Efficient Estimation of Mutual Information for Strongly Dependent Variables

Shuyang Gao
Information Sciences Institute
University of Southern California
sgao@isi.edu

Greg Ver Steeg
Information Sciences Institute
University of Southern California
gregv@isi.edu

Aram Galstyan
Information Sciences Institute
University of Southern California
galstyan@isi.edu

Abstract

We demonstrate that a popular class of non-parametric mutual information (MI) estimators based on k-nearest-neighbor graphs requires number of samples that scales exponentially with the true MI. Consequently, accurate estimation of MI between two strongly dependent variables is possible only for prohibitively large sample size. This important yet overlooked shortcoming of the existing estimators is due to their implicit reliance on local uniformity of the underlying joint distribution. We introduce a new estimator that is robust to local non-uniformity, works well with limited data, and is able to capture relationship strengths over many orders of magnitude. We demonstrate the superior performance of the proposed estimator on both synthetic and real-world data.

1 Introduction

As a measure of the dependence between two random variables, mutual information is remarkably general and has several intuitive interpretations (Cover and Thomas, 2006), which explains its widespread use in statistics, machine learning, and computational neuroscience; see (Wang et al., 2009) for a list of various applications. In a typical scenario, we do not know the underlying joint distribution of the random variables, and instead need to estimate mutual information using i.i.d. samples from that distribution. A naive approach to this problem is to first estimate the underlying probability density, and then calculate mutual information using its definition. Unfortunately, estimating joint densities from a limited number of samples is often infeasible in many practical settings.

A different approach is to estimate mutual information directly from samples in a non-parametric way, without ever describing the entire probability density. The main intuition behind such *direct* estimators is that evaluating mutual information can in principle be a more tractable problem than estimating the density over the whole state space (Pérez-Cruz, 2008). The most popular class of estimators taking this approach is the k-nearest-neighbor(kNN) based estimators. One example of such estimator is due to Kraskov, Stögbauer, and Grassberger (referred to as the KSG estimator from here on) (Kraskov et al., 2004), which has been extended to generalized nearest neighbors graphs (Pál et al., 2010).

Our first contribution is to demonstrate that kNN-based estimators suffer from a critical yet overlooked flaw. Namely, we illustrate that if the true mutual information is I, then kNN-based estimators requires exponentially many (in I) samples for accurate estimation. This means that *strong relationships* are actually *more difficult* to measure. This counterintuitive property reflects the fact that most work on mutual information estimation has focused on estimators that are good at detecting *independence* of variables rather than precisely measuring strong dependence. In the age of big data, it is often the case that many strong relationships are present in the data and we are interested in picking out only the strongest ones. MI estimation will perform poorly for this task if their accuracy is low in the regime of strong dependence.

We show that the undesired behavior of previous kNN-based MI estimators can be attributed to the assumption of local uniformity utilized by those estimators, which can be violated for sufficiently strong (almost deterministic) dependencies. As our second major contribution, we suggest a new kNN estimator that relaxes the local uniformity condition by introducing a correc-

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

tion term for local non-uniformity. We demonstrate empirically that for strong relationships, the proposed estimator needs significantly fewer samples for accurately estimating mutual information.

In Sec. 2, we introduce kNN-based non-parametric entropy and mutual information estimators. In Sec. 3, we demonstrate the limitations of kNN-based mutual information estimators. And then we suggest a correction term to overcome these limitations in Sec. 4. In Sec. 5, we show empirically using synthetic and real-world data that our method outperforms existing techniques. In particular, we are able to accurately measure strong relationships using smaller sample sizes. Finally, we conclude with related work and discussion.

2 kNN-based Estimation of Entropic Measures

In this section, we will first introduce the naive kNN estimator for mutual information, which is based on an entropy estimator due to (Singh et al., 2003), and show its theoretical properties. Next, we focus on a popular variant of kNN estimators, called KSG estimator (Kraskov et al., 2004).

2.1 Basic Definitions

Let $\mathbf{x} = (x_1, x_2, ..., x_d)$ denote a d-dimensional absolute continuous random variable whose probability density function is defined as $p : \mathbb{R}^d \to \mathbb{R}$ and marginal densities of each x_j are defined as $p_j : \mathbb{R} \to \mathbb{R}, j = 1, \ldots, d$. Shannon Differential Entropy and Mutual Information are defined in the usual way:

$$H(\mathbf{x}) = -\int_{\mathbb{R}^d} p(\mathbf{x}) \log p(\mathbf{x})\, d\mathbf{x} \tag{1}$$

$$I(\mathbf{x}) = \int_{\mathbb{R}^d} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\prod_{j=1}^{d} p_j(x_j)} d\mathbf{x} \tag{2}$$

We use natural logarithms so that information is measured in nats. For $d > 2$, the generalized mutual information is also called *total correlation* (Watanabe, 1960) or *multi-information* (Studený and Vejnarová, 1998).

Given N i.i.d. samples, $\mathcal{X} = \left\{\mathbf{x}^{(i)}\right\}_{i=1}^{n}$, drawn from $p(\mathbf{x})$, the task is to construct a mutual information estimator $\hat{I}(\mathbf{x})$ based on these samples.

