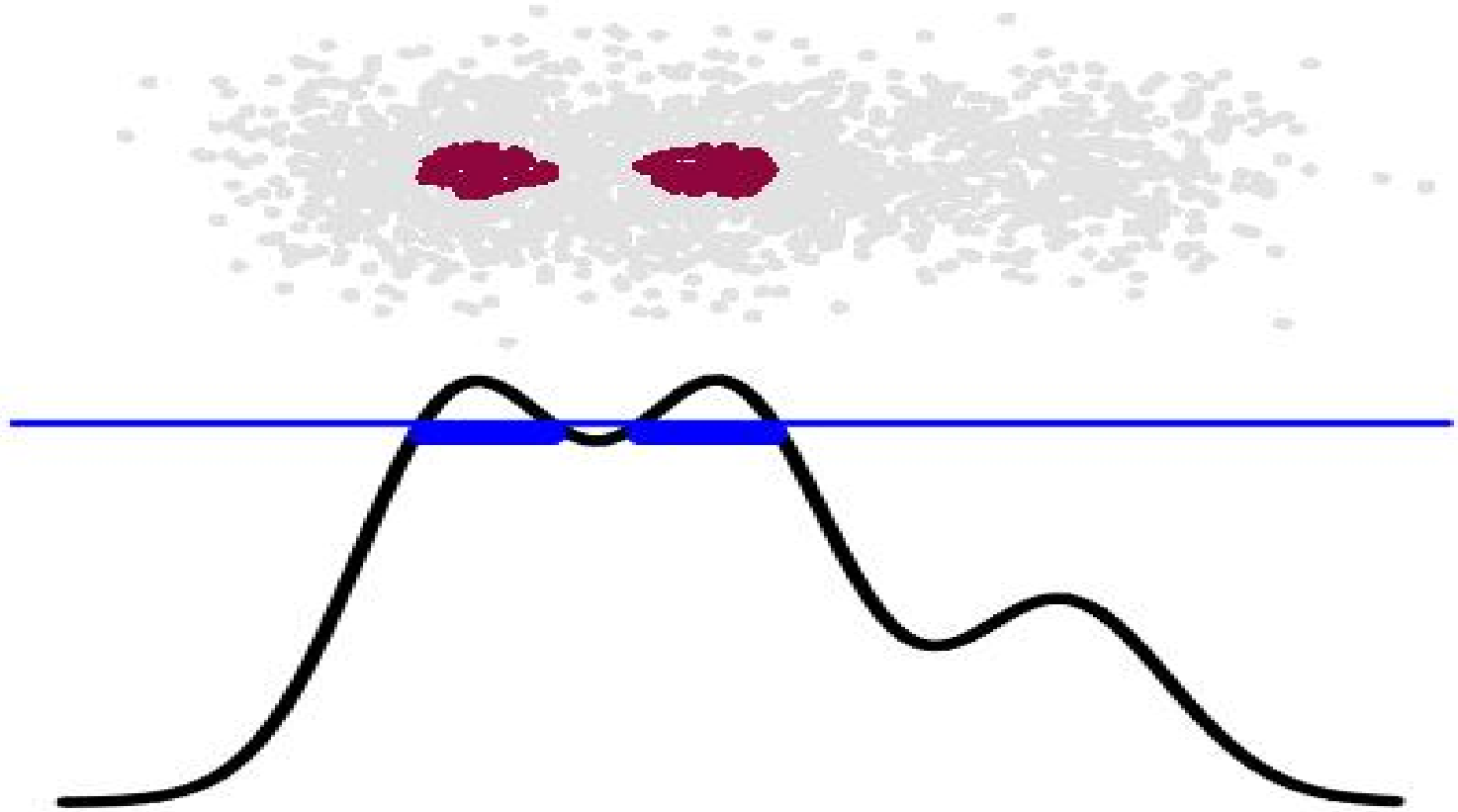
Example: Density Clustering



Example: Density Clustering



Example: Density Clustering



How to Choose the Level?

Previous Approaches for Choosing the Threshold Level

- Fix an $\alpha > 0$, e.g. $\alpha = 0.05$, and determine ρ by $P(\{h \geq \rho\}) \approx 1 - \alpha$.

How to Choose the Level?

Previous Approaches for Choosing the Threshold Level

- Fix an $\alpha > 0$, e.g. $\alpha = 0.05$, and determine ρ by $P(\{h \geq \rho\}) \approx 1 - \alpha$.
- Consider many different values for ρ and build the cluster tree.
  $\rightsquigarrow$ Stuetzle & Nugent (2010, 2010)
  Chaudhuri & Dasgupta (NIPS, 2010)
  Kpotufe & v. Luxburg (ICML, 2011).
  ...

Overview

- Consider the smallest level ρ for which level set has more than one connected component and define the corresponding clusters.
- Description of the algorithm for estimating level and clusters.
- Statistical analysis: consistency, rates, and adaptivity.
- Thoughts beyond level set clustering.

Defining Largest Clusters

Density Independent Level Sets

Definition

- $X := [0,1]^d$, μ be the Lebesgue measure on X.
- P be a μ-absolutely continuous distribution.

The density level set of P at level ρ is defined by

$$M_\rho := \operatorname{supp} \mu(\,\cdot \cap \{h \geq \rho\})\,.$$

Density Independent Level Sets

Definition

- $X := [0,1]^d$, μ be the Lebesgue measure on X.
- P be a μ-absolutely continuous distribution.

The density level set of P at level ρ is defined by

$$M_\rho := \operatorname{supp} \mu(\,\cdot \cap \{h \geq \rho\})\,.$$

Observations

- M_ρ is independent of choice of h.
- If $\mu(\partial\{h \geq \rho\}) = 0$, then $\mu(\{h \geq \rho\} \triangle M_\rho) = 0$.

Relating Connected Components

Connected Components

For closed $A \subset X$, we denote the set of all connected components (CCs) of A by $\mathcal{C}(A)$. We always assume $|\mathcal{C}(A)| < \infty$.

Lemma

Let $A \subset B$ be closed subsets of X. Then, for every $A' \in \mathcal{C}(A)$, there exists exactly one $B' \in \mathcal{C}(B)$ such that $A' \subset B'$.

Relating Connected Components

Connected Components

For closed $A \subset X$, we denote the set of all connected components (CCs) of A by $\mathcal{C}(A)$. We always assume $|\mathcal{C}(A)| < \infty$.

Lemma

Let $A \subset B$ be closed subsets of X. Then, for every $A' \in \mathcal{C}(A)$, there exists exactly one $B' \in \mathcal{C}(B)$ such that $A' \subset B'$.

Definition

For $A \subset B$, we call the unique map $\zeta := \zeta_{A,B} : \mathcal{C}(A) \to \mathcal{C}(B)$ satysfying

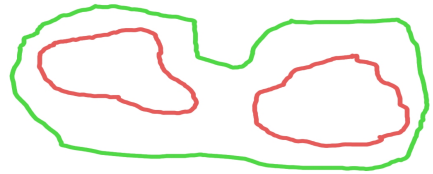
$$A' \subset \zeta(A'), \qquad A' \in \mathcal{C}(A)$$

the component relating map (CRM) between A and B.

Properties of CRMs

- $\zeta_{A,B}$ is injective if and only if the CCs of A (red) are not "glued" together in B.

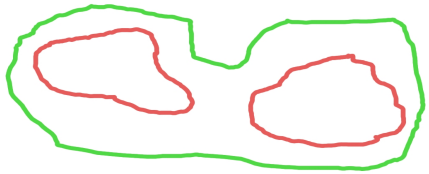




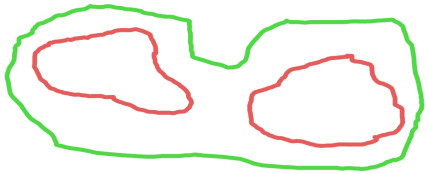
Properties of CRMs

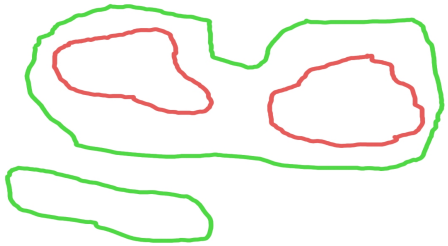
- $\zeta_{A,B}$ is injective if and only if the CCs of A (red) are not "glued" together in B.





- $\zeta_{A,B}$ is surjective if and only if the CCs of B do not "disappear" in A.

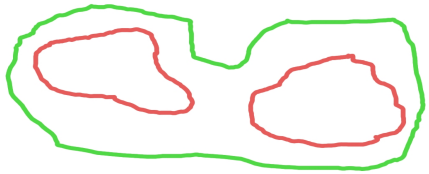




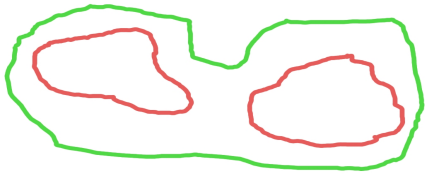
Properties of CRMs

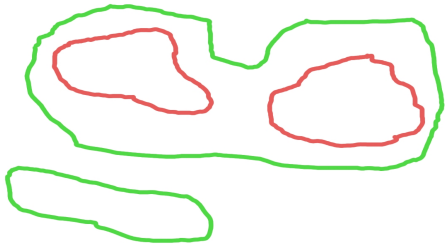
- $\zeta_{A,B}$ is injective if and only if the CCs of A (red) are not "glued" together in B.



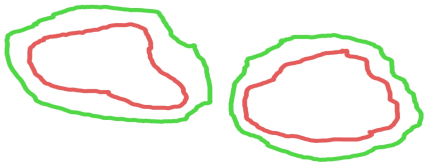


- $\zeta_{A,B}$ is surjective if and only if the CCs of B do not "disappear" in A.





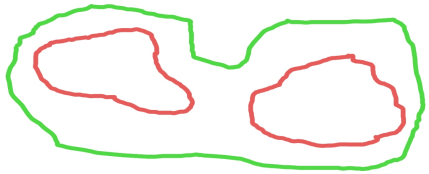
- **Consequence:**
  $\zeta_{A,B}$ is bijective if and only if A and B have the same CC-structure.



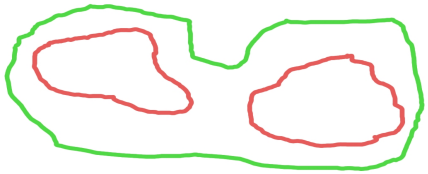
Properties of CRMs

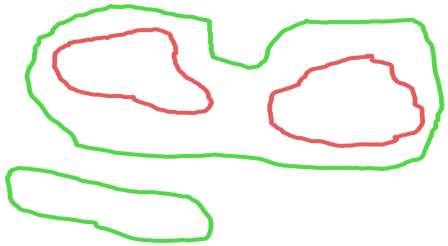
- $\zeta_{A,B}$ is injective if and only if the CCs of A (red) are not "glued" together in B.



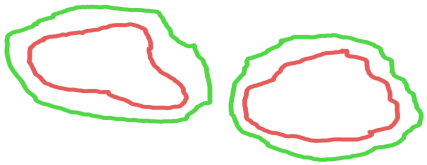


- $\zeta_{A,B}$ is surjective if and only if the CCs of B do not "disappear" in A.





- **Consequence:**
  $\zeta_{A,B}$ is bijective if and only if A and B have the same CC-structure.



- For $A \subset B \subset C$ we have $\zeta_{A,C} = \zeta_{B,C} \circ \zeta_{A,B}$.

Definition of Clusters

Definition

P has clusters between ρ^* and $\rho^{**} > \rho^*$ if, for all $\rho \in [0, \rho^{**}]$, we have:

- $1 \leq |\mathcal{C}(M_\rho)| \leq 2$
- If $|\mathcal{C}(M_\rho)| = 1$, then $\rho \leq \rho^*$.
- If $|\mathcal{C}(M_\rho)| = 2$, then $\rho \geq \rho^*$ and $\zeta : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$ is bijective.

