April, 2003

A new class of entropy estimators for multi-dimensional densities

Erik G Learned-Miller, *University of Massachusetts - Amherst*

International Conference on Acoustics, Speech, and Signal Processing, 2003

A NEW CLASS OF ENTROPY ESTIMATORS FOR MULTI-DIMENSIONAL DENSITIES

Erik G. Miller

EECS Department, UC Berkeley
Berkeley, CA 94720, USA

ABSTRACT

We present a new class of estimators for approximating the entropy of multi-dimensional probability densities based on a sample of the density. These estimators extend the classic "m-spacing" estimators of Vasicek and others for estimating entropies of one-dimensional probability densities. Unlike plug-in estimators of entropy, which £rst estimate a probability density and then compute its entropy, our estimators avoid the dif£cult intermediate step of density estimation. For £xed dimension, the estimators are polynomial in the sample size. Similarities to consistent and asymptotically ef£cient one-dimensional estimators of entropy suggest that our estimators may share these properties.

1. INTRODUCTION

The entropy $H(f)$ of a continuous probability density $f(x)$ is given by

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx,$$

as described in [1]. In this paper, we concern ourselves with the estimation of the entropy when the density $f(x)$ is unknown, but when we have a sample of size N drawn iid from this density. The estimation of entropy from a sample is an important problem, with applications in goodness-of-£t tests, parameter estimation, source-coding, econometrics, and many other areas [2].

Beirlant et al. [2] give an excellent review of standard methods of entropy estimation. A common practice is to use so-called *plug-in* estimates. In this approach, the unknown density $f(x)$ is £rst estimated from a sample using any standard density estimation technique. Subsequently, the entropy of the density estimate $\hat{f}(x)$ is computed as an estimate of the true entropy of f. While plug-in estimates work well in low-dimensions and for densities with known parametric form, the dif£cult problem of density estimation makes them impractical for small sample sizes in higher dimensions.

Funding was generously provided through ONR grant N00014-01-1-0890 under the MURI program.

Another method for estimating one-dimensional entropies is based on the order statistics of a sample. In this paper, we show how these consistent and rapidly converging estimators can be extended to multiple dimensions, resulting in effective and computationally ef£cient entropy estimators for multidimensional distributions.

2. M-SPACINGS ESTIMATES IN ONE DIMENSION

2.1. Order statistics and spacings

Consider a scalar random variable Z, and a random 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. [3]). A *spacing of order* m, or m*-spacing*, is then de£ned to be $Z^{(i+m)} - Z^{(i)}$, for $1 \leq i < i+m \leq N$. Finally, if m is a function of N, one may de£ne the m_N*-spacing* as $Z^{(i+m_N)} - Z^{(i)}$.

The m_N-spacing estimator of entropy, originally due to Vasicek [4], can now be de£ned as

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

To see where this estimator comes from, we £rst make the following observation regarding order statistics. 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

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

That is, the expected value of the probability mass of the interval between two successive elements of a sample from a random variable[1] is just $\frac{1}{N+1}$ of the total probability (1.0). This surprisingly general fact is a simple consequence of the uniformity of the random variable $P(Z)$. $P(Z)$, the random variable describing the "height" on the cumulative curve of

[1] The probability mass of the interval between two successive points is the integral of the density function between these two points.

a random draw from Z (as opposed to the function $P(z)$) is called the *probability integral transform* of Z (c.f. [5]). Thus, the key insight is that the *intervals* between successive order statistics have the same expected probability mass.

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]. De£ning $Z^{(0)}$ to be the in£mum of the support of $p(z)$ and de£ning $Z^{(N+1)}$ to be the 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

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

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. Therefore, we form the mean log density estimate using only information from the region for which we have some information, ignoring the intervals outside the range of the sample. This is equivalent to assuming that outside the sample range, the true density has the same mean log probability density as the rest of the distribution.

