

Estimating mutual information via identification risk

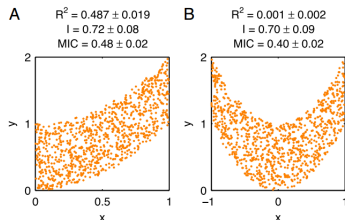
Charles Zheng

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

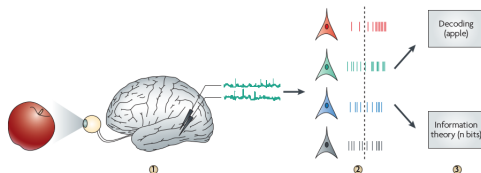


Image credits: Quiroga et al. (2009).

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:

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

- 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 $\mathbf{X}_{n \times p}$, $\mathbf{Y}_{n \times q}$.
- Fit model $Y \approx X^T B + \epsilon$, e.g. by using least-squares,

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

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

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

- 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 $i = 1, \dots, k$.

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
 $d(z, y) = (z - y)^T \hat{\Sigma}^{-1} (z - y)$.

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 (\mathbf{X}, \mathbf{Y}) and training set $(\mathbf{X}^*, \mathbf{Y}^*)$.
- 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

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

- Define the Bayes risk as the identification risk given the *true* model parameters. Hence,

$$\text{BayesRisk}_k \leq \text{IdRisk}_k.$$

- High-dimensional result.** In a certain high-dimensional asymptotic regime, there exists a limiting functional relationship

$$\text{Bayes risk} = \pi_k(\sqrt{2I(X; Y)})$$

- Resulting estimator:

$$\hat{I}_{\text{IdLoss}}(X; Y) = \frac{1}{2}(\pi_k^{-1}(\text{IdLoss}_k))^2$$

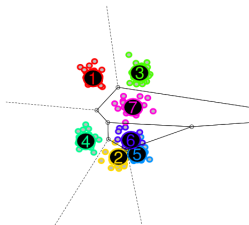
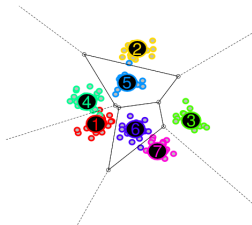
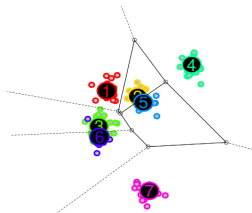
where IdLoss_k can either be the loss over a single test set of size k , or the LkoCV loss.

- Remark.* Although IdLoss_k is unbiased for IdRisk_k , g_k is nonlinear so

Gaussian example

To help think about these problems, consider a concrete example:

- Let $\mathbf{X} \sim N(0, I_d)$ and $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}, \sigma^2 I_d)$.
- We draw stimuli $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(K)} \sim N(0, I_d)$ i.i.d.
- For each stimulus $\mathbf{x}^{(i)}$, we draw observations $\mathbf{y}^{(i,j)} = \mathbf{x}^{(i)} + \epsilon^{(i,j)}$, where $\epsilon^{(i,j)} \sim N(0, \sigma^2 I_d)$.



Gaussian example

The mutual information is given by

$$I(\mathbf{X}; \mathbf{Y}) = \frac{d}{2} \log\left(1 + \frac{1}{\sigma^2}\right).$$

Define

$$Z_i = -\frac{1}{2\sigma^2} \|\mathbf{Y}^* - \mathbf{X}_i\|^2.$$

The Bayes risk can be written

$$\text{BayesRisk} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i].$$

Gaussian example: Bayes risk

- To make the problem even easier, we use another time-honored technique: the central limit theorem.
- Letting $d \rightarrow \infty$, the scores

$$Z_i = -\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{x}\|^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^d \|Y_i - x_i\|^2$$

have a jointly multivariate distribution in the limit:

$$\begin{bmatrix} Z_* \\ Z_1 \\ \vdots \\ Z_{K-1} \end{bmatrix} \xrightarrow{d} N \left(\begin{bmatrix} -\frac{d}{2} \\ -\frac{d}{2} - \frac{d}{\sigma^2} \\ \vdots \\ -\frac{d}{2} - \frac{d}{\sigma^2} \end{bmatrix}, \begin{bmatrix} \frac{d}{2} & \frac{d}{2} & \cdots & \frac{d}{2} \\ \frac{d}{2} + \frac{2d}{\sigma^2} & \frac{d}{2} + \frac{d}{\sigma^2} & \cdots & \frac{d}{2} + \frac{d}{\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{2} + \frac{d}{\sigma^2} & \frac{d}{2} + \frac{2d}{\sigma^2} & \cdots & \frac{d}{2} + \frac{2d}{\sigma^2} \end{bmatrix} \right).$$

Gaussian example: Bayes risk

Assume $(Z_*, Z_1, \dots, Z_{K-1})$ have a normal distribution with the given moments.

