Causality – in a wide sense
Lecture II

Peter Bühlmann

Seminar for Statistics
ETH Zürich

Recap from yesterday

- equivalence classes of DAGs
- estimation of equivalence classes of DAGs based on observational data
  that is: data are i.i.d. realizations from a single data-generating distribution which is faithful/Markovian w.r.t. a true underlying DAG

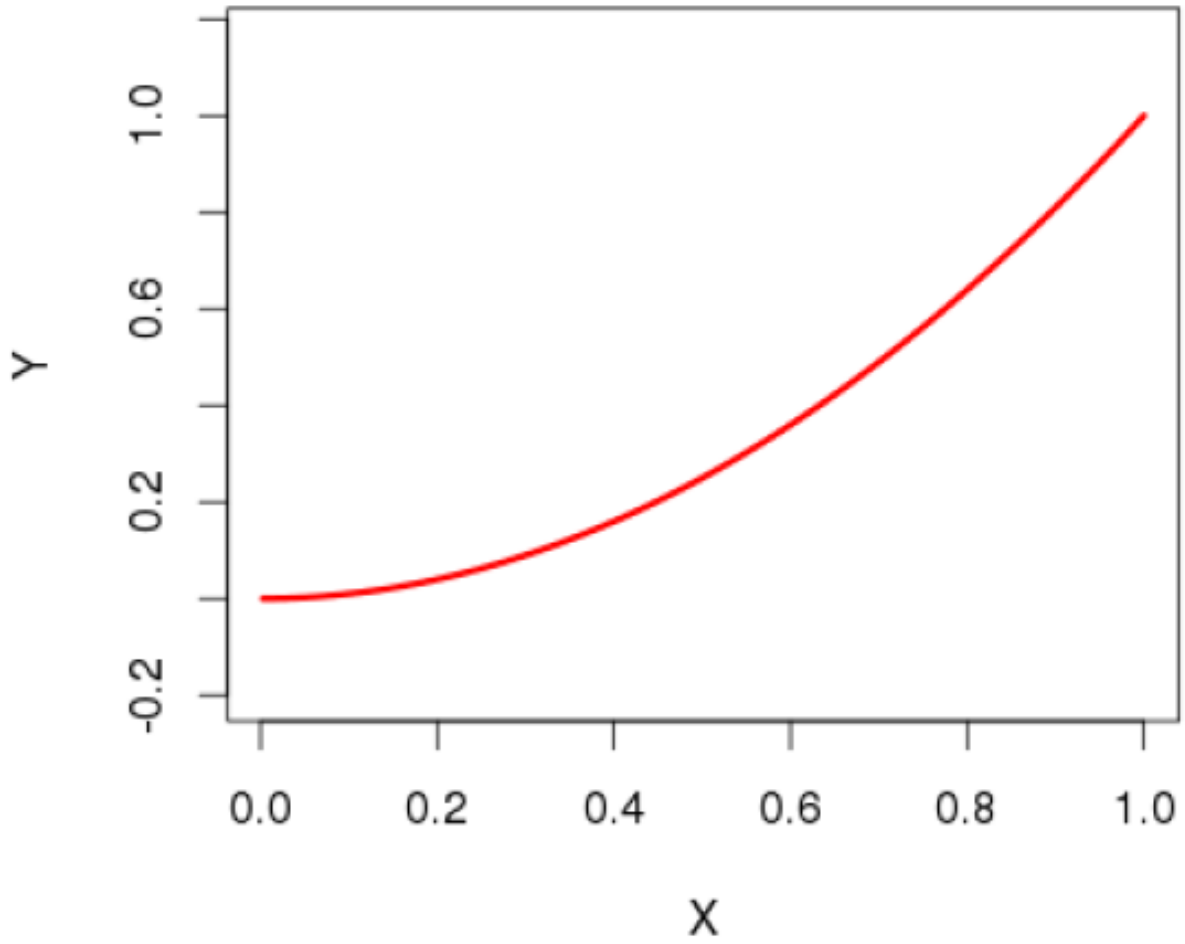
  - PC-algorithm assuming strong faithfulness conditions
  - ℓ_0-penalized Gaussian MLE assuming a weaker permutation beta min condition

Route via structural equation models: interesting conceptual extensions

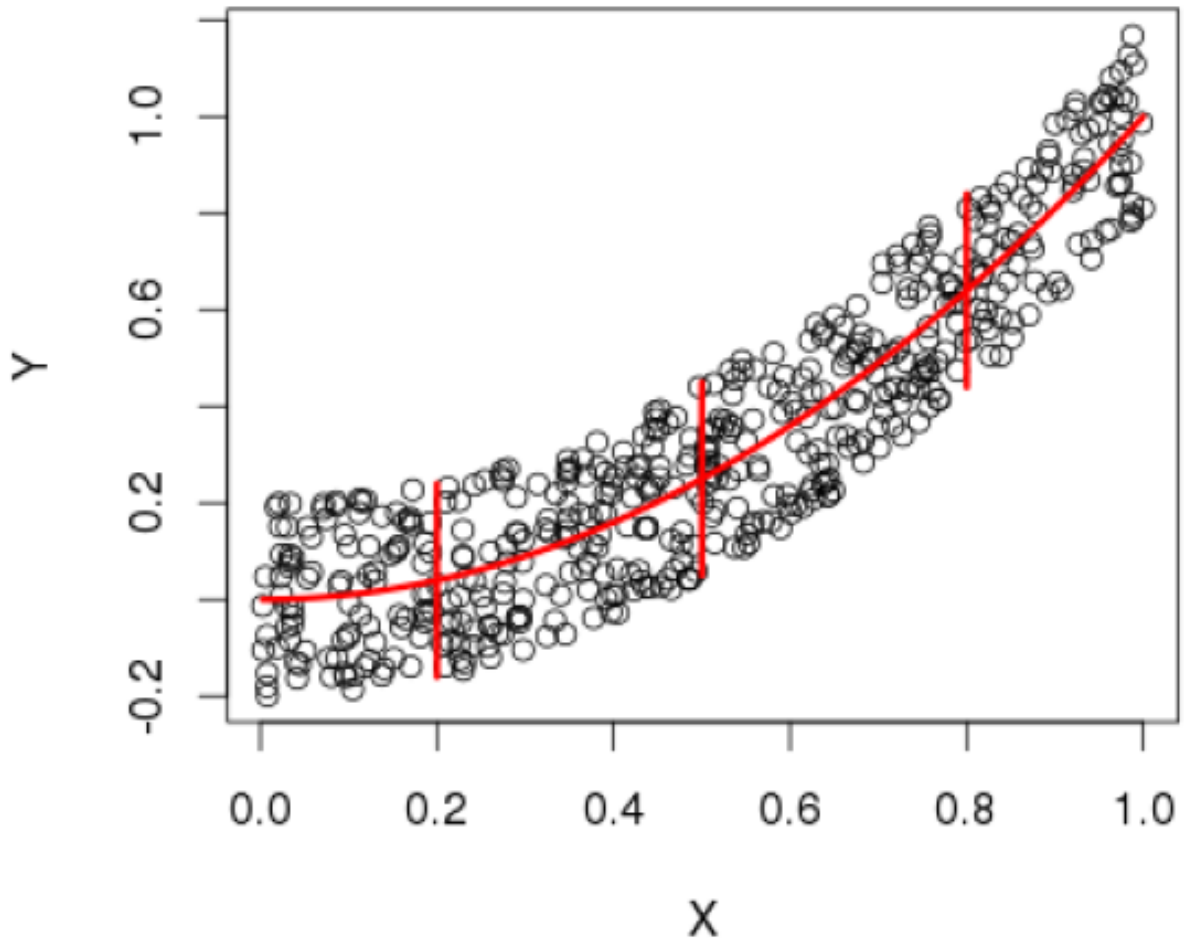
full identifiability (card(Markov equivalence class) = 1): if

- same error variances:
  $X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} B_{jk} X_k + \varepsilon_j,\ \mathrm{Var}(\varepsilon_j) \equiv \omega^2$ (Peters & PB, 2014)
- nonlinear structural equation models with additive noise:
  $X_j \leftarrow$ non-linear function $f(X_{\mathrm{pa}(j)}) + \varepsilon_j$
  Mooij, Peters, Janzing & Schölkopf (2009-2012)

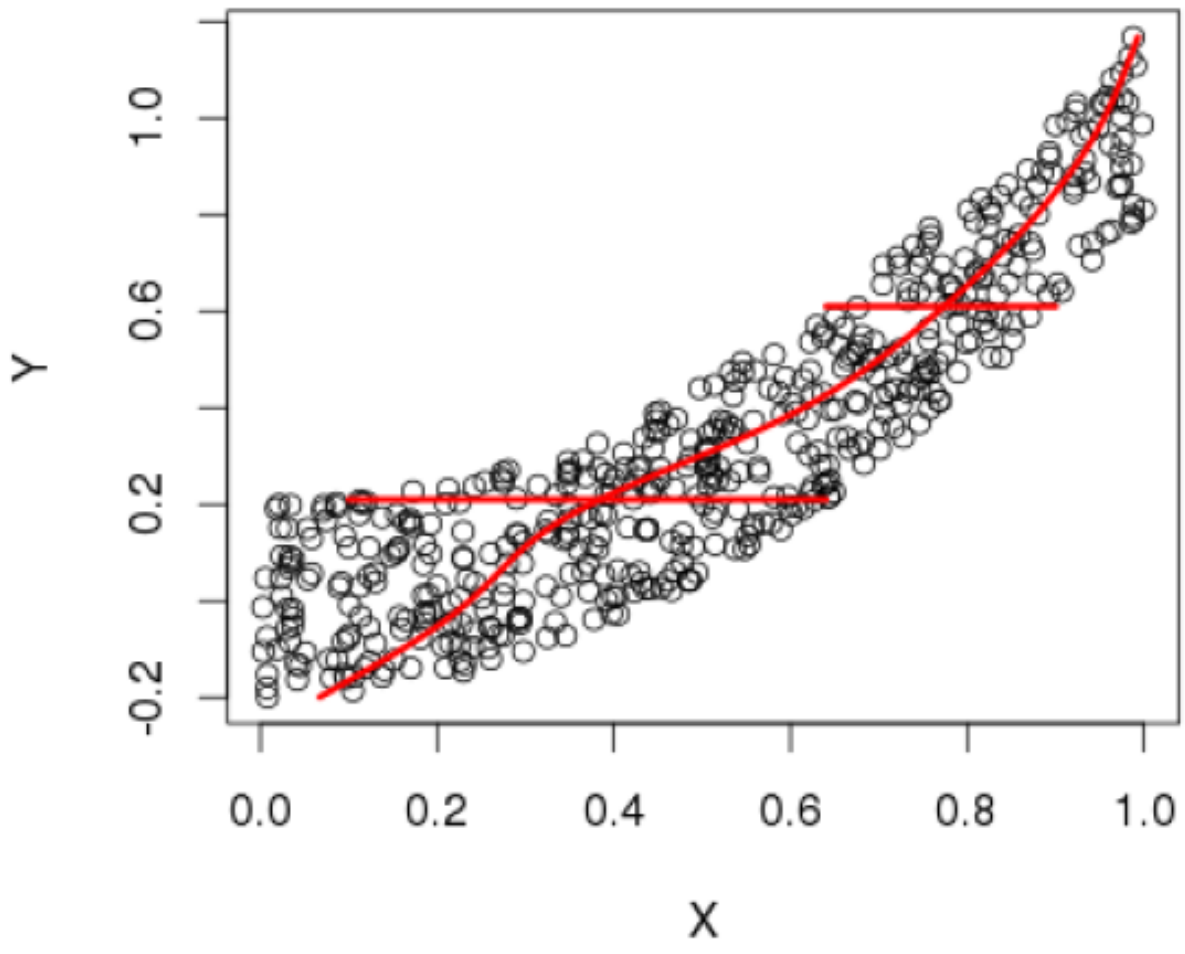
$$Y = f(X) + N_Y, \qquad N_Y \perp\!\!\!\perp X$$



$$Y = f(X) + N_Y, \qquad N_Y \perp\!\!\!\perp X$$



$$\cancel{X = g(Y) + N_X, \qquad N_X \perp\!\!\!\perp Y}$$



- nonlinear structural equation models with additive noise:
  $X_j \leftarrow$ non-linear function $f(X_{\mathrm{pa}(j)}) + \varepsilon_j$
  Mooij, Peters, Janzing & Schölkopf (2009-2012)

  $X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} f_k(X_k) + \varepsilon_j$ (CAM) (PB, Ernest & Peters, 2014)

- linear structural eqns. with non-Gaussian errors (LINGAM):
  linear SEM but all $\varepsilon_1, \ldots, \varepsilon_p$ non-Gaussian (Shimizu et al., 2006)

$$X = BX + \varepsilon$$
$$X = (I - B)^{-1}\varepsilon \quad \rightsquigarrow \text{ ICA !}$$

the real issue with causality: interventional distributions

What is Causality? ... and its relation to interventions

Causality is giving a prediction (quantitative answer) to a
"What if I do/manipulate/intervene question"

many modern applications are faced with such prediction tasks:

- genomics: what would be the effect of knocking down (the activity of) a gene on the growth rate of a plant?

  we want to predict this without any data on such a gene knock-out (e.g. no data for this particular perturbation)
- E-commerce: what would be the effect of showing person "*XYZ*" an advertisement on social media?
  no data on such an advertisement campaign for "*XYZ*" or persons being similar to "*XYZ*"
- etc.

