Learning Causal Structure with Kernel-based Dependence Measures

Kenji Fukumizu

Institute of Statistical Mathematics

Graduate University for Advances Studies

Joint work with Xiaohai Sun, Dominik Janzing, Bernhard Schölkopf, and Arthur Gretton.

November 3-4, 2007

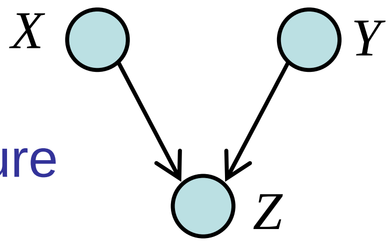
Outline

1. Introduction
2. Kernel measures for dependence
3. Kernel measures for conditional dependence
4. Causal inference with kernels
   – Kernel-based Causal Learning algorithm –
5. Conclusion

Introduction

- Conditional independence in causal learning
  - Determining independence and conditional independence is essential in causal learning.

v-structure



$$\Longleftrightarrow$$

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

and

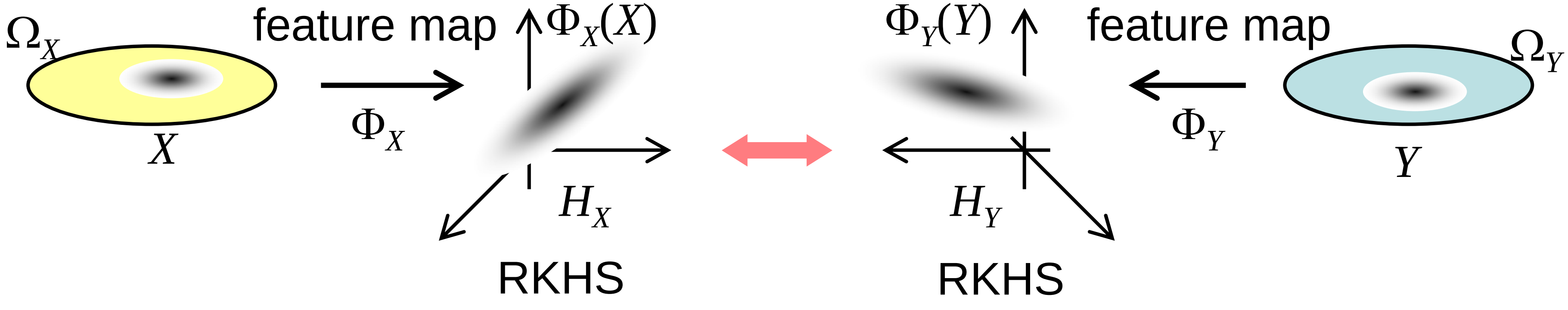
$$X \not\perp\!\!\!\perp Y \mid Z$$

  - But, in practice
    - Dependence for continuous domain is not straightforward.

      How can we estimate mutual information?
    - Many algorithms use linear statistical methods (partial correlation) or discretization.

“Kernel methods” for dependence of variables

- Positive definite kernels have been used for capturing nonlinearity of original data. *e.g*. Support vector machine.
- Kernelization: mapping data into a functional space (RKHS) and apply linear methods on RKHS.
- Recently, kernel methods have been applied for dependence analysis. Covariance structure on RKHS gives dependence and conditional dependence of the original variables.



Positive Definite Kernel and RKHS

- **Positive definite kernel (p.d. kernel)**

  Ω: set. $k : \Omega \times \Omega \to \mathbf{R}$

  k is positive definite if $k(x,y) = k(y,x)$ and for any $n \in \mathbf{N}$, $x_1, \ldots x_n \in \Omega$ the matrix $\left(k(x_i, x_j)\right)_{i,j}$ (Gram matrix) is positive semidefinite.

  – Example: Gaussian RBF kernel $k(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right)$

- **Reproducing kernel Hilbert space (RKHS)**

  k: p.d. kernel on Ω.

  $\Rightarrow$ $\exists 1\ H$: reproducing kernel Hilbert space (RKHS)

  1) $k(\cdot, x) \in H$ for all $x \in \Omega$.

  2) $\mathrm{Span}\{k(\cdot, x) \mid x \in \Omega\}$ is dense in H.

  3) $\langle k(\cdot, x), f \rangle_H = f(x)$ (reproducing property)

Feature map / feature vector

$$\Phi : \Omega \to H, \quad x \mapsto k(\cdot, x) \qquad i.e. \quad \Phi(x) = k(\cdot, x)$$

Data: $X_1, \ldots, X_N$ → $\Phi_X(X_1), \ldots, \Phi_X(X_N)$: functional data

Why RKHS?

- By the reproducing property, computation of the inner product on RKHS does not need expansion by basis functions.

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

$$f = \sum_{i=1}^{N} a_i \Phi(x_i) = \sum_i a_i k(\cdot, x_i), \qquad g = \sum_{j=1}^{N} b_j \Phi(x_j) = \sum_j b_j k(\cdot, x_j)$$

$$\Rightarrow \quad \langle f, g \rangle = \sum_{i,j} a_i b_j k(x_i, x_j)$$

The computational cost essentially depends on the sample size.
Advantageous for high-dimensional data of small sample size.

Outline

1. Introduction
2. Kernel measures for dependence
3. Kernel measures for conditional dependence
4. Causal learning with kernels
   – Kernel-based Causal Learning algorithm –
5. Conclusion

Covariance on RKHS

– Linear case (Gaussian):

$$\mathrm{Cov}[X, Y] = E[YX^T] - E[Y]E[X]^T$$: covariance matrix

– On RKHS:

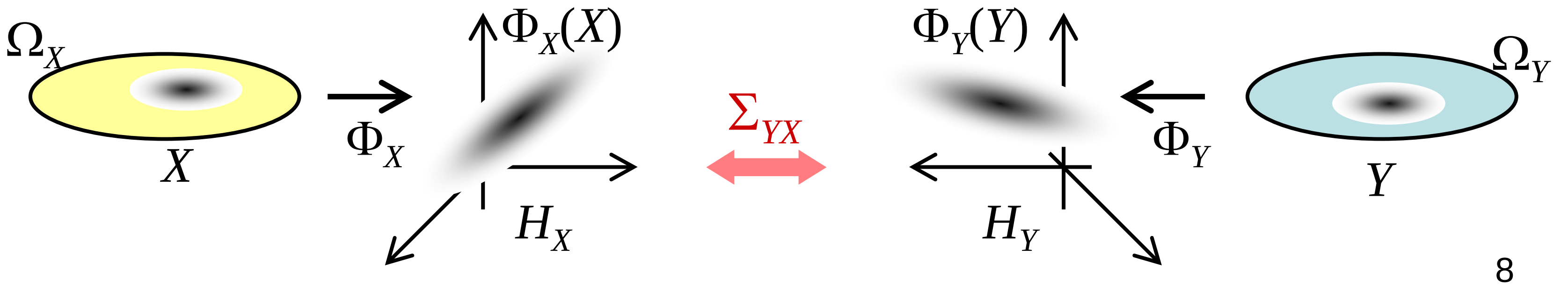
X , Y : random variables on Ω_X and Ω_Y , resp.

