Divergence matching criteria for registration, indexing and retrieval

Alfred O. Hero

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

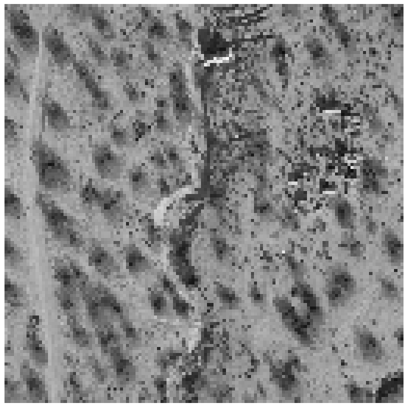
University of Michigan - Ann Arbor

hero@eecs.umich.edu

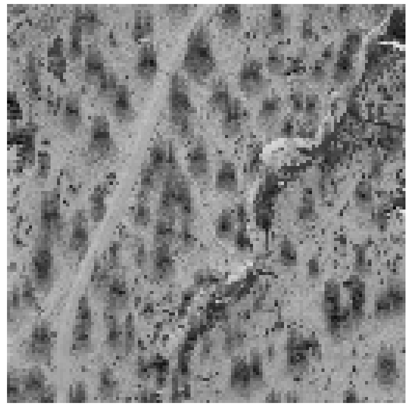
http://www.eecs.umich.edu/~hero

Outline

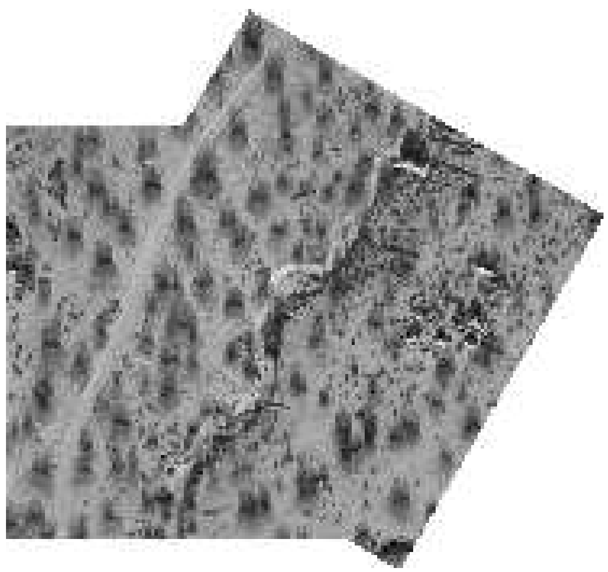
1. Statistical framework: divergence measures
2. Renyi α-divergence and error exponents
3. Indexing via divergence estimation
   (a) MLE plug-in divergence estimators
   (b) Direct divergence estimators
   (c) Graph matching and divergence estimation



(a) Image I_1



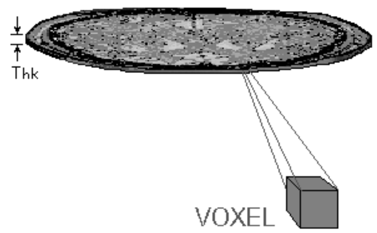
(b) Image I_0



(c) Registration result

Figure 1: A multidate image registration example

(a)



(b)

(c)

Figure 2: MRI image generation process

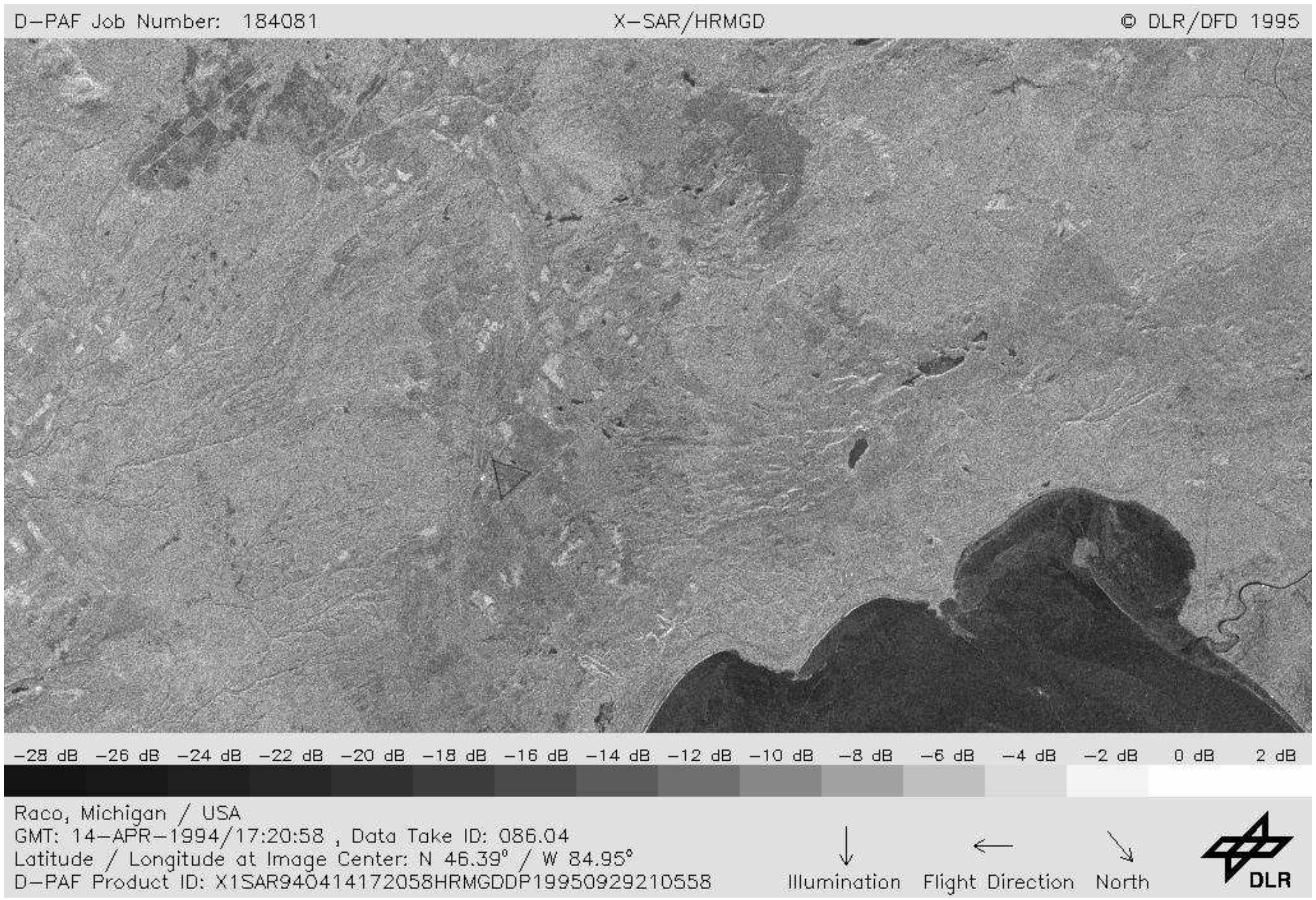


Figure 3: *X-SAR image of Raco, Michigan (DARA/ESA database)*

Statistical Framework

- X: an image
- Θ: a parameter space
- $f(X|\theta)$: an image model (likelihood)
- X_R a reference image
- $\{X_i\}$ a database of K images

$$X_R \sim f(X|\theta_R)$$

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

$\Rightarrow$ Similarity btwn X_i, X_R lies in similarity btwn models

Divergence Measures

Refs: [Csiszár:67,Basseville:SP89]

Define densities

$$f_i = f(X|\theta_i), \quad f_R = f(X|\theta_R)$$

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

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

α-Divergence: Special cases

- α-Divergence vs. Hellinger distance squared

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

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

- α-Divergence vs. Kullback-Liebler divergence

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

Clairvoyant Indexing

$$X_i \prec X_j \quad \Leftrightarrow \quad D_\alpha(f_i \| f_R) < D_\alpha(f_j \| f_R)$$

1. Image registration: $f_i = f(X|\theta_i)$ indexes poses/orientations [Viola&Wells:ICCV95]
2. Image classification: $f_i = f(X|\theta_i)$ index model classes [Stoica&etal:INRIA98]
3. Target detection: $f_i = f(X|\theta_i)$ are possible "target-present" models

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

- minimum-distance detection/estimation [Beran&Millar:AnnStat89,92]

Rényi α-divergence and Error Exponents

Observe i.i.d. sample $\underline{X} = [X^{(1)}, \ldots, X^{(n)}]$

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

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]

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

Optimal Indexing

Define

$$\Theta_{ijR} = \left\{(\theta_i, \theta_j, \theta_R) \;:\; D_\alpha(f_i \| f_R) \;<\; D_\alpha(f_j \| f_R)\right\}$$

Note: $X_i \prec X_j$ is a composite hypothesis

Define: Pairwise decision error probability $P_e(\theta_i, \theta_j, \theta_R)$

Minimum-average-P_e decision rule:

$$P((\theta_i, \theta_j, \theta_R) \in \Theta_{ijR} \mid X_i, X_j, X_R)$$

$$= \frac{1}{P(X_i, X_j, X_R)} \int\int\int dP(\theta_i, \theta_j, \theta_R) f(X_i|\theta_i) f(X_j|\theta_j) f(X_R|\theta_R) \;\geq\; \frac{1}{2}$$

Minimax-P_e decision rule:

$$\frac{1}{P^*(X_i, X_j, X_R)} \int\int\int dP^*(\theta_i, \theta_j, \theta_R) f(X_i|\theta_i) f(X_j|\theta_j) f(X_R|\theta_R) \geq \frac{1}{2}$$

dP^* is least favorable prior

Suboptimal Indexing

Estimate $D_\alpha(\theta_i \| \theta_R)$ from data

1. Parametric estimation on full model
2. Parametric estimation on reduced model
3. Nonparametric estimation

Parametric Approach: Full Model

Full model divergence estimation procedure

1. Impose known model $f(X|\theta)$ on image space
2. Precompute function $D_\alpha(\bullet\|\star)$
3. Substitute estimates $\hat{\theta}_i, \hat{\theta}_j, \hat{\theta}_R$ into D_α
4. Indexing $X_{(1)} \prec \ldots \prec X_{(K)}$ determined by

$$D_\alpha(\hat{\theta}_{(1)}\|\hat{\theta}_R) < \cdots < D_\alpha(\hat{\theta}_{(K)}\|\hat{\theta}_R)$$

Example: texture classification

Gaussian Image Model: 2D Wold decomposition (Francos:IP95,IT97)

$$X(n,m) = W(n,m) + V(n,m)$$

- $W(n,m)$: 2D Gaussian MA random field
- $V(n,m)$: 2D deterministic random field

Special case: $\alpha = 1 \Rightarrow$ KL divergence (StoicaZerubiaFrancos:INRIA98)

$$D_1(\theta_2 \| \theta_1) = \frac{1}{2}(\mu_2 - \mu_1)^T (\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_2 - \mu_1) + \frac{1}{2}\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_1\Sigma_2^{-1} - 2I\right)$$

Extension: $\alpha < 1$ [Hero&Francos&Michel:01].

Parametric Approach: Reduced Model

Reduced model divergence-estimation procedure

1. Projection of X onto lower dimensional subspace Z
2. Find induced distribution $f(Z|\theta)$ on feature space
3. Precompute function $D^Z_\alpha(\bullet\|\star)$
4. Indexing $X_{(1)} \prec \ldots \prec X_{(K)}$ determined by

$$D^Z_\alpha(\hat{\theta}_{(1)}\|\hat{\theta}_R) < \cdots < D^Z_\alpha(\hat{\theta}_{(K)}\|\hat{\theta}_R)$$

Example: SAR Target Detection

Refs:
[Kim&Hero:IP00,Reed&Yu:ASSP90,KellyAES86,Kariya&Sinha:89]

$$\mathbf{X} = [\underline{X}_1, \ldots, \underline{X}_n]$$

$\underline{X}_k$ = $m \times 1$ lexicographic ordered k-th sub-image

$$\mathbf{X} = \mathbf{S}\,\mathbf{A}\,\mathbf{B} + \mathbf{N}$$

- $\mathbf{S} = [\underline{s}_1, \ldots, \underline{s}_p]$ = an $m \times p$ matrix of target signatures (*known*)
- $\mathbf{A} = \mathrm{diag}(\underline{a})$ = a $p \times p$ diagonal matrix (*unknown*)
- $\mathbf{B}$ a $p \times n$ matrix of target locations

Multivariate Gaussian Noise Model

$\mathbf{N}$: $(m \times n)$ multivariate Gaussian matrix
w/ i.i.d. columns $(m \times 1)$ each having covariance matrix $\mathbf{R}(m \times m)$.

$$cov[\text{vec}\{\mathbf{N}\}] = \mathbf{R} \bigotimes \mathbf{I}_n = \begin{bmatrix} \begin{matrix} \mathbf{R} & O \\ O & \mathbf{R} \end{matrix} & \mathbf{O} \\ \mathbf{O} & \begin{matrix} \mathbf{R} & O \\ O & \mathbf{R} \end{matrix} \end{bmatrix} (mn \times mn)$$

Note :

$$\mathbf{ANB} \sim \mathcal{N}_{mn}\left(\underline{0}, \mathbf{ARA}^T \bigotimes \mathbf{BB}^T\right)$$

$(\mathbf{A} : m \times m,\ \mathbf{B} : n \times n)$

Detection Problem

- Parameter space $\Theta = \Theta_0 \cup \Theta_1$

$$\begin{aligned} H_0 &: \quad \mathbf{X} \sim f(X;\, \theta_0), & \theta_0 \in \Theta_0 \\ H_1 &: \quad \mathbf{X} \sim f(X;\, \theta_1), & \theta_1 \in \Theta_1 \end{aligned}$$

In this case:

$$D_1(\theta_1 \| \theta_0) = \mathrm{tr}\left\{(\mathbf{S\,A\,B})^H \mathbf{R}^{-1} (\mathbf{S\,A\,B})\right\}$$

and detection algorithm is

$$D_1(\hat{\theta}_1 \| \hat{\theta}_0) \quad \underset{H_0}{\overset{H_1}{\gtrless}} \quad \eta$$

which is equivalent to the GLRT.

Statistical Reduction via Sufficiency

Minimal Sufficient Statistic

$$\Lambda_{\theta_0,\theta_1}(\mathbf{X}) \stackrel{\text{def}}{=} \frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} = g(T(\mathbf{X}),\theta_1,\theta_0)$$

$$\{\mathbf{X} : \Lambda_{\theta_0,\theta_1}(\mathbf{X}) = \lambda\}_{\lambda>0} = \{\mathbf{X} : T(\mathbf{X}) = t\}_t$$

- $T(\mathbf{X})$ specifies orbit $\{\mathbf{X} : \Lambda_{\theta_0,\theta_1}(\mathbf{X}) = \lambda\}$ (depends on θ_0 and θ_1).

- $T(\mathbf{X})$ achieves maximal data reduction while preserving information necessary to estimate θ_0 and θ_1 and discriminate between H_0 and H_1.

- Distribution of $T(\mathbf{X})$ depends on particular values of $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$.

Statistical Reduction via Invariance

Refs: [Kariya&Sinha:89,Eaton:83,Ferguson:67]

Maximal invariant statistic

$$\begin{aligned}\tilde{\Lambda}(\mathbf{X}) &\stackrel{\text{def}}{=} \left\{\frac{f(\mathbf{X};\theta_1)}{f(\mathbf{X};\theta_0)} : \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\right\} \\ &= \left\{\tilde{h}(\mathbf{Z}) : \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\right\}\end{aligned}$$

- $\mathbf{Z} = Z(\mathbf{X})$ specifies orbit of $\{\mathbf{X} \, : \, \tilde{\Lambda}(\mathbf{X}) = \tilde{\lambda}\}$ (depends on Θ_0 and Θ_1).

- Maximal invariants can be found when Θ_0 and Θ_1 have simple topological group structure.

Group characterization of Θ_0, Θ_1

Let $\mathcal{G}$ be a group of transformations $g : \mathcal{X} \to \mathcal{X}$ acting on $\mathbf{X}$.

Assume that for each $\theta \in \Theta$ there exists a unique $\overline{\theta} = \overline{g}(\theta)$ such that

$$f_\theta(g(\mathbf{X})) = f_{\overline{\theta}}(\mathbf{X}), \qquad \overline{g}(\Theta_0) = \Theta_0, \ \overline{g}(\Theta_1) = \Theta_1$$

$\overline{g}$ is called the induced group action on Θ.

