Hyperspacings and the Estimation of Information Theoretic Quantities

Erik G. Learned-Miller
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003
elm@cs.umass.edu

Abstract

The estimation of probability densities from data is widely used as an intermediate step in the estimation of entropy, Kullback-Leibler (KL) divergence, and mutual information, and for statistical tasks such as hypothesis testing. We propose an alternative to density estimation–partitioning a space into regions whose approximate probability mass is known–that can be used for the same purposes. We call these regions *hyperspacings*, a generalization of *spacings* in one dimension. After discussing one-dimensional spacings estimates of entropy and KL-divergence, we show how hyperspacings can be used to estimate these quantities (and mutual information) in higher dimensions. Our approach outperforms certain widely used estimators based on intermediate density estimates. Using similar ideas, we also present a new *distribution-free* hypothesis test for distributional equivalence that compares favorably with the Kolmogorov-Smirnov test. Using hyperspacings, it is easily extended to multiple dimensions.

1 Introduction

Many problems in machine learning involve the estimation of information theoretic quantities such as entropy, Kullback-Leibler (KL) divergence, and mutual information in continuous probability spaces. A first step toward calculating these quantities is often estimating a probability density over one or more (possibly multivariate) random variables. Examples include estimation of joint and marginal densities in image registration problems as part of mutual information estimation [14] and estimation of marginal densities in order to estimate marginal entropies in Independent Components Analysis (ICA) problems [11].

It is not essential to estimate a **density**, however, before estimating **entropy** or other information theoretic quantities from a sample. Using methods based on *order statistics* and *spacings* (defined below), the entropy of a one-dimensional random variable can be directly estimated without an explicit density estimate [13]. These estimates are consistent and asymptotically efficient [2], and have been exploited in solutions of the ICA problem [3, 9]. Recently, the concept of spacings was extended to higher dimension and applied to the problem of entropy estimation [8]. Another class of entropy estimators that sidestep density estimation has been developed using so-called *entropic spanning graphs* [6].

Here we build on previous work with spacings and their generalization to multiple dimensions [8]. We start in Section 2 by reviewing entropy estimation in one dimension. We then introduce *Near Uniform Partitions* (NUPs), which provide a simple conceptual framework for this family of estimators and lead to a novel algorithm for KL-divergence estimation. In Section 3, we introduce hyperspacings, an attempt to create NUPs in higher dimensions. In addition to algorithms for multidimensional estimation of entropy and KL-divergence, this provides a new method for estimating mutual information. We compare our entropy estimator to a standard technique based on density estimates. In Section 4, we show how a spacings algorithm for estimating KL-divergence suggests a natural hypothesis test for whether two samples come from the same distribution. Like the Kolmogorov-Smirnov test, this test is *distribution free*, but being based on NUPs, it can be generalized to arbitrary dimension. We conclude by comparing the two tests in simulations.

2 Spacings estimates of entropy and KL-divergence

Consider a scalar random variable Z, and a random iid sample of Z denoted by $Z^1, Z^2, ..., Z^N$. The *order statistics* of a random sample of Z are simply the elements of the sample rearranged in non-decreasing order: $Z^{(1)} \leq Z^{(2)} \leq ... \leq Z^{(N)}$ (c.f. [1]). A *spacing of order m*, or *m-spacing*, is then defined[1] to be $Z^{(i+m)} - Z^{(i)}$, for $1 \leq i < i+m \leq N$. The $m-$spacing estimator of entropy, due to Vasicek [13], is defined as

$$\hat{H}_N(Z^1, ..., Z^N) = \frac{1}{N} \sum_{i=1}^{N-m} \log \left(\frac{N}{m} (Z^{(i+m)} - Z^{(i)}) \right). \tag{1}$$

To gain insight into this estimator, note that for *any random variable Z with an impulse-free density* $p(\cdot)$ *and continuous distribution function* $P(\cdot)$, the following holds. Let p^* be the N-way product density $p^*(Z^1, Z^2, ..., Z^N) = p(Z^1)p(Z^2)...p(Z^N)$. Then

$$E_{p^*}[P(Z^{(i+1)}) - P(Z^{(i)})] = \frac{1}{N+1}, \ \forall i, 1 \leq i \leq N-1. \tag{2}$$

