INSTITUT FÜR OPTISCHE SYSTEME

Hochschule Konstanz
Technik, Wirtschaft und Gestaltung

Brownbag on Neural Mutual Information estimation

Demystifying the "**Correlation of the** 21st **Century**." Terry Speed.

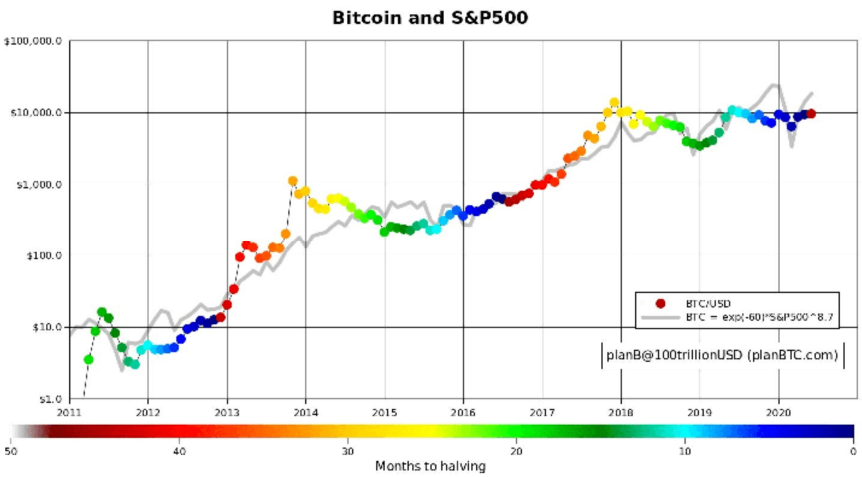
Matthias Hermann, HTWG-Konstanz, Institute for Optical Systems

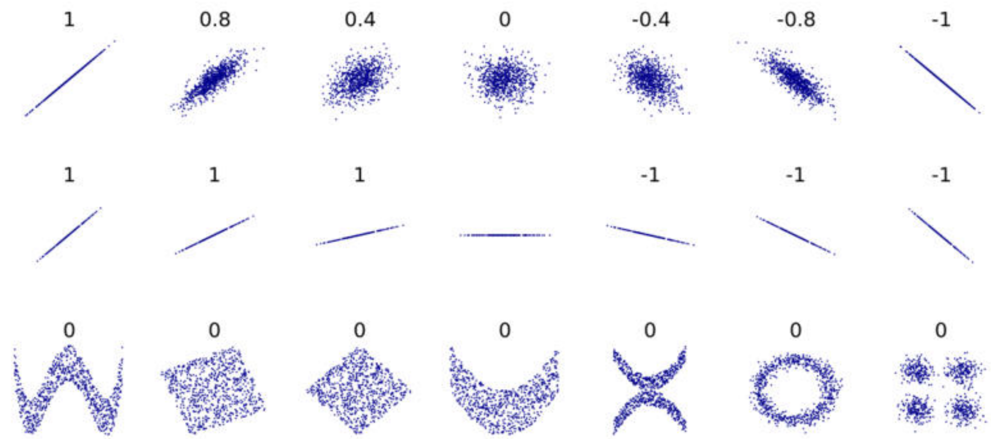
9.11.2021

Agenda

- What is mutual information?
- Estimating mutual information
  - Histogram
  - Kruskov
- Mutual Information Neural estimation (MINE)
- Application to sensor registration

Linear dependence





Bitcoin X and S&P500 Y is highly correlated.

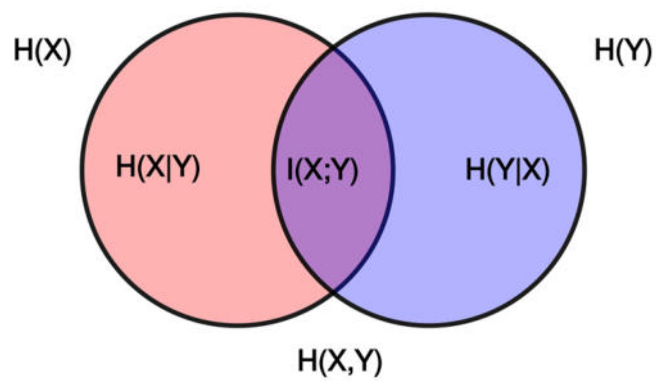
Pearson correlation coefficient measures linear relationships between X and Y. But fails with slope and non-linear relationships.

