High-dimensional causal inference, DAGs and intervention DAGs

Peter Bühlmann

Seminar für Statistik, ETH Zürich

joint work with
Marloes Maathuis, Markus Kalisch, Alain Hauser

Goal

in genomics:
if we would make an intervention at a single gene, what would be its effect on a phenotype of interest?

want to infer/predict such effects without actually doing the intervention
i.e. from observational data
or a mix of observational and interventional data

it doesn't need to be genes
can generalize to intervention at more than one variable/gene

Goal

in genomics:
if we would make an intervention at a single gene, what would be its effect on a phenotype of interest?

want to infer/predict such effects without actually doing the intervention
i.e. from observational data
or a mix of observational and interventional data

it doesn't need to be genes
can generalize to intervention at more than one variable/gene

Two examples

1. Flowering of arabidopsis thaliana
phenotype of interest: $Y =$ days to bolting (flowering)
"covariates" $X =$ gene expressions from $p = 21'326$ genes
question: infer/predict the effect of knocking-out/knocking-down (or enhancing) a single gene j on the phenotype Y?

2. Gene expressions of yeast
$p = 5360$ genes
phenotype of interest: $Y =$ expression of first gene
"covariates" $X =$ gene expressions from all other genes

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

and so on

infer/predict the effects of a single gene knock-down on all other genes

we could use linear model (fitted from observational data)

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \quad n \ll p$$

and measure the effect of $X^{(j)}$ on Y with $\hat{\beta}_j$ from e.g. Lasso or similar methods...

but regression is the “wrong approach”

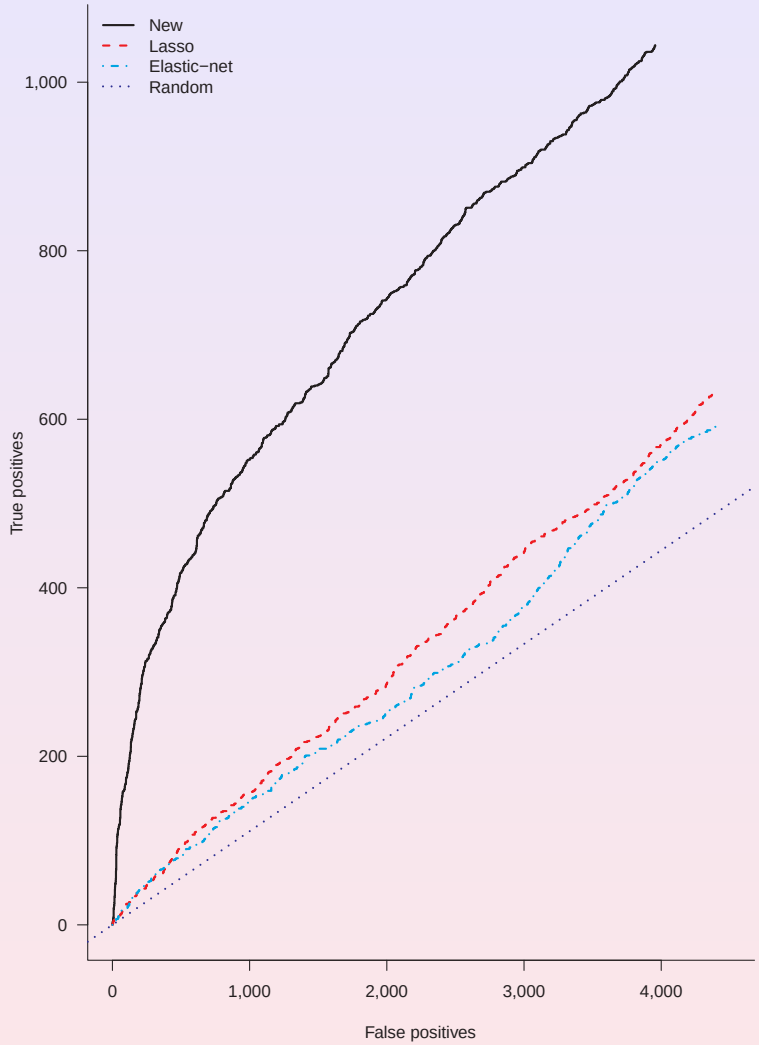
we could use linear model (fitted from observational data)

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \quad n \ll p$$

and measure the effect of $X^{(j)}$ on Y with $\hat{\beta}_j$ from e.g. Lasso or similar methods...

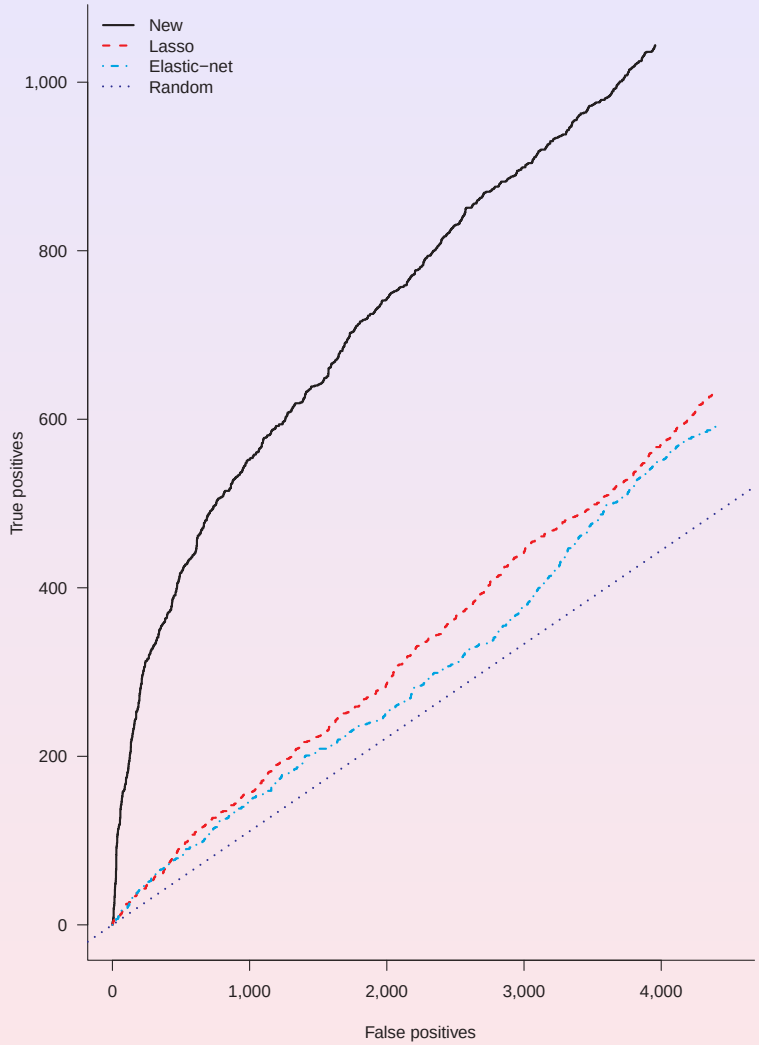
but regression is the “wrong approach”

Figure 1



↝ better than penalized regression/classification

Figure 1



$\rightsquigarrow$ better than penalized regression/classification

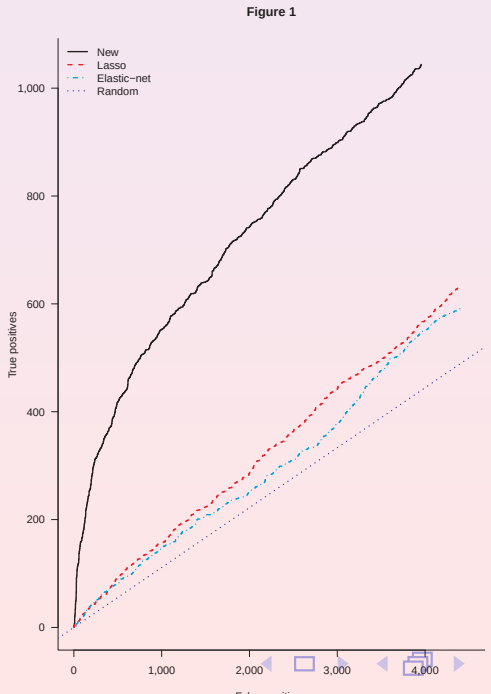
Effects of single gene knock-downs on all other genes (yeast)

(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



... “causal inference from purely observed data could have practical value in the prioritization and design of perturbation experiments”

Editorial in Nature Methods (April 2010)

Why are we doing better than regularized regression?

the problem is of intervention-type!

