Geometric Entropy Minimization

Alfred Hero

University of Michigan, Ann Arbor

July 7, 2009

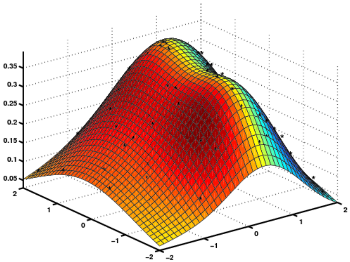
Acknowledgements

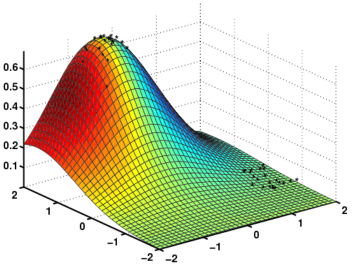
- John Gorman (SET)
- Bing Ma (UM Radiology)
- Chris Kreucher (GD)
- Huzefa Neemuchwala (Agilent)
- Jose Costa (CalTech)
- Kevin Carter (MIT Lincoln Laboratory)
- Olivier Michel (INPG Grenoble)
- Raviv Raich (Oregon State Univ)

- NSF: ITR CCR-032557
- AFOSR: FA9550-06-1-0324
- ONR: N00014-08-1-1065
- ARO: W911NF-05-1-0403
- DIGITEO, Paris France

Outline

Feature distribution supported on a smooth surface





Objective

Estimate subspace $\mathcal{S}$ along with its dimension $d \leq D$ and infer properites of sample distribution $f(y)$, $f(y) = 0$ for $y \notin \mathcal{S}$.

Unifying theme

- Inferring complexity from data
- Estimating dimension of a dataset or a distribution
- Performing dimensionality reduction for visualization
- Capturing differences/anomalies between distributions

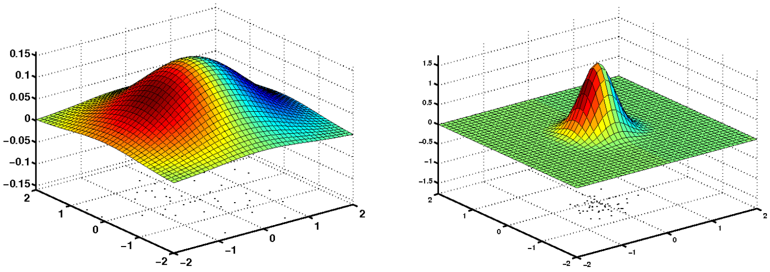
Unifying theme

- Inferring complexity from data
- Estimating dimension of a dataset or a distribution
- Performing dimensionality reduction for visualization
- Capturing differences/anomalies between distributions

Entropy and relative entropy are key tools

$$H(f) = \frac{1}{1-\alpha} \int_{\mathcal{S}} f^{\alpha}(x) dx, \quad D(f \| g) = \frac{1}{\alpha - 1} \int_{\mathcal{S}} \left(\frac{g(x)}{f(x)} \right)^{\alpha} f(x) dx$$

Feature Densities on $\mathbb{R}^2$



High entropy and low entropy feature densities

Generalized (Rényi) Entropy

Rényi entropy for a discrete r.v. X with pmf $p(x)$ (here $\alpha > 0$)

$$H_\alpha(X) = H_\alpha(p) = \frac{1}{1-\alpha}\log \sum_{x\in\mathcal{X}} p^\alpha(x) = \frac{1}{1-\alpha}\log E\left[p^{\alpha-1}(X)\right]$$

Rényi entropy for a continuous r.v. X with pdf $f(x)$

$$H_\alpha(X) = H_\alpha(f) = \frac{1}{1-\alpha}\log \int f^\alpha(x)dx = \frac{1}{1-\alpha}\log E\left[f^{\alpha-1}(X)\right]$$

Conditional Rényi entropy

$$H_\alpha(X|Y) = \int f_Y(y)\underbrace{\left(\frac{1}{1-\alpha}\log \int f^\alpha_{X|Y}(x|y)dx\right)}_{H_\alpha(X|Y=y)} dy$$

Extremal properties of Rényi entropy

- If X is discrete with finite alphabet $\mathcal{X}=\{x_1, \ldots, x_Q\}$

$$H_\alpha(X) \leq \log|\mathcal{X}| = \log Q, \quad " = " \textit{ iff } p(x_i) = \frac{1}{Q} \; \forall i$$

- If X is continous on $\mathcal{X}=\mathbb{R}$ with finite variance $\mathrm{var}(X) = E[X^2] - E^2[X]$ then $H(X)$ is maximized by a student-t density w 1 degree of freedom and identical variance.
- For X in $\mathbb{R}^d$ with given finite covariance matrix Σ Rényi entropy is maximized by multivariate Student-t density with given covariance parameter (Vignat etal [22]).

Limiting forms of Rényi entropy

- Shannon entropy limit

$$\lim_{\alpha \to 1} H_\alpha(X) = H(X) = -\int f(x) \log f(x) dx$$

- Equally likely entropy limit

$$\lim_{\alpha \to 0} H_\alpha(X) = \log Q$$

- Rarest outcome limit

$$\lim_{\alpha \to \infty} H_\alpha(X) = \log \frac{1}{\min p(x)}$$

Rényi entropy and lossless coding

- Complexity of an ensemble X = average number of bits required to optimally encode X.
- Shannon entropy $H(X)$ is optimal avg code length that minimizes redundancy
- Rényi entropy $H_\alpha(X)$ is optimal avg exponentiated code length that minimizes redundancy
- Rényi entropy $H_\alpha(X)$ increasingly sensitive to tail behavior of $f(x)$ as α decreases to zero.

