# NONPARAMETRIC DIVERGENCE ESTIMATORS FOR INDEPENDENT SUBSPACE ANALYSIS

*Barnabás Póczos*[1], *Zoltán Szabó*[2], *and Jeff Schneider*[1]

[1]School of Computer Science,
Carnegie Mellon University,
5000 Forbes Ave, 15213, Pittsburgh, PA, USA
email: {bapoczos,schneide}@cs.cmu.edu
web: http://www.autonlab.org

[2]Faculty of Informatics,
Eötvös Loránd University,
Pázmány P. sétány 1/C, H-1117 Budapest, Hungary
email: szzoli@cs.elte.hu
web: http://nipg.inf.elte.hu

## ABSTRACT

In this paper we propose new nonparametric Rényi, Tsallis, and $L_2$ divergence estimators and demonstrate their applicability to mutual information estimation and independent subspace analysis. Given two independent and identically distributed samples, a "naïve" divergence estimation approach would simply estimate the underlying densities, and plug these densities into the corresponding integral formulae. In contrast, our estimators avoid the need to consistently estimate these densities, and still they can lead to consistent estimations. Numerical experiments illustrate the efficiency of the algorithms.

## 1. INTRODUCTION

Many statistical, artificial intelligence, and machine learning problems require efficient estimation of the divergence between two distributions. This problem is challenging when the density functions are not given explicitly, or they are not members of any parametric family, and parametric approaches cannot be applied. We assume that only two finite, independent and identically distributed (i.i.d.) samples are given from the two underlying distributions.

The Rényi-$\alpha$, Tsallis-$\alpha$ , and $L_2$ divergences are important special cases of probability divergences. Divergence estimators generalize entropy estimators and can also be applied to mutual information estimation. Entropy estimators are important, e.g., in goodness-of-fit testing, parameter estimation in semi-parametric models, studying fractal random walks, and texture classification. Mutual information estimators have been used, e.g., in feature selection, clustering, causality detection, optimal experimental design, fMRI data processing, prediction of protein structures, boosting, and facial expression recognition. Both entropy estimators and mutual information estimators have been used for image registration, as well as for independent component and subspace analysis [1, 2]. For further applications, see [3].

An indirect way to obtain the desired estimates would be to use a "plug-in" estimation scheme—first, apply a consistent density estimator for the underlying densities, and then plug them into the desired formula. The unknown densities, however, are nuisance parameters in the case of divergence estimation, and we would prefer to avoid estimating them. Furthermore, density estimators usually have tunable parameters, and we may need expensive cross validation to achieve good performance.

The most relevant existing work to this paper is [4, 5], where an estimator for the KL-divergence was provided. [6, 7] investigated the Rényi divergence estimation problem but assumed that one of the two density functions is known. [8] developed algorithms for estimating the Shannon entropy and the KL divergence for certain parametric families. Recently, [9] developed methods for estimating $f$-divergences using their variational characterization properties. This approach involves solving a convex minimization problem over an infinite-dimensional function space. For certain function classes defined by reproducing kernel Hilbert spaces (RKHS), however, they were able to reduce the computational load from solving infinite-dimensional problems to solving $n$-dimensional problems, where $n$ denotes the sample size. When $n$ is large, solving these convex problems can still be very demanding. Furthermore, choosing an appropriate RKHS also introduces questions regarding model selection. An appealing property of our estimators is that we do not need to solve minimization problems over function classes; we only need to calculate certain $k$-nearest-neighbor ($k$-NN) based statistics.

Our work borrows ideas from [3] and [10], who considered Shannon and Rényi-$\alpha$ entropy estimation from a single sample. In contrast, we propose divergence estimators using two independent samples. Recently, [11, 12] proposed a method for consistent Rényi information estimation, but this estimator also uses one sample only and cannot be used for estimating divergences.

The paper is organized as follows. In the next section we introduce the Rényi, Tsallis, and $L_2$ divergences, and formally define our estimation problem. Section 3 contains our proposed divergence estimators, and here we also present our theoretical results about asymptotic unbiasedness and consistency. In Section 4 we collect the technical tools that we need for proving consistency and also present a brief sketch of the proofs. We describe how the proposed divergence estimators can be used for mutual information estimation in Section 5. We will also use these estimators for the independent subspace analysis problem (Section 6). Section 7 contains the results of our numerical experiments that demonstrate the applicability and the consistency of the estimators. Finally, we conclude with a discussion of our work.

