k-nearest neighbor estimation of entropies with confidence

Kumar Sricharan[*], Raviv Raich[+], Alfred O. Hero III[*]

[*]Department of EECS, University of Michigan, Ann Arbor 48109-2122
[+]School of EECS, Oregon State University, Corvallis, Orgon 97331-5501
Email: {kksreddy, hero}@umich.edu, raich@eecs.oregonstate.edu

***Abstract*—We analyze a *k*-nearest neighbor (*k*-NN) class of plug-in estimators for estimating Shannon entropy and Rényi entropy. Based on the statistical properties of *k*-NN balls, we derive explicit rates for the bias and variance of these plug-in estimators in terms of the sample size, the dimension of the samples and the underlying probability distribution. In addition, we establish a central limit theorem for the plug-in estimator that allows us to specify confidence intervals on the entropy functionals. As an application, we use our theory in anomaly detection problems to specify thresholds for achieving desired false alarm rates.**

***Index Terms*—entropy estimation, k-NN density estimation, plug-in estimation, central limit theorem, confidence intervals**

I. Introduction

Shannon entropy ($-\int \log f(x) f(x) dx$) and Rényi entropy ($\frac{1}{1-\alpha} \log \int f^{\alpha}(x) dx,\ \alpha \in (0,1)$) arise in applications of machine learning, signal processing and statistical estimation. Entropy based applications for image matching, image registration and texture classification are developed in [1, 2]. Entropy functional estimation is fundamental to independent component analysis in signal processing [3]. Entropy has also been used in Internet anomaly detection [4] and data and image compression applications [5]. Several entropy based nonparametric statistical tests have been developed for testing statistical models including uniformity and normality [6, 7]. Parameter estimation methods based on entropy have been developed in [8].

In these applications, the entropy must be estimated empirically from sample realizations of the underlying densities. This problem has received significant attention in the mathematical statistics community. Several estimators of Shannon entropy and Rényi entropy have been proposed for general multivariate densities f. These include consistent estimators based on entropic graphs [9, 10], gap estimators [11], nearest neighbor distances [12, 13, 14, 15], Edgeworth approximations [16], convex risk minimization [17] and kernel density estimates [18]. However, general results on rates of convergence of estimators are unavailable. Since the rate of convergence relates the number of samples to the performance of the estimator, convergence rates have great practical utility. In this paper we derive convergence rates for *data-split* versions of k-nearest neighbor (k-NN) estimators of Shannon and Rényi entropies proposed by Goria et.al. [12] and Leonenko et.al. [13] respectively.

The results in this paper improve upon existing results on k-NN estimators available in literature. Goria et.al. [12] and Leonenko et.al. [13] show that the estimators they propose are asymptotically unbiased and consistent. Liitiäinen et.al. [14] provide rates of convergence of the bias of these k-NN estimators. Evans et.al. [19] establish an upper bound on the rates of decay of the variance, while the authors of [9, 10] provide upper bounds on the ℓ_1 rate of convergence.

Our analysis improves on this work by establishing exact rates of decay of the bias and variance of data-split versions of the estimators proposed by Goria et.al. and Leonenko et.al.. Our analysis exploits a close relation between density estimation and the geometry of proximity neighborhoods in the data sample. Finally, while experimental evidence was provided supporting a Gaussian limit for k-NN estimators of Rényi entropy in Leonenko et.al. [13], our theory establishes a CLT for k-NN estimators of arbitrary functionals, including Rényi entropy. We apply these results to derive confidence intervals for Shannon and Rényi entropy.

The reminder of the paper is organized as follows. Section II formulates the problem and introduces the data-split plug-in estimator. The main results concerning the bias, variance and asymptotic distribution of these estimators are stated in Section III and the consequences of these results are discussed. We validate our theory with simulations in Section IV. In Section V, we use our theory to detect anomalies in wireless sensor networks at specified false alarm rate. Conclusions are given in Section VI. Additional details on proofs and results are given in our technical report [20].

II. Preliminaries

Notation: We will use bold face type to indicate random variables and random vectors and regular type face for constants. We denote the expectation operator by the symbol $\mathbb{E}$ and the variance operator as $\mathbb{V}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$. We denote the bias of an estimator by $\mathbb{B}$.

