# Rényi Information Divergence via Measure Transformations on Minimal Spanning Trees


**Alfred Hero**

**Dept. of EECS,**

**The University of Michigan,**

**Olivier J.J. Michel**

**École Normale Supérieure de Lyon,**


**June 2000**

## Outline

- Rényi Entropy and Rényi Divergence
- Euclidean minimal graphs
- Asymptotic Theorem
- Application

## 1. Rényi Entropy and Rényi Divergence

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

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

- Rényi Divergence of order $\nu$

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

- $f_o$ 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

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

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

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

Difficulties

- Histogram estimate of cts. entropy requires discretization correction factor
- kernel or histogram estimation is unstable esp. for large $d$
- $d$-dimensional integration in $H_\nu$ can be impractical
- convergence is slow esp. in high $d$ and asymptotic analysis is complicated
- unclear how to robustify $\hat{f}$ against outliers
- $\Rightarrow$ function $\{f(x) : x \in \mathbf{R}^d\}$ over-parameterizes entropy functional

## 2. Minimal Euclidean 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 $\gamma$) of G

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

Example:

**$n$-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

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

Other examples: TSP, Steiner Tree, K-means

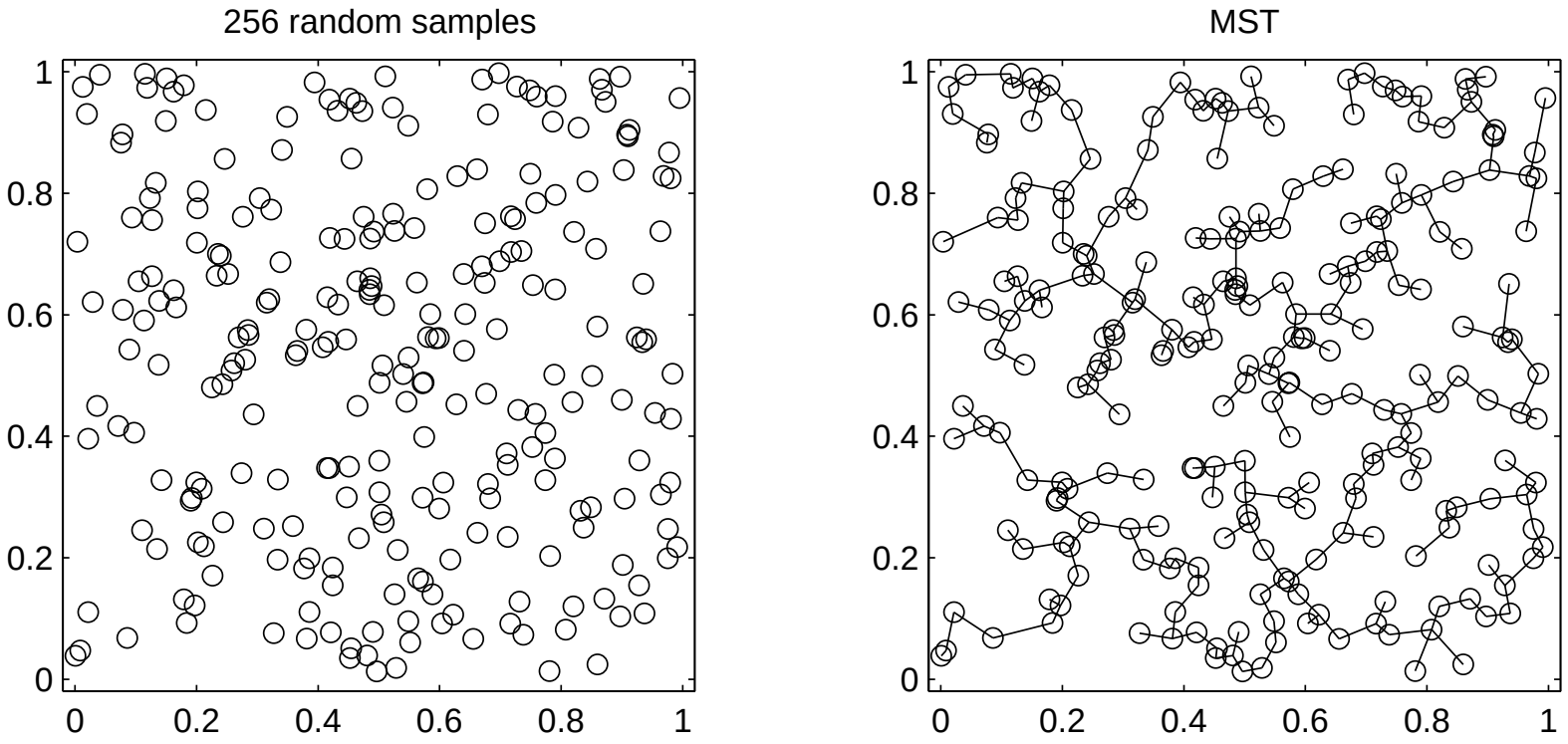


**Figure 1.** *A data set and the MST*

## 2.1. Asymptotics: the BHH Theorem and entropy estimation

**Theorem 1** *[Redmond&Yukich:96] Let $L$ be a quasi-additive Euclidean functional with power-exponent $\gamma$, 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*

