Divergence matching criteria for indexing and registration

Alfred O. Hero

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

University of Michigan - Ann Arbor

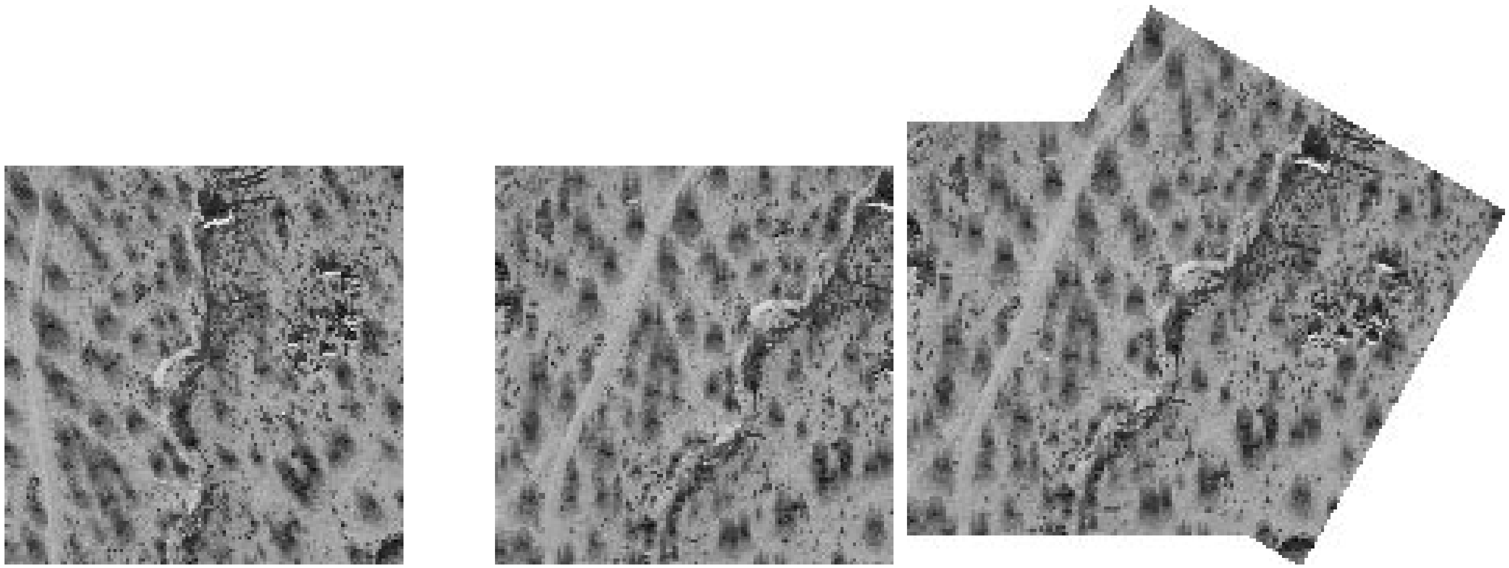
hero@eecs.umich.edu

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

Collaborators: Olivier Michel, Bing Ma, Huzefa Heemuchwala

Outline

1. Statistical framework: entropy measures, error exponents
2. α-MI indexing using single pixel gray levels
3. α-MI indexing via coincident features
4. α-entropy and α-MI estimation via MST



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

Figure 1: A multidate image registration example

Statistical Framework

- I: an image
- $Z = Z(I)$: an image feature vector over $[0,1]^d$
- I^R a reference image, feature Z^R
- $\{I^i\}$ a database of K images, features Z^i
- $f(Z^R,\, Z^i)$: joint feature density

For a pair Z^R and Z_i we have two hypotheses:

$$H_0 : f(Z^R,\, Z^i) \;=\; f(Z^R)\; f(Z^i)$$

$$H_1 : f(Z^R,\, Z^i) \;\neq\; f(Z^R)\; f(Z^i)$$

Divergence Measures

Refs: [Csiszár:67,Basseville:SP89]

Define densities

$$f_1 \;=\; f(Z^R,\, Z^i), \qquad f_0 \;=\; f(Z^R)\; f(Z^i)$$

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

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

Rényi α-Divergence: Special cases

- α-Divergence vs α-Entropy

$$H_\alpha(f_1) = \frac{1}{1-\alpha} \ln \int f_1^\alpha dx = -D_\alpha(f_1 \parallel f_0)|_{f_0 = U([0,1]^d)}$$

- α-Divergence vs. Batthacharyya-Hellinger distance

$$\begin{aligned} D_{\frac{1}{2}}(f_1 \parallel f_0) &= \ln\left(\int \sqrt{f_1 f_0} dx\right)^2 \\ D^2_{BH}(f_1 \| f_0) &= \int \left(\sqrt{f_1} - \sqrt{f_0}\right)^2 dx = 2\left(1 - \int \sqrt{f_1 f_0} dx\right) \end{aligned}$$

- α-Divergence vs. Kullback-Liebler divergence

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

Rényi α-divergence and Error Exponents

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

$$\begin{aligned} H_0 &: W_j \sim f_0(w) \\ H_1 &: W_j \sim f_1(w) \end{aligned}$$

Bayes probability of error

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

LDP gives Chernoff bound [Dembo&Zeitouni:98]

$$\liminf_{n\to\infty} \frac{1}{n} \log P_e(n) = -\sup_{\alpha\in[0,1]} \{(1-\alpha)D_\alpha(f_1 \| f_0)\}.$$

Registration via α-Mutual-Information

Ref: Viola&Wells:ICCV95

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

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

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

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

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

Registration via α-Jensen-Difference

Ref: Ma&etal:ICIP00, He&etal:SigProc01

- 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

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

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

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

Ultrasound Registration Example

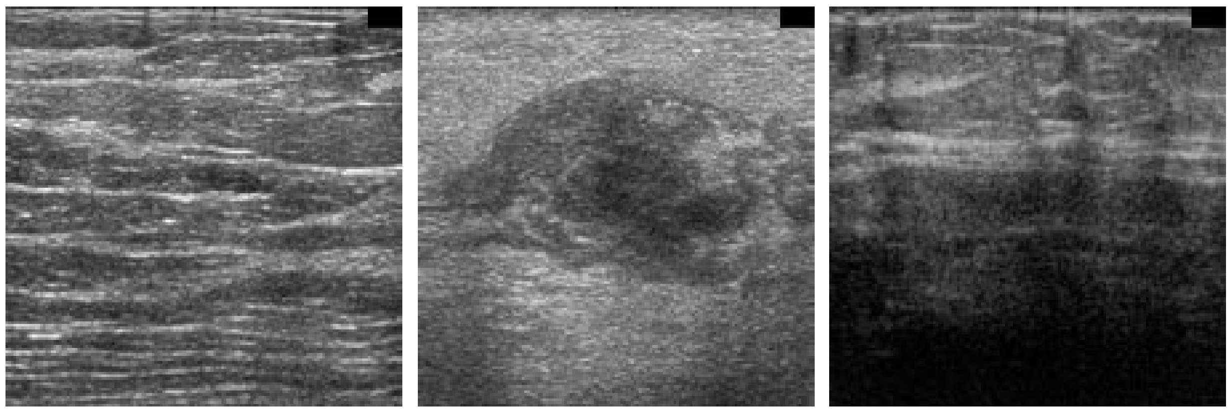
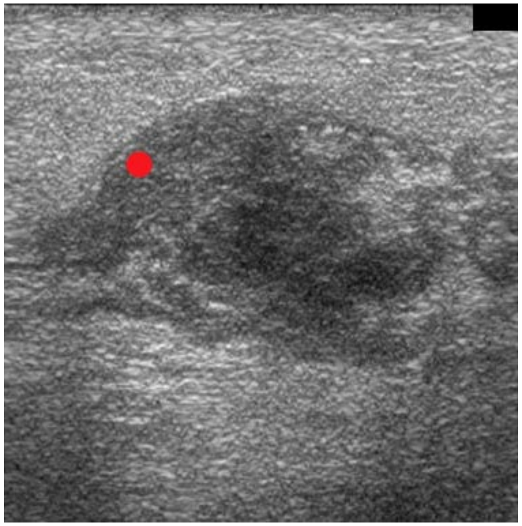
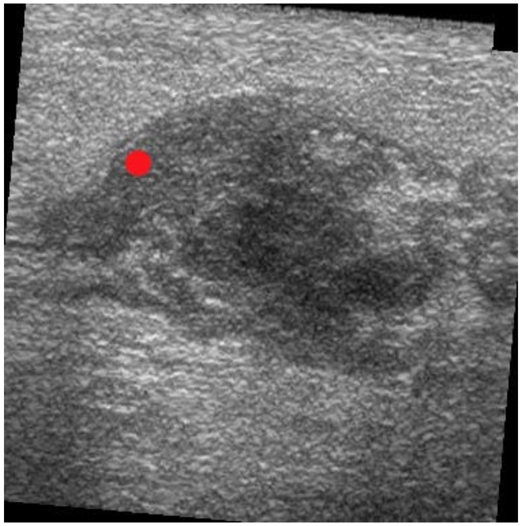


Figure 2: Three ultrasound breast scans. From top to bottom are: case 151, case 142 and case 162.