**Notation:** Let $\mathcal{B}(x,R)$ denote the closed ball around $x \in \mathbb{R}^d$ with radius $R$, and let $\mathcal{V}\big(\mathcal{B}(x,R)\big) = cR^d$ be its volume, where $c = \pi^{d/2}/\Gamma(d/2+1)$ stands for the volume of a $d$-dimensional unit ball. We use $X_n \to_p X$ and $X_n \to_d X$ to represent convergence of random variables in probability and in distribution, respectively. $F_n \to_w F$ will denote the weak convergence of distribution functions. The set where density $p$ is strictly positive is denoted by $\mathrm{supp}(p)$. If $y \in \mathbb{R}^{d_y}$, $z \in \mathbb{R}^{d_z}$ are column vectors, then $x = [y;z] \in \mathbb{R}^{d_y+d_z}$ denotes the column vector given by the concatenation of components $y$ and $z$.

## 2. DIVERGENCES

Let $p$ and $q$ be densities over $\mathbb{R}^d$, and $\alpha \in \mathbb{R} \setminus \{1\}$. The Rényi-$\alpha$ [13], Tsallis-$\alpha$ [14], and $L_2$ divergences are defined respectively as follows.

**Definition 1.**

$$\begin{aligned} R_\alpha(p\|q) &\doteq \frac{1}{\alpha-1}\log\int p^\alpha(x)q^{1-\alpha}(x)\,\mathrm{d}x, \\ T_\alpha(p\|q) &\doteq \frac{1}{\alpha-1}\left(\int p^\alpha(x)q^{1-\alpha}(x)\,\mathrm{d}x - 1\right), \\ L(p\|q) &\doteq \left(\int (p(x)-q(x))^2\,\mathrm{d}x\right)^{1/2}. \end{aligned}$$

Since

$$\lim_{\alpha\to 1} R_\alpha(p\|q) = \lim_{\alpha\to 1} T_\alpha(p\|q) = KL(p\|q) \doteq \int p\log\frac{p}{q},$$

where KL stands for the Kullback–Leibler divergence, we define $R_1(p\|q)$ and $T_1(p\|q)$ to be $KL(p\|q)$. The following equations summarize the behavior of the $R_\alpha(p\|q)$, $T_\alpha(p\|q)$ divergences as a function of $\alpha$.

**Properties 2.**

$$\begin{aligned} &\alpha < 0 \Rightarrow R_\alpha(p\|q) \le 0,\ T_\alpha(p\|q) \le 0, \\ &\alpha = 0 \Rightarrow R_\alpha(p\|q) = T_\alpha(p\|q) = 0, \\ &0 < \alpha \Rightarrow R_\alpha(p\|q) \ge 0,\ T_\alpha(p\|q) \ge 0. \end{aligned}$$

We are now prepared to formally define the goal of our paper. Given two independent i.i.d. samples from distributions with densities $p$ and $q$, respectively, we provide $L_2$-consistent estimators for the following quantities:

$$D_\alpha(p\|q) \doteq \int p^\alpha(x)q^{1-\alpha}(x)\,\mathrm{d}x, \tag{1}$$

$$L^2(p\|q) \doteq \int (p(x)-q(x))^2\,\mathrm{d}x. \tag{2}$$

Plugging these estimates into the appropriate formulae immediately leads to consistent estimators for $R_\alpha(p\|q)$, $T_\alpha(p\|q)$, and $L(p\|q)$.

## 3. THE ESTIMATORS

In this section we introduce our estimator for $D_\alpha(p\|q)$ and $L^2(p\|q)$. From now on we will assume that $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$ and rewrite (1) and (2) as

$$D_\alpha(p\|q) = \int_{\mathcal{M}} \left(\frac{q(x)}{p(x)}\right)^{1-\alpha} p(x)\,\mathrm{d}x, \tag{3}$$

$$L^2(p\|q) = \int_{\mathcal{M}} (p(x) - 2q(x) + q^2(x)/p(x))p(x)\,\mathrm{d}x. \tag{4}$$