Prepare RKHS (H_X, k_X) and (H_Y , k_Y) defined on Ω_X and Ω_Y, resp.

Define random variables on the RKHS H_X and H_Y by

$$\Phi_X(X) = k_X(\cdot, X) \qquad \Phi_Y(Y) = k_Y(\cdot, Y)$$

Define the big (possibly infinite dimensional) covariance matrix Σ_{YX} on the RKHS.



Cross-covariance operator

– Definition

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle\Phi_X(X),\cdot\rangle] - E[\Phi_Y(Y)]E[\langle\Phi_X(X),\cdot\rangle]$$

Σ_{YX} is an operator from H_X to H_Y such that

$$\langle g, \Sigma_{YX} f\rangle = E[g(Y)f(X)] - E[g(Y)]E[f(X)] \quad (= \mathrm{Cov}[f(X), g(Y)])$$

for all $f \in H_X, g \in H_Y$

– *c.f.* Euclidean case

$V_{YX} = \mathrm{E}[YX^T] - \mathrm{E}[Y]\mathrm{E}[X]^T$: covariance matrix

$$(b, V_{YX}a) = Cov[(b,Y),(a,X)]$$

Higher-order moments

Suppose X and Y are **R**-valued, and $k(x,u)$ admits the expansion

$$k(x,u) = 1 + c_1 xu + c_2 x^2 u^2 + c_3 x^3 u^3 + \cdots \qquad \text{e.g.) } k(x,u) = \exp(xu)$$

With respect to the basis $1, u, u^2, u^3, \ldots$, the random variables on RKHS are expressed by

$$\Phi(X) = k(X,u) \sim (1, c_1 X, c_2 X^2, c_3 X^3, \ldots)^T$$

$$\Phi(Y) = k(Y,u) \sim (1, c_1 Y, c_2 Y^2, c_3 Y^3, \ldots)^T$$

$$\Sigma_{YX} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ 0 & c_1^2 Cov[Y,X] & c_1 c_2 Cov[Y,X^2] & c_1 c_3 Cov[Y^3,X] & \cdots \\ 0 & c_2 c_1 Cov[Y^2,X] & c_2^2 Cov[Y^2,X^2] & c_2 c_3 Cov[Y^2,X^3] & \cdots \\ 0 & c_3 c_1 Cov[Y^3,X] & c_3 c_2 Cov[Y^3,X^2] & c_3^2 Cov[Y^3,X^3] & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

The operator Σ_{YX} contains the information on all the higher-order correlation.

Characterization of Independence

- Independence and Cross-covariance operator

If the RKHS's are "rich enough" to express all the moments,

X and Y are independent $\Leftrightarrow$ $\Sigma_{XY} = O$

$\Updownarrow$

$\mathrm{Cov}[f(X), g(Y)] = 0$

or

$E[g(Y) f(X)] = E[g(Y)] E[f(X)]$

for all $f \in H_X, g \in H_Y$

($\Rightarrow$ is always true.

$\Leftarrow$ requires some assumption

Gaussian RBF kernels gives the above equivalence.

$$k(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2 \right)$$

– *c.f.* for Gaussian variables

X and Y are independent $\Leftrightarrow$ $V_{XY} = O$ i.e. uncorrelated

Kernel Dependence Measure

– Hilbert-Schmidt Independence Criteria (HSIC)

$$HSIC(X,Y) = \left\| \Sigma_{YX} \right\|_{HS}^2$$

$$HSIC = 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

– Empirical estimator

$$HSIC_{emp}(X,Y) = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 = \mathrm{Tr}\left[G_X G_Y \right]$$

$$G_X = \left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T \right) K_X \left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T \right)$$: centered Gram matrix

$$K_X = \left(k(X_i, X_j) \right)_{i,j=1}^N$$

– Hilbert-Schmidt norm of an operator

$A: H_1 \to H_2$ operator on a Hilbert space

$\{\varphi_i\}, \{\psi_j\}$: complete orthonormal system of H_1 and H_2 (resp.).

$$\|A\|_{HS}^2 = \sum_j \sum_i \left\langle \psi_j, A\varphi_i \right\rangle^2$$

c.f. Frobenius norm of a matrix

Independence Test

- Permutation test for independence
  - Null hypothesis

    $$H_0: \quad X \perp\!\!\!\perp Y$$

  - Permutation test: simulation of the distribution of test statistics under H_0.
    - Make many samples consistent with the null hypothesis by random permutations of the original sample.

      $$\begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\ Y_1 & Y_2 & Y_3 & Y_4 & Y_5 & Y_6 & Y_7 \end{matrix} \Rightarrow \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 \\ Y_5 & Y_1 & Y_7 & Y_4 & Y_2 & Y_6 & Y_3 \end{matrix} \quad \text{independent}$$

