# Estimating mutual information using sparse regression

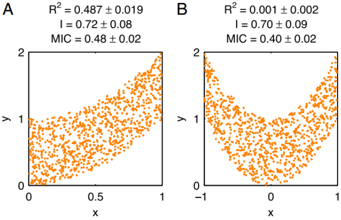
Charles Zheng

Stanford University

December 12, 2016

(Joint work with Yuval Benjamini.)

# Mutual information (Shannon 1948)




- $I(X;Y) \in [0, \infty]$. (0 if $X \perp Y$, $\infty$ 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 $\phi$, $\psi$ 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

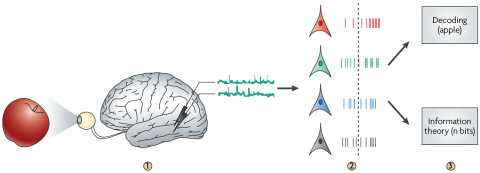



Image credits: Quiroga et al. (2009).

# 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:

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

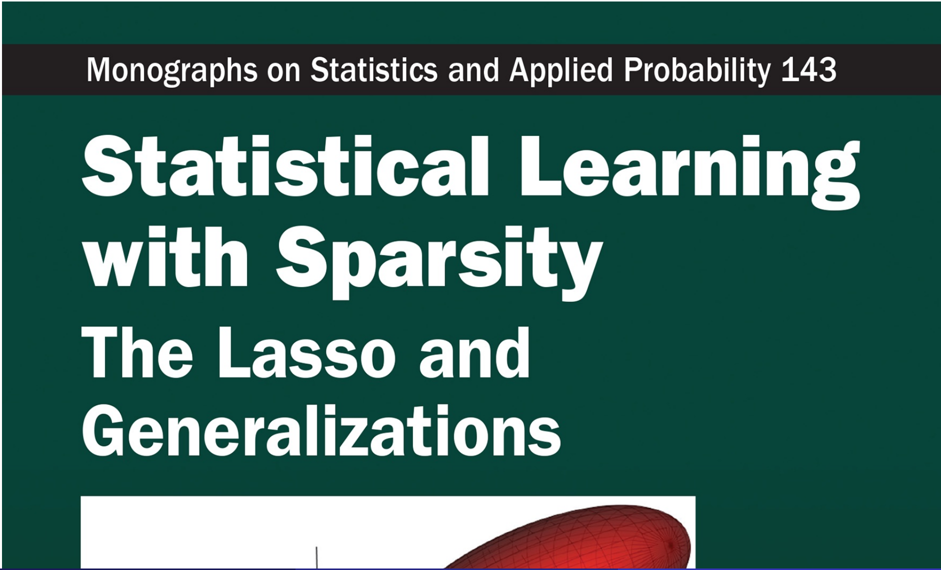
- Simply using plugging in kernel density estimate $\hat{p}(x, y)$ leads to large bias (Beirlant et al. 2001)
- Jackknifed estimate gives better result (Ivanov and Rozhkova 1981)

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

# 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.
- Other approaches: binning (Bialek et al. 1991, Paninski 2003), confusion matrix of a classifier (Treves 1997, Quiroga et al. 2009).

# Idea: Use sparsity!



# Our proposal

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

1. Estimate a (sparse) regression model 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.

# Multiple-response regression

- Pairs $(x_i, y_i)_{i=1}^n$, where $X$ is $p$-dimensional and $Y$ is $q$-dimensional.
- Data matrices $\boldsymbol{X}_{n \times p}$, $\boldsymbol{Y}_{n \times q}$.
- For each column of $Y$, fit sparse model $Y^{(i)} \approx X^T \beta^{(i)} + \epsilon$, e.g. by using elastic net (Zou 1998),

$$\hat{\beta}^{(i)} = \text{argmin}_\beta ||\boldsymbol{X}^T \beta^{(i)} - Y^{(i)}||^2 + \lambda_2 ||\beta^{(i)}||_2^2 + \lambda_1 ||\beta^{(i)}||_1$$

# 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 (L$k$oCV) 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 L$k$oCV 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:

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

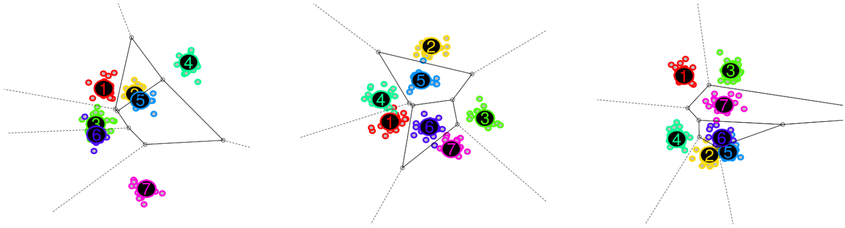
where $\text{IdLoss}_k$ can either be the loss over a single test set of size $k$, or the L$k$oCV loss.

- *Remark.* Although $\text{IdLoss}_k$ is unbiased for $\text{IdRisk}_k$, $g_k$ is nonlinear so $\hat{I}$

# Gaussian example

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

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



# Gaussian example

The mutual information is given by

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

Define

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

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 $d \to \infty$, the scores

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

have a jointly multivariate distribution in the limit:

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

# Gaussian example: Bayes risk

Assume $(Z_*, Z_1, \ldots, Z_{K-1})$ 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

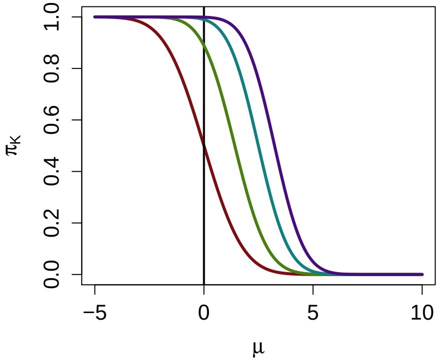
$$\text{BayesRisk} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i] \to \pi_K\left(\frac{\sqrt{d}}{\sigma}\right)$$

where

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

# Sidenote: interpretation of $\pi_K$

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

# Gaussian example: Bayes risk

Recall that

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

while

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

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

# Gaussian example: Low SNR limit

However, what if we consider a limit where the noise level $\sigma^2$ increases with $d$?
Fix some $\sigma_1^2 > 0$, and let $\sigma_d^2 = d\sigma_1^2$.
Then when $d$ is large,

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

We get

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

in the limit!

# Low SNR limit: generalization

In a sequence of gaussian models of increasing dimensionality with

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

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!**

# Low SNR theorem

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

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

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

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

# The low-SNR estimator of $I(X; Y)$

We are willing to bet that the relationship

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

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

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

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

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

# Simulation study

*Models.*

- Multiple-response logistic regression model

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

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

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

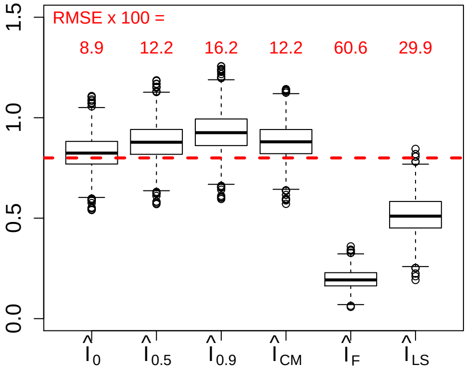
  where $B$ is a $p \times q$ matrix.

*Methods.*

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

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

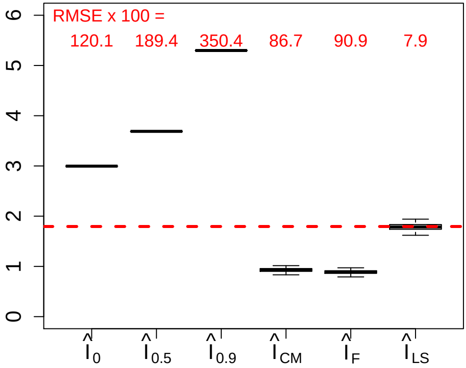
Sampling distribution of $\hat{I}$ for $\{p = 3,\ B = \frac{4}{\sqrt{3}} I_3,\ K = 20,\ r = 40\}$.
True parameter $I(X; Y) = 0.800$ *(dotted line.)*




