Entropic-graphs: Applications

Alfred O. Hero

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

University of Michigan - Ann Arbor

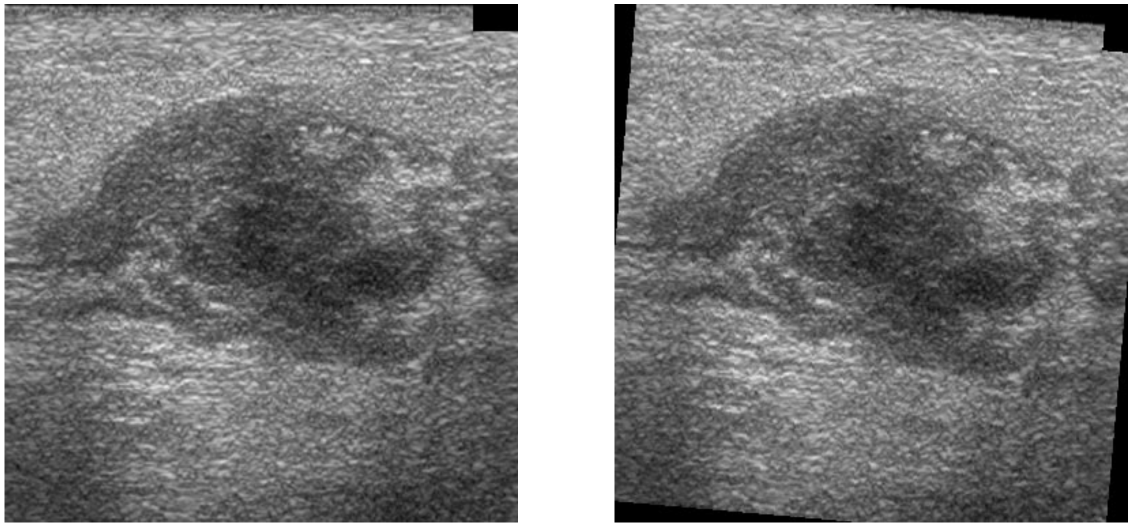
hero@eecs.umich.edu

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

Collaborators: Huzefa Heemuchwala, Jose Costa, Bing Ma, Olivier Michel

- Image registration
- Multivariate outlier rejection
- Divergence estimation

Image Registration



(a) Image X_0 (b) Image X_i

Figure 1: A multidate 3D breast-registration example

Range of UL breast Image Types

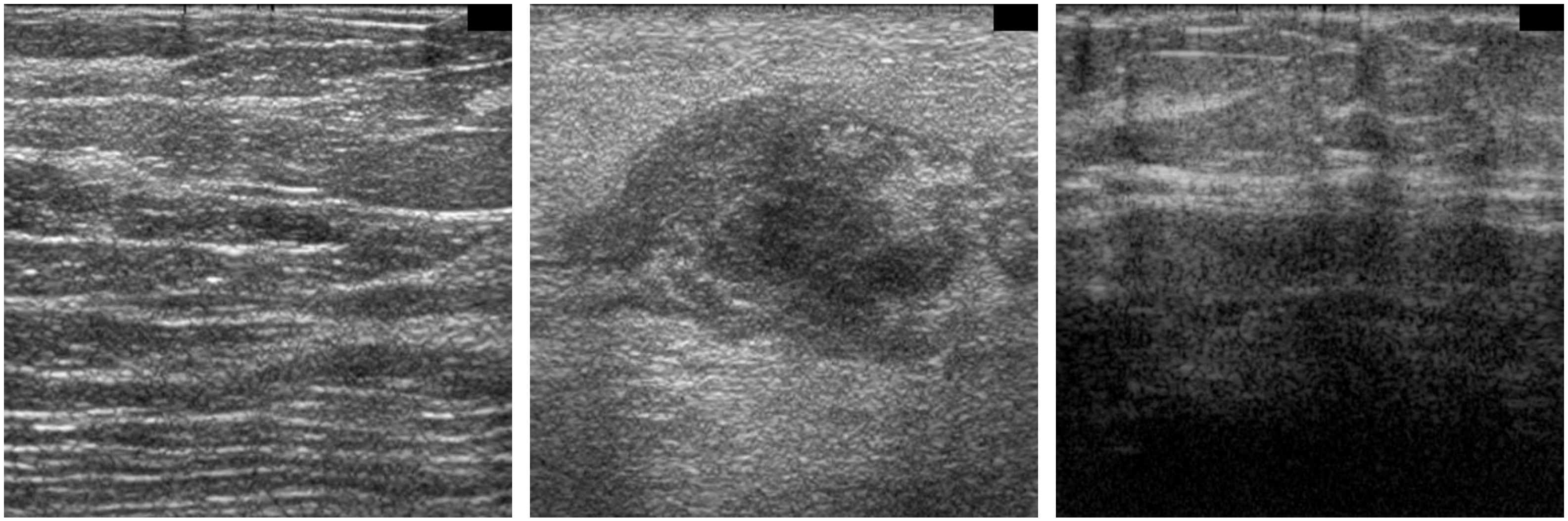


Figure 2: Three ultrasound breast scans. From top to bottom are: case 151, case 142 and case 162.

MI Registration of Gray Levels (Viola&Wells:ICCV95)

- X: a $N \times N$ UL image (lexicographically ordered)
- $X(k)$: image gray level at pixel location k
- X_0 and X_1: primary and secondary images to be registered

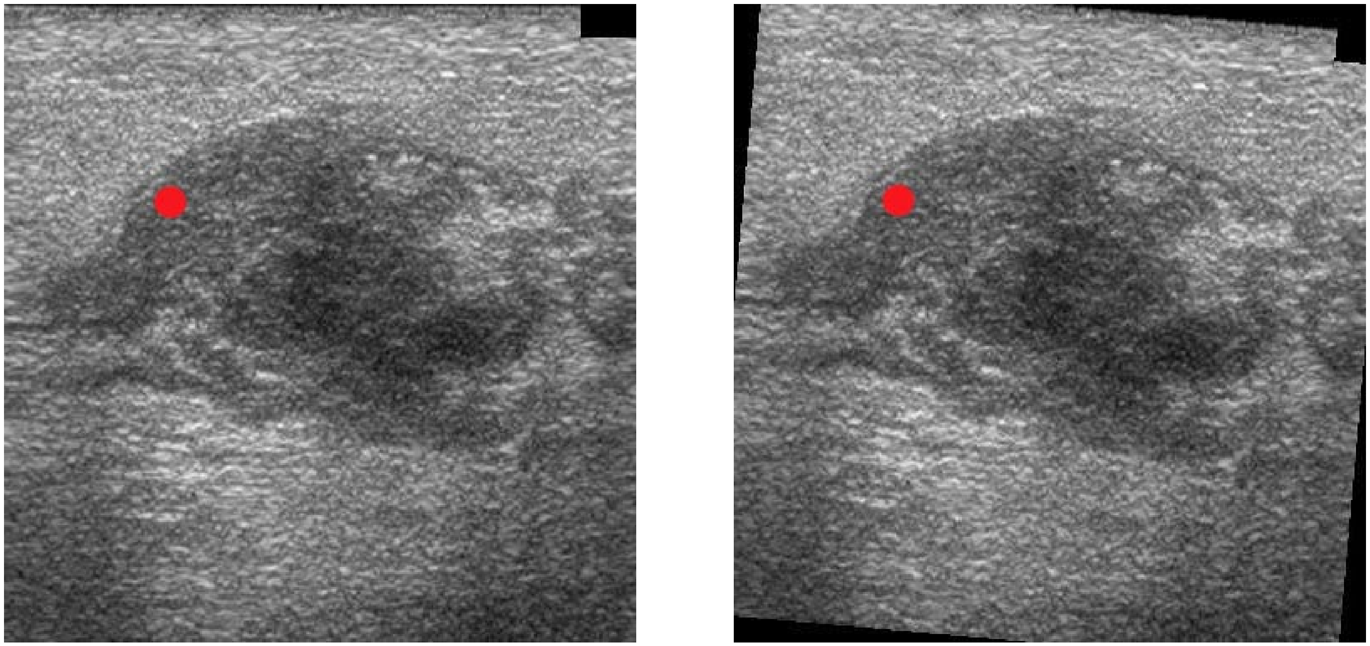
Hypothesis: $\{(X_0(k), X_i(k)\}_{k=1}^{N^2}$ are i.i.d. r.v.'s with j.p.d.f

$$f_{0,i}(x_0, x_1), \quad x_0, x_1 \in \{0, 1, \ldots, 255\}$$

Mutual Information (MI) criterion: $\mathrm{T} = \mathrm{argmax}_{\mathrm{T}_i} \hat{\mathrm{MI}}$

where $\hat{\mathrm{MI}}$ is an estimate of

$$\mathrm{MI}(f_{0,i}) = \int\int f_{0,i}(x_0, x_1) \ln f_{0,i}(x_0, x_1) / (f_0(x_0) f_i(x_1)) dx_1 dx_0. \qquad (1)$$



(a) Image I^R (b) Image I^T

Figure 3: Single Pixel Coincidences (Left and right: reference image I^R at 0^o and rotated image I^T at 8^o)

