Alpha-Divergence for Classification, Indexing and Retrieval[0] (Revised)

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

Communications and Signal Processing Laboratory
Technical Report CSPL-328

May 2001 (revised June 2002)

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

Abstract

Motivated by Chernoff's bound on asymptotic probability of error we propose the alpha-divergence measure and a surrogate, the alpha-Jensen difference, for feature classification, indexing and retrieval in image and other databases. The alpha-divergence, also known as Renyi divergence, is a generalization of the Kullback-Liebler divergence and the Hellinger affinity between the probability density characterizing image features of the query and the density characterizing features of candidates in the database. As in any divergence-based classification problem, the alpha-divergence must be estimated from the query or reference object and the objects in the database. The surrogate for the alpha-divergence, called the alpha-Jensen difference, can be simply estimated using non-parametric estimation of the joint alpha-entropy of the merged pairs of feature vectors. Two methods of alpha-entropy estimation are investigated: (1) indirect methods based on parametric or non-parametric density estimation over feature space; and (2) direct methods based on combinatorial optimization of minimal spanning trees or other continuous quasi-additive graphs over feature space. On the basis of κ-th norm error convergence rate comparisons the minimal graph entropy estimator can have better performance than an indirect entropy estimator implemented with plug-in density estimates. We illustrate these results for estimation of dependency in the plane and geo-registration of images.

1 Introduction

A database of images $\mathcal{X} = \{X_i\}_{i=1}^K$ is queried for content which is closely related to a reference image X_0. The answer to the query is a partial re-indexing of the database in decreasing order of similarity to the reference image using an index function. This content-based retrieval problem arises in geographical information systems, digital libraries , medical information processing, video indexing, multi-sensor fusion, and multimedia information retrieval [33, 37, 36, 35]. Common methods for image indexing and retrieval are color histogram matching and texture matching using cross correlation. While these methods are computationally simple they often lack accuracy and discriminatory power.

[0]Alfred Hero is with the Departments of Electrical Engineering and Computer Science (EECS), Biomedical Engineering, and Statistics at the University of Michigan Ann Arbor, MI 48109-2122. Bing Ma was with the Dept. of EECS at UM and is now with Intervideo, Inc., Fremont, CA. Olivier Michel is with the Department of Astrophysics, University of Nice, France. John Gorman is with ERIM-Veridian, Inc., Ann Arbor, MI,

There are three key ingredients to image retrieval and indexing which impact the accuracy and computation efficiency:

1. selection of image features which discriminate between different image classes yet posess invariances to unimportant attributes of the images, e.g. rigid translation, rotation and scale;
2. application of an index function that measures feature similarity and is capable of resolving important differences between images;
3. query processing and search optimization which allow fast implementation.

While these ingredients are all closely linked, this paper is primarily concerned with the appropriate choice of the feature similarity measure and its optimization. We consider the class of α-divergences, also known as Rényi divergences, and a surrogate function called the α-Jensen difference. The α-divergences can be roughly viewed as distances between the probability models underlying the query and the database of images. A special case of α-divergence is the Kullback-Liebler (KL) divergence which has been applied to indexing and image retrieval by Stoica, Zerubia and Francos [35] and Do and Vetterli [9]. A related quantity is the α-Jensen difference which is a function of the joint α-entropy of pairs of feature vectors derived from the query and images in the database. The α-Jensen difference was recently proposed independently by Ma [26, 25] for registering pairs of image modalities and by He, Ben-Hamza and Krim [15] for registering an arbitrary number of image modalities. It was proposed several years before by Michel *etal* in [27] for classifying time frequency distribution images. Another special case of the α divergence is the α-information which is a generalization of the Shannon mutual information. Although we do not explore this extension here, the α-information can be further generalized to the "f-information" which has been treated in a recent paper [31] for medical image registration and generalizes the mutual-information method of Viola and Wells [40]

Here we motivate the α-divergence for indexing by decision theoretic considerations and large deviation theory of detection and classification. A result of this paper is that use of the KL divergence ($\alpha = 1$) can be suboptimal relative to the more general α-divergence. In particular, we establish that when the feature densities are difficult to discriminate (close together in a weighted sup-norm metric) the theoretically optimal choice of α is $\alpha = 1/2$ which corresponds to the Hellinger affinity, related monotonically to the Hellinger-Battacharya distance, as contrasted to the KL divergence. We compare the local discrimination capabilities of the α-divergence and the α-Jensen difference. In particular we show that for discrimination between pairs of close feature densities the α-divergence admits a value $\alpha = 1/2$ which is universally optimal while for the α-Jensen difference the optimal value of α depends on the feature density pair.

When either the α-divergence or the α-Jensen difference are used to perform indexing, they must be estimated from the query and the database. In this paper we focus on estimation of the α-Jensen difference. When a smooth parametric model for the feature densities exists this entropy metric is a smooth non-linear function of these parameters and parametric estimation techniques such as maximum likelihood can be applied [36, 35]. For the parametric case, the entropy estimation error decreases at rate $1/\sqrt{n}$ where n is the size of the training sample of feature vectors used for density estimation. When there exists no suitable parametric model for the feature densities non-parametric density estimation methods can be applied to estimate the α-divergence or the α-Jensen difference. This technique is called a "density plug-in" method. On the other hand, minimal graph entropy estimation techniques, introduced in Hero and Michel [20], can be applied to directly estimate the α-Jensen difference. Using recent results on convergence of random graphs [16] a quantitative comparison can be made between the convergence rates of indirect density plug-in methods and direct graph entropy estimation methods. These comparisons suggest that the graph entropy methods will uniformly outperform any plug-in methods when the densities are in a smooth class of functions and when $\alpha \in [1/2, (d-1)/d]$.

Finally, we show how minimal graphs can be applied to estimation of α-divergence, when a consistent estimator of the reference density is available; estimation of the α-mutual information and α-Jensen difference. For purposes of illustration, we apply these results to a geo-registration problem.

2 Statistical Framework

Let X_0 be a reference image, called the query, and consider a database X_i, $i = 1, \ldots, K$ of images to be indexed relative to the query. Let Z_i be a feature vectors extracted from X_i. We assume that image X_i's feature vector Z_i is realization Z generated by a j.p.d.f. $f(Z|\underline{\theta})$ which depends on a vector of unknown parameters $\underline{\theta}$ lying in a specified parameter space Θ. Under this probabilistic model the k-th observed image feature vector Z_k is assumed to have been generated from model $f(Z|\underline{\theta}_k)$, where $\underline{\theta}_k$ is called the "true parameter" underlying Z_k, $k = 1, \ldots, K$. Under this statistical framework the similarity between images X_0, X_i is reduced to similarity between feature probability models $f(Z|\underline{\theta}_0), f(Z|\underline{\theta}_i)$.

2.1 Divergence Measures of Dissimilarity

Define the densities $f_i = f(Z|\theta_i)$, $i = 0, \ldots, K$. The α-divergence between f_i and f_0 of fractional order $\alpha \in [0, 1]$ is defined as [32, 7, 2]

$$\begin{aligned} D_\alpha(f_i \| f_0) &= \frac{1}{\alpha - 1} \ln \int f_0 \left(\frac{f_i(z)}{f_0(z)} \right)^\alpha dz \\ &= \frac{1}{\alpha - 1} \ln \int f_i^\alpha(z) f_0^{1-\alpha}(z) dz \end{aligned} \tag{1}$$

Note that $D_\alpha(f_i \| f_0) = D_\alpha(\theta_i \| \theta_0)$ is indexed by θ_i and θ_0.

α-Divergence: Special cases

When specialized to various values of α the α-divergence can be related to other well known divergence measures. Two of the most important examples are the Hellinger affinity $2 \ln \int \sqrt{f_i(z) f_0(z)} dz$ obtained when $\alpha = 1/2$, and is related to the Hellinger-Battacharya distance squared,

$$\begin{aligned} D_{Hellinger}(f_i \| f_0) &= \int \left(\sqrt{f_i(z)} - \sqrt{f_0(z)} \right)^2 dz \\ &= 2 \left(1 - \exp\left(\tfrac{1}{2} D_{\frac{1}{2}}(f_i \parallel f_0) \right) \right), \end{aligned}$$

and the Kullback-Liebler (KL) divergence [24], obtained when $\alpha \to 1$,

$$\lim_{\alpha \to 1} D_\alpha(f_i, f_0) = \int f_0(z) \ln \frac{f_0(z)}{f_i(z)} dz.$$

Only when $\alpha = 1/2$ is the divergence monotonically related to a true distance metric between two densities.

When the density f_0 dominates f_1 and is uniform over a compact domain $\mathcal{Z} \supset \mathrm{support}\{f_i\}$ the α-divergence reduces to the α-entropy, also known as the Rényi entropy:

$$H_\alpha(f_i) = \frac{1}{1 - \alpha} \ln \int_{\mathcal{Z}} f_i^\alpha(z) dz \tag{2}$$

3 α-divergence as an Index Function

The ordered sequence of increasing α-divergence measures $D_\alpha(f_{(1)}\|f_0),\ldots,D_\alpha(f_{(K)}\|f_0)$, induces an indexing, which we call the "true indexing," of the images

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

This indexing is unimplementable given only the Z_i's since it requires the underlying probability models f_i be known to the query processor. The non-statistical indexing problem can now be stated as: given a sequence of divergences $\{D_\alpha(\theta_i\|\theta_0)\}_{i=1}^K$ find the sequence of indices $i_1,\ldots,i_K$ which minimize $D_\alpha(\theta_{i_k}\|\theta_0)$ over the set $\{1,\ldots,K\}-\{i_1,\ldots,i_{k-1}\}$, $k=1,\ldots,K$.

Special cases of the indexing problem are

1. Content-based retrieval: the query is the density of an image object and the database consists of image densities 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: the database consists of K copies of Z_0 which are rotated, translated and possibly locally deformed. The index i_1 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 the method of Viola and Wells [40]. The 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: 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.
4. Performing parameter estimation by minimizing the Hellinger-Battacharya distance is known as minimum-Hellinger-distance-estimation (MHDE) introduced by Beran [5]. While there are obvious similarities, relations of MHDE to indexing will not be explored in this paper.

