Renyi divergence and asymptotic theory of minimal K-point random graphs

Alfred Hero

Dept. of EECS,

The University of Michigan,

Summer 1999

Outline

- Rényi Entropy and Rényi Divergence
- Euclidean k-minimal graphs
- Asymptotics for tightly coverable graphs
- Asymptotics for greedy k-minimal graphs
- Influence function for greedy k-minimal graph
- Applications

1. Rényi Entropy and Rényi Divergence

- $X \sim f(x)$ a d-dimensional random vector.
- Rényi Entropy of order ν

$$H_\nu(f) = \frac{1}{1-\nu} \ln \int f^\nu(x) dx \tag{1}$$

- Rényi Divergence of order ν

$$I_\nu(f, f_o) = \frac{1}{1-\nu} \ln \int \left(\frac{f(x)}{f_o(x)} \right)^\nu f_o(x) dx \tag{2}$$

- f_o a dominating Lebesgue density

Examples:

- Hellinger distance squared

$$I_{\frac{1}{2}}(f, f_o) = \ln \left(\int \sqrt{f(x) f_o(x)} dx \right)^2$$

- Kullback-Liebler divergence

$$\lim_{\nu \to 1} I_\nu(f, f_o) = \int f_o(x) \ln \frac{f_o(x)}{f(x)} dx.$$

2. k-Minimal graphs

A graph G of degree l consists of vertices and edges

- vertices are subset of $\mathcal{X}_n = \{x_i\}_{i=1}^n$: n points in $\mathbf{R}^d$
- edges are denoted $\{e_{ij}\}$
- for any i: $card\{e_{ij}\}_j \leq l$

Weight (with power exponent γ) of G

$$L_{\mathrm{G}}(\mathcal{X}_n) = \sum_{e \in \mathrm{G}} \|e\|^\gamma$$

Examples:

n-point Minimal Spanning Tree (MST)

Let $\mathcal{M}(\mathcal{X}_n)$ denote the possible sets of edges in the class of acyclic graphs spanning $\mathcal{X}_n$ (spanning trees).

The Euclidean Power Weighted MST achieves

$$L_{\text{MST}}(\mathcal{X}_n) = \min_{\mathcal{M}(\mathcal{X}_n)} \sum_{e \in \mathcal{M}(\mathcal{X}_n)} \|e\|^{\gamma}.$$

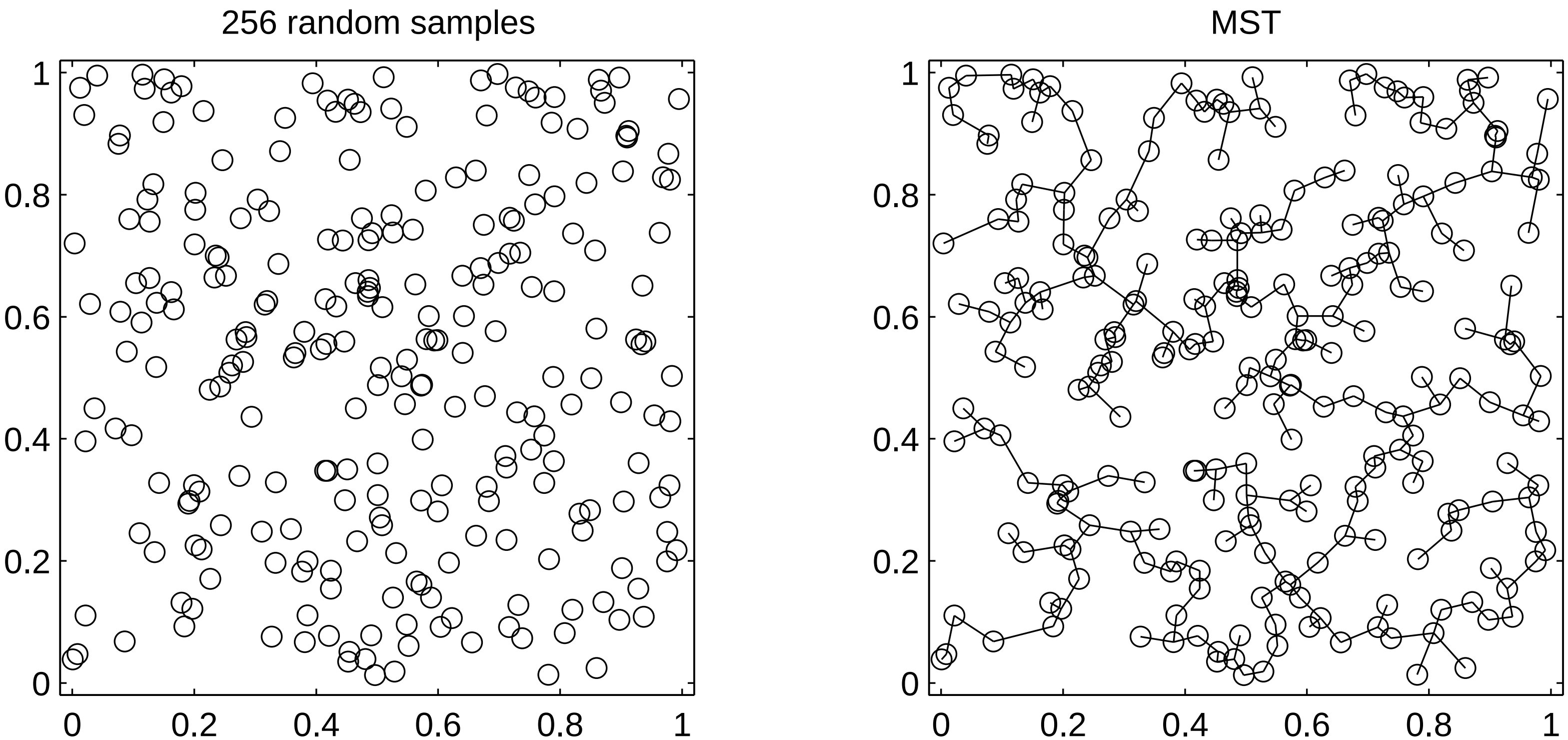

Figure 1. *A data set and the MST*

n-point Traveling Salesman Problem (TSP)

Let $\mathrm{T}(\mathcal{X}_n)$ be sets of edges in the class of graphs of degree 2 spanning $\mathcal{X}_n$.

The minimal power-weighted TSP tour achieves

$$L_{\mathrm{TSP}}(\mathcal{X}_n) = \min_{\mathrm{T}(\mathcal{X}_n)} \sum_{e \in \mathrm{T}(\mathcal{X}_n)} \|e\|^{\gamma}.$$

2.1. Quasi-additive Euclidean Functionals

L is a continuous subadditive functional if it satisfies

Null Condition: $L(\phi) = 0$, where ϕ is the null set.

