# Estimating mutual information for high-dimensional sparse relationships

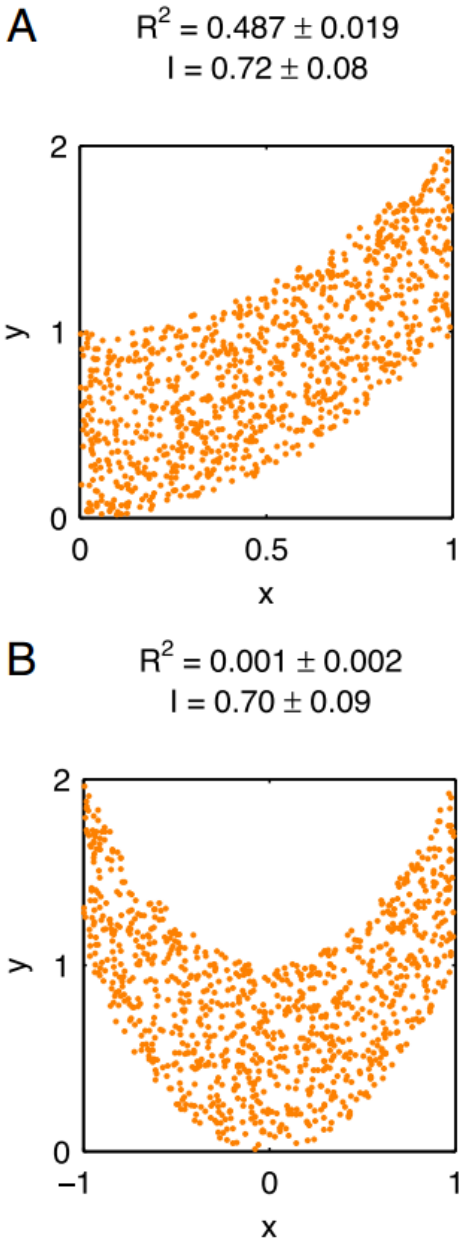
Charles Zheng

Stanford University

January 16, 2017

(Joint work with Yuval Benjamini.)

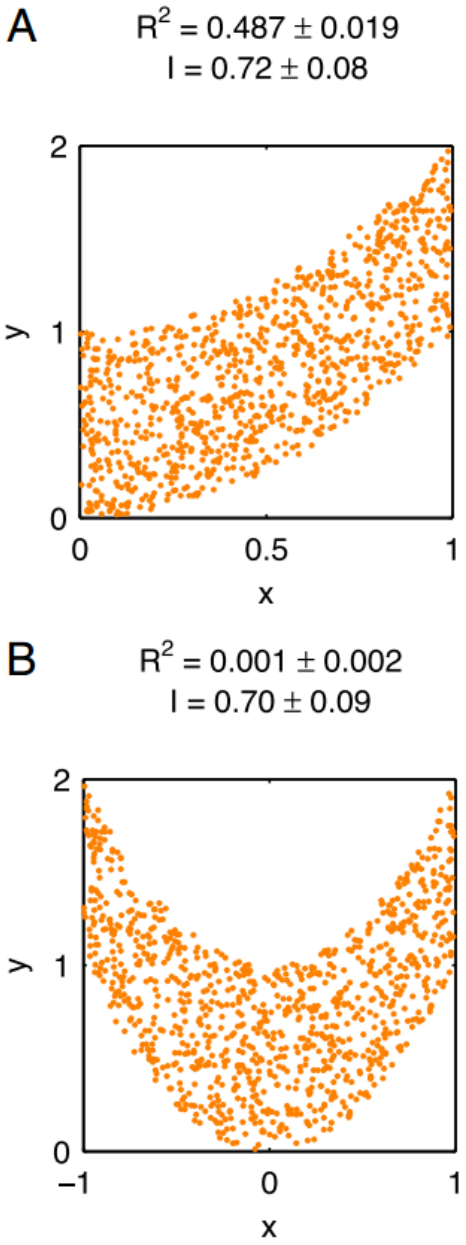
# Overview




Mutual information $I(\vec{X}; \vec{Y})$

- measures dependence between two random vectors, $\vec{X}$ and $\vec{Y}$

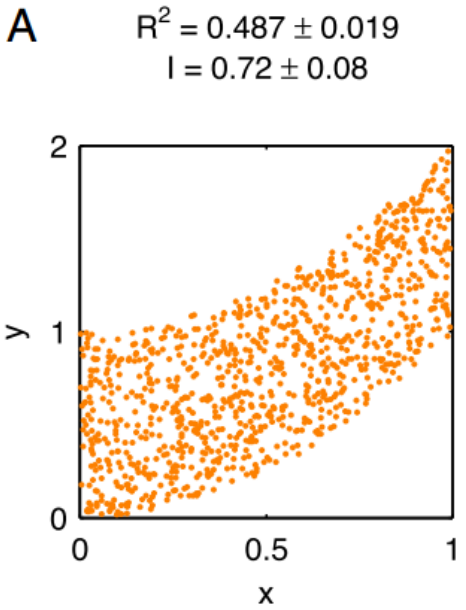
# Overview

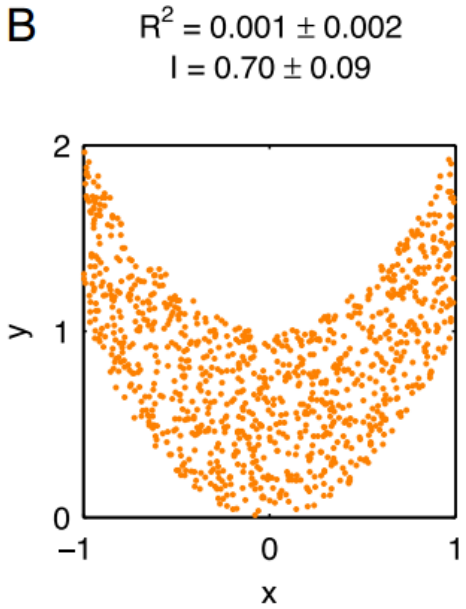



Mutual information $I(\vec{X}; \vec{Y})$

- measures dependence between two random vectors, $\vec{X}$ and $\vec{Y}$
- applies to nonlinear and multidimensional relationships (unlike correlation)

# Overview

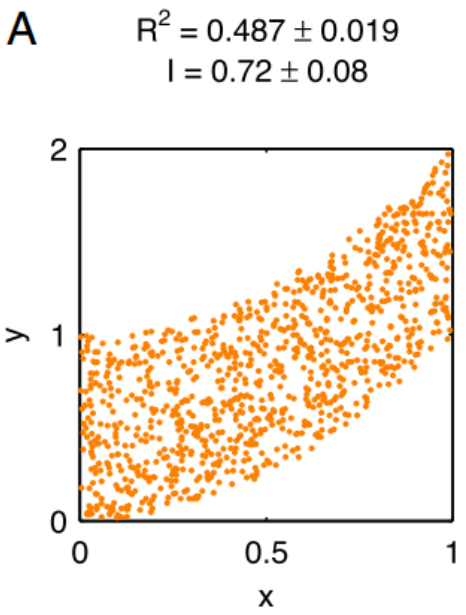


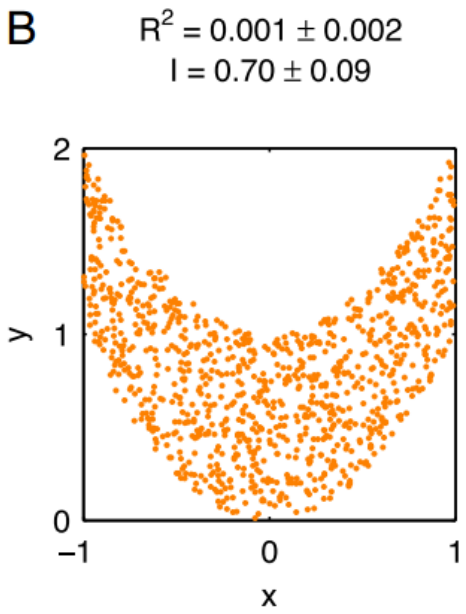




Mutual information $I(\vec{X}; \vec{Y})$

- measures dependence between two random vectors, $\vec{X}$ and $\vec{Y}$
- applies to nonlinear and multidimensional relationships (unlike correlation)
- is *difficult to estimate* in high dimensions

# Overview

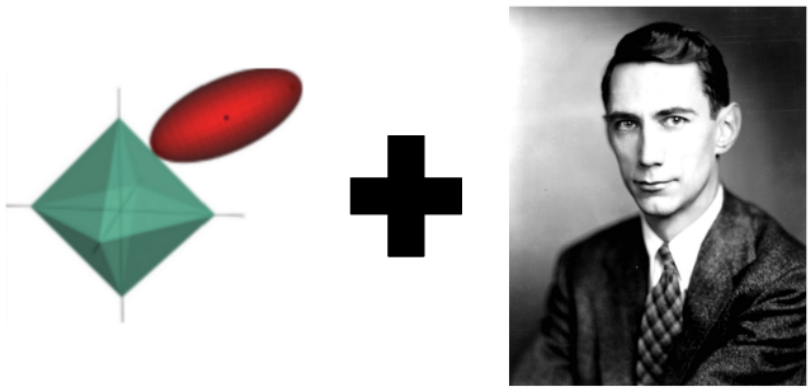


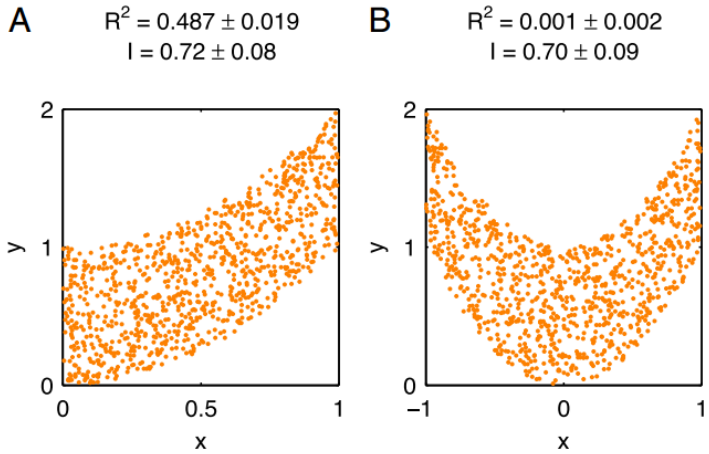

Mutual information $I(\vec{X}; \vec{Y})$

- measures dependence between two random vectors, $\vec{X}$ and $\vec{Y}$
- applies to nonlinear and multidimensional relationships (unlike correlation)
- is *difficult to estimate* in high dimensions

We combine *machine learning* (sparse estimation) with *information theory* to obtain better estimates of $I(\vec{X}; \vec{Y})$



# Mutual information $I(X; Y)$




Introduced in Shannon's 1948 paper, "A mathematical theory of communication"

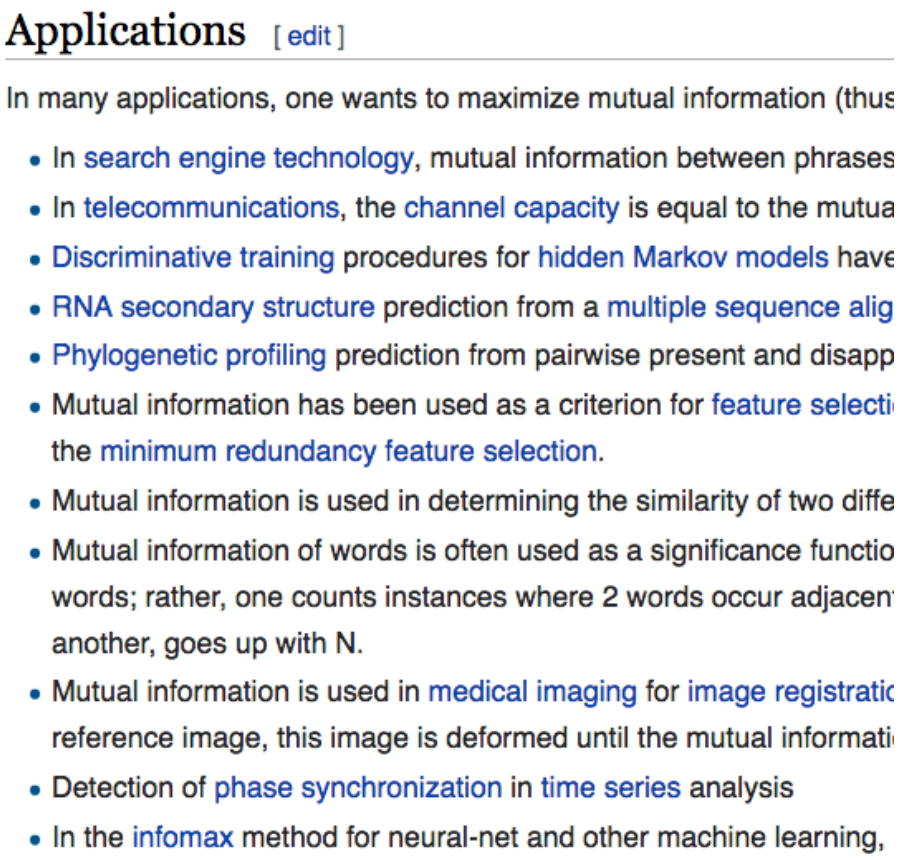
Image credit Kinney et al. 2014.

