Estimating Mutual Information by Local Gaussian Approximation

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

Estimating mutual information (MI) from samples is a fundamental problem in statistics, machine learning, and data analysis. Recently it was shown that a popular class of non-parametric MI estimators perform very poorly for strongly dependent variables and have sample complexity that scales exponentially with the true MI. This undesired behavior was attributed to the reliance of those estimators on local uniformity of the underlying (and unknown) probability density function. Here we present a novel semi-parametric estimator of mutual information, where at each sample point, densities are *locally* approximated by a Gaussians distribution. We demonstrate that the estimator is asymptotically unbiased. We also show that the proposed estimator has a superior performance compared to several baselines, and is able to accurately measure relationship strengths over many orders of magnitude.

1 Introduction

Mutual information (MI) is a fundamental measure of dependence between two random variables. While it initially arose in the theory of communication as a natural measure of ability to communicate over noisy channels (Shannon, 1948), mutual information has since been used in different disciplines such as machine learning, information retrieval, neuroscience, and computational biology, to name a few. This widespread use is due in part to the generality of the measure, which allows it to characterize dependency strength for both linear and non-linear relationships between arbitrary random variables.

Let us consider the following basic problem, where, given a set of i.i.d. samples from an unknown, absolutely continuous joint distribution, our goal is to estimate the mutual information from these samples. A naive method would be first to learn the underlying probability distribution using either parametric or non-parametric methods, and then calculate the mutual information from the obtained distribution. Unfortunately, this naive approach often fails, as it requires a very large number of samples, especially in high dimensions. A different approach is to estimate mutual information directly from samples. For instance, rather than estimating the whole probability distribution, one could estimate the density (and its marginals) only at each sample point, and then plug those estimates into the expression for mutual information. This type of direct estimators been shown to be a more feasible method for estimating MI in higher dimensions. An important and very popular class of such estimators is based on k-nearest-neighbor (kNN) graphs and their generalizations (Singh et al., 2003; Kraskov et al., 2004; Pál et al., 2010).

Despite the widespread popularity of the direct estimators, it was recently demonstrated that those methods fail to accurately estimate mutual information for *strongly dependent* variables (Gao et al., 2015). Specifically, it was shown that accurate estimation of mutual information between two strongly dependent variables requires a number of samples that scales *exponentially* with the true mutual information. This undesired behavior was contributed to the assumption of local uniformity of the underlying distribution postulated by those estimators. To address this shortcoming, (Gao et al., 2015) proposed to add a correction term to compensate for non-uniformity, based on local PCA-induced neighborhoods. Although intuitive, the resulting estimator relied on a heuristically tuned threshold parameter and had no theoretical performance guarantees (Gao et al., 2015).

Our main contribution is to propose a novel mutual information estimator based on *local Gaussian approximation*, with provable performance guarantees, and superior empirical performance compared to existing estimators over a wide range of relationship strength. Instead of assuming a uniform distribution in the local neighborhood, our new estimator assumes a Gaussian distribution *locally* around each point. The new estimator leverages previous results on *local likelihood density estimation* (Hjort and Jones, 1996;

Loader, 1996). As our main theoretical result, we demonstrate that the new estimator is asymptotically unbiased. We also demonstrate that the proposed estimator performs as well as existing baseline estimators for weak relationships, but outperforms all of those estimators for stronger relationships.

The paper is organized as follows. In the next section, we review the basic definitions of information-theoretic concepts such as mutual information and formally define our problem. In section 3, we review the limitations of current mutual information estimators as pointed out in (Gao et al., 2015). Section 4 introduces local likelihood density estimation. In Section 5 we use this density estimator to propose a novel entropy and mutual information estimator, and summarize certain theoretical properties of those estimator, which are then proved in Section 6. Section 7 provides numerical experiments demonstrating the superiority of the proposed estimator. We conclude the paper with a brief survey of related work followed by the discussion of our main results and some open problems.

2 Formal Problem Definition

In this section we briefly review the formal definition of Shannon entropy and mutual information, before formally defining the objective of our paper.

Definition 1 *Let $\mathbf{x}$ denote a d-dimensional absolutely continuous random variable with probability density function $f : \mathbb{R}^d \to \mathbb{R}$. The Shannon differential entropy is defined as*

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

Definition 2 *Let $\mathbf{x}$ and $\mathbf{y}$ denote d-dimensional and b-dimensional absolutely continuous random variables with probability density function $f_X : \mathbb{R}^d \to \mathbb{R}$ and $f_Y : \mathbb{R}^b \to \mathbb{R}$, respectively. Let f_{XY} denote the joint probability density function of $\mathbf{x}$ and $\mathbf{y}$. The mutual information between $\mathbf{x}$ and $\mathbf{y}$ is defined as*

$$I(\mathbf{x} : \mathbf{y}) = \int_{\mathbf{y}\in\mathbb{R}^b} \int_{\mathbf{x}\in\mathbb{R}^d} f_{XY}(\mathbf{x},\mathbf{y}) \log \frac{f_{XY}(\mathbf{x},\mathbf{y})}{f_X(\mathbf{x})\, f_Y(\mathbf{y})} d\mathbf{x} d\mathbf{y} \tag{2}$$

It is easy to show that

$$I(\mathbf{x} : \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x},\mathbf{y}) \ , \tag{3}$$

where $H(\mathbf{x},\mathbf{y})$ stands for the joint entropy of $(\mathbf{x},\mathbf{y})$, and can be calculated from Eq. 1 using the joint density f_{XY}. We use the natural logarithms so that information is measured in nats.

