Statistical Inference for Cluster Trees

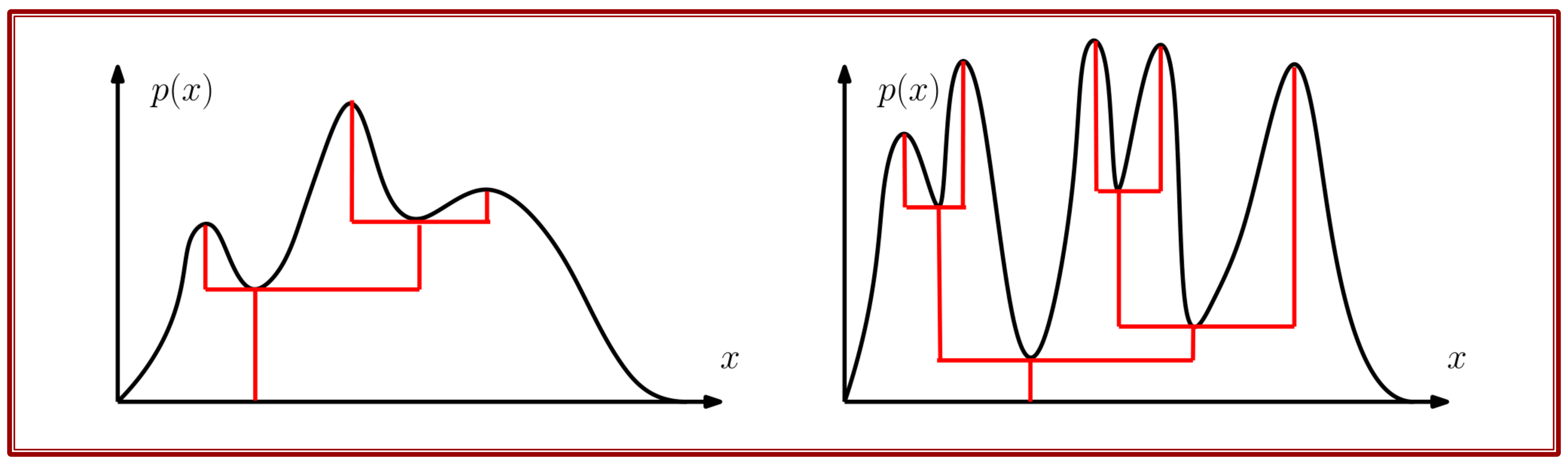
Jisu Kim, Yen-Chi Chen, Sivaraman Balakrishnan,
Alessandro Rinaldo, Larry Wasserman

Abstract

A cluster tree provides an interpretable summary of a density function by representing the hierarchy of its high-density clusters. It is estimated using the empirical tree, which is the cluster tree constructed from a density estimator. This paper assesses the statistical significance of features of an empirical cluster tree. We first study a variety of metrics that can be used to compare different trees, analyze their properties, and assess their suitability for inference. We then propose methods to construct and summarize confidence sets for the unknown true cluster tree.

Background and Definitions

- For any function f, the cluster tree of f is a function T_f, where $T_f(\lambda)$ is the set of the connected components of the upper-level set $\{x: f(x) \geq \lambda\}$.



- For two points x, y and a tree T_f, their merge height $m_f(x,y)$ is
$$m_f(x,y) = \sup\{\lambda \in \mathbb{R}: there\ exists\ C \in T_f(\lambda)\ such\ that\ x, y \in C\}$$
- An asymptotic $1-\alpha$ confidence set C_α is a collection of trees with the property that $P(T_{p_0} \in C_\alpha) = 1 - \alpha + o(1)$.
- For two trees T_f and T_g, we say $T_f \preccurlyeq T_g$ if there exists a map $\Phi: \cup_\lambda T_f(\lambda) \to \cup_\lambda T_g(\lambda)$ such that for any C_1, C_2 in $\cup_\lambda T_f(\lambda)$, $C_1 \subset C_2$ iff $\Phi(C_1) \subset \Phi(C_2)$.

