Supplementary Material for "Efficient Estimation of Mutual Information for Strongly Dependent Variables"

A Proof of Theorem 2

Notice that for a fixed sample point $\mathbf{x}^{(i)}$, its k-nearest-neighbor distance $r_k\left(\mathbf{x}^{(i)}\right)$ is always equal to or larger than the k-nearest-neighbor distance of at the same point $\mathbf{x}^{(i)}$ projected into a sub-dimension j, i.e., for any i, j, we have

$$r_k\left(\mathbf{x}^{(i)}\right) \geqslant r_k\left(\mathbf{x}_j^{(i)}\right) \tag{A.1}$$

Using Eq. A.1, we get the upper bound of $\widehat{I}_{kNN,k}(\mathbf{x})$ as follows:

$$\begin{aligned}
\widehat{I}_{kNN,k}(\mathbf{x}) &= \widehat{I}'_{kNN,k}(\mathbf{x}) - (d-1)\gamma_k \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\frac{\widehat{p}_k\left(\mathbf{x}^{(i)}\right)}{\prod\limits_{j=1}^{d}\widehat{p}_k\left(\mathbf{x}_j^{(i)}\right)} - (d-1)\gamma_k \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\frac{\frac{k}{N-1}\frac{\Gamma(d/2)+1}{\pi^{d/2}}r_k\left(\mathbf{x}^{(i)}\right)^{-d}}{\prod\limits_{j=1}^{d}\frac{k}{N-1}\frac{\Gamma(1/2)+1}{\pi^{1/2}}r_k\left(\mathbf{x}_j^{(i)}\right)^{-1}} \\
&\quad -(d-1)\gamma_k \\
&\le (d-1)\log\left(\frac{N-1}{k}\right) + \log\frac{\Gamma(d/2)+1}{\left(\Gamma(1/2)+1\right)^d} \\
&\quad -(d-1)\left(\psi(k)-\log k\right) \\
&\le (d-1)\log\left(\frac{N-1}{k}\right) + \log\frac{\Gamma(d/2)+1}{\left(\Gamma(1/2)+1\right)^d} \\
&\quad -(d-1)\left(\psi(1)-\log 1\right)
\end{aligned} \tag{A.2}$$

The last inequality is obtained by noticing that $\psi(k) - \log(k)$ is a monotonous decreasing function.

Also, we have,

$$\begin{aligned}
\log\frac{\Gamma(d/2)+1}{\left(\Gamma(1/2)+1\right)^d} &= \log\left(\Gamma(d/2)+1\right) - d\log\left(\Gamma(d/2)+1\right) \\
&< \log\left(\sqrt{2\pi}\left(\frac{d/2+1/2}{e}\right)^{d/2+1/2}\right) \\
&\quad -d\log\left(\pi^{\frac{1}{2}}+1\right) \\
&= O\left(d\log d\right)
\end{aligned}$$

The inequality above is obtained by using the bound of gamma function that,

$$\Gamma(x+1) < \sqrt{2\pi}\left(\frac{x+1/2}{e}\right)^{x+1/2}$$

Therefore, reconsidering A.2, we get the following inequality for $\widehat{I}_{kNN,k}(\mathbf{x})$:

$$\begin{aligned}
\widehat{I}_{kNN,k}(\mathbf{x}) \le &\quad (d-1)\log\left(\frac{N-1}{k}\right) + O\left(d\log d\right) \\
\le &\quad (d-1)\log\left(N-1\right) + O\left(d\log d\right)
\end{aligned}$$

Requiring that $|\widehat{I}_{kNN,k}(\mathbf{x}) - I(\mathbf{x})| \le \varepsilon$, we obtain,

$$N \ge C\exp\left(\frac{I(\mathbf{x})-\epsilon}{d-1}\right) + 1 \tag{A.3}$$

where C is a constant which scales like $O(\frac{1}{d})$. □

B Derivation of Eq. 17

The naive kNN or KSG estimator can be written as:

$$\widehat{I}_k(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\log\frac{\frac{P\left(\mathbf{x}^{(i)}\right)}{V(i)}}{\prod\limits_{j=1}^{d}\frac{P\left(\mathbf{x}_j^{(i)}\right)}{V_j(i)}} \tag{B.1}$$

where $P(\mathbf{x}^{(i)})$ is the probability mass around the k-nearest-neighborhood at $\mathbf{x}^{(i)}$ and $P(\mathbf{x}_j^{(i)})$ is the probability mass around the k-nearest-neighborhood (or $n_{x_j}(i)$-nearest-neighborhood for KSG) at $\mathbf{x}^{(i)}$ projected into j-th dimension. Also, $V(i)$ and $V_j(i)$ denote the volume of the kNN ball(or hype-rectangle in KSG) in the joint space and projected subspaces respectively.

