Painless embeddings of distributions: the function space view

Part 3 - Conditional independence with kernels

Arthur Gretton (MPI), Alex Smola (NICTA) , Kenji Fukumizu (ISM)
fukumizu@ism.ac.jp

ICML 2008 Tutorial
July 5, Helsinki, Finland

Outline of Part 3

I. Introduction

II. Conditional independence with kernels

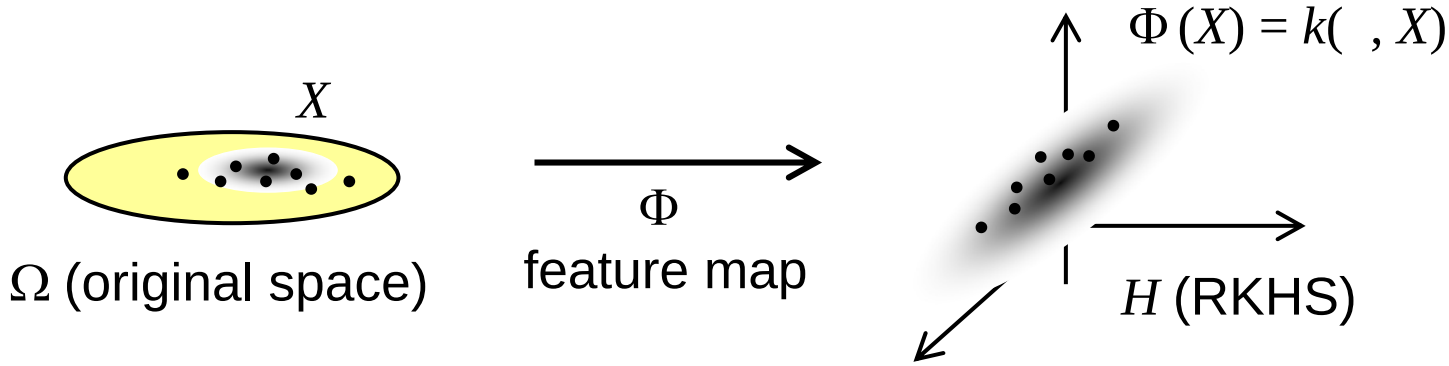
III. Application to causal inference

IV. Summary

I. Introduction

Functional Space View

Embedding into RKHS



Basic statistics on RKHS

- Mean element → characterizes a probability
- Covariance operator → independence/dependence
- Conditional covariance operator

  → conditional independence/dependence

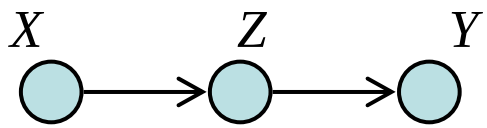
Conditional Independence

Definition

X, Y, Z: random variables with joint probability density $p_{XYZ}(x, y, z)$

X and Y are conditionally independent given Z, if

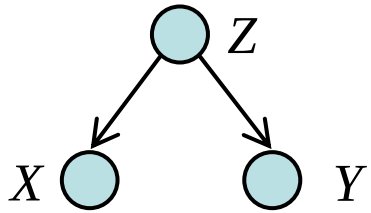
$$p_{Y|ZX}(y \mid z, x) = p_{Y|Z}(y \mid z)$$



If Z is known, the information of X is not necessary to predict Y.

or

$$p_{XY|Z}(x, y \mid z) = p_{X|Z}(x \mid z)\, p_{Y|Z}(y \mid z)$$



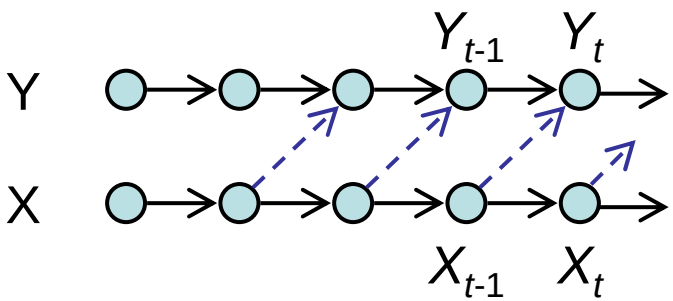
Example

- Applications in statistical inference
  - Graphical modeling:

    Separability in a graph implies conditional independence
  - Causal inference

    A formulation of causality is given by conditional independence

- Example: Time series
  - X causes Y?



