Complexity, Information and Geometry (Module 2)

Peyresque

Alfred Hero

Digiteo and University of Michigan

July, 2008

Outline of Module 2

1. Non-explicit entropy estimation
2. Random geometric graphs
3. Convergence theorem
4. Convergence rates
5. BHH theorem extensions
   - Lower dimensional manifolds
   - Pruned MST
6. Applications
   - Anomaly detection
   - Dimension estimation

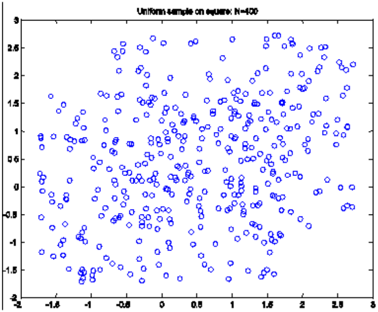
Non-explicit entropy estimation

Examples of entropy estimation methods that do not use explicit density plug-in

- Data compression (LZ, CWT) entropy estimators (Kontoyanis 1998)
- kNN estimators (Leonenko 2008) [10]
- Entropic graph estimators (Hero 1998) [9]

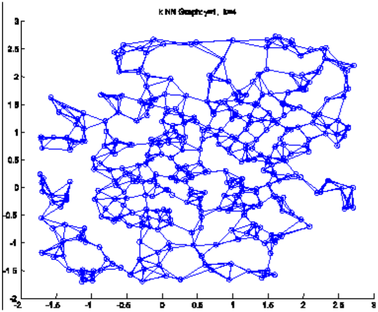
Random Euclidean graph

Uniformly distributed points in plane



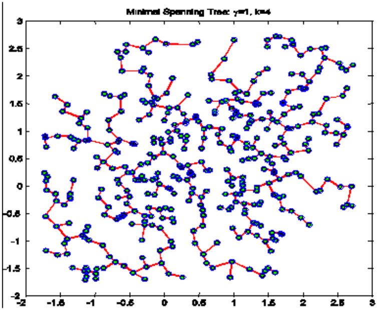
Random Euclidean graph

kNNG on uniform points in plane



Random Euclidean graph

MST on uniform points in plane



Geometric graph

A graph $\mathcal{G}$ consists of vertices $\mathcal{V}$ and edges $\mathcal{E}$ between pairs of vertices

For a geometric graph

- $\mathcal{V}$ is subset of $\mathcal{X}_n = \{x_i\}_{i=1}^n$: n points in $\mathbb{R}^d$
- Edges $e = e_{ij}$ in $\mathcal{E}$ have length related to distances between pairs x_i, x_j

A geometric graph has edge lengths $|e|$ that are constrained:

- If there is an edge between x_i and x_j then $e_{ij} = e_{ji}$, edges are undirected
- If there are edges connecting x_i, x_j and x_j, x_k then $|e_{ik}| \leq |e_{ij}| + |e_{jk}|$, edges satisfy triangle inequality

Geometric graph

The total weight or length of $\mathcal{G}$ is the (weighted) sum of its edge lengths

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

where ψ is a monotonic increasing function over $\mathbb{R}$ with $\psi(0) = 0$.

When $\psi(e) = e$ and $e = e_{ij} = \|x_i - x_j\|$ $\mathcal{G}$ is a Euclidean graph

When $\mathcal{V}$ are random points in $\mathbb{R}^d$ $\mathcal{G}$ is a random graph

Geometric graph

A path through a graph is a connected sequence of edges $e_{1,2}, e_{2,3}, \ldots, e_{p,p-1}$

A cycle exists in a graph if there exists a closed path $e_{1,2}, e_{2,3}, \ldots, e_{p,1}$

An acyclic graph $\mathcal{G}$ is a tree $\mathcal{T}$

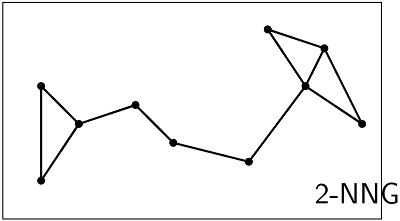
A graph $\mathcal{G}$ spans the points $\mathcal{X}_n$ if there exists an edge connecting every point in $\mathcal{X}_n$

KNN graph with power weighted edges

Let $\mathcal{N}_{k,i}(\mathcal{X}_n)$ denote the possible sets of k edges connecting point $\mathbf{x}_i$ 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 with power weighted edges

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

The Euclidean Power Weighted MST minimizes total length among spanning trees

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

Random graphs and statistical estimation

Some previous statistical uses of random graphs

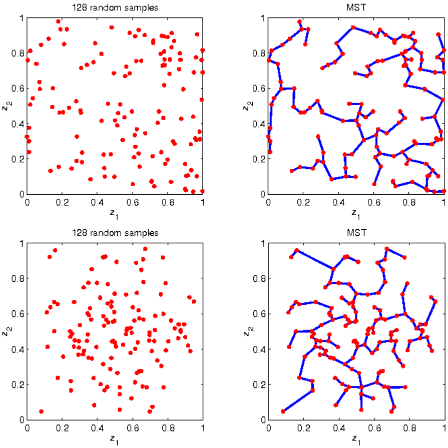
- Clustering: Zahn (1971), Toussaint (1980)
- Invariant pattern recognition: Duda&Hart (1973)
- Two sample matching: Friedman&Rafsky (1979)
- Testing for randomness: Hoffman&Jain (1983)
- Non-parametric regression: Banks (1993)