$\mathbf{Z} = Z(\mathbf{X})$ is a maximal invariant iff

1. (invariant property) $Z[g(\mathbf{X})] = Z(\mathbf{X})$ for all $g \in \mathcal{G}$ and
2. (maximal property) $Z(\mathbf{X}) = Z(\mathbf{Y}) \Rightarrow \mathbf{Y} = g(\mathbf{X})$ for some $g \in \mathcal{G}$.

Example: detection of scalar signal; a, σ^2 unknown

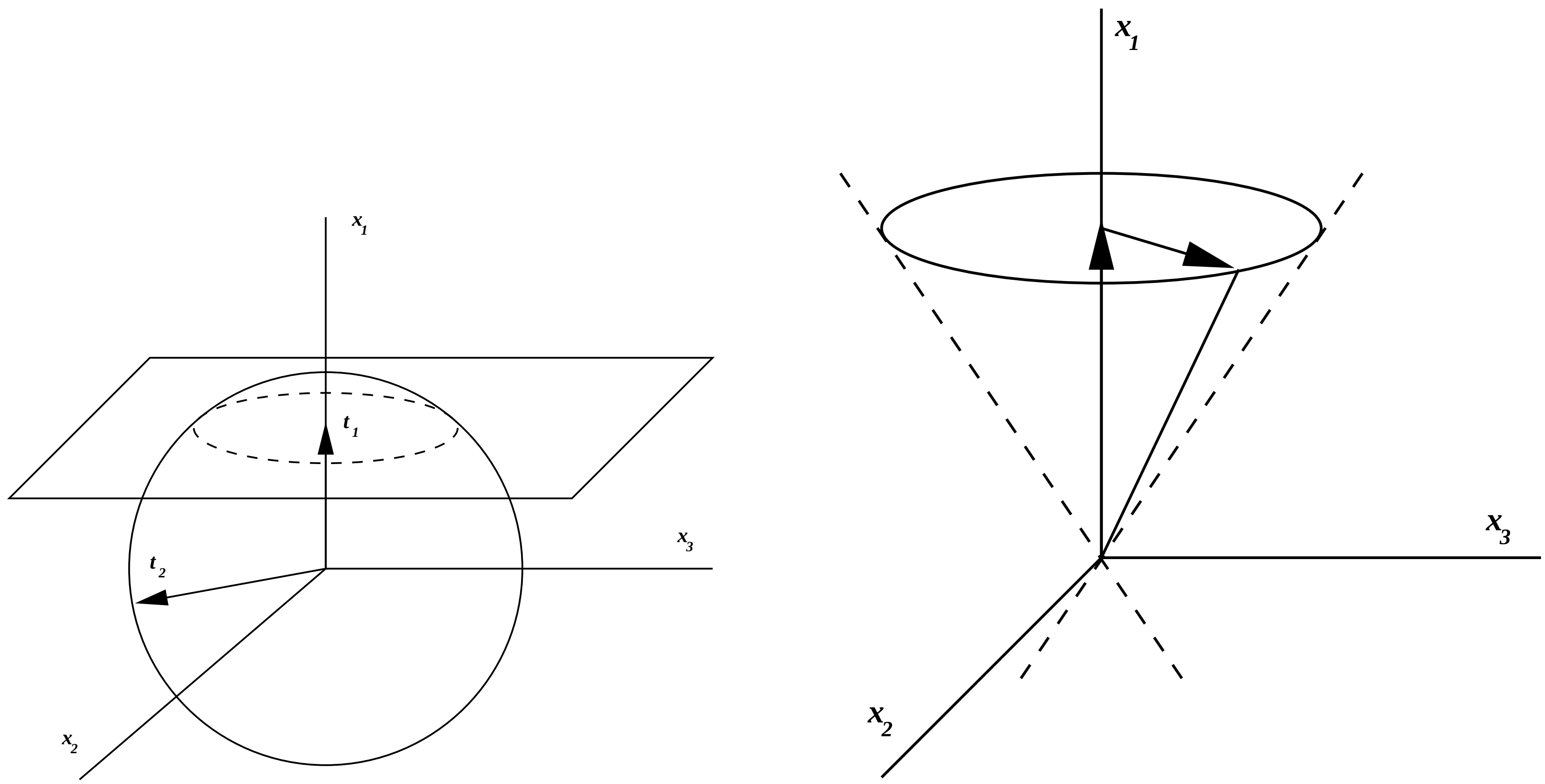
$$\underline{\mathbf{x}}^T = a[1,0,\ldots,0] + \underline{N}, \qquad \underline{N} \sim \mathcal{N}_n(0,\sigma^2\mathbf{I}).$$

$$\Theta_0 = \{a = 0, \sigma^2 > 0\}, \quad \Theta_1 = \{a \neq 0, \sigma^2 > 0\}$$

$$\log\Lambda_{\theta_0,\theta_1}(\underline{\mathbf{x}}) = \frac{a}{\sigma_1^2}\,\underline{\mathbf{e}}_1^T\underline{\mathbf{x}} + \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\,\underline{\mathbf{x}}^T\underline{\mathbf{x}} + const.$$

1. **Sufficiency orbits**: $[\underline{\mathbf{e}}_1^T\underline{\mathbf{x}}, \underline{\mathbf{x}}^T\underline{\mathbf{x}}] = [t_1, t_2]$ (circle in $\mathbf{R}^3$)

2. **Invariance orbits**: $|\underline{\mathbf{e}}_1^T\underline{\mathbf{x}}|^2/\underline{\mathbf{x}}^T\underline{\mathbf{x}} = z$ (cone in $\mathbf{R}^3$)



(a) Sufficiency orbit

(b) Invariance orbit

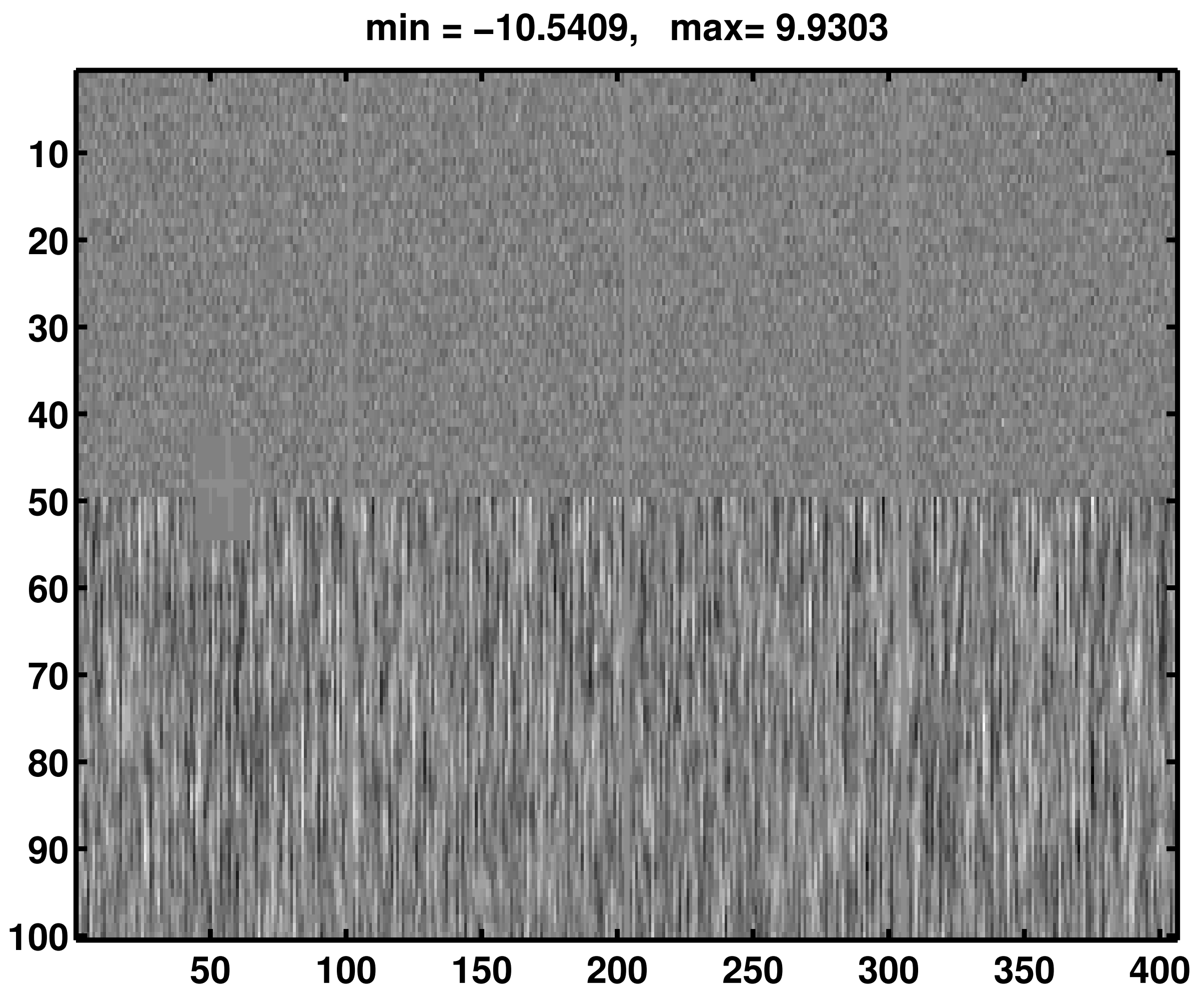


Figure 4: Deep hide target on clutter boundary

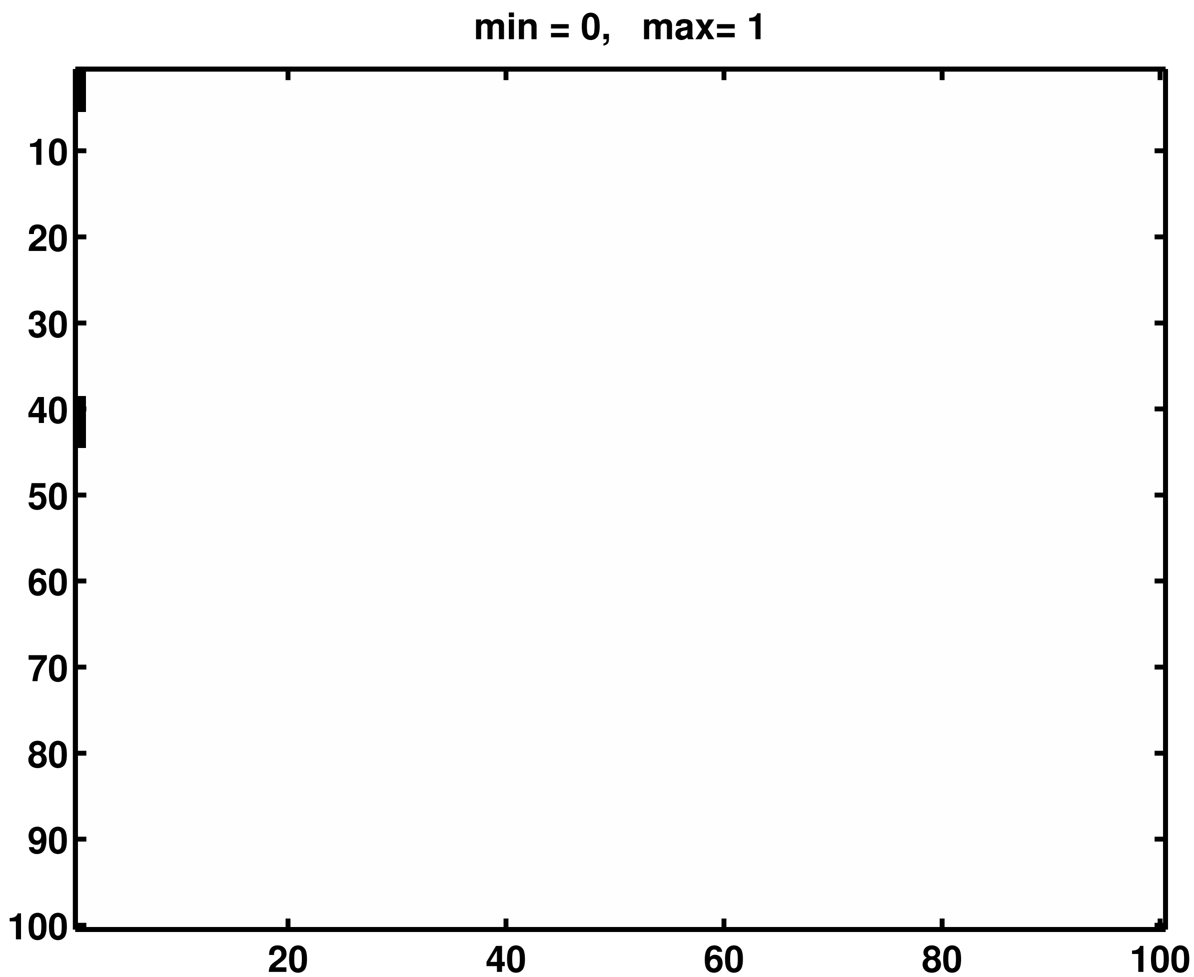


Figure 5: Transformed target matrix

Multivariate Model for Inhomogeneous Clutter

$$\mathbf{X} = a \cdot \begin{bmatrix} \underline{s}_A \\ \underline{s}_B \end{bmatrix} \underline{b}^T + \begin{bmatrix} \mathbf{N}_A \\ \mathbf{N}_B \end{bmatrix}$$

$\Rightarrow$ canonical form:

$$\mathbf{X} = a \begin{bmatrix} 1 \\ \underline{0} \\ 1 \\ \underline{0} \end{bmatrix} \cdot [1, 0, \ldots, 0] + \begin{bmatrix} \mathbf{N}_A \\ \mathbf{N}_B \end{bmatrix}$$

$$\Theta_0 = \{a = 0, \mathbf{R}_A > 0, \mathbf{R}_B > 0\}, \qquad \Theta_1 = \{a \neq 0, \mathbf{R}_A > 0, \mathbf{R}_B > 0\}$$

Smallest group of transformations leaving decision problem invariant:

$\mathcal{G}$ has group action: $g(\mathbf{X}) = \mathbf{FXH}$,

- $\mathbf{F}$ invertible $m \times m$ of form:

$$\mathbf{F} = \gamma \begin{bmatrix} \begin{matrix} \mathbf{I}_A & \mathbf{C}_A \\ \underline{0} & \Gamma_A \end{matrix} & \mathbf{O} \\ \mathbf{O} & \begin{matrix} \mathbf{I}_B & \mathbf{C}_B \\ \underline{0} & \Gamma_B \end{matrix} \end{bmatrix}$$

- $\mathbf{H}$ unitary $n \times n$ of form:

$$\mathbf{H} = \begin{bmatrix} 1 & \underline{0}^T \\ \underline{0} & \mathbf{U} \end{bmatrix}$$

Let

$$\mathbf{X} = \begin{bmatrix} \underline{x}_{A1} & \mathbf{X}_{A2} \\ \underline{x}_{B1} & \mathbf{X}_{B2} \end{bmatrix} \qquad = \begin{bmatrix} \underline{x}_{A11} & \mathbf{X}_{A12} \\ \underline{x}_{A21} & \mathbf{X}_{A22} \\ \underline{x}_{B11} & \mathbf{X}_{B12} \\ \underline{x}_{B21} & \mathbf{X}_{B22} \end{bmatrix}$$

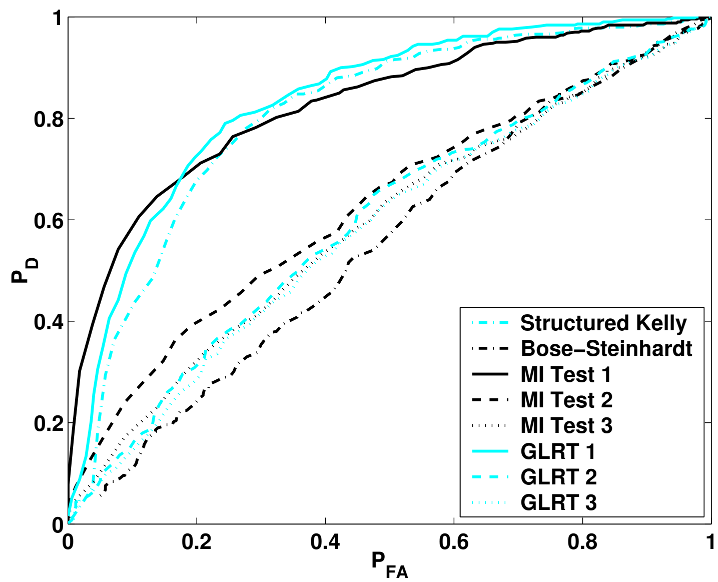
<u>Maximal Invariant</u> consists of seven terms

$$z_{1A}(\mathbf{X}) = \underline{u}_A^T \mathbf{D}_A^{-1} \underline{u}_A, \qquad z_{1B}(\mathbf{X}) = \underline{u}_B^T \mathbf{D}_B^{-1} \underline{u}_B$$

$$z_{2A}(\mathbf{X}) = \underline{x}_{A21}^T [\mathbf{X}_{A22} \mathbf{X}_{A22}^T]^{-1} \underline{x}_{A21}, \quad z_{2B}(\mathbf{X}) = \underline{x}_{B21}^T [\mathbf{X}_{B22} \mathbf{X}_{B22}^T]^{-1} \underline{x}_{B21}$$

