

# Applications of Entropic Spanning Graphs

*Alfred O. Hero, III, Bing Ma,  
Olivier J.J. Michel,  
and John Gorman*

This article presents applications of entropic spanning graphs to imaging and feature clustering applications. Entropic spanning graphs span a set of feature vectors in such a way that the normalized spanning length of the graph converges to the entropy of the feature distribution as the number of random feature vectors increases. This property makes these graphs naturally suited to applications where entropy and information divergence are used as discriminants: texture classification, feature clustering, image indexing, and image registration. Among other areas, these problems arise in geographical information systems, digital libraries, medical information processing, video indexing, multisensor fusion, and content-based retrieval.

## Introduction

Let  $X$  be an image, and let independent identically distributed (i.i.d.)  $d$ -dimensional feature vectors  $Z_1, \dots, Z_n$  be extracted from this image. Examples of such a feature vector are the position and orientation of a randomly chosen edge, a vector of samples in a textured region, or the output vector of a spatial prediction filter. Such features can be used for registering two images to each other, texture classification and segmentation, or content-based image retrieval. The basic objective of these applications can be reduced to assessing characteristics of the distribution of the feature vectors. For example, the mutual information method of image registration [53] searches through a number of coordinate transformations to find the one that minimizes the entropy of the joint feature distribution of the two images. Similarly,



many image retrieval algorithms search through a database of images to find the homologous image whose feature distribution is closest to that of the query image where closeness is measured in terms of minimum information divergence [50], [47], [11]. This article discusses minimal graph methods for estimating entropy and divergence measures associated with a set of feature vectors. Specifically, we focus on a class of graphs which span the set of feature vectors and as a byproduct produces a consistent estimator of feature entropy and divergence. We call such graphs entropic spanning graphs.

Here the relevant notion of entropy is the  $\alpha$ -entropy of the feature probability density  $f$ , also known as Rényi entropy, which for probability densities is defined as [45]

$$H_\alpha(f) = \frac{1}{1-\alpha} \ln \int_{\mathcal{Z}} f^\alpha(z) dz, \quad (1)$$

for  $\alpha \in (0,1)$ . The  $\alpha$ -entropy converges to the Shannon entropy  $-\int f(z) \ln f(z) dz$  as  $\alpha \rightarrow 1$ . A related quantity is the  $\alpha$ -divergence between two feature densities  $f_1$  and  $f_0$  of order  $\alpha \in (0,1)$  [45], [10], [3]

$$D_\alpha(f_1 \| f_0) = \frac{1}{\alpha-1} \ln \int f_1^\alpha(z) f_0^{1-\alpha}(z) dz. \quad (2)$$

$D_\alpha(f_1 \| f_0)$  is a measure of similarity or closeness of  $f_1$  and  $f_0$  in the sense that  $D_\alpha(f_1 \| f_0) \geq 0$  with equality iff  $f_1 = f_0$  almost everywhere (a.e.). When  $\alpha \rightarrow 1$  the  $\alpha$ -divergence converges to the Kullback-Leibler divergence  $KL(f_1 \| f_0) = \int f_0(z) \ln(f_0(z)/f_1(z)) dz$ . On the other hand,  $D_{1/2}(f_1 \| f_0)$  is the Hellinger affinity between  $f_1$  and  $f_0$  [6]. The Hellinger affinity is related to the Hellinger distance which is commonly used to measure differences between two probability densities [26], [31].

Nonparametric estimation of Shannon entropy has been of interest to many in nonparametric statistics, pattern recognition, model identification, image registration, and other areas [15], [28], [1], [52], [5], [53], [12]. The estimation of  $\alpha$ -entropy arises as a step towards Shannon entropy estimation, e.g., Mokkadem [37] constructed a nonparametric estimate of the Shannon entropy from a convergent sequence of  $\alpha$ -entropy estimates. However, as we will see, estimation of the  $\alpha$ -entropy is of interest in its own right. The problem arises in vector quantization where Rényi entropy is related to asymptotic quantizer distortion via the Panter-Dite factor and Bennett's integral [14], [39]. The  $\alpha$ -entropy parameterizes the Chernoff exponent governing the minimum probability of error in binary detection problems [27], [9]. It also has been used for image registration from multiple modalities via the  $\alpha$ -Jensen difference [33], [32], [16]. The most natural entropy estimation method is to substitute a nonparametric density estimator  $\hat{f}$  into the expression for entropy. This method has been widely applied to estimation of the Shannon entropy and is called "plug-in" estimation in [5]. Other methods of Shannon entropy estimation discussed in [5] include

sample spacing estimators, restricted to  $d=1$ , and estimates based on nearest neighbor distances.

## Divergence, Entropy, and Indexing

Let  $X_0$  be a reference image, called the query, and consider a database  $X_i$ ,  $i=1, \dots, K$  of images to be indexed relative to the query. Let  $Z_{in} = \{Z_{i1}, \dots, Z_{in}\}$  be  $n$  feature vectors of dimension  $d$  extracted from  $X_i$ . We assume that image  $X_i$ 's feature vectors are i.i.d. with  $Z_{i1}$  following probability density  $f_i(z)$ . Throughout we will also assume that densities are supported on the unit cube  $[0,1]^d$  in  $d$ -dimensions. Under this statistical framework the similarity between images  $X_0, X_i$  is reduced to similarity between feature densities  $f_0(z), f_i(z)$ .

### Divergence Index

The ordered sequence of increasing  $\alpha$ -divergence measures  $D_\alpha(f_i \| f_0) \leq \dots \leq D_\alpha(f_k \| f_0)$ , induces an indexing, which we call the "true divergence-indexing," of the images

$$X_i \prec X_j \Leftrightarrow D_\alpha(f_i \| f_0) < D_\alpha(f_j \| f_0).$$

Special cases of the indexing problem are the following.

▲ 1) *Content-based retrieval* [51], [47], [11]: the query is an image and the database consists of images which may "contain" the object in the sense that the object may only be found as a scaled, rotated or ortho-projected version of the query in the database. An invariant feature set is very important for this application.

▲ 2) *Image registration* [53], [41], [40], [38]: the database consists of  $K$  copies of  $Z_0$  which are rotated, translated, and possibly locally deformed. The indexing finds the pose/orientation in the database closest to that of the query. An invariant feature set is not desirable in this application. When the feature vector  $Z_i$  is defined as the set of pixel pair gray levels associated with each pair of images  $X_i, X_0$  and the mutual information criterion is applied to the pixel pair histogram one obtains an analog to the method of Viola and Wells [53]. The mutual information (MI) criterion is equivalent to the KL divergence between the joint distribution of the pixel-pair gray levels and the product of the marginal feature distributions.

▲ 3) *Target detection* [29], [42], [49]: the query is the distribution of the observations and the database is partitioned into of a family of densities  $f_i = f(Z|\theta_i)$  part of which corresponds to the "target-absent" hypothesis and the rest to "target-present." Target detection is declared if the closest density in the database is in the latter set.

As an illustrative example consider the case where  $f_0$  and  $f_1$  are multivariate Gaussian densities. The KL divergence for such a Gaussian feature model was adopted in [48] and [47]. Let  $f_0(x) = f(x; \mu_0, \Lambda_0)$  and  $f_1(x) = f(x; \mu_1, \Lambda_1)$  be  $d$ -dimensional Gaussian densities with mean vectors  $\mu_0, \mu_1$  and nonsingular covariance matrices  $\Lambda_0, \Lambda_1$ . For this model the unnor-

malized  $\alpha$ -divergence  $D_\alpha^\mu(f_1 \| f_0) = (1-\alpha)D_\alpha(f_1 \| f_0)$  of order  $\alpha$  is given by [19]

$$D_\alpha^\mu(f_1 \| f_0) = \underbrace{-\frac{1}{2} \ln \frac{|\Lambda_0|^\alpha |\Lambda_1|^{1-\alpha}}{|\alpha\Lambda_0 + (1-\alpha)\Lambda_1|}}_{\text{Term A}} + \underbrace{\frac{\alpha(1-\alpha)}{2} \Delta\mu^T (\alpha\Lambda_0 + (1-\alpha)\Lambda_1)^{-1} \Delta\mu}_{\text{Term B}} \quad (3)$$

where  $\Delta\mu = \mu_1 - \mu_0$  and  $|A|$  denotes the determinant of square matrix  $A$ . The divergence consists of two terms  $A$  and  $B$ .  $A$  is equal to zero when  $\Lambda_0 = \Lambda_1$ , and  $B$  is equal to zero when  $\mu_0 = \mu_1$ . Term  $A$  is the log of the ratio of the determinants of the geometric mean and the arithmetic means of  $\Lambda_1$  and  $\Lambda_0$  with mean weights  $\alpha$  and  $1-\alpha$ . Term  $B$  is the quadratic difference of mean vectors normalized by the arithmetic mean of  $\Lambda_1$  and  $\Lambda_0$  with mean weights  $\alpha$  and  $1-\alpha$ .

