Nonparametric Estimation of Conditional Information and Divergences

Barnabás Póczos
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
USA, 15213

Jeff Schneider
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
USA, 15213

Abstract

In this paper we propose new nonparametric estimators for a family of conditional mutual information and divergences. Our estimators are easy to compute; they only use simple k nearest neighbor based statistics. We prove that the proposed conditional information and divergence estimators are consistent under certain conditions, and demonstrate their consistency and applicability by numerical experiments on simulated and on real data as well.

1 Introduction

Conditional dependencies play a central role in machine learning and applied statistics. There are many problems where it is crucial for us to know how the relationship of two random variables changes if we observe other random variables. Correlated random variables might become independent when we observe a third random variable and the opposite situation is also possible when independent variables become dependent after observing other random variables.

Conditional mutual information (MI) can be used to capture these kind of dependencies. Although this is a fundamental problem in statistics and machine learning, interestingly, very little in known about how to estimate these quantities efficiently. *The goal of this paper* is to provide provably consistent estimators for a family of conditional mutual information and divergences. This family is large; it includes the conditional Rényi-α, Tsallis-α, Kullback-Leibler (KL), Hellinger, Bhattacharyya, Euclidean divergences and the corresponding mutual information as special cases. We also demonstrate the consistency and the applicability of the proposed estimators by numerical experiments on real as well as on simulated data sets.

The derived estimators have several potential applications. In many scientific areas (e.g., epidemiology, psychology, pharmacoinformatics, econometrics) it is crucial to discover causal relationships, to detect confounding variables, and not to infer causation from apparent correlations [Pearl, 1998, Montgomery, 2005]. We will demonstrate the applicability of our method on medical data in Section 6. In our daily life we can also easily encounter examples when people infer causation from observing correlations. Many times, however, there is a hidden factor that is responsible for this correlation. There is nonzero correlation between the reading skills of children and their shoe size. Here the underlying common factor is obviously the age. We can find many similar examples in ancient legends and folk stories too. According to a northern European legend, the stork is responsible for delivering babies to parents. Indeed, one can show that highly statistically significant correlation exists between stork populations and human birth rates across Europe [Matthews, 2000]. Conditional dependence estimators can help us identify the underlying hidden factors (confounder variables) that are responsible for these spurious relationships. Note, however, that not every variable that renders two others conditionally independent is called a confounder; conditional independence is only a necessary condition [Spirtes et al., 2001].

Conditional dependencies play a central role in Bayesian network learning as well [Zhang et al., 2011, Koller and Friedman, 2009]. It is well-known that Bayesian nets satisfy the local Markov property, that is, each variable is conditionally independent of its

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

non-descendants given its parent variables. In the structure learning problem our goal is to find an acyclic graph that is compatible with the observed data. If in a given graph satisfying the local Markov property for certain variables we find that the estimated conditional mutual information is significantly larger than zero, then this graph is not compatible with the observed data and we need to find another acyclic graph.

An indirect way to obtain the desired conditional information and divergence estimates would be to use a naïve "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. Density estimation is among the most difficult problems in statistics, and hence in many cases direct estimators, which do not apply density estimation, can achieve better performance than the "plug-in" methods. The most well-known example is the mean functional, which can be simply estimated with the empirical average, and usually we do not use sophisticated density estimators for this problem. For more complex functionals such as entropy [Leonenko et al., 2008], mutual information [Pál et al., 2010], and certain divergences [Wang et al., 2009b, Nguyen et al., 2010], empirically it was also observed that direct estimators can perform better than the "plug-in" ones. It is of great importance to know which functionals of densities can be estimated consistently without using density estimators. In this paper we show that a large family of conditional mutual information and divergences belongs to this family, and empirically also demonstrate that the proposed estimators can perform better than the naïve plug-in estimators. One might also try to use parametric approaches (e.g. mixture of Gaussians) to estimate the densities, but if the underlying density does not belong to this parametric family, then this approach leads to biased estimators, and the estimation will be inconsistent.

Although the goal of partial correlation and conditional information is similar—to describe how the dependence of random variables changes when observing other variables— the conditional information is often more informative. Partial correlation measures only "linear" association and can be zero even if there is conditional dependence between the random variables. On the contrary, the conditional information is zero iff the random variables are conditionally independent.

In machine learning, the most famous divergence and MI measures between probability distributions are the KL divergence and the Shannon information. Nonetheless, they are just the $\alpha \to 1$ limit cases of the more general Rényi-α and Tsallis-α families. For each α, these divergences behave differently, and therefore in different applications different α values (not necessarily $\alpha = 1$) might be more appropriate. For example, the Hellinger distance, which corresponds to $\alpha = 1/2$, satisfies the triangle inequality and is symmetric. The KL divergence is not symmetric and does not satisfy the triangle inequality. The Euclidean distance is always finite between distributions, while the KL divergence can be infinite. Empirically it was also observed that the convergence rates of different α-estimators depend on α and on the densities as well [Póczos and Schneider, 2011]. Therefore, even though the KL divergence and Shannon information are the most popular quantities, in certain applications other divergence and information terms can perform better and achieve faster convergence rates. Since there is no clear winner among the several (conditional) divergences and MI measures, we believe that it is important to have efficient estimators for each of them.

The paper is organized as follows. Section 2 briefly summarizes some related work. Section 3 defines the set-up of the problem and the quantities we want to estimate. In Section 4 we provide our estimators. The most important theoretical results about the consistency of the estimators are stated in Section 5. Section 5.1 contains a brief sketch of the proofs; the details are in the supplementary material [Póczos and Schneider, 2012]. Section 6 provides the results of our numerical experiments. We finish the paper with a short discussion and draw conclusions.

