Inductive Mutual Information Estimation: A Convex Maximum-Entropy Copula Approach

Yves-Laurent Kom Samo

KXY Technologies Inc.
ml@kxy.ai @Dr_YLKS @Dr_YLKS

March 20, 2021

Overview

Motivation

Whenever there is **structure in data**, there is **mutual information**. Yet mutual information (MI) estimation is a notoriously hard problem to solve (McAllester and Stratos (2020)).

- Mutual information (MI) is only a weak property of a joint distribution $P_{x,y}$.
- But virtually all MI estimators assume that we have enough data to fully characterize $P_{x,y}$.
- As a result, existing MI estimators are data-inefficient.

Contributions

Our estimator (MIND) is the only MI estimator that works for small and large sample sizes and/or input dimensions.

Key Properties:

- Data-efficiency and low-variance
- Consistency
- Marginal (or scale) invariance

The MIND Estimator

- Define $\boldsymbol{u}_z = (F_1(z_1), \ldots, F_d(z_d))$, where F_i is the CDF of z_i. u_z has uniform marginals and has CDF the copula of $\boldsymbol{z} = (z_i, \ldots, z_d)$. $\boldsymbol{u}_z$ is $\boldsymbol{z}$'s copula-uniform dual representation.
- $I(\boldsymbol{y}; \boldsymbol{x}) = I(\boldsymbol{u}_y; \boldsymbol{u}_x) = h(\boldsymbol{u}_y) + h(\boldsymbol{u}_x) - h(\boldsymbol{u}_y, \boldsymbol{u}_x)$.
- Estimate $h(\boldsymbol{u}_z)$ as the solution to the max-ent problem:

$$\begin{cases} \max\limits_{P \in \mathcal{D}_d} & h(P) \\ \text{s.t. } E_P[\phi_m(\boldsymbol{u})] = \alpha_m := E_{P_{\boldsymbol{u}_z}}[\phi_m(\boldsymbol{u})] \end{cases}, \qquad \text{(A-MIND)}$$

  over the space $\mathcal{D}_d$ of continuous distribution on $[0,1]^d$.
- ϕ_m is chosen so that $E_P(\phi_m(\boldsymbol{u}))$ reveals association between coordinates of $\boldsymbol{u}$. E.g. $E_P(\phi_s(u,v))$ is the Spearman rank correlation between u and v when $\phi_s(u,v) = 12uv - 3$.

Solving the Maximum-Entropy Problem

- **Solution:** $h_{AM}(\boldsymbol{u}_z) := -\boldsymbol{\alpha}_m^T \boldsymbol{\theta}^*$, with $\boldsymbol{\theta}^*$ solution to the CVX problem

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\boldsymbol{\alpha}_m^T \boldsymbol{\theta} + \int_{[0,1]^d} e^{\boldsymbol{\theta}^T \phi_m(\boldsymbol{u})} d\boldsymbol{u}.$$

- **Copula Entropy Error**:

$$h_{AM}(\boldsymbol{u}_z) - h(\boldsymbol{u}_z) = KL\left(P_{\boldsymbol{u}_z} || P_{AM}(\boldsymbol{u}_z)\right).$$

- **Mutual Information Error:**

$$\begin{aligned} I_{AM}(\boldsymbol{y};\boldsymbol{x}) - I(\boldsymbol{y};\boldsymbol{x}) = & KL\left(P_{\boldsymbol{u}_y} || P_{AM}(\boldsymbol{u}_y)\right) + KL\left(P_{\boldsymbol{u}_x} || P_{AM}(\boldsymbol{u}_x)\right) \\ & - KL\left(P_{\boldsymbol{u}_{x,y}} || P_{AM}(\boldsymbol{u}_{x,y})\right). \end{aligned}$$

Can be null even when m is small!

Sample Estimator Properties

- **Constraint Estimator:** $\hat{\alpha}_{m,n} = \frac{1}{n}\sum_{i=1}^{n} \phi_m\left(\frac{\mathrm{rg}(\boldsymbol{z}_i)}{n+1}\right)$, where $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ are i.i.d samples and $\mathrm{rg}(\boldsymbol{z}_i)$ are coordinatewise ranks. $\hat{\alpha}_{m,n}$ is consistent and has $O(1/n)$ MSE rate.
- **Consistency**: Overall, MIND is a consistent (in m and n) estimator of $I(\boldsymbol{y}; \boldsymbol{x})$ so long as $(\phi_m)_m$ are universal approximators of continuous functions on $[0,1]^d$ for any d (e.g. polynomials of degree m).
- **Data-Effiency:** We may achieve perfect estimation even for a small m (Corollary 3.1) and MSE rate is $O(1/n)$.