and not of association-type... which could be well-addressed by regression techniques

intervention = causality

(defined in mathematical terms)

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 association

in contrast to: (high-dimensional) regression

$$Y = \sum_{j=1}^{p} \beta_j X^{(j)} + \varepsilon,$$

$$\mathrm{Var}(X^{(j)}) \equiv 1 \text{ for all } j$$

$|\beta_j|$ measures the importance 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 are not (cannot be) kept fixed

in contrast to: (high-dimensional) regression

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

$|\beta_j|$ measures the importance 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

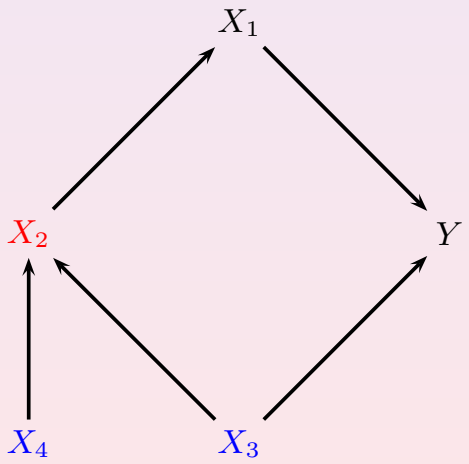
$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these are not (cannot be) kept fixed

Intervention calculus (a review)

"dynamic" notion of importance:
if we set a variable $X^{(j)}$ to a value x (intervention)
$\rightsquigarrow$ some other variables $X^{(k)}$ $(k \neq j)$ and maybe Y will change

we want to quantify the "total" effect of
$X^{(j)}$ on Y including "all changed" $X^{(k)}$ on Y

a graph or influence diagram will be very useful



for simplicity: just consider DAG's
(for ancestral graphs with hidden variables: work in progress)

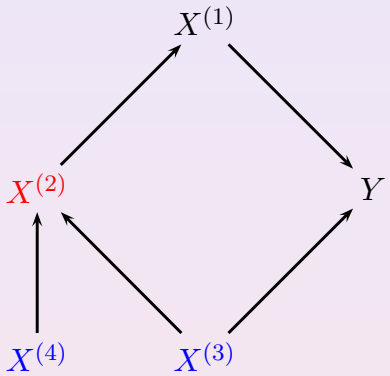
assume Markov condition for DAG: recursive factorization of joint distribution

$$P(Y, X^{(1)}, \ldots, X^{(p)}) = P(Y|X^{(\mathrm{pa}(Y))}) \prod_{j=1}^{p} P(X^{(j)}|X^{(\mathrm{pa}(j))})$$

for intervention calculus: use truncated factorization (e.g. Pearl)

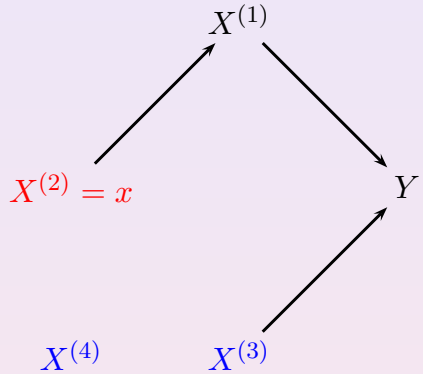
assume Markov property for causal DAG:

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 for $\mathrm{do}(X^{(2)} = x)$:

$$
\begin{aligned}
& P(Y, X^{(1)}, X^{(3)}, X^{(4)} | \mathrm{do}(X^{(2)} = x)) \\
= \quad & P(Y | X^{(1)}, X^{(3)}) P(X^{(1)} | X^{(2)} = x) P(X^{(3)}) P(X^{(4)})
\end{aligned}
$$

$$
\begin{aligned}
& P(Y | \mathrm{do}(X^{(2)} = x)) \\
= \quad & \int P(Y, X^{(1)}, X^{(3)}, X^{(4)} | \mathrm{do}(X^{(2)} = x)) dX^{(1)} dX^{(3)} dX^{(4)}
\end{aligned}
$$

the truncated factorization is a mathematical consequence of the Markov condition (with respect to the causal DAG) for the probability distribution P

the intervention distribution $P(Y|\mathrm{do}(X^{(2)} = x))$ can be calculated from

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

intervention effect:

$$\mathbb{E}[Y|\mathrm{do}(X^{(2)} = x)] = \int y P(y|\mathrm{do}(X^{(2)} = x))dy$$

intervention effect at x_0 : $\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(2)} = x)]|_{x=x_0}$

in the Gaussian case: $Y, X^{(1)}, \ldots, X^{(p)} \sim \mathcal{N}_{p+1}(\mu, \Sigma)$,

$$\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(2)} = x)] \equiv \theta_2 \text{ for all } x$$

when having no unmeasured confounder (variable):

intervention effect (as defined) = causal effect

causal effect = effect from a randomized trial
(but we want to infer it without a randomized study...
because often we cannot do it, or it is too expensive)

when having no unmeasured confounder (variable):

intervention effect (as defined) = causal effect

causal effect = effect from a randomized trial
(but we want to infer it without a randomized study...
because often we cannot do it, or it is too expensive)

An important characterization

recap, Gaussian case: $\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)] \equiv \theta_j$ for all x

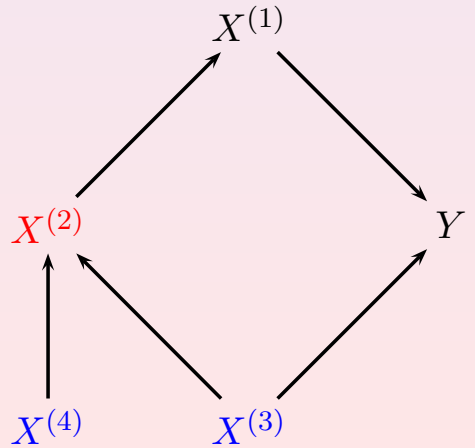
for $Y \notin \mathrm{pa}(j)$:

θ_j is the regression parameter in

$$Y = \theta_j X^{(j)} + \sum_{k \in \mathrm{pa}(j)} \theta_k X^{(k)} + \text{error}$$

only need parental set and regression

$j = 2$, $\mathrm{pa}(j) = \{3, 4\}$



in the Gaussian case:

causal inference =
regression when conditioning on the right variables

Inferring intervention effects from data

main problem: inferring parental set (or DAG) from data because regression is easy

outline

1. inferring DAG from observational data
2. inferring DAG from intervention data or from observational *and* intervention data

Inferring DAG from observational data

$\rightsquigarrow$ impossible: can only infer equivalence class of DAG's (several DAGs can encode exactly the same conditional independence relationships)

and we cannot estimate causal/intervention effects from observational data

the usual statistical inference principle doesn't work:
observational probability distribution $P \Rightarrow$ parameter $\theta(P)$

here:

$$P \text{ and graph } G \Rightarrow \text{ causal effect } \theta(P, G)$$

impossible to estimate causal/intervention effects from observational data

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

conceptual "procedure":

- probability distribution P from a DAG, generating the data
  $\rightsquigarrow$ true underlying equivalence class of DAG's
- find all DAG-members of true equivalence class:
  $G_1, \ldots, G_m$
- for every DAG-member G_r, and every variable $X^{(j)}$:
  single intervention effect $\theta_{r,j}$
  summarize them by

$$\underbrace{\Theta = \{\theta_{r,j};\ r = 1, \ldots, m;\ j = 1, \ldots, p\}}_{\text{population quantity}}$$

