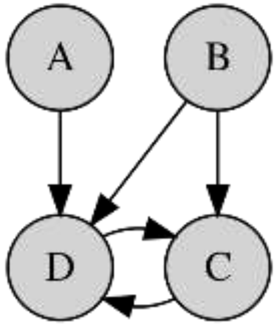
Probabilistic Graphical Model Workshop: Sparsity, Structure and High-dimensionality



Structure Learning of Bayesian Networks with p Nodes from n Samples when n<<p

Joe Suzuki
(Osaka Univ.)

March 25, 2016

Road map

Main Topic: efficient BNSL (30 mins)

1. BNSL definition
2. BNSL with B&B
3. BNSL with B&B for MDL
4. BNSL with B&B for MAP
5. Experiments
6. Discussion on n<<p
7. Summary

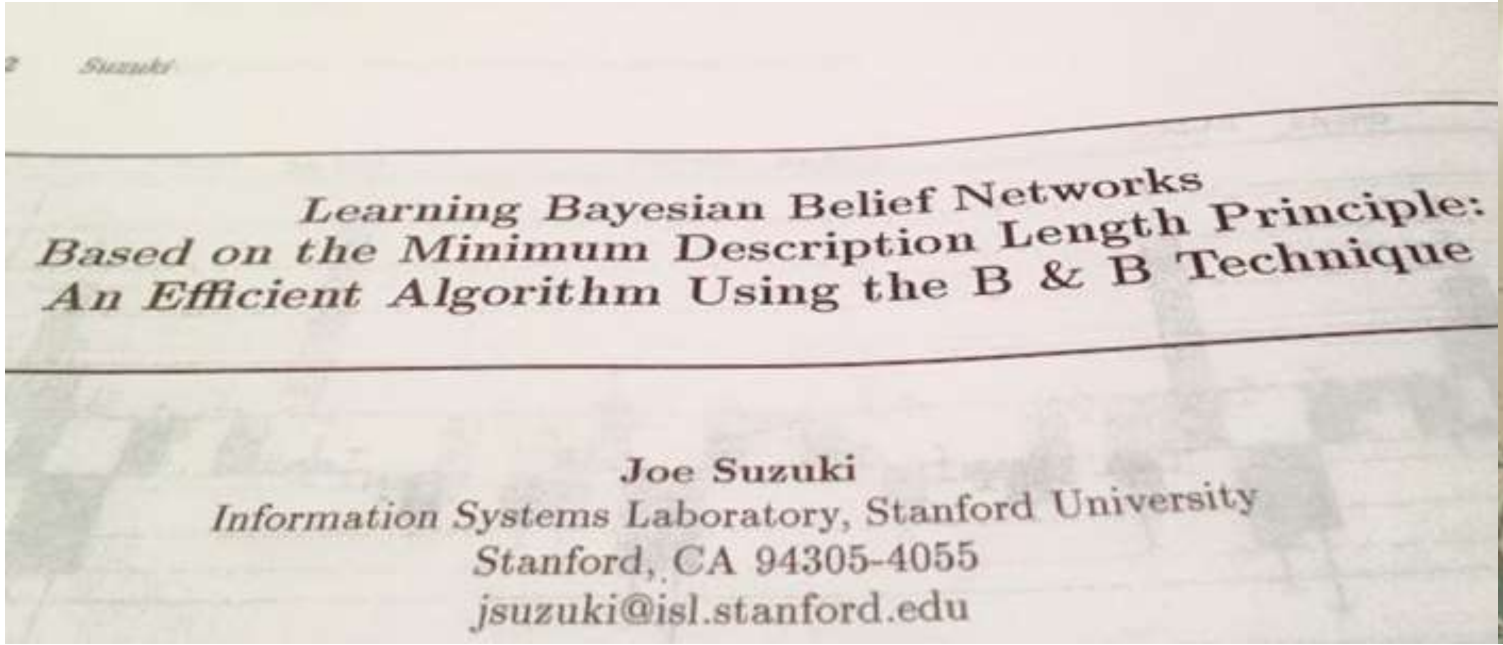
Bonus Topic: why HSIC? (15 mins)

Assumptions
Discrete Random Variables
Prior over Structures is Uniform

BNSL with Branch & Bound (Suzuki 1996)

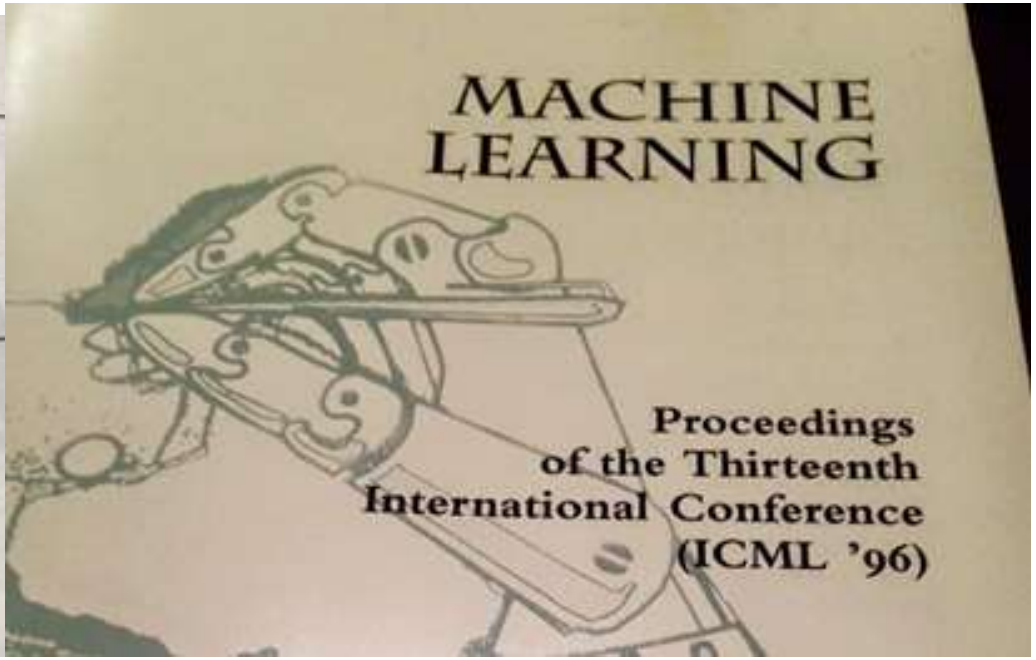
Efficient Computation for Bayesian Network Structure Learning
BNSL is to take exponential time with p (# of variables)

