From Kernels to Causal Inference

Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Tübingen, Germany

http://www.tuebingen.mpg.de/~bs
Cargo cults—religious practices in pacific tribes around World War II, trying to obtain wealth (the "cargo") by building mock landing strips etc.

…the term cargo cult… is also idiomatically used (in the words of wikipedia) "to mean any group of people who imitate the superficial exterior of a process or system without having any understanding of the underlying substance".

(source: http://philosophyisfashionable.blogspot.com/)
**Statistical Implications of Causality**

Reichenbach’s *Common Cause Principle* links *causality* and *probability*:

(i) if $X$ and $Y$ are statistically dependent, then there is a $Z$ causally influencing both;

(ii) $Z$ screens $X$ and $Y$ from each other (given $Z$, the observables $X$ and $Y$ become independent)
Functional Causal Model *(Pearl et al.)*

- Set of observables $X_1, \ldots, X_n$
- directed acyclic graph $G$ with vertices $X_1, \ldots, X_n$
- Semantics: parents = direct causes
- $X_i = f_i(\text{ParentsOf}_i, \text{Noise}_i)$, with jointly independent $\text{Noise}_1, \ldots, \text{Noise}_n$.
- this model can be shown to satisfy Reichenbach’s principle
- Independence of noises follows from causal sufficiency (by Reichenbach)

- the above entails a joint distribution $P(X_1, \ldots, X_n)$. **Questions:**
  (1) What can we say about it?
  (2) Can we recover $G$ from $P$?
Theorem: the following are equivalent:

- Existence of a functional causal model
- Local Causal Markov condition: $X_j$ statistically independent of non-descendants, given parents (i.e.: every information exchange with its non-descendants involves its parents)
- Global Causal Markov condition: d-separation (characterizes the set of independences implied by local Markov condition)
- Factorization $P(X_1, \ldots, X_n) = \prod_j P(X_j \mid \text{Parents}_j)$ (conditionals as causal mechanisms generating statistical dependence)

(subject to technical conditions)
Causal Inference from Observational Data (Spirtes, Glymour, Scheines, Pearl, …)

• **Question**: given $P(X_1, ..., X_n)$, can we infer $G$?

• **Answer**:
  - Impossible without additional information
  - Possible with *interventions* (subject to philosophical debates re. counterfactuals)
  - Assuming *faithfulness* (and suitable assumptions to make conditional independence testing feasible), can estimate a Markov equivalence class containing $G$ by running conditional independence tests (PC, FCI, …).

Unfortunately, this is useless for the very simplest graphs (2 observables).

• What kind of assumptions can help us in the 2-observable case?
Restricting the Functional Model

- general functional model

\[ X_i = f_i(\text{Parents}_i, \text{Noise}_i) \]

Note: if \( \text{Noise}_i \) can take \( N \) different values, then it could switch randomly between mechanisms \( f_i^1(\text{Parents}_i), \ldots, f_i^N(\text{Parents}_i) \)

- additive noise model

\[ X_i = f_i(\text{Parents}_i) + \text{Noise}_i \]

(\( \text{Noise}_i \) jointly independent)
Causal Inference with Additive Noise, 2-Variable Case

Forward model:

\[ y := f(x) + n, \text{ with } x \perp n \]

Identifiability: when is there a backward model of the same form?


Identifiability Result \((Hoyer, Janzing, Mooij, Peters, Schölkopf, 2008)\)

**Theorem 1** Let the joint probability density of \(x\) and \(y\) be given by

\[
p(x, y) = p_n(y - f(x))p_x(x),
\]

where \(p_n, p_x\) are positive probability densities on \(\mathbb{R}\). If there is a backward model

\[
p(x, y) = p_n(x - g(y))p_y(y),
\]

then, denoting \(\nu := \log p_n\) and \(\xi := \log p_x\) and assuming sufficient differentiability, the triple \((f, p_x, p_n)\) must satisfy the following differential equation for all \(x, y\) with \(\nu''(y - f(x))f'(x) \neq 0\):

