Alpha-Divergence for Image Indexing and Retrieval[1]

Alfred O. Hero, Bing Ma, and Olivier Michel

July 9, 2001

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 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/Battacharya distance 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. We analyze convergence rates and establish that the bias convergence rates of the MST entropy estimator can be better than that of an indirect estimator implemented with minimax adaptive kernel density estimation. We illustrate the MST estimator for geo-registration of images.

1 Indexing and Retrieval

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 [1, 2, 3, 4]. 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.

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;

[0]Alfred Hero heroeecs.umich.edu 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., Mountain View, CA. Olivier Michel omichelunice.fr is with the Department of Astrophysics, University of Nice, France.

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. The α-Jensen difference is a function of the joint α-entropy of pairs of feature vectors derived from the query and images in the database. We motivate the α-divergence for indexing by decision theoretic considerations and large deviation theory of detection and classification. A special case of α-divergence is the Kullback-Liebler (KL) divergence which has been applied to indexing and image retrieval [4, 5]. A result of this paper is that use of the KL divergence 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 optimal choice of α is $\alpha = 1/2$ which corresponds to the Hellinger-Battacharya distance as contrasted to the KL divergence.

The α-divergence must be estimated from the query and the database. When a parametric model for the feature densities exists the α-divergence is a non-linear function of these parameters and parametric estimation techniques such as maximum likelihood can be applied [3, 4]. On the other hand, when one uses the α-Jensen difference minimal graph estimation techniques can be directly applied. We analyze the convergence of non-parametric indirect methods based on minimax density estimation and direct methods based on graph entropy estimation via the MST. A comparison of the asymptotic rates of convergence of the bias of each one of these estimates suggests that the graph entropy methods will outperform the density estimation methods when the feature density are not smooth functions.

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 [6, 7, 8]

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

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-Battacharya distance squared, obtained when $\alpha = 1/2$,

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

and the Kullback-Liebler (KL) divergence [9], 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 $\alpha = 1/2$ does the divergence become 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 [10]. 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 [11]. 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 [12, 2] 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:

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

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 [13]:

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

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 $\sup_{\alpha\in[0,1]} D_\alpha^u(\overline{f}_1\|\overline{f}_0)$ 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 $D_\alpha^u(\overline{f}_1\|\overline{f}_0)$.

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

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-Battacharya distance, respectively [14]. 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-Battacharya 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 6 in Appendix B.

Proposition 1

$$D_\alpha^u(f_0\|f_1) \;=\; \frac{\alpha(1-\alpha)}{2}\int \frac{(f_0(x)-f_1(x))^2}{f_1+f_2}\,dx + o(\Delta^2), \tag{5}$$

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 $\sup_{\alpha\in[0,1]} D_\alpha^u(f_1\|f_0)$. A consequence of (5) is that to order Δ^2 the optimum value of α in the Chernoff exponent is $1/4$.

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 [3, 4] 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 $D_\alpha^u(f_1\|f_0) = D_\alpha^u(f(x;\mu_1,\Lambda_1)\|f(x;\mu_0,\Lambda_0))$ of order α is given by (see Proposition 7 in Appendix C).

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

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,

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

so that locally the Chernoff exponent increases in the trace norm of the differences between the feature covariances and, as expected, $\alpha = 1/4$ 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. Three general classes of methods can be identified: parametric estimators, non-parametric estimators based on density estimation, and non-parametric estimators based on direct estimation. The first two methods can be classified as *density plug-in* techniques where parametric or non-parametric density estimates 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 since an analytical form of the divergence can often be derived over the parametric class of densities considered for which maximum likelihood or other parameter estimates can be substituted into to the divergence formula. This approach was adopted under a multivariate Gaussian image model by Stoica *etal* [4] 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 (31) as $\alpha \to 1$, and the authors in [4] 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 the non-parametric plug-in technique is applied to estimate α-entropy it yields the estimate

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

where $\hat{f}$ is an empirical estimate of the density. 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 [15]; kernel density estimation plug-in [16]; and sample-spacing density estimator plug-in [17]. The reader is referred to [18] for a comprehensive overview of work in non-parametric estimation of Shannon entropy. The main difficulties with non-parametric 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 [19, 20, 21]. These methods are based on constructing minimal graphs spanning the feature vectors in the feature space. The overlall length of these minimal graphs can be used to construct a strongly consistent estimator of entropy for Lebesgue continuous densities. In particular, let $\mathcal{Z}^{(n)} = \{Z^{(1)}, \ldots, Z^{(n)}\}$ and define $L = L(\mathcal{Z}^{(n)}) = \min_{e\in\mathcal{E}} \sum_e |e|^\gamma$ 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 over random points $\mathcal{Z}^{(n)}$ as $n \to \infty$ has been studied for over half a decade [22, 23, 24] and, based on these studies, in [20] we gave conditions under which

$$\hat{H}_\alpha(\mathcal{Z}^{(n)}) = \ln L(\mathcal{Z}^{(n)})/n^\alpha - \ln \beta_1 \tag{8}$$

is an asymptotically unbiased and almost surely consistent estimator of the un-normalized α-entropy of f where $\alpha = (d-\gamma)/d$ and β_1 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. 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 [21].

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 is a bottleneck for a large number of feature samples.

When f_0 is known the α-divergence can be estimated by minimal graph methods using the measure transformation method outlined [21]. 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 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 (8) and indirect density plug-in entropy estimators (7) converge to zero as a function of the number n of i.i.d. observations [20, 16]. Here we attack a harder problem: comparing the asymptotic convergence rates of the bias. It is likely that this analysis can be extended to estimator variance and mean-square error but we do not explore this extension here.

Define the integral

$$I_\alpha(f) = \int f^\alpha(x)dx$$

and define the indirect and direct estimators of $I_\alpha(f)$ based on the n i.i.d. observations $\{Z_i\}$ with common marginal density f supported on $[0,1]^d$

$$\begin{aligned} I_\alpha(\hat{f}) &= \int \hat{f}^\alpha(x)dx && (9)\\ \hat{I}_\alpha &= L(Z_1,\ldots,Z_n)/n^{(d-\gamma)/d} && (10) \end{aligned}$$

where $(d-\gamma)/d = \alpha$. The asymptotic bias of the entropy estimators can be expressed in terms of the bias of the density integral estimators via a standard perturbation analysis

$$\hat{H}_\alpha - H_\alpha(f) = \frac{1}{1-\alpha}\,\frac{\hat{I}_\alpha - I_\alpha(f)}{I_\alpha(f)} + o(\Delta),$$

where $\Delta = |\hat{I}_\alpha - I_\alpha(f)|$. Thus as a function of n the asymptotic rate of convergence of the bias $E\hat{H}_\alpha - H_\alpha(f)$ of $\hat{H}_\alpha$ will be identical to that of the bias $E\hat{I}_\alpha - I_\alpha(f)$ of $\hat{I}_\alpha$. We will therefore focus on the latter integral estimators.