Definition of Clusters

Definition

P has clusters between ρ^* and $\rho^{**} > \rho^*$ if, for all $\rho \in [0, \rho^{**}]$, we have:

- $1 \leq |\mathcal{C}(M_\rho)| \leq 2$
- If $|\mathcal{C}(M_\rho)| = 1$, then $\rho \leq \rho^*$.
- If $|\mathcal{C}(M_\rho)| = 2$, then $\rho \geq \rho^*$ and $\zeta : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$ is bijective.

Using the CRMs $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$, we then define the clusters of P by

$$A_i^* := \bigcup_{\rho \in (\rho^*, \rho^{**}]} \zeta_\rho(A_i)\,, \qquad i \in \{1, 2\}\,,$$

where A_1 and A_2 are the two connected components of $M_{\rho^{**}}$.

Illustration of Clusters I

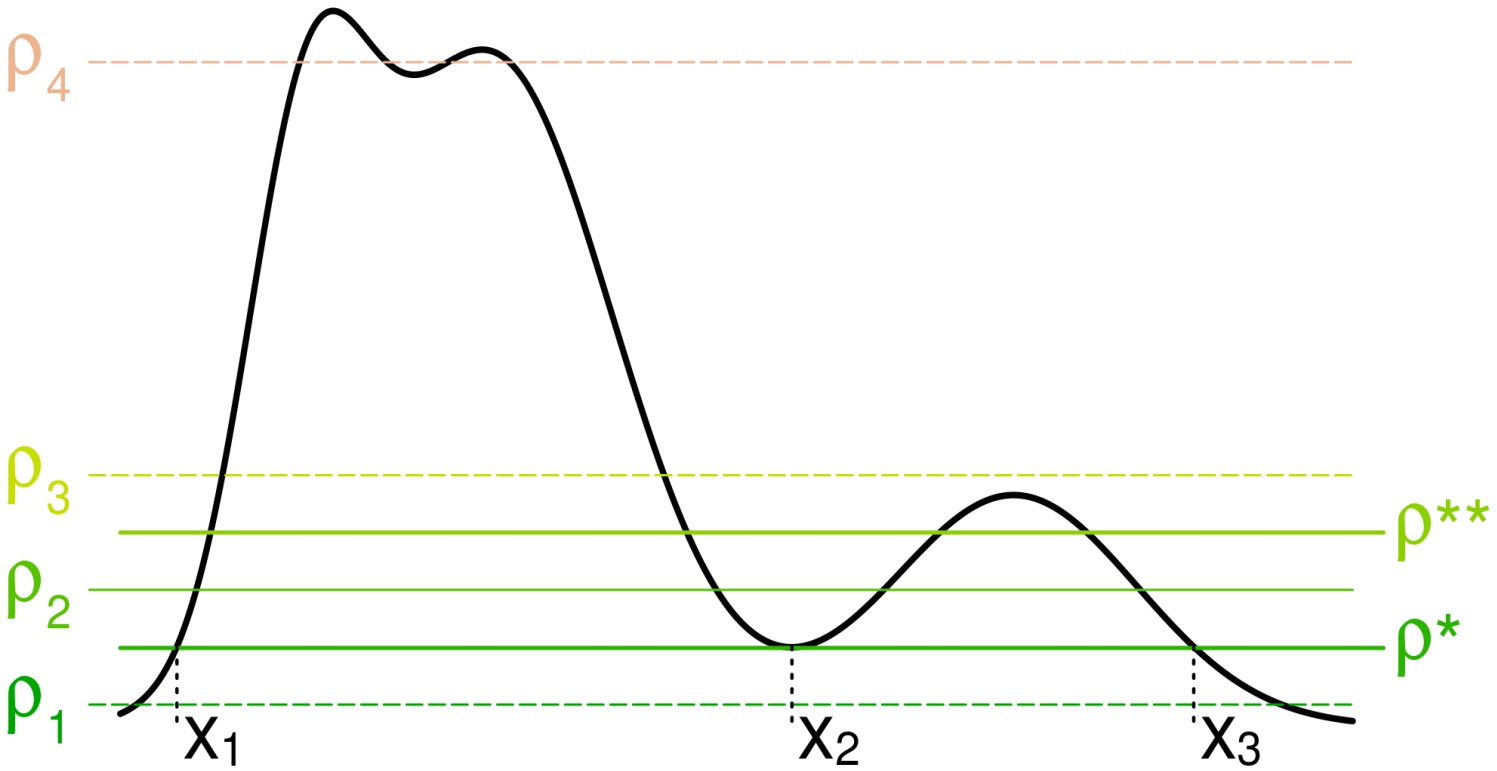


Illustration of Clusters II

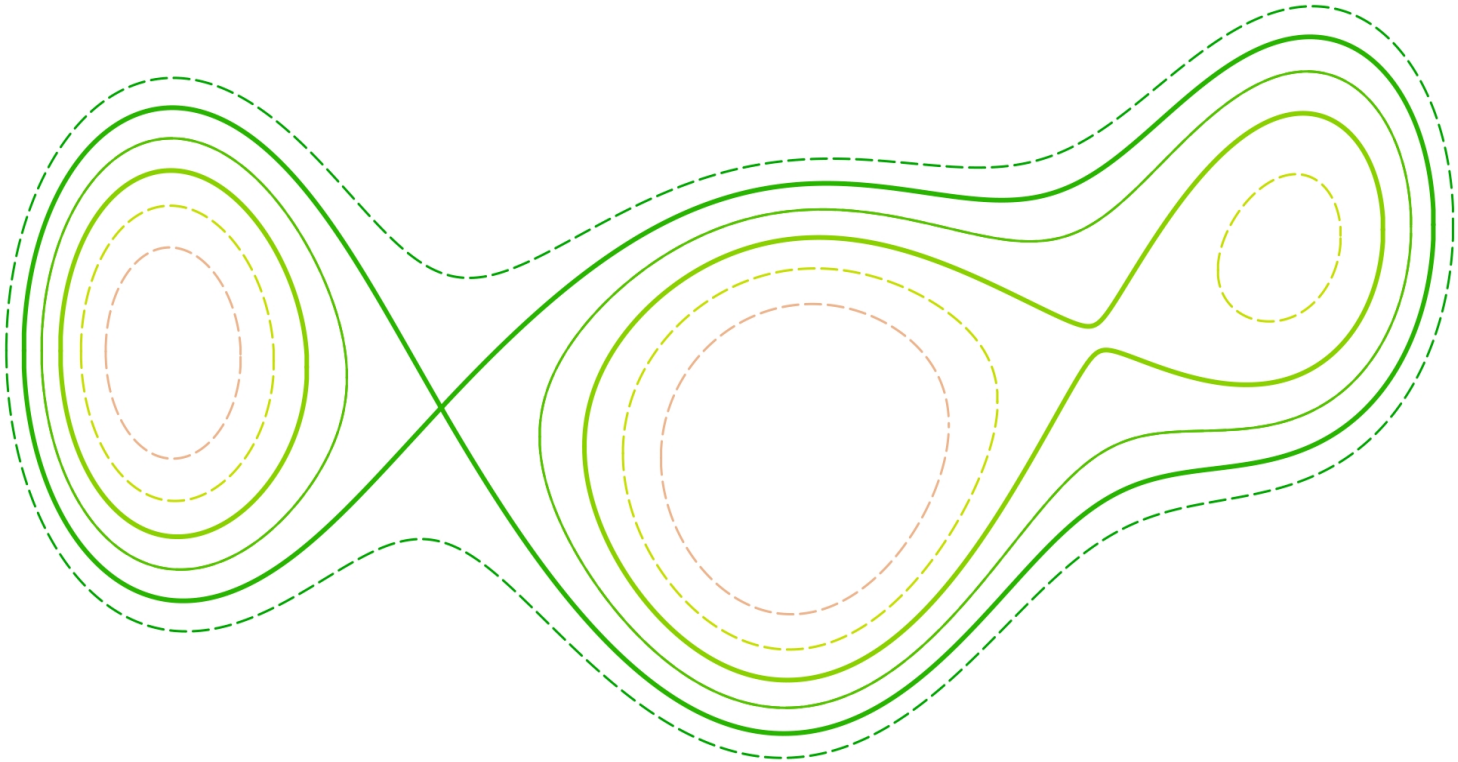
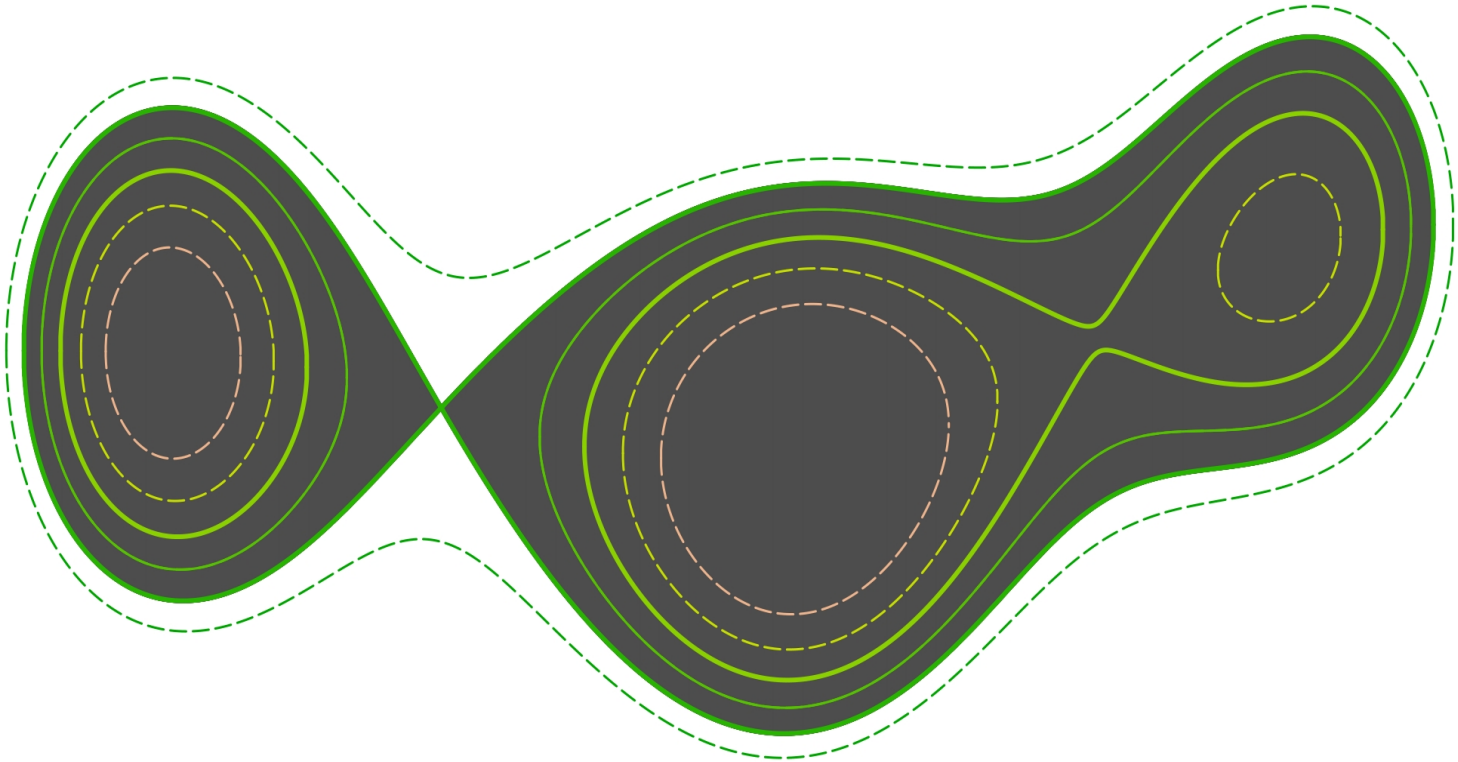


Illustration of Clusters II



Dealing with Uncertainty

Two Types of Uncertainty

Observation

Let $L_\rho := \{\hat{h} \geq \rho\}$ be a level set estimate based on data D. Then L_ρ has the following types of uncertainty compared to true level set M_ρ:

- vertical uncertainty caused by estimation error of $\hat{h}$.
- horizontal uncertainty caused by approximation error of $\hat{h}$.

