WEIGHTED k-NN GRAPHS FOR RÉNYI ENTROPY ESTIMATION IN HIGH DIMENSIONS

Kumar Sricharan, Alfred O. Hero III

Department of EECS, University of Michigan, Ann Arbor, MI 48109

{kksreddy,hero}@umich.edu,

ABSTRACT

Rényi entropy is an information-theoretic measure of randomness which is fundamental to several applications. Several estimators of Rényi entropy based on k-nearest neighbor (k-NN) based distances have been proposed in literature. For d-dimensional densities f, the variance of these Rényi entropy estimators of f decay as $O(M^{-1})$, where M is the sample size drawn from f. On the other hand, the bias, because of the curse of dimensionality, decays as $O(M^{-1/d})$. As a result the bias dominates the mean square error (MSE) in high dimensions. To address this large bias in high dimensions, we propose a weighted k-NN estimator where the weights serve to lower the bias to $O(M^{-1/2})$, which then ensures convergence of the weighted estimator at the parametric rate of $O(M^{-1/2})$. These weights are determined by solving a convex optimization problem. We subsequently use the weighted estimator to perform anomaly detection in wireless sensor networks.

***Index Terms*—** Rényi entropy estimation, weighted k-NN graphs, curse of dimensionality, parametric convergence rate

1. INTRODUCTION

In information theory, Rényi entropy is a generalization of Shannon entropy and is used to quantify the randomness of a system. Rényi entropy finds use in several applications in signal processing and statistical estimation. Entropy based applications for image matching, image registration and texture classification are developed in [1]. Entropy functional estimation is fundamental to independent component analysis in signal processing [2]. Entropy has also been used in Internet anomaly detection [3] and data and image compression applications [4]. Several entropy based nonparametric statistical tests have been developed for testing statistical models including uniformity and normality [5]. For further applications, see Leonenko et.al. [6].

In many of these applications, the entropy has to be empirically estimated from sample realizations of the underlying density. Several estimators of Rényi entropy have been proposed, including the k-NN estimators of Leonenko et.al. [6] and the entropic graph estimators of Hero et.al. [7].

Formally, the Rényi α entropy of a d-dimensional density f with support $\mathcal{S}$ is defined as $H_\alpha(f) = (1-\alpha)^{-1} I_\alpha$ where $I_\alpha = \int_{\mathcal{S}} f^\alpha(x)dx$. We consider the problem of estimating $I_\alpha(f)$ from M i.i.d realizations $\{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$. The k-NN based estimator $\hat{I}_{M,k,\alpha}$ of Leonenko et.al. [6] is given by

$$\hat{I}_{M,k,\alpha} = \frac{1}{M}\sum_{i=1}^{M} C_{k,k+1-\alpha}(c_d(M-1)(r_{k,M-1}^{(i)})^d)^{1-\alpha},$$

where $C_{a,b} = \Gamma(a)/\Gamma(b)$ and $r_{k,M-1}^{(i)}$ is the k-th nearest neighbor distance from $\mathbf{X}_i$ to some other sample $\mathbf{X}_j$ and c_d is the unit ball volume in d dimensions.

Leonenko et.al. showed that the estimator $\hat{I}_{M,k,\alpha}$ is consistent. Liitiäinen et.al. [8] then showed that the bias of this estimator is of order $O(M^{-1/d})$ while the variance is of order $O(M^{-1})$. For moderate to large dimensions d, the contribution of the bias therefore dominates the MSE. To partially address this problem, Liitiäinen et.al. considered a weighted k-NN estimator with reduced bias of $o(M^{-1/d})$ and variance of $O(M^{-1})$. In this paper, we extend Liitiäinen et.al.'s work by determining weights which will reduce the bias of the weighted estimator to $O(M^{-1/2})$.

The rest of this paper is organized as follows. In Section 2, we discuss Liitiäinen et.al.'s weighted k-NN estimator. In Sections [3-4], we analyze bias and variance of the weighted k-NN estimator and subsequently solve a convex optimization problem to determine weights which will lower the bias to $O(M^{-1/2})$ and thereby equalize the contribution of the bias and variance to the MSE. In Section 5, we show that the estimator, when suitably normalized, converges asymptotically to $\mathbb{N}(0,1)$. We show simulation results that illustrate the effectiveness of the proposed method in Section 6. We then apply our proposed estimator to detection of anomalies in wireless sensor networks. We give our conclusions in Section 7.

