REGO: Rank-based Estimation of Rényi Information using Euclidean Graph Optimization

Barnabás Póczos
Dept. of Computing Science
University of Alberta
Edmonton, AB
Canada, T6G 2E8

Sergey Kirshner
Dept. of Statistics
Purdue University
West Lafayette, IN
USA, 47907-2066

Csaba Szepesvári
Dept. of Computing Science
University of Alberta
Edmonton, AB
Canada, T6G 2E8

Abstract

We propose a new method for a non-parametric estimation of Rényi and Shannon information for a multivariate distribution using a corresponding copula, a multivariate distribution over normalized ranks of the data. As the information of the distribution is the same as the negative entropy of its copula, our method estimates this information by solving a Euclidean graph optimization problem on the empirical estimate of the distribution's copula. Owing to the properties of the copula, we show that the resulting estimator of Rényi information is strongly consistent and robust. Further, we demonstrate its applicability in image registration in addition to simulated experiments.

1 Introduction

Numerous problems in machine learning require measuring the strength of relation between random variables. From many different measures of dependence between variables, Shannon mutual information is often the natural choice because of its connection to information theory. There is an abundance of applications which involve the estimation of mutual information ranging from information theory via feature selection, clustering, image registration, independent component analysis (ICA) and independent subspace analysis (ISA), causality detection, Bayesian active learning, optimal experiment design, to graphical model structure learning. However, mutual information of the continuous random variables is difficult to estimate when the functional form of their dependence is not known.

We propose a novel non-parametric approach to estimation of Rényi differential information and as one of its limiting cases, Shannon differential mutual information. Our approach is based on an alternative formulation of the information involving the copula of the joint distribution, a multivariate distribution obtained by replacing each random variable by the value of its distribution function. This transformation standardizes the marginals to be uniform on $[0, 1]$ while preserving many of the distribution's dependence properties including its concordance measures and its information.

Our approach consists of two steps: mapping the observations into their copula domain, and computing the entropy of the transformed samples. For the copula transformation (in lack of the true marginals) we employ the marginals' empirical distribution functions. This non-parametric transformation serves three purposes. One, it simplifies the estimation problem by removing the terms corresponding to the marginal entropies (and potential errors associated with their estimation). Two, as it is based entirely on the data ranks, it is robust to outliers. Three, the resulting support is bounded making it possible to use with estimators not designed for unbounded support. The next step of the algorithm, inspired by the pioneering work of Hero and Michel (1998), estimates the information by solving a Euclidean graph optimization (EGO) problem (e.g., minimum spanning tree, k-nearest neighbor graphs) on the complete graph whose nodes correspond to the transformed sample. We prove that our estimator, which we call REGO after "rank-based EGO", is strongly consistent and has nice robustness properties. A crucial feature of REGO is that it avoids estimation of densities, nuisance parameters from the

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

point of view of information estimation. To the best of our knowledge, REGO is the only information estimator with these properties. In addition, the estimator runs in time $O(dn \log n)$, i.e., it can be used for large samples. Its implementation is straightforward and the source-code will be made publicly available. In our experiments, in low-dimensional problems the estimator proved to be competitive with alternatives, while in higher-dimensional problems (in the lack of competitors) it showed a remarkably quick convergence. We have also tested its performance on an image registration task with and without outliers. This experiment showed that the idea of working based on ranks is indeed essential to expect good performance in realistic situations.

The paper is organized as follows. In Section 2 we formally define our problem and give a short overview of alternative approaches. Our estimator REGO is described in Section 3. In Section 4 we prove REGO's strong consistency, and investigate its robustness properties. We validate our approach on simulated data and on image registration, a fundamental problem requiring the estimation of mutual information, in Sections 5 and 6, respectively. We conclude in Section 7.

2 Problem Setup

Let $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ be a sequence of independent, identically distributed (i.i.d.) random variables. We assume that $\mathbf{X} = (X_1, \ldots, X_d)^T$ has a density $f_{\mathbf{X}}$ underlying the Lebesgue-measure. We further assume that the marginals of X_j $(1 \leq j \leq d)$ assume densities f_{X_j} with respect to the Lebesgue-measure on the real-line. We consider the problem of estimating Rényi (1961) information $I_\alpha(\mathbf{X})$ underlying $\mathbf{X}$, given the sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. (Rényi information plays an important role amongst measures of information, see, e.g., Hero et al., 2002). Here $\alpha > 0$ is a parameter to be selected by the user. For $\alpha > 0$, $\alpha \neq 1$, it is defined as

$$I_\alpha(\mathbf{X}) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} f_{\mathbf{X}}^{\alpha}(\mathbf{x}) \left(\prod_{j=1}^{d} f_{X_j}(x_j) \right)^{1-\alpha} d\mathbf{x}, \tag{1}$$

while, $I_1(\mathbf{X})$ is defined through the limit $I_1(\mathbf{X}) = \lim_{\substack{\alpha \to 1 \\ \alpha \neq 1}} I_\alpha(\mathbf{X})$. As it is well known, $I_1(\mathbf{X})$ is equal to the Shannon mutual information:

$$I_1(\mathbf{X}) = \int_{\mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) \log \frac{f_{\mathbf{X}}(\mathbf{x})}{\prod_{j=1}^{d} f_{X_j}(x_j)} d\mathbf{x}. \tag{2}$$