3.1 Un-normalized α-Divergence and the Chernoff Error Exponent

Here we argue appropriateness of the α-divergence on the basis of large deviations theory results on the exponential rate of decay of the Bayes-optimal classifier between two densities. Note that the Bayes classification error probability below is different from that defined by Vasconcelos [38, 37] in that here the decision error is averaged over an ensemble of image models. Define the un-normalized α-divergence as the -log integral in the definition (2) of the α-divergence:

$$D^u_\alpha(f_1\|f_0) = -\ln\int f^\alpha(Z|\underline{\theta}_1)f^{1-\alpha}(Z|\underline{\theta}_0)\,dZ = (1-\alpha)D_\alpha(f_1\|f_0)$$

Assume that from an i.i.d. sequence of images $X^{(1)},\ldots,X^{(n)}$ we extract feature vectors $\underline{Z}=[Z^{(1)},\ldots,Z^{(n)}]$ each having density $f(Z|\underline{\theta})$ for some $\underline{\theta}\in\Theta$. Consider testing the hypotheses

$$\begin{aligned} H_0 &: \underline{\theta}\in\Theta_0 \\ H_1 &: \underline{\theta}\in\Theta_1 \end{aligned}$$

where Θ_0 and Θ_1 partition the parameter space Θ. In the context of image retrieval the parameter range Θ_1 could cover the K densities of the images in the database while parameter range Θ_0 covers densities outside of the database. In this case testing H_0 vs. H_1 is tantamount to testing whether the query lies in the database (H_1) or not (H_0). If H_1 is decided then sequential hypothesis testing could subsequently be performed to completely search the database for specific query matches by successive refinement of the parameter space Θ_1 over a depth $\log_2(K)$ binary tree.

Let $f(\underline{\theta})$ be a prior over Θ and assume that $P(H_1) = \int_{\Theta_1} f(\underline{\theta})d\underline{\theta}$ and $P(H_0) = 1 - P(H_1)$ are both positive. Then for any test of H_0 vs. H_1 define the average probability of error

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

where $\beta(n)$ and $\alpha(n)$ are Type II and Type I errors of the test, respectively, which depend on $\underline{\theta}$ in general. The α-divergence measure can be related to the minimum attainable probability of error through the Chernoff bound of large deviations theory [8]:

$$\liminf_{n\to\infty} \frac{1}{n} \ln P_e(n) = - \sup_{\alpha\in[0,1]} D^u_\alpha(\overline{f}_1 \| \overline{f}_0), \tag{3}$$

where $\overline{f}_1(Z) = \int_{\Theta_1} f(Z|\underline{\theta})f(\underline{\theta})d\underline{\theta}$ and $\overline{f}_0(Z) = \int_{\Theta_0} f(Z|\underline{\theta})f(\underline{\theta})d\underline{\theta}$. The quantity $\sup_{\alpha\in[0,1]} D^u_\alpha(\overline{f}_1 \| \overline{f}_0)$ in (3) is called the *Chernoff exponent* which gives the asymptotically optimal rate of exponential decay of the error probability for testing H_0 vs H_1. The optimal $\alpha = \alpha_o$ which attains the maximum in (3) is obtained by finding the value of α which maximizes $D^u_\alpha(\overline{f}_1 \| \overline{f}_0)$.

$$\alpha_o = \text{argmax}_{\alpha\in[0,1]} \int \overline{f}_1^{\alpha}(Z)\overline{f}_0^{1-\alpha}(Z)\, dZ \tag{4}$$

3.2 Selection of α

We have empirically determined that for an image indexing problem arising in georegistration (see Section 5) the value of α leading to highest resolution seems to cluster around either 1 or $1/2$ corresponding to the KL divergence and the Hellinger affinity respectively [25]. The determining factor appears to be the degree of differentiation between the densities $\{f_i\}_{i=0}^K$. If the densities are very similar, i.e. difficult to discriminate, then the indexing performance of the Hellinger affinity distance ($\alpha = 1/2$) was observed to be better that the KL divergence ($\alpha = 1$). This is consistent with the asymptotic local analysis below.

A locally optimum α can be explored by asymptotic analysis of the Chernoff exponent. Specifically, the following is a direct result of Proposition 5 in Appendix A.

Proposition 1

$$D^u_\alpha(f_0 \| f_1) \quad = \quad \frac{\alpha(1-\alpha)}{2} \int \frac{(f_0(x) - f_1(x))^2}{f_0(x) + f_1(x)}\, dx + o(\Delta^2), \tag{5}$$

where $\Delta \in [0,1]$ *is*

$$\Delta = 2 \sup_x \frac{|f_1(x) - f_0(x)|}{f_1(x) + f_0(x)}.$$

Recall that the detection error probability decreases exponentially with Chernoff exponent $\sup_{\alpha\in[0,1]} D^u_\alpha(f_1 \| f_0)$. A consequence of (5) is that to order Δ^2 the optimum value of α in the Chernoff exponent is $1/2$.

As an illustrative example consider the case where f_0 and f_1 are multivariate Gaussian densities. The KL information for such a Gaussian feature model was adopted in [36, 35] for performing image indexing. Let $f(x; \mu, \Lambda)$ be a real d-dimensional normal density with mean vector μ and non-singular covariance matrix Λ. The un-normalized α-divergence $D^u_\alpha(f_1\|f_0) = D^u_\alpha(f(x;\mu_1,\Lambda_1)\|f(x;\mu_0,\Lambda_0))$ of order α is given by (see Proposition 6 in Appendix B).

$$D^u_\alpha(f(x;\mu_1,\Lambda_1)\|f(x;\mu_0,\Lambda_0)) = \underbrace{-\tfrac{1}{2}\ln\frac{|\Lambda_0|^\alpha|\Lambda_1|^{1-\alpha}}{|\alpha\Lambda_0+(1-\alpha)\Lambda_1|}}_{\text{Term A}} + \underbrace{\frac{\alpha(1-\alpha)}{2}\Delta\mu^T(\alpha\Lambda_0+(1-\alpha)\Lambda_1)^{-1}\Delta\mu}_{\text{Term B}} \quad (6)$$

where $\Delta\mu = \mu_1 - \mu_0$.

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 Λ_1 and Λ_0 with mean weights α and $1-\alpha$. Term B is the quadratic difference of mean vectors normalized by the arithmetic mean of Λ_1 and Λ_0 with mean weights α and $1-\alpha$.

An asymptotic expansion yields the following expression for the case that $\Delta\mu = 0$, i.e. equal means,

$$D^u_\alpha(f_1\|f_0) = \frac{\alpha(1-\alpha)}{4}\mathrm{tr}(\Lambda_1-\Lambda_0)^2 + o(\mathrm{tr}(\Lambda_1-\Lambda_0)^2).$$

so that locally the Chernoff exponent increases in the trace norm of the differences between the feature covariances and, as expected, $\alpha = 1/2$ is optimal.

4 Divergence and Entropy Estimation

In practice the image model parameters $\underline{\theta}_k$'s are unknown so that the actual relative ordering of α-divergences $\{D_\alpha(\underline{\theta}_k\|\underline{\theta}_0)\}_{k=1}^K$ is also unknown. The statistical problem of indexing can be stated as follows: based on a single realization $X_k = X_k^{(1)}$ of the k-th image, $k = 0,\ldots,K$, estimate the actual rank ordering of α-divergences $\{D_\alpha(\underline{\theta}_k|\underline{\theta}_0)\}_{k=1}^K$ between feature distributions. Divergence estimation is closely related to entropy estimation which has a long history in the statistics and information theory communities.

Estimation of entropy is an important problem that arises in statistical pattern recognition, adaptive vector quantization, image registration and indexing, and other areas. Non-parametric estimation of Shannon entropy has been of interest to many in non-parametric statistics, pattern recognition, model identification, image registration and other areas [14, 23, 1, 39, 4, 40, 10]. Estimation of α-entropy arises as a step towards Shannon entropy estimation, e.g., Mokkadem [28] constructed a non-parametric estimate of the Shannon entropy from a convergent sequence of α-entropy estimates. However, estimation of the α-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 [11, 30]. The α-entropy parametrizes the Chernoff exponent governing the minimum probability of error in binary detection problems [22, 6]. It also has been used for image registration from multiple modalities via the α-Jensen difference [26, 25, 15]. The most natural estimation method is to substitute a non-parametric 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 [4]. Other methods of Shannon entropy estimation discussed in [4] include sample spacing estimators, restricted to $d = 1$, and estimates based on nearest neighbor distances.

Three general classes of methods can be identified: parametric estimators, non-parametric estimators based on density or function estimation, and non-parametric estimators based on direct estimation. The first two methods are classified as *plug-in* techniques where a parametric or non-parametric density estimate $\hat{f}$ or function estimate $\widehat{f^\alpha}$ are simply plugged into the divergence formula. When an accurate parametric model and good parameter estimates are available parametric plug-in estimates of divergence are attractive due to their $1/\sqrt{n}$ convergence properties. An analytical parametric form of the divergence can often be derived over the parametric class of densities considered and maximum likelihood can be used to estimate parameters in the divergence formula. This approach was adopted under a multivariate Gaussian image model by Stoica *etal* [35] for image retrieval. For Gaussian f_1 and f_0 the KL divergence $D_1(f_1\|f_0)$ has a simple closed form expression, which can be derived as the limit of (30) as $\alpha \to 1$, and the authors in [35] proposed using maximum likelihood or least squares estimates of the mean and covariance parameters of each image.

Non-parametric plug-in divergence estimates do not benefit from closed form parametric expressions for divergence but avoid pitfalls of model dependent estimates. For example, when a non-parametric estimate of $\hat{f}$ or of $\widehat{f^\alpha}$ are available the following plug-in estimates of α-entropy are natural

$$H_\alpha(\hat{f}) = \frac{1}{1-\alpha} \ln \int \hat{f}^\alpha(z)dz \tag{7}$$

$$H_\alpha(\widehat{f^\alpha}) = \frac{1}{1-\alpha} \ln \int \widehat{f^\alpha}(z)dz. \tag{8}$$

For the special case of estimation of Shannon entropy $\lim_{\alpha\to 1} H_\alpha(f) = -\int f(z)\ln f(z)dz$ recent non-parametric estimation proposals have included: histogram estimation plug-in [12]; the more general kernel density estimation plug-in [1]; and sample-spacing density estimator plug-in [13]. The reader is referred to [4] for a comprehensive overview of work in non-parametric estimation of Shannon entropy. The main difficulties with non-parametric 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; the high dimensional integration in (7) might be difficult.