It is sometime useful to represent entropy and mutual information as the following expectations:

$$H(\mathbf{x}) = \mathbb{E}_X[-\log f(\mathbf{x})] \tag{4}$$

$$I(\mathbf{x} : \mathbf{y}) = \mathbb{E}_{XY}\left[\log \frac{f_{XY}(\mathbf{x},\mathbf{y})}{f_X(\mathbf{x})\, f_Y(\mathbf{y})}\right] \tag{5}$$

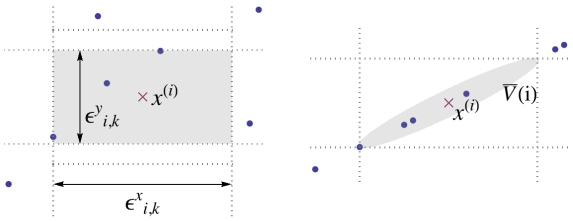
Assume now we are given N i.i.d. samples $(\mathcal{X},\mathcal{Y}) = \{(\mathbf{x},\mathbf{y})^{(i)}\}_{i=1}^n$ from the unknown joint distribution f_{XY}. Our goal is then to construct a mutual information estimator $\hat{I}(\mathbf{x} : \mathbf{y})$ based on those samples.

3 Limitations of Nonparametric MI Estimators

As pointed out in Section 1, one of the most popular class of mutual information estimators is based on k-nearest neighbor (kNN) graphs and their generalizations (Singh et al., 2003; Kraskov et al., 2004; Pál et al., 2010). However, it was recently shown that for strongly dependent variables, those estimators tend to underestimate the mutual information (Gao et al., 2015). To understand this problem, let us focus on kNN-based estimator as an example. The kNN estimator assumes uniform density within the kNN rectangle (containing k-nearest neighbors), as shown in Figure 1(a). Generally speaking, this assumption can be made valid for any relationship as long as we have sufficient number of samples. However, for limited sample size, this assumption becomes problematic when the relationship between the two variables becomes sufficiently strong. In fact, as shown in Fig. 1(b), the obtained *local* neighborhood induced by kNN is beyond the support of the probability distribution (shaded area).

This undesired behavior is closely related to the so-called *boundary effect* that occurs in nonparametric density estimation problem. Namely, for strongly dependent random variables, almost all the sample points are close to the boundary of the support (as illustrated in Figure 1(b)), making the density estimation problem difficult.

To relax the local uniformity assumption in kNN-based estimators, (Gao et al., 2015) proposed to replace the axis-aligned rectangle with a PCA-aligned rectangle *locally*, and use the volume of this rectangle for estimating the unknown density at a given point. Mathematically, the above revision was implemented by introducing a novel term that accounted for local non-uniformity. It was shown the the revised estimator significantly outperformed the existing estimators for strongly dependent variables. Nevertheless, the estimator suggested in (Gao et al., 2015) relied on a heuristic for determining when to use the correction term, and did not have any theoretical guarantees. In the remaining of this paper, we suggest a novel estimator based on *local gaussian approximation*, as more general approach to overcome the above limitations. The main idea is that, instead



(a) (b)

Figure 1: For 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, $k = 3$, (shaded area), and (b) for points drawn from a distribution over two strongly correlated variables, $k = 4$, (the area within dotted lines).

of assuming a uniform distribution around the local kNN- or a PCA-aligned rectangle, we approximate the unknown density at each sample point by a local *Gaussian* distribution, which is estimated using the k-nearest neighborhood of that point. In addition to demonstrating superior empirical performance of the proposed estimator, we also show that it is asymptotically unbiased.

4 Local Gaussian Density Estimation

In this section, we introduce a density estimation method called local Gaussian density estimation, or LGDE (Hjort and Jones, 1996), which serves as the basic building block for the proposed mutual information estimator.