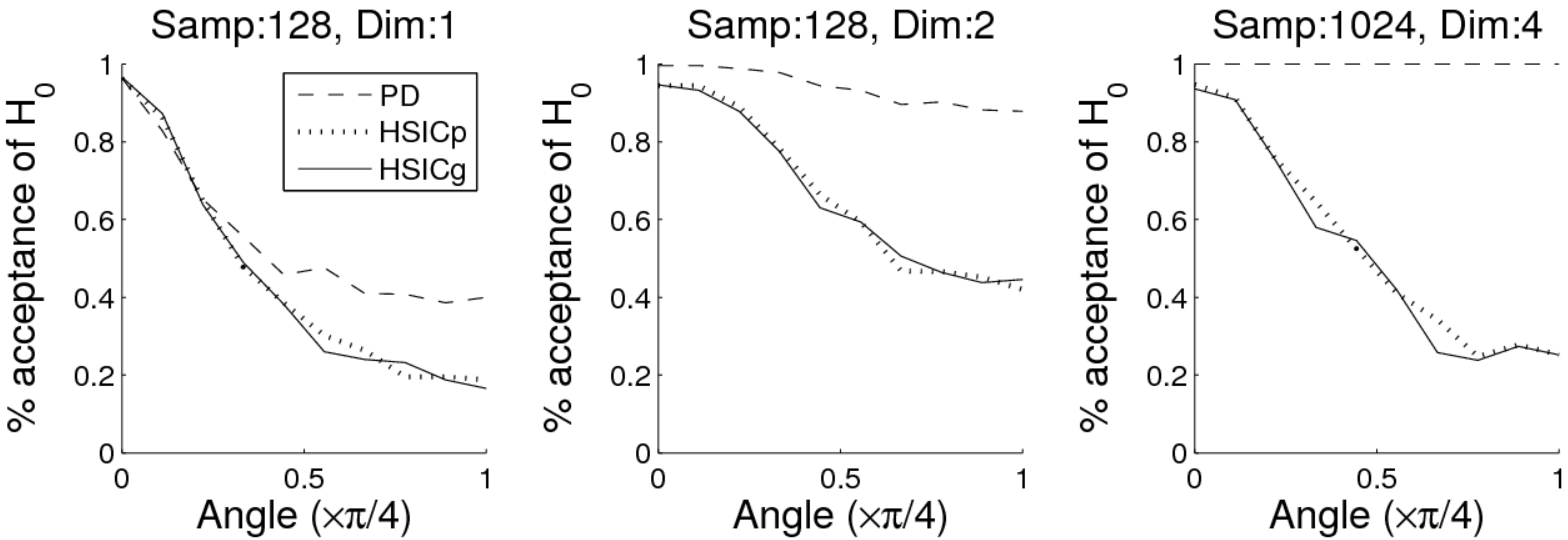
    - Compute the values of test statistics (dependence measure) for the samples.
    - Compute the critical region for a prescribed significance level.

Experiments of independence test

- Synthesized data: two d-dimensional samples

$$(X_1^{(1)},\dots,X_d^{(1)}),\dots,(X_1^{(N)},\dots,X_d^{(N)}) \qquad (Y_1^{(1)},\dots,Y_d^{(1)}),\dots,(Y_1^{(N)},\dots,Y_d^{(N)})$$

  - H_0: X and Y are independent
  - Significance level = 5%



strength of dependence

Power Divergence (Ku&Fine05, Read&Cressie)

- Make partition $\{A_j\}_{j \in J}$: Each dimension is divided into q parts so that each bin contains almost the same number of data.

- Power-divergence

$$T_N = 2I^{\lambda}(X,m) = N \frac{2}{\lambda(\lambda+2)} \sum_{j \in J} \hat{p}_j \left\{ \left(\hat{p}_j \Big/ \prod_{k=1}^{N} \hat{p}_{j_k}^{(k)} \right)^{\lambda} - 1 \right\}$$

I^0 = MI

I^2 = Mean Square Conting.

$\hat{p}_j$: frequency in A_j

$\hat{p}_r^{(k)}$: marginal freq. in r-th interval

- Null distribution under independence

$$T_N \quad \Rightarrow \quad \chi^2_{q^N - qN + N - 1} \qquad (N \to \infty)$$

- Estimation for high-dimensional data is difficult.

Outline

1. Introduction

2. Kernel measures for dependence

3. Kernel measures for conditional dependence

4. Causal inference with kernels
   – Kernel-based Causal Learning algorithm –

5. Conclusion

Conditional Covariance on RKHS

- Conditional Cross-covariance operator

  X, Y, Z : random variables on Ω_X, Ω_Y, Ω_Z (resp.).

  (H_X, k_X), (H_Y, k_Y), (H_Z, k_Z) : RKHS defined on Ω_X, Ω_Y, Ω_Z (resp.).

  – Conditional cross-covariance operator $H_X \rightarrow H_Y$

$$\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}$$

  – c.f. For Gaussian variables

    Conditional covariance of Y given X is equal to

$$V_{YX|Z} \equiv V_{YX} - V_{YZ}V_{ZZ}^{-1}V_{ZX}$$

(conditional covariance matrix)

Conditional independence with kernels

<u>Theorem</u>

Define the augmented variable $\tilde{X} = (X, Z)$ and define a kernel on $\Omega_X \times \Omega_Z$ by

$$k_{\tilde{X}} = k_X k_Z$$

Under some richness assumption, which is satisfied by Gaussian RBF kernels,

$$\Sigma_{Y\tilde{X}|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

$$\Sigma_{Y\tilde{X}|Z} = O \quad \Leftrightarrow \quad \Sigma_{\tilde{Y}X|Z} = O \quad \Leftrightarrow \quad \Sigma_{\tilde{Y}\tilde{X}|Z} = O \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid Z$$

Kernel conditional dependence measure

– Hilbert-Schmidt conditional independent criterion

$$HSCIC(X,Y \mid Z) = \left\| \Sigma_{\tilde{Y}\tilde{X}|Z} \right\|_{HS}^{2}$$

– Empirical measure

$$HSCIC_{emp}(X,Y \mid Z) = \left\| \hat{\Sigma}^{(N)}_{\tilde{Y}\tilde{X}} - \hat{\Sigma}^{(N)}_{\tilde{Y}Z} \left(\hat{\Sigma}^{(N)}_{ZZ} + \varepsilon_N I \right)^{-1} \hat{\Sigma}^{(N)}_{Z\tilde{X}} \right\|_{HS}^{2}$$

$$= \mathrm{Tr}\Big[G_X G_Y - 2 G_X (G_Z + N\varepsilon_N I_N)^{-1} G_Z G_Y + G_Z (G_Z + N\varepsilon_N I_N)^{-1} G_X (G_Z + N\varepsilon_N I_N)^{-1} G_Z G_Y \Big]$$

Consistency

If the regularization coefficient satisfies

$$\varepsilon_N \to 0 \qquad N^{1/3}\varepsilon_N \to \infty,$$

then

$$HSCIC_{emp} \to HSCIC \qquad (N \to \infty)$$

Conditional Independence Test

- Permutation test with the kernel measure

$$T_N = \left\| \hat{\Sigma}_{YX|Z}^{(N)} \right\|_{HS}^2$$

– If Z takes values in a finite set $\{1, \ldots, L\}$,

set $A_\ell = \{i \mid Z_i = \ell\}$ $(\ell = 1, \ldots, L)$,

otherwise, partition the values of Z into L subsets $C_1, \ldots, C_L$, and set

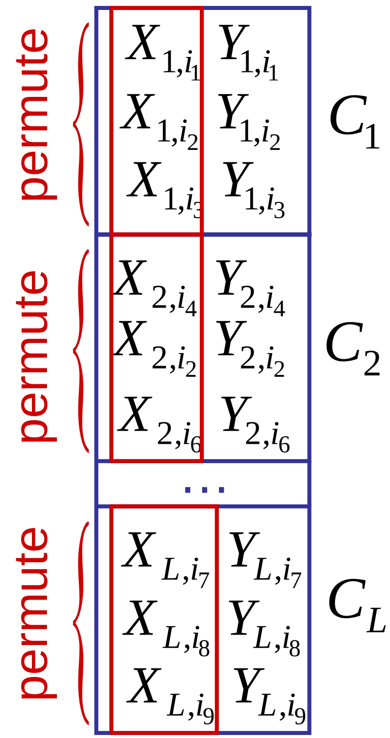
$$A_\ell = \{i \mid Z_i \in C_\ell\} \quad (\ell = 1, \ldots, L).$$

