ON THE STATISTICAL ESTIMATION OF RÉNYI ENTROPIES

Elia Liitiäinen, Amaury Lendasse and Francesco Corona

Helsinki University of Technology
Department of Information and Computer Science
P.O. Box 5400, 02015 HUT, Finland
elia.liitiainen@hut.fi

ABSTRACT

Estimating entropies is important in many fields including statistical physics, machine learning and statistics. While the Shannon logarithmic entropy is the most fundamental, other Rényi entropies are also of importance. In this paper, we derive a bias corrected estimator for a subset of Rényi entropies. The advantage of the estimator is demonstrated via theoretical and experimental considerations.

1. INTRODUCTION

Entropy is a measure of randomness originating from information theory [1]. While many possible definitions of entropy exist, it is well-known that the so-called Shannon logarithmic entropy is the only one that satisfies a set of rather intuitive axioms [1, 2]. Consequently, Shannon entropy is the most natural from information theoretic point of view and it has turned out to be an important concept for science in general [3]. However, it was shown in [2] that if one of the underlying axioms is relaxed, entropy is no longer uniquely defined and a whole family of possible measures of randomness exist. Such quantities are called Rényi entropies.

While Rényi entropies do not necessarily have as strong a theoretical background as the Shannon entropy, they have nevertheless turned out to be useful in many fields (e.g. [4, 5, 6]). If deep information theoretic properties are not needed, a Rényi entropy is usually a valid choice for applications. Moreover, there exists many situations where the Shannon entropy is not the only one with an intuitive interpretation.

In this paper, we consider estimating Rényi entropies from an independent identically (i.i.d) distributed sample $(X_i)_{i=1}^M$. Denoting the common density by p, this amounts to estimating

$$\frac{1}{1-\beta}\log\int p(x)^\beta dx$$

for some $\beta > 0$. We will first restrict ourselves to $0 < \beta < 1$ and the Shannon entropy $\beta = 1$ (understood as the limit $\beta \to 1$) is discussed in the last section. This is due to the fact that $\beta > 1$ is currently not covered by the theory behind our approach (even if an extension seems possible).

While various approaches for the estimation of Rényi entropies exist, the work closest to ours are estimators based on random graphs [4, 7]. Our proposal relies heavily on the well-known connection between nearest neighbor distances and information theory [7]. The estimator is based on our theoretical work on the boundary effect [8], which allows removing the error term caused by the boundary cut-off.

The paper is divided into three major parts. In Sections 2-4 we discuss the theory behind our proposal for $0 < \beta < 1$ and relation to earlier work. After that, simulations are made to demonstrate the practical usefulness of the theoretical considerations. Interestingly, the experiments indicate that the boundary corrected estimator is accurate even in cases, where our theoretical assumptions do not hold.

In Section 6, we consider extending the technique to estimating the Shannon entropy ($\beta = 1$) using a similar boundary correction as the one employed for Rényi entropies. We will show that a straightforward extension is possible, even though its practical validity is not clear. Unfortunately, a boundary correction for logarithmic distances does not currently exist in the literature (albeit it is straightforward to derive) and thus the discussion at this part is not yet on a completely rigorous basis.

2. RÉNYI ENTROPY

Many learning and data-analysis problems involve an observed independent identically (i.i.d) distributed sample

$$(X_i)_{i=1}^M$$

with the common probability density p on a state-space $\mathcal{X} \subset \Re^n$. In this paper, we consider estimating information theoretic properties of the density p using the sample $(X_i)_{i=1}^M$. More specifically, we consider finite sample approximations to the Rényi entropy

$$H_\beta(p) = \frac{1}{1-\beta}\log\int_{\mathcal{X}} p(x)^\beta dx \qquad (1)$$

for $\beta > 0$. Moreover, even though it is possible to choose $\beta > 1$, we have to restrict ourselves to the case $\beta < 1$ (for $\beta = 1$ see Section 6) because of the assumptions made by us in [8]. As a remark, the case $\beta > 1$ is definitely of interest too (see e.g. [6]), but to handle it a slight extension to the results in [8] would be needed. Possibly the most important entropy (except for the Shannon entropy $\beta = 1$) is the one with $\beta = 0.5$ as $H_{0.5}(p)$ has a straightforward connection to the Hellinger distance between p and the uniform density.