Consider N i.i.d. samples $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ drawn from an unknown density $f(\mathbf{x})$, where $\mathbf{x}$ is a d-dimensional continuous random variable. The central idea behind LGDE is to locally approximate the unknown probability density at point $\mathbf{x}$ using a Gaussian parametric family $\mathcal{N}_d(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x}))$, where $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\Sigma}(\mathbf{x})$ are the ($\mathbf{x}$-dependent) mean and covariance matrix of each local approximation. This intuition is formalized in the following definition:

Definition 3 (Local Gaussian Density Estimator) *Let $\mathbf{x}$ denote a d-dimensional absolutely continuous random variable with probability density function $f(\mathbf{x})$, and let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ be N i.i.d. samples drawn from $f(\mathbf{x})$. Furthermore, let $K_{\mathbf{H}}(\mathbf{x})$ be a product kernel with diagonal bandwidth matrix $\mathbf{H} = diag(h_1, h_2, ..., h_d)$, so that $\mathbf{K_H}(\mathbf{x}) = h_1^{-1}K\left(h_1^{-1}x_1\right)h_2^{-1}K\left(h_2^{-1}x_2\right)...h_d^{-1}K\left(h_d^{-1}x_d\right)$, where $K(\cdot)$ can be any one-dimensional kernel function. Then the Local Gaussian Density Estimator, or LGDE, of $f(\mathbf{x})$ is given by*

$$\widehat{f}(\mathbf{x}) = \mathcal{N}_d(\mathbf{x}; \boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x})) \ , \tag{6}$$

Here $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ are different for each point $\mathbf{x}$, and are obtained by solving the following optimization problem,

$$\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x}) = \arg\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \mathcal{L}(\mathbf{x}, \mu, \Sigma) \ , \tag{7}$$

where $\mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the local likelihood function defined as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{K_H}(\mathbf{x}_i - \mathbf{x}) \log \mathcal{N}_d(\mathbf{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &- \int \mathbf{K_H}(\mathbf{t} - \mathbf{x}) \mathcal{N}_d(\mathbf{t}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, d\mathbf{t} \end{aligned} \tag{8}$$

The first term in the right hand side of Eq. 8 is the localized version of Gaussian log-likelihood. One can see that without the kernel function, Eq. 8 becomes similar to the global log-likelihood function of the Gaussian parametric family. However, since we do not have sufficient information to specify a global distribution, we make a local smoothness assumption by adding this kernel function. The second term of right hand side in Eq. 8 is a penalty term to ensure the consistency of the density estimator.

The key difference between kNN density estimator and LGDE is that the former assumes that the density is locally uniform over the neighborhood of each sample point, whereas the latter method relaxes *local uniformity* to *local linearity*[1], which allows to compensates for the boundary bias. In fact, any non-uniform parametric probability distribution is suitable for fitting a local distribution under the local likelihood, and the Gaussian distribution used here is simply one realization.

I Theorem 1 below establishes the consistency property of this local Gaussian estimator; for a detailed proof see (Hjort and Jones, 1996).

Theorem 1 ((Hjort and Jones, 1996)) *Let $\mathbf{x}$ denote a d-dimensional absolutely continuous random variable with probability density function $f(\mathbf{x})$, and let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ be N i.i.d. samples drawn from $f(\mathbf{x})$. Let $\widehat{f}(x)$ be the Local Gaussian Density Estimator with diagonal bandwidth matrix $diag(h_1, h_2, ..., h_d)$, where the diagonal elements h_i-s satisfy the following conditions:*

$$\lim_{N\to\infty} h_i = 0 \ , \ \lim_{N\to\infty} N h_i = \infty, i = 1, 2, \ldots, d. \tag{9}$$

Then the following holds:

$$\lim_{N\to\infty} \mathbb{E}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})| = 0 \tag{10}$$

$$\lim_{N\to\infty} \mathbb{E}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 = 0 \tag{11}$$

[1] To elaborate on the local linearity, we note that Gaussian distribution is essentially a special case of Elliptical distribution $f(\mathbf{x}) = k * g((\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$. Therefore, the local Gaussian approximation actually assumes a rotated hyper-ellipsoid locally at each point.

The above theorem states that LGDE is asymptotically unbiased and L2-consistent.

5 LGDE-based Estimators for Entropy and Mutual Information

We now introduce our estimators for entropy and mutual information that are inspired by the local density estimation approach defined in the previous section.

Let us again consider N i.i.d samples $(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{x}, \mathbf{y})^{(i)}\}_{i=1}^{N}$ drawn from an unknown joint distribution f_{XY}, where $\mathbf{x}$ and $\mathbf{y}$ are random vectors of dimensionality d and b, respectively. Let us construct the following estimators for entropy,

$$\widehat{H}(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N}\log \widehat{f}(\mathbf{x}_i), \tag{12}$$

and mutual information

$$\widehat{I}(\mathbf{x}:\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\log \frac{\widehat{f}(\mathbf{x}_i, \mathbf{y}_i)}{\widehat{f}(\mathbf{x}_i)\,\widehat{f}(\mathbf{y}_i)} \tag{13}$$

where $\widehat{f}(\mathbf{x})$, $\widehat{f}(\mathbf{y})$, $\widehat{f}(\mathbf{x}, \mathbf{y})$ are the local Gaussian density estimators for $f_X(\mathbf{x})$, $f_Y(\mathbf{y})$, $f_{XY}(\mathbf{x}, \mathbf{y})$ respectively, defined in the previous section.

Recall that the entropy and mutual information can be written as appropriately defined expectations; see Eqs. 4 and 5. Then the proposed estimator simply replaces the expectation by the sample averages, and then plugs in density estimators from Section 4 into those expectations.