Regression – the "statistical workhorse": the wrong approach

example:
Y = growth rate of Arabidopsis Thaliana
X = gene expressions
What would happen if we knock out a gene (expression) X_j?

we could use linear model (fitted from n observational data)

$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon, \quad \mathrm{Var}(X_j) \equiv 1 \text{ for all } j$$

$|\beta_j|$ measures the effect of variable X_j in terms of "association"

i.e. change of Y as a function of X_j when keeping all other variables X_k fixed

$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these others are not (cannot be) kept fixed

Regression – the "statistical workhorse": the wrong approach

example:
Y = growth rate of Arabidopsis Thaliana
X = gene expressions
What would happen if we knock out a gene (expression) X_j?

we could use linear model (fitted from n observational data)

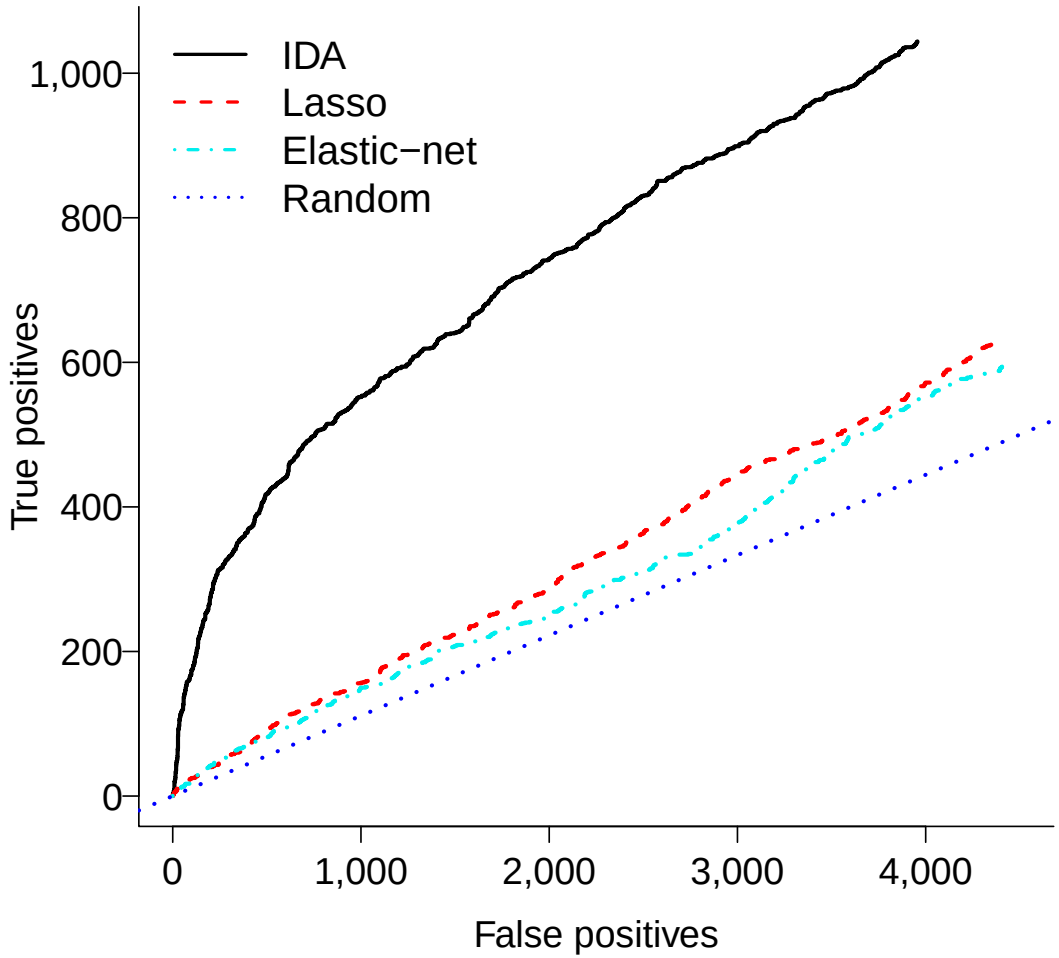
$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon, \quad \mathrm{Var}(X_j) \equiv 1 \text{ for all } j$$

$|\beta_j|$ measures the effect of variable X_j in terms of "association"

i.e. change of Y as a function of X_j when keeping all other variables X_k fixed

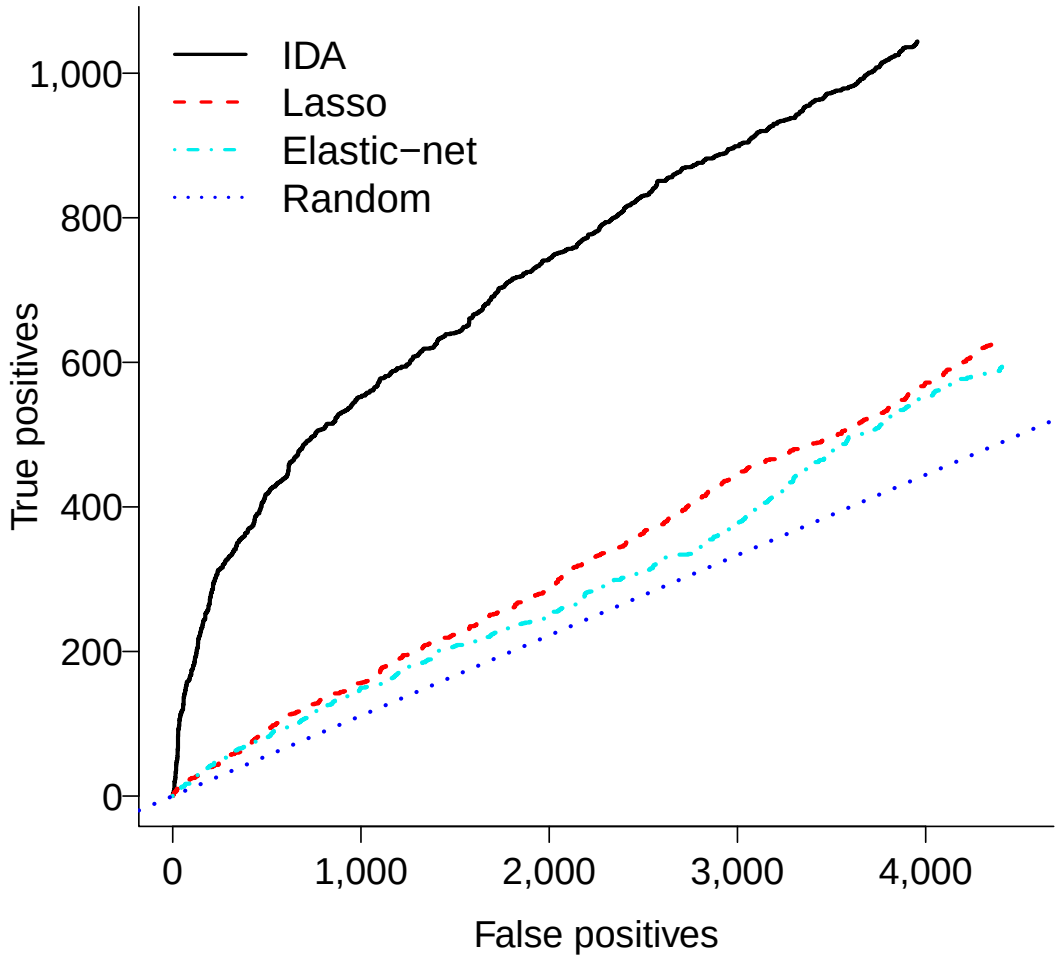
$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these others are not (cannot be) kept fixed

and indeed:



$\rightsquigarrow$ can do much better than (penalized) regression!

and indeed:



$\rightsquigarrow$ can do much better than (penalized) regression!

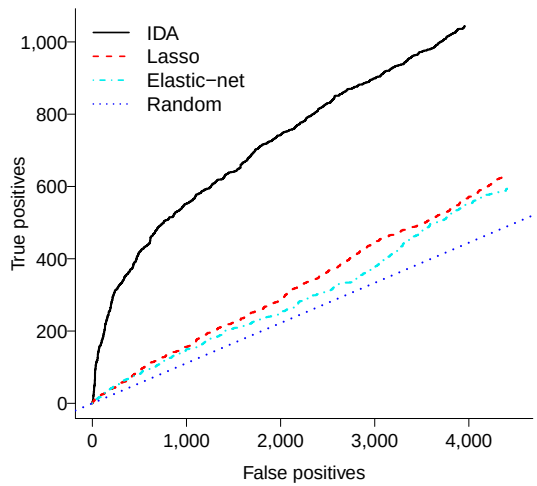
Effects of single gene knock-downs on all other genes (yeast)

(Maathuis, Colombo, Kalisch & PB, 2010)

- $p = 5360$ genes (expression of genes)
- 231 gene knock downs $\rightsquigarrow 1.2 \cdot 10^6$ intervention effects
- the truth is "known in good approximation" (thanks to intervention experiments)

goal: prediction of the true large intervention effects based on observational data with no knock-downs

$n = 63$
observational data



A bit more specifically

- univariate response Y
- p-dimensional covariate X

question:
what is the effect of setting the jth component of X to a certain value x:

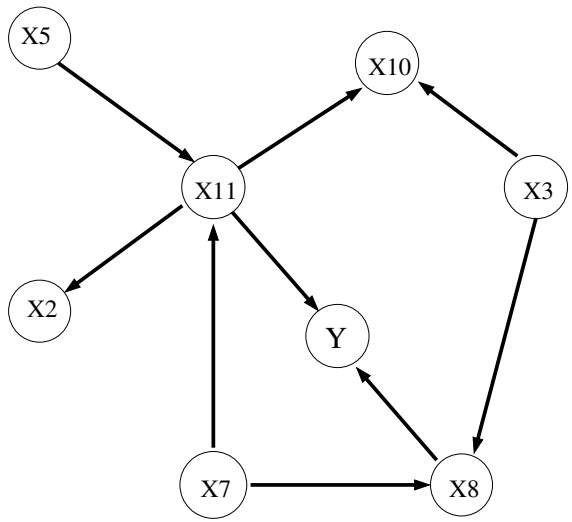
$$\mathrm{do}(X_j = x)$$

$\rightsquigarrow$ this is a question of intervention type

not the effect of X_j on Y when keeping all other variables fixed (regression effect)
Reichenbach, 1956; Suppes, 1970; Rubin, 1978; Dawid, 1979; Holland, Pearl, Glymour, Scheines, Spirtes,...

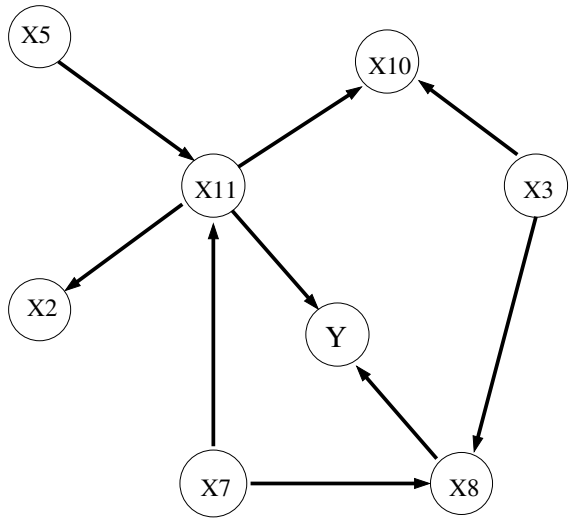
we need a “dynamic notion of importance”:

if we intervene at X_j, its effect propagates through other variables X_k $(k \neq j)$ to Y



Graphs, structural equation models and causality

intuitively:
the concept of causality in terms of graphs is plausible



in a DAG:
a directed arrow $X \to Y$ says that "X is a direct cause of Y"

- What about indirect causes? (when propagating through many variables)
  How do we link "causality" to graphs?
- What is a quantitative model for a graph structure?

Structural equation models (SEMs)

consider a DAG D ("acyclicity" for simplicity)
encoding the "causal influence diagram":
the direct causes are encoded by directed arrows

$\rightsquigarrow$ D is called the causal graph (because it is assumed to encode the direct causal relationships)

a quantitative model on the causal graph describing the quantitative behavior of the system:

structural equation model (with structure D):

$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j),\ j = 1, \ldots, p$$
$$\varepsilon_1, \ldots, \varepsilon_p \text{ independent}$$

where $\mathrm{pa}(j) = \mathrm{pa}_D(j)$ are the parents of node j

Linear SEM

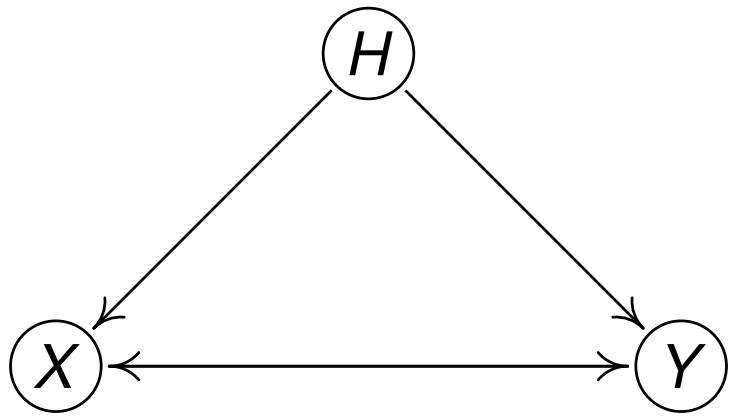
linear structral equation model (with structure D):

$$X_j \leftarrow \sum_{k \in \mathrm{pa}(j)} B_{jk} X_k + \varepsilon_j, \; j = 1, \ldots, p$$

$\varepsilon_1, \ldots, \varepsilon_p$ independent

if we knew the parental sets it is simply linear regression on the appropriate covariates

so far: no hidden “confounding” variables



$\rightsquigarrow$ see Lecture III

Local Markov property

Given P with density p from a SEM
because of independence of $\varepsilon_Y, \varepsilon_1, \ldots, \varepsilon_p$
$\rightsquigarrow$ the local Markov property holds!

and if P has continuous density: global Markov property holds!
(correspondence between conditional independence and separation in graphs)

Causality and SEM

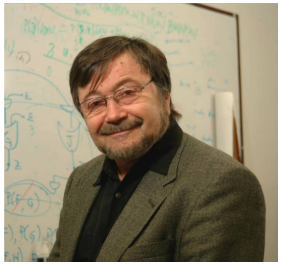
the SEM is a model for describing the "true" underlying mechanistic behavior of the system with the random variables $Y, X_1, \ldots, X_p$

having access to such a mechanistic model, one can make predictions of interventions, manipulations, perturbations

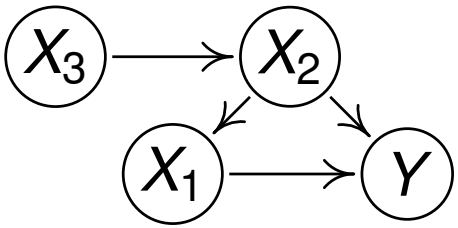
and this is the core task of causality

Modeling interventions: do-interventions

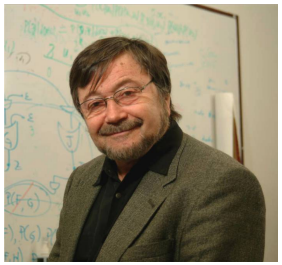
Pearl's do-interventions



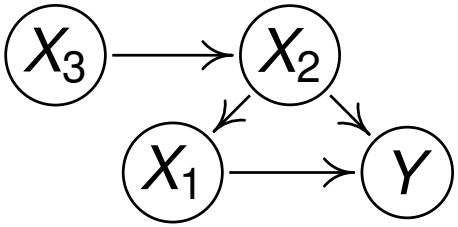
Judea Pearl



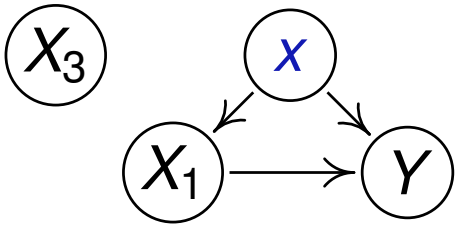
Pearl's do-interventions



Judea Pearl



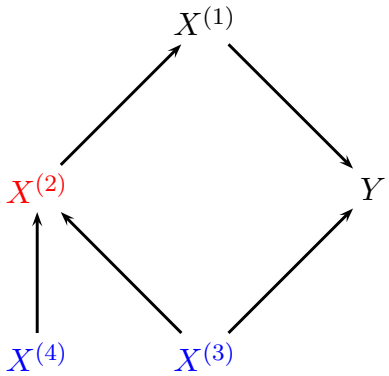
$$\mathrm{do}(X_2 = x) \rightsquigarrow$$



$$X_1 \leftarrow f_1(X_2 = x, \varepsilon_1),$$
$$X_2 \leftarrow x,$$
$$X_3 \leftarrow \varepsilon_3$$
$$Y \leftarrow f_Y(X_1, X_2 = x, \varepsilon_Y)$$

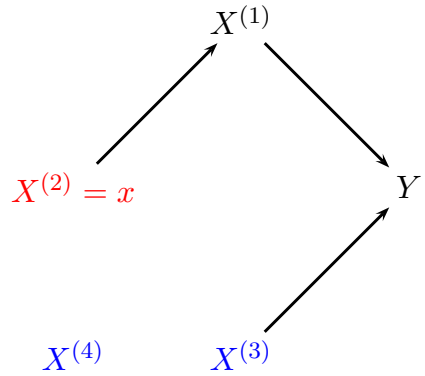
assume Markov property (rec. factorization) for causal DAG:

non-intervention



$$p(Y, X_1, X_2, X_3, X_4) = p(Y|X_1, X_3) \times p(X_1|X_2) \times p(X_2|X_3, X_4) \times p(X_3) \times p(X_4)$$

intervention $\mathrm{do}(X_2 = x)$



$$p(Y, X_1, X_3, X_4|\mathrm{do}(X_2 = x)) = p(Y|X_1, X_3) \times p(X_1|X_2 = x) \times p(X_3) \times p(X_4)$$

truncated factorization

truncated factorization for $\mathrm{do}(X_2 = x)$:

$$
\begin{aligned}
& p(Y, X_1, X_3, X_4 | \mathrm{do}(X_2 = x) \\
= \quad & p(Y | X_1, X_3) p(X_1 | X_2 = x) p(X_3) p(X_4)
\end{aligned}
$$

$$
\begin{aligned}
& p(Y | \mathrm{do}(X_2 = x)) \\
= \quad & \int p(Y, X_1, X_3, X_4 | \mathrm{do}(X_2 = x)) dX_1 dX_3 dX_4
\end{aligned}
$$

note that $\mathrm{do}(X_2 = x)$ does not change the factors

$$p(x_j | x_{\mathrm{pa}(j)})$$

this is an assumption!
and is called structural autonomous assumption

the intervention distribution $P(Y|\mathrm{do}(X_2 = x))$ can be calculated from

- observational data distribution
  $\rightsquigarrow$ need to estimate conditional distributions
- an influence diagram (causal DAG)
  $\rightsquigarrow$ need to estimate structure of a graph/influence diagram

with a SEM and (for example) do-interventions:

with $\mathrm{do}(X_j = x)$, for every j and x, we obtain a different distribution of $Y, X_1, \ldots, X_p$

can generate many interventional distributions!

