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

Dávid Pál*, Barnabás Póczos†, Csaba Szepesvári*

* Department of Computer Science, University of Alberta, Edmonton, Canada

† School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

dpal@cs.ualberta.ca, bapoczos@cs.cmu.edu, szepesva@cs.ualberta.ca

Introduction

- Long history of non-parametric estimators of entropy and mutual information based on nearest neighbor graphs (Kozachenko and Leonenko, 1987, Hero and Michel, 1999, Goria et al., 2005, Leonenko et al., 2008b, Wang et al., 2009).
- We are the first to correctly prove almost sure consistency and rates of convergence (Leonenko et al., 2008a, Kozachenko and Leonenko, 1987, Wang et al., 2009).
- We use the mathematical machinery of additive Euclidean functionals (Yukich, 1998, Steele, 1997, Koo and Lee, 2007).
- Computationally more efficient than minimum spanning tree estimator (Hero and Michel, 1999, Póczos et al., 2010).

Definitions

The Rényi entropy and the Rényi mutual information of order α of d real-valued random variables $\mathbf{X} = (X^1, X^2, \ldots, X^d)$ with joint density $f : \mathbb{R}^d \to \mathbb{R}$ and marginal densities $f_i : \mathbb{R} \to \mathbb{R}$, $1 \le i \le d$ are for $\alpha \neq 1$ respectively defined by

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

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

The limits $H_1(f) = \lim_{\alpha \to 1} H_\alpha(f)$ and $I_1(f) = \lim_{\alpha \to 1} I_\alpha(f)$ are the Shannon entropy and the Shannon mutual information respectively.

Generalized Nearest Neighbor Graphs

Fix a finite non-empty set S of positive integers; e.g. $S = \{1, 2, \ldots, k\}$ or $S = \{k\}$. Given a finite set V of points in $\mathbb{R}^d$ we define a generalized nearest neighbor graph $NN_S(V)$ as a directed graph on V where for each point $\mathbf{x} \in V$ and each $i \in S$ there is an edge from $\mathbf{x}$ to its i-th nearest neighbor in V. We define

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

Theorem. *(Constant γ) If $\mathcal{U}_n$ is an i.i.d. sample of size n from the uniform distribution over $[0,1]^d$ then for any $p \in [0,d]$ and any S there exists a constant $\gamma > 0$ such that*

$$\lim_{n\to\infty} \frac{L_p(\mathcal{U}_n)}{n^{1-p/d}} = \gamma \qquad a.s.$$

Using this theorem we can estimate the constant γ to arbitrary precision.

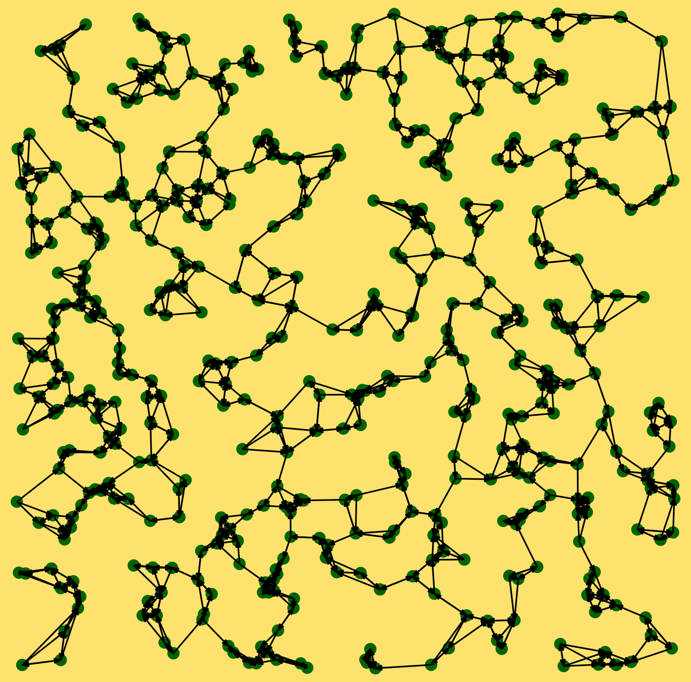

Figure 1: *Nearest neighbor graph $NN_S(\mathcal{U}_n)$ on a sample consisting of $n = 200$ points drawn i.i.d. from the uniform distribution over $[0,1]^2$ and with $S = \{1,2,3\}$.*

