Beyond Hartigan Consistency

Merge Distortion Metric for Hierarchical Clustering

J. Eldridge, M. Belkin, Y. Wang

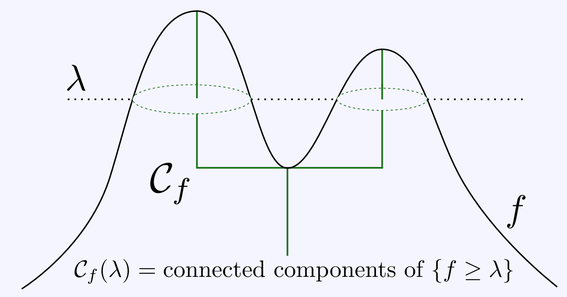
The Ohio State University

Abstract

Does a hierarchical clustering of points sampled from a density converge to the infinite tree underlying the probability distribution? Historically the answer has been "yes" – *if* the clustering method is so-called "Hartigan consistent". However Hartigan consistency is not sufficiently powerful to guarantee that the estimated tree resembles the true tree. We introduce two properties – minimality and separation – which together are stronger than Hartigan consistency and ensure convergence to the true tree in a sense that matches our intuition. Furthermore, we define a merge distortion metric for quantifying the distance between the true tree and a discrete estimate. Convergence in this metric is equivalent to our properties of minimality and separation.

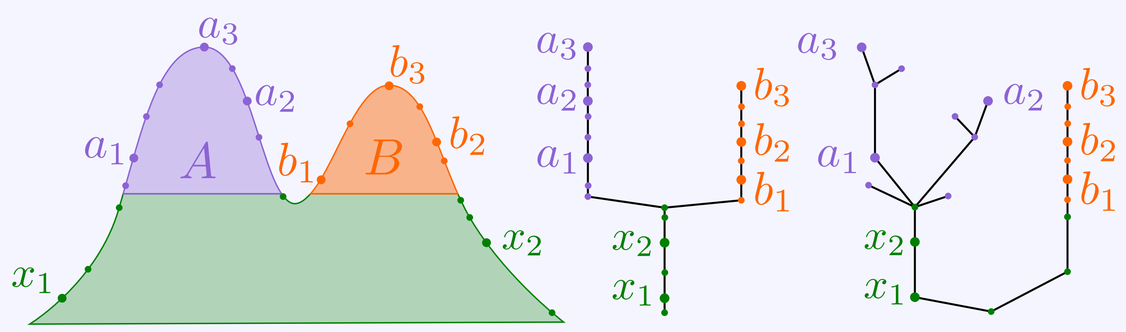
The Cluster Tree

- The *natural clusters* of a density f are the connected components of $\{f \geq \lambda\}$ for any $\lambda > 0$.
- The collection of all such clusters forms an infinite tree, called the *density cluster tree* of f.
- By sampling from f and applying a hierarchical clustering algorithm, a discrete estimate of the true density cluster tree is produced.



Hartigan Consistency

- Hartigan (1981) introduced a notion in which an estimator of the cluster tree is said to be *consistent* with the true tree.
- Consistency ensures that clusters are *well-separated*.
- **Definition**: Draw $X_n \sim f$. Let A_n (resp. B_n) be the smallest empirical cluster containing $A \cap X_n$ (resp. $B \cap X_n$). The clustering method is *Hartigan consistent* if $\Pr(A_n \cap B_n = \emptyset) \to 1$ as $n \to \infty$.
- While desirable, Hartigan consistency alone is *not sufficient*.
- For example, neither of the estimated trees shown to the right violates Hartigan consistency, yet they are very different from one another.



- In particular, Hartigan consistency permits *over-segmentation*.
- Such undesirable configurations are not ruled out precisely because Hartigan consistency lacks a strong notion of *connectedness*.

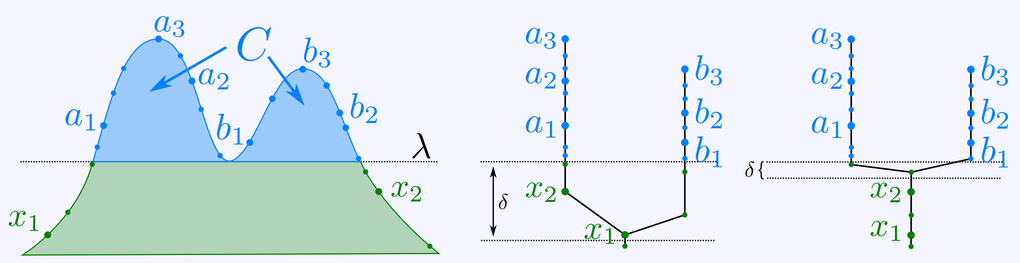
Hartigan consistency lacks a notion of *connectedness* and thus allows over-segmentation.