Now our local nonuniform correction method replaces the volume $V(i)$ in Eq. B.1 with the corrected volume $\overline{V}(i)$, thus, our estimator is obtained as follows:

$$\begin{aligned}
\widehat{I}_{LNC,k}(\mathbf{x}) &= \frac{1}{N}\sum_{i=1}^{N}\log\frac{\frac{P\left(\mathbf{x}^{(i)}\right)}{\overline{V}(i)}}{\prod\limits_{j=1}^{d}\frac{P\left(\mathbf{x}_j^{(i)}\right)}{V_j(i)}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\log\frac{\frac{P\left(\mathbf{x}^{(i)}\right)}{V(i)}\times\frac{V(i)}{\overline{V}(i)}}{\prod\limits_{j=1}^{d}\frac{P\left(\mathbf{x}_j^{(i)}\right)}{V_j(i)}} \\
&= \widehat{I}_k(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N}\log\frac{\overline{V}(i)}{V(i)}
\end{aligned} \tag{B.2}$$

C Empirical Evaluation for $\alpha_{k,d}$

Suppose we have a uniform distribution on the d dimensional (hyper)rectangle with volume V. We sample k points from this uniform distribution. We perform PCA using these k points to get a new basis[10].

[10] In practice, we recommend k to be larger than $2 * d$.

After rotating into this new basis, we find the volume, $\bar{V}$, of the smallest rectilinear rectangle containing the points. By chance, we will typically find $\bar{V} < V$, even though the distribution is uniform. This will lead to us to (incorrectly) apply a local non-uniformity correction. Instead, we set a threshold $\alpha_{k,d}$ and if $\bar{V}/V$ is above the threshold, we assume that the distribution is locally uniform. Setting α involves a trade-off. If it is set too high, we will incorrectly conclude there is local non-uniformity and therefore over-estimate the mutual information. If we set α too low, we will lose statistical power for "medium-strength" relationships (though very strong relationships will still lead to values of $\bar{V}/V$ smaller than α).

In practice, we determine the correct value of $\alpha_{k,d}$ empirically. We look at the probability distribution of $\bar{V}/V$ that occurs when the true distribution is uniform. We set α conservatively so that when the true distribution is uniform, our criteria rejects this hypothesis with small probability, ϵ. Specifically, we do a number of trials, N, and set $\hat{\alpha}_{k,d}$ such that $\sum_{i=1}^{N} \mathbf{I}\left(\frac{\bar{V_i}}{V_i} < \hat{\alpha}_{k,d}\right)/N < \epsilon$ where ϵ is a relatively small value. In practice, we chose $\epsilon = 5 \times 10^{-3}$ and $N = 5 \times 10^5$. The following algorithm describes this procedure:

Algorithm C.1 Estimating $\alpha_{k,d}$ for LNC

```
Input: parameter $d$ (dimension), $k$ (nearest neighbor), $N$, $\epsilon$
Output: $\hat{\alpha}_{k,d}$
set array a to be NULL
repeat
    Randomly choose a uniform distribution supported on $d$ dimensional (hyper) rectangle, denote its volume to be $V$
    Draw $k$ points from this uniform distribution, get the correcting volume $\bar{V}$ after doing PCA
    add the ratio $\frac{\bar{V}}{V}$ to array a
until above procedure repeated $N$ times
$\hat{\alpha}_{k,d} \leftarrow \lceil \epsilon N \rceil$ th smallest number in a
```

Figure C.1 shows empirical value of $\hat{\alpha}_{k,d}$ for different (k, d) pairs. We can see that for a fixed dimension d, $\hat{\alpha}_{k,d}$ grows as k increases, meaning that $\bar{V}$ must be closer to V to accept the null hypothesis of *uniformity*. We also find that $\hat{\alpha}_{k,d}$ decreases as the dimension d increases, indicating that for a fixed k, $\bar{V}$ becomes much smaller than V when points are drawn from a uniform distribution in higher dimensions.

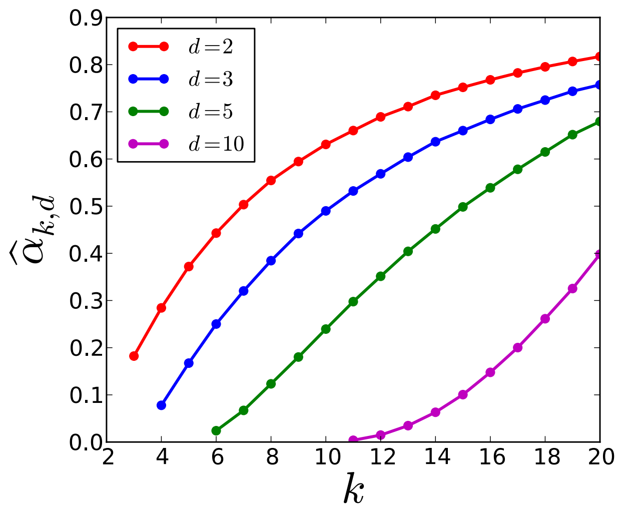

Figure C.1: $\widehat{\alpha}_{k,d}$ as a function of k. k ranges over $[d, 20]$ for each dimension d.