The next two theorems state that the proposed estimators are asymptotically unbiased.

Theorem 2 (Asymptotic Unbiasedness of Entropy Estimator) *If the conditions in Eq. 9 hold, then the entropy estimator given by Eq. 12 is asymptotically unbiased, i.e.,*

$$\lim_{N\to\infty} \mathbb{E}\widehat{H}(\mathbf{x}) = H(\mathbf{x}) \tag{14}$$

Theorem 3 (Asymptotic Unbiasedness of MI Estimator) *If the conditions in Eq. 9 hold, then the mutual information estimator given by Eq. 13 is asymptotically unbiased:*

$$\lim_{N\to\infty} \mathbb{E}\widehat{I}(\mathbf{x}:\mathbf{y}) = I(\mathbf{x}:\mathbf{y}) \tag{15}$$

We provide the proofs of the above theorems in the next section.

6 Proofs of the Theorems

Before getting to the actual proofs, we first introduce the Lebesgue's dominated convergence theorem.

Theorem 4 (Lebesgue dominated convergence theorem) *Let $\{f_N\}$ be a sequence of functions, and assume this sequence converges point-wise to a function f, i.e., $f_N(\mathbf{x}) \to f(\mathbf{x})$ for any $\mathbf{x} \in R^d$. Furthermore, let us assume that f_N is dominated by an integrable function g, e.g., we have for any $\mathbf{x}$*

$$|f_N(\mathbf{x})| \le g(\mathbf{x})$$

Then we have

$$\lim_{N\to\infty}\int_{\mathbf{x}\in\mathbf{X}} |f_N(\mathbf{x}) - f(\mathbf{x})|\, d\mathbf{x} = 0$$

6.1 Proof of Theorem 2

Consider N i.i.d. samples $\left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N}$ drawn from the probability density $f(\mathbf{x})$, and let $F_N(\mathbf{x})$ denote the empirical cumulative distribution function.

Let us define the following two quantities:

$$H_1 = -\frac{1}{N}\sum_{i=1}^{N}\ln \mathbb{E}\widehat{f}(\mathbf{x}_i) \tag{16}$$

$$H_2 = -\frac{1}{N}\sum_{i=1}^{N}\ln f(\mathbf{x}_i) \tag{17}$$

Then we have,

$$\begin{aligned}&\mathbb{E}|\widehat{H}(\mathbf{x}) - H(\mathbf{x})| \\ &= \mathbb{E}|(\widehat{H} - H_1) + (H_1 - H_2) + (H_2 - H)| \\ &\le \mathbb{E}|\widehat{H} - H_1| + \mathbb{E}|H_1 - H_2| + \mathbb{E}|H_2 - H|\end{aligned} \tag{18}$$

We now procced to show that each of the terms in Eq. 18 individually converges to 0 in the limit $N \to \infty$, which will then yield Eq. 14.

First, we note that according to the mean value theorem, for any $\mathbf{x}$, there exist $t_{\mathbf{x}}$ and $t'_{\mathbf{x}}$ in $(0, 1)$, such that

$$\begin{aligned}&\ln \widehat{f}(\mathbf{x}) = \ln \mathbb{E}\widehat{f}(\mathbf{x}) + \\ &\left(\widehat{f}(\mathbf{x}) - \mathbb{E}\widehat{f}(\mathbf{x})\right)\ln\left(t_{\mathbf{x}}\widehat{f}(\mathbf{x}) + (1 - t_{\mathbf{x}})\,\mathbb{E}\widehat{f}(\mathbf{x})\right)\end{aligned} \tag{19}$$

and

$$\begin{aligned}&\ln \mathbb{E}\widehat{f}(\mathbf{x}) = \ln f(\mathbf{x}) + \\ &\left(\mathbb{E}\widehat{f}(\mathbf{x}) - f(\mathbf{x})\right)\ln\left(t'_{\mathbf{x}} f(\mathbf{x}) + (1 - t'_{\mathbf{x}})\,\mathbb{E}\widehat{f}(\mathbf{x})\right)\end{aligned} \tag{20}$$