We first deal with the density plug-in estimator $I(\hat{f})$. Define the class $\Sigma_d(\kappa,c)$ of Hölder continuous density functions over $[0,1]^d$

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

where $p_x^k(z)$ is the Taylor polynomial (multinomial) of f of order k expanded about the point x. As κ becomes large the class $\Sigma_d(\kappa,c)$ contains functions which are increasingly smooth. For example, $\Sigma_d(0,c)$ contains all uniformly bounded functions with bound c and $\Sigma_d(\infty,c)$ contains all infinitely differentiable functions.

For the density plug-in estimator (9) it makes sense to consider a minimax optimal density estimation strategy which minimizes the worst case estimator mean integrated square error (MISE) over the densities lying in $\Sigma_d(\kappa,c)$ [25]. The minimax estimator can be implemented as an adaptive kernel density estimator, e.g. a piecewise multinomial with bin size that decreases in n at a specified optimal rate. The resultant MISE has the fastest possible rate of convergence over all $\Sigma_d(\kappa,c)$ and the rates of convergence of the squared bias and the variance of the density estimate are identical. The following proposition is established in Appendix A.

Proposition 2 *Assume that the Lebesgue density f is in the Hölder class $\Sigma_d(\kappa, c)$ and that $\int f^{\alpha-1}(z)dz < \infty$. Then, if $\hat{f}$ is a minimax MISE density estimator*

$$\sup_{f \in \Sigma_d(\kappa,c)} \left| E[I_\alpha(\hat{f})] - I_\alpha(f) \right| \; = \; n^{-\kappa/(2\kappa+d)} \; C_{\kappa,c}(1+o(1)) \tag{11}$$

where $C_{\kappa,c}$ is a constant depending on κ, c.

For the direct minimal-graph estimator (10) 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 $d = 2$ [24, 23]. it has been shown [23] 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 Appendix D we establish that the convergence rate of the MST length function is upper bounded by $n^{-1/(d+1)}$ for arbitrary density f satisfying certain bounded variation constraints. More specifically, the following follows directly from Proposition 9

Proposition 3 *Assume that the Lebesgue density f supported on $\mathcal{S} \subset [0,1]^d$ satisfies the property that f^ν is of bounded variation for $\nu = (d-\gamma)/d$, $\nu = (d-\gamma-1)/d$ and $\nu = \frac{1}{2} - \gamma/d$. Then for $d \geq 2$, $1 \leq \gamma < d$ and $\hat{I}_\alpha$ the estimator of $I_\alpha(f)$ based on the MST length function of order γ*

$$\left| E[\hat{I}_\alpha] - I_\alpha(f) \right| \; \leq \; n^{-1/(d+1)} \; K_{L,f}(1+o(1)) \tag{12}$$

where for $d > 2$ the rate constant $K_{L,f}$ is monotone increasing in $\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx$ and $\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx$. For $d = 2$, $K_{L,f}$ is also monotone increasing in $\int_{\mathcal{S}} f^{\frac{\frac{1}{2}-\gamma}{d}}(x)dx$.

As the proof of Proposition 9 depends on the rate of convergence of means result of Yukich [23, Thm. 5.2], which is restricted to the class of continuous quasi-additive functionals satisfying an "add-one bound," it is unknown whether a similar bound holds for graph estimators other than the MST. A faster rate of convergence bound, which is $O(n^{-1/d})$, is available for piecewise constant f, see Proposition 8 in Appendix D.

Observe that as compared to Proposition 2 the convergence rate in Proposition 3 does not depend on stringent Hölder continuity conditions of Proposition 2. Also note that the bounded variations assumptions imply that the rate constant $K_{L,f}$ is finite.On the other hand, while the rate in Proposition 2 is exact it is unknown whether the rate in Proposition 3 is optimal, i.e. if $\leq$ in (12) is actually equality but we believe it is true for $d = 2$. By slight modification of the proof of Proposition 9 in Appendix D, it can be shown that the greedy algorithm introduced in Hero and Michel [20] attains this rate of convergence if the cell size m is selected as the following function of n: $m = n^{1/[d(d+1)]}$.

A comparison between the convergence rate (11) and the convergence rate bound (12) indicates that the direct estimator has bias which converges with faster asymptotic rate in n when:

$$\kappa < \frac{d}{d-1} \tag{13}$$

Thus, for arbitrary $d \geq 2$ the bias of direct estimator $\hat{I}_\alpha$ converges faster for $\kappa < 1$, i.e. when f can be a non-smooth function satisfying the bounded variation constraints of Proposition 3.

4.1.1 Extensions to Estimator Mean Square Error Rates

Using concentration inequalities and Borel-Cantelli in a similar manner as in [23, Sec. 6.2] it is likely that the asymptotic convergence rate of the MSE of the MST estimator $\hat{I}_\alpha$ of $I(f)$ is identical to the convergence rate of the bias. It is conjectured that, at least for $d = 2$, this is also the minimax rate of MSE convergence of the MST estimator over the class of density functions satisfying the assumptions of Proposition 3. Since the minimax density estimator $\hat{f}$ achieves identical convergence rates for squared bias and variance, we also conjecture that the density plug-in estimator $I(\hat{f})$ has minimax MSE convergence rate identical to that of the bias. If these conjectures are true, our conclusion (13) will extend to the MSE performance of the MST and plug-in estimators.

4.2 α-Jensen Difference Index Function

We here propose an alternative index function based on the Jensen entropy difference. 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 [8]:

$$\triangle H_\alpha(\beta, f_0, f_1) \stackrel{\triangle}{=} H_\alpha(\beta f_0 + (1-\beta)f_1) - [\beta H_\alpha(f_0) + (1-\beta)H_\alpha(f_1)], \quad \alpha \in (0,1). \tag{14}$$

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 $\mathcal{Z}_1 = \{Z_0^{(i)}\}_{i=1,\ldots,n_0}$ 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 (14) 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 (8) 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 $\{\triangle H_\alpha(\beta, f_0, f_i)\}_{i=1}^K$. The problem of registering a query image to a database of images which are generated by rigid transformations of a reference image is an important example of this simplifying situation.

4.3 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 5 and 6) obtained in Appendix B:

$$D_\alpha(f_0 \| f_1) = \frac{\alpha}{4} E_{f_{\frac{1}{2}}}\left[\left(\frac{f_0 - f_1}{f_{\frac{1}{2}}}\right)^2\right] + o(\Delta^2) \tag{15}$$

$$\triangle H_\alpha(\beta, f_0, f_1) = \frac{\alpha\beta(1-\beta)}{2}\left\{E_{\tilde{f}^\alpha_{\frac{1}{2}}}\left[\left(\frac{f_0 - f_1}{f_{\frac{1}{2}}}\right)^2\right] + \frac{\alpha}{1-\alpha}\left(E_{\tilde{f}^\alpha_{\frac{1}{2}}}\left[\frac{f_0 - f_1}{f_{\frac{1}{2}}}\right]\right)^2\right\} + o(\Delta^2) \tag{16}$$

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, and Δ_3 is a term that decreases in the difference $f_0 - f_1$..