– Repeat the following process B times: ($b = 1, \ldots, B$)

1. Generate pseudo cond. independent data $D^{(b)}$ by permuting X data within each A_ℓ.
2. Compute $T_N^{(b)}$ for the data $D^{(b)}$.

⟶ Approximate null distribution under cond. indep. assumption

– Set the threshold by the $(1\text{-}\alpha)$-percentile of the empirical distributions of $T_N^{(b)}$.



Outline

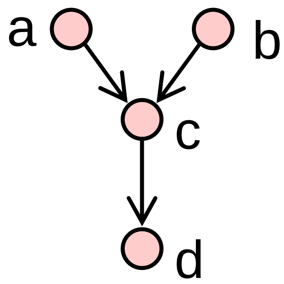
1. Introduction
2. Kernel measures for dependence
3. Kernel measures for conditional dependence
4. Causal inference with kernels
   – Kernel-based Causal Learning algorithm –
5. Conclusion

Causal Inference from Non-Experimental Data

- Constraint-based method
  - Determine the (cond.) independence of the underlying probability.
  - Relatively efficient for hidden variables.
- Score-based method
  - Structure learning of Bayesian network
  - Able to use informative prior.
  - Optimization in huge search space.
  - Many methods assume discrete variables (discretization) or parametric model.
- Kernel-based Causal Learning
  - Constraint-based method. A variant of Inductive Causation (IC)

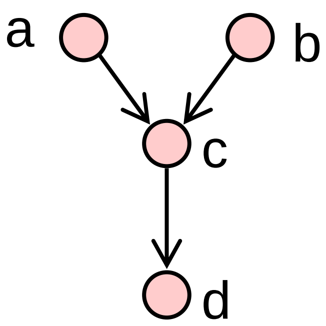
Fundamental Assumptions

- Causal Markov Condition
  - Causal relation is expressed by a DAG, and the probability generating data is consistent with the graph.

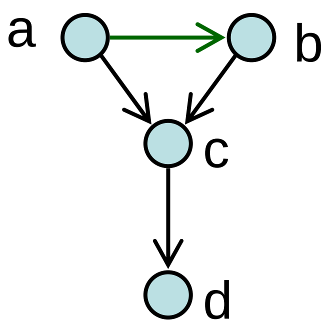


$$p(X) = p(X_a)p(X_b)p(X_c \mid X_a, X_b)p(X_d \mid X_c)$$

- Causal Faithfulness Condition
  - The inferred DAG (causal structure) must express all the independence relations.



true



unfaithful

This includes the true probability as a special case, but the structure does not express $a \perp\!\!\!\perp b$

Inductive Causation

- IC algorithm (Verma&Pearl 90)

  Input – V: set of variables, D: dataset of the variables.

  Output – DAG (specifies an equivalence class, directed partially)

1. For each $(a,b) \in V \times V \ (a \neq b)$, search for $S_{ab} \subset V \setminus \{a,b\}$ such that
   $$X_a \perp\!\!\!\perp X_b \mid S_{ab}$$
   Construct an undirected graph (skeleton) by making an edge between a and b if and only if no set S_{ab} can be found.
2. For each nonadjacent pair (a,b) with $a - c - b$, direct the edges by $a \rightarrow c \leftarrow b$ if $c \notin S_{ab}$
3. Orient as many of undirected edges as possible on condition that neither new v-structures nor directed cycles are created.

Kernel-based Causal Leaning

- Limitations of the previous implementations of IC
  - Linear / discrete assumptions in Step 1.

    e.g. PC-algorithm (Spirtes & Glymour 91) uses partial correlation and χ^2 test.

    Difficulty in testing conditional independence for continuous variables.

    → kernel method!

  - Errors of the skeleton in Step 1 cannot be recovered in the later steps.

    → voting method for direction

Note: The error in Step 1 is inevitable by statistical tests.

■ KCL algorithm (Sun et al. ICML07, Sun et al. 2007)

– Dependence measure: $\hat{\mathbb{H}}_{YX}^{(N)} = HSIC = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^{2}$

– Conditional dependence measure: $\hat{\mathbb{H}}_{YX|Z}^{(N)} \equiv \dfrac{\left\| \hat{\Sigma}_{\tilde{Y}\tilde{X}|Z}^{(N)} \right\|_{HS}^{2}}{\left\| C_{ZZ} \right\|_{HS}^{2}}$

where the operator $C_{ZZ} : H_Z \to H_Z$ is defined by

$$\langle f, C_{ZZ} g \rangle = E\left[f(Z) g(Z) \right]$$

Motivation: make $\left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^{2}$ and $\left\| \hat{\Sigma}_{\tilde{Y}\tilde{X}|Z}^{(N)} \right\|_{HS}^{2}$ comparable

<u>Theorem</u>

If $(X, Y) \perp\!\!\!\perp Z$, $\left\| \hat{\Sigma}_{\tilde{Y}\tilde{X}|Z}^{(N)} \right\|_{HS}^{2} = \left\| C_{ZZ} \right\|_{HS}^{2} \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^{2}$

Outline of KCL algorithm: IC algorithm is modified as follows.

KCL-1: Skeleton by statistical tests with the kernel measure $\hat{\mathbb{H}}^{(N)}_{YX|Z}$

(1) Permutation tests of conditional independence $X \perp\!\!\!\perp Y \mid S_{XY}$

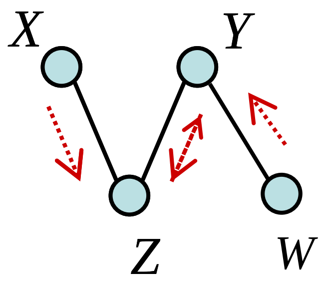
(2) Connect X and Y if no such S_{XY} exists.

The candidates of S_{XY} should be restricted → explained later.

