Inferring causal directions by evaluating the complexity of conditional distributions

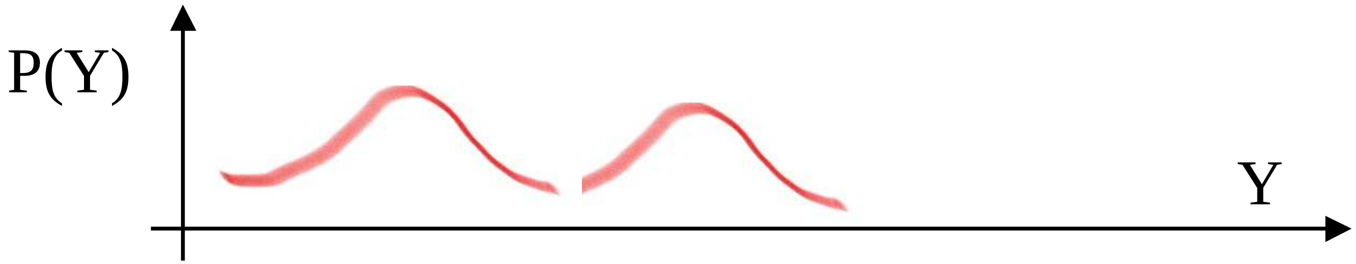
Xiaohai Sun[1], Dominik Janzing[1,2], and Bernhard Schölkopf [1]