Two Types of Uncertainty

Observation

Let $L_\rho := \{\hat{h} \geq \rho\}$ be a level set estimate based on data D. Then L_ρ has the following types of uncertainty compared to true level set M_ρ:

- vertical uncertainty caused by estimation error of $\hat{h}$.
- horizontal uncertainty caused by approximation error of $\hat{h}$.

Notation

Given some $A \subset X$, we write

$$A^{+\delta} := \{x \in X : d(x, A) \leq \delta\}$$
$$A^{-\delta} := X \setminus (X \setminus A)^{+\delta}.$$

Two Types of Uncertainty

Observation

Let $L_\rho := \{\hat{h} \geq \rho\}$ be a level set estimate based on data D. Then L_ρ has the following types of uncertainty compared to true level set M_ρ:

- vertical uncertainty caused by estimation error of $\hat{h}$.
- horizontal uncertainty caused by approximation error of $\hat{h}$.

Notation

Given some $A \subset X$, we write

$$A^{+\delta} := \{x \in X : d(x, A) \leq \delta\}$$
$$A^{-\delta} := X \setminus (X \setminus A)^{+\delta}.$$

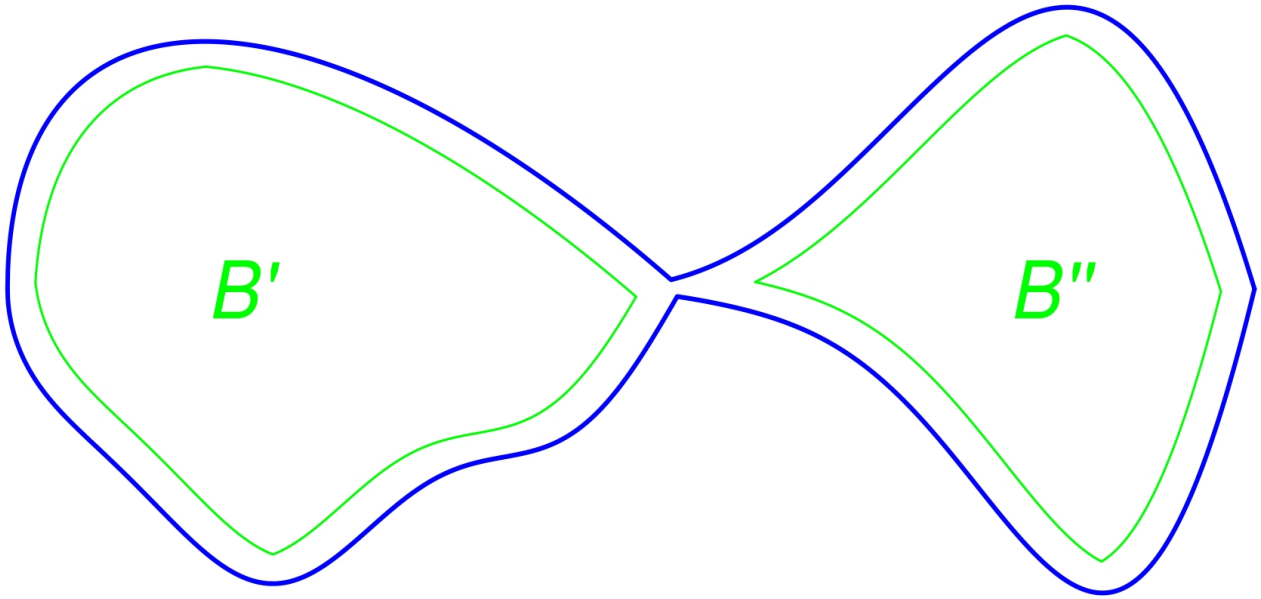
Uncertainty Control Assumption

We know $\delta, \varepsilon > 0$ such that with high probability we have

$$M_{\rho+\varepsilon}^{-\delta} \subset L_\rho \subset M_{\rho-\varepsilon}^{+\delta}.$$

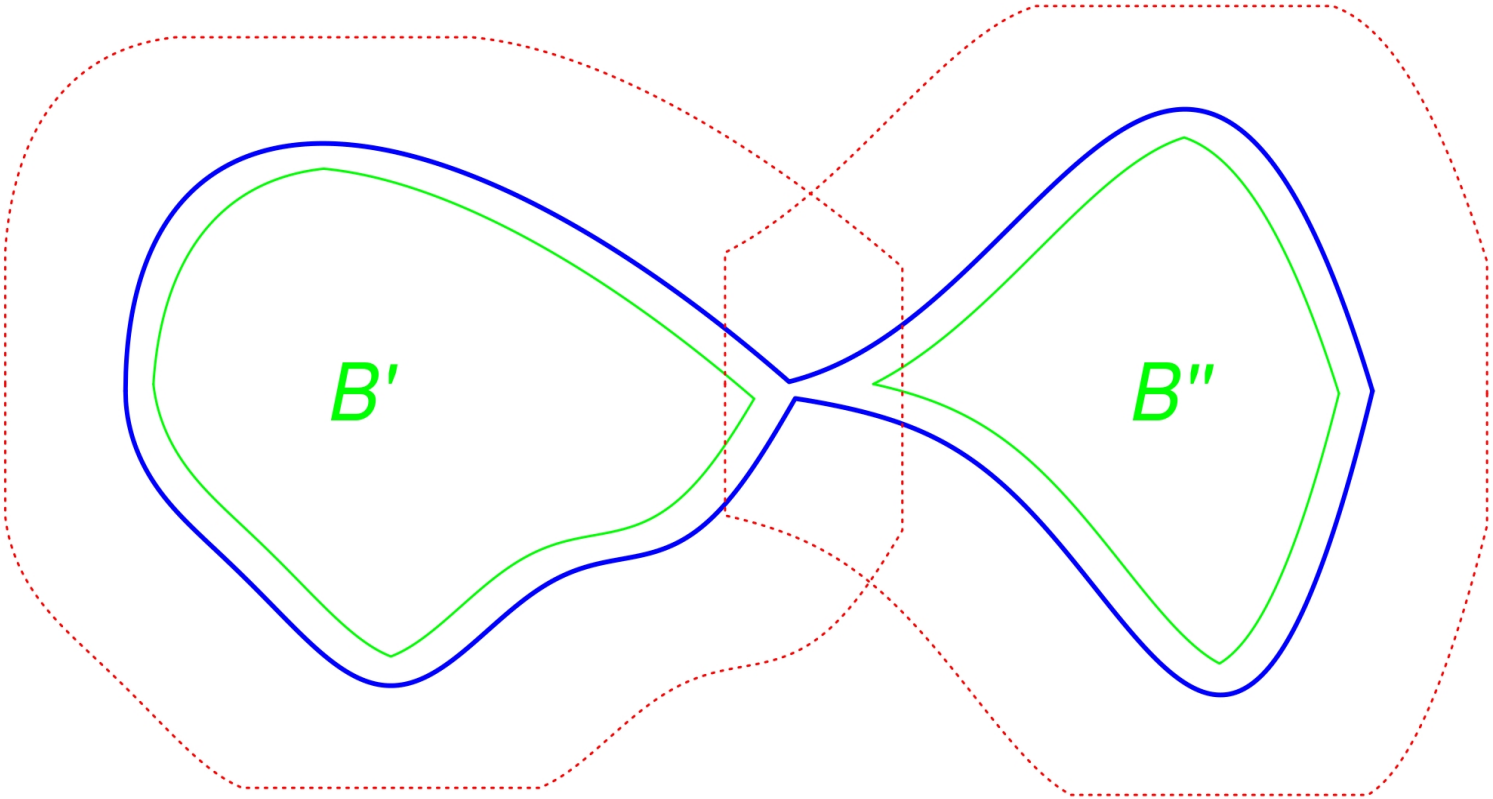
Vertical Uncertainty: The need to Glue

- Cutting δ tubes from B may change topological structure.



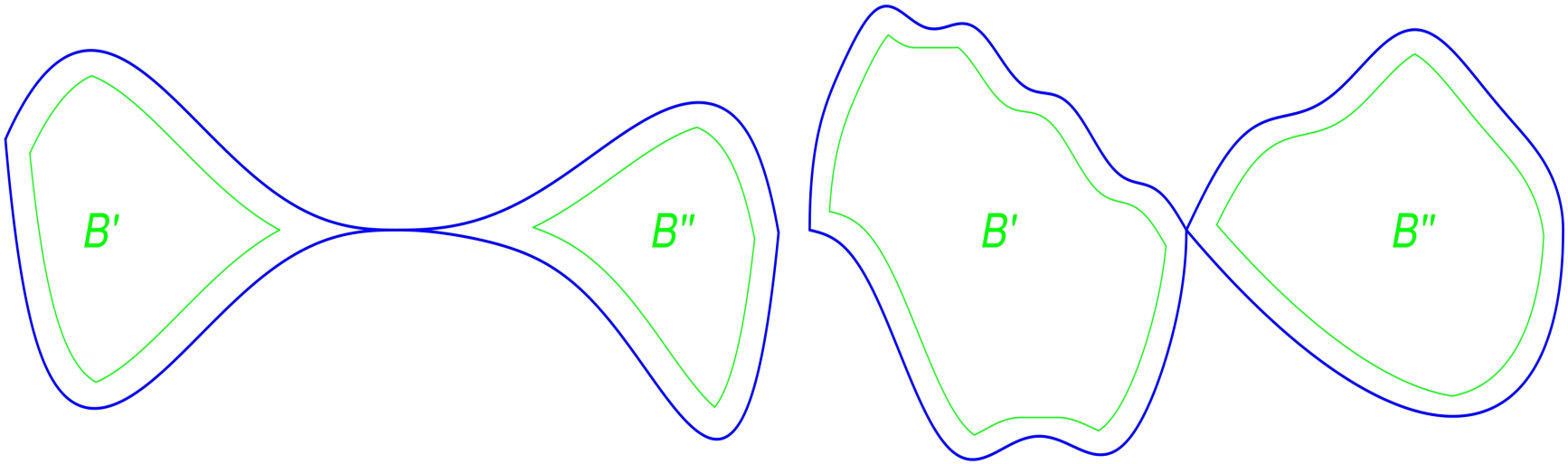
Vertical Uncertainty: The need to Glue

- Cutting δ tubes from B may change topological structure.
- This may be prevented by considering τ-connected components instead.



Vertical Uncertainty: Thickness Assumption

- Different shapes need different τ.

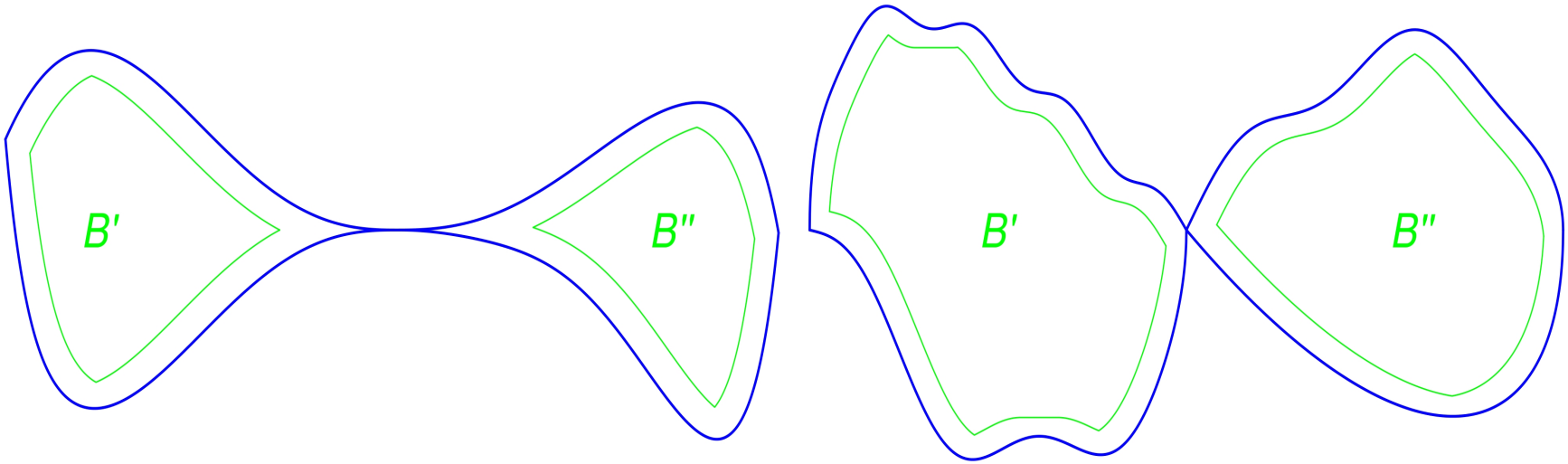


Vertical Uncertainty: Thickness Assumption

- Different shapes need different τ.
- The function

$$\psi^*(\delta) := \sup_{\rho<\rho^{**}} \sup_{x\in M_r} d(x, M_\rho^{-\delta})$$

controls thickness of bridges and cusps.