Notation: Unless otherwise stated, each vector in this paper will be a column vector. The dimension of x, y, z will be denoted by d_x, d_y, and d_z, respectively. $[x; y]$ will denote the $d_x + d_y$ dimensional column vector, whose first d_x coordinates correspond to x and the rest to y. The dimension of this vector will be denoted by d_{xy}. $|\Sigma|$ will denote the absolute value of the determinant of $\Sigma \in \mathbb{R}^{d\times d}$. Superscript ^{T} stands for the transposition. We use the $X_n \to_p X$ and $X_n \to_d X$ notations for the convergence of random variables in probability and in distribution, respectively. $F_n \to_w F$ will denote the weak convergence of distribution functions. $\mathcal{V}(\mathcal{M})$ stands for the volume of set $\mathcal{M}$. The size of the index set $\mathcal{J}$ is denoted by $|\mathcal{J}|$. $L_1(\mathcal{M})$ denotes the set of Lebesgue measurable functions having finite integrals over $\mathcal{M}$.

2 Related work

There is an increasing literature on the estimation of information theoretic quantities for continuous vari-

ables. Our work borrows ideas from Leonenko et al. [2008] and Goria et al. [2005], who considered Shannon and Rényi-α entropy estimation. Using Euclidean functionals [Steele, 1997, Yukich, 1998], Hero and Michel [1999] derived a strongly consistent estimator for the Rényi entropy. Póczos et al. [2010], Pál et al. [2010] combined these ideas with copula methods and proposed methods for Rényi mutual information estimation.

Wang et al. [2009b], Pérez-Cruz [2008] provided an estimator for the KL-divergence, and Póczos and Schneider [2011] proposed estimators for Rényi and Tsallis divergences. Hero et al. [2002a,b] also investigated the Rényi divergence estimation problem but assumed that one of the two density functions is known. Gupta and Srivastava [2010] developed algorithms for estimating the Shannon entropy and the KL divergence for certain parametric families, and Nguyen et al. [2009, 2010] developed methods for estimating f-divergences and likelihood ratio. Recently, Sricharan et al. [2010] proposed kth nearest neighbor based methods for estimating non-linear functionals of density, but in contrast to our approach, they were interested in the case where k increases with the sample size. Further information and useful reviews of several different divergences can be found, e.g., in Cichocki et al. [2009], and Wang et al. [2009a]. Other interesting nonparametric dependence measures include the kernel mutual information [Gretton et al., 2003], the Schweizer-Wolf measure [Schweizer and Wolff, 1981], and the distance based correlation [Székely et al., 2007].

All of the above mentioned quantities only consider the non-conditional problems, and we know very little about how to estimate the conditional versions of these quantities. Recently, Fukumizu et al. [2008] proposed a new method for estimating conditional dependence based on reproducing kernel Hilbert spaces (RKHS). There also exist methods for conditional independence tests (see e.g., Bouezmarni et al. [2009], Su and White [2008], Zhang et al. [2011]). However, these methods cannot be used for estimating conditional divergences or mutual information.

3 Conditional Mutual Information and Conditional Divergences

Definition 1 (Divergences). *Let p, q be $\mathbb{R}^d \supseteq \mathcal{M} \to \mathbb{R}$ density functions, and $\alpha \in \mathbb{R} \setminus \{1\}$. The Rényi-$\alpha$, Tsallis-$\alpha$, Kullback–Leibler, Bhattacharyya, squared Hellinger, and squared Euclidean divergences are defined respectively as follows.*

$$D_\alpha^R(p\|q) \doteq \frac{1}{\alpha-1}\log\int_{\mathcal{M}} p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x,$$

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

$$D^{KL}(p\|q) \doteq \int_{\mathcal{M}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x,$$

$$D^B(p\|q) \doteq -\log\int_{\mathcal{M}} p^{1/2}(x)q^{1/2}(x)\mathrm{d}x,$$

$$D^H(p\|q) \doteq 1-\int_{\mathcal{M}} p^{1/2}(x)q^{1/2}(x)\mathrm{d}x,$$

$$D^E(p\|q) \doteq \int_{\mathcal{M}} p^2(x)+q^2(x)-2p(x)q(x)\mathrm{d}x.$$

These quantities are nonnegative, and they are zero iff $p=q$ almost surely. These expressions can be used to measure the "distance" between two distributions. As a special case, when $p=p_{X,Y}$ is the joint density of random variables (X,Y), and $q=p_Xp_Y$ is the product of the marginal densities, then these divergences can be used to measure the mutual information.

Definition 2 (Mutual information). *Let $p_{X,Y}$ be the joint density of random variables X,Y with marginal densities p_X and p_Y, respectively. The Rényi-α and Shannon mutual information are defined respectively as follows:*

$$I_\alpha^R(X,Y) \doteq \frac{1}{\alpha-1}\log\iint p_{X,Y}^\alpha(x,y)(p_X(x)p_Y(y))^{1-\alpha}\mathrm{d}x\mathrm{d}y,$$

$$I^S(X,Y) \doteq \iint p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\mathrm{d}x\mathrm{d}y.$$

The Tsallis-α, Bhattacharyya, Hellinger, and Euclidean MI can be defined similarly.

These quantities are nonnegative, and they are zeros iff X and Y are independent from each other. In what follows we will define the conditional versions of divergences and mutual information.