Synthetic Benchmark

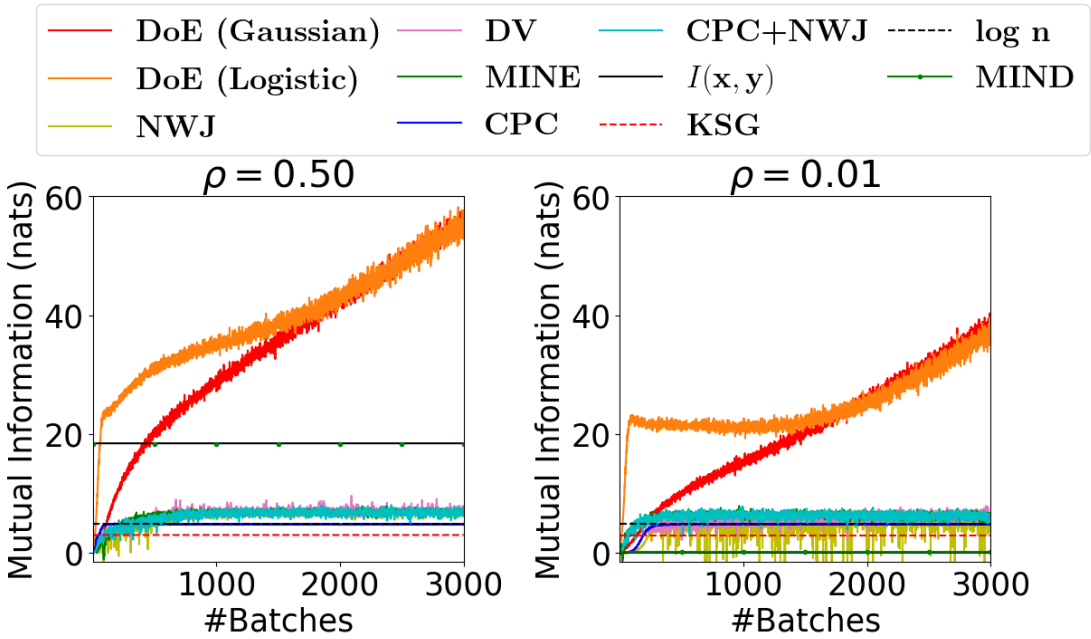

Figure: Estimation of the mutual information between two 128-dimensional vectors $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{y} = (y_1, \ldots, y_d)$ from a draw of 1000 i.i.d. samples. (x_i, y_i) are i.i.d. Gaussians with mean zero, unit marginal variance, and correlation ρ.

Real-Life Application

$I(\boldsymbol{y};\boldsymbol{x})$	$\bar{R}^2$	$R\bar{M}SE$	d	n
1.50	0.95	18,531	80	1460

Table: Mutual information and highest performances achievable in the 'House Prices: Advanced Regression Techniques' Kaggle challenge.

$$\bar{R}^2\left(P_{\boldsymbol{y},\boldsymbol{x}}\right) = 1 - e^{-2I(\boldsymbol{y};\boldsymbol{x})}$$

$$\mathrm{R\bar{M}SE}\left(P_{\boldsymbol{y},\boldsymbol{x}}\right) = e^{-I(\boldsymbol{y};\boldsymbol{x})}\sqrt{\mathrm{Var}\left(\boldsymbol{y}\right)}$$

Reference

Reference

- Csiszar, I. (1975). "I-divergence geometry of probability distributions and minimization problems." In The Annals of Probability (Ann. Probab.).
- Belghazi, M. I. et al. (2018). "Mutual Information Neural Estimation." In International Conference on Machine Learning (ICML).
- McAllester, D. Stratos, K. (2020). "Formal limitations on the measurement of mutual information." In International Conference on Artificial Intelligence and Statistics (AISTATS).

Code: `https://github.com/kxytechnologies/kxy-python`