Estimation of Rényi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs

Dávid Pál
Department of Computing Science
University of Alberta
Edmonton, AB, Canada
dpal@cs.ualberta.ca

Barnabás Póczos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
poczos@ualberta.ca

Csaba Szepesvári
Department of Computing Science
University of Alberta
Edmonton, AB, Canada
szepesva@ualberta.ca

Abstract

We present simple and computationally efficient nonparametric estimators of Rényi entropy and mutual information based on an i.i.d. sample drawn from an unknown, absolutely continuous distribution over $\mathbb{R}^d$. The estimators are calculated as the sum of p-th powers of the Euclidean lengths of the edges of the 'generalized nearest-neighbor' graph of the sample and the empirical copula of the sample respectively. For the first time, we prove the almost sure consistency of these estimators and upper bounds on their rates of convergence, the latter of which under the assumption that the density underlying the sample is Lipschitz continuous. Experiments demonstrate their usefulness in independent subspace analysis.

1 Introduction

We consider the nonparametric problem of estimating Rényi α-entropy and mutual information (MI) based on a finite sample drawn from an unknown, absolutely continuous distribution over $\mathbb{R}^d$. There are many applications that make use of such estimators, of which we list a few to give the reader a taste: Entropy estimators can be used for goodness-of-fit testing (Vasicek, 1976; Goria et al., 2005), parameter estimation in semi-parametric models (Wolsztynski et al., 2005), studying fractal random walks (Alemany and Zanette, 1994), and texture classification (Hero et al., 2002b,a). Mutual information estimators have been used in feature selection (Peng and Ding, 2005), clustering (Aghagolzadeh et al., 2007), causality detection (Hlaváckova-Schindler et al., 2007), optimal experimental design (Lewi et al., 2007; Póczos and Lőrincz, 2009), fMRI data processing (Chai et al., 2009), prediction of protein structures (Adami, 2004), or boosting and facial expression recognition (Shan et al., 2005). Both entropy estimators and mutual information estimators have been used for independent component and subspace analysis (Learned-Miller and Fisher, 2003; Póczos and Lőrincz, 2005; Hulle, 2008; Szabó et al., 2007), and image registration (Kybic, 2006; Hero et al., 2002b,a). For further applications, see Leonenko et al. (2008); Wang et al. (2009a).

In a naïve approach to Rényi entropy and mutual information estimation, one could use the so called "plug-in" estimates. These are based on the obvious idea that since entropy and mutual information are determined solely by the density f (and its marginals), it suffices to first estimate the density using one's favorite density estimate which is then "plugged-in" into the formulas defining entropy

and mutual information. The density is, however, a nuisance parameter which we do *not* want to estimate. Density estimators have tunable parameters and we may need cross validation to achieve good performance.

The entropy estimation algorithm considered here is *direct*—it does not build on density estimators. It is based on k-nearest-neighbor (NN) graphs with a fixed k. A variant of these estimators, where each sample point is connected to its k-th nearest neighbor only, were recently studied by Goria et al. (2005) for Shannon entropy estimation (*i.e.* the special case $\alpha = 1$) and Leonenko et al. (2008) for Rényi α-entropy estimation. They proved the *weak* consistency of their estimators under certain conditions. However, their proofs contain some errors, and it is not obvious how to fix them. Namely, Leonenko et al. (2008) apply the generalized Helly-Bray theorem, while Goria et al. (2005) apply the inverse Fatou lemma under conditions when these theorems do not hold. This latter error originates from the article of Kozachenko and Leonenko (1987), and this mistake can also be found in Wang et al. (2009b).

The first main contribution of this paper is to give a correct proof of consistency of these estimators. Employing a very different proof techniques than the papers mentioned above, we show that these estimators are, in fact, *strongly* consistent provided that the unknown density f has bounded support and $\alpha \in (0, 1)$. At the same time, we allow for more general nearest-neighbor graphs, wherein as opposed to connecting each point only to its k-th nearest neighbor, we allow each point to be connected to an arbitrary subset of its k nearest neighbors. Besides adding generality, our numerical experiments seem to suggest that connecting each sample point to all its k nearest neighbors improves the rate of convergence of the estimator.

The second major contribution of our paper is that we prove a finite-sample high-probability bound on the error (*i.e.* the rate of convergence) of our estimator provided that f is Lipschitz. According to the best of our knowledge, this is the very first result that gives a rate for the estimation of Rényi entropy. The closest to our result in this respect is the work by Tsybakov and van der Meulen (1996) who proved the root-n consistency of an estimator of the *Shannon* entropy and only in *one* dimension.