The problems with the above methods can be summarized by the basic observation: on the one hand parameterizing the divergence and entropy functionals with infinite dimensional density function models is a costly over-parameterization, while on the other hand artificially enforcing lower dimensional density parametrizations 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 non-parametric density estimation [18, 20, 19]. 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 specify a strongly consistent estimator of entropy for Lebesgue continuous densities. In particular, let $\mathcal{Z}^{(n)} = \{Z^{(1)}, \ldots, Z^{(n)}\}$ and define the Euclidean functional of order γ: $L_\gamma = L_\gamma(\mathcal{Z}^{(n)}) = \min_{e\in\mathcal{E}} \sum_e |e|^\gamma$ the overall length of a graph spanning n i.i.d. vectors $Z^{(i)}$ in $\mathbf{R}^d$ each with density f. Here $\gamma \in (0, d)$ is real, e are edges in a graph connecting pairs of $Z^{(i)}$'s and the minimization is over some suitable subsets $\mathcal{E}$ 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 tour. The asymptotic behavior of L_γ over random points $\mathcal{Z}^{(n)}$ as $n \to \infty$ has been studied for over half a decade [3, 41, 34] and, based on these studies, in [20] we gave conditions under which

$$\hat{H}_\alpha(\mathcal{Z}^{(n)}) = \ln L_\gamma(\mathcal{Z}^{(n)})/n^\alpha - \ln \beta_{L_\gamma,d} \tag{9}$$

is an asymptotically unbiased and almost surely consistent estimator of the un-normalized α-entropy of f where $\alpha = (d-\gamma)/d$ and $\beta_{L_\gamma,d}$ is a constant bias correction depending on the graph minimality criterion over $\mathcal{E}$ but independent of f.

As shown in [20], optimal pruning of greedy implementations of the minimal graph can robustify the entropy estimator against outliers from contaminating distributions[1]. This procedure consists of constructing the k-minimal graph which is defined as the minimum weight minimal graph spanning any k out of the n points in the realization of

[1]Note that the overly strong BV condition assumed in [20] can be considerably weakened by replacing BV space with Sobolev space and applying Lemma 2 of [16]

f. Divergence $D_\alpha(f_1\|f_0)$ between the observed feature density f and a reference feature density f_0 can be estimated similarly via performing a preprocessing step before implementing the minimal-graph entropy estimator. In this preprocessing step one applies a measure transformation on the feature space which converts the reference density to a uniform density over the unit cube [19].

As contrasted with density-based estimates of entropy, minimal graph entropy estimators enjoy the following properties: they can have faster asymptotic convergence rates (see next sub-section), especially for non-smooth densities and for low dimensional feature spaces; they completely bypass the complication of chosing and fine tuning parameters such as histogram bin size, density kernel width, complexity, and adaptation speed; the α parameter in the α-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 may be a bottleneck for a large number of feature samples, for example the MST or the k-NNG are "almost linear" algorithms of complexity $O(n \log n)$.

When f_0 is known the α-divergence can be estimated by minimal graph methods using the measure transformation method outlined [19]. For unknown f_0 and unknown f_1 the existence of consistent minimal-graph estimators of $D_\alpha(f_1\|f_0)$ is an open problem. The sequel of this paper will be concerned with an alternative index function, called the α-Jensen difference, which is a function of the joint entropy of the query and candidate image feature sets. This function can be estimated using the entropy estimation techniques discussed above.

4.1 Robust Entropy Estimation: the k-MST

In Hero and Michel [20] we established strong convergence results for a greedy approximation to the following minimal k-point Euclidean graph problem. Assume that one is given a set $\mathcal{X}_n = \{x_1, \ldots, x_n\}$ of n random points in the unit cube $[0,1]^d$ of dimension $d \geq 2$ following a Lebesgue density f. Fix k and denote by $\mathcal{X}_{n,k}$ a k-point subset of $\mathcal{X}_n$, $0 < k \leq n$. The minimal k-point Euclidean graph problem is to find the subset of points $\mathcal{X}_{n,k}$ and the set of edges connecting these points such that the resultant graph has minimum total weight $L(\mathcal{X}_{n,k})$. This problem arises in many combinatorial optimization problems, see references in [20] for a partial list. In addition to these problems, in [20] we noted that the weight function k-minimal graph could be useful for robust estimation of the Rényi entropy of order ν, where $\nu = (d-\gamma)/d \in (0,1)$ is specified by the dimension $d \geq 2$ and a weight exponent $\gamma \in (0,d)$ on the Euclidean distance between vertices of the graph. The intuition is that the k-MST tends to only accept those points that are most clustered near each other, thereby the rejected points tend to be stragglers outside of the k cluster. In Hero and Michel [20] we established almost sure (a.s.) convergence of the normalized weight function of a greedy approximation to the a class of k-point minimal graphs including the k-MST. This normalized weight function converges to a limit which, for $k \ll n$, is a close approximation to the entropy integral in (9). The influence function was investigated in Hero and Michel [20] and quantitatively establishes the k-MST as a robust estimator of entropy.

The greedy approximation was introduced in [20] to reduce the exponential complexity of the exact k-MST algorithm yet retain its outlier resistant properties. This greedy approximation involves the following partitioning heuristic: 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 and, when $k = \alpha n$, for which the log of the normalized weight function $\hat{L}_{n,k}/k^\nu$ converges to an approximation of the Rényi entropy of order ν. The computational advantage of the greedy algorithm comes from its divide-and-conquer multi-resolution structure: it only requires solving the difficult non-linear minimal graph construction on cells containing smaller numbers of points. When $\alpha = 1$ 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 multi-resolution approximation to the k-MST.

4.2 Entropy Estimator Convergence Comparisons

It can be shown that when f is a density supported on the unit cube $[0,1]^d$ the bias and variance of direct minimal-graph entropy estimators (9) and indirect density plug-in entropy estimators (7) converge to zero as a function of the number n of i.i.d. observations [20, 1]. Here we attack a harder problem: comparing the asymptotic convergence rates of the mean κ-th norm of the error.

We treat entropy estimates of the form $\hat{H}_\alpha = (1-\alpha)^{-1} \ln \hat{I}_\alpha$, where $\hat{I}_\alpha$ is a consistent estimator of the integral

$$I(f^\alpha) = \int f^\alpha(x)dx.$$

Given non-parametric function estimates $\widehat{f^\alpha}$ of f^α based on n i.i.d. observations $X_1, \ldots, X_n$ from f, define the function plug-in estimator $I(\widehat{f^\alpha})$. Define the minimal-graph estimator $\hat{I}_\alpha = L_\gamma(X_1, \ldots, X_n)/(\beta_{L_\gamma,d} n^\alpha)$, where $\gamma \in (0,d)$ is selected such that $\alpha = (d-\gamma)/d$ and L_γ is continuous quasi-additive. Let $\hat{H}_\alpha$ and $\hat{I}_\alpha$ denote estimates computed by either the plug-in or the minimal-graph estimators. A standard purturbation analysis of $\ln(z)$ establishes

$$|\hat{H}_\alpha - H_\alpha(f)| \quad = \quad \frac{1}{1-\alpha} \frac{|\hat{I}_\alpha - I(f^\alpha)|}{I(f^\alpha)} + o(|\hat{I}_\alpha - I(f^\alpha)|).$$

Thus as a function of n the asymptotic $\mathcal{L}_\kappa$ rate of convergence of $\hat{H}_\alpha - H_\alpha(f)$ will be identical to that of $\hat{I}_\alpha - I(f^\alpha)$.

We first deal with the density plug-in estimator $I(\hat{f})$ when f^α is a smooth function in the sense of belonging to a Besov space. The Besov space $B^\sigma_{p,q}(\mathbb{R}^d)$ is a space of functions $f \in \mathcal{L}_p$ whose smoothness increases in σ and q (For more detailed definitions see [16]). For scalar $C > 0$ define the Besov ball of radius C

$$B^\sigma_{p,q}(C) = \left\{ f \in B^\sigma_{p,q}(\mathbb{R}^d) : \|f\|_{\sigma,p,q} \le C \right\}$$

where $\|f\|_{\sigma,p,q}$ is the Besov norm.

The following proposition is established in [16].

Proposition 2 *Assume that the Lebesgue density f with support $\mathcal{S} \in [0,1]^d$ is such that $f^\alpha \in B^\sigma_{p,q}(C)$. Then, for any plug-in estimator $I(\widehat{f^\alpha})$*

$$\sup_{f^\alpha \in B^\sigma_{p,q}(C)} \left[E \left| I(\widehat{f^\alpha}) - I(f^\alpha) \right|^2 \right]^{1/2} \;\ge\; O\left(n^{-\sigma/(2\sigma+d)} \right). \tag{10}$$

For the minimal-graph entropy estimator exact convergence rates are more difficult to establish. Even in the relatively simple case of a uniform density f, exact rates are presently known only for $d = 2$ [34, 41]. it has been shown [41] that when f is a uniform density the MST length functional L_γ converges to the integral $I_\alpha(f)$ with rate upper bounded by $O(n^- - 1/d)$, which is exact for $d = 2$. In [16] we establish convergence rates for a Sobolev class of densities which contains $B^1_{p,1}(\mathbb{R}^d)$. Functions in this space have the property that they can be uniformly approximated in $\mathcal{L}_p$-norm by continuously differentiable functions.

Using these convergence results, the following can be established [16, Proposition 8].

Corollary 1 *Let $d \ge 2$ and $\alpha \in [1/2, (d-1)/d]$. Assume that the Lebesgue density f having support $\mathcal{S} \subset [0,1]^d$ is such that $f^\alpha \in B^1_{p,1}(C)$, $p > d$. Then, for any plug-in estimator $I(\widehat{f^\alpha})$*

$$\sup_{f^\alpha \in B^1_{p,1}(C)} \left[E \left| I(\widehat{f^\alpha}) - I(f^\alpha) \right|^2 \right]^{1/2} \;\ge\; O\left(n^{-\frac{1}{2+d}} \right), \tag{11}$$

while for the minimal-graph estimator $\hat{I}_\alpha$

$$\sup_{f^\alpha \in B^1_{p,1}(C)} \left[E \left| \hat{I}_\alpha - I(f^\alpha) \right|^2 \right]^{1/2} \;\le\; O\left(n^{-\frac{\alpha(d+1-d/p)}{\alpha(d+1-d/p)+1}\frac{1}{d}} \right). \tag{12}$$

Using the above we conclude that, under the stated assumptions, the minimal graph-estimators always have faster convergence rate than plug-in estimators for $\alpha \in [1/2, (d-1)/d]$.