Potential outcome model

Neyman (1923), Rubin (1974)

$Y_i(t) =$ response for unit/individual i under treatment

$Y_i(c) =$ response for unit/individual i under control

observed is (usually) only under control (or under treatment) but not both

$\rightsquigarrow$ missing data problem

"fact": the approach with do-interventions and the one with the potential outcome model are equivalent (under "natural" assumptions): 148 pages!

Single World Intervention Graphs (SWIGs):
A Unification of the Counterfactual and Graphical Approaches to Causality

Thomas S. Richardson
University of Washington

James M. Robins
Harvard University

Working Paper Number 128
Center for Statistics and the Social Sciences
University of Washington

30 April 2013

the approach with graphs is perhaps easier when many variables are present

Total causal effects

often one is interested in the distribution of $P(Y|\mathrm{do}(X_j = x))$ or

$$p(y|\mathrm{do}(X_j = x)) \text{ density}$$

$$\mathbb{E}[Y|\mathrm{do}(X_j = x)] = \int y p(y|\mathrm{do}(X_j = x)) dy$$

the total causal effect is defined as

$$\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X_j = x)]$$

measuring the "total causal importance" of variable X_j on Y

if we know the entire SEM, we can easily simulate the distribution $P(Y|\mathrm{do}(X_j = x))$
this approach requires global knowledge of the graph structure, edge functions/weights and error distributions

Total causal effects

often one is interested in the distribution of $P(Y|\mathrm{do}(X_j = x))$ or

$$p(y|\mathrm{do}(X_j = x)) \text{ density}$$

$$\mathbb{E}[Y|\mathrm{do}(X_j = x)] = \int y p(y|\mathrm{do}(X_j = x)) dy$$

the total causal effect is defined as

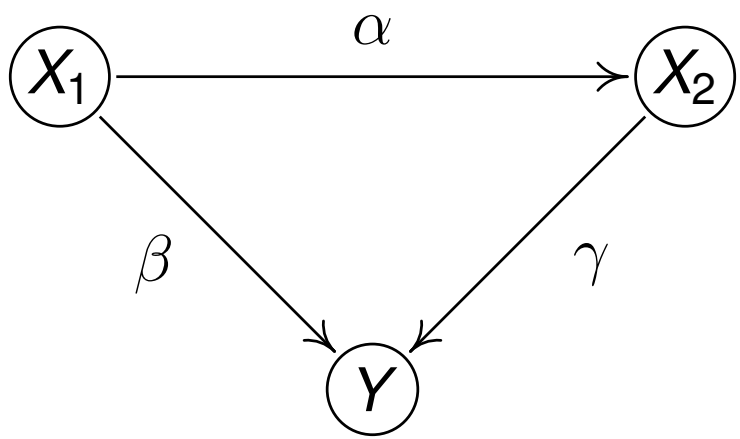
$$\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X_j = x)]$$

measuring the "total causal importance" of variable X_j on Y

if we know the entire SEM, we can easily simulate the distribution $P(Y|\mathrm{do}(X_j = x))$
this approach requires global knowledge of the graph structure, edge functions/weights and error distributions

Example: linear SEM

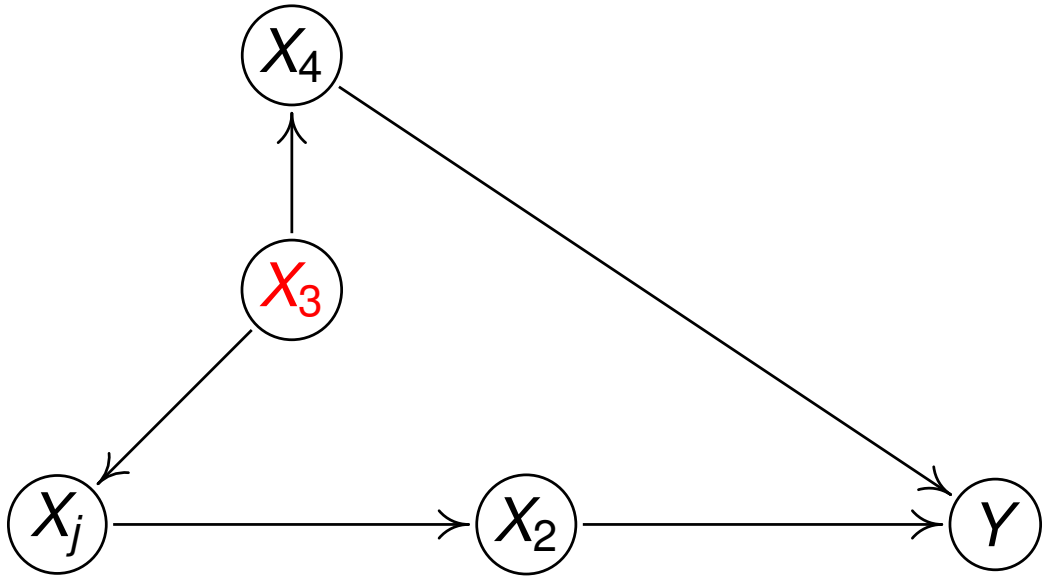
directed path p_j from X_j to Y
causal effect on p_j by product of corresponding edge weights
total causal effect $= \sum_{p_j} \gamma_j$



total causal effect from X_1 to Y: $\alpha\gamma + \beta$

needs the entire structure and edge weights of the graph

alternatively, we can use the backdoor adjustment formula: consider a set S of variables which block the "backdoor paths" of X_j to Y: one easy way to block these paths is $S = \mathrm{pa}(j)$



$\mathrm{pa}(j) = \{3\}$

backdoor adjustment formula (cf. Pearl, 2000): if $Y \notin \mathrm{pa}(j)$,

$$p(y|\mathrm{do}(X_j = x)) = \int p(y|X_j = x, X_S) dP(X_S)$$

$$\mathbb{E}[Y|\mathrm{do}(X_j) = x)] = \int y p(y|\mathrm{do}(X_j = x)) dy$$

$$= \int y p(y|X_j = x, X_S) dP(X_S) dy = \int \mathbb{E}[Y|X_j, X_S] dP(X_S)$$

for linear SEM: run regression of Y versus X_j, X_S
$\rightsquigarrow$ total causal effect of X_j on Y is regression coefficient β_j

only local structural information is required, namely e.g.
$S = \mathrm{pa}(j)$
often much easier to obtain/estimate than the entire graph

consequences: for total causal effect $\mathrm{do}(X_j = x)$, it is sufficient to know

- $\mathrm{pa}(j)$ — local graphical structure search
- $\mathbb{E}[Y|X_j = x, X_{\mathrm{pa}(j)}]$ — nonparametic regression

Henckel, Perkovic & Maathuis (2019) discuss efficiency for total causal effect estimation
with or without backdoor adjustment, possibly with a set $S \neq \mathrm{pa}(j)$, when the graph is known/given

Marginal integration (with $S = \mathrm{pa}(j)$)

recall that (for $Y \notin \mathrm{pa}(j)$)

$$\mathbb{E}[Y|\mathrm{do}(X_j = x)] = \int \mathbb{E}[Y|X_j = x, X_{\mathrm{pa}(j)}]dP(X_{\mathrm{pa}(j)})$$

estimation of the right-hand side has been developed for additive models!
cf. Fan, Härdle & Mammen (1998)
additive regression model:

$$Y = \mu + \sum_{j=1}^{d} f_j(X_j) + \varepsilon,$$
$$\mathbb{E}[f_j(X_j)] = 0 \text{ (for identifiability)}$$

$\rightsquigarrow \int \mathbb{E}[Y|X_j = x, X_{\setminus j}]dP(X_{\setminus j}) = \mu + f_j(x)$

asymp. result (Fan, Härdle & Mammen, 1998; Ernest & PB, 2015):

- regression function $\mathbb{E}[Y|X_j = x, X_{\text{pa}(j)} = x_{\text{pa}(j)}]$ exists and has bounded partial derivatives up to order 2 with respect to x and up to order $d > |\text{pa}(j)|$ w.r.t. $x_{\text{pa}(j)}$
- other regularity conditions

then, for kernel estimators with appropriate bandwidth choice:

$$\widehat{\mathbb{E}}[Y|\text{do}(X_j = x)] - \mathbb{E}[Y|\text{do}(X_j = x)] = O_P(n^{-2/5})$$

only one-dimensional variable x for the intervention

quite "nice" since the SEM is allowed to be very nonlinear with non-additive errors etc... (but smooth regression functions)

Ernest & PB (2015):

$$Y \leftarrow \exp(X_1) \times \cos(X_2 X_3 + \varepsilon_Y)$$

would be hard to model nonparametrically
$\rightsquigarrow$ instead, we rely on smoothnes of conditional expectations only

the approach by plugging-in a kernel estimator is a bit subtle in terms of choosing bandwidths (in "direction" x and $x_{\mathrm{pa}(j)}$)
one actual implementation is with boosting kernel estimation
(Ernest & PB, 2015)

Gene expressions in Arabidposis Thaliana (Wille et al., 2004)

$p = 38$, $n = 118$
graph estimated by CAM: causal additive model
Marginal integration with parental sets as in Ernest & PB (2015)

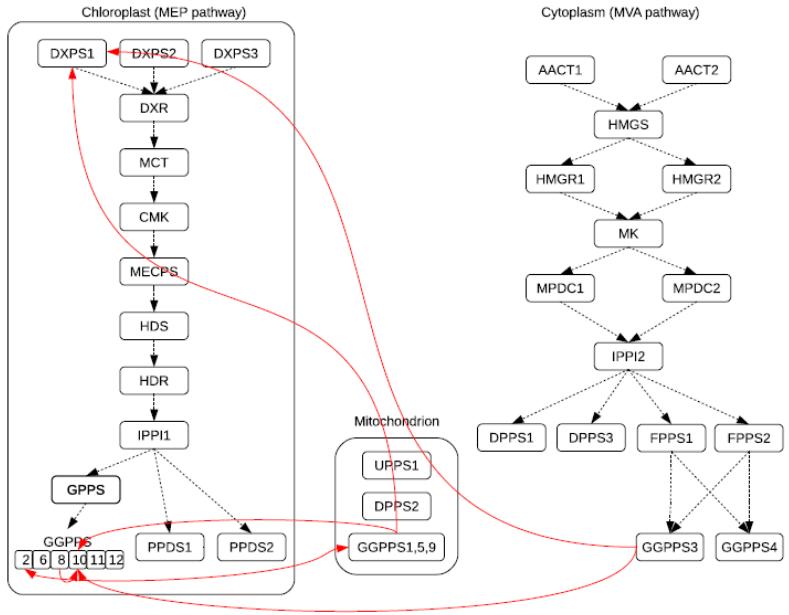


FIG 13. *Stable edges (with stability selection) for the Arabidopsis thaliana dataset. The dotted arcs represent the metabolic network, the red solid arcs the stable total causal effects found by the* est S-mint *method.*

none of the found strong total effects are against the metabolic order

one pathway: parental sets are the three closest ancestors according to metabolic order (Ernest & PB, 2015)

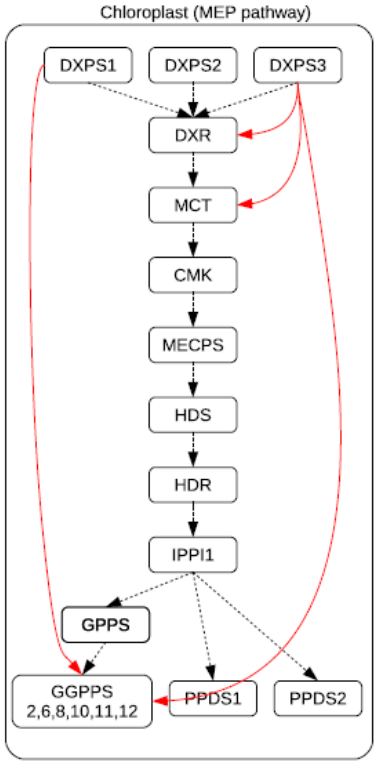


FIG 14. *Stable edges (with stability selection) for the MEP pathway in the Arabidopsis thaliana dataset. The dotted arcs represent the metabolic network whereas the red solid arcs denote the top ranked causal effects found by* S-mint *with adjustment sets chosen from the order of the metabolic network structure by considering all ancestors up to three levels back.*

from simulations: for marginal integration, the sensitivity on the correctness of the parental set is (fortunately) not so big

