ON RENYI'S ENTROPY ESTIMATION WITH ONE-DIMENSIONAL GAUSSIAN KERNELS

Septimia Sarbu

Department of Signal Processing
Tampere University of Technology
PO Box 527 FI-33101 Tampere, Finland
septimia.sarbu@tut.fi

ABSTRACT

Rényi's entropies play a significant role in many signal processing applications. Plug-in kernel density estimation methods have been employed to estimate such entropies with good results. However, they become computationally intractable in higher dimensions, because of the requirement to store intermediate probability density values for a large number of data points. We propose a method to reduce the number of the samples in a plug-in kernel density estimation method for Rényi's entropies of real exponents and to improve the result of the standard plug-in kernel density method. To this end, we derive a univariate estimator, using an Hermite expansion of sums of Gaussian kernels and a hierarchical clustering of the samples. On simulated data from a univariate Gaussian distribution, our method performs better than a k-nearest neigbour algorithm and other kernel density estimation methods.

Index Terms— Rényi's entropy estimation, Gaussian kernels, Hermite expansion, hierarchical clustering

1. INTRODUCTION

The estimation of Rényi's entropy and divergence is an important problem in information theory, because of their role in a wide range of information-theoretic and signal processing applications. Csiszár [1] relates Rényi's entropy to cutoff rates in channel and block coding. Rényi's entropy has been employed to measure the complexity and information content of deterministic signals using their time-frequency function as a probability density of the signal's energy [2]. It also appears in the definition of Rényi's information dimension, which is involved in coding theorems in compressed sensing [3]. Quadratic order Rényi's entropy plays a fundamental role in statistical learning [4]. Rényi's divergence has been applied in clustering [5], in training for time series prediction [6], in blind deconvolution [7] and machine learning [8].

The rich body of work on the estimation of Rényi's and Shannon's entropies in the multivariate case includes methods based on kernel density estimation, k-nearest neighbour distances and Euclidean graphs. As Shannon's entropy is a limiting case of Rényi's α-entropy, when $\alpha \to 1$, we describe prior work on both types of entropy, for completness. A weighted ensemble of kernel density estimation is proposed in [9], to estimate multivariate entropy functionals, such as Shannon's entropy with applications to hypothesis testing and Rényi's entropy with applications to Panter-Dite factor estimation in vector quantization. The information potential is derived as a variation on kernel density estimation in [8], with applications to Rényi's entropy estimation. Wang et. al. [10] introduce a k-nearest neighbour estimator for multidimensional Kullback-Leibler divergence. A different k-nearest neighbour estimator is derived for Rényi's entropy of multidimensional densities in [11]. In the context of manifold learning, Rényi's entropy is estimated using the method of the minimal spanning tree, which is a subset of continuous quasiadditive Euclidean graphs, for multivariate densities [12]. For $\alpha \in (0, 1)$, Rényi's entropy is estimated as the logarithm of the total edge weight of a minimal k-point Euclidean graph created with a greedy algorithm [13]. A minimax rate-optimal functional estimation method is proposed in [14] to estimate Shannon's entropy and Rényi's entropy with $\alpha \in \left(0, \frac{3}{2}\right)$. The number of samples required to estimate Rényi's entropy, for discrete distributions with k symbols, is a particular type of a function of k, according to the order of the entropy, α, [15]: it is superlinear for all $\alpha < 1$, approximately linear for real $\alpha > 1$ that are not integers, and sub-linear for integer $\alpha > 1$.

