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
Conditional independence in causal learning

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

\[ X \perp Y \quad \text{and} \quad X \not\perp Y \mid Z \]

- 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)**

\( \Omega \): set. \( k : \Omega \times \Omega \rightarrow \mathbb{R} \)

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

- Example: Gaussian RBF kernel \( k(x, y) = \exp\left(-\frac{||x - y||^2}{\sigma^2}\right) \)

**Reproducing kernel Hilbert space (RKHS)**

\( k \): p.d. kernel on \( \Omega \).

\( \Longleftrightarrow \exists H \) : reproducing kernel Hilbert space (RKHS)

1) \( k(\cdot, x) \in H \) for all \( x \in \Omega \).
2) \( \text{Span}\{k(\cdot, x) \mid x \in \Omega\} \) is dense in \( H \).
3) \( \langle k(\cdot, x), f \rangle_H = f(x) \) (reproducing property)
Feature map / feature vector

\[ \Phi : \Omega \rightarrow H, \quad x \mapsto k(\cdot, x) \quad \text{i.e.} \quad \Phi(x) = k(\cdot, x) \]

Data: \( X_1, \ldots, X_N \rightarrow \Phi(X_1), \ldots, \Phi(X_N) : \text{functional data} \)

Why RKHS?

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

\[ \langle \Phi(x), \Phi(y) \rangle = k(x, y) \]

\[ f = \sum_{i=1}^{N} a_i \Phi(x_i) = \sum_{i} a_i k(\cdot, x_i), \quad g = \sum_{j=1}^{N} b_j \Phi(x_j) = \sum_{j} b_j k(\cdot, x_j) \]

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

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):

- On RKHS:
  \(X, Y\) : random variables on \(\Omega_X\) and \(\Omega_Y\), resp.
  Prepare RKHS \((H_X, k_X)\) and \((H_Y, k_Y)\) defined on \(\Omega_X\) and \(\Omega_Y\), resp.
  Define random variables on the RKHS \(H_X\) and \(H_Y\) by
  \[ \Phi_X(X) = k_X(\cdot, X) \quad \Phi_Y(Y) = k_Y(\cdot, Y) \]

Define the big (possibly infinite dimensional) covariance matrix \(\Sigma_{YX}\) on the RKHS.
Cross-covariance operator

- Definition

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

\[ \Sigma_{YX} \text{ is an operator from } H_X \text{ to } H_Y \text{ such that} \]

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

for all \( f \in H_X, g \in H_Y \)

- c.f. Euclidean case

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

\[ (b, V_{YX} a) = \text{Cov}[(b, Y), (a, X)] \]
Higher-order moments

Suppose $X$ and $Y$ are $\mathbb{R}$-valued, and $k(x,u)$ admits the expansion

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

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

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

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

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

The operator $\Sigma_{YX}$ 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,

\[ X \text{ and } Y \text{ are independent } \iff \Sigma_{XY} = O \]

\( \Leftrightarrow \) is always true.
\( \iff \) requires some assumption

Gaussian RBF kernels gives the above equivalence.

\[ k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right) \]

- c.f. for Gaussian variables

\[ X \text{ and } Y \text{ are independent } \iff V_{XY} = O \quad \text{i.e. uncorrelated} \]
Kernel Dependence Measure

- Hilbert-Schmidt Independence Criteria (HSIC)

\[ \text{HSIC}(X, Y) = \| \Sigma_{YX} \|_{HS}^2 \]

\[ \text{HSIC} = 0 \quad \iff \quad X \perp Y \]

- Empirical estimator

\[ \text{HSIC}_{\text{emp}}(X, Y) = \| \hat{\Sigma}_{YX}^{(N)} \|_{HS}^2 = \text{Tr}[G_X G_Y] \]

\[ G_X = \left( I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) K_X \left( I_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T \right) : \text{centered Gram matrix} \]

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

- Hilbert-Schmidt norm of an operator

\[ A : H_1 \to H_2 \quad \text{operator on a Hilbert space} \]

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

\[ \| A \|_{HS}^2 = \sum_j \sum_i \langle \psi_j, A \varphi_i \rangle^2 \quad \text{c.f. Frobenius norm of a matrix} \]
Independence Test

Permutation test for independence

- Null hypothesis
  \[ H_0: \quad X \perp Y \]

- 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.
  - 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
  \begin{align*}
  (X^{(1)}_1, \ldots, X^{(1)}_d), \ldots, (X^{(N)}_1, \ldots, X^{(N)}_d) & \quad \quad (Y^{(1)}_1, \ldots, Y^{(1)}_d), \ldots, (Y^{(N)}_1, \ldots, Y^{(N)}_d)
  \end{align*}

- $H_0$: $X$ and $Y$ are independent
- Significance level = 5%
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
  
  \[
  T_N = 2I^\lambda (X,m) = N \frac{2}{\lambda(\lambda + 2)} \sum_{j \in J} \hat{p}_j \left\{ \left( \frac{\hat{p}_j}{\prod_{k=1}^{N} \hat{p}_{j_k}} \right)^{\lambda} - 1 \right\}
  \]

  \( I^0 = \text{MI} \)
  
  \( I^2 = \text{Mean Square Conting.} \)

  \( \hat{p}_j \): frequency in \( A_j \)
  
  \( \hat{p}_{j_k}^{(k)} \): marginal freq. in \( r \)-th interval

- Null distribution under independence

  \[
  T_N \implies \chi^2_{q^N - qN + N - 1} \quad (N \to \infty)
  \]

- 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 \( \Omega_X, \Omega_Y, \Omega_Z \) (resp.).