### Choice of $\alpha$ -Parameter

The  $\alpha$ -divergence is directly related to the exponential rate of decay of the Bayes-optimal binary hypothesis test between two densities  $f_0$  and  $f_1$  [19]. Specifically, given an i.i.d. sample  $Z_n = \{Z_1, \dots, Z_n\}$ , the Chernoff bound asserts that the probability of error  $P_e(n)$  of the optimal Bayes test of  $H_0: Z_i$  has density  $f_0(z)$  versus  $H_1: Z_i$  has density  $f_1(z)$ , then

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln P_e(n) = - \sup_{\alpha \in [0,1]} (1-\alpha)D_\alpha(f_1 \| f_0). \quad (4)$$

The quantity on the right in (4) is called the Chernoff exponent which is the asymptotically optimal rate of exponential decay of the error probability for testing  $H_0$  versus  $H_1$ . For indexing applications this relation suggests that the maximizing  $\alpha$  in (4) is an optimal value, and using it makes the  $\alpha$ -divergence indexing measure the most relevant to feature classification. It can be shown that the maximizing value approaches  $\alpha = 1/2$  when  $f_1$  is close to  $f_0$  [19].

### Entropy Index

An alternative index function is based on the so-called Jensen entropy difference. This index function was independently proposed by Ma [32] and He et al. [16] for image registration problems. It was proposed earlier by Michel et al. in [35] for classifying time frequency distribution images. Let  $f_0$  and  $f_1$  be two densities and  $\beta \in [0,1]$  be a mixture parameter. The  $\alpha$ -Jensen difference is the difference between the  $\alpha$ -entropies of the mixture  $f = \beta f_0 + (1-\beta)f_1$  and the mixture of the  $\alpha$ -entropies of  $f_0$  and  $f_1$  [3]:

$$\Delta H_\alpha(\beta, f_0, f_1) = H_\alpha(\beta f_0 + (1-\beta)f_1) - [\beta H_\alpha(f_0) + (1-\beta)H_\alpha(f_1)], \quad (5)$$

## Entropic spanning graphs span a set of feature vectors such that the normalized spanning length of the graph converges to the entropy of the feature

where  $\alpha \in (0,1)$ . As the  $\alpha$ -entropy  $H_\alpha(f)$  is strictly concave in  $f$  Jensen's inequality asserts that  $\Delta H_\alpha(\beta, f_0, f_1) \geq 0$  with equality iff  $f_0 = f_1$  (a.e.).

The  $\alpha$ -Jensen difference can be motivated as an index function as follows. Assume that two sets of labeled feature vectors  $Z_0 = \{Z_{0i}\}_{i=1, \dots, n_0}$  and  $Z_1 = \{Z_{1i}\}_{i=1, \dots, n_1}$  are extracted from images  $X_0$  and  $X_1$ , respectively, and assume that each of these sets consists of independent realizations from densities  $f_0$  and  $f_1$ . Define the union  $Z = Z_0 \cup Z_1$  containing  $n = n_0 + n_1$  unlabeled feature vectors. Any consistent entropy estimator constructed on the unlabeled  $Z_i$ s will converge to  $H_\alpha(\beta f_0 + (1-\beta)f_1)$  as  $n \rightarrow \infty$  where  $\beta = \lim_{n \rightarrow \infty} n_0/n$ .

For some indexing problems the marginal entropies  $\{H_\alpha(f_i)\}_{i=1}^K$  over the database are all identical so that the indexing function  $\{H_\alpha(\beta f_0 + (1-\beta)f_i)\}_{i=1}^K$  is equivalent to  $\{\Delta H_\alpha(\beta, f_0, f_i)\}_{i=1}^K$ .

### Comparisons of $\alpha$ -Jensen Difference and $\alpha$ -Divergence

There are several interesting properties of  $D_\alpha(f_1 \| f_0)$  and  $\Delta H_\alpha(\beta, f_0, f_1)$  which are discussed in [19].

▲ For  $f_1$  close to  $f_0$  the discrimination capability of the  $\alpha$ -divergence  $D_\alpha(f_1 \| f_0)$  is locally independent of  $\alpha$  while that of the  $\alpha$ -Jensen difference  $\Delta H_\alpha(\beta, f_0, f_1)$  depends on  $\alpha$ .

▲ When  $\alpha$  approaches zero, tail differences between the two densities  $f_0$  and  $f_1$  become most influential.

▲ When  $\alpha$  approaches one, central differences between the two densities become highly pronounced in  $\Delta H_\alpha(\beta, f_0, f_1)$ . Therefore, if the feature densities differ in regions where there is much mass one should choose  $\alpha$  close to one to ensure locally optimum discrimination.

▲  $\Delta H_\alpha(\beta, f_0, f_1)$  has maximal discriminative capability when  $\beta = 1/2$ , i.e., when the two images yield the same number of feature vectors.

### Entropic Spanning Graphs

The aforementioned ideal indexing scheme is of course unimplementable since one never knows the underlying feature densities exactly. Implementation thus requires estimation of the entropy or divergence. Most current nonparametric entropy and divergence estimation techniques are based on estimation of the density function followed by substitution of these estimates into the entropy or divergence functionals (1) and (2). The reader is re-



ferred to [5] for a comprehensive overview of previous work in nonparametric estimation of Shannon entropy. The main difficulties of nonparametric plug-in methods are due to the infinite dimension of the spaces in which the unconstrained densities lie; specifically, density estimator performance is poor without stringent smoothness conditions; no unbiased density estimators generally exist; density estimators have high variance and are sensitive to outliers; and the high dimensional integration required to evaluate the entropy might be difficult.

The problems with plug-in methods can be summarized by this basic observation: on the one hand parameterizing the scalar entropy functional with an infinite dimensional density function is a costly over-parameterization, while on the other hand artificially

enforcing lower dimensional density parameterizations can produce significant bias in the estimates. This observation has motivated us to develop direct methods which accurately estimate the entropy without the need for performing artificial low dimensional parameterizations or nonparametric density estimation [21]-[23]. These methods are based on constructing minimal graphs spanning the feature vectors in the feature space. The overall length of these minimal graphs can be used to construct a strongly consistent estimator of entropy for densities without singular (dirac delta) components. In particular, let  $\mathcal{Z}_n = \{Z_1, \dots, Z_n\}$  and define