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

2.2. Lowering the variance of the estimate

The estimate $\hat{H}_{simple}$ has both intuitive and theoretical appeal[3], but it has relatively high variance since while the expectation of the interval probabilities (2) is $\frac{1}{N+1}$, their variance is high.

This problem can be mitigated, and asymptotically eliminated completely, by considering $m-$spacing estimates of entropy, such as

$$\hat{H}_{m-spacing}(Z^1, ..., Z^N) \equiv \tag{4}$$
$$\frac{m}{N-1} \sum_{i=0}^{\frac{N-1}{m}-1} \log \left(\frac{N+1}{m} (Z^{(m(i+1)+1)} - Z^{(mi+1)}) \right).$$

By letting

$$m \to \infty, \frac{m}{N} \to 0, \tag{5}$$

this estimator also becomes consistent [2]. It is typical to set $m = \sqrt{N}$.

The intuition behind this estimator is that by considering m-spacings with larger and larger values of m, the variance of the probability mass of these spacings, relative to their expected values, gets smaller and smaller. 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 modi£cation of (4) in which the $m-$spacings overlap:

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

reduces the asymptotic variance of the spacings estimator. Note that this is equivalent asymptotically to Vasicek's estimator [4]. Weak and strong consistency have been shown given (5) by various authors under a variety of general conditions. For details of these results, see [2]. Perhaps the most important property of this estimator is that it is asymptotically ef£cient, as shown in [7].

3. EXTENDING SPACING ESTIMATORS TO MULTIPLE DIMENSIONS

Ultimately, m-spacings estimators of entropy are based on the intuition that sums of small random intervals (based on order statistics) have consistent behavior. While there is no clear extension of *order statistics* to higher dimensions, there are methods for generating random regions of multi-dimensional spaces with constant expected probability mass. Such methods will allow us to extend the notion

[3]The addition of a small constant renders this estimator weakly consistent for bounded densities under certain tail conditions ([6]).