3. THE CONNECTION OF RÉNYI ENTROPY AND NEAREST NEIGHBOR DISTANCES

The definition of the nearest neighbor of a point is based on the use of a proximity measure to determine similarity between points. The index of the nearest neighbor of the point X_i in Euclidean metric is

$$N[i,1] = \text{argmin}_{1 \le j \le M, j \neq i} \|X_i - X_j\|.$$

The k-th nearest neighbor is defined recursively by

$$N[i,k] = \text{argmin}_{1 \le j \le M, j \notin \{i, N[i,1], \ldots, N[i,k-1]\}} \|X_i - X_j\|,$$

that is, the closest point after removal of the preceeding neighbors. The corresponding distances are defined as $d_{i,k,M} = \|X_i - X_{N[i,k]}\|$. We also set

$$\delta_{M,k,\alpha} = \frac{1}{M} \sum_{i=1}^{M} d_{i,k,M}^{\alpha}, \tag{2}$$

which is the empirical α-moment for the distances to the k-th nearest neighbor. A well-known result in random geometry [9, 10, 11, 12] states that asymptotically $\delta_{M,k,\alpha}$ has a rather simple form for large M:

Theorem 1. *Under appropriate regularity conditions (see [10, 12]), we have the convergence ($\alpha > 0$)*

$$M^{\alpha/n} E[\delta_{M,k,\alpha}] \to V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)}{\Gamma(k)} \int_{\mathcal{X}} p(x)^{1-\alpha/n} dx \tag{3}$$

when $M \to \infty$ with everything else fixed. Here V_n denotes the volume of the unit ball in $\Re^n$ and Γ refers to the Gamma function.

An elegant aspect of Equation (3) is its connection to the Rényi entropy (1):

$$\begin{aligned} H_{1-\alpha/n}(p) &\approx \hat{H}_{1,1-\alpha/n}(p) \\ &= \frac{n}{\alpha} \log\left(\frac{M^{\alpha/n} V_n^{\alpha/n}}{\Gamma(1+\alpha/n)} \delta_{M,1,\alpha}\right), \end{aligned} \tag{4}$$

where the substitution of $M^{\alpha/n}\delta_{M,1,\alpha}$ to $E[M^{\alpha/n}\delta_{M,1,\alpha}]$ is known to be accurate for a large M [7, 13]. This simple estimate is essentially the state of the art in nearest neighbor entropy estimation [7]. While there exists various works on the consistency of the estimator, the systematic error was not well-understood until the following theorem was proven by us in [8] characterizing the difference of the left and right hand side in the limit (3). For information on technicalities related to the proof, see [8]. Some basic knowledge of manifolds is required, as the second term involves an integral over the boundary manifold $\partial\mathcal{X}$.

Theorem 2. *Suppose that $\mathcal{X}$ is an open and bounded set and the boundary $\partial\mathcal{X}$ is a twice differentiable compact manifold. Moreover, assume that the density function p is twice continuously differentiable on $\mathcal{X}$ and strictly above zero. Then we have for large M (with $k \ge 1$, $n \ge 2$ and $\alpha > 0$ fixed)*

$$\begin{aligned} E[\delta_{M,k,\alpha}] &= V_n^{-\alpha/n} \frac{\Gamma(k+\alpha/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n)} \int_{\mathcal{X}} p(x)^{1-\alpha/n} dx + \\ &+ (D - V_n^{-\alpha/n-1/n}) \frac{\Gamma(k+\alpha/n+1/n)\Gamma(M)}{\Gamma(k)\Gamma(M+\alpha/n+1/n)} \times \\ &\times \int_{\partial\mathcal{X}} p(x)^{1-\alpha/n-1/n} dS + o(M^{-\alpha/n-1/n}), \end{aligned} \tag{5}$$

where the second term in the right hand side contains the integral of $p(x)^{1-\alpha/n-1/n}$ over the boundary $\partial\mathcal{X}$. The constant D does not depend on k or M. The remainder term $o(M^{-\alpha/n-1/n})$ approaches zero faster than $M^{-\alpha/n-1/n}$ and thus it is neglible for large M.