Single-Pixel Scatterplot $(Z_j^R, Z_j^T)_{j=1}^p$

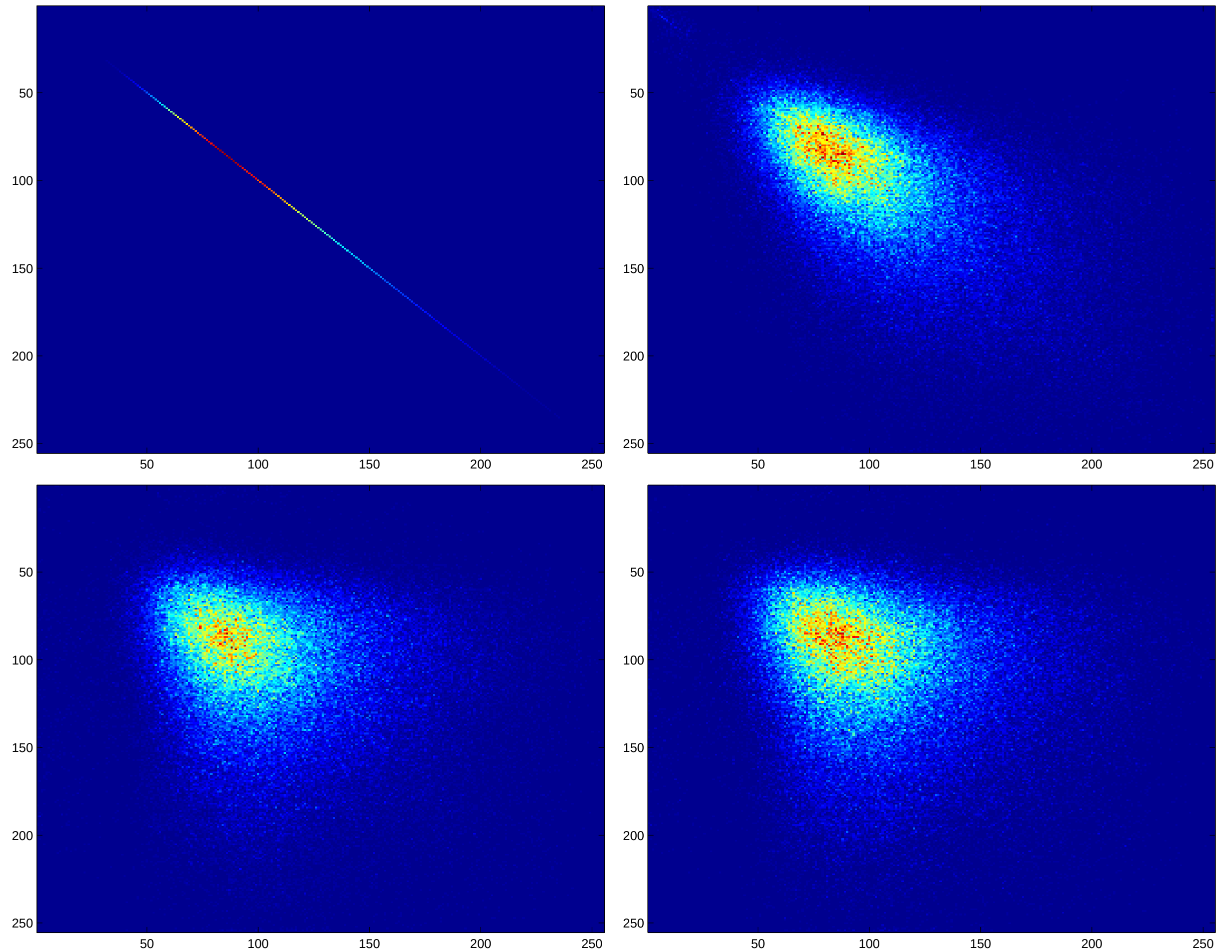


Figure 4: Grey level scatterplots. 1st Col: target=reference slice. 2nd Col: target = reference+1 slice.

α-MI Registration of Coincident Features

- X: a $N \times N$ UL image (lexicographically ordered)
- $Z = Z(X)$: a general image feature vector in a P-dimensional feature space

Let $\{Z_0(k)\}_{k=1}^K$ and $\{Z_i(k)\}_{k=1}^K$ be features extracted from X_0 and X_i at K identical spatial locations

α-MI coincident-feature criterion

$$\mathrm{T} = \mathrm{argmax}_{\mathrm{T}_i} \hat{\mathrm{MI}}_\alpha$$

where $\hat{\mathrm{MI}}_\alpha$ is an estimate of

$$\mathrm{MI}_\alpha(f_{0,i}) = \frac{1}{\alpha - 1} \log \int\int f_{0,i}^\alpha(z_0, z_1) f_0^{1-\alpha}(z_0) f_i^{1-\alpha}(z_1) dz_1 dz_0. \tag{2}$$

Why α-MI?

Special cases:

- α-MI vs. Shannon MI

$$\lim_{\alpha \to 1} \mathrm{MI}_\alpha(f_{0,i}) = \int\int f_{0,i} \ln f_{0,i}/(f_0 f_i) dz_1 dz_0.$$

- α-MI vs. Hellinger Mutual Affinity

$$\mathrm{MI}_{\frac{1}{2}}(f_{0,i}) \quad = \quad -\ln\left(\int\int \sqrt{f_{0,i} f_0 f_i}\, dz_0 dz_1\right)^2$$

- α-MI vs. Batthacharyya-Hellinger information

$$\int\int \left(\sqrt{f_{0,i}} - \sqrt{f_0 f_i}\right)^2 dz_0 dz_1 = 2\left(1 - \exp\{-\mathrm{MI}_{\frac{1}{2}}(f_{0,i})\}\right)$$

α-MI and Decision Theoretic Error Exponents

$$
\begin{aligned}
H_0 &: \quad Z_0(k), Z_i(k) \ \text{ independent} \\
H_1 &: \quad Z_0(k), Z_i(k) \ \text{ o.w.}
\end{aligned}
$$

Bayes probability of error

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

Chernoff bound

$$
\liminf_{n\to\infty} \frac{1}{n} \log P_e(n) = - \sup_{\alpha\in[0,1]} \left\{(1-\alpha)\mathrm{MI}_\alpha(f_{0,i})\right\}.
$$

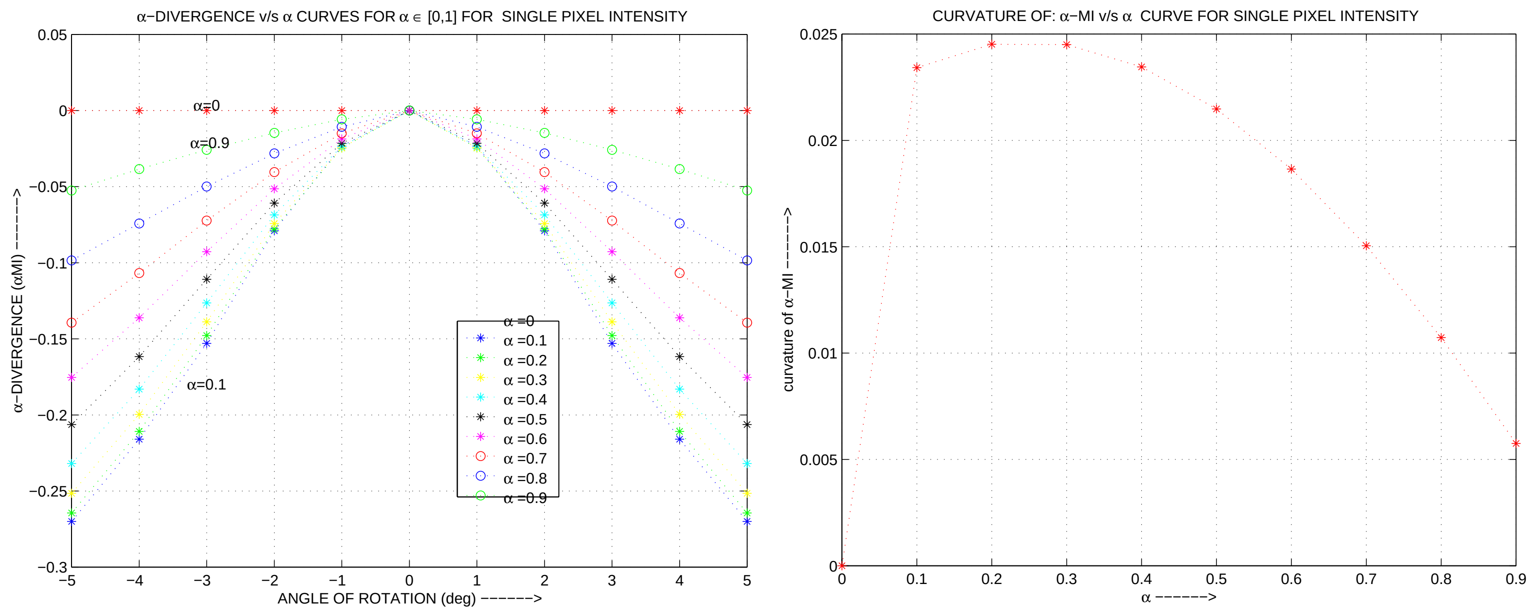


Figure 5: Left: α-Divergence as function of angle. Right: Resolution of α-Divergence as function of alpha

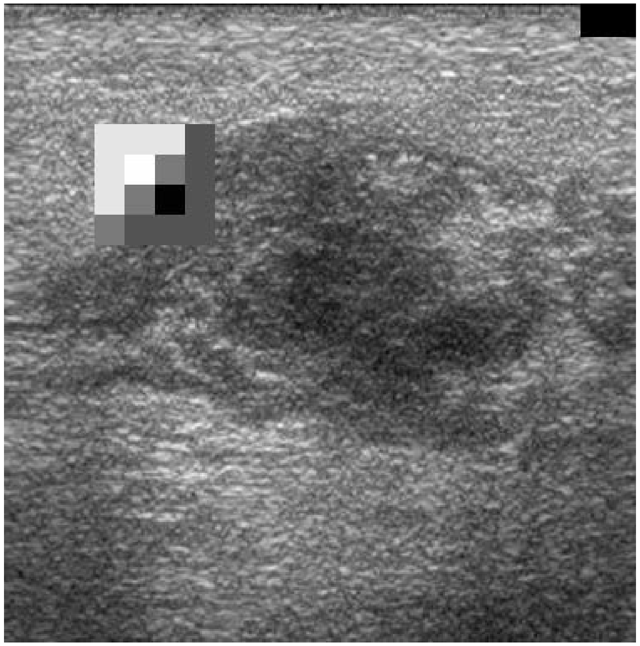
Higher Level Features

Disadvantages of single-pixel features:

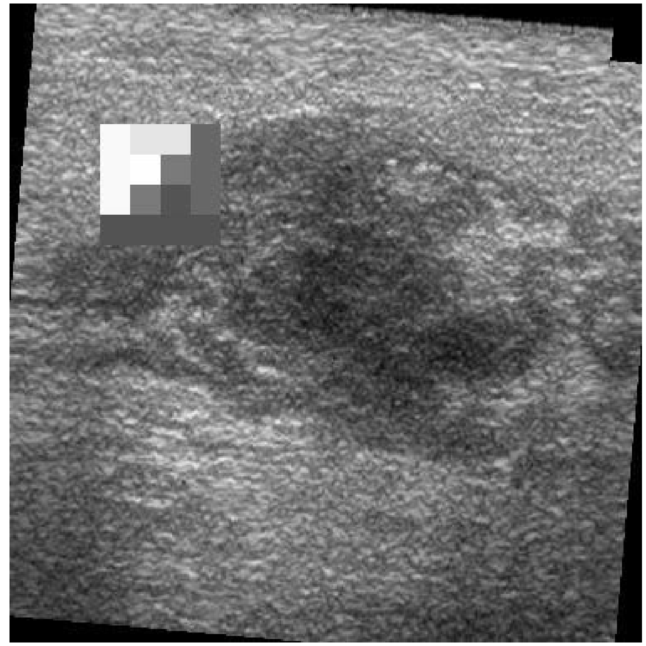
- Only depends on histogram of single pixel pairs
- Insensitive to spatial reording of pixels in each image
- Difficult to select out grey level anomalies (shadows, speckle)
- Spatial discriminants fall outside of single pixel domain

