Divergence matching criteria for registration, indexing and retrieval

Alfred O. Hero

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

University of Michigan - Ann Arbor

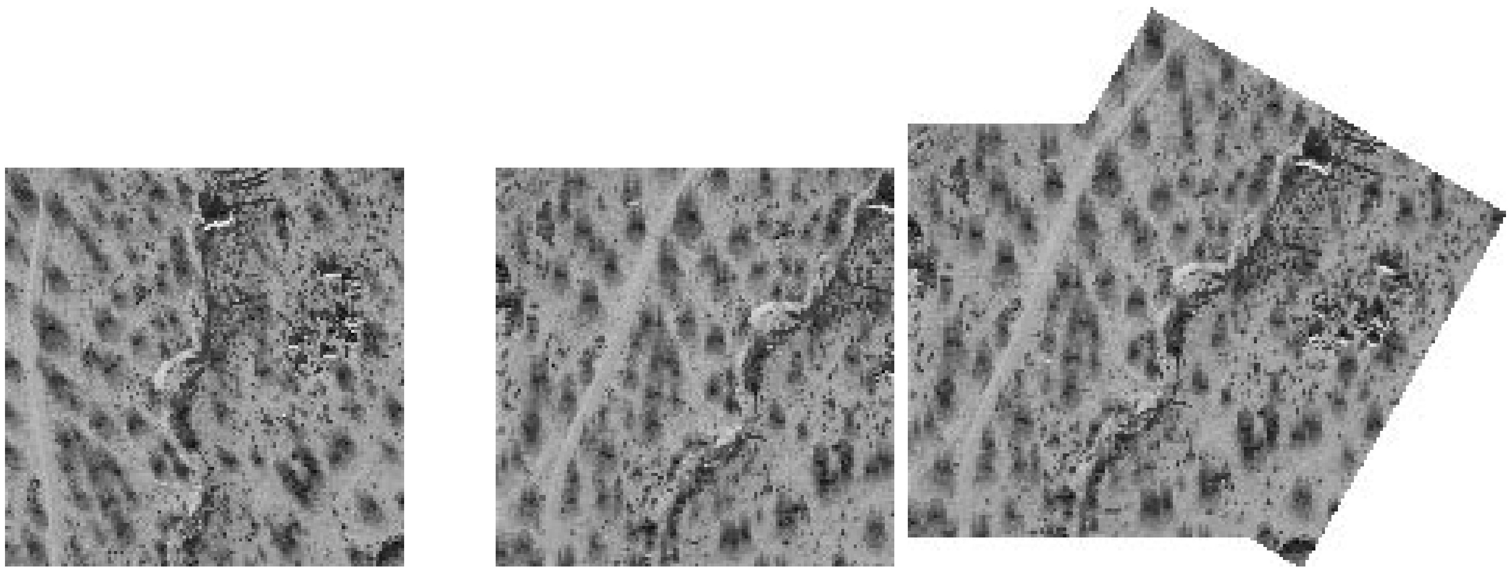
hero@eecs.umich.edu

http://www.eecs.umich.edu/˜hero

Collaborators: Olivier Michel, Bing Ma, John Gorman

Outline

1. Statistical framework: entropy measures, error exponents
2. Registration, indexing and retrieval
3. α-entropy and α-MI estimation
4. Graph theoretic entropy estimation methods



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

Figure 1: A multidate image registration example

Statistical Framework

- X: an image
- $Z = Z(X)$: an image feature vector
- Θ: a parameter space
- $f(z|\theta)$: feature density (likelihood)
- X_R a reference image
- $\{X^{(i)}\}$ a database of K images

$$Z^R = Z(X^R) \quad \sim \quad f(z|\theta_R)$$

$$Z^i = Z(X^i) \quad \sim \quad f(z|\theta_i), \quad i = 1\ldots,K$$

$\Rightarrow$ Similarity btwn X^i, X^R lies in similarity btwn models

Divergence Measures

Refs: [Csiszár:67,Basseville:SP89]

Define densities

$$f_i = f(z|\theta_i), \qquad f_R = f(z|\theta_R)$$

The Rényi α-divergence of fractional order $\alpha \in [0,1]$ [Rényi:61,70]

$$\begin{aligned} D_\alpha(f_i \parallel f_R) = D(\theta_i \| \theta_R) &= \frac{1}{\alpha - 1} \ln \int f_R \left(\frac{f_i}{f_R} \right)^\alpha dx \\ &= \frac{1}{\alpha - 1} \ln \int f_i^\alpha f_R^{1-\alpha} dx \end{aligned}$$

Rényi α-Divergence: Special cases

- α-Divergence vs. Batthacharyya-Hellinger distance

$$D_{\frac{1}{2}}(f_i \parallel f_R) \quad = \quad \ln\left(\int \sqrt{f_i f_R} dx\right)^2$$

$$D_{BH}^2(f_i \| f_R) \quad = \quad \int \left(\sqrt{f_i} - \sqrt{f_R}\right)^2 dx = 2\left(1 - \int \sqrt{f_i f_R} dx\right)$$

- α-Divergence vs. Kullback-Liebler divergence

$$\lim_{\alpha \to 1} D_\alpha(f_i, f_R) = \int f_R \ln \frac{f_R}{f_i} dx.$$

Rényi α-divergence and Error Exponents

Observe i.i.d. sample $\underline{W} = [W_1, \ldots, W_n]$

$$\begin{aligned} H_0 &: \quad W_j \sim f(w|\theta_0) \\ H_1 &: \quad W_j \sim f(w|\theta_1) \end{aligned}$$

Bayes probability of error

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

LDP gives Chernoff bound [Dembo&Zeitouni:98]

$$\liminf_{n\to\infty} \frac{1}{n} \log P_e(n) = -\sup_{\alpha\in[0,1]} \{(1-\alpha)D_\alpha(\theta_1\|\theta_0)\}\,.$$

Indexing via α-divergence

Refs: Vasconcelos&Lippman:DCC98, Stoica&etal:ICASSP98, Do&Vetterli:ICIP00

$$\begin{aligned} H_0 &: \quad Z_i^R \sim f(z|\theta_0) \\ H_1 &: \quad Z_i^R \sim f(z|\theta_1) \end{aligned}$$

Clairvoyant indexing rule:

$$X^{(i)} \prec X^{(j)} \quad \Leftrightarrow \quad D_\alpha(f_i \| f_R) < D_\alpha(f_j \| f_R)$$

Indexing problem: find θ_i attaining $\min_{\theta_i \neq \Theta_R} D_\alpha(\theta_i \| \theta_R)$

1. Image classification: f_i index model classes [Stoica&etal:INRIA98]
2. Target detection: f_R is noise reference and f_i are target references. Declare detection if $\min_{\theta_i \neq \Theta_R} D_\alpha(\theta_i \| \theta_R) > \text{threshold}$

Registration via α-Mutual-Information

Ref: Viola&Wells:ICCV95

1. Reference X^R and target X^T.
2. Set of rigid transformations $\{\mathrm{T}^i\}$
3. Derived feature vectors

$$Z^R = Z(X^R), \qquad Z^i = Z(\mathrm{T}^i(X^T))$$

$$
\begin{aligned}
H_0 &\quad : \quad \{Z_j^R, Z_j^i\} \text{ independent} \\
H_1 &\quad : \quad \{Z_j^R, Z_j^i\} \text{ dependent}
\end{aligned}
$$

Error exponent is α-MI (Pluim&etal:SPIE01, Neemuchwala&etal:ICIP01)

$$
\mathrm{MI}_\alpha(Z^R, Z^i) = \frac{1}{\alpha - 1} \ln \int f^\alpha(Z^R, Z^i)\left(f(Z^R) f(Z^i)\right)^{1-\alpha} dZ^R dZ^i.
$$

