Estimating mutual information for high-dimensional sparse relationships

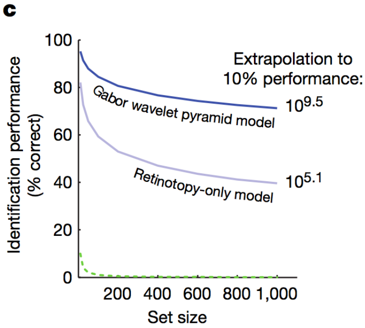
Charles Zheng

Stanford University

January 24, 2017

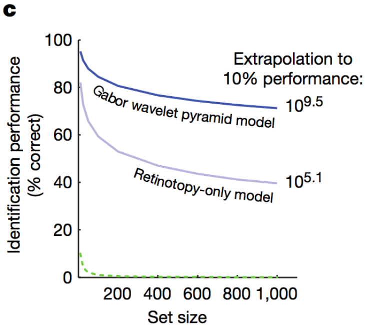
(Joint work with Yuval Benjamini, Hebrew University.)

Introduction



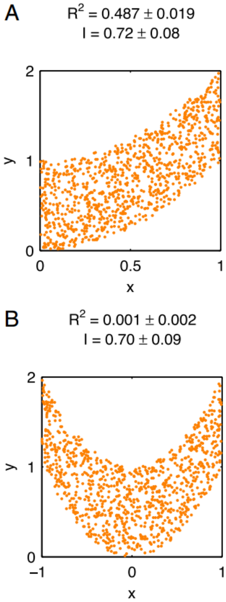
- Much of my work has been inspired by use of machine learning in encoding/decoding models in fMRI (Kay et al. 2008, Nishimoto et al. 2011)

Introduction



- Much of my work has been inspired by use of machine learning in encoding/decoding models in fMRI (Kay et al. 2008, Nishimoto et al. 2011)
- E.g.: Extrapolating classification accuracy curves (Z., Achanta, and Benjamini 2016)

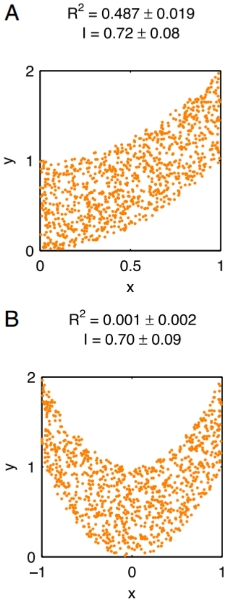
This talk



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

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

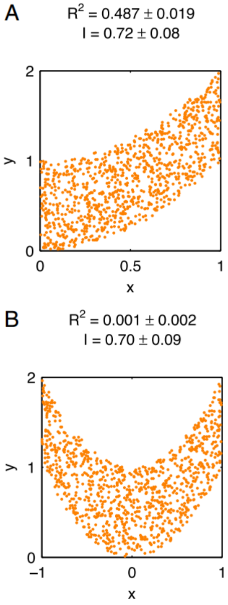
This talk



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)

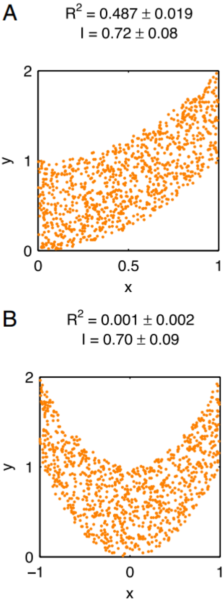
This talk



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

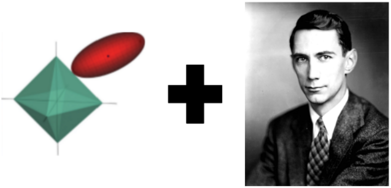
This talk



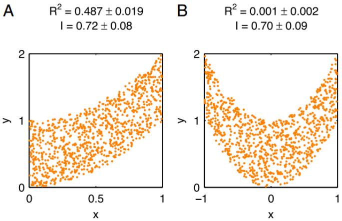
Mutual information $I(\vec{X}; \vec{Y})$

- 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"

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

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 $I(X; Y)$

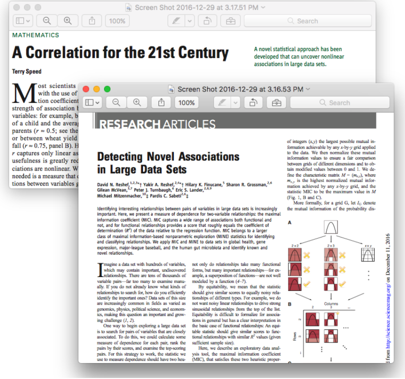
Mutual information has since been applied to many areas outside of information theory

