ICML' 14 workshop on
causal modeling & machine learning

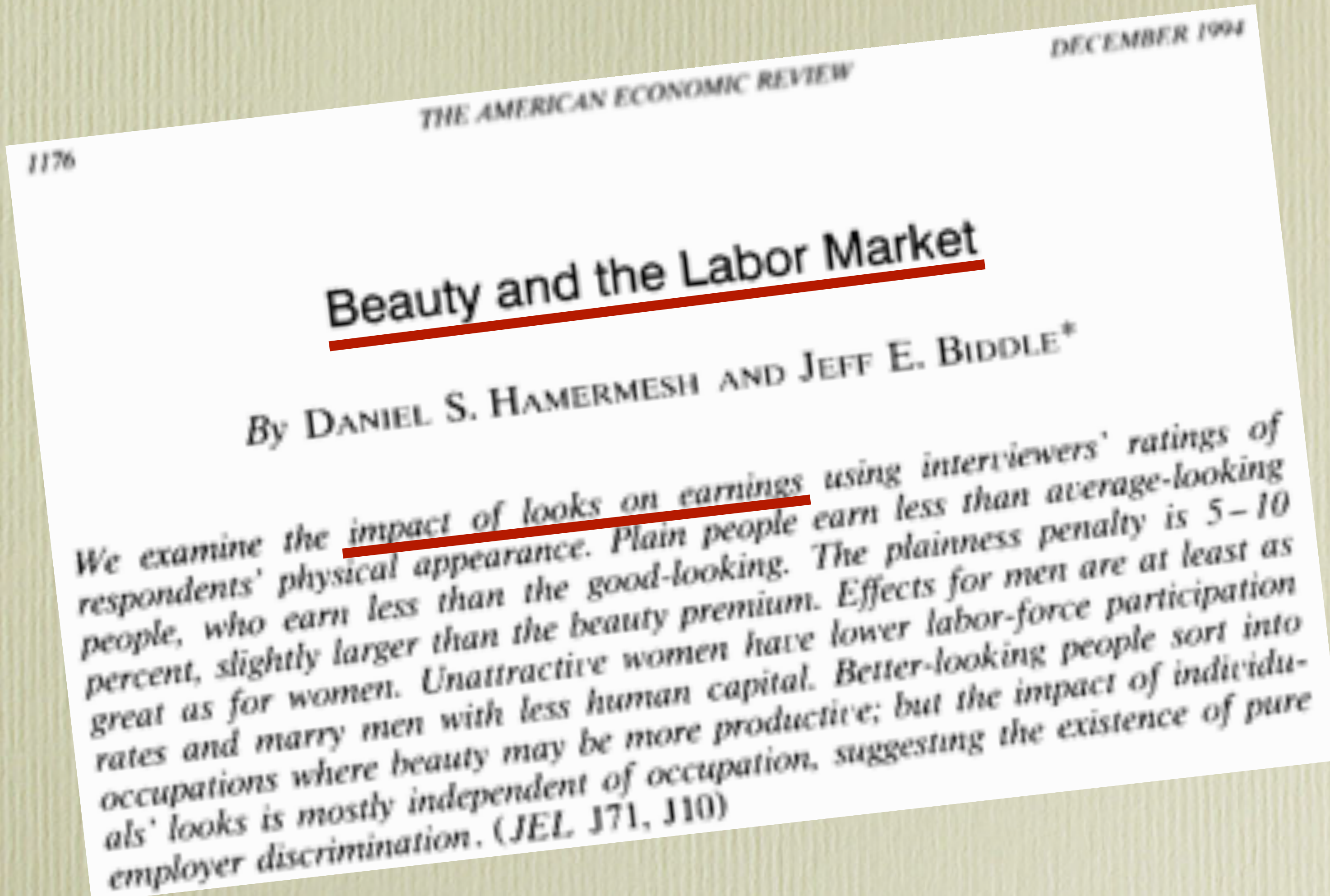
Learning causal knowledge and learning based on causal knowledge

Kun Zhang

Max-Planck Institute for Intelligent Systems
Tübingen, Germany

MAX-PLANCK-GESELLSCHAFT

Causal vs. associational knowledge



1176 THE AMERICAN ECONOMIC REVIEW DECEMBER 1994

Beauty and the Labor Market

By Daniel S. Hamermesh and Jeff E. Biddle*

We examine the impact of looks on earnings using interviewers' ratings of respondents' physical appearance. Plain people earn less than average-looking people, who earn less than the good-looking. The plainness penalty is 5–10 percent, slightly larger than the beauty premium. Effects for men are at least as great as for women. Unattractive women have lower labor-force participation rates and marry men with less human capital. Better-looking people sort into occupations where beauty may be more productive; but the impact of individuals' looks is mostly independent of occupation, suggesting the existence of pure employer discrimination. (JEL J71, J10)

Causal vs. associational knowledge


1176 THE AMERICAN ECONOMIC REVIEW DECEMBER 1994

Beauty and the Labor Market

The Telegraph

Home News World Sport Finance Comment Culture Travel Life Women Fashion

USA | Asia | China | Europe | Middle East | Australasia | Africa | South America | Central Asia

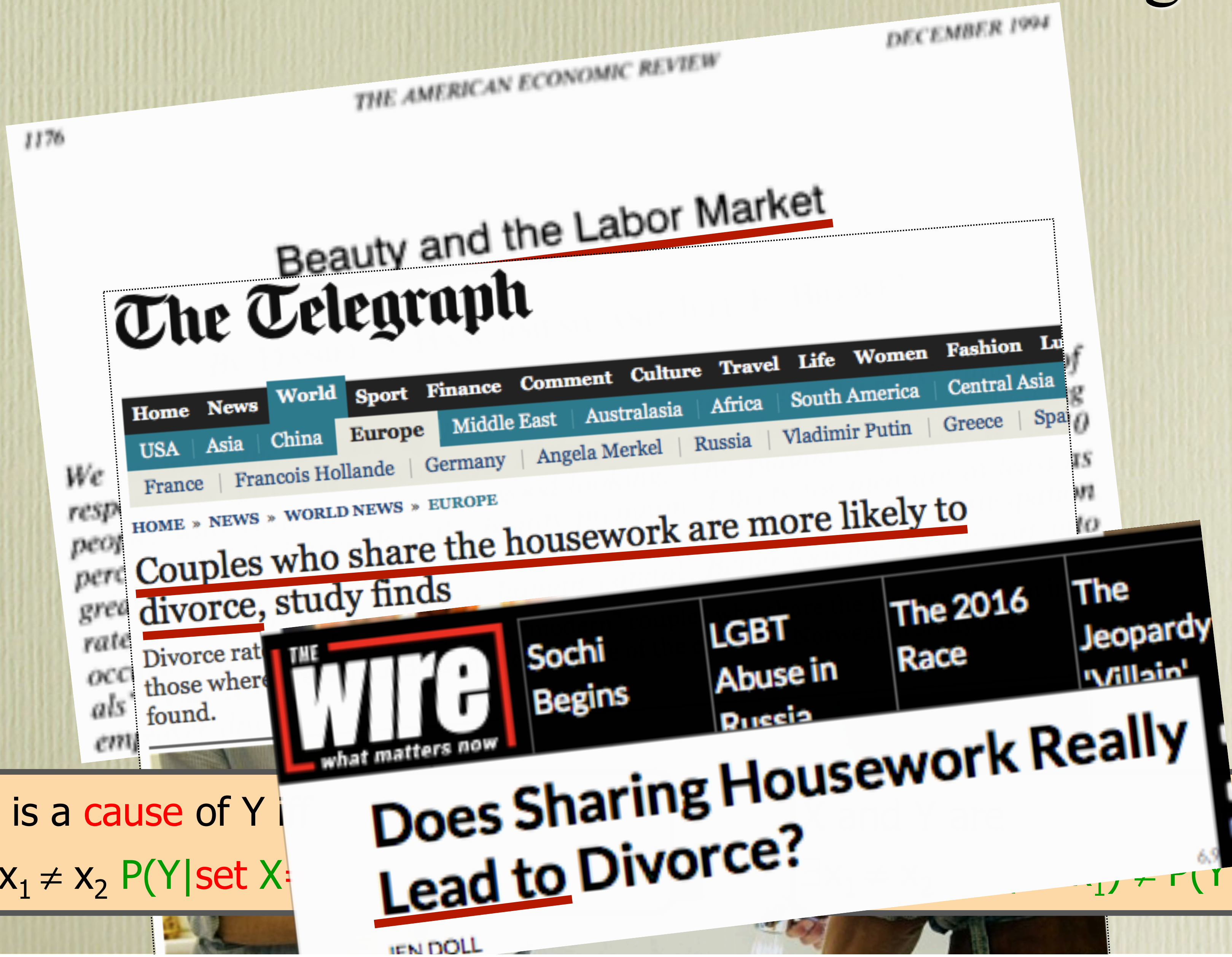
France | Francois Hollande | Germany | Angela Merkel | Russia | Vladimir Putin | Greece