Ultrasound Registration Example

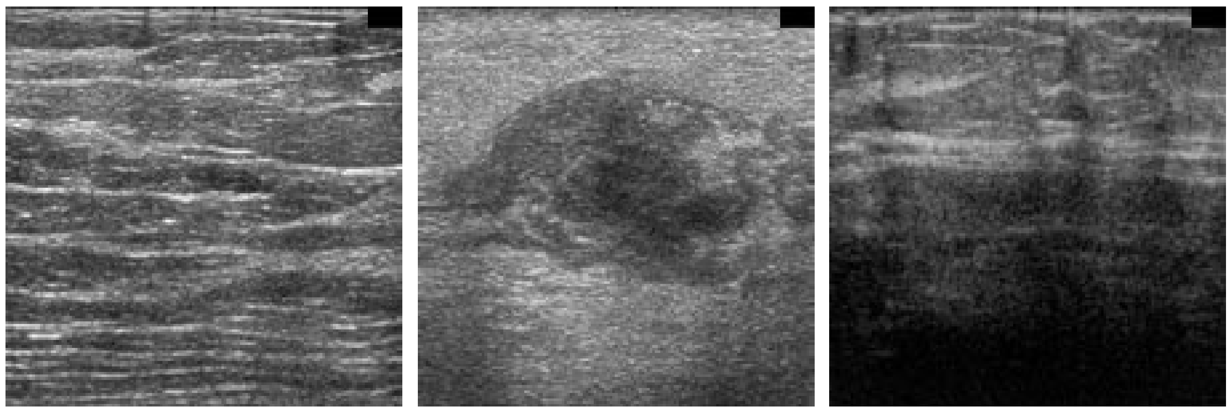


Figure 2: Three ultrasound breast scans. From top to bottom are: case 151, case 142 and case 162.

MI Scatterplots

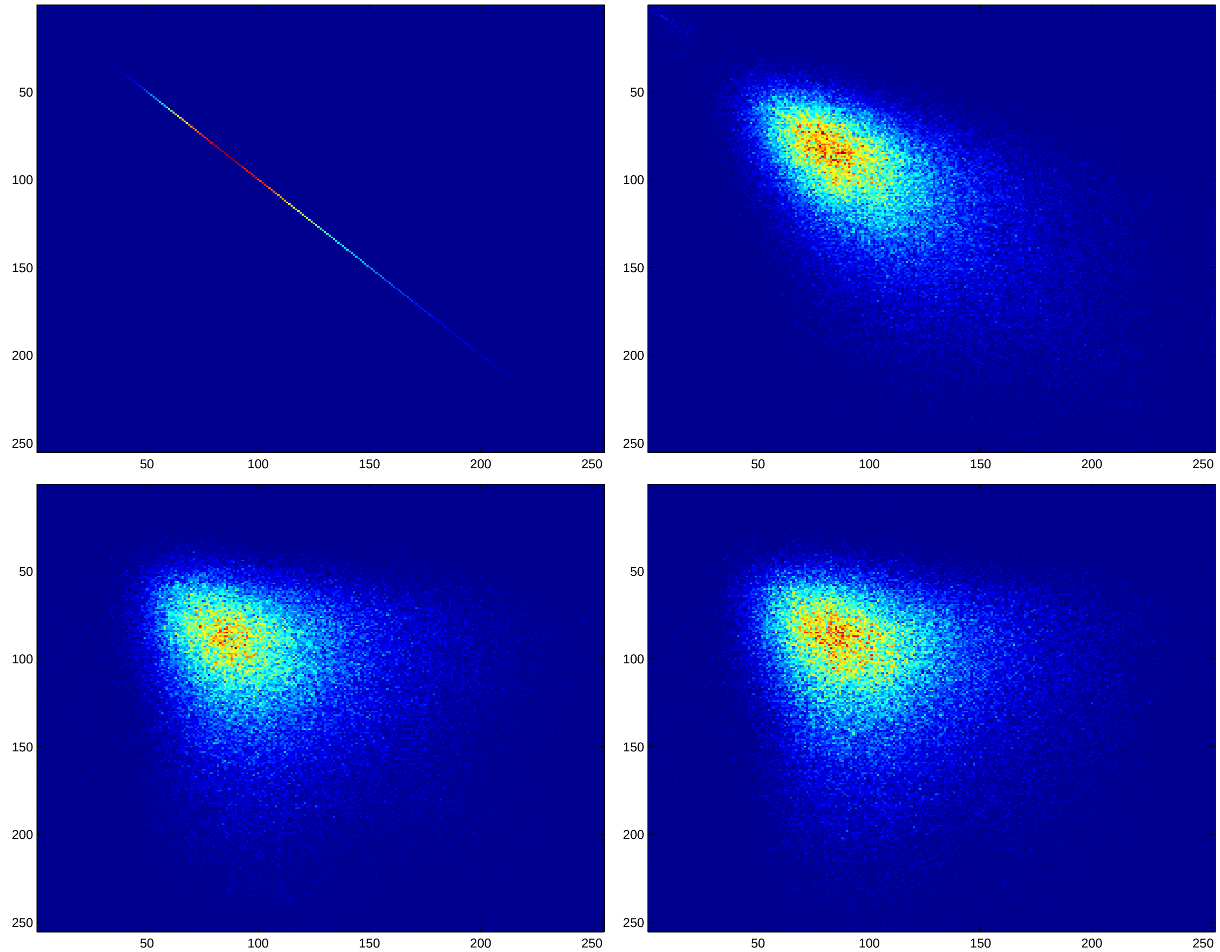


Figure 3: MI Scatterplots. 1st Col: target=reference slice. 2nd Col: target = reference+1 slice.