2. LINEAR WEIGHTED ESTIMATOR

Define the linear weighted k-NN graph estimator $\hat{I}^w_{M,k,\alpha}$ with weight vector $w = \{w(l)\}, l = \{1, \ldots, k\}$ as

$$\hat{I}^w_{M,k,\alpha} = \sum_{l=1}^{k} w_l \hat{I}_{M,l,\alpha}.$$

Liitiäinen et.al. show that the bias of the estimator $\hat{I}_{M,k,\alpha}$ is given by $Bias(\hat{I}_{M,k,\alpha}) = r_k M^{-1/d} + o(M^{-1/d})$ where $r_k = \kappa C_{(k+1-\alpha),(k+1-\alpha+1/d)}$ and κ is a constant which depends on the underlying density f. They then suggest choosing weight

Acknowledgement: This work is partially funded by the Air Force Office of Scientific Research, grant number FA9550-09-1-0471.

vector $w_f = \{w_f(l)\}$, $l = \{1, \ldots, k\}$ with minimum l_2 norm that satisfies $\sum_{l=1}^{k} w_f(l) = 1$ and $\sum_{l=1}^{k} w_f(l) C_{l,\alpha} = 0$. In theory, the bias of $\hat{I}^{w_f}_{M,k,\alpha}$ for this choice of weights will indeed reduce to $o(M^{-1/d})$ while the variance will continue to decay at the rate $O(M^{-1})$.

There are two issues with the proposed correction - (i) in simulations, we found the bias of the weighted estimator to increase for small to moderate sample sizes; and (ii) even though the bias is reduced to $o(M^{-1/d})$, this can continue to be much greater in comparison to the variance. In the next section, we explain why the bias increases for Liitiäinen et.al.'s weighted estimator. We then suggest an improvement that will reduce bias to $O(M^{-1/2})$.

3. ANALYSIS OF BIAS AND VARIANCE

Using the theory we have developed in [9] on boundary compensated k-NN graphs [10], we can *extend* the results of Liitiäinen et.al. as follows. We can show that for densities which are strictly bounded away from 0 on their support and are $\lfloor d/2 \rfloor + 1$ times continuously differentiable, the bias and variance of $\hat{I}_{M,k,\alpha}$ (defined with respect to boundary compensated k-NN graphs) are given by

$$\begin{aligned} Bias(\hat{I}_{M,k,\alpha}) &= \sum_{i \in \mathcal{I}} c_i \left(\frac{k}{M}\right)^{i/d} + o\left(\sqrt{\frac{k}{M}}\right), \\ Var(\hat{I}_{M,k,\alpha}) &= c_v \left(\frac{1}{M}\right) + o\left(\frac{1}{M}\right), \end{aligned}$$

where c_i, c_v are constants that depend on the underlying density f and $\mathcal{I} = \{1, \ldots, (\lfloor d/2 \rfloor + 1)\}$. Using these results it is easily shown that the bias of the weighted graph estimator $\hat{I}^{w}_{M,k,\alpha}$ is given by

$$Bias(\hat{I}^{w}_{M,k,\alpha}) = \sum_{i \in \mathcal{I}} c_i \gamma_w(i) M^{-i/d},$$

where $\gamma_w(i) = \sum_{l=1}^{k} w(l) l^{i/d}$. We note that for the choice of weight vector w_f of Liitiäinen et.al. , $\gamma_{w_f}(1) = 0$. If the magnitude of the weight coefficients $\{w_f(l)\}$ are large, then the coefficients in the bias expansion $\gamma_{w_f}(i)$, $i > 1$ will be quite large as well. This can therefore result in increased bias for Liitiäinen et.al.'s estimator for moderate sample sizes.