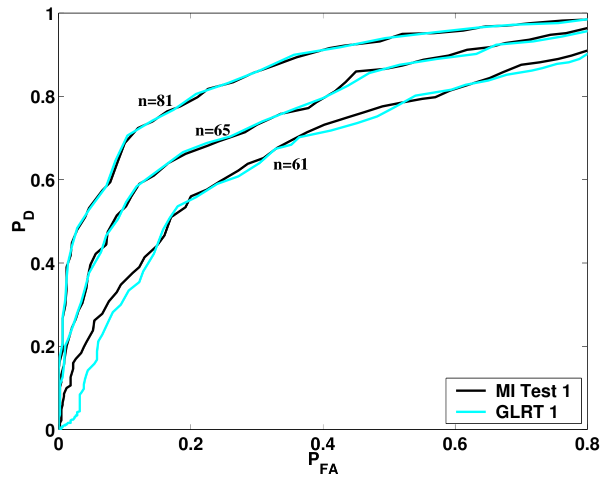
$$z_{3A}(\mathbf{X}) = \frac{\underline{u}_A \underline{u}_A^T}{\|\underline{u}_A\|^2}, \qquad z_{3B}(\mathbf{X}) = \frac{\underline{u}_B \underline{u}_B^T}{\|\underline{u}_B\|^2}$$

$$z_4(\mathbf{X}) = \frac{\underline{u}_A}{\underline{u}_B}$$

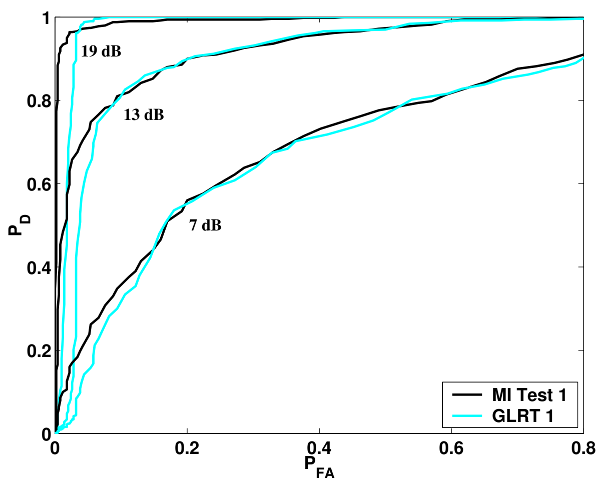


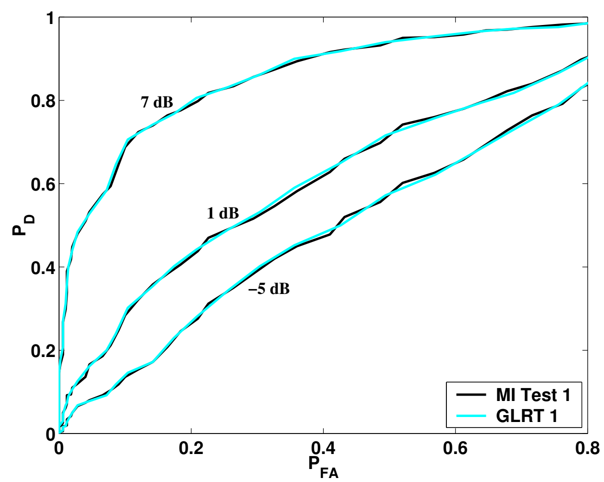
Scenario

$$SNR = 22dB \; (SNR_A = 11dB \;/\; SNR_B = 22dB)$$
$$m_A = 50 \;/\; m_B = 50 \;,\; n = 51$$



$(SNR = 7dB)$





$(n = 61)$ $(n = 81)$

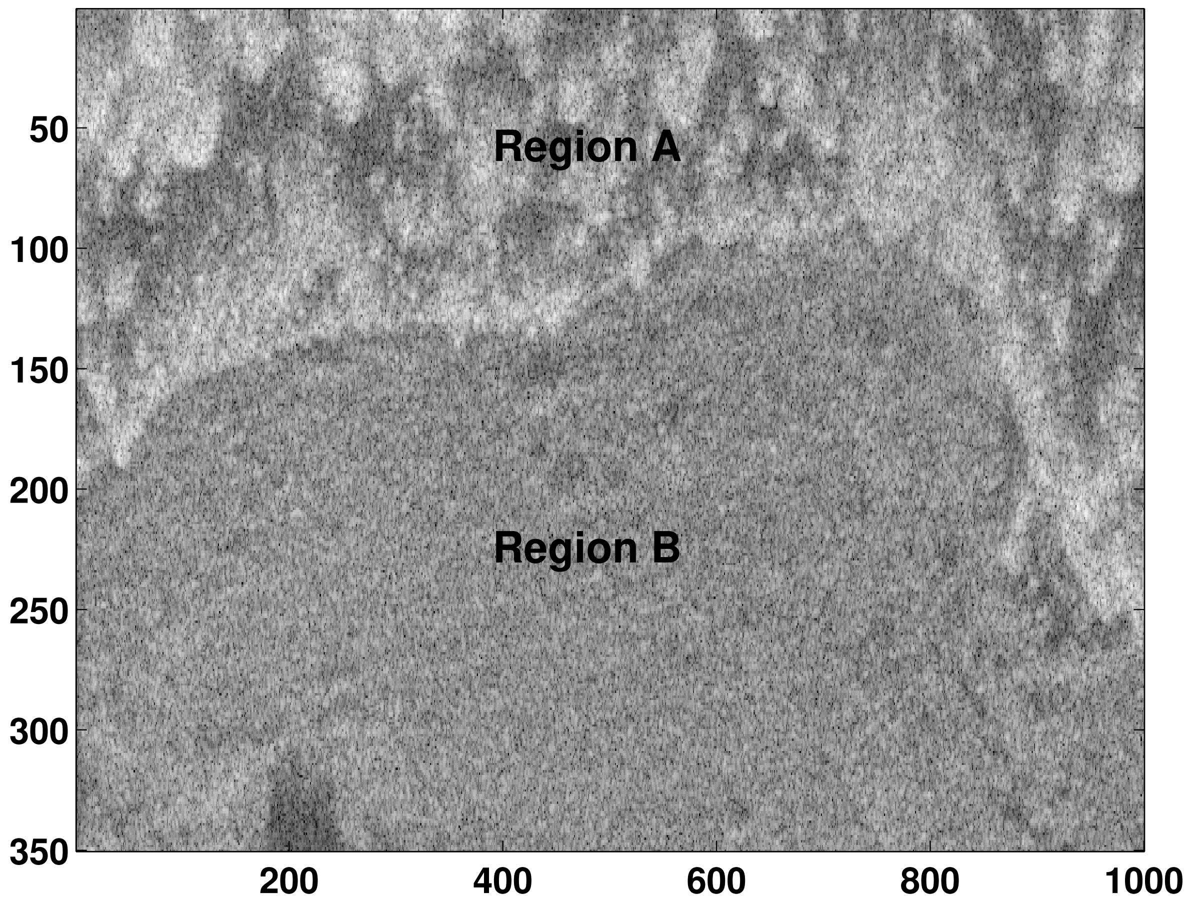


Figure 6: SAR clutter image with a single target (e) straddling the boundary at column 305.

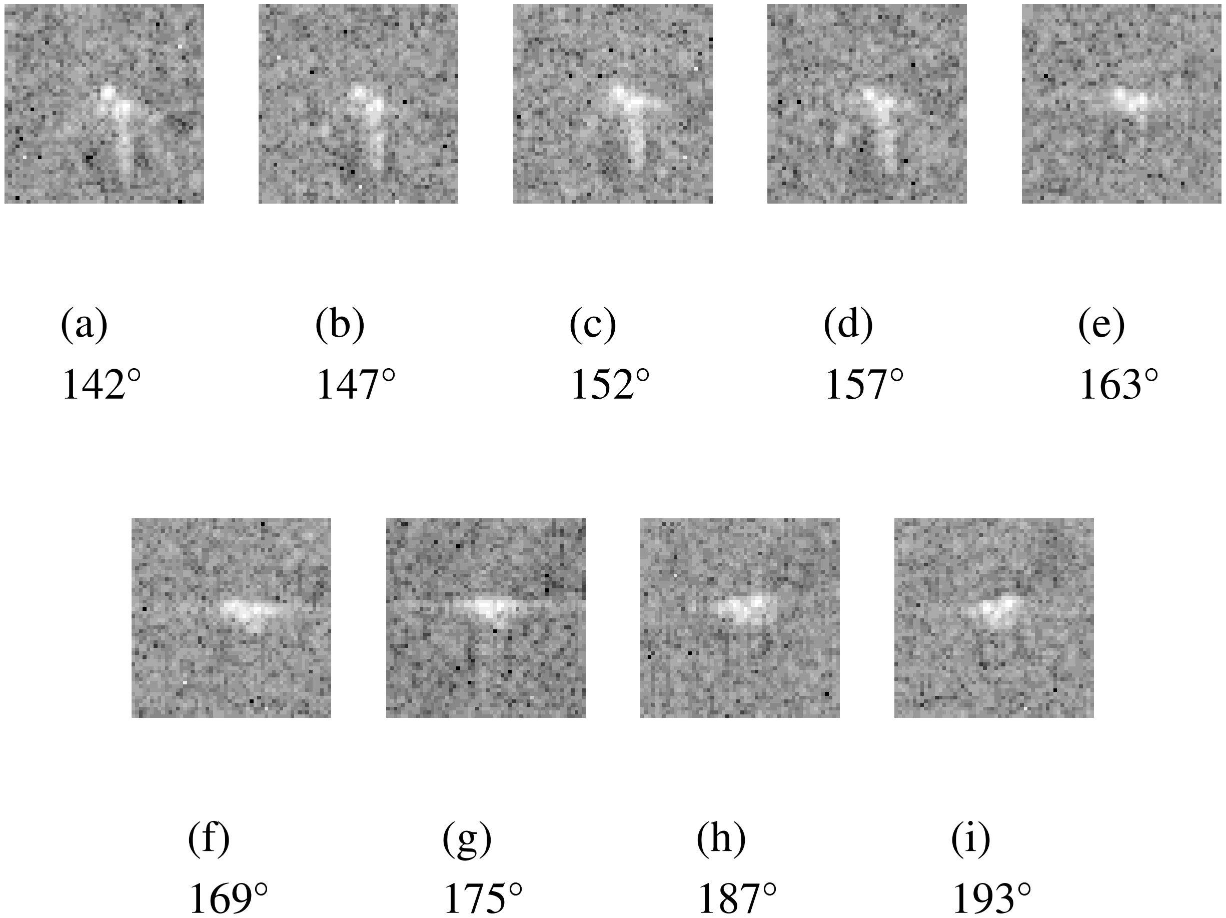


Figure 7: SLICY canonical target images (54×54) at elevation 39° and different azimuth angles. Image in (e) is inserted in Figure 6.

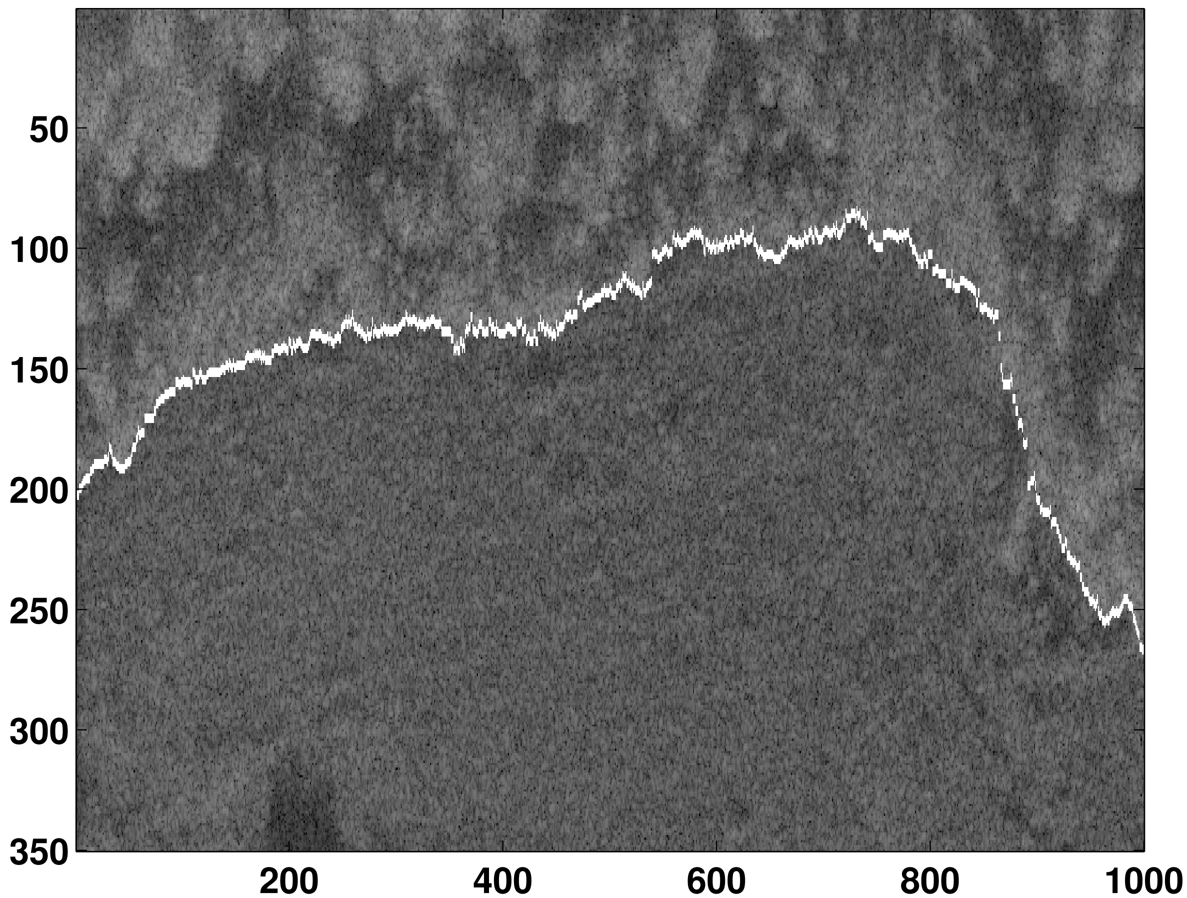


Figure 8: Hand-extracted boundary superimposed on the initial image in Figure 6.