The third contribution is a *strongly* consistent estimator of Rényi mutual information that is based on NN graphs and the empirical copula transformation (Dedecker et al., 2007). This result is proved for $d \geq 3$ [1] and $\alpha \in (1/2, 1)$. This builds upon and extends the previous work of Póczos et al. (2010) where instead of NN graphs, the minimum spanning tree (MST) and the shortest tour through the sample (*i.e.* the traveling salesman problem, TSP) were used, but it was only conjectured that NN graphs can be applied as well.

There are several advantages of using k-NN graph over MST and TSP (besides the obvious conceptual simplicity of k-NN): On a serial computer the k-NN graph can be computed somewhat faster than MST and much faster than the TSP tour. Furthermore, in contrast to MST and TSP, computation of k-NN can be easily parallelized. Secondly, for different values of α, MST and TSP need to be recomputed since the distance between two points is the p-th power of their Euclidean distance where $p = d(1 - \alpha)$. However, the k-NN graph does not change for different values of p, since p-th power is a monotone transformation, and hence the estimates for multiple values of α can be calculated without the extra penalty incurred by the recomputation of the graph. This can be advantageous *e.g.* in intrinsic dimension estimators of manifolds (Costa and Hero, 2003), where p is a free parameter, and thus one can calculate the estimates efficiently for a few different parameter values.

The fourth major contribution is a proof of a finite-sample high-probability error bound (*i.e.* the rate of convergence) for our mutual information estimator which holds under the assumption that the copula of f is Lipschitz. According to the best of our knowledge, this is the first result that gives a rate for the estimation of Rényi mutual information.

The toolkit for proving our results derives from the deep literature of Euclidean functionals, see, (Steele, 1997; Yukich, 1998). In particular, our strong consistency result uses a theorem due to Redmond and Yukich (1996) that essentially states that any quasi-additive power-weighted Euclidean functional can be used as a strongly consistent estimator of Rényi entropy (see also Hero and Michel 1999). We also make use of a result due to Koo and Lee (2007), who proved a rate of convergence result that holds under more stringent conditions. Thus, the main thrust of the present work is show-

[1] Our result for Rényi entropy estimation holds for $d = 1$ and $d = 2$, too.

ing that these conditions hold for p-power weighted nearest-neighbor graphs. Curiously enough, up to now, no one has shown this, except for the case when $p = 1$, which is studied in Section 8.3 of (Yukich, 1998). However, the condition $p = 1$ gives results only for $\alpha = 1 - 1/d$.

Unfortunately, the space limitations do not allow us to present any of our proofs, so we relegate them into the extended version of this paper (Pál et al., 2010). We instead try to give a clear explanation of Rényi entropy and mutual information estimation problems, the estimation algorithms and the statements of our converge results.

Additionally, we report on two numerical experiments. In the first experiment, we compare the empirical rates of convergence of our estimators with our theoretical results and plug-in estimates. Empirically, the NN methods are the clear winner. The second experiment is an illustrative application of mutual information estimation to an Independent Subspace Analysis (ISA) task.

The paper is organized as follows: In the next section, we formally define Rényi entropy and Rényi mutual information and the problem of their estimation. Section 3 explains the 'generalized nearest neighbor' graphs. This graph is then used in Section 4 to define our Rényi entropy estimator. In the same section, we state a theorem containing our convergence results for this estimator (strong consistency and rates). In Section 5, we explain the copula transformation, which connects Rényi entropy with Rényi mutual information. The copula transformation together with the Rényi entropy estimator from Section 4 is used to build an estimator of Rényi mutual information. We conclude this section with a theorem stating the convergence properties of the estimator (strong consistency and rates). Section 6 contains the numerical experiments. We conclude the paper by a detailed discussion of further related work in Section 7, and a list of open problems and directions for future research in Section 8.

2 The Formal Definition of the Problem

Rényi entropy and Rényi mutual information of d real-valued random variables[2] $\mathbf{X} = (X^1, X^2, \ldots, X^d)$ with joint density $f : \mathbb{R}^d \to \mathbb{R}$ and marginal densities $f_i : \mathbb{R} \to \mathbb{R}$, $1 \leq i \leq d$, are defined for any real parameter α assuming the underlying integrals exist. For $\alpha \neq 1$, Rényi entropy and Rényi mutual information are defined respectively as[3]

$$H_\alpha(\mathbf{X}) = H_\alpha(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} f^\alpha(x^1, x^2, \ldots, x^d) \, \mathrm{d}(x^1, x^2, \ldots, x^d) \,, \tag{1}$$

$$I_\alpha(\mathbf{X}) = I_\alpha(f) = \frac{1}{\alpha-1} \log \int_{\mathbb{R}^d} f^\alpha(x^1, x^2, \ldots, x^d) \left(\prod_{i=1}^d f_i(x^i) \right)^{1-\alpha} \mathrm{d}(x^1, x^2, \ldots, x^d). \tag{2}$$

For $\alpha = 1$ they are defined by the limits $H_1 = \lim_{\alpha \to 1} H_\alpha$ and $I_1 = \lim_{\alpha \to 1} I_\alpha$. In fact, Shannon (differential) entropy and the Shannon mutual information are just special cases of Rényi entropy and Rényi mutual information with $\alpha = 1$.