\[
\xi''' = \xi'' \left( -\frac{\nu'''}{\nu''} + \frac{f'''}{f'} \right) - 2\nu''f''f' + \nu f''' + \frac{\nu'\nu'''f''f'}{\nu''} - \frac{\nu'(f'')^2}{f'},
\]

where we have skipped the arguments \(y - f(x), x,\) and \(x\) for \(\nu, \xi,\) and \(f\) and their derivatives, respectively. Moreover, if for a fixed pair \((f, \nu)\) there exists \(y \in \mathbb{R}\) such that \(\nu''(y - f(x))f'(x) \neq 0\) for all but a countable set of points \(x \in \mathbb{R}\), the set of all \(p_x\) for which \(p\) has a backward model is contained in a 3-dimensional affine space.

**Corollary 1** Assume that \(\nu''' = \xi''' = 0\) everywhere. If a backward model exists, then \(f\) is linear.
Causal Inference Method

Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

- Compute a function $f$ as non-linear regression of $X$ on $Y$
- Compute the residual
  \[ E := Y - f(X) \]
- check whether $E$ and $X$ are statistically independent (uncorrelated is not enough)
Experiments

Relation between altitude (cause) and average temperature (effect) of places in Germany
Our independence tests detect strong dependence. Hence the method prefers the correct direction:

\[ \text{altitude} \rightarrow \text{temperature} \]
• Generalization to post-nonlinear additive noise models: Zhang & Hyvärinen: *On the Identifiability of the Post-Nonlinear Causal Model*, UAI 2009

• Generalization to graphs with more than two vertices: Peters, Mooij, Janzing, Schölkopf: *Identifiability of Causal Graphs using Functional Models*, UAI 2011

• Generalization to two-vertex-graphs with loops: Mooij, Janzing, Heskes, Schölkopf: *Causal discovery with Cyclic additive noise models*, NIPS 2011
Independence-based Regression \cite{Mooij2009} 

- Problem: many regression methods assume a particular noise distribution; if this is incorrect, the residuals may become dependent.

- Solution: minimize dependence of residuals rather than maximizing likelihood of data in regression objective.

- Use RKHS distance between kernel mean embeddings/Hilbert-Schmidt-norm of cross-covariance operator between two RKHSes as a dependence measure.


Kernel Independence Testing \citep{Gretton2007}

$k$ bounded p.d. kernel; \( p, q \) Borel probability measures

Define the kernel mean map

\[
\mu : p \mapsto \mathbb{E}_{x \sim p}[k(x, .)].
\]

**Theorem:** If \( k \) is universal, \( \mu \) is injective.

**Corollary:** \( x \perp y \iff \Delta := \|\mu(p_{xy}) - \mu(p_x \times p_y)\| = 0. \)

- For \( k((x, y), (x', y')) = k_x(x, x')k_y(x, x') \):
  \( \Delta^2 = \text{HS-norm of cross-covariance operator between the two RKHSes.} \)

- Why does this characterize independence: \( x \perp y \) iff

\[
\sup_{f, g \in \text{RHKS unit balls}} \text{cov}(f(x), g(y)) = 0
\]

\citep{Bach2002}
# The Kernel Trick

Substituting a Mercer kernel for the dot product, we implicitly perform a linear algorithm in a feature space nonlinearly related to input space.

<table>
<thead>
<tr>
<th>in feature space</th>
<th>in input space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptron</td>
<td>Potential function (Aizerman et al., 1964)</td>
</tr>
<tr>
<td>Optimal Hyperplane</td>
<td>SV Machine (Boser et al., 1992)</td>
</tr>
<tr>
<td>PCA</td>
<td>Kernel PCA (Schölkopf et al., 1998)</td>
</tr>
<tr>
<td>any dot product algorithm</td>
<td>Neuro-Fuzzy Quantum Belief KernelProp (...)</td>
</tr>
</tbody>
</table>

B. Schölkopf, NIPS’98
Detection of Confounders \textit{(Janzing et al., 2009)}

Given $P(X, Y)$, infer whether

\begin{itemize}
  \item $X \rightarrow Y$
  \item $Y \rightarrow X$
  \item $X \leftarrow T \rightarrow Y$ for some (possibly) unobserved variable $T$
\end{itemize}

- Confounded additive noise (CAN) models
  \[ X = u(T) + E_X \]
  \[ Y = v(T) + E_Y \]
  with functions $u, v$ and $E_X, E_Y, T$ jointly independent
  Note: includes the case
  \[ Y = f(X) + E \]
  by setting $u = id$ and $E_X = 0$.