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

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

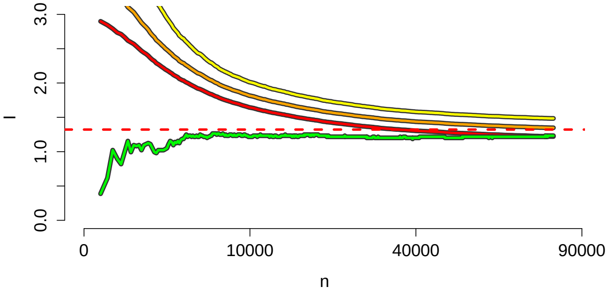
Sampling distribution of $\hat{I}$ for $\{p = 50,\ B = \frac{4}{\sqrt{50}} I_{50},\ K = 20,\ r = 8000\}$.
True parameter $I(X;Y) = 1.794$ *(dashed line.)*




Non-parametric methods extremely biased.

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

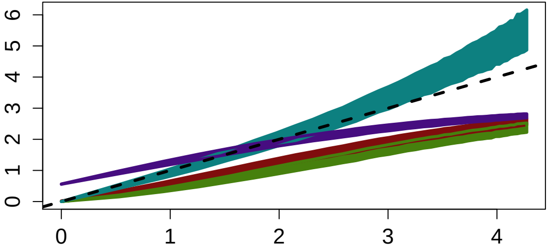
Estimation path of $\hat{I}_{LS}$ and $\hat{I}_\alpha$ as $n$ ranges from 10 to 8000.
$\{p = 10,\ B = \frac{4}{\sqrt{10}} I_{10},\ K = 20\}$. True parameter $I(X;Y) = 1.322$ *(dashed line.)*




Legend: [green] $= \hat{I}_{LS}$, [red] $= \hat{I}_0$, [orange] $= \hat{I}_{0.5}$, [yellow] $= \hat{I}_{0.9}$.

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

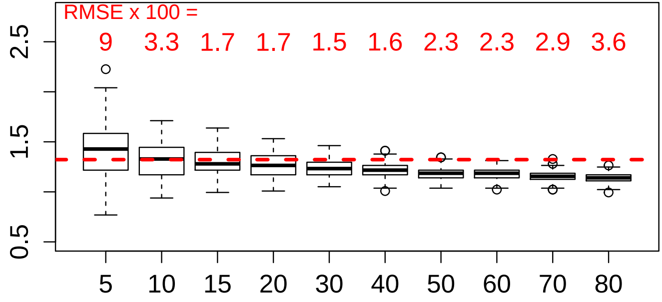
**Estimated $\hat{I}$ vs true $I$.**




Bands depict 80% central percentiles.
Legend: teal $= \hat{I}_{LS}$, purple $= \hat{I}_0$, red $= \hat{I}_{CM}$, green $= \hat{I}_F$.

# 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 $K = \{5, 10, 15, 20, \ldots, 80\}$, $r = N/k$.
True parameter $I(X; Y) = 1.322$ *(dashed line.)*

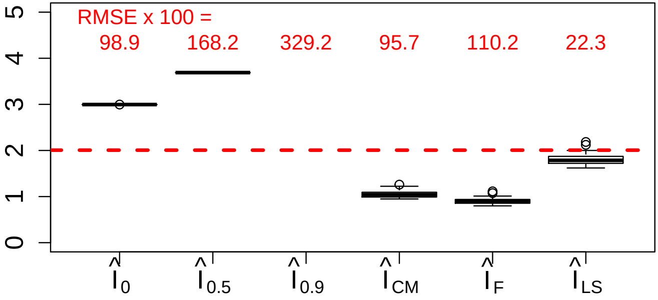



Decreasing variance as $K$ increases. Bias at large and small $K$.

# 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 $\hat{I}$.**



# Conclusions

- 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.

# References


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