The goal of this paper is to present estimators of Rényi entropy (1) and Rényi information (2) and study their convergence properties. To be more explicit, we consider the problem where we are given i.i.d. random variables $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ where each $\mathbf{X}_j = (X_j^1, X_j^2, \ldots, X_j^d)$ has density $f : \mathbb{R}^d \to \mathbb{R}$ and marginal densities $f_i : \mathbb{R} \to \mathbb{R}$ and our task is to construct an estimate $\widehat{H}_\alpha(\mathbf{X}_{1:n})$ of $H_\alpha(f)$ and an estimate $\widehat{I}_\alpha(\mathbf{X}_{1:n})$ of $I_\alpha(f)$ using the sample $\mathbf{X}_{1:n}$.

3 Generalized Nearest-Neighbor Graphs

The basic tool to define our estimators is the generalized nearest-neighbor graph and more specifically the sum of the p-th powers of Euclidean lengths of its edges.

Formally, let V be a finite set of points in an Euclidean space $\mathbb{R}^d$ and let S be a finite non-empty set of positive integers; we denote by k the maximum element of S. We define the *generalized*

[2] We use superscript for indexing dimension coordinates.

[3] The base of the logarithms in the definition is not important; any base strictly bigger than 1 is allowed. Similarly as with Shannon entropy and mutual information, one traditionally uses either base 2 or e. In this paper, for definitiveness, we stick to base e.

nearest-neighbor graph $NN_S(V)$ as a directed graph on V. The edge set of $NN_S(V)$ contains for each $i \in S$ an edge from each vertex $\mathbf{x} \in V$ to its i-th nearest neighbor. That is, if we sort $V \setminus \{\mathbf{x}\} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{|V|-1}\}$ according to the Euclidean distance to $\mathbf{x}$ (breaking ties arbitrarily): $\|\mathbf{x} - \mathbf{y}_1\| \leq \|\mathbf{x} - \mathbf{y}_2\| \leq \cdots \leq \|\mathbf{x} - \mathbf{y}_{|V|-1}\|$ then $\mathbf{y}_i$ is the i-th nearest-neighbor of $\mathbf{x}$ and for each $i \in S$ there is an edge from $\mathbf{x}$ to $\mathbf{y}_i$ in the graph.

For $p \geq 0$ let us denote by $L_p(V)$ the sum of the p-th powers of Euclidean lengths of its edges. Formally,

$$L_p(V) = \sum_{(\mathbf{x},\mathbf{y}) \in E(NN_S(V))} \|\mathbf{x} - \mathbf{y}\|^p \,, \tag{3}$$

where $E(NN_S(V))$ denotes the edge set of $NN_S(V)$. We intentionally hide the dependence on S in the notation $L_p(V)$. For the rest of the paper, the reader should think of S as a fixed but otherwise arbitrary finite non-empty set of integers, say, $S = \{1, 3, 4\}$.

The following is a basic result about L_p. The proof can be found in Pál et al. (2010).

Theorem 1 (Constant γ)**.** *Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ be an i.i.d. sample from the uniform distribution over the d-dimensional unit cube $[0,1]^d$. For any $p \geq 0$ and any finite non-empty set S of positive integers there exists a constant $\gamma > 0$ such that*

$$\lim_{n\to\infty} \frac{L_p(\mathbf{X}_{1:n})}{n^{1-p/d}} = \gamma \qquad a.s. \tag{4}$$

The value of γ depends on d, p, S and, except for special cases, an analytical formula for its value is not known. This causes a minor problem since the constant γ appears in our estimators. A simple and effective way to deal with this problem is to generate a large i.i.d. sample $\mathbf{X}_{1:n}$ from the uniform distribution over $[0,1]^d$ and estimate γ by the empirical value of $L_p(\mathbf{X}_{1:n})/n^{1-p/d}$.

4 An Estimator of Rényi Entropy

We are now ready to present an estimator of Rényi entropy based on the generalized nearest-neighbor graph. Suppose we are given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ from a distribution μ over $\mathbb{R}^d$ with density f. We estimate entropy $H_\alpha(f)$ for $\alpha \in (0,1)$ by

$$\widehat{H}_\alpha(\mathbf{X}_{1:n}) = \frac{1}{1-\alpha} \log \frac{L_p(\mathbf{X}_{1:n})}{\gamma n^{1-p/d}} \qquad \text{where} \quad p = d(1-\alpha), \tag{5}$$

and $L_p(\cdot)$ is the sum of p-th powers of Euclidean lengths of edges of the nearest-neighbor graph $NN_S(\cdot)$ for some finite non-empty $S \subset \mathbb{N}^+$ as defined by equation (3). The constant γ is the same as in Theorem 1.