A. Plug-in estimators

We are interested in estimating entropy functionals $G(f)$ of d-dimensional multi-variate densities f with bounded support $\mathcal{S}$, where $G(f)$ has the form

$$G(f) = \int g(f(x))f(x)d\mu(x) = \mathbb{E}[g(f(x))].$$

Here, μ denotes the Lebesgue measure and $\mathbb{E}$ denotes statistical expectation w.r.t density f. We require that the density f be uniformly bounded away from 0 and finite on the support $\mathcal{S}$, i.e., there exist constants ϵ_0, ϵ_∞ such that $0 < \epsilon_0 < \epsilon_\infty < \infty$ such that $\epsilon_0 \leq f(x) \leq \epsilon_\infty$ $\forall x \in \mathcal{S}$. We assume that i.i.d realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ are available from the density f.

The plug-in estimator is constructed using a data splitting approach as follows. The data sample is randomly subdivided into two parts $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ and $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$ of N and M points respectively. In the first stage, we estimate the k-NN density estimator $\hat{\mathbf{f}}$ at the N points $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ using the M realizations $\{\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{N+M}\}$. Subsequently, we use the N samples $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ to approximate the functional $G(f)$ to obtain the plug-in estimator:

$$\hat{\mathbf{G}}(f) \quad = \quad \frac{1}{N}\sum_{i=1}^{N} g(\hat{\mathbf{f}}(\mathbf{X}_i)).$$

Let $d(X, Y)$ denote the Euclidean distance between points X and Y and $\mathbf{d}_X^{(k)}$ denote the Euclidean distance between a point X and its k-th nearest neighbor amongst $\mathbf{X}_{N+1}, .., \mathbf{X}_{N+M}$. The k-NN region is $\mathbf{S}_k(X) = \{Y : d(X, Y) \leq \mathbf{d}_X^{(k)}\}$ and the volume of the k-NN region is $\mathbf{V}_k(X) = \int_{\mathbf{S_k}(X)} dZ$. The standard k-NN density estimator [21] is defined as $\hat{\mathbf{f}}(X) = \frac{k-1}{M\mathbf{V}_k(X)}$.

Let $\hat{\mathbf{H}}$ be the Shannon entropy estimate $\hat{G}(f)$ with the choice of functional $g(x) = -\log(x)$. Let $\hat{\mathbf{I}}_\alpha$ be the estimate of the Rényi α-integral estimate $\hat{G}(f)$ with the choice of functional $g(x) = x^{\alpha-1}$. Define $\tilde{\mathbf{H}} = \hat{\mathbf{H}} + [\log(k-1) - \Psi(k-1)]$ and $\tilde{\mathbf{I}}_\alpha = [(\Gamma(k + (1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}]^{-1}\hat{\mathbf{I}}_\alpha$. Also define the Rényi entropy estimator to be $\tilde{\mathbf{H}}_\alpha = (1-\alpha)^{-1}\log(\tilde{\mathbf{I}}_\alpha)$. We note that the estimators $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_\alpha$ correspond to data-split versions of the Shannon and Rényi entropy estimators of Goria et.al. [12] and Leonenko et.al. [13] respectively.

III. Main results and consequences

The bias of $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_\alpha$ was previously derived by Liitiäinen et.al. [14]. Because $[(\Gamma(k+(1-\alpha))/\Gamma(k))(k-1)^{\alpha-1}] \rightarrow 1$ and $\Psi(k-1) - \log(k-1) \rightarrow 0$ as $k \rightarrow \infty$, the estimators $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_\alpha$ will have identical variance up to leading terms as $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}_\alpha$ respectively. Likewise, $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_\alpha$, when suitably normalized, will converge to the same distribution as the estimators $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}_\alpha$ respectively.

We now state the main theorems corresponding to the bias, variance and asymptotic distribution of $\hat{\mathbf{H}}$ ($g(x) = -\log(x)$) and $\hat{\mathbf{H}}_\alpha$ ($g(x) = x^{\alpha-1}$) and sketch the proofs for these theorems. We assume that k grows logarithmically in M, i.e. $k = \Theta(\log(M))$. We assume that the density f has continuous partial derivatives of the third order. Let $\mathbf{Y}$ denote a random variable with density f and define $c(X) = \Gamma^{(2/d)}((d+2)/2)f^{-2/d}(X)tr[\nabla^2(f(X))]$.

A. Bias and Variance

This theorem on the bias of the estimator is due to Liitiäinen et.al. [14]

Theorem III.1. *The bias of the plug-in estimator $\hat{\mathbf{G}}(f)$ is given by*

$$\begin{aligned} \mathbb{B}(\hat{\mathbf{G}}(f)) \quad = \quad & c_1\left(\frac{k}{M}\right)^{1/d} + c_2\left(\frac{1}{k}\right) \\ & + o\left(\frac{1}{k} + \left(\frac{k}{M}\right)^{1/d}\right), \end{aligned}$$

where c_1 and c_2 are constants which depend on the underlying density f and the choice of functional g only.