Non-causality

$$p(Y_t \mid Y_{t-1}, X_{t-1}) = p(Y_t \mid Y_{t-1}) \; \mathbf{?}$$

$$\Updownarrow$$

$$Y_t \perp\!\!\!\perp X_{t-1} \mid Y_{t-1} \quad \mathbf{?}$$

II. Conditional independence with kernels

Review: Conditional Independence for Gaussian Variables

- Conditional covariance of Gaussian variables
  - (X, Y, Z) : multidimensional jointly Gaussian variable
  - Conditional covariance matrix

$$V_{YX|Z} \equiv \mathrm{Cov}[Y, X \mid Z = z] = V_{YX} - V_{YZ} V_{ZZ}^{-1} V_{ZX}$$

V_{XY} etc.: covariance matrix

Note: $V_{YX|Z}$ does not depend on the value of z

- Conditional independence for Gaussian variables

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{XY|Z} = O \qquad \text{i.e.} \quad V_{YX} - V_{YZ} V_{ZZ}^{-1} V_{ZX} = O$$

Conditional Covariance on RKHS

Conditional Cross-covariance operator

X, Y, Z : random variables taking values on $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ (resp.).

$(H_{\mathcal{X}}, k_{\mathcal{X}})$, $(H_{\mathcal{Y}}, k_{\mathcal{Y}})$, $(H_{\mathcal{Z}}, k_{\mathcal{Z}})$: RKHS defined on $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ (resp.).

– Conditional cross-covariance operator $H_{\mathcal{X}} \to H_{\mathcal{Y}}$

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

(Σ_{YX} etc.: covariance operators)

$$c.f. \quad V_{YX|Z} = V_{YX} - V_{YZ}V_{ZZ}^{-1}V_{ZX}$$

Note: Σ_{ZZ}^{-1} may not exist. But, we have the decomposition

$$\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2} \qquad \text{with} \quad \| W_{YX} \| \leq 1$$

Rigorously, define $\Sigma_{YX|Z} \equiv \Sigma_{YX} - \Sigma_{YY}^{1/2} W_{YZ} W_{ZX} \Sigma_{XX}^{1/2}$

$*$ $A^{1/2} = U\Lambda^{1/2}U^T$ if $A = U\Lambda U^T$.

Conditional Covariance and Conditional Covariance Operator

- Cond. cov. operator expresses cond. covariance

Theorem (FBJ'06, Sun et al. '07)

X, Y, Z : random variables taking values on $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ (resp.).

$(H_{\mathcal{X}}, k_{\mathcal{X}})$, $(H_{\mathcal{Y}}, k_{\mathcal{Y}})$, $(H_{\mathcal{Z}}, k_{\mathcal{Z}})$: RKHS defined on $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ (resp.).

Assume $k_{\mathcal{Z}}$ is a characteristic kernel.

$$\left\langle g, \Sigma_{YX|Z} f \right\rangle = E\left[Cov[g(Y), f(X) \mid Z]\right]$$

or

$$\Sigma_{YX|Z} = E_Z\left[\int \Phi_{\mathcal{Y}}(Y) \otimes \Phi_{\mathcal{X}}(X) dP(X, Y \mid Z)\right] - E_Z\left[\int \Phi_{\mathcal{Y}}(Y) \otimes \Phi_{\mathcal{X}}(X) dP(X \mid Z) dP(Y \mid Z)\right]$$

$$= \mu_{XY} - \mu_{E_Z[Y \perp\!\!\!\perp X|Z]}$$

– *c.f.* for Gaussian variables

$$a^T V_{XY|Z} b = Cov[a^T X, b^T Y \mid Z]$$

(not dependent on the value of z)

Conditional Independence with Kernels

(FBJ2004, FBJ2006, Sun et al. 2007)

Extended variables are used.