Subadditivity: There exists a constant C_1 with the following property: For any uniform resolution $1/m$-partition $\mathcal{Q}^m$

$$L(F) \leq m^{-1} \sum_{i=1}^{m^d} L(m[(F \cap Q_i) - q_i]) + C_1 m^{d-\gamma}$$

Superadditivity: There exists a constant C_2 with the following property:

$$L(F) \geq m^{-1} \sum_{i=1}^{m^d} L(m[(F \cap Q_i) - q_i]) - C_2 m^{d-\gamma}$$

Continuity: There exists a constant C_3 such that for all finite subsets F and G of $[0,1]^d$

$$|L(F \cup G) - L(F)| \leq C_3 \left(card(G)\right)^{(d-\gamma)/d}$$

Definition 1 *A continuous subadditive functional L is said to be a quasi-additive functional when there exists a continuous superadditive functional L^* which satisfies $L(F) + 1 \geq L^*(F)$ and the approximation property*

$$|E[L(U_1, \ldots, U_n)] - E[L^*(U_1, \ldots, U_n)]| \leq C_4 n^{(d-\gamma-1)/d} \tag{3}$$

where $U_1, \ldots, U_n$ are i.i.d. uniform random vectors in $[0,1]^d$.

2.2. Asymptotics: the BHH Theorem and entropy estimation

Theorem 1 *[Redmond&Yukich:96] Let L be a quasi-additive Euclidean functional with power-exponent γ, and let $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ be an i.i.d. sample drawn from a distribution on $[0,1]^d$ with an absolutely continuous component having (Lebesgue) density $f(x)$. Then*

$$\lim_{n\to\infty} L(\mathcal{X}_n)/n^{(d-\gamma)/d} \quad = \quad \beta_{L,\gamma} \int f(x)^{(d-\gamma)/d} dx, \qquad (a.s.) \tag{4}$$

Or, letting $\nu = (d-\gamma)/d$

$$\lim_{n\to\infty} L(\mathcal{X}_n)/n^{\nu} \quad = \quad \beta_{L,\gamma} \; \exp\left((1-\nu)H_\nu(f)\right), \qquad (a.s.)$$