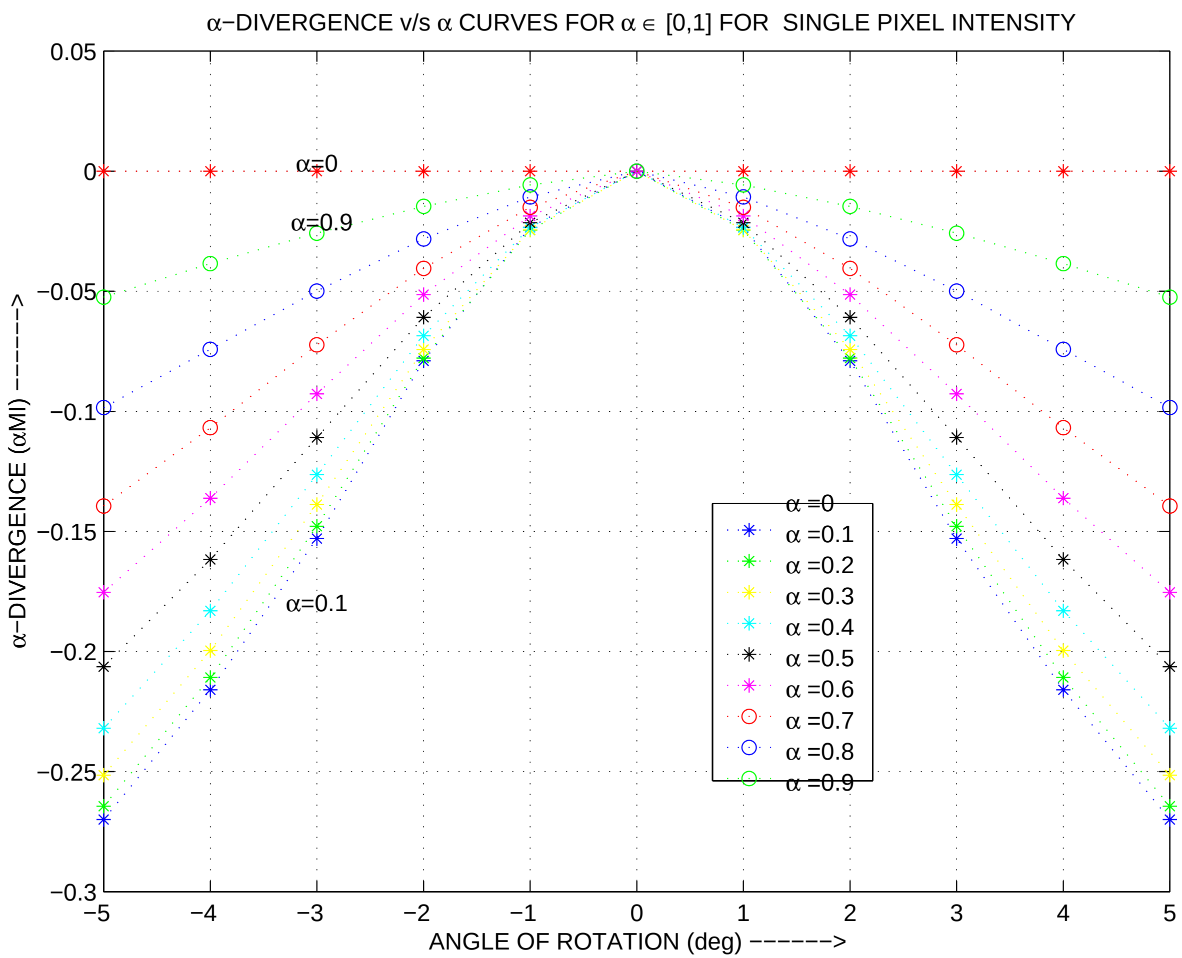


Figure 4: α-Divergence as function of angle for ultra sound image registration

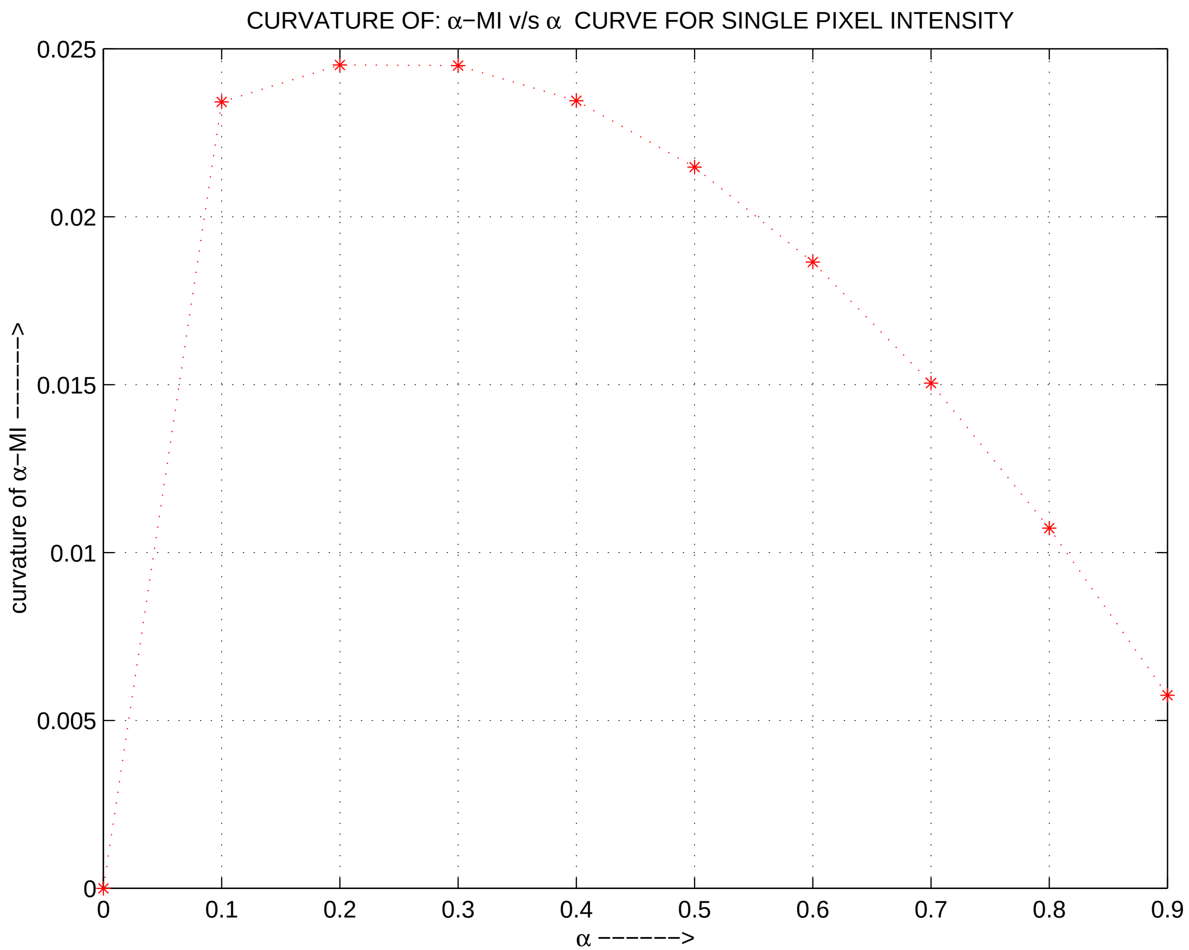


Figure 5: Resolution of α-Divergence as function of alpha

Feature Trees

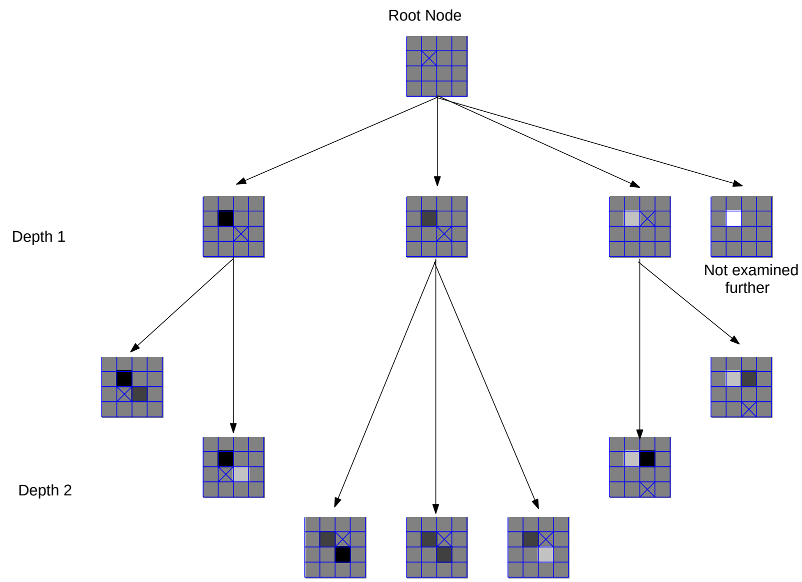


Figure 6: *Part of feature tree data structure.*

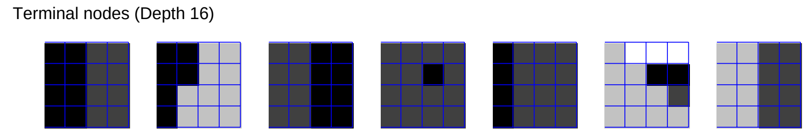


Figure 7: *Leaves of feature tree data structure.*

ICA Features

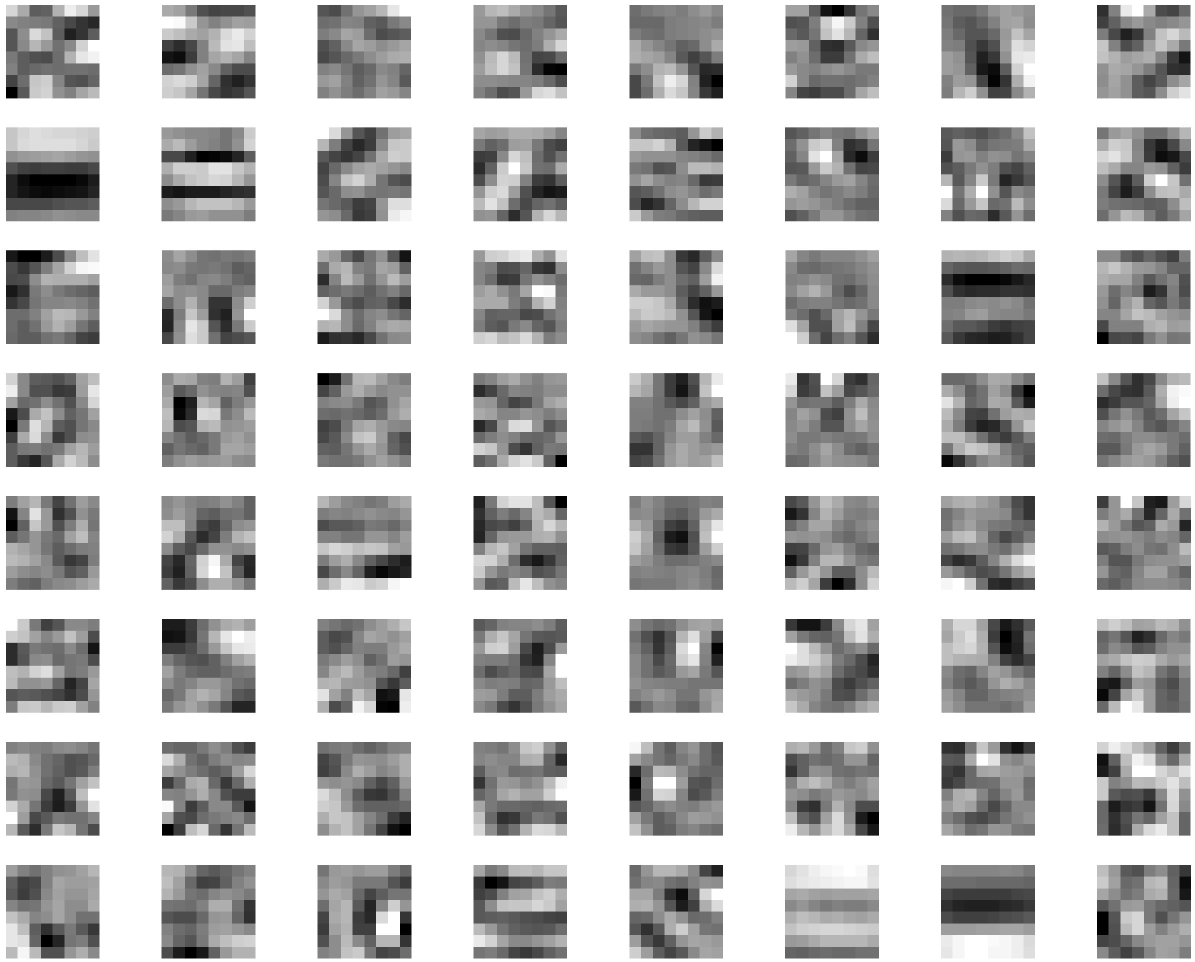


Figure 8: *Estimated ICA basis set for ultrasound breast image database*

Simple Example

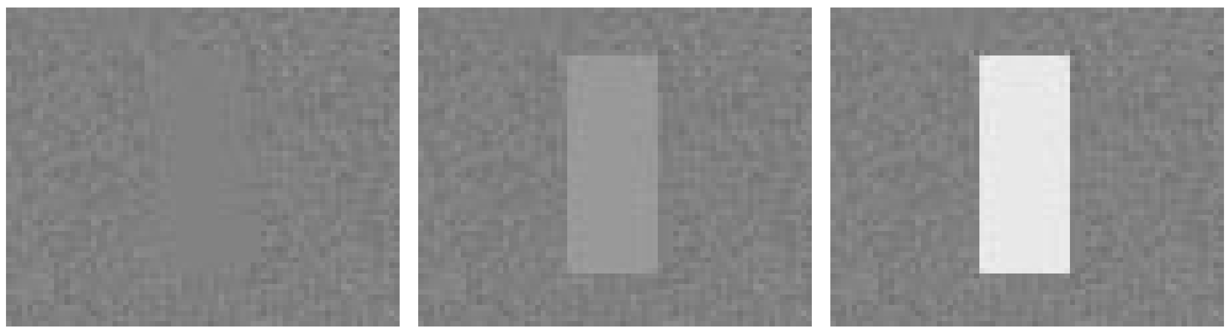


Figure 9: Bar images with contrast 1.02, 1.07 and 1.78. Background is low variance white Gaussian while bar is uniform intensity.

Single Pixel vs Feature Tag

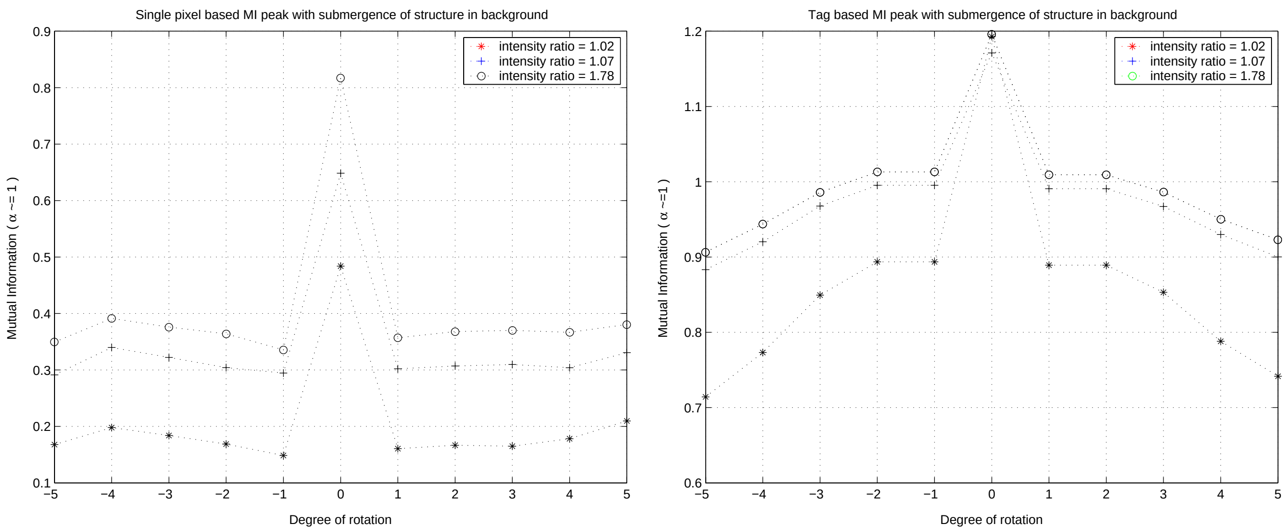


Figure 10: Upper curves are single pixel based MI trajectories while lower curves are 4×4 tag based MI trajectories for bar images.

US Registration Comparisons

	151	142	162	151/8	151/16	151/32
pixel	0.6/0.9	0.6/0.3	0.6/0.3			
tag	0.5/3.6	0.5/3.8	0.4/1.4			
spatial-tag	0.99/14.6	0.99/8.4	0.6/8.3			
ICA				0.7/4.1	0.7/3.9	0.99/7.7

Table 1: Numerator =optimal values of α and Denominator = maximum resolution of mutual α-information for registering various images (Cases 151, 142, 162) using various features (pixel, tag, spatial-tag, ICA). 151/8, 151/16, 151/32 correspond to ICA algorithm with 8, 16 and 32 basis elements run on case 151.

Methods of Divergence Estimation

- $Z = Z(X)$: a statistic (MI, reduced rank feature, etc)