KCL-2: Voting for unshielded triplets

For each triplet $X - Z - Y$ (X and Y not adjacent), compute

$$M_{XY|Z} \equiv \frac{\hat{\mathbb{H}}^{(N)}_{YX|Z}}{\hat{\mathbb{H}}^{(N)}_{YX}}, \quad M_{YZ|X}, \quad M_{ZX|Y}$$



Give a vote to the direction $X \rightarrow Z$ and $Y \rightarrow Z$ if

$$M_{XY|Z} > \max\{M_{YZ|X}, M_{ZX|Y}\}$$

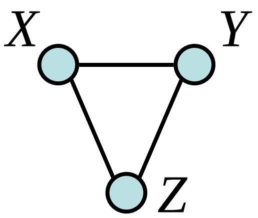
Make an arrow to each edge if a vote is given ("↔" is allowed).

KCL-3: Same as IC-3

KCL-4: Voting for shielded triplets

For each triplet $X - Z - Y$ (X and Y adjacent), compute

$$M_{XY|Z},\ M_{YZ|X},\ M_{ZX|Y}$$



Give a vote to the direction $X \rightarrow Z$ and $Y \rightarrow Z$ if

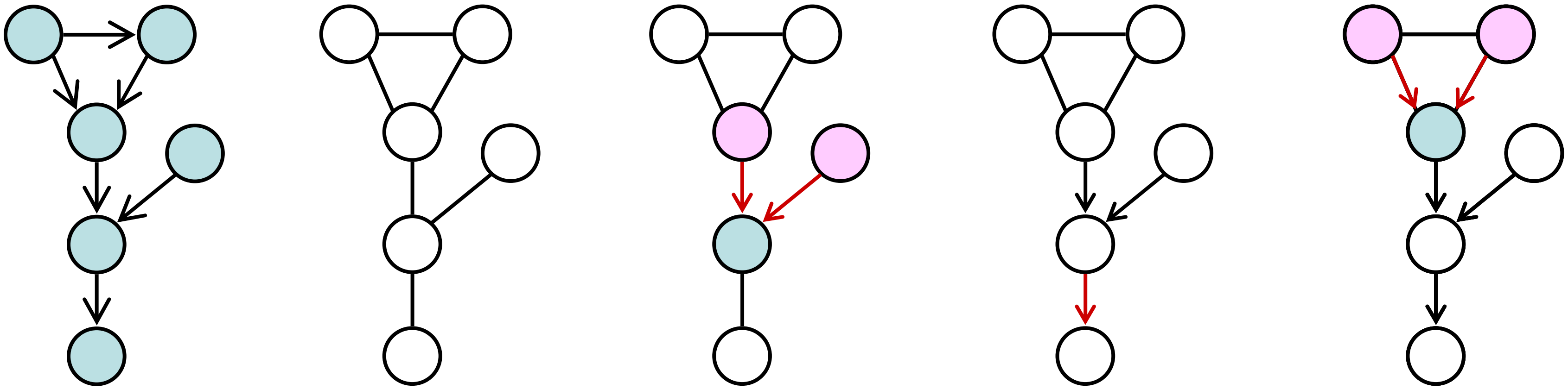
$$M_{XY|Z} > \max\{M_{YZ|X}, M_{ZX|Y}\}$$

Make an arrow to each edge if a vote is given ("$\leftrightarrow$" is allowed).

– The resulting graph is mixed: undirected $—$, directed $\longrightarrow$, or bi-directed $\longleftrightarrow$.

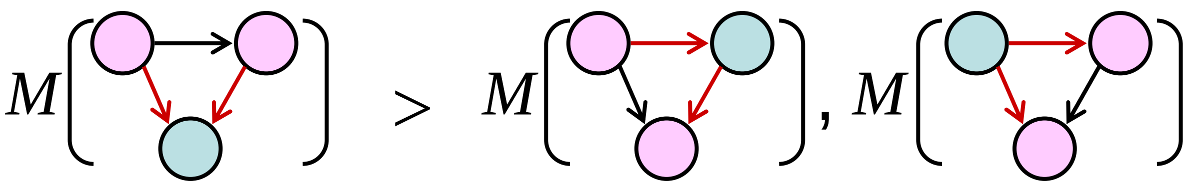
– Motivation of KCL-2 and 4:
  - By inevitable errors in statistical tests, it is preferred that the orientation process be separated from Step 1.
  - Step 4 looks for more directed edges.
    It relies on the heuristic assumption that conditioning common effect strengthens the dependence between the causes.

■ Illustration of KCL



true KCL-1 KCL-2 KCL-3 KCL-4

Heuristic assumption:



Conditioning common effect strengthens the dependence between the causes.

Details of Step 1

Auxiliary partially directed graphs are used for restricting conditioning variables S_{XY} .

- Initialize G by a complete undirected graph.
- 1(a): Unconditional independence tests

  For all pairs (X,Y), apply permutation tests for $X \perp\!\!\!\perp Y$ with $\hat{\mathbb{H}}_{YX}^{(N)}$

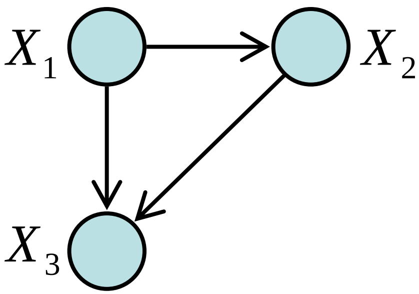
  Remove $X - Y$ if the independence is accepted.
- 1(b): Auxiliary graph

  Orient G by majority votes on all triplets $X - Y - Z$.
- 1(c): Cond. indep. tests $X \perp\!\!\!\perp Y \mid S_{XY}$ with $\hat{\mathbb{H}}_{YX|Z}^{(N)}$ in the auxiliary graph.

  S_{XY}: only variables in the directed (incl. undirected) path between X and Y.
- 1(d): Change the directed edges into undirected ones to make a skeleton G.
- 1(e): Repeat (a)-(d) until nothing changes.