Tree Metrics

Three metrics on cluster trees

1. l_∞ metric: $d_\infty(T_p, T_q) = \sup|p(x) - q(x)|$
2. Merge distortion metric: $d_M(T_p, T_q) = \sup|m_p(x,y) - m_q(x,y)|$
3. Modified merge distortion metric: $d_{MM}(T_p, T_q) = \sup|d_{T_p}(x,y) - d_{T_q}(x,y)|$, where $d_{T_p}(x,y) = p(x) + p(y) - 2m_p(x,y)$

- d_∞ and d_M are equivalent, and d_{MM} is sandwiched by d_∞. (Lemma 1)
- Large sample behavior of d_∞ is well known, while d_{MM} is not point-wise Hadamard differentiable, i.e. statistically unstable (Appendix C)

Confidence Sets

- Let biased density $p_h(x) = E[\hat{p}_h(x)]$, then T_{p_h} and T_{p_0} are equivalent if h is small enough and p_0 is regular enough (Lemma 2).
- We compute confidence set for T_{p_h}, since T_{p_h} is estimated at rate $O_P(n^{-1/2})$ while T_{p_0} is estimated at rate $O_P(n^{-2/(4+d)})$.

A data-driven Confidence set

We use bootstrap to compute $1-\alpha$ confidence set C_α.

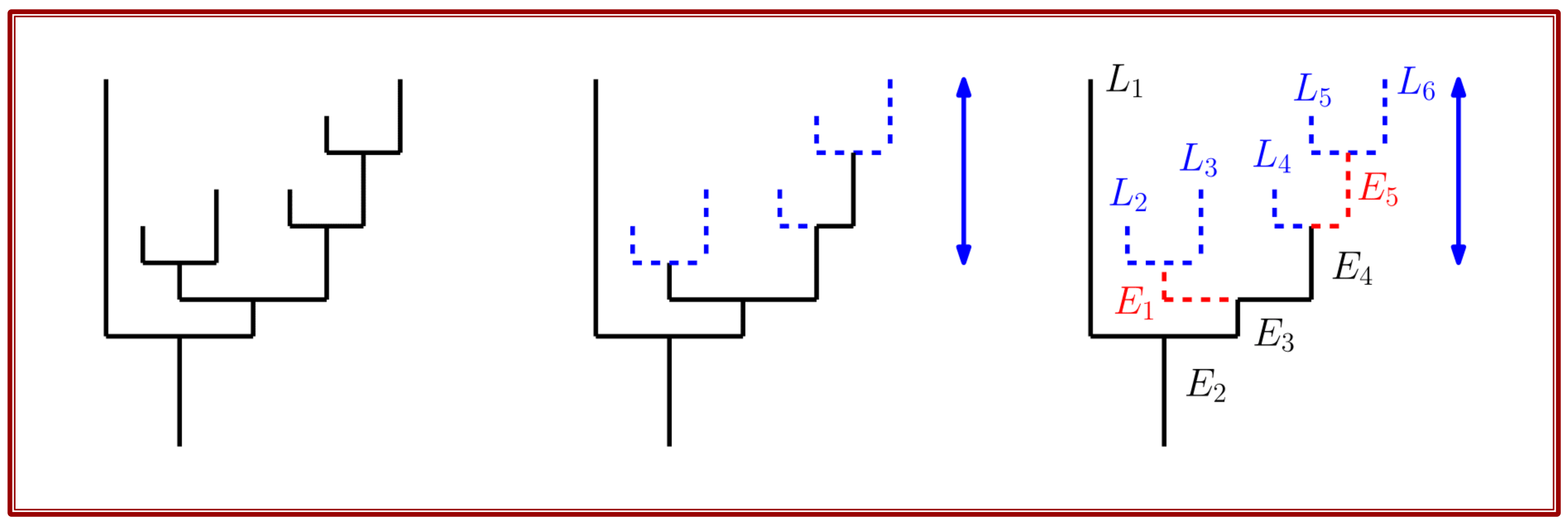
1. Generate B bootstrap samples.
2. On each bootstrap sample, we compute the KDE and the associated cluster tree. Denote those trees as $\{\tilde{T}^1_{p_h}, \ldots, \tilde{T}^B_{p_h}\}$.
3. We estimate t_α, $1-\alpha$ quantile of $d_\infty(T_{p_h}, T_{\hat{p}_h})$, by $\hat{t}_\alpha = \hat{F}^{-1}(1-\alpha)$, where $\hat{F}(s) = \frac{1}{B}\sum_{i=1}^{n} 1(d_\infty(\tilde{T}^i_{p_h}, T_{\hat{p}_h}) < s)$.
4. The data-driven confidence set is $C_\alpha = \{T: d_\infty(T, T_{\hat{p}_h}) \leq \hat{t}_\alpha\}$.

