Robust entropy estimation via pruned minimal spanning trees

Alfred Hero
University of Michigan - Ann Arbor
http://www.eecs.umich.edu/ hero/hero.html

OUTLINE

1. Entropy estimation
2. Minimal spanning trees (MST)
3. Beardwood, Halton, Hammersley (BHH) theorem for MST
4. Minimal k-point spanning trees and greedy approximations
5. Extension of BHH theorem to k-MST
6. Quantitative robustness of k-MST via influence function
7. Applications to clustering and image registration

1. Entropy Estimation

Let $\{X_i\}_{i=1}^n$ be a *point cloud* in $\mathbb{R}^d$, $d \geq 1$.

- $\{X_i\}_{i=1}^n$ are i.i.d.random vectors with **unknown** p.d.f. $f(x)$, $x \in \mathbb{R}^d$.

Define Renyi Entropy of fractional order $\nu \in (0,1)$

$$H_\nu(X) = \frac{1}{1-\nu} \ln \int_{\mathbb{R}^d} f^\nu(x) dx$$

- $\nu = 1/2$:

$$H_{\frac{1}{2}}(X) = 2 \ln \int_{\mathbb{R}^d} \sqrt{f(x)} dx \qquad (Hellinger)$$

- $\nu = 1$:

$$\lim_{\nu \to 1} H_\nu(X) = -\int_{\mathbb{R}^d} f(x) \ln f(x) dx \qquad (Shannon)$$

Objective: Non-parametric estimation of Renyi entropy of $f(x)$ based on realization $\{x_i\}_{i=1}^n$.

Entropy estimation applications:

- Lyapounov exponents of fractal and other non-linear processes (Takens)
- pattern recognition and pattern matching (D. Geman)
- image registration for multiple MRI studies (Collignon, Meyer)
- determining optimal cell density for adaptive VQ (Gersho, Neuhoff)
- quadtree termination rules for non-linear regression trees (Breiman, Hastie)
- stopping rules for projection pursuit regression (Freidman)
- Error exponent estimation from empirical measurements

Current non-parametric entropy estimation methods are based on density estimation

$$\hat{H}_\nu = \frac{1}{1-\nu} \ln \int_{\mathbb{R}^d} \hat{f}^\nu(x) dx$$

<u>Difficulties</u>

- kernel or histogram estimation is unstable esp. for large d
- d-dimensional integration in H_ν can be impractical
- asymptotic analysis is complicated
- unclear how to robustify $\hat{f}$ against outliers
- $\Rightarrow$ function $\{f(x) : x \in \mathbb{R}^d\}$ over-parameterizes entropy functional

2. Minimal Spanning Trees (MST)

For n points $x_i \in \mathbb{R}^d$ define the complete graph $\mathcal{G}$ by

- n vertices x_i
- $\binom{n}{2}$ edge weights $e = e_{ij}$

Total weight of graph:

$$L_n = \sum e_{ij}$$

A *spanning tree* $\mathsf{T}_n = \mathsf{T}(x_1, \ldots, x_n)$ is an acyclic subgraph of $\mathcal{G}$. It has weight

$$L_n = \sum_{e \in \mathsf{T}_n} e$$

The *minimal spanning tree* (MST) $\mathsf{T}_n^* = \mathsf{T}^*(x_1, \ldots, x_n)$ is the spanning tree having minimum weight

$$L_n^* = \min_{\mathsf{T}_n} \sum_{e \in \mathsf{T}_n} e$$

Previous statistical applications of MST techniques

- Clustering: Zahn (1971), Toussaint (1980)
- Invariant pattern recognition: Duda&Hart (1973)
- Testing for randomness: Hoffman&Jain (1983)
- Non-parametric regression: Banks (1993)

Examples

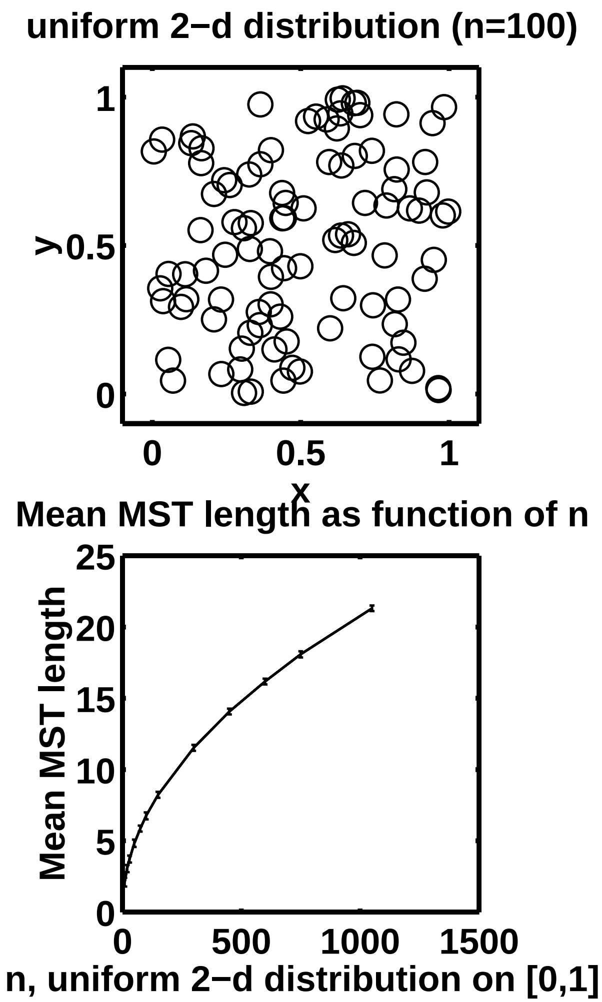


Figure 1: 2D Uniform sample study.

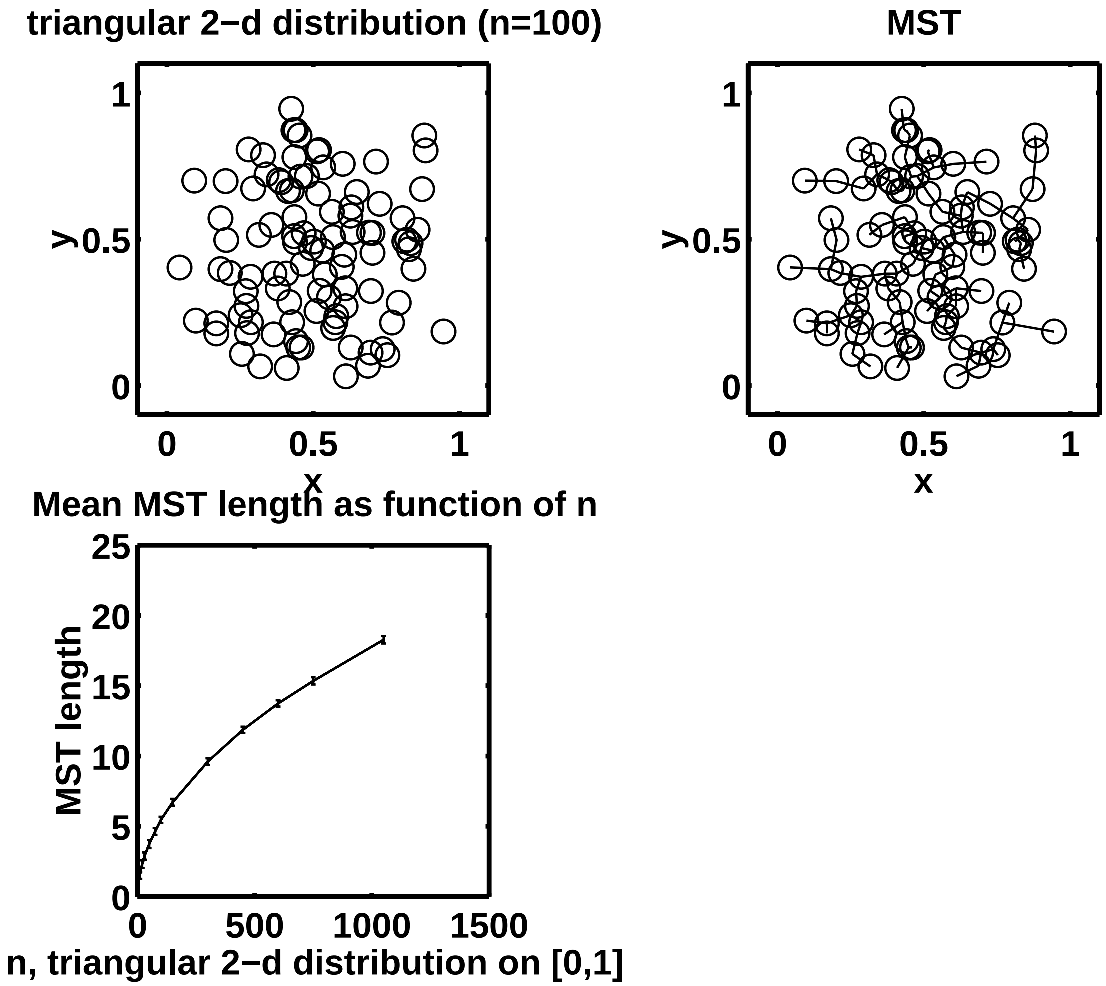


Figure 2: 2D Triangular sample study.