(a) Image X_1



(b) Image X_0

Figure 3: Single Pixel Coincidences (Left and right: 0^o and 8^o rotations)

Grey Level Scatterplots

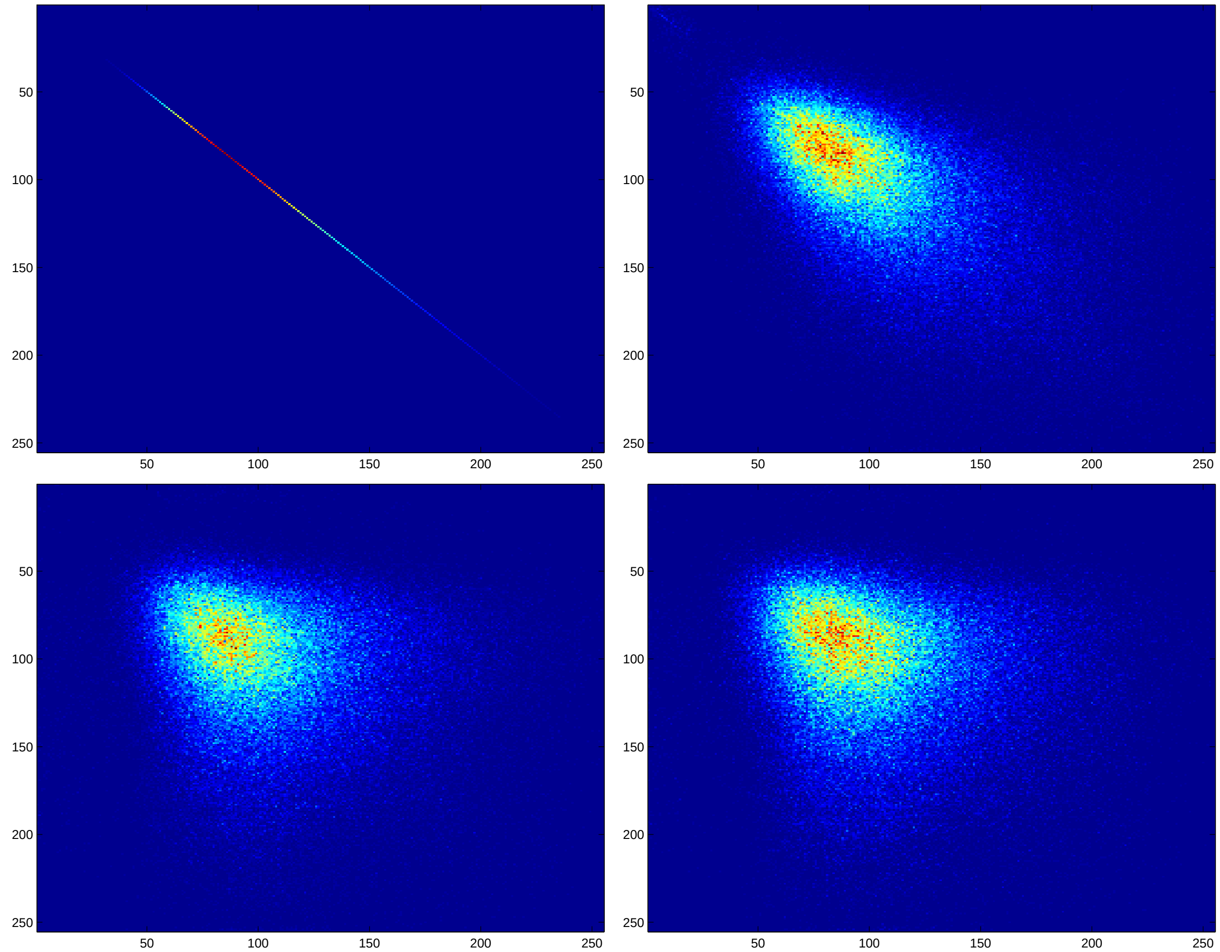


Figure 4: Grey level scatterplots. 1st Col: target=reference slice. 2nd Col: target = reference+1 slice.

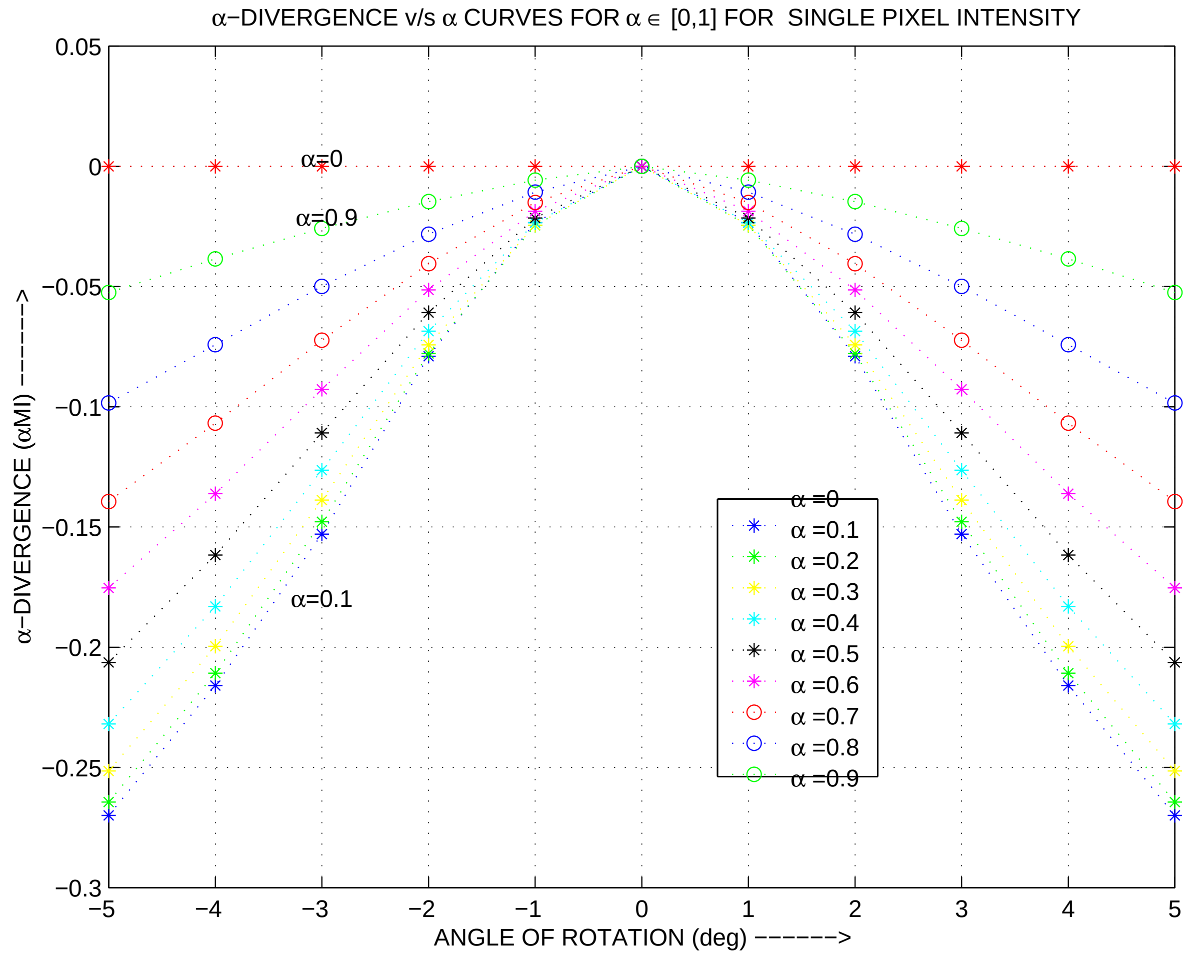


Figure 5: α-Divergence as function of angle for ultra sound image registration of image 142

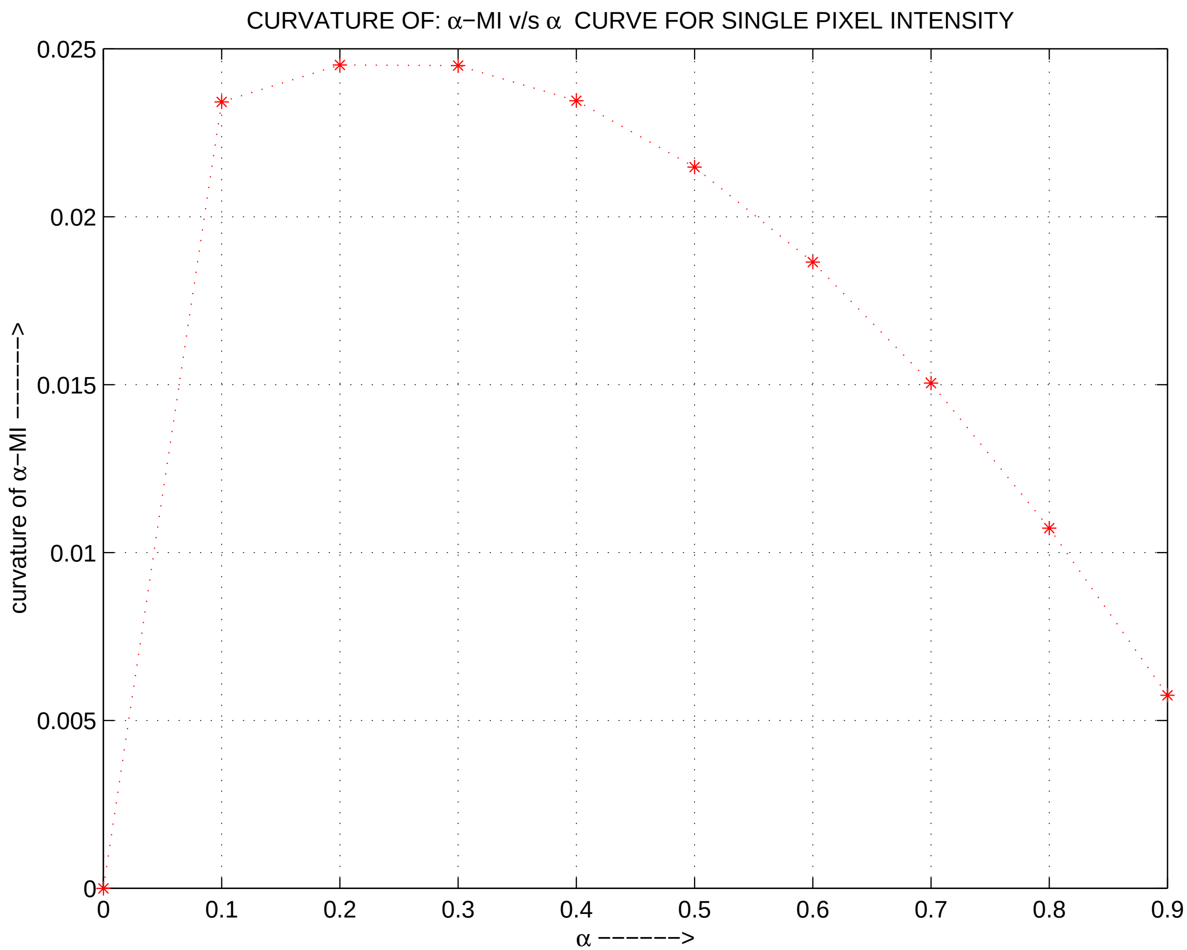


Figure 6: Resolution of α-Divergence as function of alpha

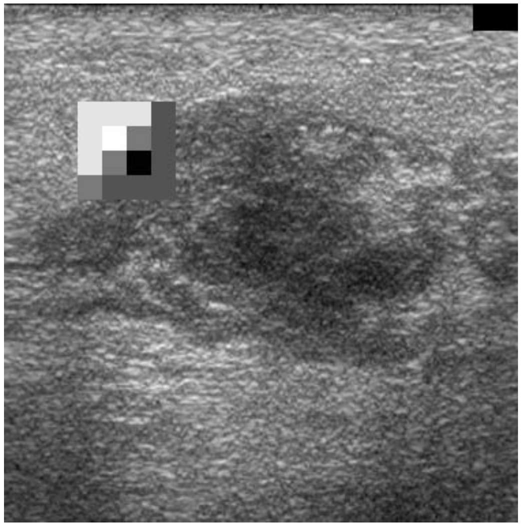
Higher Level Features

Disadvantages of gray level features:

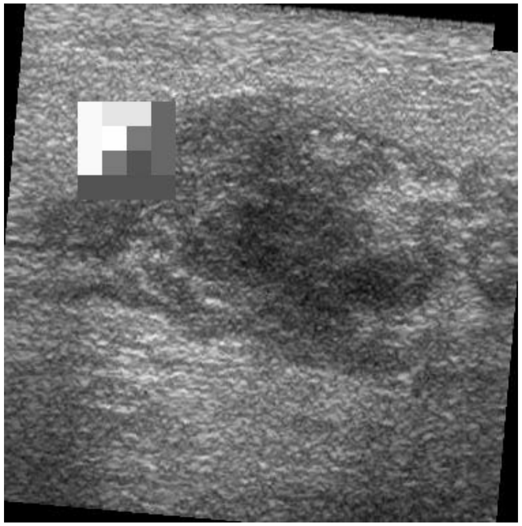
- Only depends on histogram of single pixel pairs
- Insensitive to spatial reording of pixels in each image
- Difficult to select out grey level anomalies (shadows, speckle)
- Spatial discriminants fall outside of single pixel domain

- **Alternative**: Spatial-feature-based indexing

Local Tags



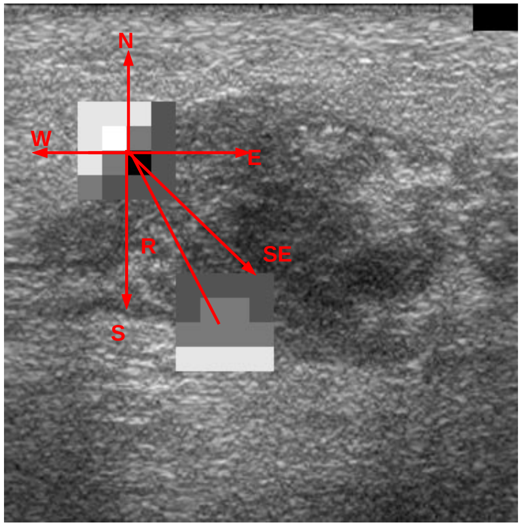
(a) Image X_0



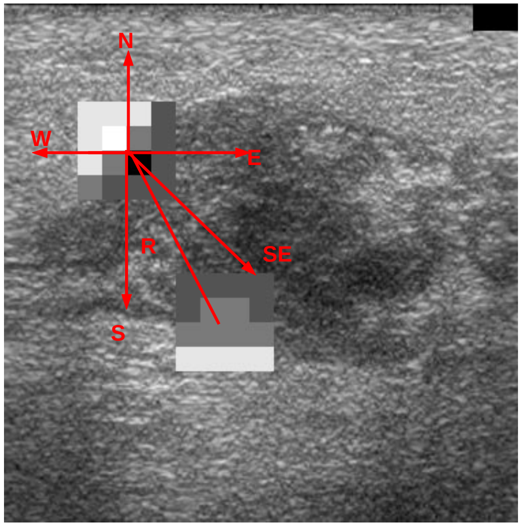
(b) Image X_i

Figure 7: Local Tag Coincidences

Spatial Relations Between Local Tags



(a) Image X_0



(b) Image X_i

Figure 8: Spatial Relation Coincidences

Feature: spatial tags via feature trees

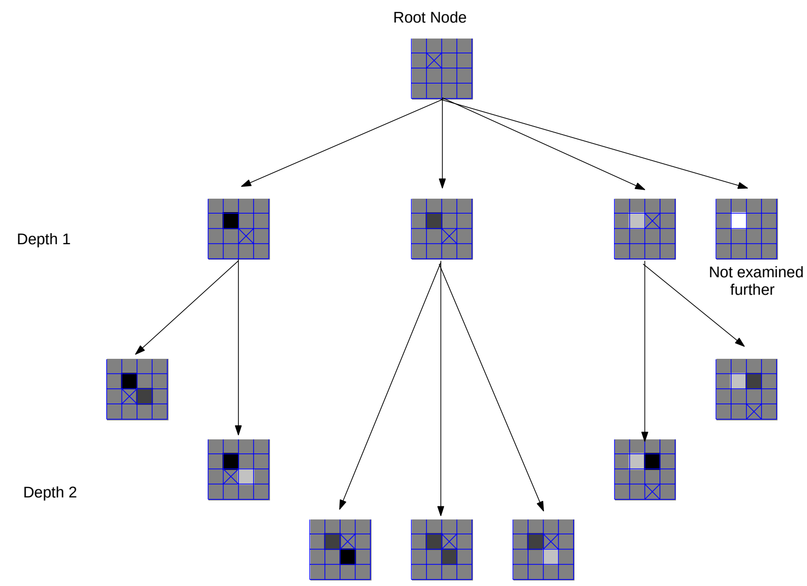


Figure 9: *Part of feature tree data structure.*

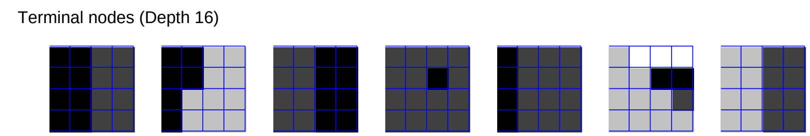


Figure 10: *Leaves of feature tree data structure.*

Feature: projection-coefficient wrt ICA basis

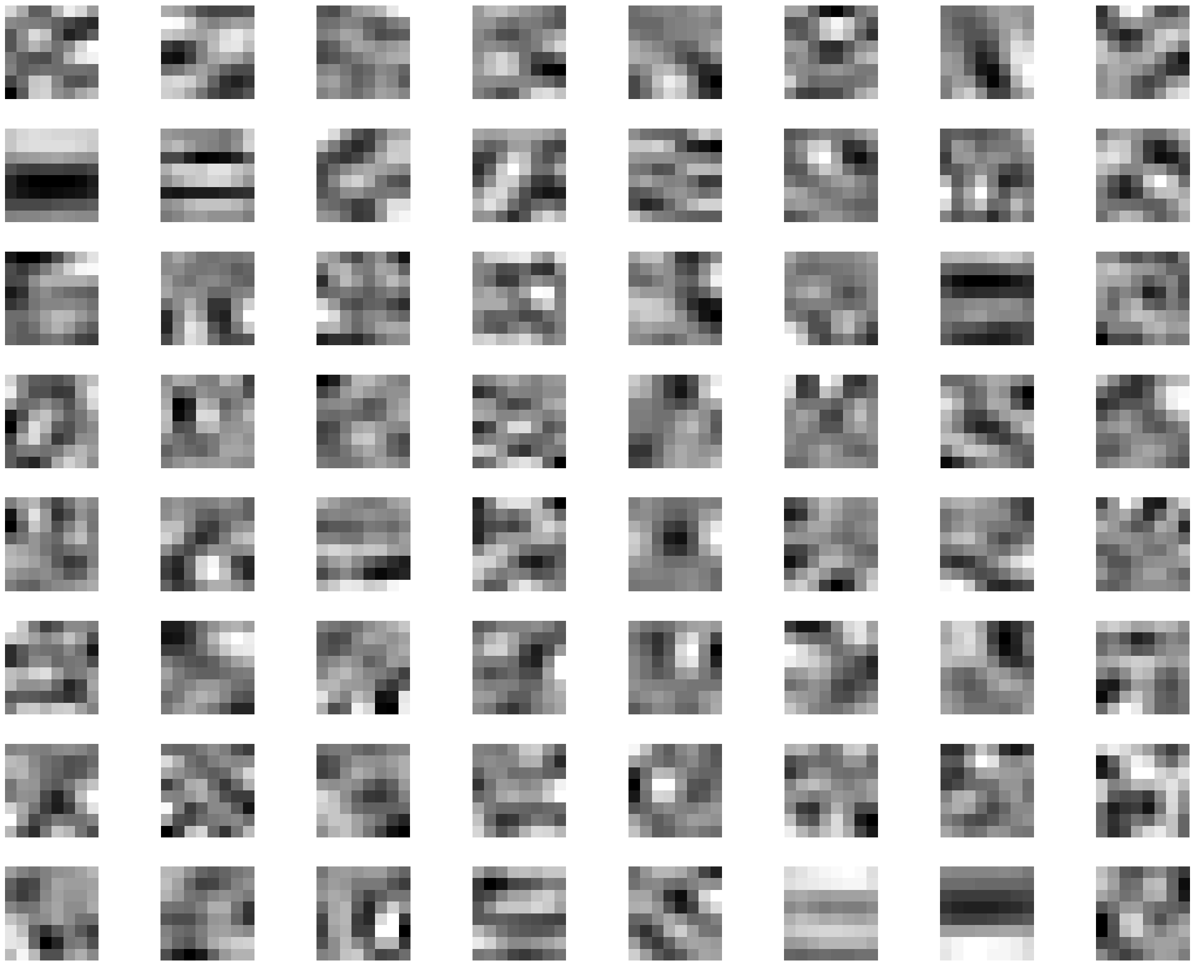


Figure 11: *Estimated ICA basis set for ultrasound breast image database*

Simple Example

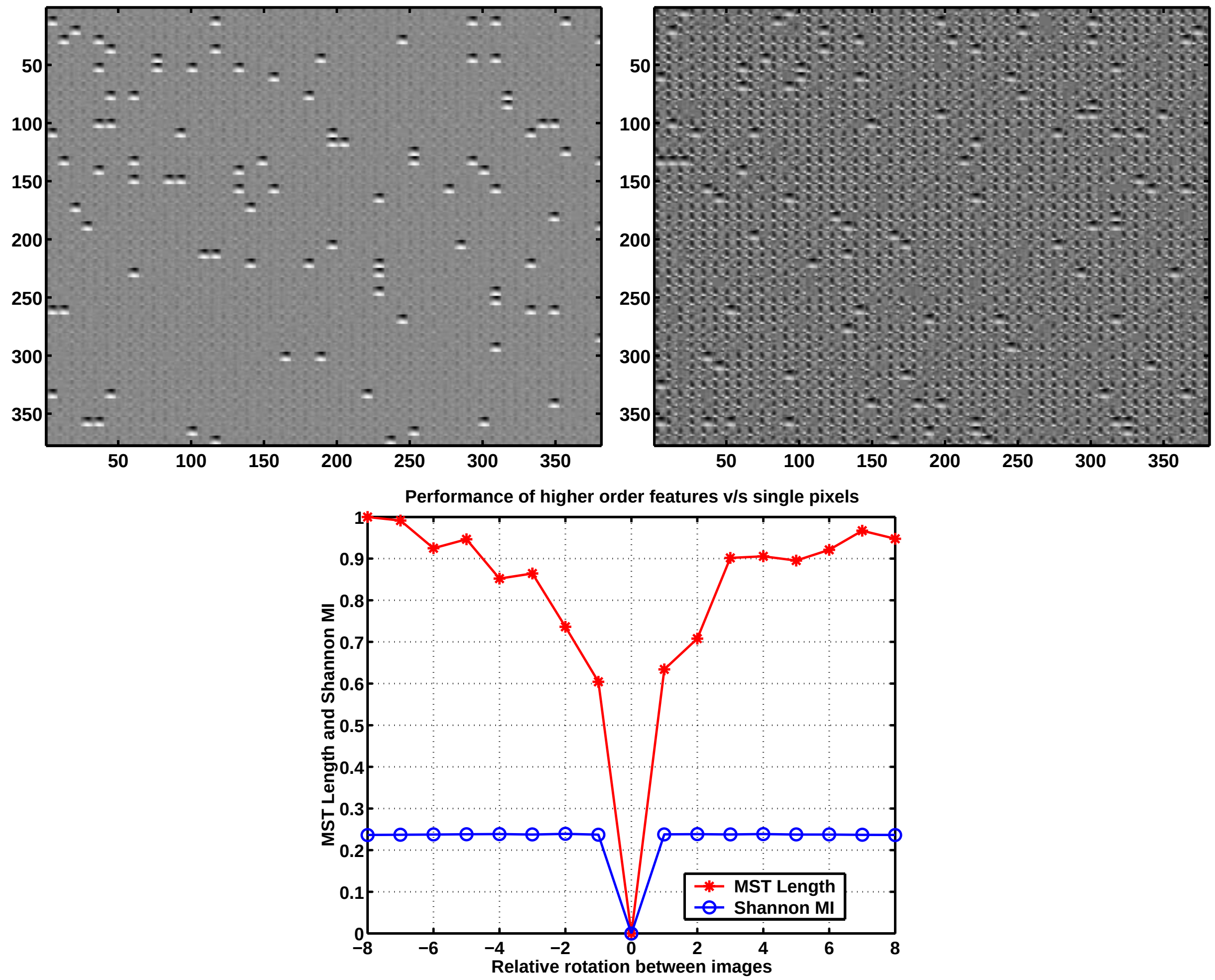


Figure 12: 10 basis composite images with noise and MI vs Feature divergence objective functions

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.