To date, there are two approaches to estimate information that we know of: *Plug-in* estimators attempt to estimate the densities $f_{\mathbf{X}}$, f_{X_1}, ..., f_{X_d} and then use the definition of information directly (the integral is computed approximately using some numerical method). From the point of view of information estimation, the densities however play the role of a nuisance parameter. Density estimators (based on histograms or kernel density estimators) will have tuneable parameters and may need to use cross-validation for model selection to achieve good performance. Therefore, they will be expensive, and their performance might be sensitive to the choice of the density estimation method. The alternative approach uses *direct* (not plug-in based) estimators of the entropy, and then computes the estimate of information based on the well-known identity $I_1(\mathbf{X}) = -H_1(\mathbf{X}) + \sum_{i=1}^{d} H_1(X_i)$. The drawback is that this approach works only when $\alpha = 1$ (here H_1 is the entropy of its argument, see (3)). Direct estimators of the entropy build a graph embedded in $\mathbb{R}^d$ whose edges span the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and compute some statistics based on the lengths of the edges. These estimators include that of Kozachenko and Leonenko (1987); Leonenko et al. (2008); Kraskov et al. (2004); Kybic (2006); Hero and Michel (1998). The first four estimators all use nearest-neighbor graphs, while Hero and Michel (1998), building on previous results from the literature of probabilistic geometric optimization, propose to use the solution of certain Euclidean Geometric Optimization (EGO) problems. Strong consistency (i.e., the estimates converge to the true value with probability one) is known to hold for the estimator of Hero and Michel (1998), while Kozachenko and Leonenko (1987); Leonenko et al. (2008) prove mean-square and weak consistency, respectively. More information and a review on other methods for entropy estimation can be found in Beirlant et al. (1997).

REGO differs from these algorithms by using rank statistics only.[1] Further, we do not have to assume $\alpha = 1$, and we do not have to estimate the entropy of the marginals for the estimation of $I_\alpha(\mathbf{X})$. According to our knowledge, this is the first non-parametric method that estimates Rényi's mutual information in a direct manner.

3 The Algorithm

Let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be an i.i.d. sample, which serves as the input to the algorithm. The algorithm consists of two steps: In the first step, the input is mapped into the unit hypercube so that the marginals become approximately uniform. In the second step the transformed sample is sent to an algorithm that

[1] It is interesting to notice that Kraskov et al. (2004) actually discussed transforming the marginals to make them uniform, but they have dropped this idea in favor of making the marginals normal with zero mean, unit variance.

estimates the α-entropy of it. For this second step we suggest to use an estimator based on Euclidean Graph Optimization. In what follows, we explain both steps in more details and give the rationale behind them.

3.1 Step 1: Empirical Distributional Transformation

The first step is to compute $\hat{\mathbf{Z}}_{1:n} = (\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_n)$, where $\hat{\mathbf{Z}}_t = (\hat{Z}_{t1}, \ldots, \hat{Z}_{td})^T$, $\hat{Z}_{tj} = F_{nj}(X_{tj})$, where F_{nj} is the empirical distribution function underlying $X_{1:n}^{(j)} = (X_{1j}, \ldots, X_{nj})$: $F_{nj}(x) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}_{\{X_{tj} \le x\}}$. Note that $n\hat{Z}_{tj}$ is just the rank of X_{tj} in $X_{1:n}^{(j)}$, thus $(\hat{Z}_{tj}; 1 \le t \le n)$ can be computed in $O(nd \log n)$-time by sorting the elements of the vector $X_{1:n}^{(j)}$.

Rationale Note that for large n we expect F_{nj} to be close to F_j, the distribution of X_j. Thus, we expect $\hat{Z}_{tj}$ ($\hat{Z}_j = F_{nj}(X_j)$) to be close to $Z_{tj} = F_j(X_{tj})$ (resp., $Z_j = F_j(X_j)$). Now, if F_j is continuous, Z_{tj} (Z_j) will be uniformly distributed. Further, F_j being a measurable invertible mapping, $I_\alpha(\mathbf{Z}) = I_\alpha(\mathbf{X})$, by the isomorphism theorem (Vajda, 1989). Now, since the marginals of $\mathbf{Z}$ are uniform, $I_\alpha(\mathbf{Z}) = -H_\alpha(\mathbf{Z})$, where

$$H_\alpha(\mathbf{Z}) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f_{\mathbf{Z}}^{\alpha}(\mathbf{z})\, d\mathbf{z},$$

when $\alpha \neq 1$ and

$$H_1(\mathbf{Z}) = -\int_{\mathcal{X}} f_{\mathbf{Z}}(\mathbf{z}) \log f_{\mathbf{Z}}(\mathbf{z})\, d\mathbf{z}, \tag{3}$$

when $\alpha = 1$. Here $f_{\mathbf{Z}}$ stands for the density underlying $\mathbf{Z}$.

The joint distribution of $\mathbf{Z}$, when restricted to $[0,1]^d$ is called the *copula* of $\mathbf{X}$, because it shows only the couplings between the components of $\mathbf{X}$ and does not depend on the marginals of $\mathbf{X}$. $H_\alpha(\mathbf{Z})$ above depends only on the copula of $\mathbf{X}$ and is called the copula's entropy. The approach of working with the rank-order statistics to estimate dependence goes back at least to Spearman (1904), but surprisingly, for some reason unknown to us , although the relationship of a random variable's mutual information and its copula's entropy must be well known, to the best of our knowledge the approach of estimating the mutual information based on the rank-order statistics have never been suggested or explored before. Yet we think, this step is highly advantageous and should always be used when estimating information.