$$\ddot{X} = (X, Z), \quad \ddot{Y} = (Y, Z) \qquad k_{\ddot{x}} = k_x k_z, \quad k_{\ddot{y}} = k_y k_z$$

Theorem (FBJ'06, Sun et al. '07)

Assume the kernels $k_{\ddot{x}}$, k_y, and k_z are characteristic. Then,

$$X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad \Sigma_{Y\ddot{X}|Z} = O \qquad \left(\Leftrightarrow \quad \Sigma_{\ddot{Y}X|Z} = O \Leftrightarrow \Sigma_{\ddot{Y}\ddot{X}|Z} = O\right)$$

- *c.f.* for Gaussian variables, $X \perp\!\!\!\perp Y \mid Z \quad \Leftrightarrow \quad V_{XY|Z} = O$
- With characteristic kernels, comparison between the (conditional) mean elements on RKHS characterizes conditional independence.
- Why is the "extended variable" needed?

$$\Sigma_{YX|Z} = O \quad \Rightarrow \quad p(x,y) = \int p(x \mid z)\, p(y \mid z) p(z) dz$$

$$\Sigma_{Y[X,Z]|Z} = O \quad \Rightarrow \quad p(x,y,z') = \int p(x,z' \mid z)\, p(y \mid z) p(z) dz$$

where $p(x,z' \mid z) = p(x \mid z)\delta(z'-z)$

Measure of Conditional Independence

- Hilbert-Schmidt norm of cond. cov. operator

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

With characteristic kernels, $HSCIC = 0 \Leftrightarrow X \perp\!\!\!\perp Y \mid Z$

- Empirical estimation is painless!

$(X_1, Y_1, Z_1), \ldots, (X_N, Y_N, Z_N)$: data

$$\Sigma_{\ddot{X}Z} \rightarrow \hat{\Sigma}_{\ddot{X}Z}^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \left(k_{\ddot{X}}(\cdot, \ddot{X}_i) - \hat{m}_{\ddot{X}} \right) \otimes \left(k_Z(\cdot, Z_i) - \hat{m}_Z \right), \quad \Sigma_{ZZ}^{-1} \rightarrow \left(\hat{\Sigma}_{ZZ}^{(N)} + \varepsilon_N I \right)^{-1}$$

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

$$HSCIC_{emp} = \mathrm{Tr} \Big[\tilde{K}_{\ddot{X}} \tilde{K}_{\ddot{Y}} - 2 \tilde{K}_{\ddot{X}} \left(\tilde{K}_Z + N\varepsilon_N I_N \right)^{-1} \tilde{K}_Z \tilde{K}_{\ddot{Y}} + \tilde{K}_Z \left(\tilde{K}_Z + N\varepsilon_N I_N \right)^{-1} K_{\ddot{X}} \left(\tilde{K}_Z + N\varepsilon_N I_N \right)^{-1} \tilde{K}_Z \tilde{K}_{\ddot{Y}} \Big]$$

Normalized Cond. Covariance

- Normalized conditional cross-covariance operator

$$W_{YX|Z} = \Sigma_{YY}^{-1/2}\Sigma_{YX|Z}\Sigma_{XX}^{-1/2} = \Sigma_{YY}^{-1/2}\left(\Sigma_{YX} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)\Sigma_{XX}^{-1/2}$$

Recall: $\Sigma_{YX} = \Sigma_{YY}^{1/2} W_{YX} \Sigma_{XX}^{1/2}$

$$HSNCIC = \left\| W_{\ddot{X}\ddot{Y}|Z} \right\|_{HS}^2$$

$$HSNCIC_{emp} = \mathrm{Tr}\left[R_{\ddot{X}} R_{\ddot{Y}} - 2R_{\ddot{X}} R_{\ddot{Y}} R_Z + R_{\ddot{X}} R_Z R_{\ddot{Y}} R_Z\right]$$

$$R_{\ddot{X}} \equiv \tilde{K}_{\ddot{X}}\left(\tilde{K}_{\ddot{X}} + N\varepsilon_N I_N\right)^{-1} \text{ etc.}$$