For the first term in Eq. 18, we use Eq. 19 to obtain

$$\begin{aligned}
&\mathbb{E}\left|\widehat{H}-H_1\right| \\
&=\mathbb{E}\left|\int\left[\ln\widehat{f}\left(\mathbf{x}\right)-\ln\mathbb{E}\widehat{f}\left(\mathbf{x}\right)\right]dF_N\left(\mathbf{x}\right)\right| \\
&=\mathbb{E}\left|\int\frac{|\widehat{f}(\mathbf{x})-\mathbb{E}\widehat{f}(\mathbf{x})|}{t_{\mathbf{x}}\widehat{f}\left(\mathbf{x}\right)+\left(1-t_{\mathbf{x}}\right)\mathbb{E}\widehat{f}\left(\mathbf{x}\right)}dF_N\left(\mathbf{x}\right)\right| \\
&\leq\frac{1}{1-t}\mathbb{E}\left|\int\frac{|\widehat{f}(\mathbf{x})-\mathbb{E}\widehat{f}(\mathbf{x})|}{\mathbb{E}\widehat{f}\left(\mathbf{x}\right)}dF_N\left(\mathbf{x}\right)\right| \\
&=\frac{1}{1-t}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{|\widehat{f}(\mathbf{x}_i)-\mathbb{E}\widehat{f}(\mathbf{x}_i)|}{\mathbb{E}\widehat{f}\left(\mathbf{x}_i\right)}\right) \\
&=\frac{1}{1-t}\mathbb{E}\left(\mathbb{E}\left(\frac{|\widehat{f}(\mathbf{u})-\mathbb{E}\widehat{f}(\mathbf{u})|}{\mathbb{E}\widehat{f}\left(\mathbf{u}\right)}\right)|\mathbf{x}=\mathbf{u}\right) \\
&=\frac{1}{1-t}\int|\widehat{f}\left(\mathbf{u}\right)-\mathbb{E}\widehat{f}\left(\mathbf{u}\right)|\frac{\widehat{f}\left(\mathbf{u}\right)}{\mathbb{E}\widehat{f}\left(\mathbf{u}\right)}d\mathbf{u} \qquad (21)
\end{aligned}$$

where t is the maximum value among all $t_{\mathbf{x}}$. Using Theorem 1, we have $|\widehat{f}\left(\mathbf{u}\right)-\mathbb{E}\widehat{f}\left(\mathbf{u}\right)|\rightarrow 0$ as $N\rightarrow\infty$. Furthermore, it is possible to show that $\exists N_0$, so that for any $N>N_0$ one has $|\widehat{f}\left(\mathbf{u}\right)-\mathbb{E}\widehat{f}\left(\mathbf{u}\right)|\frac{\widehat{f}(\mathbf{u})}{\mathbb{E}\widehat{f}(\mathbf{u})}<2f\left(\mathbf{u}\right)$. Thus, using Theorem 4, we obtain

$$\lim_{N\rightarrow\infty}\mathbb{E}|H-H_1|=0 \qquad (22)$$

Similarly, using Eq. 20, $\mathbb{E}\left|H_1-H_2\right|$ can be written as

$$\begin{aligned}
&\mathbb{E}\left|H_1-H_2\right| \\
&=\mathbb{E}\left|\int\left[\ln\mathbb{E}\widehat{f}\left(\mathbf{x}\right)-\ln f\left(\mathbf{x}\right)\right]dF_N\left(\mathbf{x}\right)\right| \\
&=\mathbb{E}\left|\int\frac{|\mathbb{E}\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{t'_{\mathbf{x}}f\left(\mathbf{x}\right)+\left(1-t'_{\mathbf{x}}\right)\mathbb{E}\widehat{f}\left(\mathbf{x}\right)}dF_N\left(\mathbf{x}\right)\right| \\
&\leq\frac{1}{t'}\mathbb{E}\left|\int\frac{|\mathbb{E}\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{f\left(\mathbf{x}\right)}dF_N\left(\mathbf{x}\right)\right| \\
&=\frac{1}{t'}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{|\mathbb{E}\widehat{f}(\mathbf{x}_i)-f(\mathbf{x}_i)|}{f\left(\mathbf{x}_i\right)}\right) \\
&=\frac{1}{t'}\int f\left(\mathbf{x}\right)\frac{|\mathbb{E}\widehat{f}(\mathbf{x})-f(\mathbf{x})|}{f\left(\mathbf{x}\right)}d\mathbf{x} \\
&=\frac{1}{t'}\int|\mathbb{E}\widehat{f}\left(\mathbf{x}\right)-f\left(\mathbf{x}\right)|d\mathbf{x} \qquad (23)
\end{aligned}$$

where t' is the minimum value among all $t'_{\mathbf{x}}$.

Invoking Theorem 1 again, we observe that the last term in Eq. 23 $|\mathbb{E}\widehat{f}\left(\mathbf{x}\right)-f\left(\mathbf{x}\right)|\rightarrow 0$ as $N\top\infty$, and is bounded by $2f(\mathbf{x})$ for sufficiently large N (e.g., when when $\widehat{f}(\mathbf{u})$ and $\mathbb{E}\widehat{f}(\mathbf{u})$ are sufficiently close). Therefore, by Theorem 4, we have

$$\lim_{N\rightarrow\infty}\mathbb{E}\left|H_1-H_2\right|=0 \qquad (24)$$

Finally, for the last term in Eq. 18, we note that

$$\mathbb{E}H_2=-\frac{1}{N}\mathbb{E}\sum_{i=1}^{N}\ln f\left(\mathbf{x}_i\right)=\mathbb{E}[-\ln f\left(\mathbf{x}\right)] \qquad (25)$$

Thus, $\mathbb{E}H_2$ is simply the entropy in Definition 1; see Eq. 4. Therefore,

$$\lim_{N\rightarrow\infty}\mathbb{E}\left|H_2-H\right|=0 \qquad (26)$$

Combining Eqs. 22, 24, 26 and 18, we arrive at Eq. 14, which concludes the proof.