Feature-based Indexing: Challenges

- How to best select discriminating features?
  - *Require training database of images to learn feature set*
  - Apply cross-validation...
  - ...bagging, boosting, or randomized selection?
- How to compute α-MI for multi-dimensional features?
  - *Tag space is of high cardinality:* $256^{16} \geq 10^{32}$
  - *ICA projection-coefficient space is multi-dimensional continuum*
  - Soln 1: partition feature space and count coincidences...
  - Soln 2: apply kernel density estimation and ...
  - ...plug in to the α-MI or α-Jensen formula
  - Soln 3: estimate α-MI or α-Jensen directly via MST.

Methods of Entropy/Divergence Estimation

- $Z = (Z^R, Z^T)$: a statistic (feature pair)
- $\{Z_i\}$: n i.i.d. realizations from $f(Z)$

Objective: Estimate

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

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

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

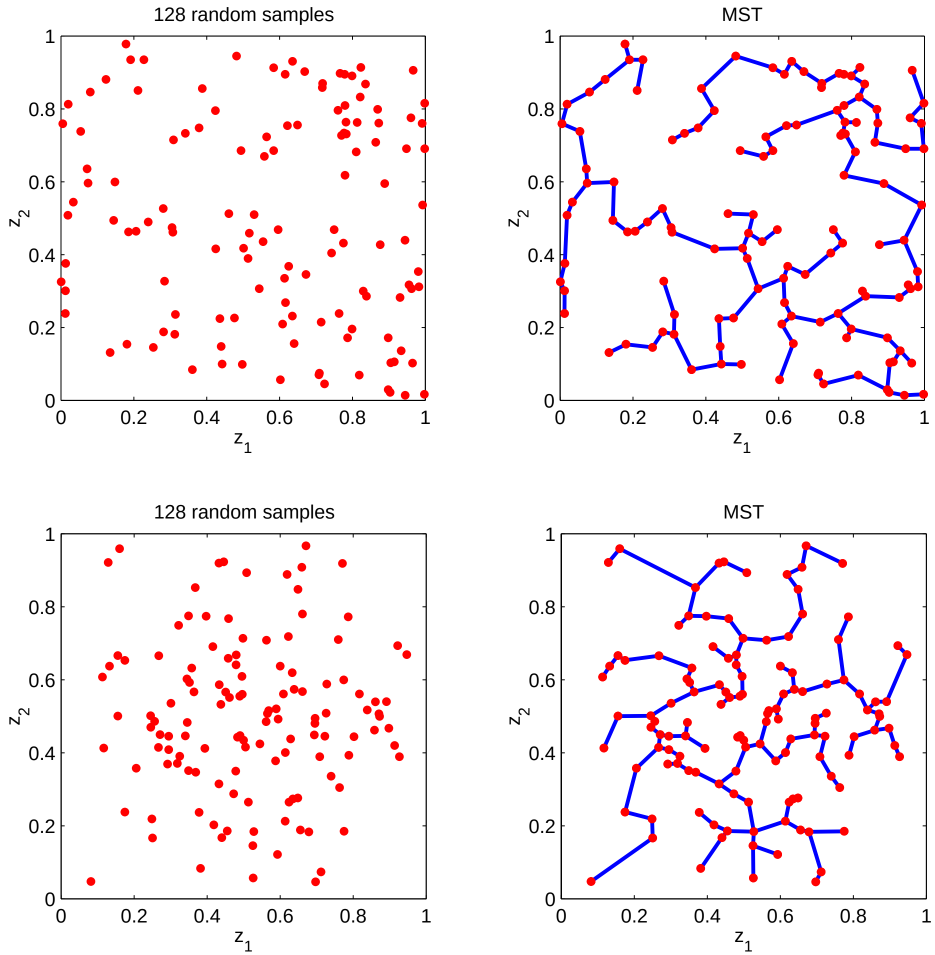

128 random samples
MST
128 random samples
MST
z_1
z_2
0
0.2
0.4
0.6
0.8
1

Figure 13:

Convergence of MST

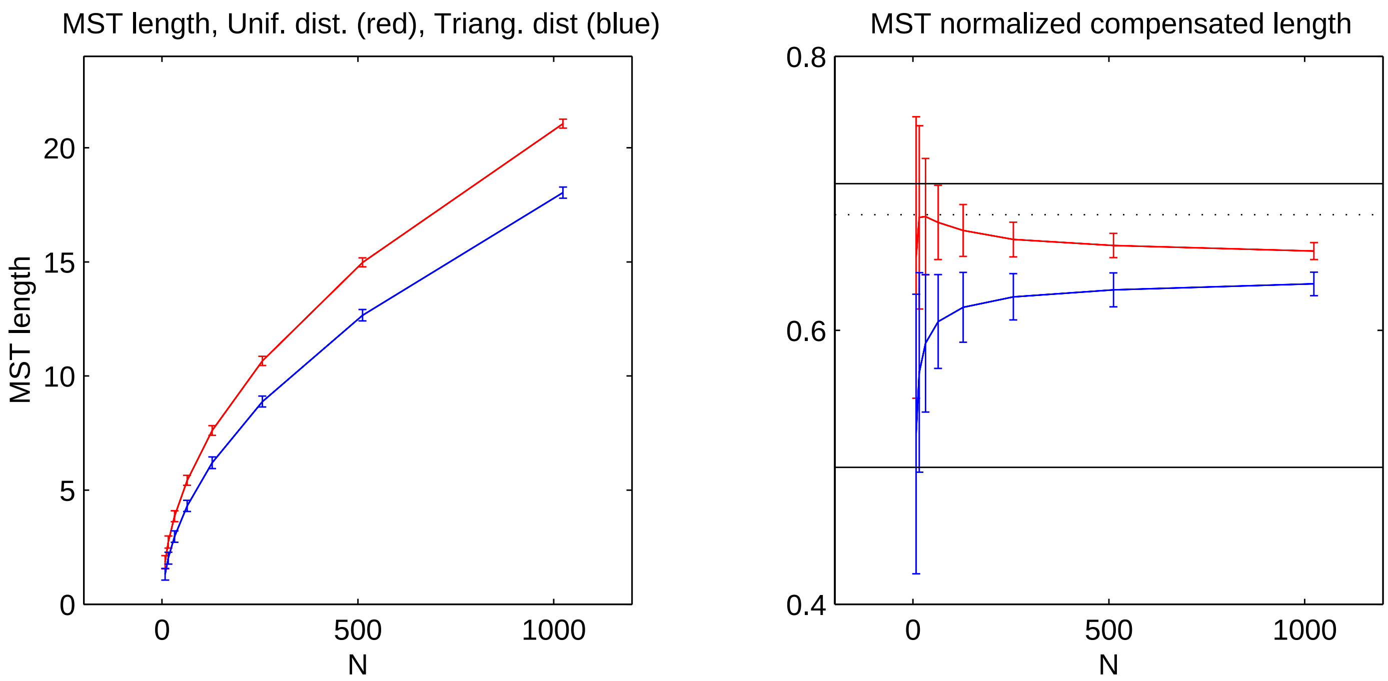


Figure 14:

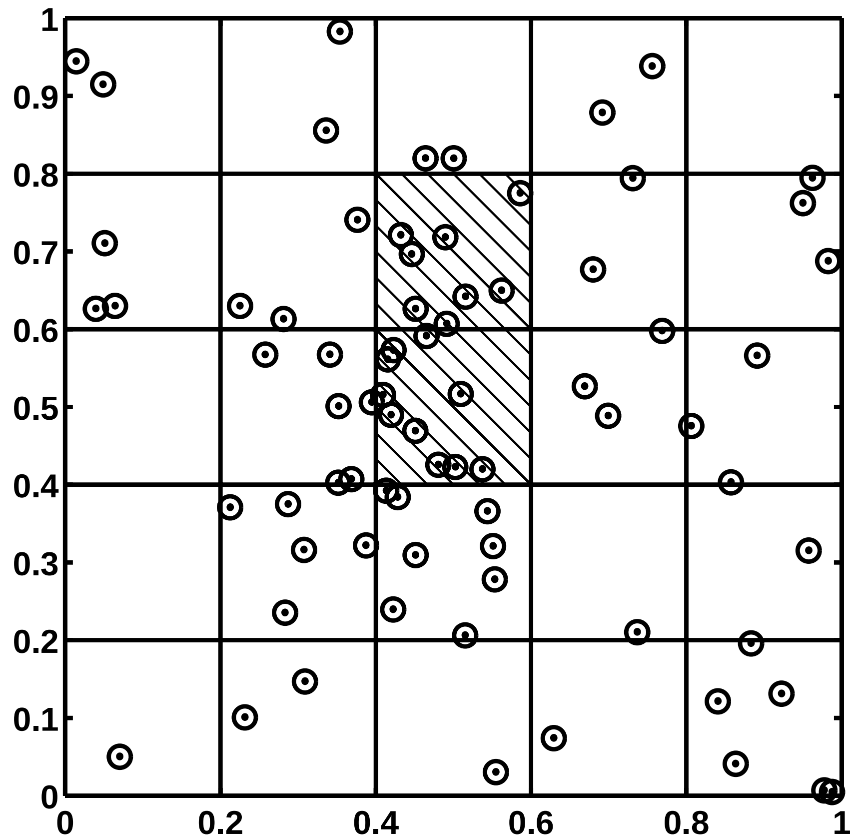


Figure 15: *Continuous quasi-additive euclidean functional satisfies "self-similarity" property on any scale.*

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 $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ 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*

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

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

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

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

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

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

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

$$\sup_{f^\alpha \in \Sigma_d(\kappa,c)} E^{1/p}\left[\left|\frac{L_\gamma(X_1,\ldots,X_n)}{n^\alpha} - \beta_{L_\gamma,d}\int f^\alpha(x)dx\right|^p\right] \leq O\left(n^{-2/(3d)}\right)$$

Conclude: minimal-graph estimator converges faster for

$$\kappa < \frac{2d}{3d-4}$$

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:

$$m^{-d} = n^{-1/3}$$

Computational Acceleration of MST

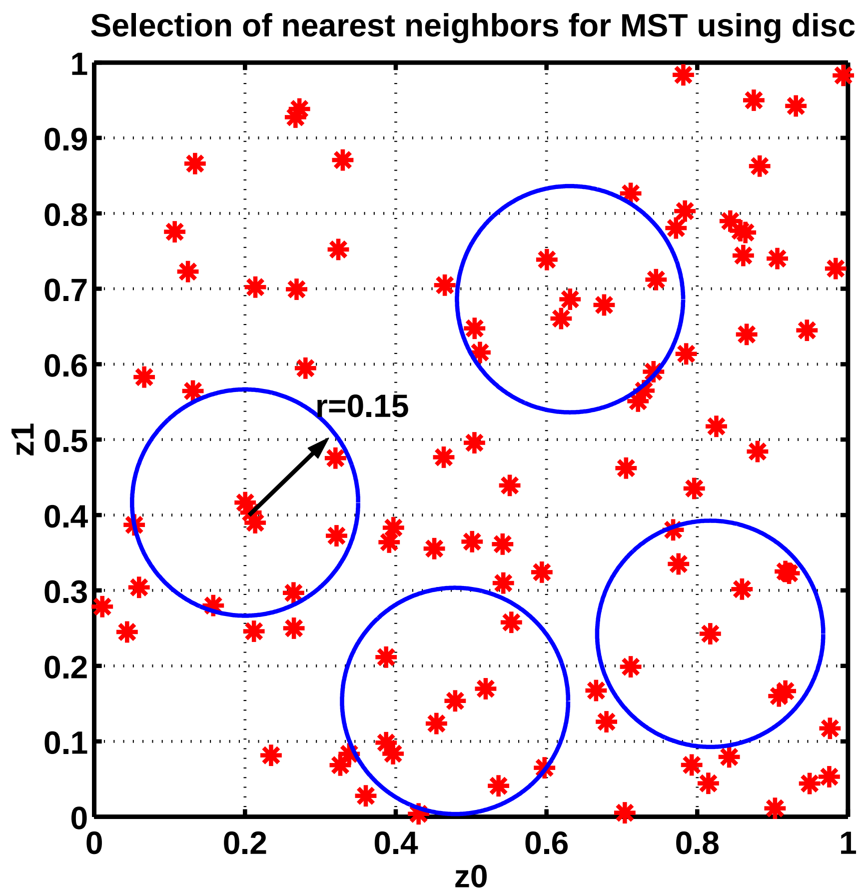


Figure 16: *Acceleration of Kruskal's MST algorithm from $n^2 \log n$ to $n \log n$.*

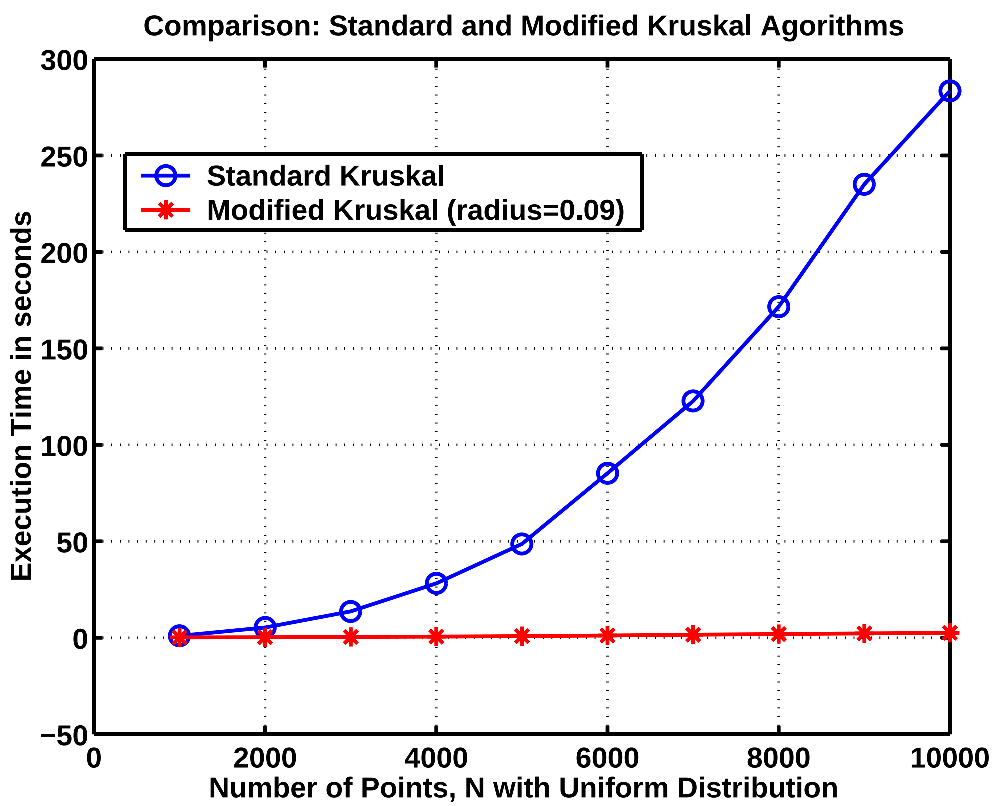


Figure 17: *Comparison of Kruskal's MST to our $n \log n$ MST algorithm.*

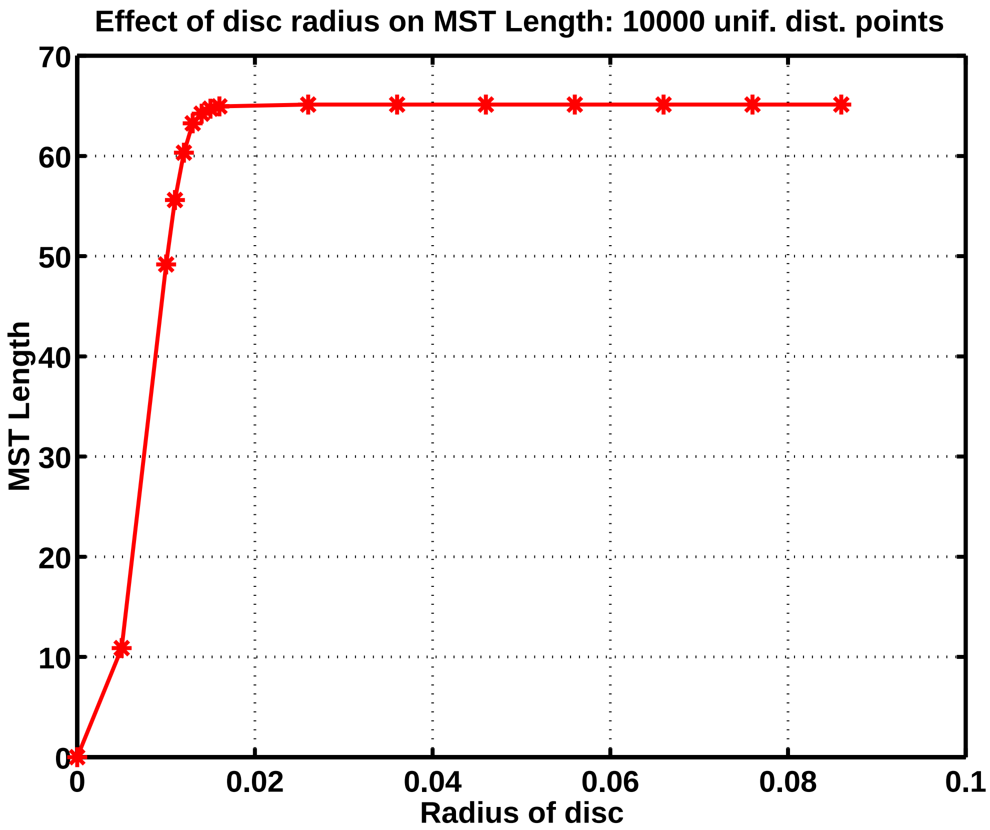


Figure 18: *Bias of $n \log n$ MST algorithm as function of radius parameter.*

Application of MST to US image Registration

1. Extract features from reference and transformed target images:

$$X_m = \{X_i\}_{i=1}^{m} \quad and \quad Y_n = \{Y_i\}_{i=1}^{n_y}$$

2. Construct MST on union of X_m and Y_n

$$L_\gamma(X_m \cup Y_n)$$

3. Minimize L_γ over transformations producing Y_n.

Note: This minimizer converges to minimizer of α-Jensen difference

$$L_\gamma(X_m \cup Y_n)/(m+n)^\alpha \;\to\; \beta_{L_\gamma,d} \exp\left((1-\alpha) H_\alpha(\varepsilon f_x + (1-\varepsilon) f_y)\right), \qquad (a.s.)$$

