Estimating mutual information via identification risk

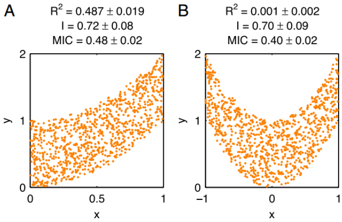
Charles Zheng

Stanford University

December 11, 2016

(Joint work with Yuval Benjamini.)

Mutual information (Shannon 1948)



- $I(X;Y) \in [0, \infty]$. (0 if $X \perp Y$, ∞ if $X = Y$ and X continuous.)
- Symmetry: $I(X;Y) = I(Y;X)$.
- Data-processing inequality

$$I(X;Y) \geq I(\phi(X); \psi(Y))$$

equality for ϕ, ψ bijections

Image credit Kinney et al. 2014.

Applications of $I(X;Y)$

- Feature selection (Peng et al. 2005, Fleuret 2004, Bennesar et al. 2015)
- Structure learning for graphical models using conditional mutual information $I(X;Y|Z)$ (Vastano and Swinney 1988, Cheng et al. 1997, Bach and Jordan 2002)
- Quantifying information capacity of neurons

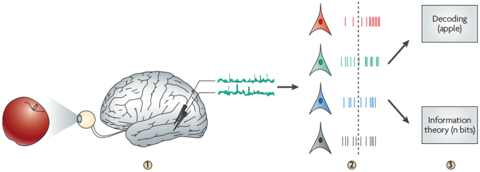


Image credits: Quiroga et al. (2009).

How to estimate $I(X;Y)$

Suppose we observe pairs $(X_i, Y_i)_{i=1}^n$ iid from density $p(x,y)$

- Definition of mutual information:

$$I(X;Y) = \int \log\left(\frac{p(x,y)}{p(x)p(y)}\right) p(x,y)dxdy$$

- Simply using plugging in kernel density estimate $\hat{p}(x,y)$ leads to large bias (Beirlant et al. 2001)
- Jackknifed estimate gives better result (Ivanov and Rozhkova 1981)

$$\hat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{\hat{p}_{-i}(x_i,y_i)}{\hat{p}_{-i}(x_i)\hat{p}_{-i}(y_i)}\right)$$

Problems in high dimensions

- Density estimation is known to have exponential complexity with respect to dimensionality.
- Many applications with high-dimensional X, Y.
  - Gene expression time series
  - Functional magnetic resonance imaging
- One approach is to assume joint multivariate normality of X, Y, but this reduces mutual information to a linear statistic.

Lower bounds on $I(X;Y)$ are still tractable

- Using data-processing inequality,

$$I(X;Y) \geq I(S(X), T(Y))$$

  where S and T are dimension-reducing inequalities (Bialek et al. 1991, Paninski 2003)
- When Y is discrete, can be also be estimated from confusion matrix of a classifier (Treves 1997, Quiroga et al. 2009)
- **Our contribution**: A novel approach for approximate lower bounds of $I(X;Y)$, specialized for the high-dimensional setting.

Our proposal

Suppose we observe pairs $(X_i, Y_i)_{i=1}^n$ iid from density $p(x, y)$.

1. Estimate a regression model (e.g. linear model or GLM) for $\mathbf{E}[y|x]$.
2. Estimate the noise model for Y.
3. Estimate the *identification risk* p using cross-validation.
4. Relate the identification risk to mutual information $I(X; Y)$:

$$I(X; Y) \approx f(p)$$

   where f is a function that we derive theoretically.

(We will stick with Gaussian linear model throughout talk.)

Multiple-response regression

- Pairs $(x_i, y_i)_{i=1}^n$, where X is p-dimensional and Y is q-dimensional.
- Data matrices $\boldsymbol{X}_{n\times p}$, $\boldsymbol{Y}_{n\times q}$.
- Fit model $Y \approx X^T B + \epsilon$, e.g. by using least-squares,

$$\hat{B}_{p\times q} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

- Estimate noise covariance $\hat{\Sigma}$ via residuals:

$$\hat{\Sigma}_{q\times q} = \frac{1}{n-k} (\boldsymbol{Y} - \boldsymbol{X}B)^T (\boldsymbol{Y} - \boldsymbol{X}B).$$

- In higher dimensions, used regularized estimators for $\hat{B}$ and $\hat{\Sigma}$ (e.g. Zou et al. 2008, Ledoit 2003).

Regression vs Identification loss

- Independent *test set* $(x_i^*, y_i^*)_{i=1}^k$.
- Use model to predict $\hat{y}_i^* = (x_i^*)^T \hat{B}$ for $i = 1, \ldots, k$.

Two ways to evaluate the predictive accuracy of the regression model:

- Regression (mean squared-error) loss:

$$\text{MSE} = \frac{1}{k} \sum_{i=1}^k ||y_i^* - \hat{y}_i^*||^2.$$

- Identification loss:

$$\text{IdLoss}_k = \frac{1}{k} \sum_{i=1}^k (1 - I\{\hat{y}_i^* \text{ is nearest neighbor of } y_i^*\}).$$

where "nearest neighbor" is with respect to Mahalanobis distance $d(z, y) = (z - y)^T \hat{\Sigma}^{-1}(z - y)$.

Cross-validated loss

Leave-k-out cross-validation (LkoCV) can be used for both squared-error loss and identification loss.

- Start with a dataset $(x_i, y_i)_{i=1}^N$.
- Let $n = N - k$. Consider all $\binom{N}{k}$ partitions of the dataset into a test set $(\boldsymbol{X}, \boldsymbol{Y})$ and training set $(\boldsymbol{X}^*, \boldsymbol{Y}^*)$.
- For each partition, compute the loss.
- Define the LkoCV loss as the average loss over $\binom{N}{k}$ partitions.

Computational note. One can subsample to avoid computing all $\binom{N}{k}$ partitions. In particular, if $m = N/k$, then one can use m-fold cross-validation which uses m partitions that have disjoint test sets.

Identification loss and mutual information

- Define the identification risk as the expected identification loss

$$\text{IdRisk}_k = \mathbf{E}[\text{IdLoss}_k]$$

- **Theorem.** (Z., Benjamini 2016) For every $k \geq 2$, there exists a function g_k such that

$$I(X;Y) \geq g_k(\text{IdRisk}_k).$$

  $\Box$.

- Resulting estimator:

$$\hat{I}_{IdLoss}(X;Y) = g_k(\text{IdLoss}_k)$$

  where IdLoss_k can either be the loss over a single test set of size k, or the LkoCV loss.

- *Remark.* Although IdLoss_k is unbiased for IdRisk_k, g_k is nonlinear so $\hat{I}_{IdLoss}$ may be biased.