- Estimate $(u(T), v(T))$ using dimensionality reduction
- If $E_X$ or $E_Y$ is close to zero, output 'no confounder'
- Identifiability result for small noise

\textit{Janzing, Peters, Mooij, Schölkopf: Identifying latent confounders using additive noise models. UAI 2009}
Inferring deterministic causal relations (Daniusis et al., 2010)

- Idea: If $X \rightarrow Y$ then $f$ and the density $p_X$ are chosen independently ”by nature” (cf. Lemeire & Dirckx, 2007; Janzing & Schölkopf, 2010).

- Hence, peaks of the density $p_X$ do not correlate with the slope of $f$.

- Then, peaks of $p_Y$ necessarily correlate with the slope of $f^{-1}$.

Daniusis, Janzing, Mooij, Zscheischler, Steudel, Zhang, Schölkopf: Inferring deterministic causal relations, UAI 2010
Causal independence implies anticausal dependence

Assume that $f$ is a monotonously increasing bijection of $[0, 1]$. View $p_x$ and $\log f'$ as RVs on the prob. space $[0, 1]$ w. Lebesgue measure.

Postulate (independence of mechanism and input):

$$\text{Cov}(\log f', p_x) = 0$$

Note: this is equivalent to

$$\int_0^1 \log f'(x)p(x)dx = \int_0^1 \log f'(x)dx,$$

since

$$\text{Cov}(\log f', p_x) = E[\log f' \cdot p_x] - E[\log f'] E[p_x] = E[\log f' \cdot p_x] - E[\log f']$$

Proposition:

$$\text{Cov}(\log f^{-1}', p_y) \geq 0$$

with equality iff $f = Id$. 
$u_x, u_y$ uniform densities for $x, y$
$v_x, v_y$ densities for $x, y$ induced by transforming $u_y, u_x$ via $f^{-1}$ and $f$

Equivalent formulations of the postulate:

**Additivity of Entropy:**
$$S(p_y) - S(p_x) = S(v_y) - S(u_x)$$

**Orthogonality (information geometric):**
$$D(p_x \parallel v_x) = D(p_x \parallel u_x) + D(u_x \parallel v_x)$$

which can be rewritten as
$$D(p_y \parallel u_y) = D(p_x \parallel u_x) + D(v_y \parallel u_y)$$

**Interpretation:**
irregularity of $p_y = $ irregularity of $p_x + $ irregularity introduced by $f$
80 Cause-Effect Pairs
## 80 Cause-Effect Pairs – Examples