Denote the standard deviation of $\hat{I}_{M,i,\alpha}$ by $\sigma_i = \sqrt{c_v/M} + o(1/\sqrt{M})$. Also denote the covariance between $\hat{I}_{M,i,\alpha}$ and $\hat{I}_{M,j,\alpha}$ by σ_{ij}. We can then bound the variance of the weighted estimator $\hat{I}^{w}_{M,k,\alpha}$ using Cauchy-Schwartz as follows

$$\begin{aligned} Var(\hat{I}^{w}_{M,k,\alpha}) &= Var(\sum_{l=1}^{k} w_l \hat{I}_{M,l,\alpha}) \\ &= \sum_{l=1}^{k} w^2(l)\sigma_l^2 + \sum_{l,m=1}^{k} 1_{\{l \neq m\}} w(l) w(m) \sigma_{l,m} \\ &\leq \sum_{l=1}^{k} w^2(l)\sigma_l^2 + \sum_{l,m=1}^{k} 1_{\{l \neq m\}} |w(l) w(m)| \sigma_l \sigma_m \\ &= \left(\sum_{l=1}^{k} |w(l)| \sigma_l\right)^2 = \frac{||w||_1^2 c_v}{M} + o\left(\frac{1}{M}\right). \end{aligned} \quad (3.1)$$

4. OPTIMAL WEIGHT SELECTION

Using the results presented in Sec. 3 on the bias and variance, we seek a weight vector w that (i) ensures that the bias of the weighted estimator is $O(M^{-1/2})$ and (ii) has minimum possible l_1 norm $||w||_1$ in order to reduce the contribution of the higher order terms in the bias and to reduce the variance of the weighted estimator. Let w_o be the solution to the optimization problem

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & ||w||_1 \\ \text{subject to} \quad & \gamma_w(0) = 1, \\ & |\gamma_w(i) M^{-i/d}| \leq \epsilon, \ i \in \mathcal{I}. \end{aligned} \quad (4.1)$$

The optimization problem defined above is convex. Because we are minimizing the l_1 norm of w, for moderately large values k of the length of the weight vector w, the solution to the optimization problem will be sparse [11].

Furthermore, the norm of the solution w_o monotonically decreases with increasing ϵ. We therefore seek to choose ϵ to be as large as possible while ensuring that the squared bias of the optimized weighted k-NN graph estimator $\hat{I}^{w_o}_{M,k,\alpha}$ will be of the same order $O(M^{-1})$ as the variance. Thus we choose $\epsilon = \sqrt{c_0/M}$ where c_0 is a bound on c_v/c_i^2, $i \in \mathcal{I}$.

5. CLT

In this section, we prove that, for any choice of weight vector w, the weighted estimator of Section 2 converges in distribution to $\mathbb{N}(0,1)$ when suitably normalized. In Appendix E, [9] we show that if the interchangeable processes $\{\mathbf{Y}_i\}$, $i = \{1, \ldots, M\}$ satisfies the conditions (i) $\mathbb{E}[\mathbf{Y}_i] = 0$; (ii) $Var[\mathbf{Y}_i] = 1$; (iii) $Cov[\mathbf{Y}_i, \mathbf{Y}_j] = o(1)$, and (iv) $Cov[\mathbf{Y}_i^2, \mathbf{Y}_j^2] = o(1)$, then $(1/\sqrt{M}) \sum_{i=1}^{M} \mathbf{Y}_i$ converges in distribution to $\mathbb{N}(0,1)$. Define

$$\mathbf{Z}_i = \sum_{l=1}^{k} w(l) C_{l,(l+1-\alpha)} (c_d (M-1) (r^{(i)}_{l,M-1})^d)^{1-\alpha},$$

and let $\mathbf{Y}_i = (\mathbf{Z}_i - \mathbb{E}[\mathbf{Z}_i])/\sqrt{Var[\mathbf{Z}_i]}$. We see that $\mathbf{Y}_i$ is indeed an interchangeable process which trivially satisfies conditions (i) and (ii). Condition (iii) follows directly from (3.1). It

is similarly possible to establish (iv) using the moment properties of $\hat{I}_{M,k,\alpha}$ [9] and the Cauchy-Schwartz inequality. We then have

$$\lim_{M\to\infty} Pr\left(\frac{\hat{I}^w_{M,k,\alpha} - \mathbb{E}[\mathbf{Z}_1]}{\sqrt{Var[\mathbf{Z}_1]/M}} \leq \alpha\right) = Pr(\mathbf{Z} \leq \alpha),$$

where $\mathbf{Z}$ is a standard normal random variable. The above result can be used to specify confidence intervals on I_α using the estimate $\hat{I}^w_{M,k,\alpha}$.