HOME » NEWS » WORLD NEWS » EUROPE

Couples who share the housework are more likely to divorce, study finds

Divorce rates are far higher among "modern" couples who share the housework than in those where the woman does the lion's share of the chores, a Norwegian study has found.

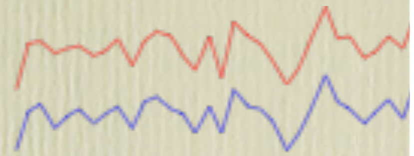
Causal vs. associational knowledge



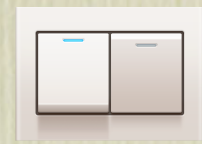
X is a cause of Y i … f

$\exists x_1 \neq x_2$ P(Y|set X= … $) \neq$ P(Y|X=x_2)

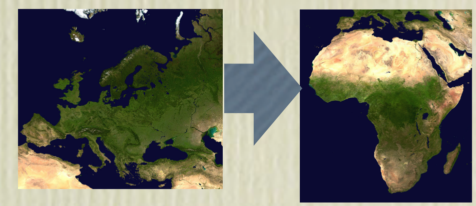
Use associational or causal knowledge?



- Make passive predictions in stationary environments ?



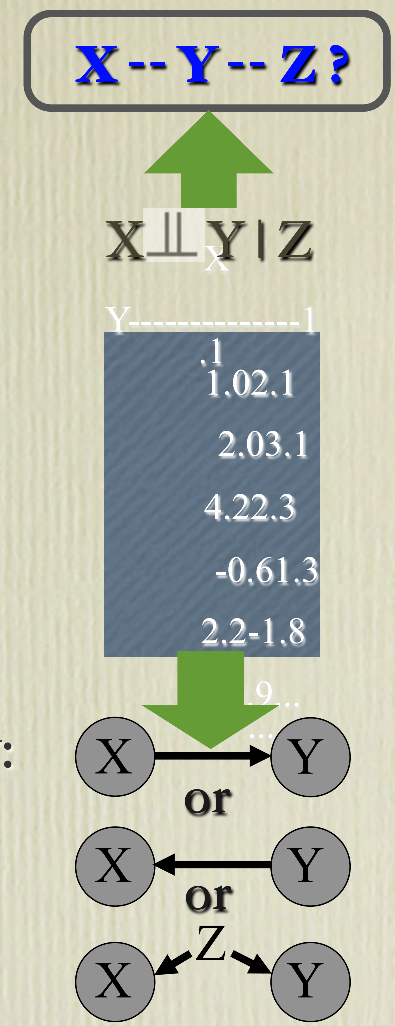
- For manipulation and control (e.g., make advertisements) ?



- Make predictions in non-stationary environments ?
- Associational information easy to calculate
- Causal knowledge usually difficult to find
  - interventions might be expensive or even impossible
  - causal discovery: find causal knowledge from passively observational data

Outline

- Constraint-based causal discovery
  - key issue: conditional independence test
- Functional causal model based
  - key issue: identifiability & applicability
  - two types of independence lead to identifiability: cause $\perp\!\!\!\perp$ noise; P(cause) $\perp\!\!\!\perp$ transformation
- Implications of causality in machine leaning (semi-supervised learning and domain adaptation)



Causal structure vs. statistical independence
(Spirtes, Pearl, et al.)

Causal Markov condition: each variable is ind. of its non-descendants (non-effects) conditional on its parents (direct causes)

causal structure (causal graph)

$Y \rightarrow X \rightarrow Z$

$Y - X - Z$?

Statistical independence(s)

$Y \perp\!\!\!\perp Z \mid X$

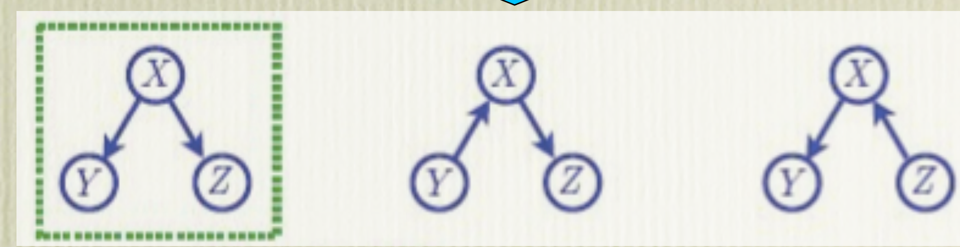
Faithfulness: all observed (conditional) independencies are entailed by Markov condition in the causal graph

Recall: $Y \perp\!\!\!\perp Z \Leftrightarrow P(Y|Z)=P(Y)$; $Y \perp\!\!\!\perp Z|X \Leftrightarrow P(Y|Z,X)=P(Y|X)$

Constraint-based causal discovery

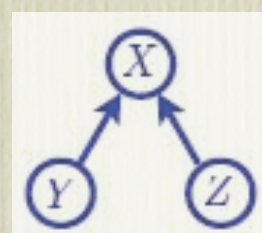
- uses (conditional) independence <u>constraints</u> to find candidate causal structures
- example: PC algorithm (Spirtes & Glymour, 1991)
- Markov equivalence class

$$Y \perp\!\!\!\perp Z \mid X$$

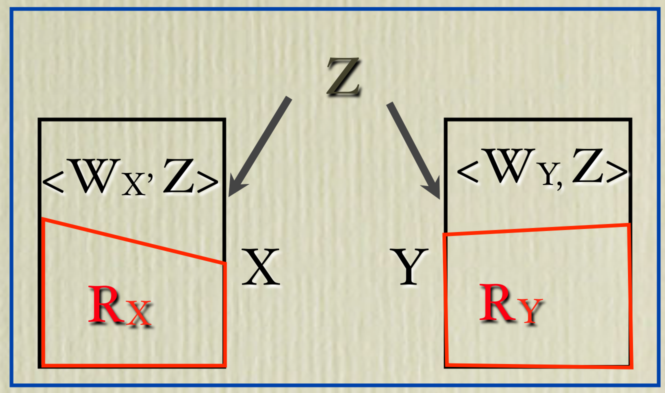


- pattern Y—X—Z
  - same adjacencies
  - → if all agree on orientation; — if disagree
- might be unique: v-structure