6.2 Proof of Theorem 3

For mutual information estimation, we use Eq. 3 to get

$$\begin{aligned}
\mathbb{E}|\widehat{I}\left(\mathbf{x}:\mathbf{y}\right)-I\left(\mathbf{x}:\mathbf{y}\right)| &\leq \mathbb{E}|H\left(\mathbf{x}\right)-\widehat{H}\left(\mathbf{x}\right)| \\
&+ \mathbb{E}|H\left(\mathbf{y}\right)-\widehat{H}\left(\mathbf{y}\right)| \\
&+ \mathbb{E}|H\left(\mathbf{x},\mathbf{y}\right)-\widehat{H}\left(\mathbf{x},\mathbf{y}\right)| \qquad (27)
\end{aligned}$$

Using Theorem 2, we see that all three terms on the right hand side in Eq. 27 converge to zero as $N\rightarrow\infty$, therefore $\lim_{N\rightarrow\infty}\mathbb{E}|\widehat{I}\left(\mathbf{x}:\mathbf{y}\right)-I\left(\mathbf{x}:\mathbf{y}\right)|=0$, thus concluding the proof.

7 Experiments

7.1 Implementation Details

Our main computational task is to maximize the local likelihood function in Eq. 8. Since computing the second term on the right hand side of Eq. 8 requires integration that can be time-consuming, we choose the kernel function $K(\cdot)$ to be a Gaussian kernel, $\mathbf{K_H}(\mathbf{t}-\mathbf{x})=\mathcal{N}_d(\mathbf{t};\mathbf{x},\mathbf{H})$ so that the integral can be performed analytically, yielding

$$\int\mathbf{K_H}\left(\mathbf{t}-\mathbf{x}\right)\mathcal{N}_d\left(\mathbf{t};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)d\mathbf{t}=\mathcal{N}_d\left(\mathbf{x};\boldsymbol{\mu},\mathbf{H}+\boldsymbol{\Sigma}\right) \qquad (28)$$

Thus, Eq. 8 reduces to

$$\begin{aligned}
\mathcal{L}\left(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma}\right) &= \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}_d\left(\mathbf{x}_i;\mathbf{x},\mathbf{H}\right)\log\mathcal{N}_d\left(\mathbf{x}_i;\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \\
&- \mathcal{N}_d\left(\mathbf{x};\boldsymbol{\mu},\mathbf{H}+\boldsymbol{\Sigma}\right) \qquad (29)
\end{aligned}$$

Maximizing Eq. 29 is a constrained non-convex optimization problem with the condition that the covariance matrix $\boldsymbol{\Sigma}$ is positive semi-definite. We use Cholesky parameterization to enforce the positive semi-definiteness of $\boldsymbol{\Sigma}$, which allows to reduce our constrained optimization problem into an unconstrained one. Also, since we would like to preserve the local structure of the data, we select the bandwidth to be close to the distance between pair of k-nearest points (averaged over all the points).

We use Newton-Ralphson method to do the maximization although the function itself is not exactly concave. The full algorithm for our estimator is given in Algorithm 1 which takes Algorithm 2 as a subroutine. Note that in Algorithm 2, the Wolfe condition is a set of inequalities in performing quasi-Newton methods (Wolfe, 1969).

Algorithm 1 Mutual Information Estimation with Local Gaussian Approximation

Input: points $(\mathbf{x}, \mathbf{y})^{(1)}, (\mathbf{x}, \mathbf{y})^{(2)}, ..., (\mathbf{x}, \mathbf{y})^{(N)}$
Output: $\widehat{I}(\mathbf{x}; \mathbf{y})$
Calculate entropy $\widehat{H}(\mathbf{x})$ using samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}...,\mathbf{x}^{(N)}$
Calculate entropy $\widehat{H}(\mathbf{y})$ using samples $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}...,\mathbf{y}^{(N)}$
Calculate joint entropy $\widehat{H}(\mathbf{x}, \mathbf{y})$ using input samples $(\mathbf{x}, \mathbf{y})^{(1)}, (\mathbf{x}, \mathbf{y})^{(2)}, ..., (\mathbf{x}, \mathbf{y})^{(N)}$
Return estimated mutual information $\hat{I} = \widehat{H}(\mathbf{x}) + \widehat{H}(\mathbf{y}) - \widehat{H}(\mathbf{x}, \mathbf{y})$