3.2 Step 2: Entropy Estimation based on Euclidean Graph Optimization

The next step of the algorithm estimates the α-entropy of $\mathbf{Z}$ based on the sample $\hat{\mathbf{Z}}_{1:n}$. For this, we propose to use *Euclidean Graph Optimization* (EGO) which we briefly review now.

EGO entropy estimators expect an i.i.d. sample and produce an estimate of the α-entropy ($0 < \alpha, \alpha \neq 1$) of the underlying common distribution. Although in our case the sample is not i.i.d., we will show that these estimator can still be used. In our description of EGO estimators we follow Steele (1996) with slight changes so that the correspondence to entropy estimation becomes more transparent.

Let $\mathfrak{G}$ be a system of graphs on n nodes numbered from 1 to n (specific examples will be given later), for a graph $G \in \mathfrak{G}$ let $E(G) \subset \{1, \ldots, n\}^2$ be its edge-set. Let G_n be the complete graph on n nodes. Thus, each graph $G \in \mathfrak{G}$ is a subgraph of G_n. Define

$$L_n(\hat{\mathbf{Z}}_{1:n}) = \min_{G \in \mathfrak{G}} \sum_{(i,j) \in E(G)} \|\hat{\mathbf{Z}}_i - \hat{\mathbf{Z}}_j\|^p, \tag{4}$$

the minimal total p-power weighted edge length above the graphs in $\mathfrak{G}$, where $p(= p_\alpha) = d(1-\alpha)$. Define the estimate of the entropy as

$$H_n(\hat{\mathbf{Z}}_{1:n}) = \frac{1}{1-\alpha} \log \frac{L_n(\hat{\mathbf{Z}}_{1:n})}{\gamma_{d,\alpha}\, n^\alpha}. \tag{5}$$

To emphasize the dependence of H_n on $\mathfrak{G}$, we will sometimes write $H_n = H_n(\hat{\mathbf{Z}}_{1:n}; \mathfrak{G})$. Above $\gamma_{d,\alpha} > 0$ is a universal constant (i.e., its value does not depend on the distribution of the data).[2] Possible systems of graph sets $\mathfrak{G}$ include $\mathfrak{G}_{\mathrm{ST}}$, all spanning trees of G_n; $\mathfrak{G}_{\mathrm{H}}$, the set of all Hamiltonian cycles of G_n, $\mathfrak{G}_{\mathrm{R}(k)}$ ($k > 0$), the set of all subgraphs of G_n where the out-degree of each node is k; and more. The functional $L_n(\hat{\mathbf{Z}}_{1:n})$ is known as *Euclidean functional*. Different choices of $\mathfrak{G}$ lead to different optimization problems. When $\mathfrak{G} = \mathfrak{G}_{\mathrm{ST}}$, computing L_n amounts to finding the (p-weighted) minimal spanning tree (MST). When $\mathfrak{G} = \mathfrak{G}_{\mathrm{H}}$, we need to solve a traveling salesman problem (TSP), while when $\mathfrak{G} = \mathfrak{G}_{\mathrm{R}(k)}$, L_n can be computed by finding the k-nearest neighbors (k-NN) for each node and summing up the p-power edge-lengths. Note that the complexity of computing L_n for the MST (k-NN problem) is $O(n^2)$ with Prim's algorithm (Prim, 1957) (resp., $O(kn \log n)$ using Dickerson and Eppstein, 1996).

Rationale The following theorem shows that the above construction is indeed a good procedure to estimate entropies:

Theorem 1 (Steele (1988), Yukich (1998)). *Let $d \ge 2$, $0 < \alpha < 1$. Let $\mathbf{Z}, \mathbf{Z}_1, \ldots, \mathbf{Z}_n$ be i.i.d. random variables, supported on $[0,1]^d$ with density $f = f_{\mathbf{Z}}$.*

[2] Specific values of γ, if needed, can be estimated using Monte-Carlo.

Assume that $\mathfrak{G} \in \{\mathfrak{G}_{\mathrm{ST}}, \mathfrak{G}_{\mathrm{H}}, \mathfrak{G}_{\mathrm{R}(k)}\}$ *and consider the estimator* $H_n = H_n(\mathbf{Z}_{1:n}; \mathfrak{G})$. *Then* $H_n \to H_\alpha(\mathbf{Z})$ *almost surely as* $n \to \infty$.

The theorem was proven by Steele for the case of $\mathfrak{G} = \mathfrak{G}_{\mathrm{ST}}$. The $\mathfrak{G}_{\mathrm{H}}$, the $\mathfrak{G}_{\mathrm{R}(k)}$ with $\alpha = 1 - 1/d$, and many other cases were investigated by Yukich. Motivated by his work Pál et al. (2010) proved the theorem for a more general class of k-NN graphs including $\mathfrak{G}_{\mathrm{R}(k)}$, and applicable for any $\alpha \in (0, 1)$.