When I was young, …


2 Suzuki

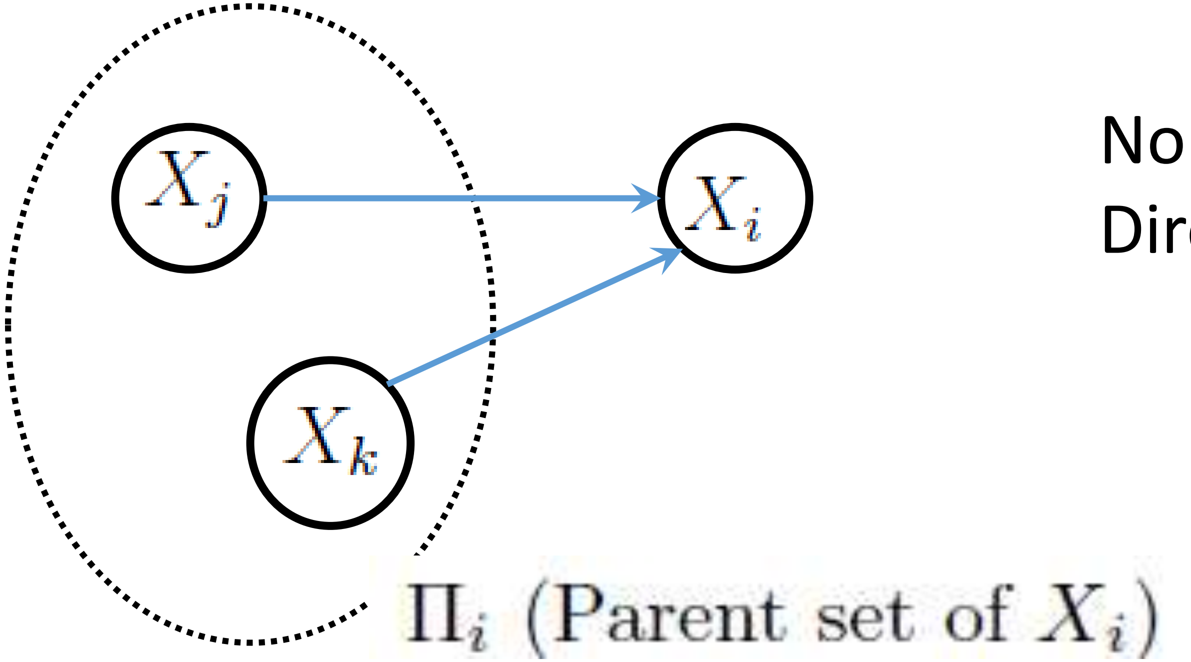
Learning Bayesian Belief Networks
Based on the Minimum Description Length Principle:
An Efficient Algorithm Using the B & B Technique

Joe Suzuki
Information Systems Laboratory, Stanford University
Stanford, CA 94305-4055
jsuzuki@isl.stanford.edu



Bayesian Network (BN)

Factorization $P(X_1, \cdots, X_p) = \prod_{i=1}^{p} P(X_i | \Pi_i)$



No Loop

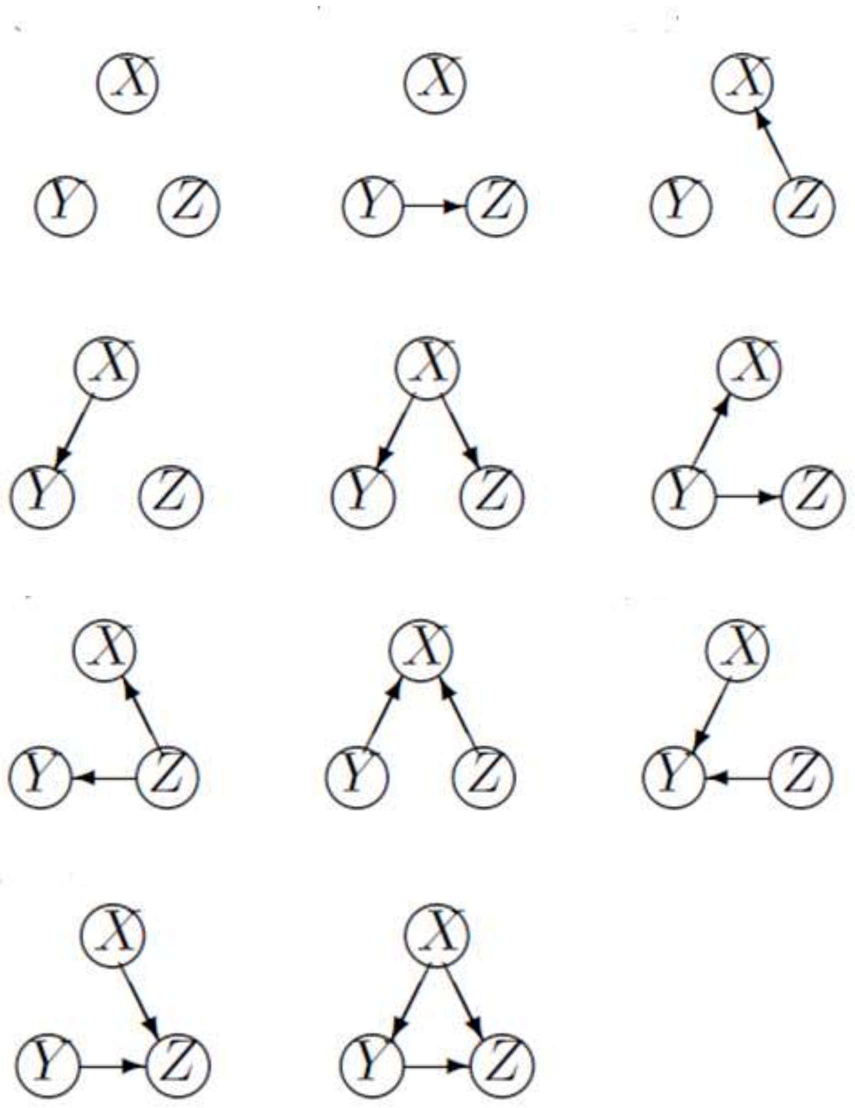
Directed Acyclic Graph (DAG)

BNSL
(BN Structure Learning)