This is consistent at a rate $O(((\log n)^7/n)^{1/6})$ (Theorem 3).

Probing the Confidence set

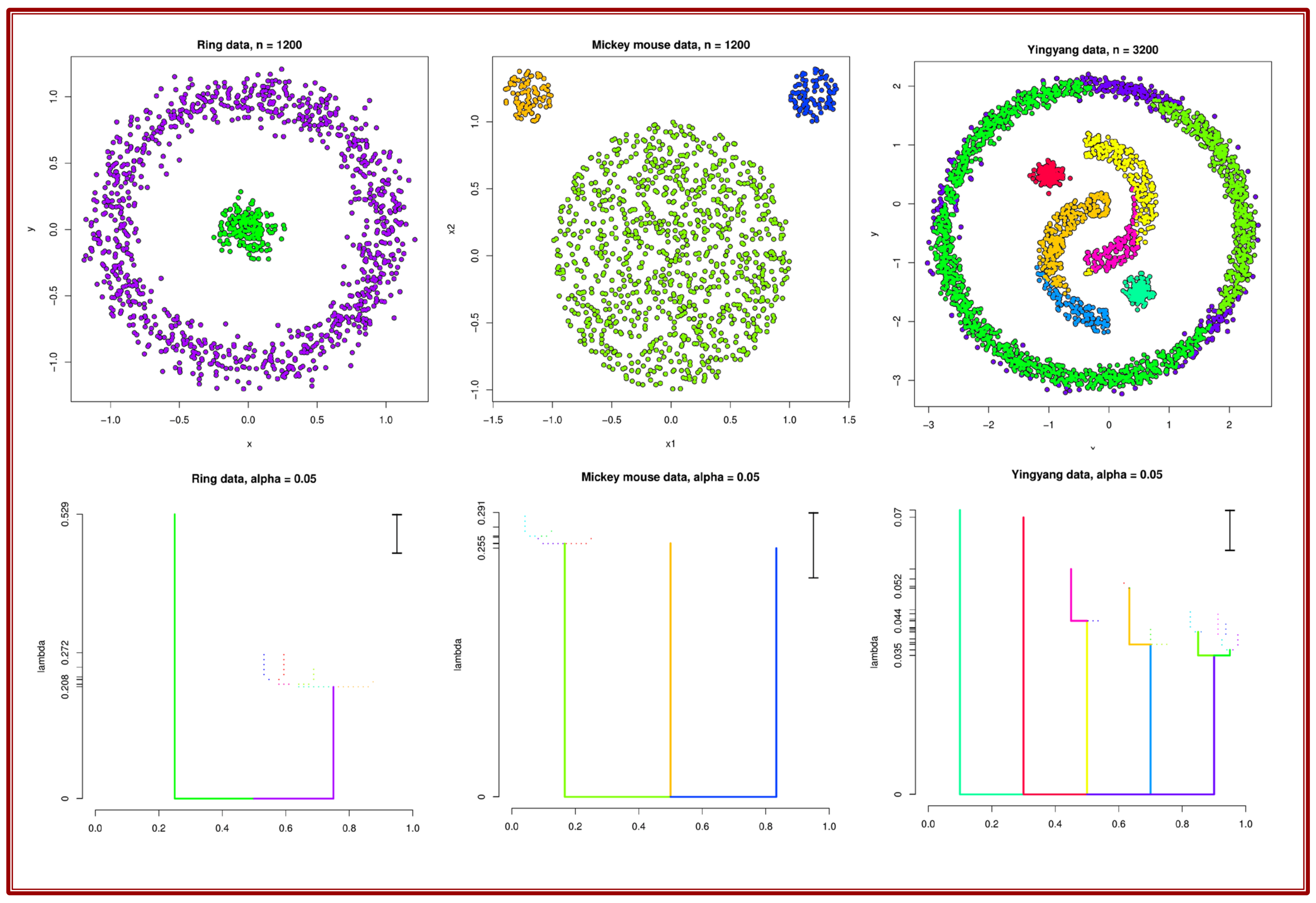
We propose two pruning schemes to find trees, that are simpler than the empirical tree T_{p_h} and are in the confidence set.

1. Pruning only leaves: Remove all leaves of length less than $2\hat{t}_\alpha$.
2. Pruning leaves and internal branches: Iteratively remove all branches of cumulative length less than $2\hat{t}_\alpha$.



Experiments

Simulated data



GvHD dataset