Definition 3 (Conditional Rényi-α divergence). *Let X,Y,Z be random variables, $X\in\mathbb{R}^{d_x}$, $Y\in\mathbb{R}^{d_y}$, $d_x=d_y$, $Z\in\mathbb{R}^{d_z}$. Denote the densities of Z by $p_0(Z)$, and let $p_1(x|z)$, $p_2(y|z)$ be the conditional densities of X given Z, and Y given Z, respectively. Let $\alpha>0$, $\alpha\neq 1$. We define the conditional Rényi-α divergence as*

$$D_\alpha^R(p_1\|p_2;p_0) \doteq \frac{1}{\alpha-1}\log Q_1, \doteq \frac{1}{\alpha-1}\log\int p_0(z)\int p_1^\alpha(v|z)p_2^{1-\alpha}(v|z)\mathrm{d}v\mathrm{d}z,$$

where $p(v,z)\doteq p_0(z)p_1(v|z)$, $q(v,z)\doteq p_0(z)p_2(v|z)$, and $Q_1\doteq\mathbb{E}_{(V,Z)\sim p}\left[\frac{q^{1-\alpha}(V,Z)}{p^{1-\alpha}(V,Z)}\right]$.

Lemma 4. *$D_\alpha^R(p_1\|p_2;p_0)\geq 0$, and it is zero iff $p_1^\alpha(v|z)p_0(z)=p_2^\alpha(v|z)p_0(z)$ for almost all v,z.*

For a proof see the supplementary material.

Another definition could be the following expression:

$$D_\alpha^R(p_1\|p_2;p_0) \doteq \frac{1}{\alpha-1}\int p_0(z)\log\int p_1^\alpha(v|z)p_2^{1-\alpha}(v|z)\mathrm{d}v\mathrm{d}z,$$

but in this paper we do not consider this problem.

The conditional Kullback–Leibler divergence is defined as follows:

Definition 5. *Conditional Kullback–Leibler divergence*

$$D^{KL}(p_1\|p_2;p_0) \doteq \int p_0(z)\int p_1(v|z)\log\frac{p_1(v|z)}{p_2(v|z)}\mathrm{d}v\mathrm{d}z = \mathbb{E}_{(V,Z)\sim p}\left[\log\frac{p(V,Z)}{q(V,Z)}\right] \doteq Q_2.$$

Similarly, one can define the conditional Tsallis-α, Bhattacharyya, Hellinger, and Euclidean divergences too.

These conditional divergences measure how far the $p_1(\cdot|z)$, $p_2(\cdot|z)$ conditional densities are from each other on average w.r.t. the $p_0(z)$ distribution.

Having defined these quantities, we introduce the conditional mutual information as the divergence between the conditional joint densities and the product of the conditional marginal densities:

Definition 6. *Conditional Rényi mutual information*

$$I_\alpha^R(X,Y|Z) \doteq \frac{1}{\alpha-1}\log Q_3,$$

where

$$Q_3 \doteq \iiint \frac{p_Z(z)p_{X,Y|Z}^\alpha(x,y|z)}{(p_{X|Z}(x|z)p_{Y|Z}(y|z))^{\alpha-1}}\mathrm{d}x\mathrm{d}y\mathrm{d}z = \mathbb{E}_{(X,Y,Z)\sim p_{X,Y,Z}}\left[\frac{p_{X,Z}^{1-\alpha}(X,Z)p_{Y,Z}^{1-\alpha}(Y,Z)}{p_{X,Y,Z}^{1-\alpha}(X,Y,Z)p_Z^{1-\alpha}(Z)}\right].$$

The conditional mutual information measures how far the $p_{X,Y|Z=z}(\cdot,\cdot|z)$ and the $p_{X|Z=z}(\cdot|z)p_{Y|Z=z}(\cdot|z)$ densities are from each other on average w.r.t. the $p_Z(z)$ measure. One can similarly define the conditional Tsallis-α, Bhattacharyya, Hellinger, and Euclidean information. The conditional Shannon information is defined as follows:

$$I^S(X,Y|Z) = \iint p_{X,Y,Z}(x,y,z)\log\frac{p_{X,Y,Z}(x,y,z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)}\mathrm{d}x\mathrm{d}y = H(X,Z)+H(Y,Z)-H(X,Y,Z)-H(Z),$$

where H stands for the Shannon entropy. In turn, this problem reduces to the task of entropy estimation, for which there are existing tools [Leonenko et al., 2008]. Note however, that this decomposition is not possible for the other conditional divergences or mutual information.

It is well-known that $I_\alpha^R(X,Y) \to I^S(X,Y)$, and $D_\alpha^R(p_1\|p_2) \to D^{KL}(p_1\|p_2)$ as $\alpha\to 1$. The following theorem states that this also holds for the conditional versions of these quantities.

Theorem 7 (The $\alpha\to 1$ limit case)**.** *When $\alpha\to 1$, then $I_\alpha^R(X,Y|Z) \to I^S(X,Y|Z)$, and $D_\alpha^R(p_1\|p_2;p_0)\to D^{KL}(p_1\|p_2;p_0)$.*

Similar theorems hold for the (conditional) Tsallis information and divergences as well. For a proof see the supplementary material.