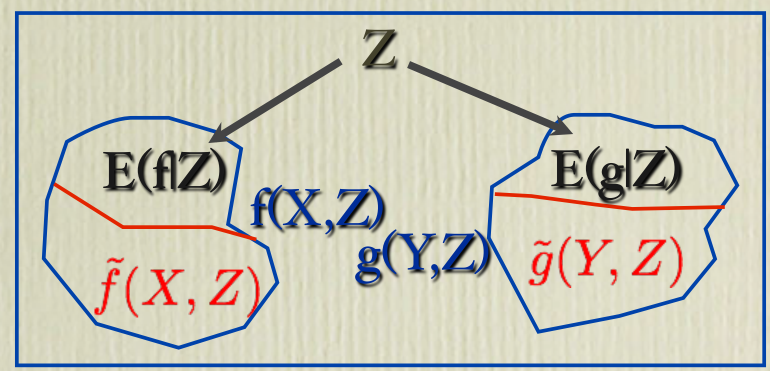
$$Y \perp\!\!\!\perp Z$$



Characterization of CI: from linear-Gaussian case to general case



- Linear Gaussian case: partial correlation $\rho_{XY\cdot Z} = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z$
- General case (Daudin, 1980): $X \perp\!\!\!\perp Y|Z \Longleftrightarrow \mathrm{E}(\tilde{f}\tilde{g}) = 0, \ \forall f \in L^2_{XZ}, \ g \in L^2_{YZ}$



- With kernels (Fukumizu et al. 2008): under some "richness" assumption on RKHS (with characteristic kernels), use RKHS $\mathcal{H}$ instead of L^2

$$\mathrm{E}(\tilde{f}) = 0, \ \mathrm{E}(\tilde{g}) = 0:$$
$$\tilde{f}(X,Z) = f(X,Z) - \mathbb{E}(f|Z), \quad \text{for } f \in L^2_{XZ}$$
$$\tilde{g}(Y,Z) = g(Y,Z) - \mathbb{E}(g|Z), \quad \text{for } g \in L^2_{YZ}$$

Kernel-based CIT (KCI-test, Zhang et al., 2011): framework

fundamental theorem on the asymp. dstr of $\mathrm{Tr}(K_X K_Y)/n$ if $f(X)$ & $g(Y)$ are uncorrelated $\forall f(X) \in \mathcal{H}_X$ & $\forall g(Y) \in \mathcal{H}_Y$

characterization of CI: CI $\Leftrightarrow$ uncorrelatedness of functions in certain spaces

$X \leftarrow (X,Z)$, $Y \leftarrow (Y,Z)$, $\mathcal{H}_X$, $\mathcal{H}_Y$: residual spaces

unconditional independence testing as a direct application

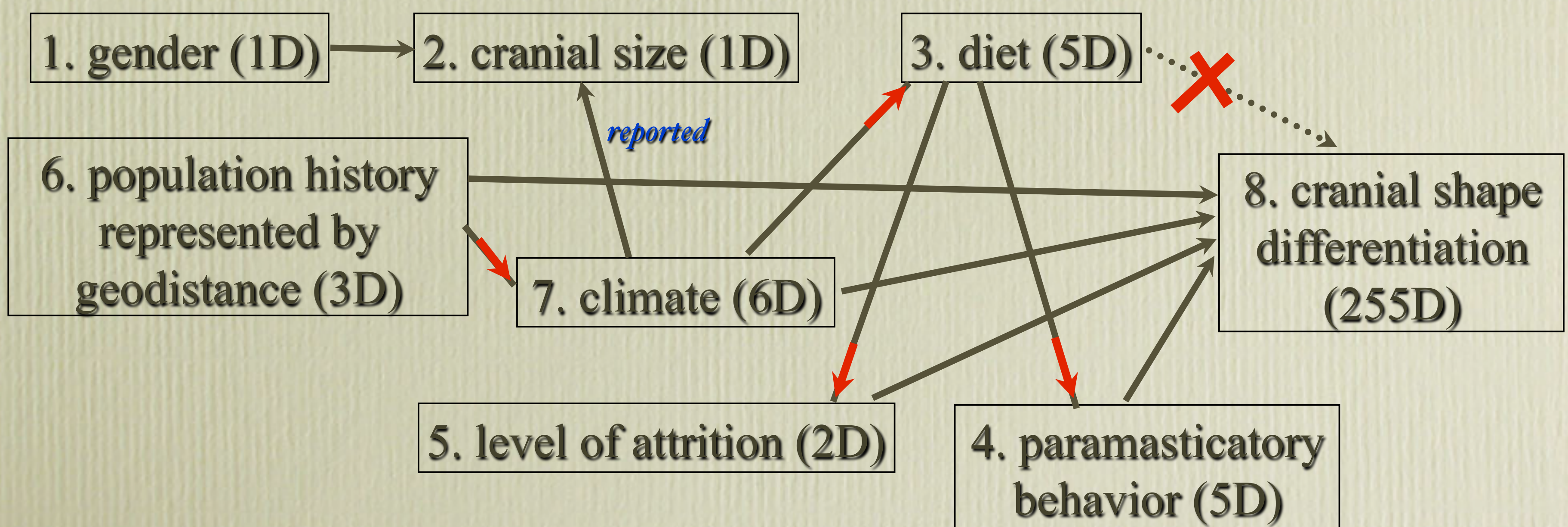
KCI-test

1. nice characterization of CI with kernels;
2. the first time the null distribution with kernels has been derived;
3. good applicability !

