Estimating mutual information using sparse regression

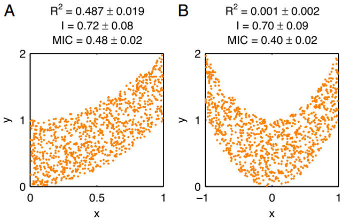
Charles Zheng

Stanford University

December 12, 2016

(Joint work with Yuval Benjamini.)

Mutual information (Shannon 1948)



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

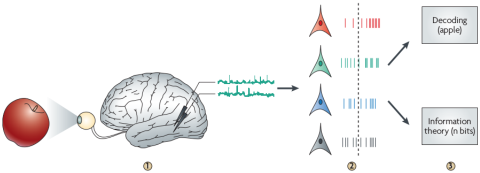


Image credits: Quiroga et al. (2009).

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:

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

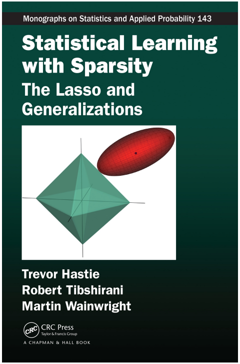
- Simply using plugging in kernel density estimate $\hat{p}(x, y)$ leads to large bias (Beirlant et al. 2001)
- Jackknifed estimate gives better result (Ivanov and Rozhkova 1981)

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

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

Idea: Use sparsity!



- Suppose that $Y \approx f(X) + \epsilon$, where f depends *sparsely* on X.
- Can we exploit the sparsity to obtain an estimate of $I(X; Y)$?

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. Use the identification risk to obtain a lower bound for the mutual information $I(X; Y)$:
$$I(X; Y) \geq f(p)$$
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),

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

- Or, fit a *random forest* model for each column of Y (Breiman 2001)

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:

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

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

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

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

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

- **Theorem.** (Z., Benjamini 2016) There exists a function g_k such that

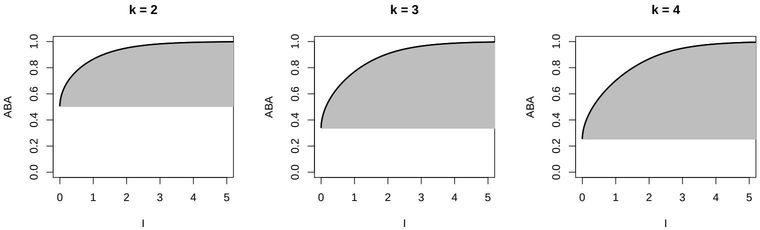
$$I(X;Y) \geq g_k(\text{IdRisk}_k).$$

- Resulting estimator:

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

Functions

Illustration of $C_k = g_k^{-1}$

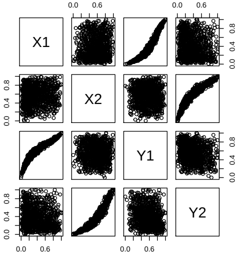


As information increases, the maximal identification risk goes to 0. [note: pictures need to be rotated]

Section 2

Applications

Simulation



- Generate data: $(Y_1, Y_2) = f(X_1, X_2, \epsilon)$ where f is nonlinear.
- $n = 1000$.
- Compare Nearest-Neighbor estimator (Mnatsakov et al, 2008, implemented in `FNN`) with our method using *Random Forest*.
- Add extra noise dimensions $X_3, X_4, \ldots$.

Simulation Results

True $I(X; Y) = 4.615$.

Extra dim	NN	RF $k = 10$	RF $k = 20$
0	**4.445**	3.989	3.924
1	3.040	**3.645**	3.610
2	1.773	**3.249**	3.182

Application to gene expression time series

to be contd

Section 3

Theory

Functional formulation

Identification risk $\text{IdRisk}_k[p(x,y)]$ and mutual information $\text{I}[p(x,y)]$ are both *functionals* of $p(x,y)$.

$$\text{IdAcc}_k[p(x,y)] = 1 - \frac{1}{k}\int p_X(x_1)\dots p_X(x_k) \max_{i=1}^{k} p(y|x_i) dx_1 \dots dx_k dy.$$

$$\text{I}[p(x,y)] = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

where $\text{IdAcc}_k = 1 - \text{IdRisk}_k$.

Problem formulation

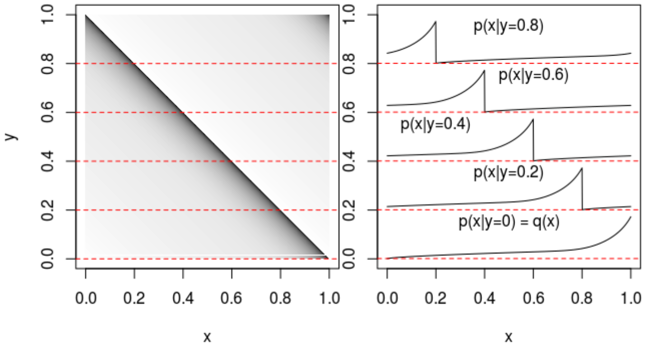
Take $\iota > 0$, and fix $k \in \{2, 3, ...\}$. Let $p(x, y)$ be a joint density (where (X, Y) could be random vectors of any dimensionality.) Supposing

$$\mathrm{I}[p(x, y)] \leq \iota,$$

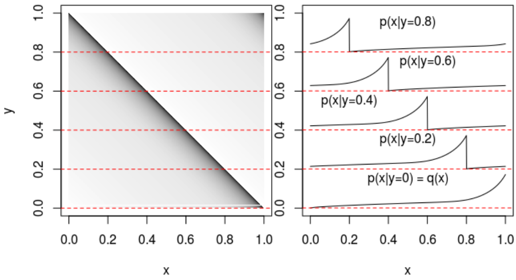
then can we find an upper bound on $\mathrm{IdAcc}_k[p(x, y)]$?