- $\{Z_i\}$: n i.i.d. realizations from $f(Z;\theta)$

Objective: Estimate $\hat{D}_\alpha(f_i \| f_R)$ from Z_i's

1. Parametric density estimation methods
2. Non-parametric density estimation methods
3. Non-parametric minimal-graph estimation methods

Non-parametric estimation methods

Given i.i.d. sample $X = \{X_1, \ldots, X_n\}$

Density "plug-in" estimator

$$H_\alpha(\hat{f}_n) = \frac{1}{1-\alpha} \ln \int_{\mathbf{R}^d} \hat{f}^\alpha(x) dx$$

Previous work limited to Shannon entropy $H(f) = -\int f(x) \ln f(x) dx$

- Histogram plug-in [Gyorfi&VanDerMeulen:CSDA87]
- Kernel density plug-in [Ahmad&Lin:IT76]
- Sample-spacing plug-in [Hall:JMS86] $(d = 1)$
  - Performance degrades as density f becomes non smooth
  - Unclear how to robustify $\hat{f}$ against outliers
  - d-dimensional integration might be difficult
  - $\Rightarrow$ function $\{f(x) : x \in \mathbf{R}^d\}$ over-parameterizes entropy functional

Direct α-entropy estimation

- MST estimator of α-entropy [Hero&Michel:IT99]:

$$\hat{H}_\alpha = \frac{1}{1-\alpha} \ln L_\gamma(X_n)/n^{-\alpha}$$

  - Direct entropy estimator: faster convergence for nonsmooth densities
  - Parameter α is varied by varying interpoint distance measure
  - Optimally pruned k-MST graphs robustify $\hat{f}$ against outliers
  - Greedy multi-scale MST approximations reduce combinatorial complexity

Minimal Graphs: Minimal Spanning Tree (MST)

Let $M_n = M(X_n)$ denote the possible sets of edges in the class of acyclic graphs spanning X_n (spanning trees).