The following theorem is our main result about the estimator $\widehat{H}_\alpha$. It states that $\widehat{H}_\alpha$ is strongly consistent and gives upper bounds on the rate of convergence. The proof of theorem is in Pál et al. (2010).

Theorem 2 (Consistency and Rate for $\widehat{H}_\alpha$)**.** *Let $\alpha \in (0,1)$. Let μ be an absolutely continuous distribution over $\mathbb{R}^d$ with bounded support and let f be its density. If $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is an i.i.d. sample from μ then*

$$\lim_{n\to\infty} \widehat{H}_\alpha(\mathbf{X}_{1:n}) = H_\alpha(f) \qquad \textit{a.s.} \tag{6}$$

Moreover, if f is Lipschitz then for any $\delta > 0$ with probability at least $1 - \delta$,

$$\left|\widehat{H}_\alpha(\mathbf{X}_{1:n}) - H_\alpha(f)\right| \leq \begin{cases} O\left(n^{-\frac{d-p}{d(2d-p)}} (\log(1/\delta))^{1/2-p/(2d)}\right), & \textit{if } 0 < p < d-1 \,; \\ O\left(n^{-\frac{d-p}{d(d+1)}} (\log(1/\delta))^{1/2-p/(2d)}\right), & \textit{if } d-1 \leq p < d \,. \end{cases} \tag{7}$$

5 Copulas and Estimator of Mutual Information

Estimating mutual information is slightly more complicated than estimating entropy. We start with a basic property of mutual information which we call *rescaling*. It states that if $h_1, h_2, \ldots, h_d : \mathbb{R} \to \mathbb{R}$ are arbitrary strictly increasing functions, then

$$I_\alpha(h_1(X^1), h_2(X^2), \ldots, h_d(X^d)) = I_\alpha(X^1, X^2, \ldots, X^d) \,. \tag{8}$$

A particularly clever choice is $h_j = F_j$ for all $1 \le j \le d$, where F_j is the cumulative distribution function (c.d.f.) of X^j. With this choice, the marginal distribution of $h_j(X^j)$ is the uniform distribution over $[0,1]$ assuming that F_j, the c.d.f. of X^j, is continuous. Looking at the definition of H_α and I_α we see that

$$I_\alpha(X^1, X^2, \ldots, X^d) = I_\alpha(F_1(X^1), F_2(X^2), \ldots, F_d(X^d)) = -H_\alpha(F_1(X^1), F_2(X^2), \ldots, F_d(X^d)) \; .$$

In other words, calculation of mutual information can be reduced to the calculation of entropy provided that marginal c.d.f.'s $F_1, F_2, \ldots, F_d$ are known. The problem is, of course, that these are not known and need to be estimated from the sample. We will use empirical c.d.f.'s $(\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_d)$ as their estimates. Given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ from distribution μ and with density f, the empirical c.d.f's are defined as

$$\widehat{F}_j(x) = \frac{1}{n} |\{i : 1 \le i \le n,\ x \le X_i^j\}| \qquad \text{for } x \in \mathbb{R},\ 1 \le j \le d \; .$$

Introduce the compact notation $\mathbf{F} : \mathbb{R}^d \to [0,1]^d$, $\widehat{\mathbf{F}} : \mathbb{R}^d \to [0,1]^d$,

$$\mathbf{F}(x^1, x^2, \ldots, x^d) = (F_1(x^1), F_2(x^2), \ldots, F_d(x^d)) \qquad \text{for } (x^1, x^2, \ldots, x^d) \in \mathbb{R}^d \; ; \quad (9)$$

$$\widehat{\mathbf{F}}(x^1, x^2, \ldots, x^d) = (\widehat{F}_1(x^1), \widehat{F}_2(x^2), \ldots, \widehat{F}_d(x^d)) \qquad \text{for } (x^1, x^2, \ldots, x^d) \in \mathbb{R}^d \; . \quad (10)$$