Previous work The first published work that we know of where EGO was suggested to be used for entropy estimation is due to Hero and Michel (1998, 1999a). More precisely, they consider the problem of estimating entropy in the presence of a contamination with a *known* distribution, but an unknown fraction $0 \le \epsilon < 1$. For this they proposed to use an estimator that is based on solving the so-called k-MST, which, for $1 \le k \le n$ is defined as the problem of finding the MST based on k out of the n points that must be picked by the algorithm optimally. The ideal value of k is $k = (1-\varepsilon)n$, but since ε is unknown, the value of k must be picked based on the sample. Although no algorithm was proposed for doing this, the experimental result in these works are quite convincing and in fact this is what motivated us to choose EGO as the method of entropy estimation. For $k(= k_n) = (1-\varepsilon)n$, Hero and Michel (1999a) proved convergence to the minimum conditional Rényi entropy over the sets A in the domain which have a mass of at least $1-\varepsilon$, i.e., for which $\mathbb{P}(\mathbf{X} \in A) \ge 1 - \varepsilon$ (for $\varepsilon = 0$ this result reduces to Theorem 1). Note that although this algorithm has a robust aspect to it, it is not trying to address the classical problem of robust statistics where the contamination is assumed to come from an *arbitrary distribution* and the proportion of contamination is assumed to be negligible compared to the sample size. (As argued by Huber (2009), in case of proportional contamination, one should not use a robust procedure, but a procedure that removes the contaminating observations, which is a different problem than making an estimator robust.) In fact, when $k = k_n$, $k_n/n \nrightarrow 1$, the algorithm is *not* a consistent estimator of the Rényi entropy. Hero and Michel (1999a) put a great amount of work into making their algorithm computationally efficient for a fixed k (the unmodified k-MST problem is hard), but since the algorithm needs to search for the right value of k it can be expected to be n times more expensive than a straightforward EGO algorithm. Hero and Michel (1999b) gives an application of k-MST to two-sample divergence estimation. Recently, Jiménez and Yukich (2005) suggested and analyzed EGO procedures for ϕ-divergence estimation between an unknown distribution that generates the sample and a known distribution.

4 Theory

In the first part of this section we prove results on the asymptotics of our estimator, while in the second part we prove results that show that our estimator is robust.

4.1 Strong consistency

The goal of this section is to show the strong consistency of our algorithm. The analysis starts from the observation that (under appropriate conditions) it follows immediately from Theorem 1 that $H_n(\mathbf{Z}_{1:n}, \mathfrak{G}) \to H_\alpha(\mathbf{Z})$ $(\alpha > 0, \alpha \ne 1)$ where $\mathbf{Z}_{1:n} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)$ with $Z_{tj} = F_{X_j}(X_{tj})$. The difficulty in showing that $H_n(\hat{\mathbf{Z}}_{1:n}, \mathfrak{G})$ also converges to H_α is twofold: First, since $F_{nj} \ne F_{X_j}$, $\hat{\mathbf{Z}}_t \ne \mathbf{Z}_t$. Second, the sample $\hat{\mathbf{Z}}_1, \dots, \hat{\mathbf{Z}}_n$ is not an i.i.d. sample (since F_{nj} is estimated based on $\mathbf{X}_{1:n}$). Hence, Theorem 1 is not directly applicable to $H_n(\hat{\mathbf{Z}}_{1:n}, \mathfrak{G})$. Nevertheless, we can still prove the following theorem:

Theorem 2. *Let* $d \ge 3$, $1/2 < \alpha < 1$. *Let* $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ *be i.i.d. random variables, supported on* $[0,1]^d$ *with density* $f = f_{\mathbf{X}}$. *Assume that* $\mathfrak{G} \in \{\mathfrak{G}_{\mathrm{ST}}, \mathfrak{G}_{\mathrm{H}}, \mathfrak{G}_{\mathrm{R}(k)}\}$ *and consider the corresponding estimator* $H_n = H_n(\hat{\mathbf{Z}}_{1:n}; \mathfrak{G})$ *obtained by running the algorithm of Section 3.1 on* $\mathbf{X}_{1:n} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$. *Then* $-H_n \to I_\alpha(\mathbf{X})$ *almost surely as* $n \to \infty$.

We prove the theorem in three steps. First we show that $\mathbf{Z}_i$ and $\hat{\mathbf{Z}}_i^{(n)}$ are close to each other for large n, where, for the sake of precision the upper index n was introduced to denote that $\hat{\mathbf{Z}}_i^{(n)}$ is obtained based on the empirical distribution of the marginals using $\mathbf{X}_{1:n}$.

Lemma 1. *Almost surely*

$$\limsup_n (d \log(2dn^2)/(2n))^{-1/2} \max_{1 \le t \le n} \|\mathbf{Z}_t - \hat{\mathbf{Z}}_t^{(n)}\| \le 1$$

Proof. Fix n. Let $F_j = F_{X_j}$. By a union bound argument and the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990), for any $0 < \delta < 1$, $\max_{1 \le j \le d} \|F_j - F_{nj}\|_\infty \le \sqrt{\frac{\log(2d/\delta)}{2n}}$ holds with probability at least $1-\delta$. Since for any $1 \le t \le n$, $\|\mathbf{Z}_t - \hat{\mathbf{Z}}_t^{(n)}\|^2 \le d \max_{1 \le j \le d} \|F_j - F_{nj}\|_\infty^2$, the probability of the event

$$\mathcal{E}_n = \left\{ \omega : \max_{1 \le t \le n} \|\mathbf{Z}_t - \hat{\mathbf{Z}}_t^{(n)}\| > \sqrt{\frac{d \log(2dn^2)}{2n}} \right\}$$

is at most $1/n^2$. Hence, $\sum_{n=1}^\infty \mathbb{P}(\mathcal{E}_n) < \infty$. Thus, by the first Borel-Cantelli lemma, the probability that the event $\mathcal{E}_n$ happens for infinitely many n is zero, and therefore the conclusion of the lemma must hold. □

Our second results concerns the stability of the EGO functional L_n:

Lemma 2. *Fix $d \geq 3$, $1/2 < \alpha < 1$. Let L_n be any of the EGO functionals where $\mathfrak{G} = \mathfrak{G}_n$ is any graph set on n nodes such that the graphs in $\mathfrak{G}_n$ have at most $O(n)$ edges. Then there exists a function $g : \mathbb{N} \times \mathbb{R} \to \mathbb{R}^+$ such that for any $n \geq 1$, $\mathbf{z}_{1:n}$, $\hat{\mathbf{z}}_{1:n}$,*

$$|L_n(\mathbf{z}_{1:n}) - L_n(\hat{\mathbf{z}}_{1:n})| \leq g(n, \max_{1 \leq i \leq n} \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|)$$

and that $g_n \stackrel{\text{def}}{=} g(n, \sqrt{d \log(2dn^2)/(2n)}) = o(n^\alpha)$.