X	Y	Z
x_1	y_1	z_1
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	z_n

Discrete Variables

Structure with Maximum Posterior Probability



p=3 (eleven Markov equivalent classes)

samples x^n and model G

Marginalizing over parameters θ with weights $w(\theta|G)$

$$P(x^n|G) = \int P(x^n|\theta, G)w(\theta|G)d\theta$$

$P(G)$: prior over models G (Uniform)

$$P(G|x^n) \propto P(G)P(x^n|G) \rightarrow \max$$

Score of X given Y

$c(x)$: occurrency of $X = x$

$c(x,y)$: occurrency of $(X,Y) = (x,y)$

$a(x), a(x,y) > 0$

Dirichlet conjugate prior

$$w(\theta) = K \prod_x \theta(x)^{a(x)-1}$$

$$Q^n(X) := \int \prod_x \theta(x)^{c(x)} w(\theta) d\theta \propto \int \prod_x \theta(x)^{c(x)+a(x)-1} d\theta$$

$$w(\theta) \propto \prod_x \prod_y \theta(x|y)^{a(x,y)-1}$$

$$Q^n(X|Y) \propto \int \prod_x \prod_y \theta(x|y)^{c(x,y)} w(\theta) d\theta = \prod_y \{ \frac{\Gamma(a(y)))}{\Gamma(c(y)+a(y))} \prod_x \frac{\Gamma(c(x,y)+a(x,y)))}{\Gamma(c(x,y)+a(x,y))}$$

Gamma Function: an Extention of Factorial

$$\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt$$

$\Gamma(u+1) = u\Gamma(u)$ for $u > 0$

$\Gamma(1) = 1$

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1) \\ &= \cdots = (n-1)(n-2)\cdots 1 \cdot \Gamma(1) = (n-1)!\end{aligned}$$

$$\begin{aligned}\frac{\Gamma(n+a)}{\Gamma(a)} &= \frac{(n+a-1)\Gamma(n+a-1)}{\Gamma(a)} \\ &= \cdots = (n+a-1)\cdots a\end{aligned}$$

for integer $n \geq 0$ and real $a > 0$

For $i = 1, \cdots, p$

$X_i = x \in \{1, \cdots, \alpha_i\}$

$\Pi_i = y \in \{1, \cdots, \beta_i\}$

$c_i(x)$: occurrency of $X_i = x$

$c_i(x, y)$: occurrency of $(X_i, \Pi_i) = (x, y)$

Score of X_i given Π_i

$$Q^n(X_i|\Pi_i)$$

Description Length of X_i given Π_i

$$-\log Q^n(X_i|\Pi_i) \approx H^n(X_i|\Pi) + \frac{(\alpha_i - 1)\beta_i}{2} \log n$$

$$H^n(X_i|\Pi) := \sum_y \sum_x -c_i(x, y) \log \frac{c_i(x, y)}{c_i(y)}$$

Formulation of Structure Learning

Silander-Milymaki (2006)


1. For each (X, S) s.t. $X \notin S \subseteq \{X_1, \cdots, X_p\}$, compute

$$R^n(X|S) := \max_{U \subseteq S} Q^n(X|U)$$

Parent Set

2. Order $X_1, \cdots, X_p$ s.t.

$$\prod_{i=1}^{p} R^n(X_i|\{X_1, \cdots, X_{i-1}\})$$

Order of Variables

(X_i does not have to depend on all the $X_1, \cdots, X_{i-1}$)

BNSL computational complexity

p	DAG cardinality	# of Markov equivalent classes
2	2	1
3	25	11
4	543	129
5	29281	5921
⋮	⋮	⋮

Increases exponentially with p (# of variables)

David M. Chickering
"Learning Bayesian Networks is NP-Complete"

- NP-Completeness only implies existence of one computationally hard instance
- No such an instance was found

when n is a constant (p may be large)

(n: # of samples, p: # of variables)

T-BNSL (p):=max_n T-BNSL(p,n)

We assume that n is a constant (small) and p is large (n << p)
While David assumes that n should be large compared with p.

Parent set with maximum score

$$R^n(X|S) := \max_{U \subseteq S} Q^n(X|U)$$
$$= \max\{Q^n(X|S), \max_{Y \in S} R^n(X|S \backslash \{Y\})\}$$

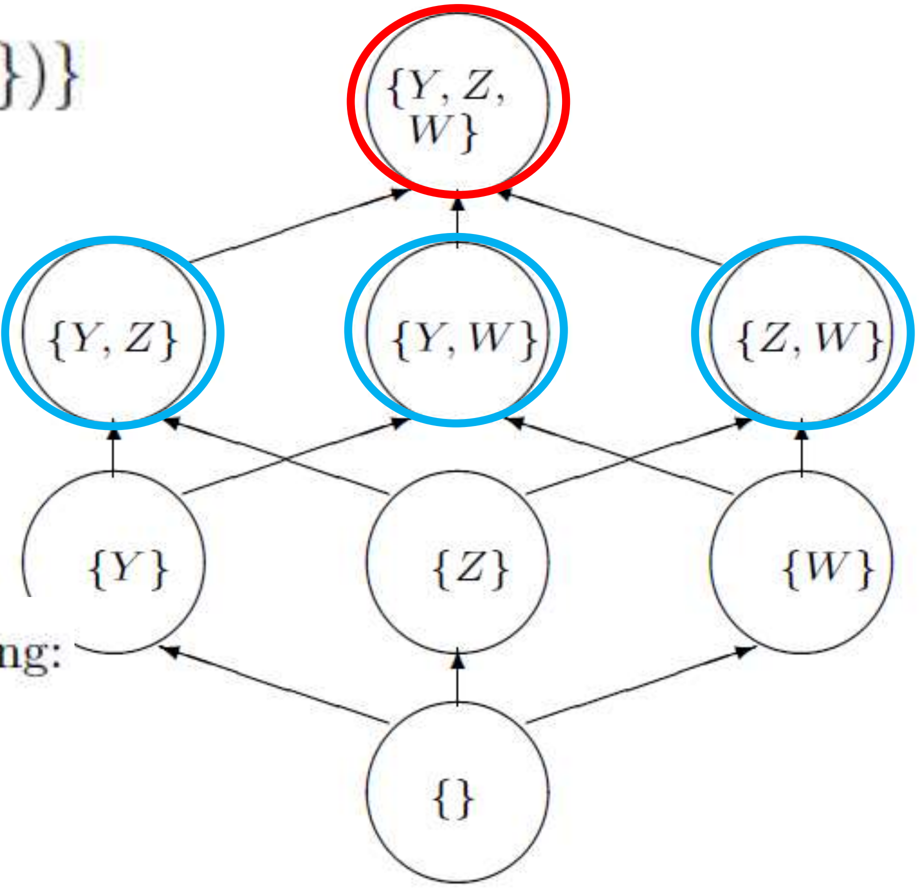
$R^n(X|Y,Z,W)$ is one of the following:

- $Q^n(X|Y,Z,W)$
- $R^n(X|Y,Z)$
- $R^n(X|Z,W)$
- $R^n(X|Z,W)$

We wants to avoid
The computation

$R^n(X|Y,Z)$ is one of the following:

- $Q^n(X|Y,Z)$
- $R^n(X|Y)$
- $R^n(X|Z)$

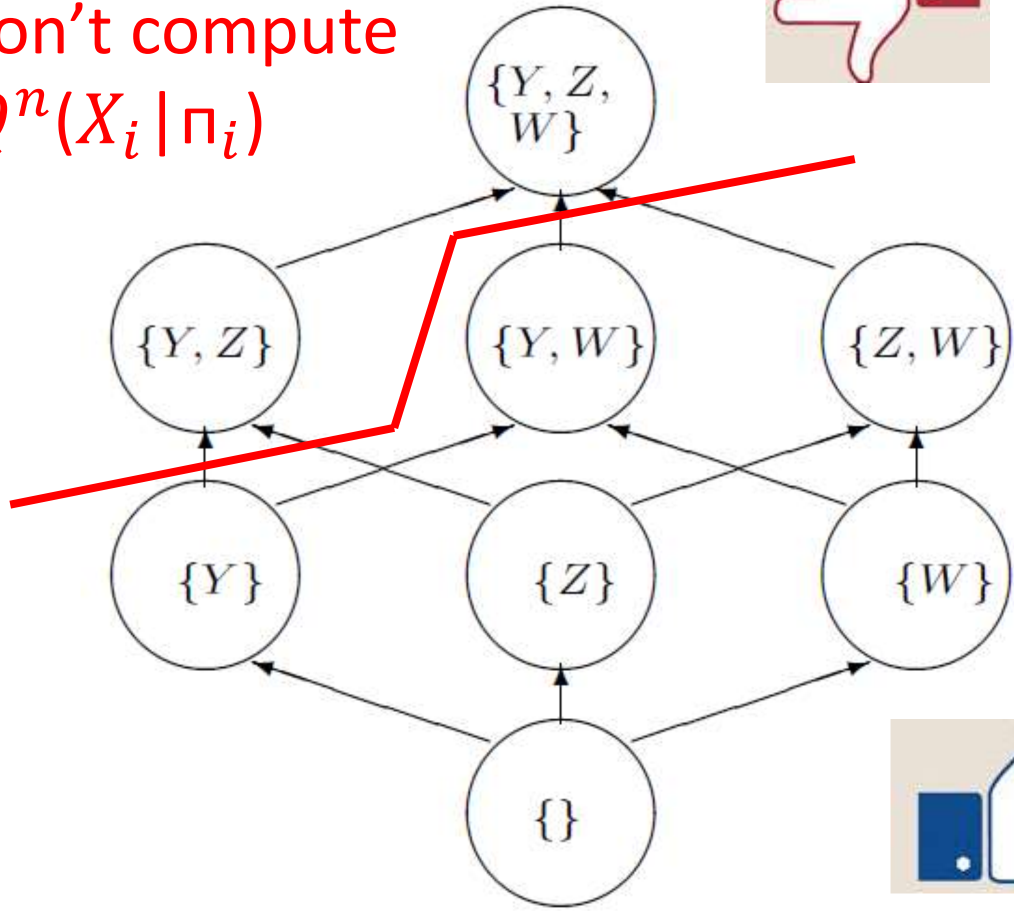


Cut computation for searching further if no optimal solution exists (B&B)

- An optimal solution is guaranteed
- Cut searching further
  if no optimal solution exists
- Take into account the overhead

The basic idea was first proposed for minimizing the MDL (Suzuki, 1996)

Don't compute $Q^n(X_i \mid \Pi_i)$



Compute $Q^n(X_i \mid \Pi_i)$

BNSL B&B for MDL (Suzuki, 1996)

$$X_i = X \text{ and } \Pi_i \subseteq \{Y, Z, W, \cdots\}$$

$$X = x \in \{1, \cdots, \alpha\}$$
$$Y = y \in \{1, \cdots, \beta\}$$
$$Z = z \in \{1, \cdots, \gamma\}$$
$$W = w \in \{1, \cdots, \delta\}$$

$$-\log Q^n(X|Y) \approx H^n(X|Y) + \frac{(\alpha-1)\beta}{2}\log n$$

$$-\log Q^n(X|YZ) \approx H^n(X|YZ) + \frac{(\alpha-1)\beta\gamma}{2}\log n$$

$$\boxed{H^n(X|Y) \le \frac{(\alpha-1)(\gamma-1)\beta}{2}\log n}$$

$$\Longrightarrow \left\{ \begin{array}{l} H^n(X|Y) + \frac{(\alpha-1)\beta}{2}\log n \le H^n(X|YZ) + \frac{(\alpha-1)\beta\gamma}{2}\log n \\ H^n(X|Y) + \frac{(\alpha-1)\beta}{2}\log n \le H^n(X|YZW) + \frac{(\alpha-1)\beta\gamma\delta}{2}\log n \end{array} \right. ,$$

{Y} may be optimal but {Y,Z} is not

$\#\Pi_i$: the # of variables that the parent set Π_i contains

From $H^n(X|Y) \leq n \log \alpha$ and $\gamma \geq 2$

$$
\begin{aligned}
& \#\Pi_i \geq \log n \ , \ n \geq 4 \\
\Longrightarrow \quad & \beta \geq n \ , \ n \geq 2 \\
\Longrightarrow \quad & \log \alpha \leq \frac{(\alpha-1)\beta}{2} \log n \\
\Longrightarrow \quad & H^n(X|Y) \leq \frac{(\alpha-1)(\gamma-1)\beta}{2} \log n
\end{aligned}
$$