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

- The divergence criterion $D_\alpha(f_0\|f_1)$ locally depends on α only through a scale factor, while the Jensen difference criterion $\triangle H_\alpha(\beta, f_0, f_1)$ is more strongly dependent on α.
  1. When α approaches 0, tail differences between the two densities f_0 and f_1 are much more influential on $\triangle H_\alpha(\beta, f_0, f_1)$ than on $D_\alpha(f_0\|f_1)$.
  2. 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 for locally optimum discrimination.
- The ratio, β, of the numbers of feature vectors extracted from the two to-be-registered images does not influence $D_\alpha(f_0\|f_1)$, while this ratio does affect $\triangle H_\alpha(\beta, f_0, f_1)$. Furthermore, $\triangle H_\alpha(\beta, f_0, f_1)$ has the maximal discriminative capability for $\beta = \frac{1}{2}$, i.e., when two images yield the same number of feature vectors.

4.4 Estimation of Dependency in the Plane

One indexing application is to rank order images accoring 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. A related measure is the mutual α-entropy difference

$$\begin{aligned} \Delta_0 H_\alpha(X,Y) &= H_\alpha(X,Y) - H_\alpha(X) - H_\alpha(Y) \\ &= \frac{1}{1-\alpha} \ln \frac{\int f^\alpha(X,Y) dX dY}{\int f^\alpha(X) dX \int f^\alpha(Y) dY} \end{aligned} \tag{17}$$

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 $L(\{(X_i, Y_i)\}_{i=1}^n)\}_{i=1}^n)/n^\alpha$ converges w.p.1 to the numerator of (17) times the scale factor β_1, $\alpha = (d-\gamma)/d$. Furthermore, let $\{\pi(i)\}_{i=1}^n$ be a permutation function, selected at random. Then $L(\{(X_{\pi(i)}, Y_{\pi(i)})/n^\alpha$ converges w.p.1 to the denominator of (17) 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

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

which does not depend on the factor β_1. 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 $L_\pi \stackrel{\text{def}}{=} L(\{X_{\pi(i)}, Y_{\pi(i)}\}_{i=1}^n)$ be replaced by a sample average $\overline{L_\pi} = 1/|\Pi| \sum_{\pi \in \Pi} L_\pi$ 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.

For this image registration problem the set of EO images are generated from the *a priori* digital elevation model (DEM)[1] 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 1 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.

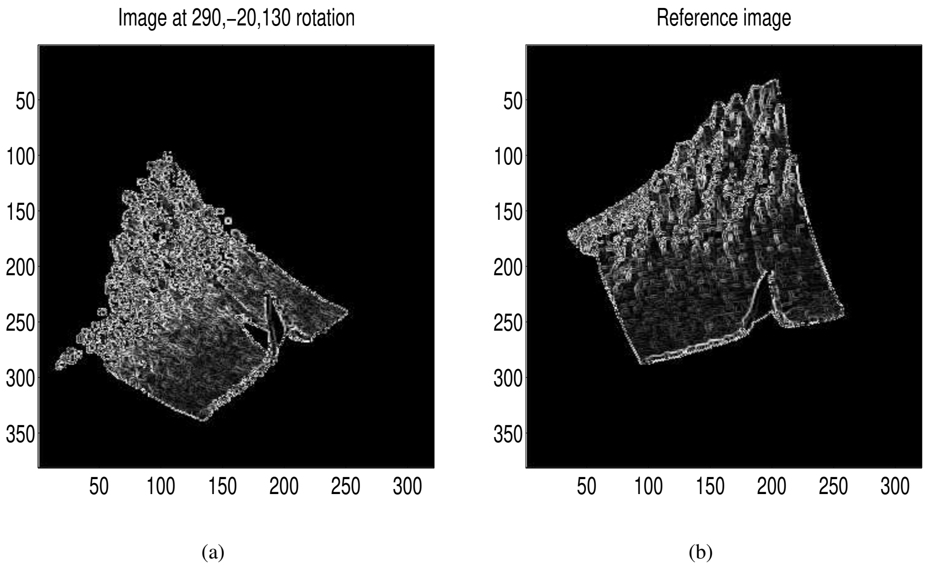
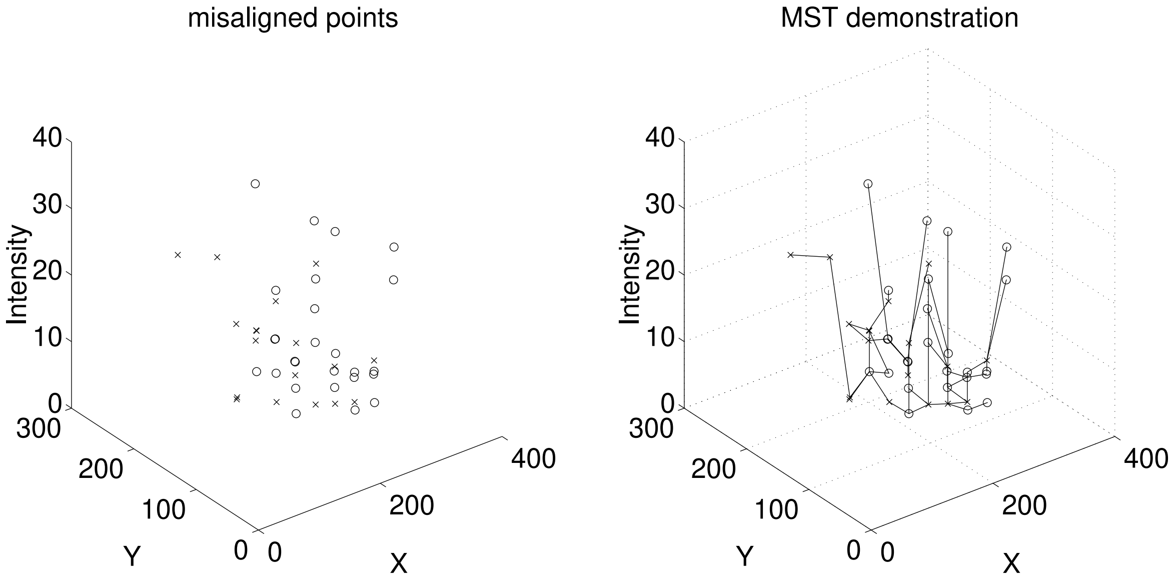


Figure 1: Misaligned EO and reference images

[1]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 [14]. 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.