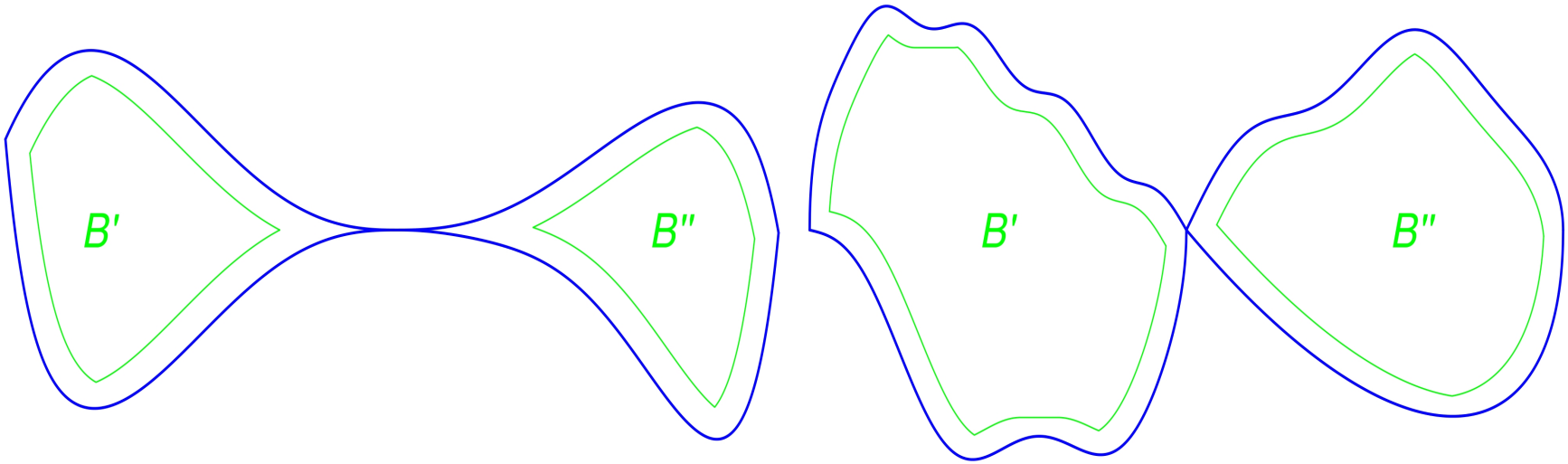


Vertical Uncertainty: Thickness Assumption

- Different shapes need different τ.
- The function

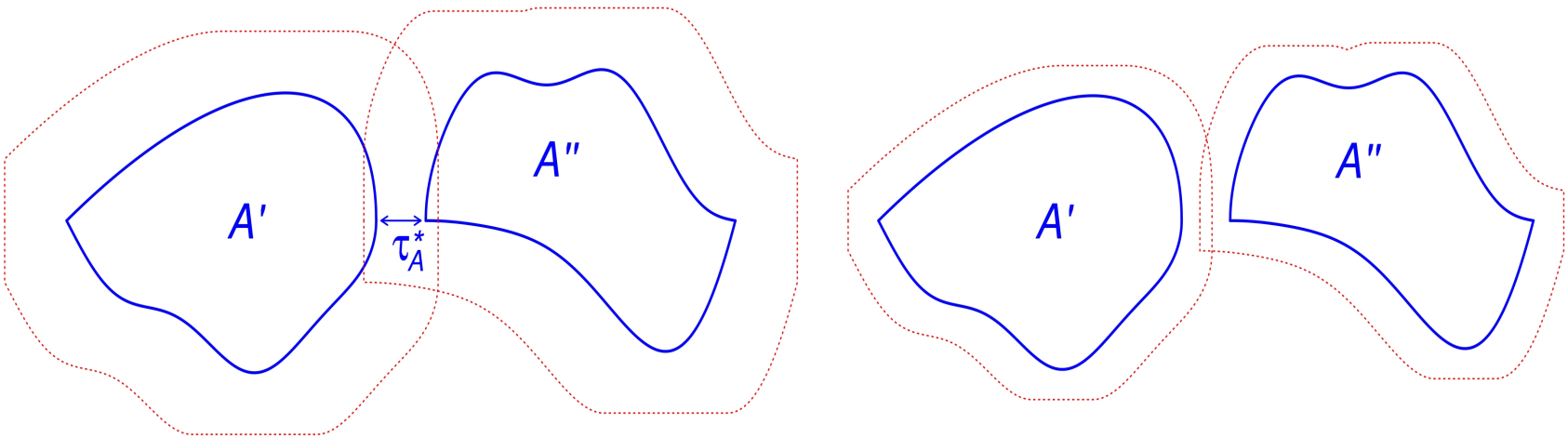
$$\psi^*(\delta) := \sup_{\rho<\rho^{**}} \sup_{x\in M_r} d(x, M_\rho^{-\delta})$$

  controls thickness of bridges and cusps.
- We assume $\psi^*(\delta) < c\,\delta^\gamma$ for some $\gamma \in (0, 1]$ and all $\delta \in (0, \delta_{\text{thick}}]$.



Vertical Uncertainty: The Danger of Glueing

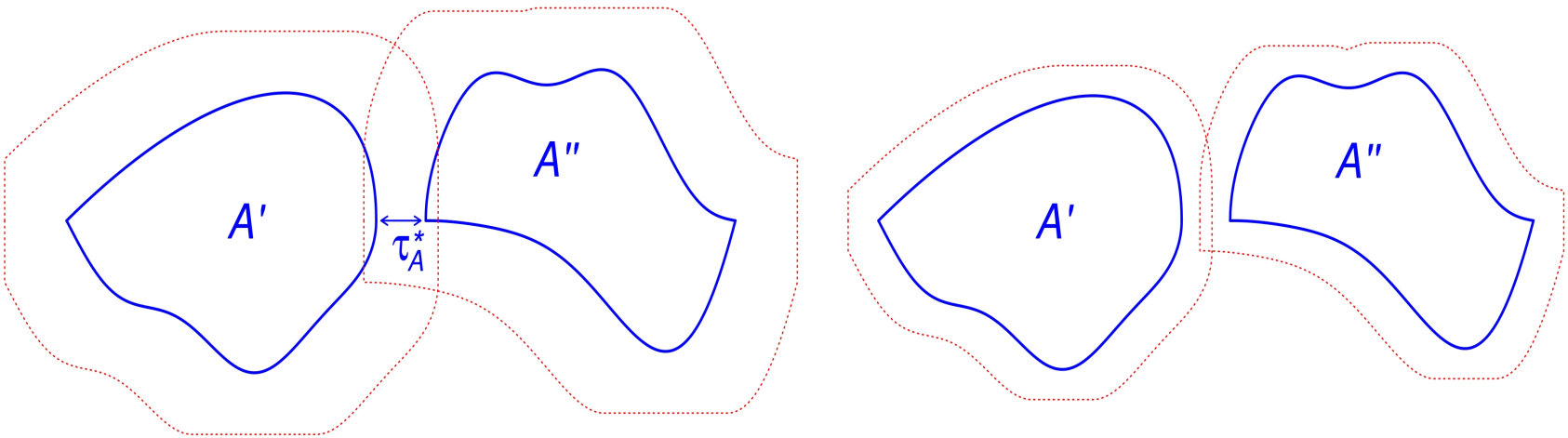
- Glueing by τ may change component structure.
- Glueing is fine for $\tau < \tau_A^* := d(A', A'')$.



Vertical Uncertainty: The Danger of Glueing

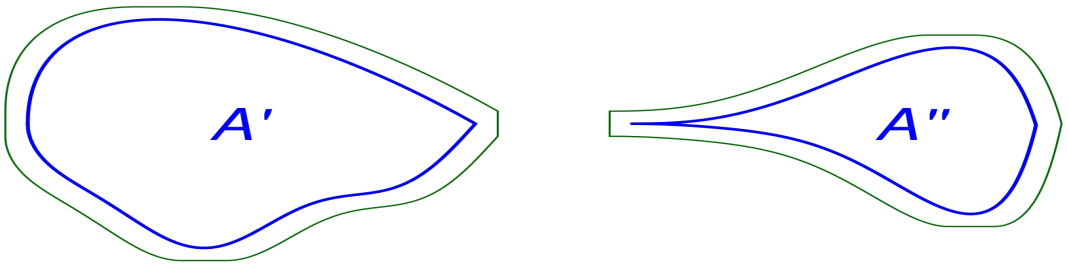
- Glueing by τ may change component structure.
- Glueing is fine for $\tau < \tau_A^* := d(A', A'')$.
- We consider the function

$$\tau^*(\varepsilon) := \frac{1}{3}\tau^*_{M_{\rho^*+\varepsilon}}$$

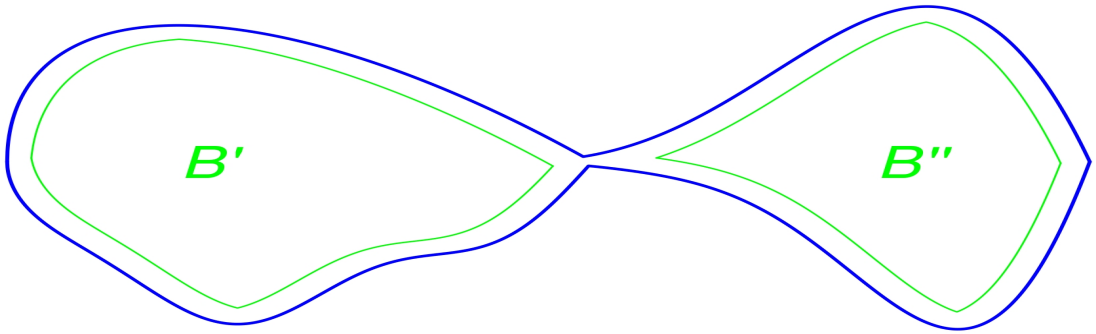


Vertical Uncertainty: Controlling the Distance

Fix $\rho - \varepsilon > \rho^*$ and $M_{\rho-\varepsilon} = A' \cup A''$. We guarantee $L_\rho \subset M_{\rho-\varepsilon}^{+\delta}$

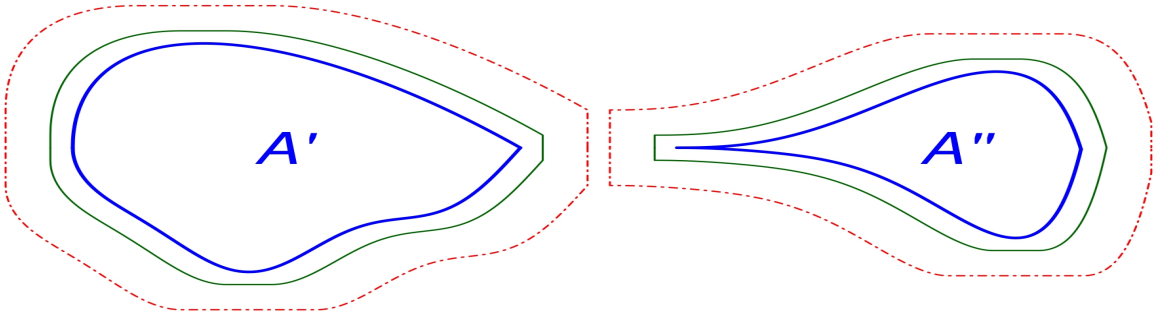


Fix $\rho + \varepsilon < \rho^*$ and $M_{\rho+\varepsilon} = A$. We guarantee $B' \cup B'' = M_{\rho+\varepsilon}^{-\delta} \subset L_\rho$