(3)

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

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

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

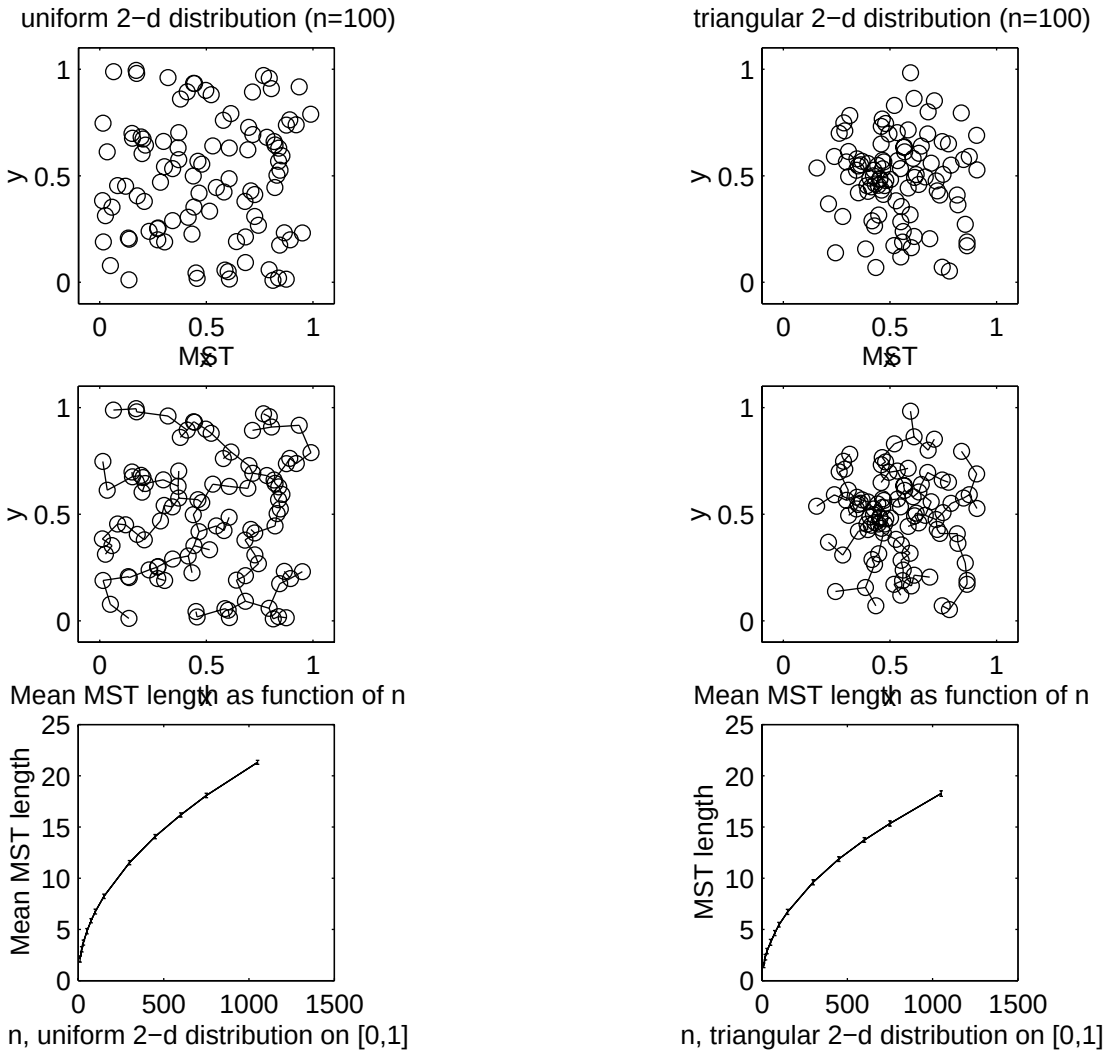


**Figure 2.** *2D Triangular vs. Uniform sample study for MST.*

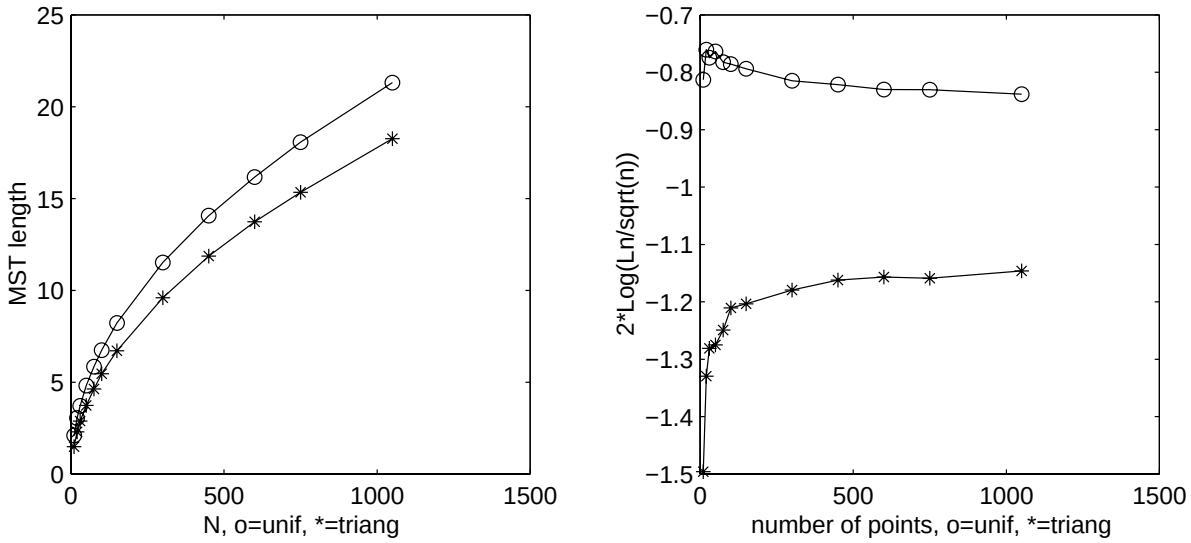


**Figure 3.** *MST and log MST weights as function of number of samples for 2D uniform vs. triangular study.*

## 2.2. 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

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

*Proof*

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

$$\begin{aligned} y^1 &= G(x^1) \\ y^2 &= G(x^2|x^1) \\ \vdots & \quad \vdots \\ y^d &= G(x^d|x^{d-1}, \ldots, x^1) \end{aligned} \tag{4}$$

   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:

$$h(y) = \frac{f(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))}{g(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))} \tag{5}$$