- Kernel-free expression. With characteristic kernels,

$$\| W_{\ddot{Y}\ddot{X}|Z} \|_{HS}^2 = \iiint \left(\frac{p_{XYZ}(x,y,z) - p_{X|Z}(x\,|\,z)\, p_{Y|Z}(y\,|\,z)\, p_Z(z)}{p_{XZ}(x,z)\, p_{YZ}(y,z)} \right)^2 p_{XZ}(x,z)\, p_{YZ}(y,z)\, dxdydz$$

("Conditional" mean square contingency)

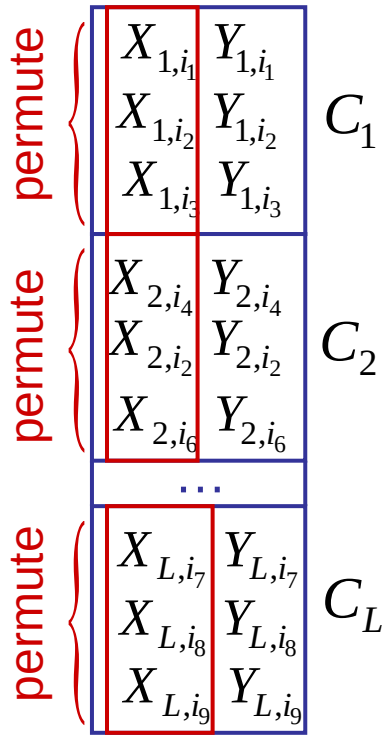
Conditional Independence Test

- **Background**
  - There are no good methods for conditional independence test on non-Gaussian continuous variables. (*e.g.* discretizing all variables).
- **Permutation test with the kernel measure**

$$T_N = HSCIC_{emp} \quad \text{or} \quad T_N = HSNCIC_{emp}$$

  - Partition the values of Z into $C_1, \ldots, C_L$, and define $A_\ell = \{i \mid Z_i \in C_\ell\}$ $(\ell = 1, \ldots, L)$.
  - Resampling (for b = 1,2,…)
    1. Generate pseudo cond. indep. sample $D^{(b)}$ by permuting X within each A_ℓ.
    2. Compute $T_N^{(b)}$ for the sample $D^{(b)}$.

    ⟶ Approximate the null distribution by samples.
  - Set the threshold for the significance level (*e.g.* 5%).



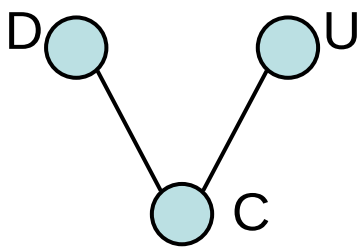
Application to Graphical Modeling

– Three continuous variables of medical measurements. N = 35. (Edwards 2000, Sec.3.1.4)

Creatinine clearance (C), Digoxin clearance (D), Urine flow (U)

Kernel mehod (permut. test)			Linear method		
	HSN(C)IC	P-val.		(partial) cor.	P-val.
$D \perp\!\!\!\perp U \mid C$	1.458	0.924	Parcor(D,U|C)	0.4847	0.0037
$C \perp\!\!\!\perp D$	0.776	<0.001	Cor(C,D)	0.7754	0.0000
$C \perp\!\!\!\perp U$	0.194	0.117	Cor(C,U)	0.3092	0.0707
$D \perp\!\!\!\perp U$	0.343	0.023	Cor(D,U)	0.5309	0.0010

– Suggested undirected graphical model by kernel method

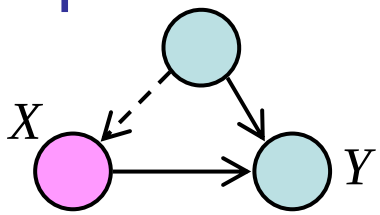


The conditional independence $D \perp\!\!\!\perp U \mid C$ coincides with the medical knowledge.

III. Application to causal inference

Causal Inference

- With manipulation – intervention



manipulate observation

X is a cause of Y?