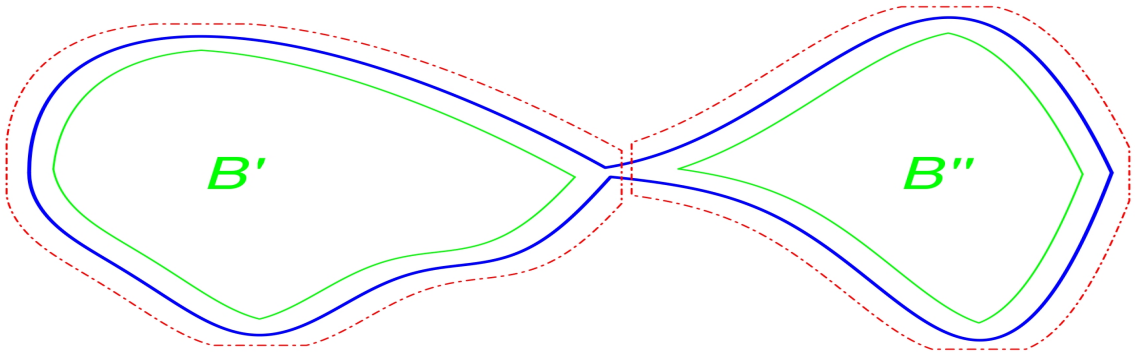


Vertical Uncertainty: Controlling the Distance

Fix $\rho - \varepsilon > \rho^*$ and $M_{\rho-\varepsilon} = A' \cup A''$. We guarantee $L_\rho \subset M_{\rho-\varepsilon}^{+\delta}$

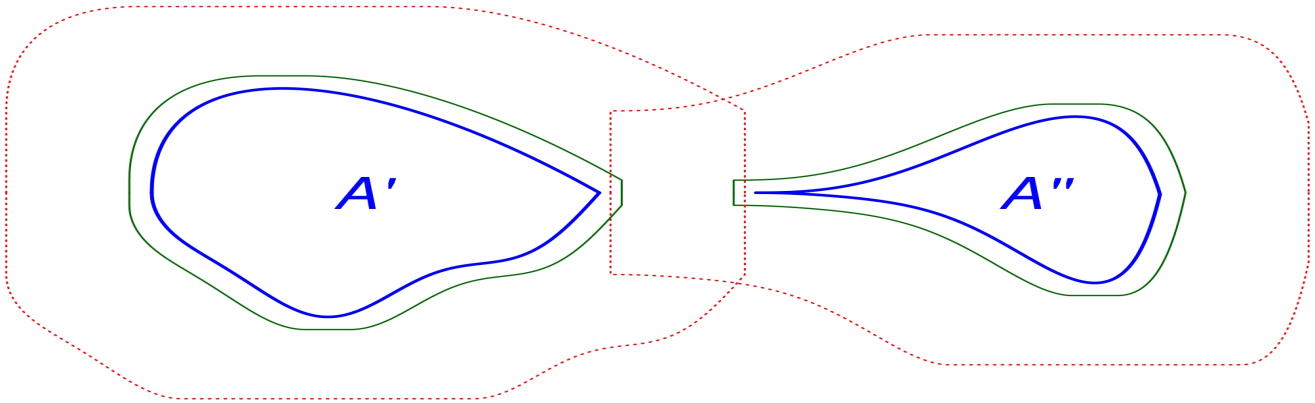


Fix $\rho + \varepsilon < \rho^*$ and $M_{\rho+\varepsilon} = A$. We guarantee $B' \cup B'' = M_{\rho+\varepsilon}^{-\delta} \subset L_\rho$

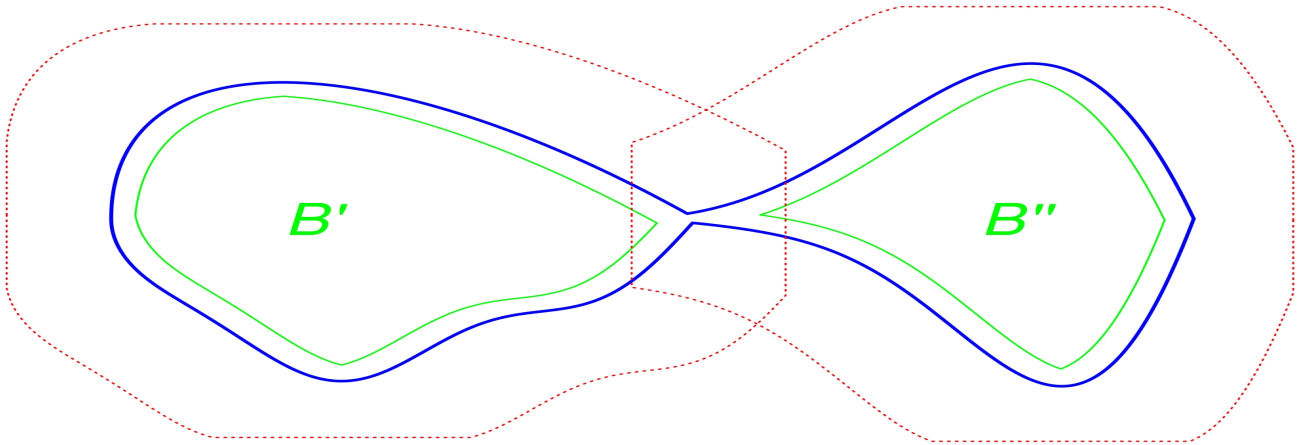


Vertical Uncertainty: Controlling the Distance

Fix $\rho - \varepsilon > \rho^*$ and $M_{\rho-\varepsilon} = A' \cup A''$. We guarantee $L_\rho \subset M_{\rho-\varepsilon}^{+\delta}$

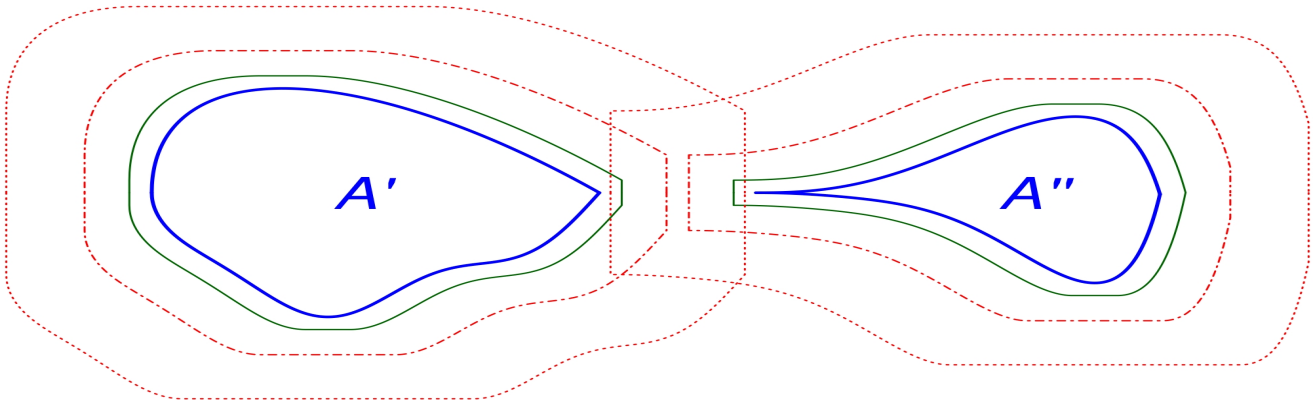


Fix $\rho + \varepsilon < \rho^*$ and $M_{\rho+\varepsilon} = A$. We guarantee $B' \cup B'' = M_{\rho+\varepsilon}^{-\delta} \subset L_\rho$

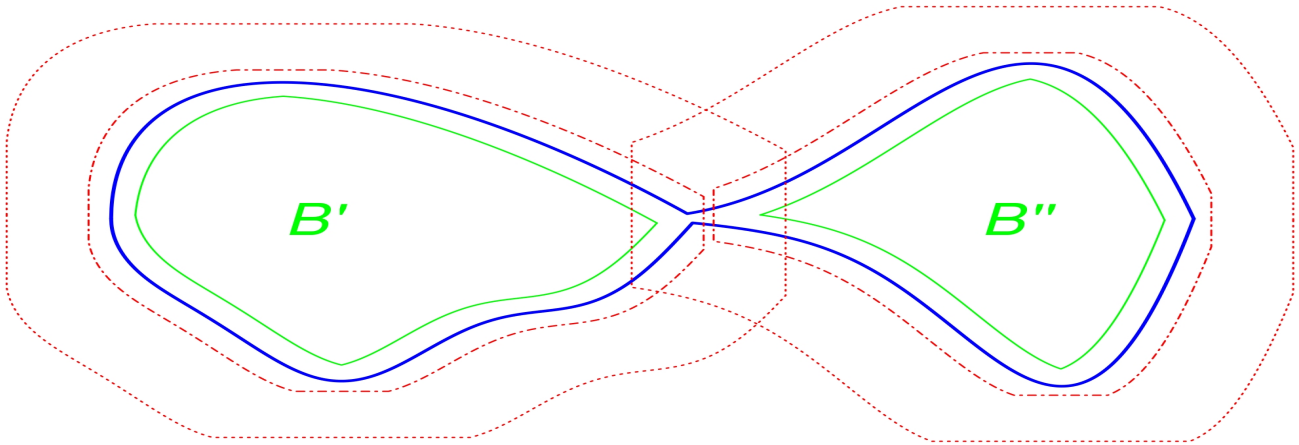


Vertical Uncertainty: Controlling the Distance

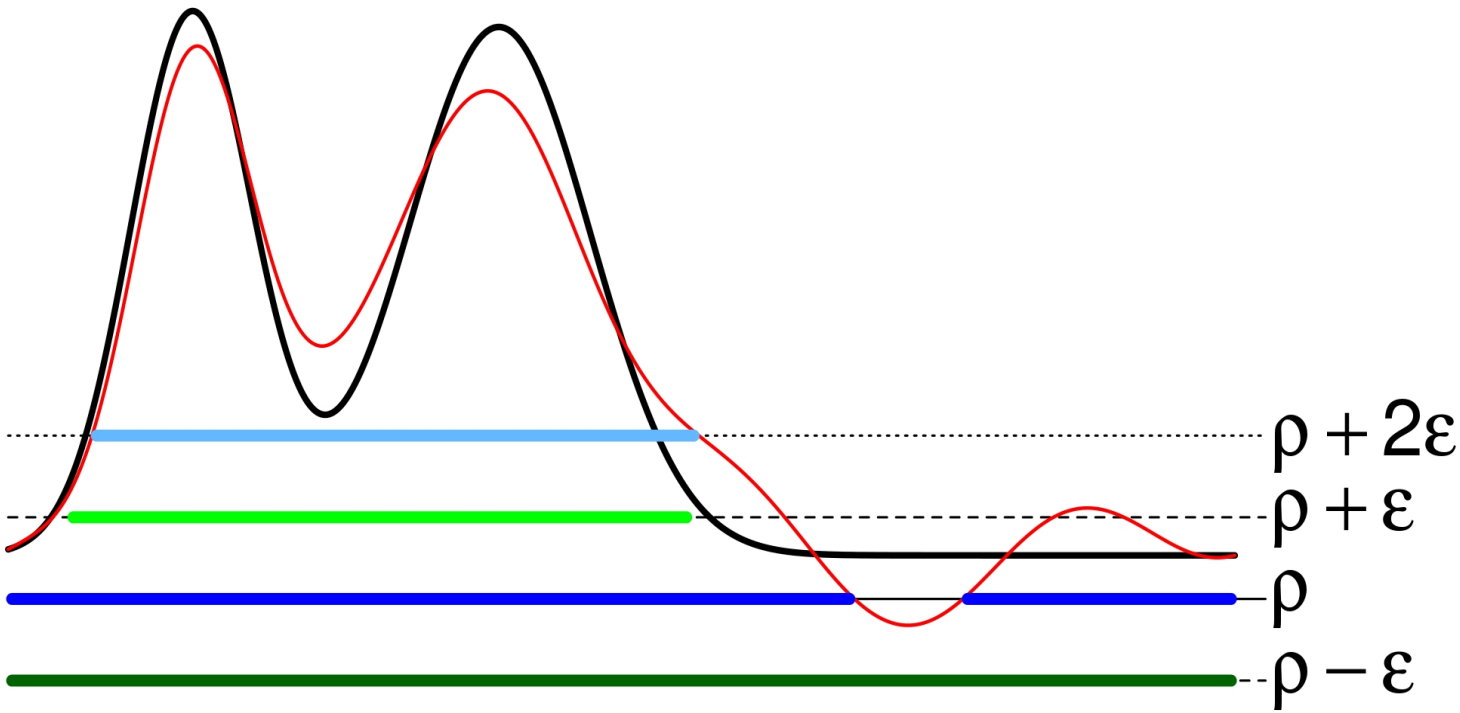
Fix $\rho - \varepsilon > \rho^*$ and $M_{\rho-\varepsilon} = A' \cup A''$. We guarantee $L_\rho \subset M_{\rho-\varepsilon}^{+\delta}$