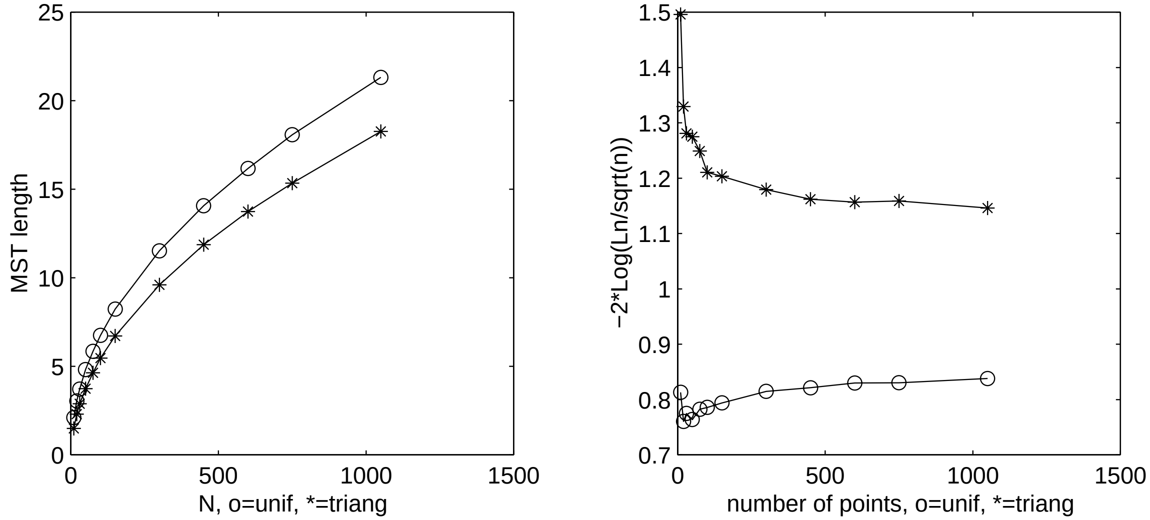


Figure 3: MST Length and log length comparisons.

3. Asymptotic theory of MST entropy estimator

Let $e_{ij} = \|x_i - x_j\|$ and specialize L_n to weighted norm

$$L_n = \sum e_{ij}^{\gamma}, \qquad \gamma \in (0, d)$$

Steele's (1988) version of the Beardwood, Halton, Hammersley (1959) Theorem

> Let $\{X_i\}_{i=1}^{n}$ be an i.i.d sequence of random variables with p.d.f. $f(x)$ having compact support in $\mathbb{R}^d$, $d > \gamma > 0$. Then
>
> $$L_n^*/n^{(d-\gamma)/d} \;\rightarrow\; \beta_{L,\gamma} \int_{\mathbb{R}^d} f^{(d-\gamma)/d}(x)dx \qquad (w.p.1)$$

Thus, as $n \to \infty$

$$\widehat{H}_\nu(X) = \frac{1}{1-\nu}\left(\ln L_n^*/n^\nu - \ln \beta_{L,\gamma}\right) \;\to\; H_\nu(X), \qquad (w.p.1)$$

where:

$$\nu = (d-\gamma)/d$$

Ingredients behind proof

First assume $\gamma = 1$ and $f(x) =$ uniform over unit cube $[0,1]^d$

1. Sphere packing bound on min nearest neighbor distances:

$$e_j^{nn} = min_i e_{ij} \le \frac{c}{n^{1/d}}$$

   $c = 2\sqrt{d} =$ const

2. By chaining nearest neighbors this gives bound on MST length

$$L_n = \sum_{ij=1}^{n-1} e_{ij} \le \sum_{j=1}^{n} e_j^{nn} \le n \, \frac{c}{n^{1/d}} = c \, n^{(d-1)/d}$$

3. Next use fact that L_n^* is "quasi-additive" and continuous (Redmond and Yukich (1996)):

   For any partition of $[0,1]^d$ into cubes Q_j of side $1/m$

$$L_n^*(F) = \sum_{j=1}^{m^d} L_n^*(F \cap Q_j) + o\left(m^{d-1}\right)$$

4. The above can be used to show that for uniform $f(x)$

$$L_n^*/n^{(d-1)/d} \;\to\; \beta_{L,\gamma}$$

This last result generalizes to blocked densities

$f(x) = \frac{1}{m^d} \Sigma_{i=1}^{m^d} \alpha_i I_{Q_i}(x)$

via quasi-additivity:

$$L_n^*([0,1]^d) \;=\; \sum_{j=1}^{m^d} L_n^*(Q_j) + o\left(m^{d-1}\right)$$

Indeed

$$\overbrace{\text{MST over } \alpha_i n \text{ pts in } [0, m^{-d}]}^{}\ \underbrace{}_{L_n^*(Q_j)} \Big/ (n\alpha_i)^{(d-1)/d} \;\to\; \frac{1}{m}\beta_{L,\gamma}$$

and therefore:

$$L_n^*([0,1]^d)/n^{(d-1)/d} \;\rightarrow\; \frac{1}{m^d}\sum_{j=1}^{m^d}\left(\frac{\alpha_i}{m^d}\right)^{(d-1)/d} = \int_{[0,1]^d}[f(x)]^{(d-1)/d}dx$$

