Entropic graphs: theory

Alfred O. Hero

Dept. EECS, Dept Biomed. Eng., Dept. Statistics

University of Michigan - Ann Arbor

hero@eecs.umich.edu

http://www.eecs.umich.edu/~hero

Collaborators: O. Michel, B. Ma, H. Neemuchwala, J. Costa,

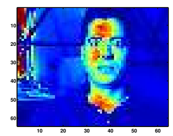
1. Motivating Examples in Imaging and Computer Vision
2. Entropic Feature Similarity/Dissimilarity Measures
3. Entropic Euclidean Graphs over Feature Space
4. Asymptotics of Entropic Graphs
5. Entropic Graphs for Pattern Matching

SOURCE: c:\hero\seminars\Distinguished\Entropic\ipm1_slides.tex

I. Motivating Examples in Imaging and Computer Vision

- Image retrieval and indexing
- Multimodality image fusion
- Inference on shape manifolds
- Image registration

Image Retrieval



QUERY

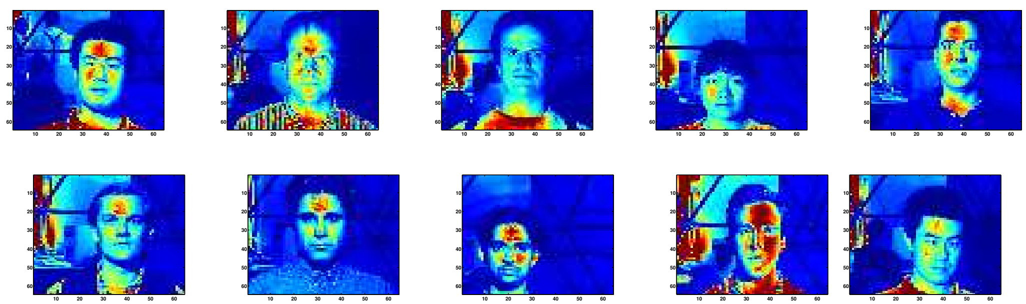


Figure 1: Yale face database `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`

DATABASE

Feature Vectors in Feature Space

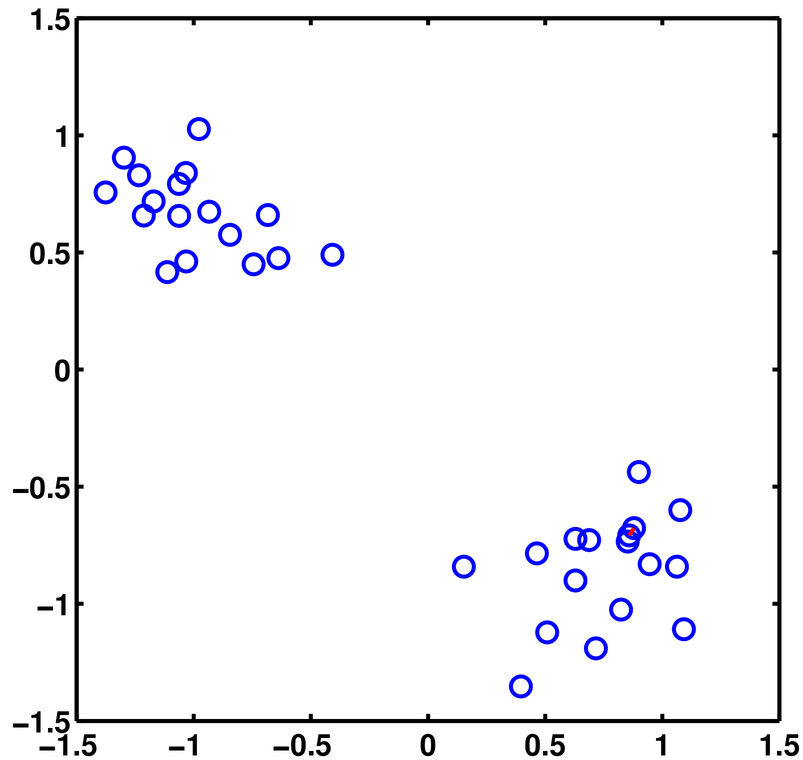
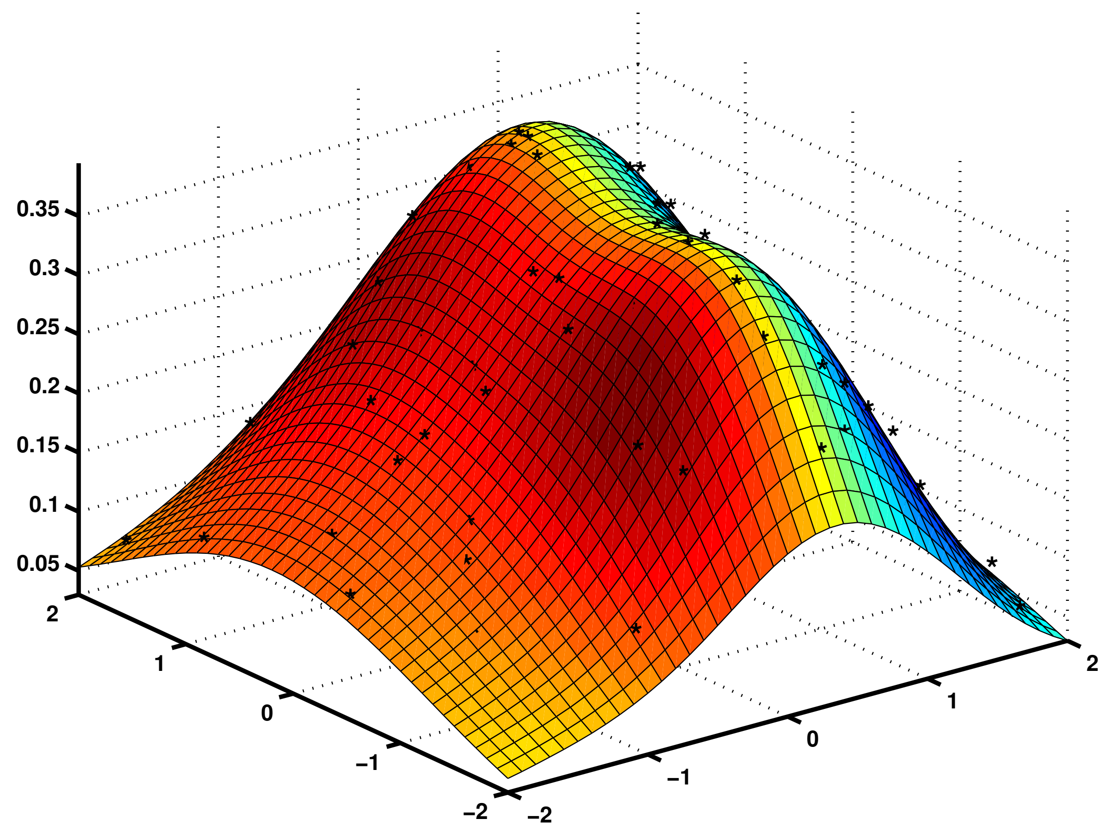


Figure 2: Vectors of projection coefficients extracted from two different images.

Inference on Shape Manifolds

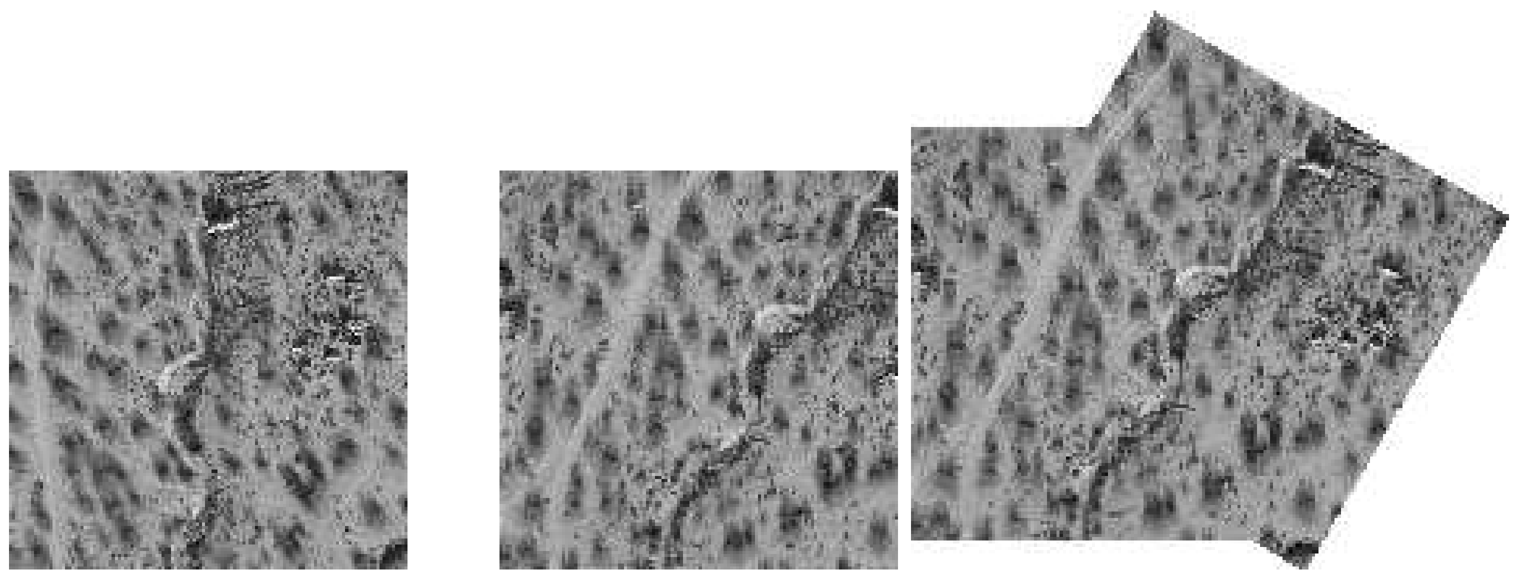


Multi-modality face Retrieval

Database of Visible/IR images

http://www.equinoxsensors.com/products/HID.html

Image Registration



(a) Image I_1 (b) Image I_0 (c) Registration result

Figure 3: A multidate image registration example

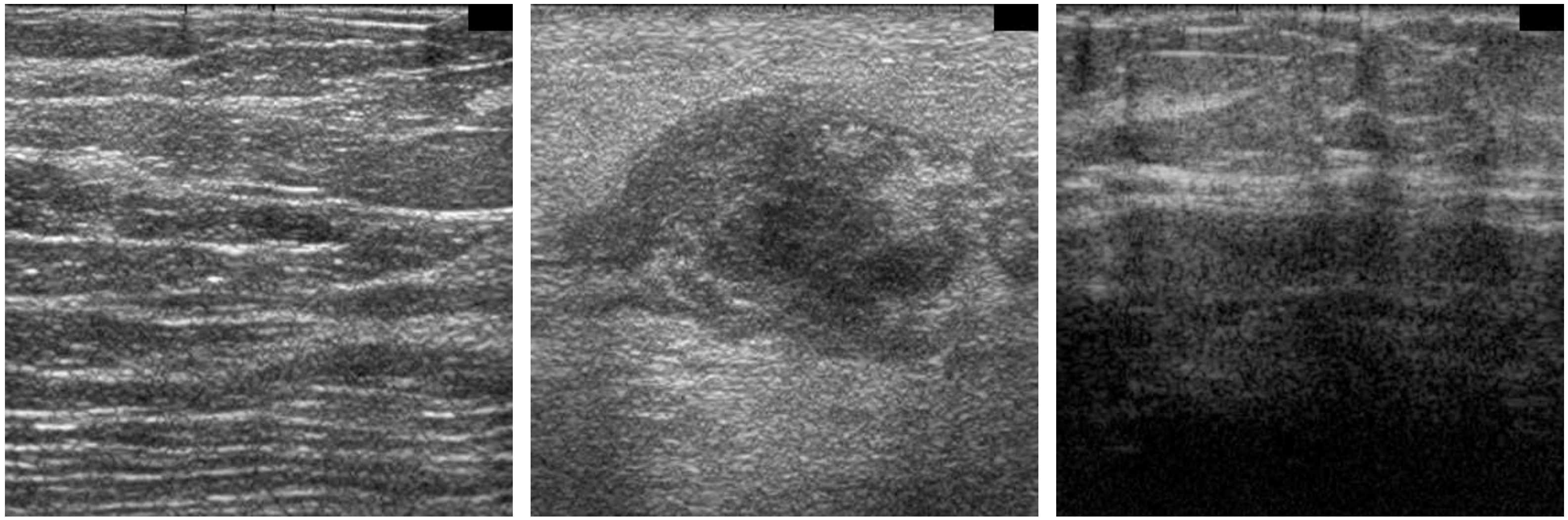
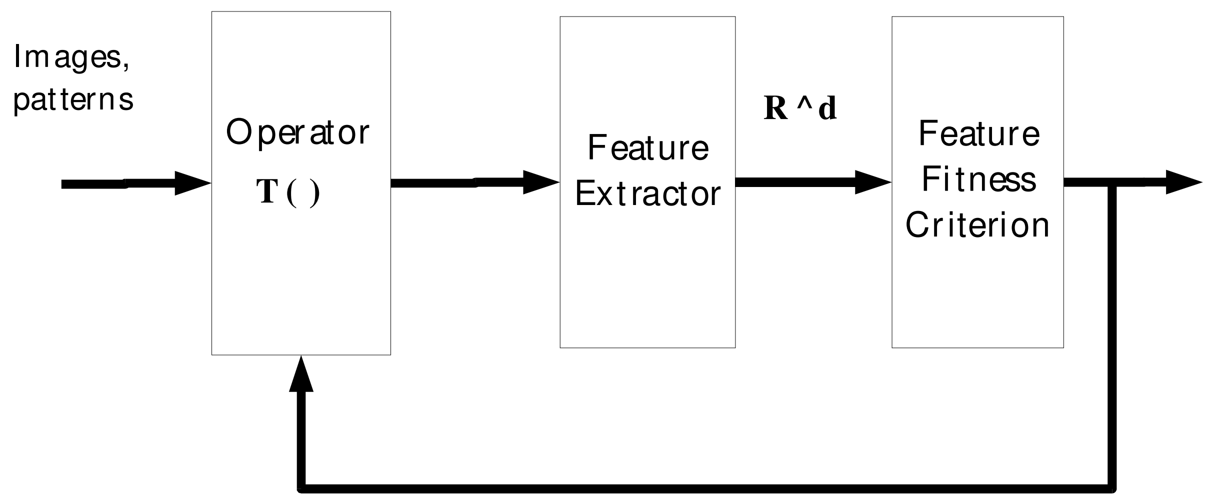