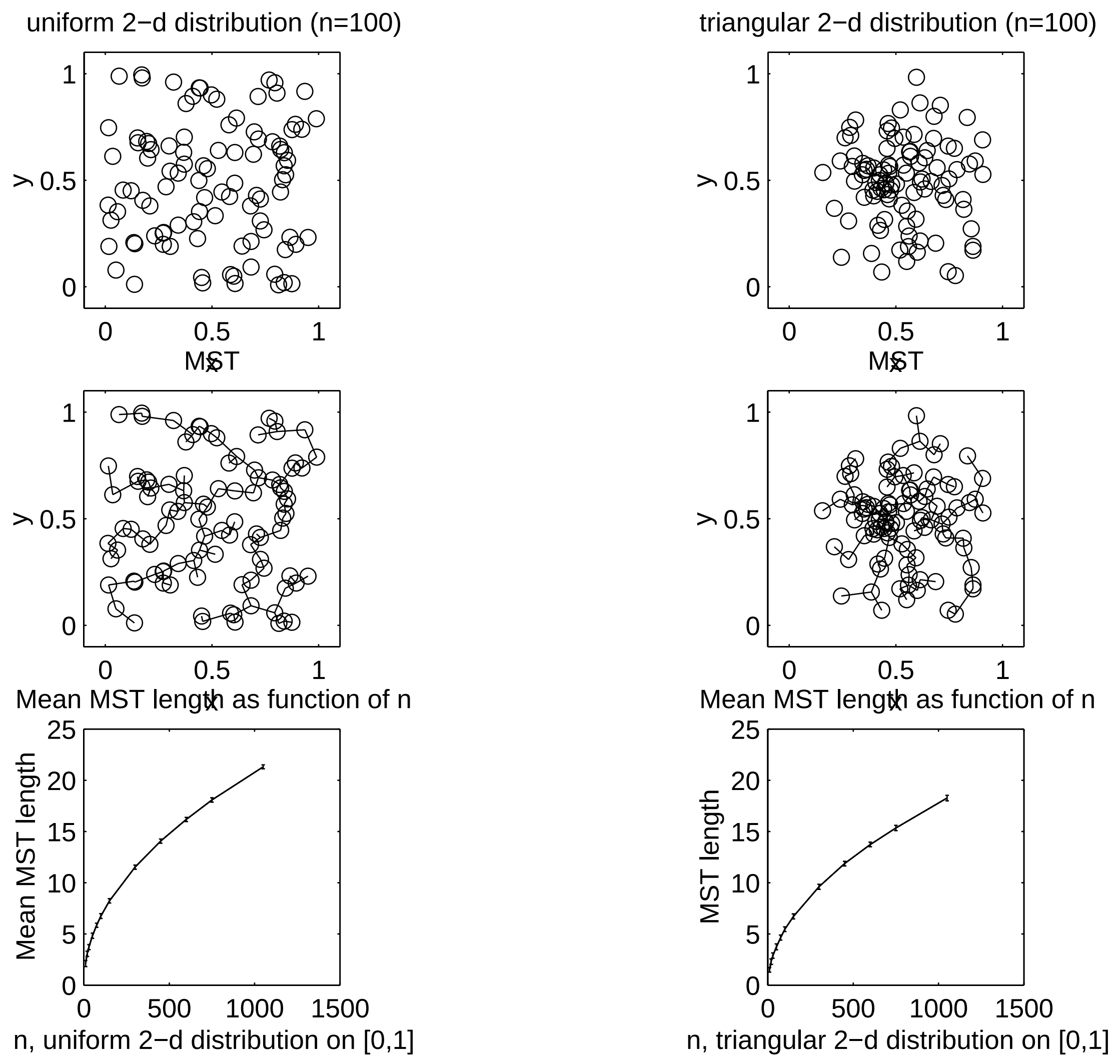

Figure 2. *2D Triangular vs. Uniform sample study for MST.*

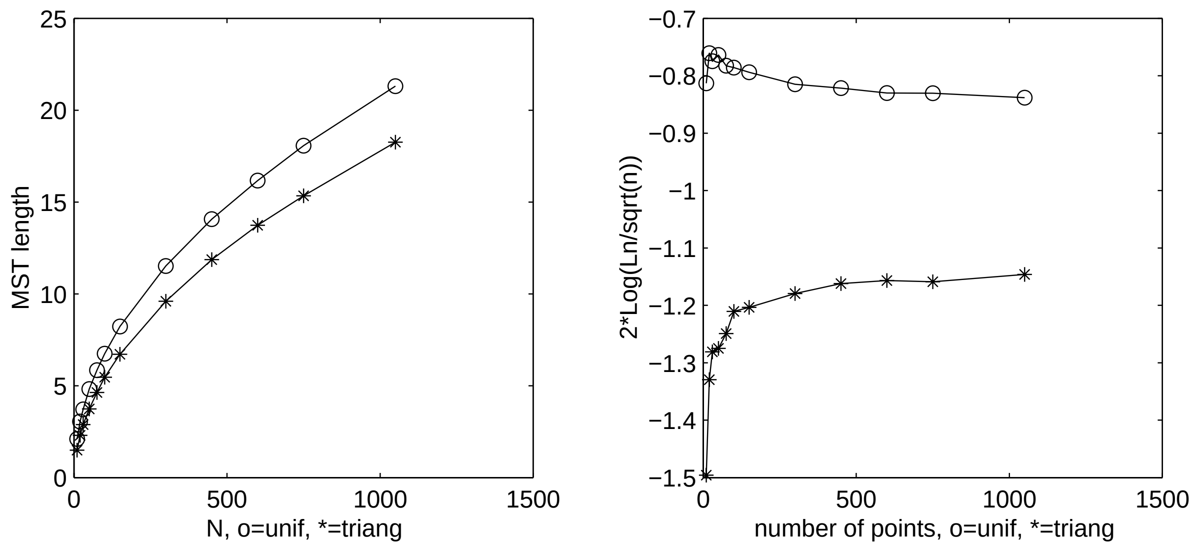

Figure 3. *MST and log MST weights as function of number of samples for 2D uniform vs. triangular study.*

2.3. I-Divergence and Quasi-additive functions

- $g(x)$: a reference density on $\mathbf{R}^d$
- Assume $f \ll g$, i.e. for all x such that $g(x) = 0$ we have $f(x) = 0$.
- Make measure transformation $dx \to g(x)dx$ on $[0,1]^d$. Then for $\mathcal{Y}_n$ = transformed data [Hero&Michel:HOS99]

$$\lim_{n\to\infty} L(\mathcal{Y}_n)/n^\nu \quad = \quad \beta_{L,\gamma}\ \exp\left((1-\nu)I_\nu(f,g)\right), \qquad (a.s.)$$

Proof

1. Make transformation of variables
   $x = [x^1, \ldots, x^d]^T \to y = [y^1, \ldots, y^d]^T$

$$\begin{aligned} y^1 &= G(x^1) \\ y^2 &= G(x^2 | x^1) \\ \vdots & \quad \vdots \\ y^d &= G(x^d | x^{d-1}, \ldots, x^1) \end{aligned} \tag{5}$$

   where $G(x^k | x^{k-1}, \ldots, x^1) = \int_{-\infty}^{x^k} g(\tilde{x}^k | x^{k-1}, \ldots, x^1) d\tilde{x}^k$

2. Induced density $h(y)$, of the vector y, takes the form:

$$h(y) = \frac{f(G^{-1}(y^1), \ldots, G^{-1}(y^d | y^{d-1}, \ldots, y^1))}{g(G^{-1}(y^1), \ldots, G^{-1}(y^d | y^{d-1}, \ldots, y^1))} \tag{6}$$

   where G^{-1} is inverse CDF and $x^k = G^{-1}(y^k | x^{k-1}, \ldots, x^1)$.

3. Then we know

$$\hat{H}_\nu(\mathcal{Y}_n) \to \frac{1}{1-\nu} \ln \int h^\nu(y)dy \quad (a.s.)$$

4. By Jacobian formula: $dy = \left|\frac{dy}{dx}\right| dx = g(x)dx$ and

$$\frac{1}{1-\nu} \ln \int h^\nu(y)dy = \frac{1}{1-\nu} \ln \int \left(\frac{f(x)}{g(x)}\right)^\nu g(x)dx = I(f,g)$$

3. Outlier Sensitivity of minimal n-point graphs

Assume f is a mixture density of the form

$$f = (1 - \epsilon) f_1 + \epsilon f_o, \tag{7}$$

where

- f_o is a known outlier density
- f_1 is an unknown target density
- $\epsilon \in [0, 1]$ is unknown mixture parameter