Proof: From the work done by Liitiäinen et.al. [14], it follows that

$$\mathbb{E}[\tilde{\mathbf{H}}] \quad = \quad H + c_1(k/M)^{1/d} + o((k/M)^{1/d})$$

$$\mathbb{E}[\tilde{\mathbf{I}}_\alpha] \quad = \quad I_\alpha + c_1(k/M)^{1/d} + o((k/M)^{1/d}).$$

We conclude the proof by observing that $[(\Gamma(k + (1 - \alpha))/\Gamma(k))(k-1)^{\alpha-1}] = 1 + (\alpha(\alpha-1)/2k) + O(1/k^2)$ and $\Psi(k-1) = \log(k-1) - 1/2k + O(1/k^2)$ as $k \rightarrow \infty$. It follows that $c_2 = \mathbb{E}[f^2(\mathbf{Y})g''(f(\mathbf{Y}), \mathbf{Y})/2]$. ■

Theorem III.2. *The variance of the plug-in estimator $\hat{\mathbf{G}}(f)$ is given by*

$$\mathbb{V}(\hat{\mathbf{G}}(f)) \quad = \quad c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right),$$

where $c_4 = \mathbb{V}[g(f(\mathbf{Y}), \mathbf{Y})]$ *and* $c_5 = \mathbb{V}[f(\mathbf{Y})g'(f(\mathbf{Y}), \mathbf{Y})]$.

Proof: Define the set $\mathcal{S}'$ to be the set of points $X \in \mathcal{S}$ whose $2k$-NN ball $\mathbf{S}_{2k}(X)$ lies in the interior of the density. Define

$$\tilde{\mathbf{G}}(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{\mathbf{X}_i \in \mathcal{S}'\}} g(\hat{\mathbf{f}}(\mathbf{X}_i)).$$

We have shown in Appendices B and D, [20] that

$$\mathbb{V}(\tilde{\mathbf{G}}(f)) = c_4\left(\frac{1}{N}\right) + c_5\left(\frac{1}{M}\right) + o\left(\frac{1}{M} + \frac{1}{N}\right).$$

The principal idea in establishing this result involves Taylor series expansions of the functional $g(\hat{\mathbf{f}}(X), X)$ about the true value $g(f(X), X)$, and subsequently using the moment properties of k-NN density estimates.

Observe that $Pr(\mathbf{Y} \in \mathcal{S}') = O((k/M)^{1/d})$. From the work done by Evans et.al. [19], we can state that $\mathbb{V}(\hat{\mathbf{G}}(f) - \tilde{\mathbf{G}}(f)) = O(k^5/M)$. Because $k = \Theta(\log(M))$, we have $O(k^5/M) \times O((k/M)^{1/d}) = o(1/M)$. The theorem follows by using the Cauchy-Schwartz inequality. ■

Our result is an improvement on the results of Evans et.al. in that we are able to provide the exact leading terms for the variance.

B. Central limit theorem

In addition to the results on bias and variance shown in the previous section, we show that our plug-in estimator, appropriately normalized, weakly converges to the normal distribution. We study the asymptotic behavior of the plug-in estimates under the following limiting conditions: (a) $k/M \to 0$, (b) $k \to \infty$, and (c) $N \to \infty$. As shorthand, we will collectively denote the above limiting assumptions by $\Delta \to 0$.

Theorem III.3. *The asymptotic distribution of the normalized plug-in estimator* $\hat{\mathbf{G}}(f)$ *is given by*

$$\lim_{\Delta \to 0} Pr\left(\frac{\hat{\mathbf{G}}(f) - \mathbb{E}[\hat{\mathbf{G}}(f)]}{\sqrt{\mathbb{V}[\hat{\mathbf{G}}(f)]}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha),$$

where $\mathbf{Z}$ *is a standard normal random variable.*

Proof: Define the random variables $\{\mathbf{Y}_{M,i}; i = 1, \ldots, N\}$ for any fixed M as

$$\mathbf{Y}_{M,i} = \frac{g(\hat{\mathbf{f}}(\mathbf{X}_i)) - \mathbb{E}[g(\hat{\mathbf{f}}(\mathbf{X}_i))]}{\mathbb{V}[g(\hat{\mathbf{f}}(\mathbf{X}_i)]}.$$