where $\mathcal{M} = \mathrm{supp}(p)$. Let $X_{1:N} \doteq (X_1,\dots,X_N)$ be an i.i.d. sample from a distribution with density $p$, and similarly let $Y_{1:M} \doteq (Y_1,\dots,Y_M)$ be an i.i.d. sample from a distribution having density $q$. Let $\rho_k(x)$ denote the Euclidean distance of the $k$th nearest neighbor of $x$ in the sample $X_{1:N}\setminus\{x\}$, and similarly let $\nu_k(x)$ denote the Euclidean distance of the $k$th nearest neighbor of $x$ in the sample $Y_{1:M}\setminus\{x\}$. We will prove that the following estimators are consistent under certain conditions:

$$\widehat{D}_\alpha(X_{1:N}\|Y_{1:M}) \doteq \frac{1}{N}\sum_{n=1}^{N}\left(\frac{(N-1)\rho_k^d(X_n)}{M\nu_k^d(X_n)}\right)^{1-\alpha} B_{k,\alpha}, \tag{5}$$

where $B_{k,\alpha} \doteq \frac{\Gamma(k)^2}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$, $k > |1-\alpha|$, and $d$ is the dimension of $X_n$ and $Y_m$.

$$\widehat{L}^2(X_{1:N}\|Y_{1:M}) \doteq \frac{1}{N}\sum_{n=1}^{N}\left[\frac{k-1}{(N-1)c\rho_k^d(X_n)} - \frac{2(k-1)}{Mc\nu_k^d(X_n)} + \frac{(N-1)c\rho_k^d(X_n)}{(Mc\nu_k^d(X_n))^2}\,\frac{(k-2)(k-1)}{k}\right], \tag{6}$$

where $k-2>0$ and $c = \pi^{d/2}/\Gamma(d/2+1)$.

Let $p$, $q$ be bounded away from zero, bounded from above, and uniformly continuous density functions. Let $\mathcal{M} = \mathrm{supp}(p)$ be a finite union of bounded convex sets. We have the following main theorems.

**Theorem 3** (Asymptotic unbiasedness). *If $k > |1-\alpha|$, then $\lim_{N,M\to\infty}\mathbb{E}[\widehat{L}^2] = L^2$, $\lim_{N,M\to\infty}\mathbb{E}[\widehat{D}_\alpha] = D_\alpha$, i.e., the estimators are asymptotically unbiased.*

**Theorem 4** ($L_2$ consistency). *If $k > 2|1-\alpha|$, then we have that $\lim_{N,M\to\infty}\mathbb{E}\left[(\widehat{L}^2 - L^2)^2\right] = 0$ and $\lim_{N,M\to\infty}\mathbb{E}\left[(\widehat{D}_\alpha - D_\alpha)^2\right] = 0$, i.e., the estimators are $L_2$ consistent.*

## 4. CONSISTENCY PROOFS

### 4.1 General Tools

We will need a couple of lemmas to be able to prove the main theorems of the previous section. This section collects these tools. The sketch of the proofs will be given in Section 4.3.

**Lemma 5** (Lebesgue (1910)). *If $\mathbb{R}^d \supseteq \mathcal{M}$ is a Lebesgue measurable set, and $g \in L_1(\mathcal{M})$, then for any sequence of $R_n \to 0$, $\delta > 0$, and for almost all $x \in \mathcal{M}$, there exists an $n_0(x,\delta) \in \mathbb{Z}^+$ such that if $n > n_0(x,\delta)$, then*

$$g(x) - \delta < \frac{\int_{\mathcal{B}(x,R_n)} g(t)\ \mathrm{d}t}{\mathcal{V}(\mathcal{B}(x,R_n))} < g(x) + \delta. \tag{7}$$

We will also need a stronger property; namely, we want it to hold uniformly over $x \in \mathcal{M}$. If $g$ is uniformly continuous on $\mathcal{M}$, then the following lemma also holds:

**Lemma 6.** *Let $g \in L_1(\mathcal{M})$ be uniformly continuous function on $\mathcal{M}$. Then for any $R_n \to 0$ series and $\delta > 0$, there exists $n_0 = n_0(\delta) \in \mathbb{Z}^+$ (independent of $x$!) such that if $n > n_0$, then for almost all $x \in \mathcal{M}$*

$$g(x) - \delta < \frac{\int_{\mathcal{B}(x,R_n)\cap\mathcal{M}} g(t)\ \mathrm{d}t}{\mathcal{V}(\mathcal{B}(x,R_n)\cap\mathcal{M})} < g(x) + \delta. \tag{8}$$

