

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

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)

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

$|\beta_j|$ 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**

↪ 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)

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

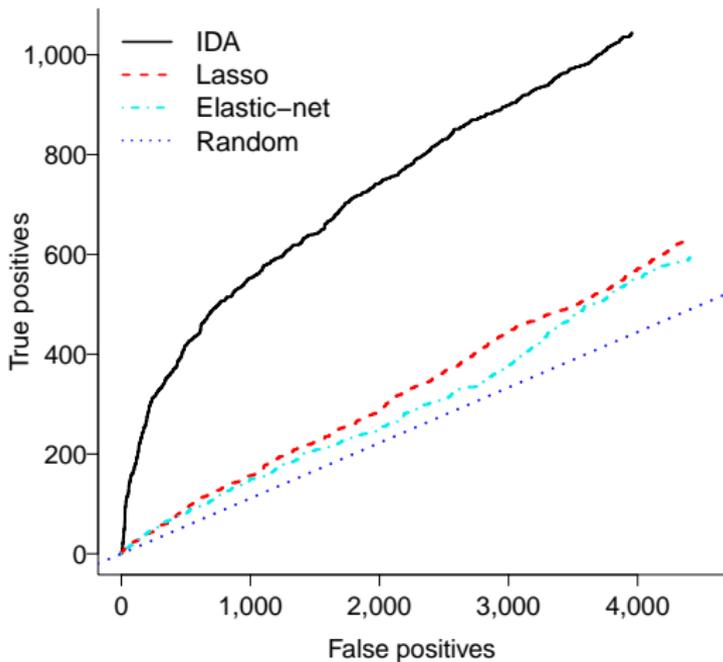
$|\beta_j|$ measures the effect of variable X_j in terms of “association”

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

↪ not very realistic for intervention problem

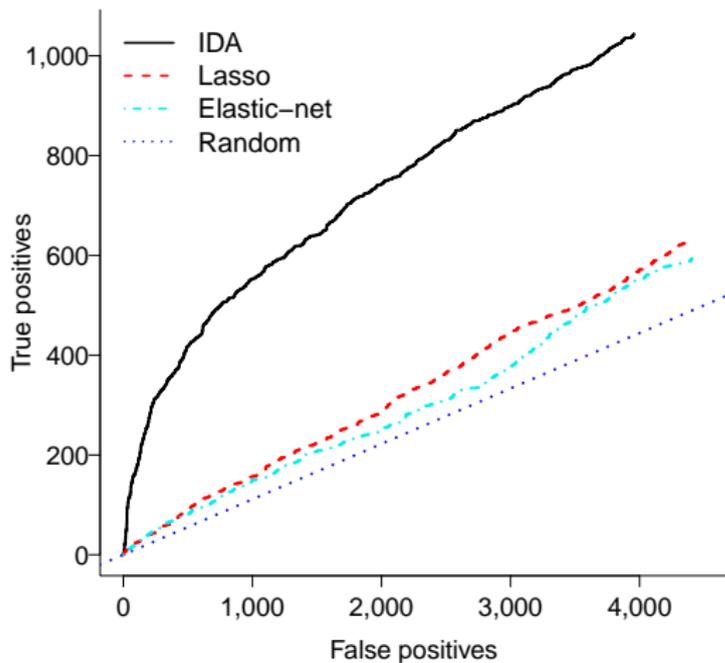
if we change e.g. one gene, some others will also change and these others are not (cannot be) kept fixed

and indeed:



↪ can do much better than (penalized) regression!

and indeed:



~> can do much better than (penalized) regression!

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

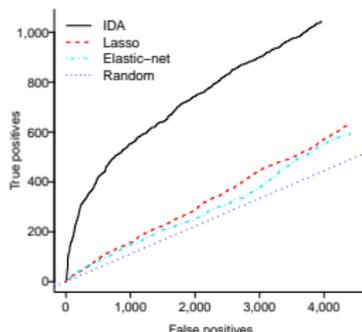
(Maathuis, Colombo, Kalisch & PB, 2010)

- $p = 5360$ genes (expression of genes)
- 231 gene knock downs $\leadsto 1.2 \cdot 10^6$ 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 j th component of X to a certain value x :