   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

$$f = (1 - \epsilon) f_1 + \epsilon f_o, \tag{6}$$

where

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

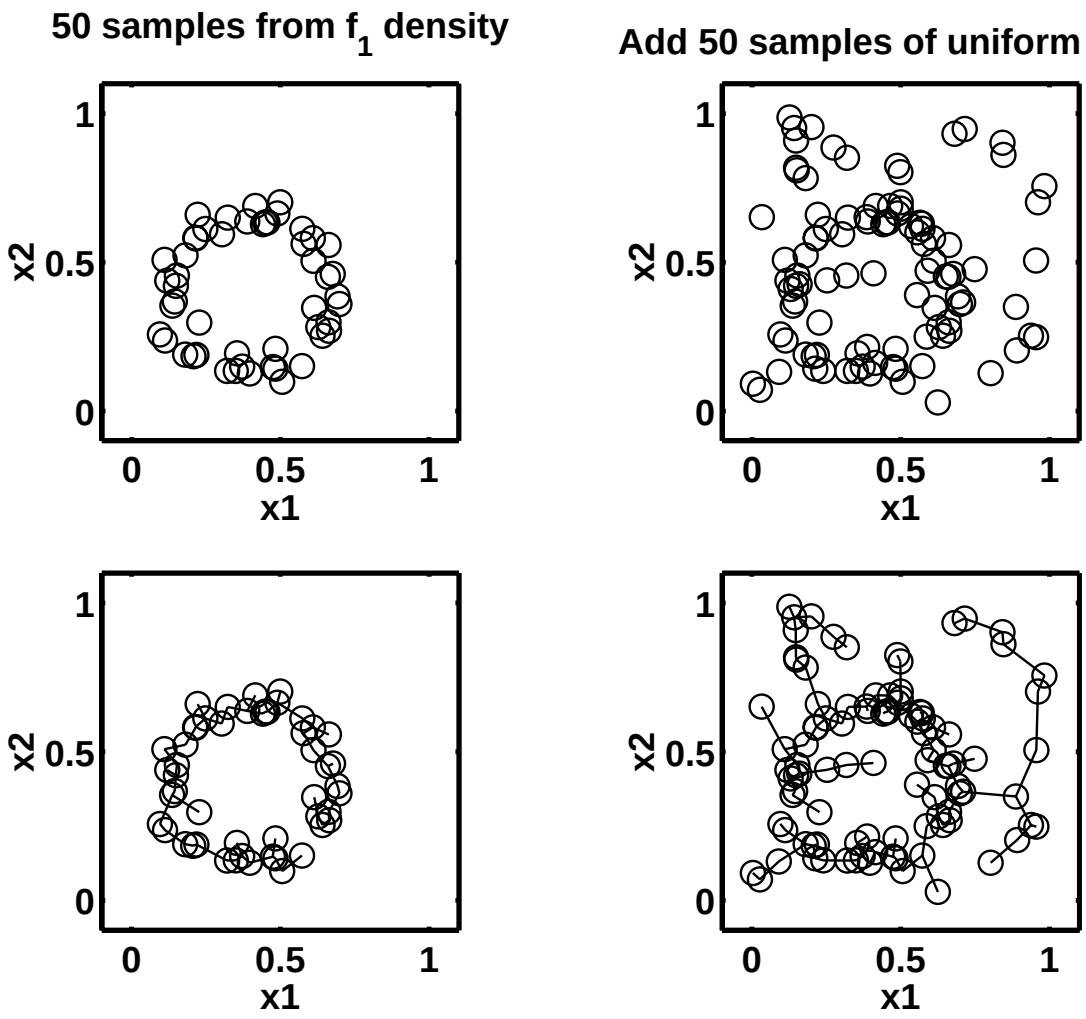


**Figure 4.** *1st row: 2D torus density with and without the addition of uniform "outliers." 2nd row: corresponding MST's.*