# Applications of $I(X;Y)$

Mutual information has since been applied to many areas outside of information theory

# Applications of $I(X;Y)$

Mutual information has since been applied to many areas outside of information theory



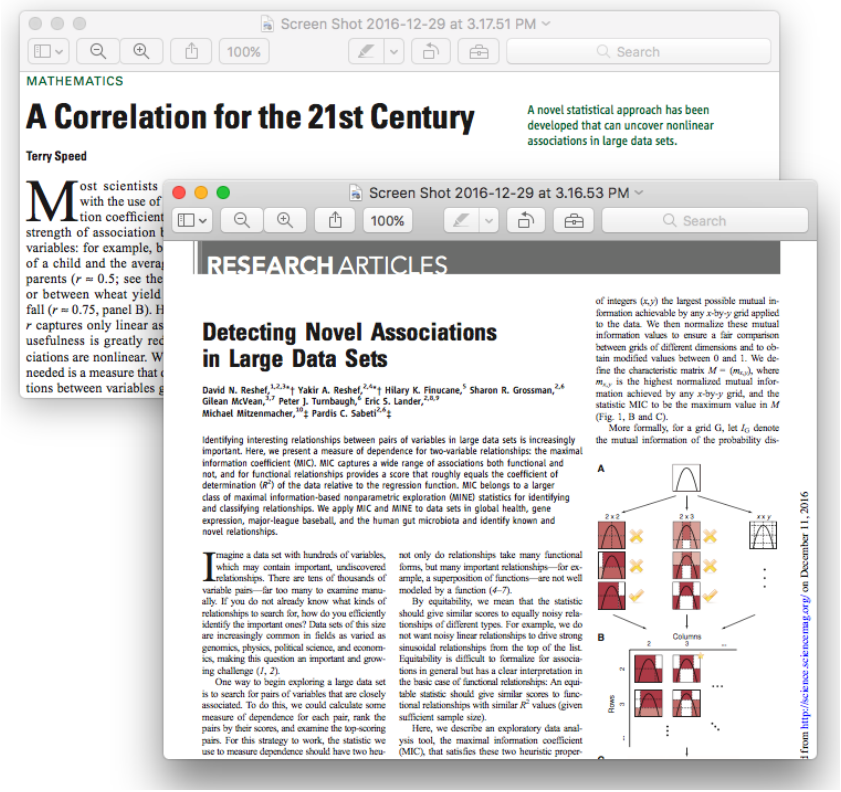
## Applications [ edit ]

In many applications, one wants to maximize mutual information (thus

- In search engine technology, mutual information between phrases
- In telecommunications, the channel capacity is equal to the mutua
- Discriminative training procedures for hidden Markov models have
- RNA secondary structure prediction from a multiple sequence alig
- Phylogenetic profiling prediction from pairwise present and disapp
- Mutual information has been used as a criterion for feature selecti the minimum redundancy feature selection.
- Mutual information is used in determining the similarity of two diffe
- Mutual information of words is often used as a significance functio words; rather, one counts instances where 2 words occur adjacen another, goes up with N.
- Mutual information is used in medical imaging for image registratic reference image, this image is deformed until the mutual informati
- Detection of phase synchronization in time series analysis
- In the infomax method for neural-net and other machine learning,

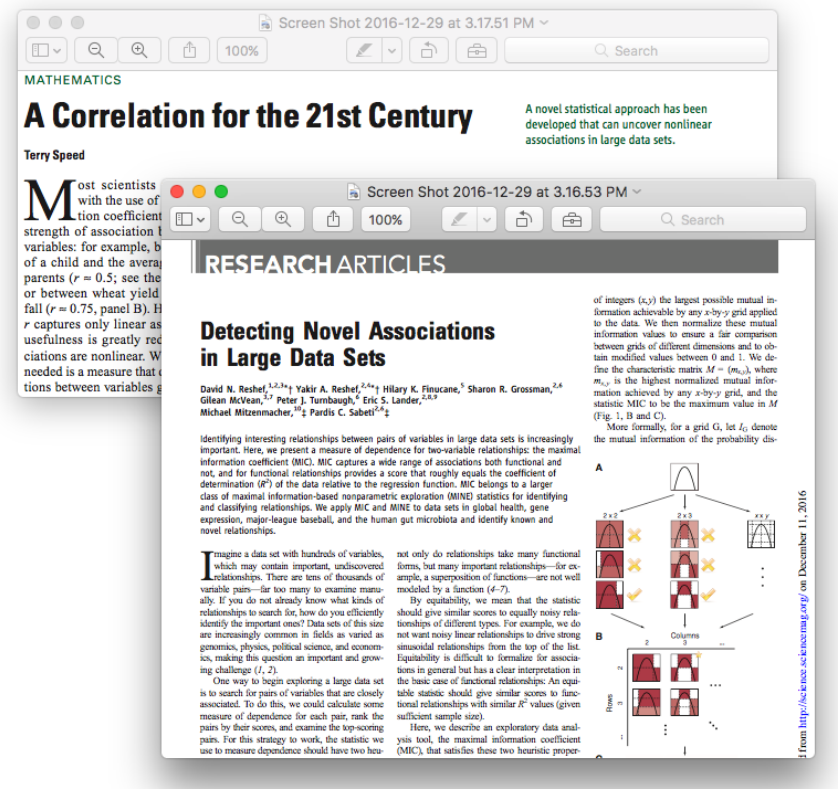
Engineering, biology, computer science, physics, medicine

# Comparing $I(X; Y)$ with Pearson correlation



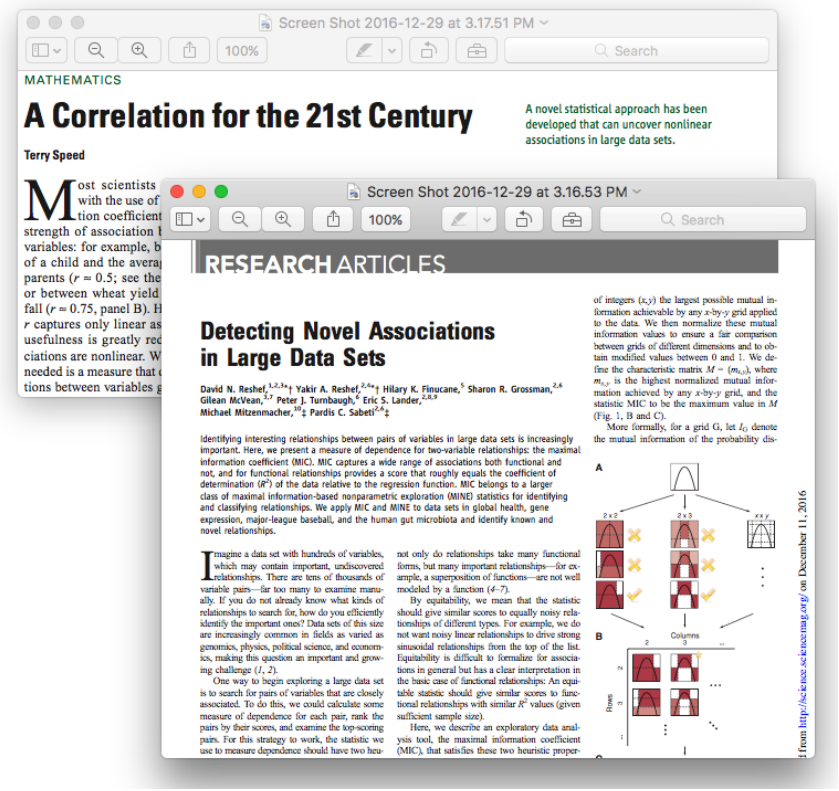

- In many applications scientists are interested in *dependence*, not *correlation* (Reshef et al. 2011, Speed 2011).

# Comparing $I(X; Y)$ with Pearson correlation




- In many applications scientists are interested in *dependence*, not *correlation* (Reshef et al. 2011, Speed 2011).
- Only mutual information (and derived quantities) measures dependence directly

# Comparing $I(X; Y)$ with Pearson correlation




- In many applications scientists are interested in *dependence*, not *correlation* (Reshef et al. 2011, Speed 2011).
- Only mutual information (and derived quantities) measures dependence directly

# Problems with mutual information

- Hard to interpret (compared to $R^2$)
- Hard to estimate (compared to $R^2$)

# Can we make $I(X;Y)$ easier to interpret?

- Define the "informational correlation" (Linfoot 1957)

$$\text{Cor}_{Info}(X,Y) = \sqrt{1 - e^{-2I(X;Y)}}$$

# Can we make $I(X; Y)$ easier to interpret?

- Define the "informational correlation" (Linfoot 1957)

$$\text{Cor}_{Info}(X, Y) = \sqrt{1 - e^{-2I(X;Y)}}$$

- Then $\text{Cor}_{Info}(X, Y) \in [0, 1]$.
- For $(X, Y)$ bivariate normal,

$$|\text{Cor}_{Pearson}(X, Y)| = \text{Cor}_{Info}(X, Y)$$

# Can we make $I(X; Y)$ easier to interpret?

- Define the "informational correlation" (Linfoot 1957)

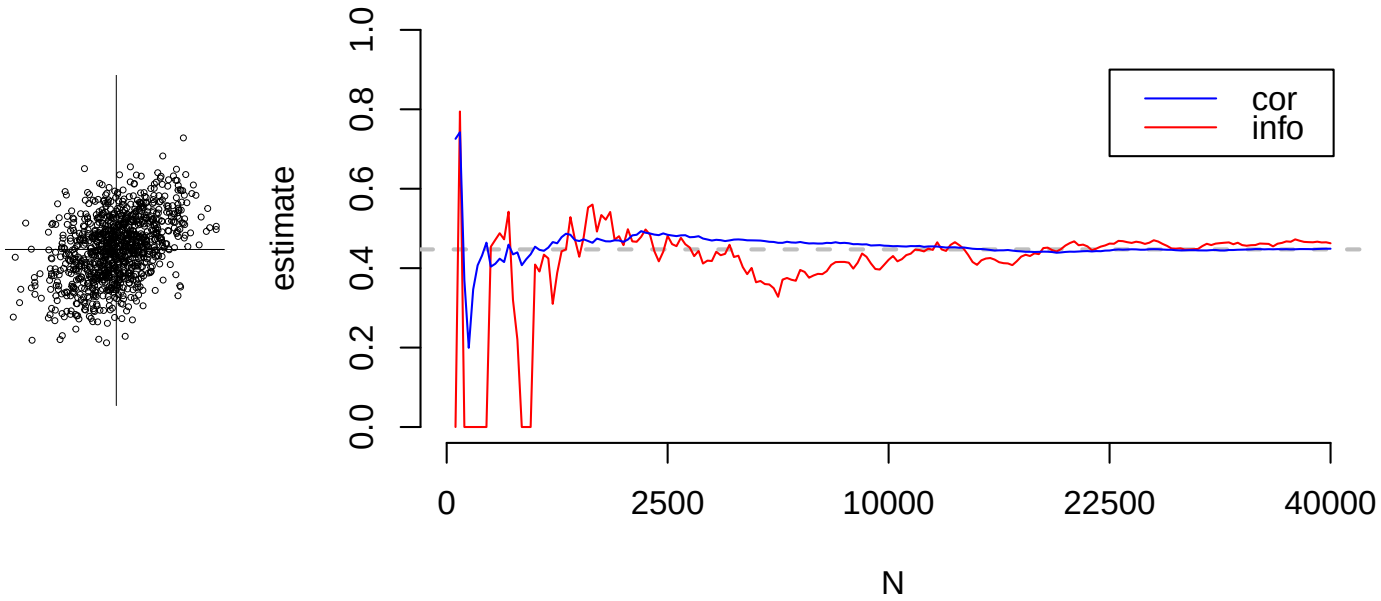
$$\text{Cor}_{Info}(X, Y) = \sqrt{1 - e^{-2I(X;Y)}}$$

- Then $\text{Cor}_{Info}(X, Y) \in [0, 1]$.
- For $(X, Y)$ bivariate normal,

$$|\text{Cor}_{Pearson}(X, Y)| = \text{Cor}_{Info}(X, Y)$$

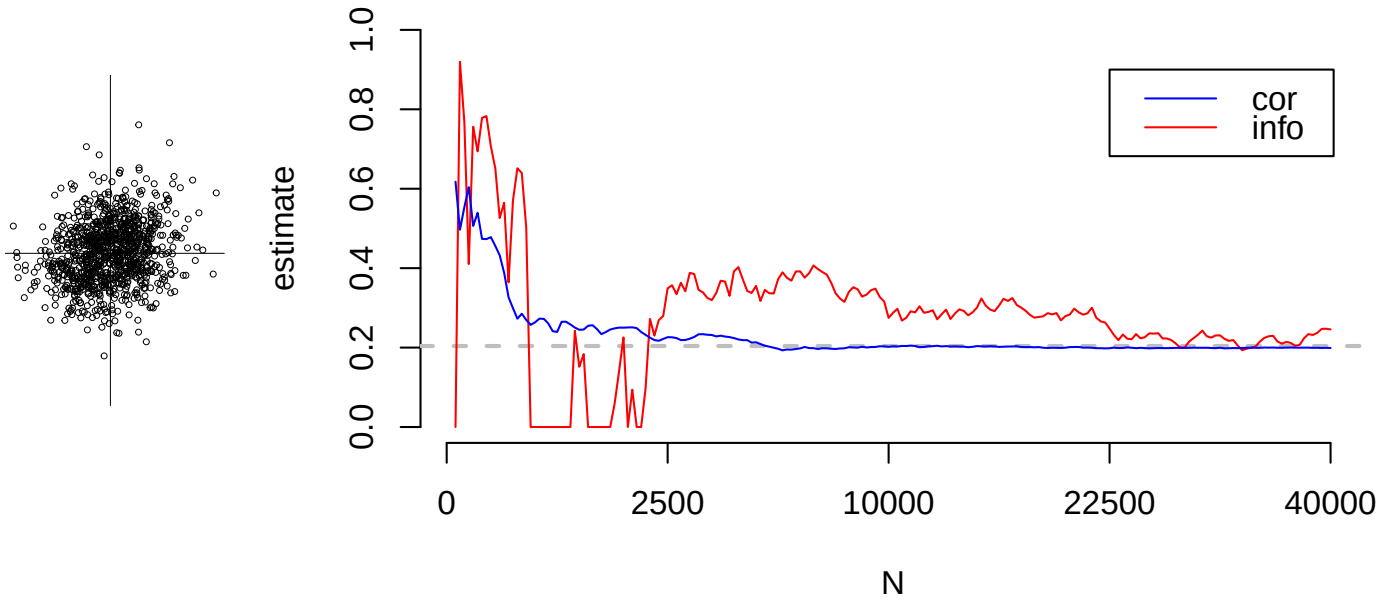
# Difficulty of estimating $I(X;Y)$

Example with $\text{Cor}_{Pearson}(X,Y) = \text{Cor}_{Info}(X,Y) = 0.44$.