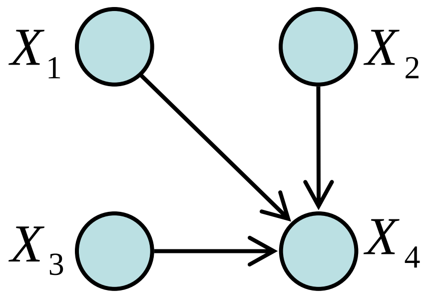
Experiments with Simple Networks

(A)



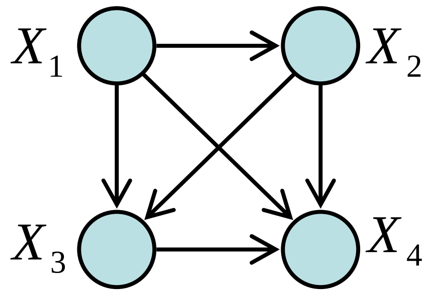
$$P(X_1 = 1) = 0.6$$
$$P(X_2 = X_1 \mid X_1) = 0.8$$
$$X_3 = \text{NoisyOR}(X_1, X_2)$$

(B)



$$P(X_1 = 1) = 0.6$$
$$P(X_2 = 1) = 0.5$$
$$P(X_3 = 1) = 0.4$$
$$X_4 = \text{NoisyOR}(X_1, X_2, X_3)$$

(C)



$$P(X_1 = 1) = 0.6$$
$$P(X_2 = X_1 \mid X_1) = 0.8$$
$$X_3 = \text{NoisyOR}(X_1, X_2)$$
$$X_4 = \text{NoisyOR}(X_1, X_2, X_3)$$

$$X_{n+1} = \text{NoisyOR}(X_1, \ldots, X_n)$$
$$\Leftrightarrow \quad P(X_{n+1} = 1 \mid X_1, \ldots, X_n) = 0.8 \times \left(1 - 0.2^{X_1 + \cdots + X_n}\right) + 0.2$$

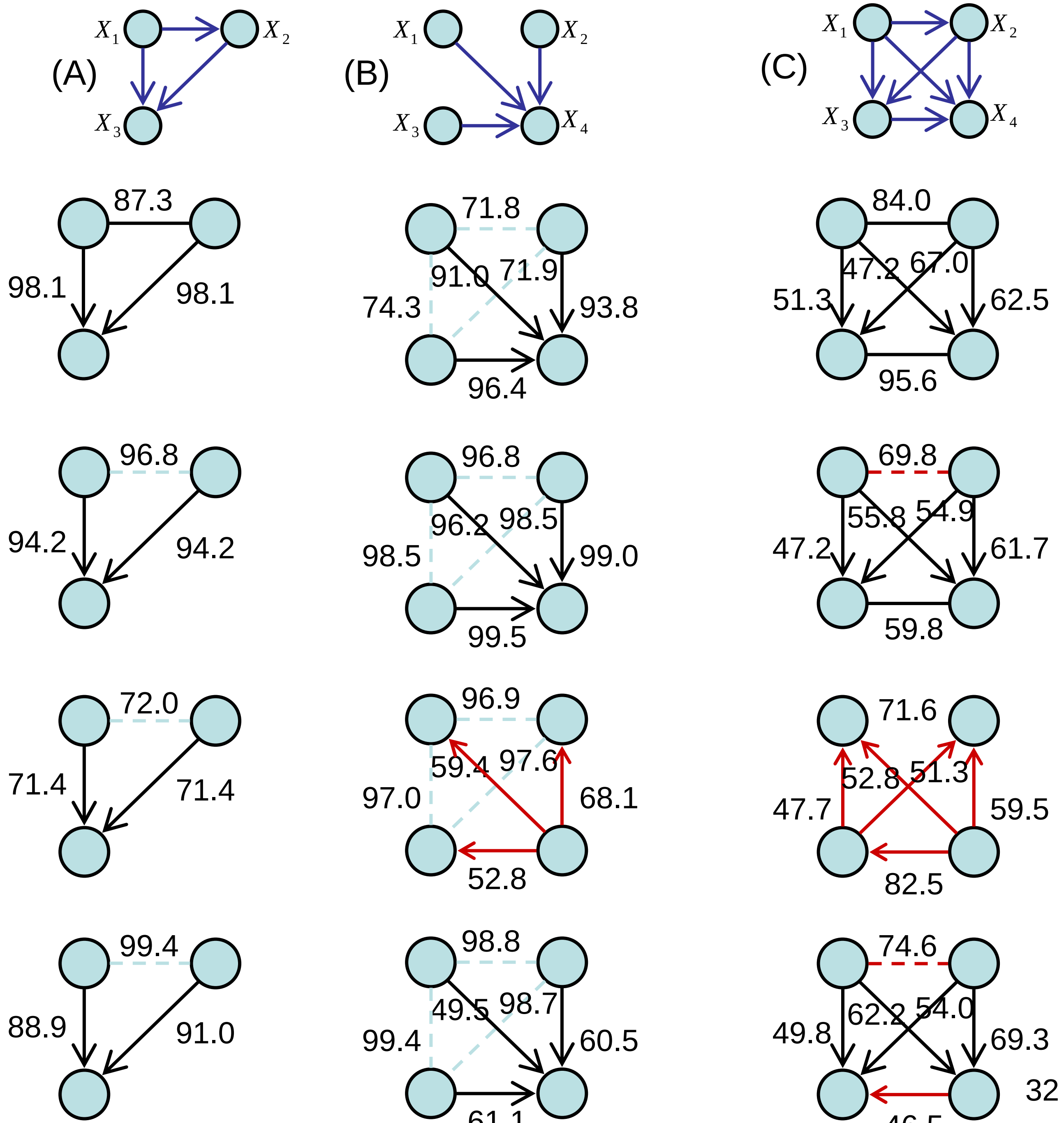
- Results (200 data, 1000 runs)

KCL

PC

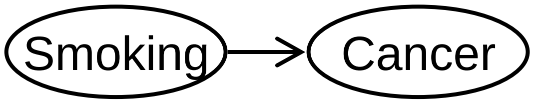
BN-PC (MI is used) [Cheng et al. '02]

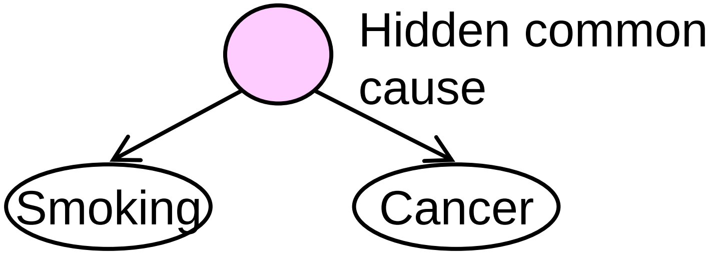
BDe (Score-based) [Heckerman et al. '97]



Hidden Common Cause

– One of the difficulties in causal leaning is possible existence of common hidden causes.





