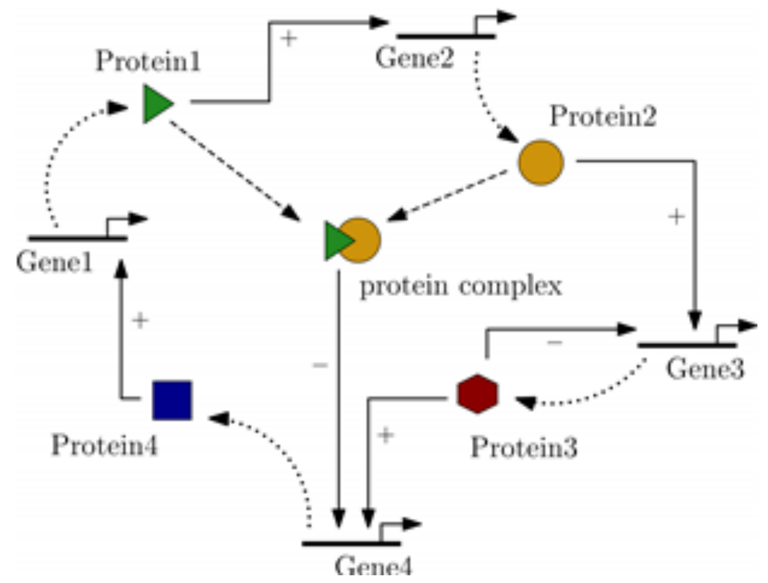
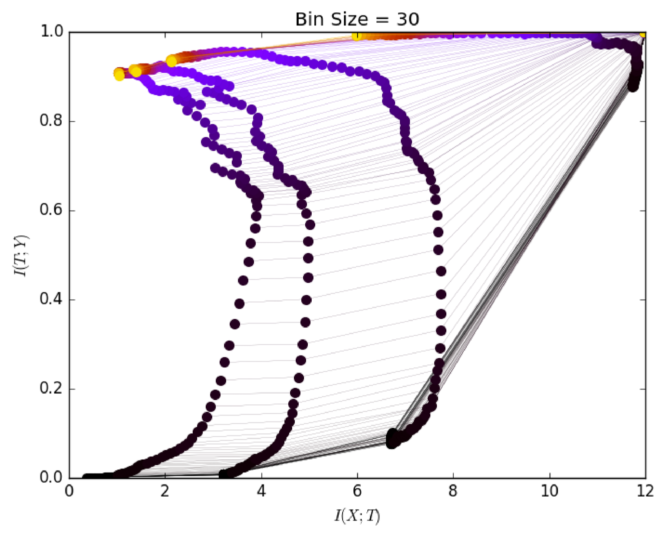
Estimating Mutual Information for Discrete-Continuous Mixtures

Weihao Gao†, Sreeram Kannan*, Sewoong Oh†, Pramod Viswanath†
†University of Illinois at Urbana-Champaign and *University of Washington