# Difficulty of estimating $I(X;Y)$

Example with $\text{Cor}_{Pearson}(X,Y) = \text{Cor}_{Info}(X,Y) = 0.2$.



# 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:

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

# 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:

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

- Kernel density estimate approaches estimate $p(x, y)$ (Beirlant et al. 2001, Ivanov and Rozhkova 1981)
- Nearest neighbor estimators rely on distance-based computations (Mnatsakanov et al. 2008, Goria et al. 2005, Singh et. al. 2003)

# How to estimate $I(X;Y)$

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

- **Plug-in estimate**:

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

- Kernel density estimate approaches estimate $p(x, y)$ (Beirlant et al. 2001, Ivanov and Rozhkova 1981)
- Nearest neighbor estimators rely on distance-based computations (Mnatsakanov et al. 2008, Goria et al. 2005, Singh et. al. 2003)

# Problems in high dimensions

- Density estimation is known to have *exponential complexity* with respect to dimensionality.
  - E.g. to get the same precision, you need 10 observations for univariate $X, Y$ but 1000 for trivariate $\vec{X}, \vec{Y}$.

# Problems in high dimensions

- Density estimation is known to have *exponential complexity* with respect to dimensionality.
  - E.g. to get the same precision, you need 10 observations for univariate $X, Y$ but 1000 for trivariate $\vec{X}, \vec{Y}$.
- Many applications with high-dimensional $X$, $Y$.
  - Gene expression time series
  - Functional magnetic resonance imaging

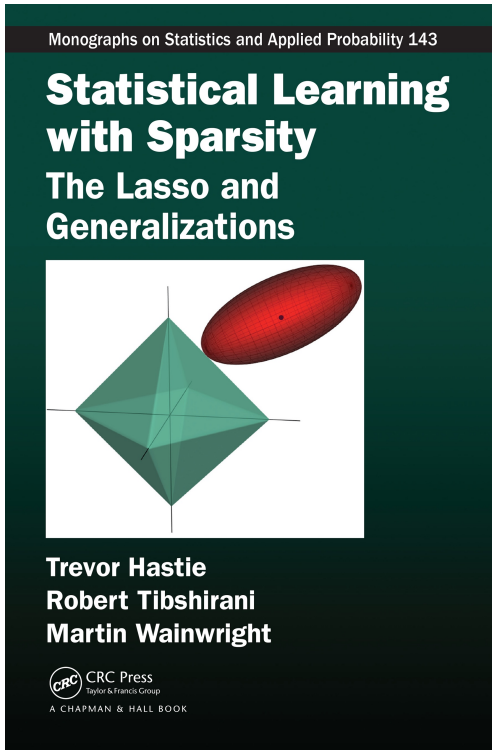
# Problems in high dimensions

- Density estimation is known to have *exponential complexity* with respect to dimensionality.
  - E.g. to get the same precision, you need 10 observations for univariate $X, Y$ but 1000 for trivariate $\vec{X}, \vec{Y}$.
- 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.

# Problems in high dimensions

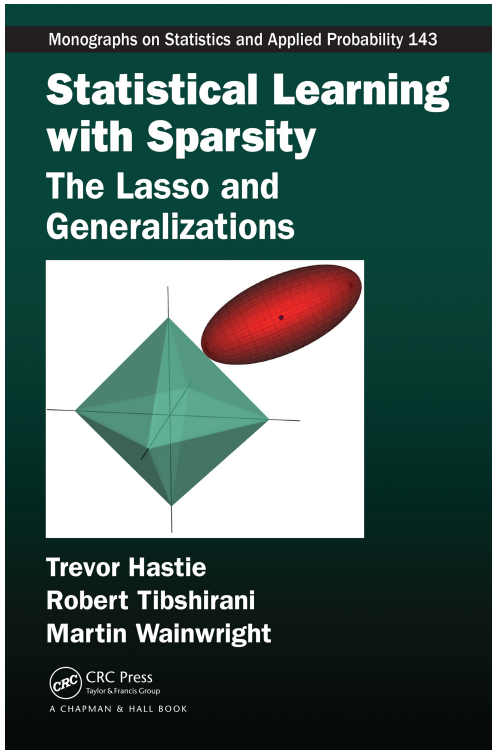
- Density estimation is known to have *exponential complexity* with respect to dimensionality.
  - E.g. to get the same precision, you need 10 observations for univariate $X, Y$ but 1000 for trivariate $\vec{X}, \vec{Y}$.
- Many applications with high-dimensional $X$, $Y$.
  - Gene expression time series
  - 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)

# First idea: Use sparsity!



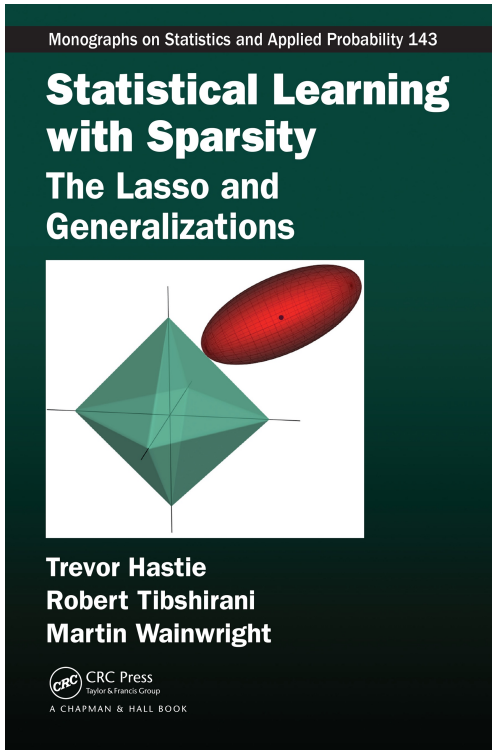

- *Sparsity* refers to existence of low-dimensional structure hidden in high-dimensional data.

# First idea: Use sparsity!



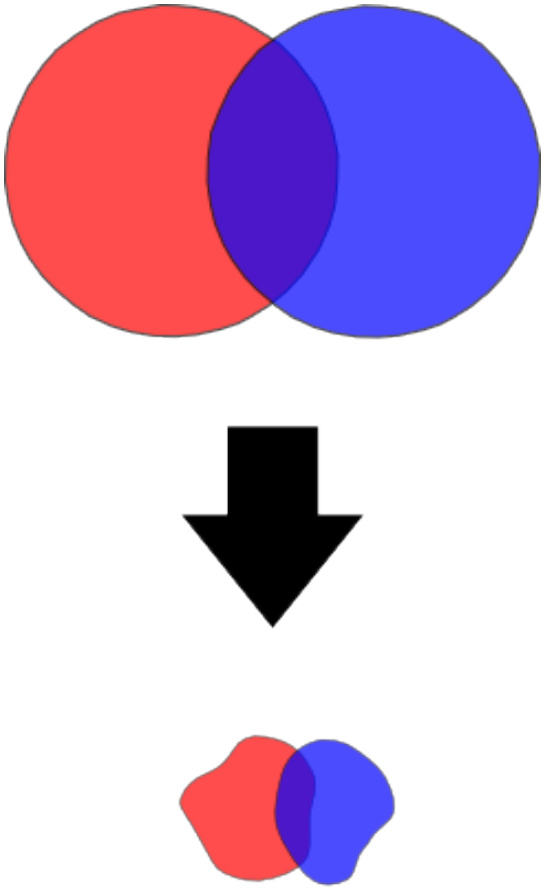

- *Sparsity* refers to existence of low-dimensional structure hidden in high-dimensional data.
- E.g. suppose $X$ is 100-dimensional but $Y$ is only a function of $(X_5, X_9)$.

# First idea: Use sparsity!



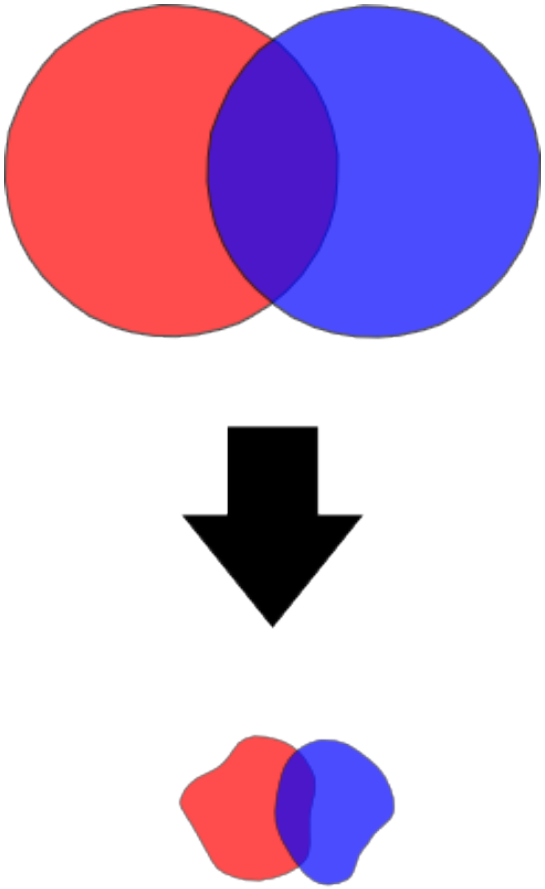

- *Sparsity* refers to existence of low-dimensional structure hidden in high-dimensional data.
- E.g. suppose $X$ is 100-dimensional but $Y$ is only a function of $(X_5, X_9)$.
- Can we exploit sparsity to obtain a good estimate of $I(X; Y)$ even under low sample sizes?

# Dimension reduction vs. sparsity?

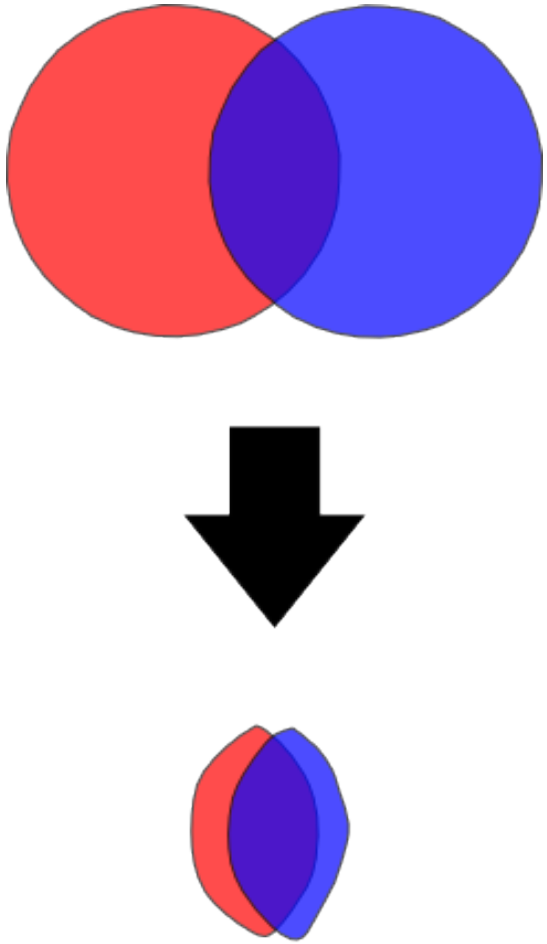


Unsupervised dimension reduction

# Dimension reduction vs. sparsity?



Unsupervised dimension reduction



Sparsity = supervised dim. reduction

# Second idea: link prediction accuracy to mutual information

- If $I(X;Y) > 0$, then $X$ carries information about $Y$ and vice-versa.

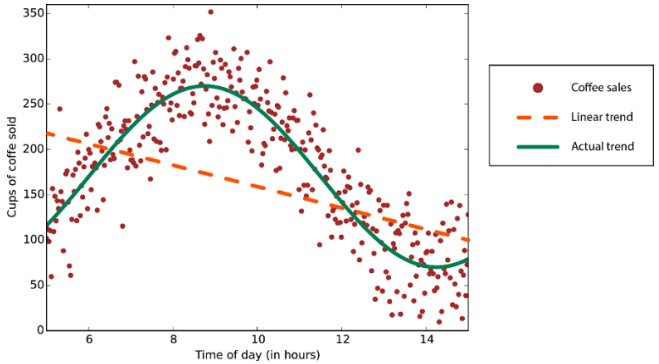
# Second idea: link prediction accuracy to mutual information

- If $I(X;Y) > 0$, then $X$ carries information about $Y$ and vice-versa.
- Therefore, we can *predict* $Y$ from $X$ (or $X$ from $Y$)

# Second idea: link prediction accuracy to mutual information

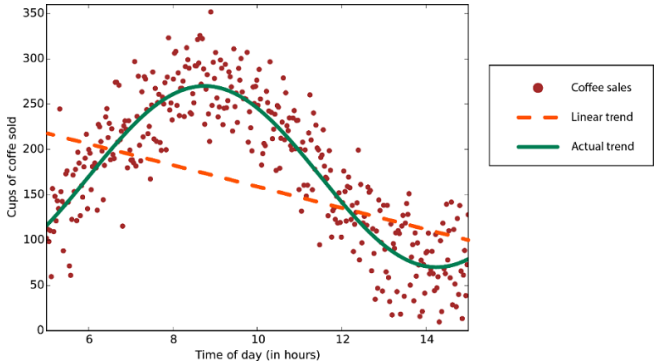
- If $I(X;Y) > 0$, then $X$ carries information about $Y$ and vice-versa.
- Therefore, we can *predict* $Y$ from $X$ (or $X$ from $Y$)
- We know that often *prediction accuracy* implies a lower bound for *mutual information* (e.g. Fano 1952)

# Background: Regression



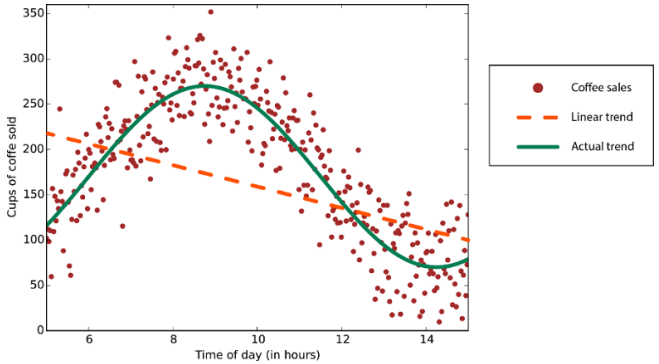

- Suppose you observe $(\vec{X}^{(i)}, Y^{(i)})_{i=1}^{n}$ where $Y^{(i)} = f(\vec{X}^{(i)}) + \epsilon$, where $f$ is an unknown function and $\epsilon$ is noise. (Also, assume $\mathbf{E}[\epsilon] = 0$.)