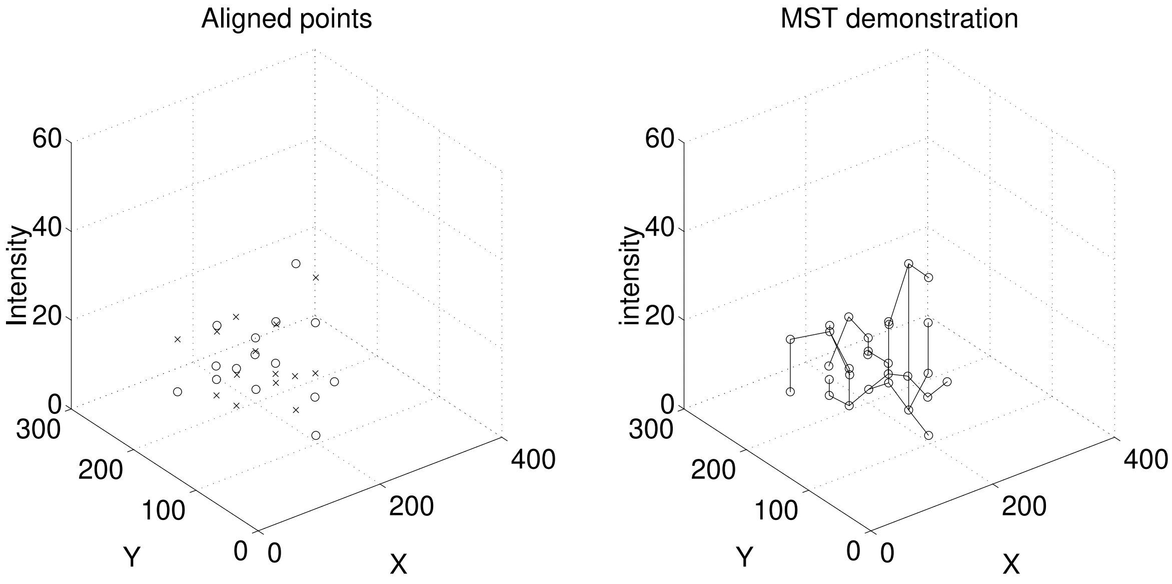
Figure 2 illustrates the MST-based registration procedure over the union of the reference and candidate image features for misaligned images, while Figure 3 shows the same for aligned images. In both Figures 2(a) and 3(a), circle points denote the pixels from Image X_1 and cross points denote the pixels from Image X_0. From Figures 2(a) and 3(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 2(b) and 3(b)).



(a) (b)

Figure 2: 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 4. 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 5.



(a)

(b)

Figure 3: MST demonstration for aligned images

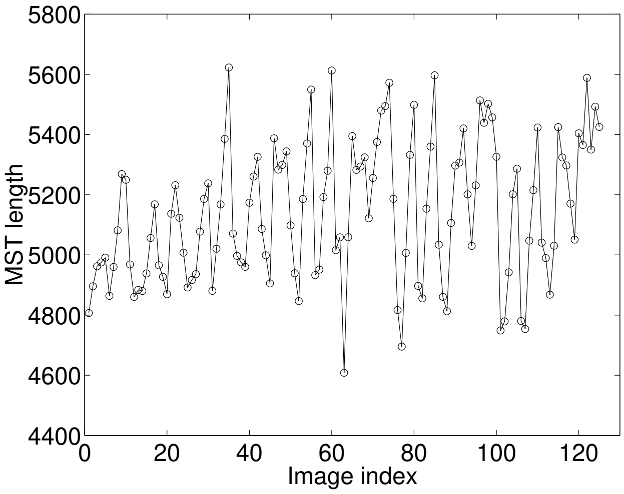
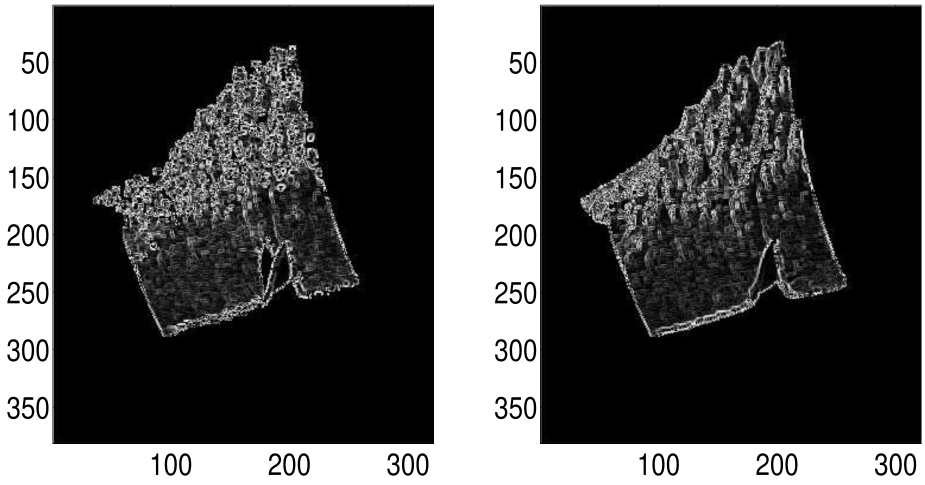


Figure 4: MST length for different test-reference image pairs



(a) Matching image

(b) Reference image

Figure 5: Co-registered EO-terrain maps

6 Conclusion

In this paper we have introduced a new criterion, the α-Jensen difference for performing indexing and image retrieval. This criterion was motivated by the α-divergence which characterizes the rate of decay of the optimal Bayes decision rule. 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 non-smooth feature densities. It will be important to compare these estimates to estimators of α-divergence which is more directly related to Bayes esimation performance and is less sensitive to α and β then the α-Jensen difference.

Appendix A

Here we establish Proposition 2.

First we recall relevant properties of minimax density estimation with respect to a mean integrated squared error (MISE) criterion [25]. As the class of kernel estimators known as the piecewise polynomial density estimators (PPE) are MISE minimax over the class $\Sigma_d(\kappa, c)$ we specialize to the case of a PPE $\hat{f}$ with optimal bin-width $\delta_n = an^{-1/(2\kappa+d)}$ for some positive constant a. The minimax MISE over Hölder class $\Sigma_d(\kappa, c)$

$$\sup_{f \in \Sigma_d(\kappa,c)} E\left[\int (\hat{f}(x) - f(x))^2 dx\right] = Cn^{-2\kappa/(2\kappa+d)}(1 + o(1)) \tag{18}$$

for a positive constant C. The squared integrated bias and the integrated variance of the minimax estimator $\hat{f}$ decay at identical minimax rates $n^{-2\kappa/(2\kappa+d)}$. Relation (18) implies that there exist positive constants C_1, C_2 such that for all $f \in \Sigma_d(\kappa, c)$

$$|\hat{f}(x) - f(x)| \leq C_1 n^{-\kappa/(2\kappa+d)}, \quad (w.p.1), \tag{19}$$

except possibly on a subset of $[0,1]^d$ of measure zero, and for some $f \in \Sigma_d(\kappa, c)$

$$|\hat{f}(x) - f(x)| \geq C_2 n^{-\kappa/(2\kappa+d)}, \quad (w.p.1) \tag{20}$$

over some subset of $[0,1]^d$ of positive measure.