Algorithm 2 Entropy Estimation with Local Gaussian Approximation

Input: points $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, ..., \mathbf{u}^{(N)}$
Output: $\widehat{H}(\mathbf{u})$
Initialize $\widehat{H}(\mathbf{u}) = 0$
for each point $\mathbf{x}^{(i)}$ **do**
  initialize $\boldsymbol{\mu} = \boldsymbol{\mu}_0$, $\boldsymbol{L} = \boldsymbol{L}_0$
  while not $\mathcal{L}(\mathbf{x}^{(i)}, \boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{L} * \boldsymbol{L}^T)$ converge **do**
    Calculate $\mathcal{L}(\mathbf{x}^{(i)}, \boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{L} * \boldsymbol{L}^T)$
    Calculate gradient vector $\mathbf{G}$ of $\mathcal{L}(\mathbf{x}^{(i)}, \boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{L} * \boldsymbol{L}^T)$, with respect to $\boldsymbol{\mu}$, $\boldsymbol{L}$
    Calculate Hessian matrix of $\mathbf{H}$ of $\mathcal{L}(\mathbf{x}^{(i)}, \boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{L} * \boldsymbol{L}^T)$, with respect to $\boldsymbol{\mu}$, $\boldsymbol{L}$
    Do Hessian modification to ensure the positive semi-definiteness of $\mathbf{H}$
    Calculate descent direction $\mathbf{D} = -\alpha \mathbf{H}^{-1}\mathbf{G}$, where we compute α to satisfy Wolfe condition
    Update $\boldsymbol{\mu}, \boldsymbol{L}$ with $(\boldsymbol{\mu}, \boldsymbol{L}) + \mathbf{D}$
  end while
  $\widehat{f}(\mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{L} * \boldsymbol{L}^T)$
  $\widehat{H}(\mathbf{u}) = \widehat{H}(\mathbf{u}) - \frac{\log f(\mathbf{x}^{(i)})}{N}$
end for

In a single step, evaluating the gradient and Hessian in Algorithm 2 would take $O(N)$ time because Eq. 8 is a summation over all the points. However, for points that are far from the current point $\mathbf{x}^{(i)}$, the kernel weight function is very close to zero and we can ignore those point and do the summation only over a local neighborhood of $\mathbf{x}^{(i)}$.

7.2 Experiments with synthetic data

Functional relationships We test our MI estimator for near-functional relationships of form $Y = f(X) + \mathcal{U}(0, \theta)$, where $\mathcal{U}(0, \theta)$ is the uniform distribution over the interval $(0, \theta)$, and X is drawn randomly uniformly from $[0, 1]$. Similar relationships were studied in (Reshef et al., 2011), (Kinney and Atwal, 2014) and (Gao et al., 2015).

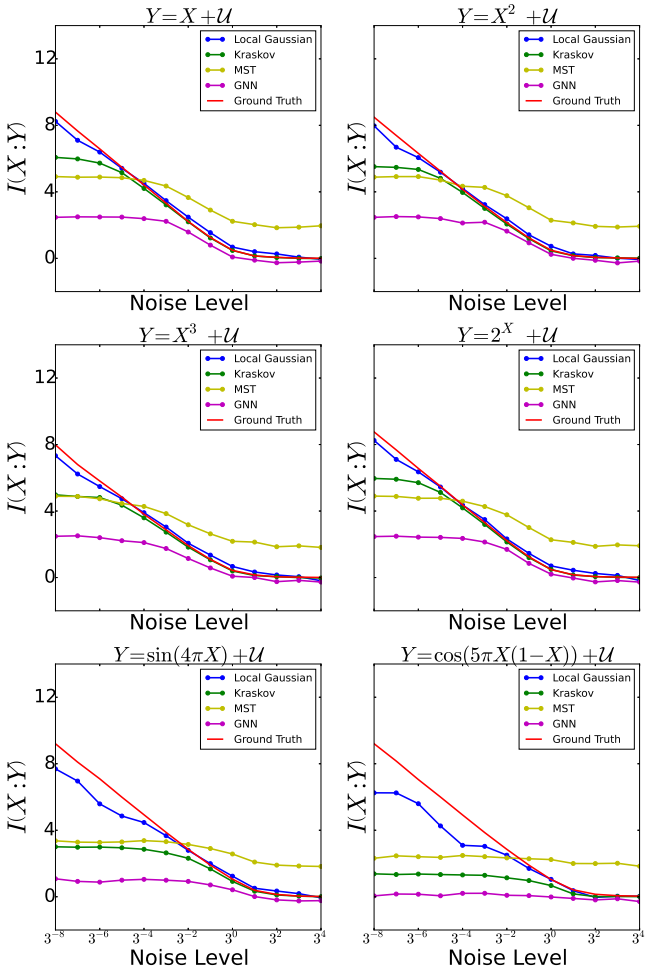


Figure 2: Functional relationship test for mutual information estimators. The horizontal axis is the value of θ which controls the noise level; the vertical axis is the mutual information in nats. For the Kraskov and GNN estimators we used nearest neighbor parameter $k = 5$. For the local Gaussian estimator, we choose the bandwidth to be the distance between a point and its 5rd nearest neighbor.

We compare our estimator to several baselines that include the kNN estimator proposed by (Kraskov et al., 2004), an estimator based on generalized nearest-neighbor graphs (GNN) (Pál et al., 2010), and minimum spanning tree method (MST) (Yukich and Yukich, 1998). We evaluate those estimators for six different functional relationships as indicated in Figure 2. We use $N = 2500$ sample points for each relationship. To speed up the optimization, we limited the summation in Eq. 29 to only k nearest neighbors, thus reducing the computational complexity from $O(N)$ to $O(k)$ in every iteration step of Algorithm 2.