# Background: Regression



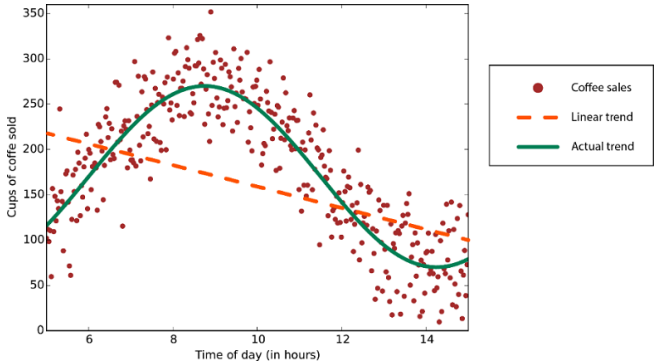

- Suppose you observe $(\vec{X}^{(i)}, Y^{(i)})_{i=1}^n$ where $Y^{(i)} = f(\vec{X}^{(i)}) + \epsilon$, where $f$ is an unknown function and $\epsilon$ is noise. (Also, assume $\mathbf{E}[\epsilon] = 0$.)
- The goal in regression is to recover the unknown function $f$.

# Background: Regression




- Suppose you observe $(\vec{X}^{(i)}, Y^{(i)})_{i=1}^n$ where $Y^{(i)} = f(\vec{X}^{(i)}) + \epsilon$, where $f$ is an unknown function and $\epsilon$ is noise. (Also, assume $\mathbf{E}[\epsilon] = 0$.)
- The goal in regression is to recover the unknown function $f$.
- In *linear regression*, we assume $f$ is linear.

# Background: Regression




- Suppose you observe $(\vec{X}^{(i)}, Y^{(i)})_{i=1}^{n}$ where $Y^{(i)} = f(\vec{X}^{(i)}) + \epsilon$, where $f$ is an unknown function and $\epsilon$ is noise. (Also, assume $\mathbf{E}[\epsilon] = 0$.)
- The goal in regression is to recover the unknown function $f$.
- In *linear regression*, we assume $f$ is linear.
- if we do not assume a particular form for $f$, we can use *nonparametric regression*.

# Background: Sparse regression

- When $\vec{X}$ is high dimensional, classical regression techniques perform poorly.

# Background: Sparse regression

- When $\vec{X}$ is high dimensional, classical regression techniques perform poorly.
- If the true function $f$ only depends on a small number of components in $\vec{X}$, we can still do well if we use *sparse* regression methods.

# Background: Sparse regression

- When $\vec{X}$ is high dimensional, classical regression techniques perform poorly.
- If the true function $f$ only depends on a small number of components in $\vec{X}$, we can still do well if we use *sparse* regression methods.

| | *Classical* | *Sparse* |
|---|---|---|
| *Linear* | Ordinary Least-Squares (Gauss 1975?) | Elastic net (Zou 2008) |
| *Nonpar.* | LOWESS (Cleveland 1979) | Random forests (Breiman 2001) |

# 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. Assess the *prediction accuracy* of the model using *identification risk*
3. Use the identification risk to obtain a lower bound for the mutual information $I(X; Y)$

# 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 2008),

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

- 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:

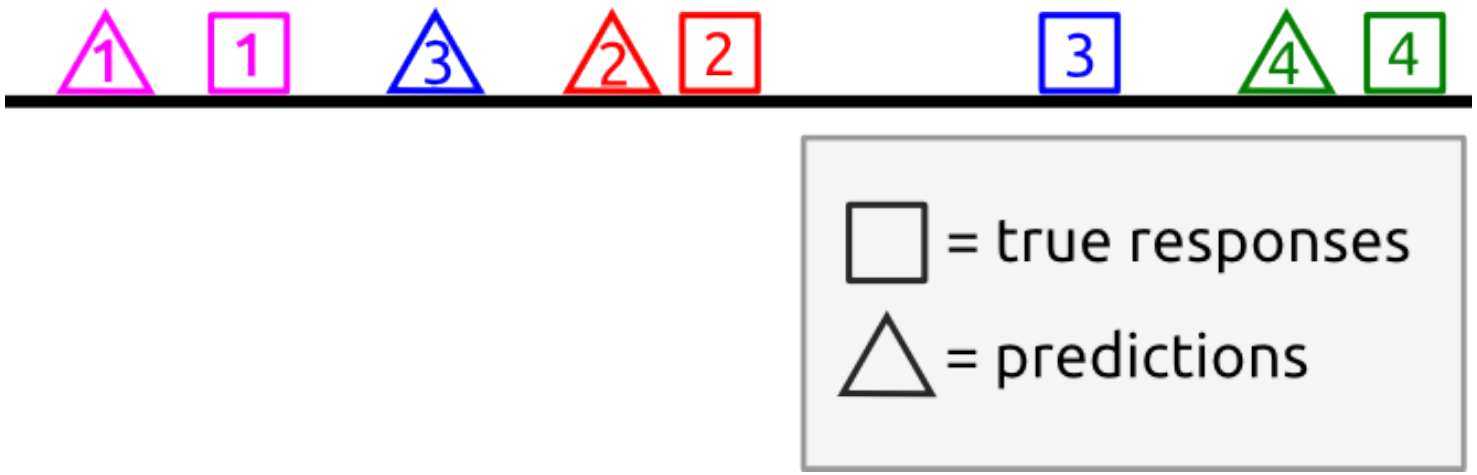
- Regression (mean squared-error) loss:

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

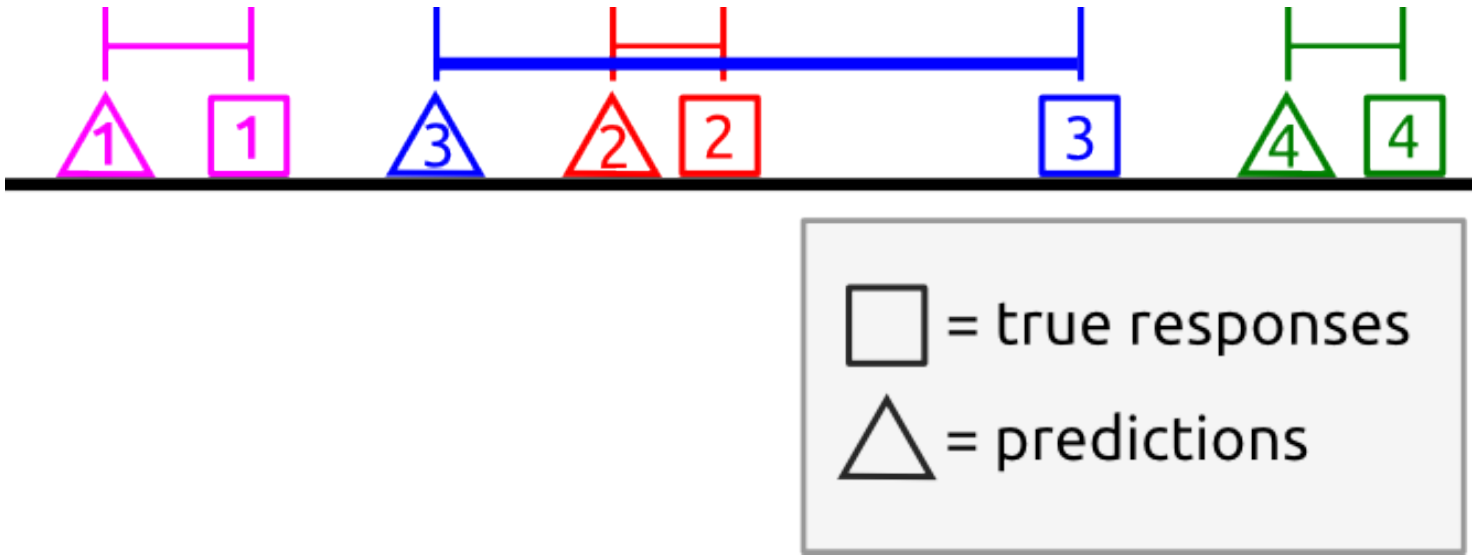
- Identification loss (Kay 2008):