To understand the result properly, it is essential to observe that ($\gamma > 0$)

$$\frac{\Gamma(M)}{\Gamma(M+\gamma)} = M^{-\gamma} + O(M^{-\gamma-1});$$

in practice the approximation is usually very accurate.

The derivation of our novel Rényi entropy estimator in the next section relies on Theorem 2. For practical purposes, the conditions of Theorem 2 are restrictive, even though in [8] the theorem is shown to hold also if $\mathcal{X}$ is a polytope. However, it is necessary for a good estimator to handle probability distributions with boundaries as otherwise it does not perform well even for simple models like uniformly distributed points.

To demonstrate the need for improvement, we state that Theorem 2 applied to Equation (4) together with a Taylor expansion yields for a fixed $0 < \alpha < n$,

$$E[\hat{H}_{1,1-\alpha/n}(p)] = H_{1-\alpha/n}(p) + cM^{-1/n} + o(M^{-1/n}), \tag{6}$$

where the constant c does not depend on M. The term $o(M^{-1/n})$ approaches zero faster than $M^{-1/n}$ and thus it is neglible for a large M. Consequently, on expectation the estimator $\hat{H}_{1,1-\alpha/n}(p)$ has an error term which depends on

$M^{-1/n}$ and it can be verified that when $n \geq 3$ such a term converges very slowly to zero with respect to M. This implies that the estimator $\hat{H}_{1,1-\alpha/n}(p)$ is sensitive to the curse of dimensionality.

4. THE NOVEL BOUNDARY CORRECTED ESTIMATOR

The theoretical result in Equation (5) is not only useful in the bias analysis of existing nearest neighbor estimators, but it also suggests an elegant bias correction to improve the rate of convergence with respect to M. To see how such a correction can be implemented, let us choose the scalar weights $(w_k)_{k=1}^l$ in such a way that

$$\sum_{k=1}^{l} \frac{w_k \Gamma(k+\alpha/n)}{\Gamma(k)} = V_n^{\alpha/n} \tag{7}$$

and

$$\sum_{k=1}^{l} \frac{w_k \Gamma(k+\alpha/n+1/n)}{\Gamma(k)} = 0. \tag{8}$$

Such weights can always be chosen if $l > 2$. Now with this choice, we have (keeping everything except M fixed) by an application of the expansion (5),

$$\frac{\Gamma(M+\alpha/n)}{\Gamma(M)} E[\sum_{k=1}^{l} w_k \delta_{M,k,\alpha}] = \int_{\mathcal{X}} p(x)^{1-\alpha/n} dx + o(M^{-1/n}).$$

The boundary term vanishes in the weighted average. This motivates our proposal for the estimation of Rényi entropies:

$$\hat{H}_{2,1-\alpha/n}(p) = \frac{n}{\alpha} \log(\sum_{k=1}^{l} w_k \delta_{M,k,\alpha}) \quad (\text{for } n \geq 2).$$

The advantage compared to $\hat{H}_{1,1-\alpha/n}(p)$ is characterized by

Theorem 3. *With weights satisfying Equations (7) and (8), we have under the conditions of Theorem 2*

$$E[\hat{H}_{2,1-\alpha/n}(p)] = H_{1-\alpha/n}(p) + o(M^{-1/n}).$$

The error term $o(M^{-1/n})$ goes to zero faster than $M^{-1/n}$.

The proof of Theorem 3 is a straightforward application of Theorem 2 and linearization of logarithm at 1. A comparison to Equation (6) reveals that our estimator $\hat{H}_{2,1-\alpha/n}(p)$ will have a lower bias for large M for a large class of densities, while it shares the same consistency properties as the k-nearest neighbor estimator [7].