## 3.1. $k$-Minimal 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

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

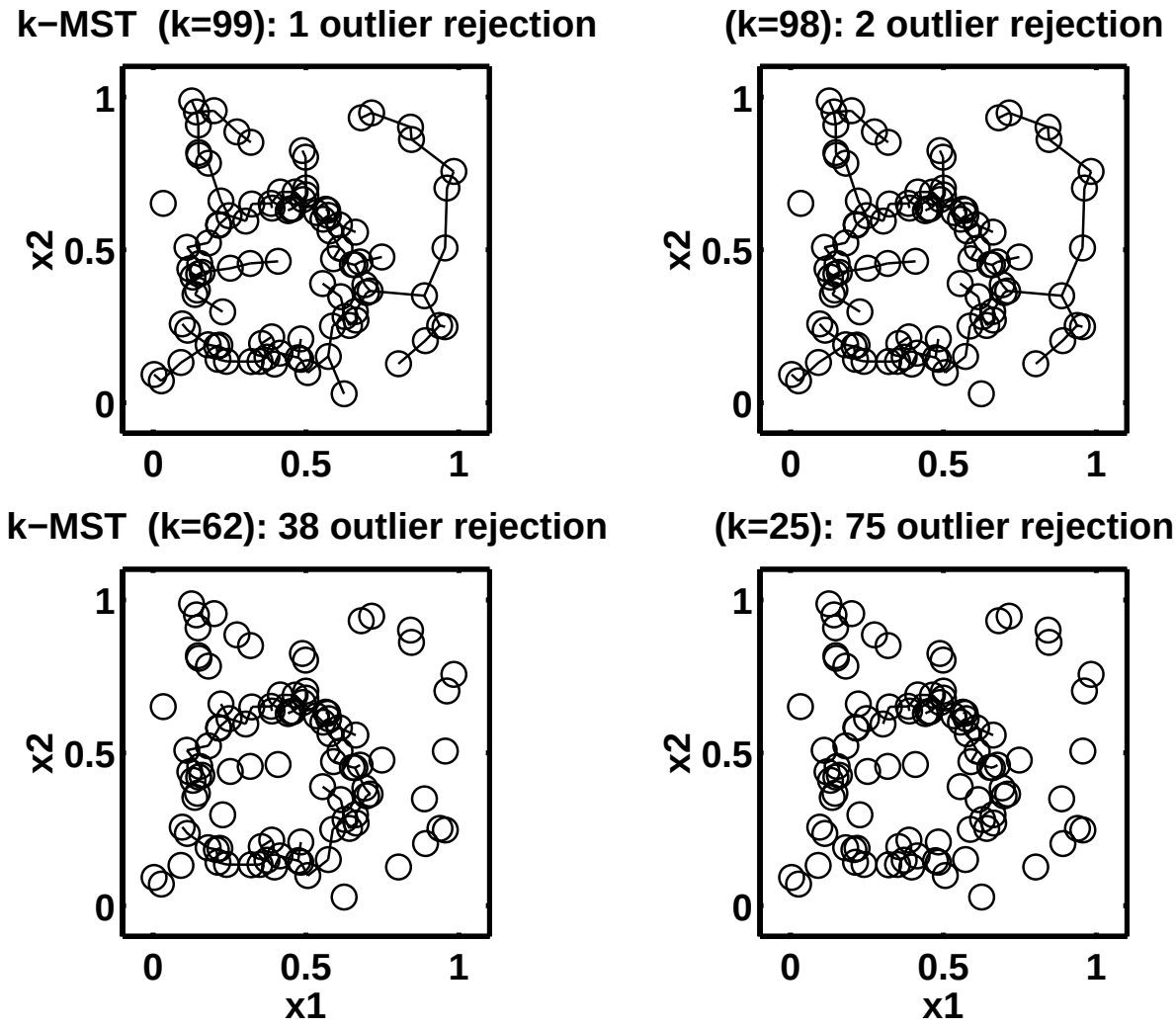


**Figure 5.** *k-MST for 2D torus density with and without the addition of uniform "outliers".*

## 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

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

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

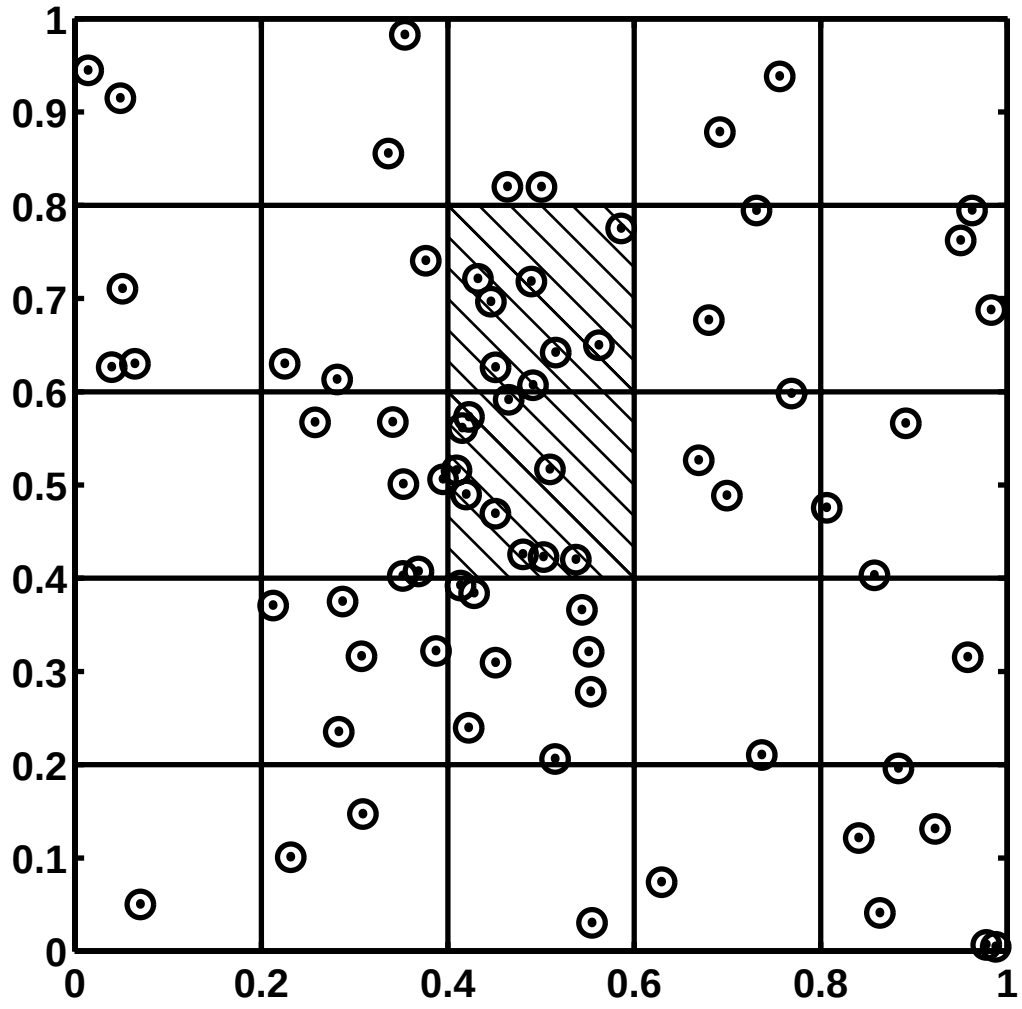


**Figure 6.** *A sample of* $75$ *points from the mixture density* $f(x) = 0.25 f_1(x) + 0.75 f_o(x)$ *where* $f_o$ *is a uniform density over* $[0, 1]^2$ *and* $f_1$ *is a bivariate Gaussian density with mean* $(1/2, 1/2)$ *and diagonal covariance* $\mathrm{diag}(0.01)$. *A smallest subset* $B_k^m$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

## 4.1. Application: Robust Density Estimation/Classification

Estimation Problem: Estimate $f_1(x)$ given sample from mixture

$$f(x) = (1-\epsilon)f_1(x) + \epsilon f_0(x)$$

- $f_0(x)$= known contaminating density

Classification problem: decide between

$$\begin{aligned} H_0 &: f(x) = f_0(x) \\ H_1 &: f(x) = (1-\epsilon)f_1(x) + \epsilon f_0(x), \quad \epsilon \in [0,1] \end{aligned}$$