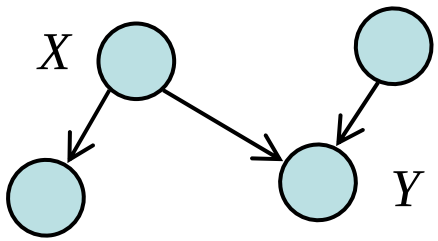
Easier. (*do*-calculus, Pearl 1995)

- No manipulation / with temporal information

$X(t)$ $Y(t)$: observed time series

$X(1), \ldots, X(t)$ are a cause of $Y(t+1)$?

- No manipulation / no temporal information



Causal inference is harder.

Causality of Time Series

Causality by conditional independence

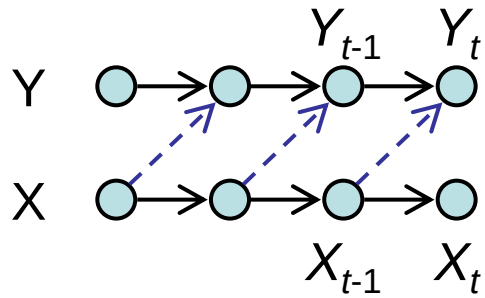
- Extended notion of Granger causality (linear AR)

  X is NOT a cause of Y if

$$p(Y_t \mid Y_{t-1},...,Y_{t-p}, X_{t-1},..., X_{t-p}) = p(Y_t \mid Y_{t-1},...,Y_{t-p})$$

$$\Leftrightarrow$$

$$Y_t \perp\!\!\!\perp X_{t-1},..., X_{t-p} \mid Y_{t-1},...,Y_{t-p}$$



- Kernel measures for causality

$$HSCIC = \left\| \hat{\Sigma}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$$

$$HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$$

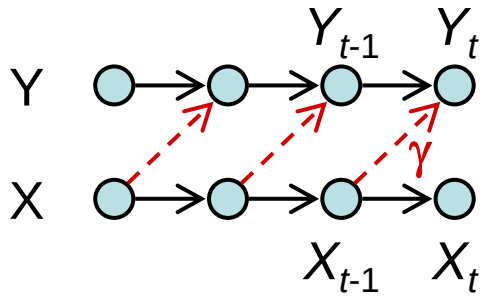
$$\mathbf{X}_{\mathrm{p}} = \{(X_{t-1,} X_{t-2}, \cdots, X_{t-p}) \in \mathbf{R}^p \mid t = p+1,..., N\}$$

$$\mathbf{Y}_{\mathrm{p}} = \{(Y_{t-1,} Y_{t-2}, \cdots, Y_{t-p}) \in \mathbf{R}^p \mid t = p+1,..., N\}$$

Example

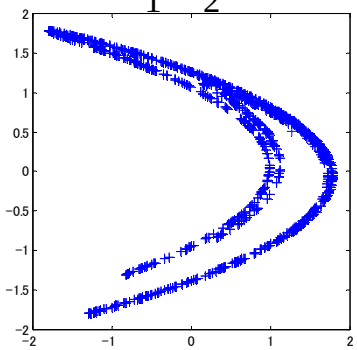
- Coupled Hénon map
  - X, Y:

$$\begin{cases} x_1(t+1) = 1.4 - x_1(t)^2 + 0.3x_2(t) \\ x_2(t+1) = x_1(t) \end{cases}$$

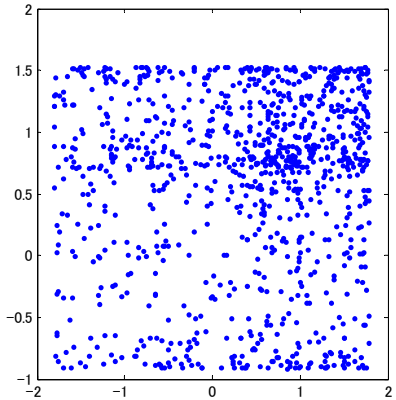


$$\begin{cases} y_1(t+1) = 1.4 - \left\{ \gamma x_1(t) y_1(t) + (1-\gamma) y_1(t)^2 \right\} + 0.1 y_2(t) \\ y_2(t+1) = y_1(t) \end{cases}$$