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

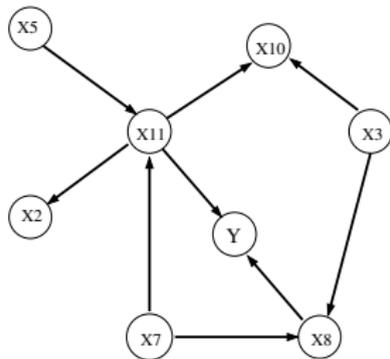
↪ this is a question of **intervention type**

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”:

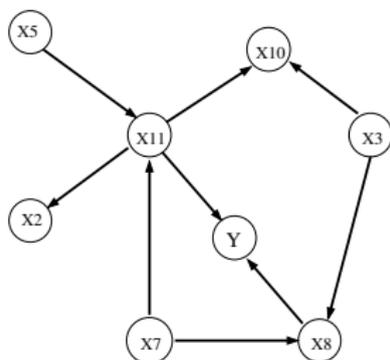
if we intervene at X_j , its effect propagates through other variables X_k ($k \neq j$) to Y



Graphs, structural equation models and causality

intuitively:

the concept of causality in terms of graphs is plausible



in a DAG:

a directed arrow $X \rightarrow Y$ 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

$\leadsto D$ 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):

$$X_j \leftarrow f_j(X_{\text{pa}(j)}, \varepsilon_j), \quad j = 1, \dots, p$$

$\varepsilon_1, \dots, \varepsilon_p$ independent

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

Linear SEM

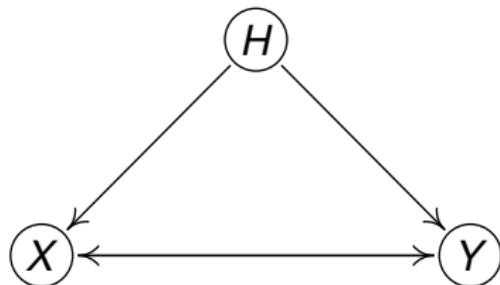
linear structural equation model (with structure D):

$$X_j \leftarrow \sum_{k \in \text{pa}(j)} B_{jk} X_k + \varepsilon_j, \quad j = 1, \dots, p$$

$\varepsilon_1, \dots, \varepsilon_p$ independent

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

so far: no hidden “confounding” variables



~> see Lecture IV

Local Markov property

Given P with density p from a SEM
because of independence of $\varepsilon_Y, \varepsilon_1, \dots, \varepsilon_p$
 \leadsto 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, \dots, X_p

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

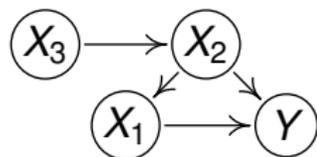
and this is the **core task of causality**

Modeling interventions: do-interventions

Pearl's do-interventions



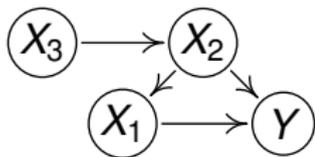
Judea Pearl



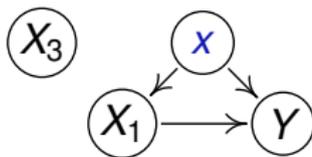
Pearl's do-interventions



Judea Pearl



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



$$X_1 \leftarrow f_1(X_2 = x, \varepsilon_1),$$

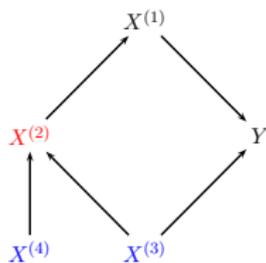
$$X_2 \leftarrow x,$$

$$X_3 \leftarrow \varepsilon_3$$

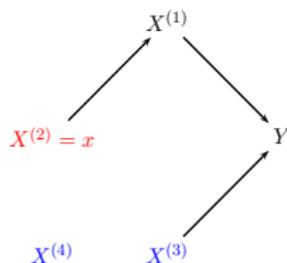
$$Y \leftarrow f_Y(X_1, X_2 = x, \varepsilon_Y)$$

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

non-intervention



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



$$\begin{aligned} p(Y, X_1, X_2, X_3, X_4) = & \\ & p(Y|X_1, X_3) \times \\ & p(X_1|X_2) \times \\ & p(X_2|X_3, X_4) \times \\ & p(X_3) \times \\ & p(X_4) \end{aligned}$$

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

truncated factorization

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

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

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

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

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

this is an assumption!

and is called **structural autonomous assumption**

the intervention distribution $P(Y|\text{do}(X_2 = x))$ 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 $\text{do}(X_j = x)$, for every j and x , we obtain a different distribution of Y, X_1, \dots, X_p

can generate many **interventional** distributions!