- **Alternative**: Aggregate spatial features

Local Tags



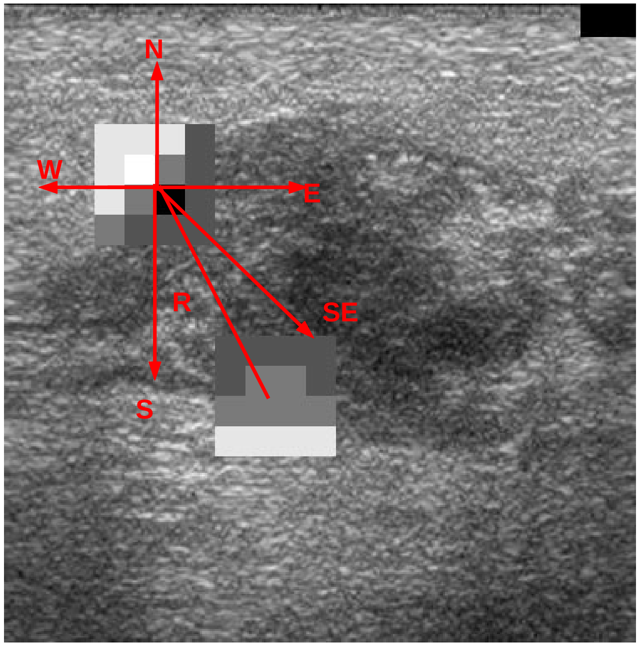
(a) Image I^R



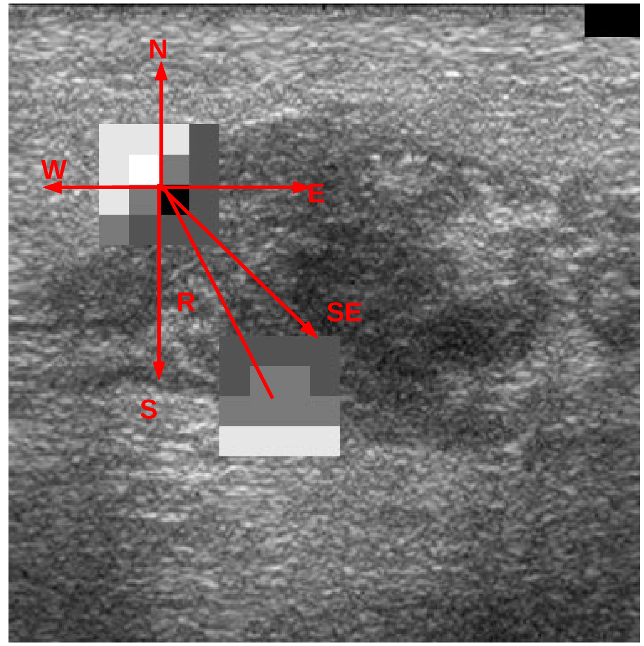
(b) Image I^T

Figure 6: Local Tag Coincidences

Spatial Relations Between Local Tags



(a) Image I^R



(b) Image I^T

Figure 7: Spatial Relation Coincidences

Feature Coincidence Tree of Local Tags

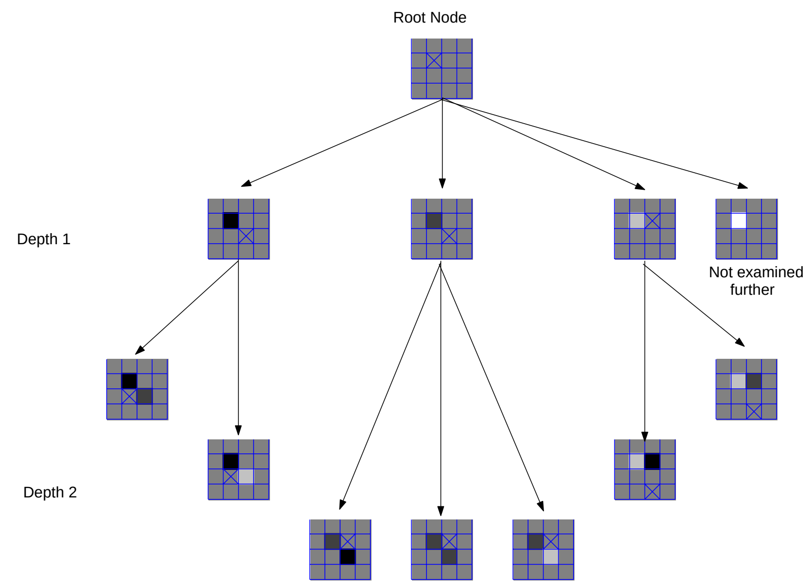


Figure 8: *Part of feature tree data structure.*

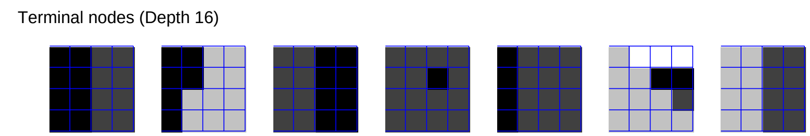


Figure 9: *Leaves of feature tree data structure.*

Forests of Randomized Feature Trees

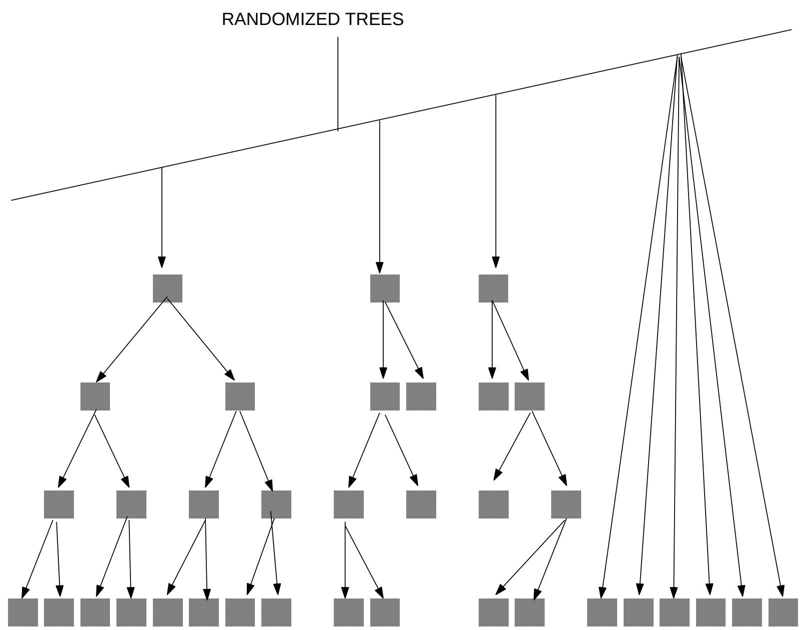


Figure 10: *Forest of randomized trees*

Registration criterion:

$$\mathrm{T} = \mathrm{argmax}_{\mathrm{T}_i} \sum_{t=1}^{\#\,trees} \hat{\mathrm{MI}}_{\alpha}(t)$$

US Registration Comparisons

	151	142	162
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

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).

ICA Features

Decomposition of $M \times M$ tag images $Y(k)$ acquired at $k = 1, \ldots, K$ spatial locations

$$Y(k) = \sum_{p=1}^{P} a_{kp} S_p$$

- $\{S_k\}_{k=1}^{P}$: statistically independent components
- a_{kp}: projection coefficients of tag $Y(k)$ onto component S_p
- $\{S_k\}_{k=1}^{P}$ and P: selected via MLE and MDL
- Feature vector for coincidence processing:

$$Z(k) = [a_{k1}, \ldots, a_{kP}]^T$$

ICA feature basis for US breast images

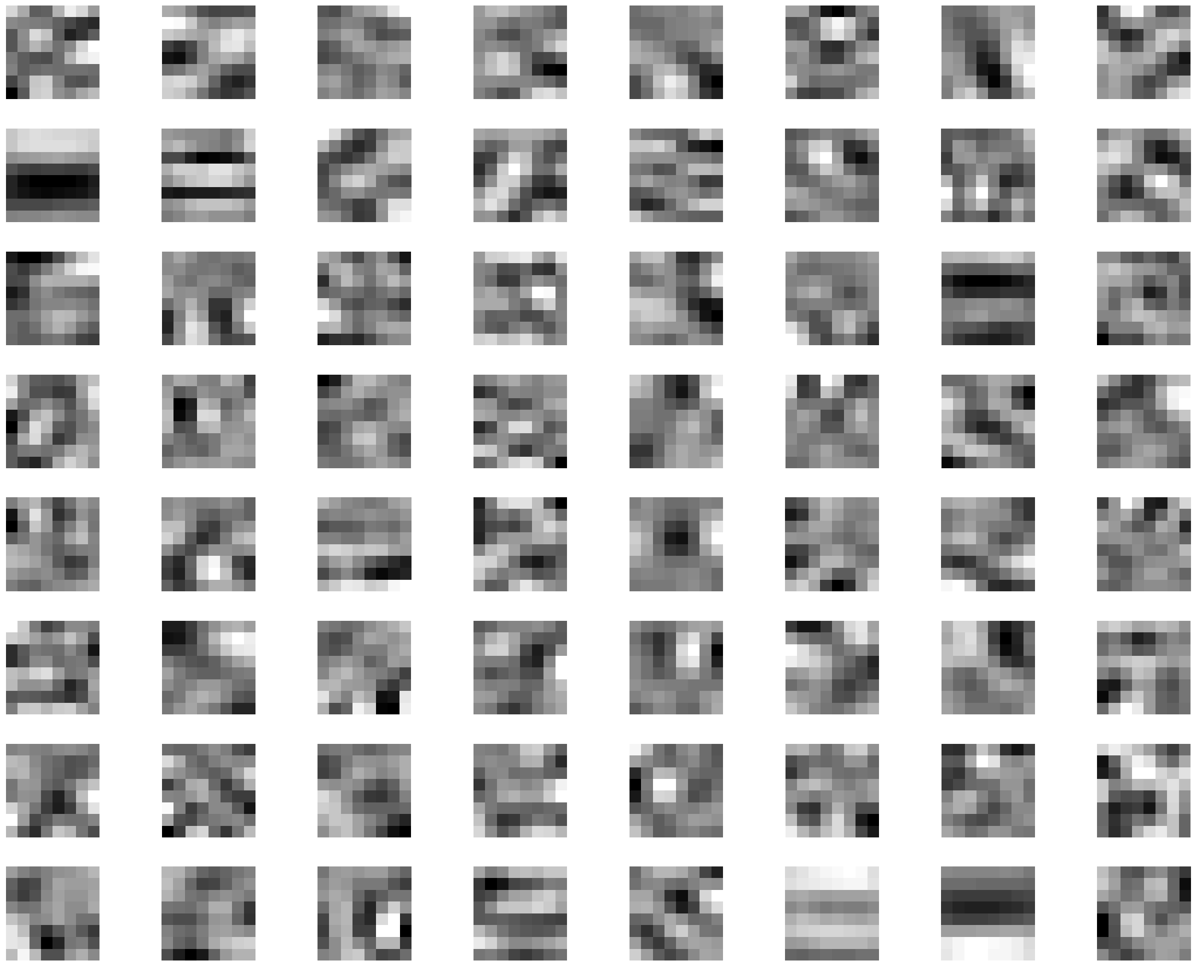


Figure 11: *Estimated ICA basis set for ultrasound breast image database*

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 density estimation and ...
  - ... plug into the α-MI
  - Soln 3: estimate α-MI directly

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 “plug-in” methods
3. Non-parametric minimal-graph estimation methods

Minimal Graphs: Minimal Spanning Tree (MST)