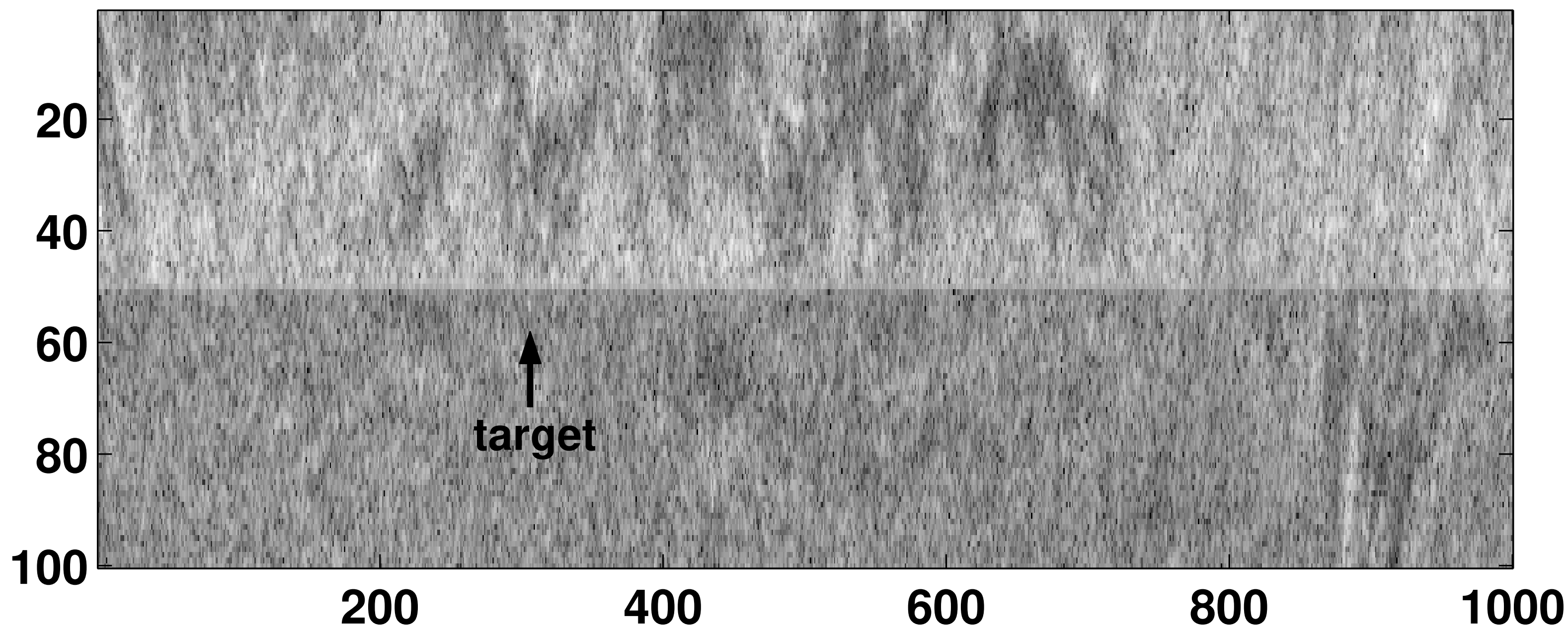


Figure 9: Image realigned along the extracted boundary. SLICY target is located at column 305 with $|a| = 0.015$. This target is just above the minimal detectable threshold for the three tests investigated in Figure **??**.

Minimum detectable thresholds

Test	$\|a\|$	
	$(n-1=250)$	$(n-1=200)$
MI test 1	1.454×10^{-2}	0.609×10^{-1}
GLR 1	1.462×10^{-2}	1.042×10^{-1}
Structured Kelly	1.407×10^{-2}	1.049×10^{-1}

Table 1: Minimum detectable amplitudes for detection of the target at the correct location.

Non-parametric Approaches to 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 $\hat{D}_\alpha(f_1\|f_0)$ from Z_i's

1. Density estimation methods
2. Minimal graph methods

Indirect Methods via Density estimation

$$\hat{D}_\alpha(f_1\|f_0) = \frac{1}{\alpha-1}\ln\int_{\mathbf{R}^d}\hat{f_1}^{\alpha}(x)\hat{f}_0^{1-\alpha}dx$$

Previous work only for Shannon entropy estimation ($\alpha = 1$)

- Histogram plug-in [Gyorfi&VanDerMeulen:CSDA87]
- Kernel plug-in [Ahmad&Lin:IT76]
- Sample-spacing plug-in [Hall:JMS86]
  - No unbiased density estimators exist
  - Performance degrades for unsmooth densities
  - Unclear how to robustify $\hat{f}_1, \hat{f}_0$ against outliers
  - d-dimensional integration might be difficult
  - $\Rightarrow$ function $\{f(x) : x \in \mathbf{R}^d\}$ over-parameterizes entropy functional

Minimal graph methods

- MST estimator of α-entropy [Hero&Michel:IT99]
- MST estimator of α-diverence [Hero&Michel:HOS99]
- k-MST robust estimator [Hero&Michel:IT99]
  - Direct entropy estimator: faster convergence for unsmooth densities
  - Parameter α is varied by varying interpoint distance measure
  - Optimally pruned graphs robustify $\hat{f}_1, \hat{f}_0$ against outliers
  - Greedy MST approximations are possible
  - Combinatorial optimization is bottleneck for large number of samples

Minimal Graphs

A graph G of degree l consists of vertices and edges

- vertices are subset of $\mathcal{X}_n = \{x_i\}_{i=1}^n$: n points in $\mathbf{R}^d$
- edges are denoted $\{e_{ij}\}$
- for any i: $card\{e_{ij}\}_j \leq l$

Weight (with power exponent γ) of G

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

Examples:

Minimal Spanning Tree (MST)

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

The Euclidean Power Weighted MST achieves

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

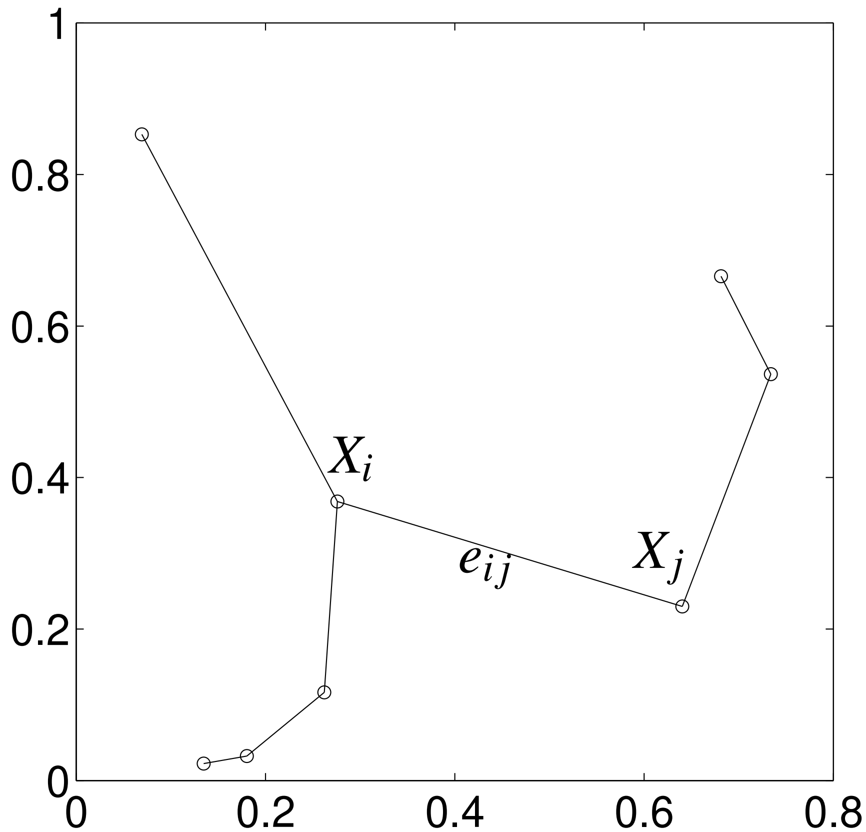


Figure 10: A MST example

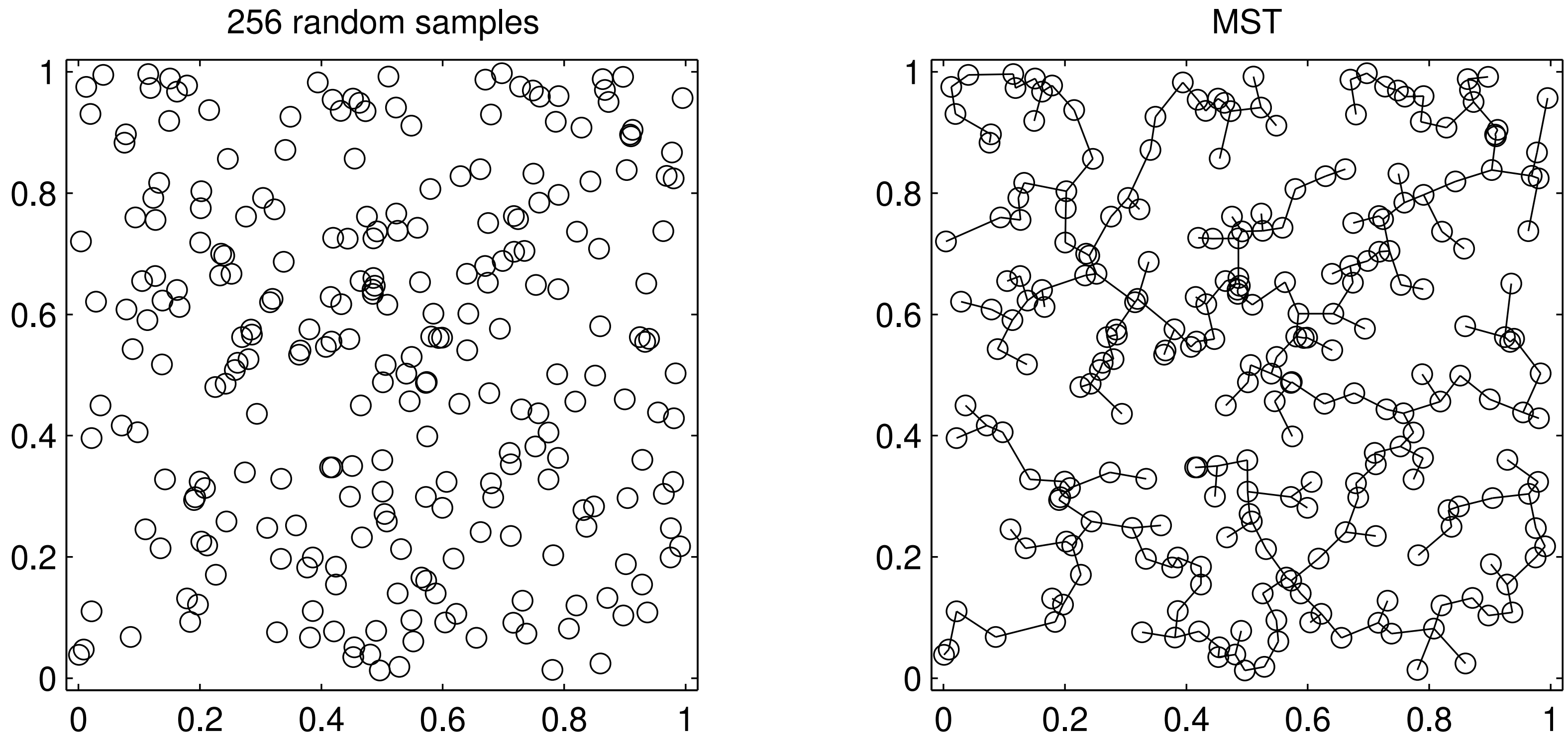


Figure 11: *A sample data set and the MST*

Traveling Salesman Problem (TSP)

Let $\mathrm{T}(\mathcal{X}_n)$ be sets of edges in the class of graphs of degree 2 spanning $\mathcal{X}_n$.

The minimal power-weighted TSP tour achieves

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

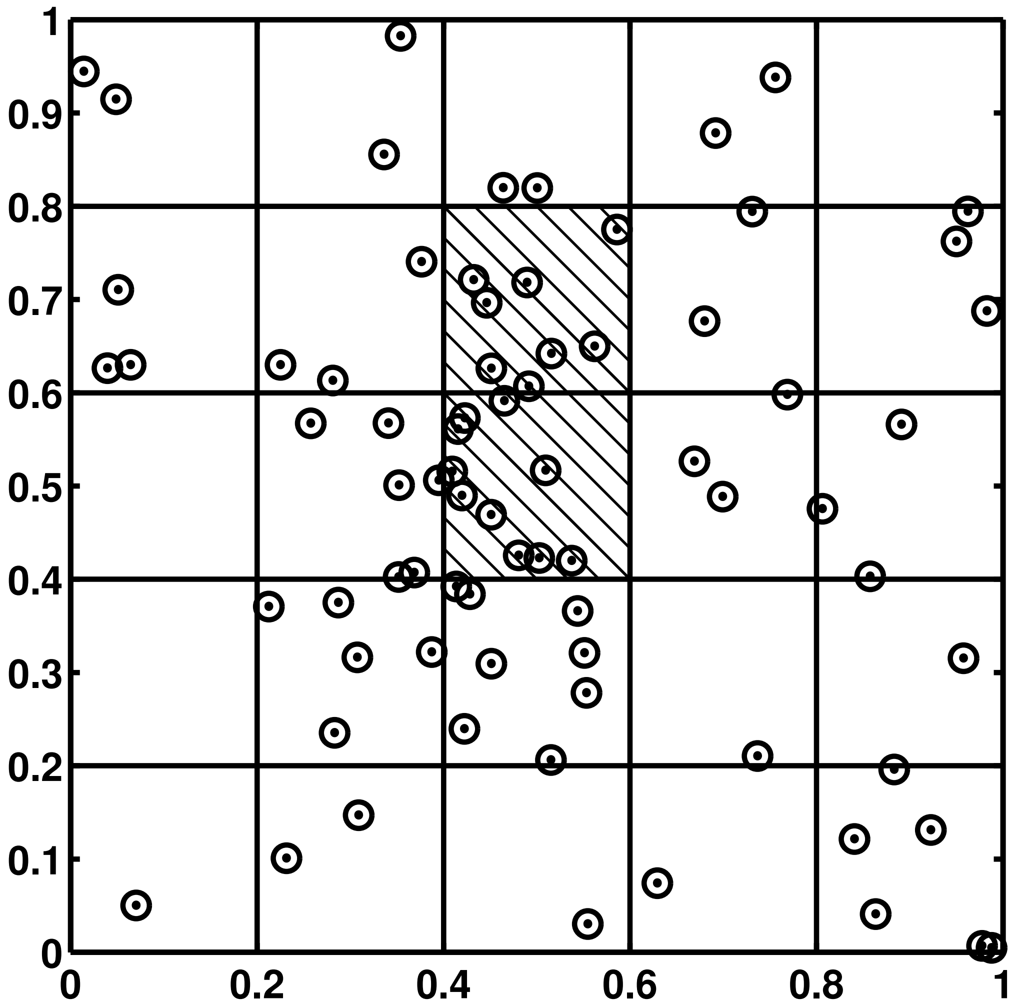


Figure 12: *A sample of 75 points from the mixture density* $f(x) = 0.25 f_1(x) + 0.75 f_o(x)$ *where* f_o *is a uniform density over* $[0,1]^2$ *and* f_1 *is a bivariate Gaussian density with mean* $(1/2, 1/2)$ *and diagonal covariance* $\mathrm{diag}(0.01)$. *A smallest subset* B_k^m *is the union of the two cross hatched cells shown for the case of* $m = 5$ *and* $k = 17$.

Quasi-additive Euclidean Functionals

L is a continuous subadditive functional [Yukich:98] if it satisfies

Null Condition: $L(\phi) = 0$, where ϕ is the null set.

Subadditivity: There exists a constant C_1 with the following property: For any uniform resolution $1/m$-partition Q^m

$$L(F) \leq m^{-1} \sum_{i=1}^{m^d} L(m[(F \cap Q_i) - q_i]) + C_1 m^{d-\gamma}$$

Superadditivity: There exists a constant C_2 with the following property:

$$L(F) \geq m^{-1} \sum_{i=1}^{m^d} L(m[(F \cap Q_i) - q_i]) - C_2 m^{d-\gamma}$$

Continuity: There exists a constant C_3 such that for all finite subsets F and G of $[0,1]^d$

$$|L(F \cup G) - L(F)| \leq C_3 \,(card(G))^{(d-\gamma)/d}$$

Definition 1 *A continuous subadditive functional L is said to be a quasi-additive functional when there exists a continuous superadditive functional L^* which satisfies $L(F) + 1 \geq L^*(F)$ and the approximation property*

$$|E[L(U_1, \ldots, U_n)] - E[L^*(U_1, \ldots, U_n)]| \leq C_4 n^{(d-\gamma-1)/d} \qquad (1)$$

where $U_1, \ldots, U_n$ are i.i.d. uniform random vectors in $[0,1]^d$.