We can compute

$$\text{BayesRisk}_k = \Pr[Z_* < \max_{i=1}^{K-1} Z_i]$$

by writing

$$Z_i = \frac{\text{Cov}(Z_*, Z_i)}{\text{Var}(Z_*)}(Z_* - \mathbf{E}Z_*) + \sqrt{\text{Var}(Z_i) - \frac{\text{Cov}(Z_*, Z_i)^2}{\text{Var}(Z_*)}} W_i,$$

where W_i are i.i.d. standard normal.

This yields

$$\Pr[Z_* < \max_{i=1}^{K-1} Z_i] = \Pr[N(\mu, \nu^2) < \max_{i=1}^{K-1} W_i]$$

where

$$\mu = \frac{\mathbf{E}[Z_* - Z_i]}{\sqrt{\frac{1}{2}\text{Var}(Z_i - Z_j)}}, \quad \nu^2 = \frac{\text{Cov}(Z_* - Z_i, Z_* - Z_j)}{\frac{1}{2}\text{Var}(Z_i - Z_j)}$$

for $i \neq j \neq K$.

Gaussian example: Bayes risk

Finally, we get

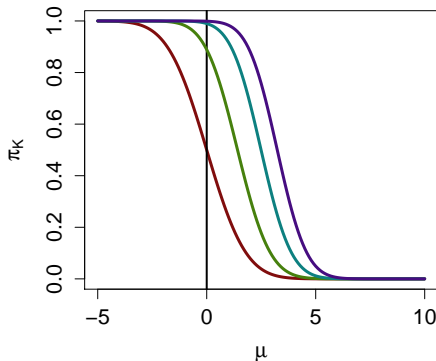
$$\text{BayesRisk} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i] \rightarrow \pi_K \left(\frac{\sqrt{d}}{\sigma} \right)$$

where

$$\pi_K(\mu) = 1 - \int_{-\infty}^{\infty} \phi(z - \mu)(1 - \Phi(z))^{K-1} dz.$$

Sidenote: interpretation of π_K

The function $\pi_K(\mu)$ gives the probability that a $N(\mu, 1)$ variable is smaller than the minimum of $K - 1$ other $N(0, 1)$ variables (all independent.) Hence $\pi_K(0) = \frac{K-1}{K}$ due to symmetry. (This is also the misclassification rate from pure guessing.)



Legend: $K = \{ \text{2}, \text{9}, \text{99}, \text{999} \}$

Gaussian example: Bayes risk

Recall that

$$I(\mathbf{X}; \mathbf{Y}) = \frac{d}{2} \log\left(1 + \frac{1}{\sigma^2}\right),$$

while

$$\text{BayesRisk}_k = \pi_K(\sqrt{d}/\sigma).$$

Hence Bayes risk is *not* a function of $I(\mathbf{X}; \mathbf{Y})$!

Gaussian example: Low SNR limit

However, what if we consider a limit where the noise level σ^2 increases with d ?

Fix some $\sigma_1^2 > 0$, and let $\sigma_d^2 = d\sigma_1^2$.

Then when d is large,

$$I(\mathbf{X}; \mathbf{Y}) = \frac{d}{2} \log\left(1 + \frac{1}{d\sigma_1^2}\right) \approx \frac{d}{2} \frac{1}{d\sigma_1^2} = \frac{1}{2\sigma_1^2}.$$

We get

$$\text{Bayes risk} = \pi_k(\sqrt{2I(\mathbf{X}; \mathbf{Y})})$$

in the limit!

Low SNR limit: generalization

In a sequence of gaussian models of increasing dimensionality with

$$\lim_{d \rightarrow \infty} I(\mathbf{X}; \mathbf{Y}) \rightarrow \iota < 0,$$

we get an exact relationship between the limiting mutual information and the average Bayes error,

$$\text{BayesRisk}_k = \pi_K(\sqrt{2\iota}).$$

This limiting relationship holds more generally!