Causal analysis of archeology data

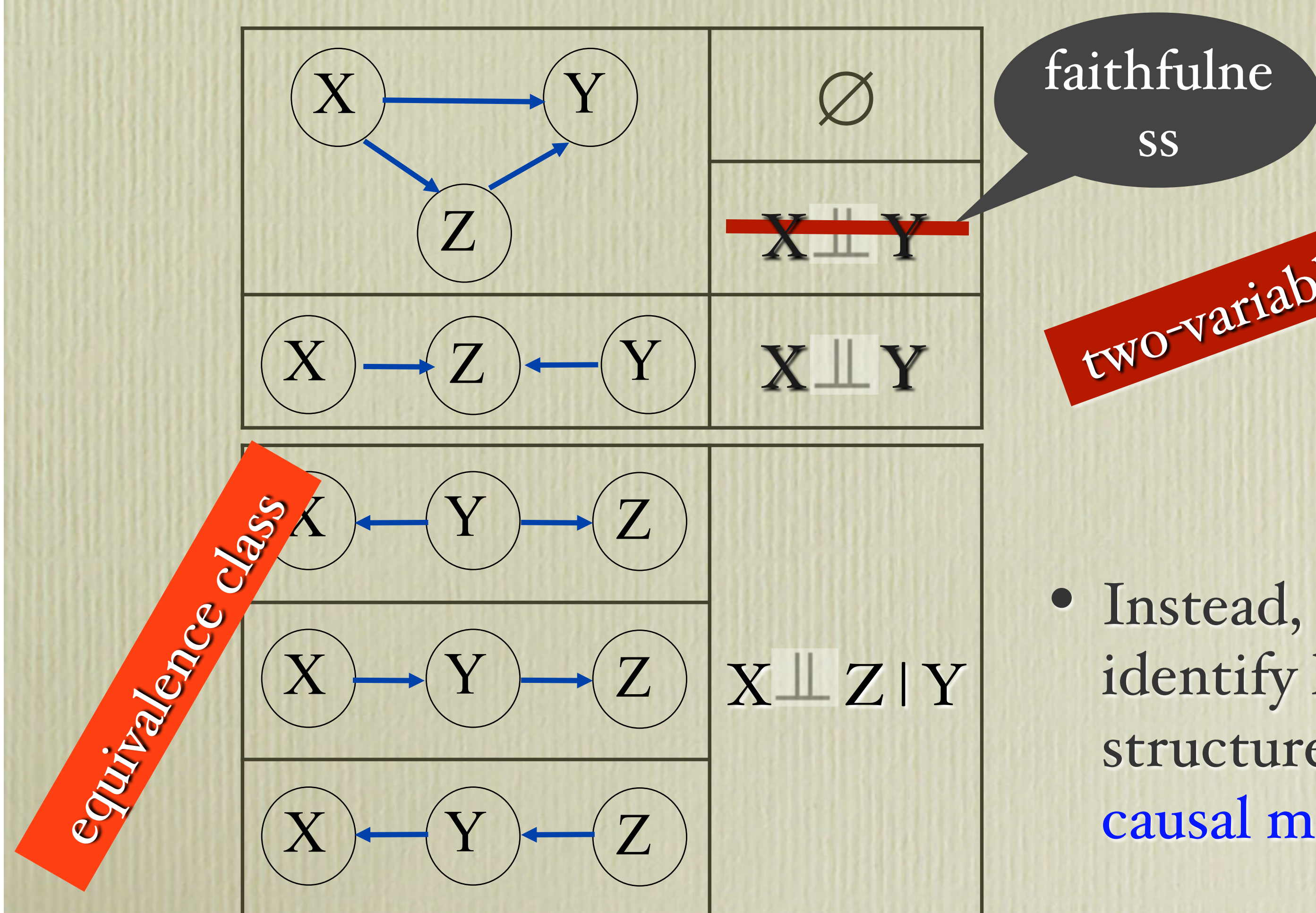
Thanks to collaborator Marlijn Noback

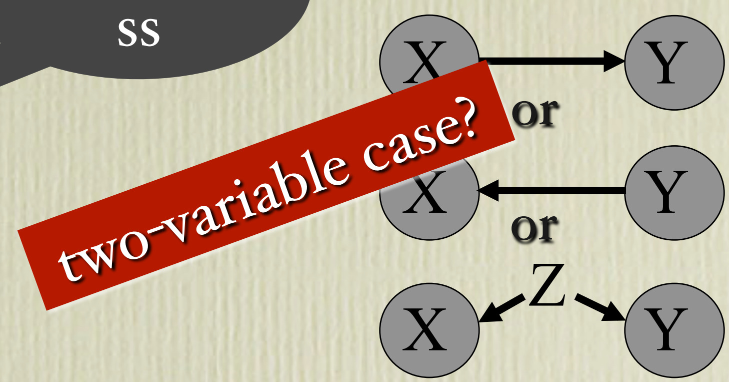
- 8 variables of 250 skeletons collected from different locations
- different dimensions (from 1 to 255) with nonlinear dependence
- **PC + KCI-test** seems to be a good choice
- Some have been reported; some are new; all seem reasonable



Constraint-based method: An inverse problem

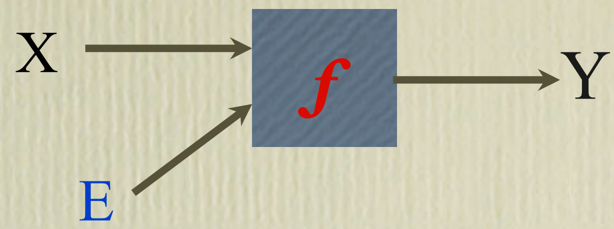
- {local causal structures} → {conditional independences}





- Instead, try to directly identify local causal structures with functional causal models

Causality is about data-generating process



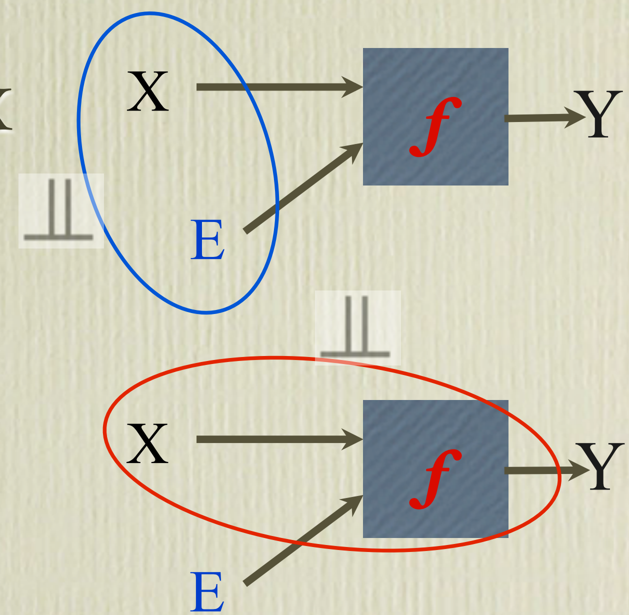
- Effect generated from cause with independent noise, represented with functional causal model : Y = f(X, E)

- Generating process for X is independent from that generates Y from X, which involves E and f

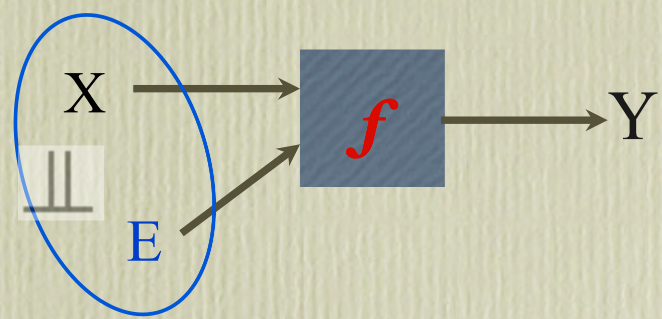
- How to describe the independence between X and E and that between X and f ?

    - X and E: statistical independence

    - X and f: "independence" between p(X) and some property of transformation f



Approach type 1: Enforce independence between X and E (with constrained f)

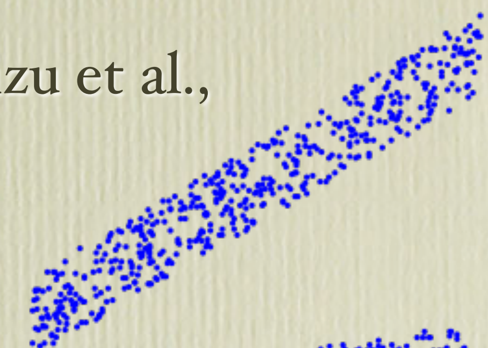


- Why useful?
  - structural constraints on f guarantees identifiability
  - identifiability guarantees asymmetry
  - in practice f can usually be approximated with a well-constrained form !

(Generally) identifiable FCMs with independent noise

- linear non-Gaussian acyclic causal model (Shimizu et al., '06)

$$Y = aX + E$$



- additive noise model (Hoyer et al., '09; Zhang & Hyvärinen, '09b)