2.2 Naive kNN Estimator

Entropy Estimation The naive kNN entropy estimator is as follows:

$$\widehat{H}'_{kNN,k}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n} \log \widehat{p}_k\left(\mathbf{x}^{(i)}\right) \tag{3}$$

where

$$\widehat{p}_k\left(\mathbf{x}^{(i)}\right) = \frac{k}{n-1}\frac{\Gamma(d/2+1)}{\pi^{d/2}} r_k\left(\mathbf{x}^{(i)}\right)^{-d} \tag{4}$$

$r_k\left(\mathbf{x}^{(i)}\right)$ in Eq. 4 is the Euclidean distance from $\mathbf{x}^{(i)}$ to its kth nearest neighbor in $\mathcal{X}$. By introducing a correction term, an asymptotic unbiased estimator is obtained:

$$\widehat{H}_{kNN,k}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n} \log \widehat{p}_k\left(\mathbf{x}^{(i)}\right) - \gamma_k \tag{5}$$

where

$$\gamma_k = \frac{k^k}{(k-1)!}\int_0^\infty \log(x)\, x^{k-1} e^{-kx} dx = \psi(k) - \log(k) \tag{6}$$

where $\psi(\cdot)$ represents the digamma function.

The following theorem shows asymptotic unbiasedness of $\widehat{H}_{kNN,k}(\mathbf{x})$ according to (Singh et al., 2003).

Theorem 1 (kNN entropy estimator, asymptotic unbiasedness, (Singh et al., 2003))**.** *Assume that* $\mathbf{x}$ *is absolutely continuous and* k *is a positive integer, then*

$$\lim_{n\to\infty} \mathbb{E}\left[\widehat{H}_{kNN,k}(\mathbf{x})\right] = H_{kNN,k}(\mathbf{x}) \tag{7}$$

i.e., this entropy estimator is asymptotically unbiased.

From Entropy to Mutual Information To construct a mutual information estimator from an entropy estimator is straightforward by combining entropy estimators using the identity (Cover and Thomas, 2006):

$$I(\mathbf{x}) = \sum_{i=1}^{d} H(x_i) - H(\mathbf{x}) \tag{8}$$

Combining Eqs. 3 and 8, we have,

$$\widehat{I}'_{kNN,k}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{\widehat{p}_k\left(\mathbf{x}^{(i)}\right)}{\widehat{p}_k\left(\mathbf{x}_1^{(i)}\right)\widehat{p}_k\left(\mathbf{x}_2^{(i)}\right)...\widehat{p}_k\left(\mathbf{x}_d^{(i)}\right)} \tag{9}$$

where $\mathbf{x}_j^{(i)}$ denotes the point projected into jth dimension in $\mathbf{x}^{(i)}$ and $\widehat{p}_k(\mathbf{x}_j^{(i)})$ represents the marginal kNN density estimator projected into jth dimension of $\mathbf{x}^{(i)}$.

Similar to Eq. 5, we can also construct an asymptotically unbiased mutual information estimator based on Theorem 1:

$$\widehat{I}_{kNN,k} = \widehat{I}'_{kNN,k} - (d-1)\gamma_k \tag{10}$$

Corollary 1 (kNN MI estimator, asymptotic unbiasedness)**.** *Assume that* $\mathbf{x}$ *is absolute continuous and* k *is a positive integer, then:*

$$\lim_{n\to\infty} \mathbb{E}\left[\widehat{I}_{kNN,k}\left(\mathbf{x}\right)\right] = I_{kNN,k}\left(\mathbf{x}\right) \tag{11}$$

2.3 KSG Estimator

The KSG mutual information estimator (Kraskov et al., 2004) is a popular variant of the naive kNN estimator. The general principle of KSG is that for each density estimator in different spaces, we would like to use the similar length-scales for k-nearest-neighbor distance as in the joint space so that the bias would be approximately smaller. Although the theoretical properties of this estimator is unknown, it has a relatively good performance in practice, see (Khan et al., 2007) for a comparison of different estimation methods. Unlike naive kNN estimator, the KSG estimator uses the max-norm distance instead of L2-norm. In particular, if $\epsilon_{i,k}$ is twice the (max-norm) distance to the k-th nearest neighbor of $\mathbf{x}^{(i)}$, it can be shown that the expectation value (over all ways of drawing the surrounding $N-1$ points) of the log probability mass within the box centered at $\mathbf{x}^{(i)}$ is given by this expression.

$$\mathbb{E}_{\epsilon_{i,k}}\left[\log \int_{|\mathbf{x}-\mathbf{x}^{(i)}|_\infty \le \epsilon_{i,k}/2} f(\mathbf{x})\, d\mathbf{x}\right] = \psi(k) - \psi(N) \tag{12}$$

We use ψ to represent the digamma function. If we assume that the density inside the box (with sides of length $\epsilon_{i,k}$) is constant, then the integral becomes trivial and we find that,

$$\log(f(\mathbf{x}_i)\epsilon_{i,k}^d) = \psi(k) - \psi(N). \tag{13}$$

Rearranging and taking the mean over $-\log f(\mathbf{x}^{(i)})$ leads us to the following entropy estimator:

$$\hat{H}_{KSG,k}(\mathbf{x}) \equiv \psi(N) - \psi(k) + \frac{d}{N}\sum_{i=1}^{N} \log \epsilon_{i,k} \tag{14}$$

Note that k, defining the size of neighborhood to use in local density estimation, is a free parameter. Using smaller k should be more accurate, but larger k reduces the variance of the estimate (Khan et al., 2007). While consistent density estimation requires k to grow with N (Von Luxburg and Alamgir, 2013), entropy estimates converge almost surely for any fixed k (Wang et al., 2009; Pérez-Cruz, 2008) under some weak conditions.[1]

[1]This assumes the probability density is absolutely continuous but see (Pál et al., 2010) for some technical concerns.