Fix $\rho + \varepsilon < \rho^*$ and $M_{\rho+\varepsilon} = A$. We guarantee $B' \cup B'' = M_{\rho+\varepsilon}^{-\delta} \subset L_\rho$

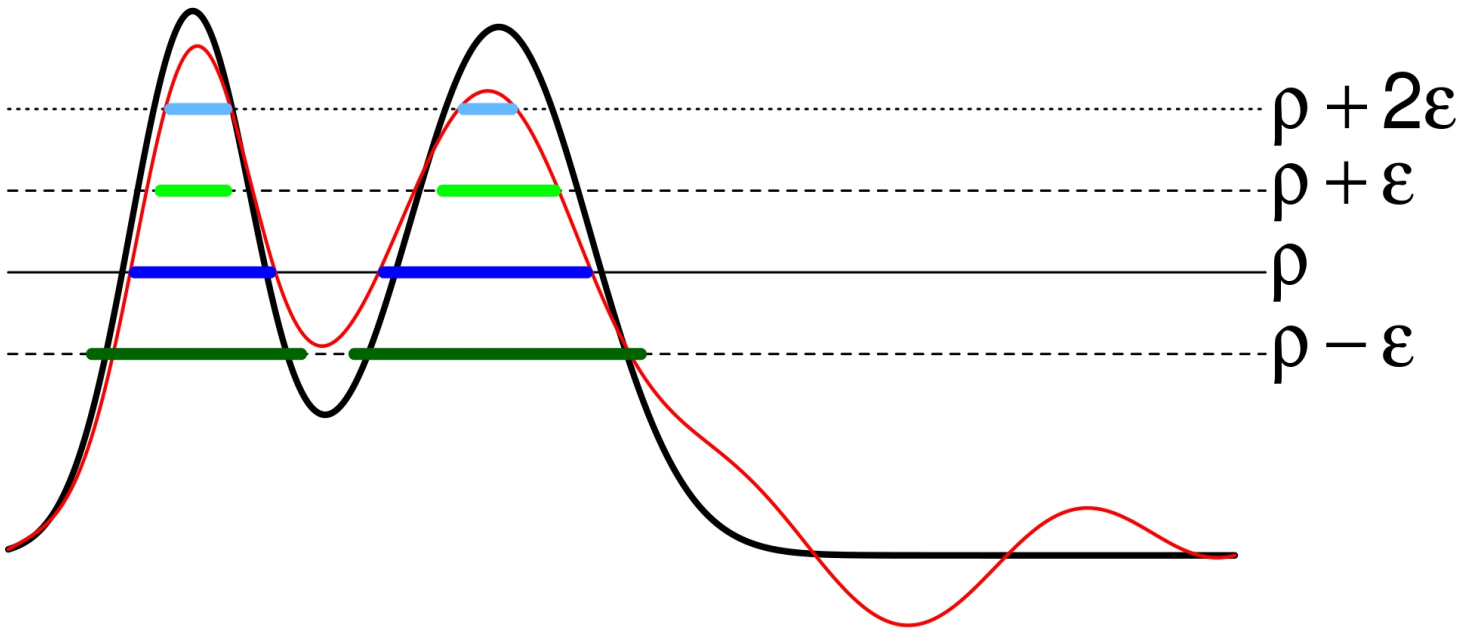


Horizontal Uncertainty: Ignoring Statistical Fluctuations



Small fluctuations of the density estimate around ρ can be detected at level $\rho + 2\varepsilon$.

Horizontal Uncertainty: Ignoring Statistical Fluctuations



Small fluctuations of the density estimate around ρ can be detected at level $\rho + 2\varepsilon$.

A Generic Algorithm

Clustering Algorithm

- Fix $\varepsilon > 0$, $\tau > 0$, and $\rho := 0$.
- **repeat**

  Identify the τ-CCs $B_1, \ldots, B_M$ of $L_{D,\rho}$ satisfying

$$B_i \cap L_{D,\rho+2\varepsilon} \neq \emptyset. \tag{1}$$

  $\rho \leftarrow \rho + \varepsilon$

  until $M \neq 1$
- **return** $\rho_D^* := \rho + \varepsilon$ and the τ-CCs $B_1, \ldots, B_M$ of $L_{D,\rho+\varepsilon}$ satisfying

$$B_i \cap L_{D,\rho+3\varepsilon} \neq \emptyset.$$

The Algorithm Works

Theorem

- Assume that $M_{\rho+\varepsilon}^{-\delta} \subset L_{D,\rho} \subset M_{\rho-\varepsilon}^{+\delta}$ for all $\rho \in [0, \rho^{**}]$.
- $\varepsilon^* \leq (\rho^{**} - \rho^*)/9$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi^*(\delta), \tau^*(\varepsilon^*)]$, and $\varepsilon \in (0, \varepsilon^*]$.

Then, the following statements are true:

The Algorithm Works

Theorem

- Assume that $M_{\rho+\varepsilon}^{-\delta} \subset L_{D,\rho} \subset M_{\rho-\varepsilon}^{+\delta}$ for all $\rho \in [0, \rho^{**}]$.
- $\varepsilon^* \leq (\rho^{**} - \rho^*)/9$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi^*(\delta), \tau^*(\varepsilon^*)]$, and $\varepsilon \in (0, \varepsilon^*]$.

Then, the following statements are true:

1. The returned level ρ_D^* satisfies both $\rho_D^* \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ and

$$\tau - \psi^*(\delta) < 3\tau^*\left(\rho_D^* - \rho^* + \varepsilon\right). \tag{2}$$

The Algorithm Works

Theorem

- Assume that $M^{-\delta}_{\rho+\varepsilon} \subset L_{D,\rho} \subset M^{+\delta}_{\rho-\varepsilon}$ for all $\rho \in [0, \rho^{**}]$.
- $\varepsilon^* \leq (\rho^{**} - \rho^*)/9$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi^*(\delta), \tau^*(\varepsilon^*)]$, and $\varepsilon \in (0, \varepsilon^*]$.

Then, the following statements are true:

1. The returned level ρ^*_D satisfies both $\rho^*_D \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ and

$$\tau - \psi^*(\delta) < 3\tau^*\left(\rho^*_D - \rho^* + \varepsilon\right). \tag{2}$$

2. Two sets $B_1(D)$ and $B_2(D)$ are returned and these sets can be ordered such that we have

$$\mu\big(B_i(D) \triangle A^*_i\big) \leq \sum_{i=1}^{2} \mu\big(A^*_i \setminus (A^i_{\rho^*_D+\varepsilon})^{-\delta}\big) + \mu\big(M^{+\delta}_{\rho^*_D-\varepsilon} \setminus \{h > \rho^*\}\big)$$

Here, $A^i_{\rho^*_D+\varepsilon} \in \mathcal{C}(M_{\rho^*_D+\varepsilon})$ are ordered in the sense of $A^i_{\rho^*_D+\varepsilon} \subset A^*_i$.

Statistical Results

Consistency

Theorem

- $X := [0,1]^d$.
- Let $\mathcal{A}_\delta$ be a partition of $[0,1]^d$ into hypercubes of length δ.
- Consider histogram-based level set estimator $L_{D,\rho}$.

Let (ε_n), (δ_n), and (τ_n) be strictly positive sequences converging to zero such that $2\delta_n^\gamma < \tau_n$ and

$$\frac{\delta_n^{2d}\varepsilon_n^2 n}{\ln \delta_n^{-1}} \to \infty\,. \tag{3}$$

Then we have with convergence in probability P^∞ for $n \to \infty$

$$\rho_D^* \to \rho^*$$
$$\mu(B_i(D) \triangle A_i^*) \to 0\,.$$

The Width of the Canyon

Definition

We say that the clusters of P have separation exponent $\kappa \in (0, \infty]$, if

$$\tau^*_{M_{\rho^*+\varepsilon}} \geq c_1 \varepsilon^{1/\kappa}, \qquad \varepsilon \in (0, \rho^{**} - \rho^*)$$

Moreover, we say that the separation exponent κ is exact, if

$$\tau^*_{M_{\rho^*+\varepsilon}} \leq c_2 \varepsilon^{1/\kappa}, \qquad \varepsilon \in (0, \rho^{**} - \rho^*)$$

The Width of the Canyon

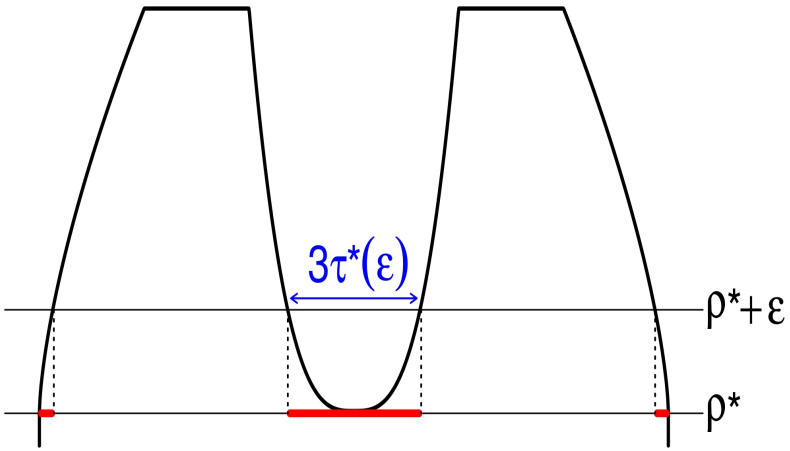
Definition

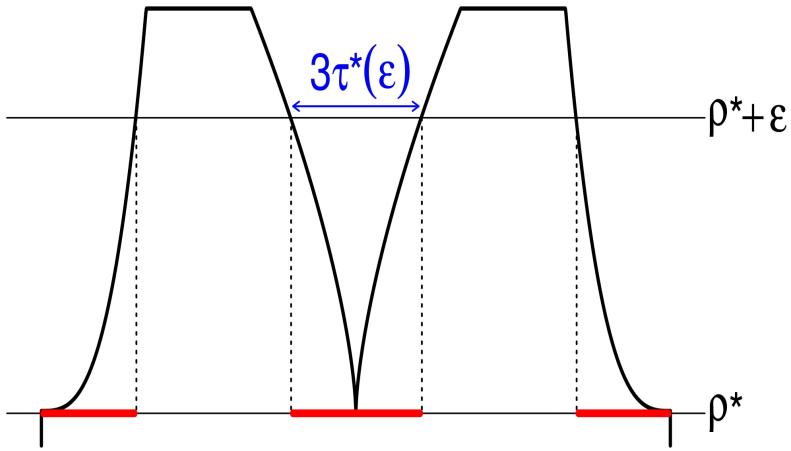
We say that the clusters of P have separation exponent $\kappa \in (0, \infty]$, if

$$\tau^*_{M_{\rho^*+\varepsilon}} \geq c_1 \varepsilon^{1/\kappa}, \qquad \varepsilon \in (0, \rho^{**} - \rho^*)$$

Moreover, we say that the separation exponent κ is exact, if

$$\tau^*_{M_{\rho^*+\varepsilon}} \leq c_2 \varepsilon^{1/\kappa}, \qquad \varepsilon \in (0, \rho^{**} - \rho^*)$$





Lemma

If h is α-Hölder-continuous, then $\kappa \geq \alpha$.