As we proceed we will frequently use the following lemma:

**Lemma 7** (Moments of the Erlang distribution). *Let $f_{x,k}(u) \doteq \frac{1}{\Gamma(k)}\lambda^k(x)u^{k-1}\exp(-\lambda(x)u)$ be the density of the Erlang distribution with parameters $\lambda(x) > 0$ and $k \in \mathbb{Z}^+$. Let $\gamma \in \mathbb{R}$ such that $\gamma + k > 0$. Then the $\gamma$th moments of this Erlang distribution can be calculated as $\int_0^\infty u^\gamma f_{x,k}(u)\ \mathrm{d}u = \lambda(x)^{-\gamma}\frac{\Gamma(k+\gamma)}{\Gamma(k)}$.*

By the Portmanteau lemma [15] we know that the weak convergence of $X_n \to_d X$ implies that $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ for every continuous bounded function $g$. However, generally it is *not* true that if $X_n \to_d X$, then $\mathbb{E}[X_n^\gamma] \to \mathbb{E}[X^\gamma]$. For this property to hold, the series $\{X_n\}_{n=1}^\infty$ of random variables should be asymptotically uniformly integrable too. The following lemma provides a sufficient condition for this.

**Lemma 8** (Limit of moments, [15]). *Let $X_n \to_d X$, $0 \le X_n, X$, and $\gamma \in \mathbb{R}$. If there exists an $\varepsilon > 0$ such that $\limsup_{n\to\infty}\mathbb{E}\left[X_n^{\gamma(1+\varepsilon)}\right] < \infty$, then the series $\{X_n\}_{n=1}^\infty$ is asymptotically uniformly integrable, and $\lim_{n\to\infty}\mathbb{E}[X_n^\gamma] = \mathbb{E}[X^\gamma]$.*

### 4.2 $k$-NN based Density Estimators

In the remainder of this paper we will heavily exploit some properties of $k$-NN based density estimators. In this section we define these estimators and briefly summarize their most important properties.

$k$-NN density estimators operate using only distances between the observations in a given sample ($X_{1:N}$, or $Y_{1:M}$) and their $k$th nearest neighbors. [16] define the $k$-NN based density estimators of $p$ and $q$ at $x$ as follows.

**Definition 9** ($k$-NN based density estimators)**.**

$$\hat{p}_k(x) = \frac{k/N}{\mathcal{V}\big(\mathcal{B}(x, \rho_k(x))\big)} = \frac{k}{Nc\rho_k^d(x)}, \tag{9}$$

$$\hat{q}_k(x) = \frac{k/M}{\mathcal{V}\big(\mathcal{B}(x, \nu_k(x))\big)} = \frac{k}{Mc\nu_k^d(x)}. \tag{10}$$

The following theorems show the consistency of these density estimators.

**Theorem 10** ($k$-NN density estimators, convergence in probability)**.** *If $k = k(N)$ denotes the number of neighbors applied at sample size $N$ in the $k$-NN density estimator, $\lim_{N\to\infty} k(N) = \infty$, and $\lim_{N\to\infty} N/k(N) = \infty$, then $\hat{p}_{k(N)}(x) \to_p p(x)$ for almost all $x$.*

**Theorem 11** ($k$-NN density estimators, almost sure convergence in sup norm)**.** *If $\lim_{N\to\infty} k(N)/\log(N) = \infty$ and $\lim_{N\to\infty} N/k(N) = \infty$, then $\lim_{N\to\infty} \sup_x \big|\hat{p}_{k(N)}(x) - p(x)\big| = 0$ almost surely.*

Note that these estimators are consistent only when $k(N) \to \infty$. We will use these density estimators in our proposed divergence estimators; however, we will keep $k$ fixed and will still be able to prove their consistency.