Let us call the maps $\mathbf{F}$, $\widehat{\mathbf{F}}$ the *copula transformation*, and the *empirical copula transformation*, respectively. The joint distribution of $\mathbf{F}(\mathbf{X}) = (F_1(X^1), F_2(X^2), \ldots, F_d(X^d))$ is called the copula of μ, and the sample $(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \ldots, \widehat{\mathbf{Z}}_n) = (\widehat{\mathbf{F}}(\mathbf{X}_1), \widehat{\mathbf{F}}(\mathbf{X}_2), \ldots, \widehat{\mathbf{F}}(\mathbf{X}_n))$ is called the empirical copula (Dedecker et al., 2007). Note that j-th coordinate of $\widehat{\mathbf{Z}}_i$ equals

$$\widehat{Z}_i^j = \frac{1}{n} \operatorname{rank}(X_i^j, \{X_1^j, X_2^j, \ldots, X_n^j\}) \; ,$$

where $\operatorname{rank}(x, A)$ is the number of element of A less than or equal to x. Also, observe that the random variables $\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \ldots, \widehat{\mathbf{Z}}_n$ are not even independent! Nonetheless, the empirical copula $(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \ldots, \widehat{\mathbf{Z}}_n)$ is a good approximation of an i.i.d. sample $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n) = (\mathbf{F}(\mathbf{X}_1), \mathbf{F}(\mathbf{X}_2), \ldots, \mathbf{F}(\mathbf{X}_n))$ from the copula of μ. Hence, we estimate the Rényi mutual information I_α by

$$\widehat{I}_\alpha(\mathbf{X}_{1:n}) = -\widehat{H}_\alpha(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \ldots, \widehat{\mathbf{Z}}_n), \quad (11)$$

where $\widehat{H}_\alpha$ is defined by (5). The following theorem is our main result about the estimator $\widehat{I}_\alpha$. It states that $\widehat{I}_\alpha$ is strongly consistent and gives upper bounds on the rate of convergence. The proof of this theorem can be found in Pál et al. (2010).

Theorem 3 (Consistency and Rate for $\widehat{I}_\alpha$)**.** *Let $d \ge 3$ and $\alpha = 1 - p/d \in (1/2, 1)$. Let μ be an absolutely continuous distribution over $\mathbb{R}^d$ with density f. If $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is an i.i.d. sample from μ then*

$$\lim_{n\to\infty} \widehat{I}_\alpha(\mathbf{X}_{1:n}) = I_\alpha(f) \qquad \textit{a.s.}$$

Moreover, if the density of the copula of μ is Lipschitz, then for any $\delta > 0$ with probability at least $1 - \delta$,

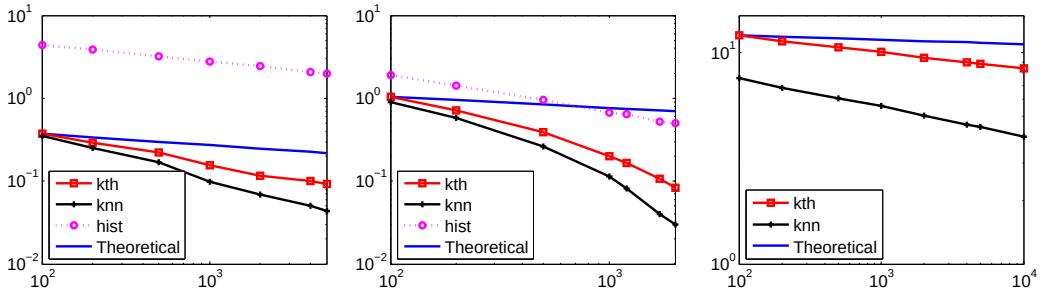
$$\left| \widehat{I}_\alpha(\mathbf{X}_{1:n}) - I_\alpha(f) \right| \le \begin{cases} O\left(\max\{n^{-\frac{d-p}{d(2d-p)}}, n^{-p/2+p/d}\} (\log(1/\delta))^{1/2} \right), & \text{if } 0 < p \le 1 \; ; \\ O\left(\max\{n^{-\frac{d-p}{d(2d-p)}}, n^{-1/2+p/d}\} (\log(1/\delta))^{1/2} \right), & \text{if } 1 \le p \le d-1 \; ; \\ O\left(\max\{n^{-\frac{d-p}{d(d+1)}}, n^{-1/2+p/d}\} (\log(1/\delta))^{1/2} \right), & \text{if } d-1 \le p < d \; . \end{cases}$$

6 Experiments

In this section we show two numerical experiments to support our theoretical results about the convergence rates, and to demonstrate the applicability of the proposed Rényi mutual information estimator, $\widehat{I}_\alpha$.

6.1 The Rate of Convergence

In our first experiment (Fig. 1), we demonstrate that the derived rate is indeed an upper bound on the convergence rate. Figure 1a-1c show the estimation error of $\widehat{I}_\alpha$ as a function of the sample size. Here, the underlying distribution was a 3D uniform, a 3D Gaussian, and a 20D Gaussian with randomly chosen nontrivial covariance matrices, respectively. In these experiments α was set to 0.7. For the estimation we used $S = \{3\}$ (kth) and $S = \{1, 2, 3\}$ (knn) sets. Our results also indicate that these estimators achieve better performances than the histogram based plug-in estimators (hist). The number and the sizes of the bins were determined with the rule of Scott (1979). The histogram based estimator is not shown in the 20D case, as in this large dimension it is not applicable in practice. The figures are based on averaging 25 independent runs, and they also show the theoretical upper bound (Theoretical) on the rate derived in Theorem 3. It can be seen that the theoretical rates are rather conservative. We think that this is because the theory allows for quite irregular densities, while the densities considered in this experiment are very nice.