Asymptotics: the BHH Theorem and entropy estimation

Theorem 1

[Beardwood&Halton&Hammersley:Camb59,Redmond&Yukich:SPA96] *Let L be a 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*

$$\lim_{n\to\infty} L(\mathcal{X}_n)/n^{(d-\gamma)/d} = \beta_{L,\gamma} \int f(x)^{(d-\gamma)/d} dx, \qquad (a.s.) \tag{2}$$

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

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

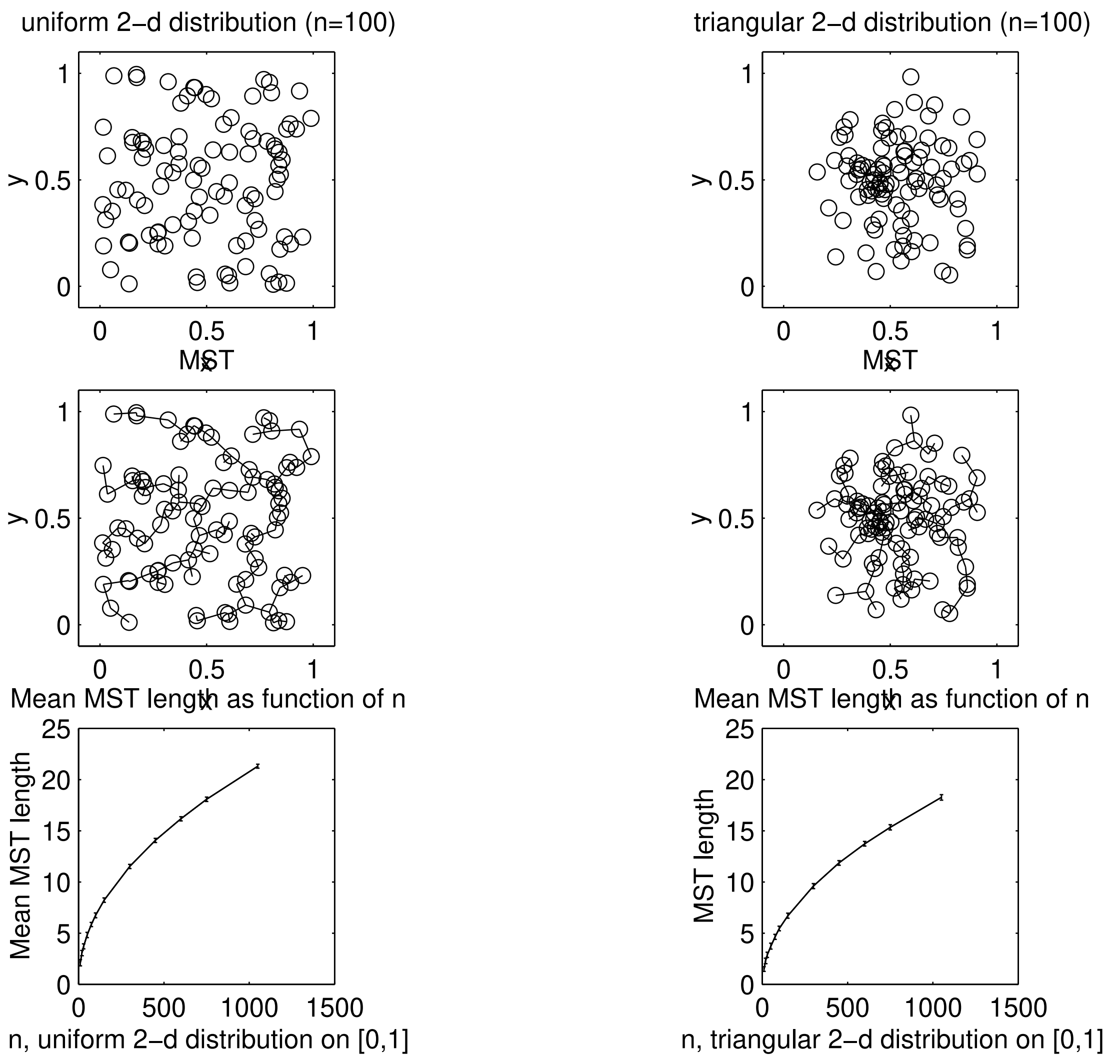


Figure 13: *2D Triangular vs. Uniform sample study for MST.*

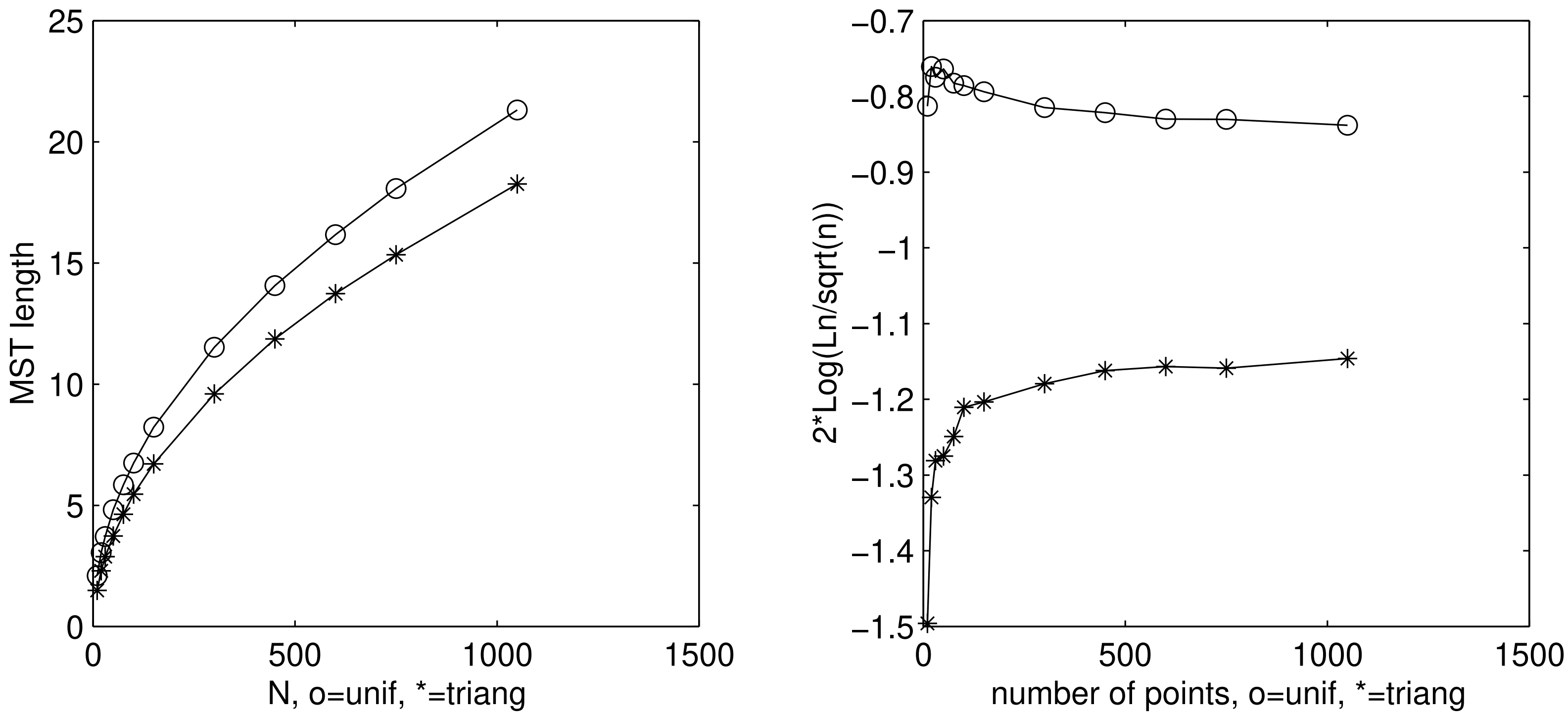


Figure 14: *MST and log MST weights as function of number of samples for 2D uniform vs. triangular study.*

Asymptotics: Indirect estimation of $H_\nu(f)$

Define the class of Holder continuous functions $\Sigma_d(\kappa, c)$ over $\mathbf{R}^d$

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

where $p_x^k(z)$ is the Taylor polynomial of f of order k expanded about the point x.

Proposition: *[Hero&Michel:99] Assume that $f \in \Sigma_d(\kappa, c)$. Then, with $\hat{f}$ any* **minimax** *density estimator*

$$\left| \int \hat{f}^{\frac{d-\gamma}{d}}(x)dx - \int f^{\frac{d-\gamma}{d}}(x)dx \right| \leq n^{-\kappa/(2\kappa+d)}\, C_{\kappa,c} \int f^{\frac{\kappa}{2\kappa+d}-\frac{\gamma}{d}}(x)dx$$

where $C_{\kappa,c}$ is a constant independent of f.

Asymptotics: Direct Estimation of $H_\nu(f)$

Proposition: *[Ma&Hero:00] Assume that for* $\nu \in (0,1)$, f^ν *is of bounded variation over* $[0,1]^d$. *Then for* $d \geq 2$

$$\left| \frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - \beta_{L,\gamma} \int f^{\frac{d-\gamma}{d}}(x)dx \right| \leq n^{-1/d} K_{L,\gamma} \int f^{\frac{d-1}{d}-\frac{\gamma}{d}}(x)dx$$

where $K_{L,\gamma}$ *is a constant independent of* f.

Conclude: Direct estimator converges faster than indirect when:

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

I-Divergence and Quasi-additive functions

- $f_0(x)$: a reference density on $\mathbf{R}^d$
- Assume $f \ll f_0$, i.e. for all x such that $f_0(x) = 0$ we have $f(x) = 0$.
- Make measure transformation $dx \to f_0(x)dx$ on $[0,1]^d$. Then for $\mathcal{Y}_n$ = transformed data [Hero&Michel:HOS99]

$$\lim_{n\to\infty} L(\mathcal{Y}_n)/n^{\nu} = \beta_{L,\gamma} \exp\left((1-\nu)D_{\nu}(f \| f_0)\right), \qquad (a.s.)$$

Outlier Sensitivity of minimal n-point graphs

Assume f is a mixture density of the form

$$f = (1-\varepsilon) f_1 + \varepsilon f_o, \tag{3}$$

where

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

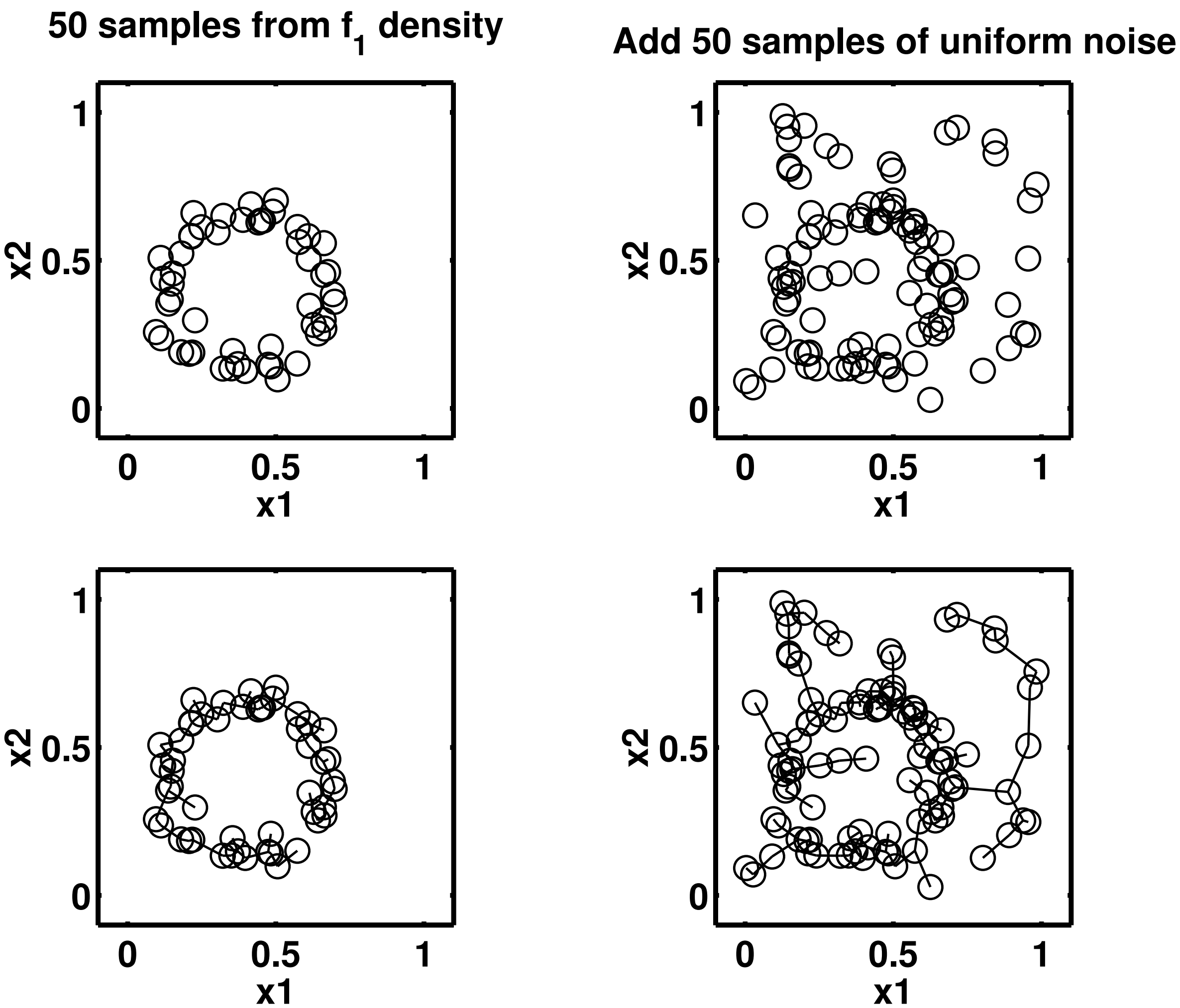


Figure 15: *1st row: 2D torus density with and without the addition of uniform "outliers." 2nd row: corresponding MST's.*

Minimal k-point Euclidean Graphs

Refs: [Zelikovsky&Lozevanu:93,Ravi&etal:SIAM96,Garg:FCS96]

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

Let $\mathrm{T}_{n,k} = \mathrm{T}(x_{i_1}, \ldots, x_{i_k})$ be a minimal graph connecting k distinct vertices $x_{i_1}, \ldots, x_{i_k}$.

The power weighted *k-minimal* graph $\mathrm{T}^*_{n,k} = \mathrm{T}^*(x_{i_1^*}, \ldots, x_{i_k^*})$ is the overall minimum weight k-point graph