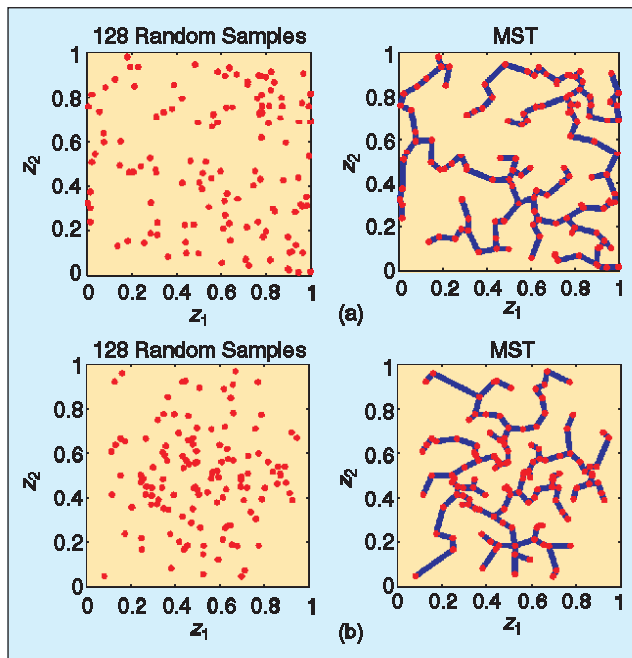
$$L(\mathcal{Z}_n) = \min_{e \in \mathcal{T}} \sum_e |e|^\gamma, \quad (6)$$

the overall length of a graph spanning  $n$  i.i.d. vectors  $Z_i$  in  $\mathbf{R}^d$  each with density  $f$ . Here the power weighting  $\gamma \in (0, d)$  is real,  $e$  are edges in a graph connecting pairs of  $Z_i$ 's,  $|e|$  denotes Euclidean ( $l_2$ ) norm of the edge, and the minimization is over some suitable subsets  $\mathcal{T}$ , e.g., spanning trees, of the  $\binom{n}{2}$  edges of the complete graph. Examples include the minimal spanning tree (MST), Steiner tree (ST), minimal matching bipartite graph, and traveling salesman problem (TSP). The asymptotic behavior of  $L(\mathcal{Z}_n)$  over random points  $\mathcal{Z}_n$  has been studied for over half a decade [4], [46].

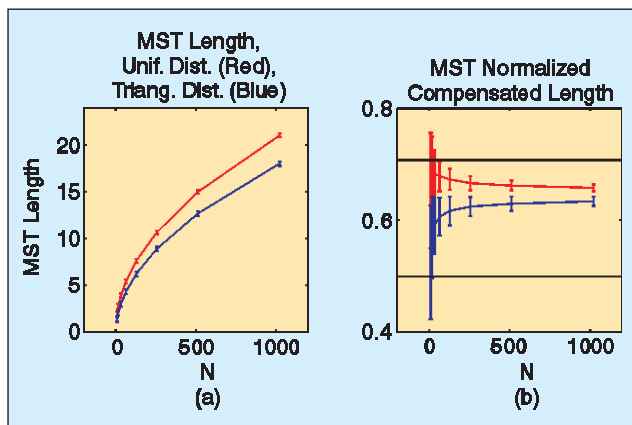
In Fig. 1 the MST is illustrated for two sets of randomly generated points in the plane, one uniformly distributed (a) and the other distributed with a more concentrated separable triangular density. The MST is defined as the minimum length graph spanning the  $n$  points. The MST length  $L_n = L(\mathcal{Z}_n)$  is plotted as a function of  $n$  in Fig. 2 for the case of uniformly and nonuniformly distributed points and for  $\gamma=1$ . It is intuitive that the length of the MST spanning the more concentrated nonuniform set of points increases at a slower rate than does the MST spanning the uniformly distributed points. This fact motivated the application of the MST as a way to test for randomness of a set of points [24]. What is more surprising is that normalizing by  $\sqrt{n}$  and taking the logarithm of these length functions produces sequences that converge (within a constant factor) to the  $\alpha$ -entropies with  $\alpha=1/2$ , as illustrated in Fig. 2(b). Furthermore, by changing the value of  $\gamma$  in (6) one can change the convergent limit to the  $\alpha$ -entropy for  $\alpha = (d-\gamma)/d$ ,  $\gamma \in (0, d)$ . Graphs for which the normalized log-length converges (a.s.) within a constant to an  $\alpha$ -entropy for some  $\alpha \in (0, 1)$  will be called entropic spanning graphs. In Fig. 2 the upper and lower horizontal lines correspond to known bounds [54] on  $\beta_{L, \gamma}$ .

We showed [23] that when a graph is "quasi-additive" [54] in  $d$ -dimensional feature space,  $d \geq 2$ , the graph is an entropic spanning graph. Specifically

$$\hat{H}_\alpha(\mathcal{Z}_n) = \frac{1}{1-\alpha} [\ln L(\mathcal{Z}_n) / n^\alpha - \ln \beta_{L, \gamma}] \quad (7)$$



▲ 1. (a) A random set of  $n=128$  uniformly distributed points in  $[0,1]^2$  and the MST spanning these points. (b) A random set of  $n=128$  points with separable triangular density and the MST spanning these points.



▲ 2. Length functions (a)  $L_n$  of MST and (b) MST divided by  $\sqrt{n}$  as function of  $n$  for the uniform and separable triangular distributed points in Fig. 1.

is an asymptotically unbiased and almost surely consistent estimator of the  $\alpha$ -entropy of  $f$  where  $\alpha = (d - \gamma) / d$  and  $\beta_{L,\gamma}$  is a constant bias correction depending on the graph minimization criterion, e.g., MST, ST, or TSP, but independent of  $f$ . The estimator  $\hat{H}_\alpha(Z_n)$  is also consistent when the power exponent function  $|e|^\gamma$  in (6) is replaced by a positive function  $g(|e|)$  which locally behaves as  $|e|^\gamma$  as  $|e| \rightarrow 0$  [46]. The fact that (7) holds for any quasi-additive graph construction opens many different possibilities for consistent graph-based entropy estimation algorithms. However, among the currently known quasi-additive algorithms the MST is the fastest (with polynomial run time) and as such we have adopted it for all of the entropy estimation applications discussed here.

As contrasted with density plug-in techniques, graph-based entropy estimators enjoy the following properties: they can have faster asymptotic convergence rates, especially for nonsmooth densities and for high dimensional feature spaces [20]; they completely bypass the complication of choosing and fine tuning parameters such as histogram bin size, density kernel width, complexity, and adaptation speed; the  $\alpha$  parameter in the  $\alpha$ -entropy function is varied by varying the interpoint distance measure used to compute the weight of the minimal graph. On the other hand, the need for combinatorial optimization is a bottleneck for large numbers of feature samples. This has motivated the development of greedy minimal graph approximations that preserve advantages such as robustness against outliers as discussed below.

### Extension to Divergence Estimation

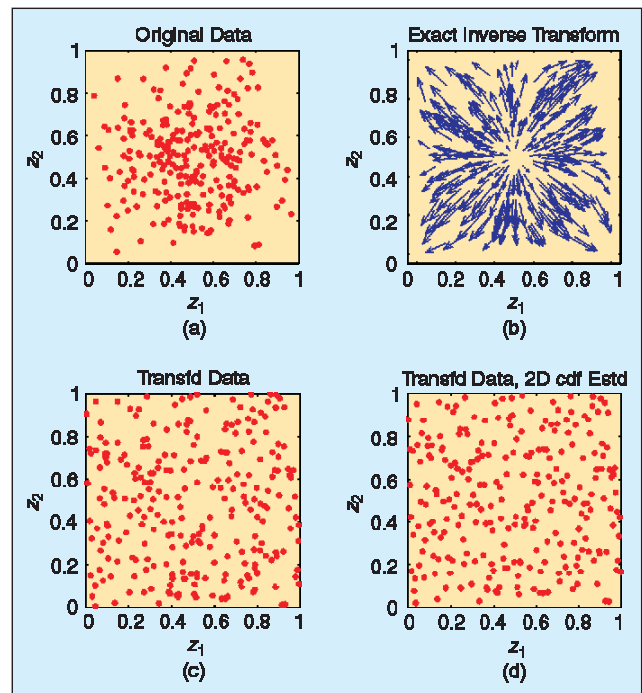
We showed in [22] how an entropic spanning graph estimation procedure can be extended to information divergence estimation by a method of measure transformation. Assume that  $f_0$  dominates  $f_1$  (a density  $h$  dominates density  $g$  if whenever  $h(z) = 0$  then  $g(z) = 0$ ) and rewrite the divergence in (2) as  $\int (f_1(z) / f_0(z))^\alpha f_0(z) dz$ . The basic idea is to apply a transformation of coordinates to the feature vectors which uniformizes the reference density  $f_0$ . We illustrate the idea behind this technique for scalar  $z$ . Assume that  $Z_n$  are  $n$  i.i.d. data points generated from density  $f_1(z)$ . Apply the coordinate transformation  $y = g(z)$  to each point in  $Z_n$  where  $g$  is an invertible function such that  $dy = f_0(z) dz$ . This produces a new set of points  $\mathcal{Y}_n$  in the transformed coordinates. By standard Jacobian formulas for change of variable of integration, the divergence integral becomes  $\int (h(y))^\alpha dy$ , where  $h(y) = f_1(z) / [dy / dz]$  is the induced density of  $\mathcal{Y}_n$ . Thus the length  $L(\mathcal{Y}_n)$  of the MST constructed on the transformed random variables  $\mathcal{Y}_n$  can be used in place of the length  $L(Z_n)$  in (7) to give a consistent estimate of the divergence (2) of  $f_1$  relative to a known reference  $f_0$ :

$$\hat{D}_\alpha(f_1 \| f_0) = \frac{1}{1 - \alpha} [\ln L(\mathcal{Y}_n) / n^\alpha - \ln \beta_{L,\gamma}] \quad (8)$$