Illustration of non-Robustness of MST

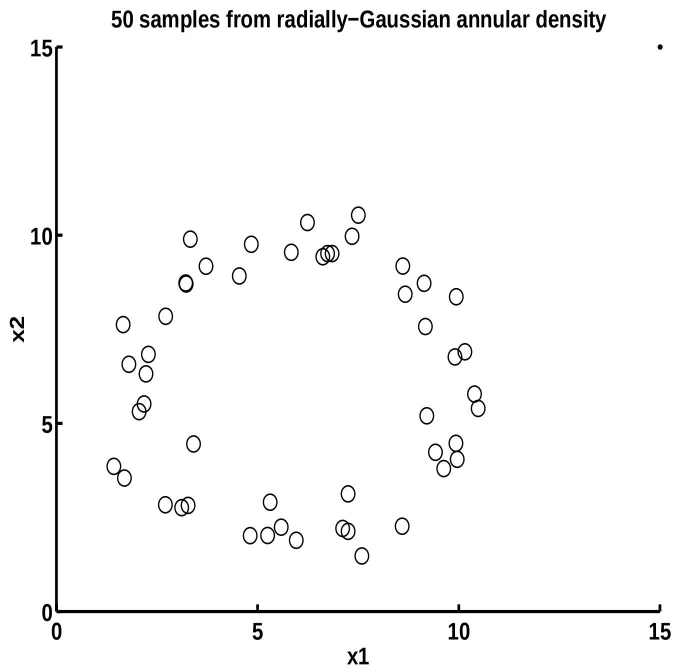
Sample from 2D mixture density

$$f(x) = (1-\epsilon) f_1(x) + \epsilon f_0(x), \qquad 0 < \epsilon \ll 1$$

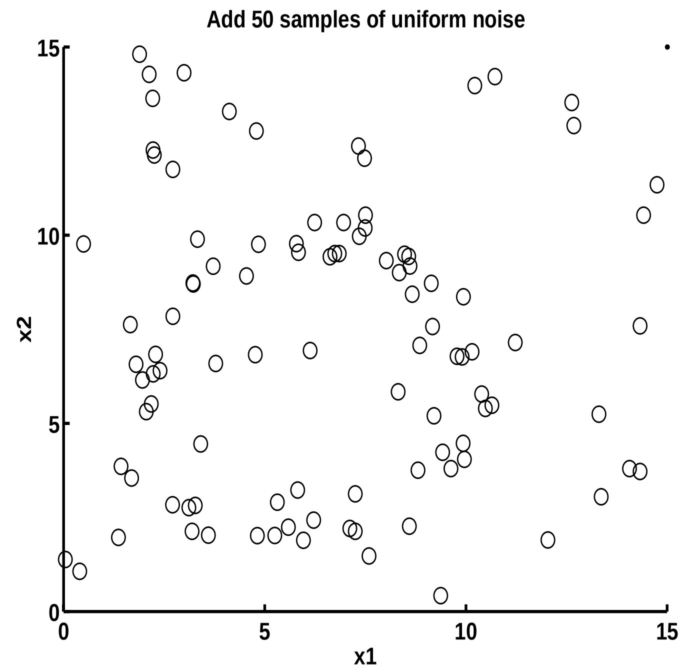
f_1 - annular Gaussian density: the target

f_0 - spatially homogeneous noise density

(a)

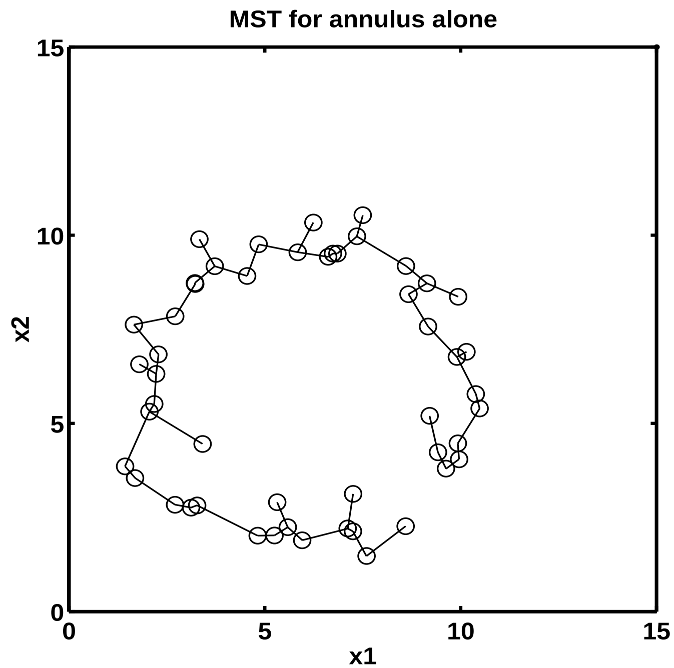


(b)

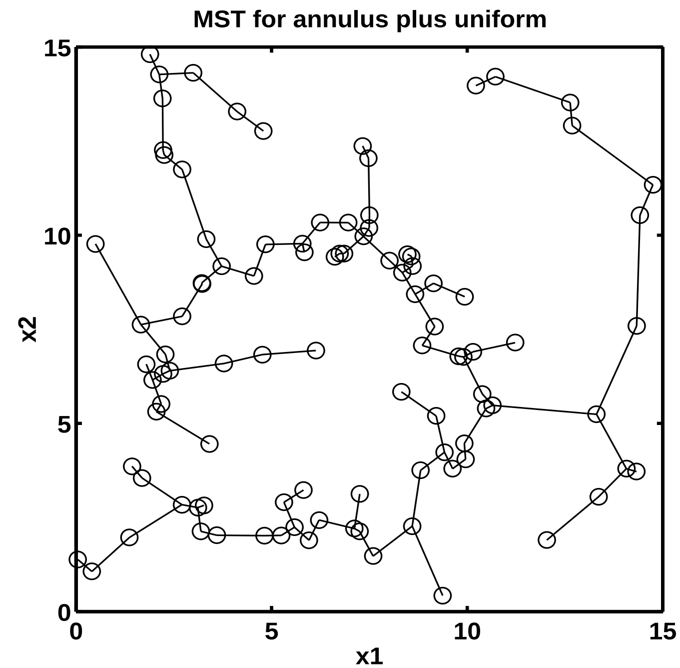


$\epsilon = 0$

$\epsilon = 0.5$


MST for annulus alone
15
10
5
0
0
5
10
15
x1
x2

(a)


MST for annulus plus uniform
15
10
5
0
0
5
10
15
x1
x2

(b)

Banks' MST Pruning Algorithm in 2D (1993)

Fix $k = \lfloor \alpha n \rfloor$, $0 < \alpha < 1$.

1. $Grow$ a full MST on $x_1, \ldots, x_n$
2. $Rank\ order$ edges: $e^*_{(1)} < \ldots < e^*_{(n)}$ in L^*_n
3. $Trim$ MST and L^*_n by eliminating $n - k$ largest edges
4. $Eliminate$ isolated trees resulting from trimmed MST
5. Use remaining $Trunk$ of tree as partial spanning tree

Difficulties:

- unclear how to choose α
- final tree is not an MST
- Asymptotic statistics of "α-trimmed" L^*_n are intractible
- cannot analyse theoretical robustness

4. k-Minimal Spanning Tree (k-MST)

Fix k, $1 \leq n$.

Let $\mathsf{T}_{n,k} = \mathsf{T}(x_{i_1}, \ldots, x_{i_k})$ be a spanning tree connecting k distinct vertices $x_{i_1}, \ldots, x_{i_k}$ of complete graph $\mathcal{G}$.

The *k-minimal spanning tree* (k-MST) $\mathsf{T}^*_{n,k} = \mathsf{T}^*(x_{i_1^*}, \ldots, x_{i_k^*})$ is the minimum weight MST among the $\binom{n}{k}$ MST's connecting subsets of k vertices of $\mathcal{G}$:

$$L^*_{n,k} = \min_{i_1,\ldots,i_k} \min_{\mathsf{T}_{n,k}} \sum_{e \in \mathsf{T}_{n,k}} e$$

Proposed k-MST pruning algorithm

1. *Grow* sequence of $n-k+1$ k-MST's on $x_1, \ldots, x_n$:
   $L^*_{n,n} > \ldots > L^*_{n,k}$
2. *Detect* breakpoint $i = i_{bk}$ in $L^*_{n,i}$ curve
3. use k-MST with $k = i_{bk}$ as partial spanning tree

Attractive Properties

- breakpoint detection is natural α selection rule
- k-MST is a MST
- asymptotics of $L^*_{n,k}$ can be studied
- can analyse theoretical robustness