impossible to estimate causal/intervention effects from observational data

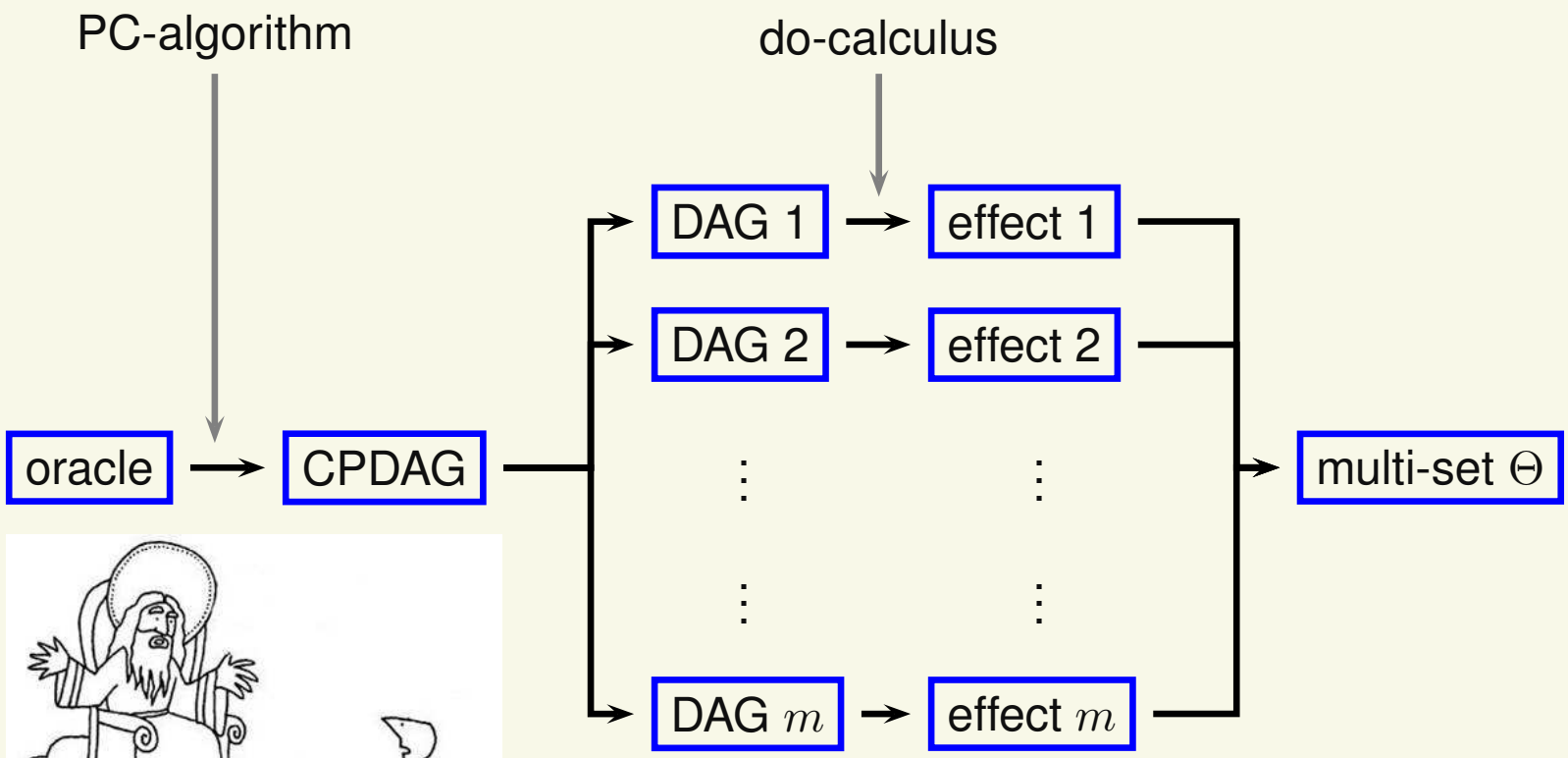
but we will be able to estimate lower bounds of causal effects

conceptual "procedure":

- probability distribution P from a DAG, generating the data
  $\rightsquigarrow$ true underlying equivalence class of DAG's
- find all DAG-members of true equivalence class:
  $G_1, \ldots, G_m$
- for every DAG-member G_r, and every variable $X^{(j)}$:
  single intervention effect $\theta_{r,j}$
  summarize them by

$$\underbrace{\Theta = \{\theta_{r,j};\ r = 1, \ldots, m;\ j = 1, \ldots, p\}}_{\text{population quantity}}$$

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

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

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

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

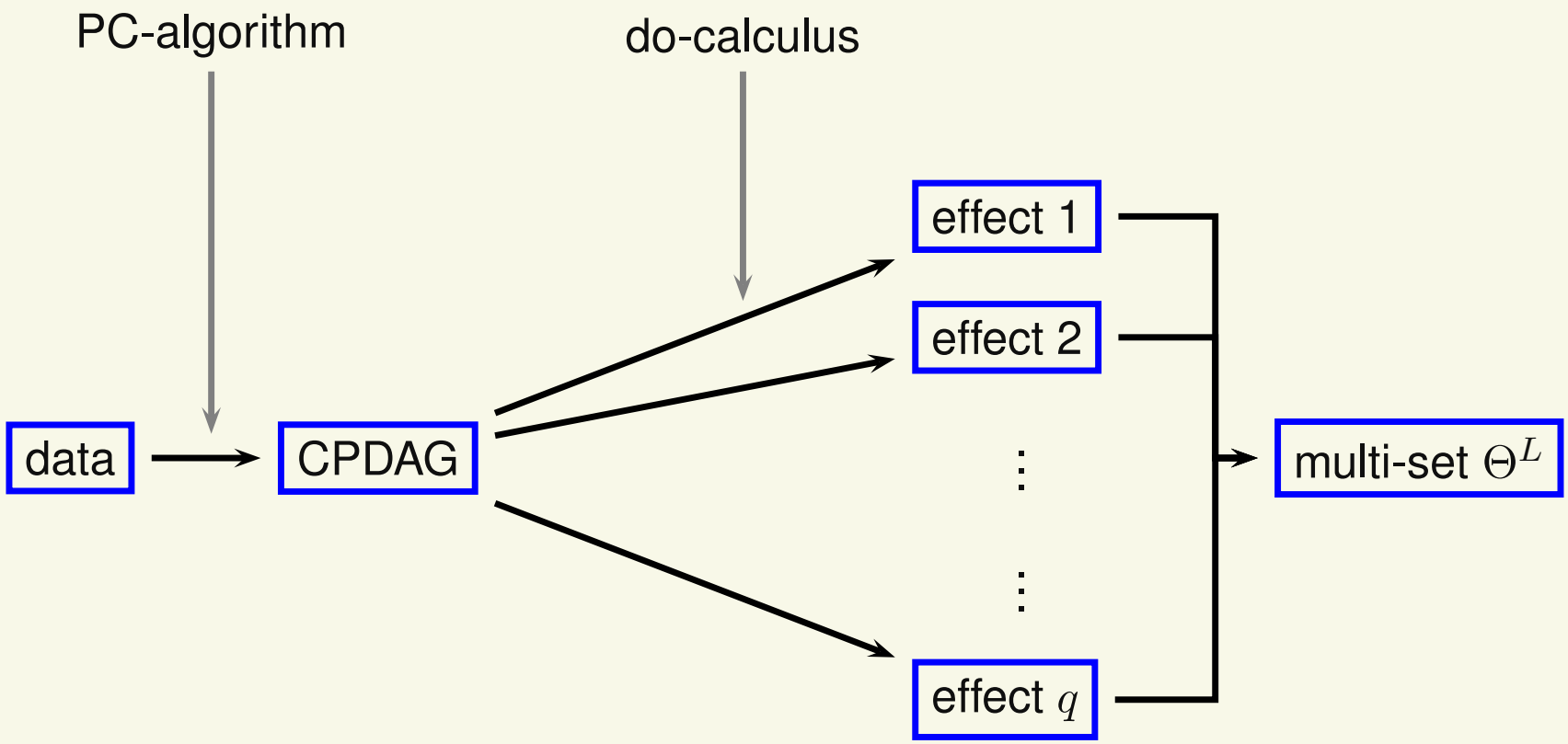
$\exists$ Computationally tractable algorithm for Θ

"local algorithm"

instead of finding all m DAG's within an equivalence class $\rightsquigarrow$
compute all intervention effects without finding all DAG's

Maathuis, Kalisch & PB (2009):