Our considerations hold only for expectation values of the estimators. Obviously to analyze consistency and rate of convergence, it is necessary to examine the variance too. However, the variance of local random variables is rather well-understood due to [13] and [14], where a law of large numbers is proven in a generality which covers the estimators analyzed in this paper. Based on those results, we can state that

$$E[\hat{H}_{2,1-\alpha/n}(p)^2] - E[\hat{H}_{2,1-\alpha/n}(p)]^2 \leq cM^{-1}$$

for a constant c which does not depend on M. A comparison to Equation (6) reveals that the standard deviation is asymptotically neglible compared to the bias of $\hat{H}_{1,1-\alpha/n}(p)$ when $n \geq 3$. Even though it is outside the scope of this paper, it would be of significant interest to characterize the variance of $\hat{H}_{2,1-\alpha/n}$ more accurately, because that would allow choosing $(w_k)_{k=1}^l$ in an optimal way.

5. SIMULATIONS

In our simulations, we compare our weighted estimator

$$\hat{H}_{2,1-1/n}(p)$$

to the original estimator

$$\hat{H}_{1,1-1/n}(p).$$

The goal is to show that the theoretical considerations have a practical significance and performance in the class of probability distributions with boundaries is improved.

When $l > 2$, the weights $(w_k)_{k=1}^l$ are not uniquely determined by Equations (7) and (8). Of the possible solutions, we simply choose the one with the smallest norm

$$\sqrt{\sum_{k=1}^{l} w_k^2}. \tag{9}$$

As the choice of the weights $(w_k)_{k=1}^l$ depends on l as well, there is another degree of freedom that needs to be addressed. Based on some initial experimentation and the magnitude of the norm (9), we decided to set $l = n + 2$ in our implementation, as it was found to be a good value when estimating the $1 - 1/n$ entropy with $n \leq 6$. The intuition behind examining the norm is that if the weights take large values, the estimator is bound to have a high variance.

5.1. Uniform Distribution

In the first experiment, we analyze uniformly distributed points in the cube $[0, 1]^n$, where n increases from 3 to 6 (the case $n \leq 2$ was not of interest to us, as we believe that

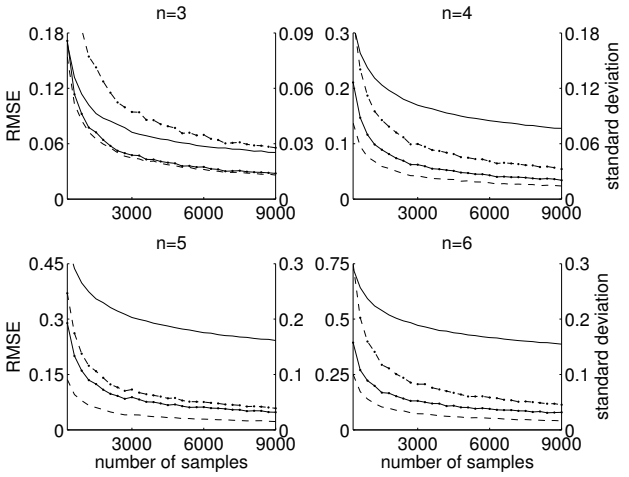

Fig. 1. Rényi entropy: RMSE and standard deviations for the experiment with uniform data. The solid lines correspond to the RMSE axis, while the two dashed lines present the standard deviations of the estimators using the second y-axis. In both cases, the lines with dots point to the bias corrected estimator.

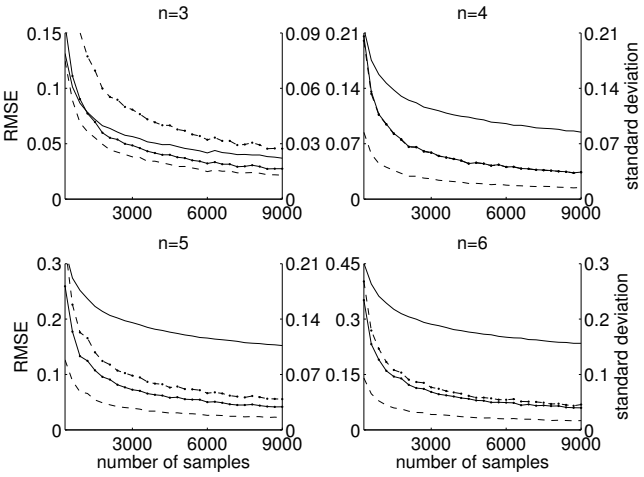

Fig. 2. Rényi entropy: RMSE and standard deviations for the experiment with truncated Gaussian data. Similarly as in Figure 1, solid lines point to RMSE and dots to the bias corrected estimator.