Rates I

Theorem

Assume a separation exponent κ. Then we have

$$\rho^*_{D,\delta} - \rho^* \preceq \left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{\gamma\kappa}{2\gamma\kappa+d}}$$

if δ_n, ε_n, and τ_n are optimally chosen. If κ is exact, the rate is exact, too.

Rates I

Theorem

Assume a separation exponent κ. Then we have

$$\rho^*_{D,\delta} - \rho^* \preceq \left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{\gamma\kappa}{2\gamma\kappa+d}}$$

if δ_n, ε_n, and τ_n are optimally chosen. If κ is exact, the rate is exact, too.

Examples

- For $d = 1$ we always have $\gamma = 1$ and the exponent becomes $\frac{\kappa}{2\kappa+1}$.
- For $d = 2$ we typically have $\gamma = 1$ and the exponent becomes $\frac{\kappa}{2\kappa+2}$.

Assumptions for Estimating the Clusters

Flatness (One-sided Polonik)

We say that P is flatness exponent $\vartheta \in (0, \infty]$, if there exists a constant $c_p > 0$ such that

$$\mu\big(\{0 < h - \rho^* < s\}\big) \leq (c_p s)^{\vartheta}, \qquad s > 0.$$

Assumptions for Estimating the Clusters

Flatness (One-sided Polonik)

We say that P is flatness exponent $\vartheta \in (0, \infty]$, if there exists a constant $c_p > 0$ such that

$$\mu\big(\{0 < h - \rho^* < s\}\big) \leq (c_p s)^{\vartheta}, \qquad s > 0.$$

Regularity of the Boundary

The clusters have an α-smooth boundary for $\alpha \in (0, 1]$, if there exists a constant $c_b > 0$ such that, for all $\rho \in (\rho^*, \rho^{**}]$ and $\delta \in (0, 1]$, we have

$$\mu\big((A_\rho^i)^{+\delta} \cap (X \setminus A_\rho^i)^{+\delta}\big) \leq c_b \delta^{\alpha}, \tag{4}$$

where A_ρ^1 and A_ρ^2 denote the two CCs of the level set M_ρ.

Rates II

Theorem

Assume a separation exponent κ, flatness exponent ϑ and α-smoothness. For optimally chosen δ_n, ε_n, and τ_n we then have

$$\mu(B_i(D) \triangle A_i^*) \preceq \begin{cases} \left(\frac{\ln n \cdot \ln \ln n}{n}\right)^{\frac{\vartheta\alpha}{2\alpha+\vartheta d}} & \text{if } \alpha \leq \vartheta\gamma\kappa \\ \left(\frac{\ln n \cdot \ln \ln n}{n}\right)^{\frac{\vartheta\gamma\kappa}{2\gamma\kappa+d}} & \text{else.} \end{cases} \tag{5}$$

Rates II

Theorem

Assume a separation exponent κ, flatness exponent ϑ and α-smoothness. For optimally chosen δ_n, ε_n, and τ_n we then have

$$\mu(B_i(D) \triangle A_i^*) \preceq \begin{cases} \left(\frac{\ln n \cdot \ln \ln n}{n}\right)^{\frac{\vartheta\alpha}{2\alpha+\vartheta d}} & \text{if } \alpha \leq \vartheta\gamma\kappa \\ \left(\frac{\ln n \cdot \ln \ln n}{n}\right)^{\frac{\vartheta\gamma\kappa}{2\gamma\kappa+d}} & \text{else.} \end{cases} \tag{5}$$

For these δ_n, ε_n, and τ_n we further have

$$\rho_{D,\delta}^* - \rho^* \preceq \left(\frac{\ln n \cdot \ln \ln n}{n}\right)^{\frac{\gamma\kappa}{2\gamma\kappa+d} \wedge \frac{\alpha}{2\alpha+d}} .$$

Picking a Candidate

Let $\Delta \subset (0,1]$ be a finite set and $n \geq 1$, $\varsigma \geq 1$. For $\delta \in \Delta$, we write $\tau_\delta := 4\delta^\gamma$ and

$$\varepsilon_\delta := \sqrt{\frac{(\varsigma + 2d \ln |\Delta| - d \ln \delta) \ln \ln n}{\delta^d n}} + \frac{2^d (\varsigma + 2d \ln |\Delta| - d \ln \delta)}{\delta^d n}, \quad (6)$$

Let us consider a width $\delta^*_{D,\Delta} \in \Delta$ that achieves the smallest level, that is

$$\delta^*_{D,\Delta} \in \arg\min_{\delta \in \Delta} \rho^*_{D,\delta}.$$

Moreover, we write

$$\rho^*_{D,\Delta} := \rho^*_{D,\delta^*_{D,\Delta}} = \min_{\delta \in \Delta} \rho^*_{D,\delta} \quad (7)$$

for the smallest returned level. Analogously, we define $\varepsilon_{D,\Delta} := \varepsilon_{\delta^*_{D,\Delta}}$.

Adaptivity

Theorem

For the interval

$$I_n := \left[\left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{1}{d}}, \left(\frac{1}{\ln \ln n} \right)^{\frac{1}{d}} \right]$$

we fix some $n^{-1/d}$-net $\Delta_n \subset I_n$ of I_n with $|\Delta_n| \leq n$. Then we have

$$\rho^*_{D,\Delta} - \rho^* \preceq \left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{\gamma\kappa}{2\gamma\kappa+d}} .$$

Adaptivity

Theorem

For the interval

$$I_n := \left[\left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{1}{d}}, \left(\frac{1}{\ln \ln n} \right)^{\frac{1}{d}} \right]$$

we fix some $n^{-1/d}$-net $\Delta_n \subset I_n$ of I_n with $|\Delta_n| \leq n$. Then we have

$$\rho^*_{D,\Delta} - \rho^* \preceq \left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{\gamma\kappa}{2\gamma\kappa+d}} .$$

If $\kappa < \infty$ is exact, the rate is exact, and for $\alpha \geq \vartheta\gamma\kappa$ we have

$$\mu(B_i(D) \triangle A^*_i) \preceq \left(\frac{\ln n \cdot \ln \ln n}{n} \right)^{\frac{\vartheta\gamma\kappa}{2\gamma\kappa+d}} .$$

Towards Axiomatic Clustering

Hierarchical Clusterings

General Assumptions

- $\mathcal{P}$ is a cone of finite measures on Polish space $(X, \mathcal{B})$.
- $\mathbb{C} : \mathcal{P} \to 2^{\mathcal{B}}$ is scale invariant, i.e.

$$\mathbb{C}(\alpha P) = \mathbb{C}(P)$$

Hierarchical Clusterings

General Assumptions

- $\mathcal{P}$ is a cone of finite measures on Polish space $(X, \mathcal{B})$.
- $\mathbb{C} : \mathcal{P} \to 2^{\mathcal{B}}$ is scale invariant, i.e.

$$\mathbb{C}(\alpha P) = \mathbb{C}(P)$$

Forest Assumption

- $\mathbb{C}(P)$ is a finite $\perp$-forest, i.e. $|\mathbb{C}(P)| < \infty$ and

$$A, A' \in c(P) \implies A \perp A' \text{ or } A \subset A' \text{ or } A \supset A',$$

where $A \perp A'$ can mean, for example:

$$A \cap A' = \emptyset$$
$$d(A, A') \geq \tau$$
$$\dots$$

Simple Forests

Relations and Structure

- For a forest F we say that A_1, A_2 are direct siblings, if $A_1 \neq A_2$ and they have the same ancestors, i.e.

$$\{A' \in F : A_1 \subset A'\} = \{A' \in F : A_2 \subset A'\}$$

- The structure of a forest F removes all chains, i.e.

$$s(F) := \left\{ A \in F \,\middle|\, A \text{ is a root or it has a direct sibling } A' \in F \right\}$$

Simple Forests

Relations and Structure

- For a forest F we say that A_1, A_2 are direct siblings, if $A_1 \neq A_2$ and they have the same ancestors, i.e.

$$\{A' \in F : A_1 \subset A'\} = \{A' \in F : A_2 \subset A'\}$$

- The structure of a forest F removes all chains, i.e.

$$s(F) := \left\{ A \in F \,\middle|\, A \text{ is a root or it has a direct sibling } A' \in F \right\}$$

Simple Forest Assumption

The map $\mathbb{C} : \mathcal{P} \to 2^{\mathcal{B}}$ is simple, i.e. for all $P \in \mathcal{P}$ we have

$$s(\mathbb{C}(P)) = \mathbb{C}(P)\,.$$

Simple Forests

Relations and Structure

- For a forest F we say that A_1, A_2 are direct siblings, if $A_1 \neq A_2$ and they have the same ancestors, i.e.

$$\{A' \in F : A_1 \subset A'\} = \{A' \in F : A_2 \subset A'\}$$

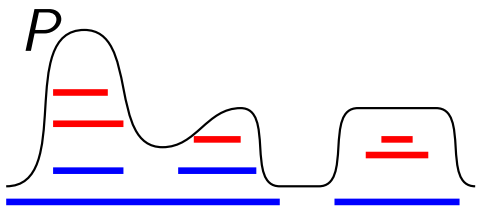
- The structure of a forest F removes all chains, i.e.

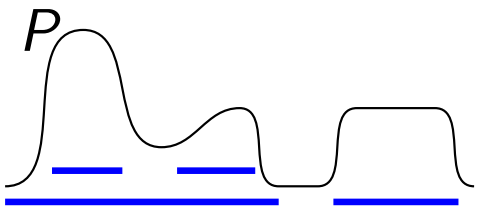
$$s(F) := \left\{ A \in F \,\middle|\, A \text{ is a root or it has a direct sibling } A' \in F \right\}$$

Simple Forest Assumption

The map $\mathbb{C} : \mathcal{P} \to 2^{\mathcal{B}}$ is simple, i.e. for all $P \in \mathcal{P}$ we have

$$s(\mathbb{C}(P)) = \mathbb{C}(P)\,.$$

$\neq \mathbb{C}(P)$ 

$\mathbb{C}(P)$ 

Distributions we are sure about

Base Sets and Distributions

- $\mathcal{Q} = (Q_A)_{A \in \mathcal{A}} \subset \mathcal{P}$ with $\operatorname{supp} Q_A = A$.
- $\mathcal{Q}$ needs to satisfy some additional technical assumptions, which are satisfied by e.g. $dQ_A = \mathbf{1}_A d\mu$.

We assume that we know the clustering for Q_A, namely

$$\mathbb{C}(Q_A) = A\,.$$

Distributions we are sure about

Base Sets and Distributions

- $\mathcal{Q} = (Q_A)_{A \in \mathcal{A}} \subset \mathcal{P}$ with $\operatorname{supp} Q_A = A$.
- $\mathcal{Q}$ needs to satisfy some additional technical assumptions, which are satisfied by e.g. $dQ_A = \mathbf{1}_A d\mu$.