An example of this procedure is shown in Fig. 3 for a two-dimensional (2-D) separable triangular reference density  $f_0$  over  $[0,1]^2$  which in this case equals the actual marginal density  $f_1$  of the observed i.i.d. points  $Z_n$ . Thus for this example the true divergence is zero. We mean  $f_0(z) = (2 - 4|z_1 - 1/2|)(2 - 4|z_2 - 1/2|)$ ,  $z = (z_1, z_2)$  when we say triangular density. A random sample of  $n = 100$  points was generated from  $f_1$ . The uniformizing transformation in this case is separable too, with each component transformation equal to the marginal cumulative density function  $F(z) = \int_0^z (2 - 4|x - 1/2|) dx$  of the one-dimensional (1-D) triangular density. We investigated both exact uniformizing transformations and estimated transformations using estimates of the one dimensional component density functions. The transformed sample is essentially uniform both for the exact and the estimated transformations. Therefore, as  $n \rightarrow \infty$ , it is expected that  $L(\mathcal{Y}_n) / n^\alpha$  will converge to  $\beta_{L,\gamma}$  and the estimated divergence (8) will converge to zero as desired.

### Robustifying Entropic Spanning Graphs

In many practical problems occasional spurious feature vectors may appear due to noise, false alarms, or small unimportant shifts and deformations during the image formation process. In such situations we are interested in robust entropy or divergence estimators which are resistant to these spurious outliers. This problem is related to



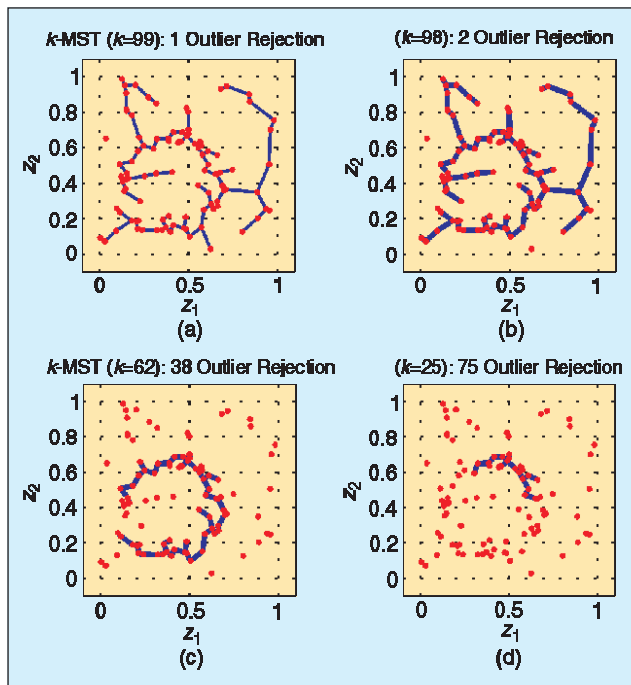
▲ 3. (a) A sample from a separable triangular pdf over the unit square. (b) A vector field indicating the action of the exact separable inverse transformation of coordinates on each sample point in top right. (c) Same sample points as in (a) after applying transformation indicated in (b). (d) Same as (c) except that estimated transformation of coordinates was implemented using  $k$ -nearest-neighbor density estimators for each of the marginals.

## The $\alpha$ -divergence is directly related to the exponential rate of decay of the Bayes-optimal binary hypothesis test between two densities $f_0$ and $f_1$ .

robust clustering for which it is common to adopt a finite mixture model to capture the incidence of points arising from different distributions [34]. For our case the appropriate mixture model is the so-called epsilon-contaminated model [25]

$$f(z) = (1 - \varepsilon)g(z) + \varepsilon h(z), \quad (9)$$

where  $\varepsilon \in [0,1]$ ,  $h$  is an unwanted outlier density, and  $g$  is the underlying density of interest. When  $n$  points are realized from the model (9) an average of  $k = (1 - \varepsilon)n$  of these points follow the distribution  $g$  while the remaining  $n - k = \varepsilon n$  are outliers generated from  $h$ . Therefore  $\varepsilon$  corresponds to the proportion  $(n - k) / n$  of outliers one might expect in a large sample from density  $f$ . It is assumed that  $\varepsilon$  is small but unknown. The target density  $g$  is also assumed unknown while the outlier density  $h$  is known and has the same support  $[0,1]^2$  as that of  $g$ .



▲ 4.  $k$ -MST for 100 points realized from an  $\varepsilon$ -mixture of 2-D annulus density  $g$  and a uniform outlier density  $h$  ( $\varepsilon = 1/2$ ). Points arising from the annulus density tend to cluster in a ring while uniform points are more widely dispersed over the unit square. Initially, as the number of points included in the  $k$ -MST graph decreases, a greater and greater number of outlier points are rejected. When  $k = 62$  (38 rejected points) the  $k$ -MST graph has successfully clustered the annular points recovering the ring Gestalt.

Under the model (9) an outlier resistant entropic spanning graph was proposed in [23] which identifies and eliminates the outlier points. First, using the measure transformation method discussed in the previous section, we transform the coordinates of the sample  $\mathcal{Z}_n$  such that  $h(z)$  is converted to a uniform distribution over  $[0,1]^d$ . This transformed sample is denoted  $\mathcal{Y}_n$  and follows a standard mixture model (9) with uniform contaminating density  $h$ . Second, iterating over  $k = n, n-1, \dots$ , we construct entropic spanning graphs over each of the  $\binom{n}{k}$   $k$ -point subsets  $\mathcal{Y}_{n,k}$  of  $\mathcal{Y}_n$ . For each value  $k$ , there will be a graph of minimum length among these  $\binom{n}{k}$  graphs. This minimal graph spans a set of points  $\mathcal{Y}_{n,k}^*$  which are “maximally clustered” among all  $k$ -point subsets. The  $n - k$  points eliminated from the span of this minimal graph are thus identified as outliers.

We illustrate this procedure in Fig. 4 for 100 realizations from a mixture density with an annular component  $g$  and a uniform component  $h$ . Here  $\varepsilon = 0.5$  corresponding to 50 realizations from each of the distributions. The annular density  $g$  has the form

$$g(z) = c \exp\left(-\frac{1}{2} 225 (\|z - [0.4, 0.4]\| - 0.25)^2\right)$$

where  $c$  is a normalizing constant and  $\|z\|^2 = z_1^2 + z_2^2$  is the magnitude squared of  $z = (z_1, z_2)$ . The constant contours of this density are circles for which the maximum contour is a circle of radius 0.25 and center  $[0.4, 0.4]$  and the other contours specify an annulus. Our objective is to estimate the  $\alpha$ -entropy of the annular density  $g$  from the 100 realizations from  $f$ . For this purpose we adopted the  $k$ -point MST ( $k$ -MST) as our entropic spanning graph algorithm. In terms of estimating this entropy, the standard MST (spanning all 100 points) is extremely sensitive to the 50 outliers which dominate the MST length function. Hence the  $k$ -MST is implemented to isolate the points from  $g$  from the outliers. The four panels in Fig. 4 illustrate the  $k$ -MST for several values of  $k$ . It is evident from the figure that as the number of points eliminated by the  $k$ -MST increases from 1 to 2 to 38 the  $k$ -MST rejects an increasing number of outliers from the contaminating density. Indeed for the case of  $k = 62$  (38 outliers rejected) the  $k$ -MST appears to have almost completely recovered the MST for the annular distribution  $f_1$ . However, as the number of rejected points increases beyond 38 to 75 the  $k$ -MST begins eliminating points which come from the desired annular distribution. The key to a practical  $k$ -MST robustification algorithm will be accurate detection of the correct number of points to reject.