Some background on Rényi entropy

- **1948** - **C. Shannon** Shannon's entropy measure published [21]
- **1961** - **A. Rényi** Rényi's α-entropies published [20]
- **1966** - **L.L. Campbell** SE and RE related by source coding arguments [1]
- **1967** - **J. Harvra** Rényi's entropy applied to classification [9]
- **1989** - **A. Mokkadem** RE used for Shannon entropy approximation [17]
- **1992** - **W. Williams** RE applied to time frequency distributions [25]
- **1994** - **B. Frieden** RE applied to signal reconstruction [7]
- **1998** - **AH** RE applied to outlier detection [13]
- **1998** - **D. Xu** RE applied to ICA [26]
- **2001** - **E. Gockay** RE applied to clustering [8]
- **2002** - **Erdogmus** RE applied to blind deconvolution [6]
- **2002** - **H. Krim** RE applied to image registration [10]
- **2003** - **C. Kreucher** RE applied to sensor management [15]
- **2004** - **S. Vinga** RE applied to DNA sequence analysis [23]
- **2004** - **J. Costa** RE applied to dimension estimation [4]
- **2005** - **H. Neemuchwala** RE applied to image retrieval [18]
- **2006** - **K. Carter** RE applied to anomaly detection [3]

Outline

Entropy minimization

Statistical parameter estimation

Define: $Y = [y_1, \ldots, y_p]$ a set of latent variables (model) and $X = [x_1, \ldots, x_n]$, $x \in \mathbb{R}^D$, data generated from model Y.
Define: empirical entropy $\hat{H}(X) = \frac{1}{n}\sum_{i=1}^{n} \phi(f(X_i|Y))$

$$\phi(u) = \begin{cases} u^{\alpha-1}/(1-\alpha), & \text{Rényi} \\ \log u, & \text{Shannon} \end{cases}$$

- Maximum likelihood estimator (p known):
  $\hat{Y} = \min_y \hat{H}(X|Y=y)$.
- Minimum description length (MDL) estimator (p unknown):
  $\hat{p}_{MDL} = \min_p \hat{H}(X, Y)$ (consistent as $n \to \infty$).

Entropy minimization

Non-parametric inference and learning

Image registration and pattern matching (Viola [24])

Estimate transformation $\mathcal{T}$ from pair of images $\{X, Y\}$,
$Y = \mathcal{T}(X) + \varepsilon$.

$$\hat{\mathcal{T}} = \text{argmin}_{\mathcal{T}} H(X, Y)$$

Entropy minimization

Non-parametric inference and learning

Image registration and pattern matching (Viola [24])

Estimate transformation $\mathcal{T}$ from pair of images $\{X, Y\}$, $Y = \mathcal{T}(X) + \varepsilon$.

$$\hat{\mathcal{T}} = \mathrm{argmin}_{\mathcal{T}} H(X, Y)$$

Anomaly detection (AH [14])

Test deviation of sample not from nominal density $f = f_o$

$$X_n \notin \mathrm{argmin}_{\mathcal{B}:P(\mathcal{B})\geq 1-\alpha} H_0(X|X \in \mathcal{B}) \int_{\mathcal{B}} f_0^{\alpha}(x)dx$$

Entropy minimization

Non-parametric inference and learning

Image registration and pattern matching (Viola [24])

Estimate transformation $\mathcal{T}$ from pair of images $\{X, Y\}$, $Y = \mathcal{T}(X) + \varepsilon$.

$$\hat{\mathcal{T}} = \mathrm{argmin}_{\mathcal{T}} H(X, Y)$$

Anomaly detection (AH [14])

Test deviation of sample not from nominal density $f = f_o$

$$X_n \notin \mathrm{argmin}_{\mathcal{B}:P(\mathcal{B})\geq 1-\alpha} H_0(X|X \in \mathcal{B}) \int_{\mathcal{B}} f_0^{\alpha}(x)dx$$

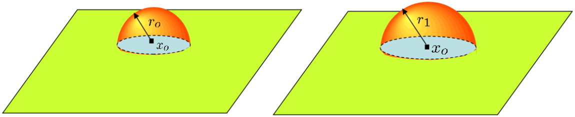
Local dimension estimation (Costa [4])

Estimate intrinsic dimension d_z of $\mathcal{S}$ in vicinity of a point $x = z$

$$\hat{d}_z = \lim_{r \to 0} \frac{dH(X|X \in B(z, r))}{d(\log r)}$$

Dimension estimation and entropy minimization

Consider growth rate of entropy over small expanding neighborhood

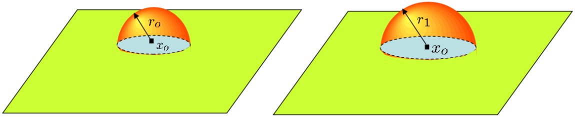


Linear equation $\mathbf{L} = d\mathbf{R} + c\mathbf{1}$ in intrinsic dimension $d \leq D$:

$$\begin{bmatrix} \log \int_{B(x_o,r_o)} \phi(f(x))dx \\ \log \int_{B(x_o,r_1)} \phi(f(x))dx \end{bmatrix} = d \begin{bmatrix} \log r_o \\ \log r_1 \end{bmatrix} + \begin{bmatrix} c(x_o) \\ c(x_o) \end{bmatrix} + \begin{bmatrix} \varepsilon_o \\ \varepsilon_1 \end{bmatrix}$$

Dimension estimation and entropy minimization

Consider growth rate of entropy over small expanding neighborhood



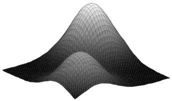
Linear equation $\mathbf{L} = d\mathbf{R} + c\mathbf{1}$ in intrinsic dimension $d \leq D$:

$$\begin{bmatrix} \log \int_{B(x_o,r_o)} \phi(f(x))dx \\ \log \int_{B(x_o,r_1)} \phi(f(x))dx \end{bmatrix} = d \begin{bmatrix} \log r_o \\ \log r_1 \end{bmatrix} + \begin{bmatrix} c(x_o) \\ c(x_o) \end{bmatrix} + \begin{bmatrix} \varepsilon_o \\ \varepsilon_1 \end{bmatrix}$$

$$\hat{d} = \mathrm{argmin}_m \min_c \|\mathbf{L} - m\mathbf{R} - c\mathbf{1}\|_2$$

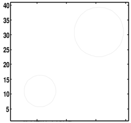
Anomaly detection and entropy minimization

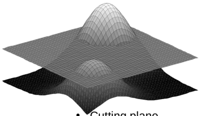
level set = minimum entropy set of nominal density $f_o(x)$