We now turn to the plug-in estimator $I_\alpha(\hat{f})$ defined in (9). Under the assumption that $\int f^{\alpha-1}(x)dx < \infty$,

$$E[I_\alpha(\hat{f})] - I_\alpha(f) = \int (E[\hat{f}^\alpha(x)] - f^\alpha(x))dx \quad = \quad \alpha \int f^{\alpha-1}(x)(E[\hat{f}(x)] - f(x))dx + o(\Delta).$$

where $\Delta = E[\sup_x |\hat{f}(x) - f(x)|]$. Using relations (19) and (20), there exists a finite constant C_3 such that

$$\left| \int f^{\alpha-1}(x)(E[\hat{f}(x)] - f(x))dx \right| \leq C_3 n^{-\kappa/(2\kappa+d)},$$

for all $f \in \Sigma_d(\kappa, c)$. Furthermore, there exists a density $f \in \Sigma_d(\kappa, c)$ such that

$$\left| \int f^{\alpha-1}(x)(E[\hat{f}(x)] - f(x))dx \right| \geq C_4 n^{-\kappa/(2\kappa+d)}$$

Therefore,

$$C_4 n^{-\kappa/(2\kappa+d)} \leq \sup_{f \in \Sigma_d(\kappa,c)} \left| \int f^{\alpha-1}(x)(E[\hat{f}(x)] - f(x))dx \right| \leq C_3 n^{-\kappa/(2\kappa+d)}$$

which finishes the proof of Proposition 2. □

Appendix B

Proposition 4 *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 $\alpha, \beta \in [0,1]$:*

$$\begin{aligned} & H_\alpha(\beta f_0 + (1-\beta) f_1) \\ = \quad & H_\alpha(f_{\frac{1}{2}}) + \frac{\alpha}{1-\alpha}\left(\beta - \frac{1}{2}\right) \frac{\int f_{\frac{1}{2}}^\alpha(x) \left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right) dx}{\int f_{\frac{1}{2}}^\alpha(x)\, dx} + \frac{\alpha}{2}\left(\frac{2\beta-1}{2}\right)^2 \frac{\int \left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 f_{\frac{1}{2}}^\alpha(x) dx}{\int f_{\frac{1}{2}}^\alpha(x) dx} \\ & - \frac{\alpha^2}{2(1-\alpha)}\left(\frac{2\beta-1}{2}\right)^2 \left(\frac{\int \frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)} f_{\frac{1}{2}}^\alpha(x) dx}{\int f_{\frac{1}{2}}^\alpha(x) dx}\right)^2 + o(\Delta^2) \end{aligned} \tag{21}$$

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

$$\Delta \stackrel{\text{def}}{=} \sup_x \frac{|f_0(x) - f_1(x)|}{f_{\frac{1}{2}}(x)}. \tag{22}$$

proof

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

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

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

$$
\begin{aligned}
f^{\alpha}_{1-\beta}(x) &= f^{\alpha}_{\frac{1}{2}}(x)\left(1+\frac{2\beta-1}{2}\,\triangle_x\right)^{\alpha} \\
&= f^{\alpha}_{\frac{1}{2}}(x)\left(1+\frac{\alpha(2\beta-1)}{2}\triangle_x+\frac{\alpha(\alpha-1)}{2}\left(\frac{2\beta-1}{2}\triangle_x\right)^2+o(\triangle_x^2)\right)
\end{aligned}
\tag{23}
$$

Taking the logarithm of both sides of (23) 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

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

This completes the proof of Proposition 4. $\square$

Proposition 5 *The following asymptotic representation of the fractional Jensen difference of two densities f_0 and f_1 holds for all $\alpha, \beta \in [0,1]$:*

$\triangle H_{\alpha}(\beta, f_0, f_1)$

$$= \frac{\alpha\beta(1-\beta)}{2}\left(\frac{\int\left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx} + \frac{\alpha}{1-\alpha}\left(\frac{\int\left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right) f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\right)^2\right) + o(\triangle^2) \quad (25)$$

where $f_{\frac{1}{2}}$ and Δ are as defined in Proposition 4.

proof

Specializing to $\beta = 0$ and $\beta = 1$ in (21) of Proposition 4 we obtain

$$\begin{aligned} H_\alpha(f_0) &= H_\alpha(f_{\frac{1}{2}}) + \frac{\alpha}{1-\alpha}\frac{\int \frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx} + \frac{\alpha}{2}\frac{\int\left(\frac{1}{2}\triangle_x\right)^2 f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx} \\ &\quad - \frac{\alpha^2}{2(1-\alpha)}\left(\frac{\int \frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\right)^2 + o(\triangle^2) \qquad (26) \\ H_\alpha(f_1) &= H_\alpha(f_{\frac{1}{2}}) - \frac{\alpha}{1-\alpha}\frac{\int \frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx} + \frac{\alpha}{2}\frac{\int\left(\frac{1}{2}\triangle_x\right)^2 f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx} \\ &\quad - \frac{\alpha^2}{2(1-\alpha)}\left(\frac{\int \frac{1}{2}\triangle_x f_{\frac{1}{2}}^{\alpha}(x)dx}{\int f_{\frac{1}{2}}^{\alpha}(x)dx}\right)^2 + o(\triangle^2) \qquad (27) \end{aligned}$$

where $\Delta_x = (f_0(x) - f_1(x))/f_{\frac{1}{2}}(x)$. Substituting (21), (26) and (27) into (14), we obtain the expression (25) for the Jensen difference. This completes the proof of Proposition 5 ☐

Proposition 6 *The α-divergence of fractional order $\alpha \in (0,1)$ between two densities f_0 and f_1 has the local representation*

$$D_\alpha(f_0\|f_1) = \frac{\alpha}{4}\int f_{\frac{1}{2}}(x)\left(\frac{f_0(x)-f_1(x)}{f_{\frac{1}{2}}(x)}\right)^2 dx + o(\triangle^2) \quad (28)$$

where $f_{\frac{1}{2}}$ and $\triangle$ are as defined in Proposition 4.

proof

Rewrite the density f_0 as

$$f_0(x) = \frac{1}{2}(f_0(x) + f_1(x)) + \frac{1}{2}(f_0(x) - f_1(x)) = f_{\frac{1}{2}}(x)(1 + \frac{1}{2}\triangle_x), \quad (29)$$

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

$$f_1(x) = f_{\frac{1}{2}}(x)(1 - \frac{1}{2}\triangle_x). \quad (30)$$

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

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

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. 6. □

Appendix C

Proposition 7 *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*

$$D_\alpha(f_1\|f_0) = \frac{1/2}{\alpha-1}\ln\frac{|\Lambda_0|^\alpha|\Lambda_1|^{1-\alpha}}{|\alpha\Lambda_0 + (1-\alpha)\Lambda_1|} + \frac{\alpha}{2}\Delta\mu^T(\alpha\Lambda_0 + (1-\alpha)\Lambda_1)^{-1}\Delta\mu \tag{31}$$

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

Proof

Start from the definition

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

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

$$\int f^\alpha(x;\mu_1,\Lambda_1) f^{1-\alpha}(x;\mu_0,\Lambda_0)dx = |\Lambda_0|^{\frac{1}{2}}\int f^\alpha(y;\Lambda_0^{-\frac{1}{2}}\Delta\mu, \Lambda_0^{-\frac{1}{2}}\Lambda_1\Lambda_0^{-\frac{1}{2}}) f^{1-\alpha}(y;0,I_d)dy, \tag{32}$$

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 (32) and performing some matrix algebra we obtain (31).