The accuracy of kernel density methods makes them appealing to estimate Rényi's entropy. However, in higher dimensions, such methods suffer either from the need to store a large number of intermediate probability values or from the need to run a large number of nested for loops to compute such probabilities. As an effort to solve this problem, we develop an estimation algorithm for one-dimensional densities, by improving on the plug-in kernel density estimator. The final aim is to develop a multivariate kernel density estimator, with high accuracy and requiring a lower number of points than the original samples. However, this is beyond the scope of the paper and we focus here on the one-dimensional case. Our aim is to reduce the number of points used for computation and to improve the estimate. To this end, we create a clustering scheme of the samples, which produces an im-

provement in the entropy estimate over the standard plug-in kernel method. Based on an Hermite expansions of the exponential kernels and using hierarchical clustering, we replace the original samples with a smaller number of strategically placed points, such that Rényi's entropy estimate is improved. The derivation starts with a sum of Gaussian kernels, as found in kernel density estimation [16], [17]. We expand each Gaussian kernel into an Hermite infinite sum, similarly to the main idea in the Fast Gauss Transform, introduced by [18] and extended in [19]. After classification, for each cluster of samples, we apply a linear approximation to transform the original sum of the kernels into a new sum using fewer points, which are the cluster centroids. We prove that our estimator tends to the plug-in kernel estimator, as $N \to \infty$, and that the latter estimator converges almost surely to the true entropy. As a result, our estimator is strongly consistent for $\alpha \in \mathbb{R}$.

The paper is organized as follows: in section 2 we provide the mathematical derivations, describe the clustering algorithm and prove consistency results. In section 3, we compare our method with the k-nearest neighbour algorithm [11] and several kernel density estimation methods, in simulation experiments, and we reserve the last section for conclusions.

2. DERIVATION OF THE ESTIMATOR AND ITS PERFORMANCE ANALYSIS

2.1. Mathematical derivations

For continuous probability density functions, Rényi's α-entropy [20] is $R_\alpha(p) = \frac{1}{1-\alpha}\log\int_x p^\alpha(x)\mathrm{d}x$.

We will use the the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}u^2}$ and the kernel density estimator $\hat{p}(x) = \frac{1}{Nh}\cdot\sum_{i=1}^{N} K\left(\frac{x-X_i}{h}\right)$.

$\Rightarrow \hat{p}^\alpha(x) = \frac{1}{\left(Nh\sqrt{2\pi}\right)^\alpha}\cdot\left[\sum_{i=1}^{N}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2}\right]^\alpha$.

Let $S_1(x) = \sum_{i=1}^{N}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2}$ and $I_1 = \int_x S_1^\alpha(x)\,\mathrm{d}x$.

Let N be the number of samples, $h = 1.06\cdot\hat{\sigma}\cdot N^{-0.2}$ the Silverman's rule for kernel bandwidth estimation, $\hat{\sigma}$ a consistent standard deviation estimate, M the number of clusters and $X_{jC}, \forall j = 1:M$, the cluster centroids.

Definition 1. *With the above notations, Rényi's entropy estimator of real order $\alpha \in \mathbb{R}$ is given by*

$$\hat{R}_\alpha(p) = \frac{1}{1-\alpha}\cdot\log\left[\frac{1}{\left(Nh\sqrt{2\pi}\right)^\alpha}\cdot I_1\right] = \frac{1}{1-\alpha}\cdot$$
$$\cdot\log\left[\frac{1}{\left(Nh\sqrt{2\pi}\right)^\alpha}\cdot\int_x\left[\sum_{j=1}^{M}e^{-\frac{1}{2}\left(\frac{x-X_{jC}}{h}\right)^2}\cdot N_j\right]^\alpha\mathrm{d}x\right].$$

The Hermite expansion of an exponential function is defined as [21]: $\mathrm{e}^{xt-\frac{t^2}{2}} = \sum_{n=0}^{+\infty}H_n(x)\cdot\frac{t^n}{n!}$.