- Density function f(x)
- Level sets

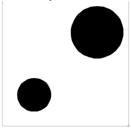
$$C(l) = \{x : f(x) = l\}$$





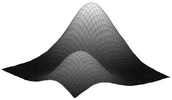
- Cutting plane
- Epigraph sets

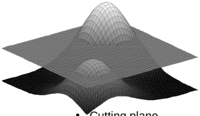
$$S(l) = \{x : f(x) \geq l\}$$



Anomaly detection and entropy minimization

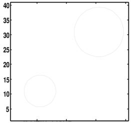
level set = minimum entropy set of nominal density $f_o(x)$





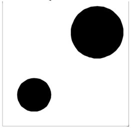
- Density function f(x)
- Level sets

$$C(l) = \{x : f(x) = l\}$$



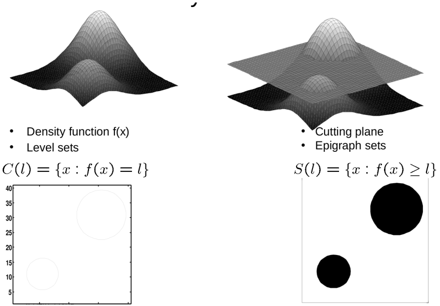
- Cutting plane
- Epigraph sets

$$S(l) = \{x : f(x) \geq l\}$$



Anomaly detection and entropy minimization

level set = minimum entropy set of nominal density $f_o(x)$



p-value: $pv(X_i) = \min_{\alpha>0} P_o(X_i \notin S_{1-\alpha})$

$$P_o(X_i \notin S_{1-\alpha}) = 1 - \int_{S_{1-\alpha}} f_o(x)dx$$

$$S_{1-\alpha} = \mathrm{argmin}_{\mathcal{B}:P_o(\mathcal{B})\geq 1-\alpha} \int_{\mathcal{B}} f_o^{\alpha}(x)dx$$

Entropy estimation

Let $h(f)$ be defined as a functional of f for given function ϕ

$$h(f) = \int \phi(f(x))dx$$

Example, $\phi(f) = f^{\alpha}/(1-\alpha)$

$$h(f) = \frac{1}{1-\alpha} \int f^{\alpha}(x)dx$$

Question: how to estimate h from empirical data?

Two methods

- Explicit density plug-in estimator

$$\hat{h} = h(\hat{f}), \quad \hat{f} = \hat{f}(X_1, \ldots, X_n)$$

- Estimation without explicit plug-in

$$\hat{h} = \hat{h}(X_1, \ldots, X_n)$$

Density plug-in estimates
Drawbacks

Drawbacks of density estimation methods for entropy estimation

- Optimal kernel bandwidth selection $\sigma = O(n^{-1/d})$ is difficult
- Datastructures for histograms are impractical in very high dimensions
- MSE convergence rate becomes logarithmic in n for large d

$$n^{-1/d} = \frac{d}{d + \log n} + O(1/d)$$

- May have few samples (fewer than dimensions)
- Density estimation in very high dimensions is fraught with difficulties

Entropy estimation without density estimation

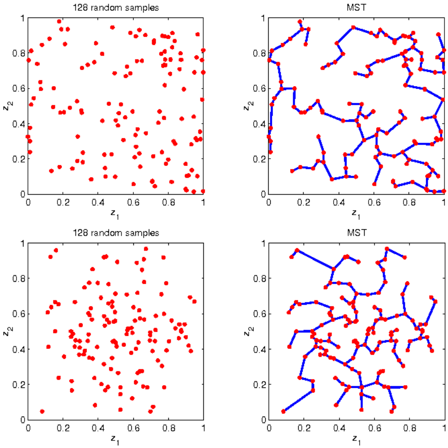
Examples of entropy estimation methods not requiring density estimation

- Data compression (LZ, CWT) entropy estimators (Kontoyanis 1998)
- kNN estimators (Leonenko 2008) [16]
- Entropic graph estimators (AH 1998) [14]

Outline

Euclidean graphs

Minimal spanning tree (MST) for uniform and triangular densities over $\mathbb{R}^D$



Rényi entropy and combinatorial optimization
MST total weight curves

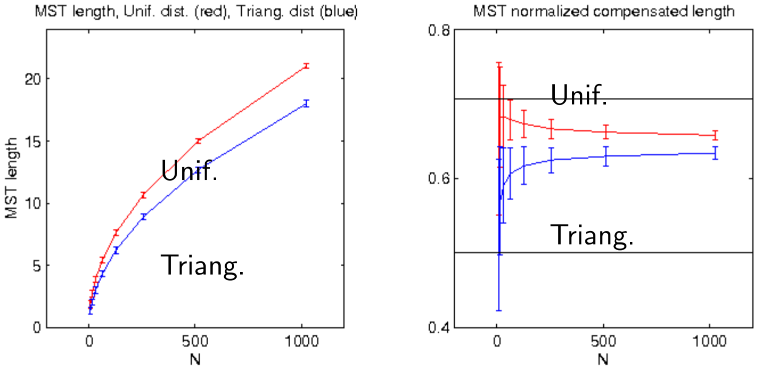


Figure: MST and log MST total weight as function of the number of samples.

Strong convergence result

BHH convergence theorem

Let $e_{ij} = \|x_i - x_j\|$ and let L_n be weighted edge length

$$L_n = \sum e_{ij}^{\gamma}, \qquad \gamma \in (0, d)$$

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*

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

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

Strong convergence result

Rényi entropy and BHH convergence theorem

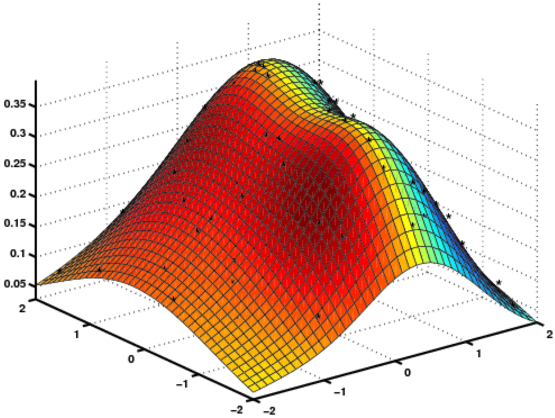
Or, letting $\alpha = (d - \gamma)/d$

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

Outline

Dimension estimation

Data collected in extrinsic dimension D but supported on a set $\mathcal{S} = \{x : f(x) > 0\}$ of dimension $d < D$



Question: how to estimate intrinsic dimension d of X?