Random graph properties

- If random graph satisfies some minimality properties there are asymptotic results on (Penrose03,Yukich98) [12],[14].
  - Average length of edges
  - Average length of monotone functionals of edges
  - Connectivity and number of components
  - The length of maximal length edge
- These results require smoothness conditions on the graph construction and the underlying density
  - Density is non-singular wrt Lebesgue measure
  - Density is bounded (lower and upper) over its support set
  - Graphs are determined by quasi-additive continuous Euclidean functionals

MST and entropy

MST for uniform and triangular densities



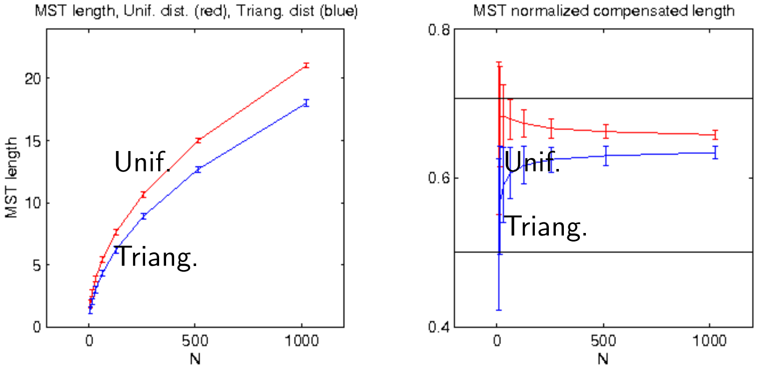


Figure: MST and log MST total weight as function of the number of samples.

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 the weight of the MST satisfies*

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

This extends to kNN, TSP, Steiner tree, minimal matching graph

Yukich's version of the Beardwood, Halton, Hammersley (1959) Theorem [14]

Let $\{X_i\}_{i=1}^n$ be an i.i.d sequence of random variables with Lebesgue p.d.f. $f(x)$ over $[0,1]^d$, $d > \gamma > 0$. If L_n is a quasi-additive continuous Eucildean functional then

$$L_n/n^{(d-\gamma)/d} \rightarrow \beta_{L,\gamma} \int_{\mathbb{R}^d} f^{(d-\gamma)/d}(x)dx \qquad (w.p.1)$$

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

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

Convergence rate

Question: What is r.m.s. rate of convergence?

Find constant r such that

$$E^{1/2}\left[\left|L_\gamma(\mathcal{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int f(x)^{(d-\gamma)/d} dx\right|^2\right] \leq O(n^{-r})$$

(**Thm 5.2 Yukich:1998**): Let L_γ be a quasi-additive continuous Euclidean functional which satisfies the add-one bound. Assume that $f(\mathbf{x})$ is uniform over $[0,1]^d$. Then for all $d \geq 2$ and $1 \leq \gamma < d$

$$\left| E[L_\gamma(\mathcal{X}_n)]/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int f(x)^{(d-\gamma)/d} dx \right| \leq O(n^{-1/d})$$

Convergence rate

Question: How to extend to non-uniform f?

1. Extend to piecewise constant "block densities" over a uniform partition $\mathcal{Q}^m$:

$$f(\mathbf{x}) = \sum_{i=1}^{m^d} \phi_i 1_{Q_i}(\mathbf{x})$$

2. Extend to space of densities sufficiently well approximated by block densities.

3. Obtain worst-case bound on rate over this space of densities.

The Hölder space of smooth functions on $\mathbb{R}^d$ is

$$\Sigma_d(\beta, L) = \left\{ g : |g(\mathbf{z}) - p_{\mathbf{x}}^{\lfloor \beta \rfloor}(\mathbf{z})| \leq L|\mathbf{z} - \mathbf{x}|^{\beta}, \mathbf{x}, \mathbf{z} \in \mathbb{R}^d \right\}.$$

• $p_{\mathbf{x}}^{k}(\mathbf{z})$ is the Taylor polynomial of g or order k expanded about the point $\mathbf{z} = \mathbf{x}$.

• $\Sigma_d(\beta, L)$ is set of Lipschitz functions with Lipschitz constant L and it contains increasingly smooth functions as β increases.

Convergence rate in BHH theorem

Costa Thesis, 2005

Corollary 13. *Let $d \geq 2$ and $1 \leq \gamma \leq d-1$. Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. random vectors with density $f \in \mathcal{F}_{\beta,L}$, $\beta \in (0,1]$. Assume also that $f^{\frac{1}{2}-\frac{\gamma}{d}}$ is integrable. Then, for any continuous quasi-additive Euclidean functional L_γ of order γ that satisfies the add-one bound (2.8), there exist positive constants c, C, depending on β, L, d and γ such that for n sufficiently large*

$$c\, n^{-\left(\frac{4\beta}{4\beta+d}\right)} \leq \sup_{f \in \mathcal{F}_{\beta,L}} \left[E \left| L_\gamma(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)/n^{(d-\gamma)/d} - \beta_{L_\gamma,d} \int_S f^{(d-\gamma)/d}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \right|^p \right]^{1/p} \leq C n^{-r_1(d,\gamma,\beta)}\,, \qquad (2.49)$$