This completes the proof of Prop. 7. □

Appendix D

Here we establish convergence rates for continuous quasi-additive power-weighted edge functionals which satisfy the "add-one bound" of Steele [24]. This class includes the minimal spanning tree (MST) through i.i.d. points $X_1, X_2, \ldots, X_n$ in $[0,1]^d$. We first provide background on quasi-additive power-weighted edge functionals. Let F be a finite subset of points in $[0,1]^d$, $d \geq 2$, and let L be a real-valued function defined on F of the form

$$L(F) = \min_{e \in \mathcal{E}} \sum_e |e(F)|^\gamma \tag{33}$$

where $\mathcal{E}$ is a suitably constrained set of graphs, e.g. spanning trees, over the points in F, e is an edge in the graph, $|e|$ is the euclidean length of e, and γ is the power weighting constant. We assume that $0 < \gamma < d$.

We state the following technical conditions for L to be a *quasi-additive power-weighted edge functional of order* γ [26, 23].

- *Null condition*: $L(\phi) = 0$, where ϕ is the null set.
- *Subadditivity*: There exists a constant C_1 with the following property: If $\mathcal{Q}^m = \{Q_i\}_{i=1}^{m^d}$ is a uniform partition of $[0,1]^d$ into m^d subcubes Q_i with edge parallel to the axes and with edge length m^{-1} and volume m^{-d} and if $\{q_i\}_{i=1}^{m^d}$ is the set of points in $[0,1]^d$ that translate each Q_i back to the origin such that $Q_i - q_i$ has the form $m^{-1}[0,1]^d$, then for every finite subset F of $[0,1]^d$,

$$L(F) \leq m^{-\gamma} \sum_{i=1}^{m^d} L\left(m[F \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \tag{34}$$

- *Superadditivity*: For the same conditions as above on Q_i, m, and q_i, there exists a constant C_2 with the following property:

$$L(F) \geq m^{-\gamma} \sum_{i=1}^{m^d} L\left(m[F \cap Q_i - q_i]\right) - C_2 m^{d-\gamma} \tag{35}$$

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

$$|L(F \cup G) - L(F)| \leq C_3 (\text{card}(G))^{(d-\gamma)/d} \tag{36}$$

  where card(G) is the cardinality of the subset G.

The functional L is said to be a *continuous subadditive functional* of order γ if it satisfies the null condition, sudadditivity and continuity. L is said to be a *continuous superadditive functional* of order γ if it satisfies the null condition, superadditivity and continuity.

For many continuous subadditive functionals L on $[0,1]^d$, e.g. the MST power weighted length, there exists a *dual* superadditive functional L^*. The dual functional satisfies two properties: 1) $L(F)+1 \geq L^*(F)$ for every finite subset F; and, 2) for i.i.d. uniform random vectors $U_1, \dots, U_n$ in $[0,1]^d$,

$$\mid E[L(U_1, \dots, U_n)] - E[L^*(U_1, \dots, U_n)] \mid \leq C_4 n^{(d-\gamma-1)/d} \tag{37}$$

with C_4 a finite constant.

A continuous subadditive functional L is said to be a *quasi-additive continuous functional* if L is continuous subadditive and there exists a continuous superadditive dual functional L^*. We point out that the dual L^* is not uniquely defined. It has been shown [27, 26] that the boundary-rooted version of L, namely, one where points may be connected to the boundary of the unit cube, usually has the requisite property (37) of the dual. These authors have displayed duals and shown continuous quasi-additivity for power weighted MST, Steiner tree, traveling salesman tour, and other length functionals.

In [23, 26] almost sure limits and convergence rates were obtained for continuous quasi-additive functionals $L(X_1, \dots, X_n)$ under the assumption of uniformly distributed points $X_1, \dots, X_n$ and an additional assumption that L satisfies the add-one bound

- *Add-one bound*:

$$\mid E[L(U_1, \dots, U_{n+1})] - E[L(U_1, \dots, U_n)] \mid \leq C_5 n^{-\gamma/d}. \tag{38}$$

This condition is satisfied by the MST length functional of order γ. Here we analyze convergence rates for general density $f(x)$. Our method of extension follows the standard practice [28, 24, 23]: we first establish convergence for piecewise constant densities and then extend to arbitrary densities.

D.1 Convergence Rates of MST Length Functionals for Block Densities

We will need the following result for the sequel.

Lemma 1 *Define $g_u(x) = x^u$, where $0 < u < 1$ and $x \geq 0$. Then*

$$g_u(x) \geq g_u(x_o) - g_u^{'}(x_o)|\Delta|$$

where $\Delta = x - x_o$ and $g_u^{'}(x) = dg_u(x)/dx$.

Proof

The Taylor expansion of g_u with remainder is

$$g_u(x) = g_u(x_o) + g_u^{'}(\zeta)\Delta$$

where ζ lies between x_o and x. The function $g_u^{'}(z) = uz^{u-1}$ is positive and monotone decreasing over $z \geq 0$. Therefore

$$g_u^{'}(\zeta)\Delta \geq \begin{cases} g_u^{'}(x)\Delta, & \Delta > 0 \\ g_u^{'}(x_o)\Delta, & \Delta < 0 \end{cases} .$$

This implies: $g_u^{'}(\zeta)\Delta \geq -g_u^{'}(x_o)|\Delta|^3$. $\square$

A density $f(x)$ over $\mathbf{R}^d$ is said to be a block density with m^d levels if for some probabilities $\{\phi_i\}_{i=1}^{m^d}$

$$f(x) = \sum_{i=1}^{m^d} \phi_i 1_{Q_i}(x)$$

where $\{Q_i\}_{i=1}^{m^d}$ is a uniform partition of the unit cube $[0,1]^d$ with partition cell edge lengths m^{-1}.

Proposition 8 *Let $X_1, \dots, X_n$ be i.i.d. sample points over $\mathbf{R}^d$ whose marginal is a block density $f(x)$ with m^d levels. Let $\mathcal{S}$ be the support of $f(x)$. Assume $d \geq 2$, $1 \leq \gamma < d$. Then for any continuous quasi-additive functional L of order γ which satisfies the add-one bound (38)*

$$\left| L(X_1, \dots, X_n)/n^{(d-\gamma)/d} - \beta_1 \int_{\mathcal{S}} f^{(d-\gamma)/d}(x)\, dx \right| \leq O(n^{-1/d}),$$

where β_1 is a constant independent of f. A more explicit form for the bound on the right hand side is

$$O(n^{-1/d}) = \left\{ \begin{array}{ll} \frac{K_1}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + o(n^{-1/d}), & d > 2 \\ \frac{1}{(nm^{-d})^{1/d}} \left[K_1 \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \beta_1 \frac{(d-\gamma)}{d} \int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx \right] + o(n^{-1/d}), & d = 2 \end{array} \right. .$$

Proof

Let n_i denote the number of samples $\{X_1, \dots, X_n\}$ falling into the partition cell Q_i. By subadditivity, we have

$$\begin{aligned} L(X_1, \dots, X_n) \qquad &\leq m^{-\gamma} \sum_{i=1}^{m^d} L\left(m[\{X_1, \dots, X_n\} \cap Q_i - q_i]\right) + C_1 m^{d-\gamma} \\ &= m^{-\gamma} \sum_{i=1}^{m^d} L(U_1, \dots, U_{n_i}) + C_1 m^{d-\gamma} \end{aligned}$$

since the samples in each partition cell Q_i are drawn independently from a conditionally uniform distribution. Note that n_i has a Binomial $B(n, \phi_i m^{-d})$ distribution.