That is, the expected value of the probability mass of the interval between two successive elements of a sample from a random variable is $\frac{1}{N+1}$. This remarkably general fact is a simple consequence of the uniformity of the random variable $P(Z)$, the *probability integral transform* of Z (c.f. [7]). Using this idea, one can develop a simple entropy estimator. We start by approximating the probability density $p(z)$ by assigning equivalent masses to each interval between points and assuming a uniform distribution of this mass across the interval.[2] Defining $Z^{(0)}$ and $Z^{(N+1)}$ to be the infimum and supremum of the support of $p(z)$, we have:

$$\hat{p}(z; Z^1, ..., Z^N) = \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}}, \tag{3}$$

for $Z^{(i)} \leq z < Z^{(i+1)}$. Then, we can write

$$\begin{aligned} H(Z) &\overset{(a)}{\approx} -\int_{-\infty}^{\infty} \hat{p}(z) \log \hat{p}(z) dz \\ &= -\sum_{i=0}^{N} \int_{Z^{(i)}}^{Z^{(i+1)}} \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} \log \frac{\frac{1}{N+1}}{Z^{(i+1)} - Z^{(i)}} dz \end{aligned}$$

[1] Here, *spacing* is defined as the length of an interval marked by order statistics. We will also refer the interval itself as a *spacing* where convenient.

[2] The notion of a density estimate aids in the intuition behind $m-$spacing estimates. However, we stress that density estimation *is not* a necessary intermediate step in our ultimate entropy estimator.

$$
\begin{aligned}
&= -\sum_{i=0}^{N} \frac{\frac{1}{N+1}}{Z^{(i+1)}-Z^{(i)}} \log \frac{\frac{1}{N+1}}{Z^{(i+1)}-Z^{(i)}} \int_{Z^{(i)}}^{Z^{(i+1)}} dz \\
&= -\frac{1}{N+1} \sum_{i=0}^{N} \log \frac{\frac{1}{N+1}}{Z^{(i+1)}-Z^{(i)}} \\
&\overset{(b)}{\approx} -\frac{1}{N-1} \sum_{i=1}^{N-1} \log \frac{\frac{1}{N+1}}{Z^{(i+1)}-Z^{(i)}} \\
&= \frac{1}{N-1} \sum_{i=1}^{N-1} \log \left((N+1)(Z^{(i+1)}-Z^{(i)}) \right) \\
&\equiv \hat{H}_{simple}(Z^1,...,Z^N). \qquad (4)
\end{aligned}
$$

The approximation (a) arises by approximating the true density $p(z)$ by $\hat{p}(z;Z^1,...,Z^N)$. The approximation (b) stems from the fact that we in general do not know $Z^{(0)}$ and $Z^{(N+1)}$, i.e. the true support of the unknown density. Instead, we form the entropy estimate using only information from the region for which we have some information.

2.1 m-spacings and overlapping m-spacings

The estimate $\hat{H}_{simple}$ has high variance inherited from the variance of the interval probabilities (2). The variance can be reduced by considering m-spacing estimates, such as

$$
\hat{H}_{mspacing}(Z^1,...,Z^N) \equiv \frac{m}{N-1} \sum_{i=0}^{\frac{N-1}{m}-1} \log \left(\frac{N+1}{m} (Z^{(m(i+1)+1)} - Z^{(mi+1)}) \right). \qquad (5)
$$

When $m,N \to \infty, \frac{m}{N} \to 0$, this estimator is consistent [2]. As m and N grow, the probability masses for m-spacings concentrate around their expected values. This property holds for *all probability distributions with continuous cumulative distribution functions*. A modification of (5) in which the m-spacings overlap,

$$
\hat{H}_{overlap}(Z^1,...,Z^N) \equiv \frac{1}{N-m} \sum_{i=1}^{N-m} \log \left(\frac{N+1}{m} (Z^{(i+m)} - Z^{(i)}) \right), \qquad (6)
$$

further reduces the asymptotic variance and is equivalent to Vasicek's estimator (1) except for constants adjusted to improve the small sample performance. There is no specific density associated with this estimator, and yet this does not diminish its performance. Next, we introduce Near Uniform Partitions, which capture some key properties of spacings.