4 The Estimation Problem and the Estimators

Now we are ready to formally define our estimation problems. Let $(X,Y,Z)\sim p_{X,Y,Z}$ random variables, $X\in\mathbb{R}^{d_x}$, $Y\in\mathbb{R}^{d_y}$, $Z\in\mathbb{R}^{d_z}$. Let us have N i.i.d. samples from the $p_{X,Y,Z}$ distribution. They are denoted by $\{(X_n;Y_n;Z_n)\}_{n=1}^N$, where $(X_n;Y_n;Z_n)\in\mathbb{R}^{d_{xyz}}$, $d_{xyz}=d_x+d_y+d_z$. Our goal is to estimate the conditional Rényi-α, Tsallis-α, Kullback–Leibler, Bhattacharyya, squared Hellinger, and squared Euclidean divergences and information. We will only show detailed calculations for $D_\alpha^R(p_1\|p_2;p_0)$, $D^{KL}(p_1\|p_2;p_0)$, and $I_\alpha^R(X,Y|Z)$. Estimators for the other quantities can be derived similarly.

We provide L_2 consistent estimators for Q_1, Q_2, and Q_3 using the $\{X_n;Y_n;Z_n\}_{n=1}^N$ sample. They immediately lead to consistent estimators for $D^{KL}(p_1\|p_2;p_3)$, $I_\alpha^R(X,Y|Z)$ and $D_\alpha^R(p_1\|p_2;p_3)$.

The estimation of Q_1. Let $\mathcal{J}_1$, $\mathcal{J}_2$ be two disjunct index sets such that $\mathcal{J}_1\cup\mathcal{J}_2=\{1,2,\dots,N\}$, and $\lim_{N\to\infty}|\mathcal{J}_i|=\infty$, $i=1,2$. Let $\rho_{yz,\mathcal{J}}(v)$ be the Euclidean distance of $v\in\mathbb{R}^{d_{yz}}$ to its kth nearest neighbor in the $\{Y_j;Z_j\}_{j\in\mathcal{J}}$ sample set. Similarly, let $\rho_{xz,\mathcal{J}}(v)$ be the Euclidean distance of $v\in\mathbb{R}^{d_{xz}}$ to its kth nearest neighbor in the $\{X_j;Z_j\}_{j\in\mathcal{J}}$ sample set. The proposed estimator of Q_1 is given as follows:

$$\widehat{Q}_1 = \frac{1}{N}\sum_{n=1}^N \frac{|\mathcal{J}_1\setminus n|^{(1-\alpha)}}{|\mathcal{J}_2\setminus n|^{(1-\alpha)}}\frac{\rho_{xz,\mathcal{J}_1\setminus n}^{d_{xz}(1-\alpha)}(X_n;Z_n)}{\rho_{yz,\mathcal{J}_2\setminus n}^{d_{yz}(1-\alpha)}(X_n;Z_n)}B,$$

where $B=\frac{\Gamma^2(k)}{\Gamma(k-\alpha+1)\Gamma(k+\alpha-1)}$.

The estimation of Q_2. Let $\mathcal{J}=\{1,2,\dots,N\}$ be an index set, and let $\rho_{yz,\mathcal{J}}(v)$ and $\rho_{xz,\mathcal{J}}(v)$ be defined as

above. The estimator of Q_2 is given as:

$$\widehat{Q}_2 = \frac{1}{N}\sum_{n=1}^{N} d_{yz} \log \frac{\rho_{yz,\mathcal{J}\setminus n}(X_n;Z_n)}{\rho_{xz,\mathcal{J}\setminus n}(X_n;Z_n)}.$$

The estimation of Q_3. Let $\mathcal{J}_i$, $(i=1,\dots,4)$ be disjunct index sets such that $\bigcup_{i=1}^4 \mathcal{J}_i = \{1,2,\dots,N\}$, and $\lim_{N\to\infty}|\mathcal{J}_i| = \infty$, $(i=1,\dots,4)$. Let $\rho_{yz,\mathcal{J}_2}(v)$ and $\rho_{xz,\mathcal{J}_3}(v)$ be defined as above, and let $\rho_{xyz,\mathcal{J}_1}(v)$ denote the Euclidean distance of $v\in\mathbb{R}^{d_{xyz}}$ to its kth nearest neighbor in the $\{X_j;Y_j;Z_j\}_{j\in\mathcal{J}_1}$ sample set. Similarly, let $\rho_{z,\mathcal{J}_4}(v)$ be the Euclidean distance of $v\in\mathbb{R}^{d_z}$ to its kth nearest neighbor in the $\{Z_j\}_{j\in\mathcal{J}_4}$ sample set. Let $c_{xyz}, c_{xy}, c_{yz}, c_z$ denote the volume of a d_{xyz}, d_{xy}, d_{yz}, d_z dimensional unit balls, respectively (*i.e.*, $c_d = \pi^{\frac{d}{2}}/\Gamma(\frac{d}{2}+1)$). Q_3 can be estimated by the following expression:

$$\widehat{Q}_3 = \frac{1}{N}\sum_{n=1}^{N} \frac{(c_{xyz}|\mathcal{J}_1\setminus n|)^{(1-\alpha)}\,\rho_{xyz,\mathcal{J}_1\setminus n}^{d_{xyz}(1-\alpha)}(X_n;Y_n;Z_n)}{(c_{xz}|\mathcal{J}_2\setminus n|)^{(1-\alpha)}\,\rho_{xz,\mathcal{J}_2\setminus n}^{d_{xz}(1-\alpha)}(X_n;Z_n)} \times \frac{(c_z|\mathcal{J}_4\setminus n|)^{(1-\alpha)}\,\rho_{z,\mathcal{J}_4\setminus n}^{d_z(1-\alpha)}(Z_n)}{(c_{yz}|\mathcal{J}_3\setminus n|)^{(1-\alpha)}\,\rho_{yz,\mathcal{J}_3\setminus n}^{d_{yz}(1-\alpha)}(Y_n;Z_n)} B^2. \quad (1)$$

5 Consistency Results

Due to the lack of space, we prove theoretical results only for the most complex estimator, $\widehat{Q}_3$. Using the same technique, similar consistency theorems can also be proven for $\widehat{Q}_1$ and $\widehat{Q}_2$. For the details, see the supplementary material. In Section 6 and in the supplementary material we will also illustrate the consistency of $\widehat{Q}_1$, $\widehat{Q}_2$, and $\widehat{Q}_3$ via numerical experiments.