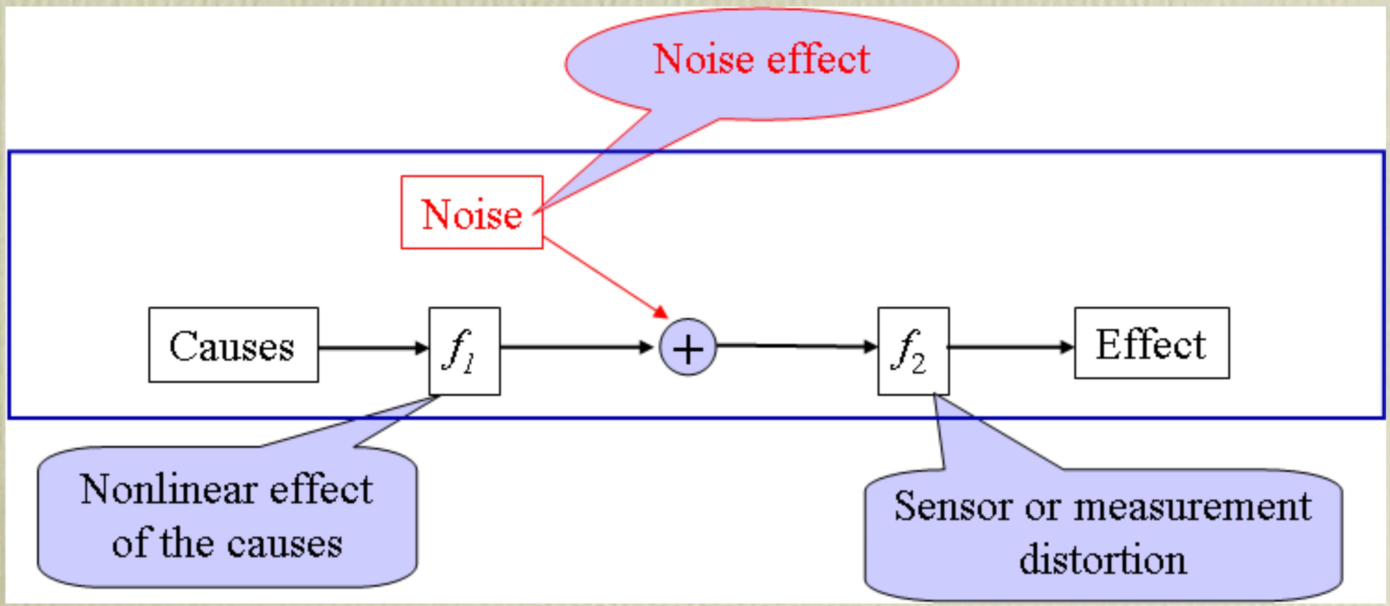
$$Y = f(X) + E$$

- post-nonlinear causal model (Zhang & Hyvärinen, '09a)

$$Y = f_2\,(\,f_1(X) + E\,)$$

Three Effects usually encountered in a causal model (Zhang & Hyvärinen, 09)

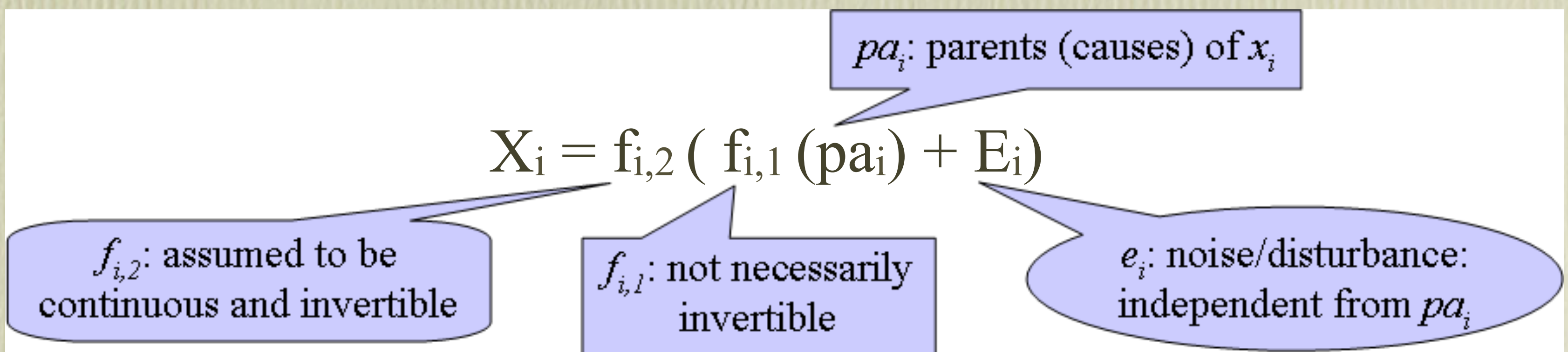
- Without prior knowledge, the assumed model is expected to be
  - general enough: adapted to approximate the true generating process
  - identifiable: asymmetry in causes and effects



- represented by <u>post-nonlinear causal model with inner additive noise</u>

PNL causal model with inner additive noise

- Acyclic data-generating process



- Two-variable case
  - $X_1 \rightarrow X_2$: $X_2 = f_{2,2}\,(\,f_{2,1}\,(X_1) + E_2)$

Identifiability in two-variable case

- Is the causal direction implied by the model unique?
- We tackle this problem by a proof of contradiction
  - Assume both $X_1 \rightarrow X_2$ and $X_1 \leftarrow X_2$ satisfy PNL model
  - One can then find all non-identifiable cases

Identifiability: A mathematical result

Notation

$t_1 \triangleq g_2^{-1}(x_1), \quad z_2 \triangleq f_2^{-1}(x_2),$
$h \triangleq f_1 \circ g_2, \quad h_1 \triangleq g_1 \circ f_2.$
$\eta_1(t_1) \triangleq \log p_{t_1}(t_1), \quad \eta_2(e_2) \triangleq \log p_{e_2}(e_2).$

- **Theorem 1**
  - Assume $x_2 = f_2(f_1(x_1) + e_2),$
    $x_1 = g_2(g_1(x_2) + e_1),$
  - Further suppose that involved densities and nonlinear functions are third-order differentiable, and that p_{e2} is unbounded,
  - For every point satisfying $\eta_2'' h' \neq 0$, we have

$$\boxed{\eta_1''' - \frac{\eta_1'' h''}{h'} = \left(\frac{\eta_2' \eta_2'''}{\eta_2''} - 2\eta_2''\right) \cdot h' h'' - \frac{\eta_2'''}{\eta_2''} \cdot h' \eta_1'' + \eta_2' \cdot \left(h''' - \frac{h''^2}{h'}\right).}$$

- Obtained by using the fact that the Hessian of the logarithm of the joint density of independent variables is diagonal everywhere (Lin, 1998)
- It is not obvious if this theorem holds in practice…

Finally: All non-identifiable cases

Log-mixed-linear-and-exponential:
$\log p_v = c_1 e^{c_2 v} + c_3 v + c_4$

$(\log p_v)' \to c \ (c \neq 0)$,
as $v \to -\infty$ or as $v \to +\infty$

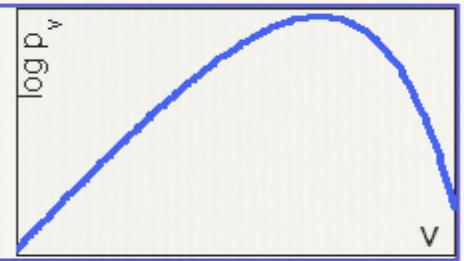
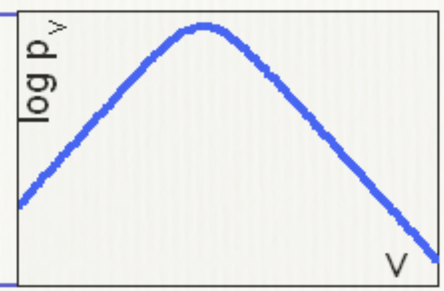


Table 1: All situations in which the PNL causal model is not identifiable.

	p_{e_2}	p_{t_1} $(t_1 = g_2^{-1}(x_1))$	$h = f_1 \circ g_2$	Remark
I	Gaussian	Gaussian	linear	h_1 also linear
II	log-mix-lin-exp	log-mix-lin-exp	linear	h_1 strictly monotonic, and $h_1' \to 0$, as $z_2 \to +\infty$ or as $z_2 \to -\infty$
III	log-mix-lin-exp	one-sided asymptotically exponential (but not log-mix-lin-exp)	h strictly monotonic, and $h' \to 0$, as $t_1 \to +\infty$ or as $t_1 \to -\infty$	—
IV	log-mix-lin-exp	generalized mixture of two exponentials	Same as above	—
V	generalized mixture of two exponentials	two-sided asymptotically exponential	Same as above	—