Extended BHH theorem

BHH for points on a Riemannian manifold

Theorem: (Costa [4],[5]) Let $(\mathcal{S}, g)$ be a compact smooth Riemann d-dimensional manifold in $\mathbb{R}^D$. Suppose $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ is a random sample on $\mathcal{S}$ with bounded density f relative to μ_g and $d \geq 2$, $1 \leq \gamma < d$. Then

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

where $\alpha = (d - \gamma)/d$.
Furthermore, the mean $E[L_\gamma(\mathcal{X}_n)]/n^\alpha$ converges to the same limit.

Dimension estimation

Implication of extended BHH theorem:

Thm: (Costa [5])

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

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

Another representation For finite n

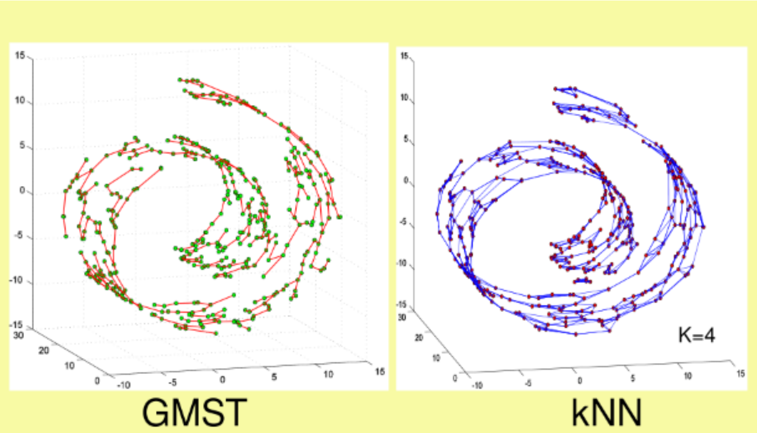
$$\log L_n = \alpha \log n + (1-\alpha) H_\alpha(X) + \log \beta_{d,L} + \varepsilon(n)$$

where $\varepsilon(n) \to 0$ 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 d of $\mathcal{S}$.

Dimension estimation

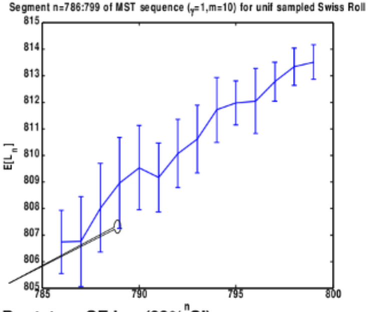
Synthetic example

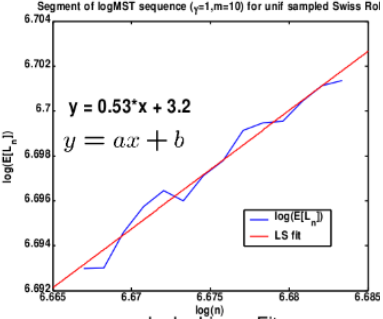


Dimension estimation

Synthetic example

Growth rate estimates of GMST





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

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

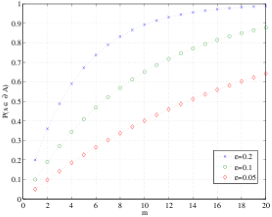
Dimension estimator bias in high dimensions

Let $X = [x_1, \ldots, x_d]$ be a random vector uniformly distrbuted in unit cube $[0, 1]^d$

Theorem: for any $\epsilon > 0$

$$P(\epsilon \leq x_i \leq 1 - \epsilon, \ \forall \ i) \leq e^{-2\epsilon d}$$

Thus, as $d \to \infty$, X escapes to the "edge" of cube with overwhelming probability - even though X uniform.



Dimension estimator bias in high dimensions
Data Depth

Let $X_1, \ldots, X_n$ be an i.i.d. sample in $\mathbb{R}^D$
Definition: (Vardi 2003) The L1 data depth of a point $X = X_i$ is

$$D_n(X) = 1 - \max\left(0, \|n^{-1} \sum_{X_i \neq X} \mathbf{e}(X_i - X)\| - n^{-1} \sum_{X_i = X}\right)$$

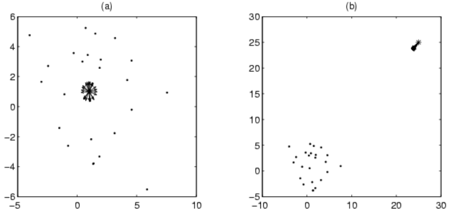


Figure: Left: a point "deep" inside data has depth ≈ 1. Right: a outlying point has depth ≈ 0 (Vardi 2003)

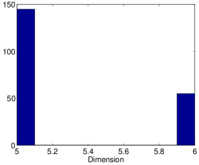
Dimension estimator bias in high dimensions

Data Depth Weighting

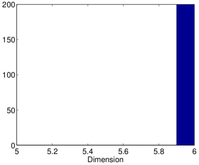
Data-depth-weighted dimension estimator: (Carter 2003 [2])

$$\hat{d} = \frac{\sum_{i=1}^{n} W_i \hat{d}_i}{\sum_{i=1}^{n} W_i}$$

$$W_i = exp(-(1 - D_n(X_i))/\sigma)$$



(a) Biased Results



(b) De-biased Results

Figure: Histograms of 200 dimension estimates obtained from 3000 i.i.d. uniform random vaectors on 6 dimensional unit sphere.