Example

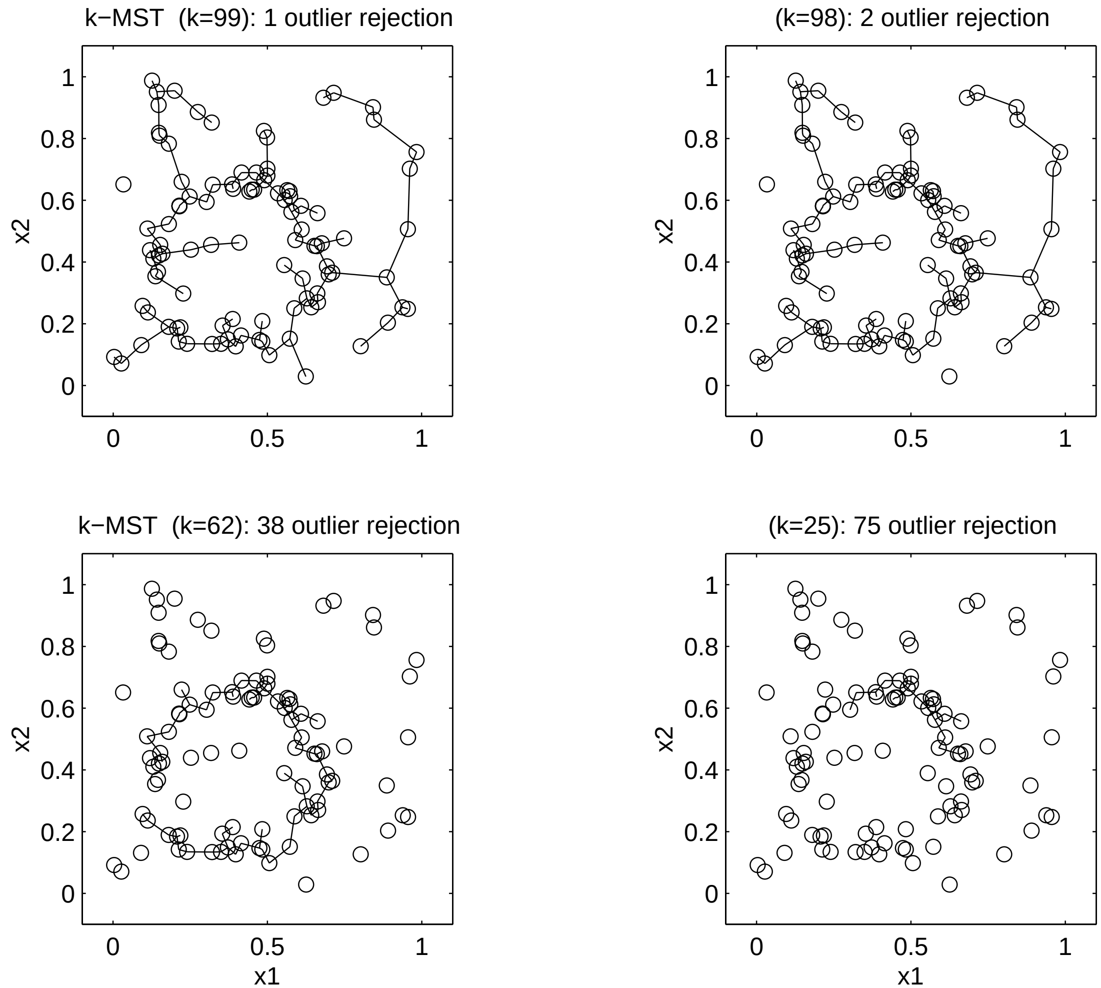


Figure 6: MST for 2D torus density with and without the addition of uniform “outliers”.

Prediction criterion

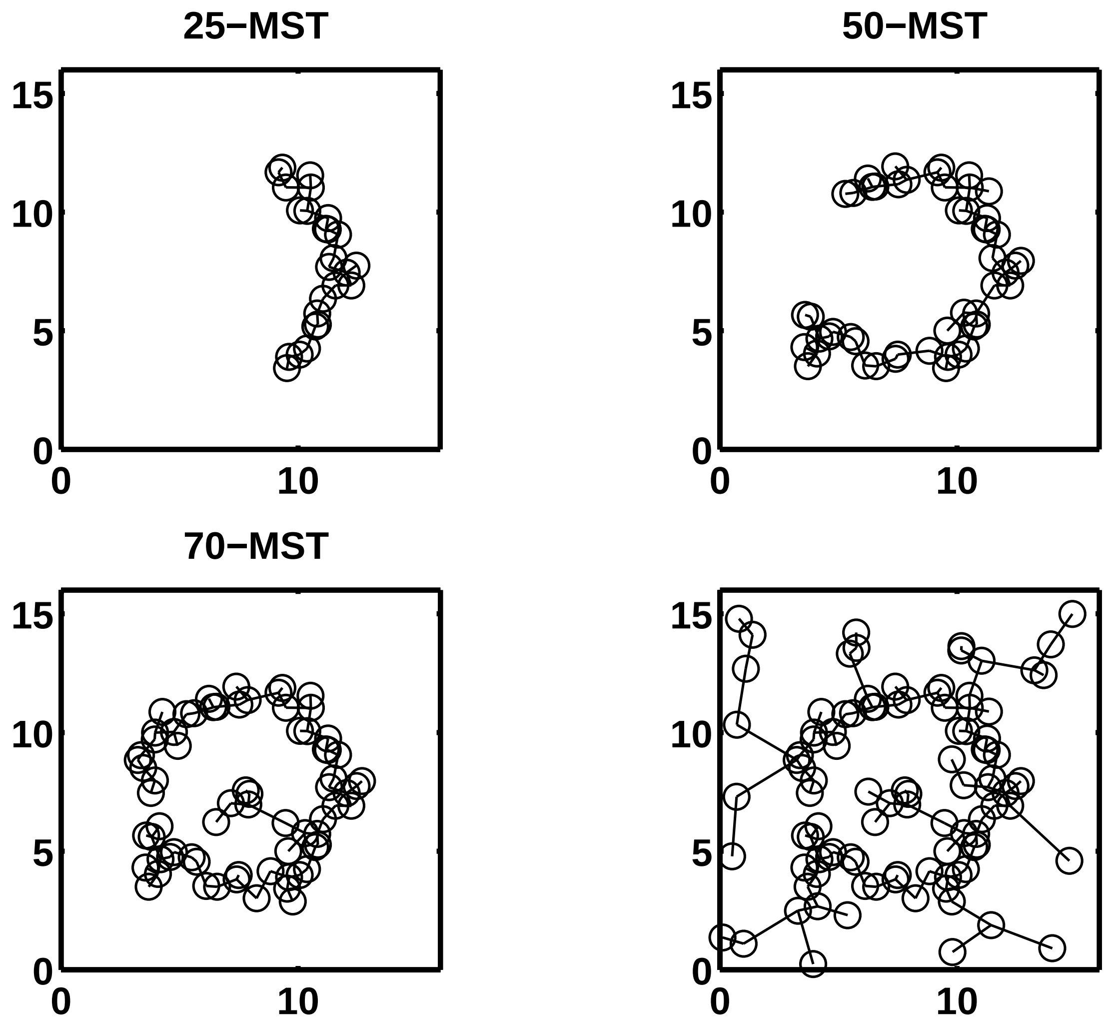


Figure 7: Annulus density examples.