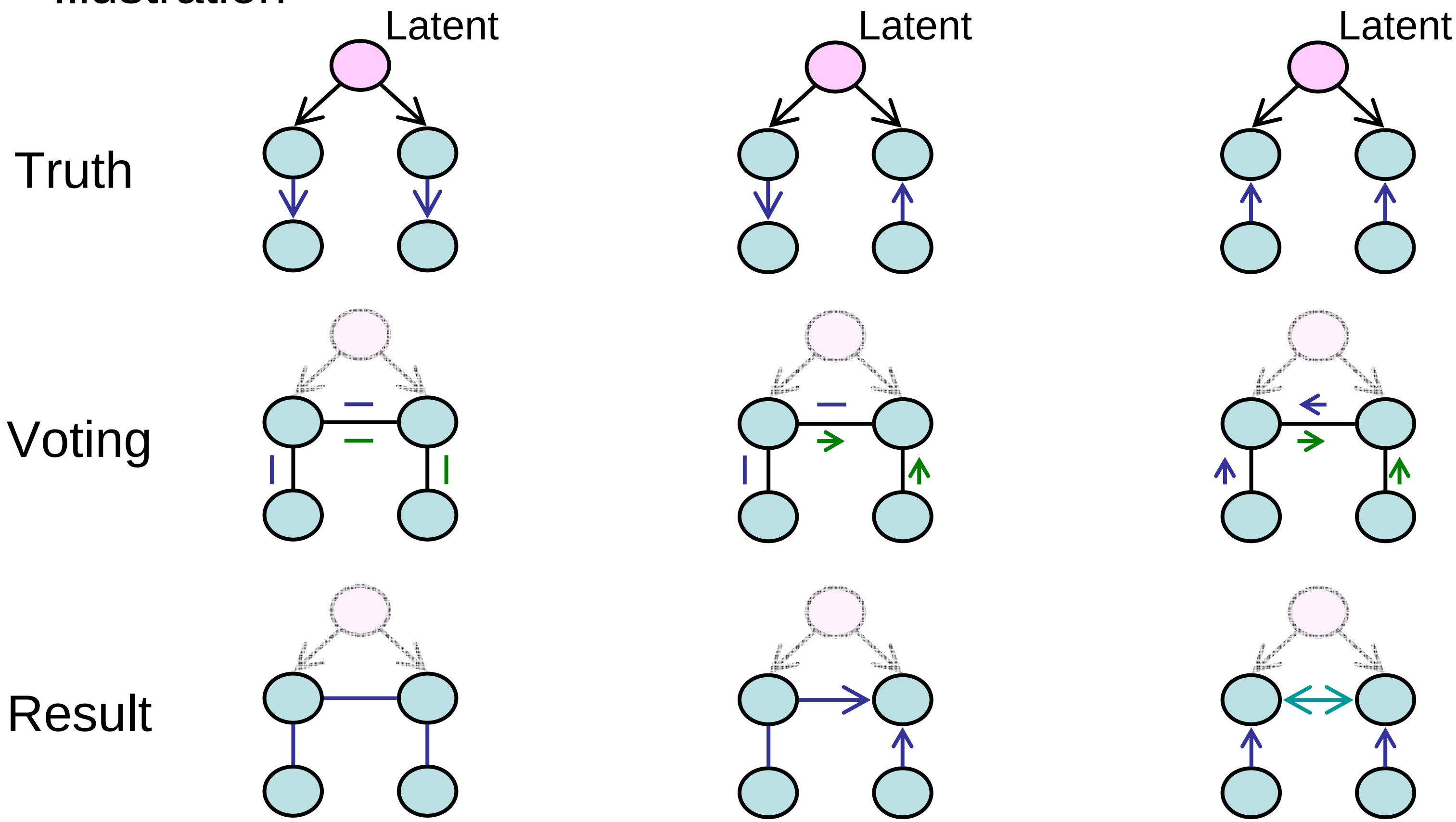
– Some methods can handle hidden variables.

FCI (Fast Causal Inference, Spirtes et al. 93) extends PC to allow hidden variables.

KCL for hidden common causes

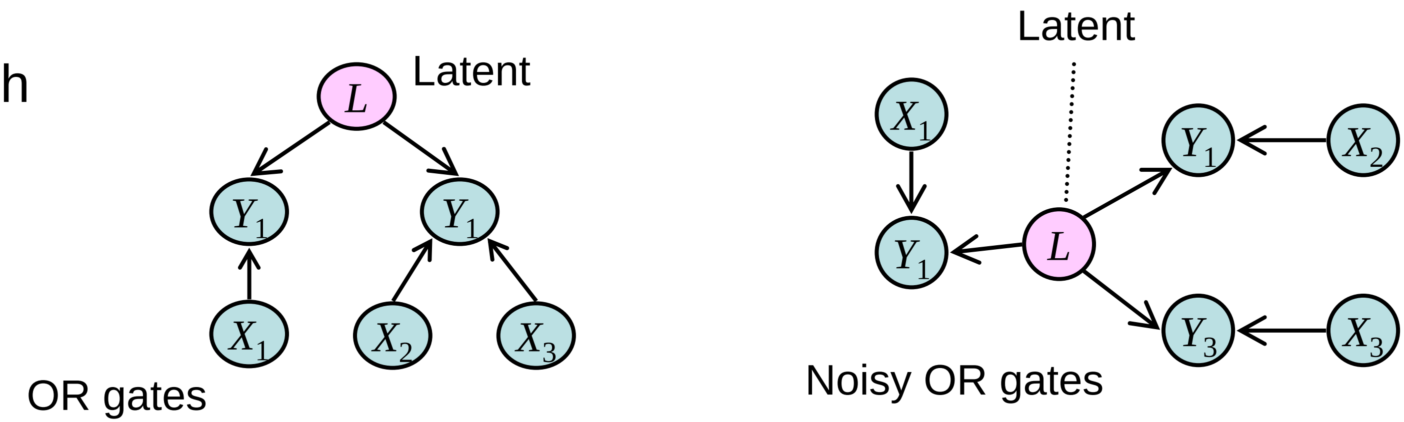
– A bi-directional arrow (↔) given by KCL may suggest existence of a hidden common cause.
Empirically verified in some situations, but no theoretical justification.