2.2 Near Uniform Partitions

Suppose we could put a grid on a probability distribution so that the integral of the distribution over each grid element was a constant. Such a "uniform partition" might be useful for estimating quantities associated with the distribution, especially expectations. Near Uniform Partitions (NUPs) are an approximation to such a grid. To construct a NUP on a space with respect to a probability distribution, we must define a set of mutually exclusive and collectively exhaustive regions on that space that are *likely to have approximately equivalent probability masses.* The following definition formalizes this idea.

Definition 2.1 (Near Uniform Partition) *Consider a probability space $(\Omega, \mathcal{F}, P)$ and the associated N-way product space $(\Omega^N, \mathcal{F}^N, P^N)$. Let $X = \{X^1, X^2, ..., X^N\} \in \Omega^N$ be a sample drawn according to P^N. Let $\mathbf{R} = \mathbf{R}(X)$ be a partition of the outcomes Ω into regions*

$\{\mathbf{R}_1, \mathbf{R}_2, \ldots \mathbf{R}_K\}$ *which depends on the sample X, and hence is a random partition. Using* $P(\mathbf{R}_i)$ *as a shorthand for* $\int_{\mathbf{R}_i} p(x)dx$*, we say that the random variable* $\mathbf{R}$ *is an* ε-δ *Near Uniform Partition if, for a random draw of* $\mathbf{R}$*,*

$$Prob\left(\max_i \left|\frac{1}{K} - P(\mathbf{R}_i)\right| > \varepsilon\right) < \delta. \tag{7}$$

Next, we make the following **claim**: *that sets of non-overlapping m-spacings form a (non-trivial) NUP for any continuous probability distribution.* Consider a sample of size $N = mK - 1$ along with the infimum and supremum of support of the distribution. To establish that the set of K non-overlapping m-spacings form a NUP, we must choose an ε and δ so that (7) holds. We start with the fact that the distribution of probability mass in an $m-$spacing is given by a beta distribution with parameters m and $N+1-m$ [1], with expectation $\frac{m}{N+1} = \frac{1}{K}$ and variance $\frac{m(N+1-m)}{(N+1)^2(N+2)}$. Applying Chebyshev's inequality gives

$$Prob\left(\left|\frac{1}{K} - P(\mathbf{R}_i)\right| > \varepsilon\right) \leq \frac{m(N+1-m)}{(N+1)^2(N+2)\varepsilon^2}. \tag{8}$$

If all of the m-spacings have probability within ε of their expectations, then the maximum deviation of these probabilities is also within ε, so we can apply the union bound to obtain

$$Prob\left(\max_i \left|\frac{1}{K} - P(\mathbf{R}_i)\right| > \varepsilon\right) \leq \frac{mK(N+1-m)}{(N+1)^2(N+2)\varepsilon^2} < \frac{1}{(N+2)\varepsilon^2}. \tag{9}$$

With large enough N, we can choose ε and δ arbitrarily small, establishing the claim. Next, we see how NUPs lead to a conceptually simple algorithm for KL-divergence estimation.

2.3 KL-divergence estimation

Starting with the definition of KL-divergence between distributions P and Q, we write

$$\begin{aligned} D(P||Q) &= \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx &&= \int_{-\infty}^{\infty} \frac{p(x)}{q(x)} \left(\log \frac{p(x)}{q(x)}\right) q(x)dx \\ &= \int_{-\infty}^{\infty} \frac{dP(x)}{dQ(x)} \left(\log \frac{dP(x)}{dQ(x)}\right) dQ(x) &&= -h(\frac{dP}{dQ}), \end{aligned}$$

where $h(\cdot)$ is the differential entropy. $\frac{dP}{dQ}$ is the density of P with respect to the underlying measure Q. This derivation can be interpreted in the following way: *By representing the probability law P in a space in which Q is uniform, the divergence between P and Q can be written as simply the negative entropy of P.* Practically, we achieve this by representing P under the NUP defined by the m-spacings of Q.