6. SIMULATIONS

We will compare the MSE for four different choices of weight vectors: The nearest neighbor estimator of Leonenko et.al. with weight $w_s = [1, 0, \ldots, 0]$, the uniform weighted estimator with weight $w_u = (1/k)[1, \ldots, 1]$, the first-order correction estimator of Liitiäinen et.al. with weight w_f, and finally the optimized weighted estimator with weight w_o. We estimate entropy for the following class of densities: 6 dimensional mixture density $f_m(p, a, b) = pf_\beta(a, b) + (1-p)f_u$; f_β: Beta density with parameters a,b; f_u: Uniform density; Mixing ratio p. In particular, we show representative results obtained by simulating samples from two densities - (i) $f_m(.8, 2, 2)$ and (ii) $f_m(0.8, 1.5, 1.5)$. The MSE error performances for these densities are shown in Fig. 1(a) and Fig. 1(b) respectively.

The observed MSE performance can be explained as follows. The performance of Liitiäinen et.al.'s first-order correction estimator is worse than Leonenko's estimator for small sample sizes, which is in agreement with our theory in Section 3.

For the density $f_m(.8, 2, 2)$, we note that the higher order co-efficients in the bias expansion c_i, $i > 2$ are identically 0. In this case, the MSE performance for the optimized weighted entropy estimator is better than Liitiäinen et.al.'s first-order correction estimator for small sample sizes because the first-order correction estimator does not account for second order terms in the bias.

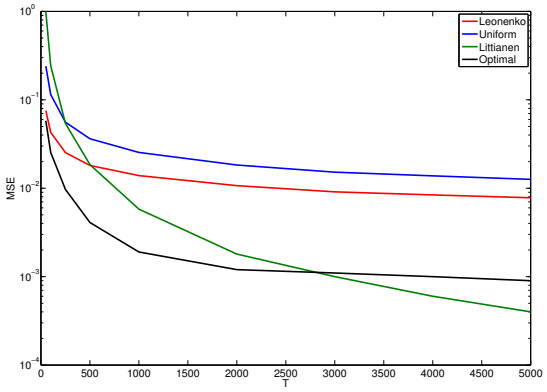
However, with increasing sample size, the contribution of the second order bias terms become negligible in comparison to the first order terms. For large sample sizes, the bias is therefore dominated by the first order terms. Because the first order bias term is explicitly set to 0 in the first-order correction estimator as compared to $\epsilon > 0$ in the optimized estimator case, the first-order correction estimator performs better with increasing sample size M as compared to the optimized weighted estimator.

On the other hand, for the density $f_m(.8, 1.5, 1.5)$, higher order co-efficients are non-zero and therefore contribute to bias. The optimized weighted estimator with higher order bias correction and lower norm therefore works better in this case for all sample sizes.

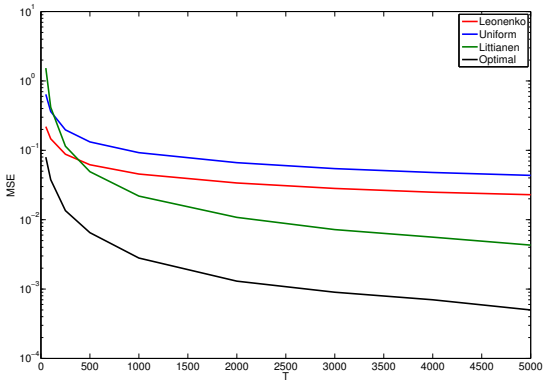
6.1. Anomaly detection

We apply our theory to the problem of anomaly detection in wireless sensor networks. The experiment 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 $\times$ 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 capture the temporal dependency between successive measurements, for each time point we form a temporal dependency discriminant by considering vectors of $d = 3$ successive time samples at each sensor and estimating the entropy by averaging over $M = 182$ spatial samples. We note that the ground truth indicator is only for



(a) MSE comparison for density $f_m(.8, 2, 2)$. A lower order bias correction suffices for this density. The optimized weighted estimator outperforms other estimators for small sample sizes, while the first-order correction estimator of Liitiäinen et.al. works better for larger sample sizes.