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

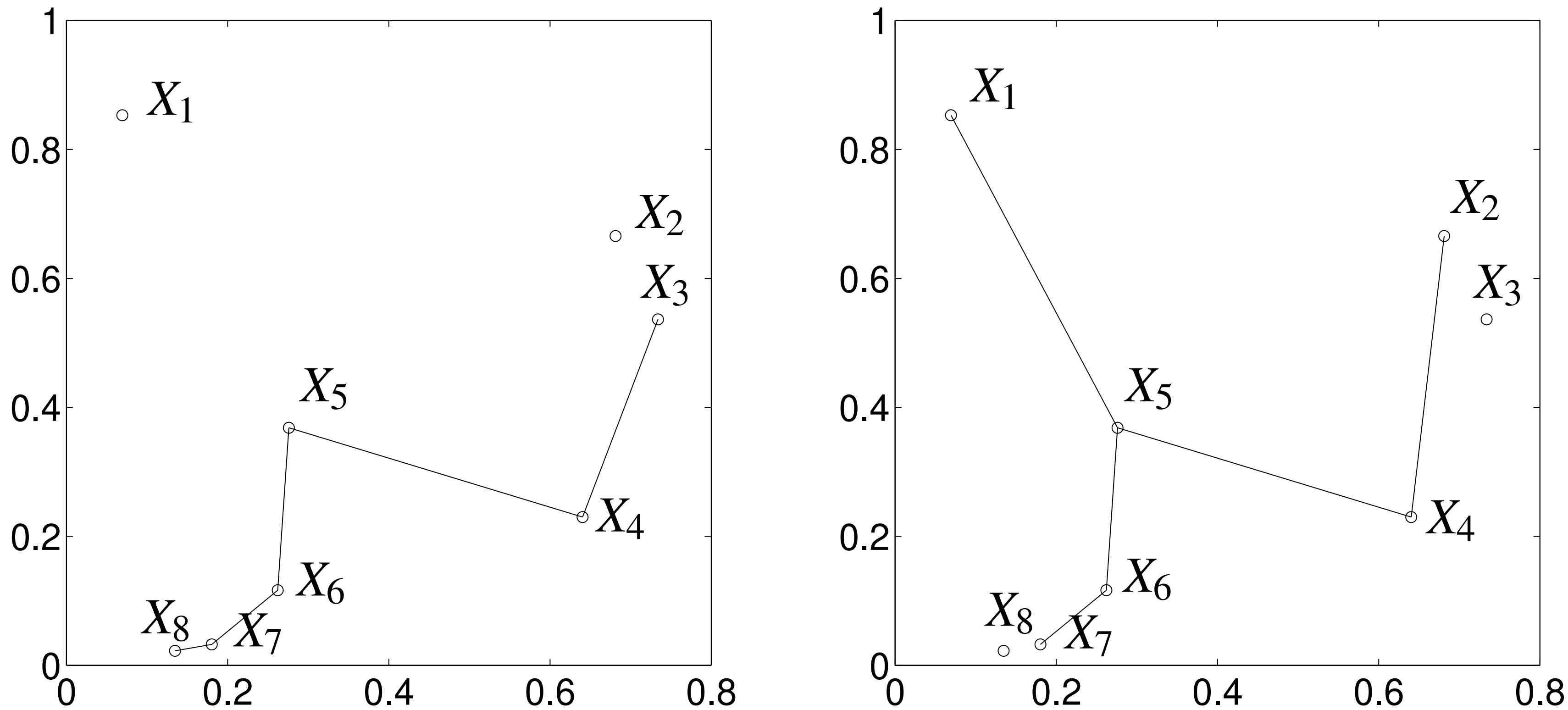


Figure 16: Examples for k-point MST's

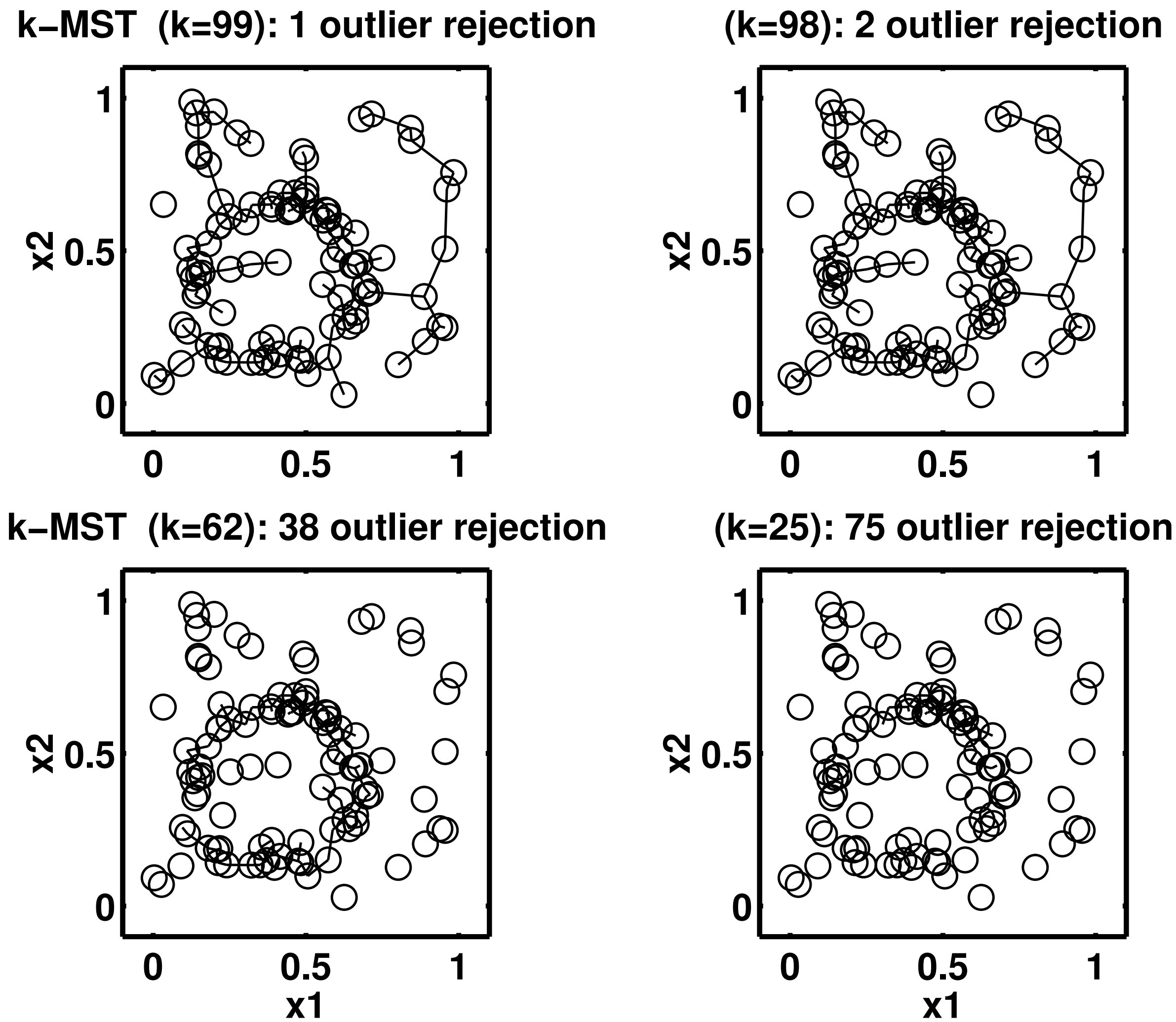
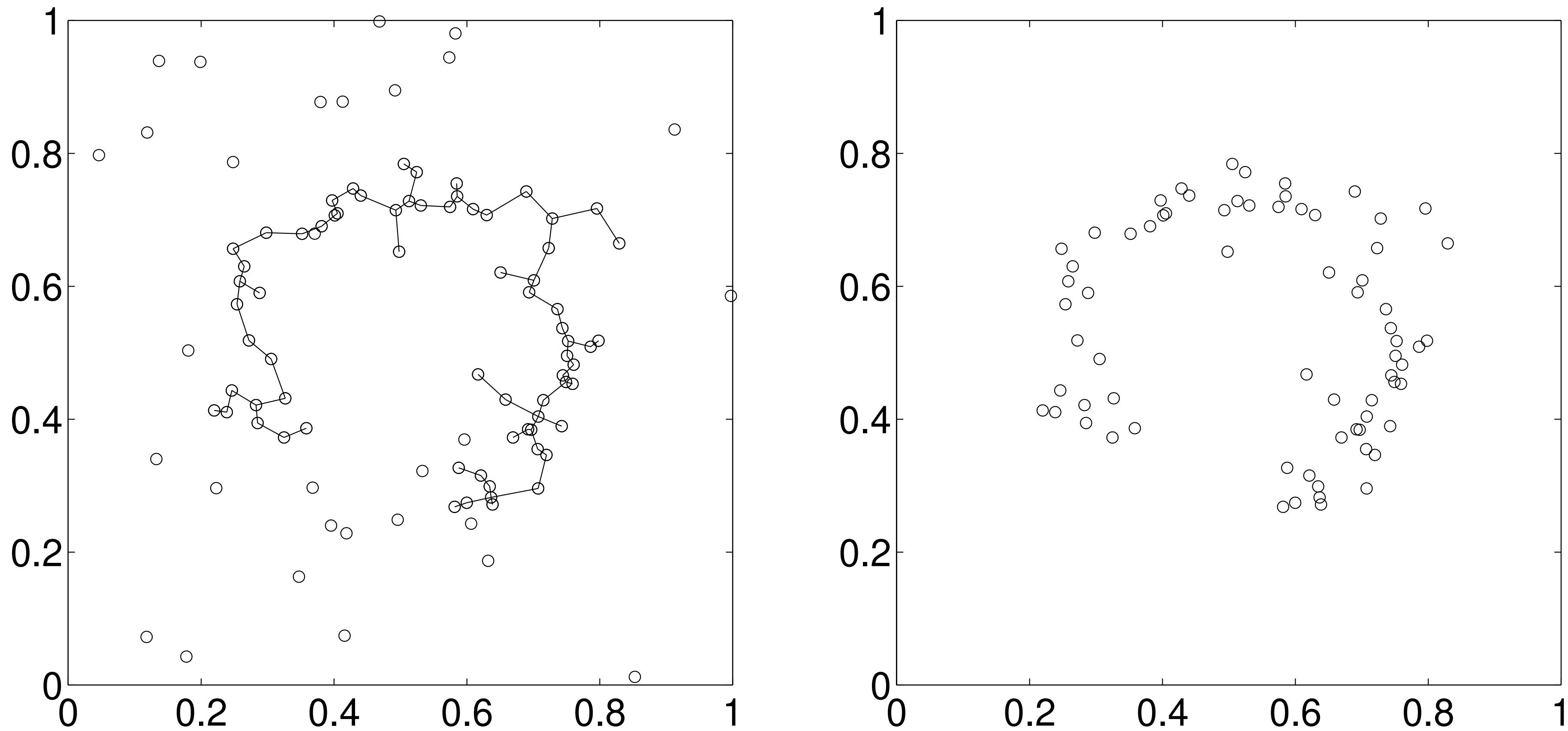


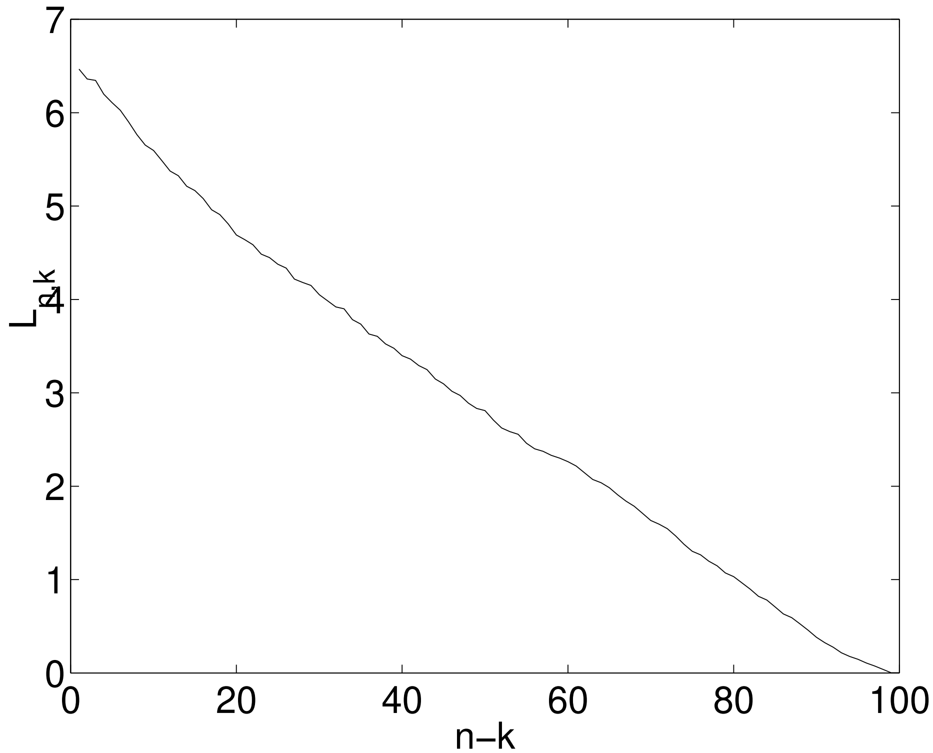
Figure 17: *k-MST for 2D torus density with and without the addition of uniform "outliers".*



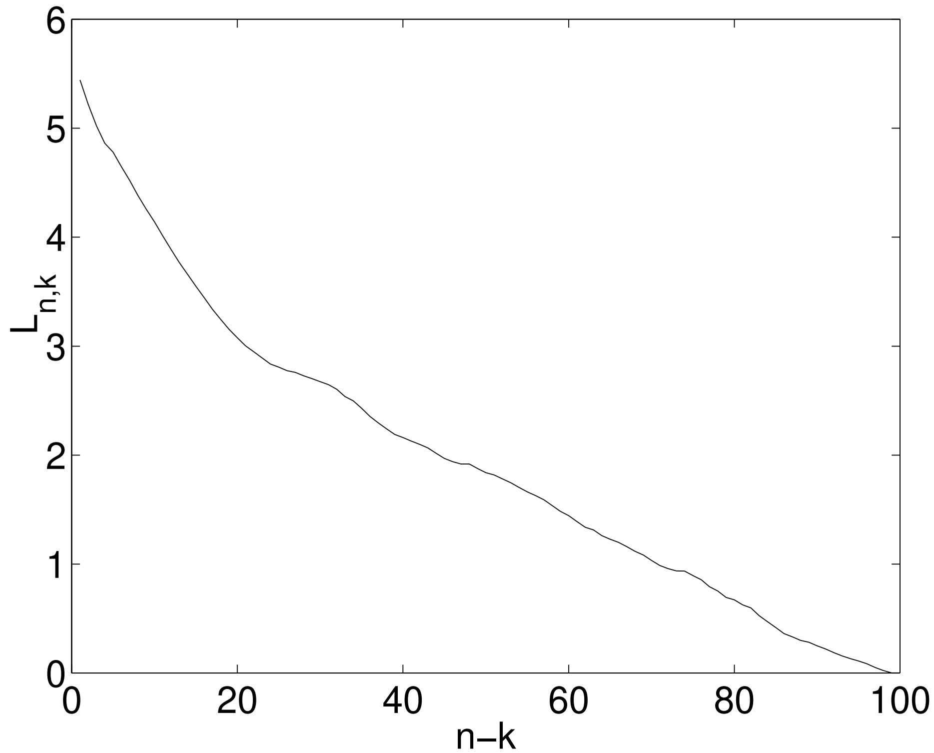
(a) *k*-MST over contaminated signal

(b) After removing outliers

Figure 18: Use *k*-MST to remove outliers



(a) k-MST length vs. number of rejected points for uniform samples



(b) k-MST length vs. number of rejected points for contaminated samples

Figure 19: k-MST length vs. number of rejected points

Application I: Nonuniform Outlier Rejection & Density Estimation

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

Objective: estimate f_1 or $D_\alpha(f_1 \| f_R)$

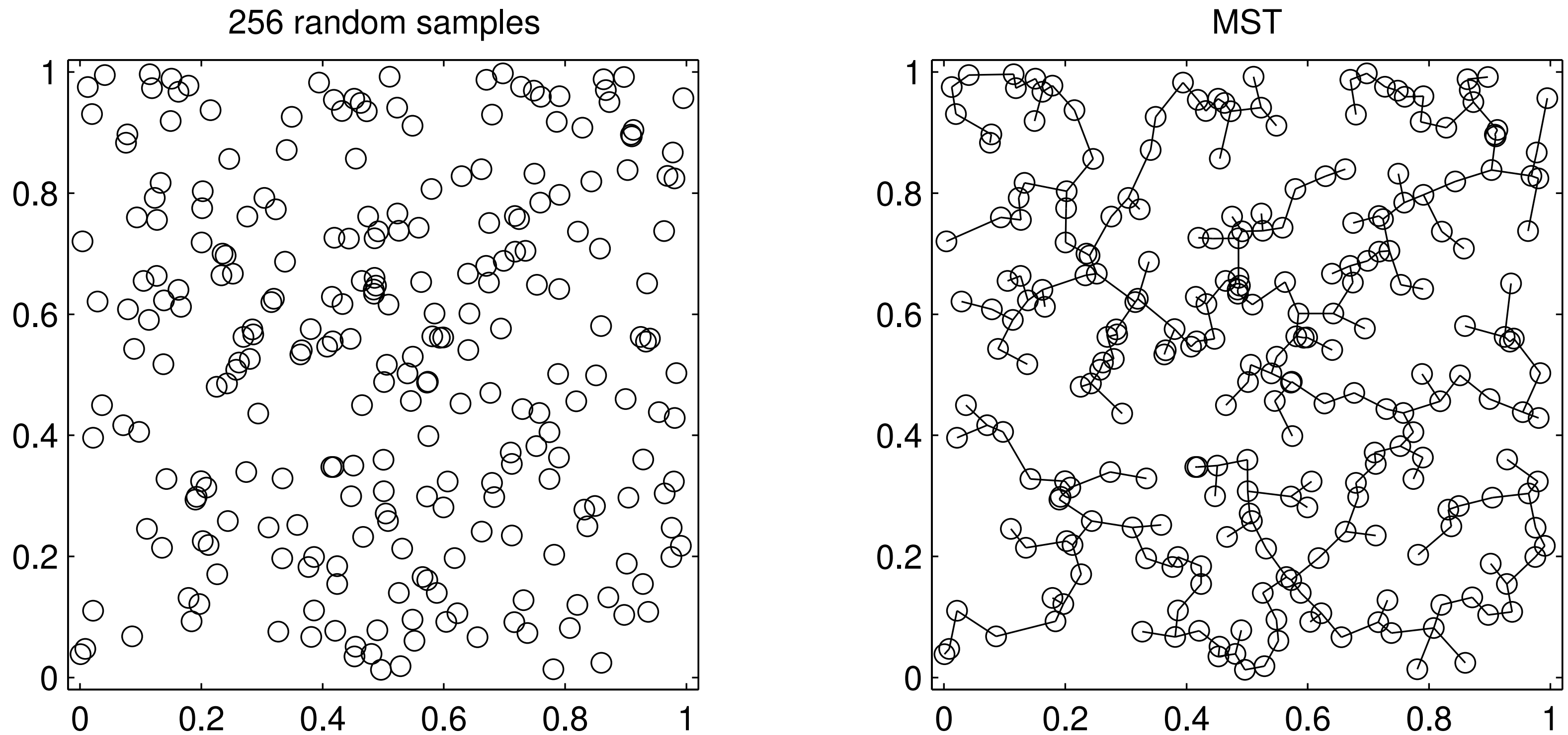


Figure 20: *Left: A scatterplot of a 256 point sample from triangle-uniform mixture density with* $\varepsilon = 0.1$. *Labels 'o' and '*' mark those realizations from the uniform and pyramid densities, respectively. Right: superimposed is the k-MST implemented directly on the scatterplot* $\mathcal{X}_n$ *with* $k = 230$.