The Euclidean Power Weighted MST achieves

$$L_{\mathrm{MST}}(X_n) = \min_{M_n} \sum_{e \in M_n} \|e\|^{\gamma}.$$

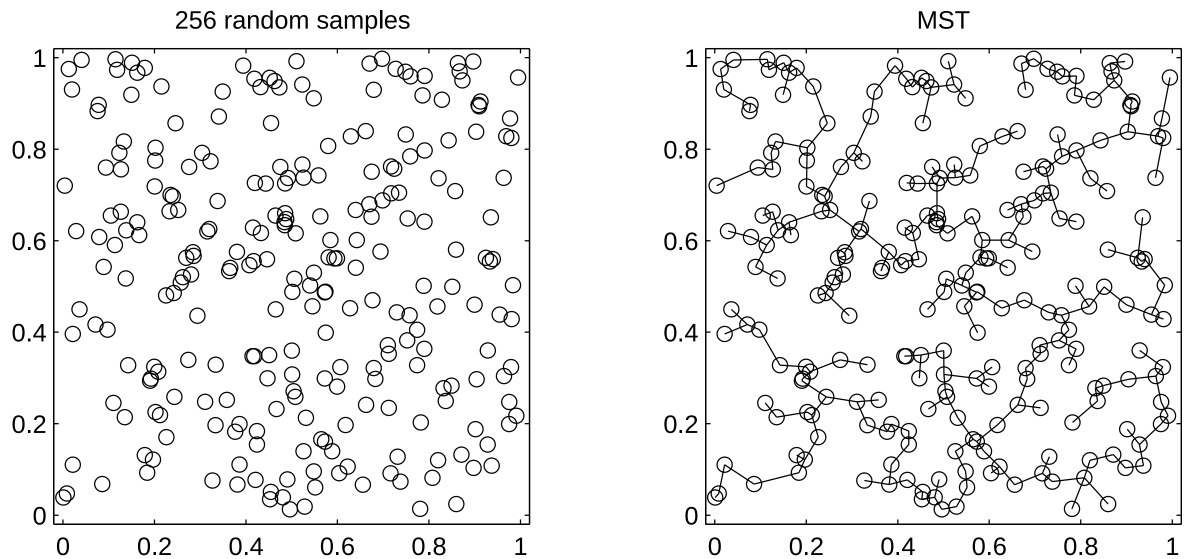


Figure 11: *A sample data set and the MST*

Minimal Graphs: Pruned MST

Fix k, $1 \leq k \leq n$.

Let $M_{n,k} = M\left(x_{i_1}, \ldots, x_{i_k}\right)$ be a minimal graph connecting k distinct vertices $x_{i_1}, \ldots, x_{i_k}$.

The k-MST $\mathrm{T}^*_{n,k} = \mathrm{T}^*(x_{i^*_1}, \ldots, x_{i^*_k})$ is minimum of all k-point MST's

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

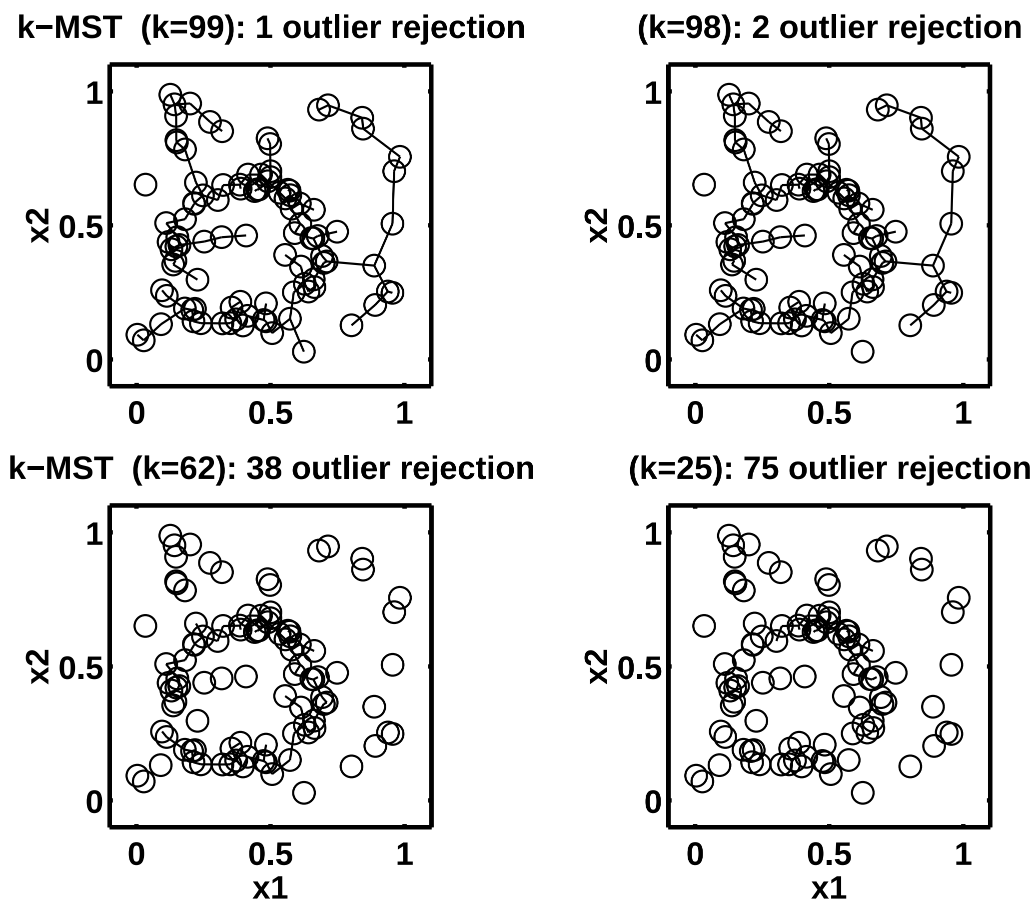


Figure 12: *k-MST for 2D torus density with and without the addition of uniform "outliers".*