\((H_X, k_X), (H_Y, k_Y), (H_Z, k_Z) : \) RKHS defined on \( \Omega_X, \Omega_Y, \Omega_Z \) (resp.).

- Conditional cross-covariance operator \( H_X \rightarrow H_Y \)

\[ \Sigma_{Y|X} \equiv \Sigma_{YY} - \Sigma_{YZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} \]

- c.f. For Gaussian variables

Conditional covariance of \( Y \) given \( X \) is equal to

\[ V_{Y|X} \equiv V_{YY} - V_{YZ} V_{ZZ}^{-1} V_{ZX} \]

(conditional covariance matrix)
Conditional independence with kernels

Theorem
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,

\[ \Sigma_{y\tilde{X}|Z} = O \quad \iff \quad X \perp Y \mid Z \]

\[ \Sigma_{y\tilde{X}|Z} = O \quad \iff \quad \Sigma_{\tilde{Y}X|Z} = O \quad \iff \quad \Sigma_{\tilde{Y}\tilde{X}|Z} = O \quad \iff \quad X \perp Y \mid Z \]
Kernel conditional dependence measure

- Hilbert-Schmidt conditional independent criterion

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

- Empirical measure

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

\[
= \text{Tr} \left[ G_X G_Y - 2G_X \left( G_Z + N\varepsilon_N I_N \right)^{-1} G_Z G_Y \right.
\]
\[
+ G_Z \left( G_Z + N\varepsilon_N I_N \right)^{-1} G_X \left( G_Z + N\varepsilon_N I_N \right)^{-1} G_Z G_Y \left. \right]
\]

Consistency

If the regularization coefficient satisfies

\[
\varepsilon_N \to 0 \quad N^{1/3} \varepsilon_N \to \infty,
\]

then

\[
HSCIC_{emp} \to HSCIC \quad (N \to \infty)
\]
Conditional Independence Test

**Permutation test with the kernel measure**

\[
T_N = \left\| \hat{\Sigma}_{XY|Z}^{(N)} \right\|_{HS}^2
\]

- If \( Z \) takes values in a finite set \( \{1, \ldots, L\} \), set

\[
A_\ell = \{ i \mid Z_i = \ell \} \quad (\ell = 1, \ldots, L),
\]

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 \( A_\ell \).
  2. Compute \( T_N^{(b)} \) for the data \( D^{(b)} \).

    → Approximate null distribution under cond. indep. assumption

- Set the threshold by the \((1-\alpha)\)-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.

\[ p(X) = p(X_a)p(X_b)p(X_c | X_a, X_b)p(X_d | X_c) \]

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

This includes the true probability as a special case, but the structure does not express \( a \perp b \).
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 \( S_{ab} \subseteq V \setminus \{a,b\} \) 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 \( a \rightarrow c \leftarrow b \), direct the edges by \( a \rightarrow c \leftarrow b \) if \( c \not\in S_{ab} \)

3. Orient as many of undirected edges as possible on condition that neither new v-structures nor directed cycles are created.
Kernel-based Causal Learning

Limitations of the previous implementations of IC

- Linear / discrete assumptions in Step 1.
  
  e.g. PC-algorithm (Spirtes & Glymour 91) uses partial correlation and $\chi^2$ test.
  
  Difficulty in testing conditional independence for continuous variables.

  $\rightarrow$ kernel method!

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

  $\rightarrow$ 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: 
  \[ \hat{H}_{YX}^{(N)} = HSIC = \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 \]

- Conditional dependence measure: 
  \[ \hat{H}_{YX|Z}^{(N)} = \frac{\left\| \hat{\Sigma}_{YX|Z}^{(N)} \right\|_{HS}^2}{\| C_{ZZ} \|_{HS}^2} \]

where the operator \( C_{ZZ} : H_Z \rightarrow H_Z \) is defined by

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

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

**Theorem**

If \( (X, Y) \perp\!\!\!\perp Z \), 
\[ \left\| \hat{\Sigma}_{YX|Z}^{(N)} \right\|_{HS}^2 = \| C_{ZZ} \|_{HS}^2 \left\| \hat{\Sigma}_{YX}^{(N)} \right\|_{HS}^2 \]
Outline of KCL algorithm: IC algorithm is modified as follows.

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

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

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

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 ("\leftrightarrow" 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 $X \to Z$ and $Y \to Z$ 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 $\to$, or bi-directed $\leftrightarrow$.

- 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: \[ M\left(\begin{array}{c}
\text{(Diagram 1)}
\end{array}\right) > M\left(\begin{array}{c}
\text{(Diagram 2)}
\end{array}\right), M\left(\begin{array}{c}
\text{(Diagram 3)}
\end{array}\right) \]

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 $X \perp Y$ with $\hat{H}^{(N)}_{YX}$
   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 $X \perp Y \mid S_{XY}$ with $\hat{H}^{(N)}_{YX|Z}$ 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) $X_1 \rightarrow X_2 \rightarrow X_3$

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

(B) $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$

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

(C) $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$

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

$X_{n+1} = \text{NoisyOR}(X_1, \ldots, X_n)$

$\iff P(X_{n+1} = 1 \mid X_1, \ldots, X_n) = 0.8 \times (1 - 0.2^{X_1 + \cdots + X_n}) + 0.2$
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

  
  ![Diagram of Truth, Voting, and Result with bi-directional arrows]

<table>
<thead>
<tr>
<th>Truth</th>
<th>Voting</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Truth Diagram]</td>
<td>![Voting Diagram]</td>
<td>![Result Diagram]</td>
</tr>
</tbody>
</table>
Experiments (200 data, 1000 runs)

Truth

Latent

OR gates

Noisy OR gates

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)
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