Applications [edit]

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

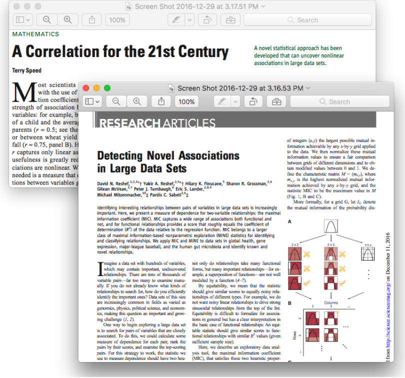
- In search engine technology, mutual information between phrases
- In telecommunications, the channel capacity is equal to the mutua
- Discriminative training procedures for hidden Markov models have
- RNA secondary structure prediction from a multiple sequence alig
- Phylogenetic profiling prediction 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



- 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.

Problems with mutual information

- Hard to interpret (compared to R^2)

Problems with mutual information

- Hard to interpret (compared to R^2)
    - Define the "informational correlation" (Linfoot 1957)

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

Problems with mutual information

- Hard to interpret (compared to R^2)
  - Define the "informational correlation" (Linfoot 1957)

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

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

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

Problems with mutual information

- Hard to interpret (compared to R^2)
  - Define the "informational correlation" (Linfoot 1957)

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

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

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

- Hard to estimate (compared to R^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:

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

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:

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

- 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

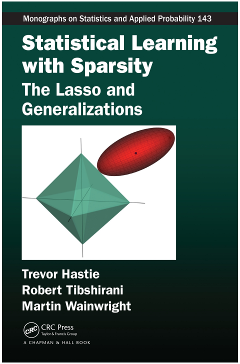
Problems in high dimensions

- 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

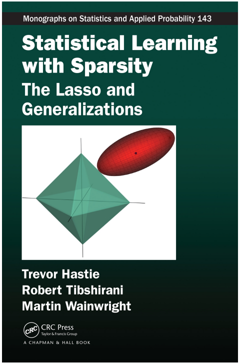
- Density estimation is known to have *exponential complexity* with respect to dimensionality.
  - E.g. to get the same precision, you need 10 observations for univariate X, Y but 1000 for trivariate $\vec{X}, \vec{Y}$.
- Many applications with high-dimensional X, Y.
  - Gene expression time series
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

New idea: Use sparsity!



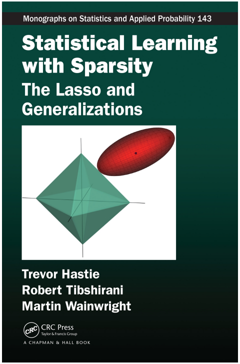
- *Sparsity* refers to existence of low-dimensional structure hidden in high-dimensional data.

New 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).

New 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?

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 $\mathbf{E}[\vec{Y}|\vec{X}]$.
2. Assess the *prediction accuracy* of the model using *identification loss* (Kay et al. 2008)
3. Use the identification loss to obtain a lower bound for the mutual information $I(X; Y)$

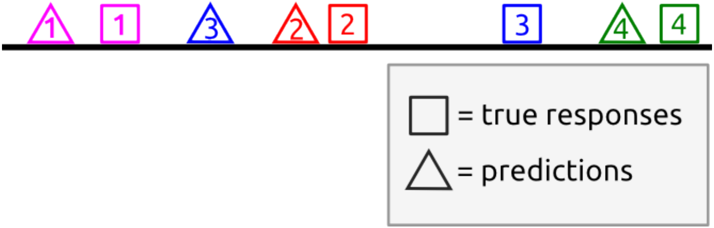
Multiple-response regression

- 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),

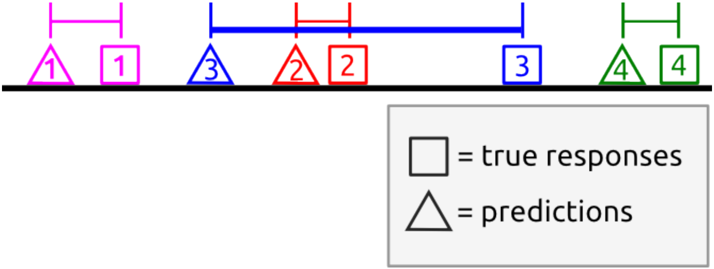
$$\hat{\beta}^{(i)} = \text{argmin}_\beta ||\boldsymbol{X}^T\beta^{(i)} - Y^{(i)}||^2 + \lambda_2||\beta^{(i)}||_2^2 + \lambda_1||\beta^{(i)}||_1$$