$p_v \propto (c_1 e^{c_2 v} + c_3 e^{c_4 v})^{c_5}$

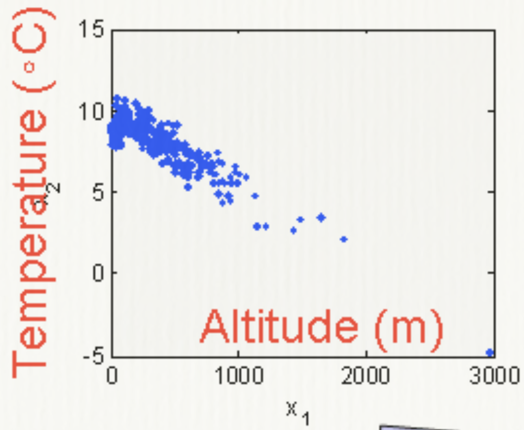
$(\log p_v)' \to c_1 \ (c_1 \neq 0)$,
as $v \to -\infty$ and
$(\log p_v)' \to c_2 \ (c_2 \neq 0)$,
as $v \to +\infty$

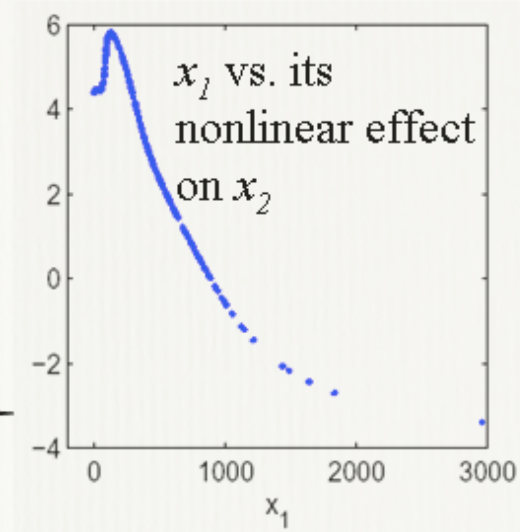


Method for distinguishing cause from effect

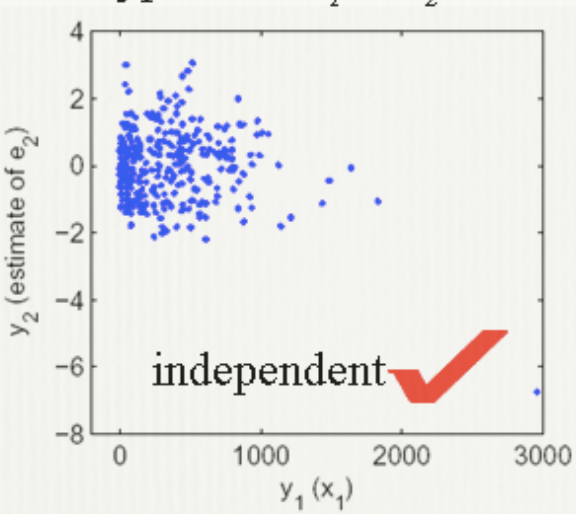
- Fit the model on both directions, estimate the noise, and test for independence
- Implemented two estimation approaches: MLP & extended warped Gaussian process regression
  - If $X_1 \rightarrow X_2$, i.e., $X_2 = f_{2,2}(f_{2,1}(X_1) + E_2)$, we have $\boxed{E_2 = f_{2,2}^{-1}(X_2) - f_{2,1}(X_1)}$ is ind. from X_1: mutual information minimization
  - $X_2 = f_{2,2}(f_{2,1}(X_1) + E_2)$, GP prior for $f_{2,1}$ and $P(E_2)$ modeled by the mixture of Gaussians: marginal likelihood maximization

Data Set 1

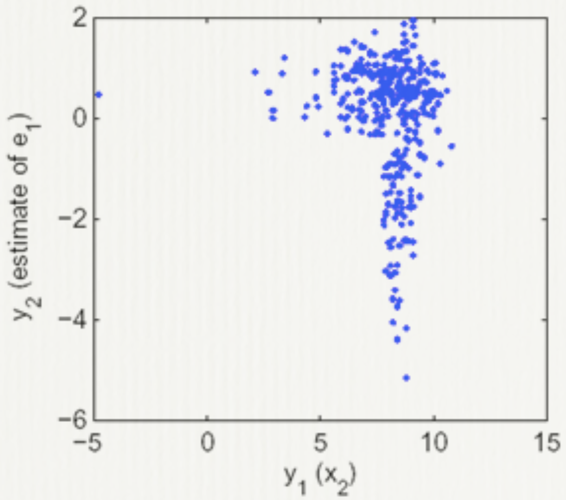


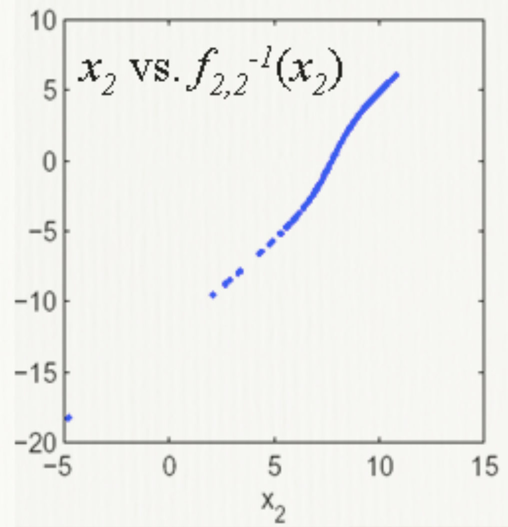


(a) y_1 vs y_2 under hypothesis $x_1 \to x_2$



(b) y_1 vs y_2 under hypothesis $x_2 \to x_1$





Independence test results on y_1 and y_2 with different assumed causal relations

Data Set	$x_1 \to x_2$ assumed		$x_2 \to x_1$ assumed	
	Threshold ($\alpha = 0.01$)	Statistic	Threshold ($\alpha = 0.01$)	Statistic
#1	2.3×10^{-3}	1.7×10^{-3}	2.2×10^{-3}	6.5×10^{-3}

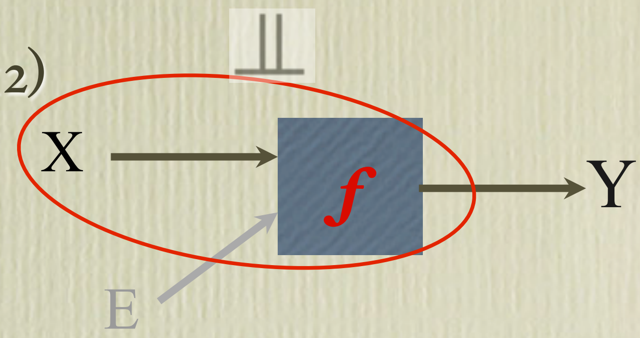
Approaches type 2: Enforcing "independence" between p(X) and complex f

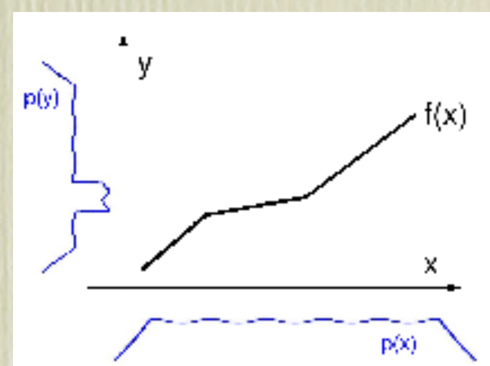
- Nonlinear deterministic case (Janzing et al. '12)
  - Y = f(X) $\Rightarrow$ p(Y) = p(X) / |f'(X)|
  - log f'(X) and p(X) uncorrelated w.r.t. a uniform reference; violated for the other direction

$$\int_0^1 \log f'(x)p(x)dx = \int_0^1 \log f'(x)\frac{p(x)}{p_0(x)}p_0(x)dx$$

$$= \int_0^1 \log f'(x)p_0(x)dx \cdot \int_0^1 p(x)dx = \int_0^1 \log f'(x)dx$$

  - Think of log f'(X) and p(X) as random processes





Performance of several methods on cause-effect pairs

- Apply different approaches for causal direction determination on 77 real cause-effect pairs, on which ground truth is known based on background info

Gaussian process latent variable model (type I)

Additive noise model (type I)

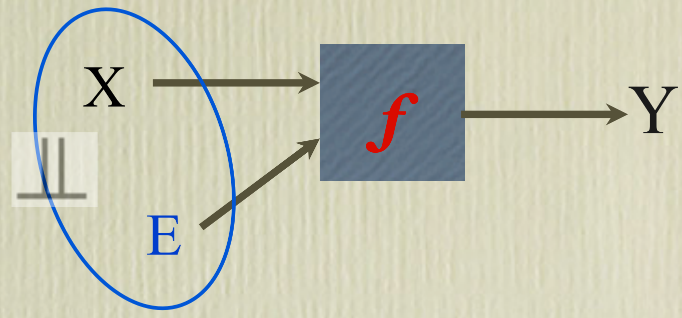
Accuracy of different methods for causal direction determination on the cause-effect pairs.