Convergence of MST

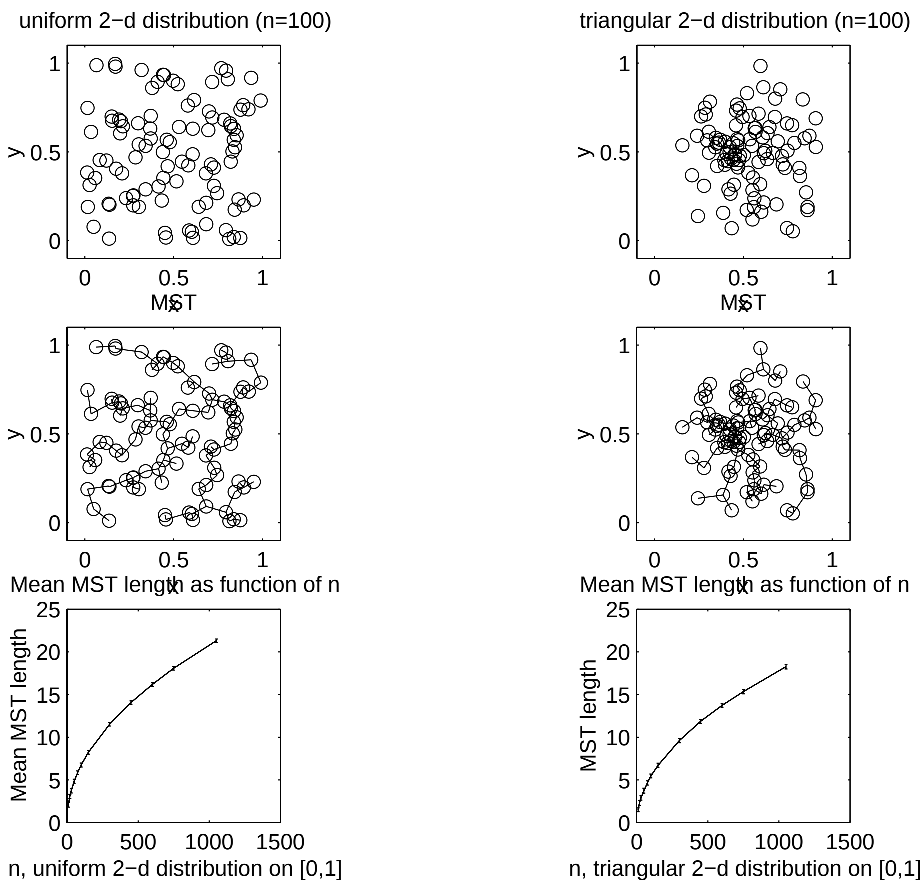


Figure 13: *2D Triangular vs. Uniform sample study for MST.*

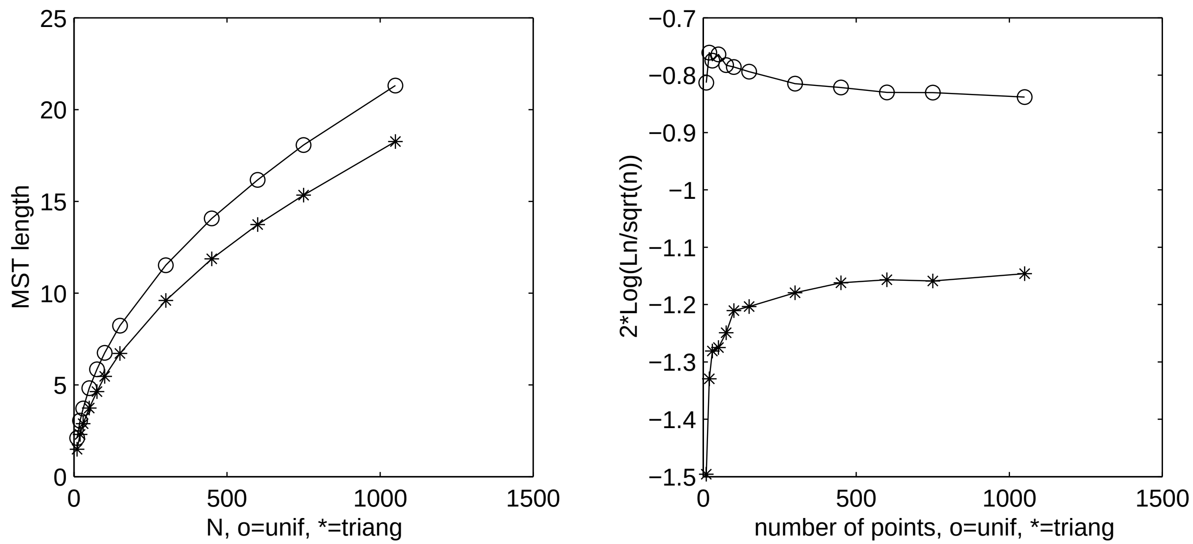


Figure 14: *MST and log MST weights as function of number of samples for 2D uniform vs. triangular study.*

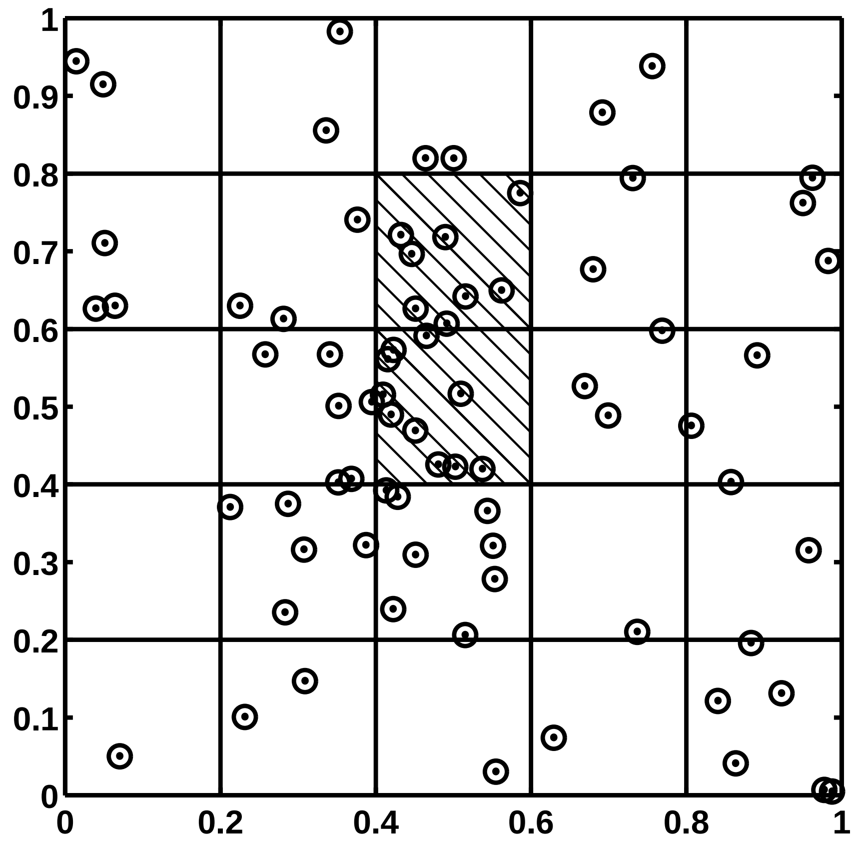


Figure 15: *Continuous quasi-additive euclidean functional satisfies "self-similarity" property on any scale.*

Asymptotics: the BHH Theorem and entropy estimation

Theorem 1

Beardwood&etal:Camb59,Steele:95,Redmond&Yukich:SPA96 *Let L be a continuous quasi-additive Euclidean functional with power-exponent γ, and let $X_n = \{X_1, \ldots, X_n\}$ 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*

$$\lim_{n\to\infty} L_\gamma(X_n)/n^{(d-\gamma)/d} = \beta_{L_\gamma,d} \int f(x)^{(d-\gamma)/d} dx, \qquad (a.s.) \tag{1}$$

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

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

Extension to Pruned Graphs

Fix $\alpha \in [0,1]$ and let $k = \lfloor \alpha n \rfloor$. Then as $n \to \infty$ (Hero&Michel:IT99)

$$L(X^*_{n,k})/(\lfloor \alpha n \rfloor)^\nu \to \beta_{L_\gamma,d} \min_{A:P(A)\geq\alpha} \int f^\nu(x|x \in A)dx \qquad (a.s.)$$

or, alternatively, with

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

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

Asymptotics: Plug-in estimation of $H_\alpha(f)$

Class of Hölder continuous functions over $[0,1]^d$

$$\Sigma_d(\kappa,c) = \left\{ f(x) : |f(x) - p_x^{\lfloor\kappa\rfloor}(z)| \leq c\,\|x-z\|^\kappa \right\}$$

Class of functions of Bounded Variation (BV) over $[0,1]^d$

$$\mathrm{BV}_d(c) = \left\{ f(x) : \sup_{\{x_i\}} \sum_i |f(x_i) - f(x_{i-1})| \leq c \right\}.$$

Proposition 1 (Hero&Ma:IT01) *Assume that* $f^\alpha \in \Sigma_d(\kappa,c)$. *Then, if* $\hat{f}^\alpha$ *is a* **minimax** *estimator*

$$\sup_{f^\alpha \in \Sigma_d(\kappa,c)} E^{1/p}\left[\left|\int \widehat{f}^\alpha(x)dx - \int f^\alpha(x)dx\right|^p\right] = O\left(n^{-\kappa/(2\kappa+d)}\right)$$

Asymptotics: Minimal-graph estimation of $H_\alpha(f)$

Proposition 2 (Hero&Ma:IT01) *Let $d \geq 2$ and $\alpha = (d-\gamma)/d \in [1/2, (d-1)/d]$. Assume that $f^\alpha \in \Sigma_d(\kappa, c)$ where $\kappa \geq 1$ and $c < \infty$. Then for any continuous quasi-additive Euclidean functional L_γ*

$$\sup_{f^\alpha \in \Sigma_d(\kappa,c)} E^{1/p}\left[\left|\frac{L_\gamma(X_1,\ldots,X_n)}{n^\alpha} - \beta_{L_\gamma,d}\int f^\alpha(x)dx\right|^p\right] \leq O\left(n^{-1/(d+1)}\right)$$