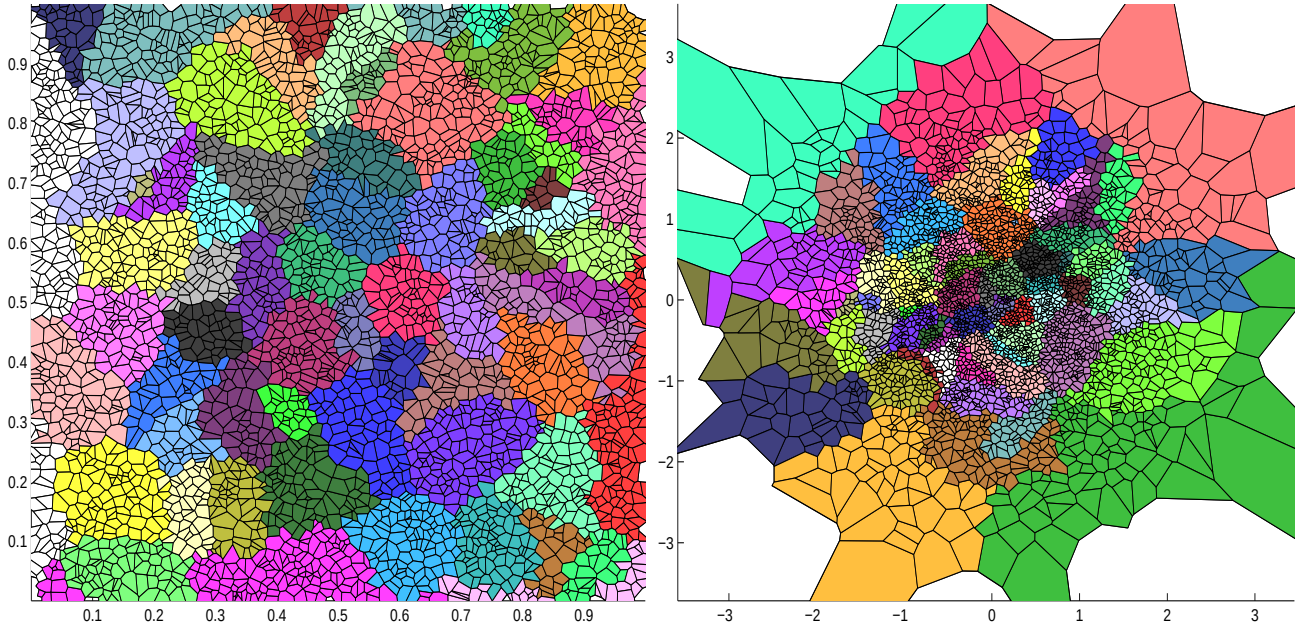

Fig. 1. Hyper-Voronoi regions for $N = 4000$ points. On the left the points were drawn from a uniform distribution over the unit square. On the right, the points were drawn from a two-dimensional Gaussian distribution with diagonal covariance. In each case, the Hyper-Voronoi regions have probability mass that is approximately linear in the number of Voronoi regions that compose them. As a rudimentary test of the m-Voronoi estimator we estimated the entropy of a two-dimensional unit variance Gaussian like the one shown at right. The true entropy in nats for this distribution is approximately 2.8379. Over 100 trials with $N = 1000$, our estimator produced a mean entropy of 3.07 with standard deviation 0.11. Presumably it is biased upward by the assumption that probability is distributed uniformly (and hence with maximum entropy) in each Hyper-Voronoi region. (For color versions of these £gures, see **http://www.eecs.berkeley.edu/˜egmil/papers/vor.pdf**.)

of spacings estimates to higher dimensions. We present two (dual) methods for generating such random regions in the next subsections.

3.1. Voronoi regions in D dimensions

Given a set of points $Z^1, Z^2, ..., Z^N$ in D dimensions, a set of *Voronoi regions* $V^1, V^2, ..., V^N$ 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. [8] is an extensive text on Voronoi regions, Voronoi diagrams, and Voronoi tessellations.

One can easily construct a density estimate of an unknown distribution f from a sample of size N in three steps, by 1) constructing the Voronoi regions, 2) assuming a £xed probability mass ($\frac{1}{N}$) for each Voronoi region, and 3) assuming uniform density over each Voronoi region. The only subtlety here is that the density becomes effectively zero for Voronoi regions which extend to in£nity. As with the spacings estimate, if we know the support of the unknown density f, we may bound these external regions and assign a £nite £xed density to them, or in the case when the support is not known, we may simply choose to ignore these external regions when estimating expectations of quantities based on the sample, just as the spacings estimator ignores the 0th and Nth intervals in the spacings estimate. (See step (b) of the $\hat{H}_{simple}$ derivation.)

If the support of f is known, then through a parallel derivation, this leads to

$$\hat{H}_{Vor-simple} \equiv \frac{1}{N} \sum_{i=1}^{N} \log \left(N A(V^i) \right), \tag{7}$$

where $A(V^i)$ is the D-dimensional volume of Voronoi region V^i. When the support is not known, an almost equivalent estimator can be used:

$$\hat{H}_{Vor-simple2} \equiv \frac{1}{N-K} \sum_{V^i s.t. A(V^i) \neq \infty} \log \left(N A(V^i) \right), \tag{8}$$

where K is the number of Voronoi regions with in£nite volume.

3.2. Delaunay regions in D dimensions

A simple variation on this theme is to use Delaunay regions instead of Voronoi regions in the estimator. Delaunay re-

gions are the duals of Voronoi regions. In two dimensions, a Delaunay region is formed by connecting the centers of three mutually adjacent Voronoi regions [8]. Due to lack of space, we cannot fully discuss the Delaunay estimators, but we note that they may be advantageous when we have a small sample N and high dimension D.