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

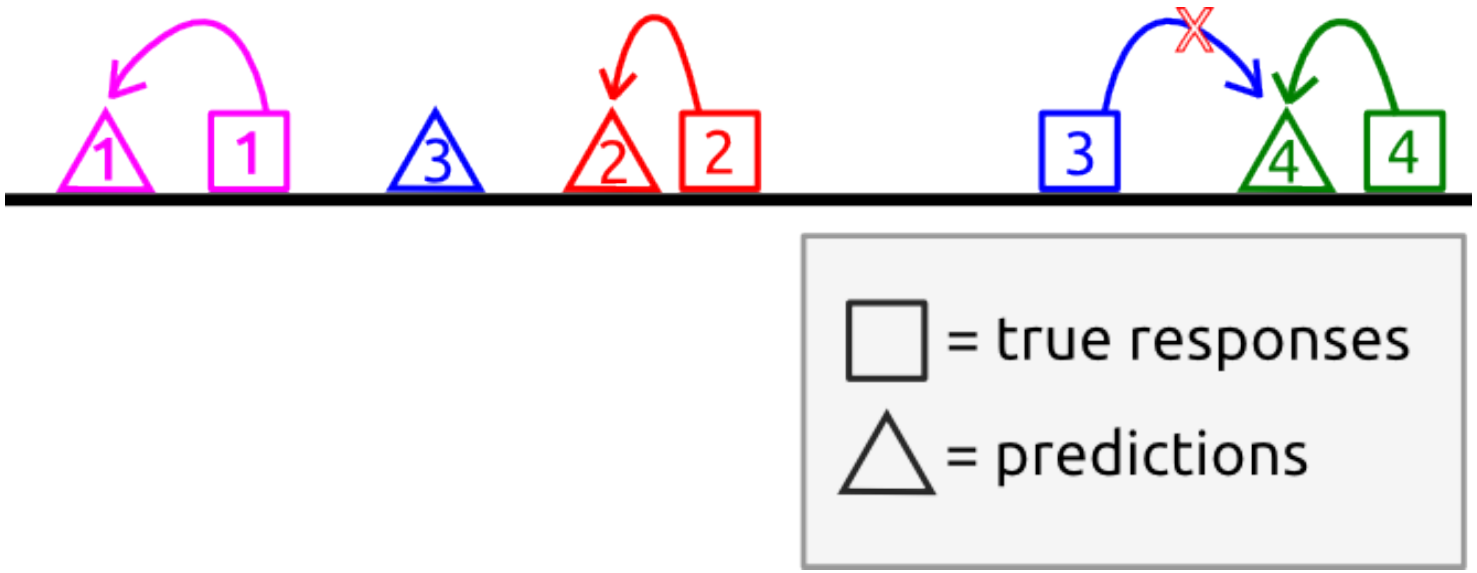
# Regression vs Identification loss



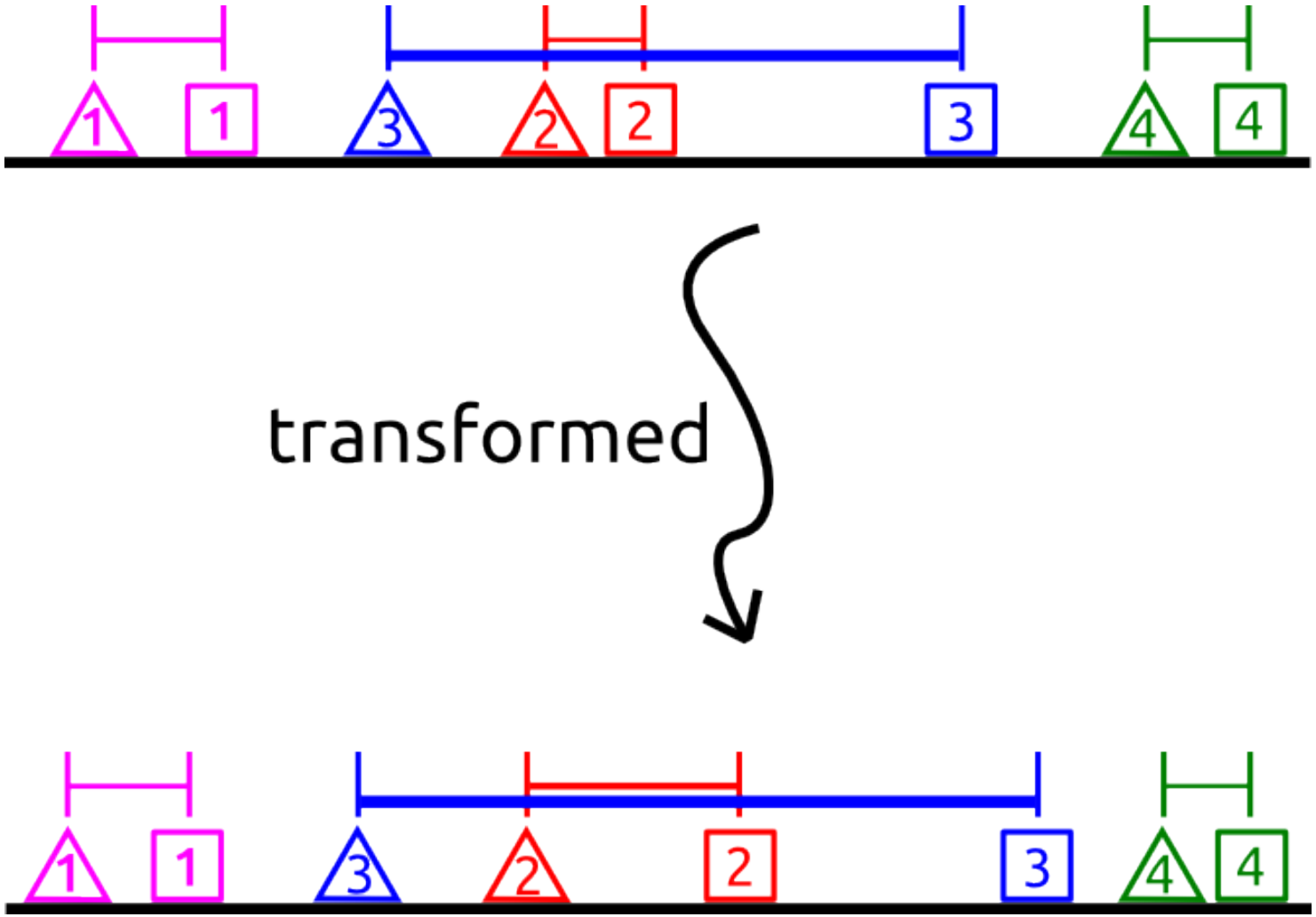
# Mean-squared error



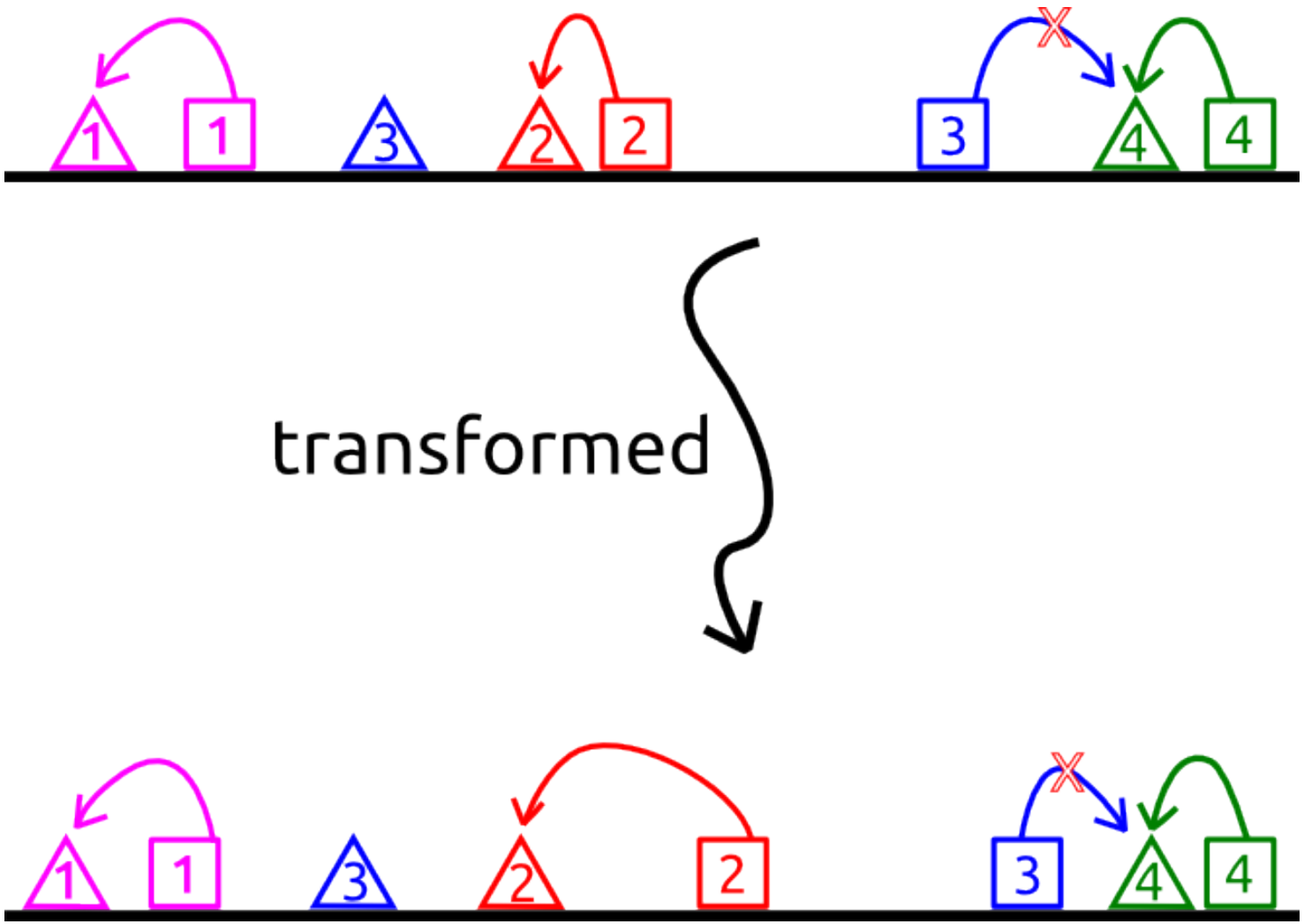
# Identification loss



# Mean-squared error changes under nonlinear scaling



# Identification loss robust under nonlinear scaling



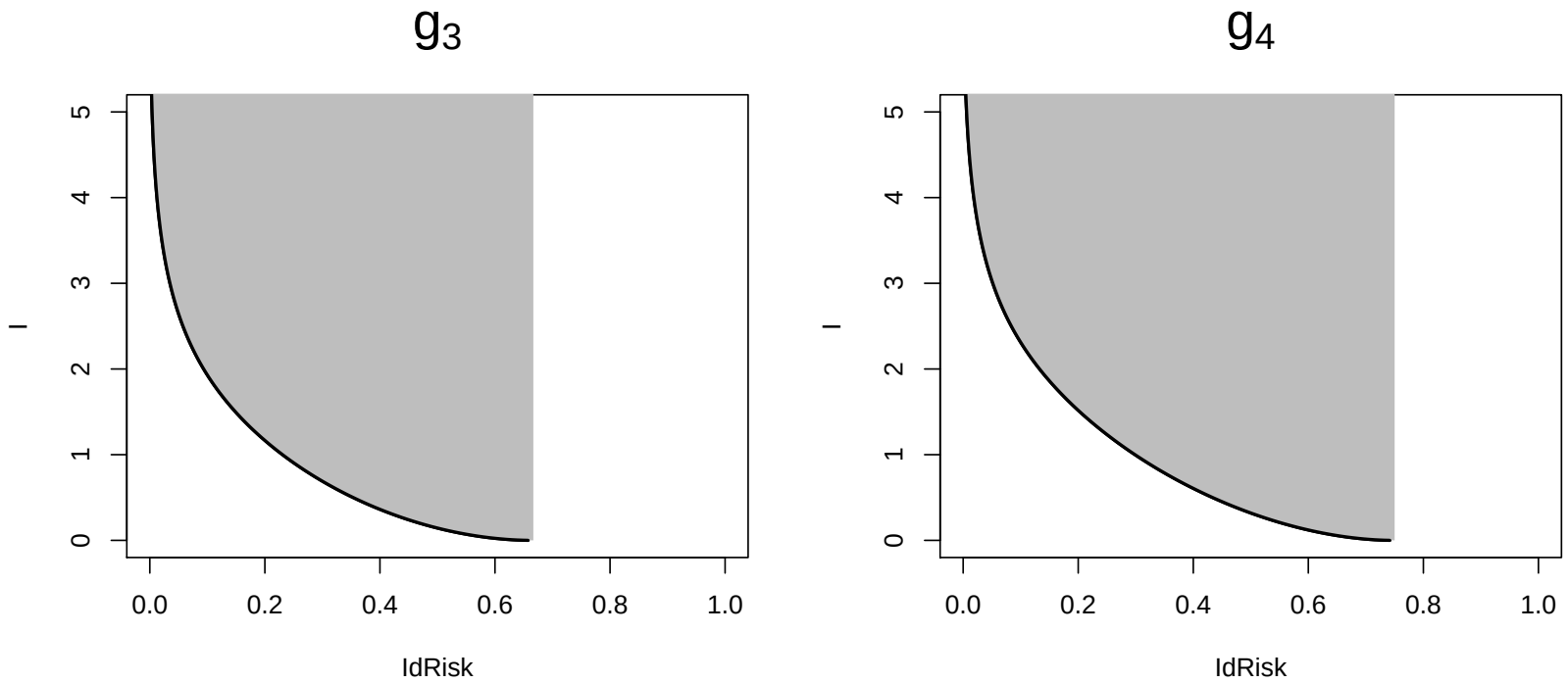
# Identification loss and mutual information

- Define the identification risk as the expected identification loss

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

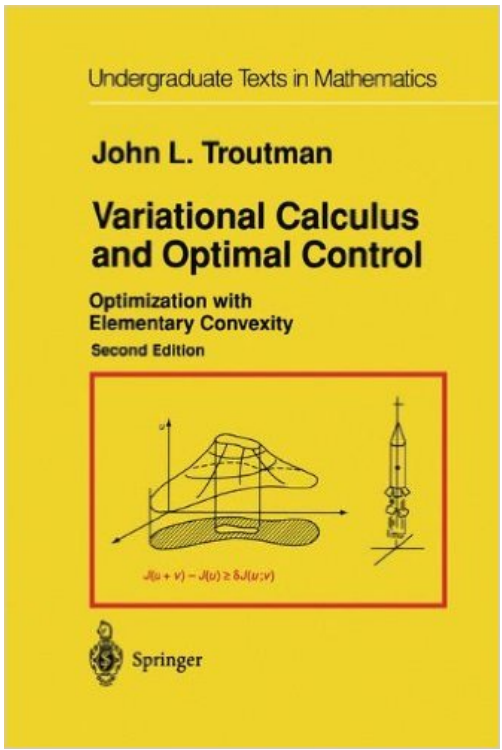
- **Theorem.** (Z., Benjamini 2017) There exists a function $g_k$ such that

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



# Proof details

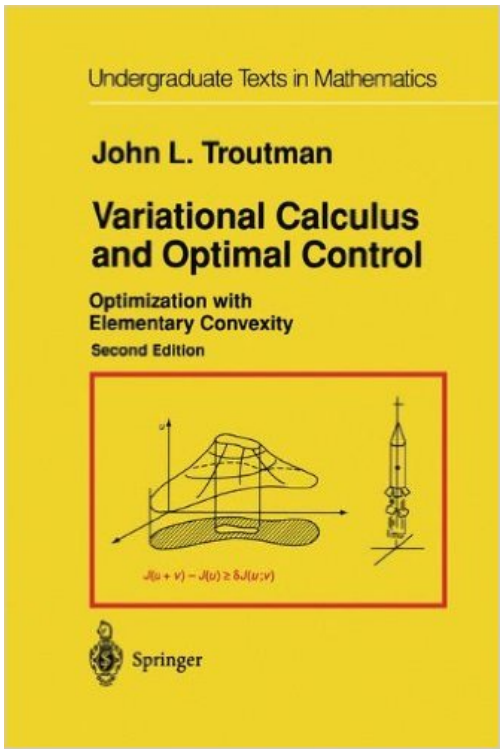
We use *calculus of variations* to obtain this result.



# Proof details

We use *calculus of variations* to obtain this result.

- Mutual information is a functional of $p(x, y)$.



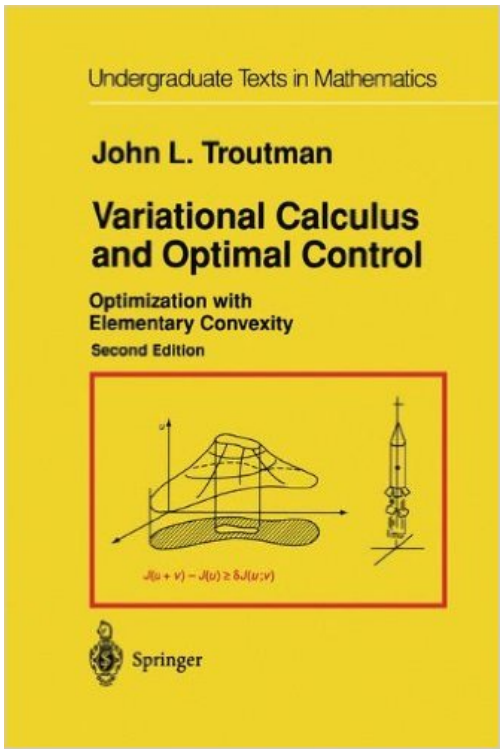

$$\mathsf{I}[p(x,y)] = \mathbf{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right].$$

# Proof details

We use *calculus of variations* to obtain this result.

- Mutual information is a functional of $p(x, y)$.