4.3 α-Jensen Difference Index Function

Here we study an alternative index function based on the Jensen entropy difference. This index function was independently proposed by Ma [25] and He *etal* [15] for image registration problems. Let f_0 and f_1 be two densities and $\beta \in [0,1]$ be a mixture parameter. The α-Jensen difference is the difference between the α-entropies of the mixture $f = \beta f_0 + (1-\beta) f_1$ and the mixture of the α-entropies of f_0 and f_1 [2]:

$$\triangle H_\alpha(\beta, f_0, f_1) \stackrel{\triangle}{=} H_\alpha(\beta f_0 + (1-\beta) f_1) - [\beta H_\alpha(f_0) + (1-\beta) H_\alpha(f_1)], \quad \alpha \in (0,1). \tag{13}$$

The α-Jensen difference is measure of dissimilarity between f_0 and f_1: as the α-entropy $H_\alpha(f)$ is concave in f it is clear from Jensen's inequality that $\triangle H_\alpha(\beta, f_0, f_1) = 0$ iff $f_0 = f_1$ a.e.

The α-Jensen difference can be motivated as an index function as follows. Assume that two sets of labeled feature vectors $\mathcal{Z}_0 = \{Z_0^{(i)}\}_{i=1,\ldots,n_0}$ and $\mathcal{Z}_1 = \{Z_1^{(i)}\}_{i=1,\ldots,n_1}$ are extracted from images X_0 and X_1, respectively. Assume that each of these sets consist of independent realizations from densities f_0 and f_1, respectively. Define the union $\mathcal{Z} = \mathcal{Z}_0 \cup \mathcal{Z}_1$ 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 $n \to \infty$ where $\beta = \lim_{n\to\infty} n_0/n$. This motivates the following consistent minimal-graph estimator of Jensen difference (13) for $\beta = n_0/n$:

$$\widehat{\triangle H}_\alpha(\beta, f_0, f_1) \stackrel{\triangle}{=} \hat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1) - \left[\beta \hat{H}_\alpha(\mathcal{Z}_0) + (1-\beta) \hat{H}_\alpha(\mathcal{Z}_1) \right], \quad \alpha \in (0,1).$$

where $\hat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1)$ is the minimal-graph entropy estimator (9) constructed on the n point union of both sets of feature vectors and $\hat{H}_\alpha(\mathcal{Z}_0)$, $\hat{H}_\alpha(\mathcal{Z}_1)$ are constructed on the individual sets of n_0 and n_1 feature vectors, respectively. We can similarly define the density-based estimator of Jensen difference based on entropy estimates of the form (7) constructed on $\mathcal{Z}_0 \cup \mathcal{Z}_1$, $\mathcal{Z}_0$ and $\mathcal{Z}_1$.

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 $\{\triangle H_\alpha(\beta, f_0, f_i)\}_{i=1}^K$. This fact was used in Ma *etal* [26] for registering a query image to a database of images which are generated by entropy-preserving rigid transformations of a reference image.

4.4 Comparisons of α-Jensen Difference and α-Divergence

The local discrimination capabilities of the α-Jensen difference and the α-divergence can easily be compared using the results (Propositions 4 and 5) obtained in Appendix A:

$$\ln D_\alpha(f_0 \| f_1) \;=\; \ln \left(E_{f_{\frac{1}{2}}} \left[\left(\frac{f_0 - f_1}{f_{\frac{1}{2}}} \right)^2 \right] \right) + C_1 + o(\Delta^2) \tag{14}$$

$$\ln \triangle H_\alpha(\beta, f_0, f_1) = \ln\left(E_{\tilde f^\alpha_{\frac{1}{2}}}\left[\left(\frac{f_0 - f_1}{f_{\frac{1}{2}}}\right)^2\right] + \frac{\alpha}{1-\alpha}\left(E_{\tilde f^\alpha_{\frac{1}{2}}}\left[\frac{f_0 - f_1}{f_{\frac{1}{2}}}\right]\right)^2\right) + C_2 + o(\Delta^2) \quad (15)$$

where $E_f[g(x)] = \int f(x)g(x)\,dx$, $\tilde f^\alpha_{\frac{1}{2}}(x) \stackrel{\triangle}{=} \frac{f^\alpha_{\frac{1}{2}}(x)}{\int f^\alpha_{\frac{1}{2}}(x)\,dx}$ is a "tilted" pdf, Δ is a term that decreases in the difference $f_0 - f_1$, and C_1, C_2 are constants independent of f_0, f_1.

There are a number of interesting properties of $\ln D_\alpha(f_0\|f_1)$ and $\ln \triangle H_\alpha(\beta, f_0, f_1)$:

- Up to order Δ^2 the leading terms in (14) and (15) are the curvatures of the log α-divergence and the log α-Jensen difference, respectively. These curvatures are a measure of the sensitivity of these index functions for discriminating between density pairs f_0, f_1. The discrimination capability of the α-divergence is locally independent of α while that of the α-Jensen difference depends on α.
- When α approaches 0, tail differences between the two densities f_0 and f_1 are influential on $\triangle H_\alpha(\beta, f_0, f_1)$.
- When α approaches 1, central differences between the two densities become highly pronounced in $\triangle H_\alpha(\beta, f_0, f_1)$. Therefore, if the feature densities differ in regions where there is a lot of mass one should choose α close to 1 to ensure locally optimum discrimination with $\triangle H_\alpha(\beta, f_0, f_1)$.
- $\triangle H_\alpha(\beta, f_0, f_1)$ has maximal discriminative capability when $\beta = \frac{1}{2}$, i.e., when two images yield the same number of feature vectors.

4.5 Estimation of α-Divergence

Here we describe extensions of the entropy estimation procedures described above to information divergence estimation. Let $g(x)$ be a reference density on $\mathbf{R}^d$ which dominates the density $f(x)$ of a sample point $x = [x^1, \ldots, x^d]^T$ in the sense that for all x such that $g(x) = 0$ we have $f(x) = 0$. As described below, when $g(x)$ is known the minimal graph estimator can be applied, achieving the rms convergence rate (12) when f and g are in a Besov space. Any plug-in estimator of the α-divergence will have slower rms convergence rate given by (11).

As introduced in Hero and Michel [19] minimal graph divergence estimation is performed by constructing a minimal graph on a transformed sample where the transformation corresponds to a change of measure which flattens the reference distribution g. For any x such that $g(x) > 0$ let $g(x)$ have the product representation $g(x) = g(x^1)g(x^2|x^1)\ldots g(x^d|x^{d-1}, \ldots, x^1)$ where $g(x^k|x^{k-1}, \ldots, x^1)$ denotes the conditional density associated with $g(x)$ of the k-th component. In what follows we will ignore the set $\{x : g(x) = 0\}$ since, as $f(x) = 0$ over this set, it has probability zero. Now consider generating the vector $y = [y^1, \ldots, y^d]^T \in \mathbf{R}^d$ by the following vector transformation

$$\begin{aligned} y^1 &= G(x^1) \\ y^2 &= G(x^2|x^1) \\ &\vdots \\ y^d &= G(x^d|x^{d-1}, \ldots, x^1) \end{aligned} \quad (16)$$

where $G(x^j|x^{j-1}, \ldots, x^1) = \int_{-\infty}^{x^j} g(\tilde x^j|x^{j-1}, \ldots, x^1)d\tilde x^j$ is the cumulative conditional distribution of the j-th component, which is monotone increasing except on the zero probability set $\{x : g(x) = 0\}$. Thus, except for this probability zero set, the conditional distribution has an inverse $x^j = G^{-1}(y^j|x^{j-1}, \ldots, x^1) = G^{-1}(y^j|y^{j-1}, \ldots, y^1)$

and it can be shown (via the standard Jacobian formula for transformation of variables) that the induced joint density, $h(y)$, of the vector y takes the form:

$$h(y) = \frac{f(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))}{g(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))} \tag{17}$$

Now let $L(\mathcal{Y}_n)$ denote the length of the greedy approximation to the MST constructed on the transformed random variables $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$. Then, by the consistency property of the MST estimator, we know that

$$\hat{H}_\alpha(\mathcal{Y}_n) \to \frac{1}{1-\alpha} \ln \int h^\alpha(y) dy \quad (a.s.). \tag{18}$$

Making the inverse transformation $y \to x$ specified by (16) in the above integral, noting that, by the Jacobian formula, $dy = g(x)dx$, and using the expression (17) for h, it is easy to see that the integral in the right hand side of (18) is equivalent to the Rényi information divergence of $f(x)$ with respect to $g(x)$

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

Hence we have established that $\hat{H}_\alpha(\mathcal{Y}_n)$ is a strongly consistent estimator of the Rényi information divergence above. The results of Hero and Ma [17] and Hero and Michel [20] can thus be easily be extended to classification against any *arbitrary* distribution f_o, and not just the uniform distribution initially studied by Hoffman and Jain [21] and Hero and Michel [18]. This extension also holds for the k-MST described in Section 4.1.

4.5.1 Application to robust divergence estimation

Here we applied the k-MST to robustly classify a triangular vs. uniform density on the plane. 256 samples were simulated from a triangle-uniform mixture density $f = (1-\epsilon)f_1 + \epsilon f_0$ where $f_1(x) = (\frac{1}{2} - |x^1 - \frac{1}{2}|)(\frac{1}{2} - |x^2 - \frac{1}{2}|)$ is a (separable) triangular shaped product density and $f_0 = 1$ is a uniform density, both supported on the unit square $x = (x^1, x^2) \in [0,1]^2$. The Rényi information divergences $I(f, f_0)$ and $I(f, f_1)$ were estimated by $\hat{H}_\alpha(\mathcal{X}_n)$ and $\hat{H}_\alpha(\mathcal{Y}_n)$, respectively, for $\alpha = \frac{1}{2}$ ($\gamma = 1$ in the k-MST construction). $\mathcal{Y}_n$ was obtained by applying the mapping $y = (y^1, y^2) = (F_1(x^1), F_1(x^2))$ to the data sample $\mathcal{X}_n$, where $F_1(u)$ is the marginal cumulative distribution function associated with the triangular density.

In a first sequence of experiments the estimates $\hat{H}_\alpha(\mathcal{X}_n)$ and $\hat{H}_\alpha(\mathcal{Y}_n)$ of the respective quantities $I(f, f_0)$ and $I(f, f_1)$ were thresholded to decide between the hypotheses $H_0 : \epsilon = 0$ vs. $H_1 : \epsilon \neq 0$ and $H_0 : \epsilon = 1$ vs. $H_1 : \epsilon \neq 1$, respectively. The receiver operating characteristic (ROC) curves are indicated in Figures 1 and 2. Note that, as expected, in each case the detection performance improves as the difference between the assumed H_0 and H_1 densities increases.