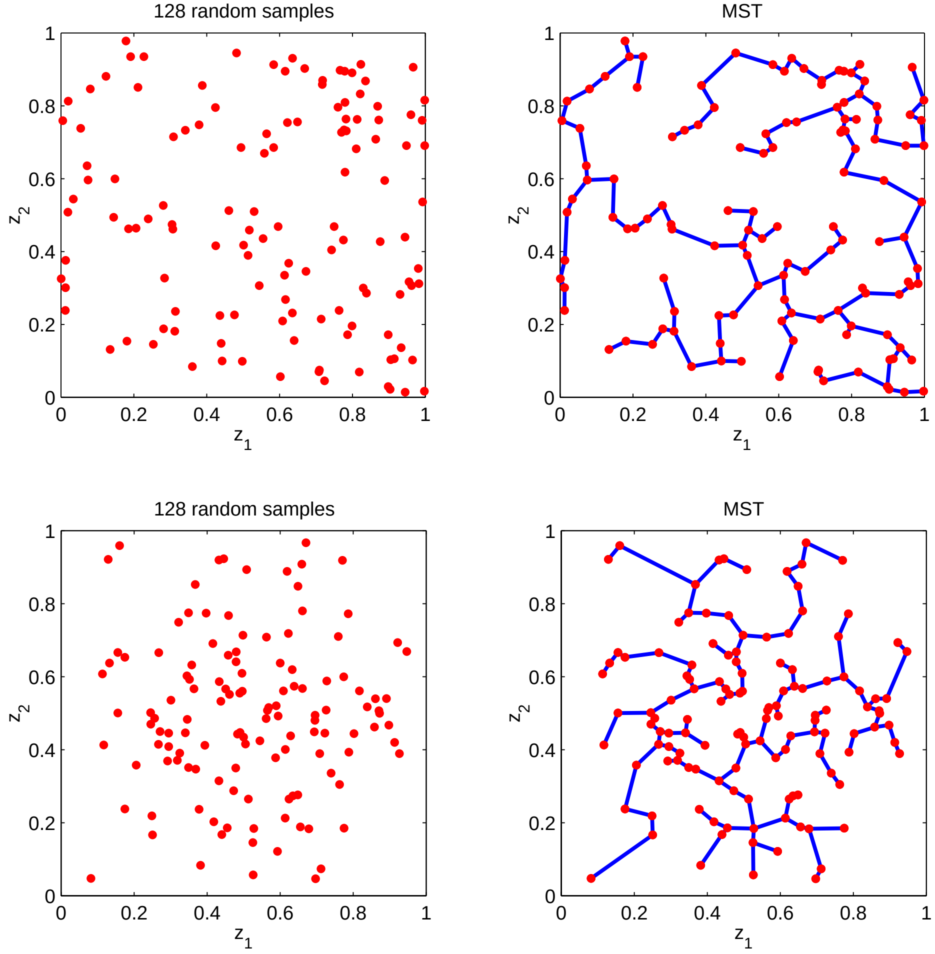


Figure 12:

Asymptotics of estimators of $H_\alpha(f)$

Define $B_p^{\sigma,q}$, the Besov space of $\ell_p(\mathbf{R}^d)$ functions with smoothness given by parameters σ and q.

Proposition 1 *Let $p > d \geq 2$ and $\alpha = (d-\gamma)/d \in [1/2, (d-1)/d]$*

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

while,

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

where $\lambda(p) = d+1-d/p$.

Note: minimal-graph estimator converges faster for all $\alpha \geq 1/2$

Extension: divergence estimation

- $g(x)$: a reference density on $\mathbf{R}^d$
- Assume $f \ll g$, i.e. for all x such that $g(x) = 0$ we have $f(x) = 0$.
- Make measure transformation $dx \to g(x)dx$ on $[0,1]^d$. Then for Y_n = transformed data

$$\lim_{n\to\infty} L(Y_n)/n^{\alpha} \quad = \quad \beta_{L_{\gamma},d} \, \exp\left((\alpha-1)D_{\alpha}(f\|g)\right), \qquad (a.s.)$$

Proof

1. Make transformation of variables $x = [x^1, \ldots, x^d]^T \rightarrow y = [y^1, \ldots, y^d]^T$

$$\begin{aligned} y^1 &= G(x^1) \\ y^2 &= G(x^2 | x^1) \\ \vdots & \quad \vdots \\ y^d &= G(x^d | x^{d-1}, \ldots, x^1) \end{aligned} \tag{3}$$

   where $G(x^k | x^{k-1}, \ldots, x^1) = \int_{-\infty}^{x^k} g(\tilde{x}^k | x^{k-1}, \ldots, x^1) d\tilde{x}^k$

2. Induced density $h(y)$, of the vector y, takes the form:

$$h(y) = \frac{f(G^{-1}(y^1), \ldots, G^{-1}(y^d | y^{d-1}, \ldots, y^1))}{g(G^{-1}(y^1), \ldots, G^{-1}(y^d | y^{d-1}, \ldots, y^1))} \tag{4}$$

   where G^{-1} is inverse CDF and $x^k = G^{-1}(y^k | x^{k-1}, \ldots, x^1)$.

3. Then we know

$$\hat{H}_\alpha(Y_n) \to \frac{1}{1-\alpha} \ln \int h^\alpha(y)dy \quad (a.s.)$$

4. By Jacobian formula: $dy = \left|\frac{dy}{dx}\right| dx = g(x)dx$ and

$$\frac{1}{1-\alpha} \ln \int h^\alpha(y)dy = \frac{1}{1-\alpha} \ln \int \left(\frac{f(x)}{g(x)}\right)^\alpha g(x)dx = D(f\|g)$$

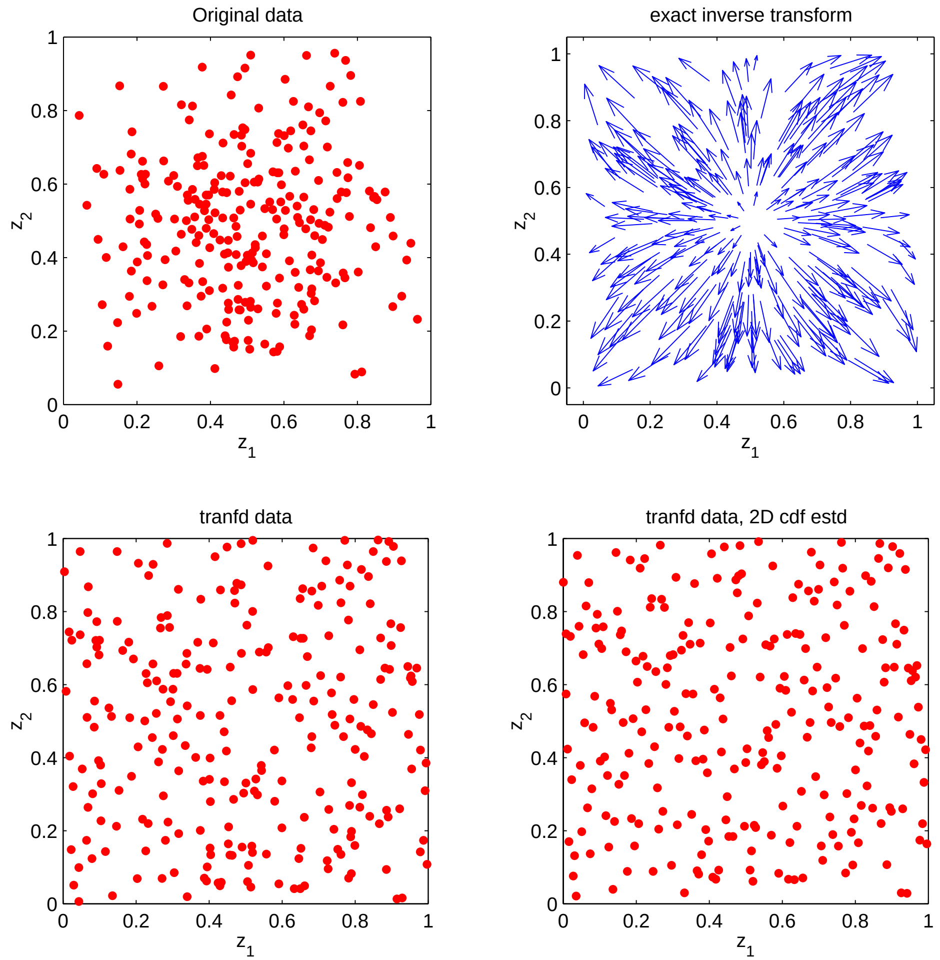


Figure 13: Top Left: i.i.d. sample from triangular distribution, Top Right: exact transformation, Bottom: after application of exact and empirical transformations.

Application: α-MI estimation

Objective: To estimate

$$\mathrm{MI}_\alpha(X,Y) = \frac{1}{\alpha-1}\ln\int f^\alpha(X,Y)(f(X)f(Y))^{1-\alpha}dX\,dY.$$

Assume that $f(X,Y)$ is such that $f^\alpha(X,Y)$ is in the the Besov space $B^1_{p,1}(\mathbb{R}^2)$, $p>2$ and $\alpha = 1/2$.

Density plug in method: rms convergence rate

$$\mathrm{M}SE^{\frac{1}{2}}(\hat{\mathrm{M}}\mathrm{I}) \geq O(n^{-1/4})$$

Measure transformation method: rms convergence rate

$$\mathrm{M}SE^{\frac{1}{2}}(\hat{\mathrm{M}}\mathrm{I}) \leq O(n^{-\alpha\lambda(p)/(1+\alpha\lambda(p))1/d}) \to_{p\to\infty} O(n^{-3/10})$$

Alternative depedency measure: α-Jensen difference

1. Extract features from reference and transformed target images:

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

2. Construct following MST function on X_m and Y_n

$$\Delta L = \ln L_\gamma(X_m \cup Y_n)/(n+m)^\alpha - \frac{m}{n+m} \ln L_\gamma(X_m)/m^\alpha - \frac{n}{n+m} \ln L_\gamma(Y_n)/n^\alpha$$

3. Minimize ΔL_γ over transformations producing Y_n.

$$(1-\alpha)^{-1}\Delta L \;\rightarrow\; H_\alpha(\varepsilon f_x + (1-\varepsilon) f_y) - \varepsilon H_\alpha(f_x) - (1-\varepsilon) H_\alpha(f_y)$$