Step 1: induce change of measure $dy = f_0(x)dx$ by transformation

$$\begin{aligned} y^1 &= F_0(x^1) \\ y^2 &= F_0(x^2|x^1) \\ &\vdots \quad\quad \vdots \\ y^d &= F_0(x^d|x^{d-1}, \ldots, x^1) \end{aligned} \tag{7}$$

Step 2: build $k$-MST on transformed variables $\{Y_i\}_{i=1}^n$

$$L^*_{n,\lfloor \alpha n \rfloor}(Y)/(\lfloor \alpha n \rfloor)^\nu \;\rightarrow\; \beta_{L,d} \min_{A:P(A)\geq\alpha} \int_A \left(\frac{f(x)}{f_0(x)}\right)^\nu f_0(x)dx$$

Robust Density Estimator: kernel estimator applied to $X_{i_1}, \ldots, X_{i_{\lfloor \alpha n \rfloor}}$

Classification rule: $L^*_n(Y)/n^\nu \underset{H_0}{\overset{H_1}{\gtrless}} \eta$

## 4.2. Application: Nonuniform Outlier Rejection

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

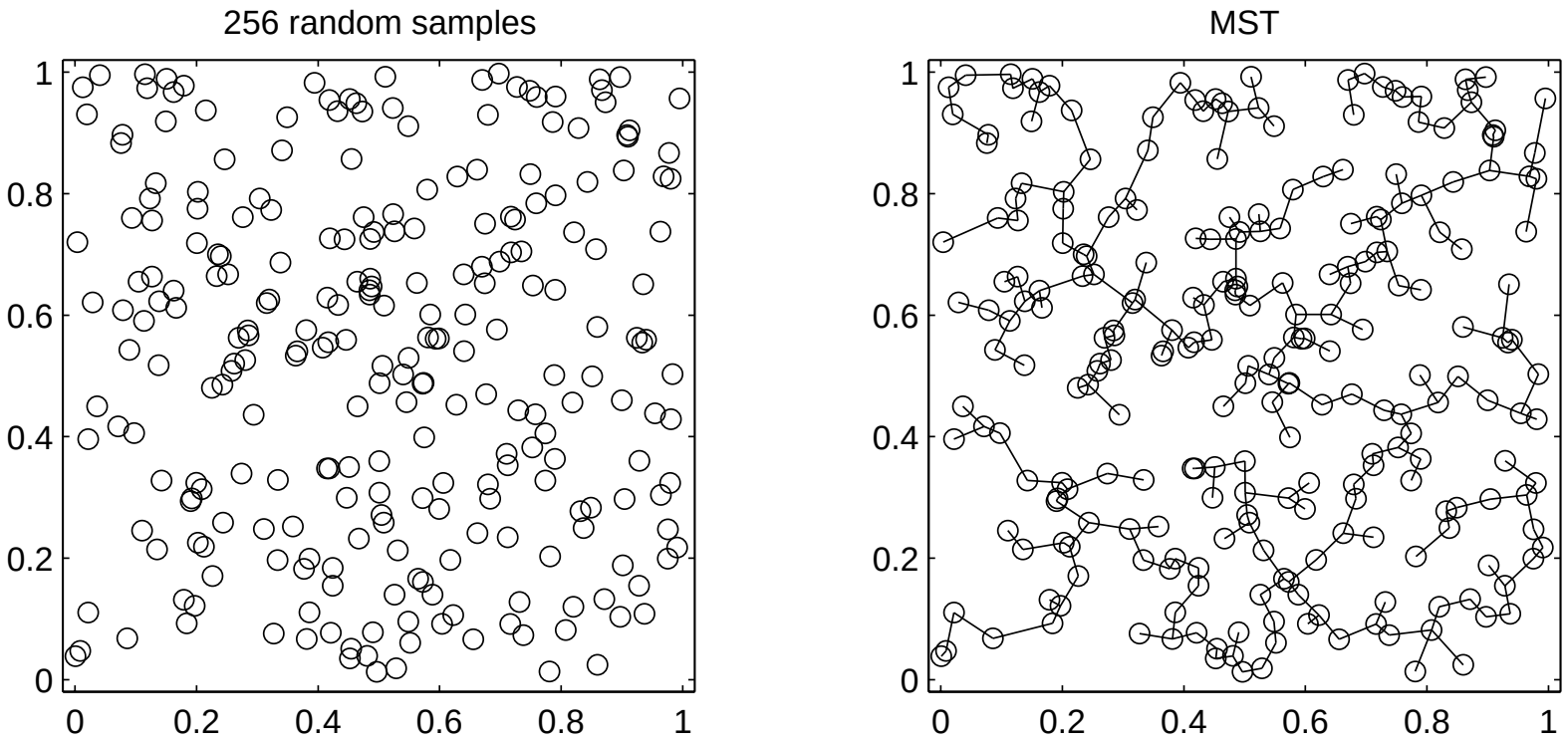


**Figure 7.** *Left: A scatterplot of a 256 point sample from triangle-uniform mixture density with* $\epsilon = 0.1$. *Labels 'o' and '*' mark those realizations from the uniform and pyramid densities, respectively. Right: superimposed is the* $k$*-MST implemented directly on the scatterplot* $\mathcal{X}_n$ *with* $k = 230$.