### 4.3 Proof Outline for Theorems 3-4

We can see from (9) that the $k$-NN estimation of $1/p(x)$ is simply $Nc\rho_k^d(x)/k$. Using Lemma 5, we can prove that the distribution of $Nc\rho_k^d(x)$ converges weakly to an Erlang distribution with mean $k/p(x)$, and variance $k/p^2(x)$ [3]. In turn, if we divide $Nc\rho_k^d(x)$ by $k$, then asymptotically it has mean $1/p(x)$ and variance $1/(kp^2(x))$. It implies that indeed (in accordance with Theorems 10–11) $k$ should converge to infinity in order to get a consistent estimator, otherwise the variance will not disappear. On the other hand, $k$ cannot grow too fast: if say $k = N$, then the estimator would be simply $c\rho_k^d(x)$, which is a useless estimator since it is asymptotically zero whenever $x \in \mathrm{supp}(p)$.

Luckily, in our case we do not need to apply consistent density estimators. The trick is that (3)–(4) have special forms; each term inside these equations has $\int p(x)p^\gamma(x)q^\beta(x)\,\mathrm{d}x$ form. In (5)–(6), each of these terms is estimated by

$$\frac{1}{N}\sum_{i=1}^{N} \left(\hat{p}_k(X_i)\right)^\gamma \left(\hat{q}_k(X_i)\right)^\beta B_{k,\gamma,\beta}, \tag{11}$$

where $B_{k,\gamma,\beta}$ is a correction factor that ensures asymptotic unbiasedness. Using Lemma 5, we can prove that the distributions of $\hat{p}_k(X_i)$ and $\hat{q}_k(X_i)$ converge weakly to the Erlang distribution with means $k/p(X_i)$, $k/q(X_i)$ and variances $k/p^2(X_i)$, $k/q^2(X_i)$, respectively [3]. Furthermore, they are conditionally independent for a given $X_i$. Therefore, "in the limit" (11) is simply the empirical average of the products of the $\gamma$th (and $\beta$th) powers of independent Erlang distributed variables. These moments can be calculated by Lemma 7. For a fixed $k$, the $k$-NN density estimator is not consistent since its variance does not vanish. In our case, however, this variance will disappear thanks to the empirical average in (11) and the law of large numbers.

While the underlying ideas of this proof are simple, there are a couple of serious gaps in it. Most importantly, from the Lebesgue lemma (Lemma 5) we can guarantee only the weak convergence of $\hat{p}_k(X_i)$, $\hat{q}_k(X_i)$ to the Erlang distribution. From this weak convergence we cannot imply that the moments of the random variables converge too. To handle this issue, we will need stronger tools such as the concept of asymptotically uniformly integrable random variables [15], and we also need the uniform generalization of the Lebesgue lemma (Lemma 6). As a result, we need to put some extra conditions on the densities $p$ and $q$ in Theorems 3–4. Due to the lack of space, we omit the details.

## 5. MUTUAL INFORMATION ESTIMATION

In this section we demonstrate that the proposed divergence estimators can also be used to estimate mutual information. Let $p : \mathbb{R}^d \to \mathbb{R}$ be the density of a $d$-dimensional distribution with $\{p_i\}_{i=1}^d$ marginal densities. The mutual information $I(p)$ is the divergence between $p$ and the product of the marginal densities ($\prod_{i=1}^d p_i$). Particularly, for the $L_2$ divergence we have that $I_L(p) \doteq L(p\|\prod_{i=1}^d p_i)$, and for the Rényi divergence it is given by $I_\alpha(p) \doteq R_\alpha(p\|\prod_{i=1}^d p_i)$. When $\alpha \to 1$, then $I_\alpha$ converges to the Shannon mutual information. If we are given a sample $X_1, \ldots, X_{2N}$ from $p$, we can estimate the mutual information as follows. We form one set of size $N$ by setting aside the first $N$ samples. We build another sample by randomly permuting the coordinates of the remaining $N$ observations independently for each coordinate to form $N$ independent instances sampled from $\prod_{i=1}^d p_i$. Using these two sets, we can estimate $I(p)$.

We note that we need to split the $2N$ sample points only because our consistency theorems require independent samples from $p$ and $q$. However, in practice we found that mutual information estimators ((5)–(6)) are consistent even if we do not do this, but instead use the full set of samples for $p$ as well as for $\prod_{i=1}^d p_i$.