As the number  $k$  of points retained increases, the sequence of MST lengths  $L(\mathcal{Y}_{n,n}^*), L(\mathcal{Y}_{n,n-1}^*), \dots, L(\mathcal{Y}_{n,k}^*)$  is monotone increasing and evolves a curve over  $k$ . As  $k$  approaches  $n$  the curve can be expected to increase more rapidly as more of the isolated “outlier” points are successively included in the MST. As these points will tend to come from the uniform distribution the average rate of increase for large  $k$  is constant. We would like to select  $k$  in



the  $k$ -MST so to eliminate as many of the uniform outlier points while eliminating as few of the other points from density  $\mathcal{g}$  as possible. If the parameter  $\varepsilon$  were known a value  $k \approx \varepsilon n$  could be chosen a priori. Otherwise, a  $k$  stopping rule can be implemented which is based on detecting the knee in the curve  $L(\mathcal{Y}_{n,k}^*)$ . Fig. 5 shows this curve for the example shown in Fig. 4. The knee detection algorithm is motivated as follows. As  $k$  decreases from  $n$  to one, more and more points are pruned from the  $k$ -MST. When the number of points retained falls below a critical threshold, points from the more concentrated  $\mathcal{g}$  distribution start to be eliminated and the slope of the curve abruptly decreases.

Once the knee  $k$  has been identified the length  $L(\mathcal{Y}_{n,k_0}^*)$  can be used for robust estimation of the  $\alpha$ -entropy of order  $\alpha$ , where as usual  $\alpha = (d - \gamma) / d \in (0, 1)$  is specified by the dimension  $d \geq 2$  and the weight exponent  $\gamma \in (0, d)$ . In [23] we established a.s. convergence of this estimate when a greedy approximation to  $k$ -point minimal entropic graph is implemented. The Huber-Hampel influence function of this robust procedure was also investigated in [23].

### Computational Issues

The computational complexity of minimal graph algorithms depends on the implementation but is generally superlinear in the number  $n$  of vertices [8]. Minimal spanning trees and  $k$ -means algorithms are of lowest complexity (complexity  $O(n^2 \log n)$  or less) among the many entropic spanning graph algorithms one might consider. We have implemented both sequential single-processor MSTs and parallel multiprocessor MSTs. While our experience has been limited to parallelization over (TCP/IP) networked workstations, we have found that parallel MST implementations, such as that proposed in [7], are stymied by high interprocessor communications overhead. There are principally two sequential implementations of the MST: Kruskal's "growing a forest of trees" algorithm [30] and Prim's "growing a single tree" algorithm [43]. Both algorithms are greedy and successively add a single edge to the graph until all points are spanned without any cycles. Using general-purpose versions of these MST algorithms computation time becomes prohibitive for more than a few thousand points. An accelerated kruskal-type of MST algorithm, only applicable to Euclidean vertices, has been developed by us [17], [18] which can compute a MST approximation for several hundred thousand points in a few seconds (C code running on a 900 MHz PC under Linux).

The  $k$ -MST discussed previously arises in many combinatorial optimization problems; see references in [23] for a partial list. Its computational complexity is exponential, which necessitates implementation of approximate schemes [2], [13], [44]. The greedy approximation used in [23] involves the partitioning heuristic used by [44]: dissect the support of the density  $f$ , assumed to be  $[0, 1]^d$ , into a set of  $m^d$  cells of equal volumes  $1/m^d$ ; rank the cells

in increasing order of numbers of points contained; starting with the highest ranked cell and continuing down the list compute the minimal spanning graphs in each cell until at least  $k$  points are covered. Stitching together these small graphs gives a graph which is an approximation to the  $k$ -minimal graph. The computational advantage of the greedy algorithm comes from its divide-and-conquer multiresolution structure: it only requires solving the difficult nonlinear minimal graph construction on cells containing smaller numbers of points. When  $k = n$  this greedy approximation reduces to a partitioning approximation to the full minimal graph spanning all of the  $n$  points. By selecting the "progressive-resolution parameter"  $m$  as a function  $m(n)$  of  $n$  we obtain an adaptive multiresolution approximation to the  $k$ -MST.

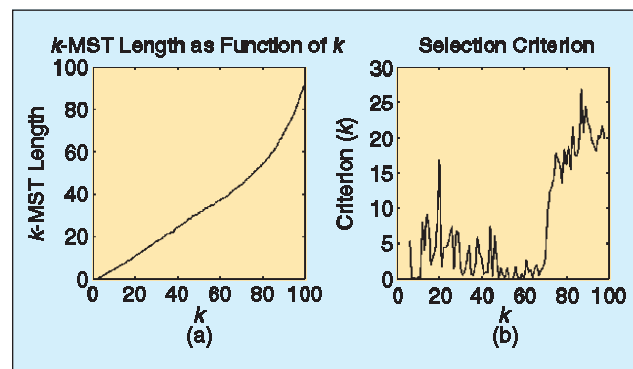
## Applications

We have implemented entropic spanning graph estimators in several application areas including: image registration of ultrasound scans [38], extraction of time-frequency skeletons from the time-frequency plane [36], robust clustering [21], pattern classification [22], and geo-registration [33]. Due to space limitations we only discuss two of these applications here.

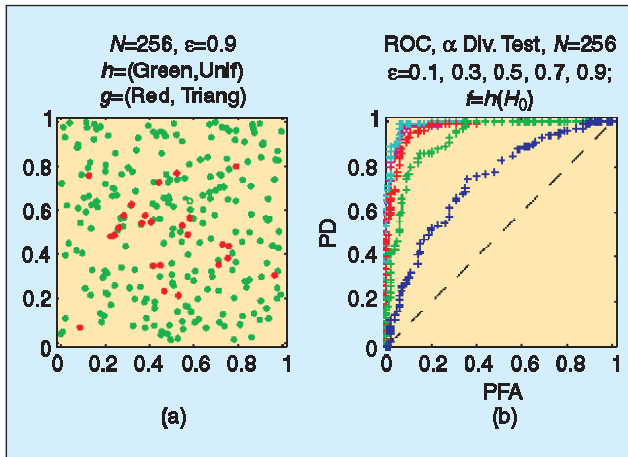
### Robust Clustering and Classification

Here we apply the  $k$ -MST to robustly cluster and classify a triangular versus uniform density. There were 256 samples simulated from a uniform-triangular mixture density  $f = (1 - \varepsilon)\mathcal{g} + \varepsilon b$  where  $\mathcal{g} = 1$  is a uniform density and  $b$  is the separable triangular shaped product density both supported on the unit square. Note that, unlike the previous annular-uniform mixture example, the "outlier" distribution  $b$  has lower entropy than the target distribution  $\mathcal{g}$  which makes the problem of clustering the realizations from  $\mathcal{g}$  more challenging.