To estimate the mutual information, in the joint $\mathbf{x}$ space we set k, the size of the neighborhood, which determines $\epsilon_{i,k}$ for each point $\mathbf{x}^{(i)}$. Next, we consider the smallest rectilinear hyper-rectangle that contains these k points, which has sides of length $\epsilon_{i,k}^{x_j}$ for each marginal direction x_j. We refer to this as the "max-norm rectangle" (as shown in Fig. 1(a)). Let n_{x_j} be the number of points at a distance less than or equal to $\epsilon_{i,k}^{x_j}/2$ in the x_j-subspace. For each marginal entropy estimate, we use $n_{x_j}(i)$ instead of k to set the neighborhood size at each point. Finally, in the joint space using a rectangle instead of a box in Eq. 14 leads to a correction term of size $(d-1)/k$ (details given in (Kraskov et al., 2004)). Adding the entropy estimators together with these choices yields the following.

$$\hat{I}_{KSG,k}(\mathbf{x}) \equiv (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi(n_{x_j}(i)) \tag{15}$$

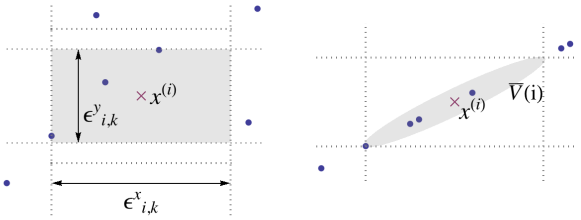

Figure 1: Centered at a given sample point, $\mathbf{x}^{(i)}$, we show the max-norm rectangle containing k nearest neighbors (a) for points drawn from a uniform distribution, $k=3$, and (b) for points drawn from a strongly correlated distribution, $k=4$.

3 Limitations of kNN-based MI Estimators

In this section, we demonstrate a significant flaw in kNN-based MI estimators which is summarized in the following theorems. We then explain why these estimators fail to accurately estimate mutual information unless the correlations are relatively weak or the number of samples is large.

Theorem 2. *For any d-dimensional absolute continuous probability density function,* $p(\mathbf{x})$*, for any* $k \ge 1$*, for the estimated mutual information to be close to the true mutual information,* $|\hat{I}_{kNN,k}(\mathbf{x}) - I(\mathbf{x})| \le \varepsilon$*, requires that the number of samples,* N*, is at least,* $N \ge C\exp\left(\frac{I(\mathbf{x})-\varepsilon}{d-1}\right)+1$*, where* C *is a constant which scales like* $O(\frac{1}{d})$*.*

The proof of Theorem 2 is shown in the Appendix A.

Theorem 3. *For any d-dimensional absolute continuous probability density function, $p(\mathbf{x})$, for any $k \geq 1$, for the estimated mutual information to be close to the true mutual information, $|\hat{I}_{KSG,k}(\mathbf{x}) - I(\mathbf{x})| \leq \varepsilon$, requires that the number of samples, N, is at least, $N \geq C \exp\left(\frac{I(\mathbf{x})-\varepsilon}{d-1}\right) + 1$, where $C = \mathbf{e}^{-\frac{k-1}{k}}$.*

Proof. Note that $\psi(n) = H_{n-1} - \gamma$, where H_n is the n-th harmonic number and $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

$$\begin{aligned} \widehat{I}_{KSG,k}(\mathbf{x}) &\leq (d-1)\psi(N) + \psi(k) - (d-1)/k \\ &\quad -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(k\right) \\ &= (d-1)(\psi(N) - \psi(k) - 1/k) \\ &= (d-1)(H_{N-1} - \gamma - H_{k-1} + \gamma - 1/k) \\ &\leq (d-1)(\log(N-1) + (k-1)/k) \end{aligned} \tag{16}$$

The first inequality is obtained from Eq. 15 by observing that $n_{x_j}(i) \geq k$ for any i, j and that $\psi(k)$ is a monotonically increasing function. And the last inequality is obtained by dropping the term $-H_{k-1} \leq 0$ and using the well-known upper bound $H_N \leq \log N + 1$. Requiring that $|\hat{I}_{KSG,k}(\mathbf{x}) - I(\mathbf{x})| < \varepsilon$, we obtain $N \geq C \exp\left(\frac{I(\mathbf{x})-\varepsilon}{d-1}\right) + 1$, where $C = e^{-\frac{k-1}{k}}$. □

The above theorems state that for any fixed dimensionality, the number of samples needed for estimating mutual information $I(\mathbf{x})$ increases exponentially with the *magnitude* of $I(\mathbf{x})$. From the point of view of determining independence, i.e., distinguishing $I(\mathbf{x}) = 0$ from $I(\mathbf{x}) \neq 0$, this restriction is not particularly troubling. However, for finding strong signals in data it presents a major barrier. Indeed, consider two random variables X and Y, where $X \sim \mathcal{U}(0,1)$ and $Y = X + \eta\mathcal{U}(0,1)$. When $\eta \to 0$, the relationship between X and Y becomes nearly functional, and the mutual information diverges as $I(X:Y) \to \log\frac{1}{\eta}$. As a consequence, the number of samples needed for accurately estimating $I(X:Y)$ diverges as well. This is depicted in Fig. 2 where we compare the empirical lower bound to our theoretical bound given by Theorem 3. It can be seen that the theoretical bounds are rather conservative, but they have the same exponentially growing rates comparing to the empirical ones.