Potential outcome model

Neyman (1923), Rubin (1974)

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

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

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

~> 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 $P(Y|\text{do}(X_j = x))$ or

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

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

the total causal effect is defined as

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

measuring the “total causal importance” of variable X_j on Y

if we know the entire SEM, we can easily simulate the distribution $P(Y|\text{do}(X_j = x))$

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 $P(Y|\text{do}(X_j = x))$ or

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

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

the total causal effect is defined as

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

measuring the “total causal importance” of variable X_j on Y

if we **know the entire** SEM, we can easily simulate the distribution $P(Y|\text{do}(X_j = x))$

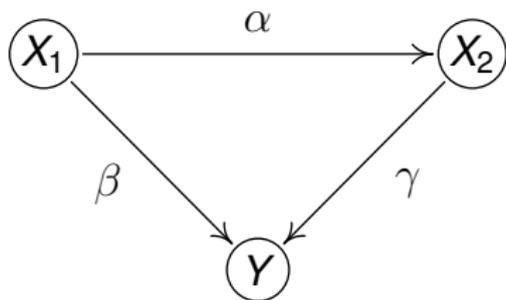
this approach requires **global knowledge** of the graph structure, edge functions/weights and error distributions

Example: linear SEM

directed path p_j from X_j to Y

causal effect on p_j by product of corresponding edge weights

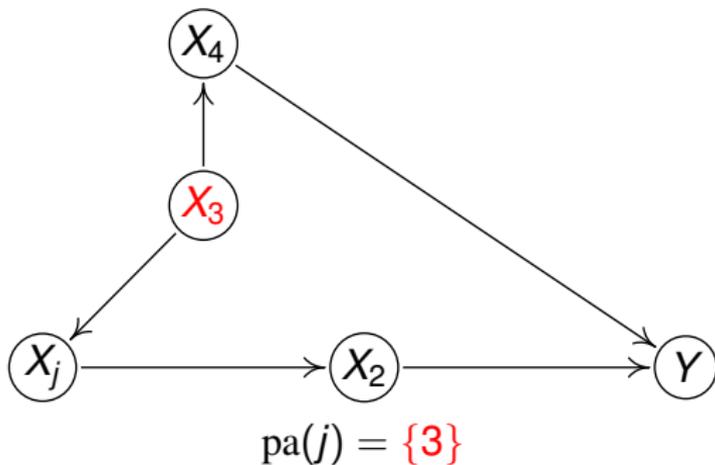
total causal effect = $\sum p_j \gamma_j$



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 $S = \text{pa}(j)$



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

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

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

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

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.

$S = \text{pa}(j)$

often much easier to obtain/estimate than the entire graph

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

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

Henckel, Perkovic & Maathuis (2019) discuss efficiency for total causal effect estimation

with or without backdoor adjustment, possibly with a set $S \neq \text{pa}(j)$, when the graph is known/given

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

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

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

estimation of the right-hand side has been developed for additive models!

cf. **Fan, Härdle & Mammen (1998)**

additive regression model:

$$Y = \mu + \sum_{j=1}^d f_j(X_j) + \varepsilon,$$

$$\mathbb{E}[f_j(X_j)] = 0 \text{ (for identifiability)}$$

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

asympt. 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

~> instead, we rely on smoothness 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 $x_{pa(j)}$)
one actual implementation is with boosting kernel estimation
(Ernest & PB, 2015)

Gene expressions in *Arabidopsis 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)

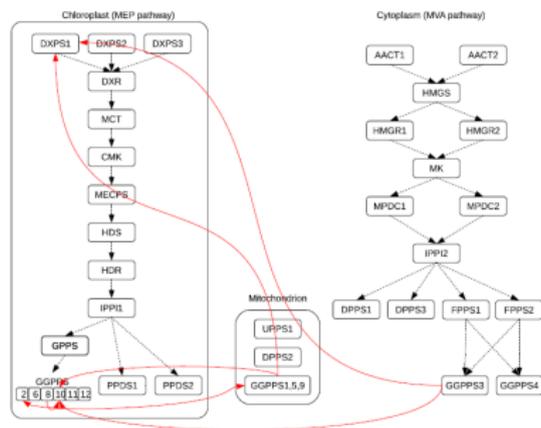


FIG 13. Stable edges (with stability selection) for the *Arabidopsis thaliana* dataset. The dotted arcs represent the metabolic network, the red solid arcs the stable total causal effects found by the est S-mint method.