where $\varepsilon = \frac{m}{m+n}$

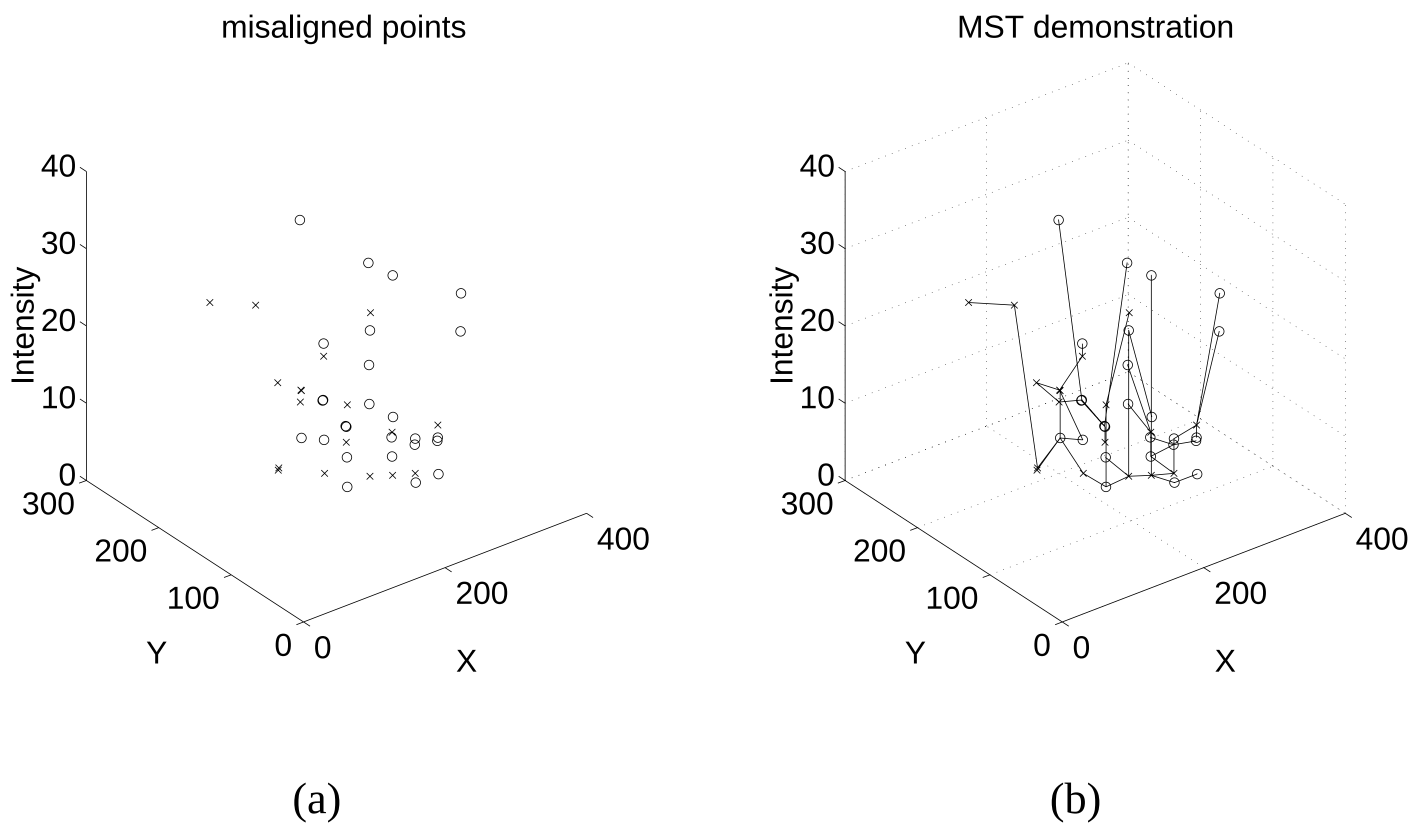
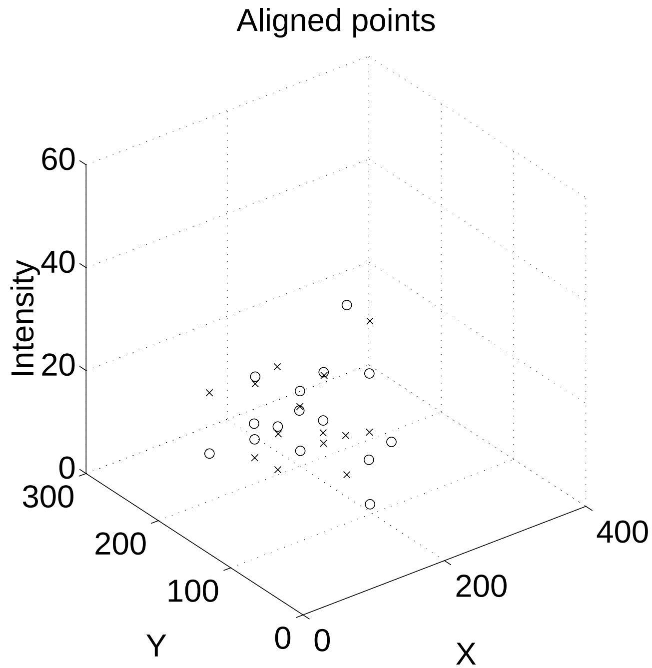
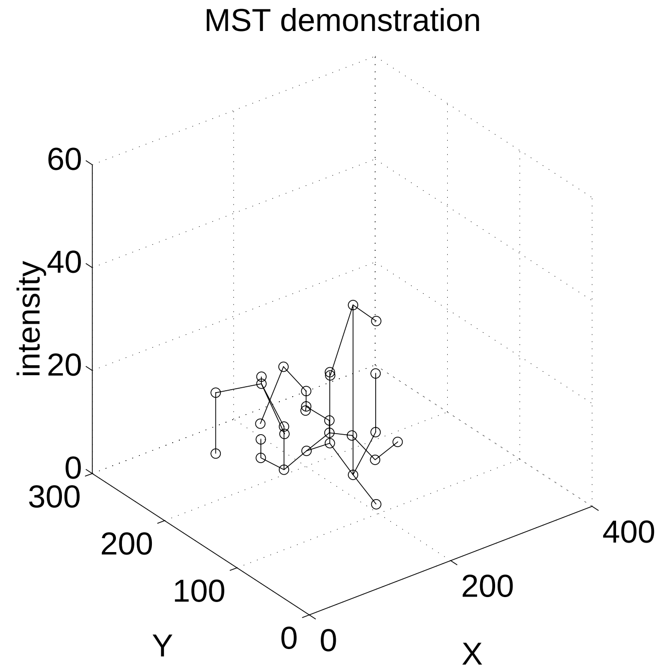


Figure 19: MST demonstration for misaligned images



(a)



(b)

Figure 20: MST demonstration for aligned images

Illustration for Case 142 and Single Pixels

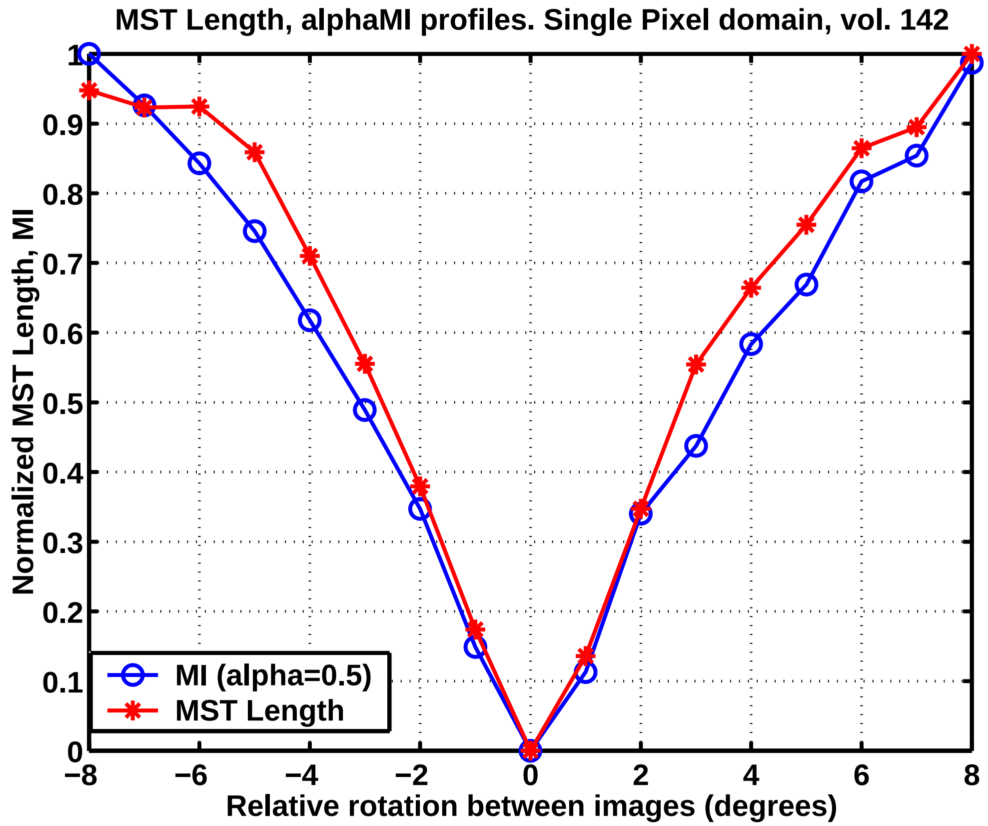


Figure 21: *Single-pixel objective function profiles for MST estimator of α-Jensen difference vs histogram plug-in estimator ($\alpha = 1/2$).*