D More Functional Relationship Tests in Two Dimensions

We have tested together twenty-one functional relationships described in Reshef et al. (2011); Kinney and Atwal (2014), we show six of them in Section 5. The complete results are shown in Figure D.1. Detailed description of the functions can be found in Table S1 of Supporting Information in Kinney and Atwal (2014).

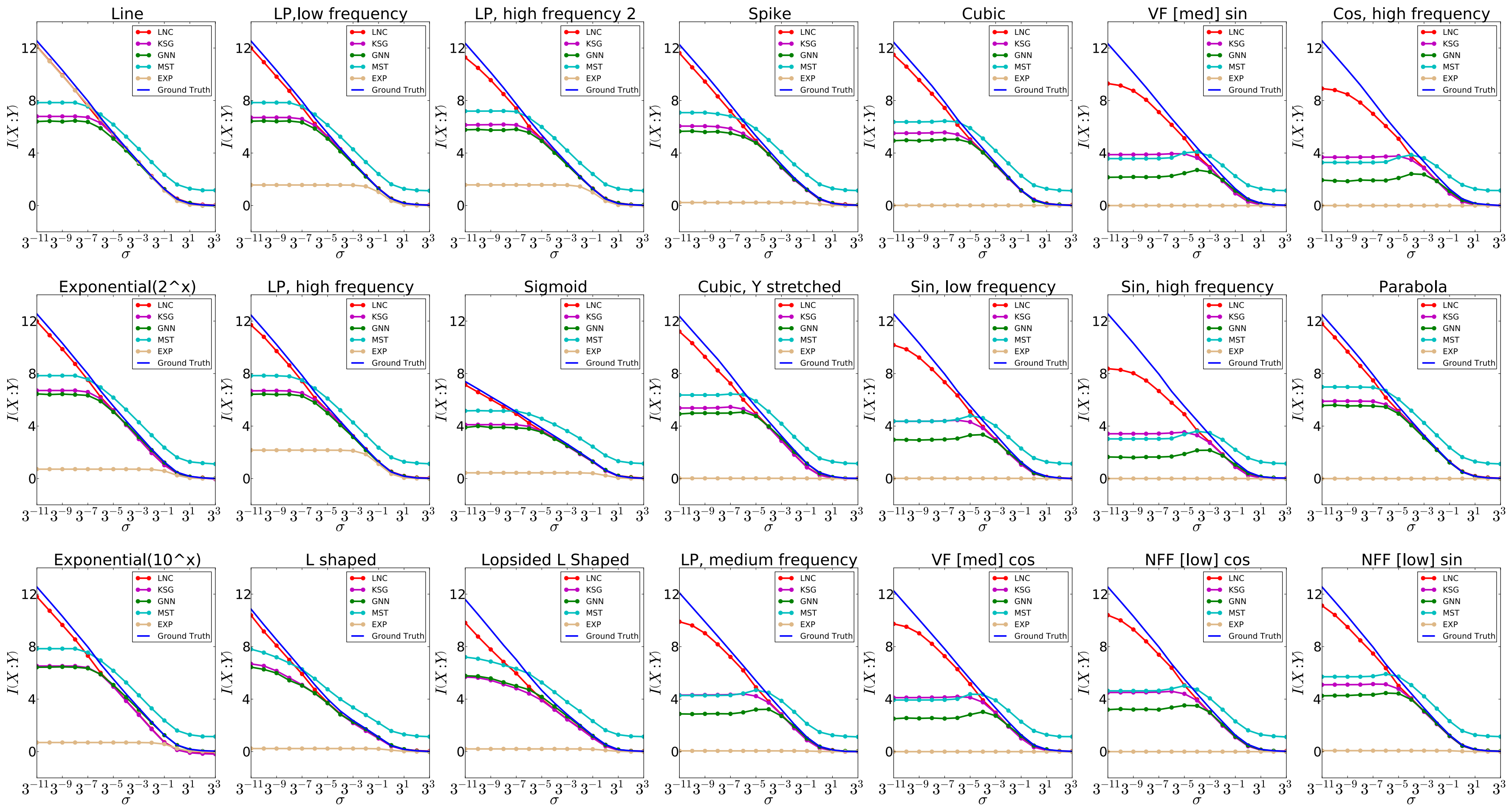

Figure D.1: Mutual Information tests of LNC, KSG, GNN, MST, EXP estimators. Twenty-one functional relationships with different noise intensities are tested. Noise has the form $U[-\sigma/2, \sigma/2]$where σ varies(as shown in X axis of the plots). For KSG, GNN and LNC estimators, nearest neighbor parameter $k = 5$. We are using $N = 5,000$ data points for each noisy functional relationship.