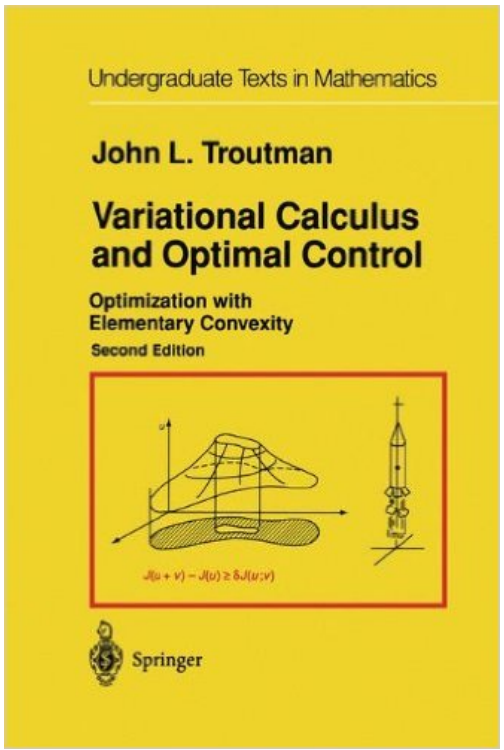
$$\mathrm{I}[p(x,y)] = \mathbf{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right].$$




- Identification risk is *lower-bounded* by another functional–the *Bayes risk*.

$$\mathrm{BayesRisk}_k[p(x,y)] = 1 - \mathbf{E}[\max_{i=1}^{k} p(Y|X_i)].$$

# Proof details




We use *calculus of variations* to obtain this result.

- Mutual information is a functional of $p(x, y)$.

$$\mathrm{I}[p(x,y)] = \mathbf{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right].$$

- Identification risk is *lower-bounded* by another functional–the *Bayes risk*.

$$\mathrm{BayesRisk}_k[p(x,y)] = 1 - \mathbf{E}[\max_{i=1}^{k} p(Y|X_i)].$$

- $g_k(u) = \inf_{p(x,y)} I[p(x,y)]$

  subject to $\mathrm{BayesRisk}_k[p(x,y)] \geq u$.

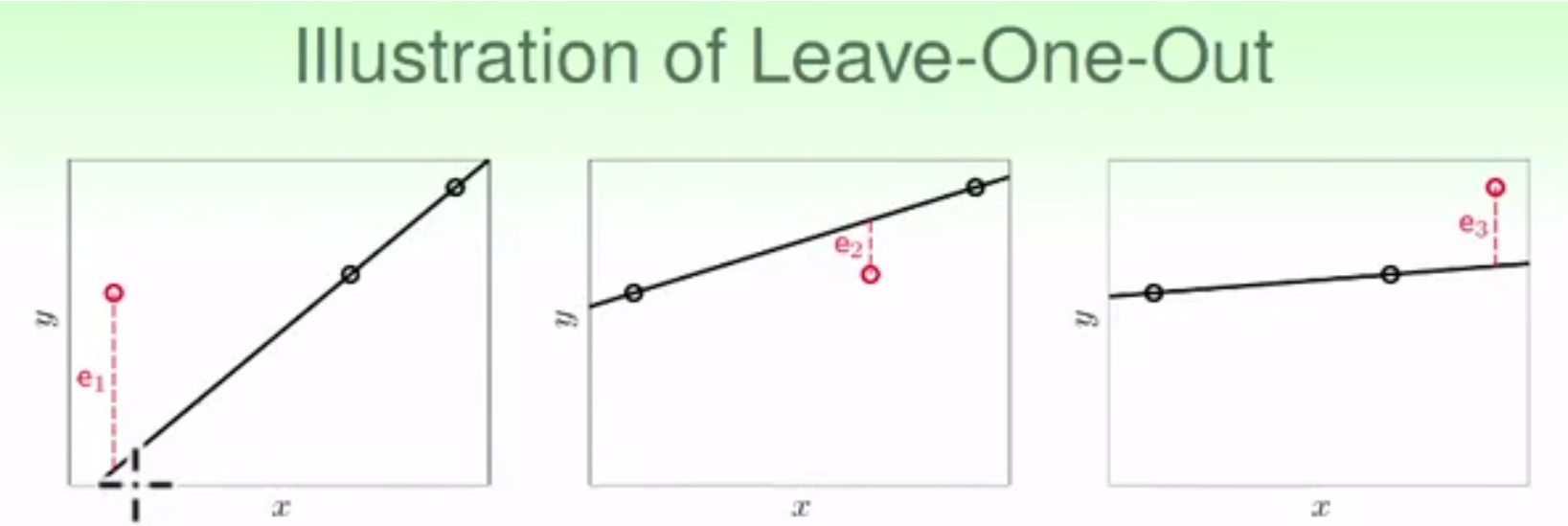
# Our proposal

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. Compute *identification loss*, $\text{IdLoss}_k$, using *leave-k-out*.
3. Estimate mutual information using

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

# What is leave-k-out cross-validation?




- Randomly hold out a subset of size $k$.
- Use remaining data to predict the held-out data.
- Obtain the average prediction error.

Image credit Hsuan-Tien Lin

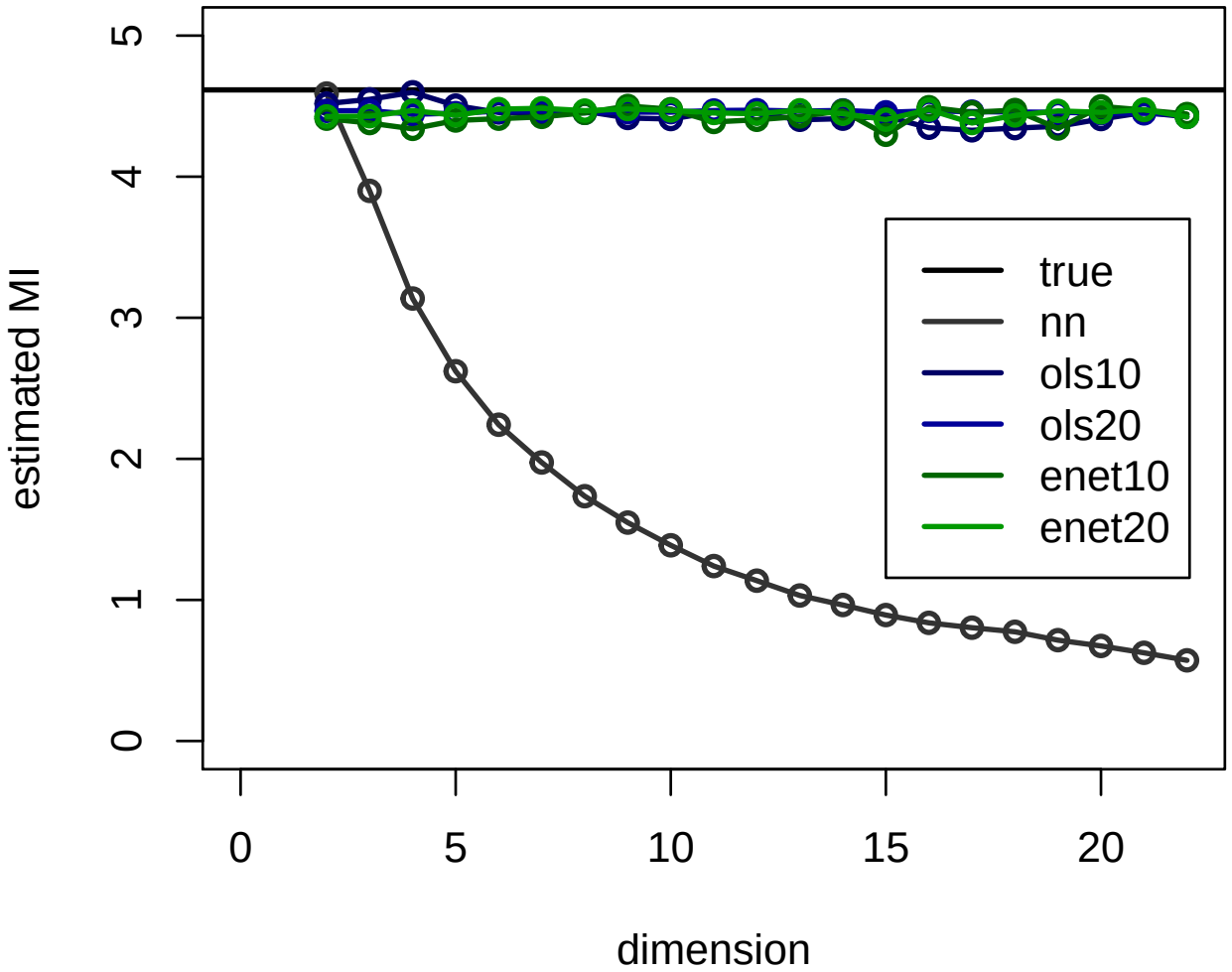
Section 2

# Applications

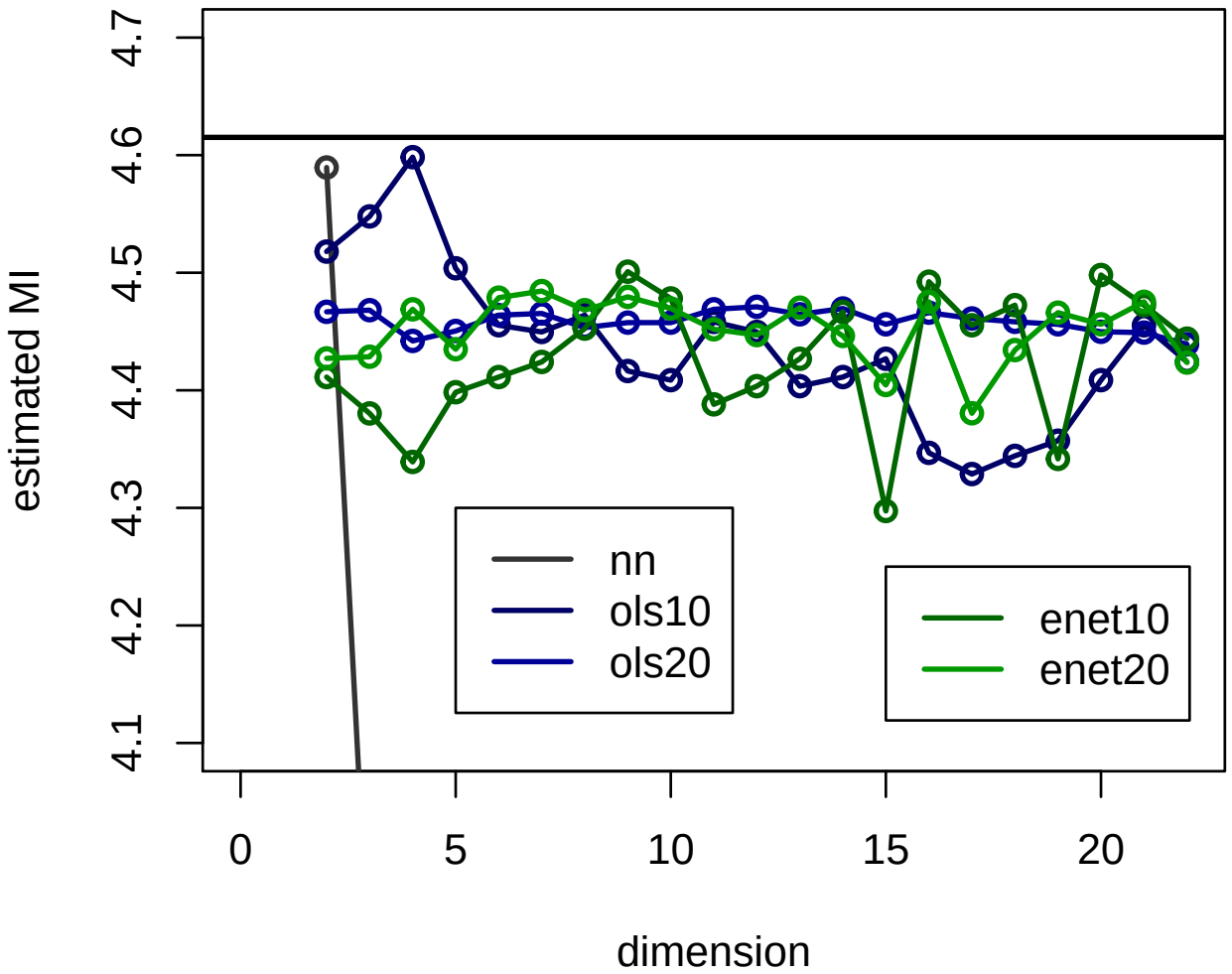
# Simulation

- Generate data: $(Y_1, Y_2) = (X_1, X_2)^T B + \epsilon$ where $B$ is a randomly generated coefficient matrix.
- Add extra noise dimensions $X_3, X_4, \ldots$.
- $n = 1000$.
- Compare Nearest-Neighbor estimator (Mnatsakov et al, 2008, implemented in `FNN`) with our method using OLS and elastic net (sparse).