The key idea here is to recognize that $\mathbf{Y}_{M,i}$ are exchangeable random variables. Blum et.al. [22] showed that for exchangeable 0 mean, unit variance random variables $\mathbf{Z_i}$, the sum $\mathbf{S}_N = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \mathbf{Z}_i$ converges in distribution to $N(0,1)$ if and only if $Cov(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $Cov(\mathbf{Z}_1^2, \mathbf{Z}_2^2) = 0$. In our case,

$$\begin{aligned} Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) &= O(1/M), \\ Cov(\mathbf{Y}_{M,i}^2, \mathbf{Y}_{M,j}^2) &= O(1/M). \end{aligned}$$

As M gets large, we then have that $Cov(\mathbf{Y}_{M,i}, \mathbf{Y}_{M,j}) \to 0$ and $Cov(\mathbf{Y}_{M,i}^2, \mathbf{Y}_{M,j}^2) \to 0$. We then extend the work by Blum et.al. to show that convergence in distribution to $N(0,1)$ holds in our case as both N and M get large. These ideas are rigorously treated in Appendix E, [20]. ■

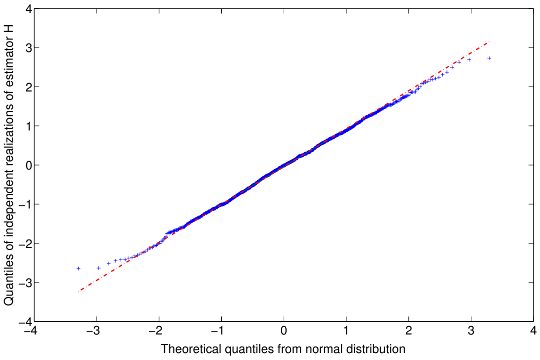

Fig. 1. q-q comparing independent realizations of the normalized Shannon estimator (L.H.S. of Theorem III.3) on the vertical axis to a standard normal population on the horizontal axis. The linearity of the points validates the central limit theorem.

The CLT for k-NN estimators of Rényi entropy was alluded to by Leonenko et.al. [13] by inferring from experimental results. Theorem III.3 formally establishes the CLT for k-NN estimators of Rényi and Shannon entropy.

IV. Simulations

We validate our theory using using the 2 dimensional mixture density $f_m = pf_\beta + (1-p)f_u$; f_β: Beta density with parameters a=4,b=4; f_u: Uniform density; Mixing ratio p = 0.8. Constants $c_i; i = 1, 2..5$ are estimated using Monte-Carlo methods [23].

We show the Q-Q plot of the normalized Shannon entropy estimate and the standard normal distribution in Fig. 1. The linear Q-Q plot validates Theorem III.3 on asymptotic normality of the plug-in estimator. Using the CLT, we plot the 95% confidence intervals for the entropy functional as a function of sample size in Fig. 2.

V. Anomaly detection in networks

We apply our theory to the problem of anomaly detection in wireless sensor networks. The experiment

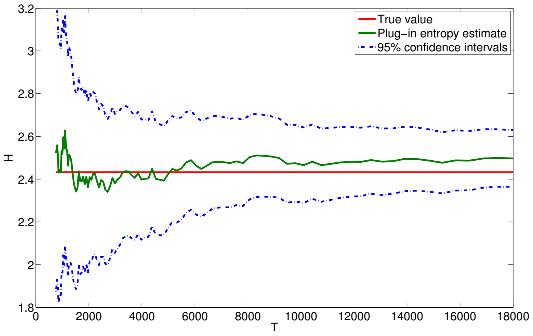

Fig. 2. Predicted confidence intervals on Shannon entropy for varying sample size T using the central limit theorem III.3. The confidence intervals decrease with sample size as expected.