Figure 4: Three ultrasound breast scans. From top to bottom are: case 151, case 142 and case 162.

System Block Diagram

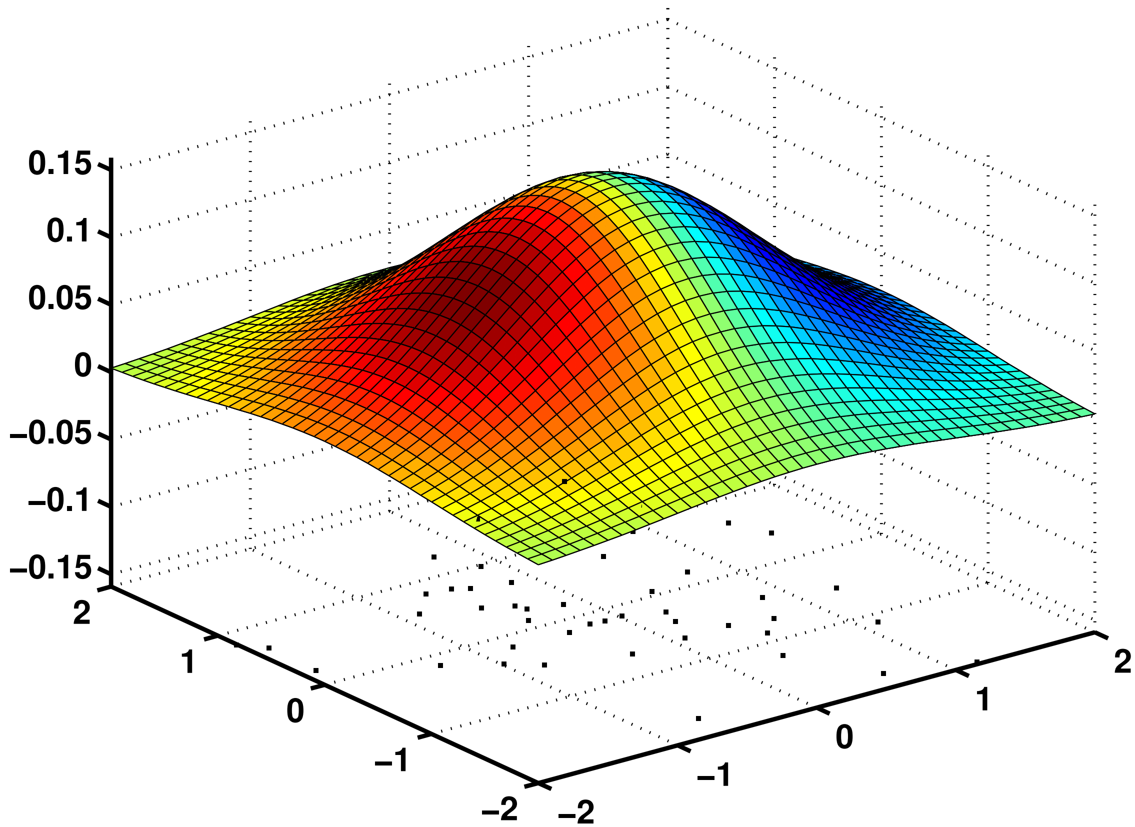


Objective: For given fitness criterion Q, find operator T which minimizes/maximizes Q

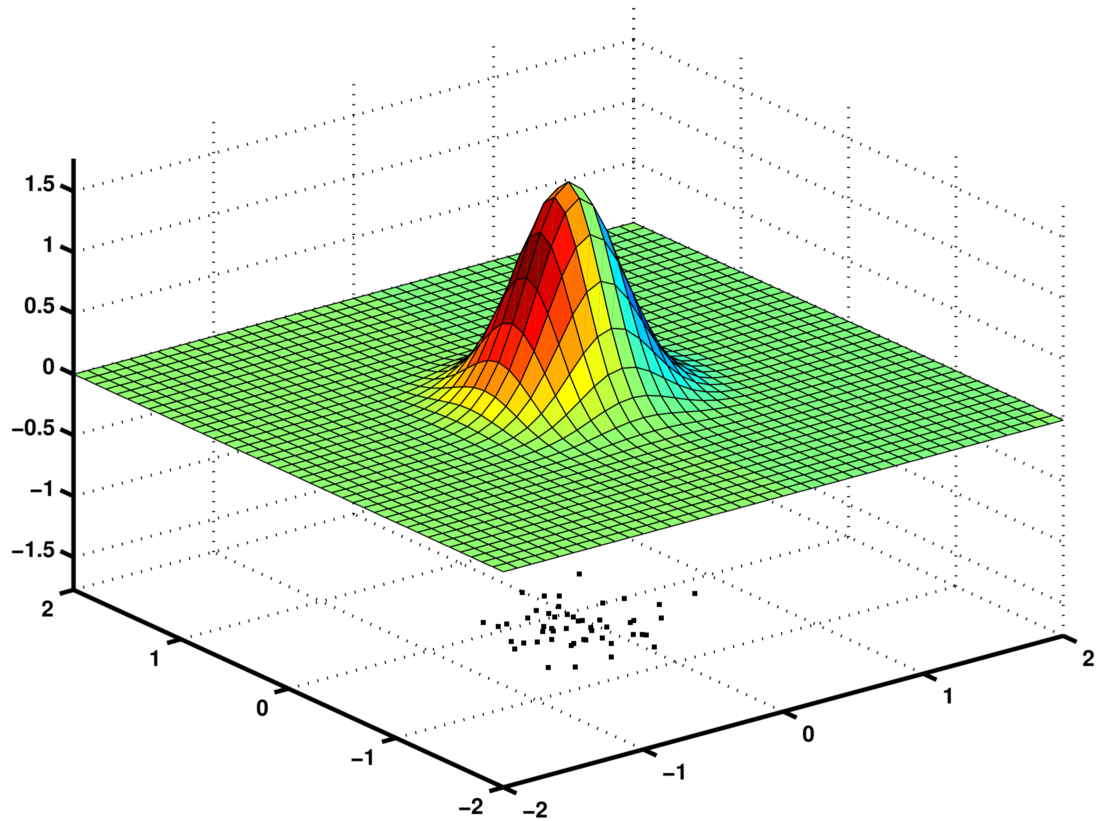
Our focus: entropic fitness criterion $Q(f)$

f: feature density over $x \in [0, 1]^d$

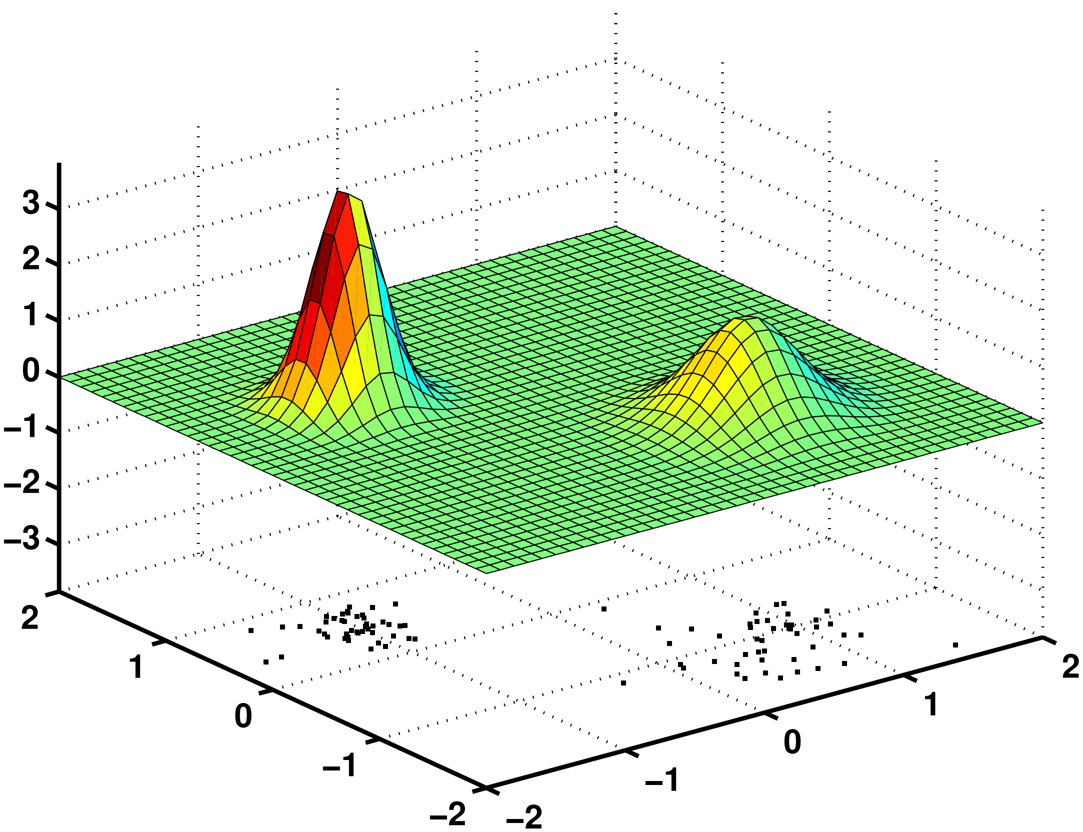
High Entropy Feature Density



Low Entropy Feature Density



High Entropy Feature Density



II. Entropic Similarity/Dissimilarity Measures

1. Shannon Entropy of feature density f

$$Q(f) = H(f) = -\int f(x) \ln f(x)\, dx$$

2. Jensen difference between feature densities f, g:

$$Q(f,g) = H(\varepsilon f + (1-\varepsilon)g) - \varepsilon H(f) - (1-\varepsilon)H(g)$$

3. KL Divergence between feature densities f, g

$$Q(f,g) = D(f\|g) = \int f(x) \ln \left(\frac{f(x)}{g(x)} \right) dx$$

4. Mutual information between feature sets $f_{X,Y}$

$$Q(f_{X,Y}) = \mathrm{MI}(X,Y) = \int \int f_{X,Y}(x,y) \ln \left(\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \right) dx$$

Issue: How to estimate entropic Q from measured data?