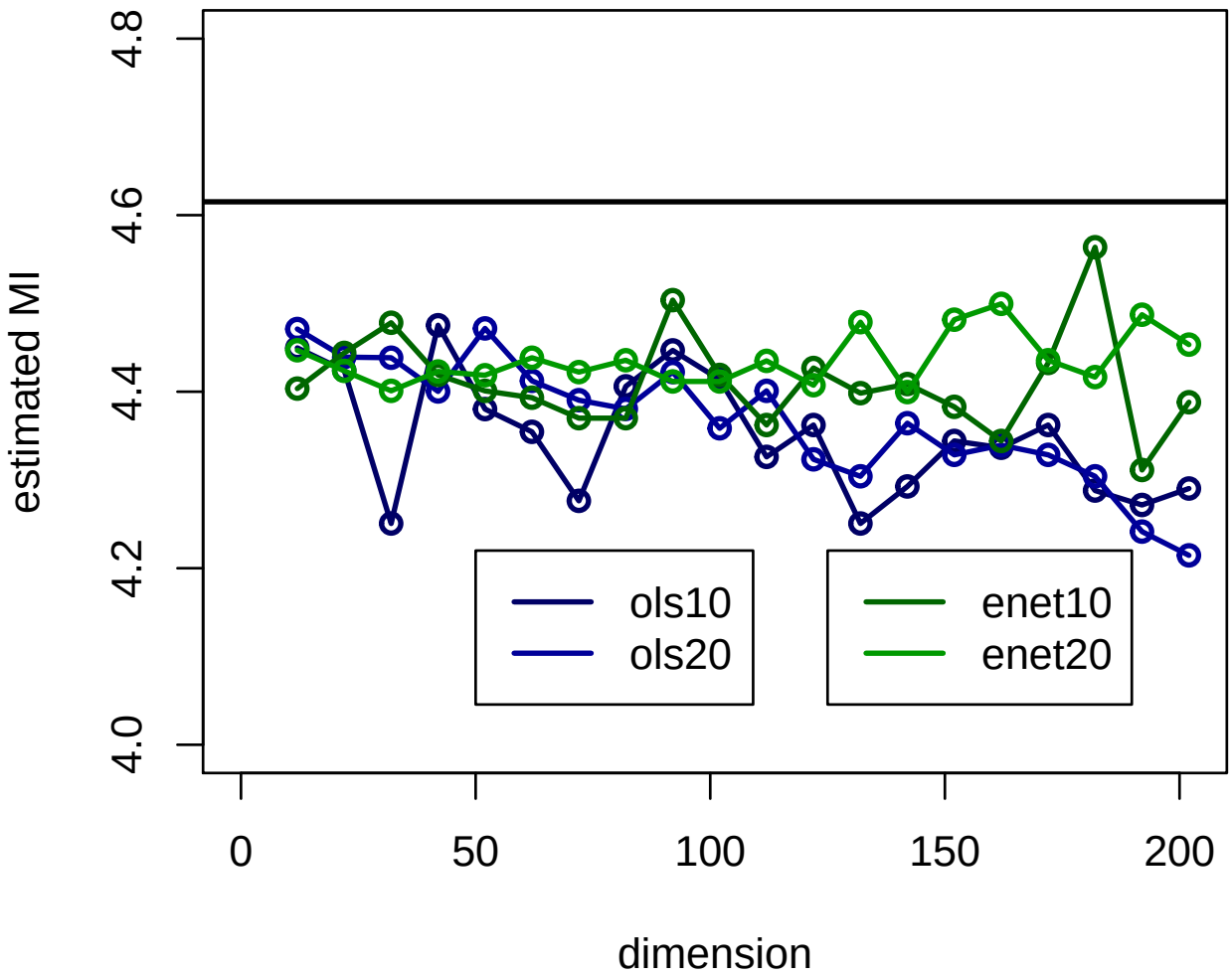
# Simulation Results - I. low dimension



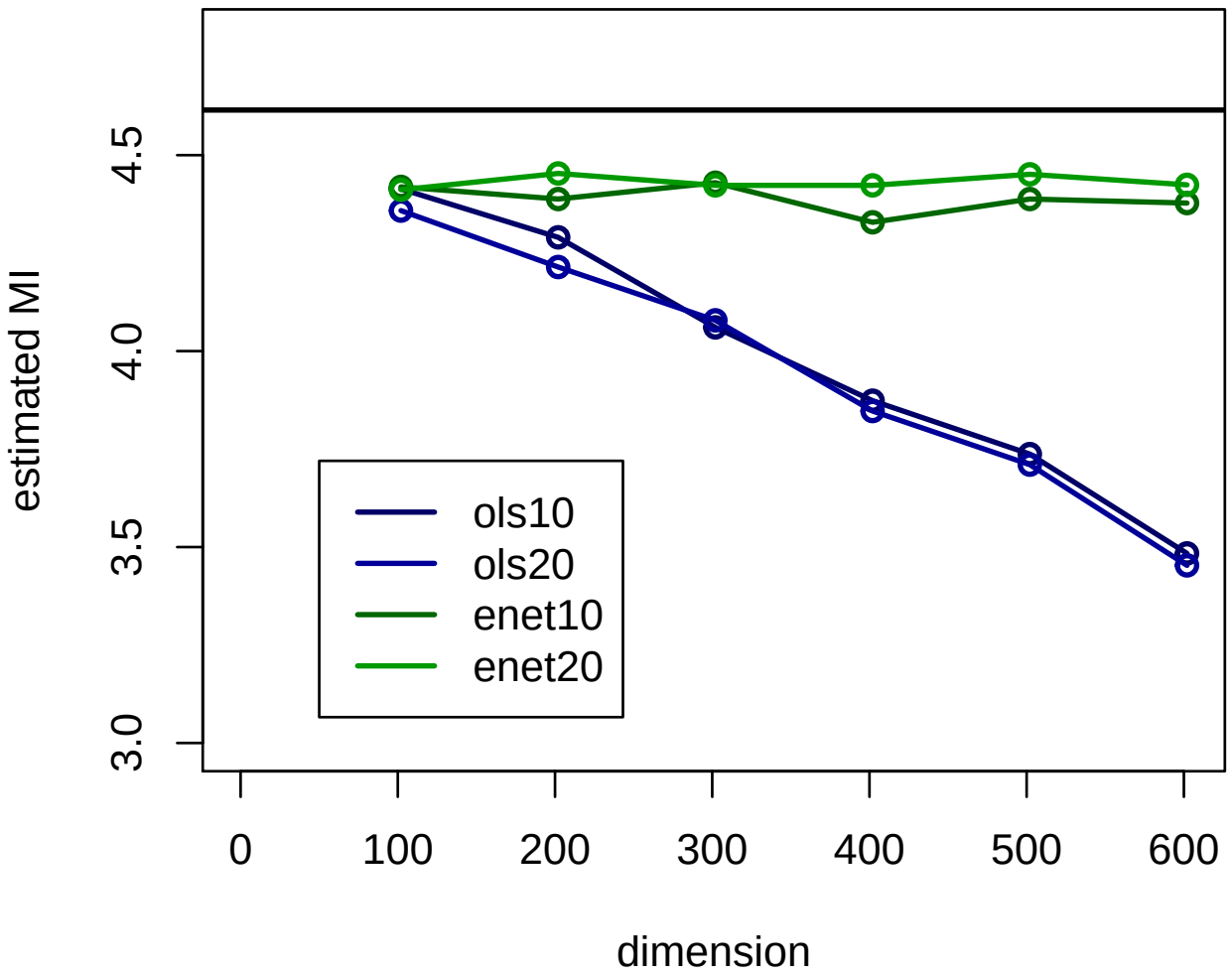
# Simulation Results - I. low dimension



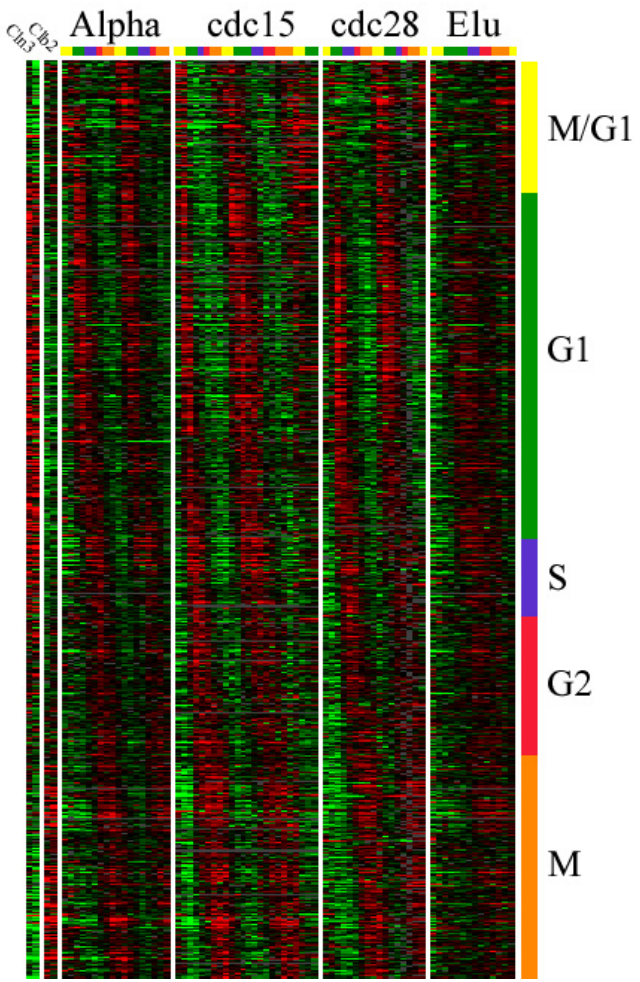
# Simulation Results - II. medium dimension



# Simulation Results - III. high dimension



# Application to gene expression time series




- Data from Spellman et al. 1998
- Expression levels of 6178 yeast genes during cell cycle
- Total 73 time points per gene

# Groups of genes

| Group | No. genes |
|---|---|
| unknown | 396 |
| cell cycle | 27 |
| DNA replication | 27 |
| transport | 19 |
| cytoskeleton | 17 |
| chromatin structure | 16 |

Total 145 different categories (only top 6 shown).

# Canonical correlations between time series

Top canonical correlation (Hotelling 1936)

| | CC | DR | Tr | Cy | CS |
|---|---|---|---|---|---|
| CC | | 1 | 1 | 1 | 1 |
| DR | | | 1 | 0.99 | 0.99 |
| Tr | | | | 0.99 | 0.98 |
| Cy | | | | | 0.98 |
| CS | | | | | |

CC = cell cycle, DR = DNA replication, Tr = transport,
Cy = cytoskeleton, CS = chromatin structure

# Sparse canonical correlations between time series

Using sparse CCA* (Witten and Tibshirani 2009).

| | CC | DR | Tr | Cy | CS |
|---|---|---|---|---|---|
| CC | | 0.96 | 0.87 | 0.92 | 0.94 |
| DR | | | 0.83 | 0.88 | 0.95 |
| Tr | | | | 0.83 | 0.78 |
| Cy | | | | | 0.90 |
| CS | | | | | |

CC = cell cycle, DR = DNA replication, Tr = transport,
Cy = cytoskeleton, CS = chromatin structure

*: using `CCApermute` in R package `PMA`

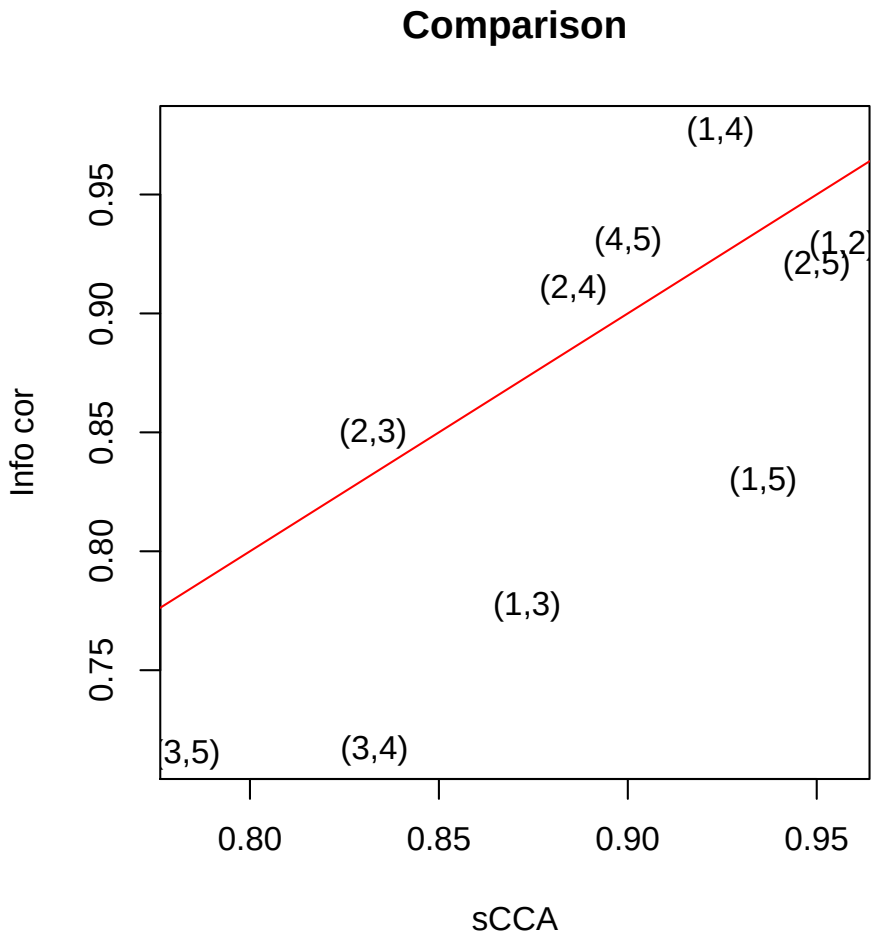
# Information correlations between time series

Taking the max of $\hat{I}(X;Y)$ and $\hat{I}(Y;X)$.

| | CC | DR | Tr | Cy | CS |
|---|---|---|---|---|---|
| CC | | 0.93 | 0.78 | 0.98 | 0.83 |
| DR | | | 0.85 | 0.91 | 0.92 |
| Tr | | | | 0.72 | 0.71 |
| Cy | | | | | 0.93 |
| CS | | | | | |

CC = cell cycle, DR = DNA replication, Tr = transport,
Cy = cytoskeleton, CS = chromatin structure