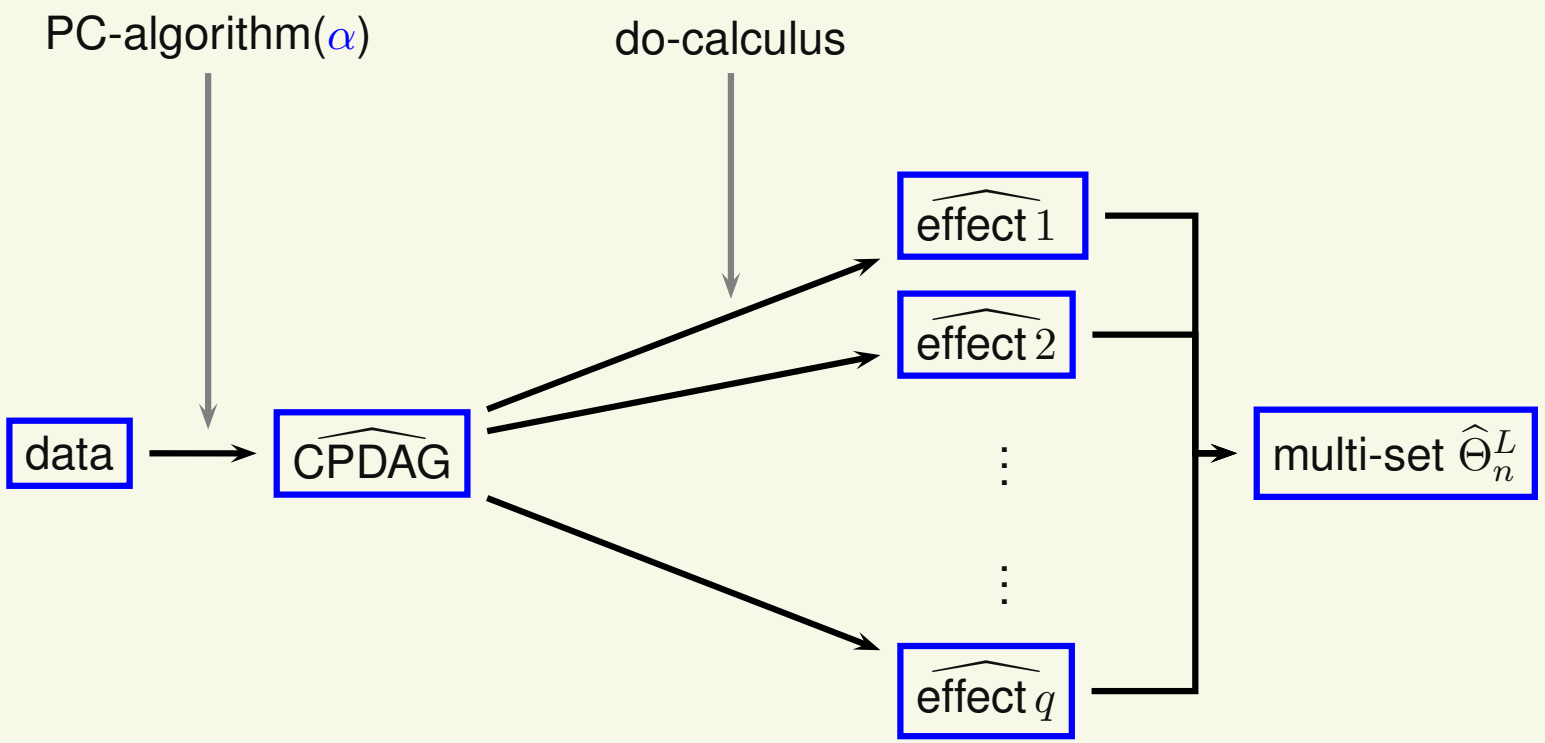
- algorithm which works on local aspects of the graph only
- prove that such a local algorithm is computing Θ



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

and PC-algorithm (Spirtes, Glymour, 1991) for estimation

IDA (local sample version)

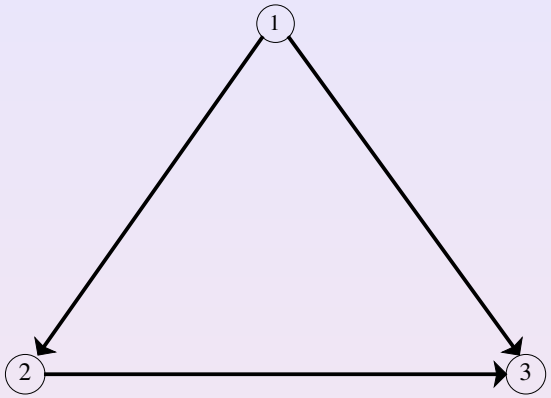


Faithfulness assumption is crucial for estimation of CPDAG

A distribution P is called faithful to a DAG G if all conditional independencies can be inferred from the graph

(can infer some conditional independencies from a Markov assumption; but we require here “all” conditional independencies)

What does it mean?



$$
\begin{aligned}
X^{(1)} &\leftarrow \varepsilon^{(1)}, \\
X^{(2)} &\leftarrow \alpha X^{(1)} + \varepsilon^{(2)}, \\
X^{(3)} &\leftarrow \beta X^{(1)} + \gamma X^{(2)} + \varepsilon^{(3)}, \\
\varepsilon^{(1)}, \varepsilon^{(2)}, \varepsilon^{(3)} &\text{ i.i.d. } \sim \mathcal{N}(0,1)
\end{aligned}
$$

enforce marginal independence of $X^{(1)}$ and $X^{(3)}$
$\beta + \alpha\gamma = 0$, e.g. $\alpha = \beta = 1,\ \gamma = -1$

$$
\Sigma = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} 3 & -2 & -1 \\ -2 & 2 & 1 \\ -1 & 1 & 1 \end{pmatrix}.
$$

failure of faithfulness due to cancellation of regression coefficients

Theorem (Kalisch & PB, 2007; Maathuis, Kalisch & PB, 2009)
triangular scheme of observations

- $Y, X^{(1)}, \ldots, X^{(p_n)} \sim \mathcal{N}_{p_n+1}(\mu_n, \Sigma_n)$ faithful to a DAG $\forall n$
- $p_n = O(n^\alpha)$ $(0 \leq \alpha < \infty)$ (high-dimensional)
- $d_n = \max_j |\mathrm{ne}(j)| = o(n)$ (sparsity)
- non-zero (partial) correlations sufficiently large ("signal strength")
  $\min\{|\rho_{n;i,j|S}|; \rho_{n;i,j|S} \neq 0,\ i \neq j, |S| \leq d_n\} \gg \sqrt{d_n \log(p_n)/n}$
- maximal (partial) correlations $\leq C < 1$ ("coherence")
  $\max\{|\rho_{n;i,j|S}|;\ i \neq j, |S| \leq d_n\} \leq C < 1$

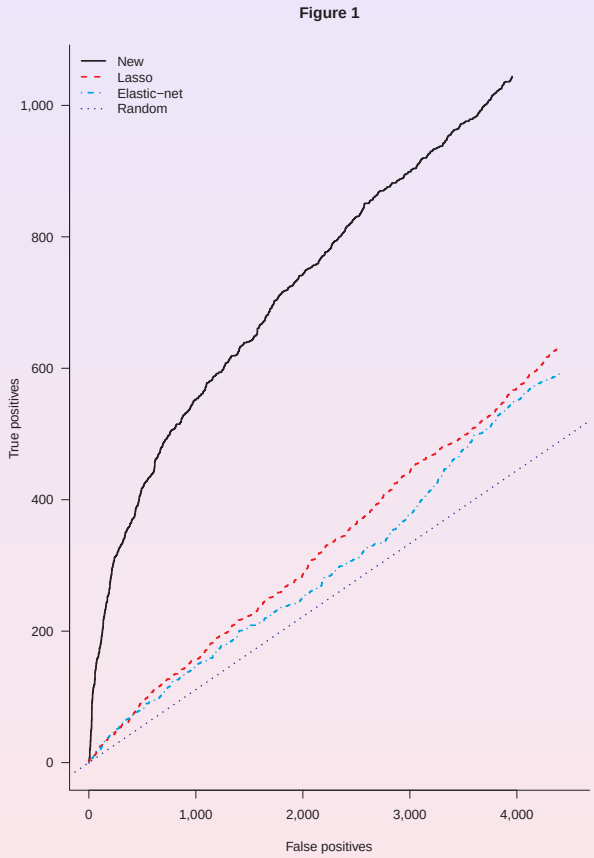
Then: for some suitable $\alpha = \alpha_n$

$$\mathbb{P}[\widehat{\mathrm{CPDAG}}(\alpha) = \text{ true CPDAG}] = 1 - O(\exp(Cn^{1-\delta}))$$

$$\mathbb{P}[\hat{\Theta}_{\mathrm{local}}(\alpha) \overset{\text{as set}}{=} \Theta] = 1 - O(\exp(Cn^{1-\delta}))$$

(i.e. consistency of lower bounds for causal effects)

How well can we do?



the real success is the prediction of causal effects on
gene interactions in yeast

where the true causal effects are "known" thanks to intervention experiments

Maathuis, Colombo, Kalisch & PB (2010)

Arabidopsis thaliana

response Y: days to bolting (flowering) of the plant
(aim: fast flowering plants)
X: gene-expression profile

observational data with $n = 47$ and $p = 21'326$

we validated the top 14 genes having largest lower bounds $\hat{\alpha}_j$:

randomized experiments with 14 mutant plants
(only 9 mutant plants survived)
$\rightsquigarrow$ found 3 significant new genes for "time to flowering"
(Stekhoven, PB and Hennig, 2010)

in short:

bounds on causal effects ($\hat{\alpha}_j$'s) based on observational data lead to interesting predictions for interventions in genomics (i.e. which genes would exhibit a large intervention effect)

and these predictions have been validated using experiments

Inference based on observational and interventional data (Hauser & PB, in progress)