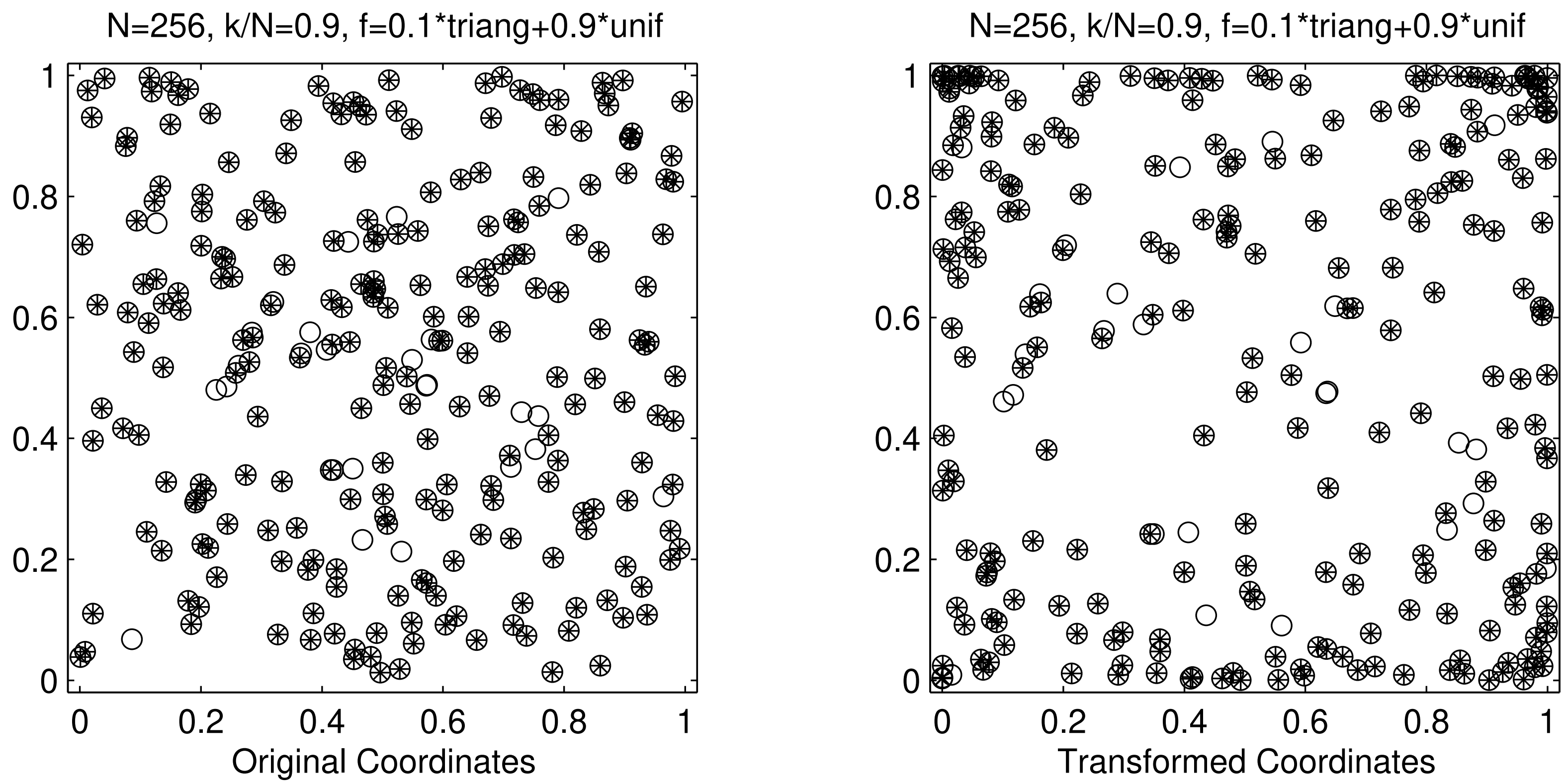


Figure 21: *Left: A sample from triangle-uniform mixture density with* $\varepsilon = 0.9$ *in the transformed domain* $\mathcal{Y}_n$. *Labels 'o' and '*' mark those realizations from the uniform and pyramid densities, respectively. Right: transformed coordinates.*

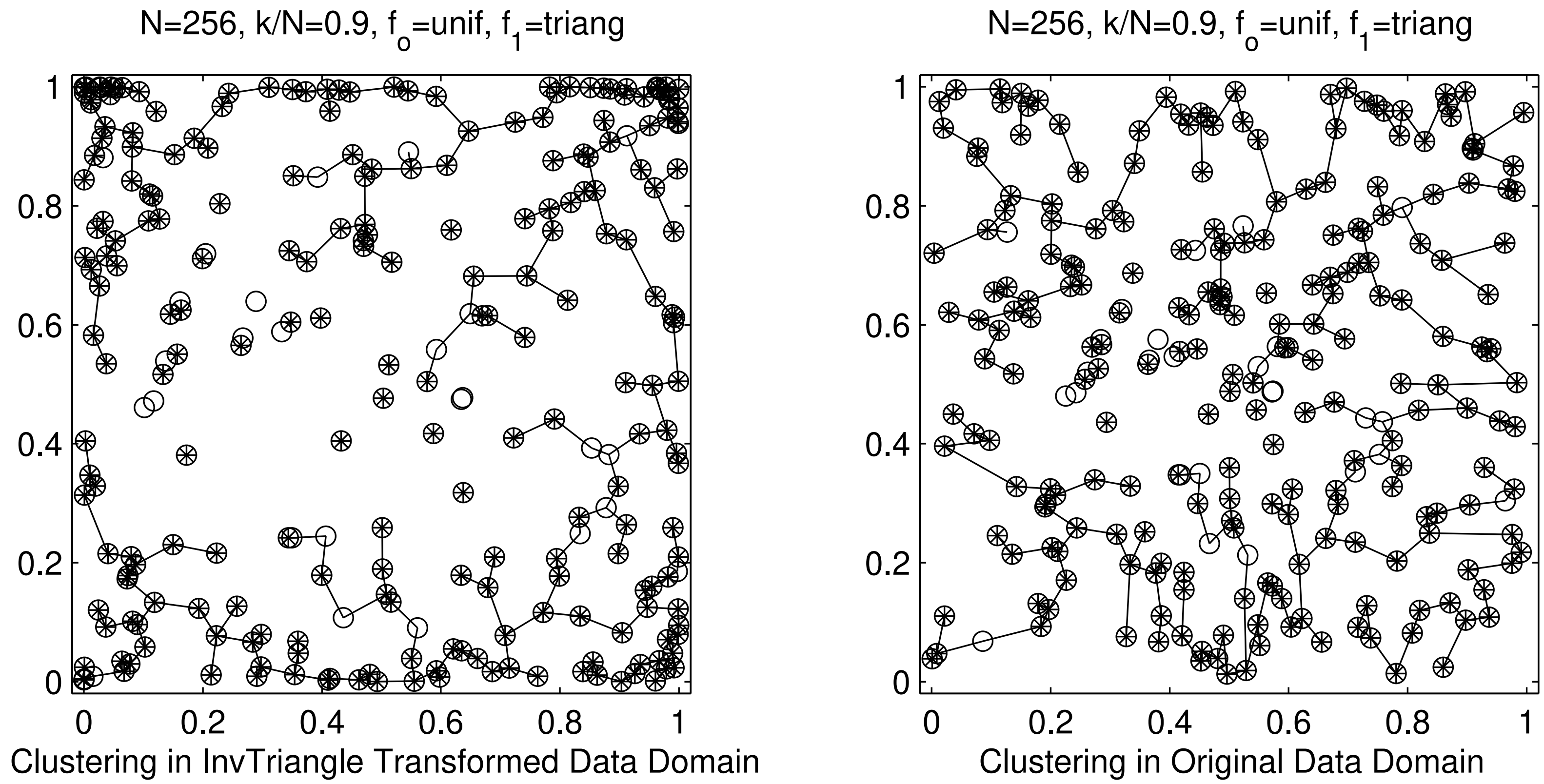


Figure 22: *Left: the k-MST implemented on the transformed scatterplot $\mathcal{Y}_n$ with $k = 230$. Right: same k-MST displayed in the original data domain.*

Application II: Image Registration

Two independent data samples from unknown distributions

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

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

Objective: find rigid transformation A, b

- Jensen's difference:

$$\Delta J_\alpha = H_\alpha(\rho f + (1-\rho)g) - \rho H_\alpha(f) - (1-\rho)H_\alpha(g) \geq 0$$

- where "=" occurs *iff* f=g (a.e.)

Indexing via Jensen's difference

- Merged sample $Z = \{\hat{A}X + \hat{b},\, Y\}$ has density

$$f_Z(z) = \rho f(\hat{A}z + \hat{b}) + (1-\rho)g(z), \quad \rho = \frac{n}{n+m}$$

$\Rightarrow$ $H_\alpha(f_Z)$ minimized for $\hat{A}, \hat{b} = A, b$

- minimum value $\min_{\hat{A},\hat{b}} H(f_Z) = H(f)$

Registration Algorithm: find $\hat{A}, \hat{b}$ to minimize MST length on merged sample

Divergence vs. Jensen: Asymptotic Comparison

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

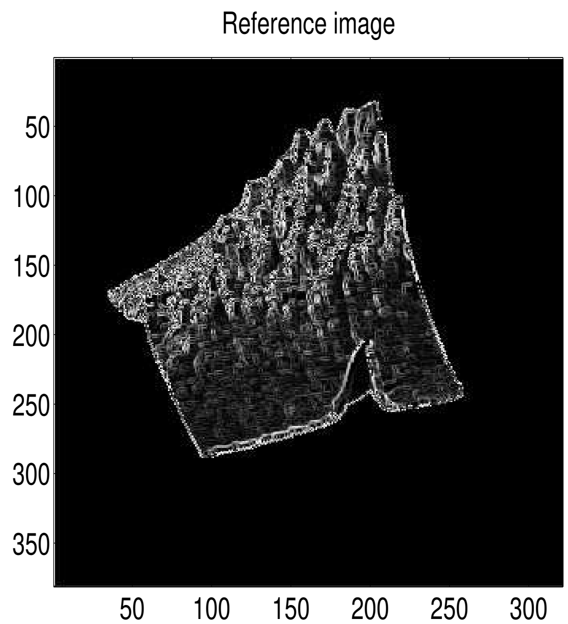
$$f_\rho \quad = \quad \rho f_1 + (1-\rho) f_0, \quad E_g[Z] = \int Z(x) g(x) dx, \quad \tilde{f}^\alpha_{\frac{1}{2}} = \frac{f^\alpha_{\frac{1}{2}}}{\int f^\alpha_{\frac{1}{2}} dx}$$

Then

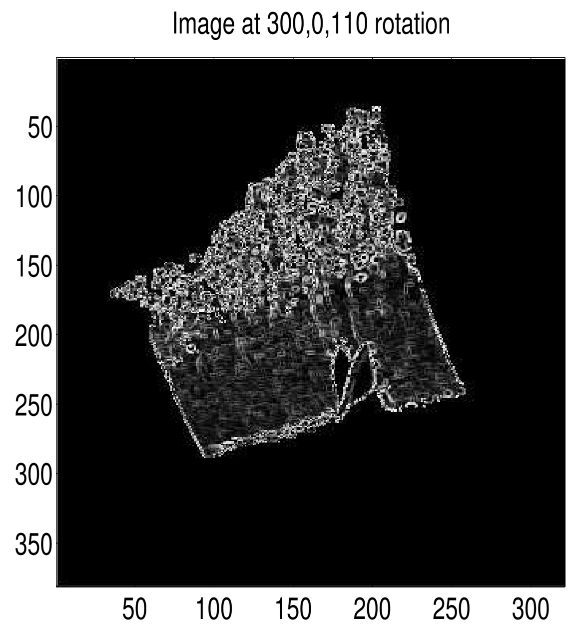
$$\Delta J_\alpha = \frac{\alpha\rho(1-\rho)}{2}\left[E_{\tilde{f}^\alpha_{\frac{1}{2}}}\left(\left[\frac{f_1 - f_0}{f_{\frac{1}{2}}}\right]^2\right) + \frac{\alpha}{1-\alpha} E_{\tilde{f}^\alpha_{\frac{1}{2}}}\left(\left[\frac{f_1 - f_0}{f_{\frac{1}{2}}}\right]\right)^2\right] + O(\Delta_3)$$

$$D_\alpha(f_1 \| f_0) = \frac{\alpha}{4} \int f_{\frac{1}{2}} \left[\frac{f_1 - f_0}{f_{\frac{1}{2}}}\right]^2 dx + O(\Delta_3)$$

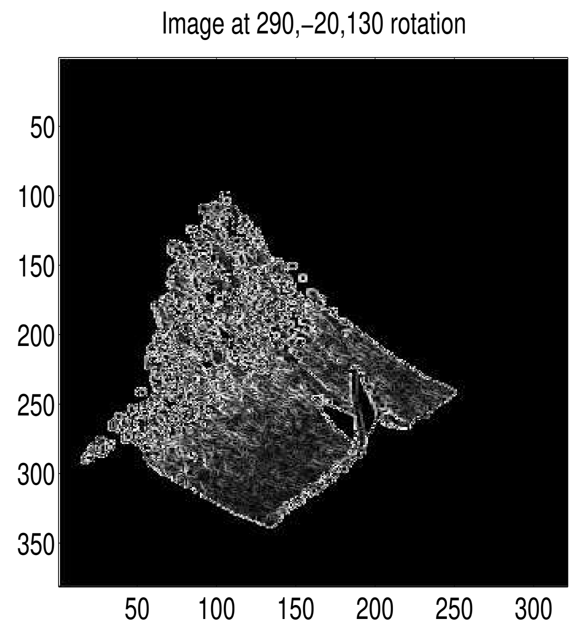
$$\Delta_3 = E_{\tilde{f}^\alpha_{\frac{1}{2}}}\left(\left|\frac{f_1 - f_0}{f_{\frac{1}{2}}}\right|^3\right)$$



(a)

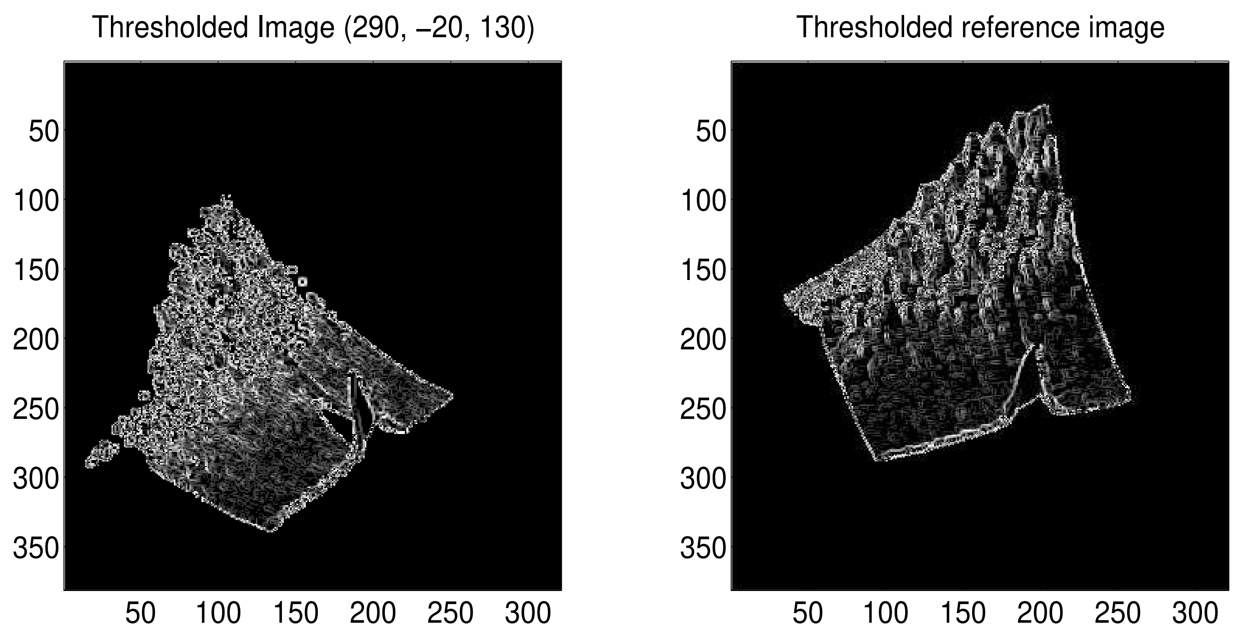


(b)