$$\Rightarrow \mathrm{e}^{-\frac{1}{2}\left(\frac{x-X_i}{h}\right)^2} = \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2}\cdot\left[\sum_{n=0}^{+\infty}H_n\left(\frac{x}{h}\right)\cdot\frac{\left(\frac{X_i}{h}\right)^n}{n!}\right]$$

$$\Rightarrow S_1 = \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2}\cdot\left[\sum_{i=1}^{N}\sum_{n=0}^{+\infty}H_n\left(\frac{x}{h}\right)\cdot\frac{\left(\frac{X_i}{h}\right)^n}{n!}\right] =$$
$$= \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2}\cdot\left[\sum_{n=0}^{+\infty}H_n\left(\frac{x}{h}\right)\cdot\frac{1}{h^n\cdot n!}\cdot\sum_{i=1}^{N}(X_i)^n\right].$$

Let $S_2 = \sum_{i=1}^{N}(X_i)^n$. We separate the N data points into M clusters, each cluster having N_j number of components, with $\sum_{j=1}^{M}N_j = N, \forall j = 1:M, \Rightarrow S_2 = \sum_{j=1}^{M}\sum_{k=1}^{N_j}(X_{jk})^n$. We create the clusters of samples to be able to use a linear approximation $(1+\gamma)^n \simeq 1+n\cdot\gamma$, with $\gamma \ll 1, \gamma \to 0$. For each cluster $j = 1:M$, we write the points as the centroid, X_{jC}, plus a small linear deviation: $X_{jk} = X_{jC}\cdot(1+\gamma_{jk})$, $\gamma_{jk} \ll 1, \gamma_{jk} \to 0, \forall k = 1:N_j$. Let $\gamma_j = \sum_{k=1}^{N_j}\gamma_{jk}, \forall j = 1:M$. We observe that $\gamma_{jk} = \frac{X_{jk}}{X_{jC}} - 1$ and $X_{jC} = \frac{\sum_{k=1}^{N_j}X_{jk}}{N_j}$

$$\Rightarrow \gamma_j = \sum_{k=1}^{N_j}\left(\frac{X_jk}{X_{jC}} - 1\right) = 0 \Rightarrow S_2 = \sum_{j=1}^{M}N_j\cdot(X_{jC})^n.$$

$$\begin{aligned}\Rightarrow S_2 &= \sum_{j=1}^{M}\sum_{k=1}^{N_j}\left[(X_{jC})^n\cdot(1+\gamma_{jk})^n\right]\\ &\simeq \sum_{j=1}^{M}\sum_{k=1}^{N_j}\left[(X_{jC})^n\cdot(1+n\cdot\gamma_{jk})\right]\\ &= \sum_{j=1}^{M}(X_{jC})^n\cdot\left(\sum_{k=1}^{N_j}1 + n\cdot\sum_{k=1}^{N_j}\gamma_{jk}\right)\\ &= \sum_{j=1}^{M}N_j\cdot(X_{jC})^n. \qquad (1)\end{aligned}$$

$$\begin{aligned}\Rightarrow S_1 &= \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2}\cdot\left(\sum_{n=0}^{+\infty}H_n\left(\frac{x}{h}\right)\cdot\frac{1}{h^n\cdot n!}\cdot S_2\right)\\ &= \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2}\cdot\left\{\sum_{j=1}^{M}\left[\sum_{n=0}^{+\infty}H_n\left(\frac{x}{h}\right)\cdot\frac{N_j\cdot(X_{jC})^n}{h^n\cdot n!}\right]\right\}\\ &= \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2}\cdot\left(\sum_{j=1}^{M}N_j\cdot HE_0\right), \qquad (2)\end{aligned}$$

where we denoted $HE_0 = \sum_{n=0}^{+\infty} \frac{H_n\left(\frac{x}{h}\right)}{n!} \cdot \left(\frac{X_{jC}}{h}\right)^n$. The definition of the Hermite expansion of an exponential expression yields $HE_0 = \mathrm{e}^{\frac{x}{h} \cdot \frac{X_{jC}}{h} - \frac{1}{2} \cdot \left(\frac{X_{jC}}{h}\right)^2}$.