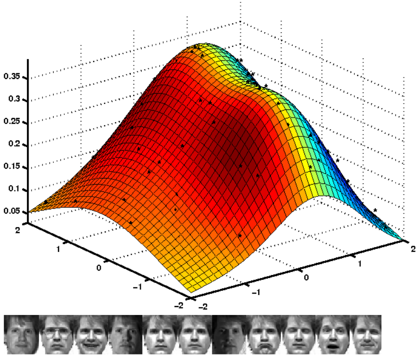
$$r_1(d,\gamma,\beta) = \frac{\alpha\,\beta}{\alpha\,\beta+1}\,\frac{1}{d}$$

Lower bound is minimax bound that is generally not attainable

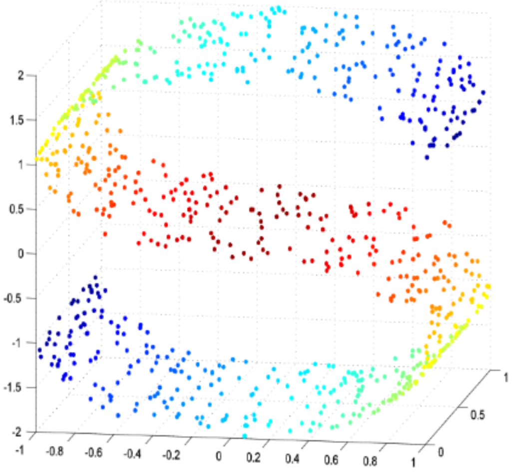
Density plug-in estimator attains a bound of order $n^{\frac{\beta}{2\beta+d}}$ which is strictly greater than entropic graph estimator upper bound for certain values of d and β.

Convergence result can be extended to non-differentiable functions by considering Sobolev spaces [6],[3]

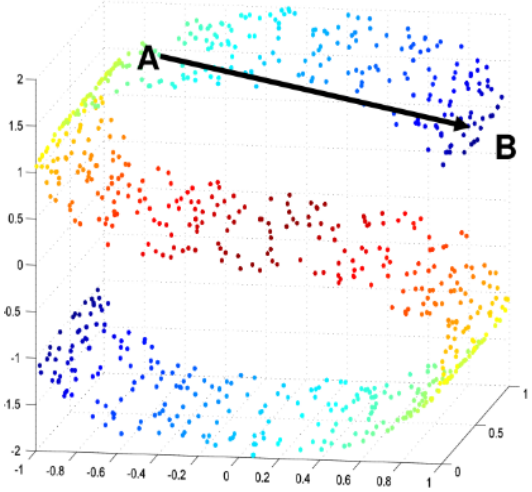
Convergence result can also be extended to greedy partitioningapproximations to any quasi-additive continuous Euclidean minimal graph [6], [3] For many natural images and signals the variability might be constrained to a surface of dimension $m < d$



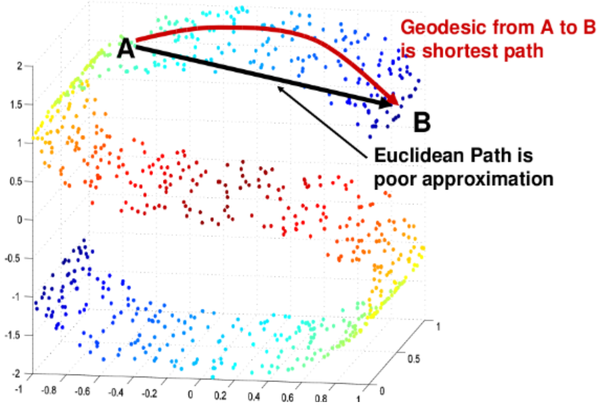
S-curve example



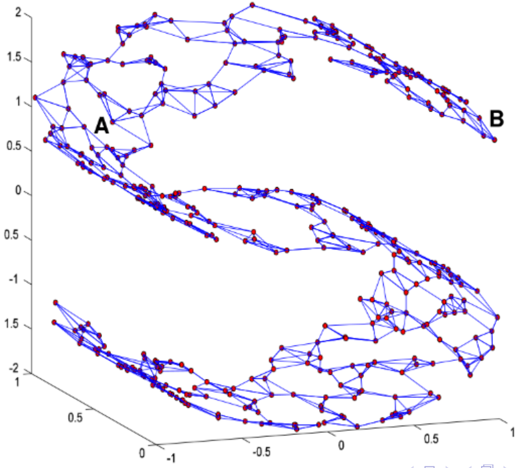
Euclidean shortest path between points A and B



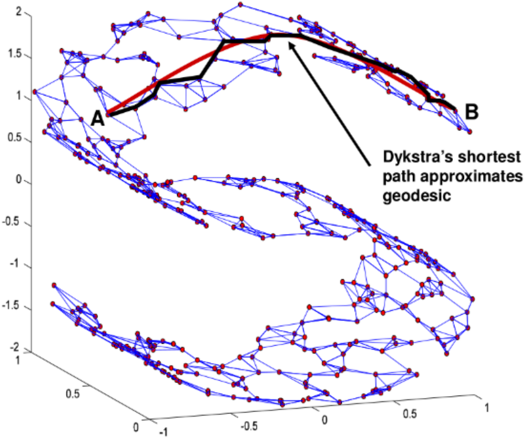
Euclidean path vs geodesic minimum distance path



KNN Graph (k=4)



KNN Graph (k=4)



Theorem: (Costa and Hero [4],[5]) Let $(\mathcal{M}, g)$ be a compact smooth Riemann m-dimensional manifold. Suppose $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ is a random sample on $\mathcal{M}$ with bounded density f relative to μ_g. Let L_γ be the total length of the MST graph or the kNN graph with lengths computed using the geodesic distance d_g. Assume $m \geq 2$, $1 \leq \gamma < m$, and define $\alpha = (m - \gamma)/m$. Then

$$\lim_{n\to\infty} \frac{L_\gamma(\mathcal{X}_n)}{n^\alpha} = \beta_{m,L_\gamma} \int_{\mathcal{M}} f^\alpha(x) d\mu_g(dx)$$

where β_{m,L_γ} is a constant independent of f and $\mathcal{M}$. Furthermore, the mean $E[L_\gamma(\mathcal{X}_n)]/n^\alpha$ converges to the same limit.