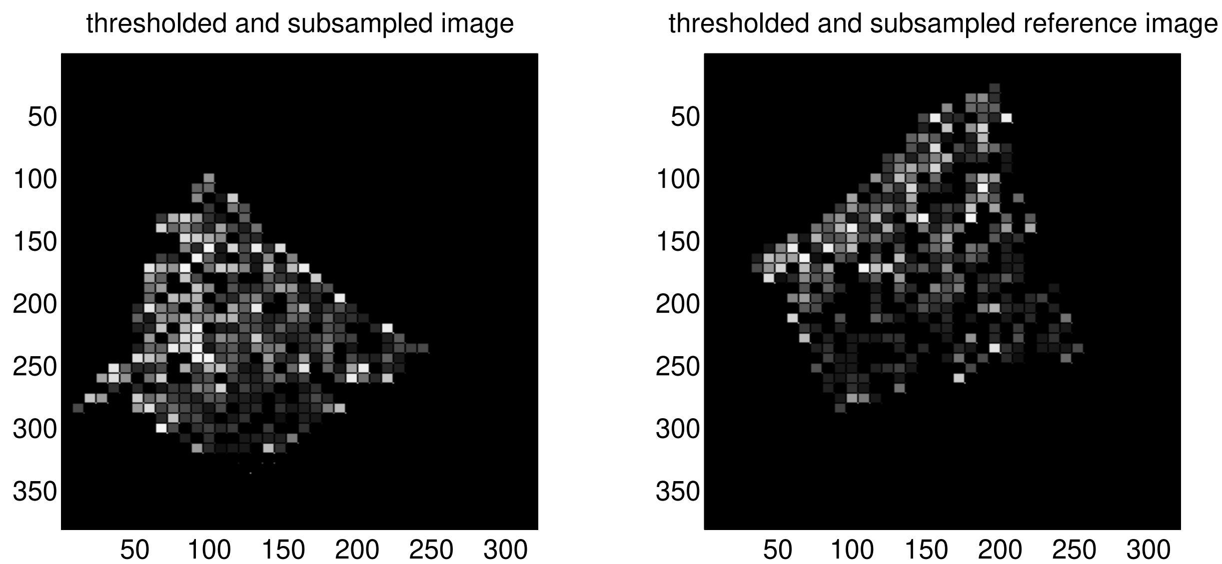
(c)

Figure 23: Reference and target SAR/DEM images



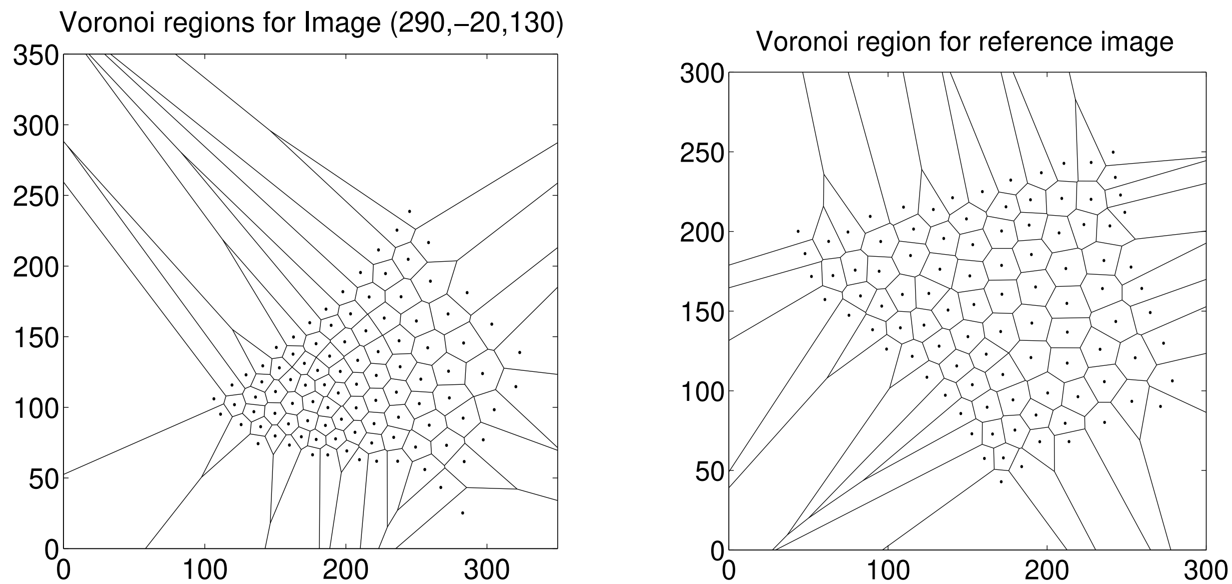
(a) (b)

Figure 24: Thresholded images



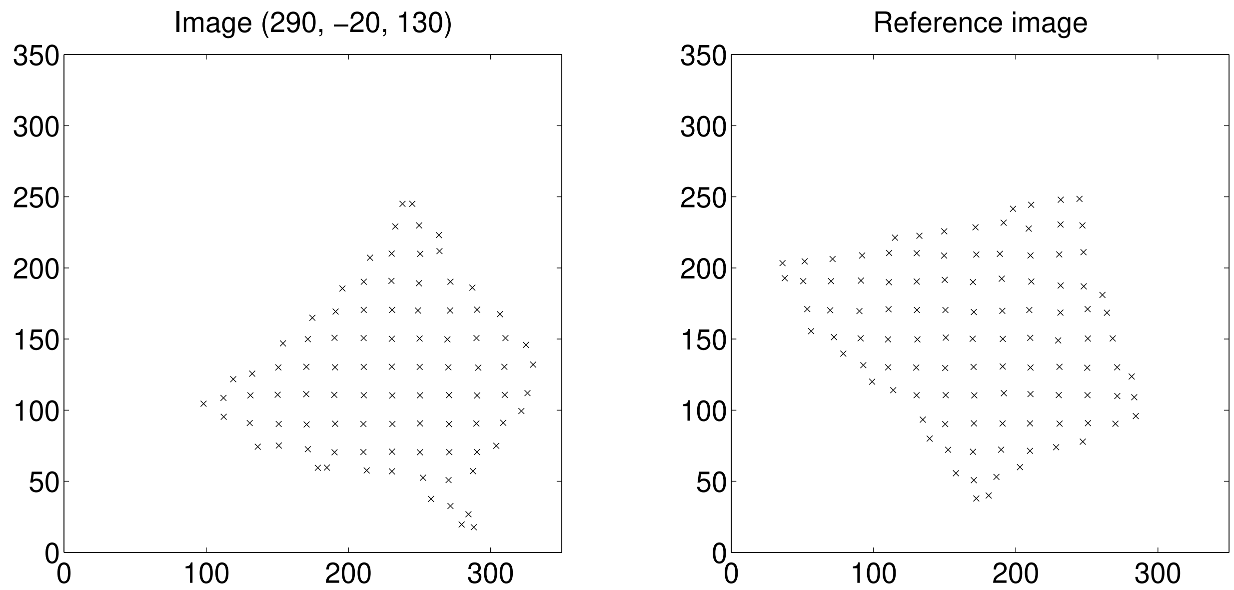
(a) (b)

Figure 25: Feature vectors extracted via uniform sampling



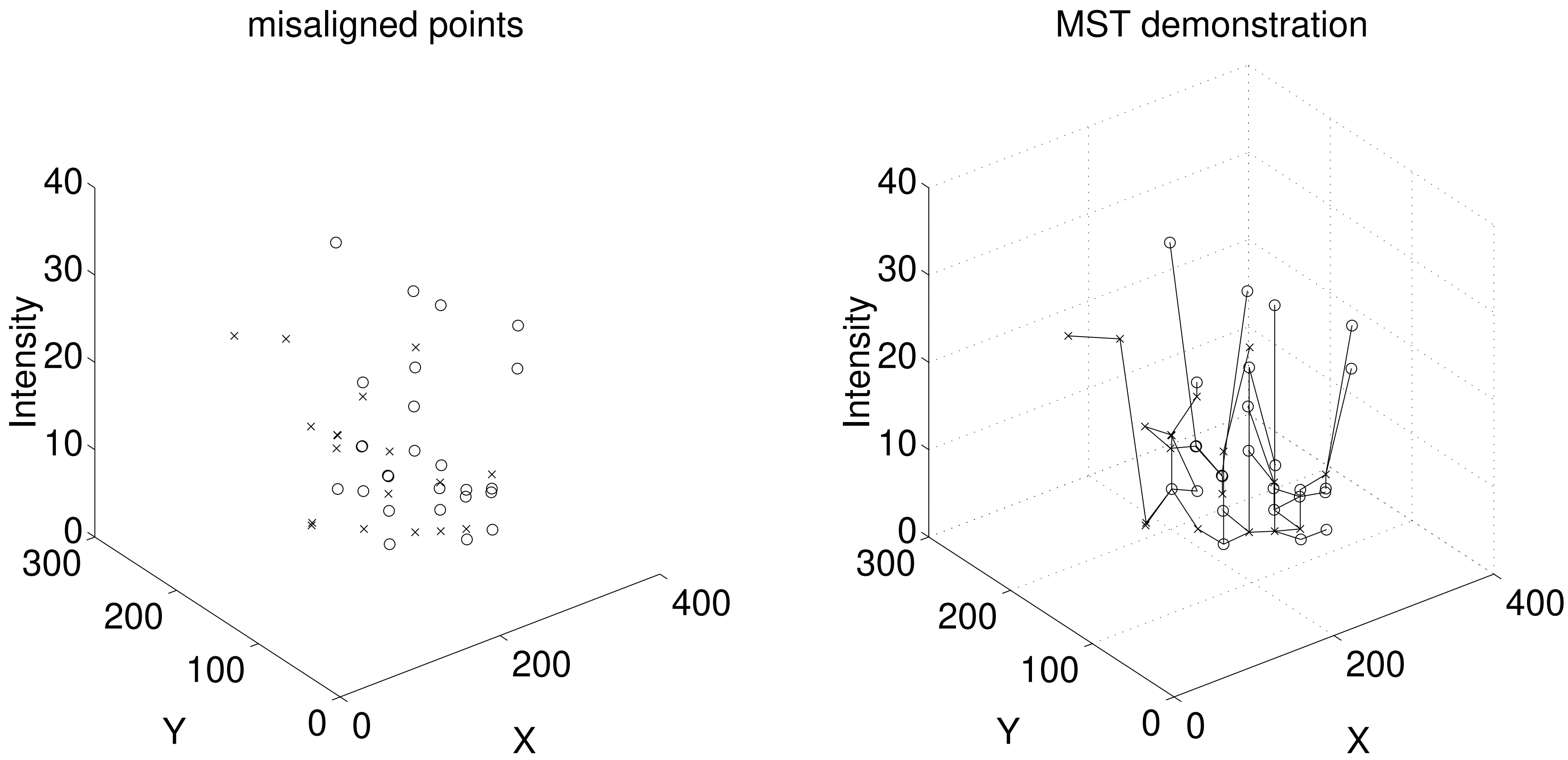
(a) (b)

Figure 26: Feature vectors via vector quantization



(a) (b)

Figure 27: Feature vectors via stratified sampling with centroid refinements



(a)

(b)

Figure 28: MST demonstration for misaligned images

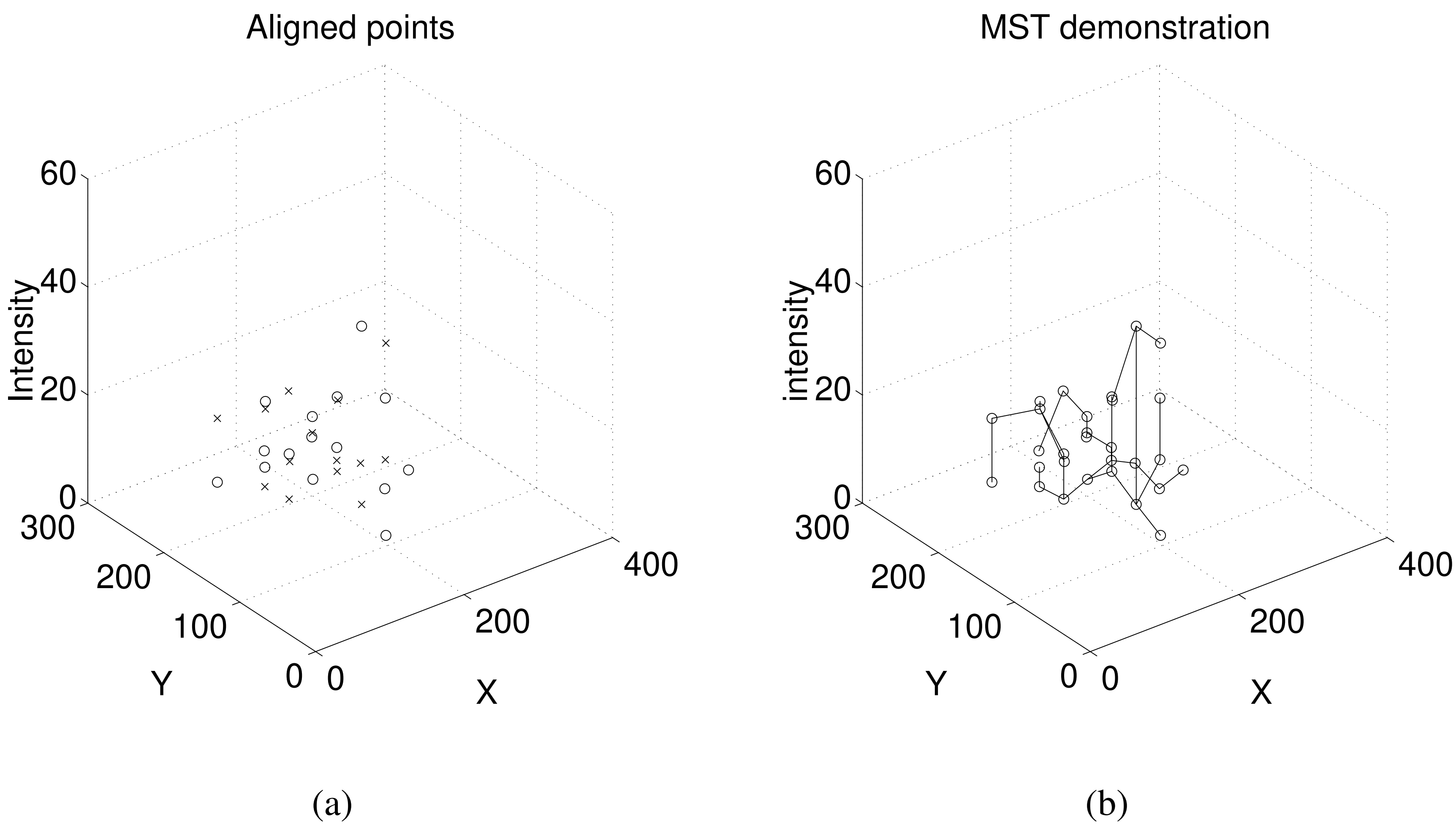


Figure 29: MST demonstration for aligned images

Conclusions

1. α-divergence is well suited to problems in indexing
2. Optimal decision theoretic indexing is possible but very high complexity
3. Suboptimal parametric indexing strategies may be improved by maximal invariant projections of image data
4. Non-parametric estimation of α-divergence is possible without density estimation
5. Non-parametric estimation of Jensen's difference is low complexity alternative to α-divergence estimation
6. Quantitative comparisons between graph matching and density estimation methods for estimating divergence are needed.