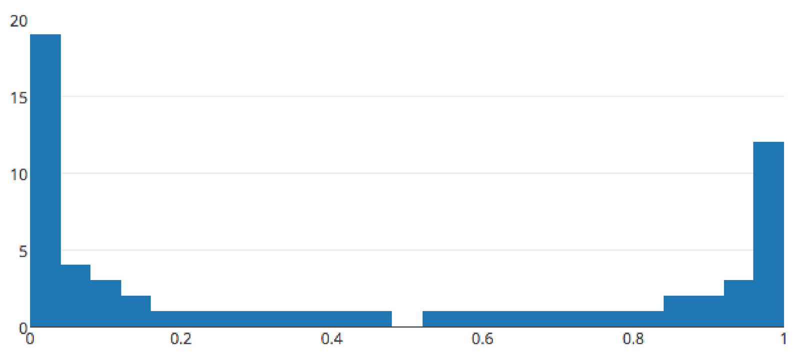
Motivation

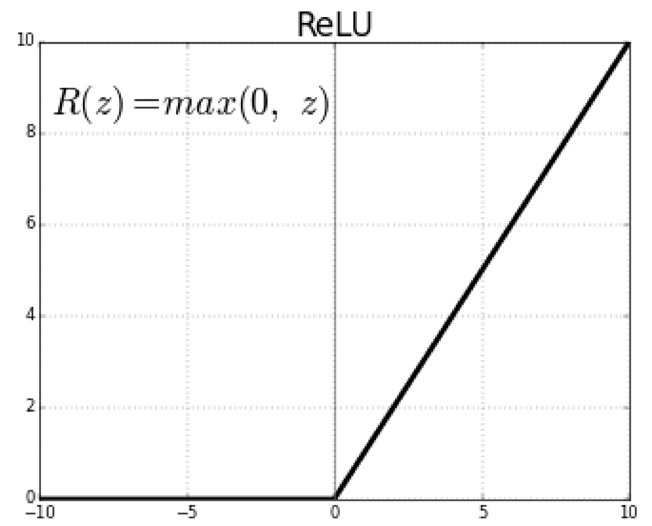
- Estimate Mutual Information $I(X;Y)$ from i.i.d. samples.
- Applications: Gene Network Inference, Information Bottleneck for Deep Neural Nets, etc.





- Calculating mutual information between genes FMR1 and UTP3 in myogenesis, answer is negative (even with many data points). Why?
- Reason: Histogram of FMR1 suggests a mixed distribution.





- Another Example: Computing information bottleneck for neural nets with ReLU, neurons admits mixed distribution.
- Goal: develop an algorithm to estimate mutual information for discrete-continuous mixtures.

General Definition of Mutual Information

- Let P_{XY} be a probability measure on the space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are both Euclidean spaces. For any measurable set $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, define $P_X(A) = P_{XY}(A \times \mathcal{Y})$ and $P_Y(B) = P_{XY}(\mathcal{X} \times B)$. Let $P_X P_Y$ be the product measure $P_X \times P_Y$. Then the mutual information $I(X;Y)$ of P_{XY} is defined as

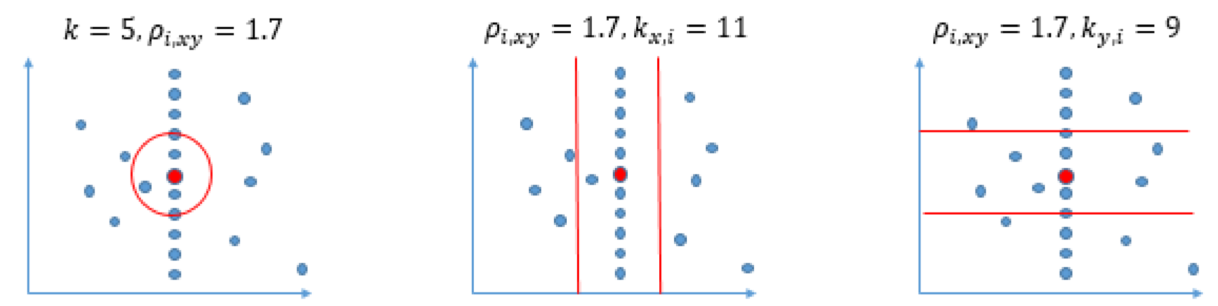
$$I(X;Y) \equiv \int_{\mathcal{X}\times\mathcal{Y}} \log \frac{dP_{XY}}{dP_X P_Y} dP_{XY},$$

where $\frac{dP_{XY}}{dP_X P_Y}$ is the Radon-Nikodym derivative.
- **Theorem.** Under certain assumptions, P_{XY} is absolutely continuous w.r.t. $P_X P_Y$, hence mutual information is well-defined for any P_{XY}.