What is the origin of this undesired behavior? An intuitive and general argument comes from looking at the assumption of local uniformity in kNN-based estimators. In particular, both naive kNN and KSG estimators approximate the probability density in the kNN ball or max-norm rectangle containing the k nearest neighbors with uniform density. If there are strong relationships (in the joint $\mathbf{x}$ space, the density becomes more singular), then we can see in Fig. 1(b) that the uniform assumption becomes problematic.

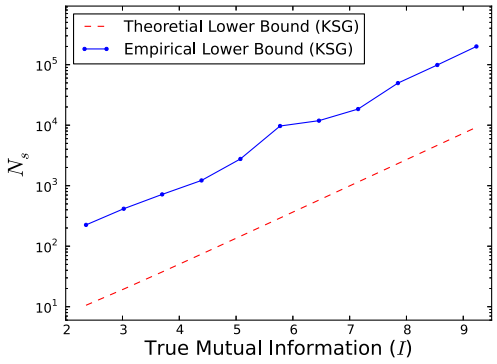

Figure 2: A semi-logarithmic plot of N_s (number of required samples to achieve an error at most ε) for KSG estimator for different values of $I(X:Y)$. We set $\varepsilon = 0.1$, $k = 1$.

4 Improved kNN-based Estimators

In this section, we suggest a class of kNN-based estimators that relaxes the local uniformity assumption.

Local Nonuniformity Correction (LNC) Our second contribution is to provide a general method for overcoming the limitations above. Considering the ball (in naive kNN estimator) or max-norm hyperrectangle (in KSG estimator) around the point $\mathbf{x}^{(i)}$ which contains k nearest neighbors, let us denote this region of the space with $\mathcal{V}(i) \subset \mathbb{R}^d$, whose volume is $V(i)$. Instead of assuming that the density is uniform inside $\mathcal{V}(i)$ around the point $\mathbf{x}^{(i)}$, we assume that there is some subset, $\bar{\mathcal{V}}(i) \subseteq \mathcal{V}(i)$ with volume $\bar{V}(i) \leq V(i)$ on which the density is constant, i.e., $\hat{p}(\mathbf{x}^{(i)}) = \frac{\mathbb{I}[\mathbf{x}\in\bar{\mathcal{V}}(i)]}{\bar{V}(i)}$. This is illustrated with a shaded region in Fig. 1(b). We now repeat the derivation above using this altered assumption about the local density around each point for $\hat{H}(\mathbf{x})$. We make no changes to the entropy estimates in the marginal subspaces. Based on this idea, we get a general correction term for kNN-based MI estimators (see Appendix B for the details of derivation):

$$\hat{I}_{LNC}(\mathbf{x}) = \hat{I}(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N}\log\frac{\bar{V}(i)}{V(i)} \tag{17}$$

where $\hat{I}(\mathbf{x})$ can be either $\hat{I}_{kNN}(\mathbf{x})$ or $\hat{I}_{KSG}(\mathbf{x})$.

If the local density in the k-nearest-neighbor region $\mathcal{V}(i)$ is highly non-uniform, as is the case for strongly

related variables like those in Fig. 1(b), then the proposed correction term will improve the estimate. For instance, if we assume that relationships in data are smooth, functional relationships plus some noise, the correction term will yield significant improvement as demonstrated empirically in Sec. 5. We note that this correction term is not bounded by N, but rather by our method of estimating $\bar{V}$. Next, we will give one concrete implementation of this idea by focusing on the modification of KSG estimator.

Estimating Nonuniformity by Local PCA With the correction term in Eq. 17, we have transformed the problem into that of finding a local volume on which we believe the density is positive. Regarding KSG estimator, instead of a uniform distribution within the max-norm rectangle in the neighborhood around the point $\mathbf{x}^{(i)}$, we look for a small, rotated (hyper)rectangle that covers the neighborhood of $\mathbf{x}^{(i)}$. The volume of the rotated rectangle is obtained by doing a localized principle component analysis around all of $\mathbf{x}^{(i)}$'s k nearest neighbors, and then multiplying the maximal axis values together in each principle component after the k points are transformed to the new coordinate system[2]. The key advantage of our proposed estimator is as follows: while KSG assumes *local uniformity* of the density over a region containing k nearest neighbors of a particular point, our estimator relies on a much weaker assumption of local linearity over the same region. Note that the *local linearity* assumption has also been widely adopted in the manifold learning, for example, *local linear embedding*(LLE) (Roweis and Saul, 2000) and *local tangent space alignment*(LTSA) (Zhang and Zha, 2002).

Testing for Local Nonuniformity One problem with this procedure is that we may find that, locally, points may occupy a small sub-volume, i.e., even if the local neighborhood is actually drawn from a uniform distribution(as shown in Fig. 1(a)), the volume of the PCA-aligned rectangle will with high probability be smaller than the volume of the max-norm rectangle, leading to an artificially large non-uniformity correction.