Dimension estimation
MNIST Digits

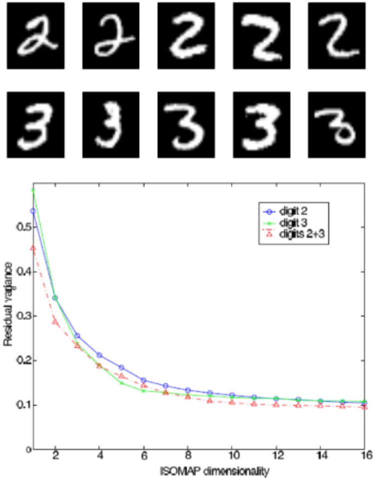


Figure: MNIST digits (48×64)and "scree" plot of spectrum

Dimension estimation
MNIST Digits

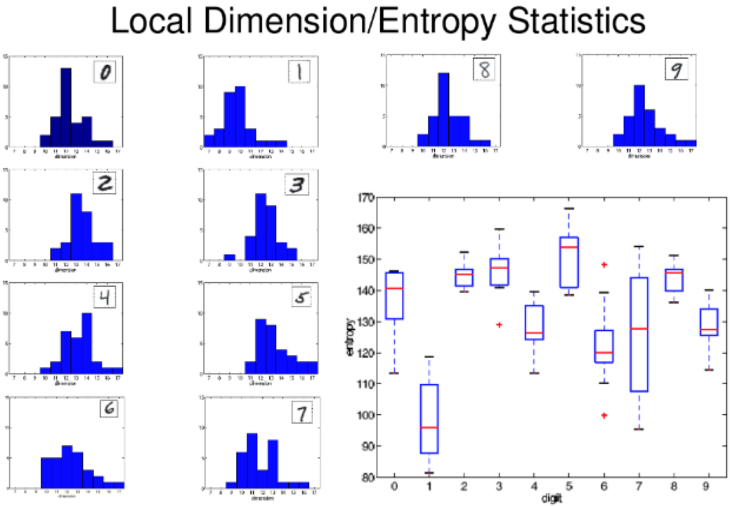


Figure: Hero and Costa [5]

Dimension estimation

Internet traffic

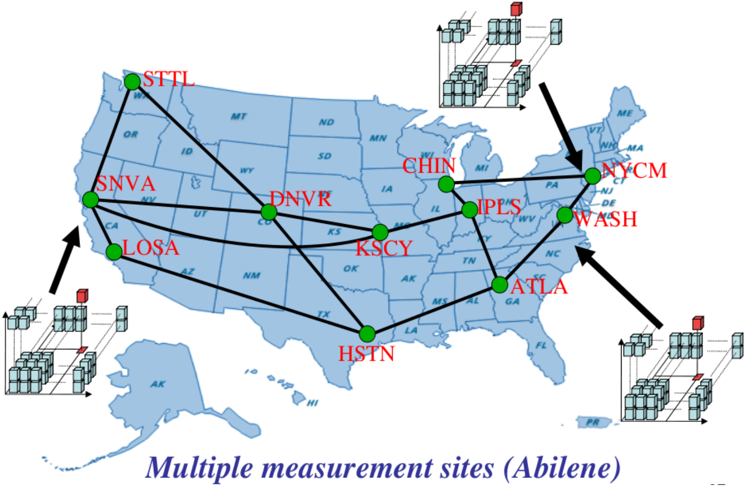
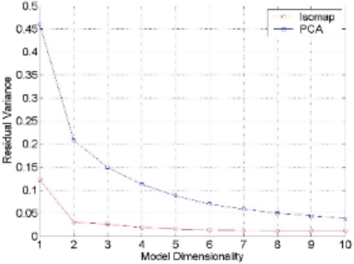


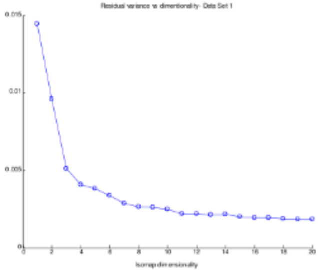
Figure: Patwari and Hero [19]

Dimension estimation

Internet traffic scree plot



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

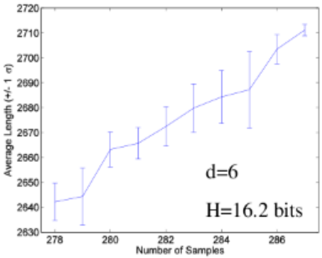


ISOMAP residual curve
for 40+ dimensional
Abilene OD link data
(Lakhina,Crovella, Diot)

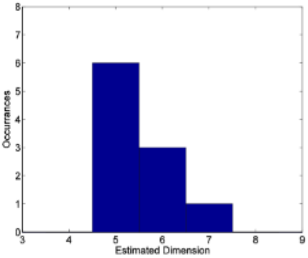
Dimension estimation

Global intrinsic dimension of internet traffic

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



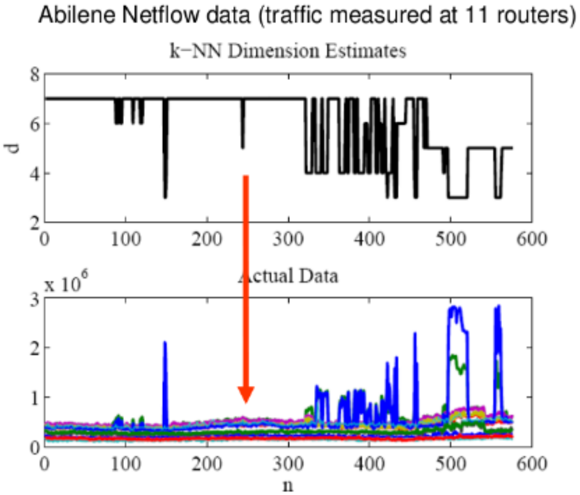
Mean GMST Length Function



Resampling histogram of d hat

Dimension estimation

Local dimension scan statistic for internet traffic



Dimension estimation

Local dimension scan statistic for internet traffic

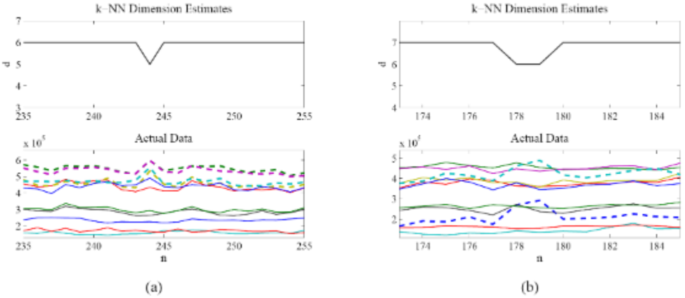


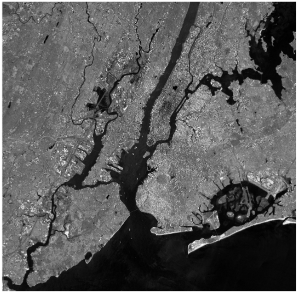
Fig. 3. Zoom shown on two non-obvious complexity changes from data in Fig. 2

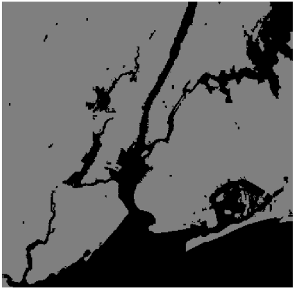
Forensic analysis: Atlanta (n=244) and Seattle (n=178,179) had high flows (almost 50% of all packets) from/to IP 128.223.216.xxx on port 119.