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**)

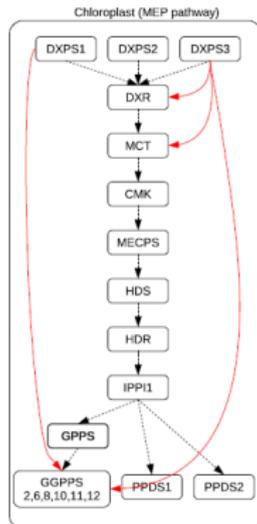


FIG 14. Stable edges (with stability selection) for the MEP pathway in the *Arabidopsis thaliana* dataset. The dotted arcs represent the metabolic network whereas the red solid arcs denote the top ranked causal effects found by S-mint with adjustment sets chosen from the order of the metabolic network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

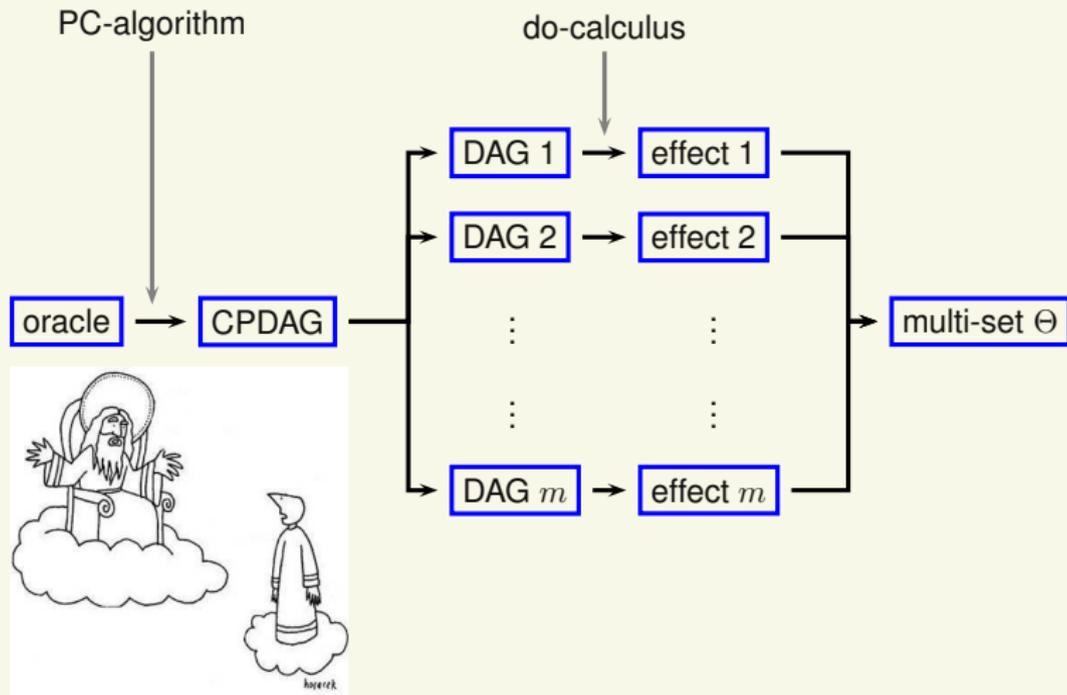
Lower bounds of total causal effects

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 (oracle version)