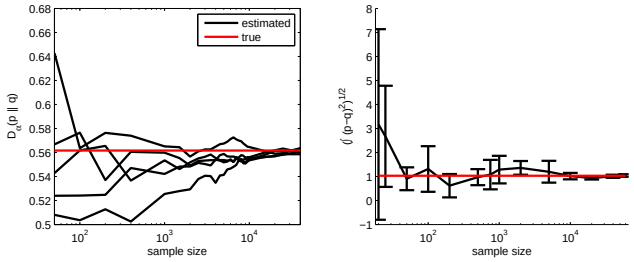
## 6. INDEPENDENT SUBSPACE ANALYSIS

In this section we briefly summarize the independent subspace analysis (ISA) problem [17]. Assume that we have $J$ hidden, independent, multidimensional $S^j \in \mathbb{R}^{d_j}$ source components ($j = 1, \ldots, J$). Suppose also that at time step $i$, only their instantaneous linear mixture is available for observation

$$O_i = AS_i, \quad (i = 1, 2, \ldots) \tag{12}$$

where $S_i = \left[S_i^1; \ldots; S_i^J\right] \in \mathbb{R}^D$ is a vector concatenated of components $S_i^j$ ($D = \sum_{j=1}^J d_j$), and $S_i^j$ denotes the $j$th hidden source component at time step $i$. We also assume that $S_i$s are i.i.d. in time $i$, and $S^j$s are non-Gaussian and jointly independent. The mixing matrix $A \in \mathbb{R}^{D\times D}$ is assumed to be invertible. The goal of the ISA problem is to estimate the original sources $S_i$ by using observations $O_i$ only. If $d_j = 1$ ($\forall j$), then the ISA problem reduces to independent component analysis (ICA) [18].

The identification of the ISA model is ambiguous. Nonetheless, the ambiguities are simple [19]: Hidden multidimensional components can be determined up to permutation and up to invertible transformation within the subspaces. In ISA, we search for the so-called demixing matrix $W \in \mathbb{R}^{D\times D}$ with which we estimate the source $S$: $\hat{S} = WO$. According to the ambiguities of the ISA problem, when the estimation is perfect, the global transform $G = WA$ is a




(a) $D_\alpha(p\|q)$ (b) $L(p\|q)$

Figure 1: Estimated vs. true divergence as a function of the sample size. The red line indicates the true divergence. The number of nearest neighbors $k$ was set to 4. (a) The results of five independent experiments are shown for estimating $D_\alpha(p\|q)$ with $\alpha = 0.7$. (b) Results for estimating $L(p\|q)$ from 50 runs. The means and standard deviations of the estimations are shown using error bars.

block-permutation matrix. This property can be measured by a simple extension of the Amari index [20] as follows. (i) Let the component dimensions and their estimations be ordered in increasing order ($d_1 \leq \ldots \leq d_J$, $\hat{d}_1 \leq \ldots \leq \hat{d}_J$), (ii) decompose $G$ into $d_i \times d_j$ blocks ($G = \left[G^{ij}\right]_{i,j=1,\ldots,J}$), and (iii) define $g^{ij}$ as the sum of the absolute values of the elements of the matrix $G^{ij} \in \mathbb{R}^{d_i \times d_j}$. Then the Amari index adapted to the ISA task of different component dimensions is defined as follows:

$$r(G) = \kappa \left[ \sum_{i=1}^{J} \left( \frac{\sum_{j=1}^{J} g^{ij}}{\max_j g^{ij}} - 1 \right) + \sum_{j=1}^{J} \left( \frac{\sum_{i=1}^{J} g^{ij}}{\max_i g^{ij}} - 1 \right) \right],$$

where $\kappa = 1/(2J(J-1))$. One can see that $0 \leq r(G) \leq 1$ for any matrix $G$, and $r(G) = 0$ if and only if $G$ is a block-permutation matrix with $d_i \times d_i$ sized nonzero blocks, and $r(G) = 1$ in the worst case.

Our proposed ISA method is as follows. According to the "ISA separation principle" [2, 17], the ISA problem can be solved by an ICA preprocessing step and then clustering the ICA elements into statistically dependent groups. For the ICA preprocessing we used the FastICA algorithm [21], and for the clustering task we estimated the pairwise mutual information of the ICA elements using the proposed Rényi ($I_\alpha$) and $L_2$ based ($I_L$) estimators ((5)–(6)). This ISA algorithm needs to know the number of subspaces $J$, but it does not need to know the true dimensions of the hidden subspaces.