To avoid this artifact, we consider a trade-off between the two possibilities: for a fixed dimension d and nearest neighbor parameter k, we find a constant $\alpha_{k,d}$, such that if $\bar{V}(i)/V(i) < \alpha_{k,d}$, then we assume local uniformity is violated and use the correction $\bar{V}(i)$, otherwise the correction is discarded for point $\mathbf{x}^{(i)}$. Note that if $\alpha_{k,d}$ is sufficiently small, then the correction term will be always discarded, so that our estimator reduces to the KSG estimator. Good choices of $\alpha_{k,d}$ are set using arguments described in Appendix C. Furthermore, we believe that as long as the expected value of $\mathbb{E}[\bar{V}(i)/V(i)] \geq \alpha_{k,d}$ in large N limit, for some properly selected $\alpha_{k,d}$, then the consistency properties of the proposed estimator will be identical to the constancy properties of the KSG estimator. A rigorous proof of this hypothesis is left as a future work.

The full algorithm for our estimator is given in Algorithm 1.

Algorithm 1 Mutual Information Estimation with Local Nonuniform Correction

Input: points $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}$, parameter d (dimension), k (nearest neighbor), $\alpha_{k,d}$
Output: $\hat{I}_{LNC}(\mathbf{x})$
Calculate $\hat{I}_{KSG}(\mathbf{x})$ by KSG estimator, using the same nearest neighbor parameter k
for each point $\mathbf{x}^{(i)}$ **do**
  Find k nearest neighbors of $\mathbf{x}^{(i)}$: $kNN_1^{(i)}$, $kNN_2^{(i)}$,..., $kNN_k^{(i)}$
  Do PCA on these k neighbors, calculate the volume corrected rectangle $\bar{V}(i)$
  Calculate the volume of max-norm rectangle $V(i)$
  if $\bar{V}(i)/V(i) < \alpha_{k,d}$ **then**
    $LNC_i = \log \frac{\bar{V}(i)}{V(i)}$
  else
    $LNC_i = 0.0$
  end if
end for
Calculate LNC^*: average value of $LNC_1, LNC_2, ..., LNC_N$
$\hat{I}_{LNC} = \hat{I}_{KSG} - LNC^*$

5 Experiments

We evaluate the proposed estimator on both synthetically generated and real-world data. For the former, we considered various functional relationships and thoroughly examined the performance of the estimator over a range of noise intensities. For the latter, we applied our estimator to the WHO dataset used previously in (Reshef et al., 2011). Below we report the results.

5.1 Experiments with synthetic data

Functional relationships in two dimensions In the first set of experiments, we generate samples from various functional relationships of the form $Y = f(X) + \eta$ that were previously studied in Reshef et al. (2011); Kinney and Atwal (2014). The noise term η

[2]Note that we manually set the mean of these k points to be $\mathbf{x}^{(i)}$ when doing PCA, in order to put $\mathbf{x}^{(i)}$ in the center of the rotated rectangle.

is distributed uniformly over the interval $[-\sigma/2, \sigma/2]$, where σ is used to control the noise intensity. We also compare the results to several baseline estimators: KSG (Kraskov et al., 2004), generalized nearest neighbor graph (GNN) (Pál et al., 2010) [3], minimum spanning trees (MST) (Müller et al., 2012; Yukich and Yukich, 1998), and exponential family with maximum likelihood estimation (EXP) (Nielsen and Nock, 2010) [4].

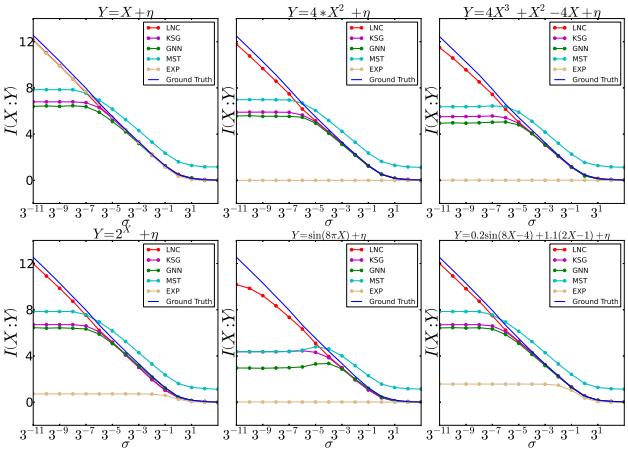

Figure 3: For all the functional relationships above, we used a sample size $N = 5000$ for each noise level and the nearest neighbor parameter $k = 5$ for LNC, KSG and GNN estimators.

Figure 3 demonstrates that the proposed estimator, LNC, consistently outperforms the other estimators. Its superiority is most significant for the low noise regime. In that case, both KSG and GNN estimators are bounded by the sample size while LNC keeps growing. MST tends to overestimate MI for large noise but still stops growing after the noise fall below a certain intensity[5]. Surprisingly, EXP is the only other estimator that performs comparably with LNC for the linear relationship (the left most plot in Figure 3). However, it fails dramatically for all the other relationship types. See Appendix D for more functional relationship tests.

Convergence rate Figure 4 shows the convergence of the two estimators $\hat{I}_{KSG}$ and $\hat{I}_{LNC}$, at a fixed (small) noise intensity, as we vary the sample size. We test the estimators on both two and five dimensional data for linear and quadratic relationships. We observe that LNC is doing better overall. In particular, for linear relationships in 2D and 5D, as well as quadratic relationships in 2D, the required sample size for LNC is several orders of magnitude less that for $\hat{I}_{KSG}$. For instance, for the 5D linear relationship KSG does not converge even for the sample size (10^5) while LNC converges to the true value with only 100 samples.