Proof. Fix $\mathfrak{G}$, α, n, $\mathbf{z}_{1:n}, \hat{\mathbf{z}}_{1:n}$. Let $p = d(1-\alpha)$ (and hence $\alpha = 1 - p/d$). For $e = (i,j)$, $1 \leq i, j \leq n$, let $c(e) = \|\mathbf{z}_i - \mathbf{z}_j\|^p$ and let $\hat{c}(e) = \|\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j\|^p$. Let $L_n = L_n(\mathbf{z}_{1:n}) = \sum_{e \in E} c(e)$ and $\hat{L}_n = L_n(\hat{\mathbf{z}}_{1:n}) = \sum_{e \in \hat{E}} c(e)$, where $E = E(G)$ for some $G \in \mathfrak{G}$ and $\hat{E} = E(\hat{G})$ for some $\hat{G} \in \mathfrak{G}$. By the optimality of the edge-set $\hat{E}$, $\hat{L}_n = \sum_{e \in \hat{E}} \hat{c}(e) \leq \sum_{e \in E} \hat{c}(e) \leq L_n + \sum_{e \in E} |\hat{c}(e) - c(e)|$. Similarly, by the optimality of E, $L_n \leq \hat{L}_n + \sum_{e \in \hat{E}} |\hat{c}(e) - c(e)|$. Hence,

$$|L_n - \hat{L}_n| \leq \max(\sum_{e \in E} |\hat{c}(e) - c(e)|, \sum_{e \in \hat{E}} |\hat{c}(e) - c(e)|).$$

Case 1: $0 < p \leq 1$. Consider an egde $e = (i,j)$: Using the triangle inequality and $(|a| + |b|)^p \leq |a|^p + |b|^p$, we get $c(e) = \|\mathbf{z}_i - \mathbf{z}_j\|^p \leq \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^p + \|\mathbf{z}_j - \hat{\mathbf{z}}_j\|^p + \|\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j\|^p \leq \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^p + \|\mathbf{z}_j - \hat{\mathbf{z}}_j\|^p + \hat{c}(e)$. By symmetry, we get

$$|c(e) - \hat{c}(e)| \leq \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|^p + \|\mathbf{z}_j - \hat{\mathbf{z}}_j\|^p.$$

Since by assumption $|E|, |\hat{E}| \leq Cn$ with some $C > 0$, $|L_n - \hat{L}_n| \leq 2Cn(\max_{1 \leq t \leq n} \|\mathbf{z}_i - \hat{\mathbf{z}}_i\|)^p$. Thus, we may choose $g(n, M) = 2CnM^p$. Hence, $g_n = O(n^{1-p/2} \log^{p/2}(n^2)) = o(n^{1-p/d})$, provided that $d > 2$.

Case 2: $1 < p < d/2$. Consider again an edge $e = (i,j)$. Let $\mathbf{x} = \mathbf{z}_i - \mathbf{z}_j$, $\hat{\mathbf{x}} = \hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j$, $g : \mathbb{R}^d \to \mathbb{R}$, $g(\cdot) = \|\cdot\|^p$. Let L be the Lipschitz constant of g over $[-1,1]^d$. Then $|c(e) - \hat{c}(e)| = |g(\mathbf{x}) - g(\hat{\mathbf{x}})| \leq L\|\mathbf{x} - \hat{\mathbf{x}}\| \leq L(\|\mathbf{z}_i - \hat{\mathbf{z}}_i\| + \|\mathbf{z}_j - \hat{\mathbf{z}}_j\|) \leq 2L \max_{1 \leq t \leq n} \|\mathbf{z}_t - \hat{\mathbf{z}}_t\|$. Thus, we may choose $g(n, M) = 2CLnM$, making $g_n = O(n^{1/2} \log^{1/2}(n)) = o(n^{1-p/d})$ where the last relation holds thanks to $p < d/2$. □

Proof of Theorem 2. Fix $\mathfrak{G}$, α. By Theorem 1, $H_n(\mathbf{Z}_{1:n}) \to H_\alpha(\mathbf{Z})$ a.s. On the other hand, by the argument in Section 3.1, $-H_\alpha(\mathbf{Z}) = I_\alpha(\mathbf{X})$. By the definition of $H_n(\hat{\mathbf{Z}}_{1:n})$, it suffices to show that $|L_n(\mathbf{Z}_{1:n}) - L_n(\hat{\mathbf{Z}}_{1:n})|/n^\alpha \to 0$ almost surely. However, this follows from Lemmas 1 and 2. □

Discussion Unfortunately, our proof technique does not work when $d = 2$ or when $\alpha < 1/2$. The latter case is maybe less interesting (the most interesting case is when $\alpha \to 1$). Let us thus discuss the case when $d = 2$: Our (wild) conjecture is that the algorithm is still strongly consistent. However, to prove this conjecture will probably require a completely different approach. One way to "patch" the algorithm is to add an extra dimension to the inputs by generating the new coordinates (say) uniformly at random, independently of each other and the sample. Although this way it becomes possible to apply our result, the variance of the resulting estimate will in general increase. An alternative is to split the sample and use the first half of it for constructing the empirical distributional transformation with which the second half is transformed and would then be the subject of the further computations. Again, this may result in a decreased accuracy.