$$\Rightarrow S1 \simeq \mathrm{e}^{-\frac{1}{2}\left(\frac{x}{h}\right)^2} \cdot \sum_{j=1}^{M} N_j \cdot \mathrm{e}^{\frac{x}{h} \cdot \frac{X_{jC}}{h} - \frac{1}{2} \cdot \left(\frac{X_{jC}}{h}\right)^2} = \\ = \sum_{j=1}^{M} N_j \cdot \mathrm{e}^{-\frac{1}{2}\left(\frac{x - X_{jC}}{h}\right)^2}. \quad (3)$$

$\Rightarrow \hat{R}_\alpha(p)$ follows from this result, as stated in definition 1.

2.2. Clustering of the data points

We perform the clustering of the samples with hierarchical clustering [22]. This method offers a division on multiple levels, meaning that clusters are created within clusters. To find one separation of points, we need to specify a distance threshold, such that all the elements within this distance from the cluster centroid belong to this collection. We use γ from the linear approximation, as a distance measure between points instead of the more widely used Euclidean distance. For each cluster j, we write the elements as $X_{jk} = X_{jC} \cdot (1 + \gamma_{jk}), \gamma_{jk} \ll 1, \gamma_{jk} = \frac{X_{jk} - X_{jC}}{X_{jC}}$. We create the distance matrix between any two data points, X_i, X_j, as $\gamma_{ij} = abs[\frac{X_j - X_i}{X_i}]$.

Since information about the data source is not available, we will employ a data-driven technique of selecting the clustering threshold. We use ideas similar to cross-validation. We start with the plug-in kernel density estimate, applied on the entire data set of size N. We compute 10 additional Rényi's entropy estimates, using different clustering thresholds. The final Rényi's estimate will be the mean of these 11 values. The procedure of selecting the partitioning of the data is similar to a 10-fold cross-validation scheme: we select uniformly at random a subset of $N-100$ samples from our data vector. We use a grid search to find the best clustering threshold, such that the optimization criterion is minimized. The optimization criterion is the standard deviation of the vector of Rényi's entropy estimates up to the current iteration. The lower limit of the grid search is $C_{thr_L} = \frac{h}{N}$, the step is equal to $\delta_{thr} = \frac{C_{thr_L}}{10}$ and the upper limit is equal to $C_{thr_U} = C_{thr_L} \cdot 50$. We initialize a vector R_E with the plug-in kernel density value, as mentioned above. At every step $i = 2 : 11$, we memorize two arrays of values: R_E that consists of the previous $i-1$ estimates and R_C that stores the entropy estimates given by the grid of clustering thresholds described above, for the current i^{th} set. The clustering threshold at the i^{th} iteration will be selected such that its corresponding element of R_C provides the minimum standard deviation for the newly formed vector of R_E. This element will be added to R_E. The final entropy estimate will be the mean over R_E.

2.3. Asymptotic analysis of the estimator

We prove that, in the limit $N \to \infty$, our estimator from definition 1 (renamed R_{hc} in the simulation experiments) tends to the plug-in kernel density estimator of Rényi's entropy, R_{kde}. We prove that R_{kde} is strongly consistent, i.e. converges almost surely to the true value, using Theorem 3.1 from [23] and the continuous mapping theorem [24]. As a result, R_{hc} converges almost surely, i.e. is strongly consistent. Let $h_N = h = 1.06 \cdot \hat{\sigma} \cdot N^{-0.2}$, where $\hat{\sigma}$ is a consistent estimator of the standard deviation of the unknown distribution. We used the built-in Matlab function for its estimation.