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

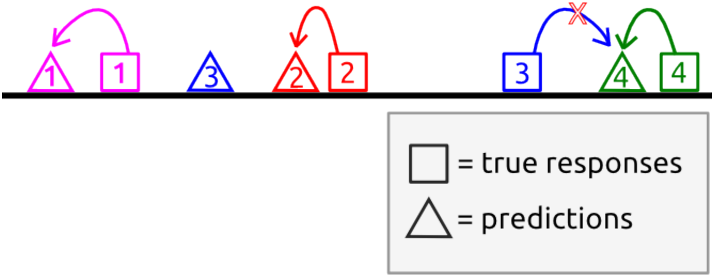
Regression vs Identification loss



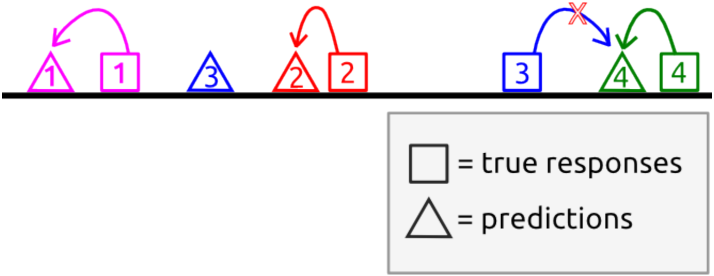
Mean-squared error



Identification loss

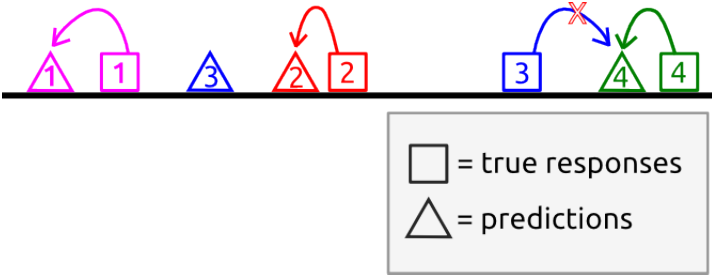


Identification loss



- First used by Kay et al. (2008) to compare accuracy of center-surround model of V1 versus Gabor filter model of V1.

Identification loss



- First used by Kay et al. (2008) to compare accuracy of center-surround model of V1 versus Gabor filter model of V1.
- We are the first to explore theoretical properties of the loss (e.g. connection to mutual information)

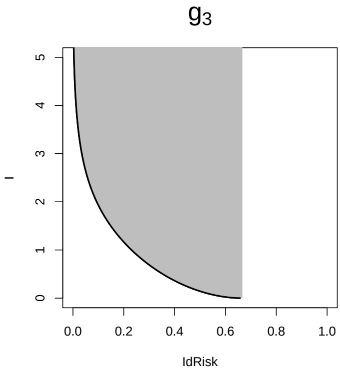
Identification loss and mutual information

- Define the identification risk as the expected identification loss

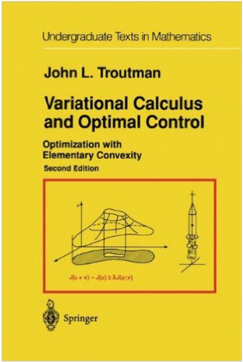
$$\text{IdRisk}_k = \mathbf{E}[\text{IdLoss}_k]$$

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

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

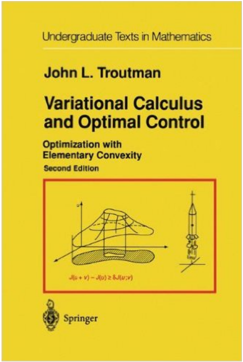


Proof details



- Variational calculus allows optimization of *functionals*.

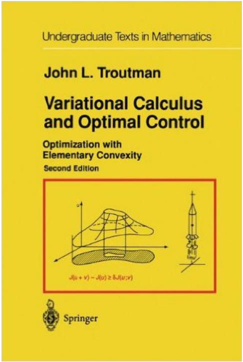
Proof details



- Variational calculus allows optimization of *functionals*.
- Mutual information is a functional of $p(x, y)$.