Let $p_{X,Y,Z}$ be bounded away from zero, bounded above, and uniformly continuous density function on $\mathcal{M} = \mathrm{supp}(p_{X,Y,Z})$ domain. Let $\mathcal{M}$ be a finite union of bounded convex sets. We have the following main theorems.

Theorem 8 (Asymptotic unbiasedness of $\widehat{Q}_3$)**.** *Let $k > \max(1-\alpha,\, \alpha-1)$. Then $\lim_{N\to\infty}\mathbb{E}\left[\widehat{Q}_3\right] = Q_3$, i.e., the estimator is asymptotically unbiased.*

Theorem 9 (L_2 consistency of estimator $\widehat{Q}_3$)**.** *If $k > 2\max(1-\alpha,\, \alpha-1)$, then $\lim_{N\to\infty}\mathbb{E}\left[(\widehat{Q}_3 - Q_3)^2\right] = 0$, i.e., the estimator is L_2 consistent.*

5.1 Proof of Consistency

We will exploit some properties of k-NN based density estimators. k-NN based density estimators use only the distances between the observations and their kth nearest neighbors. Let $X_{1:n} \doteq (X_1,\dots,X_n)$ be an i.i.d. sample from a distribution with density p. Let $\mathcal{B}(x,R)$ denote a closed ball around $x\in\mathbb{R}^d$ with radius R, and let $\mathcal{V}(B(x,R)) = c_d R^d$ be its volume, where c_d stands for the volume of a d-dimensional unit ball. Let $\rho(x)$ denote the Euclidean distance of the kth nearest neighbor of x in the sample $X_{1:n}$. Now, according to Loftsgaarden and Quesenberry [1965], the k-NN based density estimator of p at x is given as follows: $\hat{p}_k(x) = k/(nc_d\rho^d(x))$. They also showed that if $k(n)$ denotes the number of neighbors applied at sample size n, $\lim_{n\to\infty} k(n) = \infty$, and $\lim_{n\to\infty} n/k(n) = \infty$, then $\hat{p}_{k(n)}(x) \to^p p(x)$ for almost all x. Moreover, If $\lim_{n\to\infty} k(n)/\log(n) = \infty$, and $\lim_{n\to\infty} n/k(n) = \infty$, then $\lim_{n\to\infty}\sup_x|\hat{p}_{k(n)}(x) - p(x)| = 0$ almost surely. Note that these estimators are consistent only when $k(n)\to\infty$. In our proposed divergence estimators we will use these density estimators. However, we will keep k *fixed*, and will still be able to prove their consistency.

The k-NN estimation of $1/p(x)$ is simply $nc_d\rho^d(x)/k$. One can prove that the distribution of $nc_d\rho^d(x)$ converges to an Erlang distribution with mean $k/p(x)$, and variance $k/p^2(x)$. In turn, if we divide the $nc_d\rho^d(x)$ term by k, then asymptotically it has mean $1/p(x)$ and variance $1/(kp^2(x))$. It implies that indeed 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 $c_d\rho^d(x)$, which is a useless estimator since it is asymptotically zero when $x\in\mathrm{supp}(p)$.

Luckily, we do not need to apply consistent density estimators. Eq. (1) has a special form:

$$\widehat{Q}_3 = \frac{1}{N}\sum_{n=1}^{N} h_1^{\gamma}(X_n,Y_n,Z_n)h_2^{-\gamma}(X_n,Z_n) \times h_3^{-\gamma}(Y_n,Z_n)h_4^{\gamma}(Z_n), \quad (2)$$

where $\gamma \doteq (1-\alpha)$, and $h_1(x_0,y_0,z_0) = c_{xyz}|\mathcal{J}_1\setminus n|\rho_{xyz,\mathcal{J}_1}^{d_{xyz}}(x_0,y_0,z_0)$, $h_2(x_0,z_0) = c_{xz}|\mathcal{J}_2\setminus n|\rho_{xz,\mathcal{J}_2}^{d_{xz}}(x_0,z_0)$, $h_3(y_0,z_0) = c_{yz}|\mathcal{J}_3\setminus n|\rho_{yz,\mathcal{J}_3}^{d_{yz}}(y_0,z_0)$, $h_4(z_0) = c_z|\mathcal{J}_4\setminus n|\rho_{z,\mathcal{J}_4}^{d_z}(z_0)$. For the sake of brevity, let $v_1 = (x_0,y_0,z_0)$, $v_2 = (x_0,z_0)$, $v_3 = (y_0,z_0)$, $v_4 = z_0$. Using the Lebesgue lemma and the properties of Lebesgue points [Leonenko et al., 2008], we can prove that the distribution function of $h_i(v_i)$ converges weakly to the distribution function of an Erlang distribution with mean $k/p(v_i)$ and variance $k/p^2(v_i)$. Furthermore, the random variables $\{h_i(n,\gamma)\}_{i=1}^4$ are conditionally independent for a given $(X_n;Y_n;Z_n)$ (this is the reason why we split the index sets $\mathcal{J}$ into four disjunct sets). In turn, "in the limit" (2) is simply the empirical average of the product of the γth (and $-\gamma$th) powers of independent Erlang distributed variables. These moments can be calculated analytically.