Non-linear dependence

- Mutual Information is given by

$$I(\boldsymbol{X}, \boldsymbol{Y}) = H[\boldsymbol{X}] - H[\boldsymbol{X}|\boldsymbol{Y}].$$

- Specifies "how much (in bits) do we know about **X** given **Y**?".



Capturing non-linear dependencies

Given random variables **X** and **Y**, and function f

$$Y = f(X) + \sigma\epsilon$$

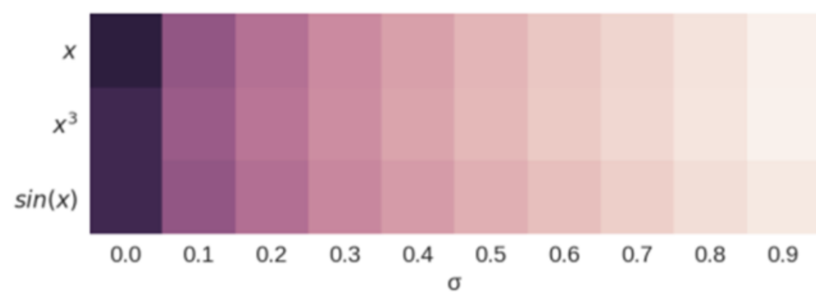
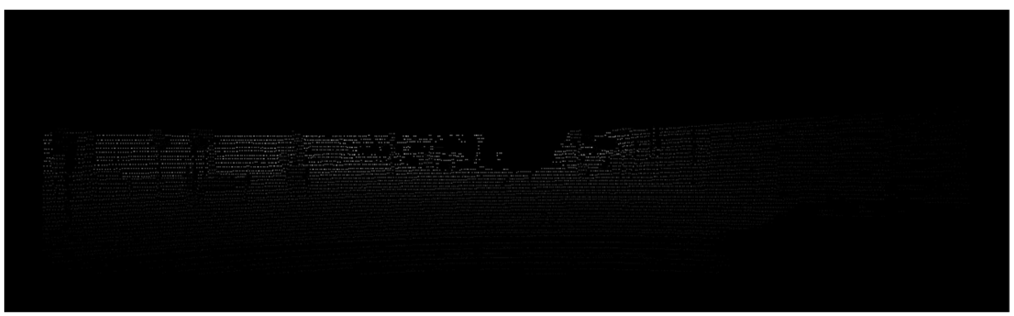


Fig. 1. Mutual information $I(\boldsymbol{X}, \boldsymbol{Y})$ measures the dependency of **X** and **Y** and is invariant to the deterministic nonlinear transformation f ("equitability").

MI is defined for arbitrary variables

- Given random variables **ProjectedLidarPoints** and **CameraImage**

$$I(\textbf{ProjectedLidarPoints}, \textbf{CameraImage}) \approx 20.89\ bits.$$



Estimating mutual information

- Definition

$$I(X,Y) = KL(P_{XY}||P_X \otimes P_Y),$$

where P_{XY} is the joint distribution and $P_X \otimes P_Y$ is the product of their marginals.

$$I(X,Y) = \underbrace{\mathrm{H}_{P_{XY}}[P_X \otimes P_Y]}_{\textit{cross-entropy}} - \underbrace{\mathrm{H}[P_{XY}]}_{\textit{joint-entropy}} = \underbrace{\mathrm{H}[P_X]}_{\textit{entropy}} + \underbrace{\mathrm{H}[P_Y]}_{\textit{entropy}} - \underbrace{\mathrm{H}[P_{XY}]}_{\textit{joint-entropy}}$$

- Naive "Histogram" approach:

$$I(X,Y) \approx I_{binned}(X,Y) = \Sigma_{y \in \boldsymbol{Y}} \Sigma_{x \in \boldsymbol{X}}\, p_{(\boldsymbol{X},\boldsymbol{Y})}(x,y) \log \frac{p_{(\boldsymbol{X},\boldsymbol{Y})}(x,y)}{p_{\boldsymbol{X}}(x) p_Y(y)}$$

- Can be problematic with empty bins!