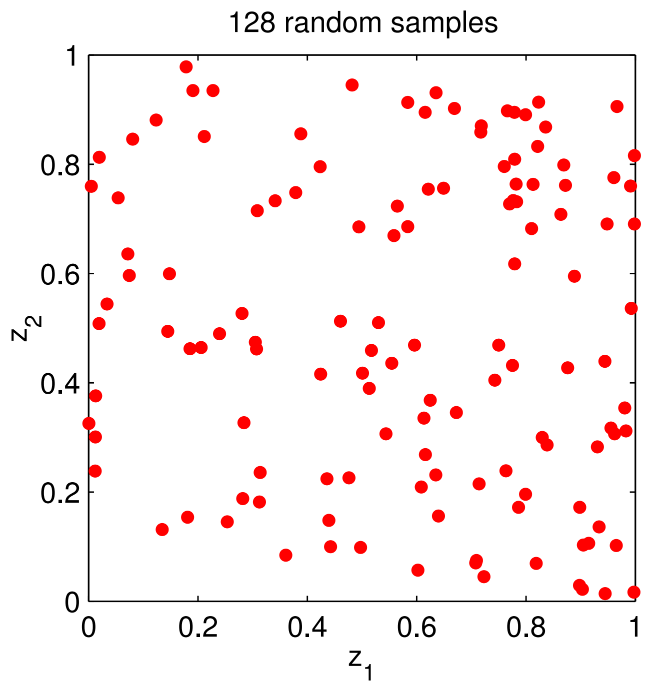
Some possibilities:

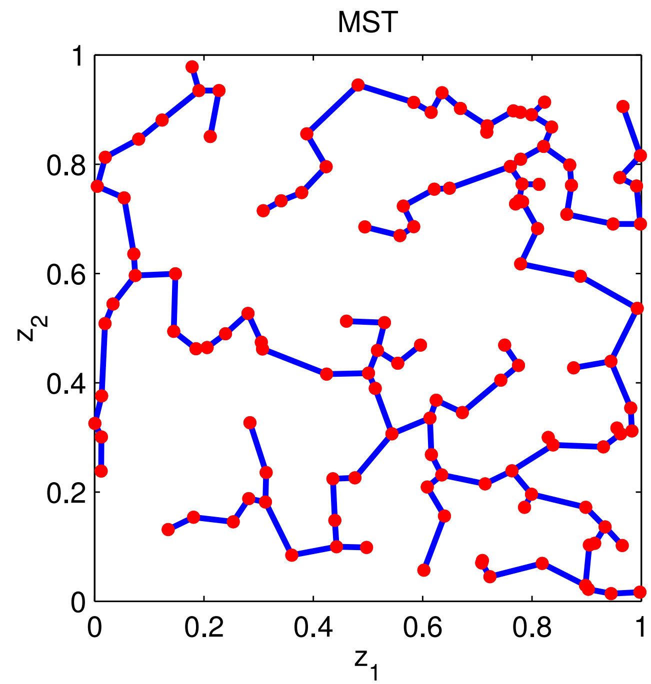
1. Assume parameteric models for f, g, $f_{X,Y}$ (Vasconcelos&Lipman:2000,Stoica&etal:1998)
2. Substitute non-parametric density estimates of f, g, $f_{X,Y}$
   (a) Quantize feature space and use histogram estimates (Beirlant&etal:1997)
   (b) Use adaptive partitioning density estimates (Vasicek:1976, Miller:2002, Gray&etal:2000)
3. Use "entropic graphs" which emulate/estimate Q (Hero&Michel:1997,Neemwuchwala,Hero&Carson:2002)

III. Entropic Euclidean Graphs

1. The minimal spanning tree (MST)
2. The k-nearest neighbor (k-NN) graph
3. Asymptotic trends

A Set of Feature Samples and a Euclidean Spanning Graph



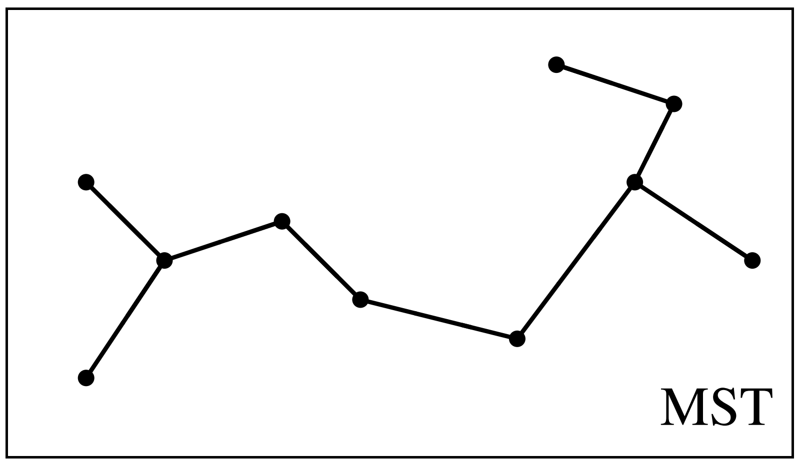


Minimal Euclidean Graphs: MST

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

The Euclidean Power Weighted MST achieves

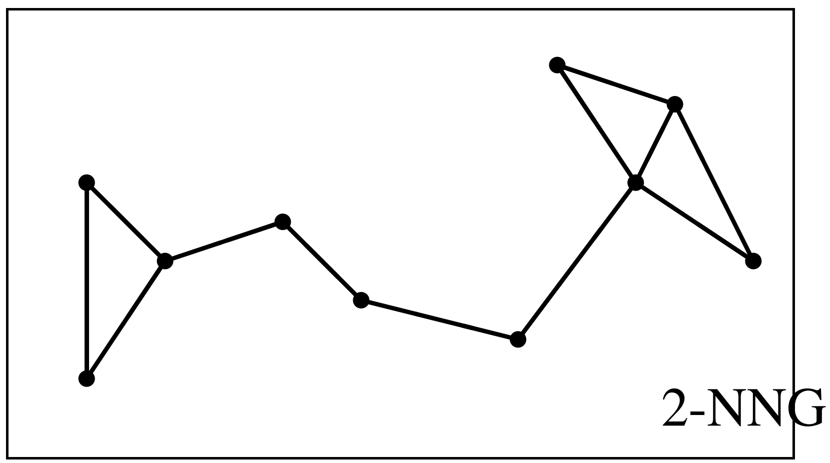
$$L_\gamma^{\mathrm{MST}}(\mathcal{X}_n) = \min_{\mathrm{T}_n} \sum_{e \in \mathrm{T}_n} \|e\|^\gamma.$$



Minimal Euclidean graphs: k-NNG

Let $\mathcal{N}_{k,i}(\mathcal{X}_n)$ denote the possible sets of k edges connecting point x_i to all other points in $\mathcal{X}_n$.
The Euclidean Power Weighted k-NNG is

$$L_\gamma^{k-NNG}(\mathcal{X}_n) = \sum_{i=1}^{n} \min_{\mathcal{N}_{k,i}(\mathcal{X}_n)} \sum_{e \in \mathcal{N}_{k,i}(\mathcal{X}_n)} |e|^\gamma$$



MST for Two Different Samples

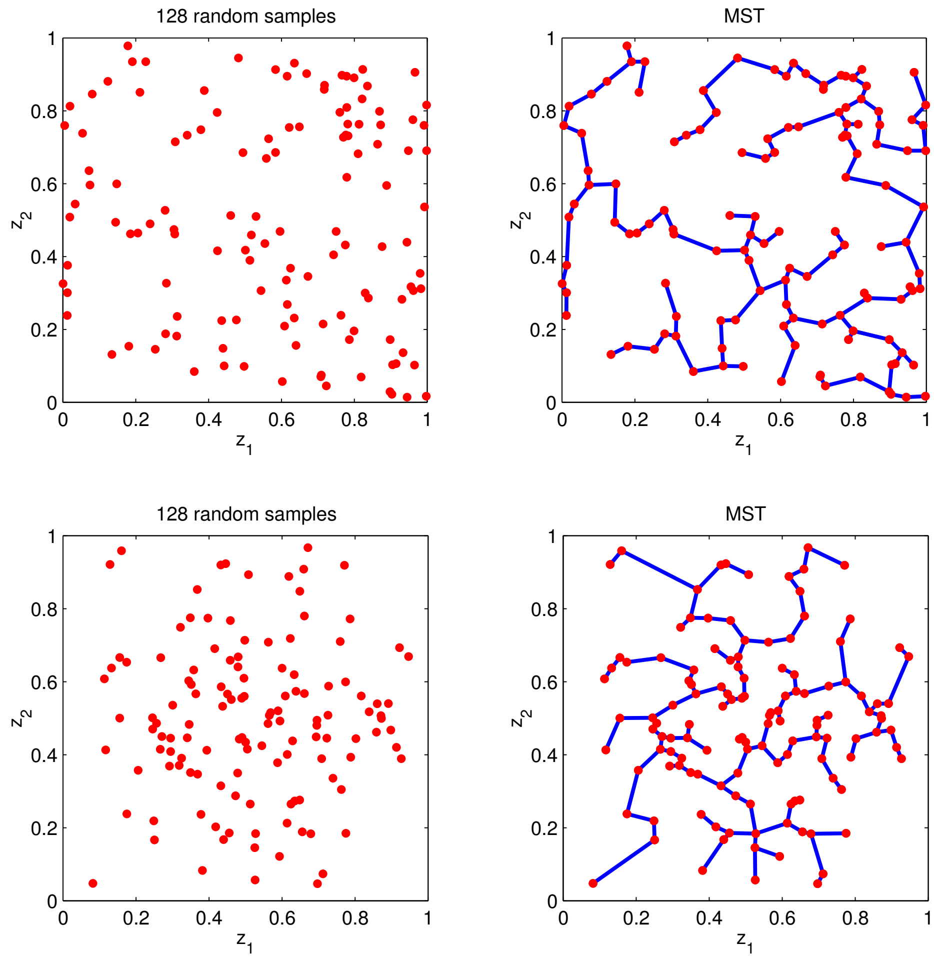


Figure 5:

Large n behavior of MST

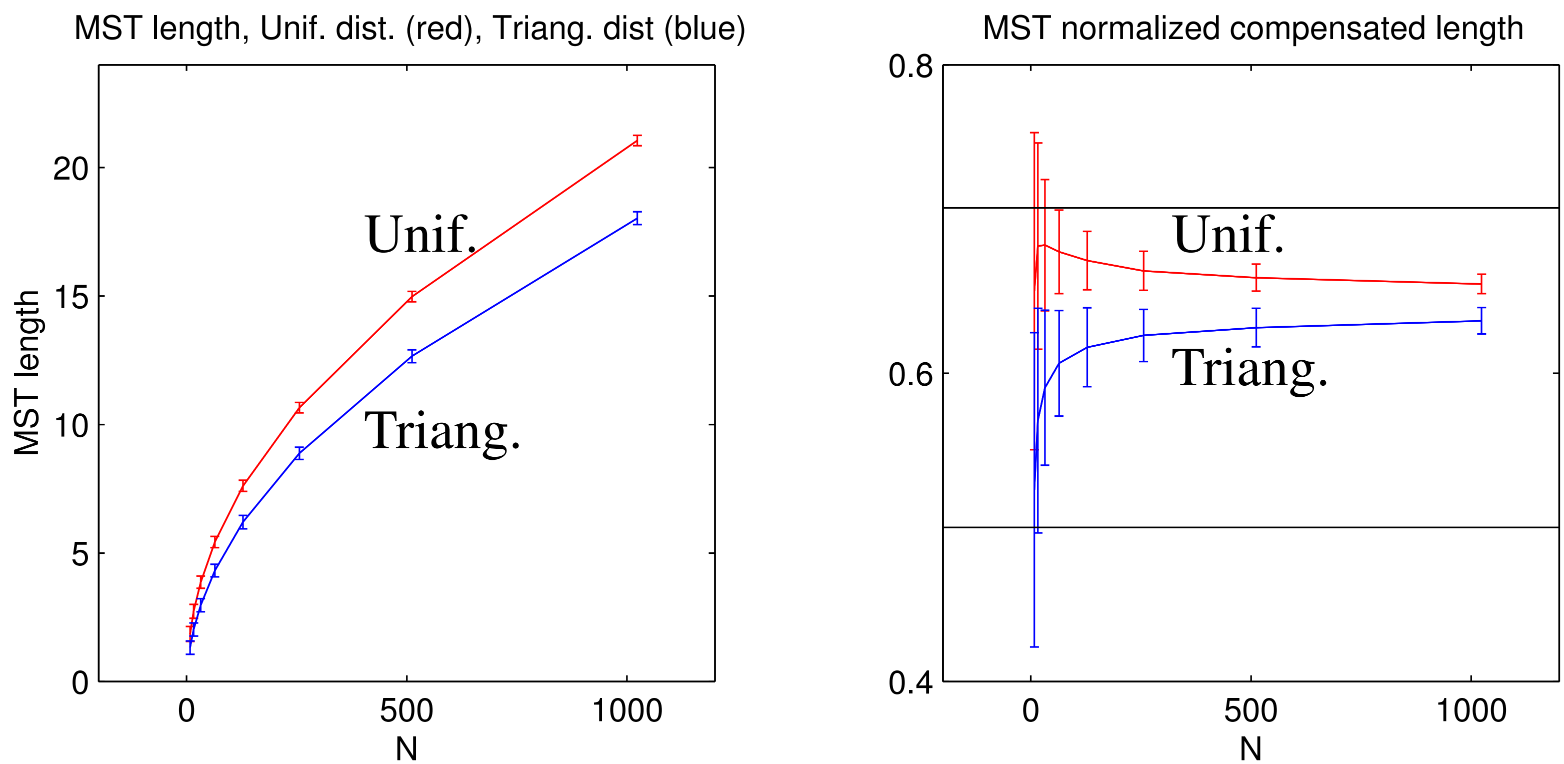


Figure: MST and log MST weights as function of the number of samples.