Finally, it is worthwhile to remark on the relatively slow convergence of LNC for the 5D quadratic example. This is because LNC, while relaxing the local uniformity assumptions, still assumes local linearity. The stronger the nonlinearity, the more samples are required to find neighborhoods that are locally approximately linear. We note, however, that LNC still converges faster than KSG.

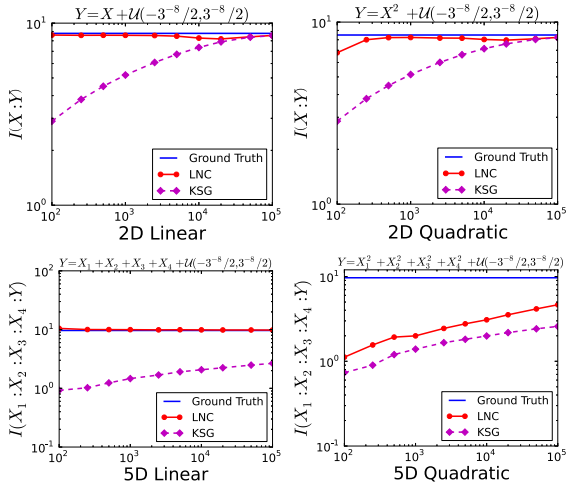

Figure 4: Estimated MI using both KSG and LNC estimators in the number of samples ($k = 5$ and $\alpha_{k,d} = 0.37$ for 2D examples; $k = 8$ and $\alpha_{k,d} = 0.12$ for 5D examples)

5.2 Experiments with real-world data

5.2.1 Ranking Relationship Strength

We evaluate the proposed estimator on the WHO dataset which has 357 variables describing various socio-economic, political, and health indicators for different countries [6]. We calculate the mutual information between pairs of variables which have at least 150 samples. Next, we rank the pairs based on their estimated mutual information and choose the top 150 pairs with highest mutual information. For these top 150 pairs, We randomly select a fraction ρ of samples for each pair, hide the rest samples and then recalculate the mutual information. We want to see how

[3] We use the online code `http://www.cs.cmu.edu/~bapoczos/codes/REGO_with_kNN.zip` for the GNN estimator.

[4] We use the Information Theoretical Estimators Toolbox (ITE) (Szabó, 2014) for MST and EXP estimators.

[5] Even if $k = 2$ or 3, KSG and GNN estimators still stop growing after the noise fall below a certain intensity.

[6] WHO dataset is publicly available at `http://www.exploredata.net/Downloads`

mutual information-based rank changes by giving different amount of less data, i.e., varying ρ. A good mutual information estimator should give a similar rank using less data as using the full data. We compare our LNC estimator to KSG estimator. Rank similarities are calculated using the standard Spearman's rank correlation coefficient described in (Spearman, 1904). Fig 5 shows the results. We can see that LNC estimator outperforms KSG estimator, especially when the missing data approaches 90%, Spearman correlation drops to 0.4 for KSG estimator, while our LNC estimator still has a relatively high score of 0.7.

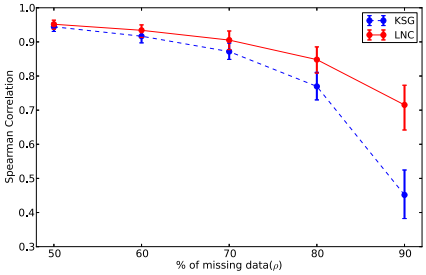

Figure 5: Spearman correlation coefficient between the original MI rank and the rank after hiding some percentage of data by KSG and LNC estimator respectively. The 95% confidence bars are obtained by repeating the experiment for 200 times.

5.2.2 Finding interesting triplets

We also use our estimator to find strong *multivariate* relationships in the WHO data set. Specifically, we search for *synergistic* triplets (X, Y, Z), where one of the variables, say Z, can be predicted by knowing both X and Y simultaneously, but not by using either variable separately. In other words, we search for triplets (X, Y, Z) such that the pair-wise mutual information between the pairs $I(X : Y)$, $I(X : Z)$ and $I(Y : Z)$ are low, but the multi-information $I(X : Y : Z)$ is relatively high. We rank relationships using the following synergy score:[7] $SS = I(X : Y : Z)/\max\{I(X : Y), I(Y : Z), I(Z : X)\}$.

We select the triplets that have synergy score above a certain threshold. Figure 6 shows two synergistic relationships detected by LNC but not by KSG. In these examples, both KSG and LNC estimators yield low mutual information for the pairs (X, Y), (Y, Z) and (X, Z). However, in contrast to KSG, our estimator yields a relatively high score for multi-information among the three variables.

For the first relationship, the synergistic behavior can be explained by noting that the ratio of the Total Energy Generation (Y) to Electricity Generation per Person (Z) essentially yields the size of the population, which is highly predictive of the Number of Female Cervical Cancer cases (X). While this example might seem somewhat trivial, it illustrates the ability of our method to extract synergistic relationships automatically without any additional assumptions and/or data preprocessing.

[7] Another measure of synergy is given by the so called "interaction information": $I(X : Y) + I(Y : Z) + I(Z : X) - I(X : Y : Z)$.