Taking expectations on both sides of the above inequality,

$$EL(X_1, \dots, X_n) \qquad \leq m^{-\gamma} \sum_{i=1}^{m^d} E\left[EL(U_1, \dots, U_{n_i}) | \, n_i \right] + C_1 m^{d-\gamma}. \tag{39}$$

For uniform samples $U_1, \dots, U_n$ in $[0,1]^d$, the following rate of convergence for quasi-additive edge functionals L satisfying the add-one bound (38) has been established [23, Ch. 5],

$$|EL(U_1, \dots, U_n) - \beta_1 n^{\frac{d-\gamma}{d}}| \leq K_1 n^{\frac{d-1-\gamma}{d}}, \tag{40}$$

where K_1 is a function of C_1, C_3 and C_5.

Using the result (40) and subadditivity (39) on L

$$EL(X_1, \dots, X_n)$$

$$
\begin{aligned}
&\leq m^{-\gamma}\sum_{i=1}^{m^d} E\left[\beta_1 n_i^{\frac{d-\gamma}{d}} + K_1 n_i^{\frac{d-\gamma-1}{d}}\right] + C_1 m^{d-\gamma} \\
&= m^{-\gamma}\beta_1 n^{\frac{d-\gamma}{d}}\sum_{i=1}^{m^d} E\left(\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right) + \ m^{-\gamma}K_1 n^{\frac{d-\gamma-1}{d}}\sum_{i=1}^{m^d} E\left(\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right) + C_1 m^{d-\gamma}
\end{aligned}
\tag{41}
$$

and similarly for the dual L^* it follows by superadditivity

$$
\begin{aligned}
&EL^*(X_1,\ldots,X_n) \\
&\quad\geq m^{-\gamma}\beta_1 n^{\frac{d-\gamma}{d}}\sum_{i=1}^{m^d} E\left(\left(\frac{n_i}{n}\right)^{\frac{d-\gamma}{d}}\right) - \ m^{-\gamma}K_1 n^{\frac{d-\gamma-1}{d}}\sum_{i=1}^{m^d} E\left(\left(\frac{n_i}{n}\right)^{\frac{d-\gamma-1}{d}}\right) - C_2 m^{d-\gamma}
\end{aligned}
\tag{42}
$$

We next develop lower and upper bounds on the expected values in (41) and (42). From Lemma 1

$$
\begin{aligned}
\left(\frac{n_i}{n}\right)^u &\geq \left[p_i + \left(\frac{n_i}{n} - p_i\right)\right]^u \\
&= p_i^u - u p_i^{u-1}\left|\frac{n_i}{n} - p_i\right|,
\end{aligned}
\tag{43}
$$

where $p_i = \phi_i m^{-d}$. In order to bound the expectation of the above inequality we use the following bound

$$
E[\left|\frac{n_i}{n} - p_i\right|] \leq \sqrt{E[\left|\frac{n_i}{n} - p_i\right|^2]} = \frac{1}{\sqrt{n}}\sqrt{p_i(1-p_i)} \leq \frac{\sqrt{p_i}}{\sqrt{n}}.
$$

Therefore, from (43),

$$
E\left[\left(\frac{n_i}{n}\right)^u\right] \geq p_i^u - u p_i^{u-\frac{1}{2}}/\sqrt{n}.
\tag{44}
$$

By concavity, Jensen's inequality yields the upper bound

$$
E\left[\left(\frac{n_i}{n}\right)^u\right] \leq E\left[\left(\frac{n_i}{n}\right)\right]^u = p_i^u
\tag{45}
$$

The upper bound (45) with $u = (d-\gamma)/d$ and $u = (d-\gamma-1)/d$ can be substituted into expression (41) to obtain

$$
\begin{aligned}
&EL(X_1,\ldots,X_n)/n^{(d-\gamma)/d} \\
&\quad\leq \beta_1\sum_{i=1}^{m^d}\phi_i^{\frac{d-\gamma}{d}} m^{-d} + \frac{K_1}{(nm^{-d})^{1/d}}\sum_{i=1}^{m^d}\phi_i^{\frac{d-\gamma-1}{d}} m^{-d} + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}} \\
&\quad= \beta_1\int_{\mathcal{S}} f^{(d-\gamma)/d}(x)dx + \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{(d-\gamma-1)/d}(x)dx + \frac{C_1}{(nm^{-d})^{(d-\gamma)/d}}.
\end{aligned}
\tag{46}
$$

Applying the bounds (45) and (44) to (42) we obtain an analogous lower bound for the mean of the dual functional L^*

$$
\begin{aligned}
&EL^*(X_1,\ldots,X_n) \\
&\quad\geq \beta_1\int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx - \frac{\beta_1}{(nm^{-d})^{1/2}}\frac{(d-\gamma)}{d}\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx \\
&\qquad - \frac{K_1}{(nm^{-d})^{1/d}}\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx - \frac{C_2}{(nm^{-d})^{(d-\gamma)/d}}
\end{aligned}
\tag{47}
$$

By definition of the dual,

$$\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} \geq \frac{EL^*(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - n^{-\frac{d-\gamma}{d}} \tag{48}$$

which when combined with (47) and (46) yields the result

$$\begin{aligned}\left|\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-1}{d}}} - \beta_1 \int f^{\frac{d-1}{d}} dx\right| \quad \leq \quad & \frac{K_1}{(nm^{-d})^{1/d}} \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \frac{\beta_1}{(nm^{-d})^{1/2}} \frac{(d-\gamma)}{d} \int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx \\ & + \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}}, \end{aligned} \tag{49}$$

where $K_2 = \max\{C_1, C_2\}$. This establishes Proposition 8. □

D.2 Convergence Rates for Arbitrary Density Functions

The total variation $V(Q)$ over a rectangle $Q \subset [0,1]^d$ of a function g on $\mathbf{R}^d$ is defined as [29]

$$V(Q) = \limsup_{\{z_i\}\in Q} \sum_i |g(z_i) - g(z_{i-1})| = V(Q), \tag{50}$$

where the limsup is taken over all countable subsets $\{z_1, z_2, \ldots,\}$ of points in Q. The function g is said to have bounded variation over Q if $V(Q) < \infty$. By convention, $V(\phi) = 0$ for ϕ the empty set.