<table>
<thead>
<tr>
<th>var 1</th>
<th>var 2</th>
<th>dataset</th>
<th>ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>pair0001 Altitude</td>
<td>Temperature</td>
<td>DWD</td>
<td>→</td>
</tr>
<tr>
<td>pair0005 Age (Rings)</td>
<td>Length</td>
<td>Abalone</td>
<td>→</td>
</tr>
<tr>
<td>pair0012 Age</td>
<td>Wage per hour</td>
<td>census income</td>
<td>→</td>
</tr>
<tr>
<td>pair0025 cement</td>
<td>compressive strength</td>
<td>concrete_data</td>
<td>→</td>
</tr>
<tr>
<td>pair0033 daily alcohol consumption</td>
<td>mcv mean corpuscular volume</td>
<td>liver disorders</td>
<td>→</td>
</tr>
<tr>
<td>pair0040 Age</td>
<td>diastolic blood pressure</td>
<td>pima indian</td>
<td>→</td>
</tr>
<tr>
<td>pair0042 day</td>
<td>temperature</td>
<td>B. Janzing</td>
<td>→</td>
</tr>
<tr>
<td>pair0047 #cars/24h</td>
<td>specific days</td>
<td>traffic</td>
<td>←</td>
</tr>
<tr>
<td>pair0064 drinking water access</td>
<td>infant mortality rate</td>
<td>UNdata</td>
<td>→</td>
</tr>
<tr>
<td>pair0068 bytes sent</td>
<td>open http connections</td>
<td>P. Daniusis</td>
<td>←</td>
</tr>
<tr>
<td>pair0069 inside room temperature</td>
<td>outside temperature</td>
<td>J. M. Mooij</td>
<td>←</td>
</tr>
<tr>
<td>pair0070 parameter</td>
<td>sex</td>
<td>Bülthoff</td>
<td>→</td>
</tr>
<tr>
<td>pair0072 sunspot area</td>
<td>global mean temperature</td>
<td>sunspot data</td>
<td>→</td>
</tr>
<tr>
<td>pair0074 GNI per capita</td>
<td>life expectancy at birth</td>
<td>UNdata</td>
<td>→</td>
</tr>
<tr>
<td>pair0078 PPFD (Photosynth. Photon Flux)</td>
<td>NEP (Net Ecosystem Productivity)</td>
<td>Moffat A. M.</td>
<td>→</td>
</tr>
</tbody>
</table>
IGCI: Deterministic Method

LINGAM: Shimizu et al., 2006

AN: Additive Noise Model (nonlinear)

PNL: AN with post-nonlinearity

GPI: Mooij et al., 2010

Accuracy (%) vs Decision rate (%)
Causal Learning and Anticausal Learning

• example 1: predict gene from mRNA sequence

• example 2: predict class membership from handwritten digit
Covariate Shift and Semi-Supervised Learning

Assumption: $p(cause)$ and mechanism $p(effect|cause)$ are independently chosen by nature;

Goal: learn $X \mapsto Y$, i.e., estimate (properties of) $p(Y|X)$

- covariate shift (i.e., $p(X)$ changes): mechanism $p(Y|X)$ is unaffected by assumption
- semi-supervised learning: impossible, since $p(X)$ contains no information about $p(Y|X)$
- transfer learning ($N_X, N_Y$ change, $\varphi$ not): could be done by additive noise model with conditionally independent noise

- covariate shift (i.e., $p(X)$ changes): need to decide if change is due to mechanism $p(X|Y)$ or cause distribution $p(Y)$ (nontrivial)
- semi-supervised learning: possible, since $p(X)$ contains information about $p(Y|X)$ — e.g., cluster assumption.

Schölkopf, Janzing, Peters, Zhang (2011)
Lens error correction \cite{schuler2011}
Inverting a nontrivial convolution model (Hirsch et al, 2011)
Inverting a nontrivial convolution model (Hirsch et al, 2011)
Shift-Invariant Kernel Mean Maps \( \mu : \mathcal{P} \mapsto \mathbb{E}_{x \sim \mathcal{P}}[k(x - \cdot)] \)

**Optics:** imaging through an aperture

- Example: \( p \) source of incoherent light, \( I \) indicator function of an aperture. In *Fraunhofer diffraction*, the intensity image is \( \propto p \ast \hat{I}^2 \).
- Set \( \hat{k} = I \ast I \geq 0 \) (this is a pd kernel by Bochner), then this equals \( \mu(p) \).