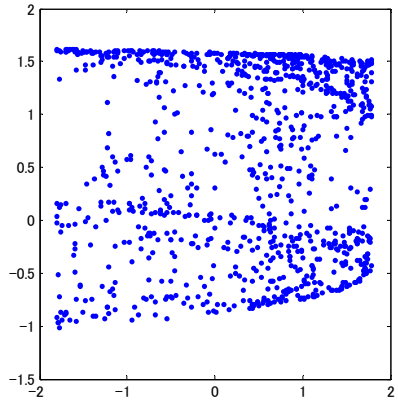
x_1-x_2



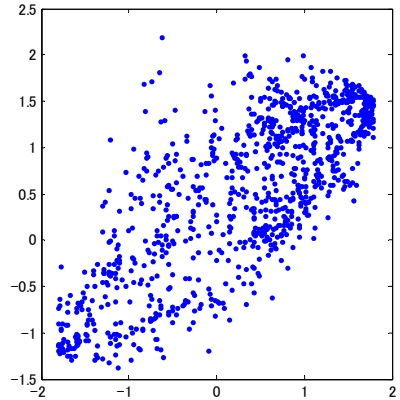
x_1-y_1



γ = 0



γ = 0.25

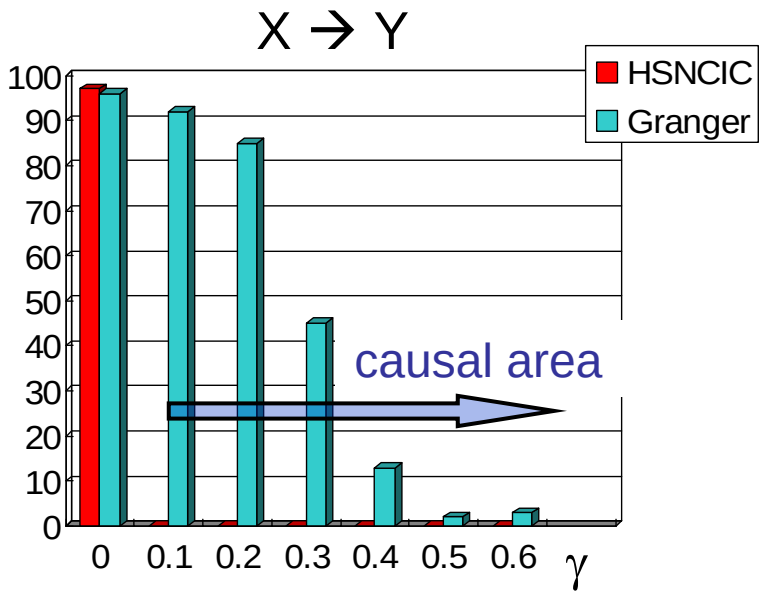


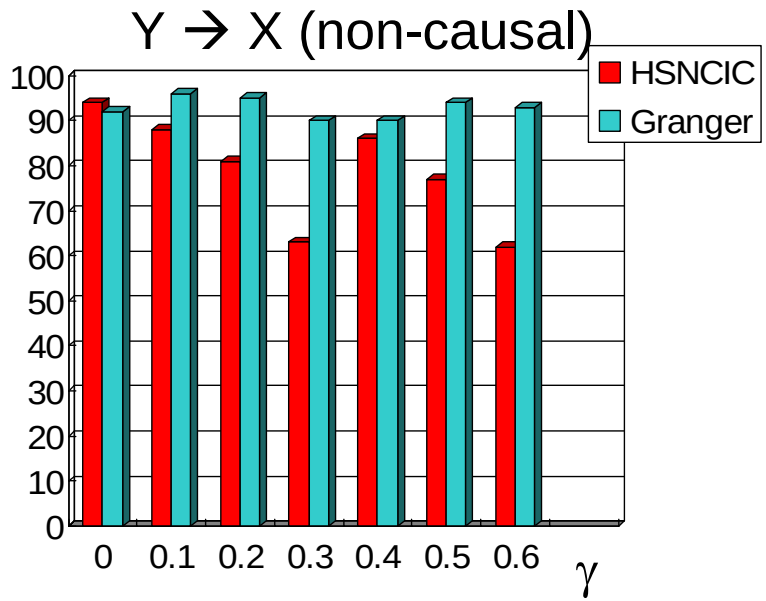
γ = 0.8

Causality of coupled Hénon map

- X is a cause of Y if $\gamma > 0$. $Y_t \not\perp\!\!\!\perp X_{t-1},\ldots,X_{t-p} \mid Y_{t-1},\ldots,Y_{t-p}$
- Y is not a cause of X for all γ. $X_t \perp\!\!\!\perp Y_{t-1},\ldots,Y_{t-p} \mid X_{t-1},\ldots,X_{t-p}$
- Permutation tests for non-causality with $HSNCIC = \left\| \hat{W}^{(N-p+1)}_{\ddot{Y}\mathbf{X_p}|\mathbf{Y_p}} \right\|^2_{HS}$
- N = 100. Significance level α = 5%

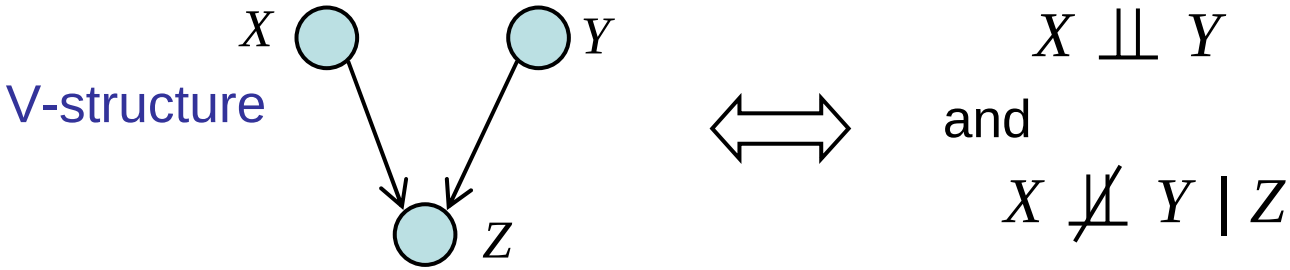
Ratio of accepting Non-Causality (/100 experiments)



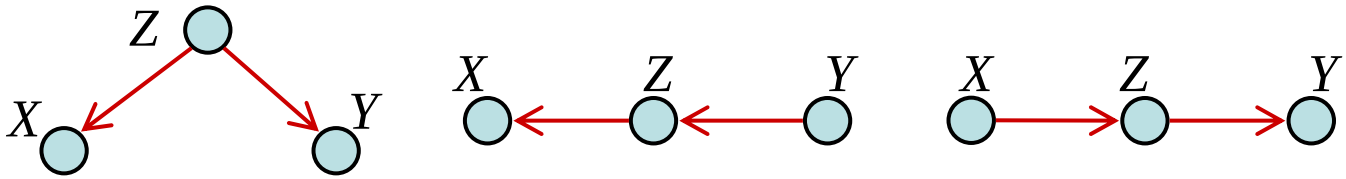


Causal Inference from Non-experimental Data

- Why is it possible?



– All the directions are not distinguishable.