Theorem. *Given an exponential family sequence model $p_d(\mathbf{x}, \mathbf{y})$, for random variates $(\mathbf{X}^{[d]}, \mathbf{Y}^{[d]}) \sim p_d(\mathbf{X}, \mathbf{Y})$, we have*

$$\lim_{d \rightarrow \infty} I(\mathbf{X}^{[d]}, \mathbf{Y}^{[d]}) = \iota < \infty$$

for some constant $\iota < \infty$; and the limiting K -class average Bayes error is given by

$$\lim_{d \rightarrow \infty} ABE = \pi_K(\sqrt{2\iota}).$$

The low-SNR estimator of $I(\mathbf{X}; \mathbf{Y})$

We are willing to bet that the relationship

$$\text{ABE} \approx \pi_K(\sqrt{2\ell})$$

holds in much greater generality than we managed to prove—namely, whenever $I(\mathbf{X}; \mathbf{Y}) \ll p$, and the scores Z_i are approximately jointly multivariate normal.

Based on these assumptions, our proposed estimator for mutual information is

$$\hat{l}_{ls}(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} \pi_K^{-1}(\widehat{\text{ABE}})^2$$

where $\widehat{\text{ABE}}$ is the test error of the classifier. (The subscript ls stands for low-SNR.)

Models.

- Multiple-response logistic regression model

$$X \sim N(0, I_p)$$

$$Y \in \{0, 1\}^q$$

$$Y_i | X = x \sim \text{Bernoulli}(x^T B_i)$$

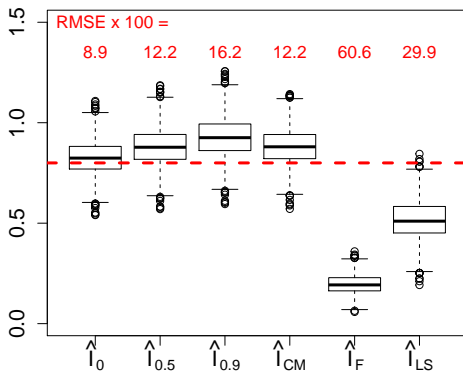
where B is a $p \times q$ matrix.

Methods.

- Nonparametric: \hat{I}_0 naive estimator, \hat{I}_α anthropic correction.
- ML-based: \hat{I}_{CM} confusion matrix, \hat{I}_F Fano, \hat{I}_{LS} low-SNR method.

Fig 1. Low-dimensional results ($q = 3$)

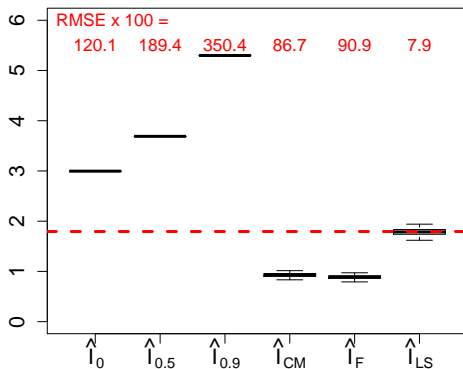
Sampling distribution of \hat{I} for $\{p = 3, B = \frac{4}{\sqrt{3}}I_3, K = 20, r = 40\}$.
True parameter $I(X; Y) = 0.800$ (dotted line.)



Naïve estimator performs best! \hat{I}_{LS} not effective.

Fig 2. High-dimensional results ($q = 50$)

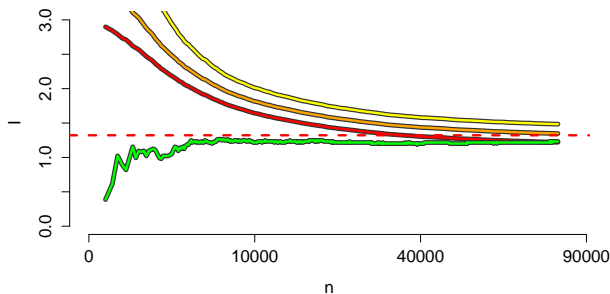
Sampling distribution of \hat{I} for $\{p = 50, B = \frac{4}{\sqrt{50}} I_{50}, K = 20, r = 8000\}$.
True parameter $I(X; Y) = 1.794$ (dashed line.)