Lower bounds of total causal effects

due to identifiability issues:
we cannot estimate causal/intervention effects from observational distribution

but we will be able to estimate lower bounds of causal effects

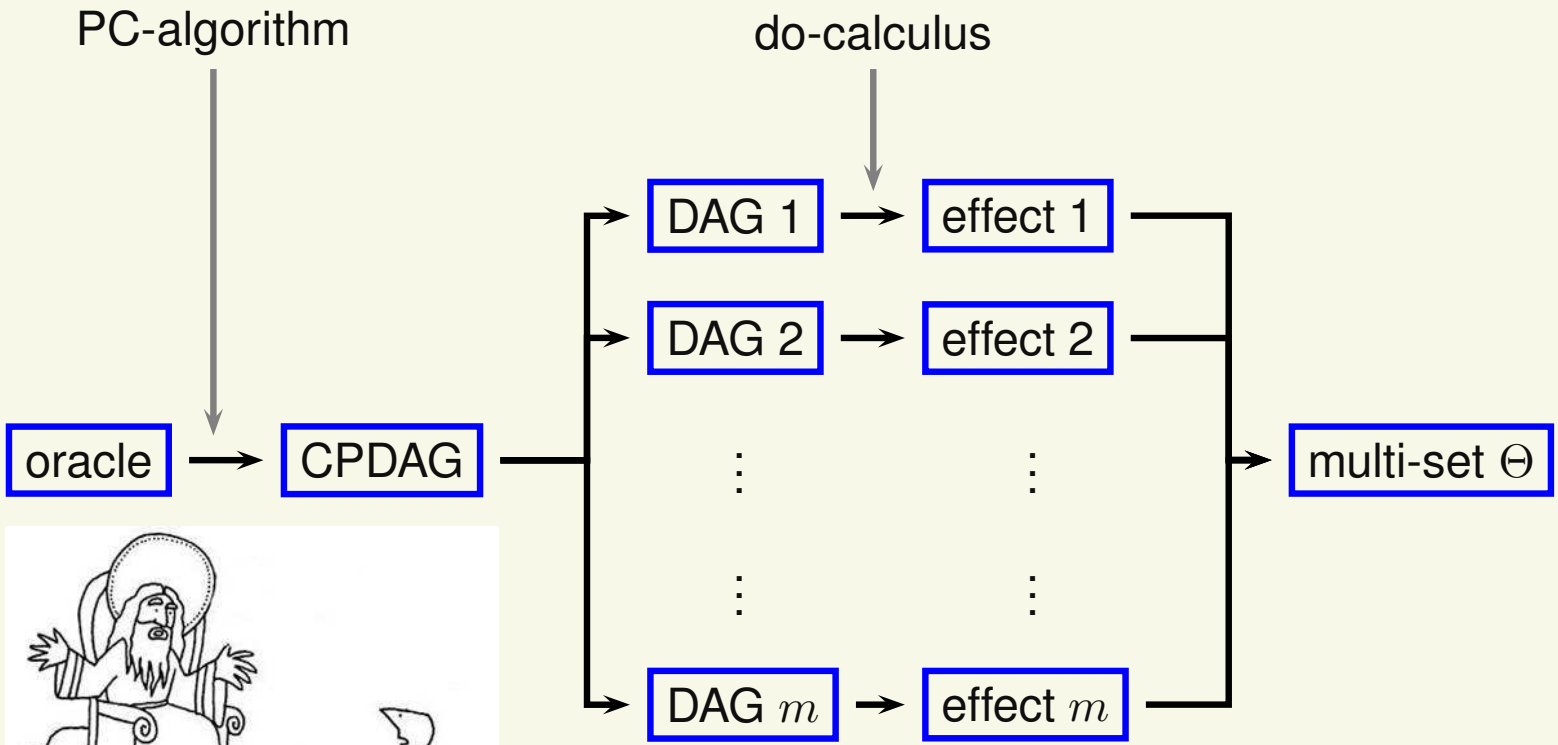
Lower bounds of total causal effects

due to identifiability issues:
we cannot estimate causal/intervention effects from observational distribution

but we will be able to estimate lower bounds of causal effects

IDA (Maathuis, Kalisch & PB, 2009)

IDA (oracle version)



If you want a single number for every variable ...

instead of the multi-set

$$\Theta = \{\theta_{r,j};\ r = 1, \ldots, m;\ j = 1, \ldots, p\}$$

minimal absolute value

e.g. for var. j: $\underbrace{|\theta_{2,j}|}_{\text{minimum}} \leq |\theta_{5,j}| \leq |\theta_{1,j}| \leq \underbrace{|\theta_{4,j}|}_{\text{true}} \leq \ldots \leq |\theta_{8,j}|$

$$\alpha_j = \min_r |\theta_{r,j}| \quad (j = 1, \ldots, p),$$

$$|\theta_{\text{true},j}| \geq \alpha_j$$

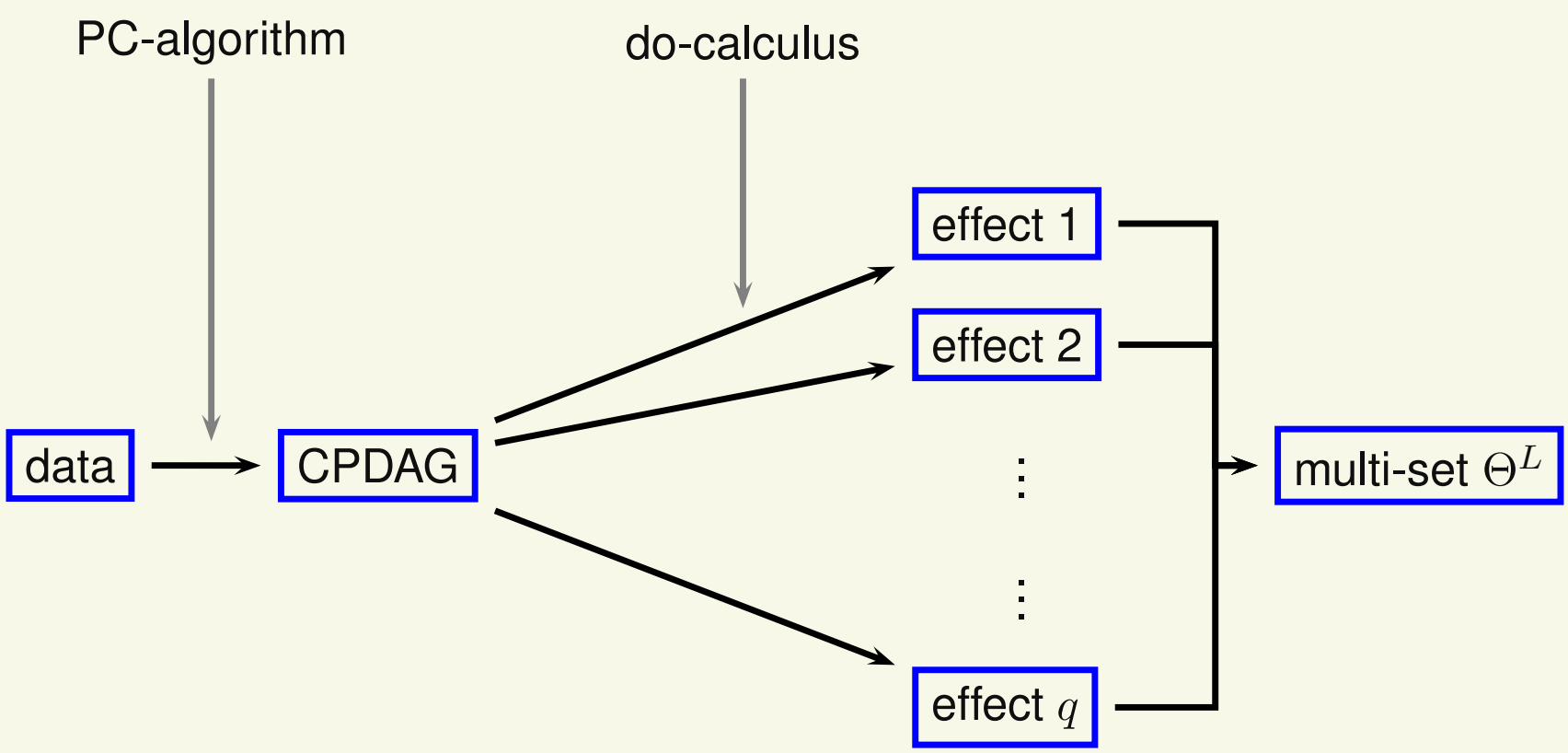
minimal absolute effect α_j is a lower bound for true absolute intervention effect

Computationally tractable algorithm

searching all DAGs is computationally infeasible if p is large (we actually can do this up to $p \approx 15 - 20$)

instead of finding all m DAGs within an equivalence class $\rightsquigarrow$ compute all intervention effects without finding all DAGs (Maathuis, Kalisch & PB, 2009)

key idea: exploring local aspects of the graph is sufficient



the local $\Theta^L = \Theta$ up to multiplicities

(Maathuis, Kalisch & PB, 2009)

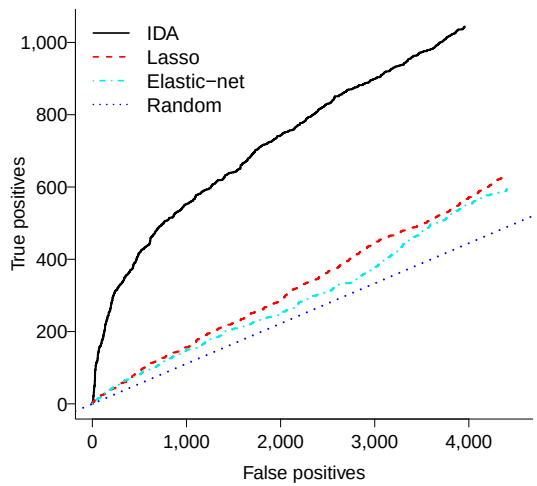
Effects of single gene knock-downs on all other genes (yeast)

(Maathuis, Colombo, Kalisch & PB, 2010)

- $p = 5360$ genes (expression of genes)
- 231 gene knock downs $\rightsquigarrow 1.2 \cdot 10^6$ intervention effects
- the truth is "known in good approximation" (thanks to intervention experiments)

goal: prediction of the true large intervention effects based on observational data with no knock-downs

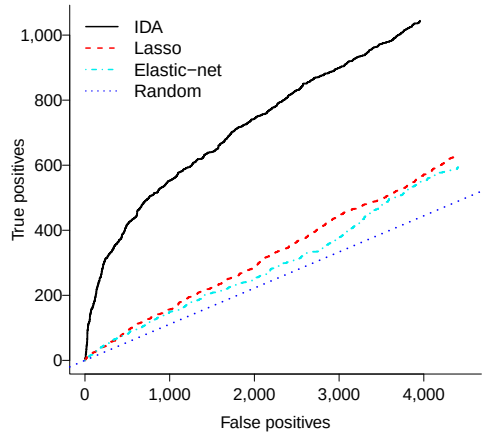
$n = 63$
observational data



Interventions and active learning

often we have observational and interventional data

example:
yeast data with $n_{\text{obs}} = 63$, $n_{\text{int}} = 231$



interventional data are very informative!
can tell the direction of certain arrows
$\rightsquigarrow$ Markov equivalence class under interventions is (much) smaller, i.e., (much) improved identifiability!

Toy problem: two (Gaussian) variables X, Y

when doing an intervention at one of them, can infer the direction

scenario I:

DAG : $X \to Y$; intervention at Y $\rightsquigarrow$ interv. DAG : $X \quad Y$

$\rightsquigarrow$ X, Y independent

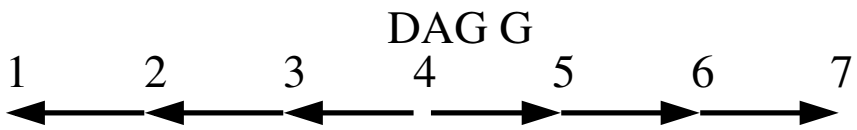
scenario II:

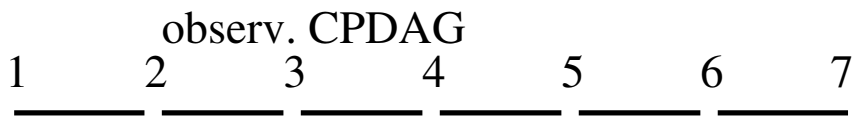
DAG : $X \leftarrow Y$; intervention at Y $\rightsquigarrow$ interv.. DAG : $X \leftarrow Y$

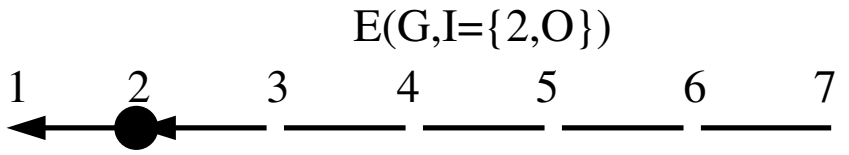
$\rightsquigarrow$ X, Y dependent

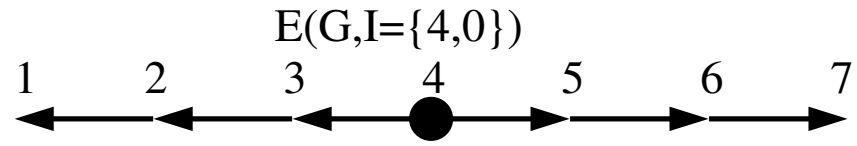
generalizes to: can infer all directions when doing an intervention at every node (which is not very clever...)

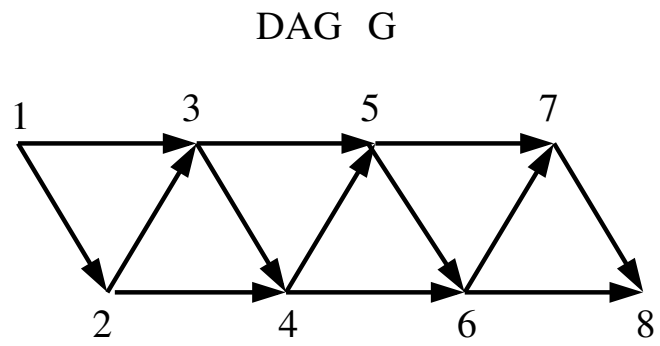
Gain in identifiability (with one intervention)

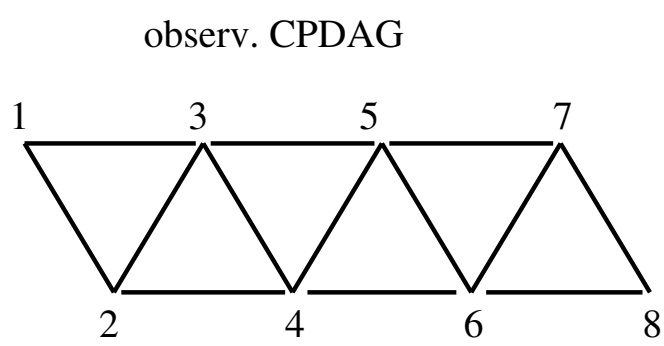


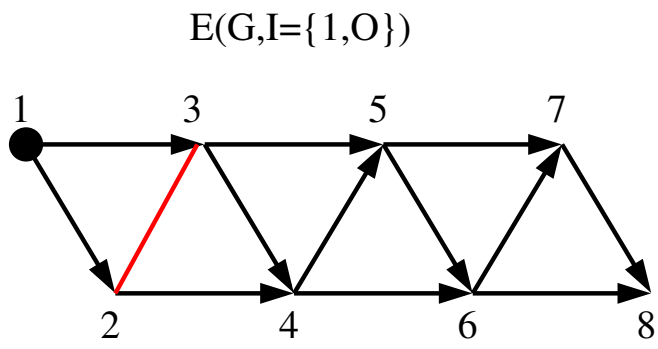


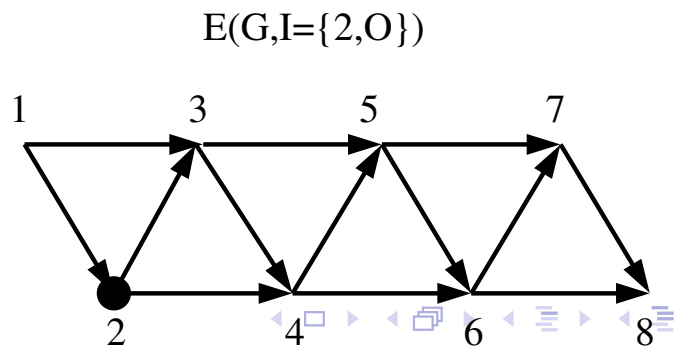












have just informally introduced interventional Markov equivalence class and its corresponding essential graph

$$\mathcal{E}(D, \underbrace{\mathcal{I}}_{\text{set of intervention variables}})$$