Reduced Problem

Rather than show the whole proof, we consider a simplified problem to illustrate the methods.



Actually, the simplified problem is equivalent to the full problem and we get the same answer (but this is non-trivial).



- $p(x, y)$ on unit square with uniform marginals.
- The conditional distributions $p(x|y)$ are just "shifted" copies of a common density, $q(x)$, on $[0, 1]$

$$p(x|y) = q(x - y + I\{x < y\})$$

- Furthermore, $q(x)$ is increasing in x.

Simplified formulae

The information and average Bayes error can be written in terms of $q(x)$.

$$\mathrm{I}[p(x,y)] = \int_0^1 q(x) \log q(x) dx$$

$$\mathrm{IdAcc}_k[p(x,y)] = \int_{[0,1]^k} \max_{i=1}^{k} q(x_i) dx_1 \cdots dx_k$$

Simplified formulae

Overload the notation and "redefine" information and average Bayes error as functionals of $q(x)$.

$$\mathrm{I}[q(x)] \stackrel{def}{=} \int_0^1 q(x) \log q(x) dx$$

$$\mathrm{IdAcc}_k[q(x)] \stackrel{def}{=} \frac{1}{k} \int_{[0,1]^k} \max_{i=1}^{k} q(x_i) dx_1 \cdots dx_k$$

Optimization problem

We now pose the question: how do we find $q(x)$ which maximizes $\mathrm{IdAcc}_k[q(x)]$ subject to $\mathrm{I}[q(x)] \leq \iota$?

- *Domain of the optimization*: Recall that $q(x)$ satisfies $q(x) \geq 0$, $\int_0^1 q(x)dx = 1$, and is increasing in x. Let $\mathcal{Q}$ denote the space of functions on $[0, 1] \to [0, \infty)$ which are increasing in x.
- *Constraints*: We have two remaining constraints, $\mathrm{I}[q(x)] \leq \iota$ and $\int_0^1 q(x)dx = 1$.

Hence the problem is

$$\text{maximize}_{q(x)\in\mathcal{Q}}\ \mathrm{IdAcc}_k[q(x)] \text{ subject to } \int_0^1 q(x)dx = 1 \text{ and } \mathrm{I}[q(x)] \leq \iota.$$

Optimization problem

$$\text{maximize}_{q(x)\in\mathcal{Q}}\ \mathrm{IdAcc}_k[q(x)] \text{ subject to } \int_0^1 q(x)dx = 1 \text{ and } \mathrm{I}[q(x)] \leq \iota.$$

- Does a solution exist? *Yes*, because the space of measures with density $q(x)$ satisfying $\mathrm{I}[q(x)] \leq \iota$ is tight, and both the constraints and objective are continuous wrt to the topology of weak convergence.
- Given a solution $q^*(x)$ exists, there exist Lagrange multipliers $\lambda \in \mathbb{R}$ and $\nu > 0$ such that q^* minimizes

$$\begin{aligned}\mathcal{L}[q(x)] &= -\mathrm{IdAcc}_k[q(x)] + \lambda \int_0^1 q(x)dx + \nu \mathrm{I}[q(x)] \\ &= \int_0^1 (-t^{k-1} + \lambda + \nu \log q(x))q(x)dx.\end{aligned}$$

Functional derivatives

- Taylor explansions are a useful trick for computing functional derivatives
- We can compute the functional derivative of $\mathcal{L}[q(x)]$ by writing

$$\begin{aligned}
&\mathcal{L}[q(x) + \epsilon\xi(x)] \\
&\quad = \int_0^1 (-t^{k-1} + \lambda + \nu \log(q(x) + \epsilon\xi(x)))(q(x) + \epsilon\xi(x))dx. \\
&\quad \approx \int (q(x) + \epsilon\xi(x))(-t^{k-1} + \lambda + \nu\{\log q(x) + \frac{\epsilon\xi(x)}{q(x)}\})dx \\
&\quad \approx \mathcal{L}[q(x)] + \int_0^1 (-t^{k-1} + \lambda + \nu(1 + \log q(x))\epsilon\xi(x)dx.
\end{aligned}$$

- Hence

$$\nabla\mathcal{L}[q](x) = -t^{k-1} + \lambda + \nu(1 + \log q(x))$$

Variational magic!

Suppose we set the functional derivative to 0,

$$0 = \nabla \mathcal{L}[q](t) = -t^{k-1} + \lambda + \nu + \nu \log q(t).$$

Then we conclude that the optimal $q^*(t)$ takes the form

$$q^*(t) = \alpha e^{\beta t^{k-1}}$$

for some $\alpha > 0$, $\beta > 0$.
From the constraint $\int q(t)dt = 1$, we get

$$q_\beta(t) = \frac{e^{\beta t^{k-1}}}{\int e^{\beta t^{k-1}} dt}.$$

Result

Theorem. For any $\iota > 0$, there exists $\beta_\iota \geq 0$ such that defining

$$q_\beta(t) = \frac{\exp[\beta t^{k-1}]}{\int_0^1 \exp[\beta t^{k-1}]},$$

we have

$$\int_0^1 q_{\beta_\iota}(t) \log q_{\beta_\iota}(t) dt = \iota.$$

Then,

$$\sup_{I(X;Y)=\iota} \text{IdAcc}_k = \int_0^1 q_{\beta_\iota}(t) t^{k-1} dt.$$