One can see from Fig. 2 that when θ is relatively large, all methods except MST produce accurate estimates of MI. However, as one decreases θ, all three baseline estimators start to significantly underestimate mutual information. In this low-noise regime, our proposed estimator outperforms the baselines, at times by a significant margin. Note also that all the estimators, including ours, perform relatively poorly for highly non-linear relationships (the last row in Figure 2). According to our intuition, this happens when the scale of the non-linearity becomes sufficiently small, so that the linear approximation of the relationship around the local neighborhood of each sample point does not hold. Under this scenario, accuracy can be recovered by adding more samples.

8 Related Work

Mutual Information Estimators Recently, there has been a significant amount of work on estimating information-theoretic quantities such as entropy, mutual information, and divergences, from i.i.d. samples. Methods include k-nearest-neighbors (Singh et al., 2003), (Kraskov et al., 2004), (Pál et al., 2010), (Póczos et al., 2011); minimum spanning trees (Yukich and Yukich, 1998); kernel density estimate (Moon et al., 1995), (Singh and Poczos, 2014); maximum likelihood density ratio (Suzuki et al., 2008); ensemble methods (Moon and Hero, 2014), Sricharan et al. (2013), etc. As pointed our earlier, all of those methods underestimate the mutual information when two variables have strong dependency. (Gao et al., 2015) addressed this shortcoming by introducing a local non-uniformity correction, but their estimator depended on a heuristically defined threshold parameter and lacked performance guarantees.

Density Estimation and Boundary Bias Density estimation is a classic problem in statistics and machine learning. Kernel density estimation and k-nearest-neighbor density estimates are the two most popular and successful non-parametric methods. However, it has been recognized that these non-parametric techniques often suffer from the problem of so-called "boundary bias". Researchers have proposed a variety of methods to overcome the bias, such as the reflection method (Schuster, 1985), (Silverman, 1986); the boundary kernel method (Zhang and Karunamuni, 2000), the transformation method (Marron and Ruppert, 1994), the pseudo-data method (Cowling and Hall, 1996) and others. All these methods are useful in some particular settings. But when it comes to mutual information estimation, how can we choose the most efficient one to use? It seems that local likelihood method (Hjort and Jones, 1996), (Loader, 1996), is a good choice for estimating the mutual information due to its ability to detect the boundary without any prior knowledge. Previous studies have already proven the power of *local regression*, which can automatically overcome the boundary bias. Methods based on *local likelihood* estimation has traditionally attracted less attention due to their computational complexity. However, advances in computational power allow us to re-consider this class of method.

9 Conclusion and Future Work

Past research on mutual information estimation has mostly focused on distinguishing weak dependence from independence. However, in the era of big data, we are often interested in highlighting the strongest dependencies among a large number of variables. When those variables are highly inter-dependent, traditional non-parametric mutual information estimators fail to accurately estimate the value due to the boundary bias.

We have addressed this shortcoming by introducing a novel semi-parametric method for estimating entropy and mutual information based on local Gaussian approximation of the unknown density at the sample points. We demonstrated that the proposed estimators are asymptotically unbiased. We also showed empirically that the proposed estimator has a superior performance compared to a number of popular baseline methods, and can accurately measure strength of the relationship even for strongly dependent variables, and limited number of samples.

There are several potential avenues for future work. First of all, we would like to validate the proposed estimator in higher-dimensional settings. In principle, the approach is general and can be applied in any dimensions. However, the optimization procedure may be computational expensive in higher dimensions, since the number of parameters scales as $O(d^2)$ with dimensionality d. An intuitive solution would be to initialize the parameters with the results obtained from the close points, which can facilitate convergence.

Another interesting issue is the bandwidth selection, which is an important problem in general density estimation problems. If the bandwidth is too large, the local Gaussian assumption may not be valid, whereas very small bandwidth will result in non-smooth densities. Ideally, we would like to choose the bandwidth in a way that preserves the local Gaussian structure in the neighborhood of each point. Another interesting extension would be choosing the bandwidth adaptively for each point.

Finally, while here we have focused on the asymptotic unbiasedness of the proposed estimator, it will be very valuable to establish theoretical results about the convergence rates of the estimators, as well as its variance in the large sample limit.

Acknowledgements

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

References

Ann Cowling and Peter Hall. On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 551–563, 1996.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS'15*, 2015.

NL Hjort and MC Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

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`.

Clive R Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996.

James Stephen Marron and David Ruppert. Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 653–671, 1994.

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.

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318–2321, 1995.

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.

Barnabás Póczos, Liang Xiong, and Jeff Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2011.

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.

Eugene F Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and methods*, 14(5):1123–1136, 1985.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379423, 1948.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

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`.

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.

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.

Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

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

Shunpu Zhang and Rohana J Karunamuni. On nonparametric density estimation at the boundary*. *Journal of nonparametric statistics*, 12(2):197–221, 2000.