In a second sequence of experiments we selected two realizations of the triangle-uniform mixture model for the values $\epsilon = 0.1$ and $\epsilon = 0.9$. For the former case the triangular is the dominating density and for the latter case the uniform is the dominating density. In each case the k-MST was implemented ($k = 90$) as a robust clustering algorithm to identify data points from the dominating densities - in the former case the k-MST was applied directly to $\mathcal{X}_n$ while in the latter case it was applied to $\mathcal{Y}_n$. The resulting k-MST quantities $\hat{H}_\alpha(\mathcal{X}_{n,k})$ and $\hat{H}_\alpha(\mathcal{Y}_{n,k})$ can be interpreted as robust estimates of the uncontaminated Rényi information divergences $I(f_1, f_0)$ and $I(f_0, f_1)$. respectively. Figure 3-5 illustrate the effectiveness of these estimates as "outlier rejection" algorithms.

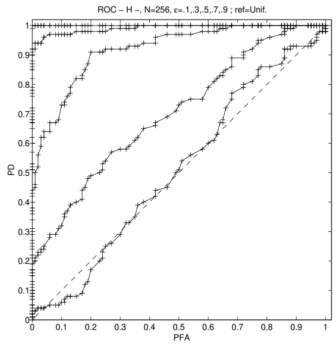


Figure 1: *ROC curves for the Rényi information divergence test for detecting triangle-uniform mixture density* $f = (1-\epsilon)f_1 + \epsilon f_0$ *(*H_1*) against the uniform hypothesis* $f = f_0$ *(*H_0*). Curves are decreasing in* ϵ *over the range* $\epsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

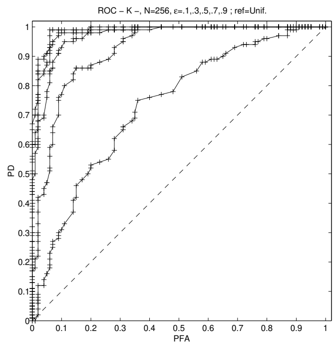


Figure 2: *Same as Figure 1 except test is against triangular hypothesis* $f = f_1$ *(*H_0*). Curves are increasing in* ϵ*.*

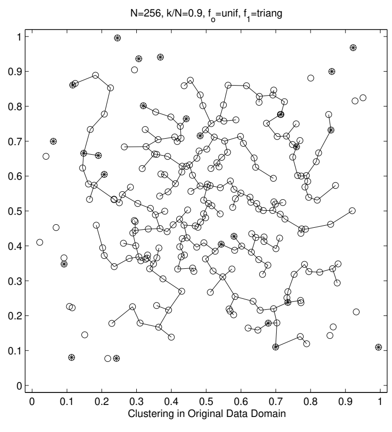


Figure 3: *A scatterplot of a 256 point sample from triangle-uniform mixture density with* $\epsilon = 0.1$. *Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Superimposed is the k-MST implemented directly on the scatterplot* $\mathcal{X}_n$ *with* $k = 230$.

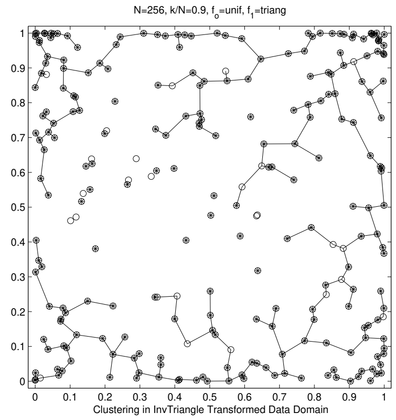


Figure 4: *A scatterplot of a 256 point sample from triangle-uniform mixture density with $\epsilon = 0.9$ in the transformed domain $\mathcal{Y}_n$. Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Superimposed is the k-MST implemented on the transformed scatterplot $\mathcal{Y}_n$ with $k = 230$*

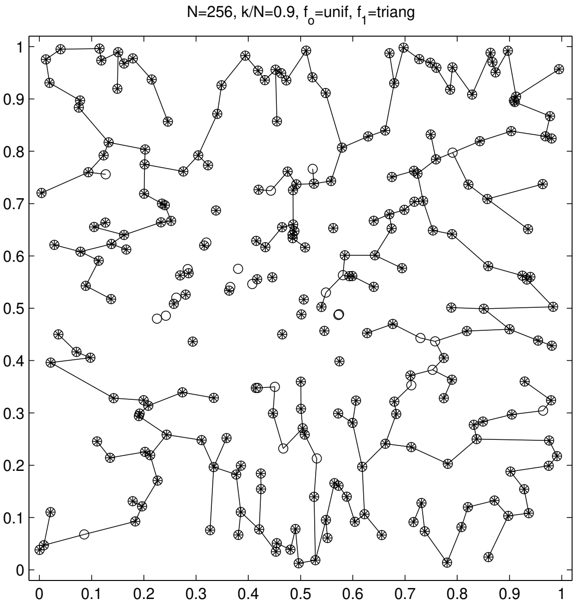


Figure 5: *Same as Figure 4 except displayed in the original data domain.*

4.6 Estimation of Dependency in the Plane

One indexing application is to rank order images according to the degree of feature dependence. For example, if two features X and Y are horizontal and vertical changes over local neighborhoods of pixels one can search for evidence of anisotropy by evaluating a measure of statistical dependence of X and Y. One possible measure is the mutual α-information

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

This quantity converges to the standard Shannon mutual information in the limit as $\alpha \to 1$ and is equal to zero if and only if X, Y are independent.

Assume that $f(X,Y)$ is such that $f^\alpha(X,Y)$ is in the the Besov space $B^1_{p,1}(\mathbb{R}^2)$. Also assume that $p > 2$ and $\alpha = 1/2$. If one could find minimax density plug-in estimates based on independent estimates of $f^\alpha(X,Y)$, $f^{1-\alpha}(X)$ and $f^{1-\alpha}(Y)$, e.g. each based on separate segements of the data sample, this would result in overall rms convergence rate no less than $O(n^{-1/4})$. Note that the plug-in estimation procedure requires a.e. positive estimates of $f^{1-\alpha}(X)$ and $f^{1-\alpha}(Y)$ or at least these marginal density estimates should dominate the joint estimate of $f^\alpha(X,Y)$.

A hybrid method of estimation of the mutual α-information which has faster rms convergence rate is the following. Step 1: generate estimates $\hat{f}(X)$ and $\hat{f}(Y)$ of the one-dimensional functions using plug-in density estimation applied to X and Y components. Such estimates will converge a.s. with rms rates $O(n^{-1/3})$. Step 2: apply the separable measure transformation $dxdy \to \hat{f}(x)\hat{f}(y)$ to the plane, as described in the previous section, and implement the MST estimator on the the (X,Y) realization imbedded into the transformed coordinates. The resultant estimator will converge a.s. to the mutual α-information with rms convergence rate bounded above by the smaller $O(n^{-\alpha(3-2/p)/(\alpha(3-2/p)+1)})$. This procedure easily generalizes to estimating the α-mutual information of a d dimensional sample $(X, Y, \ldots, Z)$ for $\alpha \in [1/2, (2d-1)/(2d)]$. In particular, if the joint density is such that $f^\alpha(X,Y) \in B^1_{p,1}(\mathbb{R}^{2d})$, for $p > 2d$, and $\alpha \in [1/2, (2d-1)/(2d)]$ then the rms convergence rates of the plug-in estimator is no less than $O(n^{-1/(2+2d)})$ and that of the hybrid estimator is no more than $O(n^{-\alpha(2d+1-2d/p)/(\alpha(2d+1-2d/p)+1)})$. For an application of multi-dimensional α-mutual information estimation to image registration see Neemuchwalla and *etal* [29].

The one dimensional density estimation step in the hybrid procedure can be circumvented by considering a related measure to the mutual α-information: the mutual α-entropy difference

$$\begin{aligned}\Delta_0 H_\alpha(X,Y) &= H_\alpha(X,Y) - H_\alpha(X) - H_\alpha(Y)\\ &= \frac{1}{1-\alpha}\ln\frac{\int f^\alpha(X,Y)dX\,dY}{\int f^\alpha(X)dX\int f^\alpha(Y)dY}\end{aligned} \tag{19}$$

which also converges to the standard Shannon mutual information in the limit as $\alpha \to 1$. Given an i.i.d. sample $\{(X_i,Y_i)\}_{i=1}^n$ the length of the MST $L_\gamma(\{(X_i,Y_i)\}_{i=1}^n)\}_{i=1}^n)/n^\alpha$ converges w.p.1 to the numerator of (19) times the scale factor $\beta_{L_\gamma,d}$, $\alpha = (d-\gamma)/d$. Furthermore, let $\{\pi(i)\}_{i=1}^n$ be a permutation function, selected at random. Then $L_\gamma(\{(X_{\pi(i)}, Y_{\pi(i)}\}_{i=1}^n))/n^\alpha$ converges w.p.1 to the denominator of (19) times the same scale factor. It can be concluded that a consistent estimator of $\Delta_0 H_\alpha(X,Y)$ is given by the ratio

$$\widehat{\Delta_0} H_\alpha(X,Y) = \frac{1}{1-\alpha}\ln\frac{L_\gamma(\{(X_i,Y_i)\}_{i=1}^n)}{L_\gamma(\{X_{\pi(i)}, Y_{\pi(i)}\}_{i=1}^n)}$$

which does not depend on the factor $\beta_{L_\gamma,d}$. By comparing this statistic to a threshold we obtain a simple test for dependence of two random variables X, Y based on n i.i.d. observations. To reduce bias for finite n it is suggested that the denominator $L_\gamma^\pi \stackrel{\text{def}}{=} L_\gamma(\{X_{\pi(i)}, Y_{\pi(i)}\}_{i=1}^n)$ be replaced by a sample average $\overline{L_\gamma^\pi} = 1/|\Pi|\sum_{\pi\in\Pi} L_\gamma^\pi$ where Π is a set of randomly selected permutation functions π.

5 Application to Geo-Registration

It is desired to register two images taken on different sensor planes by potentially different sensor modalities for geo-registration applications. Our objective is to register two types of images — a set of electro-optical(EO) images and a terrain height map. For this multisensor image registration problem, there usually exists distortions between the two types of images. The distortions are due to difference acquisition conditions of the images such as shadowing, diffraction, terrain changes over time, clouds blocking the illumination sources, seasonal variations, etc. Existence of such differences between the images to be registered requires that the registration algorithms to be robust to noise and other small perturbations in intensity values. Here we describe an application of minimal graph entropy estimation to a feature set which is the set of gray level pixels.