# Comparing sparse CCA and $Cor_{Info}$



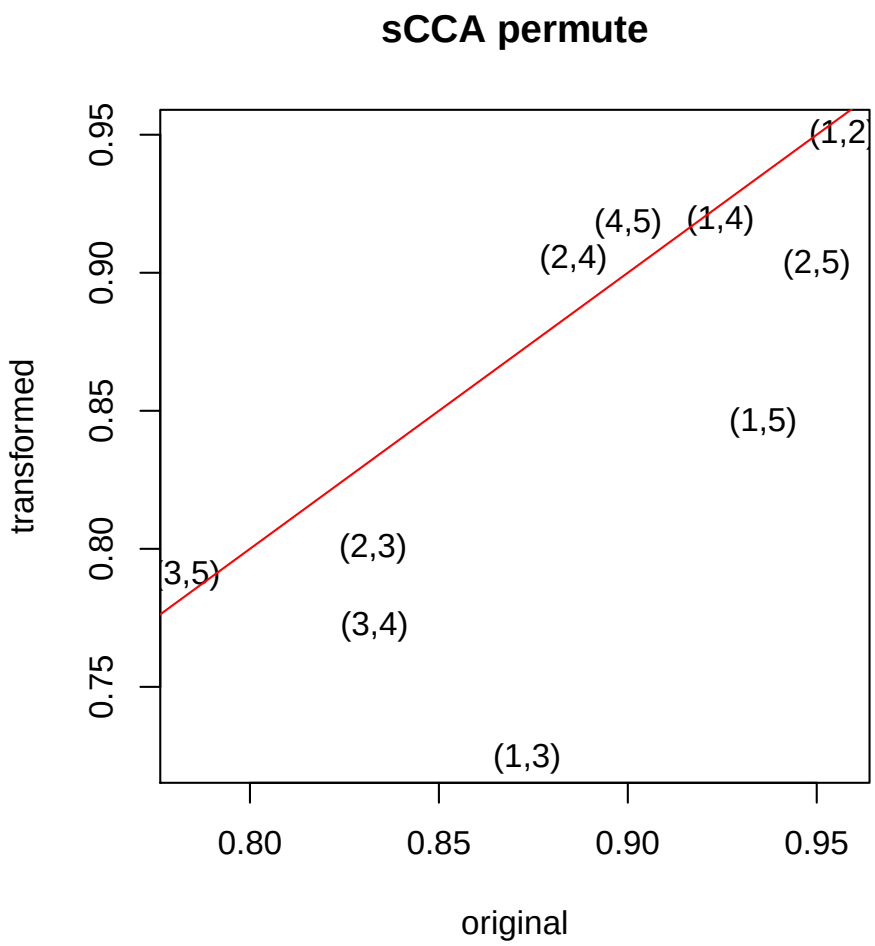

(1) cell cycle, (2) DNA replication, (3) transport,
(4) cytoskeleton, (5) chromatin structure

# Invariance properties

Transform data from each group with random rotation...

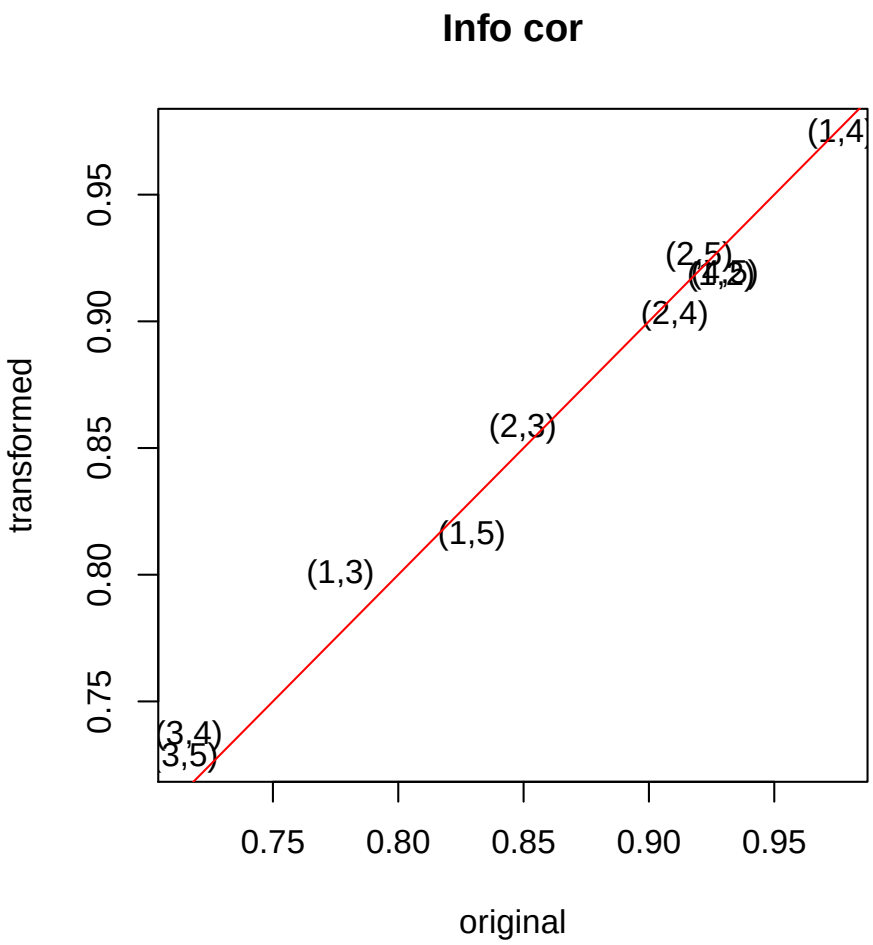
# Invariance properties

Transform data from each group with random rotation...



# Invariance properties

Transform data from each group with random rotation...



# Conclusions

- Mutual information, and derived $\text{Cor}_{Info}$ are useful measures of correlation, but hard to estimate.

# Conclusions

- Mutual information, and derived $\text{Cor}_{Info}$ are useful measures of correlation, but hard to estimate.
- Our method targets high-dimensional data with sparsity.

# Conclusions

- Mutual information, and derived $\text{Cor}_{Info}$ are useful measures of correlation, but hard to estimate.
- Our method targets high-dimensional data with sparsity.
- How to use: choose a regression model suited to the model assumptions. Our method allows you to convert the prediction accuracy of the model, $\text{IdLoss}_k$ into an estimate of $I(\vec{X}; \vec{Y})$.

# Conclusions

- Mutual information, and derived $\text{Cor}_{Info}$ are useful measures of correlation, but hard to estimate.
- Our method targets high-dimensional data with sparsity.
- How to use: choose a regression model suited to the model assumptions. Our method allows you to convert the prediction accuracy of the model, $\text{IdLoss}_k$ into an estimate of $I(\vec{X}; \vec{Y})$.
- Example application: measure of joint information between two tables which is robust to transformations.

# Related work and future directions

- What if data is high-dimensional, but not sparse? We have another method based on high-dimensional asymptotics (ZB 2016).

# Related work and future directions

- What if data is high-dimensional, but not sparse? We have another method based on high-dimensional asymptotics (ZB 2016).
- Estimating quantities related to mutual information, such as *transfer information*, *stimulus-specific information* and *redundancy* (Borst and Theunissen 1999)

# Related work and future directions

- What if data is high-dimensional, but not sparse? We have another method based on high-dimensional asymptotics (ZB 2016).
- Estimating quantities related to mutual information, such as *transfer information*, *stimulus-specific information* and *redundancy* (Borst and Theunissen 1999)
- Inferring resting-state brain networks.

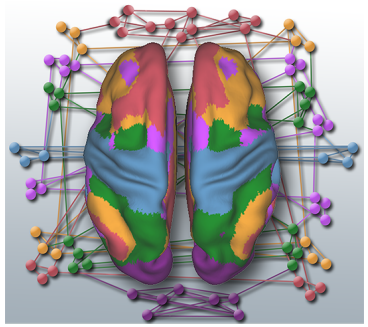


Image credit Simons Foundation

Section 3

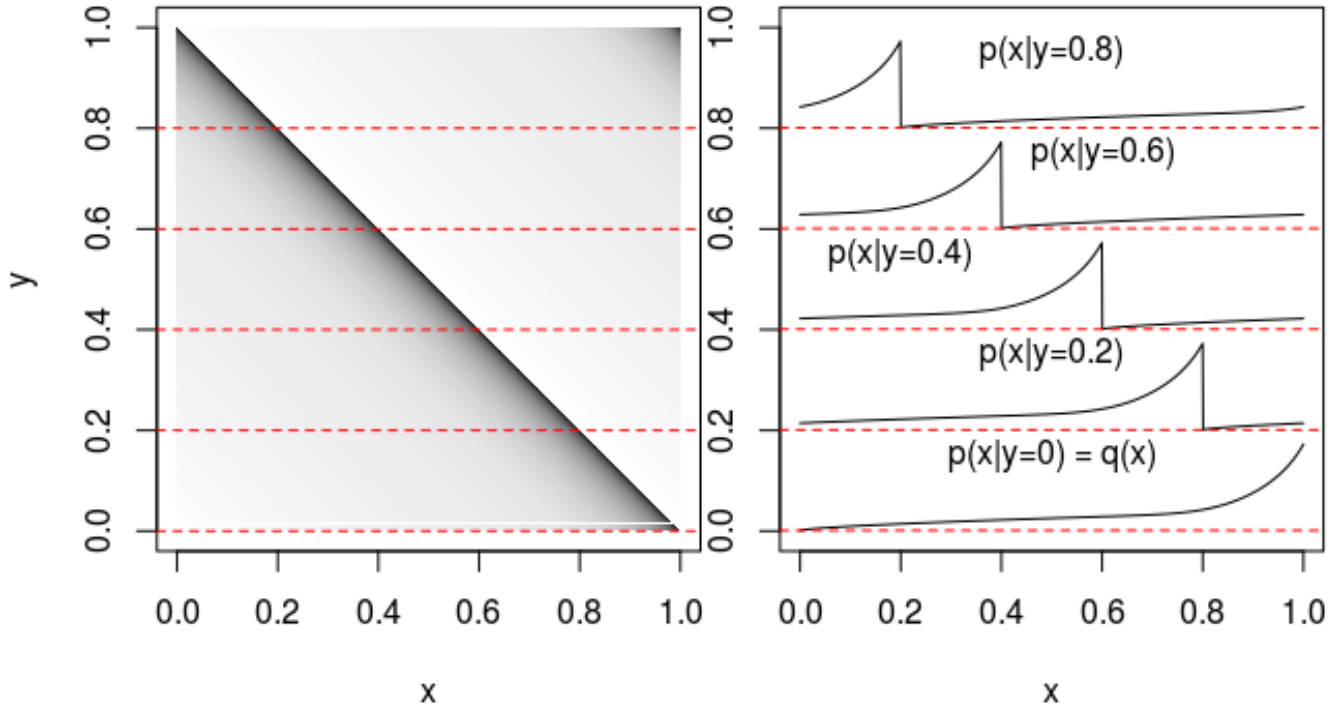
# The End

# References


- Reshef et al, 2011. "Detecting Novel Associations in Large Datasets." *Science*.
- Speed, 2011. "A correlation for the 21st century." *Science*.
- Linfoot, 1957. "An informational measure of correlation." *Information and Control*.
- Kay, 2008. "Identifying natural images from human brain activity." *Nature*.
- Mnatsakanov, et al, (2008). "K-nearest neighbor estimators of entropy." *Mathematical Methods of Statistics*
- Spellman et al., (1998). "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization." *Molecular Biology of the Cell*.
- Hotelling, H. (1936). "Relations Between Two Sets of Variates". *Biometrika.*
- Witten, Daniela M., and Robert J. Tibshirani. (2009). "Extensions of sparse canonical correlation analysis with applications to genomic data." *Statistical applications in genetics and molecular biology*

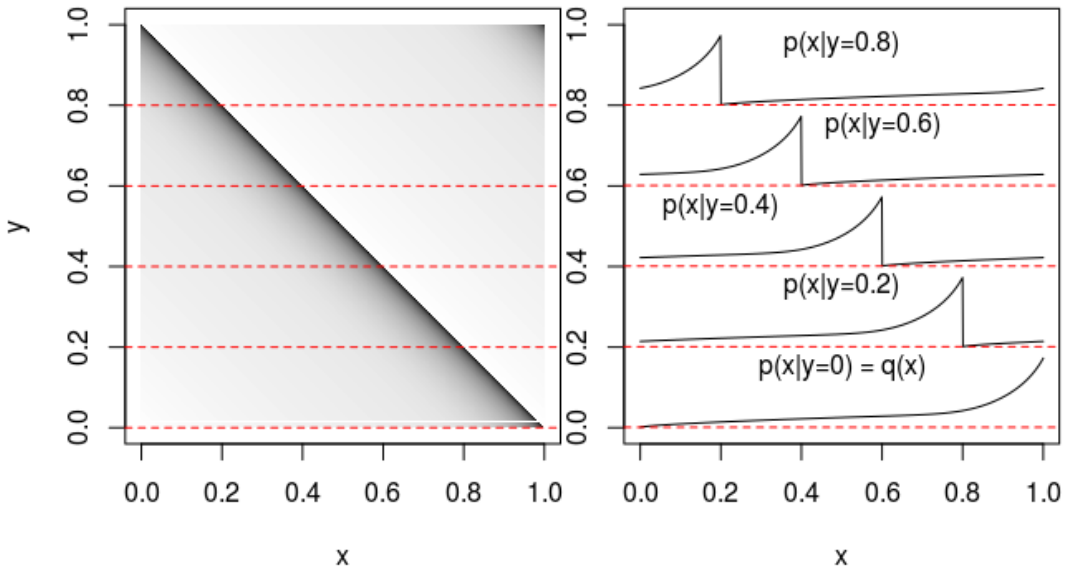
# Reduced Problem

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).

# Reduced Problem




- $p(x, y)$ 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 $q(x)$.

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

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

# Simplified formulae

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

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

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

# Optimization problem

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

- *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, $\text{I}[q(x)] \leq \iota$ and $\int_0^1 q(x)dx = 1$.

Hence the problem is

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

# Optimization problem

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

- Does a solution exist? *Yes*, because the space of measures with density $q(x)$ satisfying $\mathsf{I}[q(x)] \leq \iota$ 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

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

# 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

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

- 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,

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