More explicitly, suppose we have samples from distributions P and Q, each of size N, and we wish to estimate $D(P||Q)$. The steps of the KL-divergence algorithm are **a**) Set $m = \sqrt{N}$, **b**) Compute non-overlapping m-spacings using samples of Q, **c**) Compute the histogram of samples from P, using the m-spacings of Q as bins, and **d**) Calculate the negative entropy of this histogram to obtain a KL-divergence estimate. (This algorithm has computational complexity $O(N \log N)$ due to the sorting required to find the order statistics.)

The number of P samples in each histogram bin represents the amount of "P probability" (dP) for a fixed amount of change in Q (dQ). Our confidence that each histogram bin has a fixed amount of Q mass comes from the fact that the m-spacings of Q form a NUP. It is interesting to note that this algorithm, since it depends only on the ordering of the two samples, is completely invariant to arbitrary monotonic (non-linear!) mappings of the axes. This is a property shared by the KL-divergence of the true distributions P and Q and hence makes it an appealing property of the estimator. Note in particular that this property is *not* shared by KL-divergence estimates based on kernel density estimators.

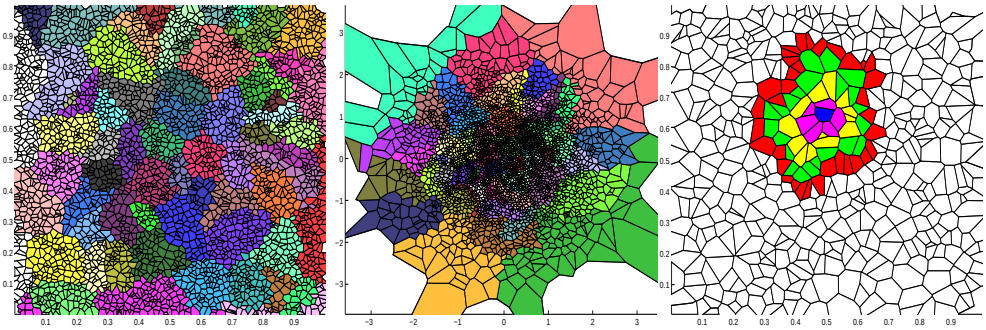

Figure 1: **Left, center:** Hyperspacings for $N = 4000$ points for the uniform and standard 2-D Gaussian distributions. In each case, the hyperspacings have probability mass that is approximately linear in the number of Voronoi regions that compose them. **Right:** A hyperspacing of radius 5. Our entropy estimator is based on overlapping hyperspacings of this form, as opposed to the non-overlapping hyperspacings shown on the left. NOTE: This figure is best viewed in color.

3 Hyperspacings: spacings in multiple dimensions

Our next goal is to extend the benefits of spacings and NUPs to higher dimensions. To do so, we need to generate regions of multidimensional spaces with predictable probability mass. Next, we present two methods for generating such regions from samples.

3.1 Voronoi regions and hyperspacings

Given a set of points $Z^1, Z^2, \ldots, Z^N$ in D dimensions, a set of *Voronoi regions* $V^1, V^2, \ldots, V^N$ is formed by associating with each point Z^i the set V^i of all points which are closer to Z^i than to any other point Z^j. Voronoi regions can be thought of as a generalization of spacings in higher dimensions. An entropy estimator similar to $\hat{H}_{simple}$ (4) is obtained by assigning to each Voronoi region a probability mass of $\frac{1}{N}$, distributing that mass uniformly in each region, and discounting regions with infinite area (or volume in higher dimensions):

$$\hat{H}_{Vor-simple} \equiv \frac{1}{N-z} \sum_{V^i s.t. A(V^i) \neq \infty} \log\left(NA(V^i)\right). \tag{10}$$

Here A is the area of region V_i, and z is the number of Voronoi regions with infinite area. Such an estimator can also be built from *Delaunay* regions, the "duals" of Voronoi regions [10]. In two dimensions, a Delaunay region is formed by connecting the centers of three mutually adjacent Voronoi regions. One important property of Delaunay tessellations is that each region is finite. The tradeoff, however, is that like spacings estimates in one dimension, only the convex hull of the sample is modeled. Most of the algorithms we present can be developed using either Voronoi or Delaunay regions, but we will usually simply refer to the Voronoi version for brevity.

