Estimation of Rényi Information Divergence via Pruned Minimal Spanning Trees[a]

Alfred Hero

Dept. of EECS,

The University of Michigan,

Olivier J.J. Michel

École Normale Supérieure de Lyon,

June 1999

[a]This work was supported in part by AFOSR under MURI grant: F49620-97-0028.

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. Random Power Weighted Minimal Spanning Trees

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

- n vertices x_i
- $\binom{n}{2}$ edge weights $e = e_{ij}$
- $\gamma \in (0, d)$: edge weight power exponent

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

$$L_n = \sum_{e \in \mathcal{T}_n} |e|^\gamma$$

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

$$L_n^* = L^*(\mathcal{X}_n) \quad = \quad \min_{\mathcal{T}_n} \sum_{e \in \mathcal{T}_n} |e|^\gamma$$

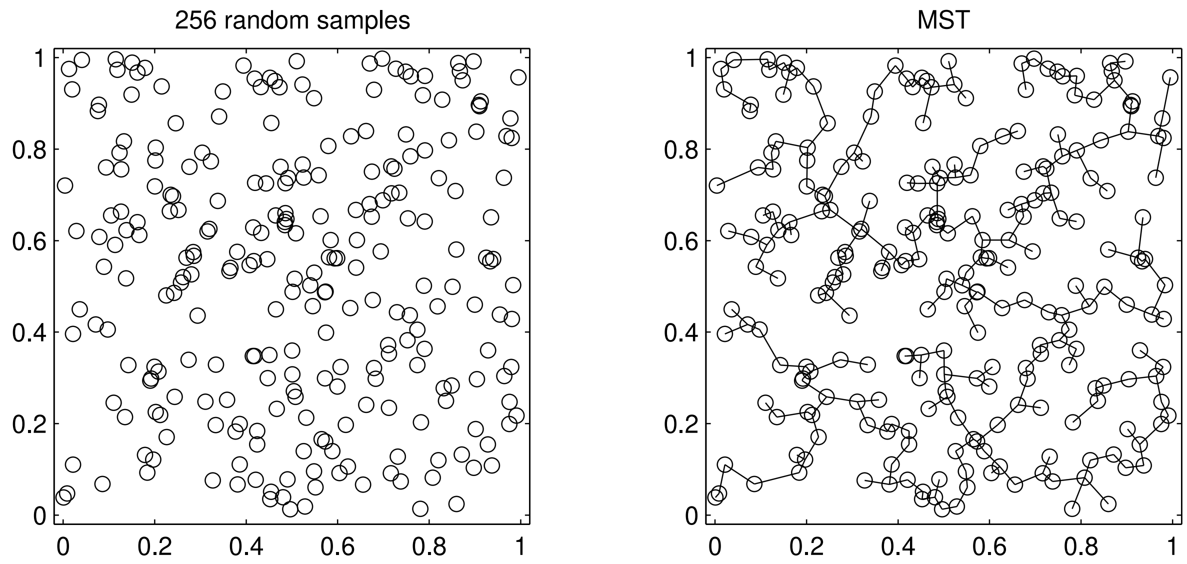

Figure 1. *A data set and the MST*

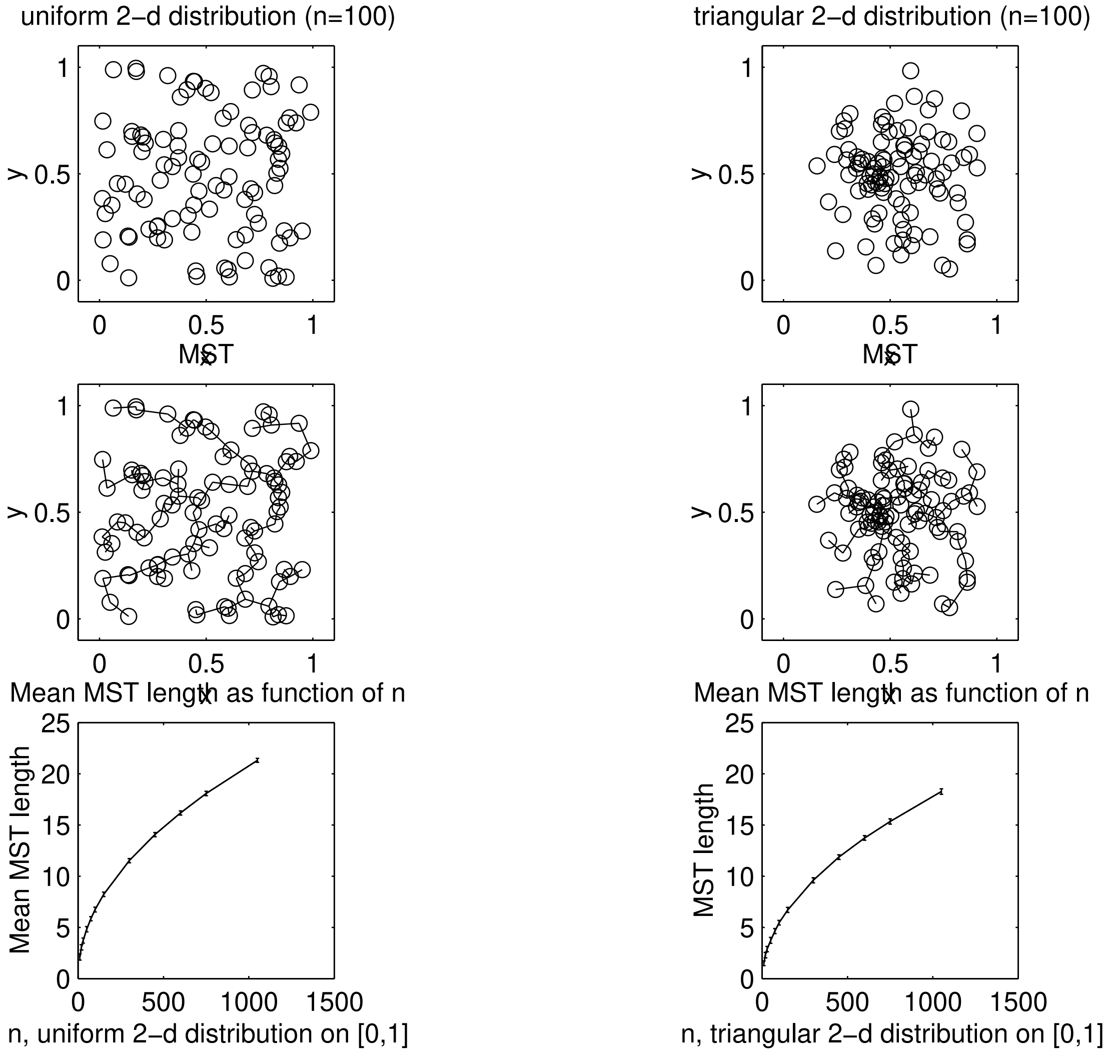

Figure 2. *2D Triangular vs. Uniform sample study.*

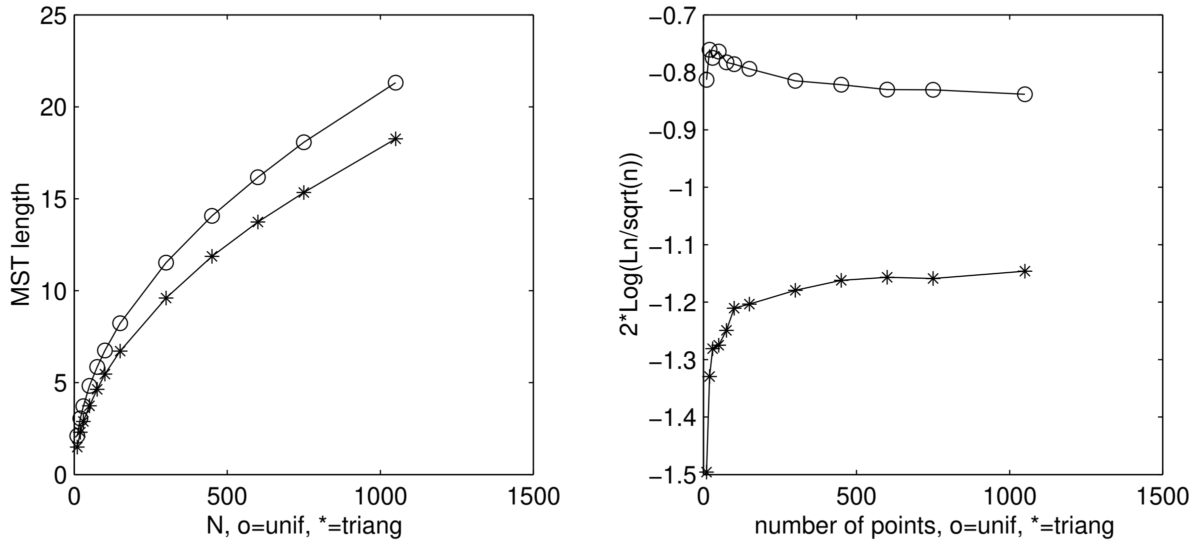

Figure 3. *MST and log MST weights as function of number of samples for 2D uniform vs. triangular study.*

3. Entropy and the MST length: The BHH Thm

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 $\mathbf{R}^d$, $d > \gamma > 0$ and define $\nu = (d-\gamma)/d$. Then [1]

$$L_n^*/n^\nu \;\to\; \beta_{L,\gamma} \int_{\mathbf{R}^d} f^\nu(x)dx \qquad (w.p.1)$$

With $H_\nu(U) = \frac{1}{1-\nu} \ln \beta_{L,\gamma}$, define

$$\hat{H}_\nu(\mathcal{X}_n) = \frac{1}{1-\nu} \ln \left(n^{-\nu} L_n^*\right) + H_\nu(U) \tag{3}$$

Then

$$\hat{H}_\nu(\mathcal{X}_n) \to \frac{1}{1-\nu} \ln \int f^\nu(x)dx \quad (a.s.)$$