Assume f is a mixture density of the form

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

where

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

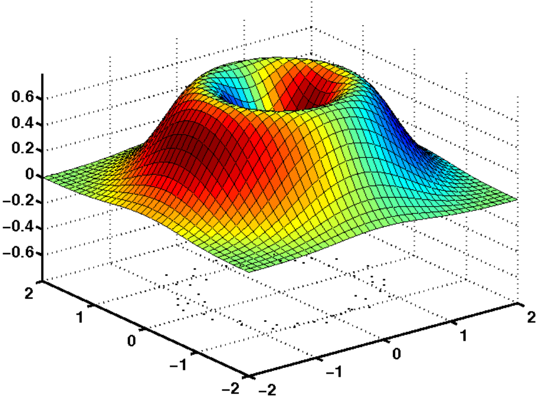
Objective: given realization $\mathcal{X}_n$ from f cluster the realizations from f_1.

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

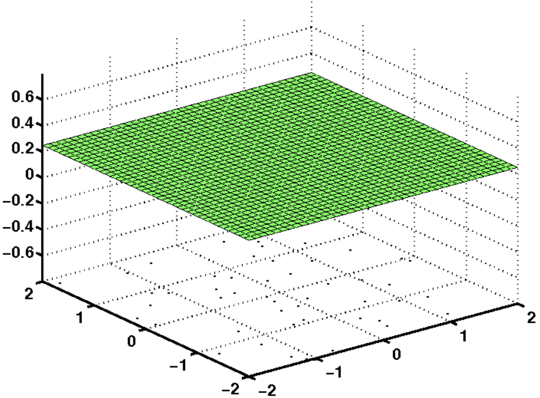
BHH theorem extensions

Example: Annulus Target Density f_1



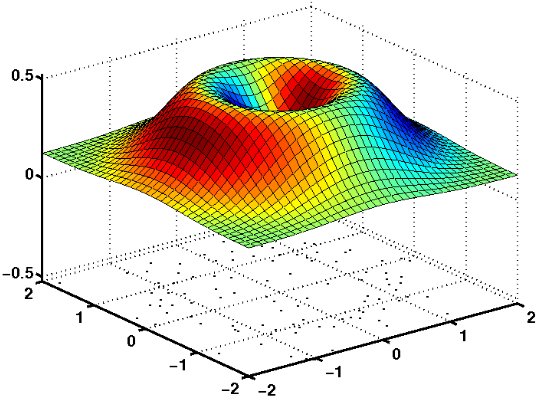
BHH theorem extensions

Uniform Outlier Density f_o



BHH theorem extensions

Mixture Density



BHH theorem extensions

k-point Minimal Spanning Tree (*k*-MST)

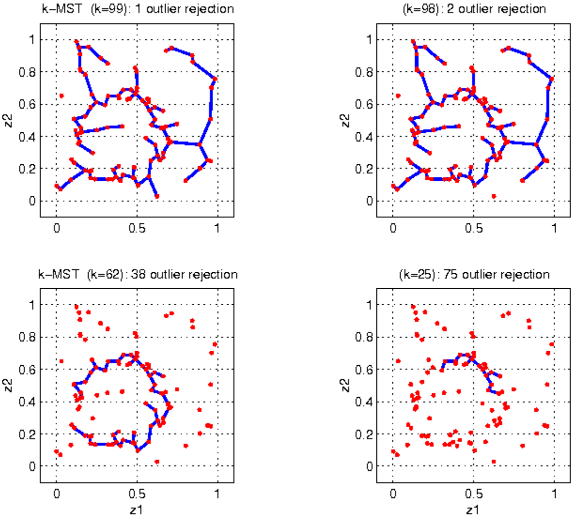


Figure: *Clustering an annulus density from uniform noise via k-MST*

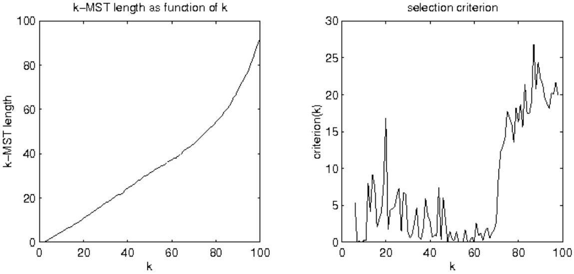


Figure: *Left: k-MST curve for 2D annulus density with addition of uniform "outliers" has a knee in the vicinity of* $n - k = 35$.