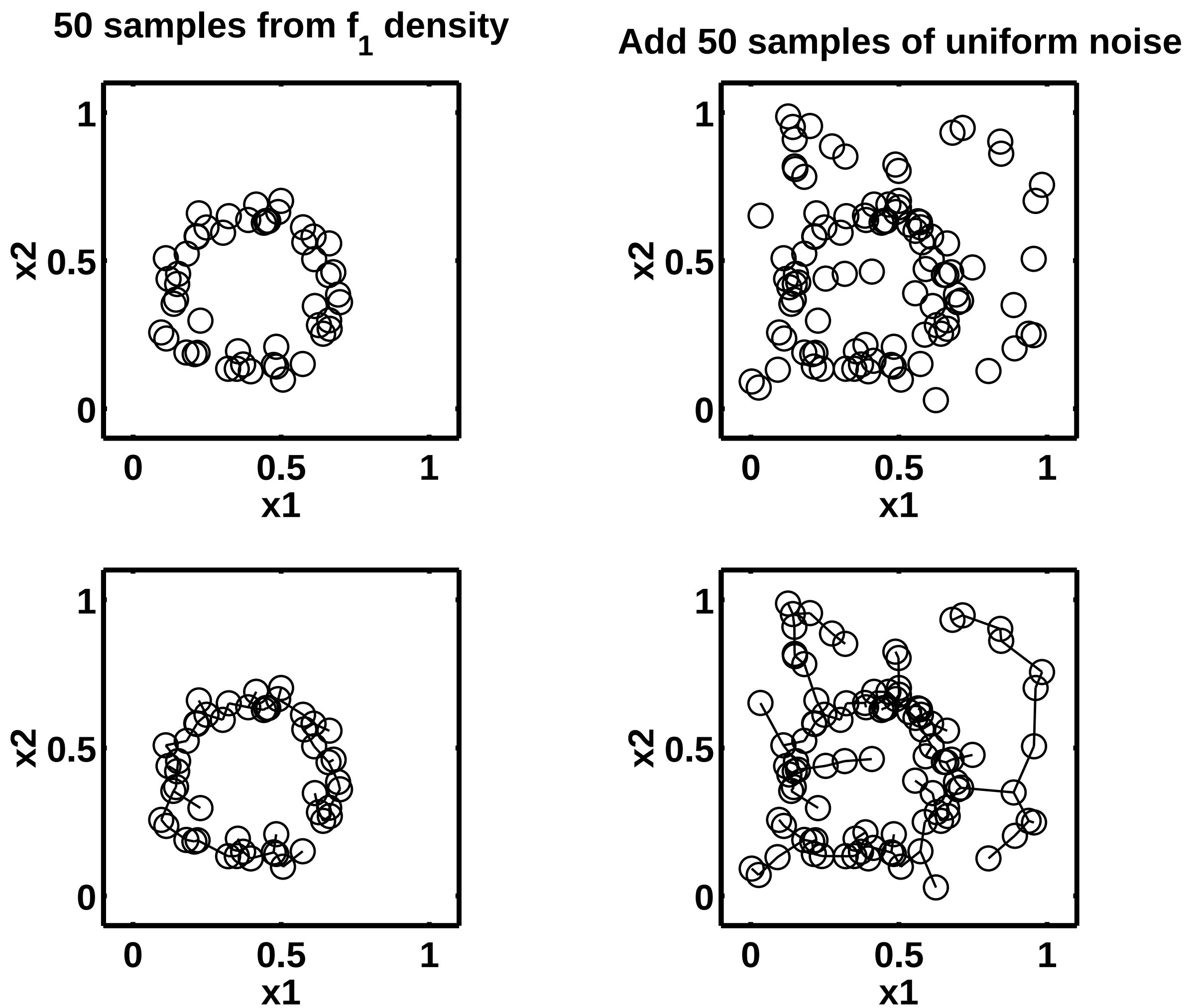

Figure 4. *1st row: 2D torus density with and without the addition of uniform "outliers." 2nd row: corresponding MST's.*

3.1. Minimal k-point Euclidean Graphs

Fix k, $1 \leq k \leq n$.

Let $\mathrm{T}_{n,k} = \mathrm{T}(x_{i_1}, \ldots, x_{i_k})$ be a minimal graph connecting k distinct vertices $x_{i_1}, \ldots, x_{i_k}$.

The power weighted *k-minimal* graph $\mathrm{T}^*_{n,k} = \mathrm{T}^*(x_{i^*_1}, \ldots, x_{i^*_k})$ is the overall minimum weight k-point graph

$$L^*_{n,k} \quad = \quad L^*(\mathcal{X}_{n,k}) = \min_{i_1,\ldots,i_k} \min_{\mathrm{T}_{n,k}} \sum_{e \in \mathrm{T}_{n,k}} \|e\|^\gamma$$

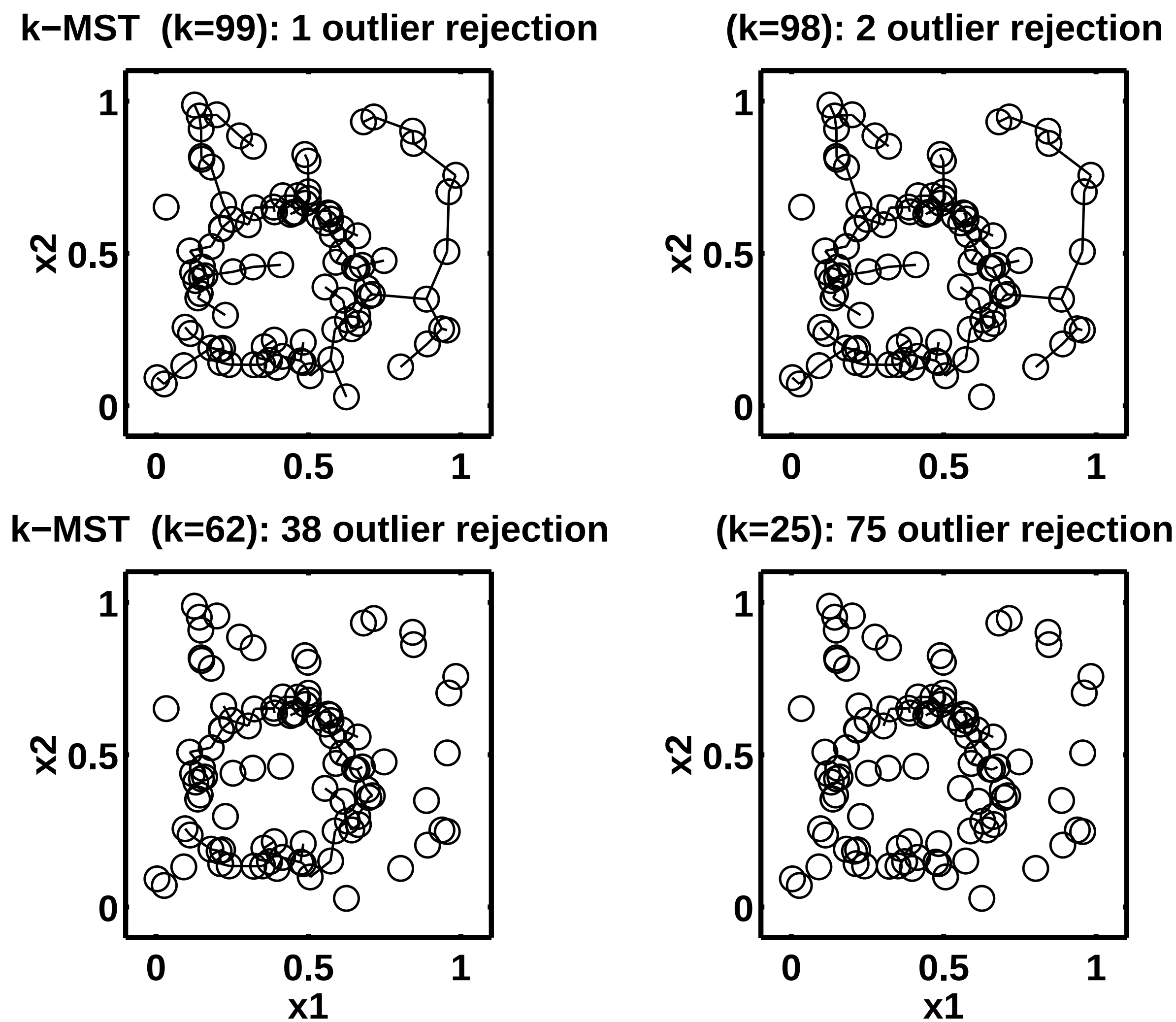

Figure 5. *k-MST for 2D torus density with and without the addition of uniform "outliers".*

4. Extended BHH Thm for k-Minimal Graphs

Fix $\alpha \in [0,1]$ and assume that the k-minimal graph is *tightly coverable*. If $k = \lfloor \alpha n \rfloor$, as $n \to \infty$ we have (Hero&Michel:IT99)

$$L(\mathcal{X}^*_{n,k})/(\lfloor \alpha n \rfloor)^\nu \to \beta_{L,\gamma} \min_{A:P(A)\geq\alpha} \int f^\nu(x|x \in A)dx \qquad (a.s.)$$

or, alternatively, with

$$H_\nu(f|x \in A) = \frac{1}{1-\nu} \ln \int f^\nu(x|x \in A)dx$$

$$L(\mathcal{X}^*_{n,k})/(\lfloor \alpha n \rfloor)^\nu \to \beta_{L,\gamma} \exp\left((1-\nu) \min_{A:P(A)\geq\alpha} H_\nu(f|x \in A)\right) \qquad (a.s.)$$