Just as the 1-spacing estimator ($\hat{H}_{simple}$) was extended to the m-spacings estimator ($\hat{H}_{mspacing}$), we can extend the basic Voronoi entropy estimator to reduce its variance. In one dimension, this was achieved by "pasting" together contiguous intervals into an m-spacing. In D dimensions, we paste together multiple Voronoi regions into *hyperspacings*. Non-overlapping hyperspacings for two different distributions are shown in the first two panels of Figure 1. A single hyperspacing, used as part of the overlapping hyperspacings estimate, is shown on the right of the figure.

	N=100				N=1000				N=5000			
Distributions	Hyper		ROT		Hyper		ROT		Hyper		ROT	
	Bias	σ	Bias	σ	Bias	σ	Bias	σ	Bias	σ	Bias	σ
1D Gaussian	**2.0**	5.9	**0.1**	5.6	**1.1**	1.8	**0.8**	1.7	**0.4**	0.7	**0.4**	0.8
1D Uniform	**4.0**	1.7	**6.7**	2.6	**1.1**	0.3	**5.0**	0.5	**0.5**	0.1	**3.8**	0.2
1D Exponential	**10.3**	9.9	**5.2**	10.9	**1.3**	2.3	**8.5**	2.7	**0.5**	1.4	**6.9**	1.5
2D Gaussian	**7.3**	2.6	**6.3**	2.6	**0.8**	1.2	**1.9**	1.1	**0.2**	0.5	**0.8**	0.5
2D Uniform	**5.3**	1.4	**4.2**	2.0	**1.1**	0.3	**4.4**	0.4	**0.6**	0.1	**3.6**	0.2
2D Exponential	**4.2**	6.9	**5.5**	7.2	**0.9**	3.2	**6.1**	3.2	**0.4**	0.9	**5.8**	0.9
2D Gauss x Exp.	**4.8**	6.8	**0.6**	6.5	**1.0**	2.0	**1.4**	1.9	**0.6**	0.9	**1.6**	0.9
2D Annulus	**12.9**	1.3	**32.3**	0.9	**6.5**	0.4	**22.2**	0.4	**3.6**	0.2	**16.1**	0.1
2D Hollow Square	**7.9**	1.1	**26.0**	1.0	**2.3**	0.4	**16.8**	0.3	**0.6**	0.2	**12.0**	0.2
3D Gaussian	**16.7**	3.7	**14.5**	4.2	**3.3**	0.7	**7.3**	0.9	**1.5**	0.5	**4.7**	0.5
3D Uniform	**11.0**	1.4	**0.5**	1.4	**2.9**	0.2	**1.8**	0.3	**1.6**	0.1	**2.1**	0.1
3D Exponential	**13.1**	5.5	**5.5**	7.9	**2.3**	1.4	**1.2**	1.7	**1.5**	0.9	**0.2**	0.9

Table 1: Results of entropy experiments. The entropy of each distribution shown on the left was estimated from samples (of size 100, 1000, and 5000) using the hyperspacings estimate based on Delaunay tessellations ("Hyper") and based on a kernel density estimate with Gaussian kernels, using Silverman's "rule-of-thumb" to estimate the kernel size ("ROT"). The bias for each estimate is the mean absolute value percentage difference from the true entropy. Distributions were chosen so that the true entropy had a value of 2 or greater. σ shows the standard deviation of each estimate as a percentage of the entropy. For $N = 100$, results are mixed between the two estimators, but for larger sample sizes, the hyperspacings estimate performs substantially better, with lower bias and similar standard deviation on both smooth and rapidly changing densities.

In constructing a hyperspacing, it is tempting to include any Voronoi region whose center is included in some Euclidean ε-ball of a particular point. However, this method of forming hyperspacings gives clusters with more constituent Voronoi regions in areas of high density than in areas of low density. Instead, we define an *adjacency metric* on the set of Voronoi regions by setting the distance between any two regions V^i and V^j to be the shortest path on the adjacency graph for the set of regions. The rightmost panel of Figure 1 shows a typical adjacency metric ball around a particular Voronoi region. The use of an adjacency metric makes the hyperspacings method of partitioning a distribution relatively insensitive to the underlying distribution, and allows the efficient computation of hyperspacings.