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

Proof details



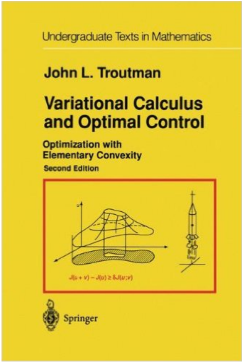
- Variational calculus allows optimization of *functionals*.
- Mutual information is a functional of $p(x, y)$.

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

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

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

Proof details



- Variational calculus allows optimization of *functionals*.
- 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.

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

- $g_k(u)$ obtained by minimizing $\mathrm{I}[p(x, y)]$ subject to $\text{BayesRisk}_k[p(x, y)] \leq u$.

Result

Theorem. (Z., Benjamini 2017) For any $\iota > 0$ and $k = 2, 3, \ldots$, 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, there exists a function g_k such that

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

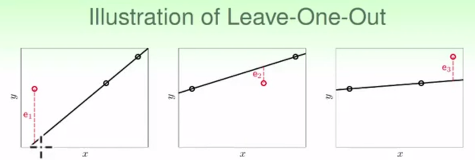
defined by

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

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 $\mathbf{E}[\vec{Y}|\vec{X}]$.
2. Compute *identification loss*, IdLoss_k, using *leave-k-out*.



3. Estimate mutual information using

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

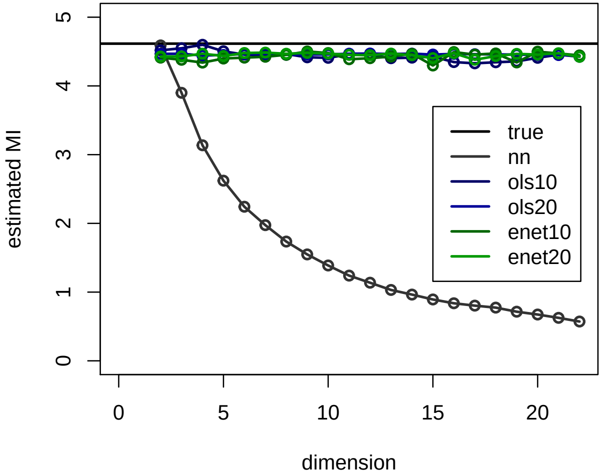
Section 2

Applications

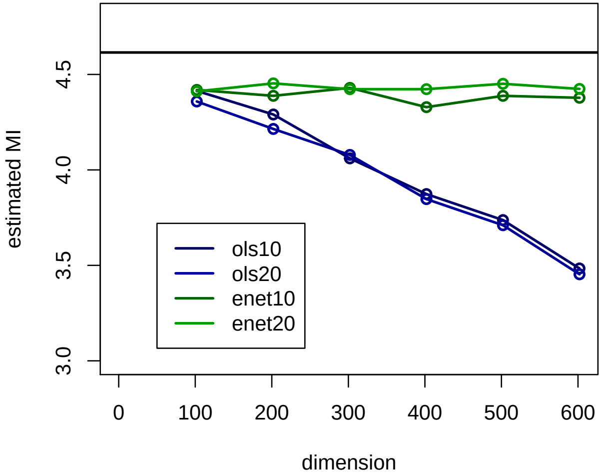
Simulation

- 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 - III. high dimension



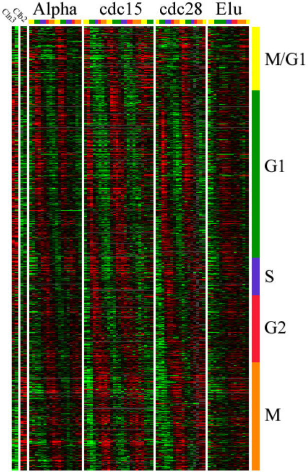
Application to gene expression time series

- Suppose we have *groups* of genes, $\vec{X}^{(1)}, \ldots, \vec{X}^{(m)}$.
- Each group may consist of a different number of genes, p_i.
- Goal: estimate the mutual information

  $$I(\vec{X}^{(i)}; \vec{X}^{(j)})$$

  between each pair, $i \neq j$.
- Conclusion: find groups which are high predictive of each other–this may have biological significance.