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 (2) 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 we can guarantee only the weak convergence of $h_i(v_i)$ 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 [van der Wart, 2007], and we need the uniform generalization of the Lebesgue lemma. As a result, we will need to put some extra conditions on the density $p_{X,Y,Z}$. The technical details can be found in the supplementary material.

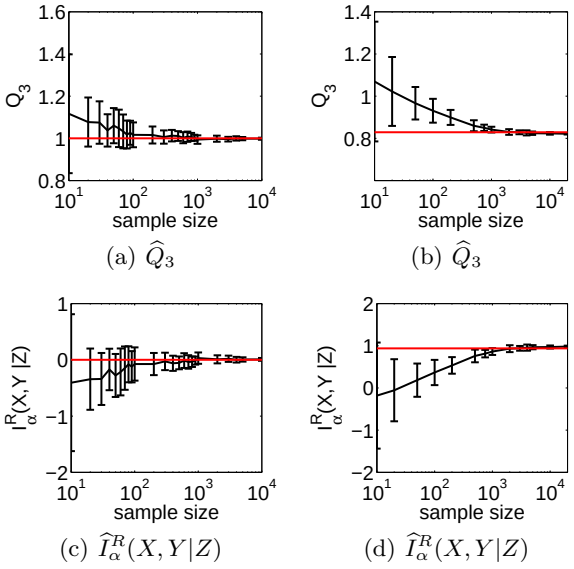
6 Numerical Experiments

In this section we demonstrate the consistency and the applicability of the proposed estimators by numerical experiments.

In the first experiment we generated samples with $I_\alpha(X,Y) > 0$ and $I_\alpha(X,Y|Z) = 0$ properties. Here, we model the situation when two random variables are dependent, but there is a third random variable that is responsible for this dependence. For this purpose, we considered the $X_n = A_n + Z_n$, $Y_n = B_n + Z_n$ random variables, where A_n, B_n, and Z_n were independent random variables with 1-dimensional normal distributions having zero means and randomly chosen covariances. Fig. 1(a) and Fig. 1(c) demonstrate the convergence of $\widehat{Q}_3$ and $\widehat{I}^R_\alpha \doteq \log(\widehat{Q}_3)/(\alpha-1)$ as a function of the sample size. We chose k=1, and $\alpha = 0.7$ in these experiments. The error bars show the standard deviation calculated from 25 independent runs. The red lines correspond to the true Q_3 and I^R_α values.

To show that the estimators can estimate mutual information in the general case too (i.e., $I(X,Y) > 0$, $I(X,Y,Z) > 0$, and $I(X,Y|Z) > 0$), we repeated the previous experiment, but this time $(X;Y;Z)$ was generated from a general 3-dimensional normal distribution with zero means and randomly selected covariance matrix ($\Sigma = CC^T$, where $C_{i,j} \sim \mathcal{N}(0,1)$). Fig. 1(b) and Fig. 1(d) demonstrate the convergence of $\widehat{Q}_3$ and $\widehat{I}^R_\alpha$ as a function of the sample size. In the supplementary material we demonstrate that the $\widehat{Q}_1$ and $\widehat{Q}_2$ estimators are consistent as well.

In the following experiment we show on image data that the mutual information can either be larger or smaller than the conditional mutual information. In other words, extra knowledge can either increase or decrease the mutual information. We chose a gray-scale image (Fig. 2(c)) of size 75×100 and considered its pixel values ($Z \in [0,255]$) as if they were samples from a distribution. We also constructed two noisy versions of Z: we set $X = Z + A$ (Fig. 2(a)) and $Y = Z + B$ (Fig. 2(b)), where A and B were independent random noise variables with uniform $U[-5,5]$ distributions. By construction, $I^R_\alpha(X,Y) > 0$, and $I^R_\alpha(X,Y|Z) = 0$, that is, the observation of Z eliminates the mutual information between X and Y. This is also confirmed by the estimated $I^R_\alpha(X,Y)$ and $I^R_\alpha(X,Y|Z)$ values (Fig. 2(d)). Here we used $\alpha = 0.75$, and tried $k = 2,5,10,30$ nearest neighbors.



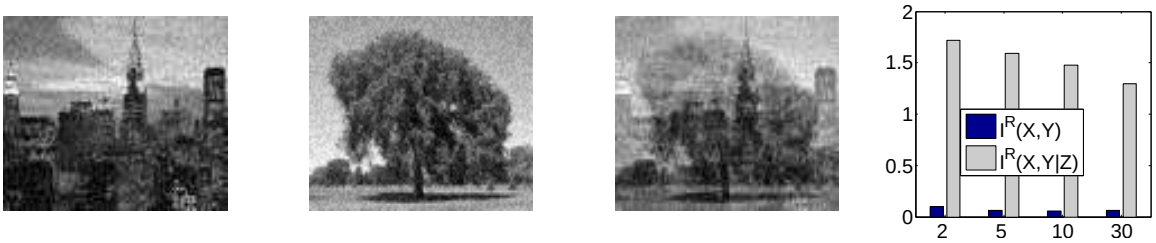
(a) $\widehat{Q}_3$ (b) $\widehat{Q}_3$ (c) $\widehat{I}^R_\alpha(X,Y|Z)$ (d) $\widehat{I}^R_\alpha(X,Y|Z)$

Figure 1: Estimated vs. true conditional mutual information as a function of sample size. (a), (c): $I_\alpha(X,Y|Z) = 0$, (b), (d): $I_\alpha(X,Y|Z) > 0$. The error bars show the standard deviation calculated from 25 independent runs. The red lines correspond to the true Q_3 and I^R_α values. We used $k = 1$ nearest neighbor.