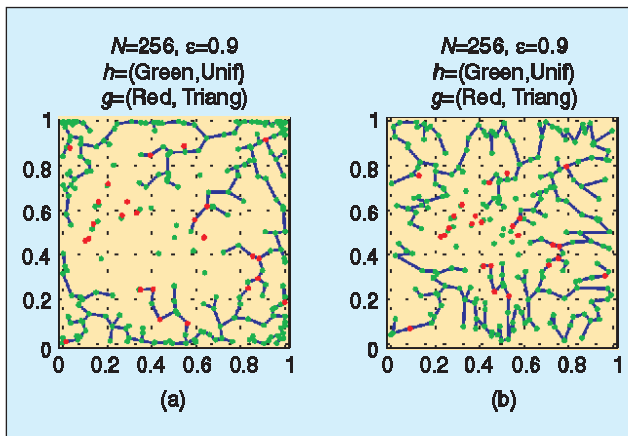
The  $\alpha$ -divergence  $D_\alpha(f, b)$  was estimated by  $\hat{H}_\alpha(\mathcal{Y}_n)$  for  $\alpha = 1/2$  ( $\gamma = 1$ ) using the MST estimator.  $\mathcal{Y}_n$  was obtained by applying the "uniformizing" coordinate transfor-



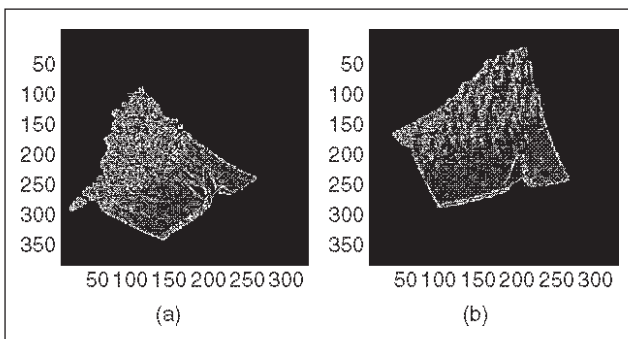
▲ 5. (a)  $k$ -MST curve for 2-D annulus density with addition of uniform "outliers" has a knee in the vicinity of  $k = 68$ . This knee can be detected using residual analysis from a linear regression line fitted to the left-most part of the curve. (b) Error residual of linear regression line.



▲ 6. (a) A scatterplot of a 256-point sample from the uniform-triangular mixture density with  $\epsilon=0.1$ . Green and red points mark those realizations from the uniform and triangular densities, respectively. (b) ROC curves for the  $\alpha$ -divergence test for detecting the uniform-triangular mixture density  $f=(1-\epsilon)g+\epsilon h$  ( $H_1$ ) against the triangular hypothesis  $f=h$  ( $H_0$ ). Curves are increasing in  $\epsilon$  over the range  $\epsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ .



▲ 7. (a) The scatterplot of Fig. 6 after applying the uniformizing coordinate transformation. Green and red points mark the transformed realizations from the uniform and triangular densities, respectively. Superimposed is the  $k$ -MST implemented on the transformed scatterplot  $\text{cal}Y_n$  with  $k = 230$ . (b) The same as (a) except displayed in the original data domain.



▲ 8. Misaligned EO and reference images: (a) image at 290,  $-20$ , 130 rotation and (b) reference image.

mation to  $Z_n$  used earlier. In a first sequence of experiments the estimate  $\hat{H}_\alpha(\mathcal{Y}_n)$  was thresholded to decide between the hypotheses  $H_0:\epsilon=0$  versus  $H_1:\epsilon \neq 0$ . Simulations were performed to generate the receiver operating characteristic (ROC) curves indicated in Fig. 6 for various values of  $\epsilon$ . Note that, as expected, in each case the detection performance improves as the difference, indexed by  $\epsilon$ , between the assumed  $H_0$  and  $H_1$  densities increases.

In a second sequence of experiments we selected two realizations of the uniform-triangular mixture model for the value  $\epsilon = 0.1$ . The  $k$ -MST procedure ( $k=0.9n$ ) was implemented on  $\mathcal{Y}_n$  as a robust algorithm to cluster data points from the uniform density. The cluster of points are defined as those points connected by the  $k$ -MST graph. The  $k$ -MST length can thus be used as a robust estimate  $\hat{H}_\alpha(\mathcal{Y}_{n,k})$  of the uncontaminated divergence  $D_\alpha(g,h)$ . Fig. 7 illustrates the effectiveness of this clustering method: within the cluster defined by the vertices of the  $k$ -MST the proportion of contaminating points from  $h$  has dropped from the original 10% to less than 4%.

### Geo-Registration Application

Multisensor image registration problems can be cast as specific cases of a more general sensor registration problem in which the imaging sensors jointly observe 2-D projections of a common three-dimensional (3-D) object. The challenges presented in multisensor image registration are severalfold. Differences between sensor viewpoints and imaging modality can cause unknown relative geometric distortions and missing pixels between image pairs. Differences in illumination and environmental conditions introduce further complications. Existence of such differences between images to be registered requires that the registration algorithms be robust to noise and other small perturbations in intensity values.

One approach to solving the multisensor image registration problem is to first geo-register the images to a common terrain model and then to refine the registration by working with the geo-registered images. In this geo-registration application, a digital elevation model (DEM) of a terrain patch (terrain height map) plays the role of the image database and the image indexing problem is that of selecting the sensor and environmental parameters (pointing angle, latitude and longitude, sun-angle, etc.) that yield the best match between the reference sensor image and a modeled or rendered version of the DEM. (DEM stores the terrain height information in a three dimensional array where each element of the array consists of the locations ( $x$  and  $y$  coordinates) and the height of the terrain at that location.) Database images are generated from the DEM by rendering a sensor's view of the model at a variety of look angles and possibly under different illumination conditions.

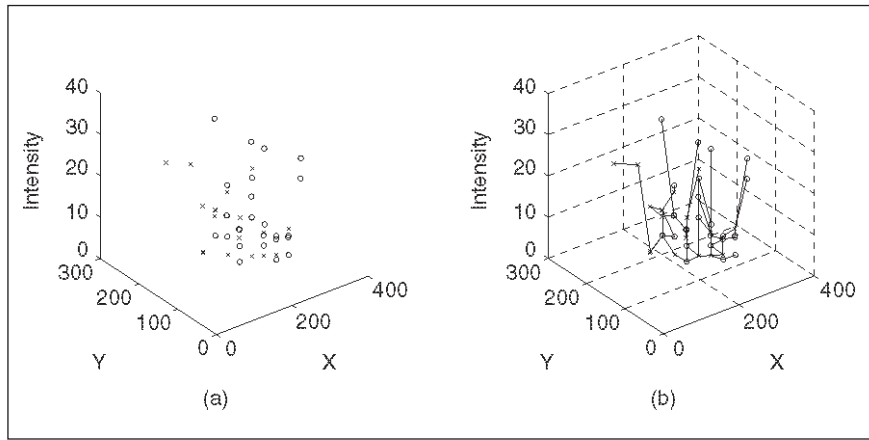
Fig. 8 shows an edge map extracted from an optical view of a terrain map (DEM) at viewing angle (290,  $-20$ , 130) as well as the edge map extracted from a reference electrooptic (EO) image that is to be geo-registered. Clearly, they are misaligned.



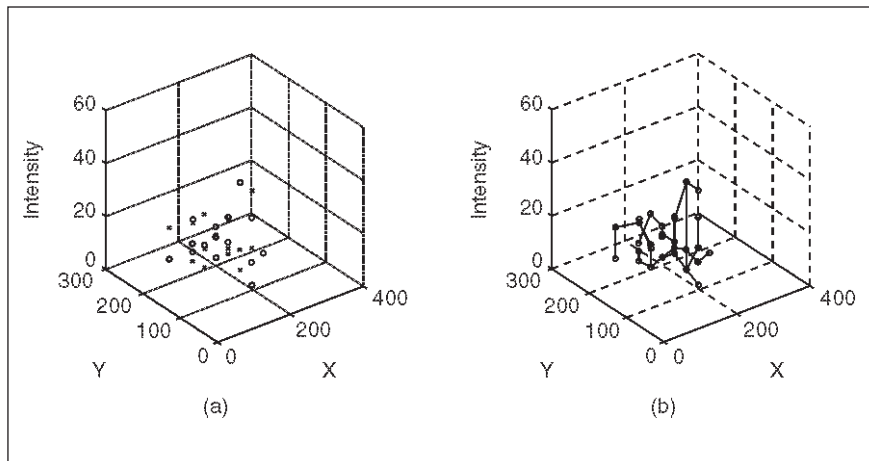
For matching criterion we implemented the  $\alpha$ -Jensen difference applied to grey level features extracted from the reference images and candidate EO images derived from the DEM database. The parameter  $\alpha$  was chosen arbitrarily as 0.5, corresponding to a MST construction minimizing the Euclidean norm in (6) without any power weighting ( $\gamma=1$ ). For illustration purposes we selected a very simple set of features via stratified sampling of the grey levels with centroid refinements. This sampling method produces a set of  $n$  three-dimensional feature vectors  $Z_i = (x_i, y_i, F(x_i, y_i))$  where  $F(x, y)$  is a sample of the gray level at planar position  $x, y$  and where  $n$  is fixed in advance. The points  $\{(x_i, y_i)\}_{i=1}^n$  approximate the centroids of Voronoi cells and  $\{F(x_i, y_i)\}_{i=1}^n$  correspond to the set of  $n$  samples of the image from which we could reconstruct the original image with minimum mean square error. For more details see [32]. When the union of features from reference and target images are rendered as points in three dimensions we obtain a point cloud of features over which the MST can be constructed and the Jensen difference estimated. Since  $n_1 = n_0 = n$  we have used  $\beta=1/2$  in the Jensen difference (5). One issue that we have not addressed here is the validity of the i.i.d. assumption on the feature vector set  $Z_n$  acquired for this example. We believe that this is a good approximation for our choice of spatially distinct features but this question deserves further investigation.