For this image registration problem the set of EO images are generated from the *a priori* digital elevation model (DEM)[2] of a terrain patch (the terrain height map) at different look angles (determined by the sensor's location) and with different lighting positions. With different sensor and light locations, we can simulate the distortions mentioned above. For example, shadows are generated by taking into account both the sensor location and the lighting location as follows. The scene is first rendered using the lighting source as the viewing location. Depth values (distance from the light source) are generated for all pixels in the scene and stored in a depth buffer. Next, the scene is rendered using the sensor's location as the viewpoint. Before drawing each pixel, its depth value as measured from the sensor is compared to the transformed depth value as measured from the light source. This comparison determines if a particular pixel is illuminated by the source. Shadows are placed on those pixels that fail this comparison.

Geo-registration of a EO reference image to DEM's in an image database is accomplished by selecting a candidate DEM image from the database and projecting it into the EO image plane of the reference image. The objective is to find the correct viewing angle such that the corresponding EO image is the best match to the EO reference image. Figure 6 shows an DEM projected into the EO image plane with viewing angles (290, -20, 130) and the reference EO image. Clearly they are not aligned.

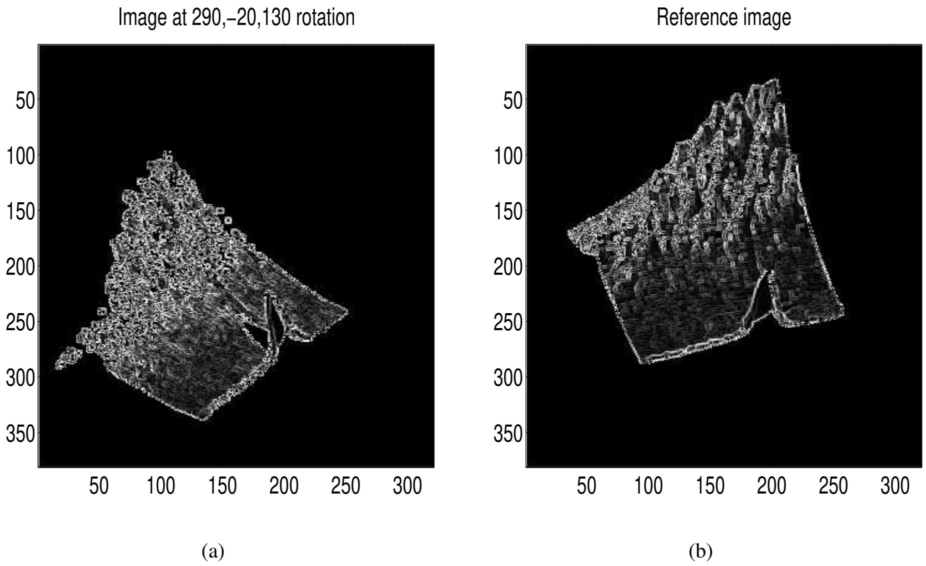
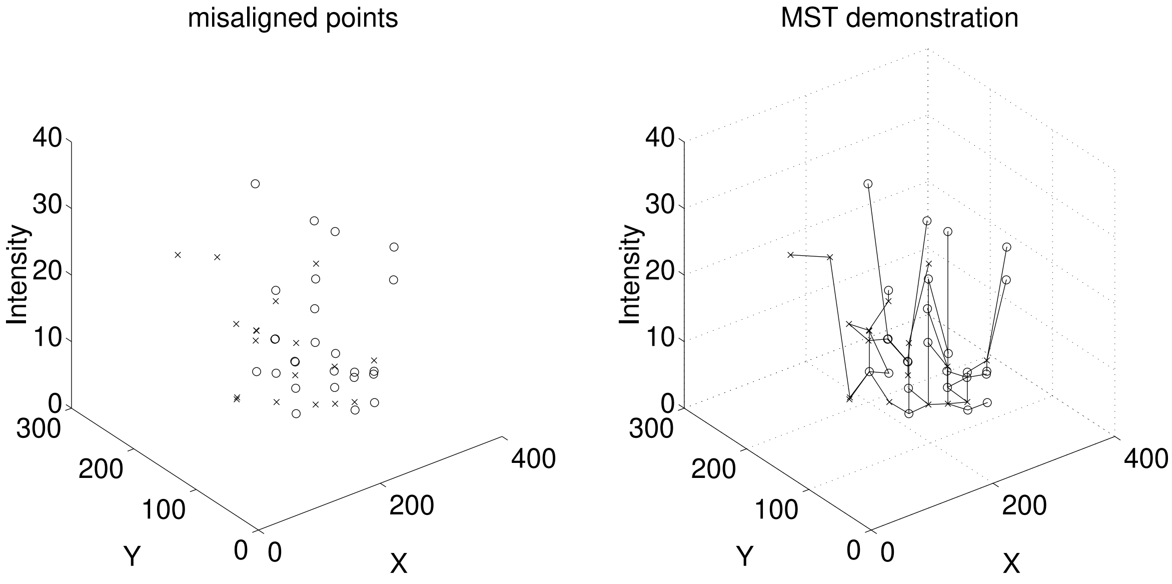


Figure 6: Misaligned EO and reference images

[2]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.

For matching criterion we use the α-Jensen difference, with α chosen arbitrarily as 0.5, applied to grey level features extracted from the reference images and candidate EO images derived from the DEM database. 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 grey level at planar position x, y. 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 [25]. 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.

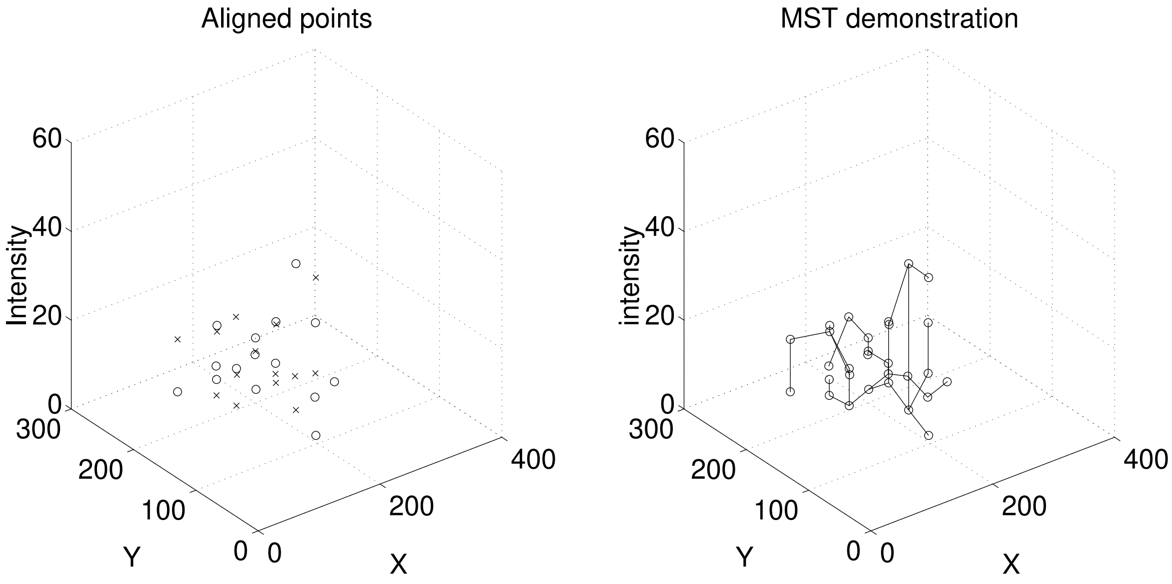
Figure 7 illustrates the MST-based registration procedure over the union of the reference and candidate image features for misaligned images, while Figure 8 shows the same for aligned images. In both Figures 7(a) and 8(a), circle points denote the pixels from Image X_1 and cross points denote the pixels from Image X_0. From Figures 7(a) and 8(a) we see that for misaligned images, the representation points have larger distances than those for aligned images. Therefore the corresponding MST for the misaligned images has a longer length than that for the aligned images (Figures 7(b) and 8(b)).



(a) (b)

Figure 7: MST demonstration for misaligned images

We repeat this MST construction process over the union of reference features and features derived from each of the images in the DEM database. The MST length can then be plotted in Figure 9. The x-axis stands for the image index, which corresponds to the viewing angles from the aircraft. The minimum of MST length indicates the best matching of the EO image and the reference image, which corresponds to the registered pair in Figure 10.



(a) (b)

Figure 8: MST demonstration for aligned images

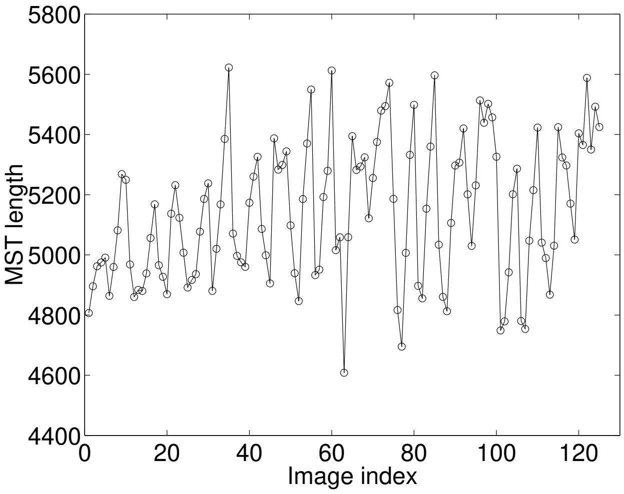
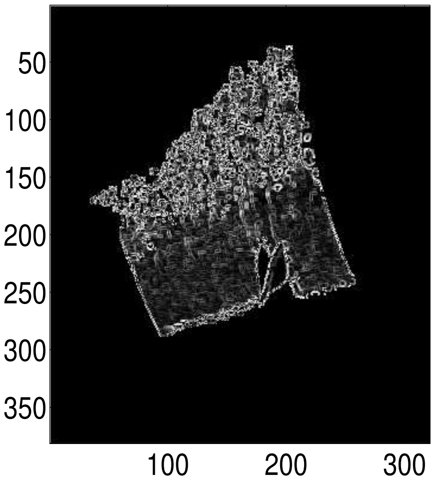
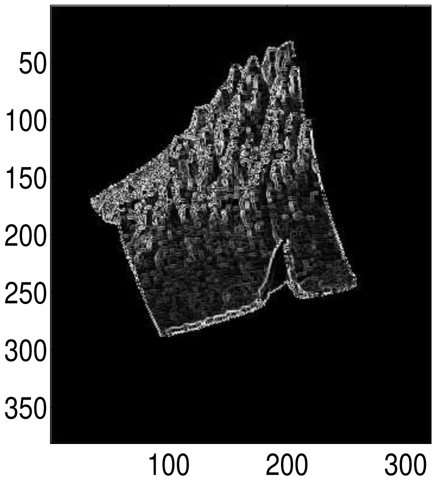


Figure 9: MST length for different test-reference image pairs



(a) Matching image



(b) Reference image

Figure 10: Co-registered EO-terrain maps

6 Conclusion

In this report we have discussed and compared α-divergence and α-entropy estimation techniques using minimal graph estimation and density plug-in methods. We have also considered the α-Jensen difference for performing indexing and image retrieval. We have investigated the estimation of α-Jensen difference using density plug-in estimators and the MST minimal graph method. We demonstrated theoretical advantages of the latter method for indexing planar features or higher dimensional features with feature densities belonging to Besov spaces.