Method	PNL-MLP	PNL-WGP-Gaussian	PNL-WGP-MoG	ANM	GPI	IGCI
Accuracy (%)	70	67	76	63	72	73

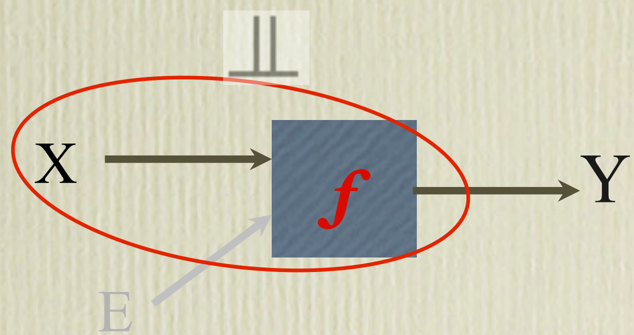
Information geometric causal inference (type II)

Two types of independence in FCMs for causal discovery: Comparison

- Independence between cause and noise:
  - Constrained f ⇒ identifiability ⇒ asymmetry
  - In practice f can usually be approximated with a simple (well constrained) form !
  - Is the f class correct?

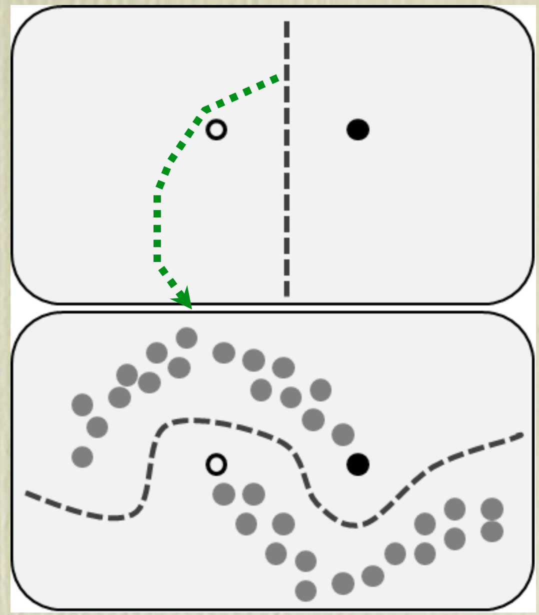


- "Independence" between P(X) and log f'(X)
  - Identifiable in the noiseless case
  - They have to be "complex" to have enough effective samples; might fail if they are simple (too smooth)
  - Noise effect ?

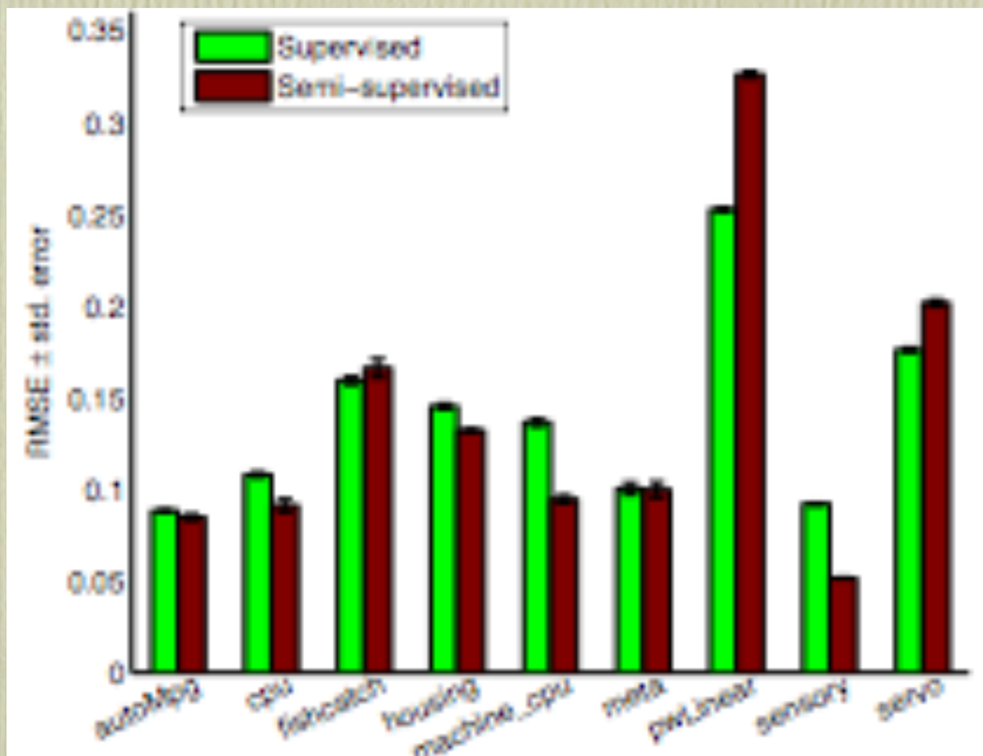


Machine learning based on causal independence: Semi-supervised learning

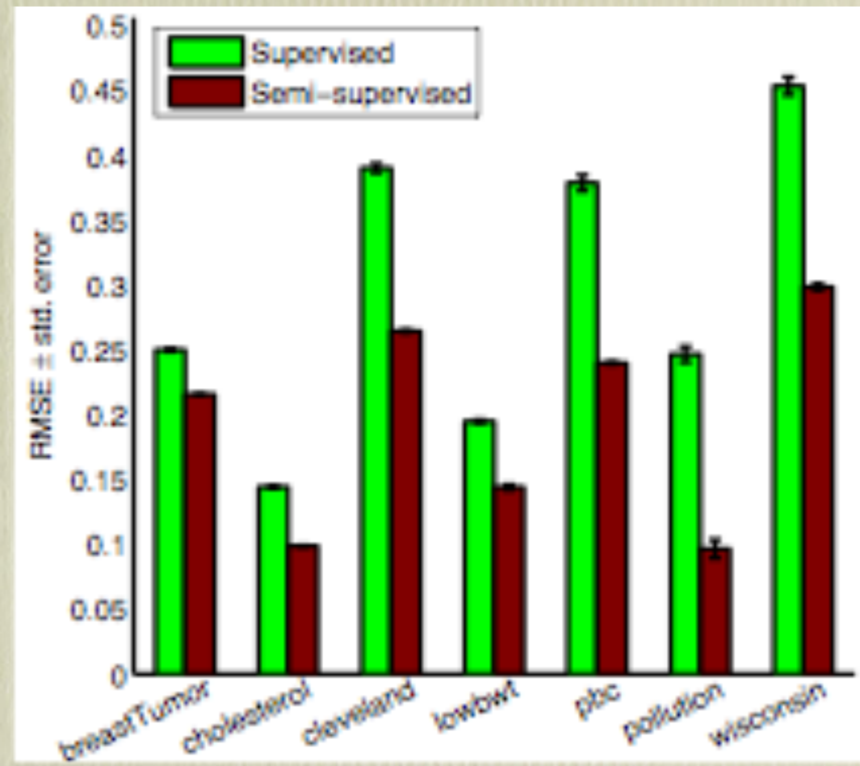
- semi-supervised learning: more precise estimate of P_X helps learn $P_{Y|X}$
- utilizes dependence between p_X and $p_{Y|X}$ (Schölkopf et al., 2012)
  - $X \rightarrow Y$: unlabeled points do not help
  - $Y \rightarrow X$: Yes