A optimal parent set contains at most log n variables

(Campos et. al. 2011)

of the possible subsets is polynomial in p if n is independent from p

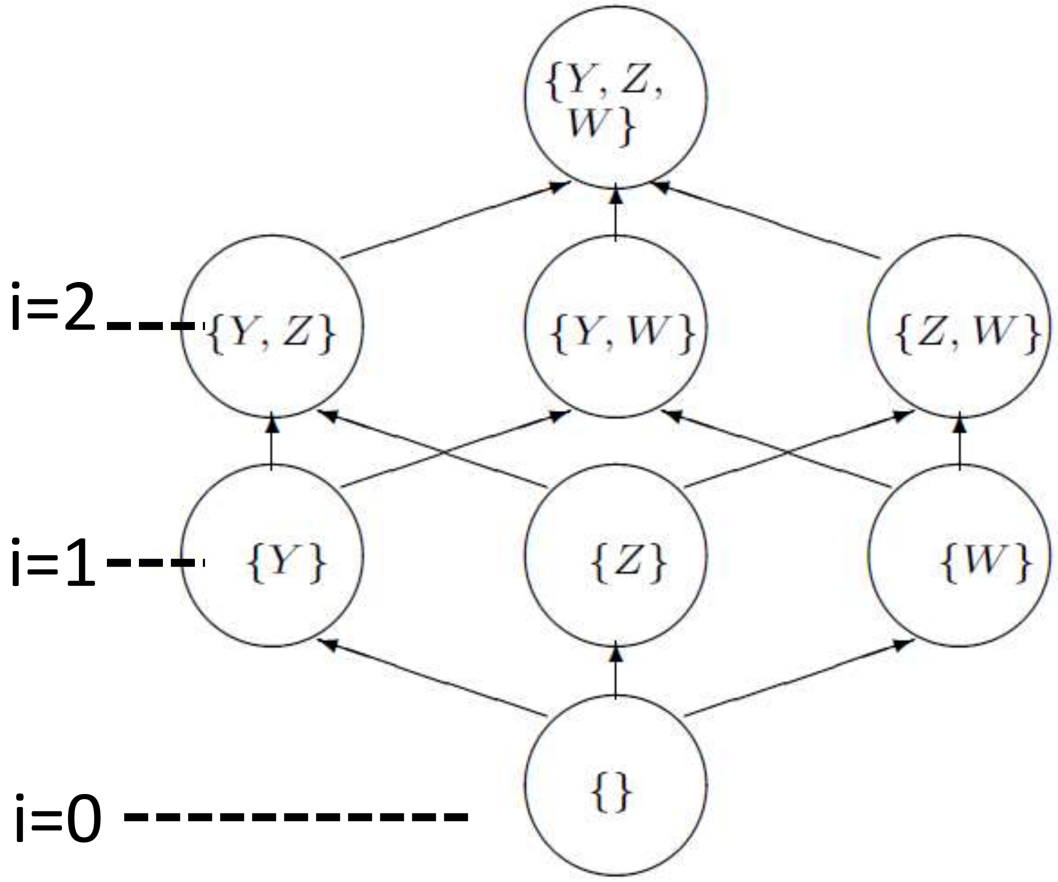
Only seeking the parent sets (Step 1) are considered

$$L = \log n$$

$$\sum_{i=0}^{L} \binom{p}{i} \leq 1 + p^L$$

If n is small,
the computation is efficient

$$n = 128 \Longrightarrow 1 + p^L = 1 + p^7$$



Conjecture: if n and p are independent, BNSL with MDL is polynomial time of p

Step	What to do
1	Parent sets
2	Order the variables

- Step 1 takes much more time than Step 2 in experiments
- Step 2 has not been proved to take polynomial time of p.

BNSL B&B for MAP (Campos, et. Al. 2011)

$$
\begin{aligned}
R^n(X|S) &= \max_{U \subseteq S} Q^n(X|U) \\
&= \max\{Q^n(X|S), \max_{Y \in S} R^n(X|S \backslash \{Y\})\}
\end{aligned}
$$

$$\max_{Y \in S} R^n(X|S \backslash \{Y\}) \geq \sup_{S \subseteq S'} Q^n(X|S')$$ UpperBound

$$\Longrightarrow S \text{ and its supersets are not optimal}$$

Find $Q^n_*(X|S)$

1. $Q^n(X|S) \leq Q^n_*(X|S)$
2. $Q^n(X|S') \leq Q^n_*(X|S)$ for all $S' \supseteq S$

$$Q^n_*(X|S) := \prod_{y:c(y)>0} \frac{\max_x a(x,y)}{a(y)}$$

$$a(x,y) = a(y) = 0.5 \Longrightarrow Q^n_*(X|S) = \alpha^{\#\{y|c(y)>0\}}$$

A Tighter Upper Bound (Suzuki 2015)

Proposed

$$Q^n_{**}(X|S) := \prod_x \prod_y \{ \frac{\Gamma(c(x,y)+a(x,y))}{\Gamma(a(x,y))} \cdot \frac{\Gamma(a(y))}{\Gamma(c(x,y)+a(y))} \}$$

$$\sup_{S' \supseteq \Pi} Q^n(X|S') \leq Q^n_{**}(X|S) \leq Q^n_*(X|S)$$

Existing

$$Q^n_*(X|S) := \prod_{y:c(y)>0} \frac{\max_x a(x,y)}{a(y)}$$

Experiments using Alarm Netwok data

(The proposed is three times faster than the existing one)

n=100

p	2^{p-1} t	R_* T_*	R_{**} T_{**}
16	32,768 3.123	11,834 2.156	3,366 0.828
18	13,1072 11.641	41,903 3.766	11,342 1.385
20	524,288 56.875	145,374 15.047	37,306 5.063
22	2,097,152 203.766	514,300 49.843	127,562 15.406
24	8,388,608 1012.234	1,698,040 211.656	411,466 63.95

n=500

p	2^{p-1} t	R_* T_*	R_{**} T_{**}
16	32,768 22.469	24,788 17.891	17,337 13.312
18	131,072 92.047	119,828 83.080	76,698 59.172
20	524,288 782.72	355,379 548.219	192,798 367.969
22	2,097,152 2992.078	1,251,530 1816.0630	705,208 1122.531
24	8,388,608 10720.192	4,380,355 6356.221	2,468,228 3928.858

What we can say from the experiments

The proposed bound cut the computation
to one-third compared with Campos 2011
10% to compared with computation w/o B&B

When the bound works most often ?
The sample size n is small
The # of variables p is large

Discussion for small n

Why B&B works so effectively?	
Bayes/MDL avoids complicated models for small n even during the search	
⇒	The search space is limited for small n
⇒	The computational effort is small for small n

Why we do not seek consistency?
If the MAP solution is obtained, Bayes/MDL does its best, and we should be satisfied.
Prior info should be tuned carefully for small n

Summary for the main topic

BNSL with B&B
MDL: started when I was young, and many other started now
MAP: recently improved for small n and large p (n<<p)

Future Task: **BNSL takes polynomial when p and n are independent**
Many people believes that the statement is false
I have much evidence that the statement is true

Why HSIC?