Unlike *m*-spacings in one dimension, however, it is difficult to prove that hyperspacings are NUPs. Instead we note the following properties of hyperspacings:

1. For a uniform distribution on the unit hypercube , a single hyperspacing (not intersecting the boundary of the hypercube) with *m* subregions has expected probability mass $\frac{m}{N}$. According to our experiments in two dimensions, the standard deviation of this mass is already less than 10% of the expected mass for a hyperspacing radius of only 4 (in the adjacency metric).
2. Every probability density with bounded partial derivatives is locally approximately uniform. Also, the probability mass in a hyperspacing does not depend upon the local height of a density, as long as it is uniform.[3] Together these imply that the probability masses of hyperspacings are asymptotically invariant to the underlying density at any particular location, as long as the density is smooth.
3. Hyperspacings on densities with unbounded derivatives can still be well behaved,

[3]This is a direct consequence of the invariance of Voronoi tessellations to scale [10].

Distribution 1	Distribution 2	Kol-Smir	Hyperspace
Uniform($\mu = 0, \sigma^2 = 1$)	Uniform($\mu = 0, \sigma^2 = 0.9$)	57	**86**
Uniform($\mu = 0.1, \sigma^2 = 1$)	Uniform($\mu = 0, \sigma^2 = 1$)	39	**81**
Normal($\mu = 0, \sigma^2 = 1$)	Normal($\mu = 0, \sigma^2 = 0.9$)	**9**	6
Normal($\mu = 0.1, \sigma^2 = 1$)	Normal($\mu = 0, \sigma^2 = 1$)	**48**	6
3-mode($\mu = 0, \sigma^2 = 1$)	3-mode($\mu = 0, \sigma^2 = 1.02$)	**81**	44
5-mode($\mu = 0, \sigma^2 = 1$)	5-mode($\mu = 0, \sigma^2 = 1.02$)	94	**99**
7-mode($\mu = 0, \sigma^2 = 1$)	7-mode($\mu = 0, \sigma^2 = 1.01$)	59	**100**

Table 2: Results of hypothesis test experiments for the 1-D test. A sample of size 1000 was drawn from the pair of distributions in each row. Under the null hypothesis, the distributions are the same. The power of the tests (shown in columns 3 and 4) are the rejection percentages for the null hypothesis out of 1000 runs at the $\alpha = 0.05$ significance level. The higher power test is shown in bold in each case.

as long as the number of hyperspacings which contain sharp transitions is small relative to the total number of hyperspacings.

In summary, we shall assume that hyperspacings are "close enough" to NUPs to be useful, and we shall let them adopt the roles of *m*-spacings in our higher dimensional estimators.

3.2 Entropy experiments and mutual information

Using overlapping hyperspacings as surrogates for overlapping *m*-spacings, we formed an entropy estimator for distributions in arbitrary dimension that is essentially equivalent to $\hat{H}_{overlap}$ (6). We conducted experiments to evaluate our entropy estimator in one, two, and three dimensions. We compared against Silverman's "rule-of-thumb" estimator [12], which is a fixed kernel estimator. Results in Table 1 show that our estimator outperforms the Silverman estimator for larger samples, and is comparable for small samples. Setting $m = \log(N)$, our 2-D estimator is $O(N \log N)$, since 2-D Voronoi regions can be computed in $O(N \log N)$ [10]. We do not yet have complexity results for higher dimensions.

It is well-known that the mutual information between two random variables can be written as $h(X) + h(Y) - h(X,Y)$, where $h(\cdot,\cdot)$ is the joint differential entropy [4]. With an estimator of both one and two-dimensional entropies, it is easy to estimate mutual information simply by computing each of the constituent entropies. We now offer another application of our KL-divergence estimators.