Ravi and 1996 proposed greedy partitioning approach to k-MST

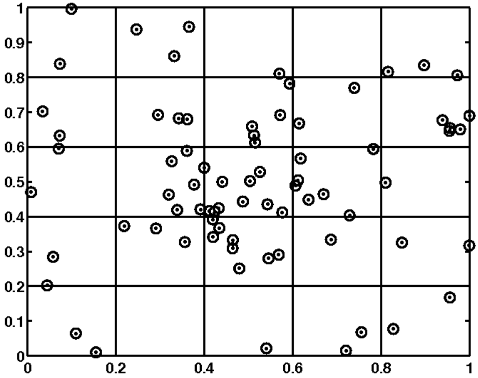


Figure: *The case of $m = 5$ and $k = 17$.*

Thm: Fix $\rho \in [0,1]$. If $k/n \to \rho$ then the length of the greedy partitioning k-MST satisfies (Hero and Michel [9])

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

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

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

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

BHH theorem extensions

Waterpouring solution=Level set of density

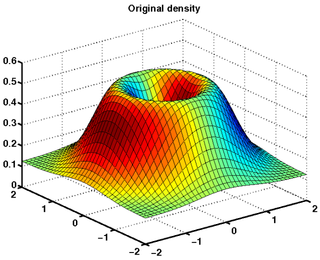


Figure: Waterpouring contruction of minimum entropy density.

Note: $P(X \in A_0) = \rho$

Consider testing hypotheses on $f(x) = (1-\epsilon)f_0(x) + \epsilon U(x)$

$$\begin{aligned} H_0 &: \epsilon = 0 \\ H_1 &: \epsilon > 0 \end{aligned}$$

based on a sample $\mathbf{X} = [X_1, \ldots, X_n]$,$X_i \in [0,1]^d$ and $\epsilon \in [0,1]$.

When f_0 and $U(x)$ are known, most powerful test of level $\alpha = 1-\rho$ is LRT

$$\Lambda(\mathbf{X}) = \frac{f(\mathbf{X}|H_1)}{f(\mathbf{X}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta$$

where η is a threshold chosen to satisfy $P(\Lambda(\mathbf{X}) > \eta | H_0) = 1-\rho$

If $U(x)$ is uniform density then

$$\Lambda(\mathbf{X}) > 0 \text{ iff } f_0(\mathbf{X}) > \gamma = \frac{\eta - \epsilon}{1 - \epsilon}$$

which is equivalent to

Definitions (Level set test)

`Decide` H_1 `if` $\mathbf{X} \notin A_0$
where A_0 is the level set satisfying $\int_{A_o} f_0(x)dx = 1 - \rho$.

Note: The decision region of the most powerful test does not depend on ϵ

$\Rightarrow$ test is **uniformly most powerful** over ϵ

For unknown f_0 the level set test can be implemented using K-MST

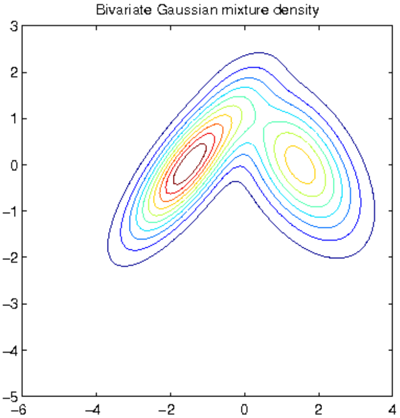


Figure: Bivariate mixture of Gaussians density

Anomaly detection

Greedy K-MST test example

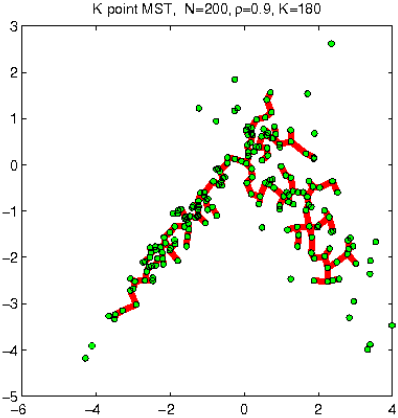


Figure: K-MST over a training realization from MoG

Anomaly detection

Greedy K-MST test example

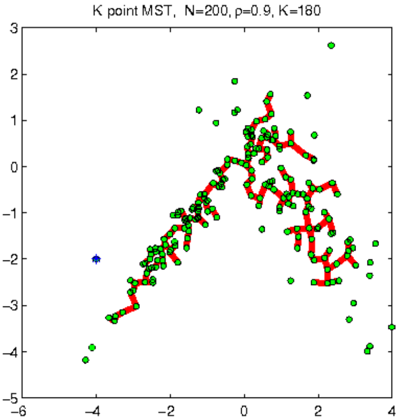


Figure: K-MST fails to capture new point (blue asterisk is outlier)

Anomaly detection

Greedy K-MST test example

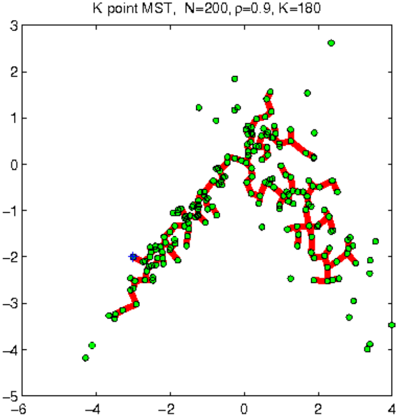


Figure: K-MST capture new point (blue asterisk is inlier)