Accepted: March 23, 2016
Appears coming week

Article

An Estimator of Mutual Information and its Application to Independence Testing

Joe Suzuki *,

Received: date ; Accepted: date ; Published: date
Academic Editor: name

[1] Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan
* Correspondence: suzuki@math.sci.osaka-u.ac.jp

What measures Independence ?

$$\rho(X,Y) := \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = 0 \; \overset{\Longleftarrow}{\not\Longrightarrow} \; X \perp\!\!\!\perp Y$$

- Mutual Information: $I(X,Y) := \sum_x \sum_y P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$

$$I(X,Y) = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$$

- Hilbert Schmidt independent criterion

$$HSIC(X,Y) = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$$

Independence Test (Whether $X \perp\!\!\!\perp Y$ or not)

Given $(x_1, y_1), \cdots, (x_n, y_n)$, estimate $I(X,Y)$, $HSIC(X,Y)$, etc.

A naive estimator of I(X,Y)

$$I_n := \sum_x \sum_y \frac{c(x,y)}{n} \log \frac{\frac{c(x,y)}{n}}{\frac{c(x)}{n} \cdot \frac{c(y)}{n}}$$

$c(x), c(y), c(x,y)$: Occurrencies of $x, y, (x,y)$ in $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y)$

$I_n > 0$ with positive probability as $n \to \infty$ even when $X \perp\!\!\!\perp Y$

$$H^n(X) := \sum_x -\frac{c(x)}{n}\log\frac{c(x)}{n} \qquad H^n(Y) := \sum_y -\frac{c_(y)}{n}\log\frac{c(y)}{n}$$

$$H^n(X,Y) := \sum_x\sum_y -\frac{c_(x,y)}{n}\log\frac{c(x,y)}{n}$$

$$I_n = \frac{1}{n}\{H^n(X) + H^n(Y) - H^n(X,Y)\}$$

$$-\log Q^n(X) \approx H^n(X) + \frac{\alpha-1}{2}\log n \qquad -\log Q^n(Y) \approx H^n(Y) + \frac{\beta-1}{2}\log n$$

$$-\log Q^n(X,Y) \approx H^n(X,Y) + \frac{\alpha\beta-1}{2}\log n$$

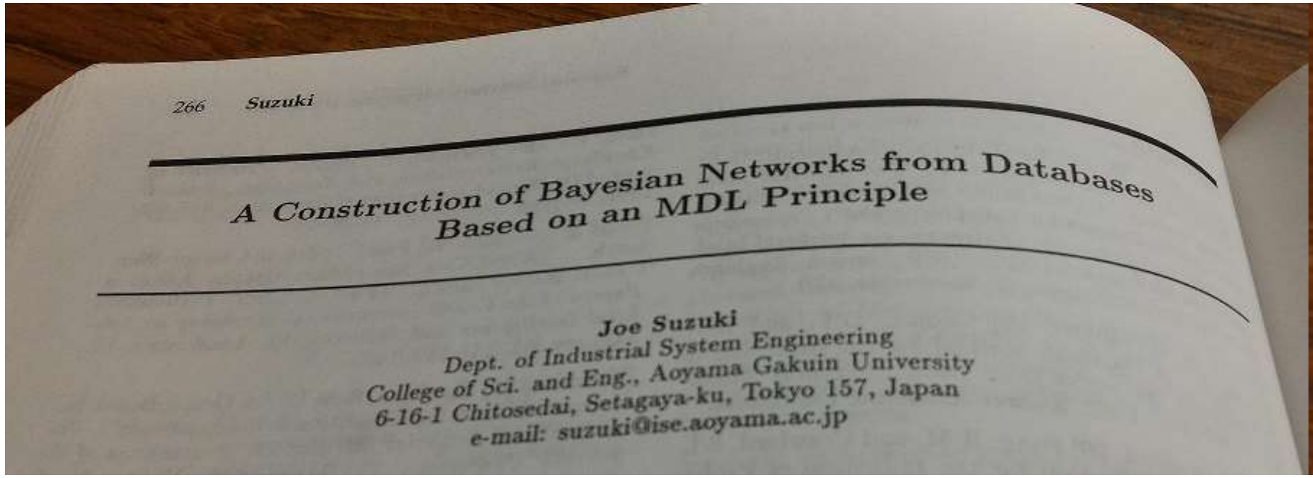
$$J_n \quad := \quad \frac{1}{n}\log\frac{Q^n(X,Y)}{Q^n(X)Q^n(Y)} = I_n - \frac{(\alpha-1)(\beta-1)}{2n}\log n$$

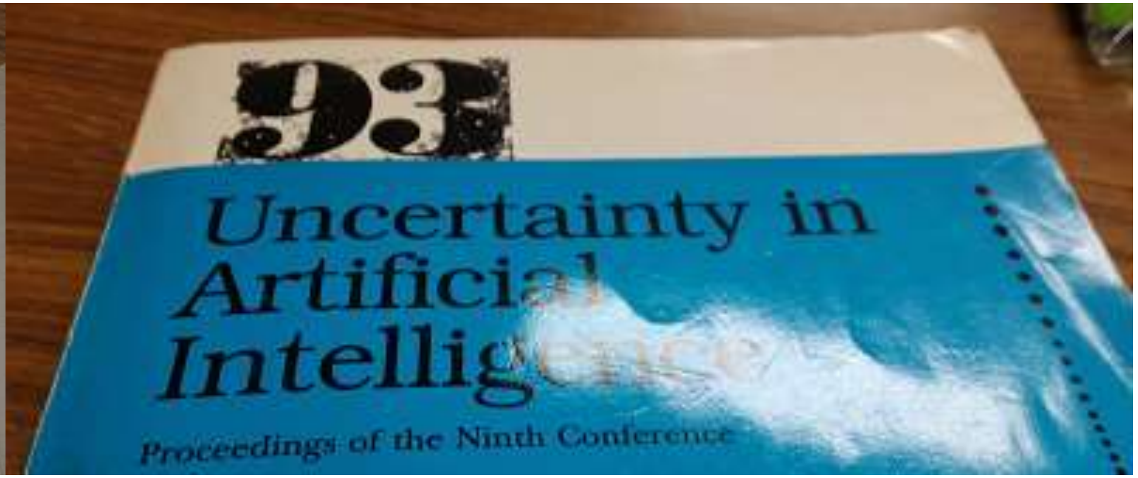
MDL/Bayes Estimators (Suzuki 1993, 2012)

$$J_n = \max\{\frac{1}{n}\log\frac{Q^n(X,Y)}{Q^n(X)Q^n(Y)}, 0\} \approx \max\{I_n - \frac{(\alpha-1)(\beta-1)}{2n}\log n, 0\}$$