(needs new definitions: Hauser & PB, 2012)

there is a minimal set of intervention variables $\mathcal{I}_{\min}$ such that $\mathcal{E}(D, \mathcal{I}_{\min}) = D$
in previous example: $\mathcal{I}_{\min} = \{2, O\}$

the size of $\mathcal{I}_{\min}$ has to do with "degree" of so-called protectedness

very roughly speaking:
the "sparser (few edges) the DAG D, the better identifiable from observational/intervention data"
in the sense that $|\mathcal{I}_{\min}|$ is small

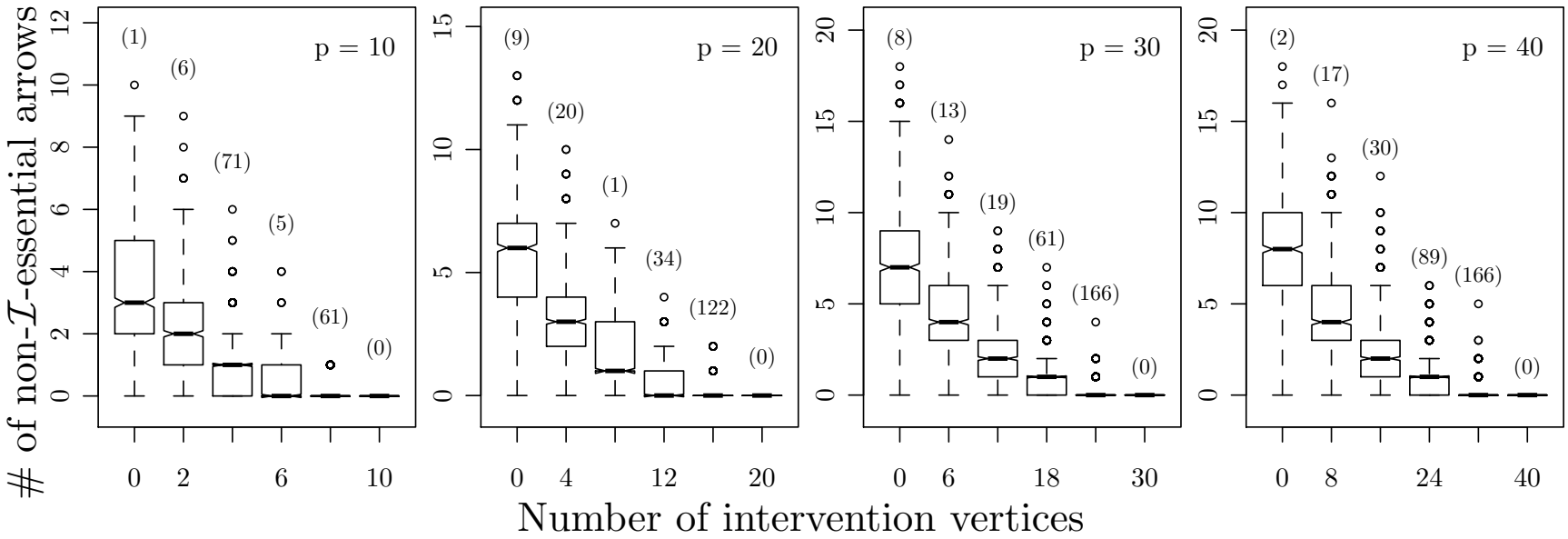
inferring $\mathcal{I}_{\min}$ from available data?

methods for efficient sequential design of intervention experiments

"active learning"

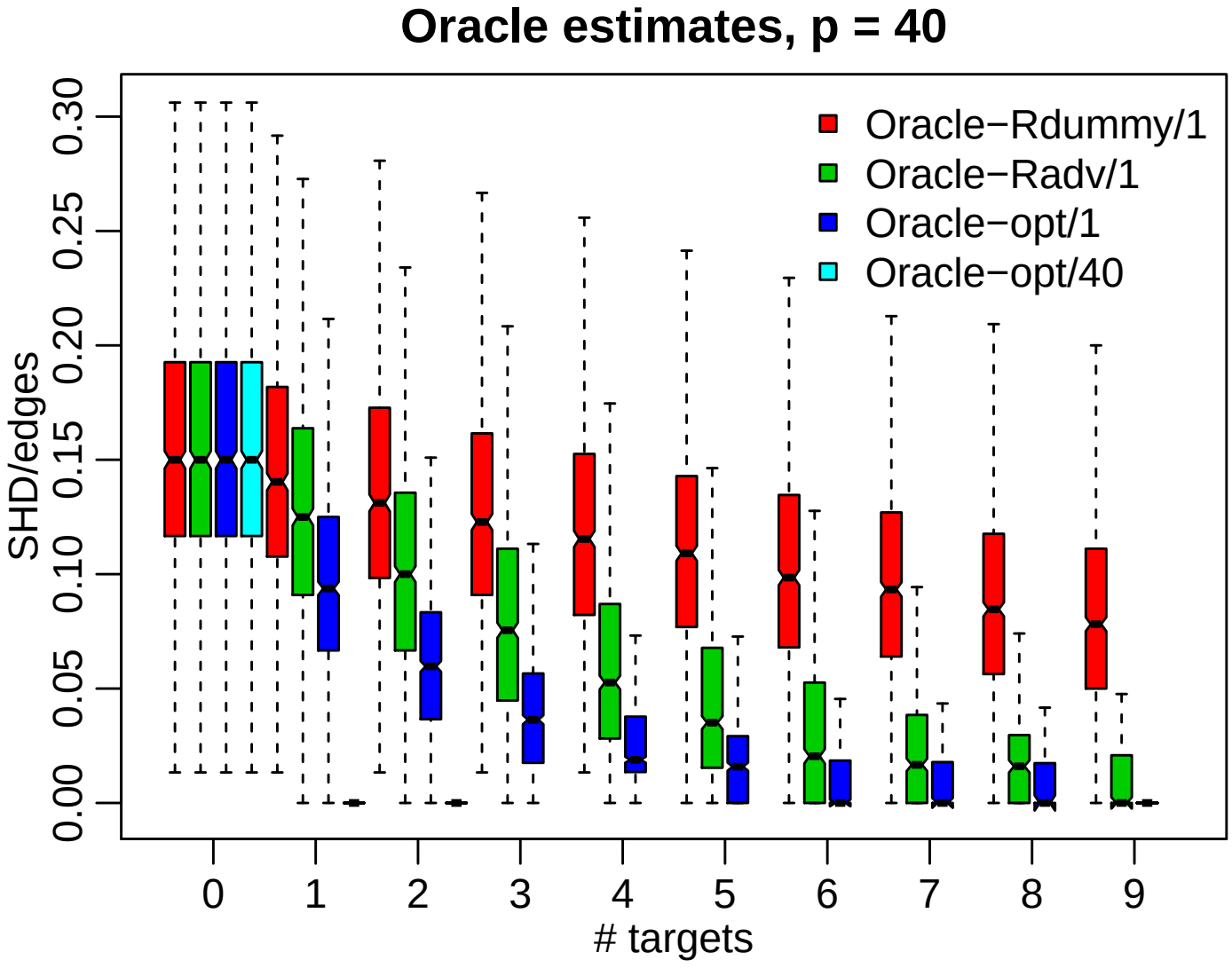
a lot of very recent work in 2019...

randomly chosen intervention variables



a few interventions (randomly placed) lead to substantial gain in identifiability

active learning: cleverly chosen intervention variables
(Eberhardt conjecture, 2008; Hauser & PB, 2012, 2014)



The model and the (penalized) MLE

consider data

$$X_{1,\text{obs}}, \ldots, X_{n_1,\text{obs}}, \quad X_{1,I_1=x_1}, \ldots, X_{n_2,I_{n_2}=x_{n_2}}$$

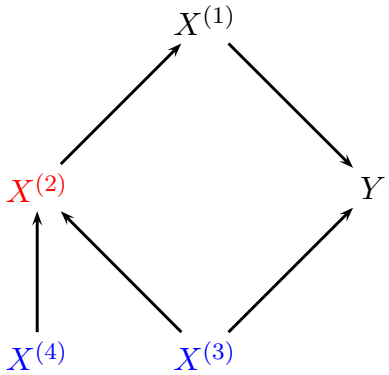
n_1 observational data
n_2 interventional data (single variable interventions)

model:

$X_{1,\text{obs}}, \ldots, X_{n_1,\text{obs}}$ i.i.d. $\sim P_{\text{obs}} = \mathcal{N}_p(0, \Sigma)$ faithful to a DAG D,
$X_{1,I_1}, \ldots, X_{n_2,I_{n_2}}$ independent, non-identically distributed
independent of $X_{1,\text{obs}}, \ldots, X_{n_1,\text{obs}}$
$X_{i,I_i=x_i} \sim P_{\text{int};I_i,x_i}$ linked to the above P_{obs} via do-calculus

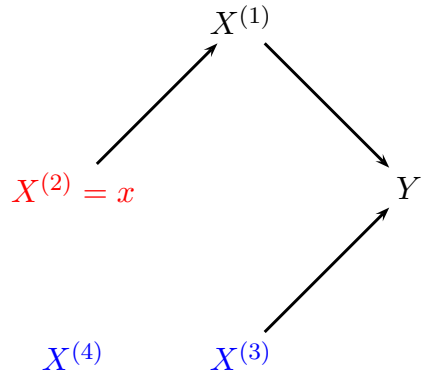
$P_{\text{int};I_i=2,x}$ given by P_{obs} and the DAG D

non-intervention



$$P(Y, X_1, X_2, X_3, X_4) = P(Y|X_1, X_3) \times P(X_1|X_2) \times P(X_2|X_3, X_4) \times P(X_3) \times P(X_4)$$

intervention $\mathrm{do}(X_2 = x)$



$$P(Y, X_1, X_3, X_4|\mathrm{do}(X_2 = x)) = P(Y|X_1, X_3) \times P(X_1|X_2 = x) \times P(X_3) \times P(X_4)$$

can write down the likelihood:

$$\hat{B}, \hat{\Omega} = \text{argmin}_{B,\Omega} - \text{log-likelihood}(B, \Omega; \text{data}) + \lambda \|B\|_0$$

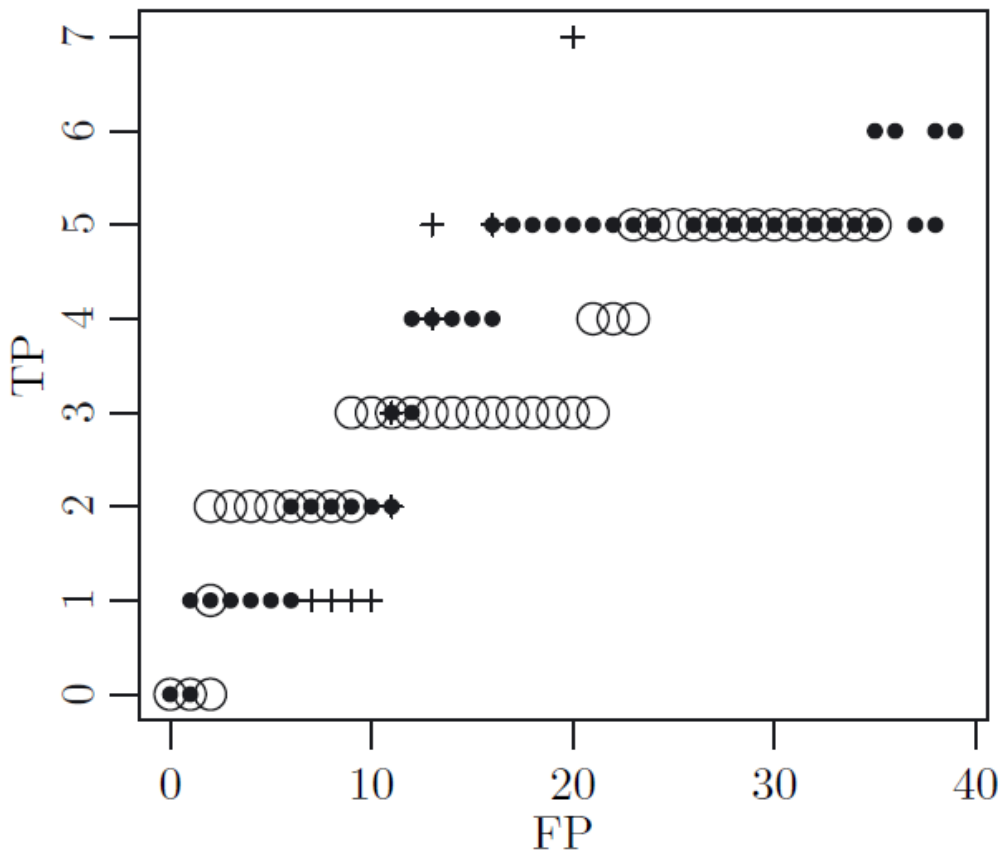
with "argmin" under the constraint that B does not lead to directed cycles

- greedy algorithm: GIES (Greedy Interventional Equivalence Search) Hauser & PB (2012, 2015)
  Wang, Solus, Yang & Uhler (2017)
- consistency of BIC (Hauser & PB, 2015) for fixed p and e.g.:
  - one data point for each intervention with do-value different from observational expectation of the intervention variable
  - no. of observational data points $n_{\text{obs}} \to \infty$

Sachs et al. (2005): flow cytometry data

$p = 11$ proteins and lipids, $n = 5846$ interventional data points
a rough assignment of interventions to single variables is "possible" (but perhaps not very good)

GIES: ◯ (with stability selection) and ● (plain GIES)
the ground-truth is according to Sachs et al. (2005)



conclusion for Sachs et al data: it is hard to see good performance with GIES and a couple of other methods

possible reasons: the interventions are not so specific, there are latent confounders, the linear SEM is heavily misspecified, the data is very noisy, the assumed ground-truth is incorrect

Open problems and conclusions

open problems:

autonomy assumption with do-interventions:
$\mathrm{do}(X_k = x)$ does not change the factors

$$p(x_j | x_{\mathrm{pa}(j)}) \ (j \neq k)$$

probably a bit unrealistic in biology applications!

other interventions which are targeted to specific X-variables (nodes in the graph), for example for jth variable:

$$X_j = \sum_{k \in \mathrm{pa}(j)} B_{jk} X_k + a_j \varepsilon_j$$

noise intervention with factor $a_j > 0$
also here: autonomy assumption that all other structural equations remain the same

environment intervention, for example

$$Y^{(e)} = \sum_{j \in \mathrm{pa}(Y)} B_{Yj} X_j^{(e)} + \varepsilon_Y \text{ for different discrete } e$$

$X^{(e)}$ changing arbitrary over e

see Lecture III
also here: the Y-structural equation has the same parameter B_Y and the same noise distribution ε_Y over all e:

an autonomy assumption

- active learning
  a trade-off between statistical estimation accuracy and identifiability
- in general: statistics for perturbation (e.g. interventional-observational) data
  see Lecture III

conclusions:

- graph-based methods are perhaps not so great for interventional data
  need specific information about interventions – not really the case in biology with “off-target effetcs”
- intervention modeling is still in its infancies
  it is over-shadowed by Pearls excellent and simple do-intervention model
- active learning is interesting and not very well developed
  poor

References

- Ernest, J. and Bühlmann, P. (2015). Marginal integration for nonparametric causal inference. Electronic Journal of Statistics 9, 3155–3194.
- Fan, J., Härdle, W. and Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. Annals of Statistics, 26, 943–971.
- Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Journal of Machine Learning Research 13, 2409-2464.
- Hauser, A. and Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. International Journal of Approximate Reasoning 55, 926–939.
- Hauser, A. and Bühlmann, P. (2015). Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs. Journal of the Royal Statistical Society: Series B 77, 291–318.
- Maathuis, M.H., Colombo, D., Kalisch, M. and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. Nature Methods 7, 247–248.
- Maathuis, M.H., Kalisch, M. and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. Annals of Statistics 37, 3133–3164.
- Pearl, J. (2000). Causality: Models, Reasoning and Inference. Springer.
- Wang, Y., Solus, L., Yang, K.D. and Uhler, C. (2017). Permutation-based Causal Inference Algorithms with Interventions. Advances in Neural Information Processing Systems (NIPS 2017).