IV. Entropic Euclidean Graph Theory

1. The Beardwood, Halton Hammersley Theorem
2. Extension to divergence estimation
3. Extension to greedy algorithms
4. Extension to K-MST

Asymptotics: the BHH Theorem

Define the MST length functional

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

Theorem 1 *[Beardwood, Halton&Hammersley:1959] Let* $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ *be an i.i.d. realization from a Lebesgue density* f *with support* $\mathcal{S} \subset [0,1]^d$.

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

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

$$\frac{1}{1-\alpha} \ln\left(L_\gamma(\mathcal{X}_n)/n^\alpha\right) \quad \to \quad H_\alpha(f) + c \qquad (a.s.)$$

Rényi Entropy and Divergence

- Rényi Entropy of order α [Rényi:61,70]

$$H_\alpha(f) = \frac{1}{1-\alpha} \ln \int_S f^\alpha(x) dx$$

- Rényi α-divergence of fractional order $\alpha \in [0,1]$

$$\begin{aligned} D_\alpha(f_1 \,\|\, f_0) &= \frac{1}{\alpha-1} \ln \int_S f_0 \left(\frac{f_1}{f_0}\right)^\alpha dx \\ &= \frac{1}{\alpha-1} \ln \int_S f_1^\alpha f_0^{1-\alpha} dx \end{aligned}$$

  – α-Divergence vs. Kullback-Liebler divergence

$$\lim_{\alpha \to 1} D_\alpha(f_1 \| f_0) = \int f_1 \ln \frac{f_1}{f_0} dx.$$

α-Divergence and Decision Theoretic Error Exponents

Let Z_i be i.i.d.:

$$\begin{aligned} H_0 &: Z_i \sim f \\ H_1 &: Z_i \sim g \end{aligned}$$

Bayes probability of error

$$P_e(n) = \beta(n)P(H_1) + \alpha(n)P(H_0)$$

Sanov bound (Blahut:1987,Dembo&Zeitouni:98)

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n}\log P_F(n) &= -\sup_{\alpha\in[0,1]} \{(1-\alpha)D_\alpha(g\|f)\} \\ \liminf_{n\to\infty} \frac{1}{n}\log P_M(n) &= -\sup_{\alpha\in[0,1]} \{(1-\alpha)D_\alpha(f\|g)\}. \end{aligned}$$

Entropic Graphs for Clustering and Outlier Rejection: k-MST

Assume f is a mixture density of the form

$$f = (1-\varepsilon) f_1 + \varepsilon f_o,$$

where

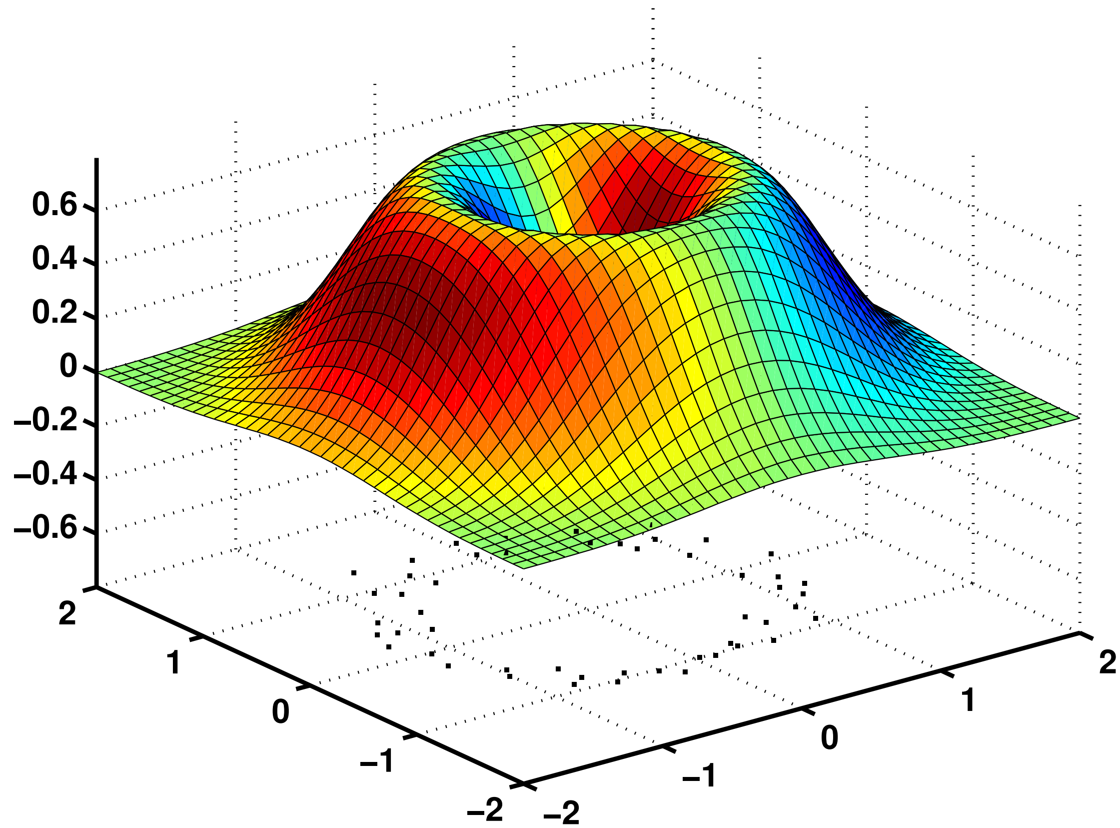
- f_o is a known ”outlier” density
- f_1 is an unknown target density
- $\varepsilon \in [0,1]$ is unknown mixture parameter

Objective: given realization $\mathcal{X}_n$ from f cluster the realizations from f_1.

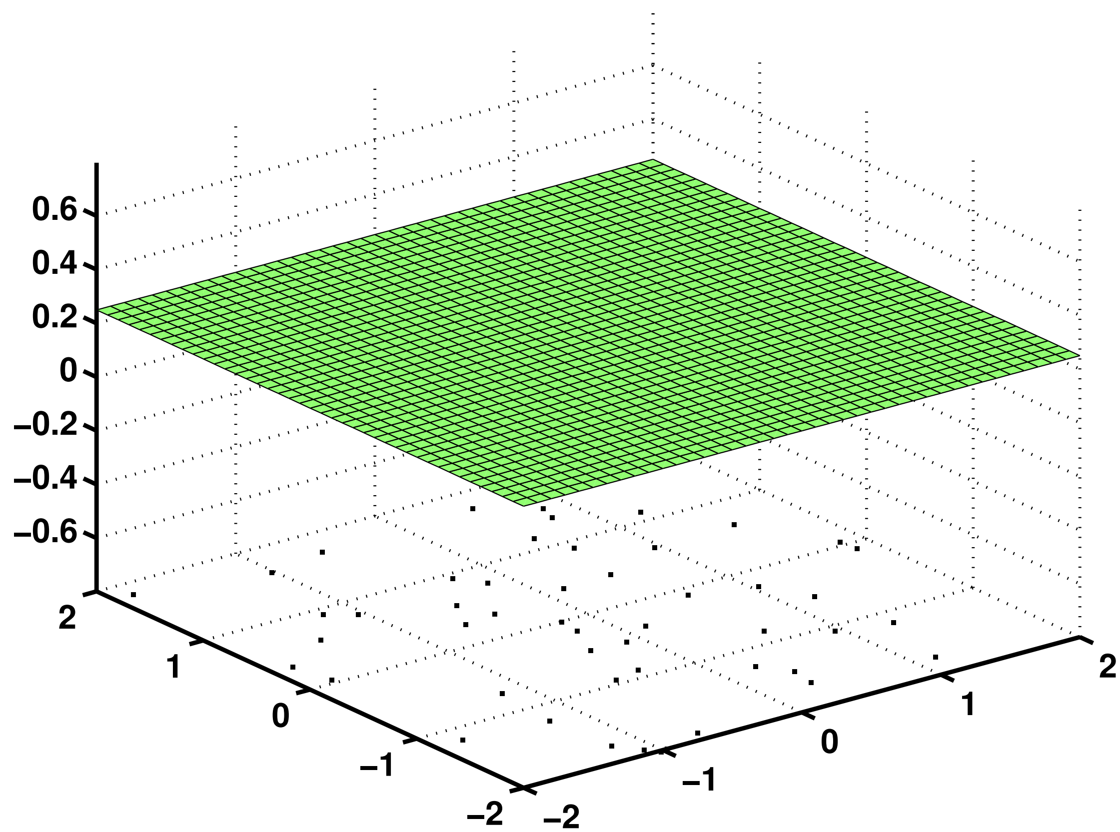
Two-step k-MST procedure:

1. Convert f_o to maxent (uniform) density via measure transformation
2. ”Prune” the MST on transformed $\mathcal{X}_n$ to eliminate vertices arising from maxent density

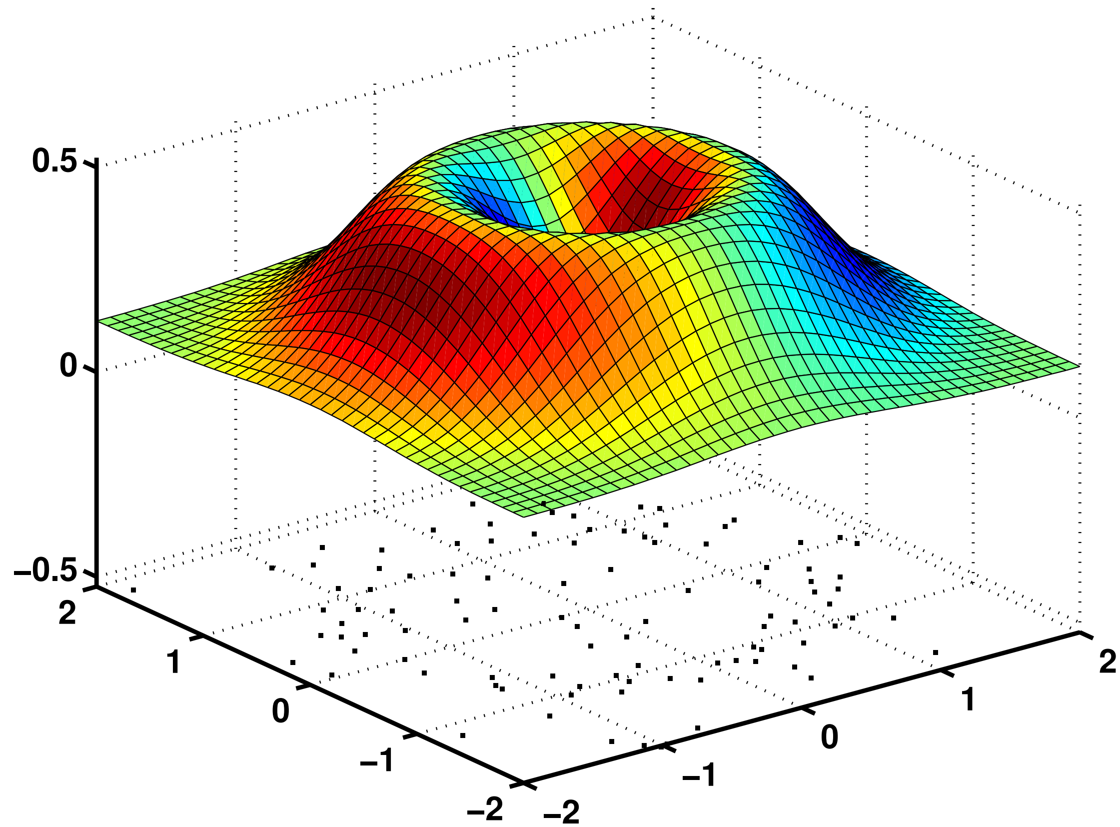
Example: Annulus Target Density f_1



Uniform Outlier Density f_o



Mixture Density



k-point Minimal Spanning Tree (k-MST)