We replace Hartigan consistency by two new notions: *Minimality* and *Separation*.

Together, Minimality and Separation are stronger than (and, in fact, imply), Hartigan consistency.

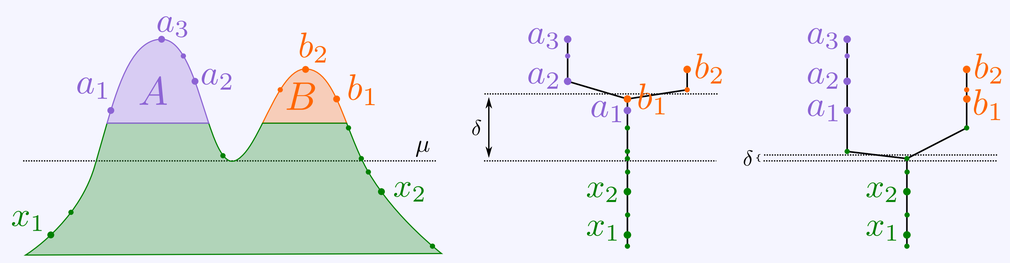
Minimality

- We introduce *minimality* to ensure connectedness.
- **Definition:** A clustering method ensures *minimality* if given any connected component C of the superlevel set $\{f \geq \lambda\}$, $C \cap X_n$ is connected at level $\lambda - \delta$ for any $\delta > 0$ as $n \to \infty$.



Separation

- We introduce *separation* as a weaker version of Hartigan consistency.
- **Definition:** A clustering method ensures *separation* if given any disjoint Clusters A and B merging at level μ, $A \cap X_n$ and $B \cap X_n$ are separated at level $\mu + \delta$ for any $\delta > 0$ as $n \to \infty$.



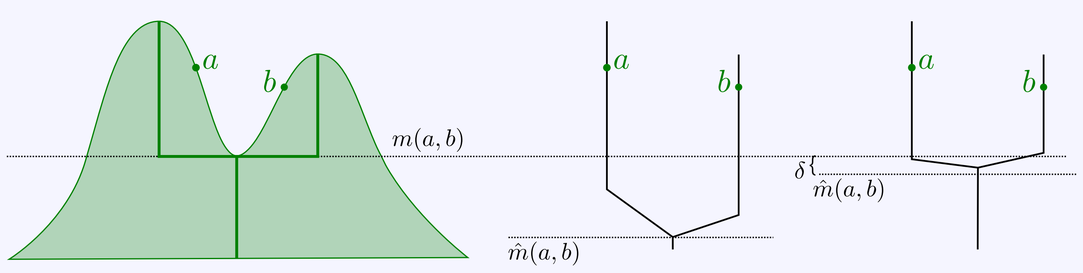
Minimality + Separation $\iff$ Convergence

Merge Distortion Metric

- We introduce a *merge distortion metric* in order to quantify the disparity between the true tree $\mathcal{C}_f$ and an estimate:

$$d(\mathcal{C}_f, \hat{\mathcal{C}}_{f,n}) = \max_{x,x' \in X_n} |m(x,x') - \hat{m}(x,x')|,$$

  where m and $\hat{m}$ are the merge height functions of the true tree and the estimated tree, respectively.
- Convergence in the merge distortion metric is *equivalent* to the combination of uniform minimality and uniform separation.



Do algorithms exist which converge?

- Convergence in our metric is a stronger property than Hartigan consistency – we must show that it is reasonable.
- We have shown that two algorithms converge in our metric:
  - *Robust single linkage* of Chaudhuri and Dasgupta (2008).
  - A split-tree-based algorithm drawing on the work of Chazal et al. (2013).

Stability

- The true density cluster tree is stable under our metric with respect to L_∞ perturbation of the density.
- An additional stability result suggests that we may confidently use the distance between two consecutive estimated trees as a criterion for when to stop sampling.