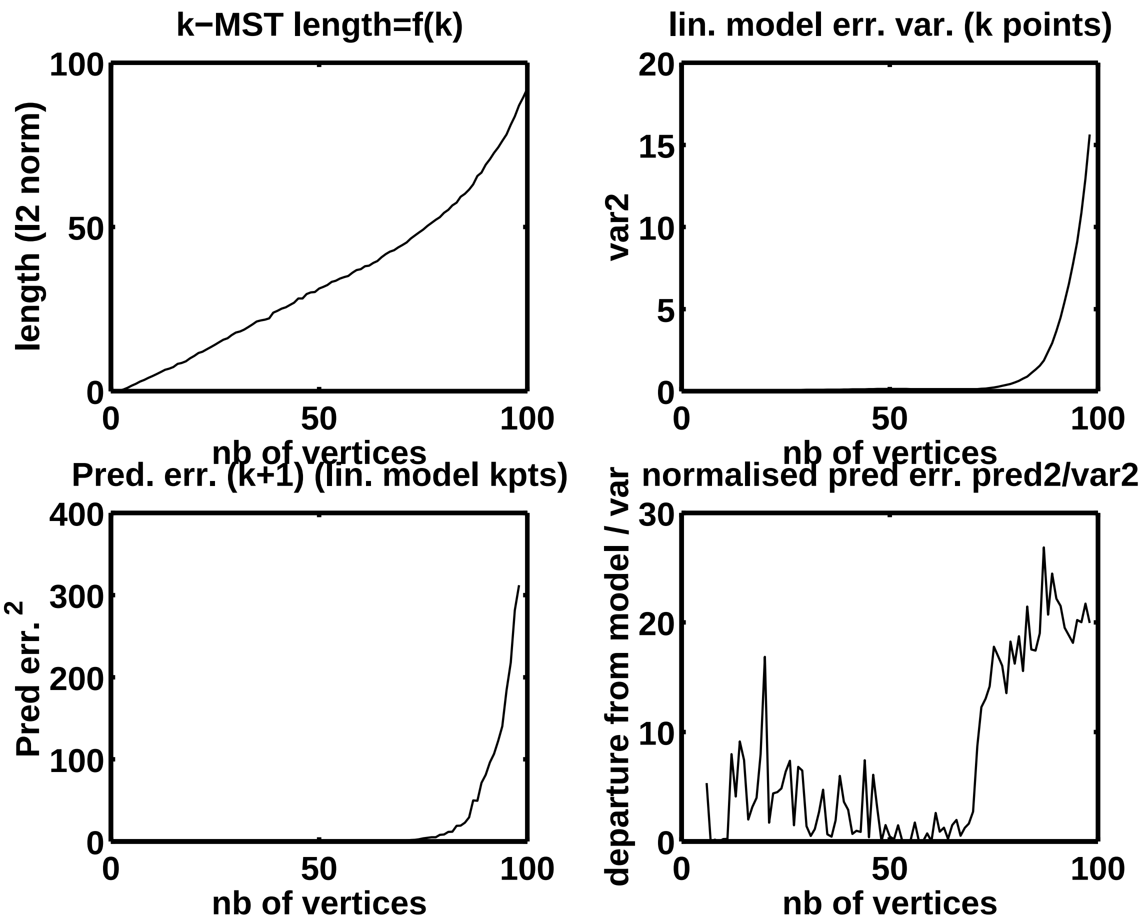


Figure 8: Prediction criterion for threshold determination of optimal pruned k-MST

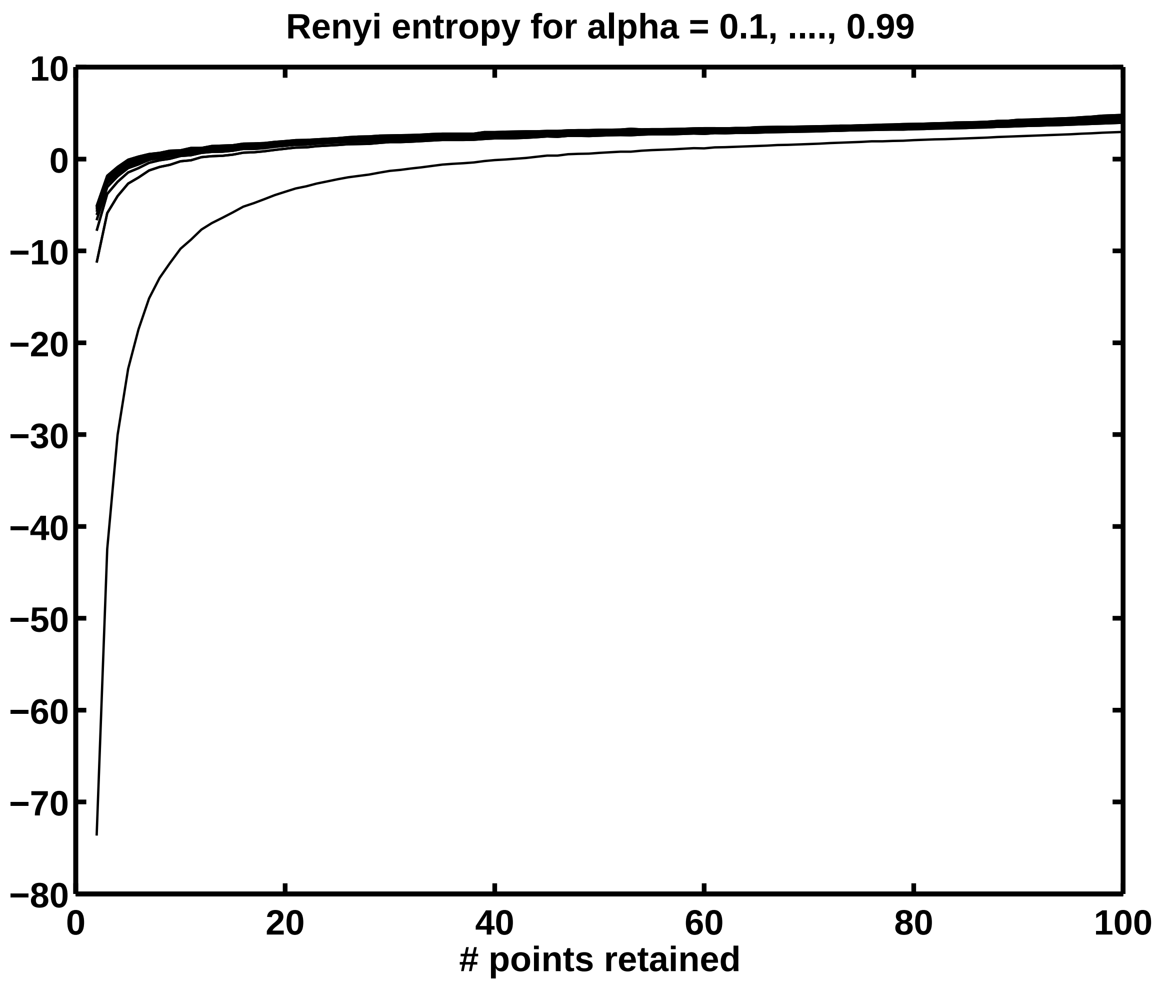


Figure 9: Renyi entropy of various orders for annulus example.

Brief history

Exact MST and k-MST algorithms

- MST for general undirected graphs
  $O(n^2)$ ops: Kruskal (1956), Dijkstra (1959) ($d = 2$)
- MST for undirected graphs on $\mathbb{R}^d$
  $O(n(n \log n)^{(1-2^{-(d+1)})})$ ops: Yao (1982)
- k-minimum MST algorithm
  $O(n^{n-k})$ ops: Zelikovsky and Lozevanu (1993)

ϵ-optimal MST and k-MST algorithms

- $L_n^{**}/L_n^* = 1 + \epsilon$, $O(\epsilon^{-k}\, n \log n)$ ops: Vaidya (1984, 1988)
- $L_{n,k}^{**}/L_{n,k}^* = O(\log k)$, $O(n^2 k^4 + n^3)$ ops: Garg&Hochbaum (1994)

- $L^{**}_{n,k}/L^{*}_{n,k} = O(k^{1/4})$: Ravi, Sundaram, Marathe, Rosenkrantz, Ravi (1994)
- $L^{**}_{n,k}/L^{*}_{n,k} \leq 3$, $O(n^2k^4 + n^3)$ ops: Hochbaum (1996)
- $L^{**}_{n,k}/L^{*}_{n,k} = 1 + \epsilon$, $n^{O(1/\epsilon)}$ ops: Mitchell (1996)
- $L^{**}_{n,k}/L^{*}_{n,k} = 1 + \epsilon$, $O(nk(\log(k))^{O(1/\epsilon)}$ ops: Arora (1997)

Greedy Approximation to k-MST

4 steps to approximation

1. user specifies a positive integer m
2. user specifies a uniform partition $\mathcal{Q}^m$ of $[0,1]^d$ having m^d cells Q_i of resolution $1/m$;
3. user runs algorithm to find the smallest subset $B_k^m = \cup_i Q_i$ of partition elements containing at least k points;
4. on this reduced subset the algorithm runs k-MST.

Ravi's Greedy Subset Selection Algorithm

Intialize: $B = \phi$, $j = 1$

Sort Q_i in decreasing order of $\mathrm{card}\{Q_i\}$

Do until $\mathrm{card}\{\mathcal{X}_n \cap B\} \geq k$

$B = B \cup Q_{(j)}$

End $j = j + 1$

Note: smallest subset found by algorithm is not unique!