Definition 2 (Tightly Coverable Graphs) *Let $\mathcal{Q}^m$, $m = 1, 2, \ldots$, be a sequence of uniform partitions of $[0,1]^d$ of resolution $1/m$. Let G be an algorithm which constructs a graph with with $k = \lfloor \alpha n \rfloor$ vertices $\mathcal{U}_{n,k} \subset \mathcal{U}_n$, an i.i.d. uniform sample over $[0,1]^d$. Define $D_k^m = \cap_{\{C \in \sigma(\mathcal{Q}^m) : \mathcal{U}_{n,k} \in C\}} C$ the minimum volume set in $\sigma(\mathcal{A}^m)$ which covers $\mathcal{U}_{n,k}$. The algorithm G is said to generate tightly coverable subgraphs if for any $\epsilon > 0$ there exists an M such that for all $m > M$*

$$\limsup_{n \to \infty} \left| \frac{card(\mathcal{U}_n \cap D^m_{\lfloor \alpha n \rfloor}) - \lfloor \alpha n \rfloor}{n} \right| \leq \epsilon, \qquad (a.s.)$$

5. Greedy Partition Algorithm

Greedy approximation to k-minimal graph (Ravi&*etal*:94)

0) specify a uniform partition $\mathcal{Q}^m$ *of* $[0,1]^d$ *having* m^d *cells* Q_i *of resolution* $1/m$*;*

1) find the smallest subset $B_k^m = \cup_i Q_i$ *of partition elements containing at least* k *points*

2) out of $\mathcal{X}_n \cap B_k^m$ *select* k *points* $\mathcal{X}_{n,k}$ *which minimize* $L(\mathcal{X}_{n,k})$.

Properties:

- Greedy algorithm is polynomial time unlike exponential time exact k-minimal algorithm.
- Greedy algorithm yields tightly coverable graphs by construction

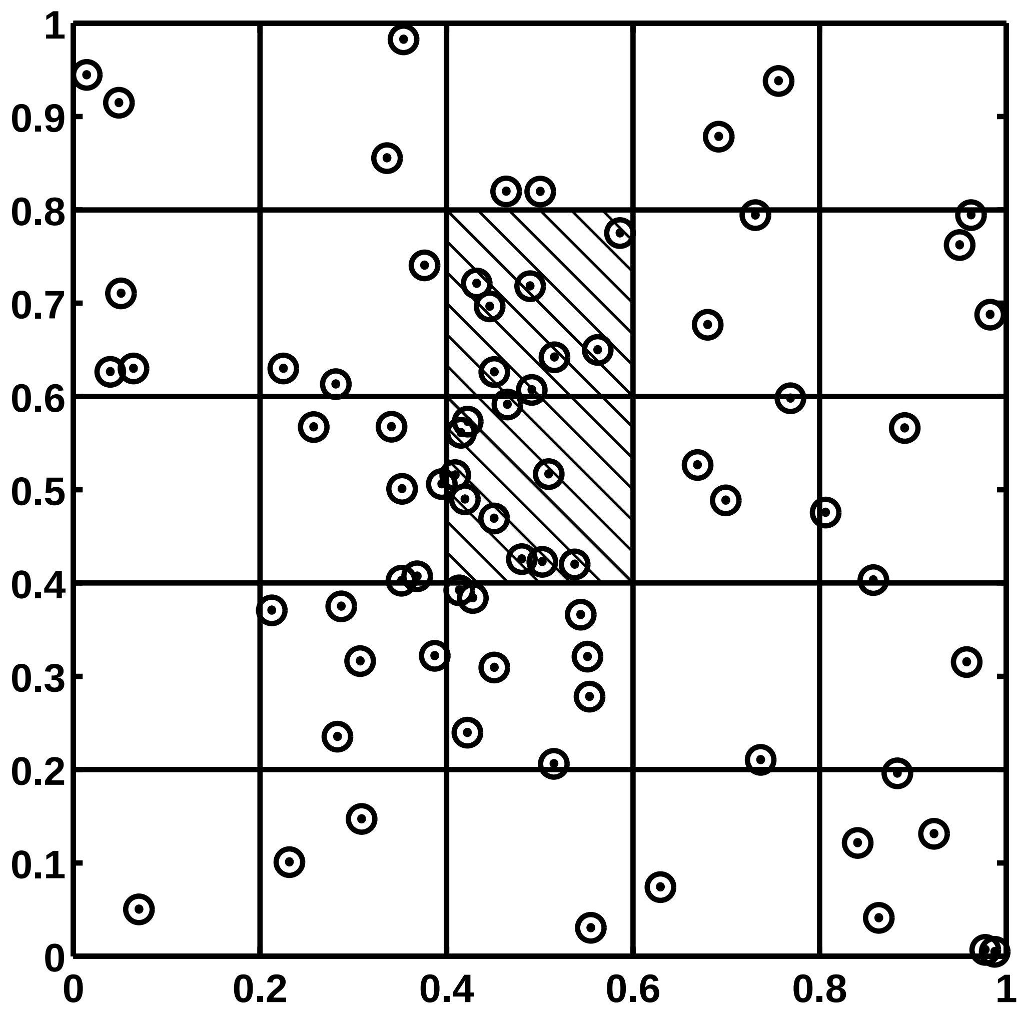

Figure 6. *A sample of* 75 *points from the mixture density* $f(x) = 0.25 f_1(x) + 0.75 f_o(x)$ *where* f_o *is a uniform density over* $[0, 1]^2$ *and* f_1 *is a bivariate Gaussian density with mean* $(1/2, 1/2)$ *and diagonal covariance* $\mathrm{diag}(0.01)$. *A smallest subset* B_k^m *is the union of the two cross hatched cells shown for the case of* $m = 5$ *and* $k = 17$.