Illustration for Case 142 and 8-D ICA Features

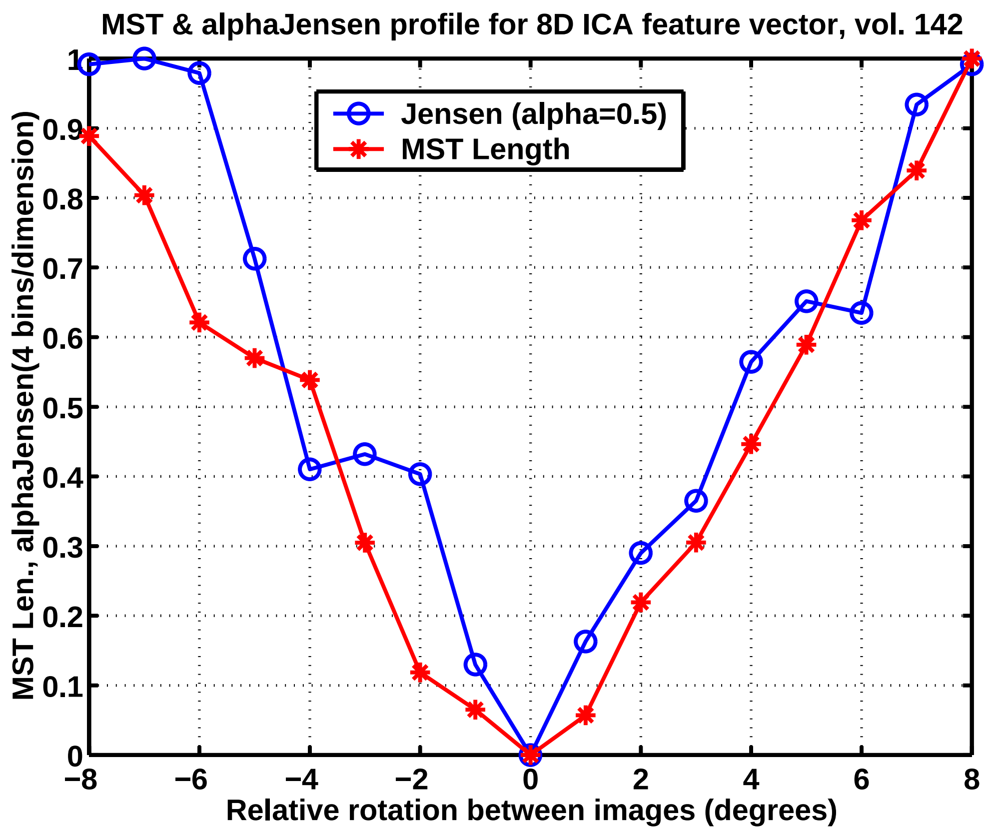


Figure 22: *8-basis ICA objective function profiles for MST estimator of α-Jensen difference vs histogram plug-in estimator of α-MI ($\alpha = 1/2$).*

Illustration for Case 142 and 64-D ICA Features

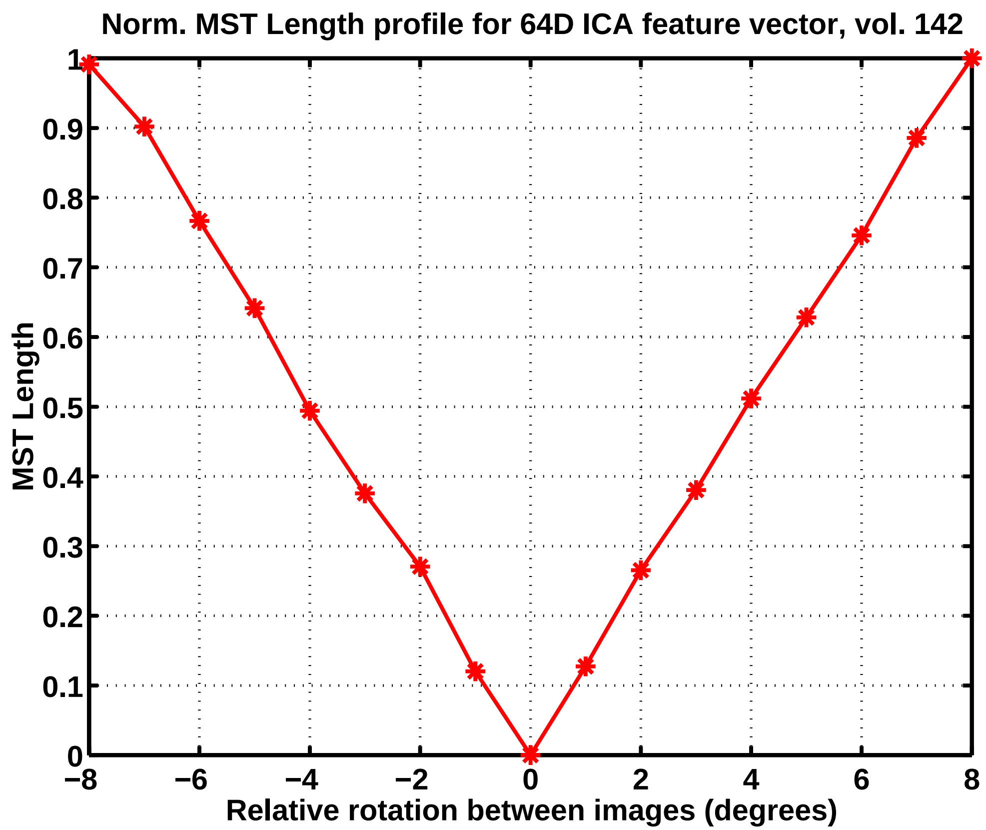


Figure 23: *64-basis ICA objective function profiles for MST estimator of α-Jensen difference.*

Extension of MST to divergence estimation

1. Let i.i.d. $\{Z_i\}_{i=1}^n$ have marginal density f_1 on $[0,1]^d$

2. Let f_0 dominate density f_1

3. Define measure transformation M on $[0,1]^d$ which takes f_0 to uniform density

Then $S_i \stackrel{\text{def}}{=} M(Z_i)$ has α-entropy:

$$(1-\alpha)H_\alpha(S_i) = \ln \int f_S^\alpha(s)ds = \ln \int \left(f_1(z)/f_0(z)\right)^\alpha f_0(z)dz$$

Conclude, for $\mathcal{S}_n = \{S_1, \ldots, S_n\}$:

$$\hat{D}_\alpha(f_1 \| f_0) = \frac{1}{1-\alpha}\left[\ln L_\gamma(\mathcal{S}_n)/n^\alpha - \ln \beta_{L,\gamma}\right] . \quad (2)$$

is a consistent estimator of α-divergence

Example

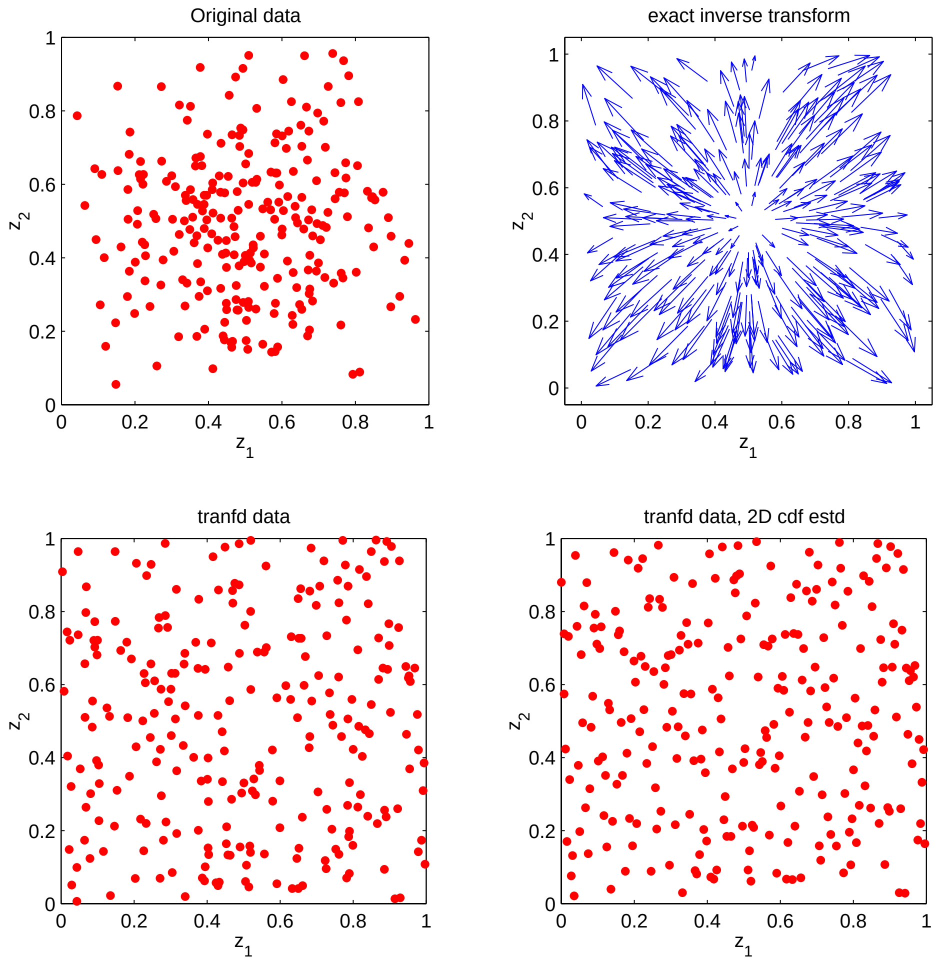


Figure 24:

MST robustification against outliers: 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

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

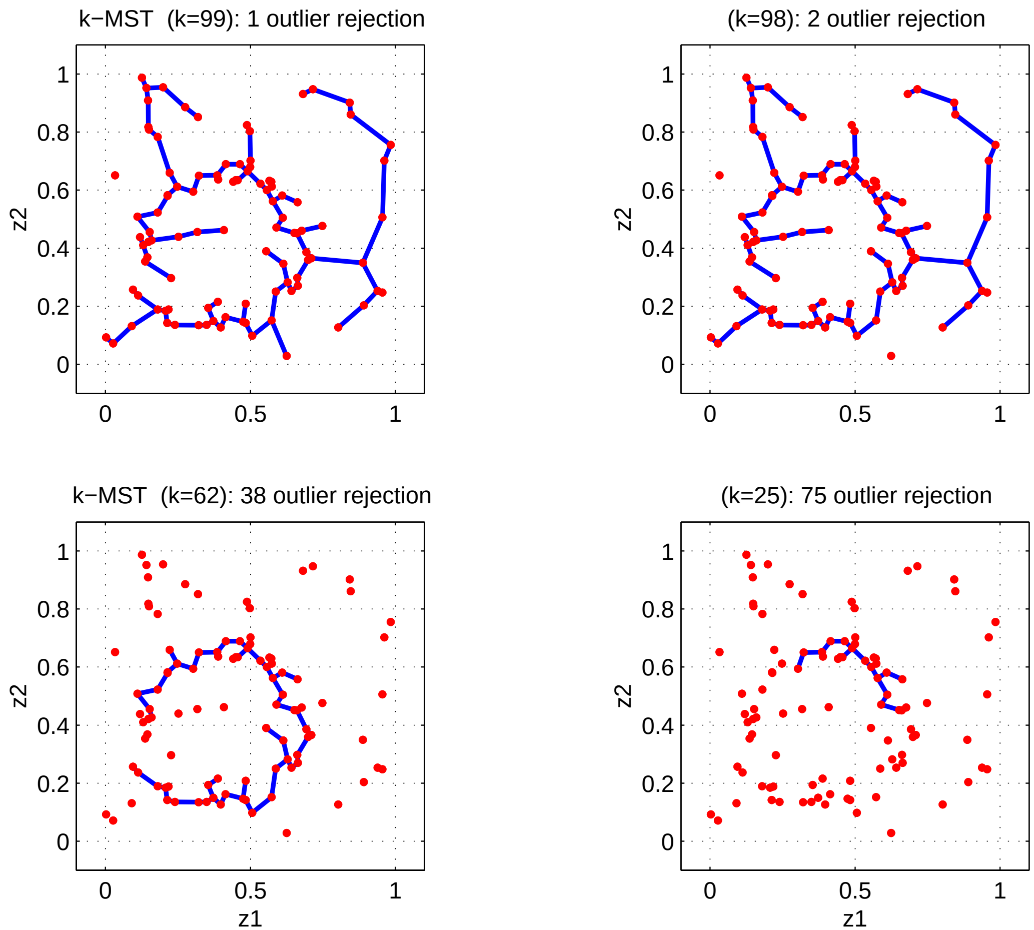


Figure 25: *k-MST for 2D annulus density with and without the addition of uniform "outliers".*