1) MPI for Biological Cybernetics, Tübingen, Germany

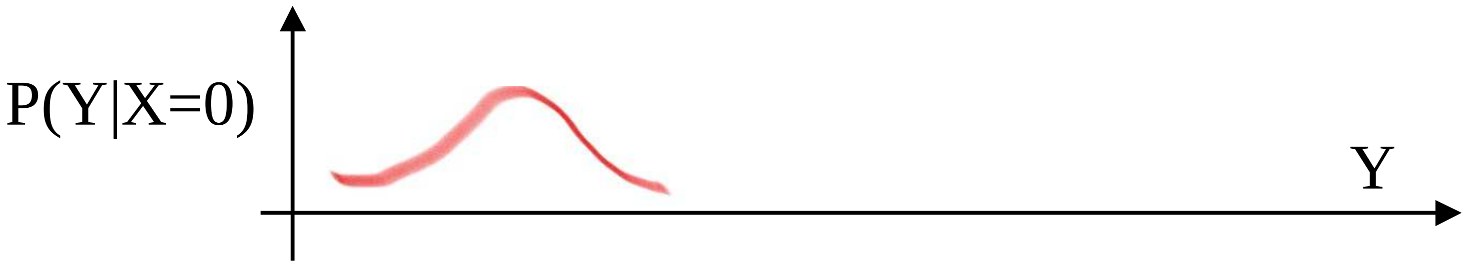
2) Universität Karlsruhe (TH), Germany

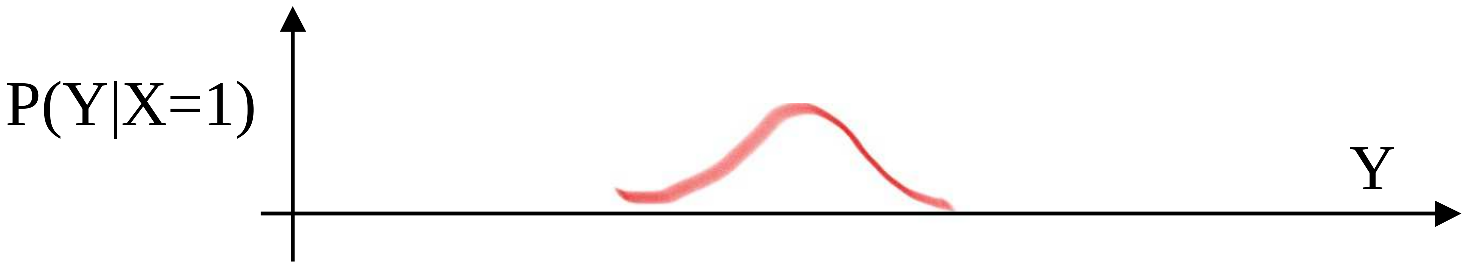
A naive approach to causal reasoning

Given the following distribution for real-valued Y

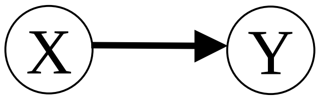


and a binary variable X such that





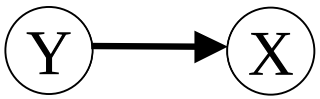
Strong evidence for a certain causal direction…



Plausible:

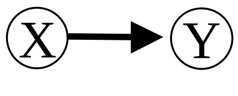
1) explains bimodality of P(Y)

2) X shifts distribution of Y : linear effect



Implausible:

1) bimodality of P(Y) remains unexplained
2) unlikely that conditioning on effect strictly seperates modes

Try to formalize why 

is more plausible

Markov kernels of a causal hypothesis

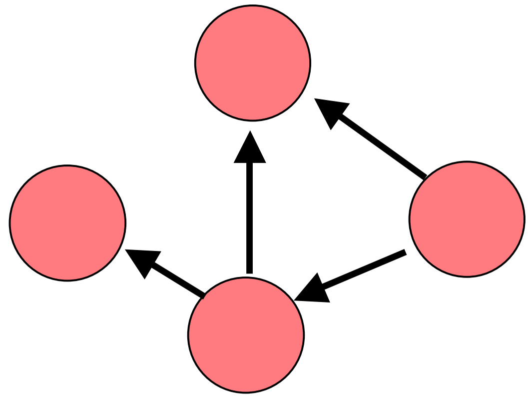
Given n random variables $X_1,...,X_n$ with joint measure P

causal hypothesis:

DAG G such that P is Markovian relative to G

$$\Rightarrow \; P(x_1,...,x_n) = \Pi_{j=1...n} \, P(x_j|pa_j)$$

causal parents of X_j



$P(x_j|pa_j)$: Markov kernels of P w.r.t. G

Principle of “most plausible” Markov kernels

Prefer causal hypotheses which lead to
“smooth” and “simple” Markov kernels

Intuition:

describes the
“physics” of the
causal mechanism

P(effect,cause)=P(effect|cause) P(cause)

leads typically to smoother terms than factorization

P(effect,cause)=P(cause|effect) P(effect)

only an abstract
mathematical
expression

How to get well-defined inference rules from these vague ideas...

1) Shimizu, Hyärinen, Kano & Hoyer 2005:
   Prefer linear effects with additive noise
   (ICA for identifying most plausible causal order)

2) Sun, Janzing & Schölkopf 2006: Prefer Markov kernels that maximize conditional entropy of effects, given their causes s.t. the observed first and second moments

3) Sun, Janzing & Schölkopf 2006: Evaluate complexity of Markov kernels using a Hilbert space norm

Defining complexity of conditional probabilities by semi-norms

1) Write $P(y|x) = \exp\left(f(y,x) - \ln z(x) \right)$ with appropriate f

2) Define complexity of $P_{Y|X}$ by $C(P_{Y|X}) := \| f \|^2$,
where $\|.\|$ is some seminorm on a Hilbert space H_{YX}
Idea: small seminorm for *smooth* f

($P_{Y|X}$ is simple if it maximizes conditional entropy
s.t. smooth constraints)

Note: $\log P_{Y|X}$ need not to be smooth,
partition function $z(x)$ may be arbitrarily complex

Properties of C

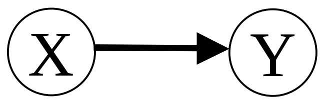
If semi-norm satisfies $\|a \otimes 1\| = \|a\| = \|1 \otimes a\|$ we have:

1) Additivity: $C(P \otimes Q) = C(P) + C(Q)$

2) Consistency: If X,Y independent then $C(P_{Y|X}) = C(P_Y)$

3) Asymmetry: $C(P_{XY}) \neq C(P_{Y|X}) + C(P_X) \neq C(P_{X|Y}) + C(P_Y)$

Consider $C(P_{Y|X}) + C(P_X)$ as complexity of the causal model



$\Rightarrow$ Prefer causal direction with smaller complexity

Construct semi-norms by penalized subspaces

Split $H = H_1 \oplus H_2$, $f = f_1 \oplus f_2$, define seminorm $\| f \| := \| f_2 \|$

Idea: Let H_1 contain extremely simple functions

(e.g. polynomials of degree 2 since they
generate gaussians with linear interaction terms:
$P(y|x) = \exp(- ay^2 - bxy - \ln z(x))$)

Kernelizing the norms (RKHS)

H_1 := span of functions $k_1((x,y),(.,.))$ with pos. semidef. k_1
H_2 := span of functions $k_2((x,y),(.,.))$

Our (preliminary) choice:

$k_2((x,y),(x',y')) := \exp(-\|(x,y)-(x',y')\|^2/\sigma^2)$
$k_1((x,y),(x',y')) := (a\langle x,x'\rangle + b)(c\langle y,y'\rangle + d)^2$

Gaussian term k_2 provides flexibility,
polynomial term k_1 allows for decay of probabilities at infinity and supports linear interactions and Gauss distributions

Mercer *kernels* k_1 k_2 have nothing to do with Markov *kernels* !

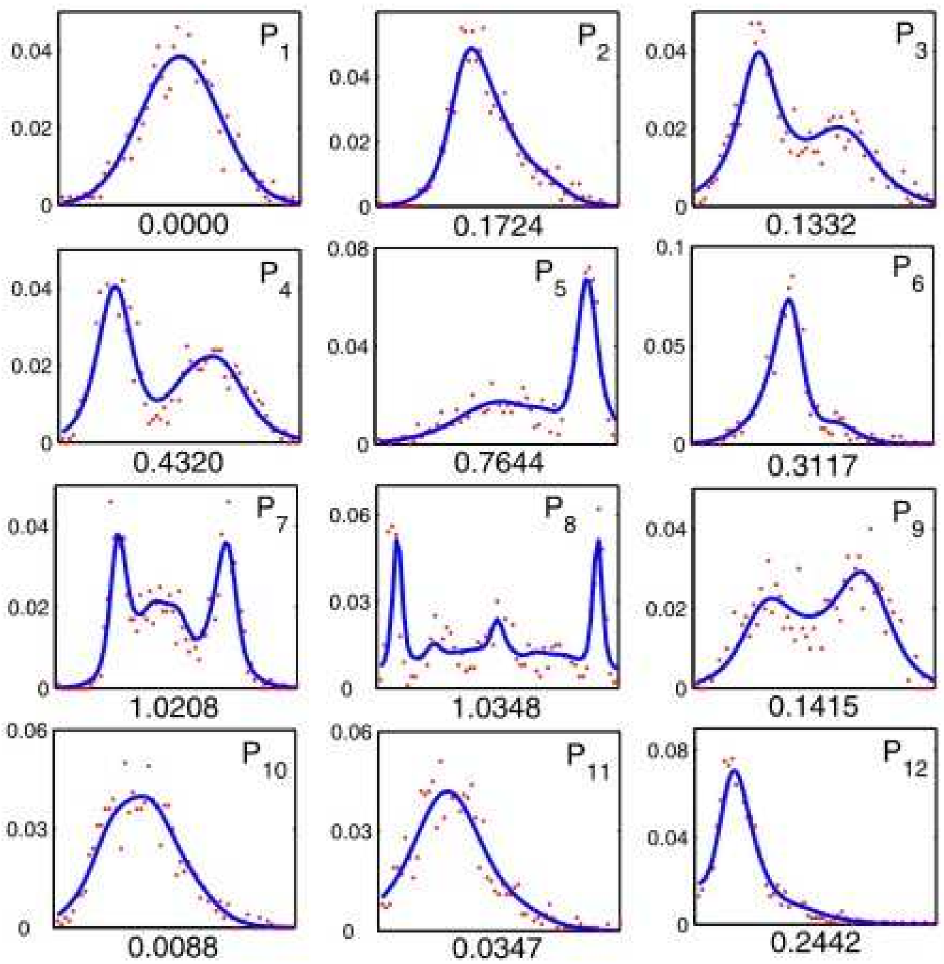
Model fit for finite dataset (regularized ML)