Estimating mutual information

- Improving estimation by k-nearest neighbor statistics [2]:

$$I(X,Y) = \underbrace{\mathrm{H}[P_X]}_{\textit{entropy}} + \underbrace{\mathrm{H}[P_Y]}_{\textit{entropy}} - \underbrace{\mathrm{H}[P_{XY}]}_{\textit{joint-entropy}}$$

$$I(X,Y) \approx I_{knn}(X,Y) = \widehat{\mathrm{H}}[P_X] + \widehat{\mathrm{H}}[P_Y] - \widehat{\mathrm{H}}[P_{XY}]$$

$$\widehat{\mathrm{H}}[X] = -\underbrace{\psi(k)}_{\textit{digamma}} + \underbrace{\psi(N)}_{\textit{digamma}} + \log \underbrace{c_d}_{\textit{volume d-ball}} + \frac{d}{N}\sum_{i}^{N} \log \underbrace{\epsilon(i)}_{\textit{distance to k-th neighbor}}$$

Estimating mutual information

- K-nearest neighbor method is problematic for estimating joint entropy

$$\widehat{\mathrm{H}}[P_{XY}]$$

as choosing the same **k** and computing

$$I_{knn}(X,Y) = \widehat{\mathrm{H}}[P_X] + \widehat{\mathrm{H}}[P_Y] - \widehat{\mathrm{H}}[P_{XY}]$$

would effectively use different scales for the joint and marginal space.

- Kraskov's method [2] corrects for that by choosing **k** dynamically!

Estimating mutual information in high dimensions

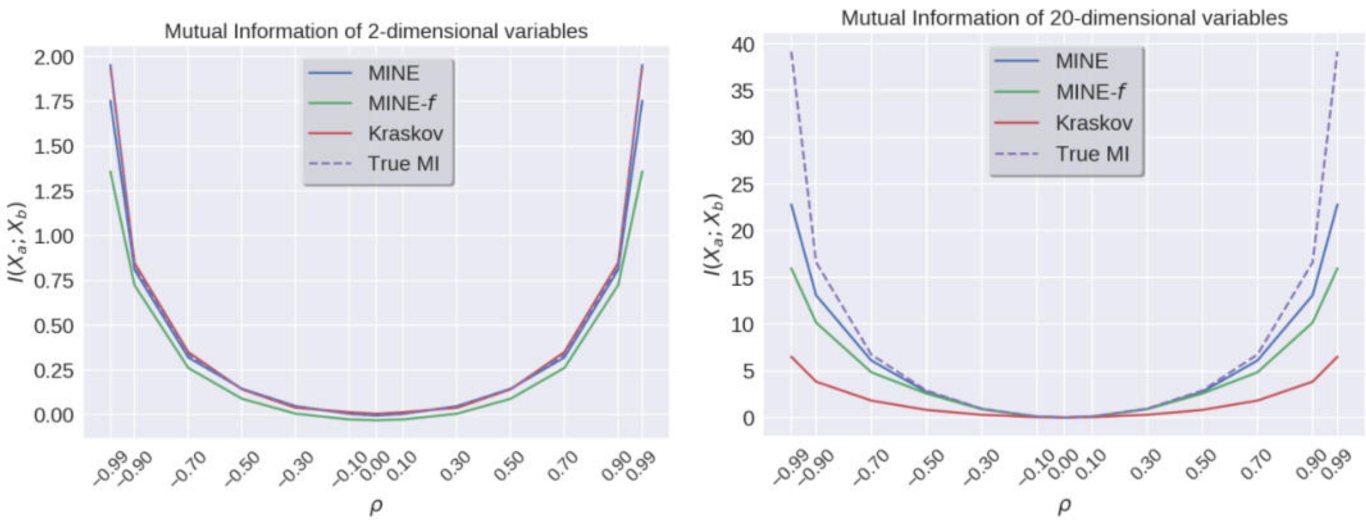


Fig. 2. Mutual information between two multivariate Gaussians with component-wise correlation $\rho \in (-1, 1)$ [1]. Kraskov's estimator [2] underestimates the true MI in high dimensions.

MINE - Neural mutual information estimation

- Maximizing the Donsker-Varadhan (DV) lower bound of the KL divergence

$$I(X,Y) = KL(P_{XY}||P_X \otimes P_Y)$$

$$\geq \sup_{f_\theta} \mathrm{DV}(J, M; f_\theta) = E_{x\sim J}[f_\theta(x)] - \log\left(E_{y\sim M}\left[e^{f_\theta(y)}\right]\right),$$

with $J = P_{XY}$ and $\mathrm{M} = P_X \otimes P_Y$ and function space $f \in F$.

- In the case of MINE [1], f_θ is a neural network and $\sup_{f_\theta} DV(\cdot; f_\theta)$ is computed by standard gradient ascent!