For large n (almost surely), $J_n = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$

$$I_n = 0 \not\Longleftrightarrow X \perp\!\!\!\perp Y$$

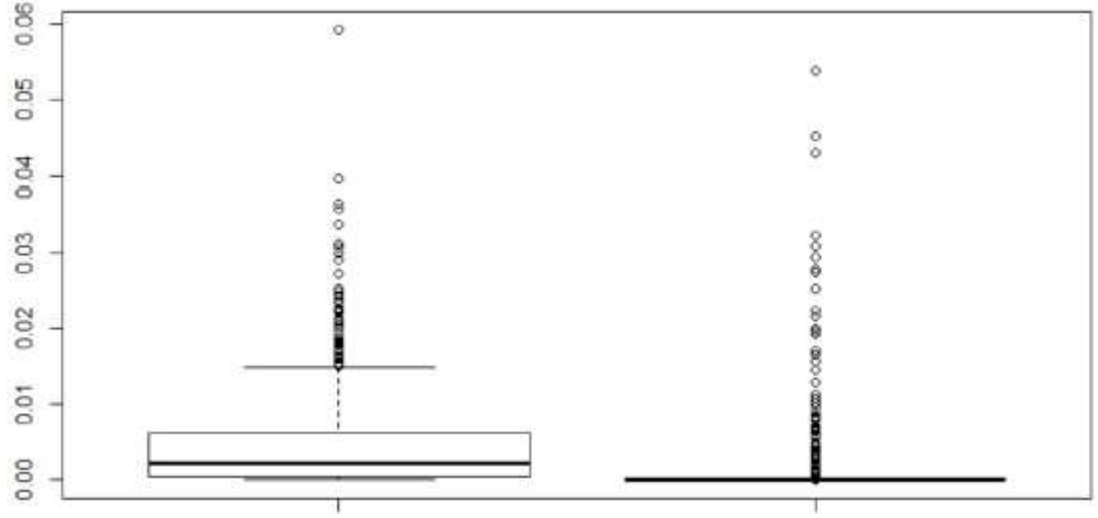




I_n overestimates but J_n captures the $I(X,Y)$

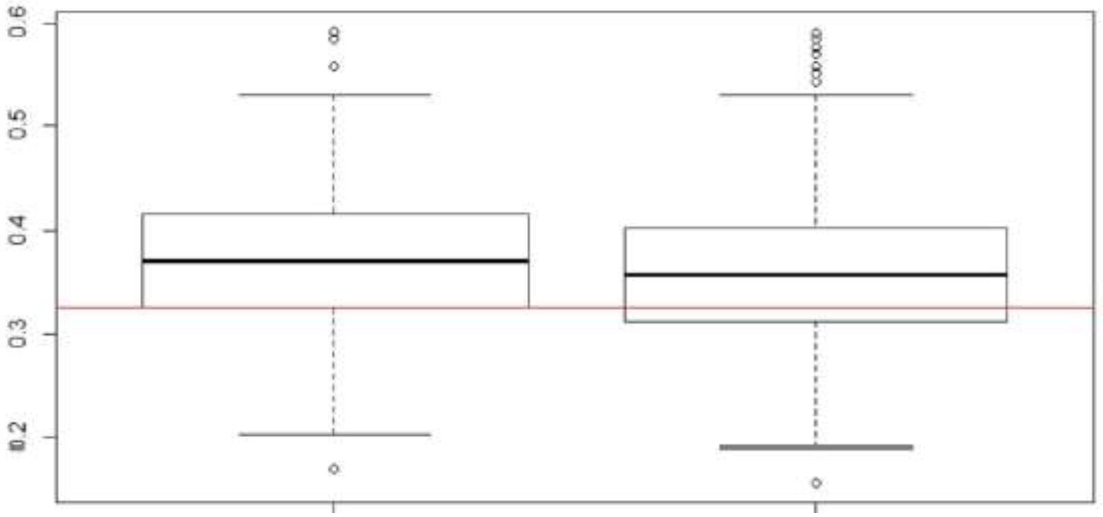
Binary sequences of length 100 with $X \perp\!\!\!\perp Y$ and $X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y$



I_n J_n

$X \not\perp\!\!\!\perp Y$ $I(X,Y) = 0.36$



I_n J_n

Why HSIC?

$$E_{XX'YY'}[k(X,X')l(Y,Y')] + E_{XX'}[k(X,X')] \cdot E_{YY'}[l(Y,Y')] - 2E_{XY}\{E_{X'}[k(X,X')]E_{Y'}[l(Y,Y')]\}$$

$HSIC(X,Y) = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$ for characteristic kernels k, l