(b) MSE comparison for density $f_m(.8, 1.5, 1.5)$. This density requires higher order bias correction. The optimized weighted entropy estimator therefore has superior MSE performance for all sample size regimes.

Fig. 1. Comparison of MSE of weighted estimators for different choices of weight vectors.

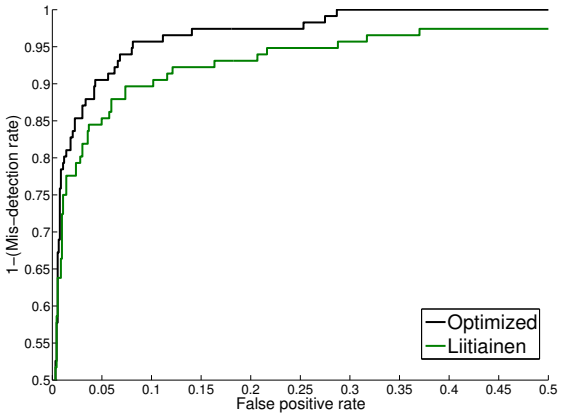


Fig. 2. Comparison of ROC curves for anomaly detection. The optimized weighted entropy estimator outperforms Liitiäinen et.al.'s first-order correction estimator.

evaluating the detecting performance and the detection scheme presented here is conducted in a completely unsupervised manner.

In order to detect anomalies, we form a running estimate of the Rényi α-entropy with $\alpha = 0.95$, of the 3-dimensional time sequence using weighted k-NN estimators with first order correction weight w_f and optimized correction weight w_o. With the choice of $k = 5$, the first order correction weight w_f was found to be $[9.4907, -1.9047, -11.0134, -7.8171, 12.2444]$ and the optimized correction weight w_o was found to be $[0.9568, 1.1795, 0.7278, -0.2381, -1.6261]$. In accordance with our theory, the l_1 norm of w_f ($||w_f||_1 = 42.47$) was found to be much higher than the l_1 norm of w_o ($||w_o||_1 = 4.728$).

We perform anomaly detection by thresholding the entropy estimate. A time sample is regarded to be anomalous if the entropy estimate exceeds a specified threshold. ROC curves corresponding to first-order correction weight w_f and optimized correction weight w_o are shown in Fig. 2. The Area under the ROC curve (AUC) was found to be 0.9538 and 0.9821 for the first-order correction estimator and the optimized weighted estimator respectively. It is clear that the detection performance using the optimized weight w_o is superior to the performance using Liitiäinen et.al.'s first-order correction weight w_f.

7. CONCLUSION

When implemented in high dimension, k-NN based estimators of Rényi entropy suffer from large bias. To address this issue, we have proposed a weighted k-NN graph estimator with an optimized weight vector determined by solving a convex optimization problem. The resulting weighted estimator has reduced bias of order $O(M^{-1/2})$ and converges at the parametric rate of $O(M^{-1/2})$. We have also established weak convergence of the suitably normalized weighted estimator to a normal random variable.

Our weighted estimator is an improvement over the estimator of Liitiäinen et.al. in that our selection of the weight vector has a smaller norm while simultaneously providing for better bias correction. The smaller norm ensures that the variance of the weighted estimator is lower and the contribution of higher order terms in the bias is negligible even for small sample sizes.

We illustrate superior performance of our weighted estimator via simulations. We also apply our proposed estimator to the problem of anomaly detection in sensor networks. Experimental results show that our proposed estimator outperforms Liitiäinen et.al.'s first-order correction estimator.

References

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

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

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

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

[5] D.V. Gokhale, "On entropy-based goodness-of-fit tests," *Computational Statistics and Data Analysis*, vol. 1, pp. 157 – 165, 1983.

[6] 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.

[7] 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.

[8] 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.

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

[10] K. Sricharan, R. Raich, and A. O. Hero, "Boundary compensated k-nn graphs," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, May 2010.

[11] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution," *Manuscript, Department of Statistics, Stanford University*, September 2004.