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...)

consider data

$$X_{1,\mathrm{obs}}, \ldots, X_{n_1,\mathrm{obs}}, \quad X_{1,I_1}, \ldots, X_{n_2,I_{n_2}}$$

n_1 observational data
n_2 intervention data (single variable interventions)

model:

$$X_{1,\mathrm{obs}}, \ldots, X_{n_1,\mathrm{obs}} \text{ i.i.d. } \sim \mathcal{N}_p(0, \Sigma),$$
faithful to a DAG G,
$X_{1,I_1}, \ldots, X_{n_2,I_{n_2}}$ independent
independent of $X_{1,\mathrm{obs}}, \ldots, X_{n_1,\mathrm{obs}}$
and arising from $\mathcal{N}_p(0, \Sigma)$ faithful to a DAG G

"arising from $\mathcal{N}_p(0, \Sigma)$ faithful to a DAG G": via the do-calculus

the intervention data have non-identical distributions

$\rightsquigarrow$ can write down the likelihood

$$-\ell(\Sigma, G;\ \text{data}) = \ldots$$

unknown quantities are Σ and G

Gaussian DAG is Gaussian linear structural equation model:

$$X^{(j)} \leftarrow \sum_{k=1}^{p} \beta_{jk} X^{(k)} + \varepsilon_j \ (j = 1, \ldots, p),\ \beta_{jk} \neq 0 \Leftrightarrow \text{ edge } k \to j$$

$$X = B^T X + \varepsilon,\ \varepsilon \sim \mathcal{N}_p(0, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)) \text{ in matrix notation}$$

$\rightsquigarrow$ reparametrization

$$(\Sigma, G) \leftrightarrow (B, \{\sigma_j^2;\ j = 1, \ldots, p\})$$

(non-zeroes of B do not lead to directed cycles)

$\rightsquigarrow$ can write down the likelihood

$$-\ell(\Sigma, G;\ \text{data}) = \ldots$$

unknown quantities are Σ and G

Gaussian DAG is Gaussian linear structural equation model:

$$X^{(j)} \leftarrow \sum_{k=1}^{p} \beta_{jk} X^{(k)} + \varepsilon_j \ (j = 1, \ldots, p),\ \beta_{jk} \neq 0 \Leftrightarrow \text{ edge } k \to j$$

$$X = B^T X + \varepsilon,\ \varepsilon \sim \mathcal{N}_p(0, \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)) \text{ in matrix notation}$$

$\rightsquigarrow$ reparametrization

$$(\Sigma, G) \leftrightarrow (B, \{\sigma_j^2;\ j = 1, \ldots, p\})$$

(non-zeroes of B do not lead to directed cycles)

thus:

$$X_{i;\text{obs}} \sim \mathcal{N}_p(0, \Sigma),\ \Sigma = (I - B)^{-T}\text{diag}(\{\sigma_j^2; j\})(I - B)^{-1}$$

and

$$\begin{aligned}
X_{i;I_i} &= X_i|\text{do}(X_i^{(I_i)} = x_i) \sim \mathcal{N}_{p-1}(\mu_{I_i}, \Sigma_{I_i}),\\
\mu_{I_i} &= (I - BR_{I_i})^{-T} Q_{I_i}^T x_i,\\
\Sigma_{I_i} &= (I - BR_{I_i})^{-T} R_{I_i}\text{diag}(\{\sigma_j^2; j\})R_{I_i}(I - BR_{I_i})^{-1}
\end{aligned}$$

$\rightsquigarrow$ explicit form of likelihood

$$-\ell(\Sigma, G;\ \text{data}) = -\ell(B, \{\sigma_j^2; j\};\ \text{data})$$

where non-zeroes of B do not lead to directed cycles

Penalized MLE

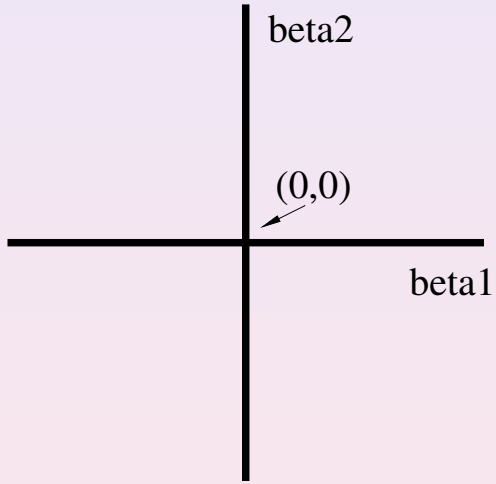
$$\begin{aligned}\hat{\Sigma},\ \hat{G} &= \text{argmin}_{\Sigma;G \text{ a DAG}} - \ell(\Sigma, G;\ \text{data}) + \lambda|G| \\ &= \text{argmin}_{B;\ \{\sigma_j^2;j\}} - \ell(B, \{\sigma_j^2;\ j\};\ \text{data}) + \lambda \underbrace{\|B\|_0}_{\sum_{ij} I(B_{ij}\neq 0)}\end{aligned}$$

under the non-convex constraint that B corresponds to "no directed cycles"

severe non-convex problem due to the "no directed cycle" constraint
($\|\cdot\|_0$-penalty rather than e.g. $\|\cdot\|_1$ doesn't make the problem much harder)

Example

$$X^{(1)} \leftarrow \beta_1 X^{(2)} + \varepsilon_1$$
$$X^{(2)} \leftarrow \beta_2 X^{(1)} + \varepsilon_2$$



no straightforward way to do convex relaxation

Properties and computation of penalized MLE

Identifiability

set of variables where interventions are performed

$$\mathcal{I} \subseteq \{1, \ldots, p\} \cup O$$

where O denotes observational

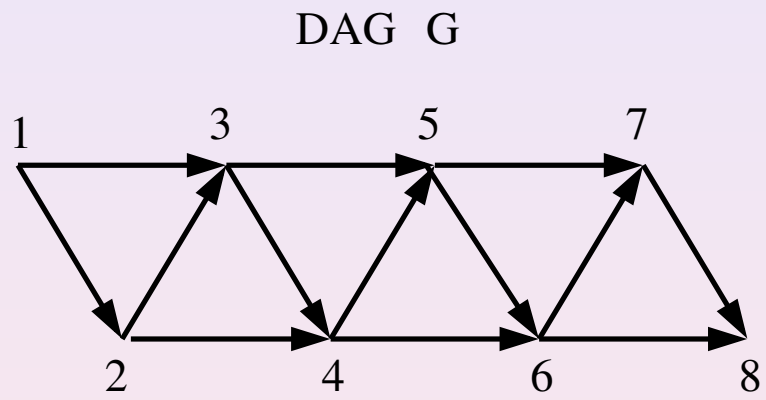
Essential graph $\mathcal{E}(G)$:
encodes the (Markov-) equivalence class under the interventions at $\mathcal{I}$, i.e.

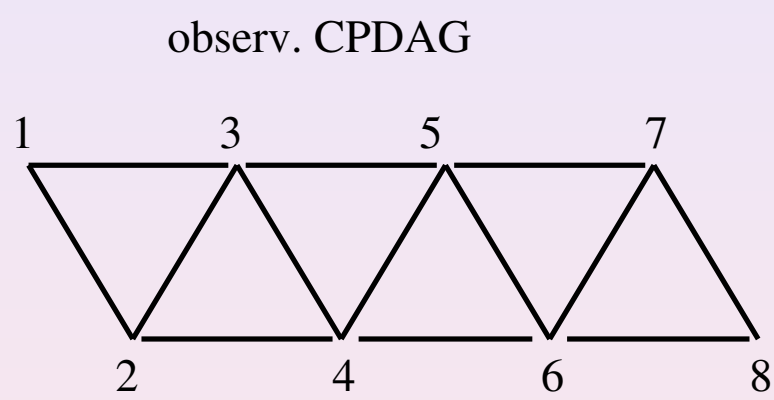
$$\mathcal{E}(G) = \cup_{G'} \{G' \sim_{\mathcal{I}} G\}$$