Anomaly detection

Greedy K-MST test example

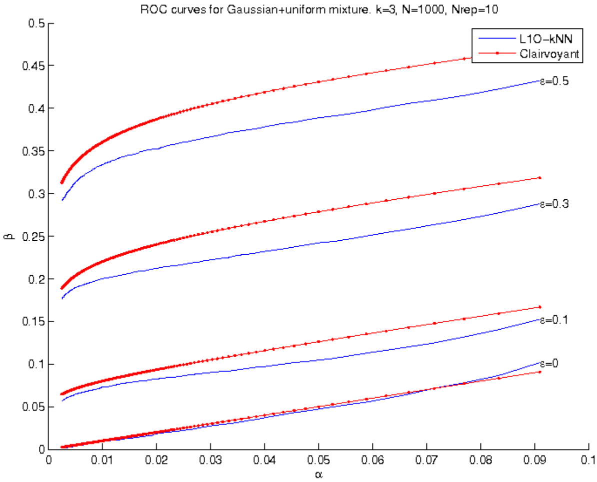


Figure: ROC curves for L1O-kNNG approximation are close to UMP curves for Gaussian example

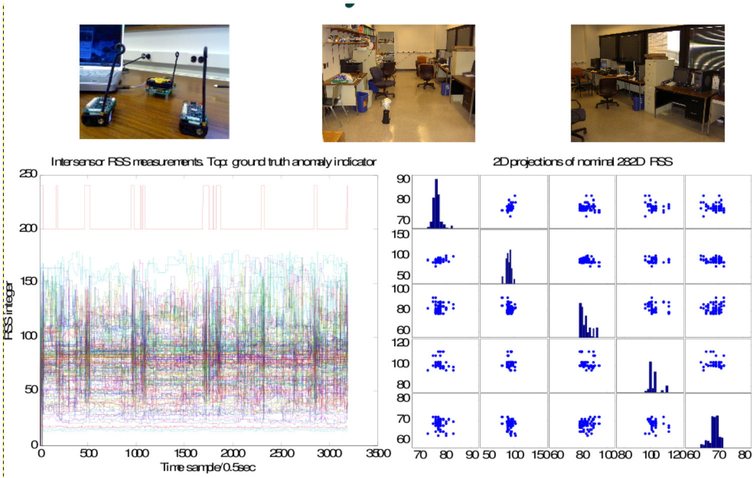


Figure: Hero [7]

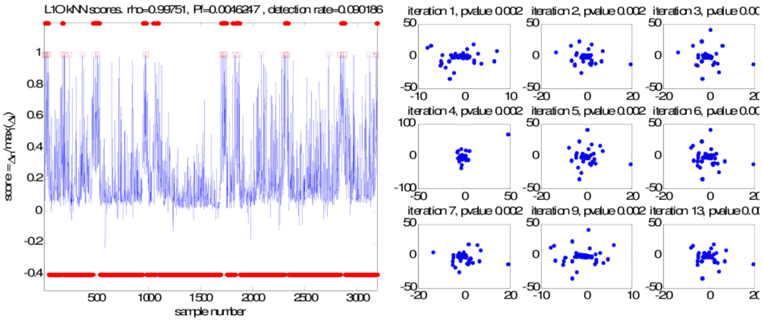
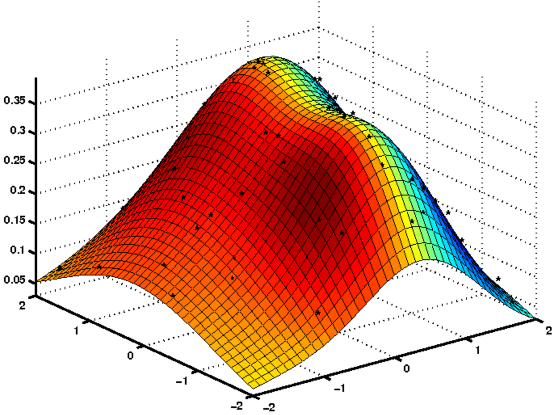


Figure: Online activity detector statistic (Left) some anomalies detected (right)

Application: Dimension estimation

Support set of unknown density $f(x)$ with realizations $\mathbf{X} = X_1, \ldots, X_n$



Question: what is dimension of the support set?

Dimension estimation

Recall form of the Costa's version of the BHH Theorem for $X \in \mathbb{R}^d$ whose density $f(x)$ is supported on smooth surface $\mathcal{M}$ of lower dimension m:

Thm: (Costa [5])

$$L_n/n^\alpha \ \rightarrow \ \beta_{L,\gamma} \int_{\mathcal{M}} f^\alpha(x) d\mu_g(x) = \beta_{L,\gamma} H_\alpha(X) \qquad (w.p.1)$$

$\alpha = (m - \gamma)/m$

Another representation For finite n

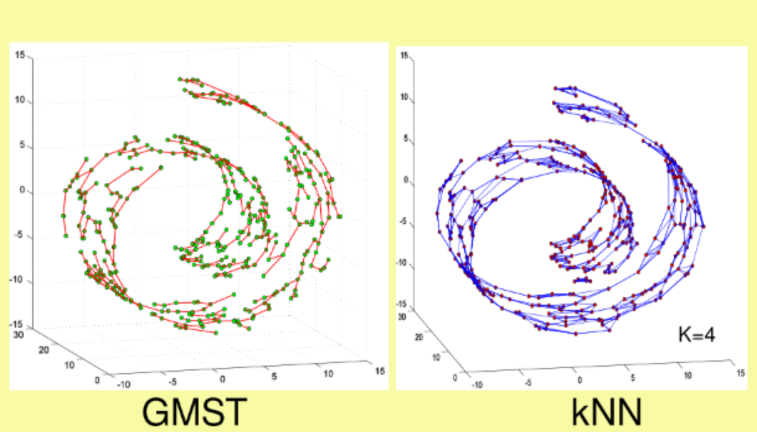
$$\log L_n = \alpha \log n + (1-\alpha) H_\alpha(X) + \log \beta_{L,\gamma} + \varepsilon(n)$$

where $\varepsilon(n) \rightarrow 0$ w.p.1.