(a) 3D uniform (b) 3D Gaussian (c) 20D Gaussian

Figure 1: Error of the estimated Rényi informations in the number of samples.

6.2 Application to Independent Subspace Analysis

An important application of dependence estimators is the Independent Subspace Analysis problem (Cardoso, 1998). This problem is a generalization of the Independent Component Analysis (ICA), where we assume the independent sources are multidimensional vector valued random variables. The formal description of the problem is as follows. We have $\mathbf{S} = (\mathbf{S}^1; \ldots; \mathbf{S}^m) \in \mathbb{R}^{dm}$, m independent d-dimensional sources, *i.e.* $\mathbf{S}^i \in \mathbb{R}^d$, and $I(\mathbf{S}^1, \ldots, \mathbf{S}^m) = 0$.[4] In the ISA statistical model we assume that $\mathbf{S}$ is hidden, and only n i.i.d. samples from $\mathbf{X} = \mathbf{AS}$ are available for observation, where $\mathbf{A} \in \mathbb{R}^{q \times dm}$ is an unknown invertible matrix with full rank and $q \geq dm$. Based on n i.i.d. observation of $\mathbf{X}$, our task is to estimate the hidden sources $\mathbf{S}^i$ and the mixing matrix $\mathbf{A}$. Let the estimation of $\mathbf{S}$ be denoted by $\mathbf{Y} = (\mathbf{Y}^1; \ldots; \mathbf{Y}^m) \in \mathbb{R}^{dm}$, where $\mathbf{Y} = \mathbf{WX}$. The goal of ISA is to calculate $\operatorname{argmin}_{\mathbf{W}} I(\mathbf{Y}^1, \ldots, \mathbf{Y}^m)$, where $\mathbf{W} \in \mathbb{R}^{dm \times q}$ is a matrix with full rank. Following the ideas of Cardoso (1998), this ISA problem can be solved by first preprocessing the observed quantities $\mathbf{X}$ by a traditional ICA algorithm which provides us $\mathbf{W}_{ICA}$ estimated separation matrix[5], and then simply grouping the estimated ICA components into ISA subspaces by maximizing the sum of the MI in the estimated subspaces, that is we have to find a permutation matrix $\mathbf{P} \in \{0, 1\}^{dm \times dm}$ which solves

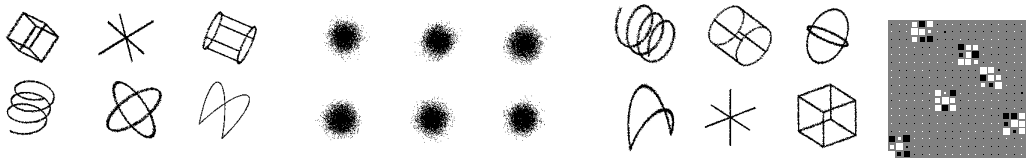
$$\max_{\mathbf{P}} \sum_{j=1}^{m} I(Y_1^j, Y_2^j, \ldots, Y_d^j) \,. \tag{12}$$

where $\mathbf{Y} = \mathbf{PW}_{ICA}\mathbf{X}$. We used the proposed copula based information estimation, $\widehat{I}_\alpha$ with $\alpha = 0.99$ to approximate the Shannon mutual information, and we chose $S = \{1, 2, 3\}$. Our experiment shows that this ISA algorithm using the proposed MI estimator can indeed provide good

[4] Here we need the generalization of MI to multidimensional quantities, but that is obvious by simply replacing the 1D marginals by d-dimensional ones.

[5] for simplicity we used the FastICA algorithm in our experiments (Hyvärinen et al., 2001)

estimation of the ISA subspaces. We used a standard ISA benchmark dataset from Szabó et al. (2007); we generated 2,000 i.i.d. sample points on 3D geometric wireframe distributions from 6 different sources independently from each other. These sampled points can be seen in Fig. 2a, and they represent the sources, $\mathbf{S}$. Then we mixed these sources by a randomly chosen invertible matrix $\mathbf{A} \in \mathbb{R}^{18\times 18}$. The six 3-dimensional projections of $\mathbf{X} = \mathbf{AS}$ observed quantities are shown in Fig. 2b. Our task was to estimate the original sources $\mathbf{S}$ using the sample of the observed quantity $\mathbf{X}$ only. By estimating the MI in (12), we could recover the original subspaces as it can be seen in Fig. 2c. The successful subspace separation is shown in the form of Hinton diagrams as well, which is the product of the estimated ISA separation matrix $\mathbf{W} = \mathbf{PW}_{ICA}$ and $\mathbf{A}$. It is a block permutation matrix if and only if the subspace separation is perfect (Fig. 2d).