Although we have not given an algorithm to estimate the Shannon information, such an estimator could be constructed by making $\alpha = \alpha_n \to 1$ as $n \to \infty$. However, details and analysis of such an estimator are left for future work.

4.2 Robustness

The purpose of this section is to study the robustness properties of our estimator. We start by analyzing the finite-sample breakdown point. This is followed by the study of the influence curve.

4.2.1 The finite-sample breakdown point

Given some sample-size n, the finite-sample breakdown point ϵ_n^* of an estimator indicates the proportion of outliers that the estimator can tolerate in that if the proportion of outliers is larger than this critical value the estimator "breaks down", i.e., it can give arbitrarily large estimates (Huber, 1981). The breakdown point of any estimator that uses the rank-order statistics of the sample and which gives uniformly bounded estimates no matter what the rank-order statistics is has a breakdown point of 1 (i.e., is maximally robust). It is known (Yukich, 1998, e.g.) that for the functionals we consider $\sup_{\hat{\mathbf{z}}_{1:n} \in [0,1]^d} L_n(\hat{\mathbf{z}}_{1:n}) \leq Cn^\alpha$ with some $C > 0$. Hence, $\sup_{\hat{\mathbf{z}}_{1:n} \in [0,1]^d} H_n(\hat{\mathbf{z}}_{1:n}) < +\infty$, i.e., our estimator gives uniformly bounded estimates irrespectively of its input. Hence, its break-down point is 1. Note that here the key is that the sample is transformed to $[0,1]^d$. Had we left out the first step of the algorithm (e.g., when considering entropy estimation), the breakdown point would become $1/n$ since by moving a single point arbitrarily far from the rest of the points we can increase L_n without any

limit. One way to deal with this issue would be to use truncated edge-weights, which does not influence the asymptotics. However, this would introduce an additional free parameter. The k-MST estimator might be another solution (Hero and Michel, 1998), though, as discussed earlier, this estimator will be biased unless $k = k_n$, $k_n/n \to 1$. It is, however, still interesting to study its robustness. In this case (assuming $k > n/2$) if we change fewer than $n - k$ points the estimate will stay bounded no matter how these points are changed. Hence, in this case $\epsilon_n^* = (n - k + 1)/n = \epsilon + 1/n$, if $k = (1 - \epsilon)n$. At this stage, it is however important to emphasize that we do not know how these entropy estimators could be used to estimate Rényi information unless our suggestion of applying an empirical distributional transformation is used as the first step of the algorithm.

4.2.2 Stability

The breakdown point measures how resistant the estimator is against outliers. However, it fails to quantify the stability of the estimator to outliers. For this, the following measure, inspired by Tukey's finite-sample influence curve, can be used: Define $\Delta_n(\mathbf{x}) = |H_{n+1}(\mathbf{X}_{1:n}, \mathbf{x}) - H_n(\mathbf{X}_{1:n})|$ the amount of change caused in the estimate by adding a single observation $\mathbf{x}$ to the sample $\mathbf{X}_{1:n}$. In general, we would like $\Delta_n(\mathbf{x}) = o(1)$ to hold a.s. independently of $\mathbf{x}$ as this indicates that the effect of a single sample becomes negligible as $n \to \infty$. (Tukey's criterium for stability was $\Delta_n(\mathbf{x}) = O(n^{-1})$, and of course, this, in general, is the best that we can hope for.) We have the following result:

Proposition 3. *Let* $\mathfrak{G} = \mathfrak{G}_{\mathrm{ST}}$. *Then* $\Delta_n(\mathbf{x}) = O(n^{-\alpha})$ *holds a.s., uniformly in* $\mathbf{x}$.

Proof. It is known (Yukich, 1998) that $\{L_n\}$ is smooth on the unit cube in the sense that for any $n, m \geq 1$, and $\mathbf{z}_{1:n+m} \subset [0,1]^d$, $|L_{n+m}(\mathbf{z}_{1:n+m}) - L_n(\mathbf{z}_{1:n})| \leq Cm^\alpha$ with some $C > 0$. Fix the sample $\mathbf{X}_{1:n}$ and let $L_n = L_n(\mathbf{X}_{1:n})$ and $L'_{n+1} = L_{n+1}(\mathbf{X}_{1:n}, \mathbf{x})$. By exploiting the smoothness property and since by Theorem 1 $L_n = \Theta(n^\alpha)$ it follows that

$$\begin{aligned}
\Delta_n(\mathbf{x}) &= \frac{1}{1-\alpha}\left|\log\frac{L'_{n+1}}{(n+1)^\alpha} - \log\frac{L_n}{n^\alpha}\right| \\
&\leq \frac{1}{1-\alpha}\frac{\left|L'_{n+1} - L_n\right|}{\min\{L_n, L'_{n+1}\}} - \frac{\alpha}{1-\alpha}\log\frac{n}{n+1} \\
&\leq O(n^{-\alpha}).
\end{aligned}$$

□

Note that the smoothness property is expected to hold for many other graphs (and in fact we only need a much weaker condition). Smoothness holds for $\alpha = 1 - 1/d$ for $L_n(\cdot; \mathfrak{G}_{R(k)})$ (Yukich, 1998).