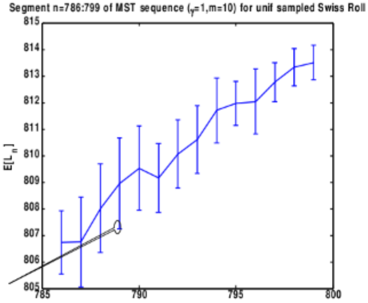
Key observation: Rate of growth of L_n in n provides a consistent estimate of α that can be used to estimate intrinsic dimension m of $\mathcal{M}$.

Dimension estimation

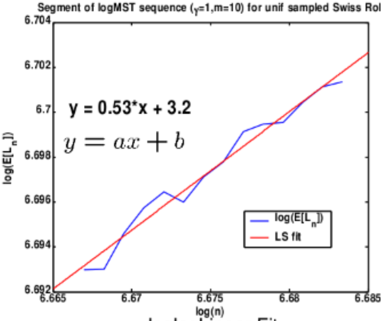
Synthetic example



Growth rate estimates of GMST



Bootstrap SE bar (83% CI)



loglogLinear Fit

$$\hat{d} = \text{round}\Bigg(\underbrace{\frac{\gamma}{1-a}}_{2.1}\Bigg) = 2 \qquad \hat{H}_\alpha(f_Y) = \frac{b - \gamma/2 \ \log \beta_{\hat{d}}}{1-a} = 7.3$$

Truth $H_\alpha(f_Y) = \log(1869) = 7.53$

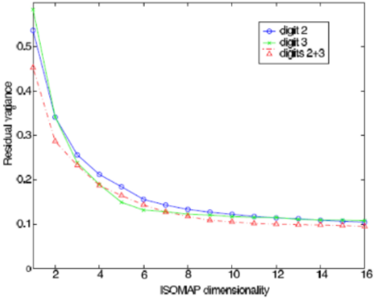


Figure: MNIST digits (48 $\times$ 64)and "scree" plot of spectrum

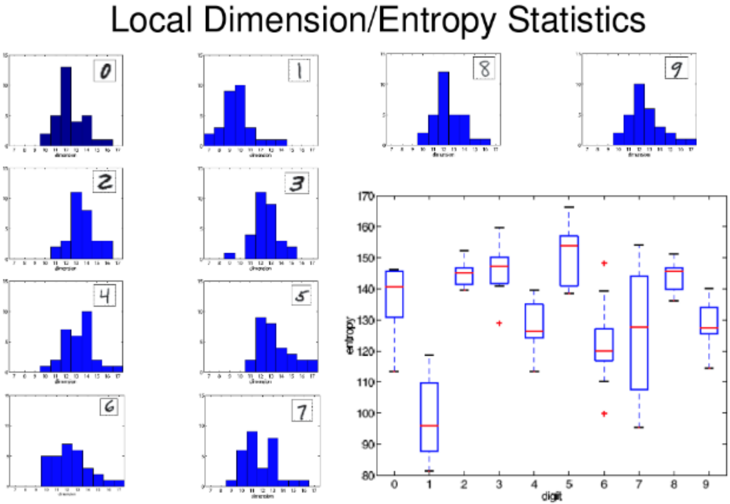


Figure: Hero and Costa [5]

Dimension estimation
Internet traffic

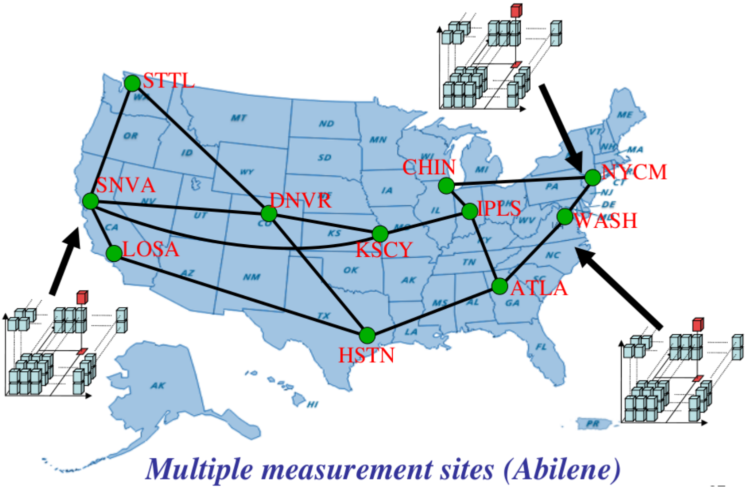
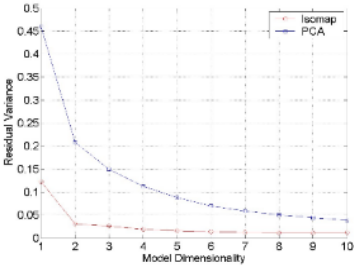