$\hat{H}_{1,1-1/n}(p)$ works usually just fine in that case) and M grows up to 9000 in steps of 300. As can be easily verified

$$H_{1-1/n}(p) = 0.$$

The uniform distribution can be viewed as the distribution with maximal randomness in the unit cube. The results of the simulations are drawn in Figure 1. We have computed estimates for the root mean square error (RMSE)

$$\sqrt{E[(H_{1-1/n}(p) - \hat{H}_{i,1-1/n}(p))^2]} \qquad (i = 1, 2)$$

by averaging over 1000 realizations to estimate the expectation inside the square root. In addition, Figure 1 includes estimates of the standard deviations

$$\sqrt{E[\hat{H}_{i,1-\alpha/n}(p)^2] - E[\hat{H}_{i,1-\alpha/n}(p)]^2} \qquad (i = 1, 2).$$

We can see that the proposed method is more accurate, even though there is a clear increase in standard deviation.

5.2. Truncated Gaussian

The experiment involves both boundaries and correlation between components introducing an additional degree of difficulty. In this case, the variables $(X_i)_{i=1}^M$ are samples from the multivariate normal distribution restricted to the unit ball. Again, the true value of the entropy is rather straightforwardly computable. In Figure 2, we have drawn the standard deviations and RMSEs averaged over 1000 realizations. Again, the bias corrected estimator is significantly more accurate.

5.3. The Gaussian Distribution

In the last experiment, we experimented with normally distributed data. Because normal distribution is invariant with respect to orthonormal rotations, there is no point in including correlations between the components of the random vectors. However, we wanted to test how the estimators behave when different components of X_i have different variances. To achieve this goal, the components were scaled so as to set the variance of the first to 0.5, the second to 0.75, third to 1 and so on (in total n components with $3 \leq n \leq 6$).

Interestingly, the Gaussian distribution is not covered by our theoretical analysis as it does not have boundaries. From Figure 3, we can see that both $\hat{H}_{1,1-1/n}(p)$ and $\hat{H}_{2,1-1/n}(p)$ are accurate, which is expected as the estimation problem is significantly easier in the absence of cut-offs. However, for large sample sizes $\hat{H}_{2,1-1/n}(p)$ is more accurate whereas for smaller sample sizes it is actually worse due to the increase in variance (in practice, normalization as a preprocessing step might alleviate the problem). The result is interesting as a faster rate of convergence for $\hat{H}_{2,1-1/n}(p)$ has not been theoretically established in this case.

6. ON THE EXTENSION TO THE SHANNON ENTROPY

It is natural to ask, whether a boundary correction exists for the Shannon entropy

$$-\int_{\mathcal{X}} p(x) \log p(x) dx$$

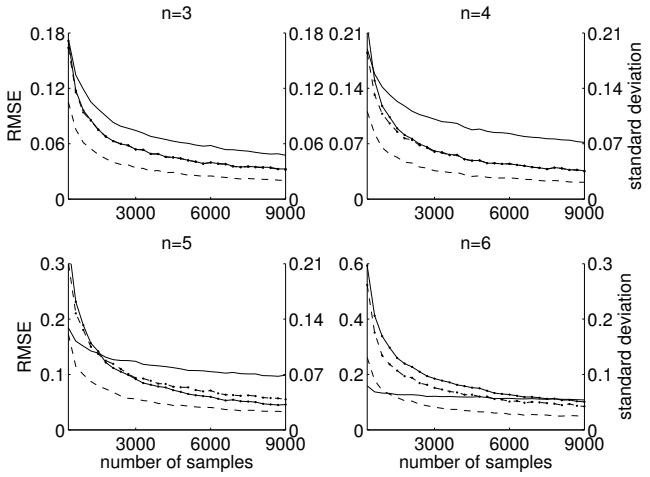

Fig. 3. Rényi entropy: RMSE and standard deviations for the experiment with Gaussian data. Solid lines point to RMSE and dots to the bias corrected estimator.