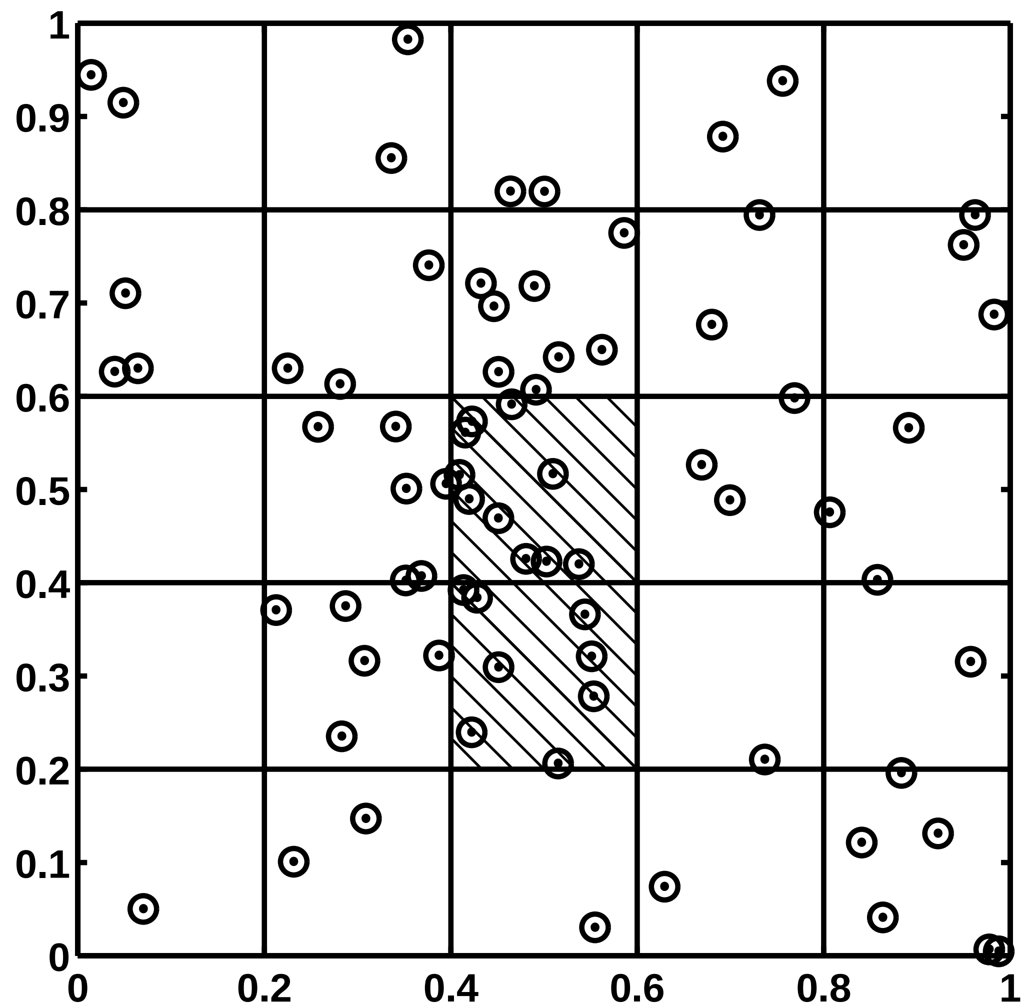

Figure 7. *Another smallest subset* B_k^m *containing at least* $k = 17$ *points for the mixture sample shown in Fig 6.*

Minimal $1/m$-Cover of Probability at least α

If for any $C \in \sigma(\mathcal{Q}^m)$ satisfying $P(C) \geq \alpha$ the set $A \in \sigma(\mathcal{Q}^m)$ satisfies

$$P(C) \geq P(A) \geq \alpha,$$

then A is called a *minimal resolution*-$1/m$ *set of probability at least* α.

The class of all such sets is denoted $\mathcal{A}_\alpha^m$.

$\Rightarrow$ all sets in $\mathcal{A}_\alpha^m$ have identical coverage probabilities $p_{\mathcal{A}_\alpha^m} \geq \alpha$.

Theorem 2 *Let $\mathcal{X}_n$ be an i.i.d. sample from a distribution having Lebesgue density $f(x)$. Fix $\alpha \in [0,1]$, $\gamma \in (0,d)$. Let $f^{(d-\gamma)/d}$ be of bounded variation over $[0,1]^d$ and denote by $v(A)$ its total variation over a subset $A \subset [0,1]^d$. Let L be a quasi-additive functional with power exponent γ as in Theorem 1. Then,*

$$\limsup_{n\to\infty} \left| L(\mathcal{X}_{n,\lfloor \alpha n \rfloor}^{G_m})/n^{\nu} - \beta_{L,\gamma} \min_{A:P(A)\geq\alpha} \int f^{\nu}(x|A)dx \right| \quad < \delta, \qquad (a.s.),$$

where

$$\begin{aligned} \delta &= 2m^{-d}\beta_{L,\gamma}\sum_{i=1}^{m^d} v(Q_i \cap \partial\mathcal{A}_{\alpha}^m) + C_3(p_{\mathcal{A}_{\alpha}^m} - \alpha)^{(d-\gamma)/d} \\ &= O(m^{\gamma-d}), \end{aligned}$$

and $p_{\mathcal{A}_{\alpha}^m}$ is the coverage probability of minimizing set $A = \mathcal{A}_{\alpha}^m$.

Main idea behind proof: for large n can index over $\mathcal{X}_{n,k} \subset \mathcal{X}_n$ by indexing over $A \subset$ Borel in $[0,1]^d$.

$$E[\min_{\mathcal{X}_{n,\lfloor \alpha n \rfloor}} L(\mathcal{X}_{n,\lfloor \alpha n \rfloor})] \approx \inf_{A: P(A) \geq \alpha} E[L(\mathcal{X}_n \cap A)],$$

Proof of Theorem uses following lemmas:

Lemma 1 *For given* $\alpha \in [0,1]$ *and a set of* n *i.i.d. points* $\mathcal{X}_n = [x_1, \ldots, x_n]^T$ *let* B_n^m *be the minimal cover of* $\lfloor \alpha n \rfloor$ *points with resolution-*$1/m$ *produced by the greedy subset selection algorithm. Then*

$$P\left(\liminf_{n\to\infty} \{\mathcal{X}_n : B_n^m \in \mathcal{A}_\alpha^m\}\right) = 1.$$

Lemma 2 *For* $\nu \in [0,1]$ *let* f^ν *be of bounded variation over* $[0,1]^d$ *and denote by* $v(A)$ *its total variation over any subset* $A \in [0,1]^d$*. Define the resolution* $1/m$ *block density approximation* $\tilde{f}(x) = \sum_{i=1}^{m^d} \theta_i I_{Q_i}(x)$ *where* $\theta_i = m^d \int_{Q_i} f(x)dx$*. Then for any* $A \in \sigma(\mathcal{Q}^m)$

$$0 \leq \int_A [\tilde{f}^\nu(x) - f^\nu(x)]dx \leq m^{-d} \sum_{i=1}^{m^d} v(Q_i \cap A).$$

Lemma 3 *Assume f is of bounded total variation $v(Q_i)$ in each partition cell $Q_i \in \mathcal{Q}^m$. Let A be any set in the class $\mathcal{A}_\alpha^m$. Then for any quasi-additive functional $L_n(B^m_{\lfloor \alpha n \rfloor}) \stackrel{\text{def}}{=} L(\mathcal{X}_n \cap B^m_{\lfloor \alpha n \rfloor})$*

$$\limsup_{n\to\infty} \left| L_n(B^m_{\lfloor \alpha n \rfloor})/n^{(d-\gamma)/d} - \beta_{L,\gamma} \int_A f^{(d-\gamma)/d}(x)dx \right|$$

$$< 2m^{-d}\beta_{L,\gamma} \sum_{i=1}^{m^d} v(Q_i \cap \partial \mathcal{A}_\alpha^m), \quad (a.s).$$