Conclude: minimal-graph estimator converges faster for

$$\kappa < \frac{d}{d-1}$$

As $\Sigma_d(1,c) \subset \mathrm{BV}_d(c)$, we have

Corollary 1 (Hero&Ma:IT01) *Let $d \geq 2$ and $\alpha = (d-\gamma)/d \in [1/2,(d-1)/d]$. Assume that f^α is of bounded variation over $[0,1]^d$. Then*

$$\sup_{f^\alpha \in \mathrm{BV}_d(c)} E^{1/p}\left[\left|\int \widehat{f}^\alpha(x)dx - \beta_{L_\gamma,d}\int f^\alpha(x)dx\right|^p\right] \quad \geq \quad O\left(n^{-1/(d+2)}\right)$$

$$\sup_{f^\alpha \in \mathrm{BV}_d(c)} E^{1/p}\left[\left|\frac{L_\gamma(X_1,\ldots,X_n)}{n^\alpha} - \beta_{L_\gamma,d}\int f^\alpha(x)dx\right|^p\right] \quad \leq \quad O\left(n^{-1/(d+1)}\right)$$

Observations

- Minimal graph rates valid for MST, k-NN graph, TSP, Steiner Tree, etc
- Analogous rate bound holds for progressive-resolution algorithm

$$L_\gamma^G(X_n) = \sum_{i=1}^{m^d} L_\gamma(X_n \cap Q_i)$$

$\{Q_i\}$ is uniform partition of $[0,1]^d$ into cell volumes $1/m^d$

- Optimal sequence of cell volumes is:

$$m^{-d} = n^{-1/(d+1)}$$

- These results also apply to greedy multi-resolution k-MST

Application: Image Registration

Two independent data samples from unknown distributions

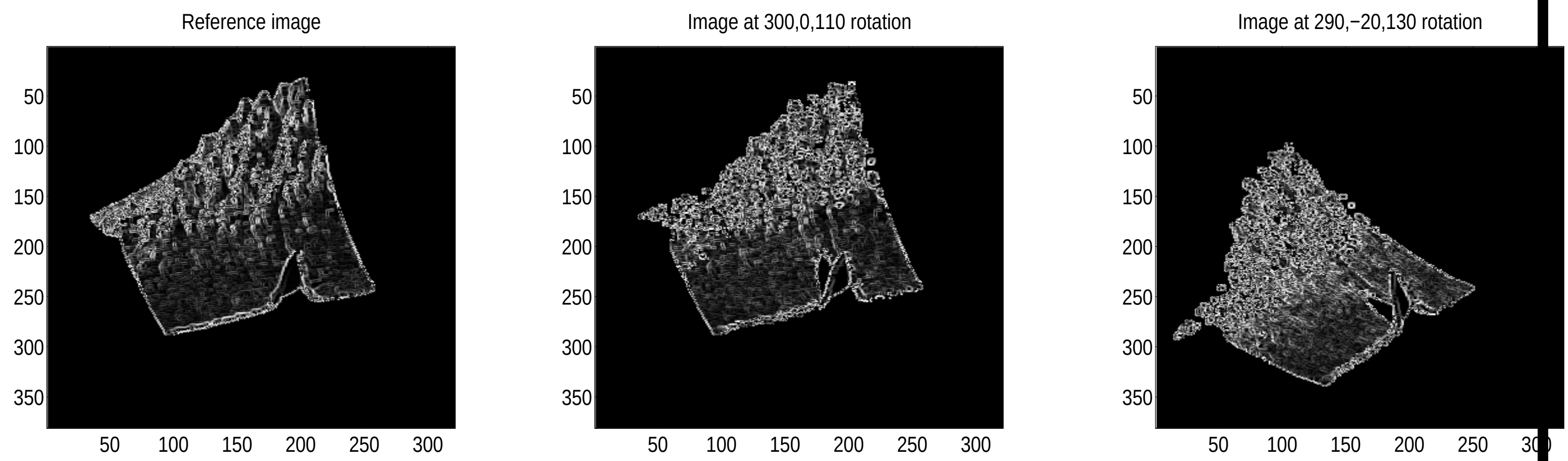
- $X = [X_1, \ldots, X_m] \sim f(x)$
- $Y = [Y_1, \ldots, Y_n] \sim g(x)$

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

Objective: find rigid transformation A, b

- Two methods:

1. α-MI of $\{(X_i, Y_i)\}_{i=1}^n$
2. α-Entropy of $\{X_i\}_{i=1}^m + \{Y_i\}_{i=1}^n$



(a) (b) (c)

Figure 16: Reference and target SAR/DEM images

$O(n^{-1/(2d+1)})$ algorithm for α-MI estimation

$$\mathrm{MI}_\alpha(X,Y) = \frac{1}{\alpha-1}\ln\int f^\alpha_{X,Y}(x,y)(f_X(x)f_Y(y))^{1-\alpha}dxdy.$$

Algorithm:

1. `Kernel estimates` $\hat{f}_X, \hat{f}_Y$ `(`$O(n^{-1/(d+2)})$`)`
2. `Uniformizing probability transformations`:
   $\tilde{X} = F_X(X),\ \tilde{Y} = F_Y(Y)$
3. `Graph entropy estimate of` $\mathrm{MI}_\alpha(X,Y)$ `(`$O(n^{-1/(2d+1)})$`)`

$$\begin{aligned}\frac{L_\gamma(\{(\tilde{X}_1,\tilde{Y}_1),\ldots,(\tilde{X}_n,\tilde{Y}_n)\})}{n^\alpha} &\to \beta_{L_\gamma,d}\int f^\alpha_{\tilde{X},\tilde{Y}}(x,y)dxdy \\ &= \beta_{L_\gamma,d}\int f^\alpha_{X,Y}(x,y)(f_X(x)f_Y(y))^{1-\alpha}dxdy \quad (w.p\end{aligned}$$

$O(n^{-1/(d+1)})$ criterion: α-Jensen difference

- Jensen's difference btwn f_0, f_1:

$$\Delta J_\alpha = H_\alpha(\varepsilon f_1 + (1-\varepsilon) f_0) - \varepsilon H_\alpha(f_1) - (1-\varepsilon) H_\alpha(f_0) \geq 0$$

- f_0, f_1 are two densities, ε satisfies $0 \leq \varepsilon \leq 1$
- Let X, Y be i.i.d. features extracted from two images

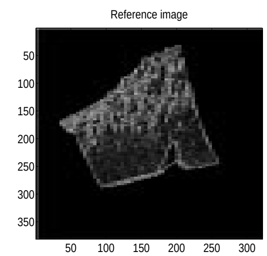
$$X = \{X_1, \ldots, X_m\}, \quad Y = \{Y_1, \ldots, Y_n\}$$

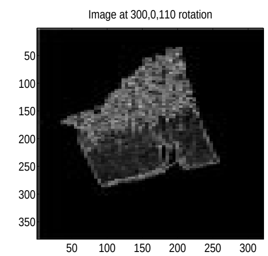
- Each realization in *unordered* sample $Z = \{X, Y\}$ has marginal

$$f_Z(z) = \varepsilon f_X(z) + (1-\varepsilon) f_Y(z), \quad \varepsilon = \frac{m}{n+m}$$

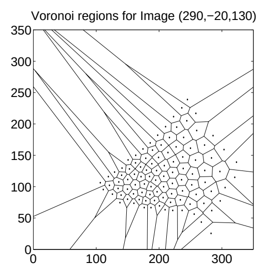
- α-Jensen difference for rigid transformation T

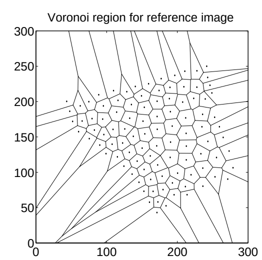
$$\Delta J_\alpha(\mathrm{T}) = H_\alpha(\varepsilon f_X + (1-\varepsilon) f_Y) - \underbrace{\varepsilon H_\alpha(f_X) - (1-\varepsilon) H_\alpha(f_Y)}_{constant}$$