(a) Original (b) Mixed (c) Estimated (d) Hinton

Figure 2: ISA experiment for six 3-dimensional sources.

7 Further Related Works

As it was pointed out earlier, in this paper we heavily built on the results known from the theory of Euclidean functionals (Steele, 1997; Redmond and Yukich, 1996; Koo and Lee, 2007). However, now we can be more precise about earlier work concerning nearest-neighbor based Euclidean functionals: The closest to our work is Section 8.3 of Yukich (1998), where the case of NN_S graph based p-power weighted Euclidean functionals with $S = \{1, 2, \ldots, k\}$ and $p = 1$ was investigated.

Nearest-neighbor graphs have first been proposed for Shannon entropy estimation by Kozachenko and Leonenko (1987). In particular, in the mentioned work only the case of NN_S graphs with $S = \{1\}$ was considered. More recently, Goria et al. (2005) generalized this approach to $S = \{k\}$ and proved the resulting estimator's weak consistency under some conditions on the density. The estimator in this paper has a form quite similar to that of ours:

$$\tilde{H}_1 = \log(n-1) - \psi(k) + \log\left(\frac{2\pi^{d/2}}{d\Gamma(d/2)}\right) + \frac{d}{n}\sum_{i=1}^{n}\log\|\mathbf{e}_i\| \,.$$

Here ψ stands for the digamma function, and $\mathbf{e}_i$ is the directed edge pointing from $\mathbf{X}_i$ to its k^{th} nearest-neighbor. Comparing this with (5), unsurprisingly, we find that the main difference is the use of the logarithm function instead of $|\cdot|^p$ and the different normalization. As mentioned before, Leonenko et al. (2008) proposed an estimator that uses the NN_S graph with $S = \{k\}$ for the purpose of estimating the Rényi entropy. Their estimator takes the form

$$\tilde{H}_\alpha = \frac{1}{1-\alpha}\log\left(\frac{n-1}{n}V_d^{1-\alpha}C_k^{1-\alpha}\sum_{i=1}^{n}\frac{\|\mathbf{e}_i\|^{d(1-\alpha)}}{(n-1)^\alpha}\right) \,,$$

where Γ stands for the Gamma function, $C_k = \left[\frac{\Gamma(k)}{\Gamma(k+1-\alpha)}\right]^{1/(1-\alpha)}$ and $V_d = \pi^{d/2}\Gamma(d/2+1)$ is the volume of the d-dimensional unit ball, and again $\mathbf{e}_i$ is the directed edge in the NN_S graph starting from node $\mathbf{X}_i$ and pointing to the k-th nearest node. Comparing this estimator with (5), it is apparent that it is (essentially) a special case of our NN_S based estimator. From the results of Leonenko et al. (2008) it is obvious that the constant γ in (5) can be found in analytical form when $S = \{k\}$. However, we kindly warn the reader again that the proofs of these last three cited articles (Kozachenko and Leonenko, 1987; Goria et al., 2005; Leonenko et al., 2008) contain a few errors, just like the Wang et al. (2009b) paper for *KL divergence* estimation from two samples. Kraskov et al. (2004) also proposed a k-nearest-neighbors based estimator for the *Shannon* mutual information estimation, but the theoretical properties of their estimator are unknown.

8 Conclusions and Open Problems

We have studied Rényi entropy and mutual information estimators based on NN_S graphs. The estimators were shown to be strongly consistent. In addition, we derived upper bounds on their convergence rate under some technical conditions. Several open problems remain unanswered:

An important open problem is to understand how the choice of the set $S \subset \mathbb{N}^+$ affects our estimators. Perhaps, there exists a way to choose S as a function of the sample size n (and d, p) which strikes the optimal balance between the bias and the variance of our estimators.

Our method can be used for estimation of Shannon entropy and mutual information by simply using α close to 1. The open problem is to come up with a way of choosing α, approaching 1, as a function of the sample size n (and d, p) such that the resulting estimator is consistent and converges as rapidly as possible. An alternative is to use the logarithm function in place of the power function. However, the theory would need to be changed significantly to show that the resulting estimator remains strongly consistent.