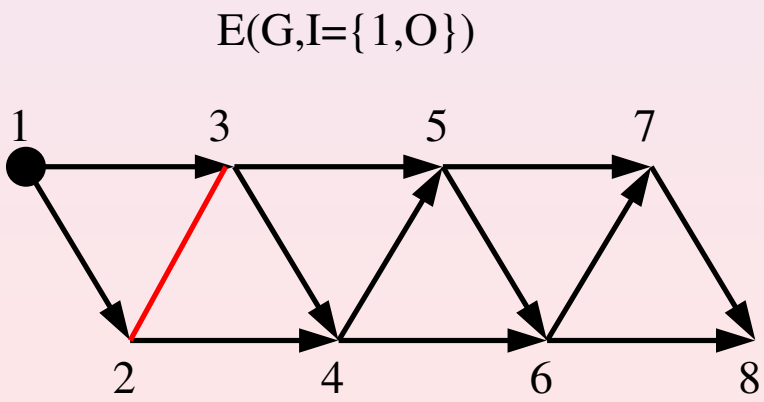
($\sim_{\mathcal{I}}$ needs to be defined...: "G' and G encode the same independence relations for all interventions $I \in \mathcal{I}$")

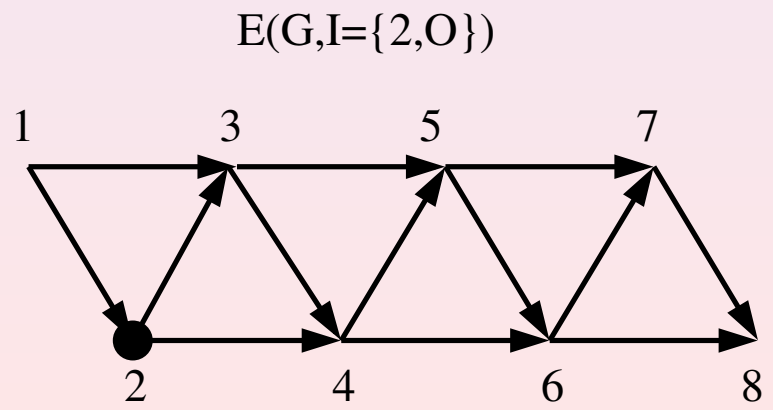
Example: DAG G (top left)

- $\mathcal{I} = \{O\}$ (observational data only) $\rightsquigarrow$ CPDAG(G) (top right)
  CPDAG equivalence class contains 26 DAG elements
- $\mathcal{I} = \{1, O\} \rightsquigarrow$ small equivalence class (bottom left)
- $\mathcal{I} = \{2, O\} \rightsquigarrow$ can recover the true DAG G (bottom right)









there is a minimal set of intervention variables $\mathcal{I}_{\min}$ such that $\mathcal{E}(G, \mathcal{I}_{\min}) = G$
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, the better identifiable from observational/intervention data"
in the sense that $|\mathcal{I}_{\min}|$ is small

Open problem 1:
Inferring $\mathcal{I}_{\min}$ from available data
(for doing the next intervention experiment)

- "optimal" sequential estimation
- optimal active learning for estimating the true underlying DAG

Estimation of equivalence class

for a penalized MLE:

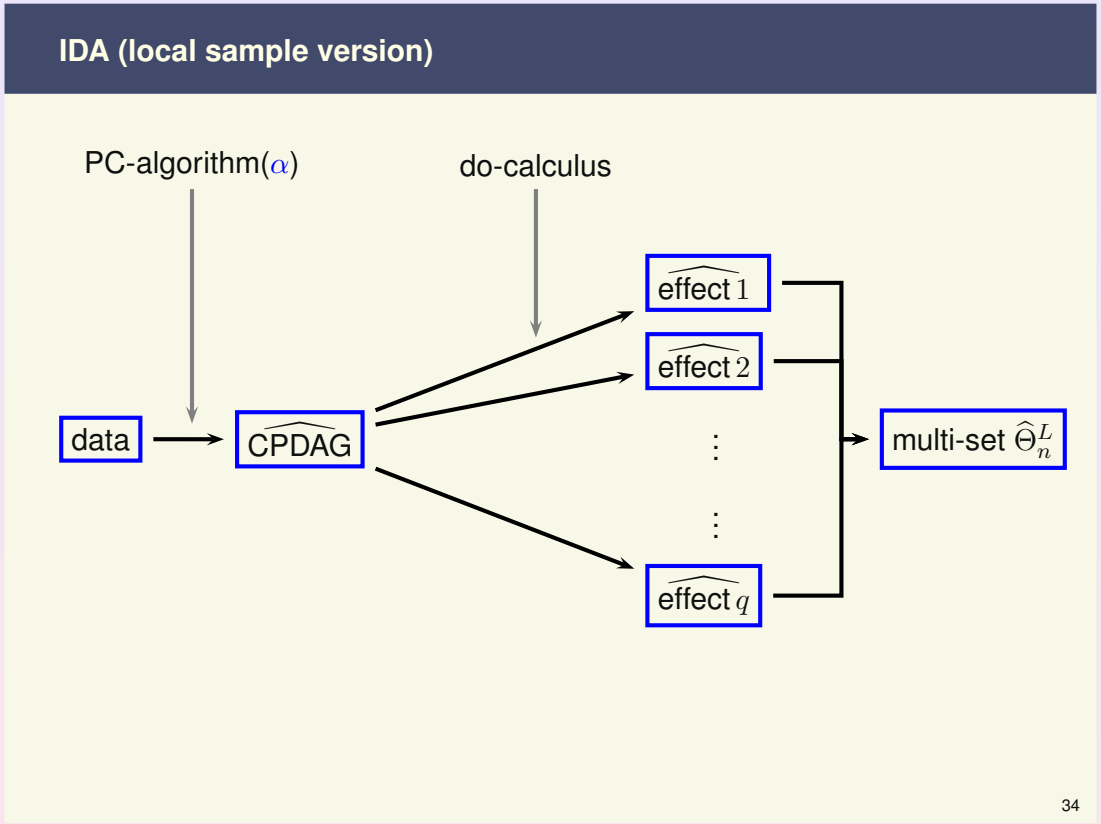
$$\hat{\Sigma},\ \hat{G} \quad = \quad \text{argmin}_{\Sigma;G \text{ a DAG}} - \ell(\Sigma, G;\ \text{data}) + \lambda|G|$$

complete it to the estimator of the equivalence class

$$\hat{\mathcal{E}}(\mathcal{I}) = \mathcal{E}(\hat{G}, \mathcal{I})$$

and in fact: every $G' \in \hat{\mathcal{E}}(\mathcal{I}) = \mathcal{E}(\hat{G}, \mathcal{I})$ leads to the same optimum of penalized likelihood

once we have equivalence class $\hat{\mathcal{E}}(\mathcal{I})$
$\rightsquigarrow$ use the local algorithm to compute all possible causal effects



but replacing

- PC-algorithm with penalized MLE
- $\widehat{\text{CPDAG}}$ with often much smaller $\hat{\mathcal{E}}(\mathcal{I})$

Asymptotic theory (Hauser & PB, in progress)

- p fix, $n \to \infty$
- true distribution $P_0 = \mathcal{N}_p(0, \Sigma_0)$ which is faithful w.r.t. to true underlying DAG G_0
- assume $\lambda = \lambda_{\mathrm{BIC}} = \log(n)/2$ from BIC (or any $\lambda = \lambda_n \to \infty$, $\lambda_n/n \to 0$)

then, for any set of intervention variables $\mathcal{I}$

$$\mathbb{P}[\hat{\mathcal{E}}(\mathcal{I}) = \mathcal{E}(G_0, \mathcal{I})] \to 1 \ (n \to \infty),$$
$$\hat{\Sigma} - \Sigma_0 = o_P(1) \ (n \to \infty)$$

$n \to \infty$ means: number of observations for every intervention experiment in $\mathcal{I} \to \infty$
i.e. repeated observations for every intervention (and maybe also observational setting)

instead of $n \to \infty$, and more realistic:

- only one observation for every intervention but intervention value far away from the observational mean:

$$\mathrm{do}(X^{(j)} = x) \text{ with } |x| \to \infty$$

- no. of observational observations $\to \infty$

technicalities: we have to deal with non-i.i.d. data
generalize results for curved exponential families (Haughton, 1988)