Example of non-uniqueness

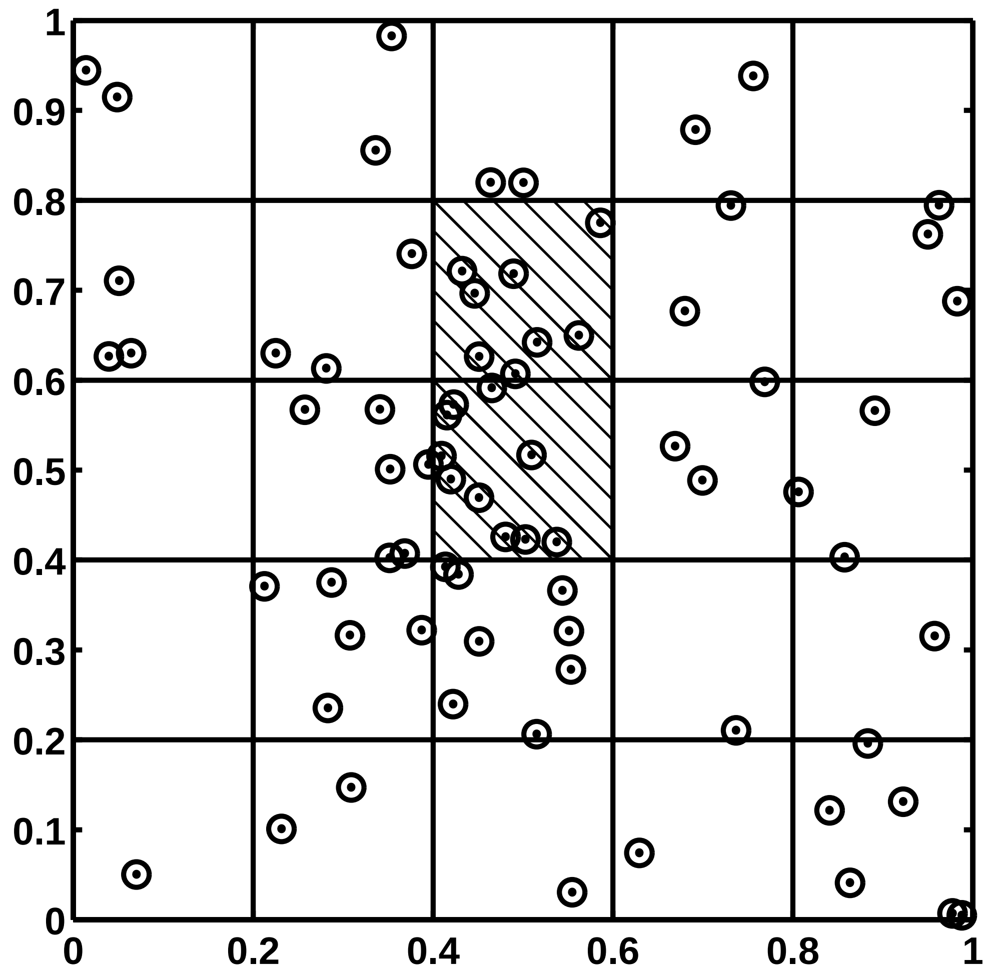


Figure 11: A sample of 75 points from the mixture density $f(x) = 0.25 f_1(x) + 0.75 f_o(x)$ where f_o is a uniform density over $[0,1]^2$ and f_1 is a bivariate Gaussian density with mean $(1/2, 1/2)$ and diagonal covariance diag(0.01). A smallest subset B_k^m is the union of the two cross hatched cells shown for the case of $m = 5$ and $k = 17$.

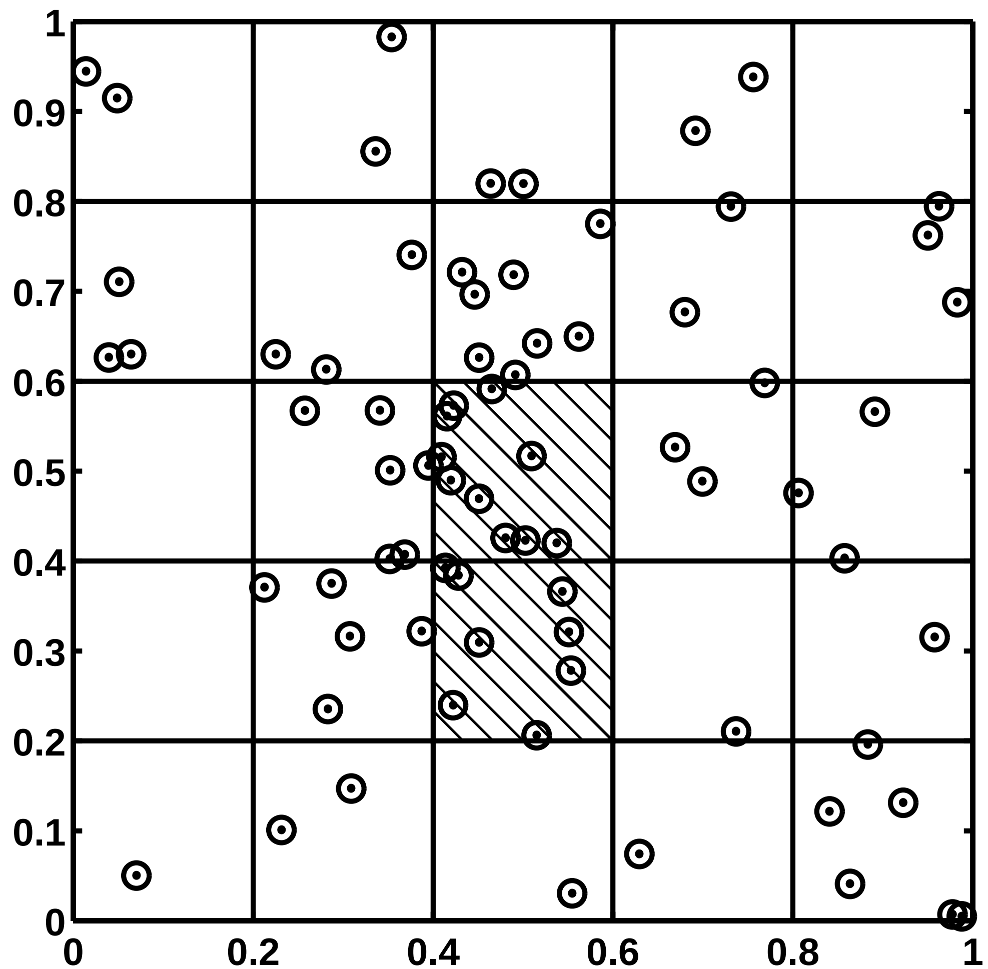


Figure 12: Another smallest subset B_k^m containing at least $k = 17$ points for the mixture sample shown in previous Fig

5. A BHH Theorem for k-MST

Fix α and let $k = \lfloor \alpha n \rfloor$ be number of retained points in k-MST

We will let $n \to \infty$ and investigate behavior of $L^*_{n,\lfloor \alpha n \rfloor}$

Two basic steps

Step 1: Index the $\binom{n}{\lfloor \alpha n \rfloor}$ vertices by unions B of the m^d partition cells

Then for large n

$$\underbrace{\min_{i_1,\ldots,i_{\lfloor \alpha n \rfloor}} L^*(X_{i_1}, \ldots, X_{i_{\lfloor \alpha n \rfloor}})}_{k\text{-MST length}} \approx \min_{B:P(B)\geq\alpha} \underbrace{L^*(\{X_1, \ldots, X_n\} \cap B)}_{\text{MST length over } B}$$

Thus

$$L^*_{n,\lfloor \alpha n \rfloor}/(\lfloor \alpha n \rfloor)^\nu \ \to \ \beta_{L,\gamma} \min_{A:P(A)\geq\alpha} \int_{\mathbb{R}^d} f^\nu(x|A)dx \qquad (w.p.1)$$

where

$$f(x|A) = \begin{cases} f(x)/P(A), & x \in A \\ 0, & o.w. \end{cases}$$

Step 2: find explicit form for constrained minimum

Rewrite limiting form as:

$$\begin{aligned} L^*_{n,\lfloor \alpha n \rfloor}/(n)^\nu &\to \beta_{L,\gamma} \min_{A:P(A)\geq\alpha} \int_A f^\nu(x)dx \qquad (w.p.1) \\ &= \beta_{L,\gamma} \min_{A:P(A)\geq\alpha} \rho(A) \end{aligned}$$

Write objective function $\rho(A)$ as (unconstrained) Lagrangian:

$$\begin{aligned} \rho(A) &= \int_A f^\nu(x)dx - \lambda\left(\int_A f(x)dx - \alpha\right) \\ &= \int_A \left(1 - \lambda f^{1-\nu}(x)\right) f^\nu(x)dx + \lambda\alpha \end{aligned}$$

where $\lambda > 0$

Minimizer $A = A_\alpha$ is now obvious

$$A_\alpha = \{x : f(x) \geq \eta\}$$

where $\eta = \lambda^{1/(\nu-1)} \geq 0$ selected s.t.