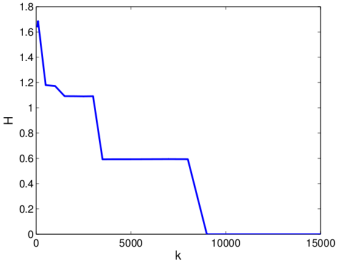
Figure: Carter [3]

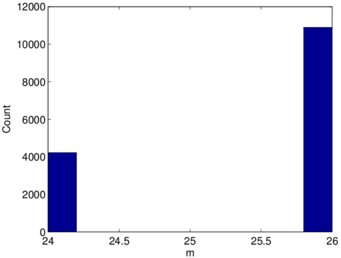
Dimension estimation

Dimension-only image segmentation



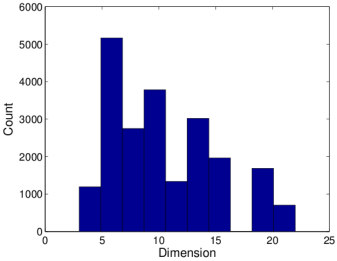


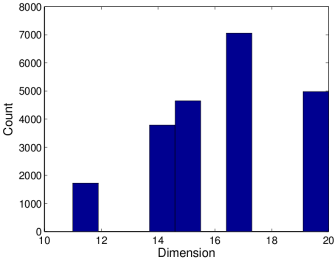




Dimension estimation

Dimension-only image segmentation





Dimension estimation
Dimension-only image segmentation

Figure: Carter [2]

Outline

BHH theorem extensions

Outlier Rejection: k-MST

Model: f is a mixture of nominal and anomalous densities

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

where

- f_1 is an "outlier" density
- f_0 is an nominal density
- $\epsilon \in [0,1]$ is unknown mixture parameter

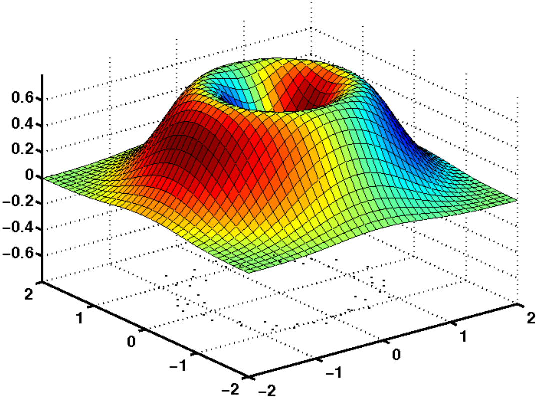
Objective: given realization $\mathcal{X}_n$ from f cluster the realizations from f_o.

Two-step k-MST procedure [14]:

1. Convert f_1 to maxent (uniform) density via measure transformation
2. "Prune" the MST on transformed $\mathcal{X}_n$ to eliminate vertices arising from maxent density

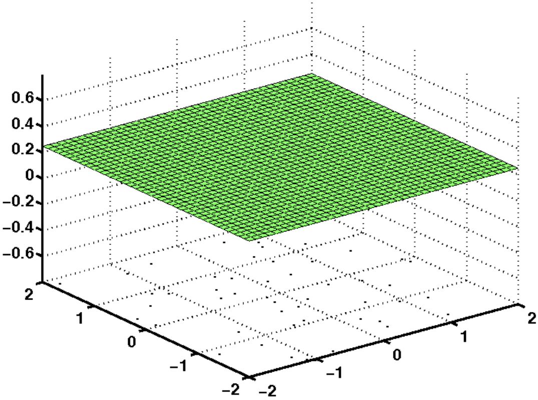
BHH theorem extensions

Example: Annulus Target Density f_1



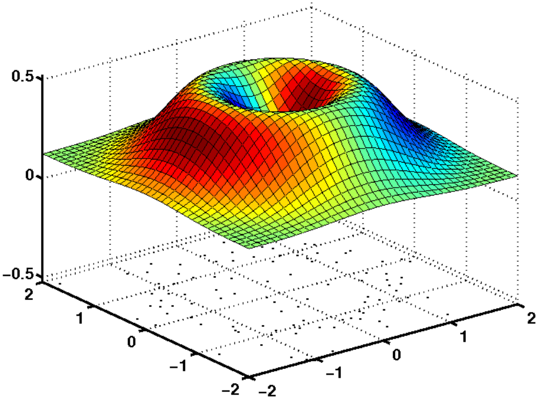
BHH theorem extensions

Uniform Outlier Density f_o



BHH theorem extensions

Mixture Density



BHH theorem extensions

k-point Minimal Spanning Tree (k-MST)

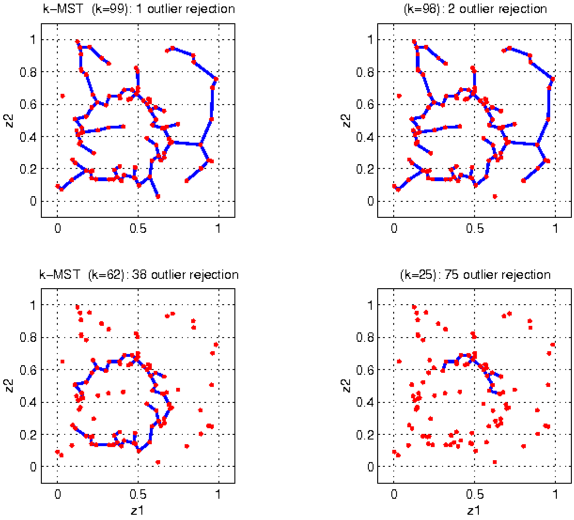


Figure: Clustering an annulus density from uniform noise via k-MST.