Methodological “thinking”

- inferring causal effects from observation data is very ambitious
  (perhaps “feasible in a stable manner” in applications with very large sample size)
- using interventional data is beneficial
  this is what scientists have been doing all the time

$\rightsquigarrow$ the agenda:

- exploit (observational-) interventional/perturbation data
- for unspecific interventions
- in the context of hidden confounding variables (Lecture III)

“my vision”: do it without graph estimation

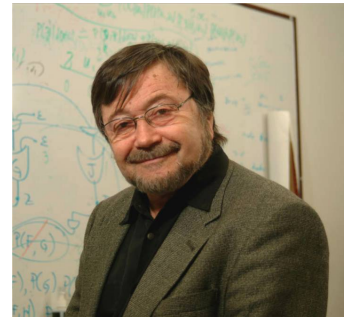
(but use graphs as a language to describe the aims)

Adversarial Robustness

machine learning, Generative Networks

e.g. Ian Goodfellow

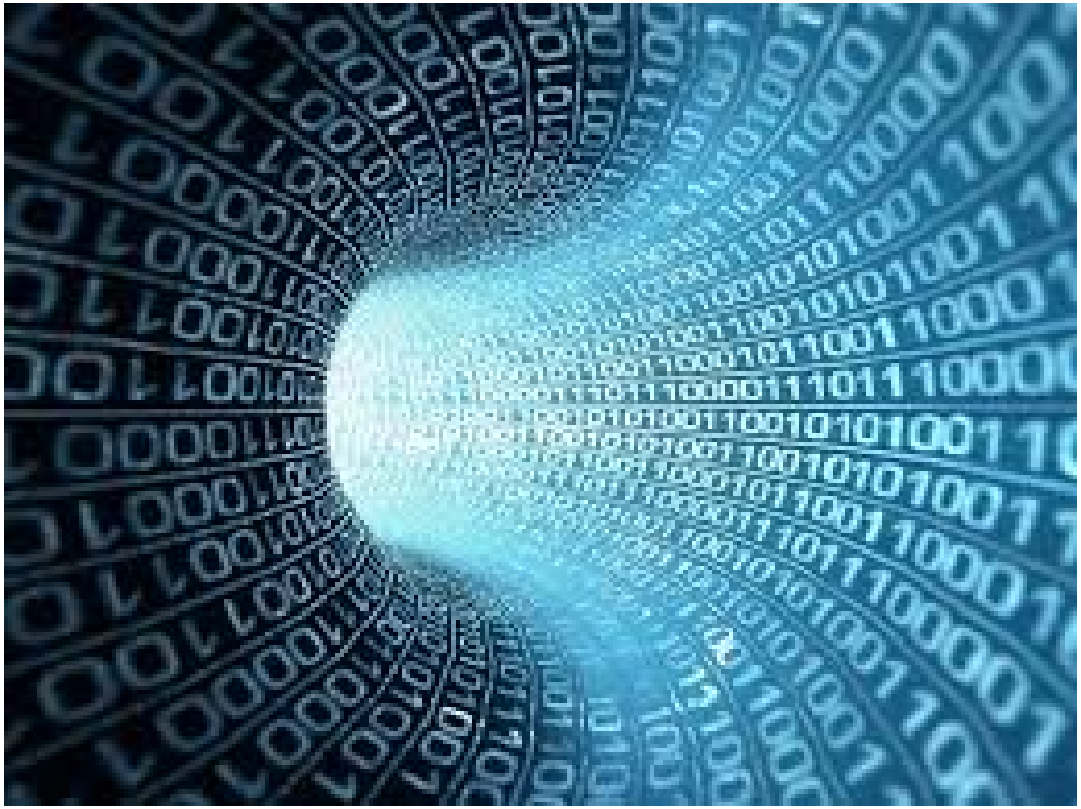
Causality



e.g. Judea Pearl

Do they have something “in common”?

Heterogeneous (potentially large-scale) data



we will take advantage of heterogeneity

often arising with large-scale data where

i.i.d./homogeneity assumption is not appropriate

It's quite a common setting...

data from different known observed environments or experimental conditions or perturbations or sub-populations $e \in \mathcal{E}$:

$$(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$$

with response variables Y^e and predictor variables X^e

examples:

- data from 10 different countries
- data from different econ. scenarios (from diff. "time blocks")

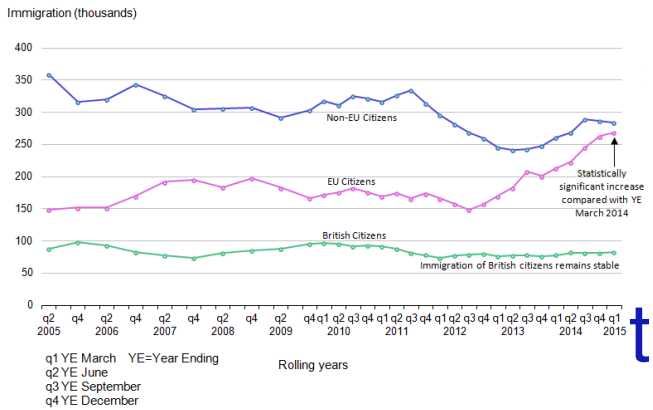
immigration in the UK

consider “many possible” but mostly non-observed environments/perturbations $\mathcal{F} \supset \underbrace{\mathcal{E}}_{\text{observed}}$

examples for $\mathcal{F}$:

- 10 countries and many other than the 10 countries
- scenarios until today and new unseen scenarios in the future

immigration in the UK



the unseen future

problem:

predict Y given X such that the prediction works well (is “robust”) for *“many possible”* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

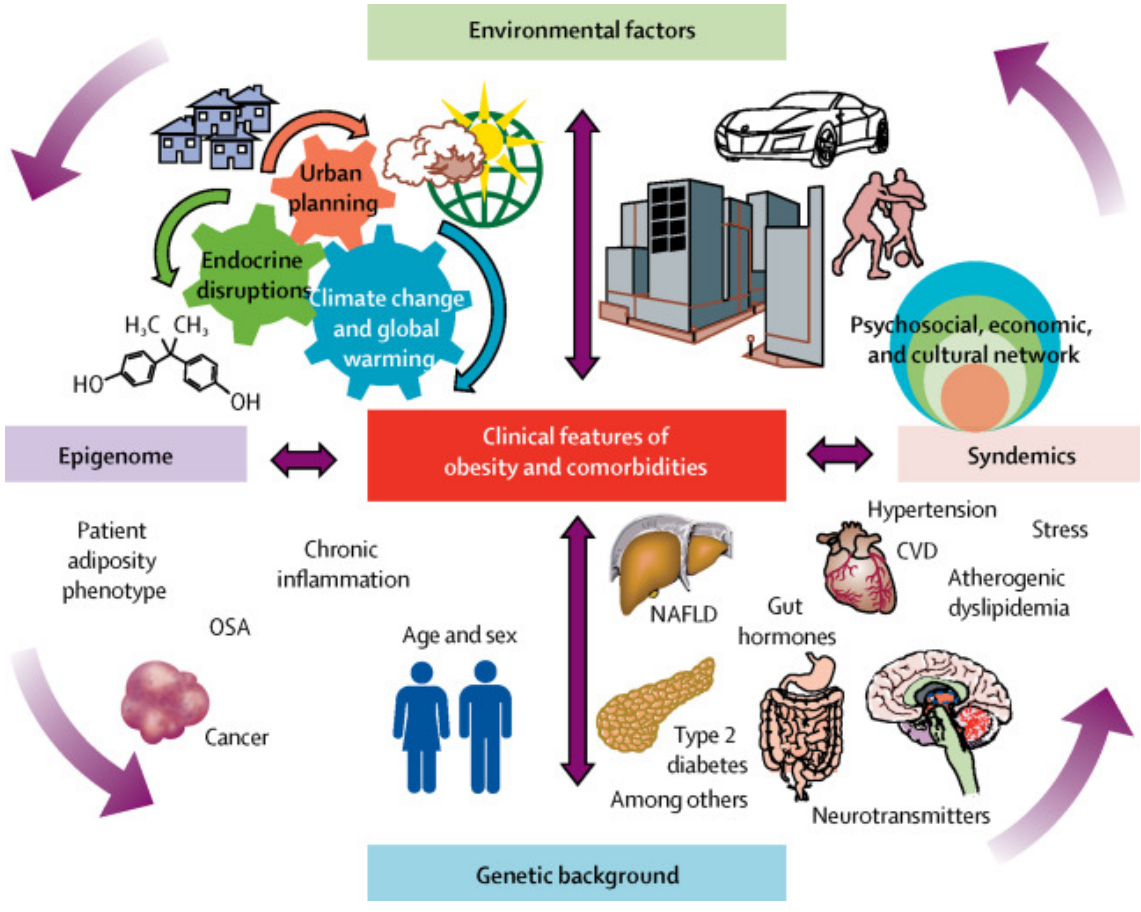
trained on designed, known scenarios from $\mathcal{E}$

trained on designed, known scenarios from $\mathcal{E}$

new scenario from $\mathcal{F}$!

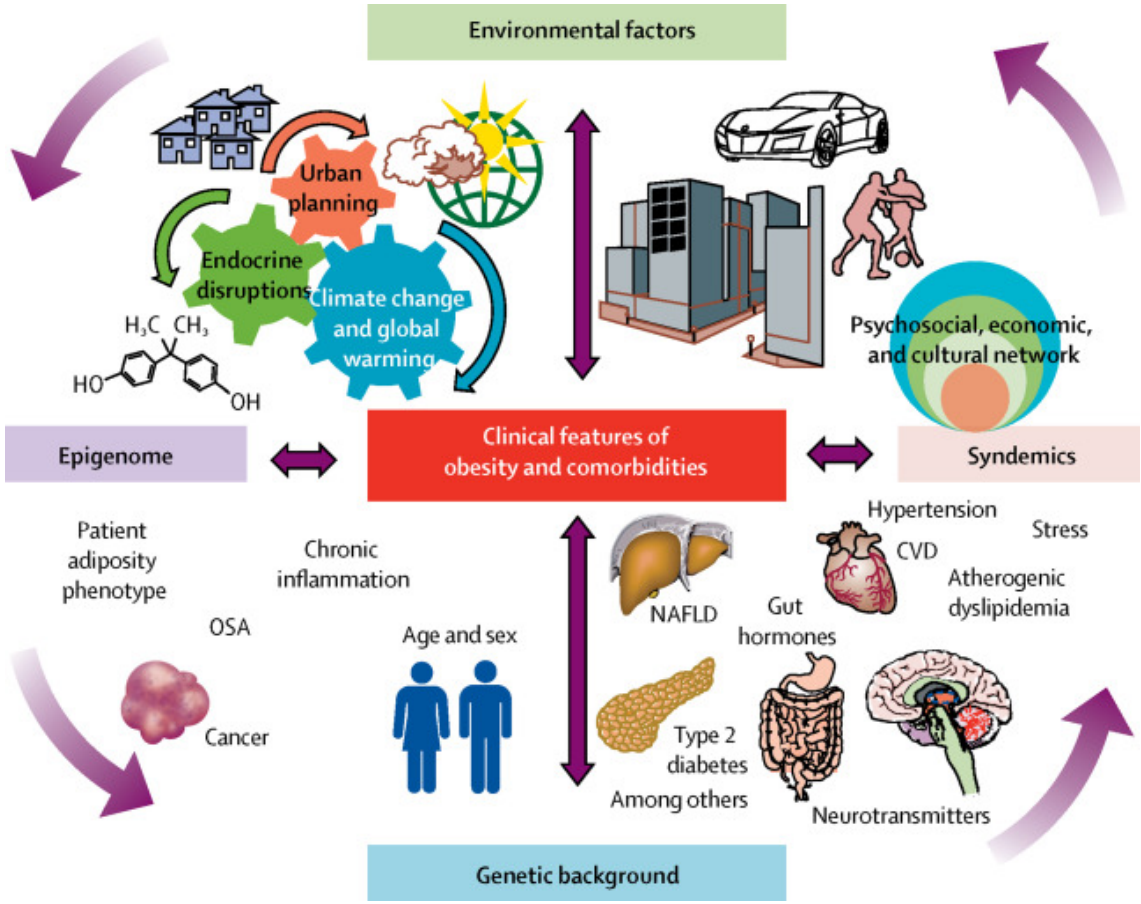
Personalized health

want to be robust across
environmental factors



Personalized health

want to be robust across unseen environmental factors



a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

and remember:

causality is predicting an answer to a

"what if I do/perturb question"!

that is: prediction for new unseen scenarios/environments

a pragmatic prediction problem:

predict Y given X such that the prediction works well (is "robust") for *"many possible"* environments $e \in \mathcal{F}$ based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness"

and remember:

causality is predicting an answer to a

"what if I do/perturb question"!