as well. As stated earlier, the logarithmic entropy is related to the Rényi entropies by the limit

$$-\int_{\mathcal{X}} p(x)\log p(x)dx = \lim_{\beta\to 1} H_\beta(p).$$

The following theorem shows that in principle an analogous boundary correction is possible. The proof is rather complicated and it is omitted here, but the reader can informally verify that the result is obtained by taking the derivative of Equation (5) with respect to α, even though such an observation is by no means valid as a rigorous proof. The actual proof is essentially the same as that of Theorem 2 with the exception that the α-moment is replaced by logarithm.

Theorem 4. *Under the assumptions of Theorem 2, we have*

$$E[\log d_{1,k}] = C_1 \frac{\Gamma(M)}{\Gamma(M+1/n)}\int_{\partial\mathcal{X}} p(x)^{1-1/n}dS - \\ - n^{-1}\int_{\mathcal{X}} p(x)\log p(x)dx + C_2 + o(M^{-1/n}).$$

The variables C_1 and C_2 are functions of M and k:

$$C_1(M,k) = \frac{V_n^{-1/n}\Gamma(k+1/n)\log V_n}{n\Gamma(k)} + \\ + \frac{V_n^{-1/n}\phi(M+1/n)\Gamma(k+1/n)}{n\Gamma(k)} + D_1(\phi(k+1/n) - \\ - \frac{\Gamma(k+1/n)\phi(M+1/n)}{\Gamma(k)}) - \frac{V_n^{-1/n}\phi(k+1/n)}{n} + \\ + \frac{D_2\Gamma(k+1/n)}{\Gamma(k)}$$

$$C_2(M,k) = \frac{1}{n}(\phi(k) - \phi(M) - \log V_n),$$

where the constants D_1 and D_2 depend only on n. ϕ refers to the digamma function.

Now if we choose the weights $(w_k)_{k=1}^l$ in such a way that

$$\sum_{k=1}^{l} w_k = n$$
$$\sum_{k=1}^{l} w_k \frac{\Gamma(k+1/n)}{\Gamma(k)} = 0$$
$$\sum_{k=1}^{l} w_k \phi(k+1/n) = 0,$$

then Theorem 4 implies that

$$\sum_{k=1}^{l} w_k E[\log d_{1,k}] = -\int_{\mathcal{X}} p(x)\log p(x)\,dx + o(M^{-1/n})$$

and terms of order $M^{-1/n}$ disappear. Thus the estimator

$$-\int_{\mathcal{X}} p(x)\log p(x)\,dx \approx \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{l} w_k \log d_{i,k} - \\ - \sum_{k=1}^{l} w_k C_2(M,k) \qquad (10)$$

would seem to have a similar speed of convergence with respect to M as the Rényi entropy estimators $\hat{H}_{2,1-\alpha/n}$.

Even if the extension seems theoretically straightforward, some practical difficulties arise due to the fact that the norm

$$\sqrt{\sum_{k=1}^{l} w_k^2} \qquad (11)$$

tends to be large. Intuitively, this may lead to high variance unless l is large. To assess the validity of the estimate (10), the experiment in Section 5.2 was repeated for the logarithmic entropy with the heuristic choice $l = 20$ for $n = 3, 4$ and $l = 30$ for $n = 5, 6$; similarly as before, among possible choices for the weights $(w_k)_{k=1}^l$, the ones with the minimal norm (11) are chosen. The result is drawn in Figure 4

As expected, an asymptotic improvement is obtained. On the other hand, especially when $n > 4$, the small sample behavior of (10) is not very good. Moreover, the increase in standard devitation has a significant impact on accuracy even for large M.