Appendix A

Proposition 3 *Let* $f_{\frac{1}{2}} \stackrel{\text{def}}{=} \frac{1}{2}(f_0 + f_1)$*. The following local representation of the fractional Rényi entropy of a convex mixture* $\beta f_0 + (1-\beta) f_1$ *holds for all* $\alpha, \beta \in [0,1]$*:*

$$
\begin{aligned}
&H_\alpha(\beta f_0 + (1-\beta) f_1) \\
&= H_\alpha(f_{\frac{1}{2}}) + \frac{\alpha}{1-\alpha}\left(\beta - \frac{1}{2}\right) \frac{\int f_{\frac{1}{2}}^\alpha(x) \left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right) dx}{\int f_{\frac{1}{2}}^\alpha(x)\, dx} + \frac{\alpha}{2}\left(\frac{2\beta-1}{2}\right)^2 \frac{\int \left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 f_{\frac{1}{2}}^\alpha(x) dx}{\int f_{\frac{1}{2}}^\alpha(x) dx} \\
&\quad - \frac{\alpha^2}{2(1-\alpha)}\left(\frac{2\beta-1}{2}\right)^2 \left(\frac{\int \frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)} f_{\frac{1}{2}}^\alpha(x) dx}{\int f_{\frac{1}{2}}^\alpha(x) dx}\right)^2 + o(\Delta^2) \qquad (20)
\end{aligned}
$$

where $\Delta \in [0,1]$ *is*

$$
\Delta \stackrel{\text{def}}{=} \sup_x \frac{|f_0(x) - f_1(x)|}{f_{\frac{1}{2}}(x)}. \qquad (21)
$$

proof

Let $f_{1-\beta}(x) = \beta f_0(x) + (1-\beta) f_1(x)$. It can be written as

$$\begin{aligned} f_{1-\beta}(x) &= \frac{1}{2}[f_0(x) + (1-\beta)(f_1(x) - f_0(x))] + \frac{1}{2}[f_0(x) + \beta(f_0(x) - f_1(x))] \\ &= f_{\frac{1}{2}}(x) + \frac{1}{2}(2\beta - 1)(f_0(x) - f_1(x)) \\ &= f_{\frac{1}{2}}(x)\left(1 + \frac{(2\beta-1)\triangle_x}{2f_{\frac{1}{2}}(x)}\right) \end{aligned}$$

where $\Delta_x = (f_0(x) - f_1(x))/f_{\frac{1}{2}}(x)$. A Taylor series expansion of $f^{\alpha}_{1-\beta}(x)$ yields

$$\begin{aligned} f^{\alpha}_{1-\beta}(x) &= f^{\alpha}_{\frac{1}{2}}(x)\left(1 + \frac{2\beta-1}{2}\triangle_x\right)^{\alpha} \\ &= f^{\alpha}_{\frac{1}{2}}(x)\left(1 + \frac{\alpha(2\beta-1)}{2}\triangle_x + \frac{\alpha(\alpha-1)}{2}\left(\frac{2\beta-1}{2}\triangle_x\right)^2 + o(\triangle_x^2)\right) \end{aligned} \tag{22}$$

Taking the logarithm of both sides of (22) and dividing by $1-\alpha$

$$\begin{aligned} &\frac{1}{1-\alpha}\ln\int f^{\alpha}_{1-\beta}(x)dx \\ =\ & \frac{1}{1-\alpha}\ln\int\left[f^{\alpha}_{\frac{1}{2}}(x) + \alpha f^{\alpha}_{\frac{1}{2}}(x)\left(\frac{2\beta-1}{2}\triangle_x\right)\right. \\ &\qquad\left. + \frac{\alpha(\alpha-1)}{2}f^{\alpha}_{\frac{1}{2}}(x)\left(\frac{2\beta-1}{2}\triangle_x\right)^2 + f^{\alpha}_{\frac{1}{2}}(x)o(\triangle_x^2)\right]dx \\ =\ & \frac{1}{1-\alpha}\ln\left\{\int f^{\alpha}_{\frac{1}{2}}(x)dx\left[1 + \frac{\alpha\int f^{\alpha}_{\frac{1}{2}}(x)\left(\frac{2\beta-1}{2}\triangle_x\right)dx}{\int f^{\alpha}_{\frac{1}{2}}(x)dx}\right.\right. \\ &\qquad\left.\left. + \frac{\frac{\alpha(\alpha-1)}{2}\int f^{\alpha}_{\frac{1}{2}}(x)\left(\frac{2\beta-1}{2}\triangle_x\right)^2dx}{\int f^{\alpha}_{\frac{1}{2}}(x)dx} + o(\triangle^2)\right]\right\} \\ =\ & \frac{1}{1-\alpha}\ln\int f^{\alpha}_{\frac{1}{2}}(x)dx + \frac{1}{1-\alpha}\ln\left[1 + \frac{\alpha\int f^{\alpha}_{\frac{1}{2}}(x)\left(\frac{2\beta-1}{2}\triangle_x\right)dx}{\int f^{\alpha}_{\frac{1}{2}}(x)dx}\right. \\ &\qquad\left. + \frac{\frac{\alpha(\alpha-1)}{2}\int f^{\alpha}_{\frac{1}{2}}(x)\left(\frac{2\beta-1}{2}\triangle_x\right)dx}{\int f^{\alpha}_{\frac{1}{2}}(x)dx} + o(\triangle^2)\right]. \end{aligned}$$

Since $\ln(1+x) = x - \frac{x^2}{2} + o(x^2)$, we have

$$\begin{aligned} H_{\alpha}(\beta f_0 + (1-\beta)f_1) &= \frac{1}{1-\alpha}\ln\int f^{\alpha}_{1-\beta}(x)dx \\ &= H_{\alpha}(f_{\frac{1}{2}}) + \frac{\alpha}{1-\alpha}\frac{2\beta-1}{2}\frac{\int\triangle_x f^{\alpha}_{\frac{1}{2}}(x)dx}{\int f^{\alpha}_{\frac{1}{2}}(x)dx} + \frac{\alpha}{2}\left(\frac{2\beta-1}{2}\right)^2\frac{\int\triangle_x^2 f^{\alpha}_{\frac{1}{2}}(x)dx}{\int f^{\alpha}_{\frac{1}{2}}(x)dx} \end{aligned}$$

$$-\frac{\alpha^2}{2(1-\alpha)}\left(\frac{2\beta-1}{2}\right)^2\frac{\left(\int\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx\right)^2}{\left(\int f_{\frac{1}{2}}^{\alpha}(x)dx\right)^2}+o(\triangle^2). \tag{23}$$

This completes the proof of Proposition 3. □

Proposition 4 *The following asymptotic representation of the fractional Jensen difference of two densities f_0 and f_1 holds for all $\alpha,\beta\in[0,1]$:*

$$\begin{aligned}
&\triangle H_\alpha(\beta,f_0,f_1)\\
&= \frac{\alpha\beta(1-\beta)}{2}\left(\frac{\int\left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}+\frac{\alpha}{1-\alpha}\left(\frac{\int\left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\right)^2\right)\\
&\qquad +o(\triangle^2)
\end{aligned} \tag{24}$$

where $f_{\frac{1}{2}}$ and Δ are as defined in Proposition 3.

proof

Specializing to $\beta=0$ and $\beta=1$ in (20) of Proposition 3 we obtain

$$\begin{aligned}
H_\alpha(f_0) &= H_\alpha(f_{\frac{1}{2}})+\frac{\alpha}{1-\alpha}\frac{\int\frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}+\frac{\alpha}{2}\frac{\int\left(\frac{1}{2}\triangle_x\right)^2 f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\\
&\quad -\frac{\alpha^2}{2(1-\alpha)}\left(\frac{\int\frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\right)^2+o(\triangle^2) &(25)\\
H_\alpha(f_1) &= H_\alpha(f_{\frac{1}{2}})-\frac{\alpha}{1-\alpha}\frac{\int\frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}+\frac{\alpha}{2}\frac{\int\left(\frac{1}{2}\triangle_x\right)^2 f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\\
&\quad -\frac{\alpha^2}{2(1-\alpha)}\left(\frac{\int\frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\right)^2+o(\triangle^2) &(26)
\end{aligned}$$

where $\Delta_x=(f_0(x)-f_1(x))/f_{\frac{1}{2}}(x)$. Substituting (20), (25) and (26) into (13), we obtain the expression (24) for the Jensen difference. This completes the proof of Proposition 4 □

Proposition 5 *The α-divergence of fractional order $\alpha\in(0,1)$ between two densities f_0 and f_1 has the local representation*

$$D_\alpha(f_0\|f_1)=\frac{\alpha}{4}\int f_{\frac{1}{2}}(x)\left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 dx+o(\triangle^2) \tag{27}$$

where $f_{\frac{1}{2}}$ and $\triangle$ are as defined in Proposition 3.

proof

Rewrite the density f_0 as

$$f_0(x) \quad = \quad \frac{1}{2}(f_0(x) + f_1(x)) + \frac{1}{2}(f_0(x) - f_1(x)) = f_{\frac{1}{2}}(x)(1 + \frac{1}{2}\triangle_x), \tag{28}$$

where $\triangle_x = (f_0(x) - f_1(x))/f_{\frac{1}{2}}(x)$. Similarly,

$$f_1(x) \quad = \quad f_{\frac{1}{2}}(x)(1 - \frac{1}{2}\triangle_x). \tag{29}$$