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

Example

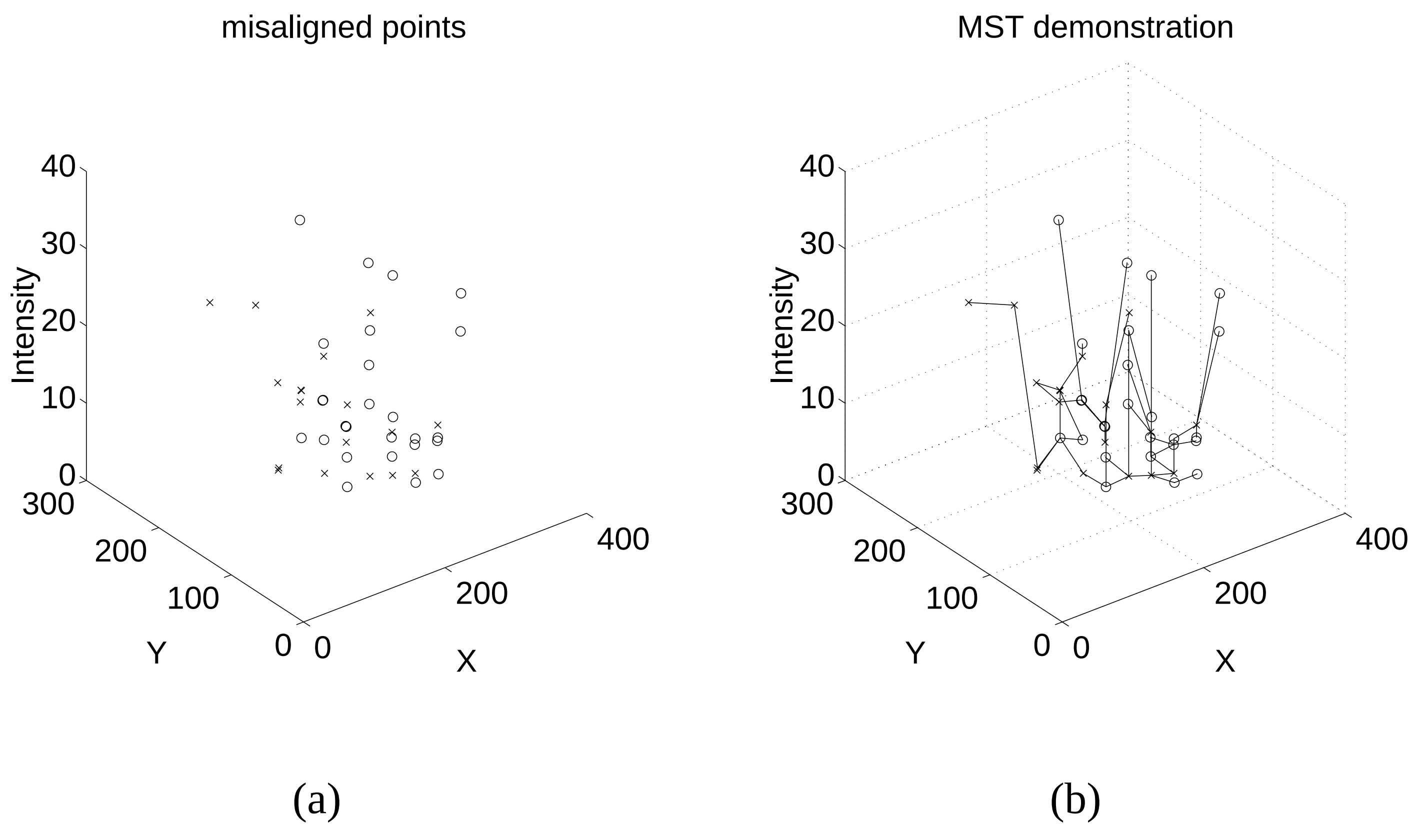


Figure 14: MST demonstration for misaligned images

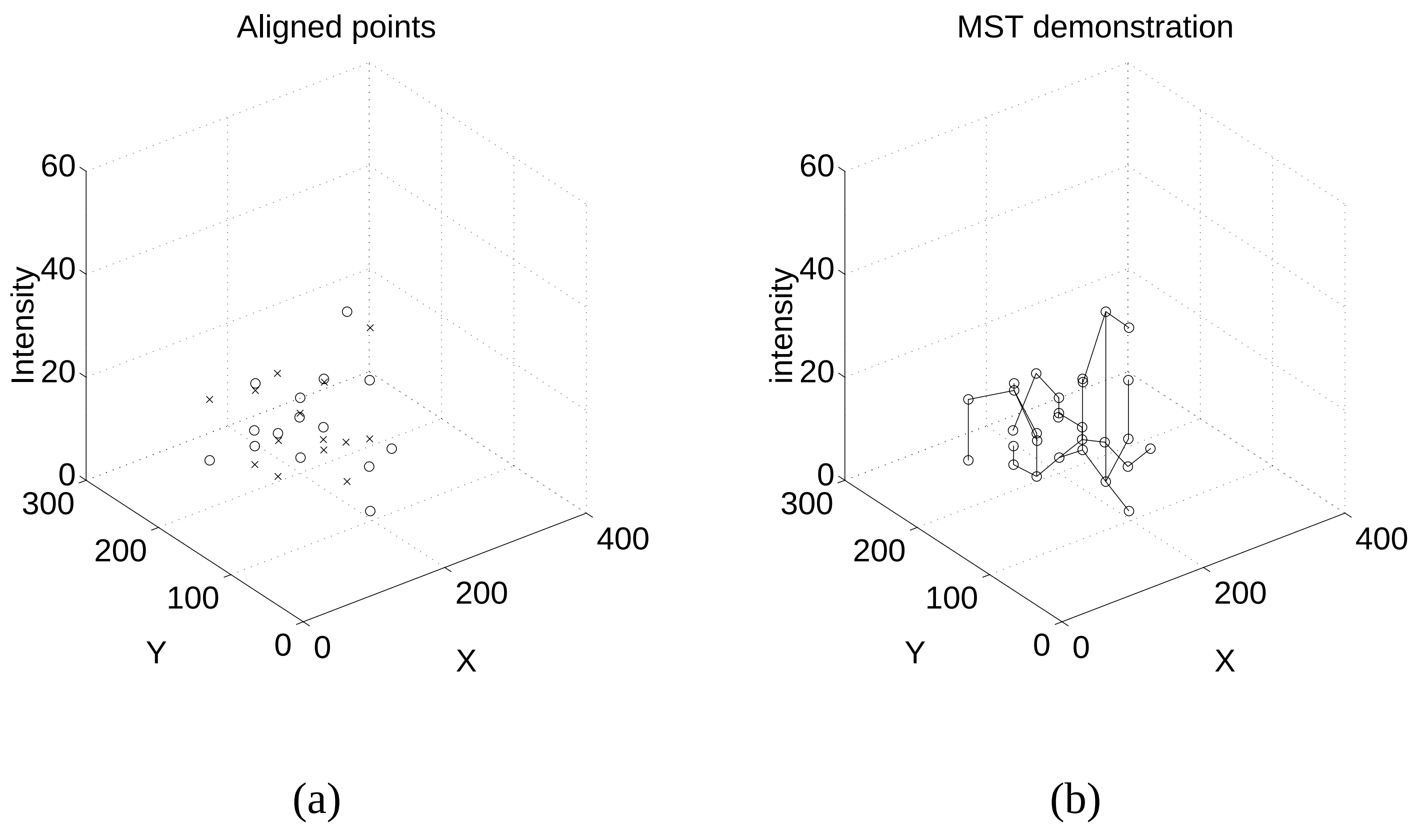


Figure 15: MST for aligned images. “x” denotes reference while “o” denotes a candidate image in the DEM database.

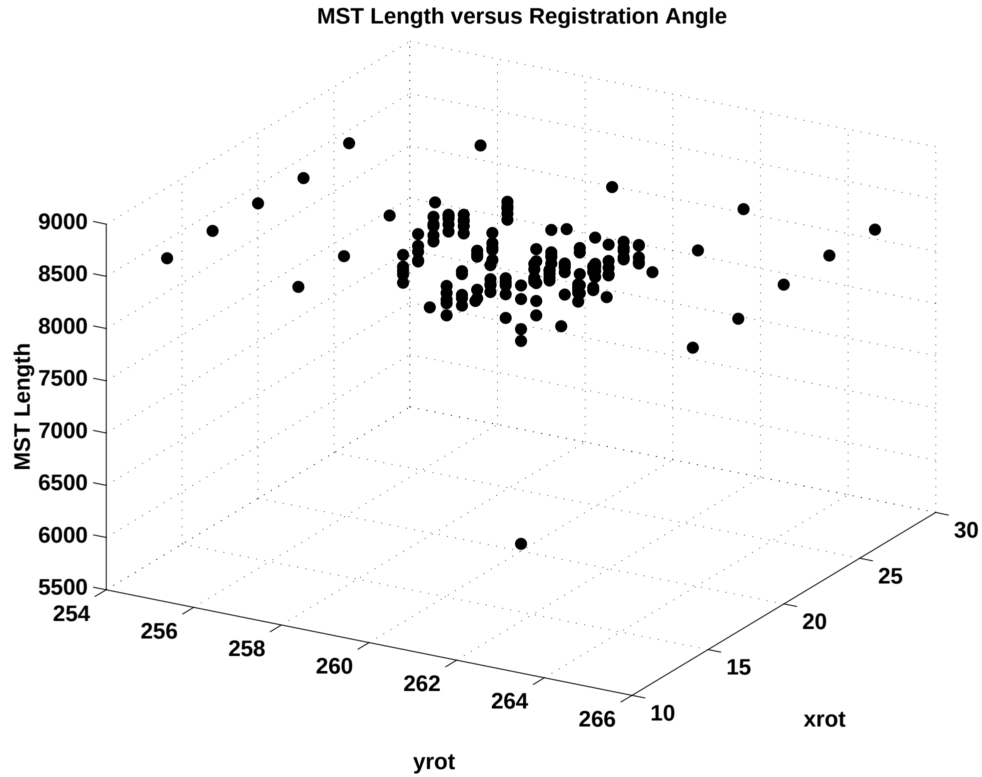
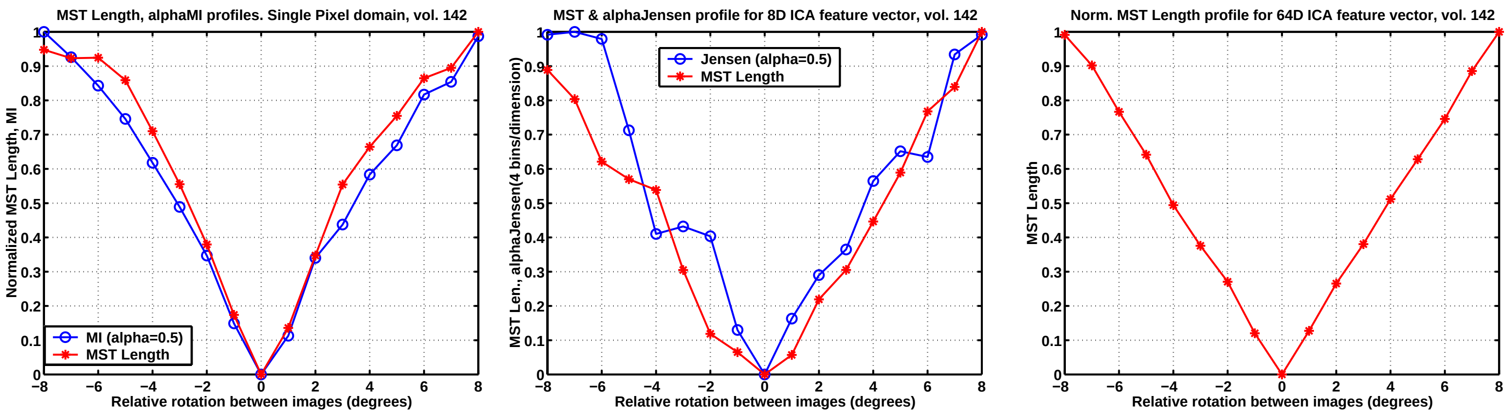


Figure 16: Scatter plot of MST length for a selection of relative rotation angles between reference DEM image and target radar image.

Experimental results for US Image Registration



(a) (b) (c)

Figure 17: Objective function profiles for histogram (L,M) and MST (L,M,R) estimators of α-Jensen difference vs histogram plug-in estimator ($\alpha = 1/2$): Single-pixel (L), 8D ICA (M), 64D ICA (R).