Sometimes, stability is measured by the so-called population influence curve. However, this measure can only be used when the statistics of interest is obtained by applying a fixed T functional to the empirical distribution function. Since our estimator is not a plug-in estimator (just like many other estimators)[3], the infinitesimal approach is not applicable and we shall not explore it further.

5 Experiments on Simulated Data

The purpose of these experiments is to check consistency in the two-dimensional case (going beyond our theoretical results), as well as to check consistency and rate of convergence in a higher dimensional setting.

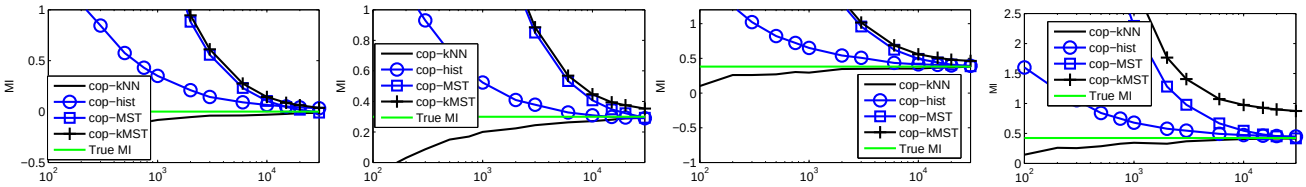
5.1 Consistency in 2D

We conjecture that our estimator is consistent for $d = 2$. Absent of proof, our goal here is to verify this empirically. In particular, we consider several REGO estimators based on either the k-NN, the MST, or the k-MST estimators (the k-MST estimator is expected to be biased). We have also implemented an information estimator that we call *cop-hist* and which applies a standard, well-tested 2D histogram based plug-in entropy estimator due to Scott (1979) on the empirical copula (i.e., the empirical distribution function of the transform obtained in the first step of our method).

The following numerical experiments support this conjecture. For the plots in Fig. 1, i.i.d samples were drawn from two random variables that are (a) independent *Beta* $(3, 4)$, (b) uniform on a square $[-0.5, 0.5]^2$ rotated by $\pi/4$, (c) a distribution with *Gamma* marginals and a t-copula, where we chose the copula and marginal parameters randomly, and (d) jointly Gaussian with $\Sigma_{11} = 7$, $\Sigma_{12} = 2$, $\Sigma_{21} = 2$, $\Sigma_{22} = 1$ covariance matrix. We used $\alpha = 0.999$ (approximating Shannon information at $\alpha = 1$), and $k = 3$ for the k-NN. As for the k-MST we used $k = \lfloor 0.95n \rfloor$ in Fig. 1(a)- 1(c), and $k = \lfloor 0.8n \rfloor$ in Fig. 1(d).

Fig.1 demonstrates that (with the exception of the k-MST, which is not expected to be consistent) as the number of samples (x-axis) increases, the estimates converge to the true Shannon information (y-axis).

[3] This also applies to the estimator of Hero and Michel (1998), who analyzed the influence function of the functional their estimator converges to in the limit. It is however questionable of the properties of this influence function to have any relevance whatsoever to the robustness of their estimator. We have found this part of the literature of robust statistics quite controversial for several reasons. Huber (2009) lists a few problems with the infinitesimal approach.



(a) Independent beta (b) Rotated uniform (c) t-copula (d) Gauss

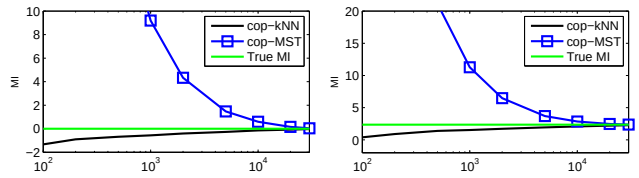
Figure 1: Consistency in 2D. The estimated MI in the number of samples. The curves are averaged over 5 runs.

5.2 Consistency in 10D

In the next experiment we demonstrate the consistency of our algorithm when the dimension is larger, in particular, for $d = 10$. The histogram based methods are not efficient when $d > 4$ due to the curse of dimensionality. Also, we already know that the k-MST is asymptotically biased, thus we do not show results for these estimators. Figure 2 demonstrates that the REGO estimators based on either the MST, or the k-NN graphs are consistent. In order to achieve reasonable estimates, in this case we need $n = 30,000$ samples, which is remarkable given that $n^{1/d} = 2.8$ only. In Fig. 2(a) and 2(b) the task was to estimate the MI between the marginals of a 10D uniform distribution with $[0,1]^{10}$ support, and a 10D distribution with Gaussian copula and Gamma distributed marginals, where the parameters of this distributions were chosen randomly.