- The Auxilary dataset $\mathrm{M} = (X, Y^*)$ is constructed by sampling y^* without replacement (shuffling).

MINE - Neural mutual information estimation

- Maximizing the Donsker-Varadhan (DV) lower bound of the KL divergence

$$I(X,Y) = KL(P_{XY} || P_X \otimes P_Y)$$

$$\geq \sup_{f_\theta} \mathrm{DV}(J, M; f_\theta) = E_{(x,y)\sim J}\left[f_\theta\left((x,y)\right)\right] - \log\left(E_{(x,y^*)\sim M}\left[e^{f_\theta\left((x,y^*)\right)}\right]\right),$$

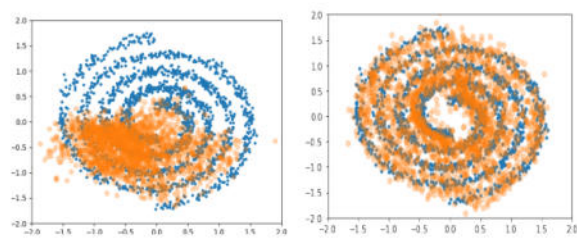
with $J = P_{XY}$ and $\mathrm{M} = P_X \otimes P_Y$ and function space $f \in F$.

- In the case of MINE [1], f_θ is a neural network and $\sup_{f_\theta} DV(\cdot; f_\theta)$ is computed by back-prob and standard gradient ascent!

MINE – a proxy “classification” problem

- We need the auxiliary datasets $J = P_{XY}$ and $\mathrm{M} = P_X \otimes P_Y$
- The dataset $J = P_{XY}$ is generated by concatenating the given training examples (**X**, **Y**).
- The dataset $\mathrm{M} = (X, Y^*)$ is constructed by sampling y^* without replacement (shuffling).

MINE - Application



(a) GAN (b) GAN+MINE

Figure 3. The generator of the GAN model without mutual information maximization after 5000 iterations suffers from mode collapse (has poor coverage of the target dataset) compared to GAN+MINE on the spiral experiment.

(a) Original data (b) GAN (c) GAN+MINE

Figure 4. Kernel density estimate (KDE) plots for GAN+MINE samples and GAN samples on 25 Gaussians dataset.

Lidar-to-Camera registration I(Lidar, Camera)

- Problem definition:
  - Variable **X** becomes the Lidar data
  - Variable **Y** becomes the camera data

  - Find unknown parameters rotation **R** and translation **t**, such that $\boldsymbol{C_{coord}} = [\boldsymbol{Rt}]\boldsymbol{L_{coord}}$, by Maximizing **I(X, Y).**

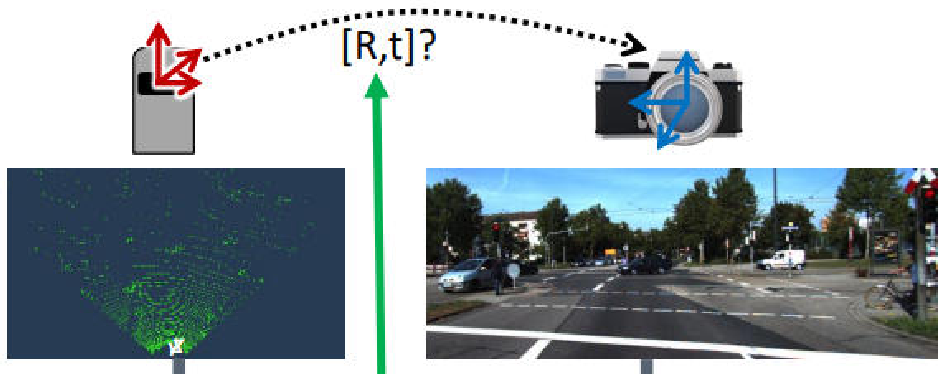


Fig. 3. Registering Lidar **X** to camera **Y** means finding the extrinsic calibration parameters **R** and **t**, where **R** is a 3d rotation matrix and t is a 3d translation vector.

Projecting Lidar data into image

Fig. 4. Visualizing projected Lidar data points in the image plane of the camera: Unregistered (left) and registered (right).

Thanks.

Matthias Hermann

Hochschule Konstanz
Institute for Optical Systems

matthias.hermann@htwg-konstanz.de

www.ios.htwg-konstanz.de

Literature

- [1] Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018, July). Mutual information neural estimation. In *International Conference on Machine Learning* (pp. 531-540). PMLR.
- [2] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.