p(x|z)p(y|z)p(z) = p(x|z)p(z|y)p(y) = p(x|z)p(z|y)p(x) = p(x,y,z)

– Constraint-based causal learning
  - Determine the cond. independence of the underlying probability.
  - Markov assumption: data is generated by a DAG.

Causal Leaning

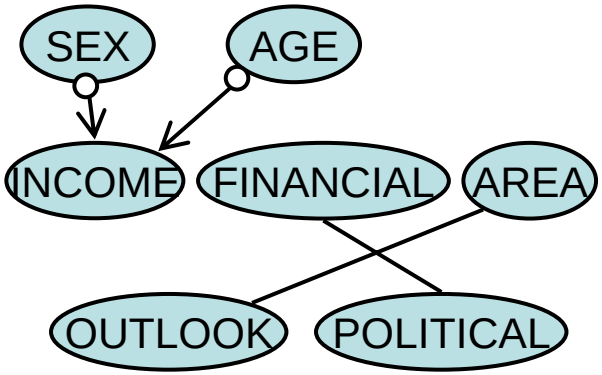
- Inductive causation (IC, Verma&Pearl 90)
  - Basic idea:
    - Make a list of all conditional independence /dependence relations among variables.
    - Make an undirected graph under Markov assumption.
    - Make directions of the edges by finding V-structure.
  - PC algorithm (Peter Sprites & Clark Glymour 1991)
    - Efficient implementation of IC.
    - Gaussian or discrete assumptions for the cond. indep. tests.
- Kernel Causal Learning (KCL, Sun et al. ICML2007)
  - Kernel test for conditional independence for both of continuous and discrete variables.
  - Make directions by voting.