Theorem 1. *The estimator $R_{hc} \to R_{kde}$, as $N \to \infty$.*

Proof. We have that $0 \leq h_N \leq 1$ and $\lim_{N\to\infty} h_N = 0$.

$$\Rightarrow \lim_{N\to\infty} C_{thr_L} = \lim_{N\to\infty} \frac{h_N}{N} = 0 \\ \Rightarrow \lim_{N\to\infty} \delta_{thr} = 0 \text{ and } \lim_{N\to\infty} C_{thr_U} = 0. \quad (4)$$

This implies that the number of clusters $M \to N$ or, equivlently $\gamma_{ik} \to 0, \forall j = 1 : M, k = 1 : N_j$, from the linear approximation. This means that we obtain the original plug-in kernel density estimate, R_{kde}. In the 10-fold cross-validation scheme, all the elements become equal to R_{kde}, as well as the initial element. As a result, the mean of this vector will be equal to the plug-in kernel estimate. Thus, $R_{hc} \to R_{kde}$, as $N \to \infty$. □

We apply the theory presented in [23], to prove that R_{kde} converges almost surely. With their notation, the unknown probability density is denoted by $f(x)$ and

$$T(f) = \int_{\mathbb{R}} \Phi(f(x))\mathrm{d}x, k = 0, \Phi(x) = x^\alpha, \alpha \in \mathbb{R} \quad (5)$$

and its kernel density estimator is

$$\hat{f}_{h_N} = \frac{1}{N} \sum_{i=1}^{N} K_{h_N}(x - X_i), \forall x \in \mathbb{R}, \\ K_{h_N}(x - X_i) = \frac{1}{h_N} K(\frac{x - X_i}{h_N}). \quad (6)$$

Then, Rényi's entropy becomes $R_\alpha(f) = \log T(f)$ and $R_{kde} = R_\alpha(\hat{f}_{h_N}) = \log T(\hat{f}_{h_N})$.

Theorem 2. *The plug-in kernel density estimator is strongly consistent, i.e. $R_{kde} \to R_\alpha(f)$ a.s., as $N \to \infty$, where a.s. stands for almost surely.*

Proof. In our simulation experiments, the probability density function is a Gaussian distribution, which satisfies conditions $(F.i - F.iii)$ from [23]. The proofs of this section are valid for any probability density f that satisfies conditions $(F.i - F.iii)$, not only for Gaussian distributions. The kernel

function K is a Gaussian distribution, which satisfies conditions $(K.i) - (K.v)$ of [23]. The function $\Phi(x) = x^\alpha, \alpha \in \mathbb{R}$ satisfies conditions $(\Phi.i) - (\Phi.ii)$ of [23]. The second derivative of Φ is equal to $\Phi''(x) = \alpha \cdot (\alpha - 1) \cdot x^{\alpha-2}$. As the probability density f and the kernel K are bounded, the domain of Φ is bounded, i.e $\Phi : [a_\Phi \ b_\Phi] \to \mathbb{R}$, $\Phi(x) = x^\alpha$, $\alpha \in \mathbb{R} \Rightarrow \sup \Phi''$ is bounded, i.e. $\sup \Phi'' \leq B_\Phi$. Thus, condition $(\Phi.iii)$ is fulfilled.

In addition, we have that $\lim_{N\to\infty} \frac{\log N}{N \cdot h_N} = 0$. Thus, all the conditions of Theorem 3.1 of [23] are satisfied and this yields $T(f) \to T(\hat{f}_{h_N})$ a.s., as $N \to \infty$. Then, as the logarithm is a continuous function, by the continuous mapping theorem [24], $R_{kde} \to R_\alpha(f)$ a.s., as $N \to \infty$. □

3. SIMULATION RESULTS

We perform simulation experiments with data generated from a univariate Gaussian distribution with mean $\mu = 5$ and variance $\sigma = 1$. We compare the newly derived estimator R_{hc} with the theoretical Rényi's entropy for a Gaussian distribution, R_{th}, with the plug-in kernel density estimator, R_{kde}, with the information potential estimator [8], R_a, and with the k-nearest neighbour estimator [11] with $k = 25$, R_{knn}. The theoretical value can be easily derived and is equal to $R_{th} = \frac{1}{2}\log 2\pi + \log\sqrt{\sigma} + \frac{1}{2\cdot(\alpha-1)}\log\alpha$. The Rényi's entropy estimates are denoted by R and the empirical mean by E.