Estimator

- Challenge: want to estimate $I(X;Y)$ in the following cases.
  - X is discrete, Y is continuous (or vice versa).
  - X or Y has many components, some are discrete, some are continuous.
  - X or Y is a mixture of continuous and discrete distributions.
  - Any mixture of above cases.
- Previous works focus on either purely discrete or purely continuous.
  - Discrete data — plug-in estimator.
  - Continuous data — k-nearest neighbor methods.
- KSG estimator for continuous data: Let $\rho_{i,xy}$ be the distance of the k-NN for (X_i, Y_i) tuple. $k_{x,i}$ is the number of samples of X_i within distance $\rho_{i,xy}$. Similarly $k_{y,i}$.

$$\widehat{I}_{\text{KSG}}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} \{\psi(k) + \log(n) - \log(k_{x,i}) - \log(k_{y,i})\}$$



- **Proposed** estimator for **Mixture** of data:

$$\widehat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} \left\{\psi(\tilde{k}_i) + \log(n) - \log(k_{x,i}) - \log(k_{y,i})\right\}$$

where $\tilde{k}_i = k$ for continuous point, $\tilde{k}_i$ = number of times that particular point was seen for discrete point.

Theorems

The proposed estimator is ℓ_2 **consistent**, i.e.

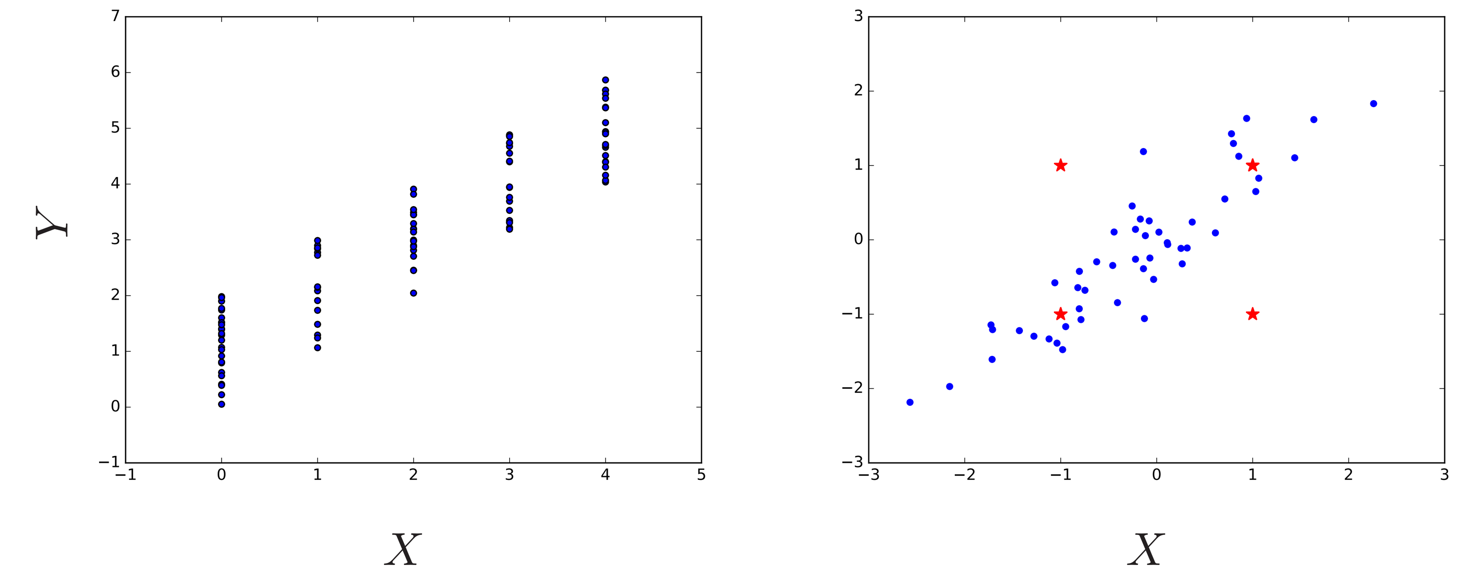
$$\lim_{n\to\infty} \mathbb{E}\left[\left(\widehat{I}(X;Y) - I(X;Y)\right)^2\right] = 0,$$