Some meta-analysis of previous experimental results



Semi-supervised regression on causal datasets ($X \to Y$)

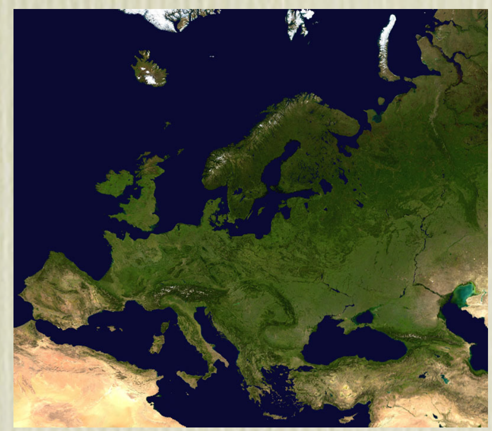


Semi-supervised regression on anticausal/confounded datasets

Machine learning based on causal independence: Domain adaptation

$(\mathbf{x}^{\mathrm{tr}}, \mathbf{y}^{\mathrm{tr}})$ source (training) → target (test) $\mathbf{x}^{\mathrm{te}}$

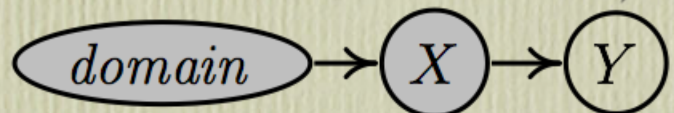
- Traditional supervised learning: $P^{te}_{XY} = P^{tr}_{XY}$
- might not be the case in practice:

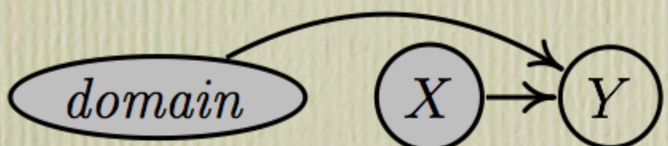


Possible situations for domain adaptation: When X→Y

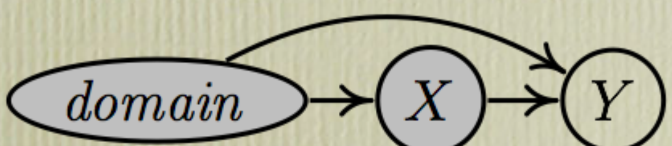
covariate shift

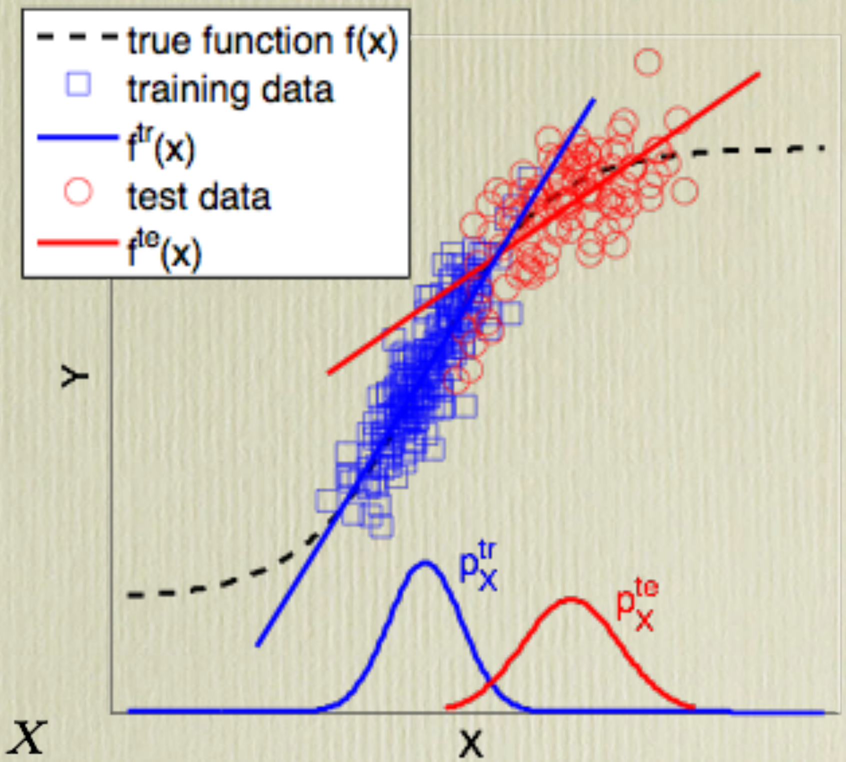
(Shimodaira00; Sugiyama etal.08; Huang etal.07, Gretton etal.08...)





☹ no clue as to find $P^{te}_{Y|X}$

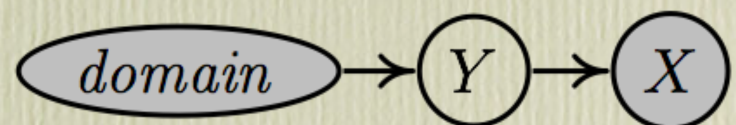




Possible situations for domain adaptation: When **Y→X** (Zhang et al., 2013)

- Y is usually the cause of X
  (especially for classification)

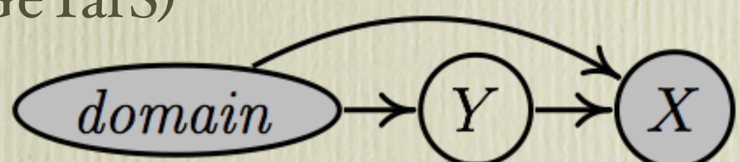
- Target shift (TarS)



- Conditional shift (ConS)

P_X^{te} helps find $P_{Y|X}^{te}$

- Generalized target shift (GeTarS)



involved parameters estimated by matching P_X

On remote sensing image classification

- two domains (area 1 & area 2)
- 14 classes

Class	Number of patterns			
	Area 1		Area 2	
	TR_1	TS_1	TR_2	TS_2
Water	69	57	213	57
Hippo grass	81	81	83	18
Floodplain grasses1	83	75	199	52
Floodplain grasses2	74	91	169	46
Reeds1	80	88	219	50
Riparian	102	109	221	48
Firescar2	93	83	215	44
Island interior	77	77	166	37
Acacia woodlands	84	67	253	61
Acacia shrublands	101	89	202	46
Acacia grasslands	184	174	243	62
Short mopane	68	85	154	27
Mixed mopane	105	128	203	65
Exposed soil	41	48	81	14
Total	1242	1252	2621	627

Misclassification rates by different methods

Problem	Unweight	CovS	TarS	LS-GeTarS
$TR_1 \to TS_2$	20.73%	20.73%	20.41%	**11.96%** ✓
$TR_2 \to TS_1$	26.36%	25.32%	26.28%	**13.56%** ✓

Summary

- Different types of independence helps in causal discovery
  - Conditional independence for constraint-based approach
  - "Independence" in FCMs gives rise to asymmetry between two variables
    - Cause & noise
    - P(cause) & transformation
    - Which one is better?
  - How to systematically make use of the info from all aspects?
- "Causal independence" could facilitate understanding & solving some machine learning tasks