Fig. 9 illustrates the MST-based registration procedure over the union of the reference and candidate image features for misaligned images, while Fig. 5 shows the same for aligned images. From Figs. 9(a) and 10(a) we see that for misaligned images, the representation points “x” and “o” are at larger distances, giving corresponding larger MST weight, than those for aligned images.

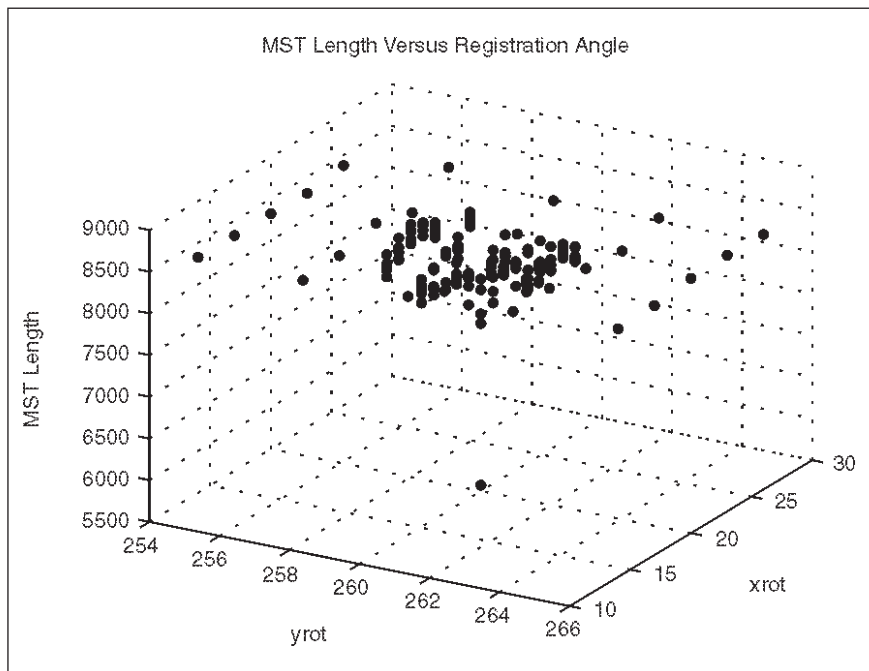
We repeat this MST construction process over the union of reference features and features derived from each of the images in the DEM data-



▲ 9. MST demonstration for misaligned images: (a) misaligned points and (b) MST demonstration.



▲ 10. MST demonstration for aligned images. “x” denotes reference while “o” denotes a candidate image in the DEM database: (a) aligned points and (b) MST demonstration.



▲ 11. Scatter plot of MST length for a selection of relative rotation angles between reference image and target radar image. The MST length surface exhibits a sharp minimum at the correct registration angle.

base. The MST length can then be plotted as a scatterplot as in Fig. 11. The minimum MST length indicates the best matching of the EO image and the reference image, which corresponds to the registered pair.

## Conclusion

In this article we have discussed theory and application of entropic spanning graphs for clustering, imaging, and entropy estimation problems. There are many open problems in this area that must be addressed. The entropic spanning graph is not a consistent estimator of entropy when the underlying density has discrete components, i.e.,  $f$  contains dirac delta functions. While worst case bounds on convergence rates of these estimators are available, a more complete comparison of plug-in versus entropic spanning graph estimators of entropy has yet to be performed. Despite the many open problems, entropic spanning graph methods are very promising due to their simplicity relative to other nonparametric techniques for clustering and feature classification.

## Acknowledgment

This work was supported in part by a NATO Collaborative Linkage Grant and AFOSR MURI Grant F49620-97-0028.

*Alfred O. Hero III* received the B.S. from Boston University in 1980 and the Ph.D. from Princeton University in 1984, both in electrical engineering. Since 1984 he has been a Professor with the University of Michigan, Ann Arbor, in the Departments of Electrical Engineering and Computer Science, Biomedical Engineering, and Statistics. His research interests are in the areas of estimation and detection theory, statistical communications, bioinformatics, signal processing, and image processing. He has served as Associate Editor for the *IEEE Transactions on Information Theory* and the *IEEE Transactions on Signal Processing*. He was Chair of the Statistical Signal and Array Processing (SSAP) Technical Committee, Treasurer of the Conference Board, and Vice President (Finance) of the IEEE Signal Processing Society. He has chaired several conferences, most recently the 2002 Workshop on Genomic Signal Processing and Statistics (GENSIPS). He is a Fellow of the IEEE, a member of Tau Beta Pi, the American Statistical Association, the Society for Industrial and Applied Mathematics, and the U.S. National Commission (Commission C) of the International Union of Radio Science (URSI). He has received the Meritorious Service Award (1998), a Best Paper Award (1998), an IEEE Third Millennium Medal (2000), and a Distinguished Lecturer appointment (2002) from the IEEE Signal Processing Society.

*Bing Ma* received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 1993 and the M.S. and Ph.D. degrees in electrical engineering and

computer science from the University of Michigan, Ann Arbor, in 1997 and 2001, respectively. She is now with M-Vision, Inc., where she is a research scientist. Her research interests include statistical signal and image processing, estimation and detection, computer vision, pattern recognition, and multisensor fusion.

*Olivier J.J. Michel* completed his studies at Ecole Normale Supérieure de Cachan, in applied physics, where he received the “Agrégation de Physique” in 1986. He received a Ph.D. degree from University Paris-XI Orsay in 1991 in signal processing. From 1991 to 2000, he worked at the physics department at Ecole Normale Supérieure de Lyon, France, as an assistant professor. Since 2000, he has been a professor at the University of Nice Sophia Antipolis. His research interest include nonstationary spectral analysis, array processing, nonlinear time series problems, information theory, and dynamical systems.

*John Gorman* received the B.S. (1983) and the M.S. degrees (1985) from Washington University in Saint Louis, both in electrical engineering, and the Ph.D. in electrical engineering from the University of Michigan in Ann Arbor (1991). He is with Science, Engineering and Technology Associates, Arlington, VA. Prior to that he was Director of Informatics and co-founder of Cytoprint, Inc. From 1985 through 2001 he was a staff member and senior scientist at the Environmental Research Institute of Michigan (ERIM) in Ann Arbor. His research interests are in the areas of detection and estimation, signal and image processing, and their application to reconnaissance, surveillance and bioinformatics. He served on the board of the 1995 IEEE International Symposium on Acoustics, Speech and Signal Processing. He is a member of the IEEE, Tau Beta Pi, Eta Kappa Nu, the American Statistical Association, and the Society for Industrial and Applied Mathematics.