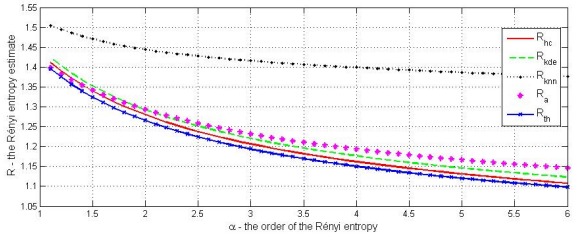

Fig. 1. *R in the order of best to worst: R_{th} - line with x marker (-x-), R_{hc} - solid line (-), R_{kde} - dashed line (–), R_a - no line with * marker (*) and R_{knn} - dotted line with dot marker (...), $\forall\alpha \in \mathbb{R}$.*

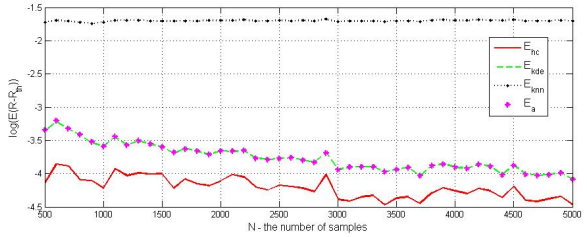

Fig. 2. *The logarithm of the empirical mean error, $\log E(R - R_{th})$, in the order of best to worst: E_{hc} - solid line (-), E_{kde} - dashed line (–), E_a - no line with * marker (*) and E_{knn} - dotted line with dot marker (...), $\forall\alpha \in \mathbb{R}$.*

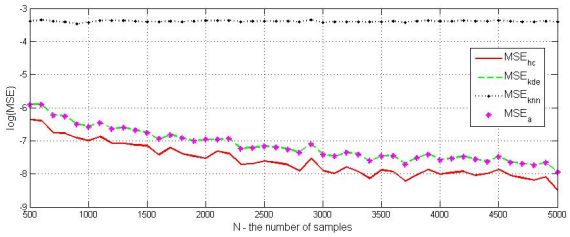

Fig. 3. *The logarithm of the empirical MSE, $\log E(R - R_{th})^2$, in the order of best to worst: MSE_{hc} - solid line (-), MSE_a - no line with * marker (*), MSE_{kde} - dashed line (–) and MSE_{knn} - dotted line with dot marker (...), $\forall\alpha \in \mathbb{R}$.*

In figure 1, we present the empirical mean of the estimators, as a function of $\alpha \in \mathbb{R}$. The size of the data sets is $N = 1000$ samples. The values are averaged over 100 experiments. Our method is the best performing algorithm.

In figure 2, we show the logarithm of the empirical mean error between the entropy estimates and the theoretical value, as a function of the sample size, N, in the case $\alpha = 2$. In figure 3, we present the logarithm of the empirical mean-square error between the entropy estimates and the theoretical value (MSE), as a function of the sample size, N, in the case $\alpha = 2$. The results are averaged over 100 experiments. We selected to display the logarithm of these values, to make the lines clearly visible in the figures, because the difference between the kernel density methods is much smaller than that of these methods and the k-nearest neighbour algorithm. Our method produces the smallest mean error and mean-squared error. These errors are decreasing, as the sample size N is increasing.

4. CONCLUSIONS

We have successfully obtained a univariate strongly consistent estimator for Rényi'e entropy of real order, using Gaussian kernels. We have reduced the number of samples required for the estimation and we have improved the plug-in kernel density estimate. Our experiments reveal that the newly proposed method is the best performing algorithm. As the sample size increases, our estimator tends to the plug-in kernel density estimator, which, in turn, tends to the true entropy. The improvement on the plug-in kernel density estimator takes place for smaller number of samples. This is beneficial in applications where the number of data points is limited. As an extension to higher dimensions, we would derive a multivariate Hermite expansion of the exponentials and then apply the hierarchical clustering of the samples.