Quantitative Performance Comparisons

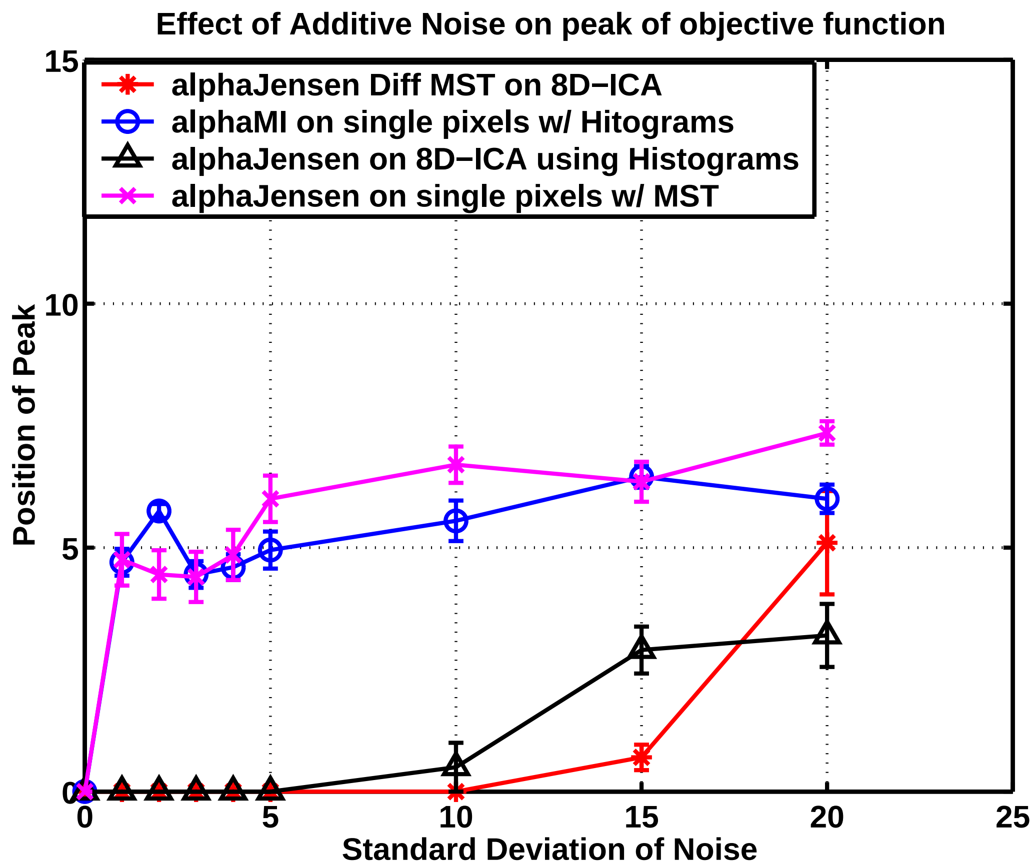


Figure 18: *Quantitative registration MSE comparisons.*

Computational Acceleration of MST

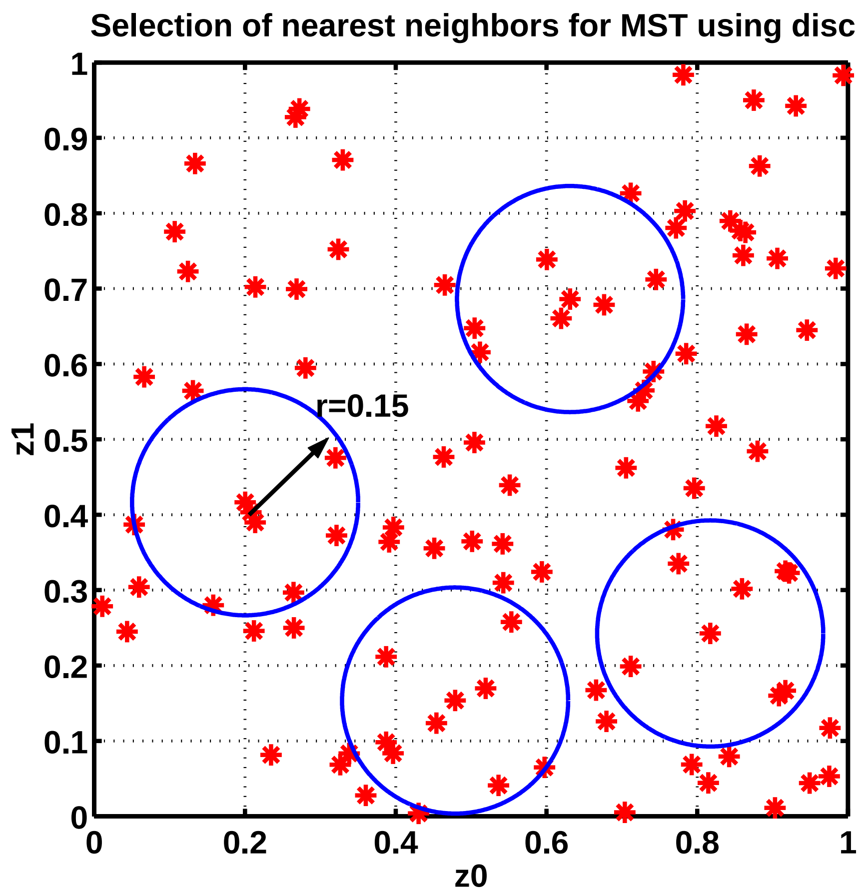


Figure 19: *Acceleration of Kruskal's MST algorithm from* $n^2 \log n$ *to* $n \log n$*.*