$P(y|x) \sim \exp(f(x,y))$ with f solution of

$$\max_g \left\{ \Sigma_i \left(g(x_i,y_i) - \Sigma_x \exp(g(x_i,y)) - \varepsilon\|g\| \right\} \right.$$

Bayesian interpretation:
prior proportional to $\exp(- \varepsilon \|g\|)$

Experiments with random data

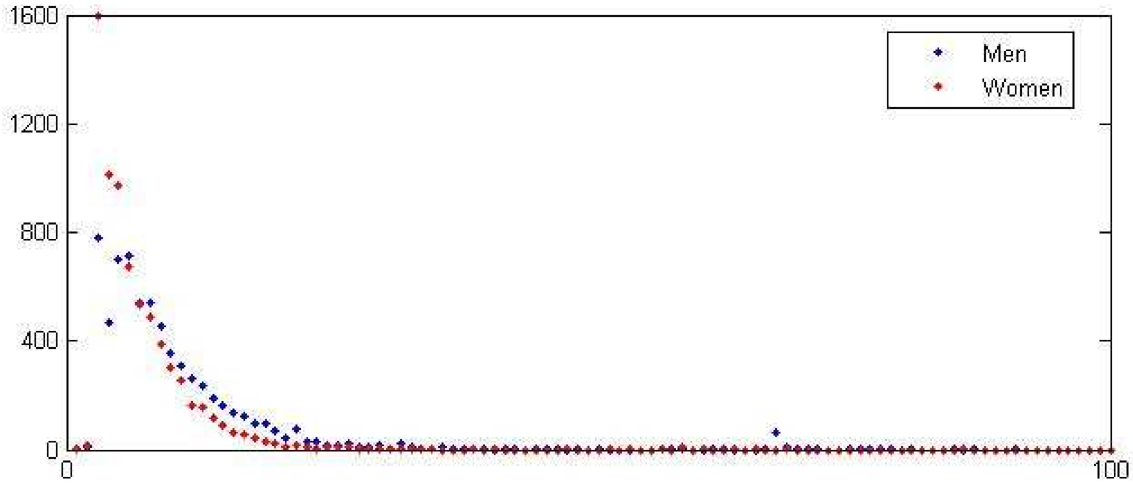


Mixtures of 1 - 5 Gauss or Gamma distributions:

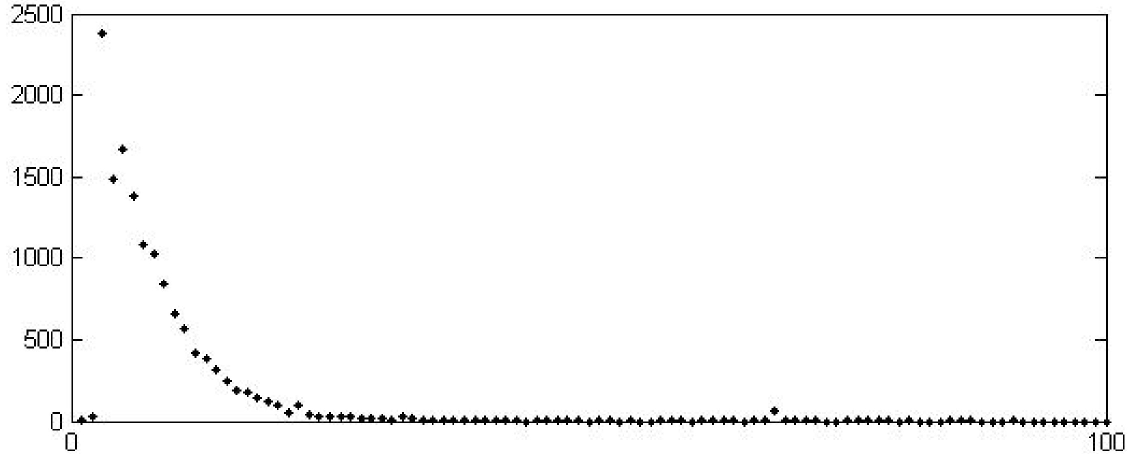
Larger complexity values than pure ensembles

(even when mixture was not obvious!)

Example with real-world data:
Income of 112 000 persons (USA, Pacific Division)



Income of
men / women



Distribution of
Income over
total population

Evaluation of Complexities:

$C(P_{Income}) = 27.57$

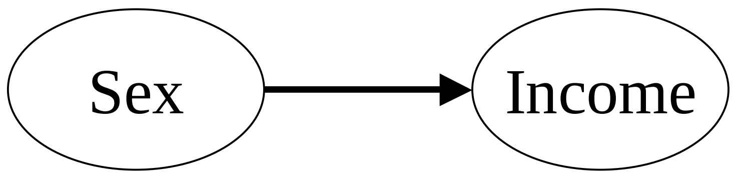
$+\ C(P_{Sex|Income}) = 0.0255$

$C(P_{Income|Sex}) = 20.29$

$C(P_{Sex}) = 0$

$$C(P_{Income}) + C(P_{Sex|Income}) \quad > \quad C(P_{Sex}) + C(P_{Income|Sex})$$

$\Rightarrow$ Prefer causal hypothesis



Example with real-world data: Age and marital status

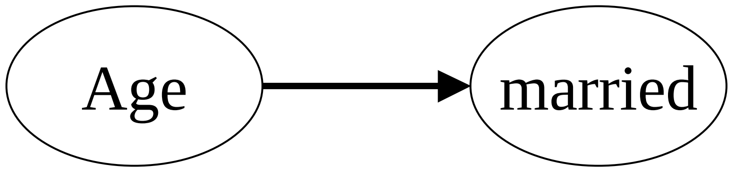
Variables: Age: natural number
Marital Status: binary: never married (yes/no)

$C(P_{Age}) = 0.0164$ $C(P_{Age \mid married}) = 0.1145$

$+\ C(P_{married \mid Age}) = 0.0082$ $C(P_{married}) = 0$

$$C(P_{age}) + C(P_{married \mid Age}) < C(P_{married}) + C(P_{Age \mid married})$$

$\Rightarrow$ Prefer causal hypothesis



Partially negative results:

Handwritten numerals (0,1) as cause
and some Karhunen-Loeve coefficients as effects

- Correct results when coefficient was strongly correlated to the class label

- Balanced results in case of weak correlations

How we would like to use our approach…

…in constraint-based approaches:
use plausibility of Markov kernels to select among Markov-equivalent graphs
(our optimization is not feasible without pre-selection!)

…in Bayesian approaches:
complexity measure provides priors for Markov kernels
(our priors take into account the structure of the value set!)

Conclusions

1) Every causal inference method could benefit from a good complexity / plausibility measure for Markov kernels (providing *additional* information)

2) We don´t claim to have the right one…

   …however:

RKHS-norms are a *flexible* way of constructing complexity measures having nice properties

Thanks for your attention !