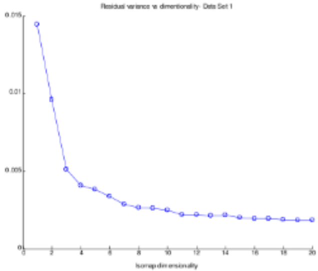
Figure: Patwari and Hero [11]

Dimension estimation

Internet traffic



Residual fitting curves
for 11x21 = 231 dimensional
Abilene Netflow data set

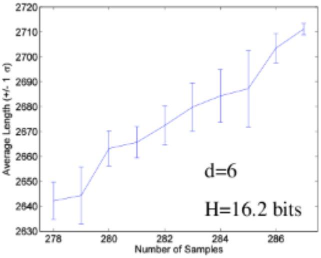


ISOMAP residual curve
for 40+ dimensional
Abilene OD link data
(Lakhina,Crovella, Diot)

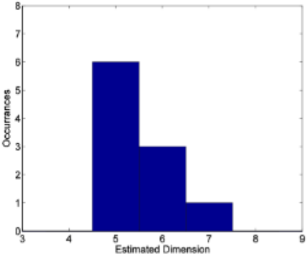
Dimension estimation

Internet traffic

- 11 routers and 21 applications = each sample lives in 231 dimensions
- 24 hour data block divided into 5 min intervals = 288 samples



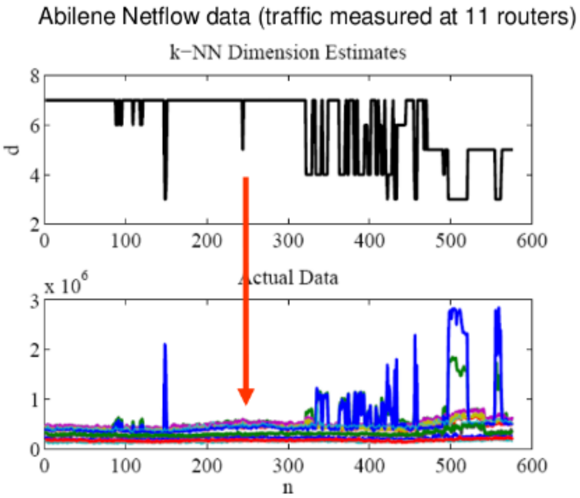
Mean GMST Length Function



Resampling histogram of d hat

Dimension estimation

Internet traffic



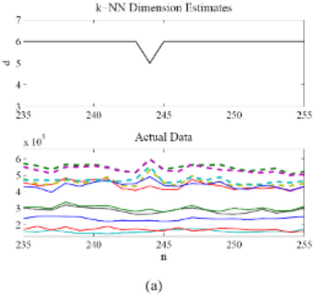


Fig. 3. Zoom shown on two non-obvious complexity changes from data in Fig. 2

Forensic analysis: Atlanta (n=244) and Seattle (n=178,179) had high flows (almost 50% of all packets) from/to IP 128.223.216.xxx on port 119.

Figure: Carter and Hero [2]

Bibliographic references for Module 2

Convergence results presented here: Hero and Costa [8], Hero and Costa [5], Hero and Michel [9]

Other relevant references

Random Euclidean graphs: Yukich [14], Penrose [12]

Original reference for BHH theorem: [1]

Dual rooted approach to convergence of MST, kNNG etc: Steele [13]

Application to dimension estimation: Costa and Hero [4], [5]

Application to anomaly detection: Hero [7]

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.

K. Carter and A. O. Hero, “Debiasing for intrinsic dimension estimation,” in *IEEE Workshop on Statistical Signal Processing*, Madison, WI, August 2007.

J. Costa, *Random Graphs for Structure Discovery in High-Dimensional Data*, PhD thesis, University of Michigan, Ann Arbor, MI, 48109, 2005.

J. Costa and A. O. Hero, “Geodesic entropic graphs for dimension and entropy estimation in manifold learning,” *IEEE Trans. on Signal Process.*, vol. SP-52, no. 8, pp. 2210–2221, August 2004.

J. Costa and A. O. Hero, “Learning intrinsic dimension and entropy of shapes,” in *Statistics and analysis of shapes*, H. Krim and T. Yezzi, editors, Birkhauser, 2005.

J. Costa, A. O. Hero, and B. Ma, “Asymptotic convergence of random graphs and entropy estimation,” Technical Report 315, Comm. and Sig.

Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, 2003.

A. O. Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," in *Proc. Neural Information Processing Systems (NIPS) Conference*, 2006.

A. O. Hero, J. Costa, and B. Ma, "Asymptotic relations between minimal graphs and alpha entropy," Technical Report 334, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, Mar, 2003.
www.eecs.umich.edu/~hero/det_est.html.

A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

N. N. Leonenko and L. P. and, "A class of rényi information estimators for multidimensional densities," *Annals of Statistics*, vol. To appear, , 2008.

N. Patwari, I. Alfred O. Hero, and A. Pacholski, "Manifold learning visualization of network traffic data," in *MineNet '05: Proceeding of the 2005 ACM SIGCOMM workshop on Mining network data*, pp. 191–196, New York, NY, USA, 2005, ACM Press.

M. Penrose, *Random geometric graphs*, Oxford University Press, 2003.

J. M. Steele, *Probability theory and combinatorial optimization*, volume 69 of *CBMF-NSF regional conferences in applied mathematics*, Society for Industrial and Applied Mathematics (SIAM), 1997.

J. E. Yukich, *Probability theory of classical Euclidean optimization*, volume 1675 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, 1998.