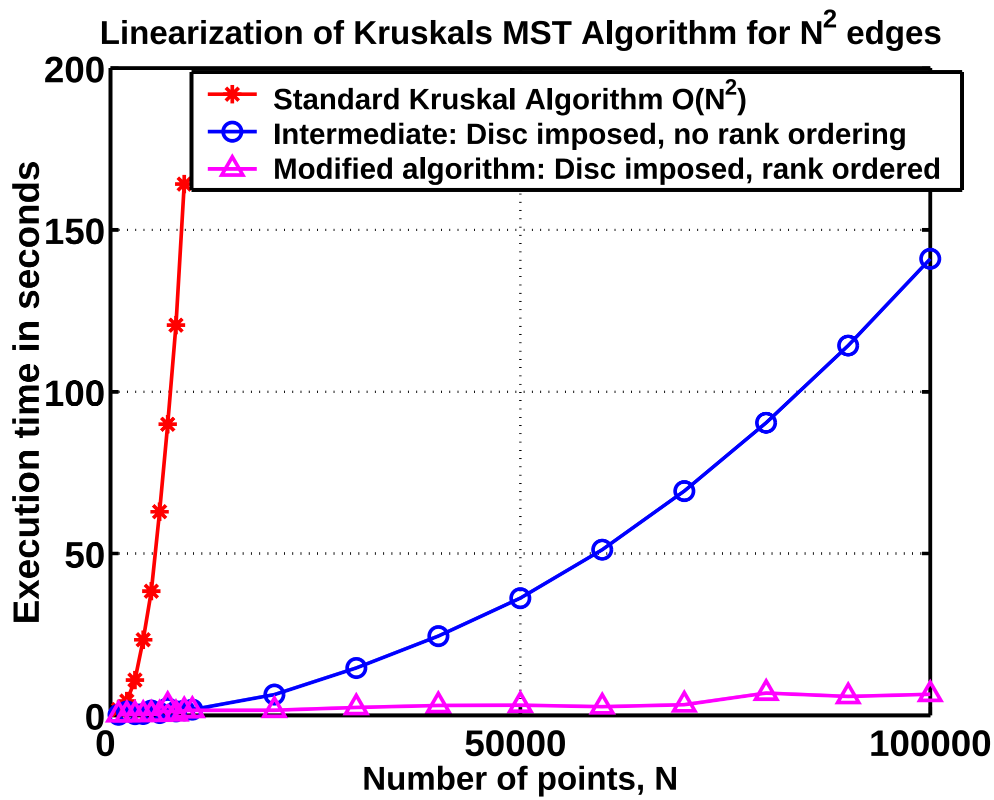


Figure 20: *Comparison of Kruskal's MST to our $n \log n$ MST algorithm.*

Outlier Sensitivity of minimal n-point graphs

Assume f is a mixture density of the form

$$f = (1-\varepsilon) f_1 + \varepsilon f_o, \tag{5}$$

where

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

Outlier rejection via K-MST

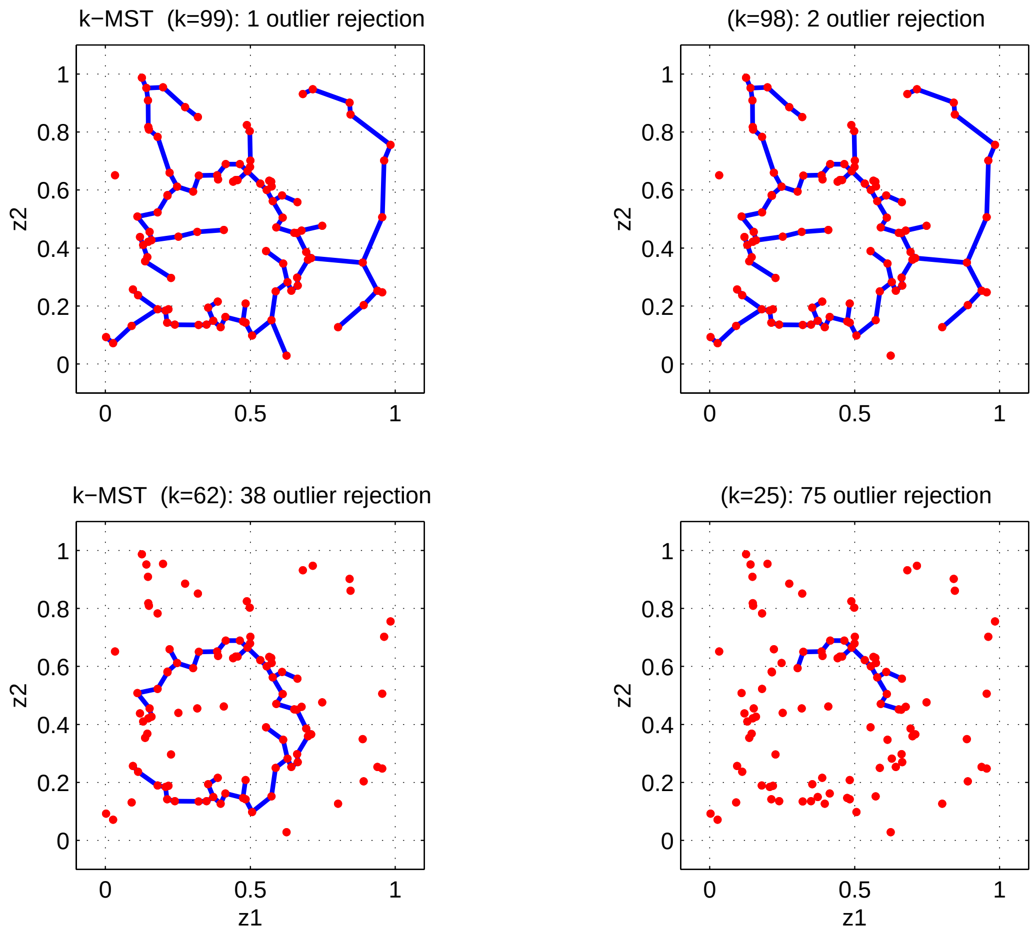


Figure 21: *k-MST for 2D annulus density.*

k-MST Stopping Rule

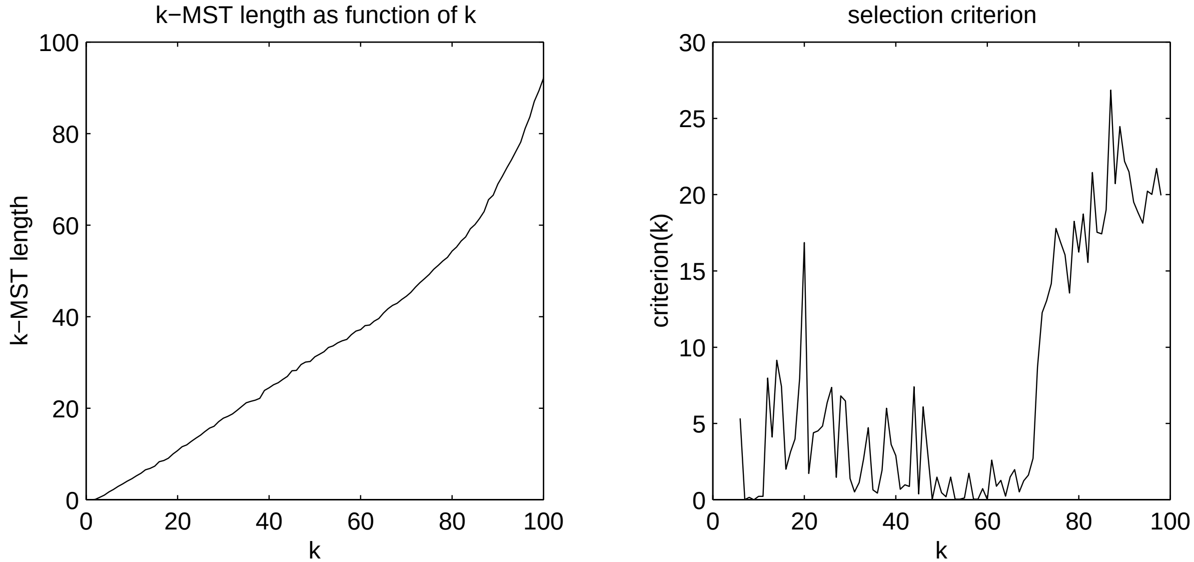


Figure 22: *Left: k-MST curve for 2D annulus density with addition of uniform "outliers" has a knee in the vicinity of* $n - k = 35$. *This knee can be detected using residual analysis from a linear regression line fitted to the left-most part of the curve. Right: error residual of linear regression line.*

Greedy partioning approximation to K-MST

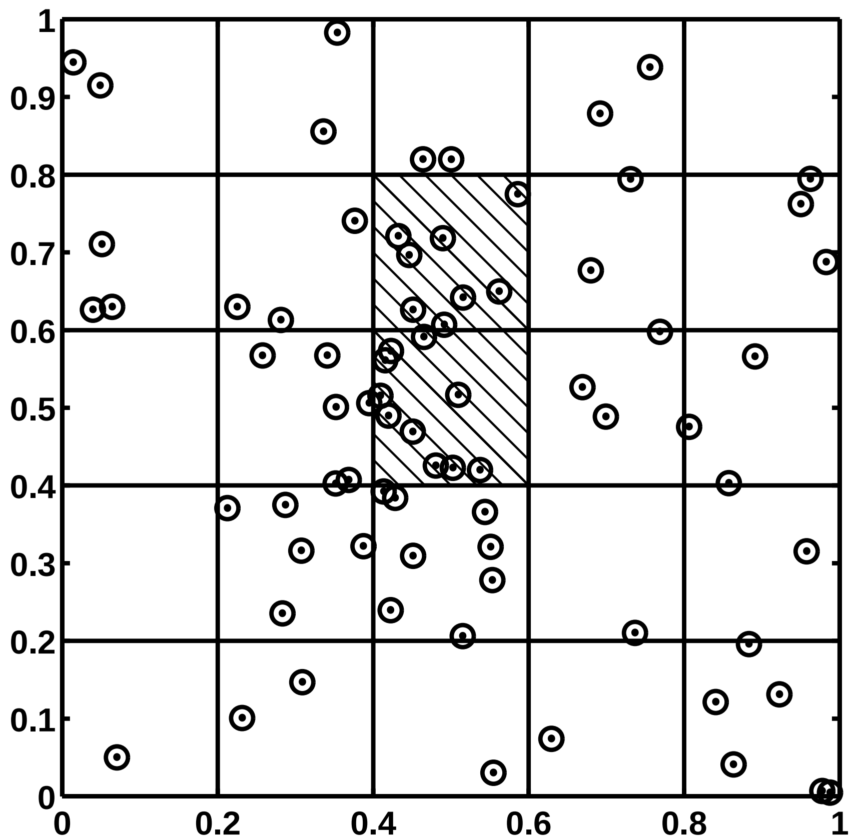


Figure 23: *A sample of* 75 *points from the mixture density* $f(x) = 0.25 f_1(x) + 0.75 f_o(x)$ *where* f_o *is a uniform density over* $[0,1]^2$ *and* f_1 *is a bivariate Gaussian density with mean* $(1/2, 1/2)$ *and diagonal covariance* $\mathrm{diag}(0.01)$. *A smallest subset* B_k^m *is the union of the two cross hatched cells shown for the case of* $m = 5$ *and* $k = 17$.

Extended BHH Theorem for Greedy K-MST

Fix $\rho \in [0,1]$ and assume that the k-minimal graph is *tightly coverable*. If $k = \lfloor \rho n \rfloor$, as $n \to \infty$ we have (Hero&Michel:IT99)

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

or, alternatively, with

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

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

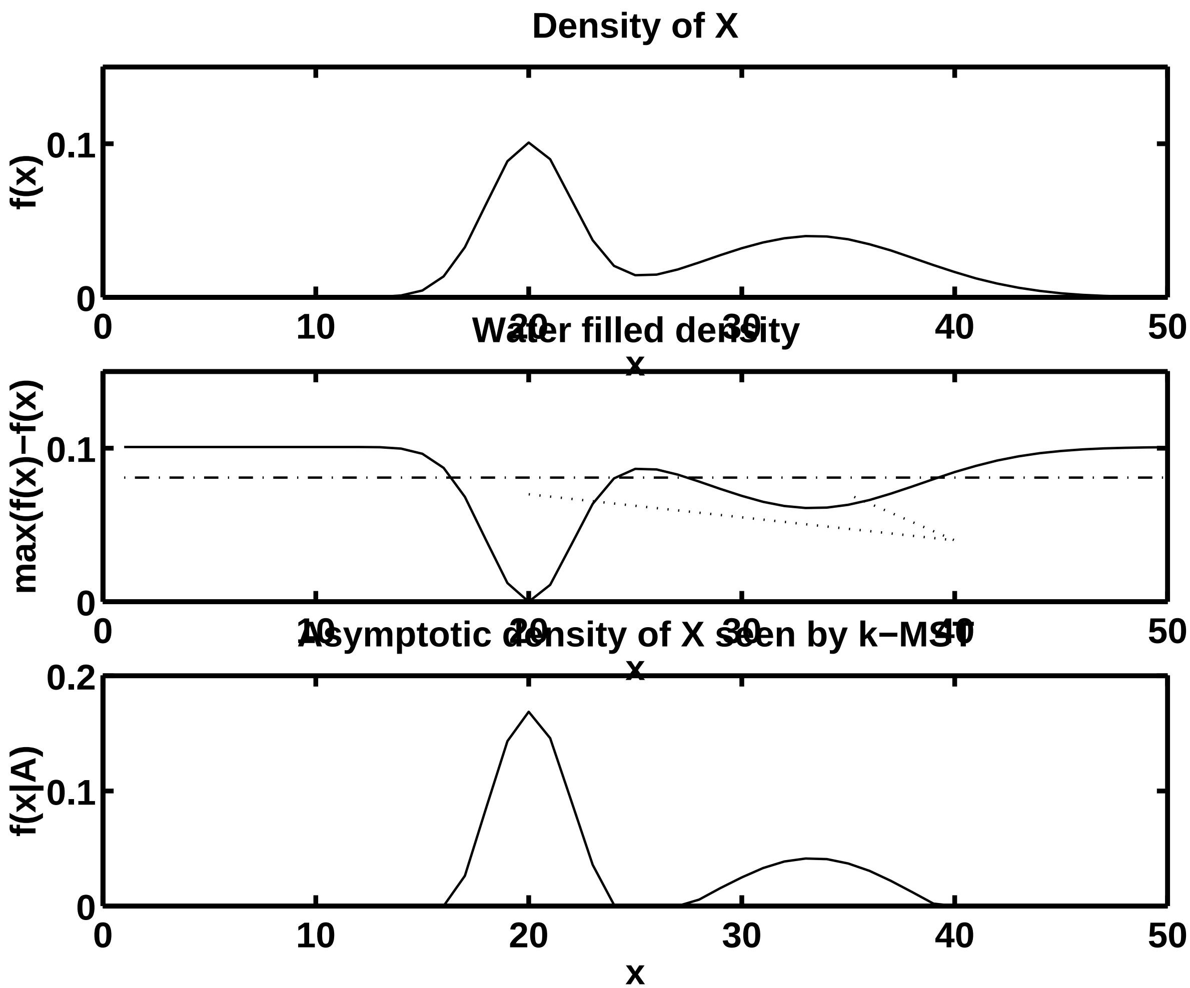


Figure 24: *Waterpouring contruction of minimum entropy density.*

k-MST Influence Function for Gaussian Feature Density

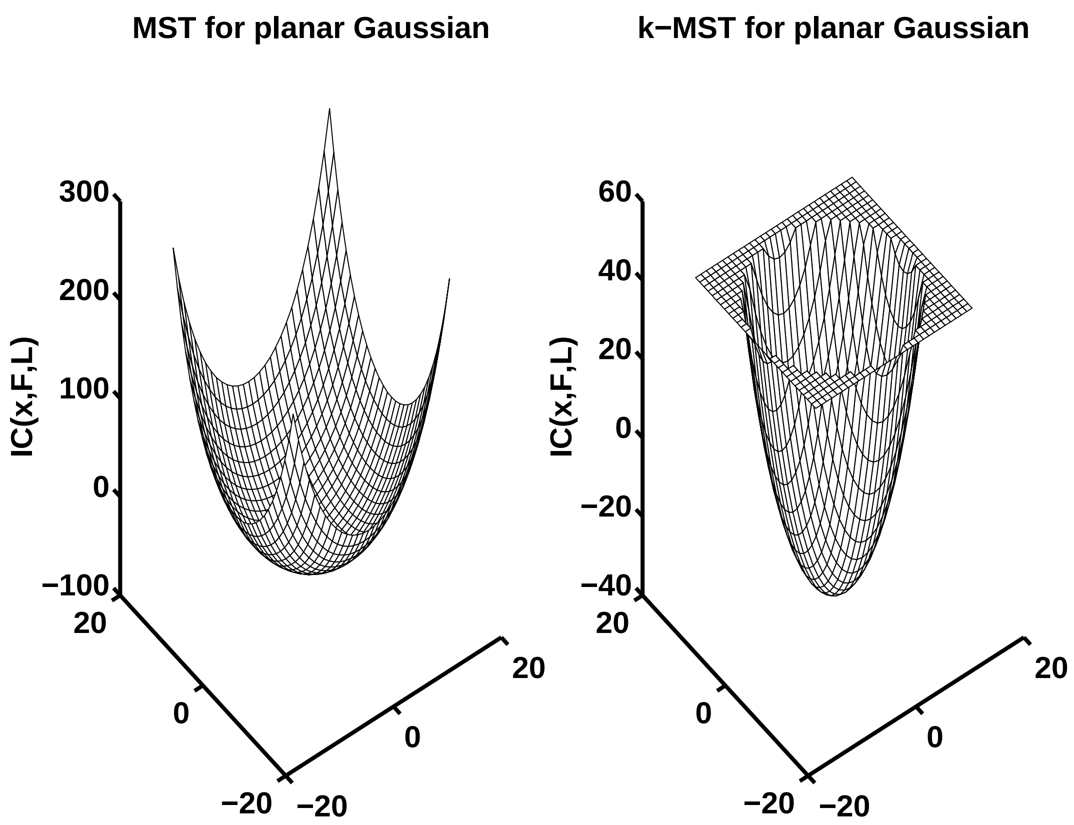


Figure 25: *MST and k-MST influence curves for Gaussian density on the plane.*

Application: testing distributions

Non-parametric density classification problem: decide between

$$
\begin{aligned}
H_0 &: \quad f(x) = f_0(x) \\
H_1 &: \quad f(x) \neq f_0(x)
\end{aligned}
$$