$$P(A_\alpha) = \alpha$$

Furthermore the minimum can be written

$$\rho(A_\alpha) = \int_{\mathbb{R}^d} f^\nu(x|A_\alpha)dx$$

$\Rightarrow f(x|A_\alpha)$ obtained by "water pouring"

This gives the result:

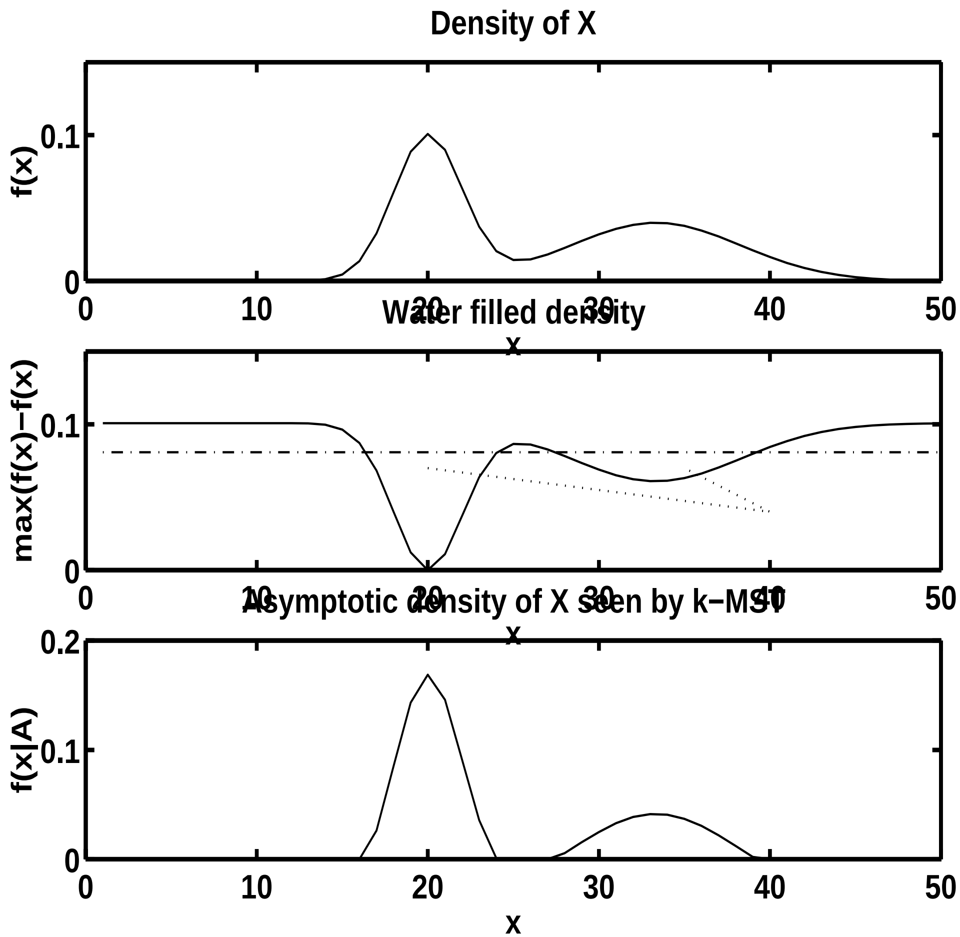
$$L^*_{n,\lfloor \alpha n \rfloor}/(\lfloor \alpha n \rfloor)^\nu \;\to\; \beta_{L,\gamma} \int_{\mathbb{R}^d} f^\nu(x|A_\alpha)dx \qquad (w.p.1)$$

where

$$f(x|A_\alpha) = \begin{cases} \frac{f(x)}{P(A)}, & x \in A_\alpha \\ 0, & o.w. \end{cases}$$

and A_α is a subset of $supp\{f(x)\}$ defined by the pair of conditions

$$A_\alpha = \{x : f(x) \geq \eta\}, \qquad \int_{A_\alpha} f(x)dx = \alpha$$


Density of X
f(x)
0.1
0
0 10 20 30 40 50
x
Water filled density
max(f(x)−f(x))
0.1
0
0 10 20 30 40 50
x
Asymptotic density of X seen by k−MST
f(x|A)
0.2
0.1
0
0 10 20 30 40 50
x

Extended BHH theorem in terms of Renyi entropy

$$\hat{R}_{\nu,\alpha} = \frac{1}{1-\nu}\left(\ln L^*_{n,\lfloor \alpha n \rfloor}/(\lfloor \alpha n \rfloor)^\nu - \ln \beta_{L,\gamma}\right) \rightarrow R_\nu(X|A_\alpha) \qquad (w.p.1)$$

where

$$R_\nu(X|A_\alpha) = \min_{A:P(A)=\alpha} R_\nu(X|A) = \frac{1}{1-\nu}\int_{\mathbb{R}^d} f^\nu(x|A_\alpha)dx,$$

Theorem 1 $\hat{R}_\nu$ *is a strongly consistent estimator of the maximum conditional Rényi entropy* $R_\nu(f|A_\alpha)$ *of order* $\nu \in (0,1)$ *as* $m, n \rightarrow \infty$*.*

Implications

1. k-MST estimator $\hat{R}_\nu$ is unbiased and has vanishing variance.
2. k-MST entropy estimator is robust to outliers. Conditional entropy of mixture $f = (1-\epsilon)f_1 + \epsilon f_0$ equals unconditional entropy of f_1 for small ϵ.
3. $\beta_{L,\gamma}$ need not be computed if only relative entropy is of interest
4. Given maximum tolerated approximation error ϵ, and an upper bound $\overline{v}$ on the total variation of f, we can specify selection rule for required partition resolution
$$1/m \approx \frac{\epsilon}{(2 + C_3)\overline{v}}.$$
5. estimates of Rényi entropy of lower orders ($\nu < 1/d$) converge faster than estimates of higher orders.

Examples

1. $X_i \sim$ uniform on unit sphere $S(0,1)^d$

Find

$$A_\alpha = \left\{ x : \|x\| \leq \left(\frac{\alpha}{|S(0,1)^d|} \right)^{\frac{1}{d}} \right\}$$

and

$$f(x|A_\alpha) = \begin{cases} \frac{1}{\alpha}, & x \in A_\alpha \\ 0, & o.w. \end{cases}$$

Implication: $L_{n,\lfloor \alpha n \rfloor}$ is linear in α

$$L_{n,\alpha n} \; = \; \alpha \cdot \beta(d,\gamma) n^{(d-\gamma)/d}, \qquad (n \text{ large})$$

2. $X_i \sim \mathcal{N}_d(0, \sigma^2 \mathbf{I})$ on $\mathbb{R}^d$

Find

$$A_\alpha = \left\{ x : \|x\| \leq \sigma \sqrt{Q^{-1}_{\chi^2}(\alpha; d)} \right\}$$

and

$$f(x|A_\alpha) = \begin{cases} \frac{1}{\alpha(2\pi\sigma)^{d/2}} e^{-\frac{x^T x}{2\sigma^2}}, & x \in A_\alpha \\ 0, & o.w. \end{cases}$$

so that $L_{n,\lfloor \alpha n \rfloor}$ is non-linear in α

$$L_{n,\lfloor \alpha n \rfloor} \;\sim\; Q_{\chi^2}(\nu Q^{-1}_{\chi^2}(\alpha; d)) \cdot (2\pi\sigma)^{\frac{\gamma}{2d}} \beta(d, \gamma) n^{(d-\gamma)/d}$$

$\nu = (d - \gamma)/d$

5. Investigation of robustness via influence function

Let F_n be the empirical distribution function of the samples $\{x_i\}_{i=1}^n$.