In [20] the following lemma was established.

Lemma 2 *For $\nu \in [0,1]$ let f^ν be of bounded variation over $[0,1]^d$ and denote by V its total variation over $[0,1]^d$. Define the resolution $1/m$ block density approximation $\phi(x) = \sum_{i=1}^{m^d} \phi_i \, 1_{Q_i}(x)$ where $\phi_i = m^d \int_{Q_i} f(x)dx$. Then*

$$0 \leq \int |\phi^\nu(x) - f^\nu(x)|dx \leq m^{-d} \sum_{i=1}^{m^d} V(Q_i). \tag{51}$$

Proposition 9 *Assume that f is a Lebesgue density and that f^ν is of bounded variation over $[0,1]^d$ for $\nu = (d-\gamma)/d$, $\nu = (d-\gamma-1)/d$ and $\nu = 1/2 - \gamma/d$. Assume also that $d \geq 2$, $1 \leq \gamma < d$. Then for any continuous quasi-additive functional L satisfying the "add-one bound" (38)*

$$\left|\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - \beta_1 \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right| \leq C\, n^{-1/(d+1)} + o(n^{-1/(d+1)}),$$

where

$$C = \begin{cases} K_1 \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \beta_1 V_0, & d > 2 \\ K_1 \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \beta_1 V_0 + \beta_1 \frac{(d-\gamma)}{d} \int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx, & d = 2 \end{cases},$$

where V_0 is the total variation of $f^{(d-\gamma)/d}$ over $[0,1]^d$.

Proof

Denote the total variation of $f^{\frac{d-\gamma}{d}}, f^{\frac{d-\gamma-1}{d}}$ and $f^{\frac{1}{2}-\frac{\gamma}{d}}$ over a subset A of $[0,1]^d$ as $V_0(A)$, $V_1(A)$ and $V_2(A)$, respectively. Also, for given m let $\phi(x)$ be the block density approximation to f defined in Lemma 2.

Using Schwartz inequality on (49) and (51)

$$\left|\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - \beta_1 \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right| \tag{52}$$

$$\begin{aligned}
&= \left|\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - \beta_1 \int_{\mathcal{S}} \phi^{\frac{d-\gamma}{d}}(x)dx + \beta_1 \int_{\mathcal{S}} \phi^{\frac{d-\gamma}{d}}(x)dx - \beta_1 \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right| \\
&\leq \left|\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - \beta_1 \int_{\mathcal{S}} \phi^{\frac{d-\gamma}{d}}(x)dx\right| + \beta_1 \left|\int \phi^{\frac{d-\gamma}{d}}(x)dx - \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right| \\
&\leq \frac{K_1}{(nm^{-d})^{1/d}} \left[\int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + m^{-d}\sum_{i=1}^{m^d} V_1(Q_i)\right] \\
&+ \frac{\beta_1}{(nm^{-d})^{1/2}} \frac{(d-\gamma)}{d} \left[\int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx + m^{-d}\sum_{i=1}^{m^d} V_2(Q_i)\right] \\
&+ \frac{K_2}{(nm^{-d})^{(d-\gamma)/d}} + n^{-\frac{d-\gamma}{d}} + \beta_1 m^{-d}\sum_{i=1}^{m^d} V_0(Q_i),
\end{aligned} \tag{53}$$

By the bounded variation assumptions the terms $\sum_{i=1}^{m^d} V_j(Q_i)$ are finite for all m, $j = 0,1,2$. These terms are upper bounded by the corresponding total variations of $f^{(d-\gamma)/d}$, $f^{\frac{d-\gamma-1}{d}}$ and $f^{\frac{1}{2}-\frac{\gamma}{d}}$ over $[0,1]^d$ which we denote V_0, V_1 and V_2.

The bound (53) is actually a family of bounds for different values of $m = 1,2,\ldots$. The rates in m, n which dominate in the bound are $(nm^{-d})^{-1/d}$ and m^{-d}, which are respectively due to the first and last terms in the sum (53). We obtain an m-independent bound on the bias term (52) by selecting $m = m(n)$ to be the increasing function of n which minimizes the maximum of these two rates

$$m(n) = \text{argmin}_m \max\left\{(nm^{-d})^{-1/d}, m^{-d}\right\}.$$

The solution $m = m(n)$ satisfies $(nm^{-d})^{1/d} = m^{-d}$, or $m = n^{1/[d(d+1)]}$ (integer part) and, correspondingly, $nm^{-d} = n^{d/(d+1)}$ and $m^{-d} = n^{-1/(d+1)}$. Therefore

$$\begin{aligned}
&\left|\frac{EL(X_1,\ldots,X_n)}{n^{\frac{d-\gamma}{d}}} - \beta_1 \int_{\mathcal{S}} f^{\frac{d-\gamma}{d}}(x)dx\right| \\
&\leq \frac{1}{n^{1/(d+1)}}\left[K_1 \int_{\mathcal{S}} f^{\frac{d-\gamma-1}{d}}(x)dx + \beta_1 V_0\right] \\
&+ \frac{\beta_1}{n^{d/[2(d+1)]}} \frac{(d-\gamma)}{d} \int_{\mathcal{S}} f^{\frac{1}{2}-\frac{\gamma}{d}}(x)dx + \frac{\beta_1}{n^{(d+2)/[2(d+1)]}} \frac{(d-\gamma)}{d} V_2 \\
&+ \frac{K_1}{n^{2/(d+1)}} V_1 + \frac{K_2}{n^{(d-\gamma)/(d+1)}} + n^{-\frac{d-\gamma}{d}}.
\end{aligned}$$

This establishes Proposition 9. □.

References

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

[2] 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/`.

[3] 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, 1998.

[4] 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, 1998.

[5] 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.

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

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

[9] S. Kullback and R. Liebler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[10] 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.

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

[12] 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/`.

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

[14] B. Ma, *Parametric and non-parametric approaches for multisensor data fusion*, PhD thesis, University of Michigan, Ann Arbor, MI 48109, 2001.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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)*, San Diego, CA, July 1998.

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

[22] 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.

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

[24] 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.

[25] A. P. Korostelev and A. B. Tsybakov, *Minimax theory of image reconstruction*, Springer-Verlag, New York, 1993.

[26] C. Redmond and J. E. Yukich, "Limit theorems and rates of convergence for Euclidean functionals," *Ann. Applied Probab.*, vol. 4, no. 4, pp. 1057–1073, 1994.

[27] C. Redmond and J. E. Yukich, "Asymptotics for Euclidean functionals with power weighted edges," *Stochastic Processes and their Applications*, vol. 6, pp. 289–304, 1996.

[28] J. M. Steele, "Growth rates of euclidean minimal spanning trees with power weighted edges," *Ann. Probab.*, vol. 16, pp. 1767–1787, 1988.

[29] F. Riesz and B. Sz.-Nagy, *Functional analysis*, Ungar, New York, 1955.