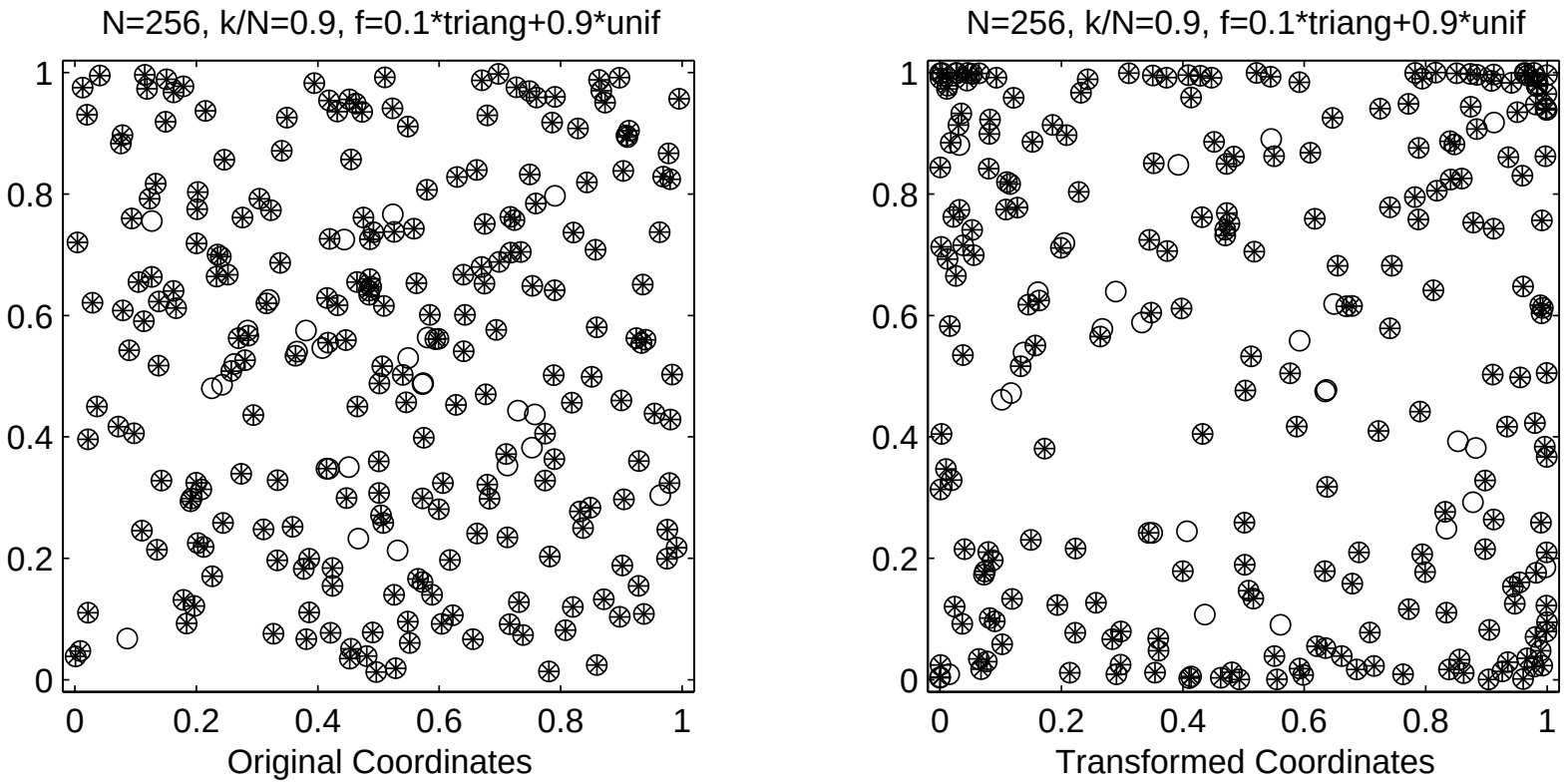


**Figure 8.** *Left: A sample from triangle-uniform mixture density with $\epsilon = 0.9$ in the transformed domain $\mathcal{Y}_n$. Labels 'o' and '*' mark those realizations from the uniform and pyramid densities, respectively. Right: transformed coordinates.*