5. ACKNOWLEDGEMENTS

We would like to thank anonymous reviewers for helpful comments and for suggesting references [12], [10], [9], [15].

6. REFERENCES

[1] Imre Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, January 1995.

[2] Richard G. Baraniuk, Patrick Flandrin, Augustus J.E.M. Janssen, and Olivier J.J. Michel, "Measuring time-frequency information content using the Rényi entropies," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1391–1409, May 2001.

[3] Yihong Wu and Sergio Verdú, "Rényi information dimension: fundamental limits of almost lossless analog compression," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3721–3748, August 2010.

[4] R.A. Morejon and J.C. Principe, "Advanced search algorithms for information-theoretic learning with kernel-based estimators," *IEEE Transactions on Neural Networks*, vol. 15, no. 4, pp. 874–884, July 2004.

[5] E. Gokcay and J.C. Principe, "Information theoretic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 158–171, February 2002.

[6] D. Erdogmus and J.C. Principe, "Generalized information potential criterion for adaptive system training," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1035–1044, September 2002.

[7] D. Erdogmus, K.E. Hild, J.C. Principe, M. Lazaro, and I. Santamaria, "Adaptive blind deconvolution of linear channels using Rényi's entropy with Parzen window estimation," *IEEE Transactions on Signal Processing*, vol. 52, no. 6, pp. 1489–1498, June 2004.

[8] J.C. Principe, M. Jordan (series editor), R. Nowak (series editor), and B. Schölkopf (series editor), *Information Theoretic Learning: Rényi's entropy and kernel perspectives*, Springer, Second edition, 2010.

[9] Kumar Sricharan, Dennis Wei, and Alfred O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4374–4388, July 2013.

[10] Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, May 2009.

[11] N. Leonenko, L. Pronzato, and V. Savani, "A class of Rényi information estimators for multidimensional densities," *The Annals of Statistics*, vol. 36, no. 5, pp. 2153–2182, 2008.

[12] Jose A. Costa and Alfred O. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2210–2221, August 2004.

[13] Alfred O. Hero and Olivier J.J. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1921–1938, September 1999.

[14] Jiantao Jiao, Kartik Vankat, Yanjun Han, and Tsachy Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, May 2015.

[15] Acharya J., Orlitsky A., Suresh A.T., and Tyagi H., "The complexity of estimating Rényi entropy," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2015)*, 4-6 January 2015.

[16] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[17] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[18] L. Greengard and J. Strain, "The fast Gauss transform," *SIAM Journal on Scientific and Statistical Computing*, vol. 12, no. 1, pp. 79–94, 1991.

[19] Yang C., Duraiswami R., Gumerov N.A., and Davis L., "Improved fast Gauss transform and efficient kernel density estimation," in *Proceedings of the 2003 ninth IEEE International Conference on Computer Vision*, 13-16 October 2003, vol. 1, pp. 664–671.

[20] Rényi A., "On measures of entropy and information," in *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1961, vol. 1, pp. 547–561.

[21] F.W.J. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark (Editors), *NIST Handbook of Mathematical Functions*, Cambridge University Press, 2010.

[22] The MathWorks, "hierarchical clustering," in *http://se.mathworks.com/help/stats/hierarchical-clustering.html*, 1994-2015.

[23] David M. Mason, Elizbar Nadaraya, and Grigol Sokhadze, "Integral functionals of the density," *Institute of Mathematical Statistics Collections*, vol. 7, pp. 153–168, 2010.

[24] H.B. Mann and A. Wald, "On stochastic limit and order relationships," *The Annals of Mathematical Statistics*, vol. 14, no. 3, pp. 217–226, 1943.