Step 1: Perform uniformizing transformation on X_n (under H_0)

Step 2: Construct MST on transformed variables Y_n

Classification rule:

$$
\hat{D}_\alpha(f \| f_0) \stackrel{\text{def}}{=} L_\gamma(Y_n)/n^\alpha \quad \underset{H_0}{\overset{H_1}{\gtrless}} \quad \eta
$$

Application: Robust density estimation

Estimate $f_1(x)$ given sample from mixture

$$f(x) = (1-\varepsilon) f_1(x) + \varepsilon f_0(x)$$

- $f_0(x)$= known contaminating density

Step 1: Perform transformation on X_n to uniformize f_0 component

Step 2: Construct k-MST on transformed variables Y_n

$$L_\gamma(Y^*_{n,k})/(\lfloor \rho n \rfloor)^\alpha \rightarrow \beta_{L_\gamma,d} \min_{A:P(A)\geq\rho} \int_A \left(\frac{f(x)}{f_0(x)}\right)^\alpha f_0(x)dx$$

Robust Density Estimator: kernel estimator applied to $X_{i_1}, \ldots, X_{i_{\lfloor \rho n \rfloor}}$

Classification Example

To test:

$$\begin{aligned} H_0 &: \quad f(x) = triangular \\ H_1 &: \quad f(x) \neq triangular \end{aligned}$$

Ground Truth:

- $f(x) = (1-\varepsilon) f_1(x) + \varepsilon f_0(x)$: mixture density
- $f_1(x)$ is uniform density on $[0,1]^2$
- $f_0(x)$ is triangular density on $[0,1]^2$

Test statistic: $\hat{D}_\alpha(f \| f_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta$

ROC curves

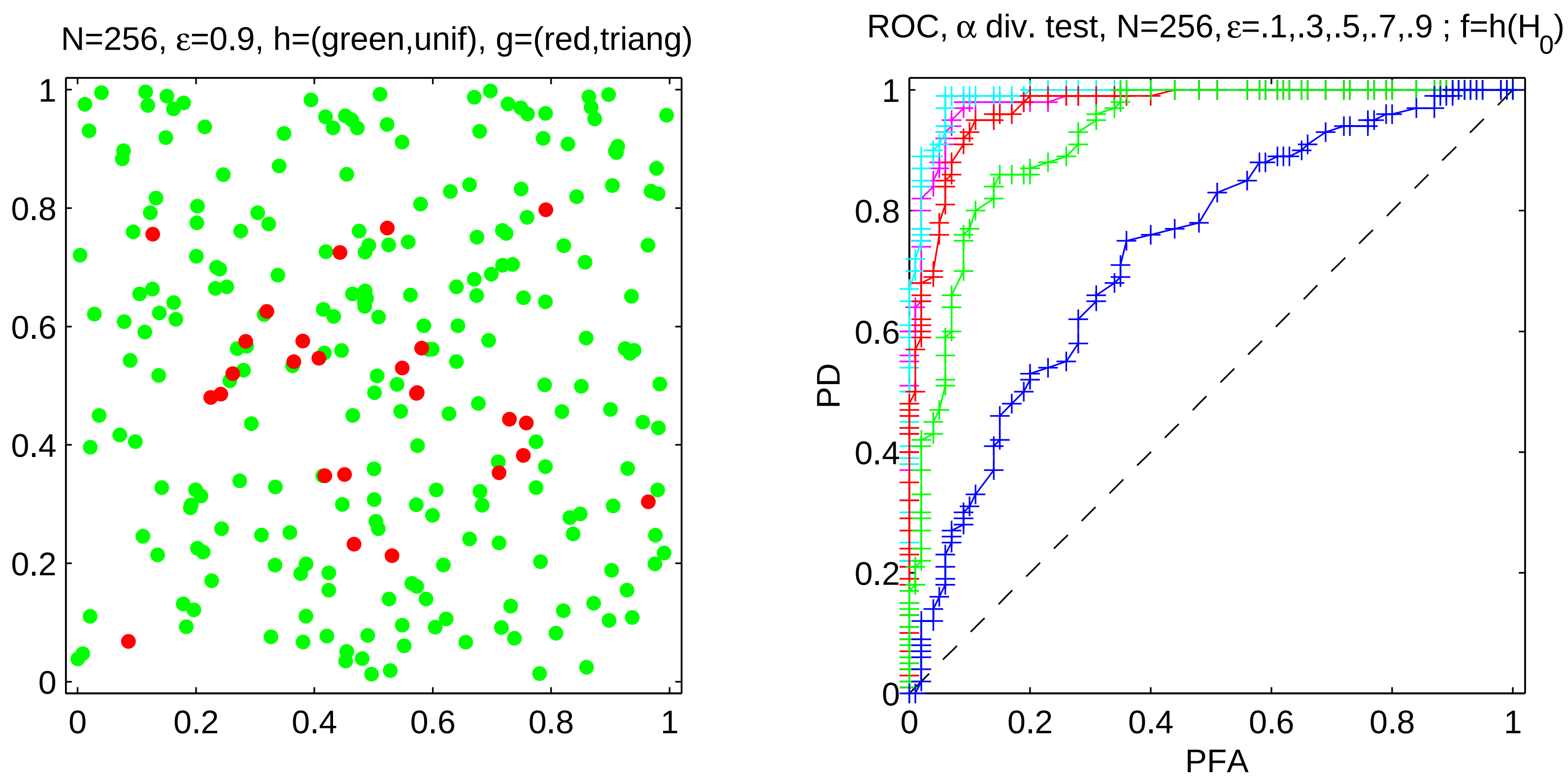


Figure 26: *Left: A sample from triangle-uniform mixture density with* $\varepsilon = 0.9$ *in the transformed domain* Y_n. *Right: ROC curves of thresholded K-MST. Curves are increasing in* ε *over the range* $\varepsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$

Outlier rejection example

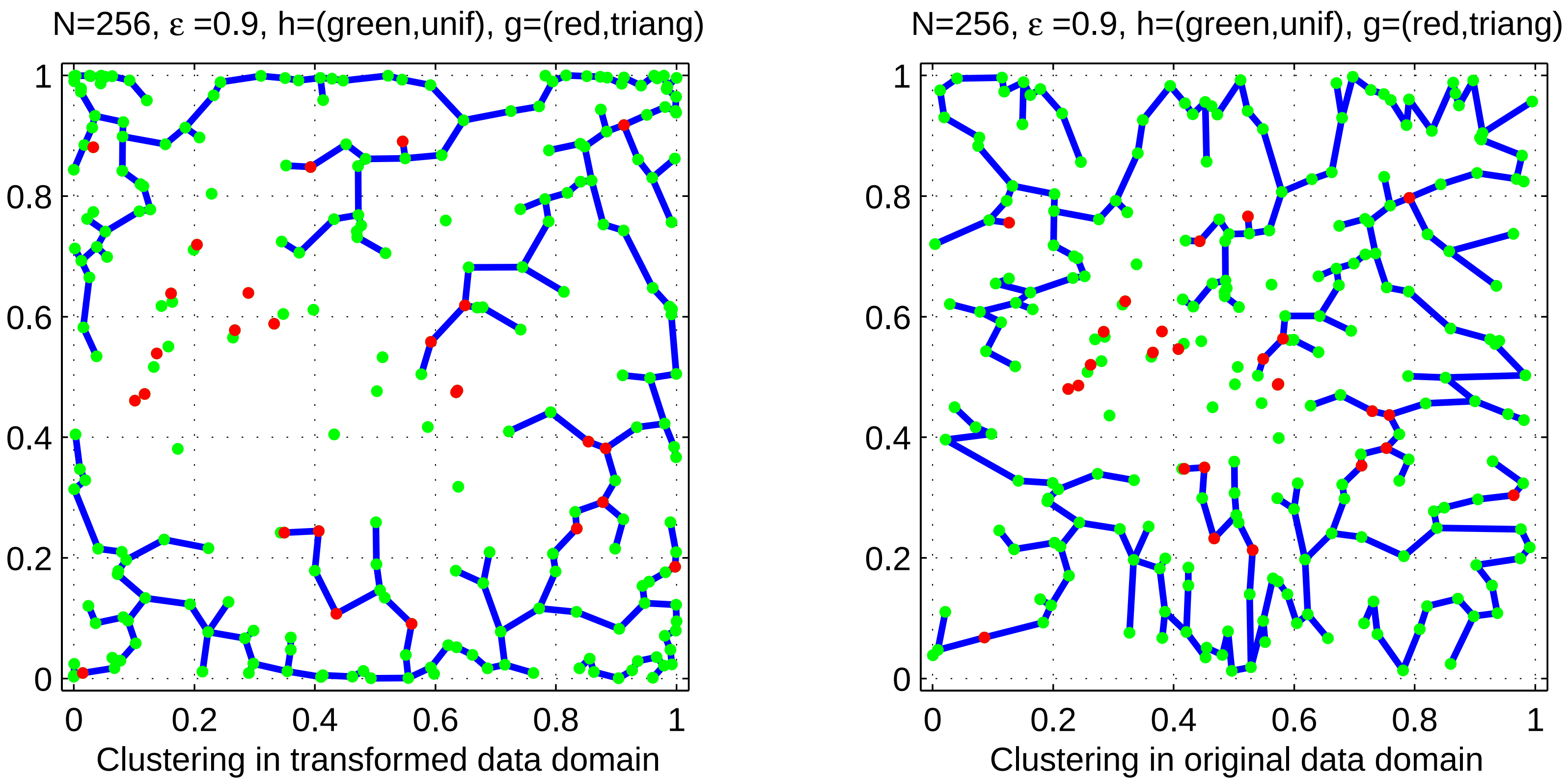


Figure 27: *Left: the k-MST implemented on the transformed scatterplot Y_n with $k = 230$. Right: same k-MST displayed in the original data domain.*

Conclusions

1. α-divergence can be justified via decision theory
2. Applicable to feature-based image registration
3. Non-parametric estimation is possible even for very high dimensions via MST
4. MST outperforms plug-in estimation when latter is feasible
5. Robustified MST can be defined via optimal pruning of MST: k-MST
6. Divergence can be estimated by preprocessing with measure tranformation