We assume that we know the clustering for Q_A, namely

$$\mathbb{C}(Q_A) = A\,.$$

Examples of Base Sets

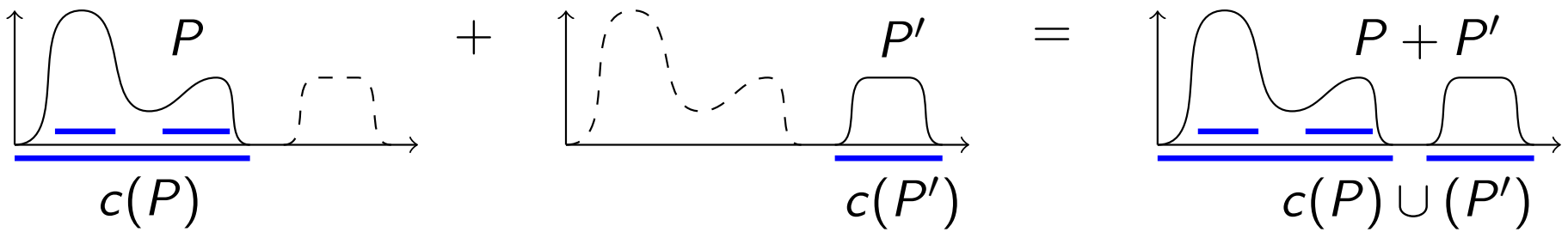
- $\mathcal{A}_{ball}$:= set of all finite connected unions of balls
- $\mathcal{A}_{conv}$:= set of all finite connected unions of convex sets
- $\mathcal{A}_{op-ba}$:= set of all finite unions with connected interior of balls

Disjoint Additivity

Additivity for Measures with Disjoint Support

For all $P, P' \in \mathcal{P}$ we have

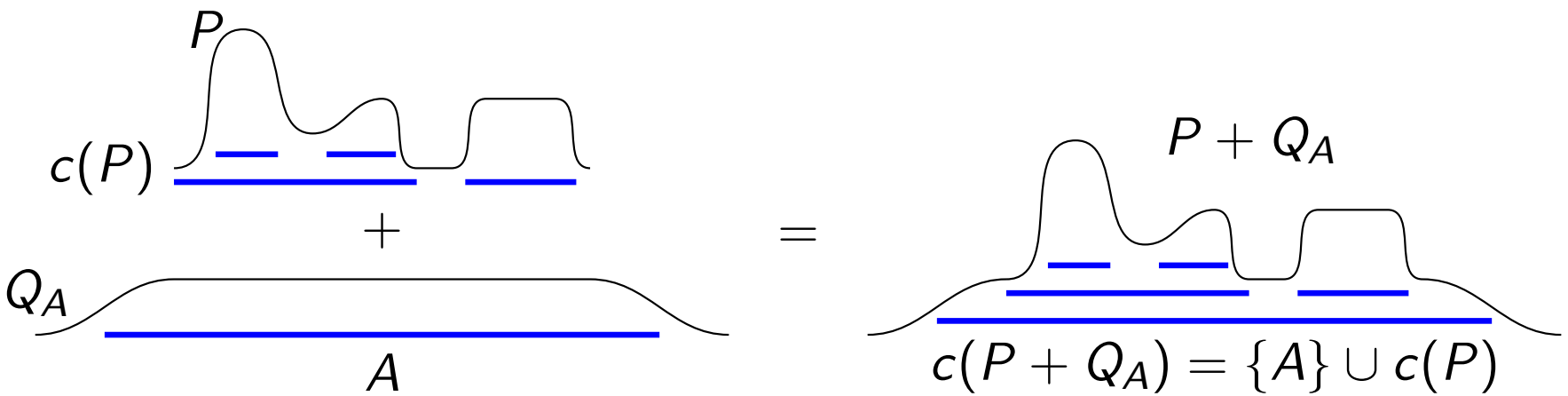
$$\operatorname{supp} P \perp \operatorname{supp} P' \implies P + P' \in \mathcal{P} \text{ and } c(P + P') = c(P) \cup c(P')$$



Hierarchical Structure

Base Additivity For all $P \in \mathcal{P}$ and $A \in \mathcal{A}$ we have

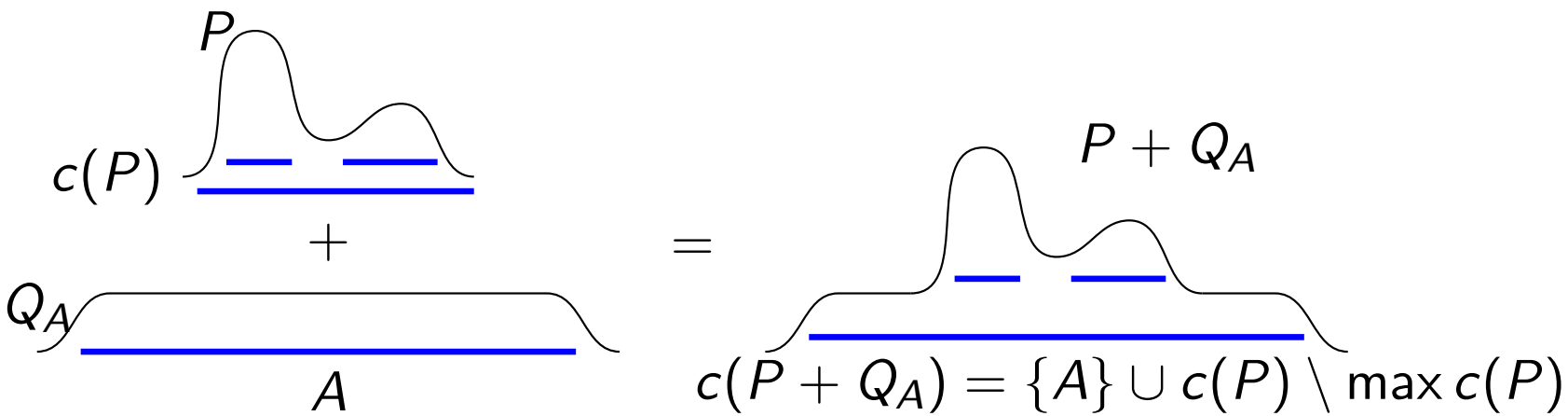
$$\operatorname{supp} P \subset A \quad \Longrightarrow \quad P + Q_A \in \mathcal{P} \text{ and } \mathbb{C}(P + Q_A) = s(\mathbb{C}(P) \cup \{A\})$$



Hierarchical Structure

Base Additivity For all $P \in \mathcal{P}$ and $A \in \mathcal{A}$ we have

$$\operatorname{supp} P \subset A \quad \Longrightarrow \quad P + Q_A \in \mathcal{P} \text{ and } \mathbb{C}(P + Q_A) = s(\mathbb{C}(P) \cup \{A\})$$



Unique Additive Extension

Theorem

There exists a unique clustering $\mathbb{C} : \mathcal{S}(\mathcal{A}) \to 2^{\mathcal{A}}$, where

$$\mathcal{S}(\mathcal{A}) := \left\{ \sum_{A \in F} \alpha_A Q_A \,\middle|\, F \subset \mathcal{A} \text{ is a } \perp\text{-forest and } \alpha_A > 0 \text{ for all } A \in F \right\}$$

It is given by

$$\mathbb{C}\Big(\sum_{A \in F} \alpha_A Q_A\Big) = s(F)\,.$$

Towards Continuity

Why Continuity?

- The distributions we can cluster so far are too simple.
- If we want to estimate the clustering from data, some version of continuity seems to be necessary.

Towards Continuity

Why Continuity?

- The distributions we can cluster so far are too simple.
- If we want to estimate the clustering from data, some version of continuity seems to be necessary.

What could "Continuity" mean?

Continuity is

$$P_n \to P \quad \Longrightarrow \quad \mathbb{C}(P_n) \to \mathbb{C}(P).$$

- The graph structure of $\mathbb{C}(P_n)$ should "converge" to that of $\mathbb{C}(P)$ $\rightsquigarrow$ graph structure should not change for $n \geq n_0$.
- Each leaf of $\mathbb{C}(P_n)$ should converge to the corresponding one of $\mathbb{C}(P)$.

Towards Continuity

Why Continuity?

- The distributions we can cluster so far are too simple.
- If we want to estimate the clustering from data, some version of continuity seems to be necessary.

What could "Continuity" mean?

Continuity is

$$P_n \to P \quad \Longrightarrow \quad \mathbb{C}(P_n) \to \mathbb{C}(P).$$

- The graph structure of $\mathbb{C}(P_n)$ should "converge" to that of $\mathbb{C}(P)$ $\rightsquigarrow$ graph structure should not change for $n \geq n_0$.
- Each leaf of $\mathbb{C}(P_n)$ should converge to the corresponding one of $\mathbb{C}(P)$.
- $P_n \to P$ needs to be specified.

Iso-Monotone Convergence

Iso-Monotonicity of Forests

Two forests F and F' are isomorphic, if there exists a bijection $\zeta : F \to F'$ such that for all $A_1, A_2 \in F$:

$$A_1 \subset A_2 \qquad \Longleftrightarrow \qquad \zeta(A_1) \subset \zeta(A_2)\,.$$

We write $F \leq F'$, if there is a (necessarily unique) monotone ζ, i.e.

$$A \subset \zeta(A)\,, \qquad\qquad A \in F.$$

Iso-Monotone Convergence

Iso-Monotonicity of Forests

Two forests F and F' are isomorphic, if there exists a bijection $\zeta : F \to F'$ such that for all $A_1, A_2 \in F$:

$$A_1 \subset A_2 \qquad \Longleftrightarrow \qquad \zeta(A_1) \subset \zeta(A_2)\,.$$

We write $F \leq F'$, if there is a (necessarily unique) monotone ζ, i.e.

$$A \subset \zeta(A)\,, \qquad\qquad A \in F.$$

Iso-Monotone Convergence

For a iso-monotone sequence of simple forests (F_n), we define

$$F_\infty := \lim_{n\to\infty} F_n := \left\{ \bigcup_{n\geq 1} \zeta_n(A) : A \in F_1 \right\}$$

This gives $F_n \leq F_\infty$.

Iso-Monotone Extension

Limits of Monotone Limits

We define $\overline{\mathcal{S}}(\mathcal{A})$ to be the set of probability measures P for which:

- $\exists(P_n) \subset \mathcal{S}(\mathcal{A})$ with $P_n(B) \nearrow P(B)$ for all $B \in \mathcal{B}$.
- $\mathbb{C}(P_n) \nearrow$.

Main Results

Unique Extension

- There exists a clustering $\mathbb{C} : \overline{\mathcal{S}}(\mathcal{A}) \to 2^{\overline{\mathcal{A}}}$ that is continuous with respect to adapted, iso-monotone sequences $(P_n) \subset \mathcal{S}(\mathcal{A})$ with $P_n \nearrow P$.
- The nodes of $\mathbb{C}(P)$ are P-almost surely unique.

Main Results

Unique Extension

- There exists a clustering $\mathbb{C} : \overline{\mathcal{S}}(\mathcal{A}) \to 2^{\overline{\mathcal{A}}}$ that is continuous with respect to adapted, iso-monotone sequences $(P_n) \subset \mathcal{S}(\mathcal{A})$ with $P_n \nearrow P$.
- The nodes of $\mathbb{C}(P)$ are P-almost surely unique.

Applications

- For $\mathcal{A}_{conv}$ and $\perp_\emptyset$ we obtain density-based cluster trees.
- Other allowed combinations of $\mathcal{A}$ and $\perp$ lead to other clusterings.
- Some clusterings can be combined, e.g. with respect to the Lebesgue measure and some lower dimensional Huasdorff measures.

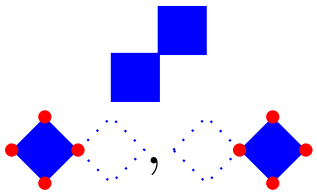
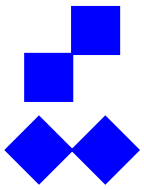
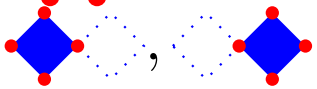
Example

In $\mathbb{R}^2$ we consider the following compact events:

$B :=$ 

$B' :=$ 

By natural approximations we get for $\perp$:

	$\mathcal{A}_1 = (\blacklozenge)$	$\mathcal{A}_2 = (\bullet)$	$\mathcal{A}_\infty = (\blacksquare)$	$\mathcal{A}_{conv}$
B				
B'				

References

I. Steinwart, **Fully adaptive density-based clustering**,
Ann. Statist., vol. 43, pp. 2132-2167, 2015.

P. Thomann, N. Schmid, and I. Steinwart **Towards an axiomatic approach to hierarchical clustering of measures**,
J. Mach. Learn. Res., 2015.