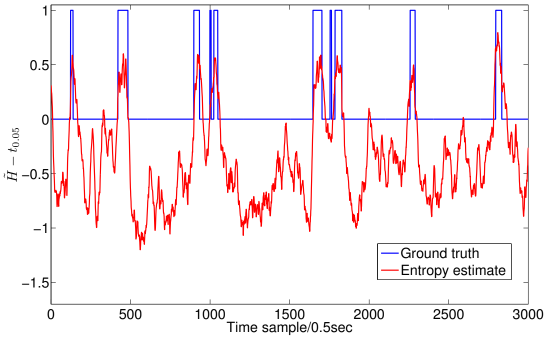

Fig. 3. Entropy estimator $\tilde{H}$ implemented as a scan statistic over time for anomaly detection in wireless ad hoc sensor network experiment. Ground truth indicator function (in blue) indicates when anomalous activity occurred. The entropy estimator detects these anomalies whenever the entropy estimate crosses the level $\alpha = 0.05$ threshold $t_{0.05}$ analytically determined by the CLT in Theorem III.3.

was set up on a Mica2 platform, which consists of 14 sensor nodes randomly deployed inside and outside a lab room. Wireless sensors communicate with each other by broadcasting and the received signal strength (RSS), defined as the voltage measured by a receiver's received signal strength indicator circuit (RSSI), was recorded for each pair of transmitting and receiving nodes. There were 14 × 13 = 182 pairs of RSSI measurements over a 30 minute period, and each sample was acquired every 0.5 sec. During the measuring period, students walked into and out of lab at random times, which caused anomaly patterns in the RSSI measurements. Finally, a web camera was employed to record activity for ground truth.

The mission of this experiment is to use the 182 RSS sequences to detect any intruders (anomalies). To remove the temperature drifts of receivers we pre-process the data by removing their local mean values. Let $y_i[n]$ be the pre-processed n-th sample of the i-th signal and denote $y[n] = (y_1[n], \dots, y_{182}[n])'$.

We now estimate the Shannon entropy for each 1-dimensional, 182 sample sequence $y[n]$ using the estimator $\tilde{\mathbf{H}}$. We detect anomalies by thresholding the entropy estimate $\tilde{H}[n]$. A time sample n is regarded to be anomalous if the entropy estimate $\tilde{H}[n]$ exceeds a specified threshold. We seek to choose the threshold appropriately for achieving a desired false alarm rate.

To this end, we estimate the entropies $\tilde{H}[n]$ for the time instants $n = 1, \dots, 50$ when no anomalies were known to have occurred and subsequently estimate the mean μ and variance σ^2 of the entropy estimates for this nominal time interval $n \in [1, 50]$. Using these estimates of the mean and variance, we use the central limit theorem III.3 to set the threshold t_α for a given false alarm rate α as $t_\alpha = \mu + z_{\alpha/2}\sigma$ where $z_{\alpha/2}$ is the z-score corresponding to coverage $1 - \alpha$. This threshold t_α is then used to detect anomalies at time instants $n > 50$.

Desired and observed false alarm rates						
Desired	.20	.10	.05	.02	.01	.005
Observed	.269	.111	.062	.026	.015	.009

The desired and corresponding observed false alarm rates are shown in the table above. The slightly higher observed false alarm rates can be attributed to the temporal dependence between the RSS sequences at successive time samples. This dependence results in marginally higher entropy estimates at non-anomalous time instants immediately preceding and succeeding anomalous time intervals as compared to entropy estimates at nominal time instants farther away from anomalous activity. This is corroborated by Fig. 3, which shows the ground truth and the normalized entropy estimator response ($\tilde{H}[n] - t_\alpha$ with false alarm rate $\alpha = 0.05$) as a function of time.

ROC curves corresponding to the entropy estimator are shown in Fig. 4 in addition to the ROC curves using the subspace method of Lakhina et.al. [4] and the covariance based estimator of Chen et.al. [24]. It is clear that the detection performance using the entropy estimator is marginally better than the subspace and covariance based methods of Lakhina et.al. and Chen et.al. respectively. The Area under the ROC curves were found to be 0.9784, 0.9722 and 0.9645 for the entropy, covariance and subspace based anomaly detection methods respectively.

VI. Conclusion

We proposed a class of data-split k-NN density plug-in estimators for estimating Shannon and Rényi entropies of densities that are bounded strictly away from 0. We

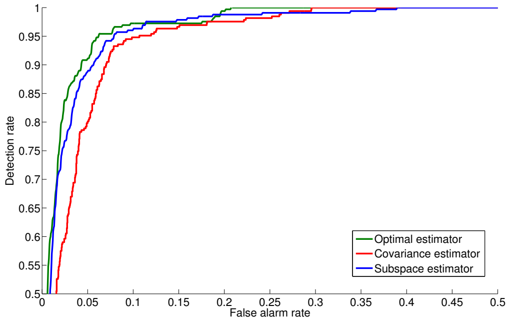

Fig. 4. ROC curves for entropy, covariance and subspace based anomaly detection. The performance of the entropy based method is the best as measured by area under the curve (0.9784 and compared to 0.9722 and 0.9645).

derived the bias, variance and mean square error of the estimator in terms of the sample size, the dimension of the samples and the underlying probability distribution. In addition, we developed a central limit theorem for these estimators and used our theory to specify confidence intervals on the entropy. Finally, we used our entropy estimator to perform anomaly detection in wireless sensor networks and used our asymptotic theory to set thresholds appropriately to achieve specified false alarm rates.