- this imaging process is *not* invertible for the class of all light sources, but it is if we restrict the class.
Causal Inference for Individual Objects *(Janzing & Schölkopf, 2010)*

Similarities between single objects also indicate causal relations:

However, if similarities are too simple there need not be a common cause:
Causal Markov Conditions

• Recall the (Local) Causal Markov condition:
  An observable is statistically independent of its non-descendants, given parents

• Reformulation:
  Given all direct causes of an observable, its non-effects provide no additional statistical information on it
Causal Markov Conditions

- **Generalization:**
  Given all direct causes of an observable, its non-effects provide no additional *statistical* information on it

- **Algorithmic Causal Markov Condition:**
  Given all direct causes of an object, its non-effects provide no additional *algorithmic* information on it
Kolmogorov complexity

(Kolmogorov 1965, Chaitin 1966, Solmonoff 1964)

of a binary string $x$

- $K(x) :=$ length of the shortest program with output $x$ (on a Turing machine)
- interpretation: number of bits required to describe the rule that generates $x$
- equality "$=\$" is always understood up to string-independent additive constants

- $K(x)$ is uncomputable
- probability-free definition of information content
Conditional Kolmogorov complexity

- $K(y \mid x)$: length of the shortest program that generates $y$ from the shortest description of the input $x$.
- number of bits required for describing $y$ if the shortest description of $x$ is given
- note: $x$ can be generated from its shortest description but not vice versa because there is no algorithmic way to find the shortest compression
Algorithmic mutual information \((Chaitin, Gacs)\)

Information of \(x\) about \(y\)

- \(I(x : y) := K(x) + K(y) - K(x, y)\)
  \[= K(x) - K(x \mid y) = K(y) - K(y \mid x)\]

- Interpretation: number of bits saved when compressing \(x, y\) jointly rather than independently

- Algorithmic independence \(x \perp y \iff I(x : y) = 0\)
Conditional algorithmic mutual information

Information that $x$ has on $y$ (and vice versa) when $z$ is given

- $I(x : y \mid z) := K(x \mid z) + K(y \mid z) - K(x, y \mid z)$

- Analogy to statistical mutual information:
  \[
  I(X : Y \mid Z) = S(X \mid Z) + S(Y \mid Z) - S(X, Y \mid Z)
  \]

- Conditional algor. independence $x \perp y \mid z :\iff I(x : y \mid z) = 0$
Algorithmic mutual information: example

\[ I(\star : \star) = K(\star) \]
Postulate: Local Algorithmic Markov Condition

Let $x_1, \ldots, x_n$ be observations (formalized as strings). Given its direct causes $pa_j$, every $x_j$ is conditionally algorithmically independent of its non-effects $nd_j$

$$x_j \perp nd_j \mid pa_j$$
Equivalence of Algorithmic Markov Conditions

For $n$ strings $x_1, \ldots, x_n$ the following conditions are equivalent:

- **Local Markov condition**
  \[ I(x_j : nd_j | pa_j) = 0 \]

- **Global Markov condition:**
  If $R$ d-separates $S$ and $T$ then $I(S : T | R) = 0$

- **Recursion formula for joint complexity**
  \[ K(x_1, \ldots, x_n) = \sum_{j=1}^{n} K(x_j | pa_j) \]

*Janzing & Schölkopf, IEEE Trans. Information Theory, 2010*
Algorithmic model of causality

- for every node $x_j$ there exists a program $n_j$ that computes $x_j$ from its parents $pa_j$

- all $n_j$ are jointly independent

- the program $n_j$ represents the causal mechanism that generates the effect from its causes

- $n_j$ are the analog of the unobserved noise terms in the statistical functional model

**Theorem:** this model implies the algorithmic Markov condition

*(Janzing & Schölkopf IEEE TIT 2010)*
Applications

Steudel, Janzing & Schölkopf, Causal Markov condition for submodular information measures, COLT, 2010

Janzing, Steudel, Justifying additive-noise based causal discovery via algorithmic information theory, Open Systems and Information Dynamics, 2010
Causality

Dominik Janzing, Jonas Peters, Kun Zhang, Joris Mooij, Olivier Stegle, Eleni Sgouritsa, Jakob Zscheischler

Image Deconvolution

Michael Hirsch, Stefan Harmeling, Christian Schuler

Kernel Means

Arthur Gretton, Kenji Fukumizu, Alex Smola, Bharath Sriperumbudur

Plug: if you are looking for a PhD position in astronomical image processing or causal inference, talk to me