BHH theorem extensions

k-MST Stopping Rule

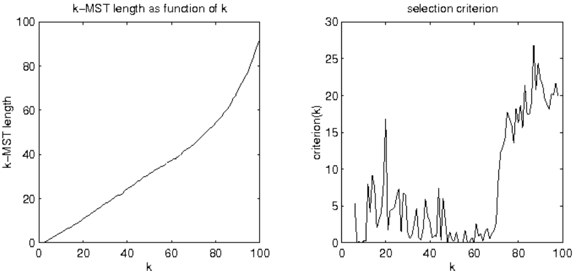


Figure: Left: k-MST curve for 2D annulus density with addition of uniform "outliers" has a knee in the vicinity of $n - k = 35$.

BHH theorem extensions

Greedy partioning approximation to k-MST

Ravi and 1996 proposed greedy partitioning approach to k-MST

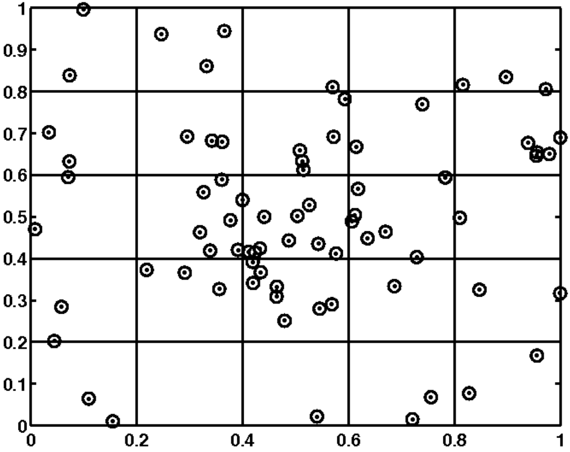


Figure: The case of $m = 5$ and $k = 17$.

BHH theorem extensions

Extended BHH Theorem for Greedy k-MST

Thm: Fix $\rho \in [0,1]$. If $k/n \to \rho$ then the length of the greedy partitioning k-MST satisfies (Hero and Michel [14])

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

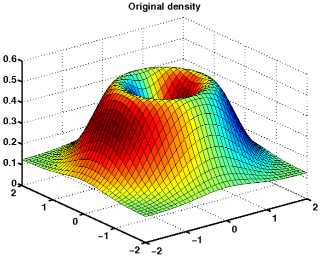
where A_o is level set of f which satisfies $\int_{A_o} f = \rho$. Alternatively, with

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

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

BHH theorem extensions

Waterpouring solution=Level set of density



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

Anomaly detection

Optimality of level set

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

$$H_0 \quad : \quad \epsilon = 0 \tag{1}$$

$$H_1 \quad : \quad \epsilon > 0 \tag{2}$$

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 $P(\Lambda(\mathbf{X}) > \eta | H_0) = 1 - \rho$

Anomaly detection

Level set estimation

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

Anomaly detection

Leave-one-out kNNG approximation to k-MST

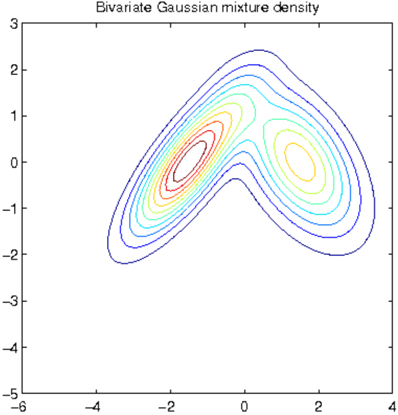


Figure: Bivariate mixture of Gaussians density

Anomaly detection

Greedy K-MST test example

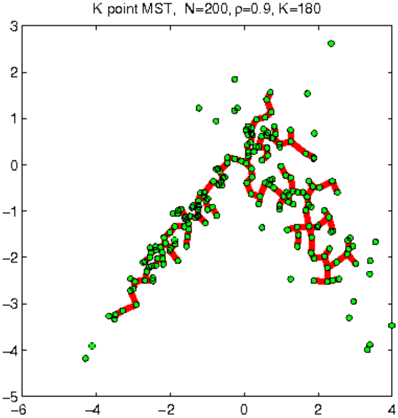


Figure: K-MST over a training realization from MoG

Anomaly detection

Greedy K-MST test example

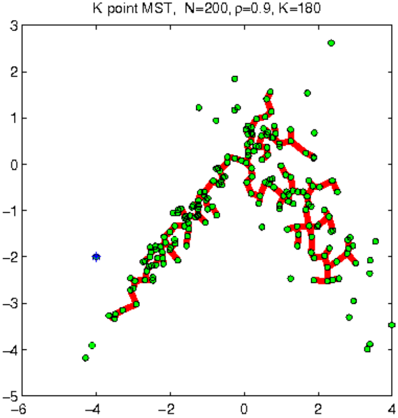


Figure: K-MST fails to capture new point (blue asterisk is outlier)

Anomaly detection

Greedy K-MST test example

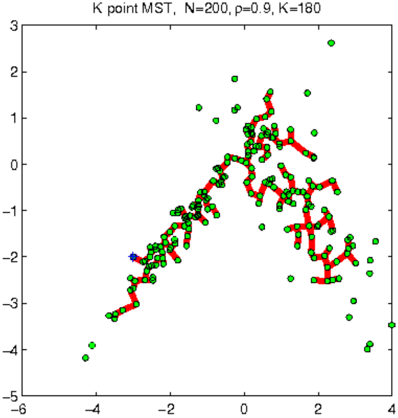


Figure: K-MST capture new point (blue asterisk is inlier)

Anomaly detection
Greedy K-MST test example

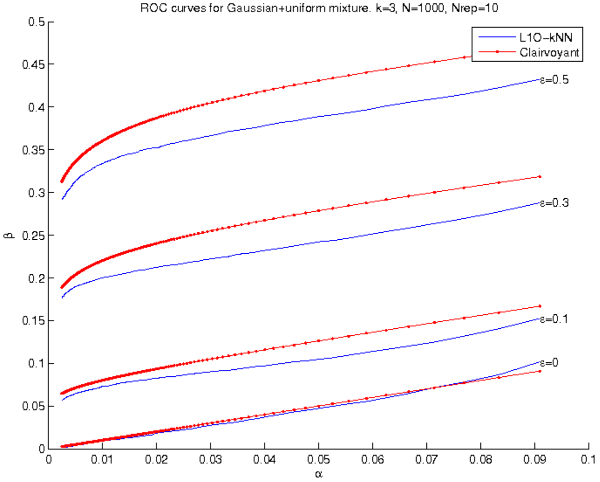


Figure: ROC curves for L1O-kNNG approximation are close to UMP curves for Gaussian example