In the proof of consistency of our mutual information estimator $\widehat{I}_\alpha$ we used Kiefer-Dvoretzky-Wolfowitz theorem to handle the effect of the inaccuracy of the empirical copula transformation (see Pál et al. (2010) for details). Our particular use of the theorem seems to restrict α to the interval $(1/2, 1)$ and the dimension to values larger than 2. Is there a better way to estimate the error caused by the empirical copula transformation and prove consistency of the estimator for a larger range of α's and $d = 1, 2$?

Finally, it is an important open problem to prove bounds on converge rates for densities that have higher order smoothness (*i.e.* β-Hölder smooth densities). A related open problem, in the context of of theory of Euclidean functionals, is stated in Koo and Lee (2007).

Acknowledgements

This work was supported in part by AICML, AITF (formerly iCore and AIF), NSERC, the PASCAL2 Network of Excellence under EC grant no. 216886 and by the Department of Energy under grant number DESC0002607. Cs. Szepesvári is on leave from SZTAKI, Hungary.

References

C. Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1:3–22, 2004.

M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh. A hierarchical clustering based on mutual information maximization. In *in IEEE ICIP*, pages 277–280, 2007.

P. A. Alemany and D. H. Zanette. Fractal random walks from a variational formalism for Tsallis entropies. *Phys. Rev. E*, 49(2):R956–R958, Feb 1994.

J. Cardoso. Multidimensional independent component analysis. *Proc. ICASSP'98, Seattle, WA.*, 1998.

B. Chai, D. B. Walther, D. M. Beck, and L. Fei-Fei. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.

J. A. Costa and A. O. Hero. Entropic graphs for manifold learning. In *IEEE Asilomar Conf. on Signals, Systems, and Computers*, 2003.

J. Dedecker, P. Doukhan, G. Lang, J.R. Leon, S. Louhichi, and C Prieur. *Weak Dependence: With Examples and Applications*, volume 190 of *Lecture notes in Statistics*. Springer, 2007.

M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17: 277–297, 2005.

A. O. Hero and O. J. Michel. Asymptotic theory of greedy approximations to minimal k-point random graphs. *IEEE Trans. on Information Theory*, 45(6):1921–1938, 1999.

A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval, 2002a. Communications and Signal Processing Laboratory Technical Report CSPL-328.

A. O. Hero, B. Ma, O. Michel, and J. Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002b.

K. Hlaváckova-Schindler, M. Paluŝb, M. Vejmelkab, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

M. M. Van Hulle. Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*, 20:964–973, 2008.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley, New York, 2001.

Y. Koo and S. Lee. Rates of convergence of means of Euclidean functionals. *Journal of Theoretical Probability*, 20(4):821–841, 2007.

L. F. Kozachenko and N. N. Leonenko. A statistical estimate for the entropy of a random vector. *Problems of Information Transmission*, 23:9–16, 1987.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

J. Kybic. Incremental updating of nearest neighbor-based high-dimensional entropy estimation. In *Proc. Acoustics, Speech and Signal Processing*, 2006.

E. Learned-Miller and J. W. Fisher. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

N. Leonenko, L. Pronzato, and V. Savani. A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, 36(5):2153–2182, 2008.

J. Lewi, R. Butera, and L. Paninski. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

D. Pál, Cs. Szepesvári, and B. Póczos. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs, 2010. http://arxiv.org/abs/1003.1954.

H. Peng and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans On Pattern Analysis and Machine Intelligence*, 27, 2005.

B. Póczos and A. Lőrincz. Independent subspace analysis using geodesic spanning trees. In *ICML*, pages 673–680, 2005.

B. Póczos and A. Lőrincz. Identification of recurrent neural networks by Bayesian interrogation techniques. *Journal of Machine Learning Research*, 10:515–554, 2009.

B. Póczos, S. Kirshner, and Cs. Szepesvári. REGO: Rank-based estimation of Rényi information using Euclidean graph optimization. In *AISTATS 2010*, 2010.

C. Redmond and J. E. Yukich. Asymptotics for Euclidean functionals with power-weighted edges. *Stochastic processes and their applications*, 61(2):289–304, 1996.

D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

C. Shan, S. Gong, and P. W. Mcowan. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.

J. M. Steele. *Probability Theory and Combinatorial Optimization*. Society for Industrial and Applied Mathematics, 1997.

Z. Szabó, B. Póczos, and A. Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.

A. B. Tsybakov and E. C. van der Meulen. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*, 23:75–83, 1996.

O. Vasicek. A test for normality based on sample entropy. *Journal of the Royal Statistical Society, Series B*, 38:54–59, 1976.

Q. Wang, S. R. Kulkarni, and S. Verdú. Universal estimation of information measures for analog sources. *Foundations and Trends in Communications and Information Theory*, 5(3):265–352, 2009a.

Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009b.

E. Wolsztynski, E. Thierry, and L. Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Process.*, 85(5):937–949, 2005.

J. E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*. Springer, 1998.