Lemma 4 *Let* $\mathcal{X}_{n,\lfloor\alpha n\rfloor}$ *be any* $\lfloor\alpha n\rfloor$ *points selected from* $B^m_{\lfloor\alpha n\rfloor}$*. Then, for any quasi-additive functional* $L_n(B^m_{\lfloor\alpha n\rfloor}) \stackrel{\text{def}}{=} L_n(\mathcal{X}_n \cap B^m_{\lfloor\alpha n\rfloor})$

$$\limsup_{n\to\infty} \left| L_n(B^m_{\lfloor\alpha n\rfloor}) - L(\mathcal{X}_{n,\lfloor\alpha n\rfloor}) \right| / n^{(d-\gamma)/d}$$
$$< C_3(p_{\mathcal{A}^m_\alpha} - \alpha)^{(d-\gamma)/d}, \quad (a.s.)$$

where $p_{\mathcal{A}^m_\alpha} = P(A^m_\alpha)$ *is the coverage probability of sets* A^m_α *in* $\mathcal{A}^m_\alpha$*.*

Interpretations of Theorem 2:

- Bound δ is tight: $\lim_{\alpha \to 1} \delta = 0$ and theorem reduces to BHH.
- Since $\sup_{x \in Q_{(q)}} f^{(d-\gamma)/d}(x) \leq v([0,1]^d)$ and $\sum_{i=1}^{m^d} v(Q_i \cap \partial \mathcal{A}_\alpha^m) \leq v([0,1]^d)$, δ can be upper bounded by

$$\delta \leq \left[2\beta_{L,\gamma} m^{-d} + C_3 m^{\gamma-d}\right] v([0,1]^d).$$

  Thus if an upper bound $\overline{v}$ on the total variation of f is available and the tolerance ϵ is given

$$|L(\mathcal{X}_{n,\lfloor \alpha n \rfloor}^{G_m})/(\lfloor \alpha n \rfloor)^\nu - \beta_{L,\gamma} \exp\{-(1-\nu)R_\nu)| < \epsilon$$

  we obtain a selection rule for required partition resolution $1/m$

$$1/m \leq \frac{\epsilon}{(2+C_3)\overline{v}}.$$

- δ decreases to zero as a function of resolution $1/m$ at overall rate

$O(m^{\gamma-d})$. Thus convergence rate in $1/m$ is fastest for small γ.

- Conjecture: as $|E[L_{MST}(\mathcal{U}_n)] - \beta_{L_{MST},\gamma} n^{(d-\gamma)/d}| = O\left(\max(1, n^{(d-\gamma-1)/d})\right)$ [Redmond&Yukich:96], rate of convergence in the limsup of Theorem 2 is at best $O(1/n^{1/d})$ and this rate can be attained only when $\gamma \leq d-1$.
- k-minimal graph entropy estimator will have fastest convergence when the Rényi order parameter ν is in the range $1/d \leq \nu < 1$.
- Theorem extends easily to I-divergence limit by measure transformation.

6. Application 1: MST Discrimination

- $f(x) = (1 - \epsilon) f_1(x) + \epsilon f_0(x)$: mixture density
- $f_1(x)$ is 2D separable triangle density on $[0, 1]^2$
- $f_0(x)$ is 2D uniform density on $[0, 1]^2$

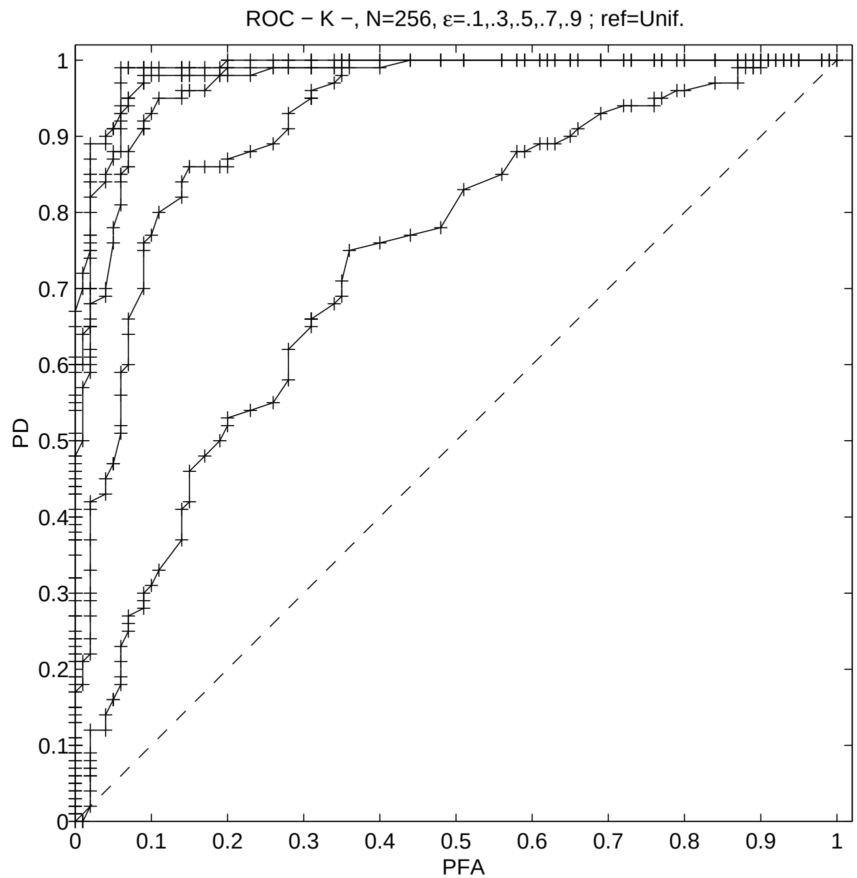

Figure 8. *ROC curves for the Rényi information divergence test for detecting triangle-uniform mixture density* $f = (1-\epsilon) f_1 + \epsilon f_0$ *(H_1) against triangular hypothesis* $f = f_1$ *(H_0). Curves are increasing in* $\epsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