4 A distribution-free test of distributional equivalence

The KL-divergence estimator described in Section 2.3 can be used to form a simple hypothesis test of distributional equivalence, i.e. whether two samples were drawn from the same distribution. The idea is to estimate the KL-divergence[4] between two samples and see whether it exceeds a particular threshold. Since KL-divergence is minimized when two distributions are equivalent, this is a natural test. Our one-dimensional hypothesis test, like our KL-divergence estimator, is *distribution-free* in the sense that an arbitrary monotonic transformation of the coordinate axes will not affect its behavior. We stress that this is true *for any sample size*. As with the Kolmogorv-Smirnov (KS) test [7], the critical values for this test *do not depend upon the distributions being tested*, since the distribution over the test statistic is equivalent under the null hypothesis, irrespective of the distribution. This

[4] We found that the *symmetric divergence*, $D(P||Q) + D(Q||P)$ produced a more powerful test than the simple KL-divergence.

is true because the test statistic is only dependent upon the *ordering of points in the two samples*, and this does not depend on the densities themselves, but only on their ratio.

To obtain critical values (at the $\alpha = 0.05$ significance level) for our hypothesis test, we computed the test statistic over 50,000 trials, using 1000 samples from each of two uniform distributions on each run. We compared the power of our test and the KS test by evaluating the rate of rejection of the null hypothesis under sampling from the pairs of distributions shown in Table 2. The power of our test was higher in the majority of cases we examined.

In addition, our test generalizes elegantly via hyperspacings to higher dimensions, although it is no longer strictly distribution-free. Our initial tests compared to 2-D extensions of the KS test [5] suggest that our test is more powerful for some distribution pairs and less for others. An example for which it outperforms [5] is in detecting the difference between a uniform distribution rotated at 45 vs. 55 degrees. Here, the power of our test was 52 vs. 18 for [5]. We note that our algorithm is $O(N \log N)$ in 2-D vs. a complexity of $O(N^2)$ for [5]. Further investigations of the multidimensional tests are left for future work.

4.1 Summary

The central idea in this paper is that spacings have certain distribution-free properties, and that these properties can be extended to higher dimensions, in an approximate fashion, via *hyperspacings*. We have presented new competitive algorithms for KL-divergence estimation and hypothesis testing in one dimension, and new multidimensional algorithms for these quantities, as well as entropy and mutual information, using hyperspacings.

References

[1] Arnold, B., Balakrishnan, N., Nagaraja, H. *A First Course in Order Stats*. Wiley & Sons, 1992.

[2] Beirlant, J., Dudewicz, E. J., Györfi, L., van der Meulen, E. C. Nonparametric entropy estimation: an overview, *International Journal of Math. Stat. Sci.* **6**, pp. 17-39, 1997.

[3] Bercher, J.-F. and Vignat, C. Estimating the entropy of a signal with applications. *IEEE Transactions on Signal Processing,* **48**, pp.1687-1694, 2000.

[4] Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley & Sons, 1991.

[5] Fasano, G. and Franceschini, A. A multidimensional version of the Kolmogorov-Smirnov test. *Monthly Notices of the Royal Astronomical Society,* **225** pp. 155-170.

[6] Hero, A., Ma, B., Michel, O., Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine (Special Issue on Mathematics in Imaging),* **19**, pp. 85-95, 2002.

[7] Manoukian, E. *Modern Concepts and Thms. of Math. Stats*. New York: Springer-Verlag. 1986.

[8] Miller, E. A new class of entropy estimators for multi-dimensional densities. *International Conference on Acoustics, Speech, and Signal Processing,* 2003.

[9] Miller, E. and Fisher, J. ICA using spacings estimates of entropy. *Fourth International Symposium on Independent Component Analysis and Blind Signal Separation,* 2003.

[10] Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, 2nd Edition,* Wiley & Sons, 1992.

[11] Pham, D. T. Blind separation of instantaneous mixtures of sources via an independent component analysis. *IEEE Transactions on Signal Processing,* **44**, pp.2768-2779, 1996.

[12] Silverman, B. W. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall. 1992.

[13] Vasicek, O. A test for normality based on sample entropy. *J. Royal Stat. Soc., Ser. B,* **38**, pp. 54-59, 1976.

[14] Viola, P. and Wells III, W. M. Alignment by maximization of mutual information. *Proceedings of IEEE International Conference on Computer Vision*, pp. 16-23, 1995.