Thus, by Taylor series expansion, we have

$$\begin{aligned} f_1^\alpha(x) &= f_{\frac{1}{2}}^\alpha(x) - \alpha f_{\frac{1}{2}}^\alpha(x)\left(\frac{\triangle_x}{2}\right) + \frac{\alpha(\alpha-1)}{2} f_{\frac{1}{2}}^\alpha(x)\left(\frac{\triangle_x}{2}\right)^2 + o(\triangle_x^2) \\ f_0^{1-\alpha}(x) &= f_{\frac{1}{2}}^{1-\alpha}(x) + (1-\alpha) f_{\frac{1}{2}}^{1-\alpha}(x)\left(\frac{\triangle_x}{2}\right) + \frac{\alpha(1-\alpha)}{2} f_{\frac{1}{2}}^{1-\alpha}(x)\left(\frac{\triangle_x}{2}\right)^2 + o(\triangle_x^2). \end{aligned}$$

Therefore

$$f_1^\alpha(x) f_0^{1-\alpha}(x) \quad = \quad f_{\frac{1}{2}}(x)\left[1 - (2\alpha-1)\frac{\triangle_x}{2} - \alpha(1-\alpha)\left(\frac{\triangle_x}{2}\right)^2 + o(\triangle_x^3)\right]$$

and

$$\begin{aligned} &D_\alpha(f_0\|f_1) = \frac{1}{\alpha-1}\ln\int f_1^\alpha(x) f_0^{1-\alpha}(x)dx \\ &= \frac{1}{\alpha-1}\ln\int\left(f_{\frac{1}{2}}(x) - (2\alpha-1)f_{\frac{1}{2}}(x)\frac{\triangle_x}{2} - \alpha(1-\alpha)f_{\frac{1}{2}}(x)\left(\frac{\triangle_x}{2}\right)^2 + f_{\frac{1}{2}}(x)o(\triangle_x^2)\right)dx \\ &= \frac{1}{\alpha-1}\ln\left(1 - \alpha(1-\alpha)\int f_{\frac{1}{2}}(x)\left(\frac{\triangle_x}{2}\right)^2 dx + o(\Delta^2)\right) \\ &= \alpha\int f_{\frac{1}{2}}(x)\left(\frac{\triangle_x}{2}\right)^2 dx + o(\Delta^2) \\ &= \frac{\alpha}{4}\int f_{\frac{1}{2}}(x)\left(\frac{f_0(x) - f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 dx + o(\Delta^2). \end{aligned}$$

This completes the proof of Prop. 5. □

Appendix B

Proposition 6 *Let $f_1(x) = f(x;\mu_1,\Lambda_1)$ and $f_0(x) = f(x;\mu_0,\Lambda_0)$ be multivariate d-dimensional Gaussian densities with vector means μ_1, μ_0 and positive definite covariance matrices Λ_1, Λ_0. The Rényi divergence of order α between f_1 and f_0 is*

$$D_\alpha(f_1\|f_0) \quad = \quad \frac{1/2}{\alpha-1}\ln\frac{|\Lambda_0|^\alpha|\Lambda_1|^{1-\alpha}}{|\alpha\Lambda_0 + (1-\alpha)\Lambda_1|} + \frac{\alpha}{2}\Delta\mu^T(\alpha\Lambda_0 + (1-\alpha)\Lambda_1)^{-1}\Delta\mu \tag{30}$$

where $\Delta\mu = \mu_1 - \mu_0$.

Proof

Start from the definition

$$D_\alpha(f_1\|f_0) \quad = \quad \frac{1}{\alpha-1}\ln\int f^\alpha(x;\mu_1,\Lambda_1)f^{1-\alpha}(x;\mu_0,\Lambda_0)dx$$

and make a change of variable $y = \Lambda_0^{-\frac{1}{2}}(x-\mu_0)$ in the integral to obtain

$$\int f^\alpha(x;\mu_1,\Lambda_1)f^{1-\alpha}(x;\mu_0,\Lambda_0)dx \quad = \quad |\Lambda_0|^{\frac{1}{2}}\int f^\alpha(y;\Lambda_0^{-\frac{1}{2}}\Delta\mu,\Lambda_0^{-\frac{1}{2}}\Lambda_1\Lambda_0^{-\frac{1}{2}})f^{1-\alpha}(y;0,I_d)dy, \tag{31}$$

where I_d is the $d\times d$ identity matrix.

By completion of the square and elementary matrix manipulations it is straightforward to show that for any d-element vector m and positive definite $d\times d$ covariance matrix A

$$\int f^\alpha(y;m,A)f^{1-\alpha}(y;0,I_d)dy$$
$$= \left(\frac{|A|^{1-\alpha}}{|(1-\alpha)A+\alpha I_d|}\right)^{\frac{1}{2}}\exp\left(-\tfrac{1}{2}\alpha(1-\alpha)\,m^T[(1-\alpha)A+\alpha I_d]^{-1}m\right)$$

Finally, identifying $m = \Lambda_0^{-\frac{1}{2}}\Delta\mu$ and $\Lambda = \Lambda_0^{-\frac{1}{2}}\Lambda_1\Lambda_0^{-\frac{1}{2}}$, substitution of the above into (31) and performing some matrix algebra we obtain (30).

This completes the proof of Prop. 6. □

References

[1] I. Ahmad and P.-E. Lin, “A nonparametric estimation of the entropy for absolutely continuous distributions,” *IEEE Trans. on Inform. Theory*, vol. IT-22, pp. 664–668, 1976.

[2] M. Basseville, “Distance measures for signal processing and pattern recognition,” *Signal Processing*, vol. 18, pp. 349–369, 1989.

[3] J. Beardwood, J. H. Halton, and J. M. Hammersley, “The shortest path through many points,” *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

[4] 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.

[5] R. J. Beran, “Minimum Hellinger distance estimates for parametric models,” *Annals of Statistics*, vol. 5, pp. 445–463, 1977.

[6] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1987.

[7] I. Csiszár, “Information-type measures of divergence of probability distributions and indirect observations,” *Studia Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.

[8] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Springer-Verlag, NY, 1998.

[9] M. N. Do and M. Vetterli, “Texture similarity measurement using Kullback-Liebler distance on wavelet subbands,” in *IEEE Int. Conf. on Image Processing*, pp. 367–370, Vancouver, BC, 2000.

[10] D. L. Donoho, "One-sided inference about functionals of a density," *Annals of Statistics*, vol. 16, pp. 1390–1420, 1988.

[11] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Inform. Theory*, vol. IT-28, pp. 373–380, 1979.

[12] L. Györfi and E. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Comput. Statist. Data Anal.*, vol. 5, pp. 425–436, 1987.

[13] P. Hall, "On powerful distributional tests based on sample spacings," *Journ. Multivar. Anal.*, vol. 19, pp. 201–225, 1986.

[14] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69–88, 1993.

[15] Y. He, A. Ben-Hamza, and H. Krim, "An information divergence measure for ISAR image registration," *Signal Processing*, Submitted, 2001.

[16] A. Hero, J. Costa, and B. Ma, "Convergence rates of minimal graphs with random vertices," *IEEE Trans. on Inform. Theory*, vol. submitted, , 2002. `http://www.eecs.umich.edu/~hero/det_est.html`.

[17] A. Hero, B. Ma, and O. Michel, "Convergence rates of random minimal graphs," *IEEE Trans. on Inform. Theory*, Submitted Aug. 2001.

[18] A. Hero and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 3459, pp. 250–261, San Diego, CA, July 1998.

[19] A. Hero and O. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," in *IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, June 1999.

[20] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

[21] R. Hoffman and A. K. Jain, "A test of randomness based on the minimal spanning tree," *Pattern Recognition Letters*, vol. 1, pp. 175–180, 1983.

[22] A. Jain, P. Moulin, M. I. Miller, and K. Ramchandran, "Performance bounds for atr based on compressed data," in *Proc. Army Workshop on Data Compression Techniques for Missile Guidance Data Links*, Huntsville, AL, Dec 1998.

[23] H. Joe, "On the estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 41, pp. 683–697, 1989.

[24] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[25] B. Ma, *Parametric and non-parametric approaches for multisensor data fusion*, PhD thesis, University of Michigan, Ann Arbor, MI 48109-2122, 2001. `http://www.eecs.umich.edu/~hero/research.html`.

[26] B. Ma, A. O. Hero, J. Gorman, and O. Michel, "Image registration with minimal spanning tree algorithm," in *IEEE Int. Conf. on Image Processing*, Vancouver, BC, October 2000.

[27] O. Michel, R. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measures," in *IEEE International time-frequency and Time-Scale Analysis Symposium*, pp. 64–67, Oct 1994.

[28] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Trans. on Inform. Theory*, vol. IT-35, no. 1, pp. 193–196, 1989.

[29] H. Neemuchwala, A. Hero, and P. Carson, "Feature coincidence trees for registration of ultrasound breast images," in *IEEE Int. Conf. on Image Processing*, Thesaloniki, Greece, October 2001.

[30] D. N. Neuhoff, "On the asymptotic distribution of the errors in vector quantization," *IEEE Trans. on Inform. Theory*, vol. IT-42, pp. 461–468, March 1996.

[31] J. P. Pluim, J. B. A. Maintz, and M. A. Viergever, "f-information measures in medical image registration," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, M. Sonka and K. M. Hanson, editors, volume 4322, pp. 579–587, 2001.

[32] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.

[33] H. Samet, *Applications of spatial data structures : computer graphics, image processing, and GIS*, Addison-Wesley, reading, MA, 1990.

[34] J. M. Steele, *Probability theory and combinatorial optimization*, volume 69 of *CBMF-NSF regional conferences in applied mathematics*, Society for Industrial and Applied Mathematics (SIAM), 1997.

[35] R. Stoica, J. Zerubia, and J. M. Francos, "Image retrieval and indexing: A hierarchical approach in computing the distance between textured images," in *IEEE Int. Conf. on Image Processing*, Chicago, Oct. 1998.

[36] 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, and Sig. Proc.*, Seattle, May 1998.

[37] N. Vasconcelos and A. Lippman, "A Bayesian framework for content-based indexing and retrieval," in *IEEE Data Compression Conference*, Snowbird, Utah, 1998. `http://nuno.www.media.mit.edu/people/nuno/`.

[38] N. Vasconcelos and A. Lippman, "Bayesian representations and learning mechanisms for content based image retrieval," in *SPIE Storage and Retrieval for Media Databases 2000*, San Jose, CA, 2000. `http://nuno.www.media.mit.edu/people/nuno/`.

[39] O. Vasicek, "A test for normality based on sample entropy," *J. Royal Statistical Society, Ser. B*, vol. 38, pp. 54–59, 1976.

[40] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proc. of 5th Int. Conf. on Computer Vision, MIT*, volume 1, pp. 16–23, 1995.

[41] J. E. Yukich, *Probability theory of classical Euclidean optimization*, volume 1675 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, 1998.