## 7. NUMERICAL EXPERIMENTS

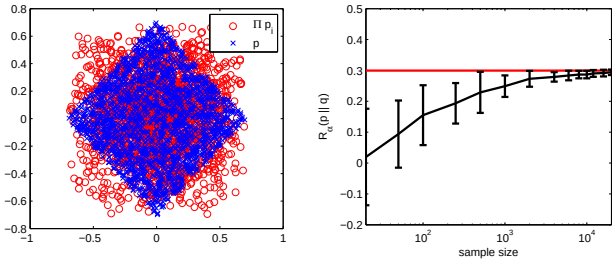
### 7.1 Demonstration of Consistency

In this section we present a few numerical experiments to demonstrate the consistency of the proposed divergence estimators. We run experiments on normal distributions because in this case the divergences have known closed-form expressions, and thus it is easy to evaluate our methods.

In Figure 1 we display the performances of the proposed $\widehat{L}$ and $\widehat{D}_\alpha$ divergence estimators when the underlying densities were zero-mean 2-dimensional Gaussians with randomly chosen covariance matrices. Our results demonstrate that when we increase the sample sizes $N$ and $M$, then the $\widehat{L}$ and $\widehat{D}_\alpha$ values converge to their true values. For simplicity, in our experiments we always set $N = M$.

### 7.2 Mutual Information Estimation

In this experiment our goal is to estimate Shannon information. For this purpose, we selected a $2d$ uniform distribution on $[-1/2, 1/2]^2$ rotated by $\pi/4$. Due to this rotation, the marginal distributions are no longer independent. Because our goal is to estimate the Shannon information, we used $R_\alpha$ and set $\alpha$ to $1 - 1/\sqrt{N}$ (so that $\alpha \to 1$). Figure 2(a) shows the original samples as well as the independent samples from the product of the marginal distributions. Figure 2(b) demonstrates the consistency of the algorithm; as we increase the sample size, the estimator approaches the Shannon information.




(a) samples from $p$ and $\prod_{i=1}^{d} p_i$ (b) $R_\alpha(p\|\prod_{i=1}^{d} p_i)$

Figure 2: Estimated vs. true Rényi information as a function of sample size. (b) The red line shows the true mutual information. The sample size was varied between 20 and 20 000, $k$ was set to 4. The error bars are calculated from 50 independent runs.

### 7.3 ISA Experiments

We illustrate the efficiency of the presented nonparametric divergence estimators for the ISA problem. We tested our algorithms on two datasets. In the *celebrities* dataset, densities of $S^j$ correspond to 2-dimensional images ($d_j = 2$), see Fig. 3(a). We chose $J = 10$ components. In our second dataset, which we call *d-spherical*, the $S^j$ components were spherical random variables. We sampled them randomly from the distribution of $VU$, where $U$ is uniformly distributed on the $d$-dimensional unit sphere, and the distribution of $V$ was set to exponential, lognormal, and uniform, respectively (see Fig. 3(b)). The dimensions of the $S^j$ components were set to $d_1 = d_2 = 6$, and $d_3 = 8$.

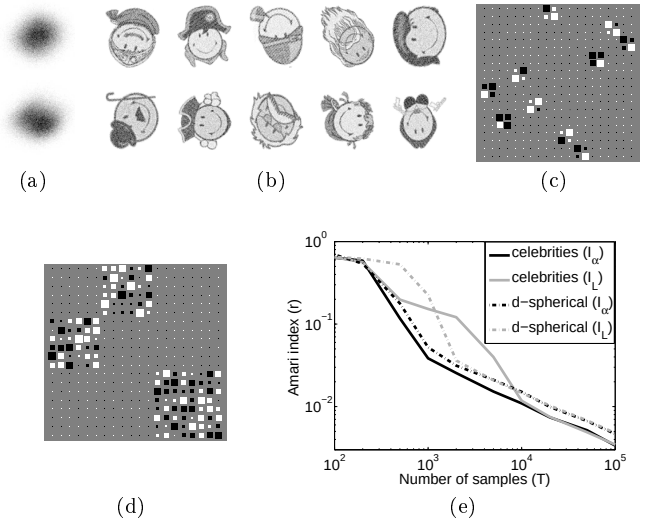
After the ICA preprocessing and the pairwise Rényi/$L_2$ information estimation steps, we clustered the ICA components by using either a greedy clustering algorithm (*celebrities* dataset), or by using the NCut spectral clustering method (*d-spherical* dataset). In the mutual information estimators, the number of neighbors was set to $k = 4$. The sample size $T$ was varied between 100 and 100 000. $\alpha$ was set to 0.997 when we estimated $I_\alpha$.