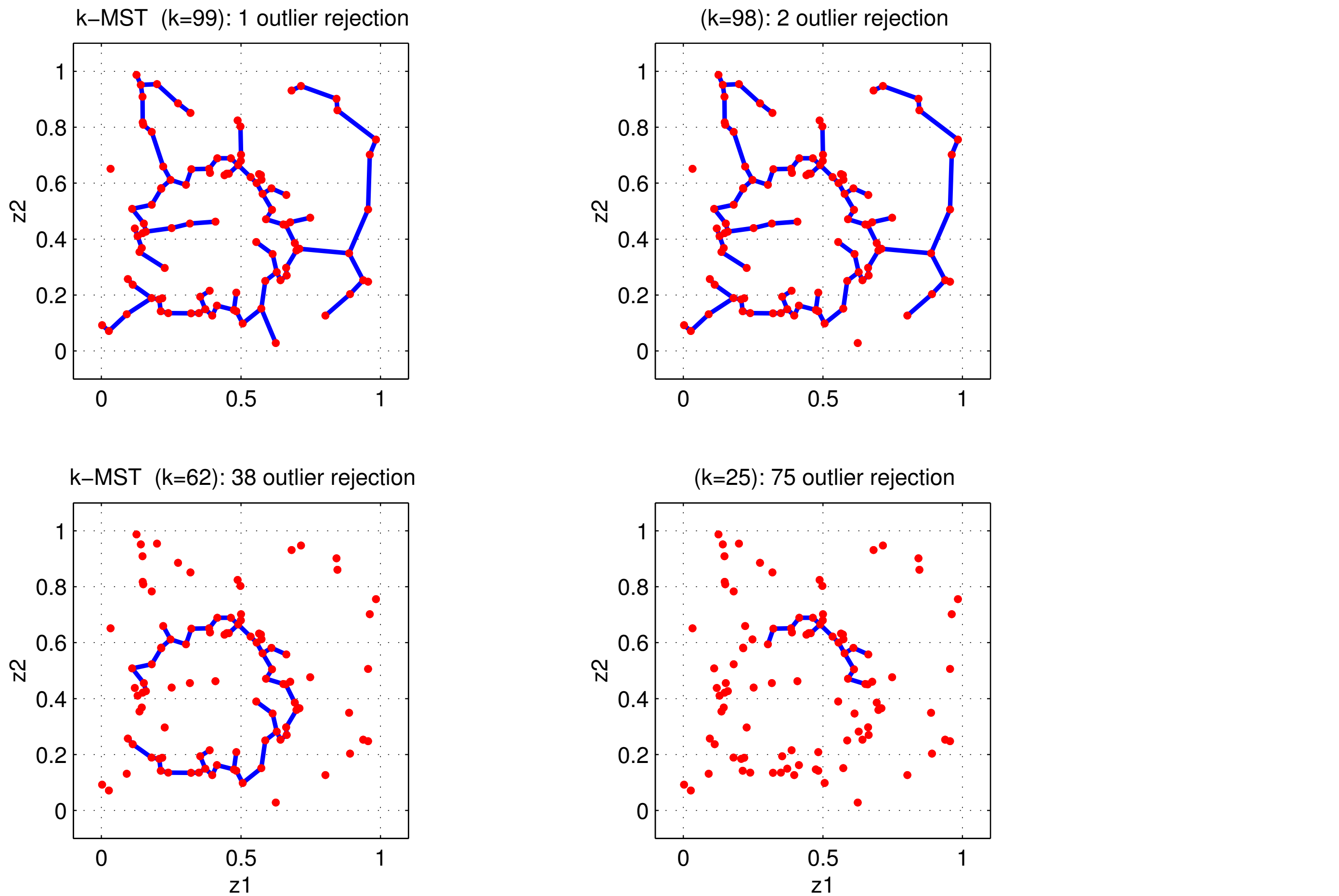


Figure 6: *Clustering an annulus density from uniform noise via k-MST.*

k-MST Stopping Rule (Hero&Michel:1997)

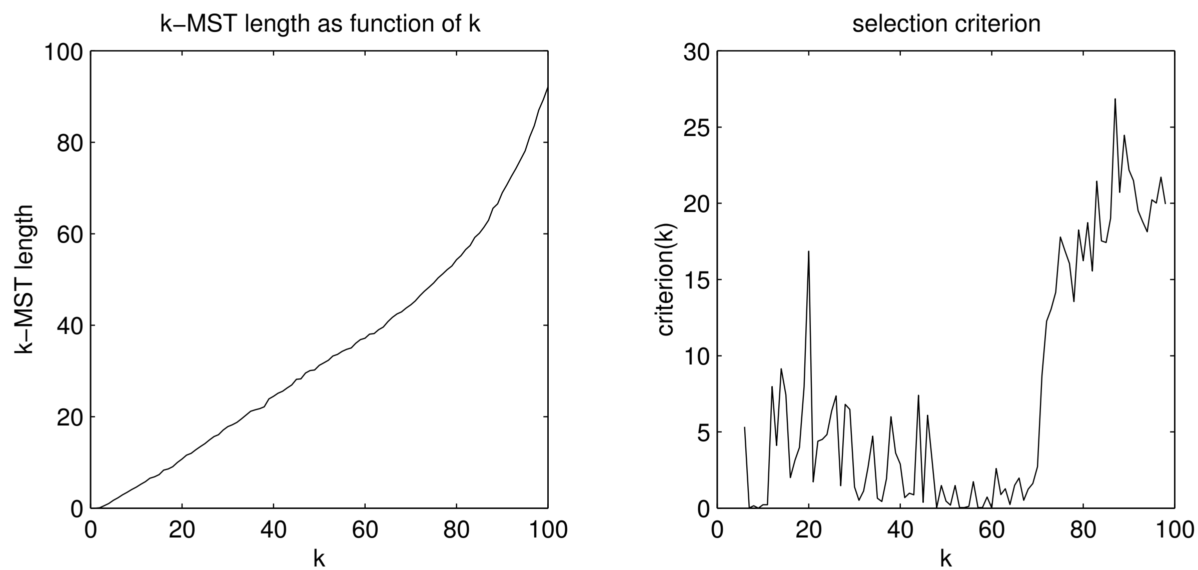


Figure 7: *Left: k-MST curve for 2D annulus density with addition of uniform "outliers" has a knee in the vicinity of* $n - k = 35$.

Greedy partioning approximation to k-MST (Ravi&etal:1996)

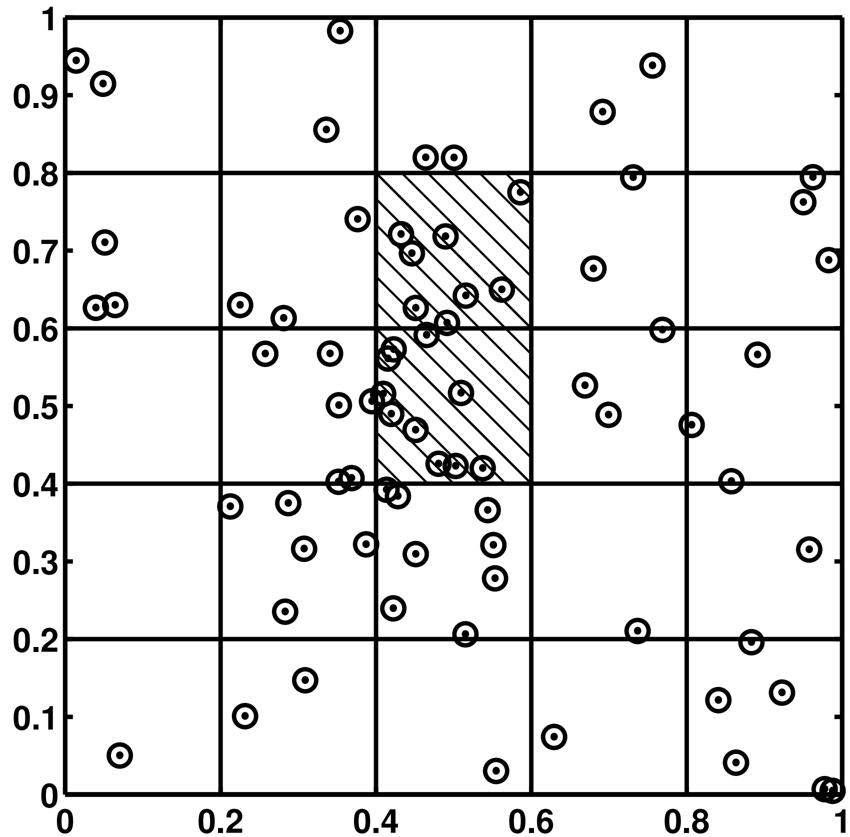


Figure 8: *A smallest subset B_k^m is the union of the two cross hatched cells shown for the case of $m = 5$ and $k = 17$.*

Extended BHH Theorem for Greedy k-MST (Hero&Michel:1999)

Fix $\rho \in [0,1]$. If $k/n \to \rho$ then the length of the greedy partitioning k-MST satisfies [Hero&Michel:IT99]

$$L_\gamma(\mathcal{X}^*_{n,k})/(\lfloor \rho n \rfloor)^\alpha \to \beta_{L_\gamma,d} \int_{\mathcal{S}} f^\alpha(x|x \in A_o)dx \qquad (a.s.)$$

where A_o is level set of f which satisfies $\int_{A_o} f = \rho$. Alternatively, with

$$H_\alpha(f|x \in A_o) = \frac{1}{1-\alpha} \ln \int_{\mathcal{S}} f^\alpha(x|x \in A_o)dx$$

$$\frac{1}{1-\alpha} \ln \left(L_\gamma(\mathcal{X}^*_{n,k})/(\lfloor \rho n \rfloor)^\alpha\right) \to \beta_{L_\gamma,d} H_\alpha(f|x \in A_o) + c \qquad (a.s.)$$

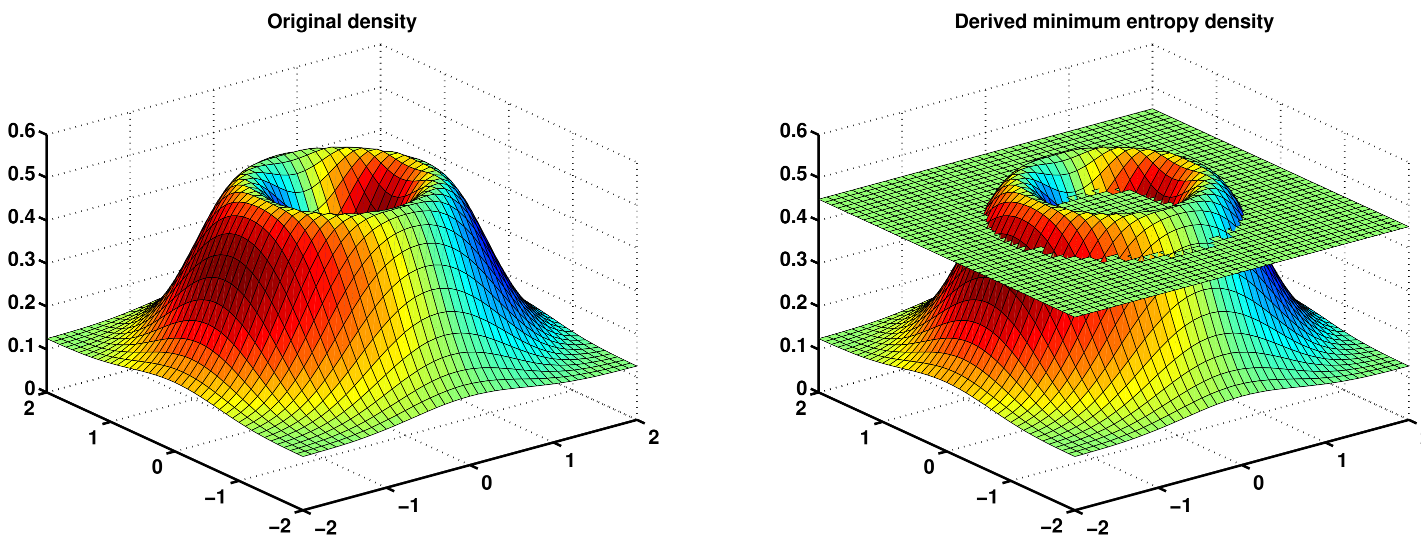


Figure 9: Waterpouring contruction of minimum entropy density.