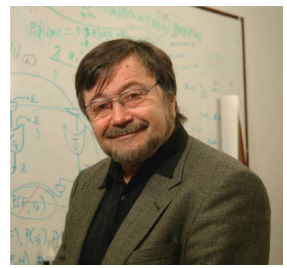
that is: prediction for new unseen scenarios/environments

a pragmatic prediction problem:
predict Y given X such that the prediction works well
(is "robust") for *"many possible"* environments $e \in \mathcal{F}$
based on data from much fewer environments from $\mathcal{E}$

for example with linear models: find

$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

it is "robustness" and also about causality



and remember:
causality is predicting an answer to a
"what if I do/perturb question"!
that is: prediction for new unseen scenarios/environments

Prediction and causality

indeed, for linear models: in a nutshell

$$\text{for } \mathcal{F} = \{\text{all perturbations not acting on } Y \text{ directly}\},$$
$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T\beta|^2 = \text{ causal parameter}$$

that is:
causal parameter optimizes
worst case loss w.r.t. “very many” unseen (“future”) scenarios

later:
we will discuss models for $\mathcal{F}$ and $\mathcal{E}$ which make these relations more precise

Prediction and causality

indeed, for linear models: in a nutshell

$$\text{for } \mathcal{F} = \{\text{all perturbations not acting on } Y \text{ directly}\},$$
$$\text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T\beta|^2 = \text{ causal parameter}$$

that is:
causal parameter optimizes
worst case loss w.r.t. “very many” unseen (“future”) scenarios

later:
we will discuss models for $\mathcal{F}$ and $\mathcal{E}$ which make these relations more precise

How to exploit heterogeneity? for causality or "robust" prediction

Invariant causal prediction (Peters, PB and Meinshausen, 2016)

a main simplifying message:

causal structure/components remain the same for different environments/perturbations

while non-causal components can change across environments

thus:

$\rightsquigarrow$ look for "stability" of structures among different environments

How to exploit heterogeneity? for causality or "robust" prediction

Invariant causal prediction (Peters, PB and Meinshausen, 2016)

a main simplifying message:

causal structure/components remain the same for different environments/perturbations

while non-causal components can change across environments

thus:
$\rightsquigarrow$ look for "stability" of structures among different environments

Invariance: a key conceptual assumption

Invariance Assumption (w.r.t. $\mathcal{E}$)

there exists $S^* \subseteq \{1, \ldots, d\}$ such that:

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is invariant across } e \in \mathcal{E}$$

for linear model setting:
there exists a vector γ^* with $\text{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\forall e \in \mathcal{E}: \quad Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*}$$

$\varepsilon^e \sim F_\varepsilon$ the same for all e

X^e has an arbitrary distribution, different across e

$\gamma^*,\ S^*$ is interesting in its own right!

namely the parameter and structure which remain invariant across experimental settings, or heterogeneous groups

Invariance: a key conceptual assumption

Invariance Assumption (w.r.t. $\mathcal{E}$)

there exists $S^* \subseteq \{1, \ldots, d\}$ such that:

$$\mathcal{L}(Y^e | X^e_{S^*}) \text{ is invariant across } e \in \mathcal{E}$$

for linear model setting:
there exists a vector γ^* with $\text{supp}(\gamma^*) = S^* = \{j;\ \gamma^*_j \neq 0\}$
such that:

$$\forall e \in \mathcal{E}: \quad Y^e = X^e \gamma^* + \varepsilon^e,\ \varepsilon^e \perp X^e_{S^*}$$

$\varepsilon^e \sim F_\varepsilon$ the same for all e

X^e has an arbitrary distribution, different across e

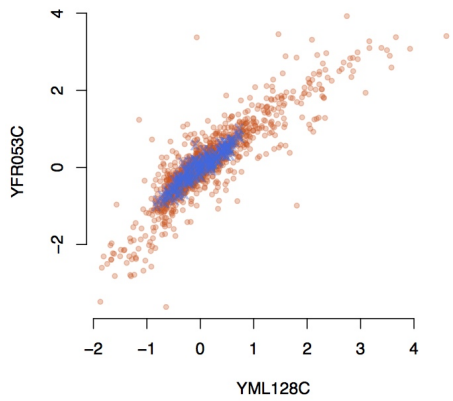
$\gamma^*,\ S^*$ is interesting in its own right!

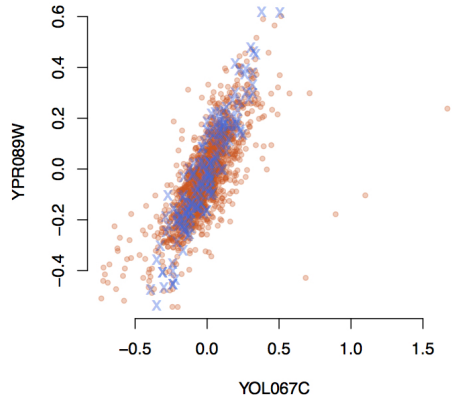
namely the parameter and structure which remain invariant across experimental settings, or heterogeneous groups

Invariance Assumption: plausible to hold with real data

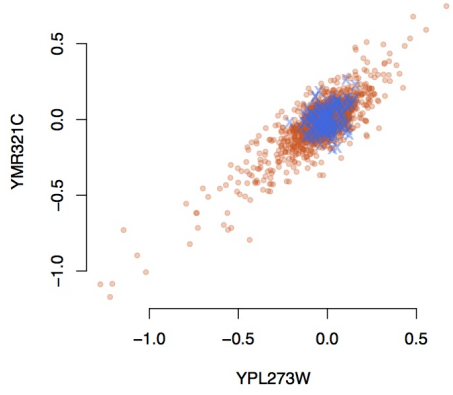
two-dimensional conditional distributions of observational (blue) and interventional (orange) data
(no intervention at displayed variables X, Y)

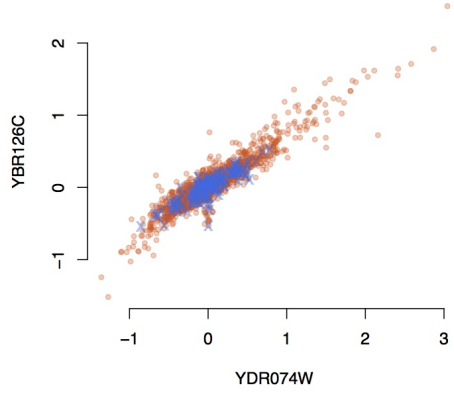
seemingly
no invariance
of conditional d.





plausible
invariance
of conditional d.





Invariance Assumption w.r.t. $\mathcal{F}$

$$\text{where } \mathcal{F} \underbrace{\supset}_{\text{much larger}} \mathcal{E}$$

now: the set S^* and corresponding regression parameter γ^* are for a much larger class of environments than what we observe!

$\rightsquigarrow$

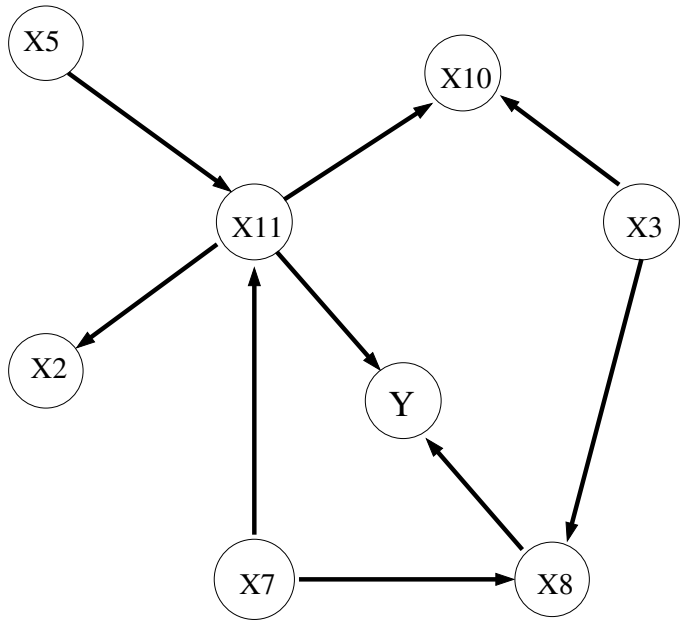
$$\gamma^*,\ S^* \text{ is even more interesting in its own right!}$$

since it says something about unseen new environments!

Link to causality

mathematical formulation with structural equation models:

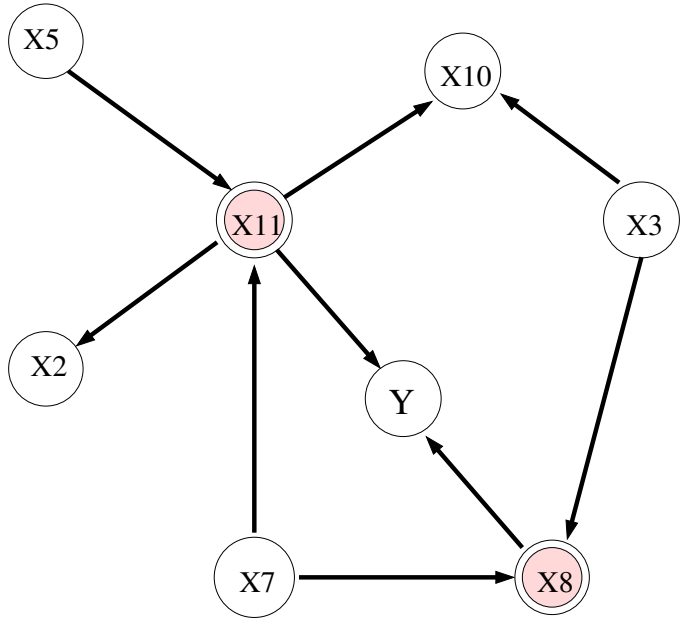
$$Y \leftarrow f(X_{\mathrm{pa}(Y)}, \varepsilon),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j) \ (j = 1, \ldots, p)$$
$$\varepsilon, \varepsilon_1, \ldots, \varepsilon_p \text{ independent}$$



Link to causality

mathematical formulation with structural equation models:

$$Y \leftarrow f(X_{\mathrm{pa}(Y)}, \varepsilon),$$
$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j) \ (j = 1, \ldots, p)$$
$$\varepsilon, \varepsilon_1, \ldots, \varepsilon_p \text{ independent}$$



(direct) causal variables for Y: the parental variables of Y

Link to causality

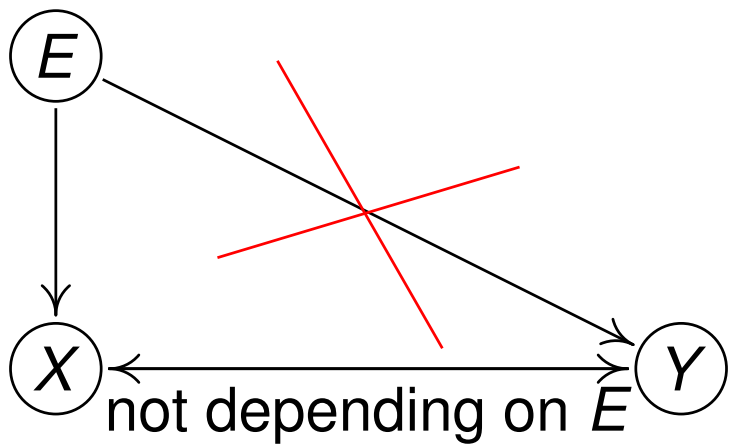
problem:
under what model for the environments/perturbations e can we have an interesting description of the invariant sets S^*?

loosely speaking: assume that the perturbations e

- do not act directly on Y
- do not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

graphical description: E is random with realizations e



Link to causality

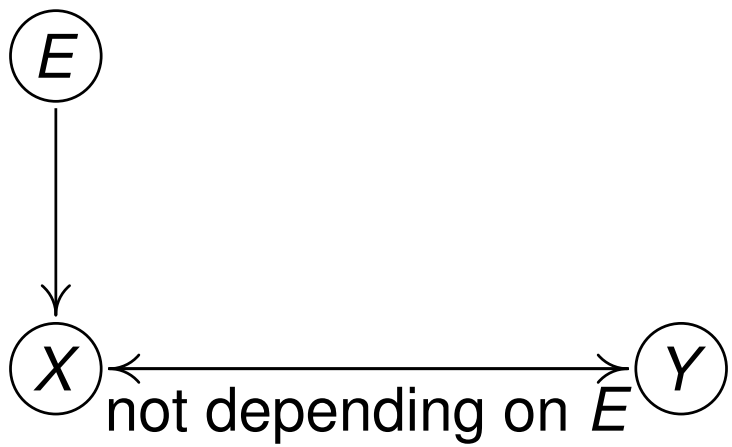
problem:
under what model for the environments/perturbations e can we have an interesting description of the invariant sets S^*?

loosely speaking: assume that the perturbations e

- do not act directly on Y
- do not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

graphical description: E is random with realizations e



Link to causality

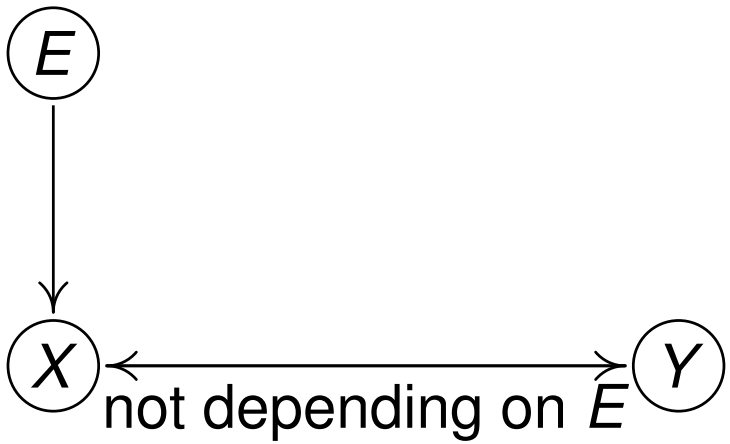
problem:
under what model for the environments/perturbations e can we have an interesting description of the invariant sets S^*?

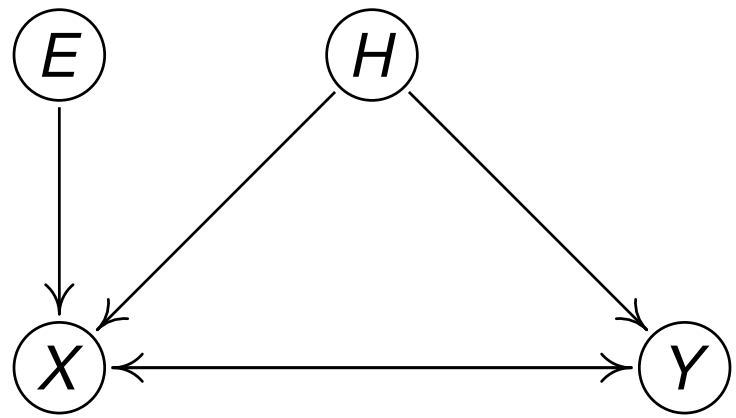
loosely speaking: assume that the perturbations e

- do not act directly on Y
- do not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

graphical description: E is random with realizations e





IV model: see Lecture III

Link to causality

easy to derive the following:

Proposition

- structural equation model for (Y, X);
- model for $\mathcal{F}$ of perturbations: every $e \in \mathcal{F}$
  - does not act directly on Y
  - does not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Then: the causal variables $\mathrm{pa}(Y)$ satisfy the invariance assumption with respect to $\mathcal{F}$

causal variables lead to invariance under arbitrarily strong perturbations from $\mathcal{F}$ as described above

Proposition

- structural equation model for (Y, X);
- model for $\mathcal{F}$ of perturbations: every $e \in \mathcal{F}$
  - does not act directly on Y
  - does not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Then: the causal variables $\mathrm{pa}(Y)$ satisfy the invariance assumption with respect to $\mathcal{F}$

as a consequence: for linear structural equation models

for $\mathcal{F}$ as above,

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2 = \underbrace{\beta^0_{\mathrm{pa}(Y)}}_{\text{causal parameter}}$$

if the perturbations in $\mathcal{F}$ would not be arbitrarily strong

$\rightsquigarrow$ the worst-case optimizer is different! (see later)

Proposition

- structural equation model for (Y, X);
- model for $\mathcal{F}$ of perturbations: every $e \in \mathcal{F}$
  - does not act directly on Y
  - does not change the relation between X and Y

but may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Then: the causal variables $\mathrm{pa}(Y)$ satisfy the invariance assumption with respect to $\mathcal{F}$

as a consequence: for linear structural equation models

for $\mathcal{F}$ as above,

$$\mathrm{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2 = \underbrace{\beta^0_{\mathrm{pa}(Y)}}_{\text{causal parameter}}$$

if the perturbations in $\mathcal{F}$ would not be arbitrarily strong
$\rightsquigarrow$ the worst-case optimizer is different! (see later)

A real-world example and the assumptions

Y: growth rate of the plant
X: high-dim. covariates of gene expressions

perturbations e: different gene knock-out experiments
$\rightsquigarrow$ e changes the expressions of some components of X

it's plausible that perturbations e

- do not directly act on Y $\surd$
- do not change the relation between X and Y ?

may act arbitrarily on X (arbitrary shifts, scalings, etc.)