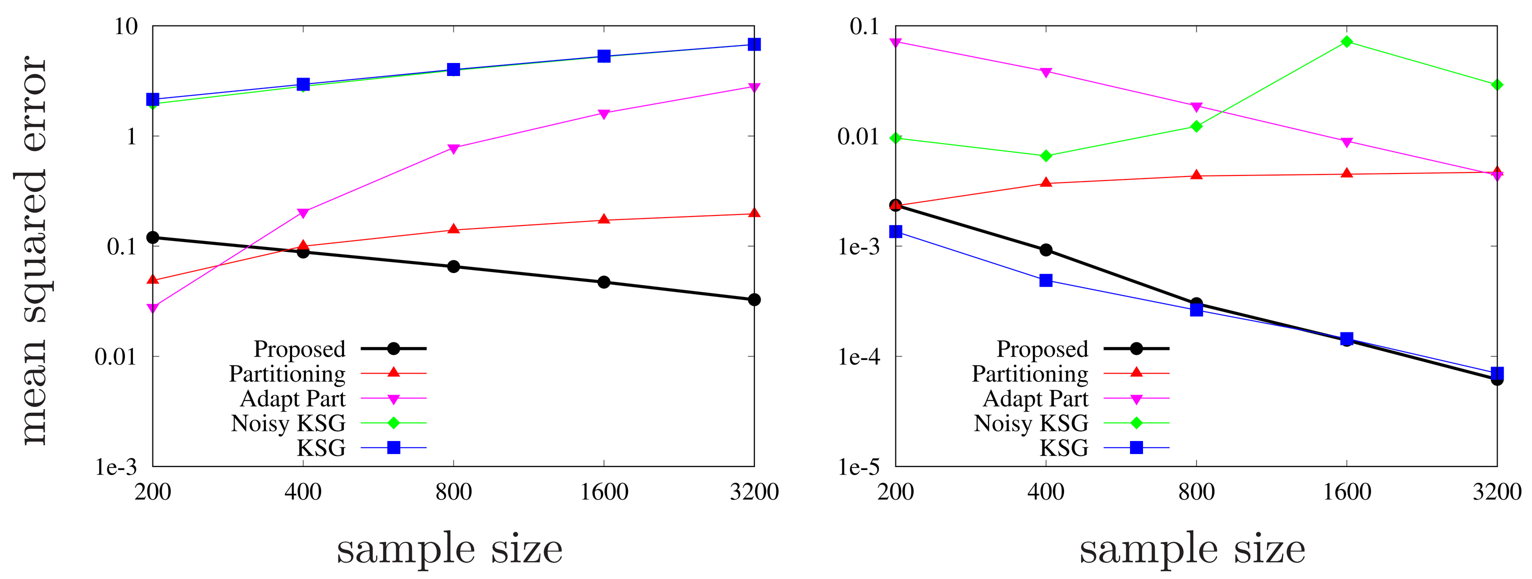
under the following assumptions:

1. k is a function of n such that $k_n \to \infty$ and $(k_n \log n)^2/n \to 0$.
2. The set of discrete points $\{(x,y) : P_{XY}(x,y,0) > 0\}$ is finite.
3. $\frac{P_{XY}(x,y,r)}{P_X(x,r)P_Y(y,r)}$ converges to $f(x,y)$ as $r \to 0$ and $f(x,y) \le C$ with prob. 1.
4. $\mathcal{X} \times \mathcal{Y}$ can be decomposed into countable disjoint sets $\{E_i\}_{i=1}^{\infty}$ such that $f(x,y)$ is uniformly continuous on each E_i.
5. $\int_{\mathcal{X}\times\mathcal{Y}} \left|\log f(x,y)\right| dP_{XY} < +\infty$.

Experiments

1. Synthetic Experiments.
   - **Left**: X is discrete and Y is continuous.
   - **Right**: (X,Y) is a mixture of continuous (blue) and discrete (red stars) distributions.
   - Plot mean squared error v.s. sample size for **proposed** estimator and other estimators.





2. Gene Network Inference.
   - 20 genes, connect with each other through a network
   - If $I(\text{Gene}_i, \text{Gene}_j)$ greater than a threshold, we claim that there are connected.
   - Dropout: $\text{Gene}_i = 0$ with probability p.
   - Plot Area Under ROC v.s. level of dropout p.