k-MST Influence Function

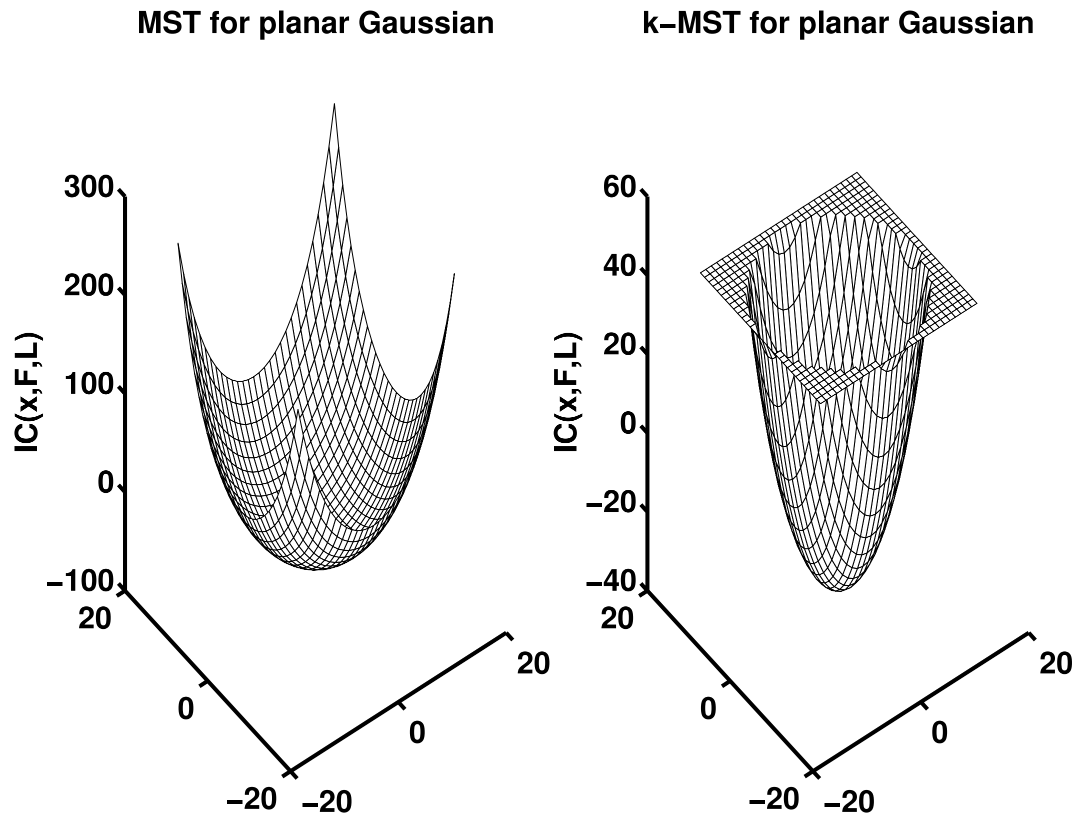
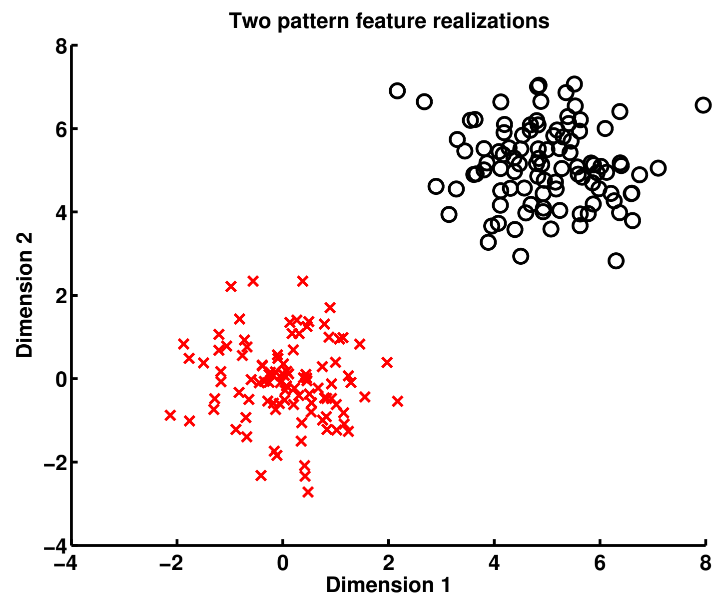


Figure 10: *MST and k-MST influence curves for Gaussian density on the plane.*

V. Entropic Graphs for Pattern Matching



Two groups of i.i.d. feature realizations on $[0,1]^d$:

- $\mathcal{X}_m = \{X_1, \ldots, X_m\}$, $X_i \sim f$
- $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$, $Y_i \sim g$
- $p = m/(m+n)$, $q = 1-p$

Objective: estimate separation of f and g using $\mathcal{X}_m$ and $\mathcal{Y}_n$

Some entropic graph estimation possibilities

Option 1. construct MST/k-NNG on pooled data $\mathcal{X}_m \cup \mathcal{Y}_n$ (Hero,Ma,Michel&Gorman:2001):

$$\ln L_\gamma(\mathcal{X}_m \cup \mathcal{Y}_n)/N^\alpha \to (1-\alpha)H_\alpha(pf+qg)+c, \quad (a.s.)$$

If subsequently subtract $\ln L(\mathcal{X}_m)/N^\alpha$ and $\ln L(\mathcal{Y}_n)/N^\alpha$ obtain estimator of α-Jensen difference (Basseville:1989,He&etal:2001)

$$\Delta(f,g) = H_\alpha(pf+qg) - pH_\alpha(f) - qH_\alpha(g)$$

Option 2: prune all single-class connections from pooled MST and compute normalized length

$$L_\gamma(\mathcal{X}_m \Delta \mathcal{Y}_n) = \frac{1}{N^\alpha}\sum_{e_{xy}} |e_{xy}|^\gamma$$

- for $\gamma = 0$ obtain ”Multivariate runs statistic” Friedman&Rafsky:1979 (FR).
- for $0 < \gamma < d$ obtain generalized FR statistic (Costa&Hero:2003)
- FR($\gamma = 0$) statistic converges a.s. to affinity (Henze&Penrose:1998)

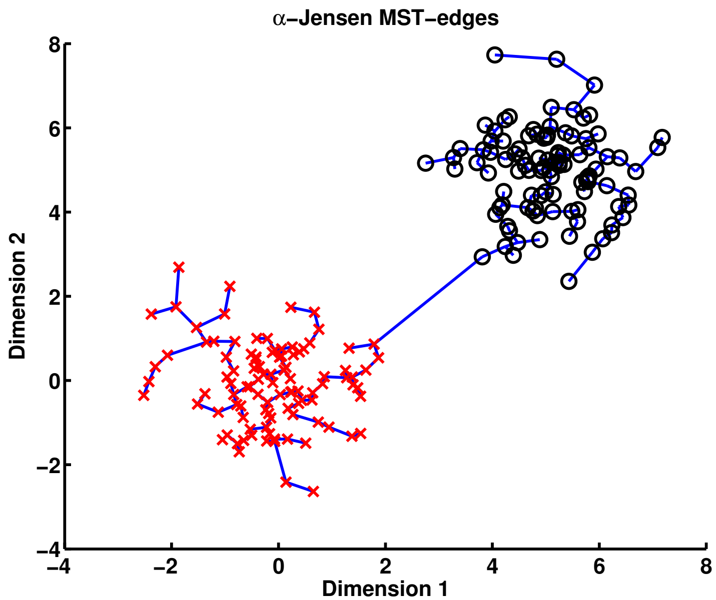
$$A_{FR}(f,g) = 2pq \int \frac{f(x)g(x)}{pf(x)+qg(x)} dx$$

This affinity is related to divergence measure:

$$D_{FR}(f\|g) = 1 - A_{FR}(f,g) = \int \frac{p^2 f^2(x) + q^2 g^2(x)}{pf(x)+qg(x)} dx$$

Option 3: implement entropic graph approximation of adaptive partition estimators of different divergence functionals (example below).

Illustration: Jensen Difference estimator



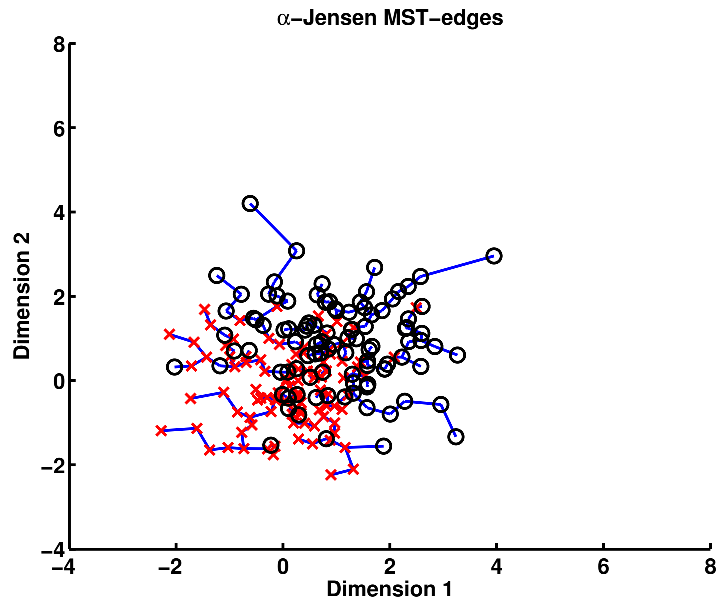
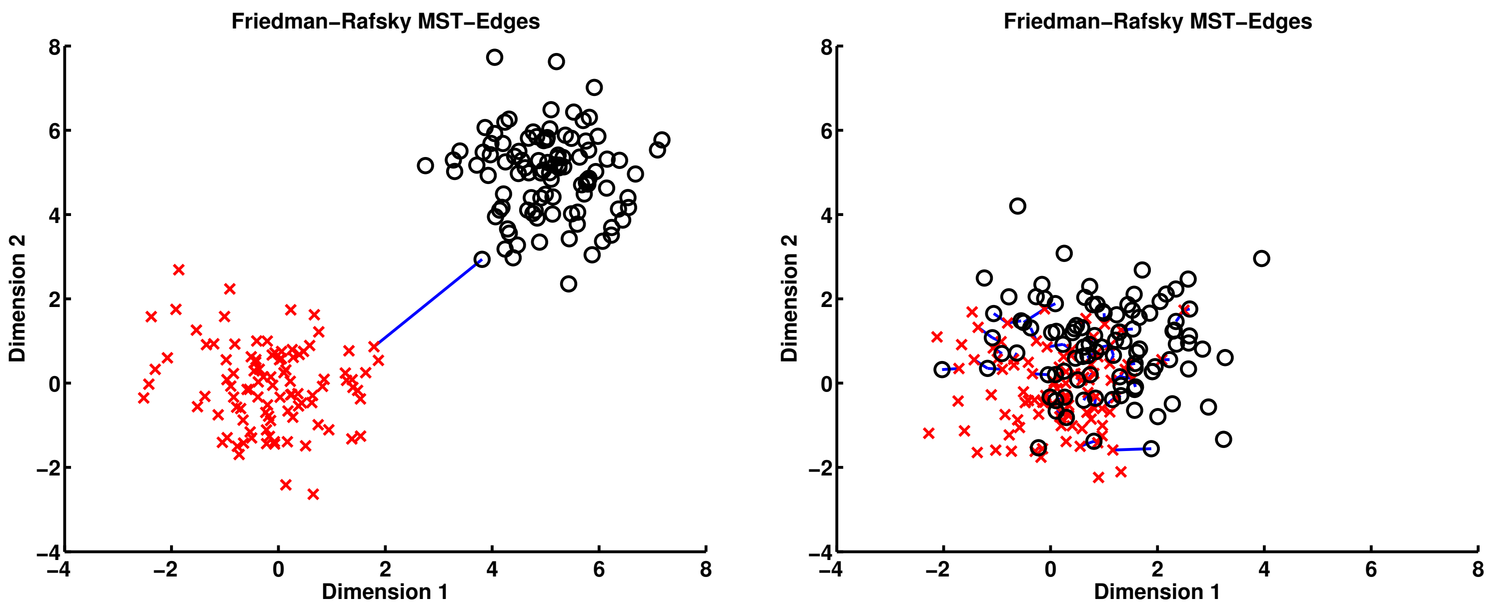