Entropy Estimator

Given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ where each $\mathbf{X}_i$ has density f, we estimate $H_\alpha(f)$ for $\alpha \in (0,1)$ by

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

Theorem (Consistency and Rate for $\widehat{H}_\alpha$)**.** *Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ be an i.i.d. sample from a distribution over $\mathbb{R}^d$ with bounded support and density f. Then,*

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

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

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

Copulas and Estimator of Mutual Information

We estimate the Rényi mutual information $I_\alpha(f)$ by

$$\widehat{I}_\alpha(\mathbf{X}_{1:n}) = \widehat{H}_\alpha(\text{Empirical Copula}(\mathbf{X}_{1:n})) .$$

Theorem (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*

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

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

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

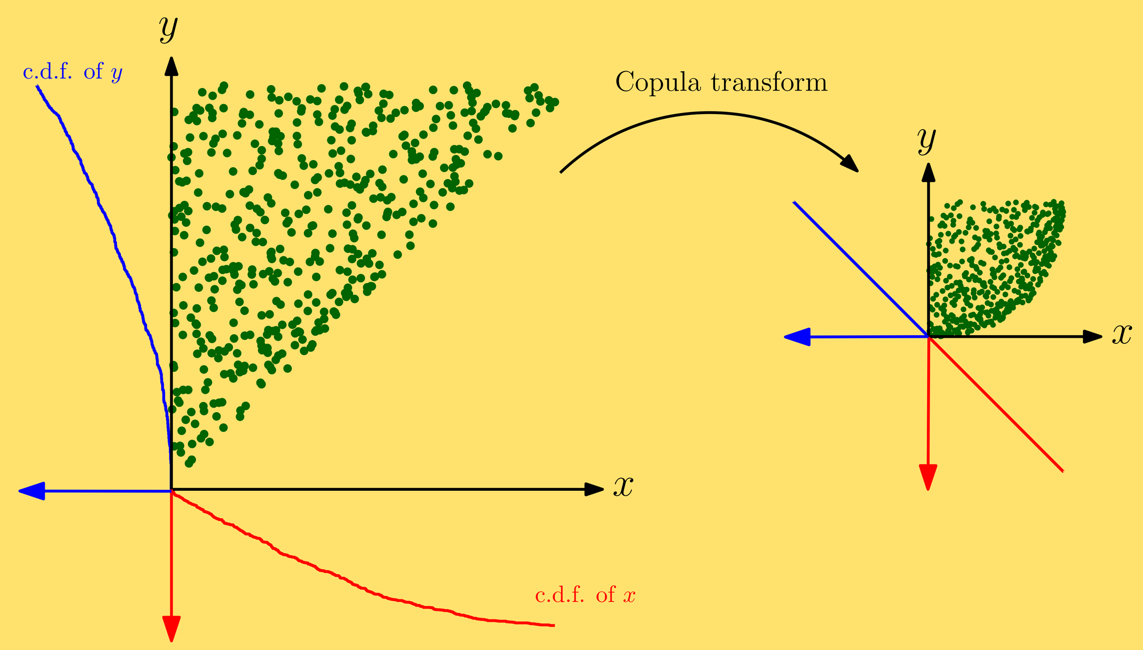

Figure 2: *Sample from the uniform distribution over a triangle with vertices $(0,0)$, $(3,0)$, $(0,3)$ and its empirical copula.*

References

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.

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.

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

N. Leonenko, L. Pronzato, and V. Savani. Estimation of entropies and divergences via nearest neighbours. *Tatra Mt. Mathematical Publications*, 39, 2008b.

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

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

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, 2009.

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

NIPS 2010, Twenty-Fourth Annual Conference on Neural Information Processing Systems, December 6–9, 2010, Vancouver, BC, Canada