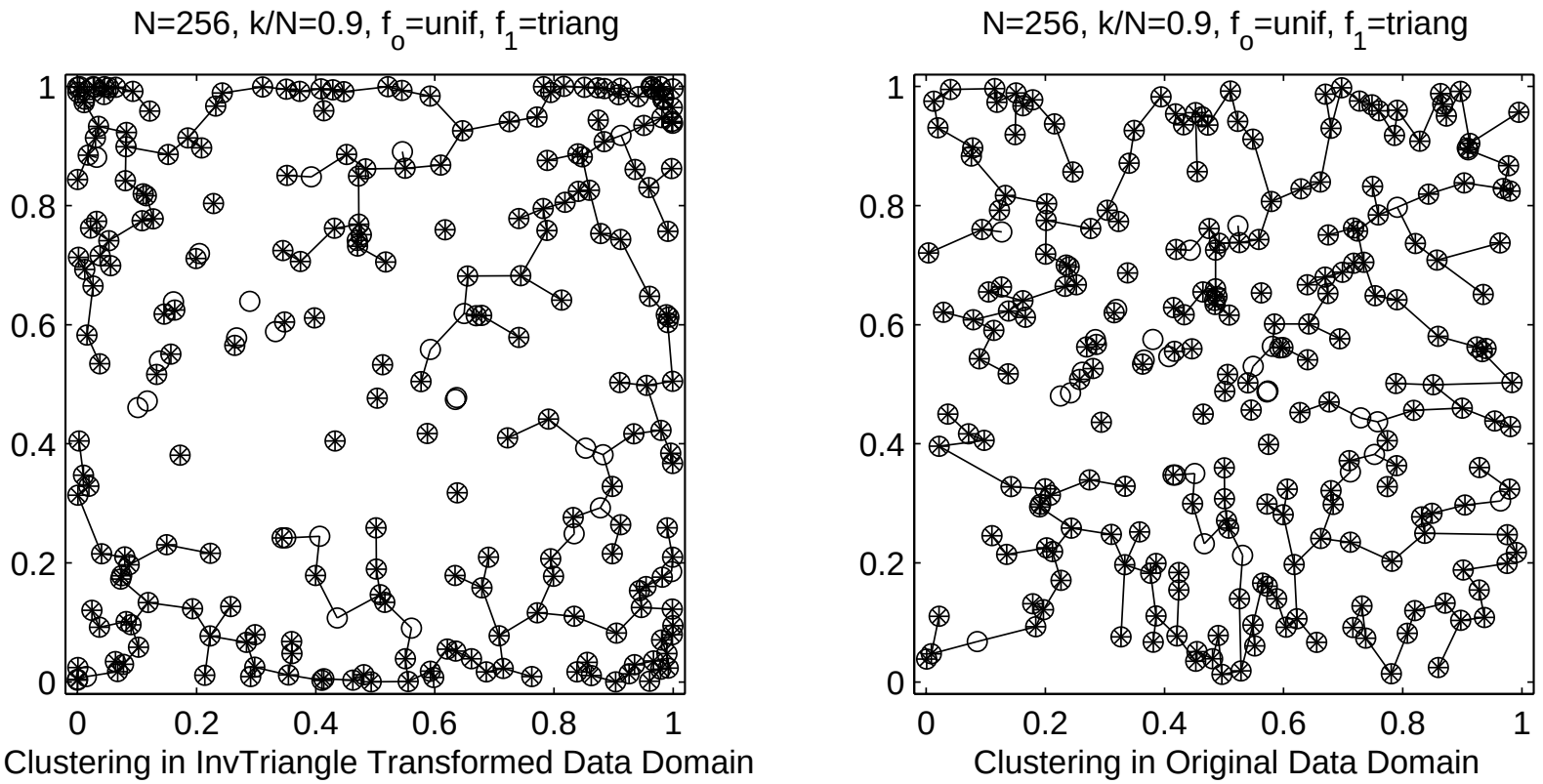


**Figure 9.** *Left: the $k$-MST implemented on the transformed scatterplot $\mathcal{Y}_n$ with $k = 230$. Right: same $k$-MST displayed in the original data domain.*

## 4.3. Application: Pattern Matching and Registration

Two independent data samples to be matched

- $X = [X_1, \ldots, X_n] \sim f(x)$
- $Y = [Y_1, \ldots, Y_m] \sim g(x)$

Suppose: $g(x) = f(Ax + b)$, $A^T A = I$

Objective: find rigid transformation $A, b$ to minimize Rényi divergence

$$I_\nu(f, g) = \frac{1}{\nu - 1} \ln \int_{\mathbf{R}^d} \left( \frac{g(x)}{f(x)} \right)^\nu f(x) dx$$

**Conclusions**

- Random quasi-additive graph weight converges to Rényi-Divergence of order $\nu$ after measure transformation
- Greedy polynomial implementations of $k$-MST have been developed for robust estimation, discrimination and pattern matching applications
- Decision threshold depends on difficult quantity $\beta_{d,\gamma}$