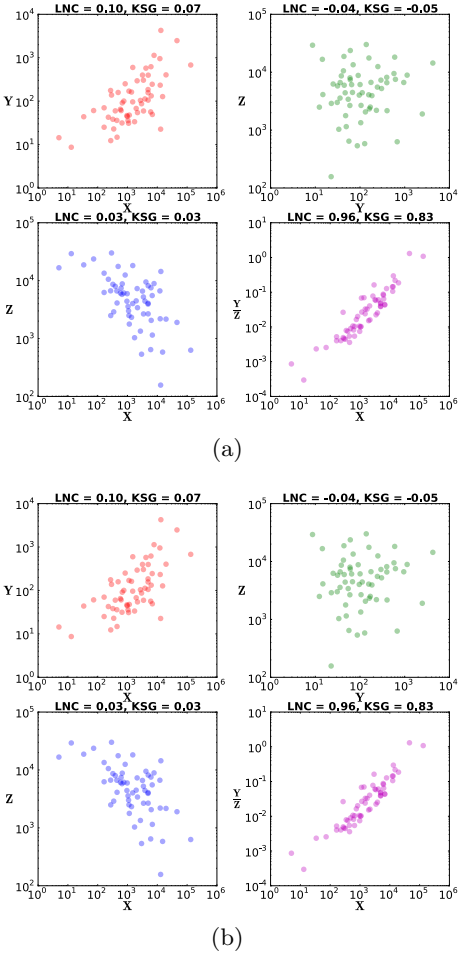

Figure 6: Two examples of synergistic triplets: $\hat{I}_{KSG}(X : Y : Z) = 0.14$ and $\hat{I}_{LNC}(X : Y : Z) = 0.95$ for the first example; $\hat{I}_{KSG}(X : Y : Z) = 0.05$ and $\hat{I}_{LNC}(X : Y : Z) = 0.7$ for the second example

In the second example, LNC predicts a strong synergistic interaction between Total Cell Phones (Y), Number of Female Cervical Cancer cases (Z), and rate of Tuberculosis deaths (X). Since the variable Z (the number of female cervical cancer cases) grows with the total population, $\frac{Y}{Z}$ is proportional to the average number of cell phones per person. The last plot indicates that a higher number of cell phones per person are predictive of lower tuberculosis death rate. One

possible explanation for this correlation is some common underlying cause (e.g., overall economic development). Another intriguing possibility is that this finding reflects recent efforts to use mobile technology in TB control.[8]

6 Related Work

MI estimators There has been a significant amount of recent work on estimating entropic measures such as divergences and mutual information from samples (see this survey (Walters-Williams and Li, 2009) for an exhaustive list). Khan et al. (2007) compared different MI estimators for varying sample sizes and noise intensity, and reported that for small samples, the KSG estimator was the best choice overall for relatively low noise intensities, while KDE performed better at higher noise intensities. Other approaches include estimators based on Generalized Nearest-Neighbor Graphs (Pál et al., 2010), minimum spanning trees (Müller et al., 2012), maximum likelihood density ratio estimation (Suzuki et al., 2008), and ensemble methods (Sricharan et al., 2013; Moon and Hero, 2014). In particular, the latter approach works by taking a weighted average of simple density plug-in estimates such as kNN or KDE. However, it is upper-bounded by the largest value among its simple density estimates. Therefore, this method would still underestimate the mutual information when it goes larger as discussed before.

It has been recognized that kNN-based entropic estimators underestimate probability density at the sample points that are close to the support boundary (Liitiäinen et al., 2010). Sricharan et al. (2012) proposed a bipartite plug-in estimator for non-linear density functionals that extrapolates the density estimates at *interior* points that are close to the boundary in order to compensate the boundary bias. However, this method requires to identify boundary and interior points, which is a difficult problem when mutual information is large, so that almost all the points are close to the boundary. Singh and Poczos (2014) used a "mirror image" kernel density estimator to escape the boundary effect, but their estimator relies on the knowledge of the support of the densities.

Mutual Information and Equitability Reshef et al. (2011) introduced a property they called "suitability" for a measure of correlation. If two variables are related by a functional form with some noise, equitable measures should reflect the magnitude of the noise while being insensitive to the form of the functional relationship. They used this notion to justify a new correlation measure called MIC. Based on comparisons with MI using the KSG estimator, they concluded that MIC is "more equitable" for comparing relationship strengths. While several problems (Simon and Tibshirani, 2014; Gorfine et al.) and alternatives (Heller et al., 2013; Székely et al., 2009) were pointed out, Kinney and Atwal (KA) showed that MIC's apparent superiority to MI was actually due to flaws in estimation (Kinney and Atwal, 2014). A more careful definition of equitability led KA to the conclusion that MI is actually more equitable than MIC. KA suggest that mutual information estimation could be improved by using more samples for estimation. However, here we showed that the number of samples required for KSG is prohibitively large, but that this difficulty can be overcome by using an improved MI estimator.