3.3. m-Voronoi and m-Delaunay estimators

Just as the 1-spacing estimator ($\hat{H}_{simple}$) was extended to the m-spacings estimator ($\hat{H}_{m-spacing}$), we can extend the basic Voronoi and Delaunay entropy estimators to reduce their variance. In one dimension, this was achieved by merely "pasting" together contiguous intervals into an m-spacing, as de£ned by the order statistics of a sample. In D dimensions, we will do this by pasting together multiple Voronoi regions into *Hyper-Voronoi regions* or multiple Delaunay regions into *Delaunay clusters*. Hyper-Voronoi regions for two different distributions are shown in Figure 2.2.

It is tempting to include in a Hyper-Voronoi region any Voronoi region whose center is included in some Euclidean ϵ-ball of a particular point. However, this method of forming Hyper-Voronoi regions gives clusters with many more constituent Voronoi regions in areas of high density than in areas of low density, since low density regions tend to have larger regions. Instead, we desire a technique which gives the same expected number of sub-regions for each Hyper-Voronoi region, irrespective of the underlying density.

To achieve this, we de£ne 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 Voronoi regions (with each edge having weight 1). The Voronoi clustering algorithm then proceeds as follows. m of the N Voronoi regions are chosen at random as Hyper-Voronoi region seeds. Then the Hyper-Voronoi regions are "grown" by adding to them all of the Voronoi regions that are adjacent to them and have not yet been assigned. That is, a Hyper-Voronoi "ball" is formed using the Voronoi adjacency metric. This process is continued (the balls are grown) until all Voronoi regions have been assigned to a Hyper-region. This is essentially the same process by which Voronoi regions themselves are de£ned, only with the adjacency metric rather than with the traditional Euclidean metric. It is for this reason that we call these clusters *Hyper-Voronoi regions*.

This leads to the types of regions shown in Figure 2.2. Note that while the Hyper-Voronoi regions do not have the same number of Voronoi region components, their probability masses will tend to be linear (and hence predictable with low variance) in the number of component Voronoi regions. Such *predictable and locally* assigned probability mass to such regions allows the reliable estimation of functionals of the underlying density, such as entropy.

Assuming now that each Hyper-Voronoi region U^i has probability mass proportional to the number of (£nite volume) regions it is composed of, that this mass is again distributed uniformly, that the number of (£nite volume) Voronoi regions in a Hyper-Region U^i is given by $C(U^i)$, and that $N = \sum_i C(U^i)$, we have the £nal form of our estimator:

$$\hat{H}_{HyperVor} \equiv \sum_{i=1}^{m} \frac{C(U^i)}{N} \log \frac{NA(U^i)}{C(U^i)}. \qquad (9)$$

We can extend this estimator to incorporate overlapping Hyper-Voronoi regions, as in the overlapping m-spacings estimate. We conjecture that these overlapping estimators have similar consistency and convergence properties to the one-dimensional overlapping $m-$spacing estimators, but the proof of these properties is left to future work.

Finally, regarding computational complexity, we note that for £xed dimension, the evaluation of (9) and its overlapping version is polynomial in N. While Hyper-Voronoi regions can be implicitly de£ned in polynomial time, the calculation of their volumes needed for (9) appears, unfortunately, to be exponential in the dimension.

4. REFERENCES

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

[2] J. Beirlant, E. Dudewicz, L. Gyor£, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[3] Barry C. Arnold, N. Balakrishnan, and H.N. Nagaraja, *A First Course in Order Statistics*, John Wiley & Sons, 1992.

[4] O. Vasicek, "A test for normality based on sample entropy," *Journal of the Royal Statistical Society, Series B*, vol. 38, pp. 54–59, 1976.

[5] Edward B. Manoukian, *Modern Concepts and Theorems of Mathematical Statistics*, New York: Springer-Verlag, 1986.

[6] P. Hall, "Limit theorems for sums of general functions of m-spacings," *Math. Proc. Camb. Phil. Soc.*, vol. 96, pp. 517–532, 1984.

[7] B. Ya Levit, "Asymptotically optimal estimation of nonlinear functionals," *Problems of Information Transmission*, vol. 14, pp. 65–72, 1978.

[8] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, Second Edition*, John Wiley & Sons, 1992.