– Illustration

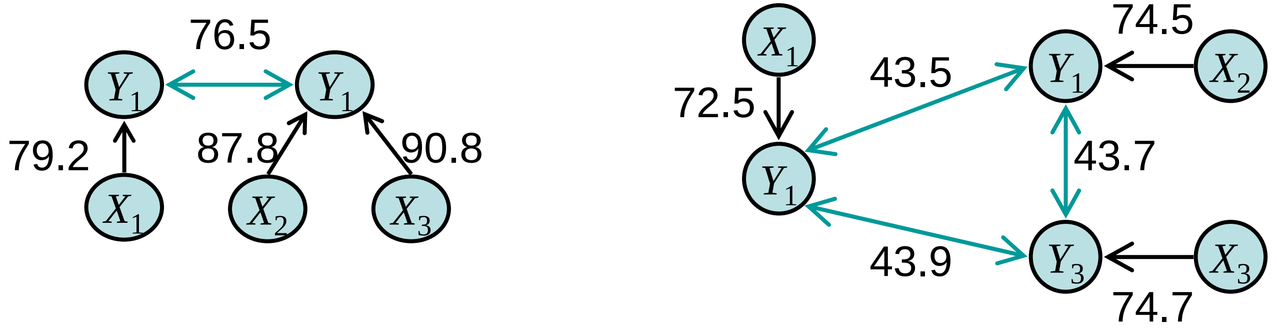


– Experiments (200 data, 1000 runs)

Truth



Result of KCL

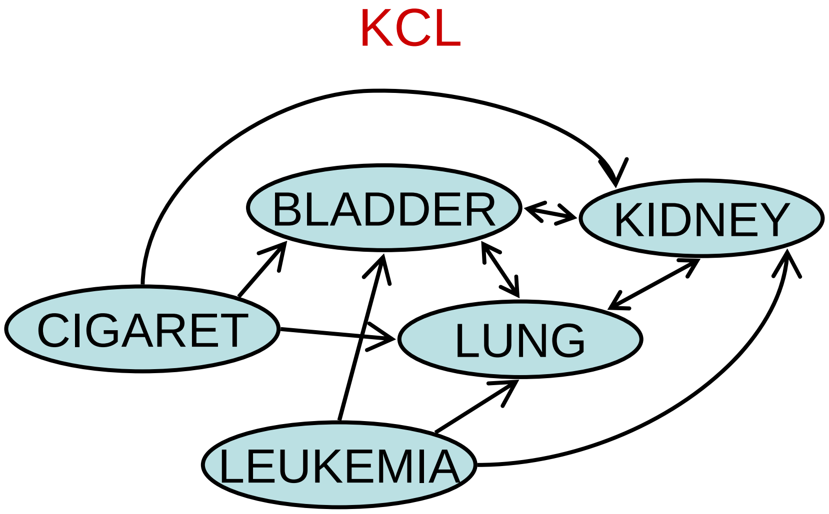


Experiments with Real Data

- Smoking and Cancer
  - Data: 5 continuous variables, $N = 44$

    CIGARET: Cigarettes sales in 43 states in US and District of Columbia

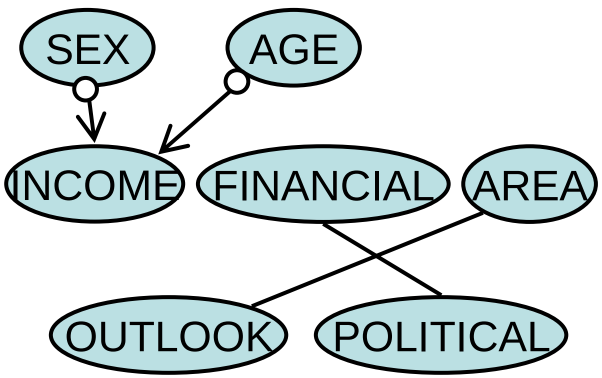
    BLADDER, LUNG, KIDNEY, LEUKEMIA: death rates from various cancers
  - Results



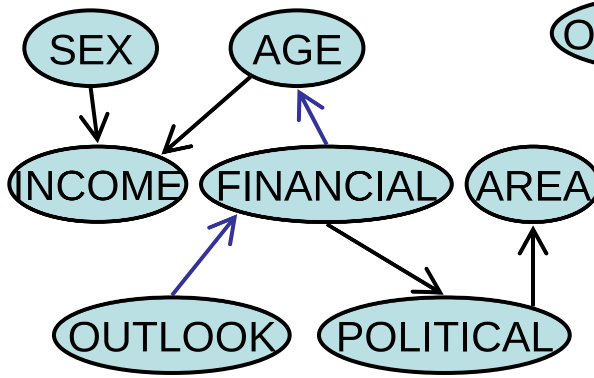
■ Montana Economic Outlook Poll (1992)

– Data: 7 discrete variables, $N = 209$

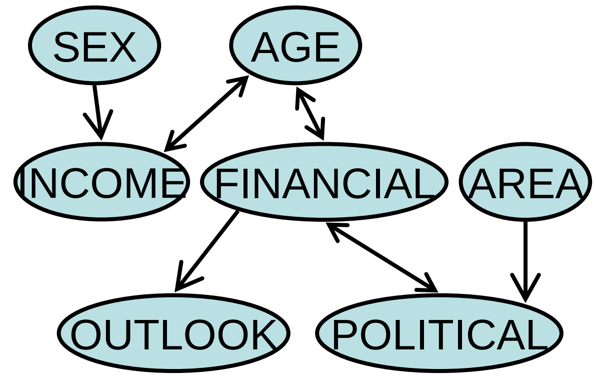
AGE (3), SEX (2), INCOME (3), POLITICAL (3), AREA (3), FINANCIAL status (3, better/same/worse than a year ago), OUTLOOK (2)



FCI



BN-PC



KCL

Conclusion

- Kernel measures of (conditional) dependence
  - Covariance and conditional covariance considered on RKHS provide criterion of independence and conditional independence, resp.
  - Kernel measures are proposed for (conditional) dependence.
- Causal inference from non-experimental data
  - Kernel-based Causal Learning (KCL) algorithm
    - Constraint-based method: A variant of Inductive Causation
      - Conditional independence tests with kernel measures
      - Voting method for orienting edges
    - KCL can handle discrete and continuous domains in a unified way.
    - More theoretical justification is required.

References

Sun, X., D. Janzing, B. Schölkopf, and K. Fukumizu. A kernel-based causal learning algorithm. *Proc. 24th Intern. Conf. Machine Learning* (*ICML2007*), pp.855-862. (2007)

Sun, X., D. Janzing, B. Schölkopf, K. Fukumizu, and A. Gretton. Learning causal structures via kernel-based statistical dependence measures. *Submitted* (2007)

Fukumizu, K., F. Bach, and M. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Leaning Research*, 5:73-99 (2004).

Fukumizu, K., F. Bach, and M. Jordan. Kernel dimension reduction in regression. Tech Report 715, Dept. Statistics, University of California, Berkeley, 2006.

Gretton, A., O. Bousquet, A. Smola and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory: 16th International Conference, ALT 2005*, pp.63-78 (2005)