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

instead of the multi-set

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

minimal absolute value

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

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

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

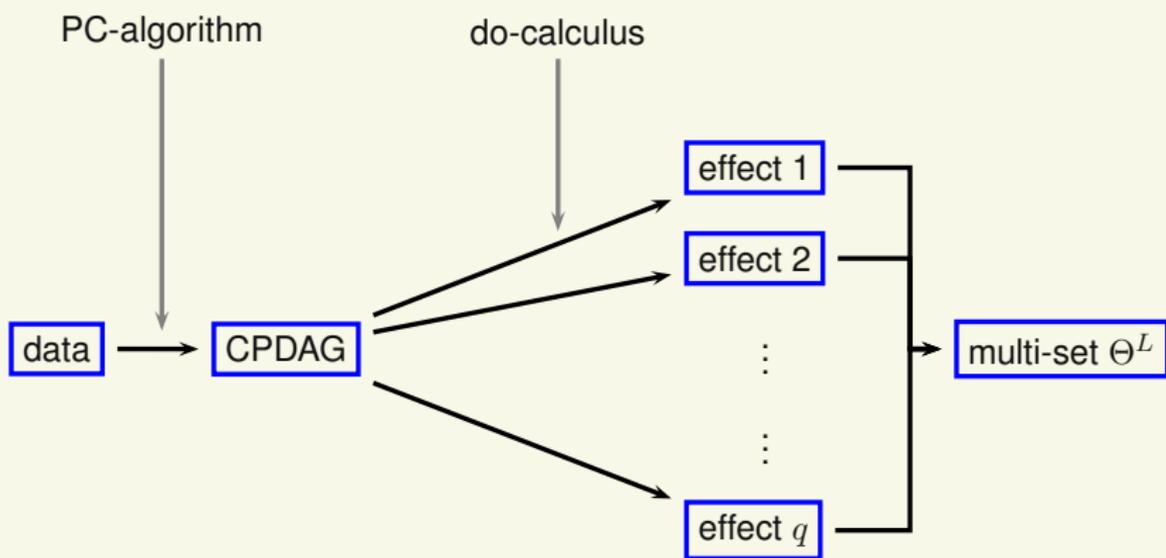
minimal absolute effect α_j is a **lower bound for true absolute intervention effect**

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)

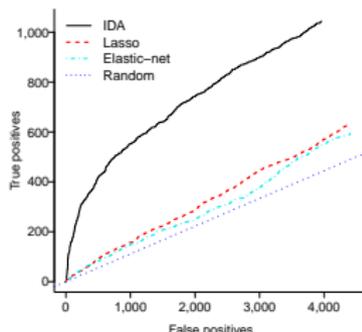
(Maathuis, Colombo, Kalisch & PB, 2010)

- $p = 5360$ genes (expression of genes)
- 231 gene knock downs $\leadsto 1.2 \cdot 10^6$ 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

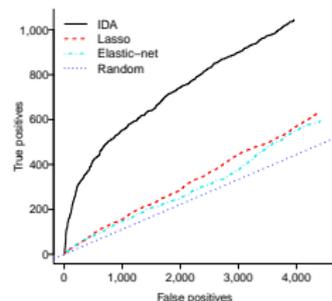


Interventions and active learning

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

~> 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:

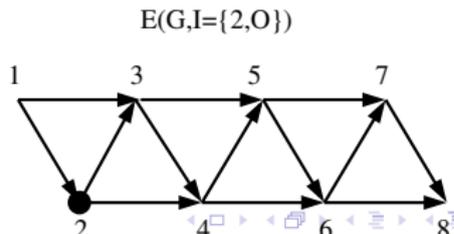
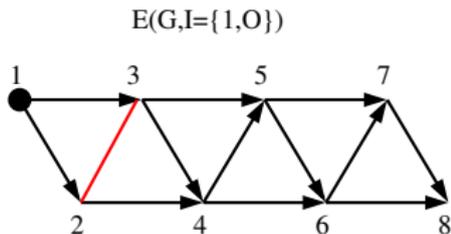
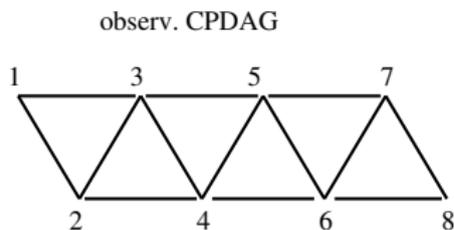
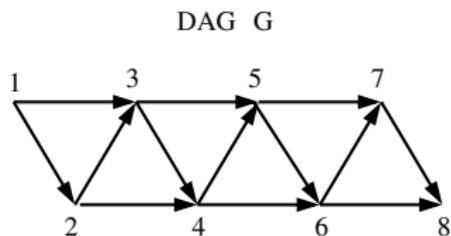
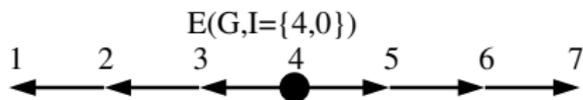
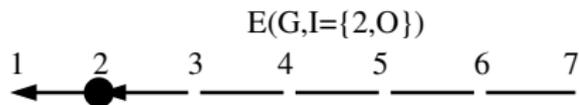
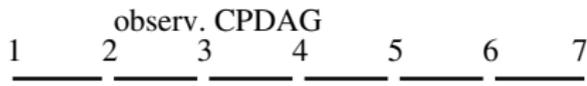
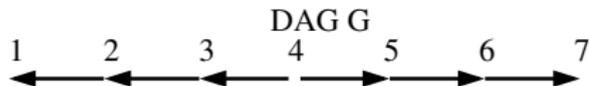
DAG : $X \rightarrow Y$; intervention at $Y \rightsquigarrow$ interv. DAG : $X \perp Y$
 $\rightsquigarrow X, Y$ independent