Application to gene expression time series



- Data from Spellman et al. 1998
- Expression levels of 6178 yeast genes during cell cycle
- Total 73 measurements 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

Invariance properties

- Transform data from each group with random rotation...

$$\tilde{bX} = \boldsymbol{X} E$$

$$\tilde{bY} = \boldsymbol{X} F$$

with $E^T E = I$, $F^T F = I$.

Invariance properties

- Transform data from each group with random rotation...

$$\tilde{bX} = \boldsymbol{X}E$$

$$\tilde{bY} = \boldsymbol{X}F$$

with $E^T E = I$, $F^T F = I$.

- CCA is invariant:

$$\text{CCA}(\boldsymbol{X}, \boldsymbol{Y}) = \text{CCA}(\tilde{bX}, \tilde{bY})$$

Invariance properties

- Transform data from each group with random rotation...

$$\tilde{bX} = \boldsymbol{X}E$$

$$\tilde{bY} = \boldsymbol{X}F$$

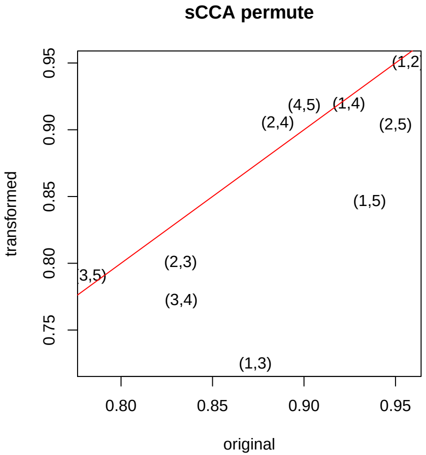
with $E^T E = I$, $F^T F = I$.

- CCA is invariant:

$$\text{CCA}(\boldsymbol{X}, \boldsymbol{Y}) = \text{CCA}(\tilde{bX}, \tilde{bY})$$

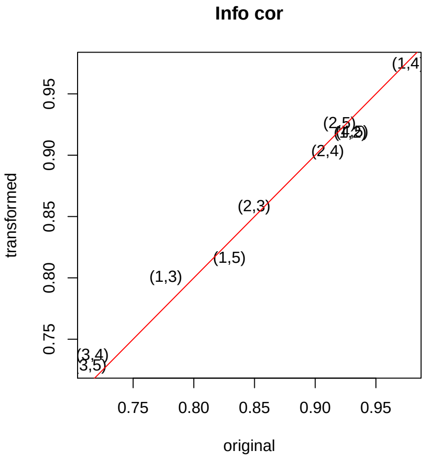
Invariance properties

However, sparse CCA is not invariant.



Invariance properties

Our method, on the other hand, is *robust* to rotation.



Conclusions

- Mutual information, and derived Cor_{Info} are useful measures of correlation, but hard to estimate.

Conclusions

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

Conclusions

- Mutual information, and derived 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, IdLoss_k into an estimate of $I(\vec{X}; \vec{Y})$.

Conclusions

- Mutual information, and derived 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, 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.

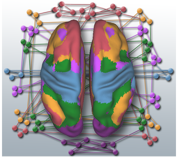


Image credit Simons Foundation

Section 3

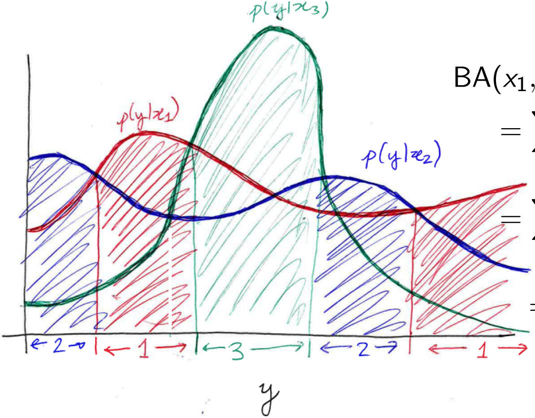
The End

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*

Intuition behind identity

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



$$\begin{aligned} &\text{BA}(x_1, x_2, x_3) \\ &= \sum_i \Pr[x_i] \Pr_{Y \sim p(y|x_i)}[Y \in \text{zone } i] \\ &= \sum_i \frac{1}{k} \text{Area under curve } i \text{ in zone } i \\ &= \frac{1}{k} \text{Area under } \max_{i=1}^{k} p(y|x_i) \end{aligned}$$

Back to main