Experiment: Montana Economic Outlook Poll

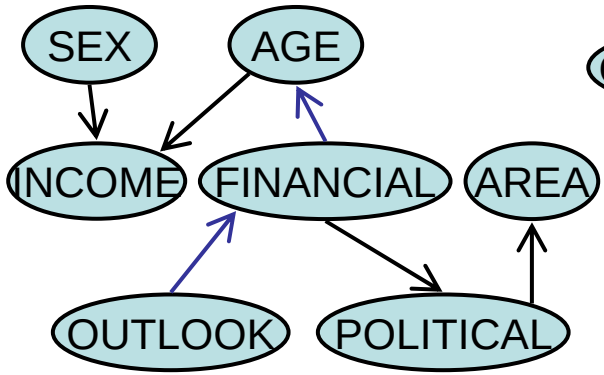
- Data: 7 discrete variables, N = 209

  AGE (3), SEX (2), INCOME (3), POLITICAL (3), AREA (3),

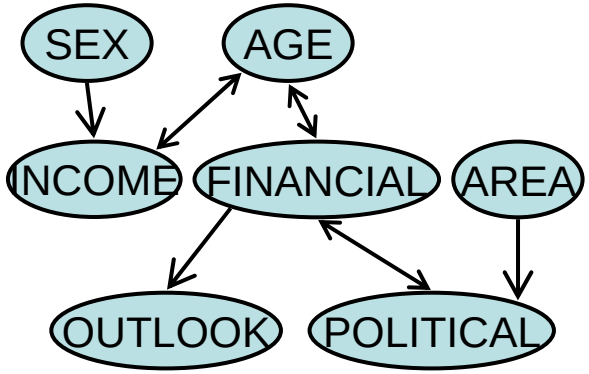
  FINANCIAL status (3, better/same/worse than a year ago), OUTLOOK (2)



FCI



BN-PC



KCL

BN-PC is a constraint-based method using mutual information (Chen et al. 2002)
FCI is the fast IC algorithm which allows hidden variables. (Spirtes et al 1993)

Summary of Part 3

- Conditional independence with kernels
  - Conditional covariance on RKHS characterizes conditional independence.
  - HS-norm for finite sample gives a kernel measure of conditional independence
  - Kernel method gives a unified method of conditional independence test for continuous and discrete variables.

- Causal inference with kernels
  - Kernel conditional independence test are applied to causal inference, such as
    - causality of time series (extension of Granger causality)
    - causal inference from non-experimental data (constrained-based causal learning).

References

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers. (2004).

Cheng, J., R. Greiner, J. Kelly, D. A. Bell, and W. Liu. *Learning Bayesian networks from data: An information-theory based approach*. Artificial Intelligence Journal, 137:43–90, 2002.

Fukumizu, K., A. Gretton, X. Sun., and B. Schölkopf. Kernel Measures of Conditional Dependence. *Advances in NIPS* 20:489-496 (2008)

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

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424-438 (1969).

Spirtes, P. and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* 9:62-72.

Spirtes, P., C. Glymour, and R. Scheines. *Causation, prediction, and search*. Springer-Verlag, New York, NY, 1993.

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)

Verma, T., J. Pearl. Equivalence and synthesis of causal models. *Proc. 6th Conf. Uncertainty in Artificial Intelligence* (*UAI1990*) pp.220-227 (1990)

Pearl, J. *Causality*. Cambridge University Press (2000)

Edwards, D. *Introduction to graphical modelling*. Springer verlag, New York (2000).