Activity detection

Sensor network activity detection experiment

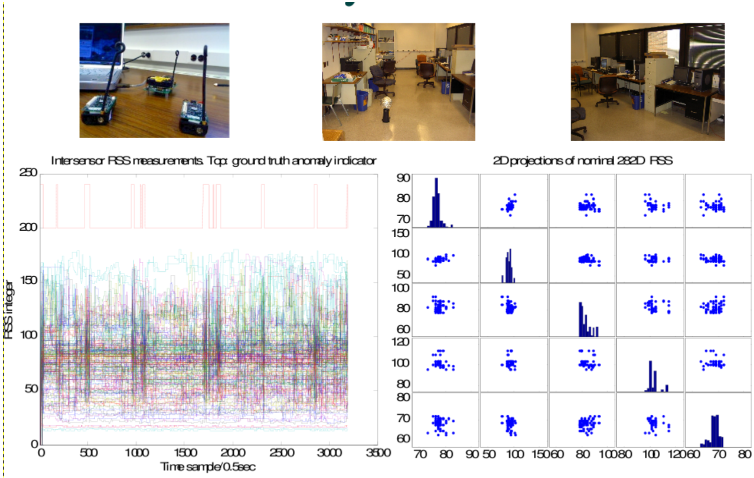


Figure: Hero [11]

Anomaly detection

Sensor network activity detection experiment

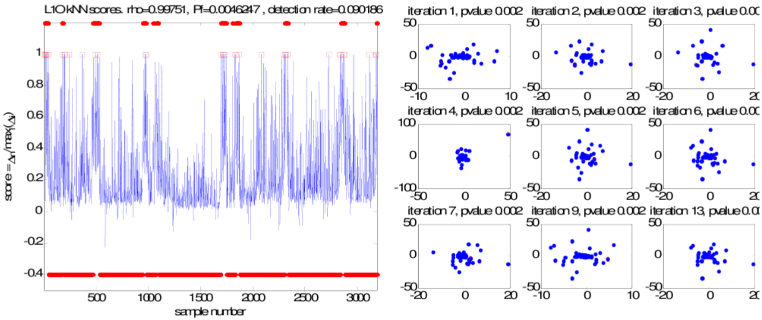


Figure: Online activity detector statistic (Left) some anomalies detected (right)

Outline

Conclusions

- Minimum entropy principle is fundamental in statistical estimation and learning
- Geometric graphs are alternatives to density plug-in estimates of entropy, topological dimension, and level sets from random samples.
- Bounds on convergence rates are available (AH and Costa [12], Costa and AH [5], AH and Michel [14]).
- Results generalize to non-Eucildean geometries such as information geometries of distributions (Carter [2]).

Bibliographic references

See following slides

- L. Campbell, "Definition of entropy by means of a coding problem," *Z. Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 6, pp. 111–118, 1966.
- K. Carter, *Dimensionality Reduction on Statistical Manifolds*, PhD thesis, University of Michigan, Dept of EECS, 2008.
- K. Carter and A. O. Hero, "Debiasing for intrinsic dimension estimation," in *IEEE Workshop on Statistical Signal Processing*, Madison, WI, August 2007.
- 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.

V. Erdogmus, J. Principe, and L. Vielva, “Blind deconvolution with minimum rényi’s entropy,” in *EUSIPCO*, Toulouse, France, 2002.

B. Frieden and A. T. Bajkova, “Reconstruction of complex signals using minimum rényi information,” in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 2298, 1994.

E. Gokcay and J. Principe, “Information Theoretic Clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 158–171, 2002.

J. Havrda and F. Chárvat, “Quantification method of classification processes,” *Kiberbetika Cislo*, vol. 1, no. 3, pp. 30–34, 1967.

Y. He, A. Ben-Hamza, and H. Krim, “An information divergence measure for ISAR image registration,” in *IEEE Int. Workshop on Statistical Signal Processing*, volume Singapore, Aug. 2001.

- 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, "Robust entropy estimation strategies based on edge weighted random graphs," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 3459, pp. 250–261, San Diego, CA, July 1998.
- 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.

C. Kreucher, K. Kastella, and A. O. Hero, “Multi-target sensor management using alpha-divergence measures,” in *3rd Workshop on Information Processing for Sensor Networks*, Palo Alto, CA, 2003.

N. Leonenko and L. Pronzato, “A class of Rényi information estimators for multidimensional densities,” *Annals of Statistics*, 2008.

A. Mokkadem, “Estimation of the entropy and information of absolutely continuous random variables,” *IEEE Trans. on Inform. Theory*, vol. IT-35, no. 1, pp. 193–196, 1989.

H. Neemuchwala, A. O. Hero, and P. Carson, “Image matching using alpha-entropy measures and entropic graphs,” *European Journal of Signal Processing (Special Issue on Content-based Visual Information Retrieval)*, vol. 85, pp. 277–296, 2005.

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.

A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.

C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. Journ.*, vol. 27, pp. 379–423, 1948.

C. Vignat, A. Hero, and J. Costa, "About closedness by convolution of the tsallis maximizers," *Physica A*, vol. 340, no. 1-3, pp. 147–152, 2004.

S. Vinga and J. Almeida, "Rényi continuous entropy of DNA sequences," *Journal of Theoretical Biology*, vol. 231, no. 3, pp. 377–388, 2004.

P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," in *Proceedings of IEEE International*

Conference on Computer Vision, pp. 16–23, Los Alamitos, CA, Jun. 1995.

W. J. Williams, M. L. Brown, and A. O. Hero, "Uncertainty, information, and time-frequency distributions," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 1566, pp. 144–156, 1991.

D. Xu, J. Principe, J. Fisher III, and H. Wu, "A novel measure for independent component analysis (ICA)," in *Acoustics, Speech, and Signal Processing, 1998. ICASSP'98. Proceedings of the 1998 IEEE International Conference on*, volume 2, 1998.