Extension of BHH 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

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

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

Water Pouring Construction

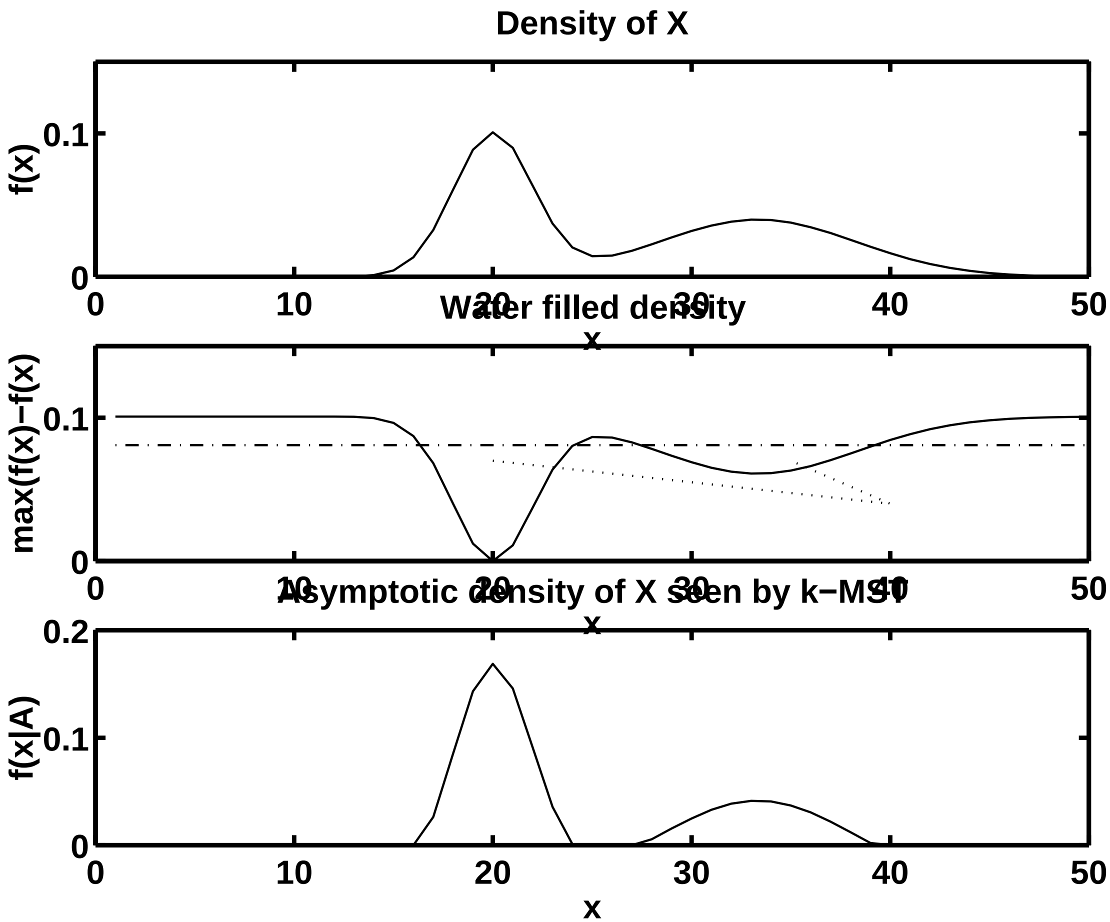


Figure 26: *Water pouring construction of* $f(x|A_o)$*. Arrows indicate the high water marks indicating high probability regions which are not truncated in the influence function.*

k-MST Influence Function for Gaussian Density

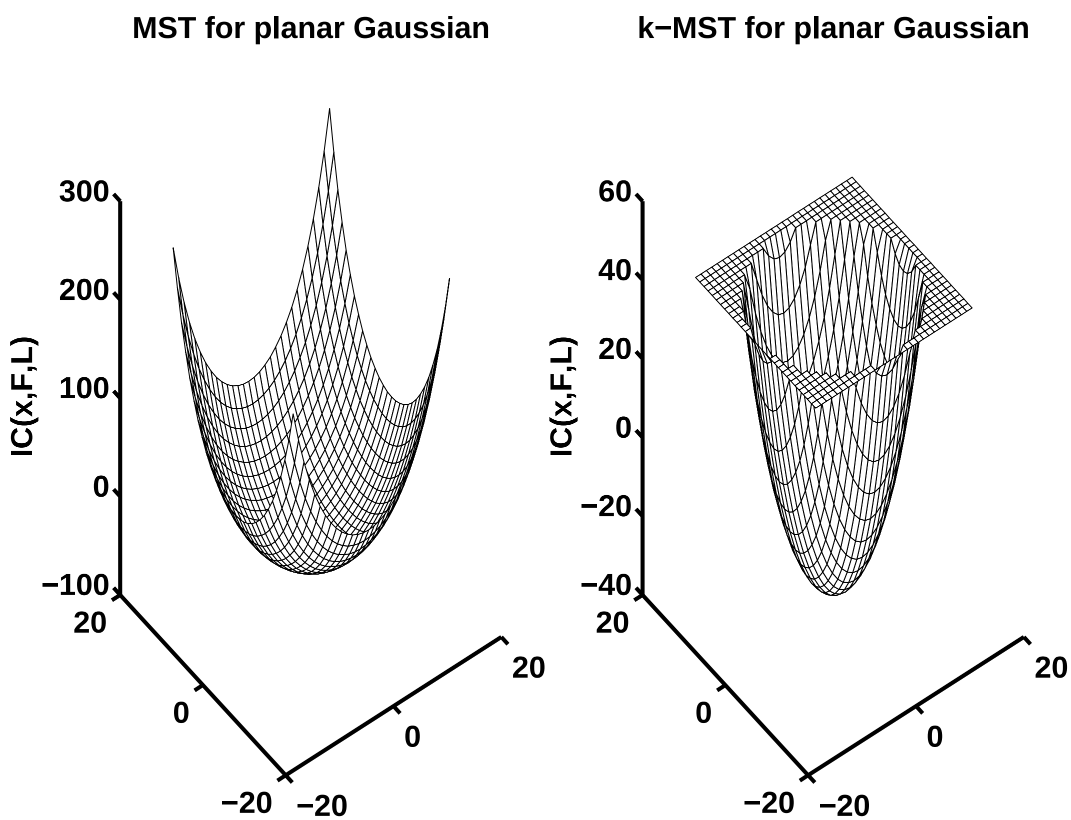


Figure 27: *MST and k-MST influence curves for Gaussian density on the plane.*

k-MST Stopping Rule

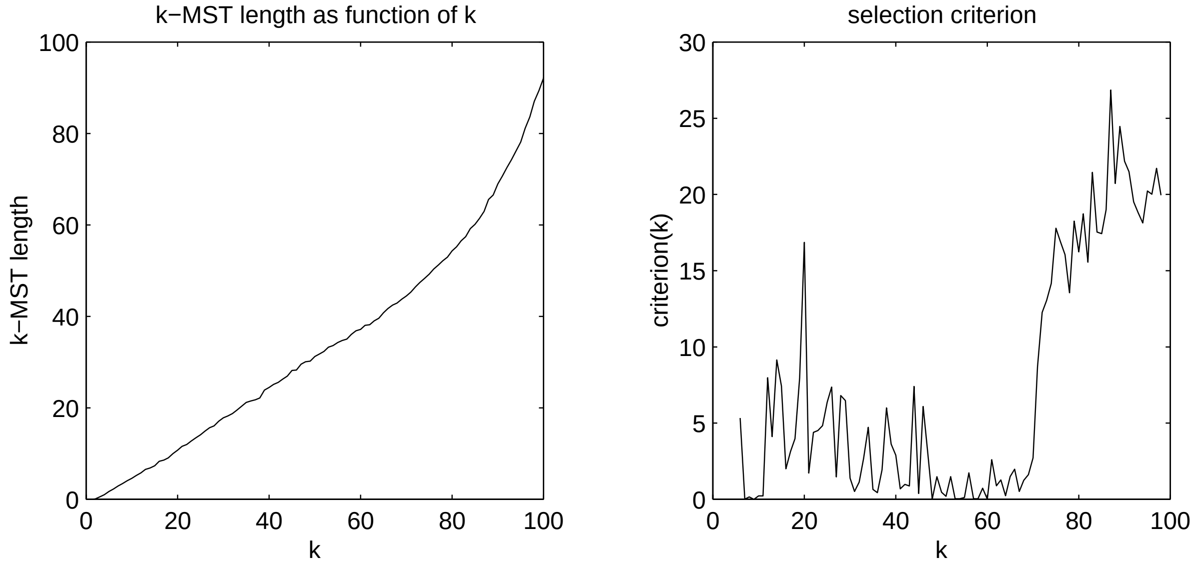


Figure 28: *Left: k-MST curve for 2D annulus density with addition of uniform "outliers" has a knee in the vicinity of* $n-k=35$*. This knee can be detected using residual analysis from a linear regression line fitted to the left-most part of the curve. Right: error residual of linear regression line.*

Conclusions

1. α-divergence for indexing can be justified via decision theory
2. Applicable to feature-based image registration
3. Non-parametric estimation is possible without density estimation via MST
4. MST outperforms plug-in estimation for non-smooth densities
5. Robustified MST can be defined via optimal pruning of MST: k-MST

Divergence vs. Jensen: Asymptotic Comparison

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

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

Then

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

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