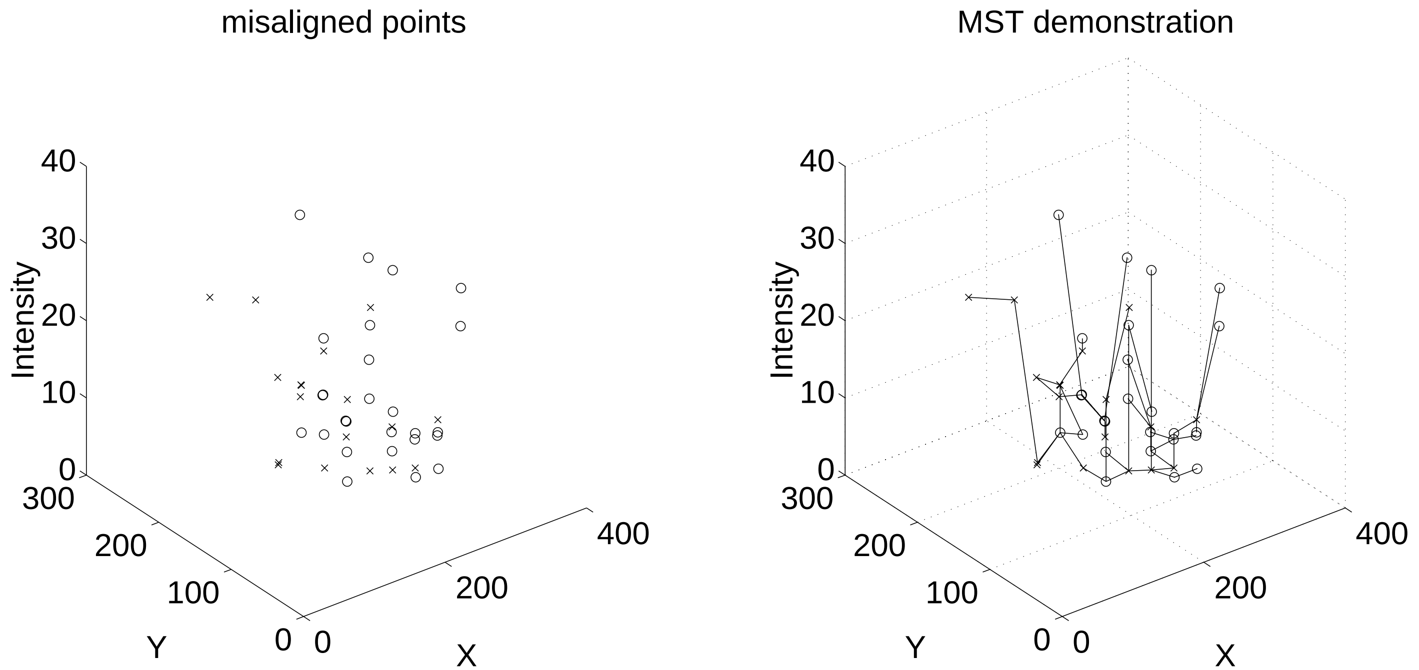
(a) (b)





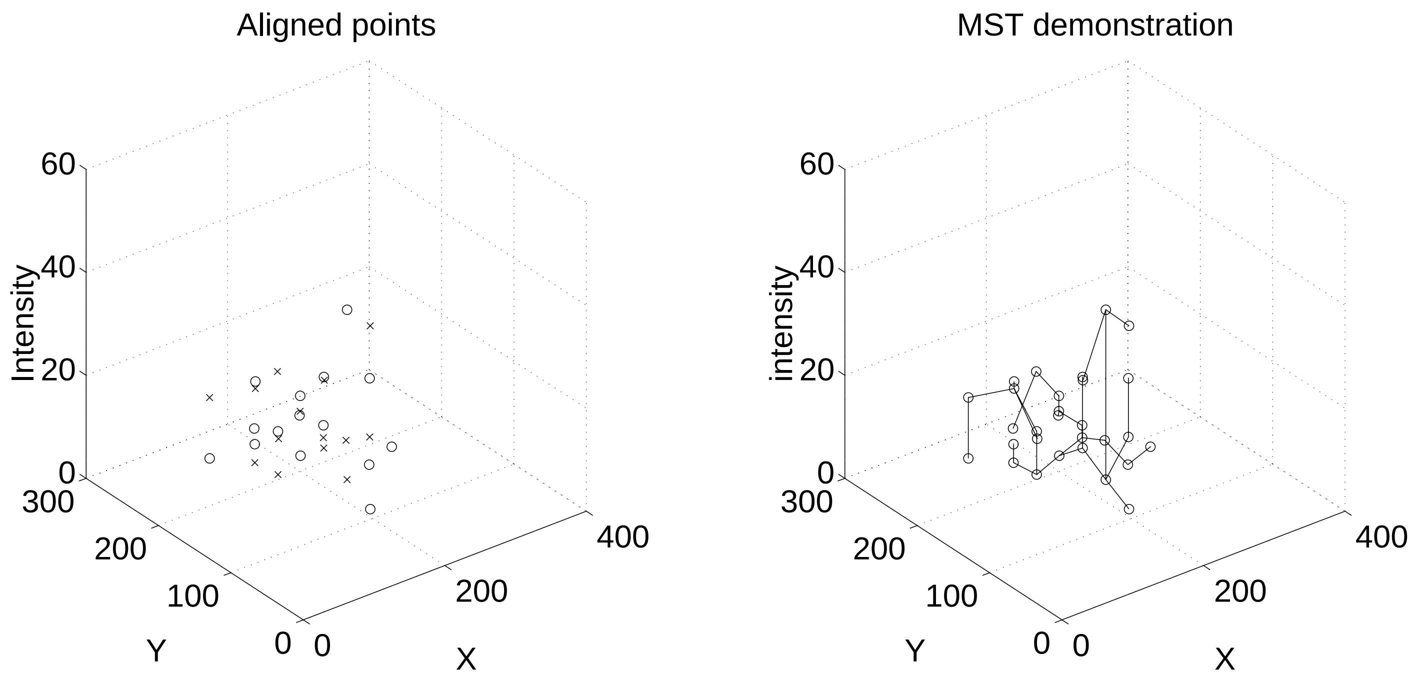
(c) (d)

Figure 17: Reference and target SAR/DEM images



(a) (b)

Figure 18: MST demonstration for misaligned images



(a) (b)

Figure 19: MST demonstration for aligned images

Conclusions

1. α-divergence for indexing can be justified via decision theory
2. Non-parametric estimation of Jensen's difference is low complexity alternative to α-divergence estimation
3. Non-parametric estimation of Jensen's difference is possible without density estimation
4. Minimal-graph estimation outperforms plug-in estimation for non-smooth densities

Divergence vs. Jensen: Asymptotic Comparison

For $\varepsilon \in [0,1]$ and g a p.d.f. define

$$f_\varepsilon \quad = \quad \varepsilon f_1 + (1-\varepsilon) f_0, \quad E_g[Z] = \int Z(x) g(x) dx, \quad \tilde{f}^\alpha_{\frac{1}{2}} = \frac{f^\alpha_{\frac{1}{2}}}{\int f^\alpha_{\frac{1}{2}} dx}$$

Then

$$\Delta J_\alpha = \frac{\alpha\varepsilon(1-\varepsilon)}{2} \left[E_{\tilde{f}^\alpha_{\frac{1}{2}}} \left(\left[\frac{f_1 - f_0}{f_{\frac{1}{2}}} \right]^2 \right) + \frac{\alpha}{1-\alpha} E_{\tilde{f}^\alpha_{\frac{1}{2}}} \left(\left[\frac{f_1 - f_0}{f_{\frac{1}{2}}} \right] \right)^2 \right] + O(\Delta)$$

$$D_\alpha(f_1 \| f_0) = \frac{\alpha}{4} \int f_{\frac{1}{2}} \left[\frac{f_1 - f_0}{f_{\frac{1}{2}}} \right]^2 dx + O(\Delta)$$