7. Application II: Nonuniform Outlier Rejection

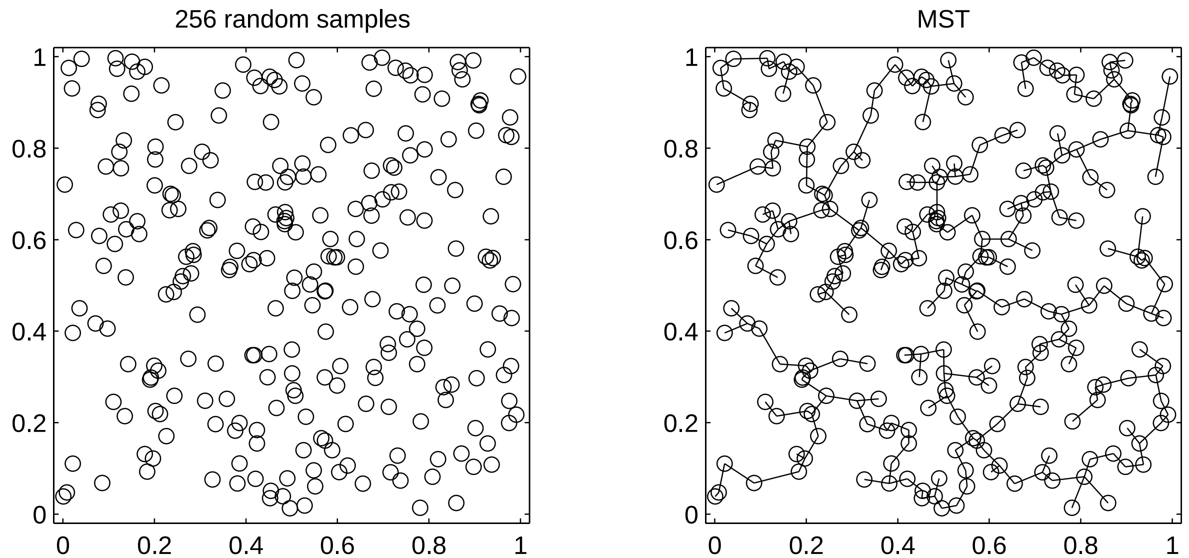

Figure 9. *Left: A scatterplot of a 256 point sample from triangle-uniform mixture density with* $\epsilon = 0.1$. *Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Right: superimposed is the* k*-MST implemented directly on the scatterplot* $\mathcal{X}_n$ *with* $k = 230$.

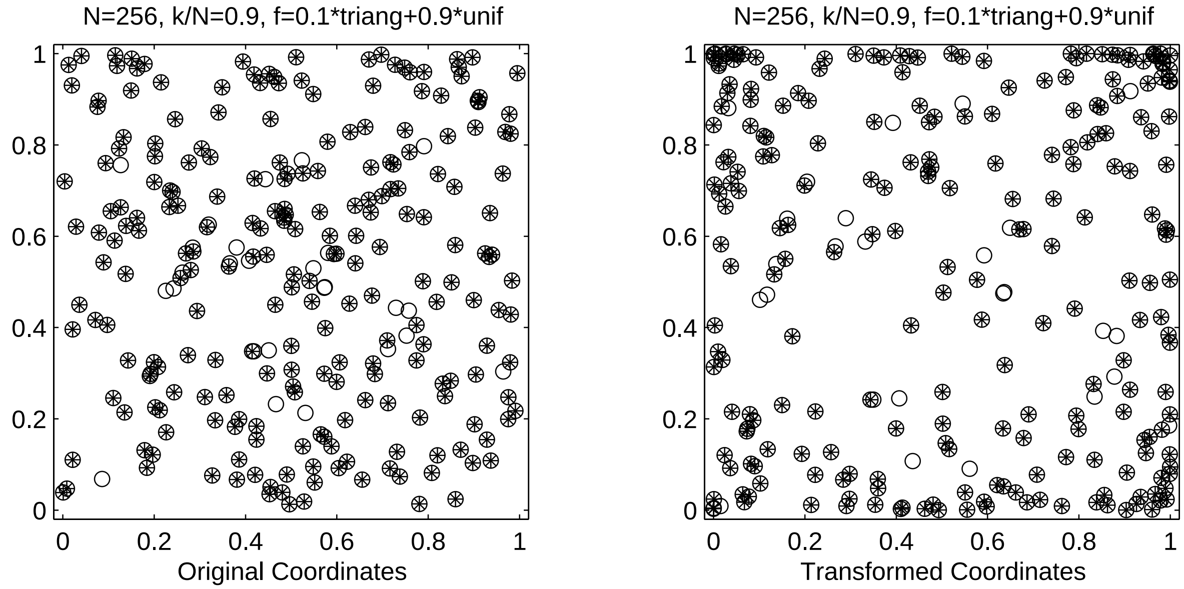

Figure 10. *Left: A sample from triangle-uniform mixture density with* $\epsilon = 0.9$ *in the transformed domain* $\mathcal{Y}_n$*. Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Right: transformed coordinates.*

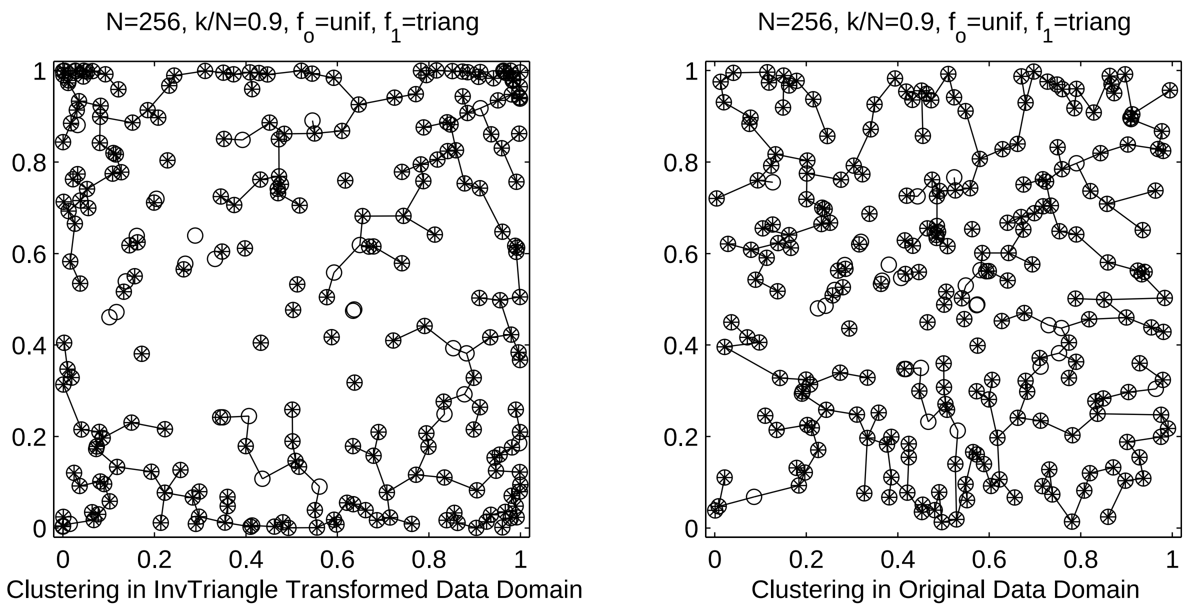

Figure 11. *Left: the k-MST implemented on the transformed scatterplot $\mathcal{Y}_n$ with $k = 230$. Right: same k-MST displayed in the original data domain.*