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”

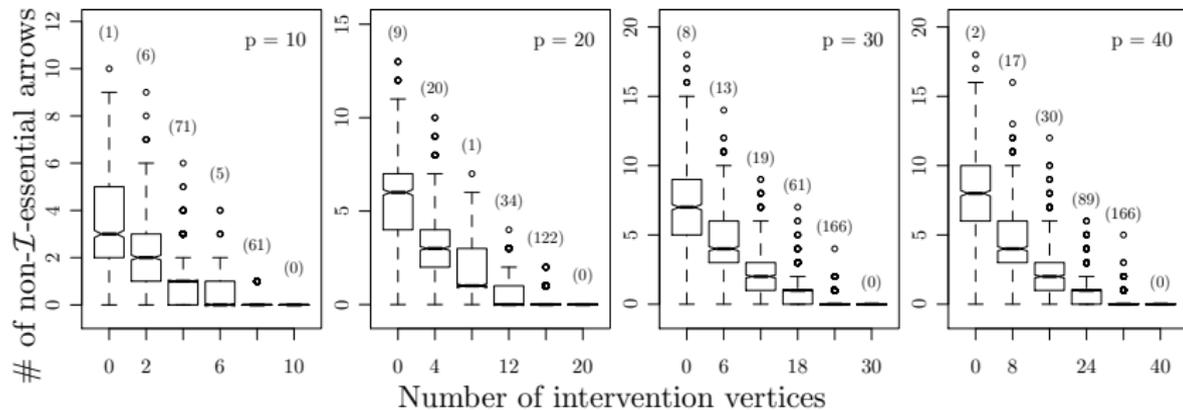
in the sense that $|\mathcal{I}_{\min}|$ is small

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

methods for efficient sequential design of intervention experiments

“active learning”

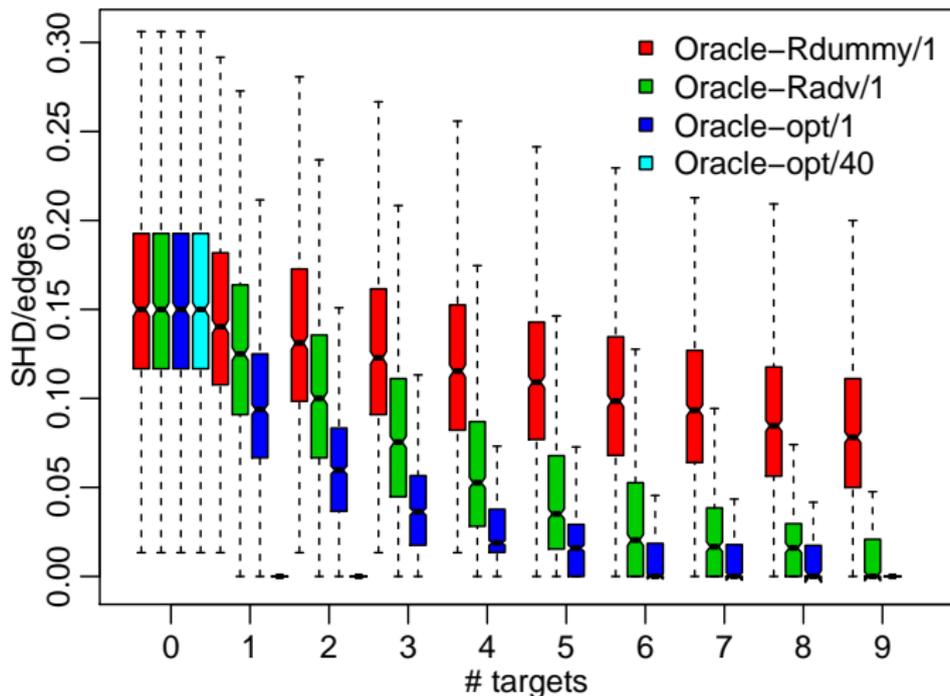
randomly chosen intervention variables



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

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

Oracle estimates, $p = 40$



The model and the (penalized) MLE

consider data

$$X_{1,\text{obs}}, \dots, X_{n_1,\text{obs}}, \quad X_{1,l_1=x_1}, \dots, X_{n_2,l_{n_2}=x_{n_2}}$$

n_1 observational data

n_2 interventional data (single variable interventions)

model:

$X_{1,\text{obs}}, \dots, X_{n_1,\text{obs}}$ i.i.d. $\sim P_{\text{obs}} = \mathcal{N}_p(0, \Sigma)$ faithful to a DAG D ,

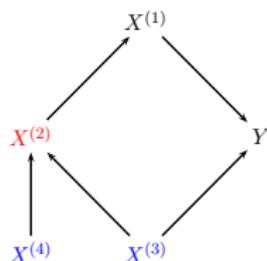
$X_{1,l_1}, \dots, X_{n_2,l_{n_2}}$ independent, non-identically distributed

independent of $X_{1,\text{obs}}, \dots, X_{n_1,\text{obs}}$

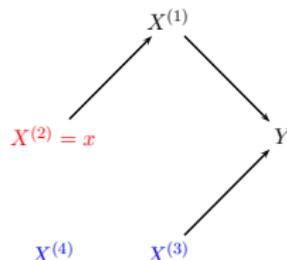
$X_{i,l_i=x_i} \sim P_{\text{int};l_i,x_i}$ linked to the above P_{obs} via do-calculus

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

non-intervention



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



$$\begin{aligned} P(Y, X_1, X_2, X_3, X_4) = & \\ & P(Y|X_1, X_3) \times \\ & P(X_1|X_2) \times \\ & P(X_2|X_3, X_4) \times \\ & P(X_3) \times \\ & P(X_4) \end{aligned}$$

$$\begin{aligned} P(Y, X_1, X_3, X_4 | \text{do}(X_2 = x)) = & \\ & P(Y|X_1, X_3) \times \\ & P(X_1|X_2 = x) \times \\ & P(X_3) \times \\ & P(X_4) \end{aligned}$$

can write down the likelihood:

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

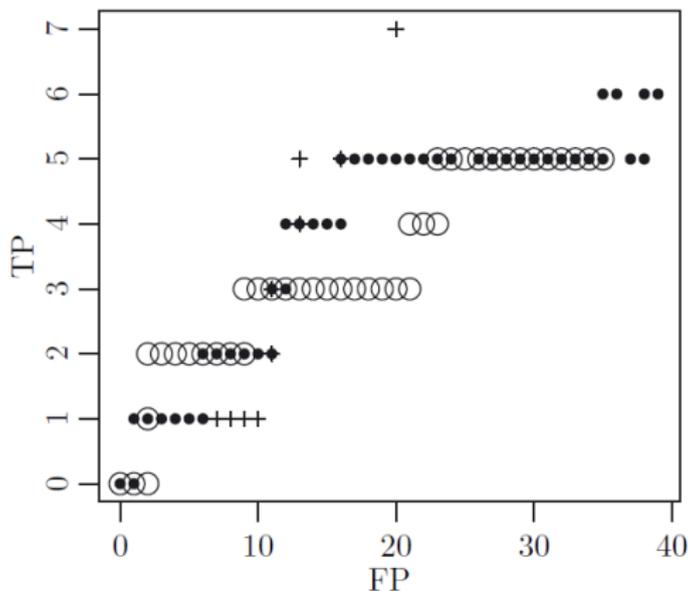
with “argmin” under the constraint that B does not lead to directed cycles

- ▶ 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 $n_{\text{obs}} \rightarrow \infty$

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:

do($X_k = x$) does **not** change the factors

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

probably a bit unrealistic in biology applications!

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

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

noise intervention with factor $a_j > 0$

also here: **autonomy assumption** that all other structural equations remain the same

environment intervention, for example

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

$X^{(e)}$ 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 Lectures III and IV

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 effects”
- ▶ intervention modeling is still in its infancies
it is over-shadowed by Pearl's excellent and simple do-intervention model
- ▶ active learning is interesting but finite-sample performance still poor

Thank you!

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)*.