Open problem 2:
statistical properties for $p \gg n$ setting

Computation

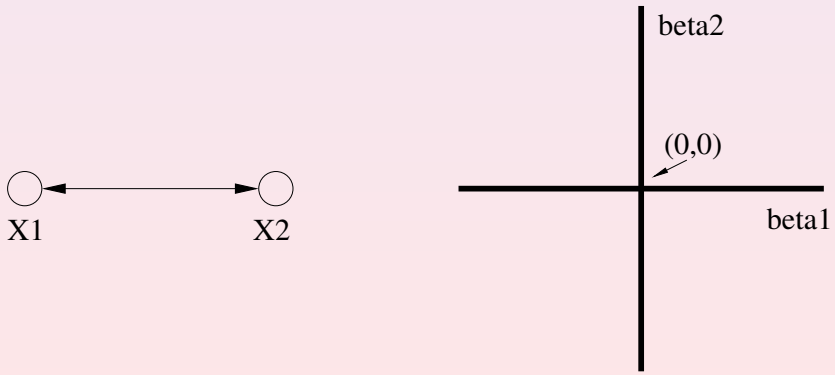
for computing ℓ_0-penalized MLE:

$$\hat{B}, \{\hat{\sigma}_j^2; j\} = \text{argmin}_{B,\{\sigma_j^2;j\}} - \ell(B, \{\sigma_j^2; j\}; \text{ data}) + \lambda \sum_{i,j} I(B_{ij} \neq 0)$$

under the non-convex constraint that B corresponds to "no directed cycles"
and then the equivalence class $\hat{\mathcal{E}}(\mathcal{I}) = \hat{\mathcal{E}}(\hat{G}, \mathcal{I})$

recall the Example:

$$X^{(1)} \leftarrow \beta_1 X^{(2)} + \varepsilon^{(1)}$$
$$X^{(2)} \leftarrow \beta_2 X^{(1)} + \varepsilon^{(2)}$$



no straightforward way to do convex relaxation

strategy: do greedy search over equivalent classes,
forward and backward

like Chickering (2002)'s greedy equivalent search for observational DAGs

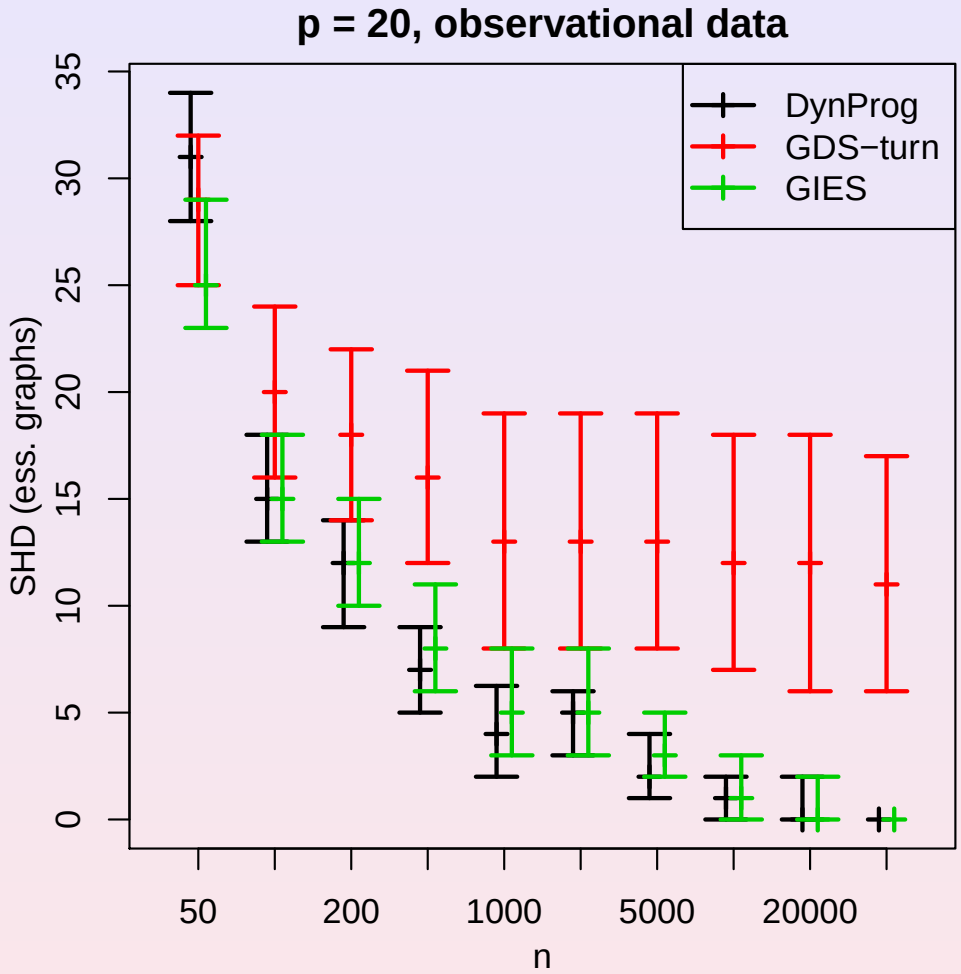
forward:

- current Markov equivalence class $\mathcal{E}$
- go to the next equivalence class $\mathcal{E}^+$ such that:
  there exist DAG G in $\mathcal{E}$ and $G^+ \in \mathcal{E}^+$ where G^+ has one more directed edge than G;
  $\mathcal{E}^+$ is such that the objective function is reduced most in one step (greedy)

backward: ... by deleting one edge...

this can be done efficiently without enumerating all members in the equivalence classes (Hauser & PB, in progress)

Performance comparison of algorithms



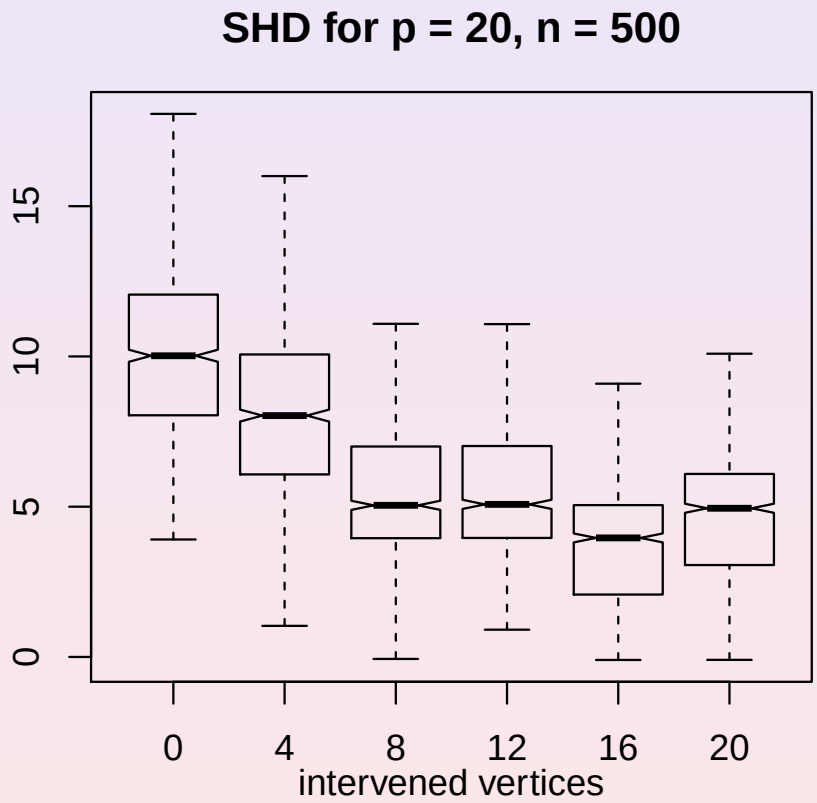
greedy equivalent (class) search is

- much better than greedy search (over DAGs)
- and for small dimension as good as exhaustive search

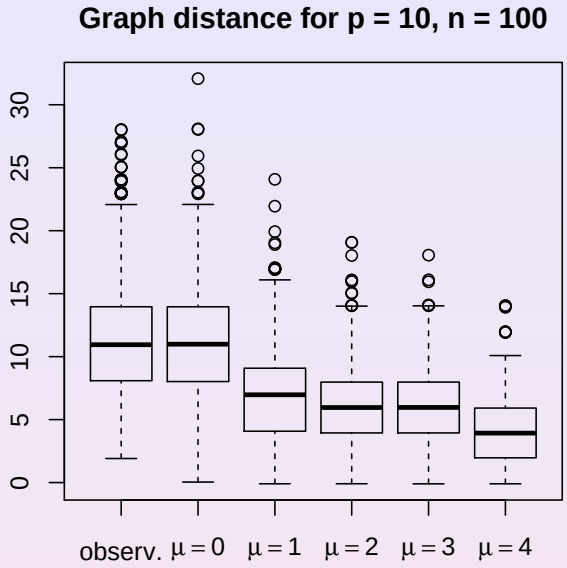
Open problem 3:
for large p: an algorithm with provable convergence property to an optimum for the ℓ_0-penalized MLE (or an ℓ_1-penalized MLE)