Illustration: Friedman-Rafsky Statistic



Entropic Graphs vs Adaptive-Partition Density Plug-in Estimates

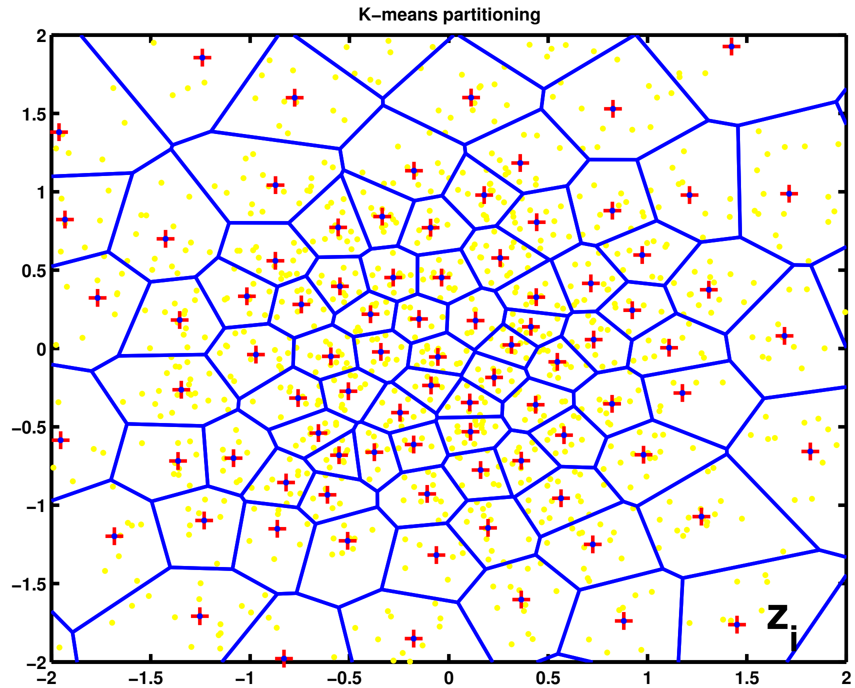
Define

$$I(f) = \int_S f^{\alpha}(x)dx$$

For N i.i.d. realizations $\{x_i\}_{i=1}^N$ from f define:

1. Π: an M-cell partition of $[0,1]^d$.
2. $\Pi(x)$: the cell in Π containing point $x \in [0,1]^d$
3. $\hat{f}_\Pi$: a partition estimator of f

$$\hat{f}_\Pi(x) = \frac{\mu(\Pi(x))}{\lambda(\Pi(x))}, \quad x \in [0,1]^d$$



For $\{z_i\}_{i=1}^N$ an i.i.d. realization **independent** of $\{x_i\}_{i=1}^N$ consider the α-entropy estimator

$$\hat{I}_\Pi \quad = \quad \frac{1}{N}\sum_{i=1}^N \hat{f}_\Pi^{\alpha-1}(z_i) = \frac{1}{N}\sum_{i=1}^N \left(\frac{\mu(\Pi(z_i))}{\lambda(\Pi(z_i))}\right)^{\alpha-1}$$

Under weak conditions on Π (Lugosi&Nobel:1996), this estimator converges a.s. to

$$E[f^{\alpha-1}(z_i)] = \int_S f^{\alpha}(x)dx = I(f)$$

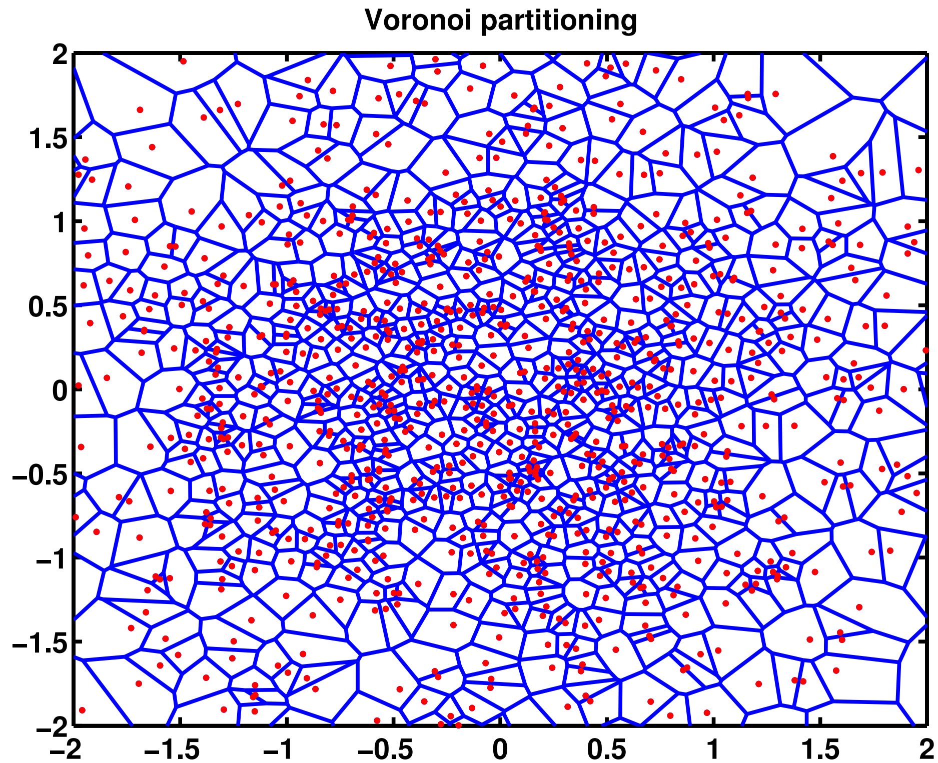
as $N \to \infty$. Equivalently,

$$\beta_{L_\gamma,d}\hat{I}_\Pi \to \beta_{L_\gamma,d}\int_S f^{\alpha}(x)dx$$

which corresponds to the a.s. limit of $L_\gamma(\mathcal{X}_N)/N^{\alpha}$.

To exploit this correspondence, (formally) specialize to:

1. $\alpha = (d-\gamma)/d$
2. Π is Voronoi partition ($\mu(\Pi(x)) \equiv 1$)
3. $z_i = x_i$, $i = 1, \ldots, N$



In this case we have:

$$\beta_{L_\gamma,d}\hat{I}_\Pi = \frac{\beta_{L_\gamma,d}}{N}\sum_i \left(\frac{1}{\lambda(\Pi(z_i))}\right)^{\alpha-1} = \frac{\beta_{L_\gamma,d}}{N}\sum_i \left(\lambda^{1/d}(\Pi(z_i))\right)^{\gamma}$$

Q. What relation between $\lambda^{1/d}(\Pi(z_i))$ and e_i would make $\beta_{L_\gamma,d}\hat{I}_\Pi$ equal to $L_\gamma(\mathcal{X}_N)/N^\alpha$?

A. When

$$\frac{\beta_{L_\gamma,d}}{N}\sum_i \left(\lambda^{1/d}(\Pi(z_i))\right)^\gamma \quad = \quad \frac{1}{N^\alpha}\sum_i e_i^\gamma$$

which occurs if we identify

$$\lambda^{1/d}(\Pi(z_i)) = \frac{n^{1/d}}{\beta_{L_\gamma,d}^{1/\gamma}}\, e_i \tag{1}$$

Heuristic: can use formal relation (1) to obtain entropic graph implementations of divergence estimators.

Example: Geometric-Arithmetic (GA) Affinity (Taneja:2001)

$$A(f,g) \quad = \quad \int (pf(x)+qg(x))^{\alpha} \left(f^{p}(x)g^{q}(x)\right)^{1-\alpha} dx$$

1. Pooled sample $\mathcal{Z}_{m+n} = \mathcal{X}_m \cup \mathcal{Y}_m$ has density $h = pf + qg$
2. Adaptive-partition plug-in estimator of $A(f,g)$ is

$$\hat{A}_{ap} = \frac{1}{N} \sum_{z_i=1}^{N} \left(\frac{\hat{f}^{p}(z_i)\hat{g}^{q}(z_i)}{\hat{h}(z_i)} \right)^{1-\alpha} \to A(f,g) \quad (a.s.)$$

3. Specialize partition to Voronoi and substitute (1):

$$\hat{A}_{eg} = \frac{1}{N} \underbrace{\sum_i \min\left\{ \left(\frac{e_i(\mathcal{Y}_n)}{e_i(\mathcal{X}_m)} \right)^{p\gamma}, \left(\frac{e_i(\mathcal{X}_m)}{e_i(\mathcal{Y}_n)} \right)^{q\gamma} \right\}}_{R_\gamma(\mathcal{X}_m \cup \mathcal{Y}_n)}$$

Planar Pattern Matching Simulation

- $\mathcal{X}_m$ realization from $\mathcal{N}_2(\underline{0}, \mathbf{I})$
- $\mathcal{Y}_n$ realization from $\mathcal{N}_2(\underline{D}, \mathbf{I})$
- Four pattern separation measures Δ investigated

$$\Delta = \begin{array}{cc} L_\gamma(\mathcal{X}_m \cup \mathcal{Y}_n)/N^\alpha, & L_0(\mathcal{X}_m \Delta \mathcal{Y}_n)/N \\ L_\gamma(\mathcal{X}_m \Delta \mathcal{Y}_n)/N^\alpha, & R_\gamma(\mathcal{X}_m \cup \mathcal{Y}_n)/N \end{array}$$

- $\gamma = 1$, $\alpha = 1/2$
- Local resolution measure

$$\rho(\Delta) = \frac{|E[\Delta|D=0] - E[\Delta|D=1]|}{\sqrt{\sigma_\Delta^2(D=0) + \sigma_\Delta^2(D=1)}}$$