6 Application to Image Registration

Image registration is an important application of mutual information estimation. We use this application to demonstrate the efficiency of our estimators, as well as their robustness to outliers. We repeated the experiment proposed in Kybic (2006) for the evaluation of MI estimators: Given two images represented as vectors $(X_{1i}, \dots, X_{ni}) \in [0,1]^n$, $i = 1, 2$, the task is to estimate



(a) 10D uniform (b) 10D Gauss copula

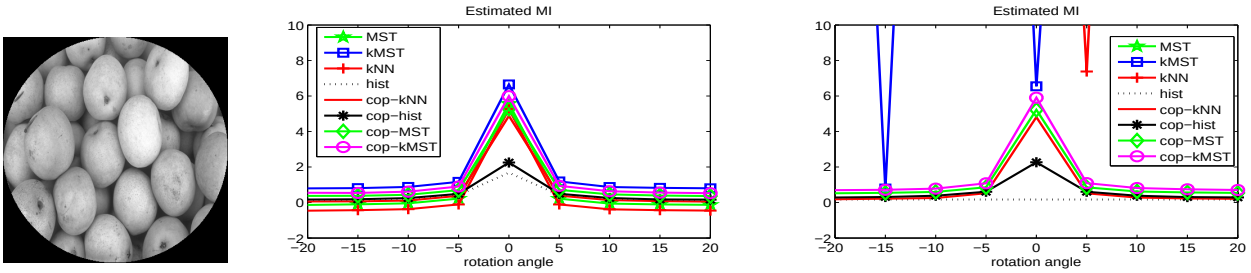
Figure 2: Consistency in 10D. The estimated MI in the number of samples. The curves are averaged over 5 runs.

the mutual information of the sample $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ where $\mathbf{X}_t = (X_{t1}, X_{t2})^T$. Following Kybic (2006), we used the 70×70 version ($n = 4,900$) of Fig. 3(a) with the pixel intensities normalized to $[0,1]$ as one of the image ($i = 1$), and the rotated version of the same image as the comparison image ($i = 2$).

We compared several estimators (k-NN, MST, k-MST, and the plug-in histogram estimator of Scott (1979)) with and without the empirical distributional transformation. Fig. 3(b) shows the results, we used $\alpha = 0.99$ in our experiments. As expected, all the estimators achieve their maximum when the rotation angle is zero: from the point of view image registration they perform perfectly. To test their sensitivity to outliers, we corrupted 200 pixels (ca. 4%) of the rotated images by random values and rerun the algorithms. This is a realistic situation in the image registration, where, in practice, we would expect even larger corruption. The results are shown on Fig. 3(c). As expected, the rank-based estimators were not influenced by the outliers. However, the response of the EGO methods based on MST, k-MST, and k-NN is so heavily influenced that they either just partly fit the figure if at all. We have also experimented with kernel density (KDE) based plug-in estimators, but its performance was very similar to that of the histogram based estimator, thus we do not show it here.

7 Discussion and conclusion

Our method is unique in that *(i)* it is strongly consistent, *(ii)* it is remarkably robust as it works only with ranks, *(iii)* it is computationally efficient, *(iv)* it converges quickly, *(v)* it works for distributions with unbounded support, *(vi)* it is insensitive to the choice of the parameters. Many of these good properties are the result of applying the probability integral transformation based on the empirical distribution function, thereby reducing the problem of estimating information to that of estimating the neg-entropy of the resulting random variable. Although in this paper we advocated the use of EGO techniques for this latter task, other methods can also be used after the transformation. Indeed, in our experiments, we have used histogram based estimators, which seemed to work



(a) Image (b) Without outliers (c) With outliers

Figure 3: Image registration experiment. (a) The original image. (b) Performance on clean data. (c) Performance in the presence of 200 outliers: only the copula based estimators gave reasonable estimates. Some of the estimators behaved so poorly that we can only partly show their response. The response of the MST estimator did not fit on the graph at all. The curves show the results of a typical run.

well in the low-dimensional cases. A few important questions remained unanswered in connection to our method: Most importantly, we could not prove the consistency of our estimator in two dimensions. While an upper bound on the rate of convergence has been derived in Pál et al. (2010), not much is known about lower rates of convergence.

References

Beirlant, J., Dudewicz, E. J., Györfi, L., and Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.*, 6:17–39.

Dickerson, M. and Eppstein, D. (1996). Algorithms for proximity problems in higher dimensions. *Comput. Geometry: Theory and Applications*, 5(5):277–291.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27.

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

Hero, A. O. and Michel, O. J. J. (1998). Robust entropy estimation strategies based on edge weighted random graphs. In *Proc. of Meeting of Intl. Soc. for Optical Engin.*, pages 250–261.

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

Hero, A. O. and Michel, O. J. J. (1999b). Estimation of Rényi information divergence via pruned minimal spanning trees. In *Proc. IEEE Workshop Higher Order Statistics*, Caesaria, Israel.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Huber, P. J. (2009). On the non-optimality of optimal procedures. In *Optimality: The Third Erich L. Lehmann Symposium*, volume 57, pages 31–46. Institute of Mathematical Statistics, Beachwood, OH, USA.

Jiménez, R. and Yukich, J. E. (2005). Statistical distances based on Euclidean graphs. In *Recent Advances in Probability and its Applications*, pages 223–239. Springer.

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

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

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

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

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

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

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Syst. Tech. Journal*, 36:1389–1401.

Rényi, A. (1961). On measure of entropy and information. In *Proc. of the Fourth Berkeley Symposium on Mathematics, Statistics, and Probability*, pages 547–561.

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Steele, J. M. (1988). Growth rates of Euclidean minimal spanning trees with power weighted edges. *Annals of Probability*, 16:1767–1787.

Steele, J. M. (1996). *Probability Theory and Combinatorial Optimization*. SIAM.

Tukey, J. W. (1970). *Exploratory Data Analysis. Mimeograph*. Addison-Wesley.

Vajda, I. (1989). *Theory of Statistical Inference and Information*. Kluwer, Boston.

Yukich, Y. E. (1998). *Probability Theory Of Classical Euclidean Optimization Problems*, volume 1675 of *Lecture Notes in Mathematics*. Springer.