Performance gain with intervention data

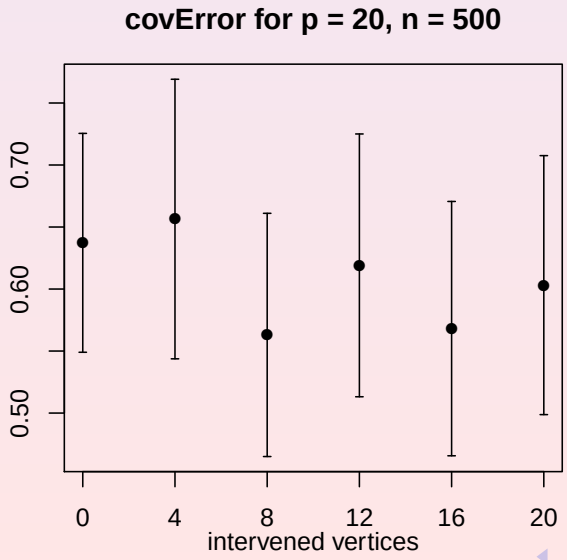
1. interventions at randomly chosen nodes:
   $\mathcal{I} = \{j_1, \ldots, j_m, O\}$ for $m = 0, 4, \ldots, 20$



2. interventions at all nodes: $\mathcal{I} = \{1, \ldots, p, O\}$
   for varying intervention values μ (observational mean = 0)



3. estimation of covariance Σ (varying no. interventions)



Beware of over-interpretation!

so far, based on current data:
we can not reliably infer the causal network
despite theorems...
(stability selection/bootstrapping yields rather unstable networks)

but apparently: we obtain stable and better ranking/prediction for intervention/causal effects than modern but conceptually wrong regression methods

Concluding remarks

observational data from one probability distribution P_{obs}

- can estimate equivalence class $\text{CPDAG}(G)$ (PC-algorithm)
- can infer lower bounds for causal effects
  (local algorithm in CPDAG space)

this doesn't involve likelihood

and because likelihood is not involved
$\rightsquigarrow$ for high-dimensional sparse setting:
- computation is feasible and provably correct
- method is statistically consistent

observational and interventional data from from different distributions

$$P_{\text{obs}}, \{P_I;\ \text{for all interventions } I\}$$

maybe only one observation for every P_I

"borrow strength from neighborhood (other interventions)": every P_I is a function of the DAG G and P_{obs} and the function is explicit thanks to the do-calculus

$$P_I = P_{I;x} = f_I(P_{\text{obs}}, G, \underbrace{x}_{\text{interv. value}}) = f_I(\Sigma, G, x)$$

likelihood is a convincing approach to "borrow strength from other interventions":

- optimization is highly non-convex
- statistical consistency is a much harder problem
- interventions $\rightsquigarrow$ better identifiability and better causal infer.

Thank you!

References:

- Hauser, A. and Bühlmann, P. (in progress). “Causal inference based on intervention and observational data”.
- 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.

Intervention data and the DAG model

observations from different distributions

$$P_{\text{obs}}, \{P_I;\ \text{for all interventions } I\}$$

maybe only one observation for every P_I

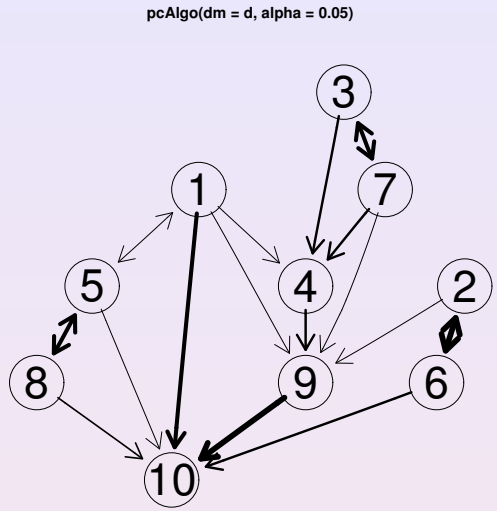
we need to "borrow strength from the neighborhood":
here: every P_I is a function of the DAG G and P and the function is explicit thanks to the do-calculus

$$P_I = P_{I;u} = f(P, G, \underbrace{u}_{\text{interv. value}}) = f(\Sigma, G, u)$$

Estimation from finite samples

difficult part: estimation of CPDAG (equivalence class of DAG's)
$\rightsquigarrow$ estimation of structure

$P \Rightarrow$ $\underbrace{\text{CPDAG}}_{\text{equiv. class of DAG's}}$



this can be inferred (statistical testing) from a list of conditional independence statements:

$$X^{(j)} \not\perp X^{(k)} | X^{(S)} \text{ for all subsets } S \subseteq \{1, \ldots, p\} \setminus \{j, k\}$$

or

$$X^{(j)} \perp X^{(k)} | X^{(S)} \text{ for some subset } S \subseteq \{1, \ldots, p\} \setminus \{j, k\}$$

so-called faithfulness assumption allows to reduce to "some subsets S"

The PC-algorithm (Spirtes & Glymour, 1991)

- crucial assumption:
  distribution P is faithful to the true underlying DAG
  i.e. all conditional (in-)dependencies can be read-off from the DAG (using the Markov property)
- less crucial but convenient:
  Gaussian assumption for $Y, X^{(1)}, \ldots, X^{(p)} \rightsquigarrow$ can work with partial correlations

strategy of the algorithm:

- estimate the skeleton first
- estimate some of the directions (using some special rules)

PC-algorithm: a rough outline for estimating the skeleton of underlying DAG

1. start with the full graph (all edges present)
2. remove edge $i - j$ if standard sample correlation $\widehat{\text{Cor}}(X^{(i)}, X^{(j)})$ is small
   by using Fisher's Z-transform and exact null-distribution of zero correlation
3. move-up to partial correlations of order 1:

$$\hat{\rho}_{i,j|k} = \frac{\hat{\rho}_{i,j} - \hat{\rho}_{i,k}\hat{\rho}_{j,k}}{\sqrt{(1-\hat{\rho}_{i,k}^2)(1-\hat{\rho}_{j,k}^2)}}$$

4. remove edge $i - j$ if standard sample partial correlation $\widehat{\text{Parcor}}(X^{(i)}, X^{(j)}|X^{(k)})$ is small for some k in the current neighborhood of i or j (thanks to faithfulness)

5. move-up to partial correlations of order 2 via recursive formula
6. remove edge $i - j$ if standard sample partial correlation $\widehat{\text{Parcor}}(X^{(i)}, X^{(j)}|X^{(k)}, X^{(\ell)})$ is small for some k, ℓ in the current neighborhood of i or j (thanks to faithfulness)
7. until removal of edges is not possible anymore, i.e. stop at minimal order of partial correlation where edge-removal becomes impossible

additional step of the algorithm needed for estimating directions yields an estimate of the CPDAG (equivalence class of DAG's)

one tuning parameter (cut-off parameter) α for truncation of estimated Z-transformed partial correlations

if the graph is "sparse" (few neighbors) $\rightsquigarrow$ few iterations only and only low-order partial correlations play a role

and thus: the estimation algorithm works for $p \gg n$ problems

more generally: assume knowledge of the skeleton (or the CPDAG) from observational data;
when doing an intervention at every variable, can infer all directions of all the arrows in the DAG (cf. He & Geng (2008))

not a very clever approach:

- want to do much less intervention experiments
- want to use information from intervention data for inferring the skeleton (or CPDAG) and the edge weights

directed Markov property on DAG
$\Leftrightarrow$ recursive factorization of joint distribution

if A and B are d-separated by $C \Rightarrow X^{(A)} \perp X^{(B)} | X^{(C)}$

Equivalence class of DAGs

- Several DAGs can encode exactly the same conditional independence relationships. Such DAGs form an equivalence class.
- Example: unshielded triple

	$X_1 \perp\!\!\!\perp X_3$	$X_1 \perp\!\!\!\perp X_3 \mid X_2$	
$X_1 \rightarrow X_2 \rightarrow X_3$	false	true	
$X_1 \leftarrow X_2 \leftarrow X_3$	false	true	← no v-structure
$X_1 \leftarrow X_2 \rightarrow X_3$	false	true	
$X_1 \rightarrow X_2 \leftarrow X_3$	true	false	← v-structure

- All DAGs in an equivalence class have the same skeleton and the same v-structures
- An equivalence class can be uniquely represented by a completed partially directed acyclic graph (CPDAG)