A real-world example and the assumptions

Y: growth rate of the plant
X: high-dim. covariates of gene expressions

perturbations e: different gene knock-out experiments
$\rightsquigarrow$ e changes the expressions of some components of X

it's plausible that perturbations e

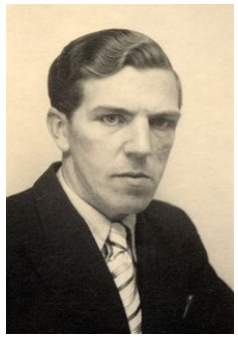
- do not directly act on Y $\surd$
- do not change the relation between X and Y ?

may act arbitrarily on X (arbitrary shifts, scalings, etc.)

Causality $\Longleftrightarrow$ Invariance

we just argued: causal variables $\Longrightarrow$ invariance

known since a long time:
Haavelmo (1943)



Trygve Haavelmo

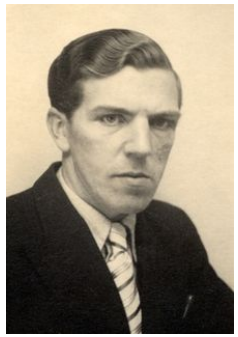
Nobel Prize in Economics 1989

(...; Goldberger, 1964; Aldrich, 1989;... ; Dawid and Didelez, 2010)

Causality $\Longleftrightarrow$ Invariance

we just argued: causal variables $\Longrightarrow$ invariance

known since a long time:
Haavelmo (1943)



Trygve Haavelmo

Nobel Prize in Economics 1989

(...; Goldberger, 1964; Aldrich, 1989;... ; Dawid and Didelez, 2010)

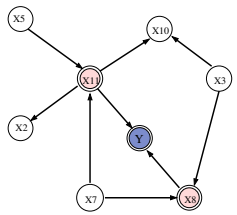
more novel: the reverse relation

causal structure, predictive robustness $\Longleftarrow$ invariance

(Peters, PB & Meinshausen, 2016)

The search for invariance and causality (Peters, PB & Meinshausen, 2016)

causal structure/variables $\Longleftarrow$ invariance



severe issues of identifiability !

can perform statistical test whether a subset S of covariates satisfies the invariance assumption

$H_0 - \mathrm{InvA}(\mathcal{E})$: $\mathcal{L}(Y^e | X_S^e)$ is invariant across $e \in \underbrace{\mathcal{E}}_{\text{observed environments}}$

in a linear model $\rightsquigarrow$ Chow (1960)

$\rightsquigarrow$ sets $S_1, \ldots, S_k$ which are statistically compatible with invariance assumption $H_0 - \mathrm{InvA}(\mathcal{E})$

making it identifiable:

$$\hat{S}(\mathcal{E}) = \bigcap \{S;\ S \underbrace{\text{ statistically compatible with } H_0 - \text{InvA}(\mathcal{E})\}}_{\text{no rejection at significance level } \alpha}$$

Theorem: (Peters, PB and Meinshausen, 2016)
assume structural equation model

- linear model for Y versus X, Gaussian errors
- $e \in \mathcal{E}$ does not act directly on Y and
  does not change the relation between X and Y

Then:

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \underbrace{S_{\text{causal}}}_{\text{pa}(Y)}] \geq 1 - \alpha$$

confidence guarantee against false positive causal selection

making it identifiable:

$$\hat{S}(\mathcal{E}) = \bigcap \{S;\ S \underbrace{\text{ statistically compatible with } H_0 - \text{InvA}(\mathcal{E})}_{\text{no rejection at significance level } \alpha}\}$$

Theorem: (Peters, PB and Meinshausen, 2016)
assume structural equation model

- linear model for Y versus X, Gaussian errors
- $e \in \mathcal{E}$ does not act directly on Y and
  does not change the relation between X and Y

Then:

$$\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq \underbrace{S_{\text{causal}}}_{\text{pa}(Y)}] \geq 1 - \alpha$$

confidence guarantee against false positive causal selection

ICP = Invariant Causal Prediction

Proof: the causal set S_{causal} leads to invariance

$$\begin{aligned}
\mathbb{P}[\hat{S}(\mathcal{E}) \subseteq S_{\text{causal}}] &= \mathbb{P}[\bigcap \{S; H_{0,S} \text{ not rejected}\} \subseteq S_{\text{causal}}] \\
&\geq \mathbb{P}[H_{0,S_{\text{causal}}} \text{ not rejected}] \geq 1 - \alpha
\end{aligned}$$

□

Conclusions

- causality can be framed as worst case risk optimization! more on that in Lecture IV
- causality can be inferred from invariance and a “stability” argument
- ICP (Invariant Causal Prediction) is a conceptual approach and method

make heterogeneity or non-stationarity your friend

(rather than your enemy)!

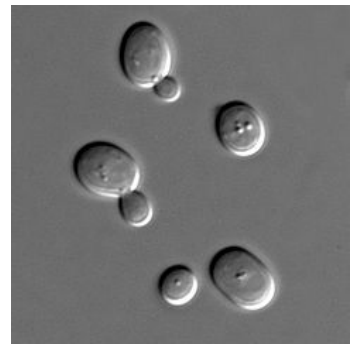
make heterogeneity or non-stationarity your friend

(rather than your enemy)!

References

- Bühlmann, P. (2018). Invariance, Causality and Robustness. To appear in Statistical Science. Preprint arXiv:1812.08233
- Meinshausen, N., Hauser, A., Mooij, J.M., Peters, J., Versteeg, P. and Bühlmann, P. (2016). Methods for causal inference from gene perturbation experiments and validation. Proceedings of the National Academy of Sciences USA 113, 7361-7368.
- Peters, J., Bühlmann, P. and Meinshausen, N. (2016). Causal inference using invariant prediction: identification and confidence intervals (with discussion). Journal of the Royal Statistical Society, Series B 78, 947-1012.
- Pfister, N., Bühlmann, P. and Peters, J. (2018). Invariant causal prediction for sequential data. Journal of the American Statistical Association, published online DOI 10.1080/01621459.2018.1491403.

Single gene deletion experiments in yeast



$d = 6170$ genes
response of interest: $Y =$ expression of first gene
"covariates" $X =$ gene expressions from all other genes

and then
response of interest: $Y =$ expression of second gene
"covariates" $X =$ gene expressions from all other genes

and so on

infer/predict the effects of unseen/new single gene deletions on all other genes

Kemmeren et al. (2014):
genome-wide mRNA expressions in yeast: $d = 6170$ genes

- $n_{obs} = 160$ "observational" samples of wild-types
- $n_{int} = 1479$ "interventional" samples
  each of them corresponds to a single gene deletion strain

for our method: we use $|\mathcal{E}| = 2$
(observational and interventional data)

training-test data splitting:

- training set: all observational and 2/3 of interventional data
- test set: other 1/3 of gene deletion interventions
  $\rightsquigarrow$ can validate predicted effects of these interventions
- repeat this for the three blocks of interventional test data

multiplicity adjustment:

since ICP is used 6170 times (once for every response var.) we use coverage $1 - \alpha/6170$ with $\alpha = 0.05$

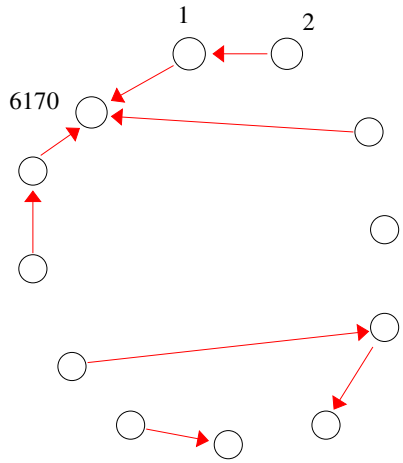
Results for inferring causal variables on a single training-test split

8 genes are "significant" ($\alpha = 0.05$ level) causal variables
(each of the 8 genes "causes" one other gene)

Results for inferring causal variables on a single training-test split

8 genes are "significant" ($\alpha = 0.05$ level) causal variables
(each of the 8 genes "causes" one other gene)

not many findings...



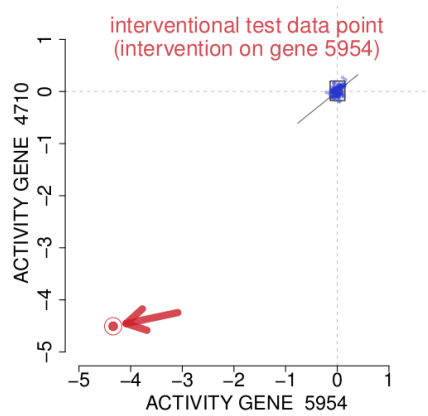
but we use a stringent criterion with Bonferroni corrected
$\alpha/6170 = 0.05/6170$ to control the familywise error rate

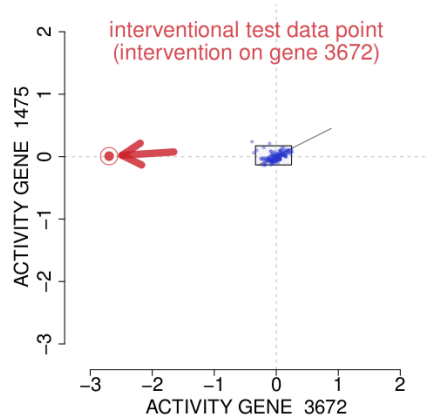
8 genes are “significant” ($\alpha = 0.05$ level) causal variables

validation:
thanks to the intervention experiments (in the test data) we can validate the method(s)

we only consider true Strong Intervention Effects (SIEs)

SIE = the observed response value associated to an intervention is in the 1%- or 99% tail of the observational data





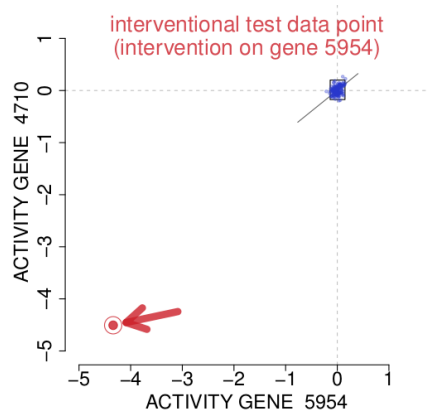
8 genes are “significant” ($\alpha = 0.05$ level) causal variables

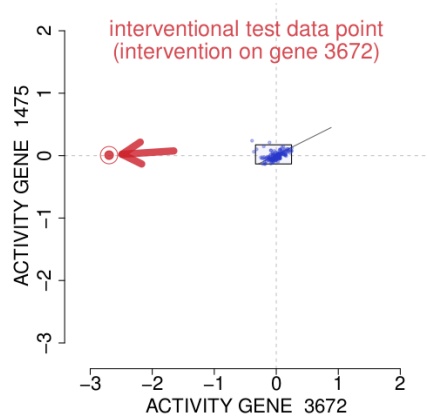
validation:
thanks to the intervention experiments (in the test data) we can validate the method(s)

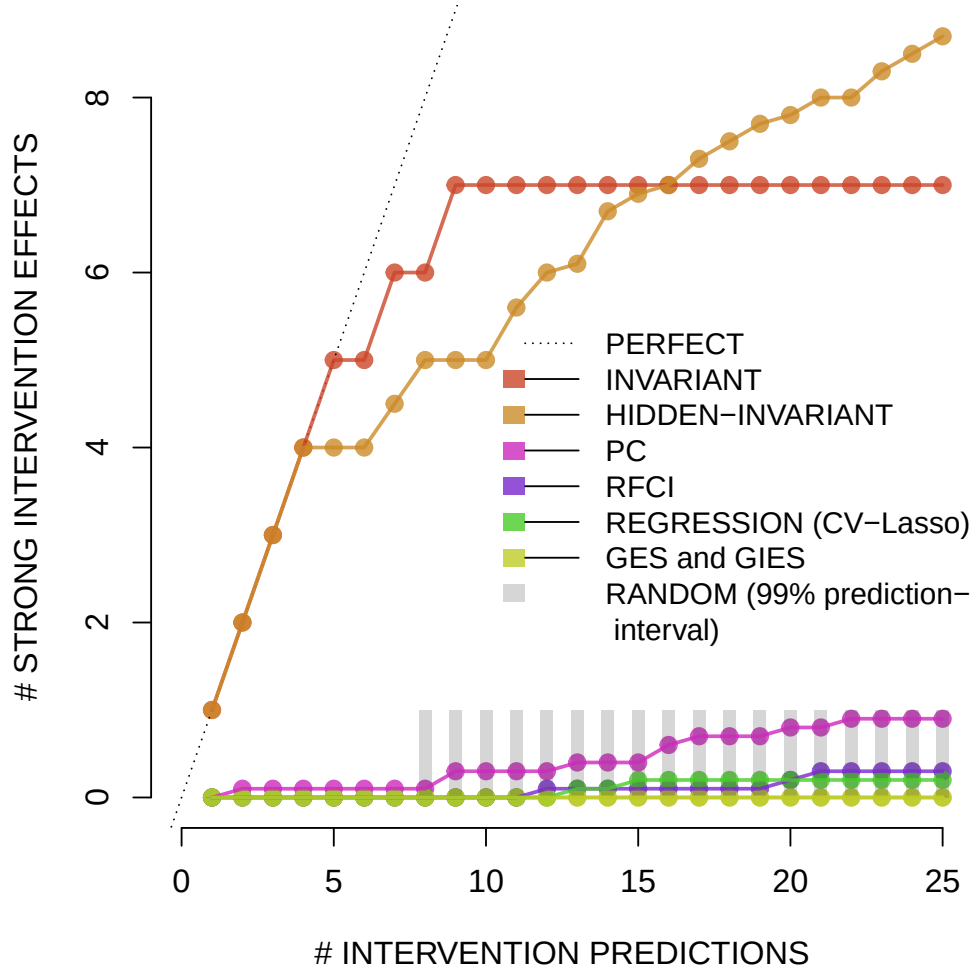
we only consider true Strong Intervention Effects (SIEs)

6 out of the 8 “significant” genes are true SIEs!

SIE = the observed response value associated to an intervention is in the 1%- or 99% tail of the observational data







I : invariant prediction method

H: invariant prediction with some hidden variables