Using the theory presented in the paper, one can specify the minimum necessary sample size required to obtain requisite accuracy in entropy estimates. This in turn can be used to predict and optimize performance in applications like structure discovery in graphical models and dimension estimation for support sets of low intrinsic dimension. See [20] for more details on these applications.

Acknowledgment

This work is partially funded by the Air Force Office of Scientific Research, grant number FA9550-09-1-0471. The authors would like to thank Neal Patwari for designing the experiment and collecting the data.

References

[1] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," vol. 19, no. 5, September 2002.

[2] H. Neemuchwala and A. O. Hero, "Image registration in high dimensional feature space," *Proc. of SPIE Conference on Electronic Imaging, San Jose*, January 2005.

[3] E. G. Miller and J. W. Fisher, "ICA using spacings estimates of entropy," *Proc. 4th Intl. Symp. on ICA and BSS*, pp. 1047–1052, 2003.

[4] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *In ACM SIGCOMM*, 2005, pp. 217–228.

[5] A. Jain, "Image data compression: A review," *Proceedings of the IEEE*, vol. 69, no. 3, pp. 349 – 389, 1981.

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

[7] E. J. Dudewicz and E. C. van der Meulen, "Entropy-based tests of uniformity." *Journal of the American Statistical Association*, vol. 76, pp. 967–974, 1981.

[8] B. Ranneby, "The maximum spacing method. an estimation method related to the maximum likelihood method." *Scandinavian Journal of Statistics*, vol. 11, pp. 93–112, 1984.

[9] A. O. Hero, J. Costa, and B. Ma, "Asymptotic relations between minimal graphs and alpha-entropy," *Technical Report CSPL-334 Communications and Signal Processing Laboratory, The University of Michigan*, March 2003.

[10] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," *ArXiv e-prints*, Mar. 2010.

[11] B. van Es, "Estimating functionals related to a density by class of statistics based on spacing," *Scandinavian Journal of Statistics*, 1992.

[12] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. N. Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Nonparametric Statistics*, 2004.

[13] N. Leonenko, L. Prozanto, and V. Savani, "A class of rényi information estimators for multidimensional densities," *Annals of Statistics*, vol. 36, pp. 2153–2182, 2008.

[14] E. Liitiäinen, A. Lendasse, and F. Corona, "A boundary corrected expansion of the moments of nearest neighbor distributions," *Random Structures and Algorithms*, vol. 37, pp. 223–247, September 2010.

[15] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *Information Theory, IEEE Transactions on*, vol. 51, no. 9, pp. 3064–3074, 2005.

[16] M. M. V. Hulle, "Edgeworth approximation of multivariate differential entropy," *Neural Computation*, vol. 17, no. 9, pp. 1903–1910, 2005.

[17] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions on*, vol. 56, no. 11, pp. 5847 –5861, 2010.

[18] P. B. Eggermont and V. N. LaRiccia, "Best asymptotic normality of the kernel density entropy estimator for smooth densities," *Information Theory, IEEE Transactions on*, vol. 45, no. 4, pp. 1321 –1326, May 1999.

[19] D. Evans, "A law of large numbers for nearest neighbor statistics," *Proceedings of the Royal Society A*, vol. 464, pp. 3175–3192, 2008.

[20] K. Sricharan, R. Raich, and A. O. H. III, "Empirical estimation of entropy functionals with confidence," *ArXiv e-prints*, Dec. 2010.

[21] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, 1965.

[22] J. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher, "Central limit theorems for interchangeable processes," *Canadian Journal of Mathematics*, June 1957.

[23] V. C. Raykar and R. Duraiswami, "Fast optimal bandwidth selection for kernel density estimation," in *Proceedings of the sixth SIAM International Conference on Data Mining*, J. Ghosh, D. Lambert, D. Skillicorn, and J. Srivastava, Eds., 2006, pp. 524–528.

[24] Y. Chen, A. Wiesel, and A. O. Hero, "Robust shrinkage estimation of high-dimensional covariance matrices," submitted to IEEE Trans. on Signal Process., preprint available in arXiv:1009.5331.