4. Outlier Sensitivity of the MST

Assume f is a mixture density of the form

$$f = (1-\epsilon) f_1 + \epsilon f_o, \tag{4}$$

where

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

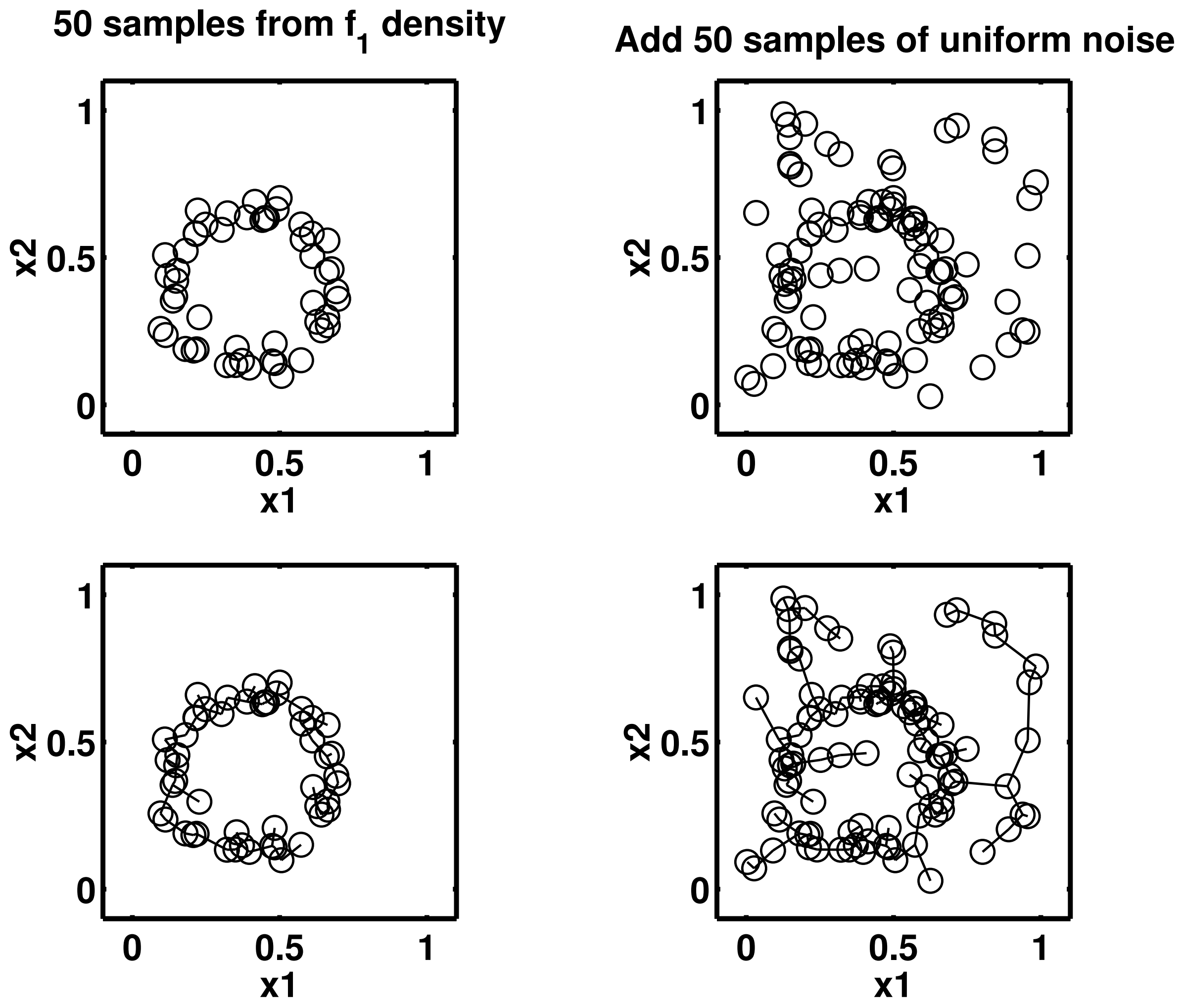

Figure 4. *1st row: 2D torus density with and without the addition of uniform "outliers." 2nd row: corresponding MST's.*

5. The k-MST

Fix k, $1 \leq n$.

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

The power weighted *k-minimal spanning tree* (k-MST) $\mathcal{T}^*_{n,k} = \mathcal{T}^*(x_{i^*_1}, \ldots, x_{i^*_k})$ is the minimum weight MST among the $\binom{n}{k}$ MST's connecting subsets of k vertices of $\mathcal{G}$:

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

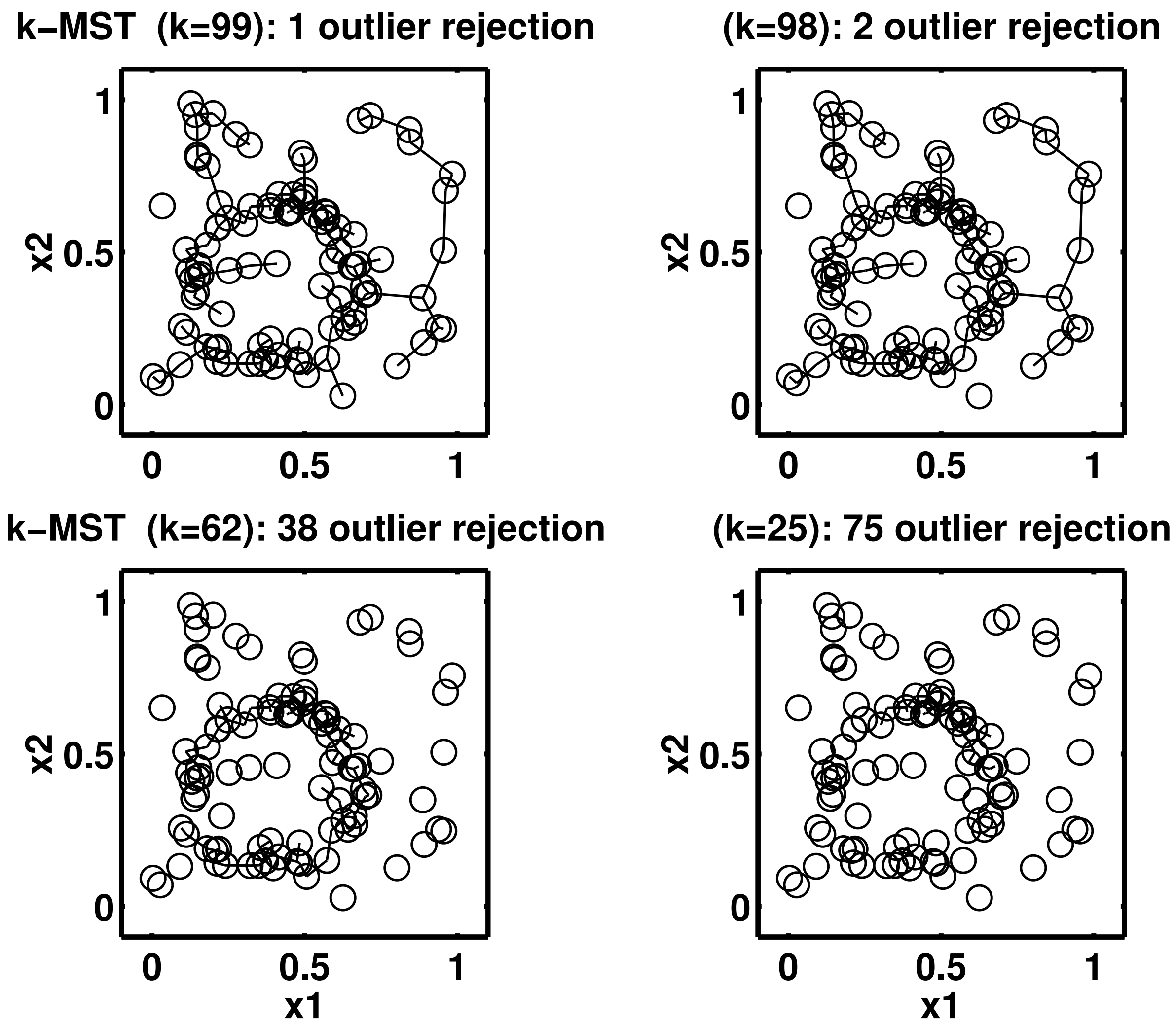

Figure 5. *k-MST for 2D torus density with and without the addition of uniform "outliers".*

6. Extended BHH Thm for k-MST

Define:

$$\hat{H}_\nu(\mathcal{X}^*_{n,k}) = \frac{1}{1-\nu} \ln \left(k^{-\nu} L(\mathcal{X}^*_{n,k})\right) + H_\nu(U)$$

Then if $k = \lfloor \alpha n \rfloor$, as $n \to \infty$ we have [2]

$$\hat{H}_\nu(\mathcal{X}^*_{n,k}) \to \min_{A:P(A)\geq\alpha} \frac{1}{1-\nu} \ln \int f^\nu(x|A)dx \quad (a.s.)$$

7. I-Divergence and the MST (k-MST) length

- $g(x)$: 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 transformation: $x = [x^1, \ldots, x^d]^T \to y = [y^1, \ldots, y^d]^T$

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

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

- 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 $x^k = G^{-1}(y^k|x^{k-1},\ldots,x^1)$.