7 Conclusion

The problem of deciding whether or not two variables are independent is a historically significant endeavor. However, modern data mining presents us with problems requiring a totally different perspective. It is not unusual to have thousands of variables which could have millions of potential relationships. We have insufficient resources to examine each potential relationship so we need an assumption-free way to pick out only the most promising relationships for further study. Many applications have this flavor including the health indicator data considered above as well as gene expression microarray data, human behavior data, to name a few. How can we select the most interesting relationships? Mutual information gives a clear and general basis for comparing the strength of otherwise dissimilar variables and relationships. While non-parametric mutual information estimators exist, we showed that strong relationships require at least exponentially many samples to accurately measure using some of these techniques. We introduced a non-parametric mutual information estimator that can measure the strength of nonlinear relationships even with small sample sizes. We have incorporated these novel estimators into an open source entropy estimation toolbox.[9] As the amount and variety of available data grows, general methods for identifying strong relationships will become increasingly necessary. We hope that the developments suggested here will help to address this need.

Acknowledgements

This research was supported in part by DARPA grant No. W911NF–12–1–0034.

[8] See *Stop TB Partnership*, `http://www.stoptb.org`.

[9] `https://github.com/BiuBiuBiLL/MIE`

References

A. R. Alizad Rahvar and M. Ardakani. Boundary effect correction in k-nearest-neighbor estimation. *Phys. Rev. E*, 83:051121, May 2011. doi: 10.1103/PhysRevE.83.051121. URL http://link.aps.org/doi/10.1103/PhysRevE.83.051121.

T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

Malka Gorfine, Ruth Heller, and Yair Heller. Comment on detecting novel associations in large data sets. *(available at http://emotion. technion. ac. il/ gorfinm/files/science6. pdf on 11 Nov. 2012)*.

Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

Shiraj Khan, Sharba Bandyopadhyay, Auroop R. Ganguly, Sunil Saigal, David J. Erickson, Vladimir Protopopescu, and George Ostrouchov. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Phys. Rev. E*, 76:026209, Aug 2007. doi: 10.1103/PhysRevE.76.026209. URL http://link.aps.org/doi/10.1103/PhysRevE.76.026209.

J. Kinney and G. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004. doi: 10.1103/PhysRevE.69.066138. URL http://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Elia Liitiäinen, Amaury Lendasse, and Francesco Corona. A boundary corrected expansion of the moments of nearest neighbor distributions. *Random Struct. Algorithms*, 37(2):223–247, September 2010. ISSN 1042-9832. doi: 10.1002/rsa.v37:2. URL http://dx.doi.org/10.1002/rsa.v37:2.

K.R. Moon and A.O. Hero. Ensemble estimation of multivariate f-divergence. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 356–360, June 2014. doi: 10.1109/ISIT.2014.6874854.

Andreas Müller, Sebastian Nowozin, and Christoph Lampert. Information theoretic clustering using minimum spanning trees. *Pattern Recognition*, pages 205–215, 2012.

F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 3621–3624. IEEE, 2010.

Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems 23*, pages 1849–1857. Curran Associates, Inc., 2010.

Fernando Pérez-Cruz. Estimation of information theoretic measures for continuous random variables. In *Proceedings of NIPS-08*, pages 1257–1264, 2008.

David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 (5500):2323–2326, 2000.

Noah Simon and Robert Tibshirani. Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *arXiv preprint arXiv:1401.7645*, 2014.

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23(3-4):301–321, 2003. doi: 10.1080/01966324.2003.10737616. URL http://dx.doi.org/10.1080/01966324.2003.10737616.

Shashank Singh and Barnabas Poczos. Generalized exponential concentration inequality for renyi divergence estimation. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 333–341, 2014. URL http://machinelearning.wustl.edu/mlpapers/papers/icml2014c1_singh14.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):pp. 72–101, 1904. ISSN 00029556. URL http://www.jstor.org/stable/1412159.

K. Sricharan, R. Raich, and A.O. Hero. Estimation of nonlinear functionals of densities with confidence. *Information Theory, IEEE Transactions on*, 58(7):4135–4159, July 2012. ISSN 0018-9448. doi: 10.1109/TIT.2012.2195549.

K. Sricharan, D. Wei, and A.O. Hero. Ensemble estimators for multivariate entropy estimation. *Information Theory, IEEE Transactions on*, 59(7):4374–4388, July 2013. ISSN 0018-9448. doi: 10.1109/TIT.2013.2251456.

Milan Studenỳ and Jirina Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. In *Learning in graphical models*, pages 261–297. Springer, 1998.

Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In Yvan Saeys, Huan Liu, Iaki Inza, Louis Wehenkel, and Yves Van de Peer, editors, *FSDM*, volume 4 of *JMLR Proceedings*, pages 5–20. JMLR.org, 2008.

Zoltán Szabó. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014. (https://bitbucket.org/szzoli/ite/).

Gábor J Székely, Maria L Rizzo, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

Ulrike Von Luxburg and Morteza Alamgir. Density estimation from unweighted k-nearest neighbor graphs: a roadmap. In *Advances in Neural Information Processing Systems*, 2013.

Janett Walters-Williams and Yan Li. Estimation of mutual information: A survey. In *Rough Sets and Knowledge Technology*, volume 5589 of *Lecture Notes in Computer Science*, pages 389–396. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-02961-5. doi: 10.1007/978-3-642-02962-2_49. URL http://dx.doi.org/10.1007/978-3-642-02962-2_49.

Q. Wang, S.R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Inf. Theor.*, 55:2392–

2405, May 2009. ISSN 0018-9448. doi: 10.1109/TIT.2009.2016060. URL `http://dl.acm.org/citation.cfm?id=1669487.1669521`.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

Joseph E Yukich and Joseph Yukich. *Probability theory of classical Euclidean optimization problems*. Springer Berlin, 1998.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2002.