$$F_n(A) \stackrel{\text{def}}{=} \frac{1}{n} \int_A \delta_{x_i}(x)dx$$

For any statistic $T_n = T(F_n)$ converging w.p.1. to $T = T(F)$ the influence curve is (Hampel 1968)

$$IC(x_o, F, T) = \lim_{s \to 0} \frac{T(\,(1-s)F + s\delta_{x_o})}{s}$$

- quantitative measure of outlier sensitivity
- gives asymptotic estimator variance (Huber 1981)

$$n\mathrm{var}(T_n) \;\to\; \int IC^2(x, F, T) f(x) dx$$

IC for k-MST

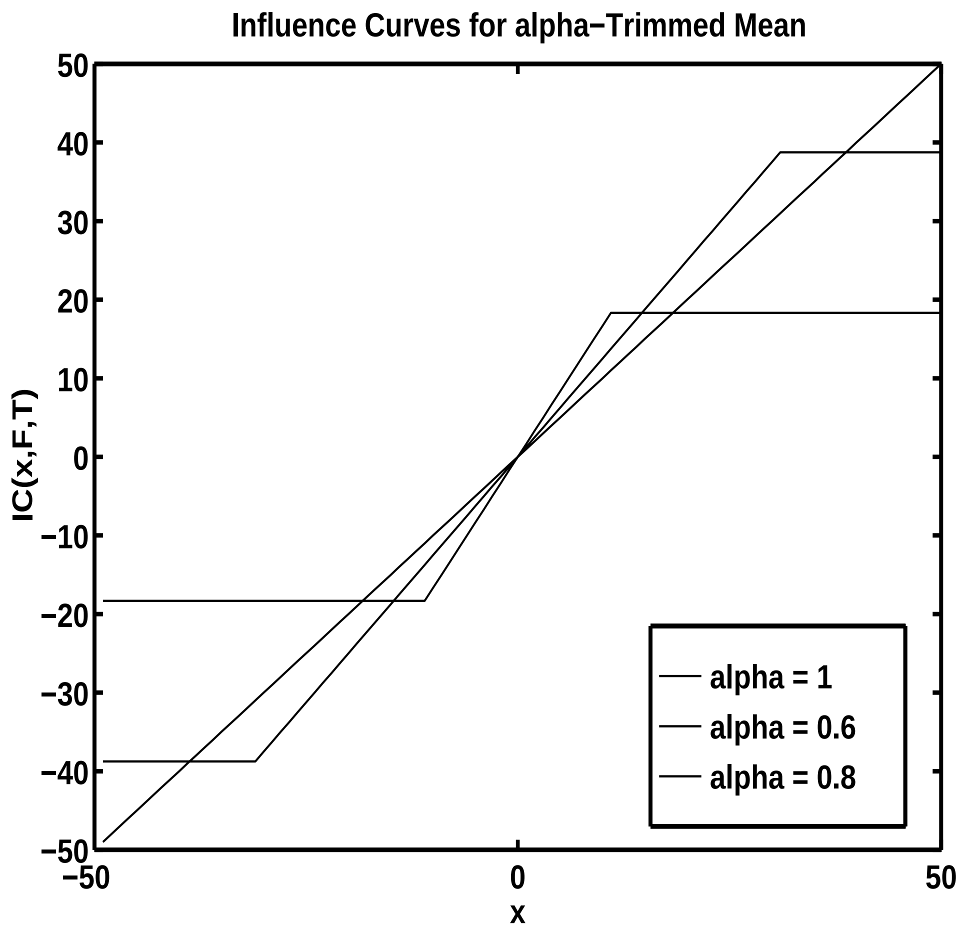
Using $T_n = L^*_{n,\lfloor \alpha n \rfloor}/(\lfloor \alpha n \rfloor)^\nu$ we find

$$IC(x_o, F, L) \;=\; \beta_{L,\gamma} \left(\frac{d}{ds} \int_{\mathbb{R}^d} f_s^\nu(x|A_\alpha) dx \,\Big|_{s=0} \right)$$

where

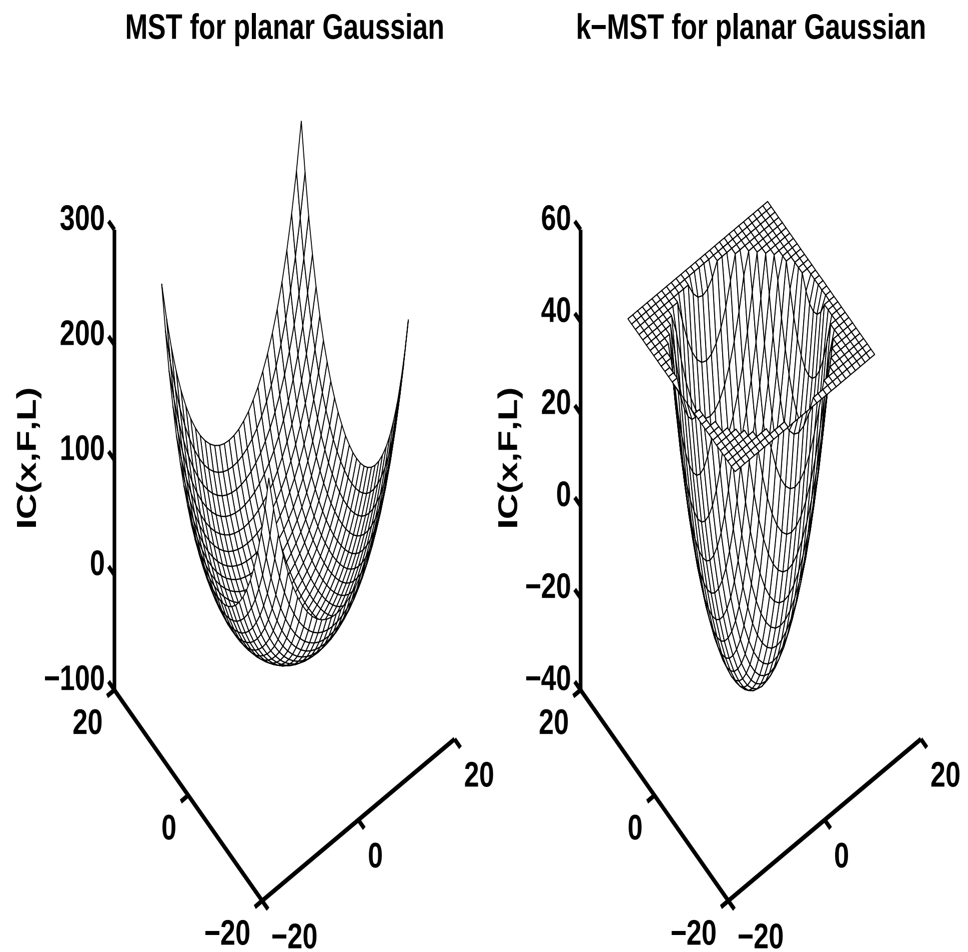
$$f_s(x) = (1-s) f(x) + s\delta_{x_o}(x)$$

$$IC(x_o, F, L) \;=\; \frac{\beta(d,\gamma)}{\alpha^\nu} \cdot \begin{cases} \nu \cdot f^{\nu-1}(x_o|A_\alpha) - (1-\nu) \cdot e^{R_\nu(X|A_\alpha)}, & x_o \in A_\alpha \\ -(1-\nu) \cdot e^{R_\nu(X|A_\alpha)}, & x_o \notin A_\alpha \end{cases}$$


Influence Curves for alpha−Trimmed Mean
IC(x,F,T)
x
50
40
30
20
10
0
−10
−20
−30
−40
−50
−50
0
50
alpha = 1
alpha = 0.6
alpha = 0.8

Observations:

- MST ($\alpha = 1$) has unbounded influence curve as $\|x_o\| \to \infty$ for non-compactly supported f (e.g. exponential family)
- k-MST has bounded influence curve for all f
- IC has similar form to IC for 1-D rank-order statistics


MST for planar Gaussian
k−MST for planar Gaussian
IC(x,F,L)
300
200
100
0
−100
20
0
−20
−20
0
20
IC(x,F,L)
60
40
20
0
−20
−40
20
0
−20
−20
0
20

Conclusions

- $k - MST$ generalizes rank order statistics (trimmed mean) to $\mathbb{R}^d$
- computational complexity appears competitive with density estimation techniques, esp. for large d
- In $k - MST$ length gives consistent entropy estimator with provable robustness
- asymptotics apply to quasiadditive weight functionals: k-TSP, k-Steiner trees, minimal matching
- Yukich's ergodic theory of MST's may provide useful extension for correlated data