- Then we know

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

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

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

8. 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 $[0,1]^2$

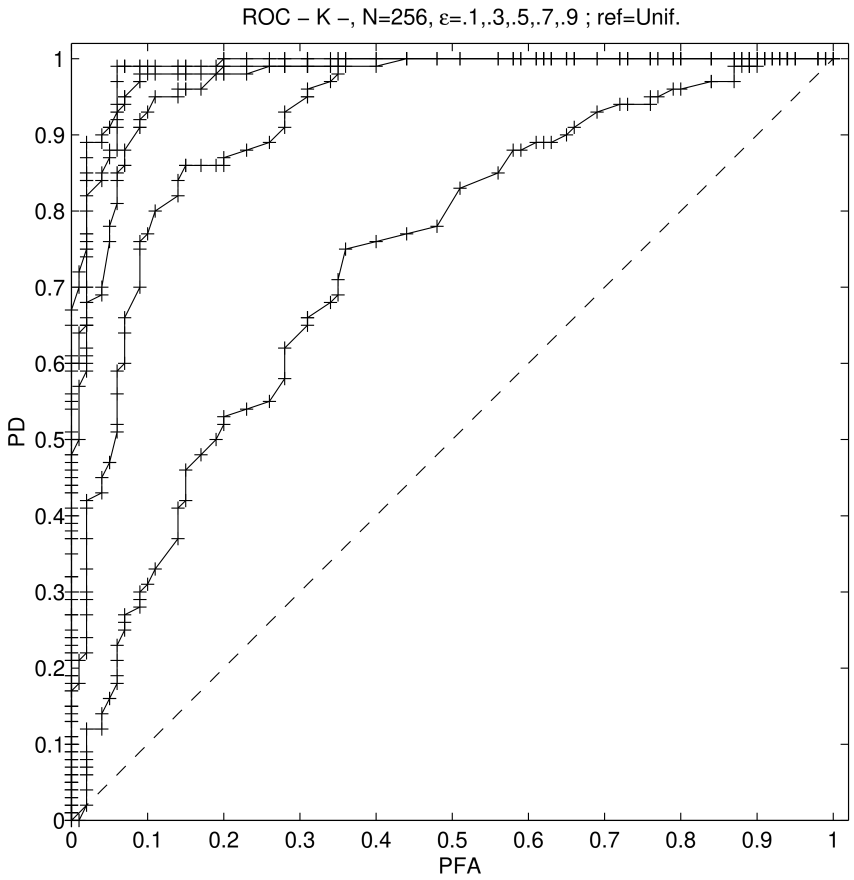

Figure 6. *ROC curves for the Rényi information divergence test for detecting triangle-uniform mixture density* $f = (1-\epsilon)f_1 + \epsilon f_0$ *(*H_1*) against triangular hypothesis* $f = f_1$ *(*H_0*). Curves are increasing in* $\epsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

9. Application II: Nonuniform Outlier Rejection

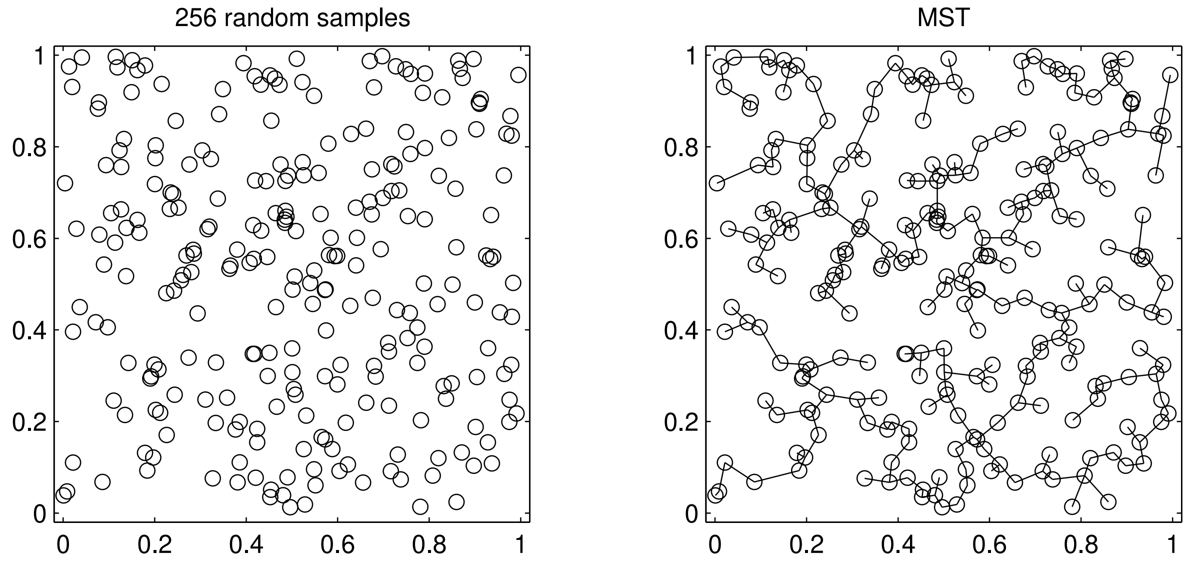

Figure 7. *Left: A scatterplot of a 256 point sample from triangle-uniform mixture density with* $\epsilon = 0.1$. *Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Right: superimposed is the* k*-MST implemented directly on the scatterplot* $\mathcal{X}_n$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