## References

- [1] I. Ahmad and P.-E. Lin, “A nonparametric estimation of the entropy for absolutely continuous distributions,” *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 664-668, 1976.
- [2] S. Arora, “Nearly linear time approximation schemes for (Euclidean TSP) and other geometric problems,” in *Proc. IEEE 38th Annu. Symp. Foundations Comput. Sci.*, 1997, pp. 554-563.
- [3] M. Basseville, “Distance measures for signal processing and pattern recognition,” *IEEE Signal Processing Mag.*, vol. 18, pp. 349-369, 1989.
- [4] J. Beardwood, J.H. Halton, and J.M. Hammersley, “The shortest path through many points,” in *Proc. Cambridge Philosophical Soc.*, vol. 55, 1959, pp. 299-327.
- [5] J. Beirlant, E.J. Dudewicz, L. Györfi, and E. van der Meulen, “Nonparametric entropy estimation: An overview,” *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17-39, June 1997.
- [6] L. Birgé and P. Massart, “Estimation of integral functions of a density,” *Ann. Statist.*, vol. 23, pp. 11-29, 1995.
- [7] S. Chung and A. Condon, “Parallel implementation of Boruvka’s minimal spanning tree algorithm,” in *Proc. 10th Int. Parallel Processing Symp.*, 1996.
- [8] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*. Englewood Cliffs, NJ: McGraw-Hill, 1990.

- [9] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1987.
- [10] I. Csiszár, "Information-type measures of divergence of probability distributions and indirect observations," *Studia Sci. Math. Hung.*, vol. 2, pp. 299-318, 1967.
- [11] M.N. Do and M. Vetterli, "Texture similarity measurement using Kullback-Liebler distance on wavelet subbands," in *IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, 2000, pp. 367-370.
- [12] D.L. Donoho, "One-sided inference about functionals of a density," *Ann. Statist.*, vol. 16, pp. 1390-1420, 1988.
- [13] N. Garg and D.S. Hochbaum, "An  $O(\log k)$  approximation for the  $k$  minimum spanning tree problem in the plane," *Algorithmica*, vol. 18, no. 1, pp. 111-121, May 1997.
- [14] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 373-380, 1979.
- [15] P. Hall and S.C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69-88, 1993.
- [16] Y. He, A. Ben-Hamza, and H. Krim, "An information divergence measure for (ISAR) image registration," in *Proc. IEEE Workshop Statistical Signal Processing*, Singapore, 2001.
- [17] H. Heemuchwala, A. Hero, and P. Carson, "Fast entropic-graph matching for image registration," in *Proc. Asilomar Conf.*, Monterey, CA, to be published.
- [18] H. Heemuchwala and A.O. Hero, "Application of entropic graphs to image registration," *Comm. and Sig. Proc. Lab. (CSPL)*, Dept. EECS, Univ. Michigan, Ann Arbor, Tech. Rep., June, 2002. Available: [http://www.eecs.umich.edu/hero/det\\_est.html](http://www.eecs.umich.edu/hero/det_est.html)
- [19] A.O. Hero, B. Ma, O. Michel, and J.D. Gorman, "Alpha-divergence for classification, indexing and retrieval," *Comm. and Sig. Proc. Lab. (CSPL)*, Dept. EECS, Univ. Michigan, Ann Arbor, Tech. Rep. 328, May 2001. Available: [http://www.eecs.umich.edu/hero/det\\_est.html](http://www.eecs.umich.edu/hero/det_est.html)
- [20] A. Hero, J. Costa, and B. Ma, "Convergence rates of minimal graphs with random vertices," submitted for publication. Available: [http://www.eecs.umich.edu/hero/det\\_est.html](http://www.eecs.umich.edu/hero/det_est.html)
- [21] A. Hero and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *Proc. Meeting Intl. Soc. Optical Engin. (SPIE)*, vol. 3459, San Diego, CA, July 1998, pp. 250-261.
- [22] A. Hero and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *Proc. IEEE Workshop Higher Order Statistics*, Caesaria, Israel, June 1999.
- [23] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal  $k$ -point random graphs," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 1921-1939, Sept. 1999.
- [24] R. Hoffman and A.K. Jain, "A test of randomness based on the minimal spanning tree," *Pattern Recognit. Lett.*, vol. 1, pp. 175-180, 1983.
- [25] P.J. Huber, *Robust Statistics*. New York: Wiley, 1981.
- [26] I.A. Ibragimov and R.Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag, 1981.
- [27] J.A. O'Sullivan, "Divergence penalty for image regularization," in *Proc. 1994 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-94)*, vol. V, 1994, pp. 541-544.
- [28] H. Joe, "On the estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 41, pp. 683-697, 1989.
- [29] H. Kim and A. Hero, "Comparison of GLR and invariant detectors under structured clutter covariance," *IEEE Trans. Image Processing*, vol. 10, pp. 1509-1520, Oct. 2001.
- [30] J.B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, pp. 48-50, 1956.
- [31] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.
- [32] B. Ma, "Parametric and non-parametric approaches for multisensor data fusion," Ph.D. dissertation, EECS Dept., Univ. Michigan, Ann Arbor, MI, 2001. Available: <http://www.eecs.umich.edu/hero/research.html>
- [33] B. Ma, A.O. Hero, J. Gorman, and O. Michel, "Image registration with minimal spanning tree algorithm," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Oct. 2000.
- [34] G.L. McLachlan and K.E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.
- [35] O. Michel, R. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measures," in *Proc. IEEE Int. Time-Frequency Time-Scale Analysis Symp.*, Oct. 1994, pp. 64-67.
- [36] O. Michel, P. Flandrin, and A. Hero, "Automatic extraction of time-frequency skeletons with minimal spanning trees," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Istanbul, Turkey, June 2000.
- [37] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Trans. Inform. Theory*, vol. IT-35, no. 1, pp. 193-196, 1989.
- [38] H. Neemuchwala, A. Hero, and P. Carson, "Feature coincidence trees for registration of ultrasound breast images," in *IEEE Int. Conf. Image Processing*, Thessaloniki, Greece, Oct. 2001.
- [39] D.N. Neuhoﬀ, "On the asymptotic distribution of the errors in vector quantization," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 461-468, Mar. 1996.
- [40] J.A. O'Sullivan, "Divergence penalty for image registration," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc.* Adelaide, Australia, 1994.
- [41] J.P. Pluim, J.B.A. Maintz, and M.A. Viergever, " $f$ -information measures in medical image registration," *Proc. Meeting Intl. Soc. Optical Engin. (SPIE)*, M. Sonka and K. M. Hanson, Eds., vol. 4322, pp. 579-587, 2001.
- [42] H.V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1988.
- [43] R.C. Prim, "Shortest connection networks and some generalizations," *Bell Syst. Tech. J.*, vol. 36, pp. 1389-1401, 1957.
- [44] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees—Short or small," *SIAM J. Discrete Math.*, vol. 9, pp. 178-200, 1996.
- [45] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, vol. 1, pp. 547-561, 1961.
- [46] J.M. Steele, *Probability Theory and Combinatorial Optimization* (CBMF-NSF Regional Conferences in Applied Mathematics, vol. 69). Society for Industrial and Applied Mathematics (SIAM), 1997.
- [47] R. Stoica, J. Zerubia, and J.M. Francos, "Image retrieval and indexing: A hierarchical approach in computing the distance between textured images," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998.
- [48] R. Stoica, J. Zerubia, and J.M. Francos, "The two-dimensional wold decomposition for segmentation and indexing in image libraries," in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc.*, Seattle, WA, May 1998.
- [49] H.L. Van-Trees, *Detection, Estimation, and Modulation Theory: Part I*. New York: Wiley, 1968.
- [50] N. Vasconcelos and A. Lippman, "A Bayesian framework for content-based indexing and retrieval," in *Proc. IEEE Data Compression Con.*, Snowbird, UT, 1998. Available: <http://nuno.www.media.mit.edu/people/nuno/>
- [51] N. Vasconcelos and A. Lippman, "Feature representations for image retrieval: Beyond the color histogram," in *Proc. Int. Conf. Multimedia Expo*, New York, 2000. Available: <http://nuno.www.media.mit.edu/people/nuno/>
- [52] O. Vasicek, "A test for normality based on sample entropy," *J. Royal Stat. Soc.*, ser. B, vol. 38, pp. 54-59, 1976.
- [53] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Computer Vision*, MIT, Cambridge, MA, vol. 1, 1995, pp. 16-23.
- [54] J.E. Yukich, *Probability Theory of Classical Euclidean Optimization (Lecture Notes in Mathematics, vol. 1675)*. Berlin, Germany: Springer-Verlag, 1998.