7. CONCLUSIONS

A bias-corrected method for the estimation of Rényi entropies was proposed. The estimators are based on weighting k-nearest neighbor estimators with the choice of weights

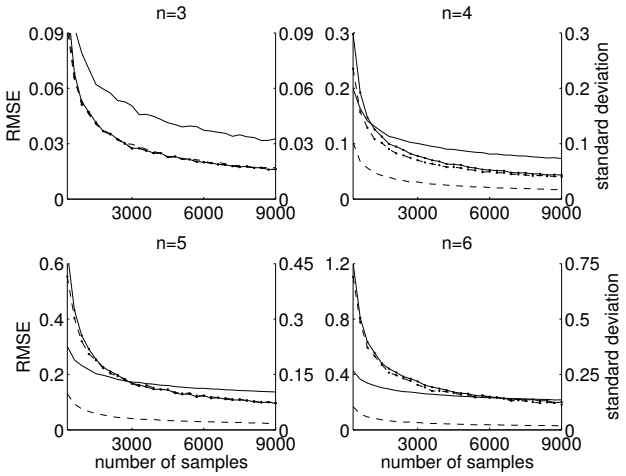

Fig. 4. Shannon entropy: RMSE and standard deviations for the experiment with truncated Gaussian data. Solid lines point to RMSE and dots to the bias corrected estimator (10).

based on theoretical considerations. The simulations show that in practice, an improvement in accuracy is obtained especially for large sample sizes. Because the theoretical assumptions were rather restrictive, further theoretical work is still needed to understand the properties of the method.

As a topic of future research, we state examining the asymptotic variance of the proposed estimators. It is possible that the weights should be chosen so as to minimize the asymptotic variance. This seems to be especially relevant for the estimation of the Shannon entropy. Finally, possibly the most important contribution of this work is the possibility for higher order bias corrections.

8. REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[2] A. Rényi, "On measures of information and entropy," in *Proceedings of the 4th Berkeley Symposium on Mathematics*, 1960, pp. 547–561.

[3] L. Brillouin, *Science and Information Theory*, Academic Press, 1962.

[4] H. Neemuchwala, A.O. Hero, and P.L. Carson, "Image matching using alpha-entropy measures and entropic graphs," *Signal Processing (Special Issue on Content-based Visual Information Retrieval)*, vol. 85, pp. 277–296, 2005.

[5] A.G. Bashkirov, "Renyi entropy as a statistical entropy for complex systems," *Theoretical and Mathematical Physics*, vol. 149, no. 2, pp. 1559–1573, 2006.

[6] D. Erdogmus and J.C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Transaction On Signal Processing*, vol. 50, no. 7, pp. 1780–1786, 2002.

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

[8] Elia Liitiäinen, Francesco Corona, and Amaury Lendasse, "A boundary corrected expansion of the moments of nearest neighbor distributions," Tech. Rep. TKK-ICS-R9, Helsinki University of Technology, TKK Reports in Information and Computer Science, Espoo, Finland, 2008.

[9] C. Redmond and J.E. Yukich, "Limit theorems and rates of convergence for euclidean functionals," *Annals of Applied Probability*, vol. 4, no. 4, pp. 1057–1073, 1994.

[10] D. Evans, A. Jones, and W. M. Schmidt, "Asymptotic moments of near neighbour distance distributions," *Proceedings of the Royal Society A*, vol. 458, no. 2028, pp. 2839–2849, 2008.

[11] M. D. Penrose and J. E. Yukich, "Weak laws of large numbers in geometric probability," *Annals of Applied Probability*, vol. 13, no. 1, pp. 277–303, 2003.

[12] A. R. Wade, "Explicit laws of large numbers for random nearest-neighbour type graphs," *Advances in Applied Probability*, vol. 39, no. 2, pp. 326–342, 2007.

[13] D. Evans, "A law of large numbers for nearest neighbour statistics," *Proceedings of the Royal Society A*, vol. 464, no. 2100, pp. 3175–3192, 2008.

[14] M. D. Penrose, "Gaussian limits for random geometric measures," *Electronic Journal of Probability*, vol. 12, pp. 989–1035, 2007.