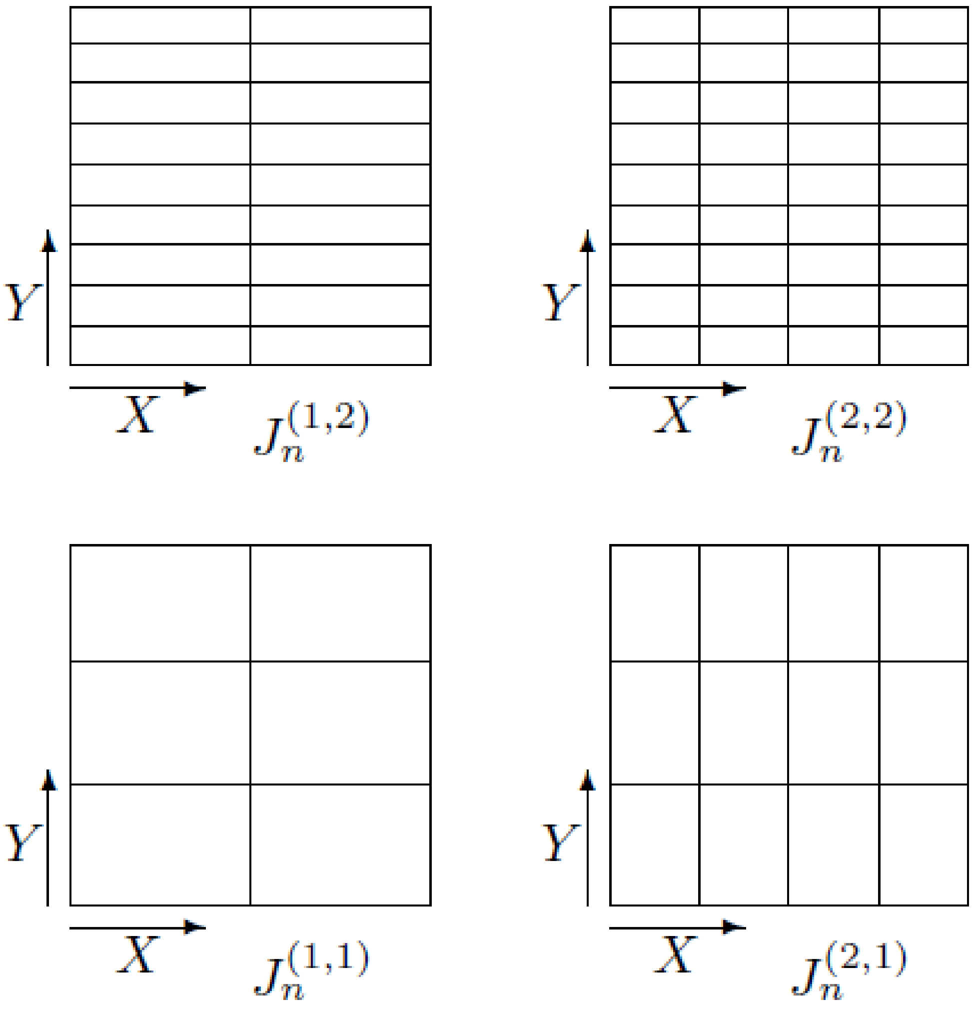
Commonly used Estimator of HSIC

$$\frac{1}{n^2}\sum_{i,j} k(x_i,x_j)l(y_i,y_j) + \frac{1}{n^4}\sum_{i,j} k(x_i,x_j)\sum_{p,q} l(y_p,y_q) - \frac{2}{n^3}\sum_{i,p,q} k(x_i,x_p)l(y_i,y_q)\}$$

- no Estimator $HSIC_n$ s.t.

  For large n (almost surely), $HSIC_n = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$
- no null hypothesis distribution is known
- $O(n^4)$ computation
- hard to extend Independence to Conditional Independence

Estimating MI for Continuous variables



$$J_n^{(u,v)} = I_n^{(u,v)} - \frac{(p^u - 1)(q^u - 1)}{2n} \log n$$

$$J_n := J_n^{(u,v)}$$

1	2	3	4	5	6	7	8
125	125	125	125	125	125	125	125

	1	2	3	4	5	6	7	8
1	75	32	12	5	1	0	0	0
2	25	41	25	18	9	7	0	0
3	15	23	32	27	14	11	1	2
4	5	17	24	22	27	19	11	0
5	5	9	19	24	23	23	17	5
6	0	3	7	18	26	26	28	17
7	0	0	6	9	19	21	45	25
8	0	0	0	2	6	18	23	76

Properties of the Proposed Method

Theorem 1 For $n \geq n_0 := \max\{p^{\frac{2}{(p-1)(1-1/q)}}, q^{\frac{2}{(q-1)(1-1/p)}}\}$, the optimal (u, v) satisfies $p^u q^v \leq n$.

Theorem 2 With probability one as $n \to \infty$, $J_n \leq 0$ if and only if X and Y are independent.

$X \in \{0, 1\}$ (equi-prob), $U \in \{0, 1\}$ (prob. $p = 0.1, 0.2, 0.3, 0.4, 0.5$)

$Y = X + U \bmod 2$

$n = 100$, $n = 200$

Repeat to compute $J_n(x^n, y^n)$ and $HSIC_n(x^n, y^n)$ 100 times

$n = 200$ (100 trials)	p=0.5		p=0.4	
	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$
HSIC	95	5	24	76
MI	94	6	19	81

$n = 100$ (100 trials)	p=0.5		p=0.4	
	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$
HSIC	95	5	49	51
MI	88	12	33	67

$X, U \sim \mathcal{N}(0, 1)$ (mutually independent)

$Y = qX + \sqrt{1 - q^2}U$, $q = 0, 0.2, 0.4, 0.6, 0.8$

$n = 100$, $n = 200$.

Repeat to compute $J_n(x^n, y^n)$ and $HSIC_n(x^n, y^n)$ 100 times

$n = 200$ (100 trials)	q=0		q=0.2		q=0.4	
	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$
HSIC	97	3	51	49	0	100
MI	95	5	58	42	4	92

$n = 100$ (100 trials)	q=0		q=0.2		q=0.4	
	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$
HSIC	93	7	74	26	11	89
MI	94	6	56	44	23	77

Cases that HSIC does not detect independences

1. $X, U \sim \mathcal{N}(0, 0.25)$ (mutually independent)

$Y = X - \lfloor X \rfloor + \lfloor U \rfloor$

Integers and fractions are independent and dependent for X, Y

$$X \in \{0, 1, \cdots, 9\} \text{ (uniform)}$$

$$Y \in \begin{cases} \{0, 2, 4, 6, 8\}, & X : even \\ \{1, 3, 5, 7, 9\}, & X : odd \end{cases}$$

	Rounding	
	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$
	100	0
	1	99

2. $$X \in \{0, 1, \cdots, 9\} \text{ (uniform)}$$

$$Y \in \begin{cases} \{0, 2, 4, 6, 8\}, & X : even \\ \{1, 3, 5, 7, 9\}, & X : odd \end{cases}$$

$X + Y$ should be even

$n = 200$ (100 trials)	parity	
	$\perp\!\!\!\perp$	$\not\perp\!\!\!\perp$
HSIC	96	4
MI	0	100

The proposed **and** HSIC require $O(nlog^2n)$ and $O(n^3)$ computations

n	100	500	1000	2000
HSIC	0.50	9.51	40.28	185.53
Proposed	0.30	0.33	0.62	1.05

Summary for the Bonus Topic

No Free Lunch for Independence Testing

No one independent test outperm other test for all the problem (correctness)

Efficiency

The proposed and HSIC require $O(nlog^2n)$ and $O(n^3)$ computations

The largest merit: for large n

$$J_n = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$$

Open Problem

Is there any estimator $HSIC_n$ s.t. for large n

$$HSIC_n = 0 \iff X \perp\!\!\!\perp Y \quad ?$$