Non-parametric methods extremely biased.

Fig 3. Dependence on n ($q = 10$)

Estimation path of \hat{l}_{LS} and \hat{l}_α as n ranges from 10 to 8000.
 $\{p = 10, B = \frac{4}{\sqrt{10}} l_{10}, K = 20\}$. True parameter $I(X; Y) = 1.322$ (dashed line.)

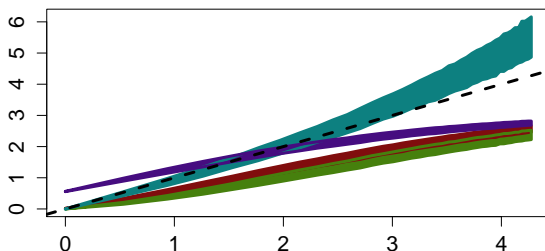


Legend: \hat{l}_{LS} , \hat{l}_0 , $\hat{l}_{0.5}$, $\hat{l}_{0.9}$.

Fig 4. Dependence on true $I(X; Y)$ ($q = 10$)

$$\{p = 10, B = [0, 200] \times \frac{1}{\sqrt{10}} l_{10}, r = 1000, K = 20\}.$$

Estimated \hat{I} vs true I .



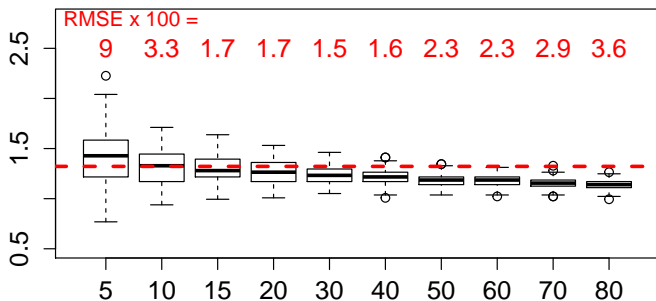
Bands depict 80% central percentiles.

Legend: ■ = \hat{I}_{LS} , ■ = \hat{I}_0 , ■ = \hat{I}_{CM} , ■ = \hat{I}_F .

Fig 5. Dependence on K given fixed N ($q = 10$)

Sampling distribution of \hat{I}_{LS} for $\{p = 10, B = \frac{4}{\sqrt{10}}I_{10}, N = 80000\}$,
and $K = \{5, 10, 15, 20, \dots, 80\}$, $r = N/k$.

True parameter $I(X; Y) = 1.322$ (dashed line.)



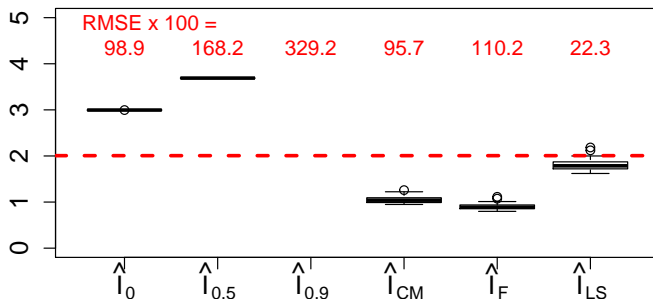
Decreasing variance as K increases. Bias at large and small K .

Fig 6. Non-identity B ($q = 40$)

$p = 20$ and $q = 40$, entries of B are iid $N(0, 0.025)$.

$K = 20$, $r = 8000$, true $I(X; Y) = 1.86$ (dashed line.)

Sampling distribution of \hat{I} .



- We derive a relationship between average Bayes error (ABE) and mutual information (MI), motivating a novel estimator \hat{I}_{LS} .
- Theory based on high dimensional, low SNR limit, where

$$\text{ABE} \leftrightarrow \text{MI}.$$

- In ideal settings for supervised learning, ABE can be estimated effectively and \hat{I}_{LS} can recover MI at much lower sample sizes than nonparametric methods.
- In simulations, \hat{I}_{LS} works better than Fano's inequality or the confusion matrix approach.

- Cover and Thomas. Elements of information theory.
- Muirhead. Aspects of multivariate statistical theory.
- van der Vaart. Asymptotic statistics.