The following experiment demonstrates that the opposite situation can also occur. Similarly to the previous case, we chose two noisy images (Fig. 2(e) and Fig. 2(f)). We considered their pixel values as if they were i.i.d. samples from two random variables X and Y, and then constructed their noisy sum: $Z = X + Y + A$, where A played the role of noise and it had uniform $U[-5,5]$ distribution. Fig. 2(h) shows that $I^R_\alpha(X,Y) \approx 0$ (i.e. the two original images were almost independent), but $I^R_\alpha(X,Y|Z) > 0$.

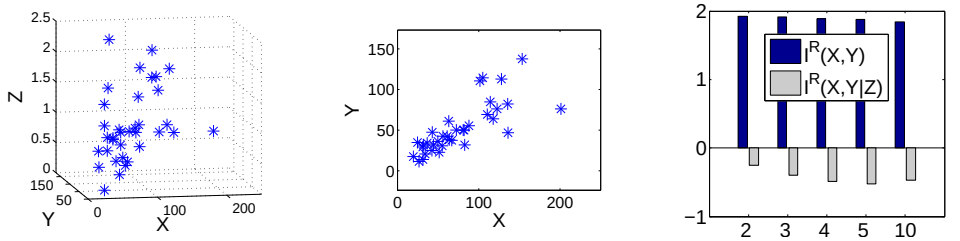
The next experiment demonstrates that the proposed estimator might be useful to detect confounder variables in medical data too. Note, however, that conver-

(a) X (b) Y (c) Z (d) I^R estimates



(e) X (f) Y (g) Z (h) I^R estimates

Figure 2: Demonstration that conditioning to a third variable (Z) can either increase or decrease the mutual information between X and Y. (a), and (b): Noisy versions of the picture in (c). (g): Noisy sum of pictures in (e) and (f). (d) and (h): Estimated $I_\alpha^R(X,Y)$ and $I_\alpha^R(X,Y|Z)$ values for $k = 2, 5, 10, 30$.



(a) Medical data (X, Y, Z) (b) Medical data (X, Y) (c) Estimated information

Figure 3: (a): The medical data set. (b): There is dependence between X and Y. (c): Estimated $I_\alpha^R(X,Y)$ and $I_\alpha^R(X,Y|Z)$ values for $k = 2, 3, 4, 5, 10$. Negative and zero values indicate independence.

gence rates of the estimators are not known yet. We used the medical data published in Edwards [2000] (Section 3.1.4.). The data were taken from 35 patients and consist of three variables: digoxin clearance (X), urine flow (Y), and creatinine clearance (Z) (Fig. 3(a)). From medical knowledge we know that X should be independent of Y given Z. It was presented in Fukumizu et al. [2008] that there is a strong linear correlation between X and Y (Fig. 3(b)), and a partial correlation based test was not able to show the conditional independence of X and Y given Z. Below we demonstrate that our method is able to detect the conditional independence of the variables. Since the dataset consists of only 35 points, we applied the following bootstrap method: We repeated each (X, Y, Z) point of the dataset 300 times. Then we added a small, uniformly distributed $U[-\epsilon, \epsilon]$ perturbation to each of these 7 000 data points, where ϵ was set to 5 percent of the mean values of the variables X, Y, and Z. Fig. 3(c) shows that the Rényi information estimator was able to detect the large dependence between variables X and Y, and the conditional information estimator shows that this dependence vanishes when we observe variable Z. We set α to 0.5, and we experimented with $k = 2, 3, 4, 5, 10$. For all of these parameters the estimated $I_\alpha^R(X,Y)$ values were larger than zero, while the estimated $I_\alpha^R(X,Y|Z)$ values were all negative indicating conditional independence.

One might wonder whether the proposed estimators are better than the naïve plug-in based estimators. To answer this question, we implemented a plug-in type conditional information estimator. It uses the kernel density estimator of Gray and Moore [2003] implemented by Ihler [2003]. For the kernel bandwidth selection, we used the Scott's factor Scott [1992]. In this example we set $\alpha = 0.8$, and $(X; Y; Z)$ was gener-

ated from a general 3-dimensional normal distribution with zero means and randomly selected covariance matrix. Fig 6 shows the estimation errors of these KDE methods using Gauss and Laplace kernels, and we also present the estimation errors of our "direct" method. In this experiment our method achieved smaller errors than its KDE based competitors. KDE methods tend to be sensitive to their parameter and bandwidth settings. Our method has only one parameter k; we did not tune it, we simply set it to $k = 2$.

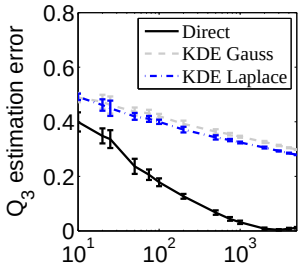
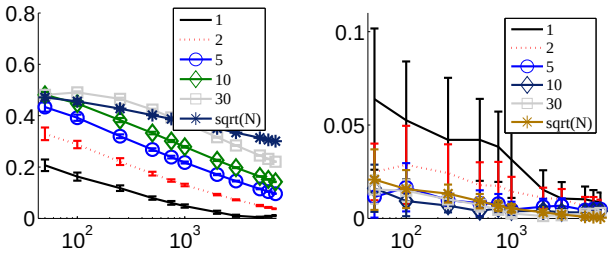


Figure 4: Comparison between our method (Direct) and KDE based plug-in estimators using Gauss and Laplace kernels. We show the estimation errors as a function of sample size. Error bars are calculated from 25 independent runs.