Pattern Matching Simulation: Convergence

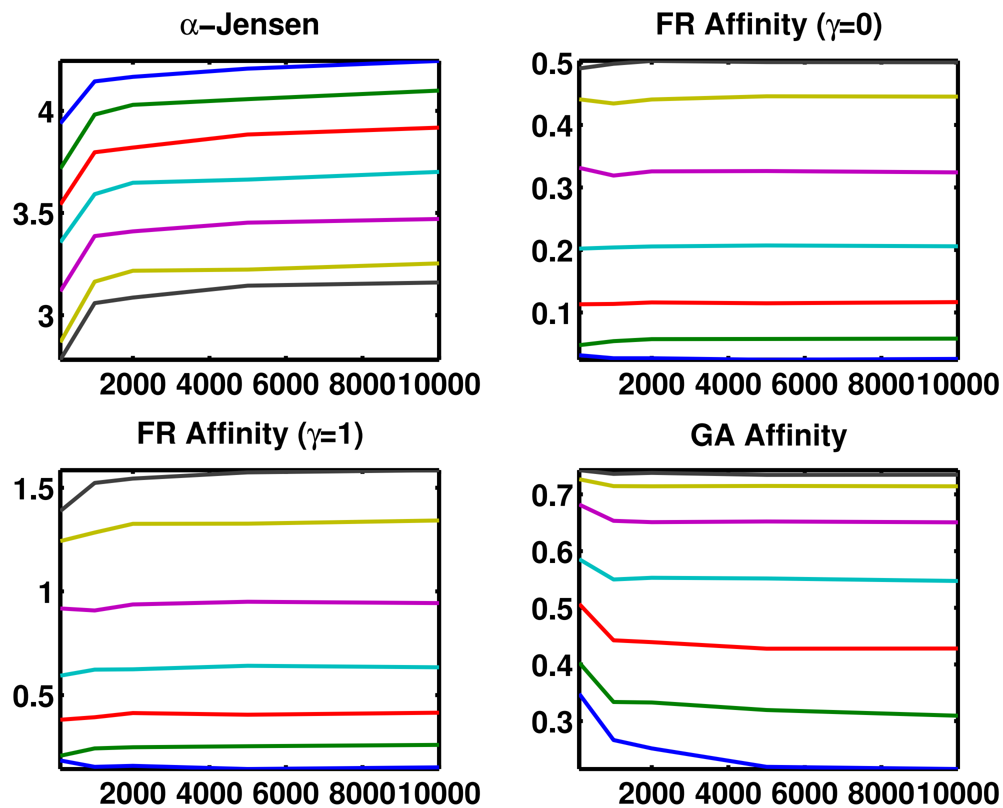
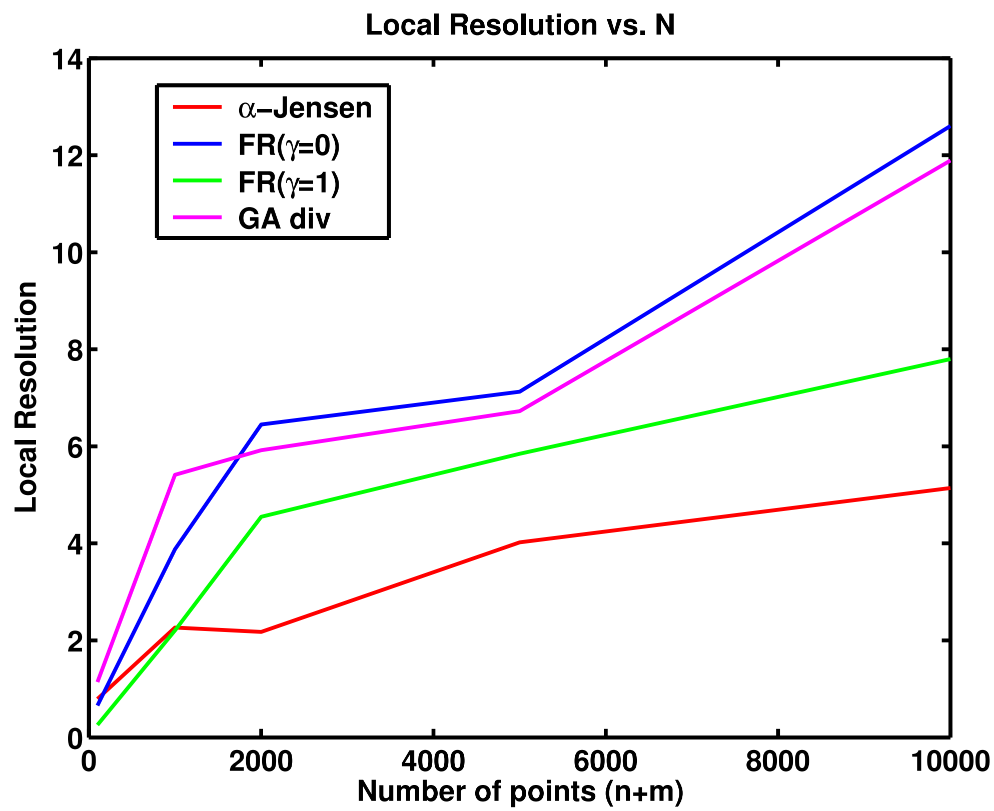


Figure 11: Convergence rates $1/\sqrt{N}$ (left) and $1/N$ (right)

Pattern Matching Simulation: Local Resolution



Conclusions

1. Entropic graphs can be used to estimate α-entropy and α-divergence
2. These methods can be applied to high dimensional feature-spaces
3. Clustering can be performed using entropic K-point graphs
4. Asymptotic theory can be used to motivate new entropic graph measures

References

1. ”Robust entropy estimation strategies based on edge weighted random graphs,” A.O. Hero and O. Michel, Proc. of Int. Soc. for Optical Engineering (SPIE) Symposium on Optical Science, San Diego, July 1998.
2. A. O. Hero, B. Ma, O. Michel and J. Gorman, ”Applications of entropic spanning graphs,” IEEE Signal Proc. Magazine (Special Issue on Mathematics in Imaging), Vol 19, No. 5, pp 85-95, Sept. 2002.
3. “Image registration using alpha-entropy measures and entropic graphs,” H. Neemuchwala, A. O. Hero and P. L. Carson, European Journal of Signal processing, to appear Sept. 2003.
4. “Parametric and non-parametric approaches for multisensor data fusion,” Bing Ma, PhD thesis in Dept. EECS, Univ. of Michigan, Jan. 2001.
5. “Asymptotic theory of greedy approximations to minimal K-point random graphs,” A. O. Hero and O. Michel, IEEE Trans. on Information Theory, Vol. IT-45, pp. 1921-1939, Sept. 1999.

6. ”Asymptotic rates of convergence of random minimal graphs,” A. O. Hero, J. Costa and B. Ma, IEEE Trans. on Inform. Theory, Submitted Aug. 2001.
7. O. Michel and A. O. Hero ”Entropic graph applications”, Proc. XI European Signal Processing Conference, Toulouse France, Sept 2002.
8. “Estimation of Rényi Information Divergence via Pruned Minimal Spanning Trees,” A.O. Hero and O. Michel, Proc. of 1999 IEEE Workshop on Higher Order Statistics, Caesaria Israel, June 1999.

Extension of BHH to Divergence Estimation?

Question: How to generalize entropic graph estimates of

$$\frac{1}{1-\alpha}\ln\int f^{\alpha}(x)dx \quad \textbf{to} \quad \frac{1}{\alpha-1}\ln\int f^{\alpha}(x)g^{1-\alpha}(x)dx \ ?$$

One possibility:

- $g(x)$: a **known** reference density on $[0,1]^d$
- Assume $f \ll g$, i.e. for all x such that $g(x)=0$ we have $f(x)=0$.
- Make measure transformation $M(x)$ such that $dx \rightarrow g(x)dx$ on $[0,1]^d$. Then for $\mathcal{Y}_n = M(\mathcal{X}_n)$

$$L_\gamma(\mathcal{Y}_n)/n^{\alpha} \quad \rightarrow \quad \beta_{L_\gamma,d}\int\left(\frac{f(x)}{g(x)}\right)^{\alpha} g(x)dx, \qquad (a.s.)$$

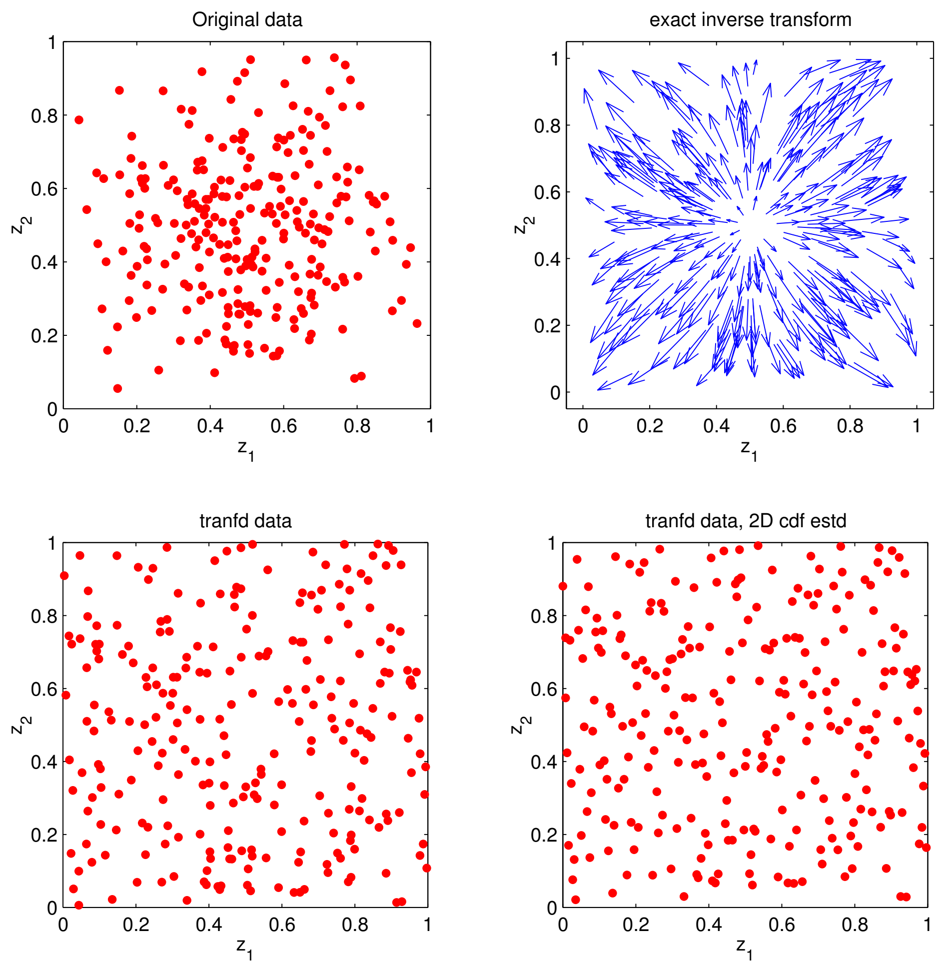


Figure 12: Top Left: i.i.d. sample from triangular distribution, Top Right: exact transformation, Bottom: after application of exact and empirical transformations.

What is the entropic graph's convergence rate?

Theorem 2 (Hero,Costa&Ma:2001) *Let* $d \geq 2$ *and* $1 \leq \gamma \leq d-1$. *Assume* $X_1, \ldots, X_n$ *are i.i.d. random vectors over* $[0,1]^d$ *with density* $f \in \Sigma_d(\beta, l)$, $\beta, l > 0$, *having support* $\mathcal{S} \subset [0,1]^d$. *Assume also that* $f^{\frac{1}{2}-\frac{\gamma}{d}}$ *is integrable. Then,*

$$O\left(n^{-r_1(d,\beta)}\right) \leq \sup_{f \in \Sigma_d(\beta,l)} E\left[\left|L_\gamma(X_1,\ldots,X_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)\mathrm{d}x\right|^p\right]^{1/p} \leq O\left(n^{-r_2(d,\beta)}\right),$$

where

$$r_1(d,\beta) = \min\{\frac{4\beta}{4\beta+d}, 1/2\} \quad r_2(d,\beta) = \frac{\alpha\beta}{\alpha\beta+1}\ \frac{1}{d}$$

and $\alpha = \frac{d-\gamma}{d}$.

Extension to Partition Approximations

$$L_\gamma^m(\mathcal{X}_n) = \sum_{i=1}^{m^d} L_\gamma(\mathcal{X}_n \cap Q_i) + b(m),$$

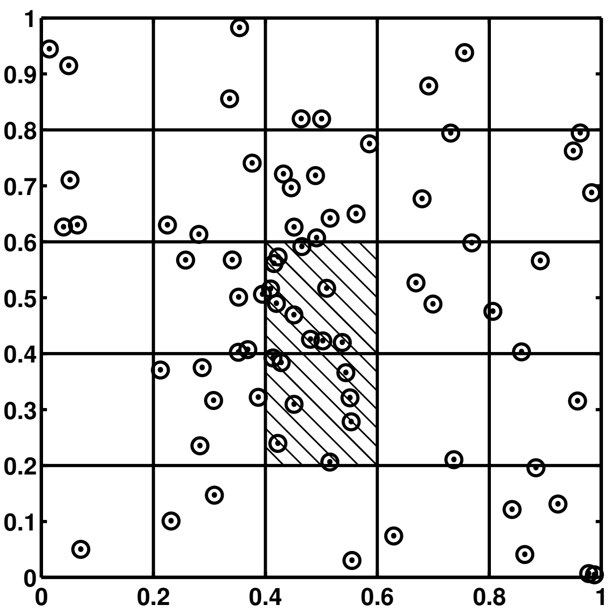


Figure 13: *Partition approximation.*

Theorem 3 (Hero,Costa&Ma:2001) *Let $L_\gamma^m(\mathcal{X}_n)$ be a partition approximation to $L_\gamma(\mathcal{X}_n)$. Under the same hypotheses as in the previous proposition, if* $b(m) = O(m^{d-\gamma})$

$$O\left(n^{-r_1(d,\beta)}\right) \quad \leq$$

$$\sup_{f \in \Sigma_d(\beta,l)} E\left[\left|L_\gamma^{m(n)}(X_1,\ldots,X_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d}\int_S f^{(d-\gamma)/d}(x)\mathrm{d}x\right|^p\right]^{1/p}$$

$$\leq O\left(n^{-r_3(d,\beta)}\right),$$

where

$$r_3(d,\beta) = \frac{\alpha\beta}{\frac{d-1}{\gamma}\ \alpha\beta + 1}\ \frac{1}{d}\ .$$

This bound is attained by choosing the progressive-resolution sequence $m = m(n) = n^{1/[d(\frac{d-1}{\gamma}\ \alpha\beta+1)]}$.