We used the Amari index to measure the performance of our algorithms. Fig. 4(e) presents how the Amari index changes as a function of the sample size. The figure shows the mean curves of 12 independent runs on the studied datasets using the $I_L$ and $I_\alpha$ estimators. The figure demonstrates that (i) both the Rényi and $L_2$ mutual information estimators can be used for solving the ISA problem. (ii) After a few thousand samples, the Amari indices decrease according to a power-law (the curves are linear on log-log scale). (iii) For small sample size, the $I_\alpha$ estimator seems to perform better than $I_L$.

Fig. 4(a) shows the first two 2-dimensional projections of the observations. Fig. 4(b) demonstrates the estimated components, and Fig. 4(c) presents the Hinton diagram of $G = WA$, this is indeed close to a block-permutation matrix.

(a) (b)

Figure 3: Illustration of the *celebrities* (a) and *d-spherical* (b) datasets.




(a) (b) (c) (d) (e)

Figure 4: The divergence estimators on ISA. Illustration of the estimations: (a)–(d), number of samples $T = 100\,000$. (a)–(c): *celebrities*, $I_\alpha$, (d): *d-spherical* dataset, $I_L$ mutual information estimation. (a): observation ($O_i$), the first two 2-dimensional projections. (b): estimated components ($\hat{S}^j$), (c)–(d): Hinton diagram of $G$s, approximately block-permutation matrices. (e): performance as a function of the sample number on log-log scale.

Fig. 4(d) shows the Hinton diagram for the experiment when our task was to separate the mixture of one 8-dimensional and two 6-dimensional $d$-spherical subspaces.

## 8. CONCLUSION AND DISCUSSION

In this paper we proposed consistent nonparametric Rényi, Tsallis, and $L_2$ divergence estimators. We demonstrated their applicability to mutual information estimation and independent subspace analysis. There are several open questions left waiting for answers. Our empirical results indicate that the conditions of our consistency theorems could be extended. Currently, we do not know the rate of the estimators either. All of our theoretical results are asymptotic, and it would be important to derive finite sample bounds too.

**Acknowledgments.** The research was partly supported by the Department of Energy (grant number DESC0002607). The European Union and the European Social Fund have provided financial support to the project under the grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003.

## REFERENCES

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

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

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

[4] Q. Wang, S. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via $k$-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5), 2009.

[5] F. Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *NIPS-2008*, volume 21, pages 1257–1264.

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

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

[8] M. Gupta and S. Srivastava. Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy*, 12:818–843, 2010.

[9] X. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE T INFORM THEORY, To appear.*, 56:5847–5861, 2010.

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

[11] B. Póczos, S. Kirshner, and Cs. Szepesvári. REGO: Rank-based estimation of Rényi information using Euclidean graph optimization. *AISTATS 2010*, pages 605–612.

[12] D. Pál, B. Póczos, and Cs. Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *NIPS 2010*, pages 1849–1857.

[13] A. Rényi. On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, 1961.

[14] T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using Frechet-derivatives, 2010. University of Applied Sciences Mittweida.

[15] A. van der Wart. *Asymptotic Statistics*. Cambridge University Press, 2007.

[16] D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics*, 36:1049–1051, 1965.

[17] J. Cardoso. Multidimensional independent component analysis. In *ICASSP'98*, volume 4, pages 1941–1944.

[18] P. Comon. Independent component analysis: a new concept? *Signal Processing*, 36:287–314, 1994.

[19] H. Gutch and F. Theis. Independent subspace analysis is unique, given irreducibility. In *ICA07*, pages 49–56.

[20] S. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind signal separation. *NIPS'96*, 8:757–763.

[21] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10:626–634, 1999.