In Theorems 8-9 we have claimed that our estimator is consistent for any fixed and large enough k. The convergence rate and finite sample performance, however, depends on k. We conjecture that the best k value depends on the actual distributions. Specifically, for normal distributions with large $I(X, Y|Z)$ conditional dependence it seems that setting $k = 1$ gives the highest convergence rate (Fig. 5(a)). However, when the dependence is small, then larger k values lead to better performances (Fig. 6). The figures show the estimation errors as a function of sample size for several k values. We also compare the estimators with $k = \sqrt{N}$, which corresponds to a plug-in type estimator with consistent k-nearest neighbor based density estimator. In Fig. 5(a) this estimator has the largest error for large sample size and the second largest for small sample size.

7 Discussion and Conclusion

We proposed new nonparametric estimators for a family of conditional divergences and mutual information. We theoretically proved the consistency of these estimators, and demonstrated their efficiency by numerical experiments on images, synthetic, and medical data. To the best of our knowledge, these are the first consistent conditional divergence and mutual information estimators that can avoid the need for density estimation. We have also shown empirically that the estimators can perform better than the naïve plug-in density estimator based variants.



(a) Dependence (b) Independence

Figure 5: We show the estimation errors as a function of sample size for different $k \in \{1, 2, 5, 10, 30, \sqrt{N}\}$ values. Error bars are calculated from 25 independent runs. (a): Normal distribution with large $I_\alpha^R(X, Y|Z) \approx 2.8$ value (b): Normal distribution with $I_\alpha^R(X, Y|Z) = 0$.

There are several open questions left waiting for answers. Currently we do not know the convergence rates of the estimators, and how they depend on the parameters k, α, the densities, and the dimensions. One challenging problem here is that there are no known convergence rates so far for the much easier unconditional entropy and divergence estimation special cases either [Leonenko et al., 2008, Wang et al., 2009b]. In turn, in order to derive a tight convergence rate for our case, one should solve those open problems first. Our numerical results indicate that the parameter k which gives the fastest convergence rate depends on the distributions. Note however, that the estimator is convergent for every (large enough) fixed k, and k does not need to converge to infinity, which is a requirement for k-NN based density estimators. In practice we got good results even when k was set to small numbers, e.g. $k = 2$. We also found that our estimator performed better than the naive, "plug-in" algorithms, which estimate the densities first either with KDE or k-NN based density estimators.

The conditions of our consistency theorems could also be extended. We also note that although our proof techniques require the $\mathcal{J}_i$ index sets to be disjunct, in practice we found that even when the index sets are totally overlapping ($\mathcal{J}_i = \mathcal{J}_j$), the estimators are still consistent, suggesting that asymptotically the correlations between the limiting Erlang distributions disappear. This observation leads to a 4-fold improvement in sample efficiency. In the future we are going to investigate these questions, and we also plan to develop new conditional independence tests based on the proposed estimators.

References

T. Bouezmarni, J. Rombouts, and A. Taamouti. A nonparametric copula based test for conditional independence with applications to Granger causality. Technical report, Universidad Carlos III, Departamento de Economia, 2009.

A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. John Wiley and Sons, 2009.

D. Edwards. *Introduction to graphical modelling*. Springer verlag, New York, 2000.

K. Fukumizu, A. Gretton, X. Sun, and B. Schoelkopf. Kernel measures of conditional dependence. In *NIPS 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.

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. Gray and A. Moore. Very fast multivariate kernel density estimation using via computational geometry. In *Joint Stat. Meeting*, 2003.

A. Gretton, R. Herbrich, and A. Smola. The kernel mutual information. In *Proc. ICASSP*, 2003.

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

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

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

A. Ihler. Kernel density estimation (kde) toolbox for matlab. `http://www.ics.uci.edu/ ihler/code/kde.html`, 2003.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

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

D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist*, 36:1049–1051, 1965.

R. Matthews. Storks deliver babies (p=0.008). *Teaching Statistics*, 22:36–38, 2000.

D. Montgomery. *Design and Analysis of Experiments*. John Wiley and Sons, 2005.

X. Nguyen, M.J. Wainwright, and M.I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37:876–904, 2009.

X. Nguyen, M.J. Wainwright, and M.I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory, To appear.*, 2010.

D. Pál, B. Póczos, and Cs. Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *NIPS 24*, 2010.

J. Pearl. Why there is no statistical test for confounding, why many think there is, and why they are almost right. UCLA Computer Science Department, Technical Report R-256, 1998.

F. Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *NIPS 21*, 2008.

B. Póczos and J. Schneider. On the estimation of alpha-divergences. In *AISTATS 2011*, 2011.

B. Póczos and J. Schneider. Nonparametric estimation of conditional information and divergences, 2012. Technical report, `http://www.autonlab.org/autonweb/library/papers.html`.

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

B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9, 1981.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2001.

K. Sricharan, R. Raich, and A. Hero. Empirical estimation of entropy functionals with confidence. Technical Report, `http://arxiv.org/abs/1012.4188`, 2010.

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

L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, 24:829–864, 2008.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794, 2007.

A. W. van der Wart. *Asymptotic Statistics*. Cambridge University Press, 2007.

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), 2009b.

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

K. Zhang, J. Peters, D. Janzing, and B. Scholkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI 2011*, 2011.