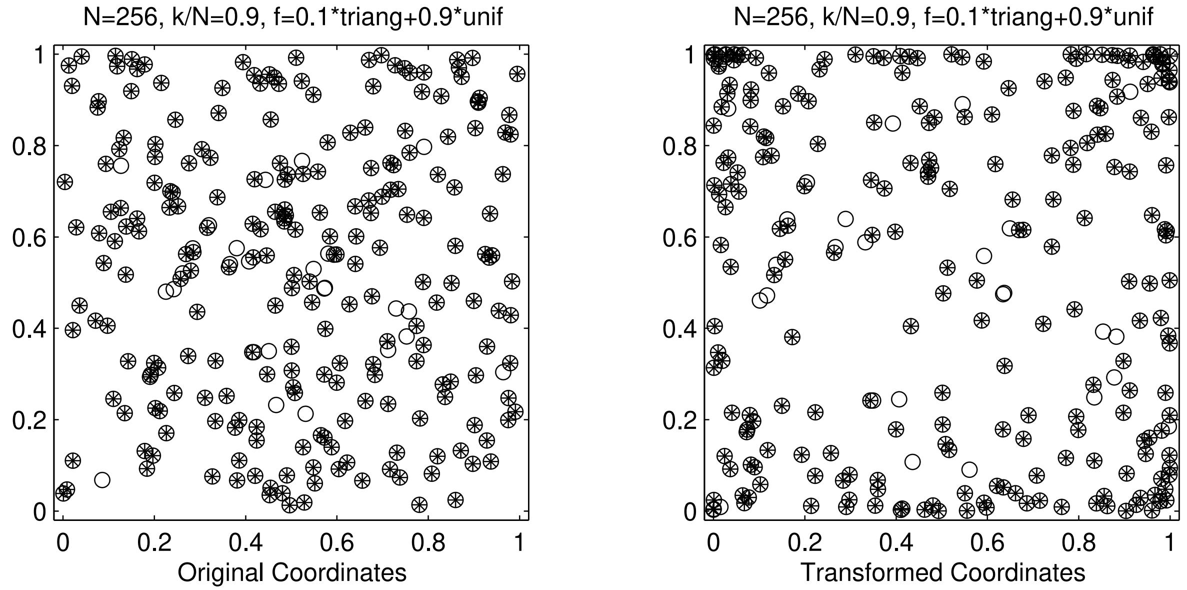

Figure 8. *Left: A sample from triangle-uniform mixture density with $\epsilon = 0.9$ in the transformed domain $\mathcal{Y}_n$. Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Right: transformed coordinates.*

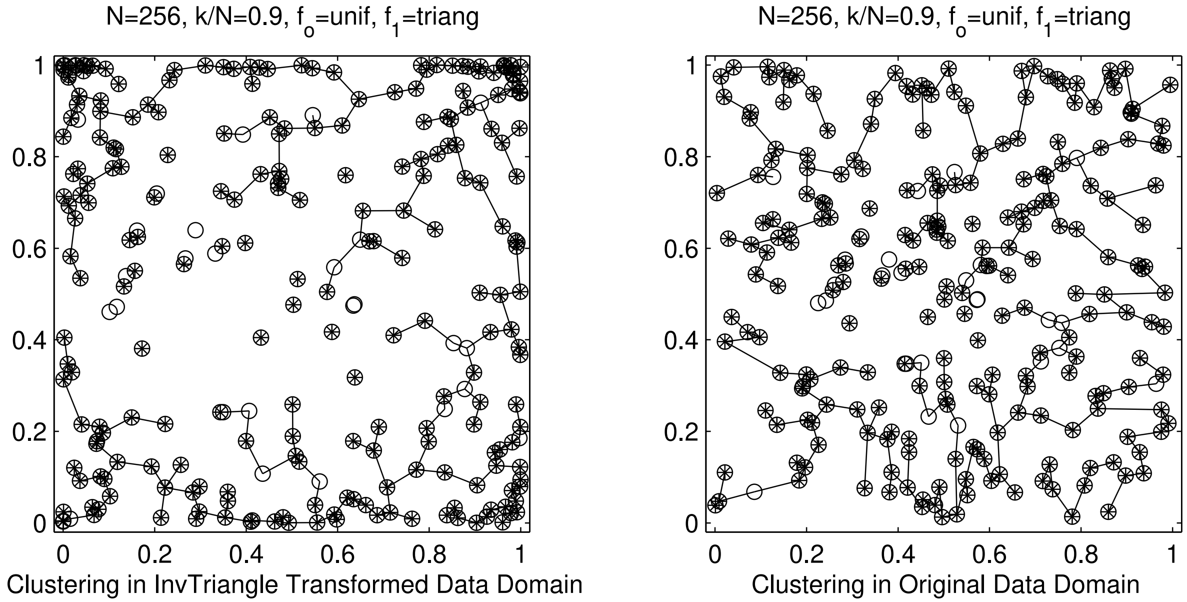

Figure 9. *Left: the k-MST implemented on the transformed scatter-plot $\mathcal{Y}_n$ with $k = 230$. Right: same k-MST displayed in the original data domain.*

References

[1] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[2] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. to appear, , 1999.