Statistics and Optimization for Causal Inference in Large-Scale Biological Systems

Peter Bühlmann

Seminar für Statistik, ETH Zürich

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
(from observations of a “steady-state system”)

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
(from observations of a “steady-state system”)

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

Examples

Policy making in economics

what would happen to an economic variable (e.g. “health costs”) when implementing a certain policy (e.g. “new health policy”) ?

James Heckman: Nobel Prize Economics 2000

Genomics

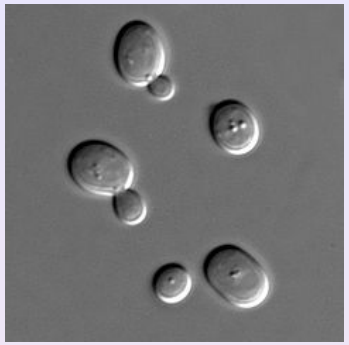
1. Flowering of arabidopsis thaliana

phenotype/response variable of interest:
$Y =$ days to bolting (flowering)
"covariates" $X =$ gene expressions from $p = 21'326$ genes

remark: "gene expression": process by which information from a gene is used in the synthesis of a functional gene product (e.g. protein)

question: infer/predict the effect of knocking-out/knocking-down (or enhancing) a single gene (expression) on the phenotype/response variable 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

$\rightsquigarrow$ consider the framework of an

intervention effect = causal effect

(mathematically defined $\rightsquigarrow$ see later)

Regression – the "statistical workhorse": the wrong approach

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

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

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

$|\beta_j|$ 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

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

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

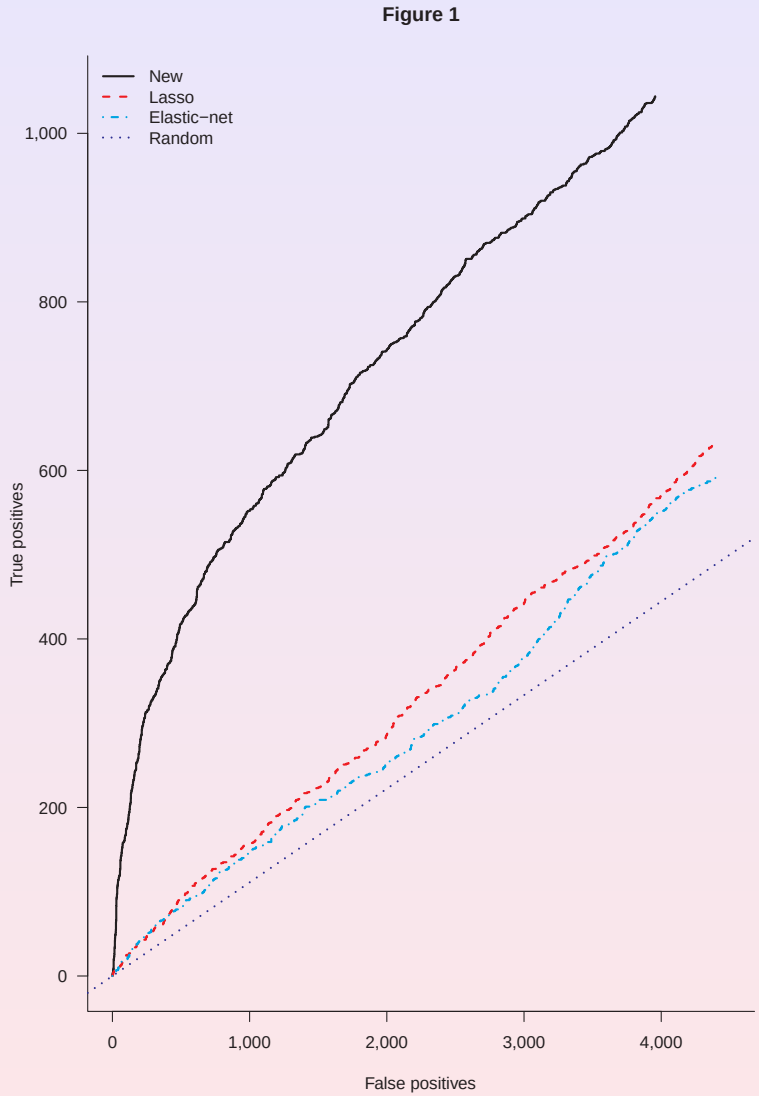
$$\mathrm{Var}(X^{(j)}) \equiv 1 \text{ for all } j$$

$|\beta_j|$ 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

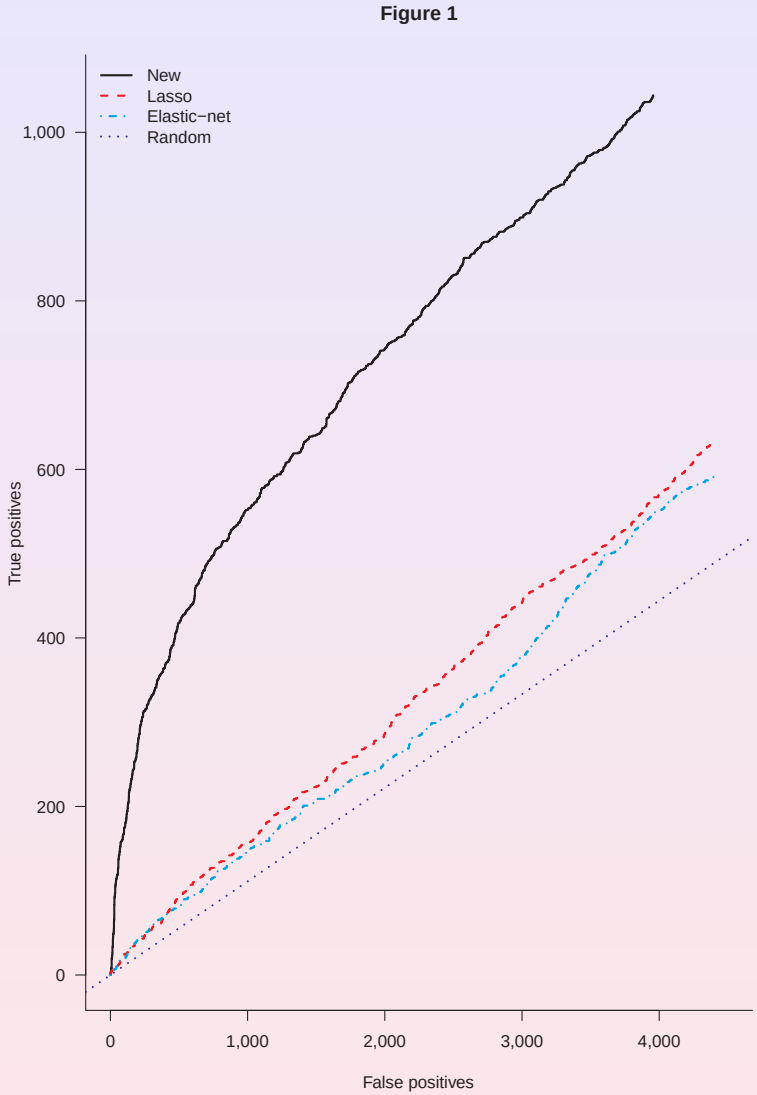
$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these others are not (cannot be) kept fixed

and indeed:



⇝ can do much better than (penalized) regression!

and indeed:



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

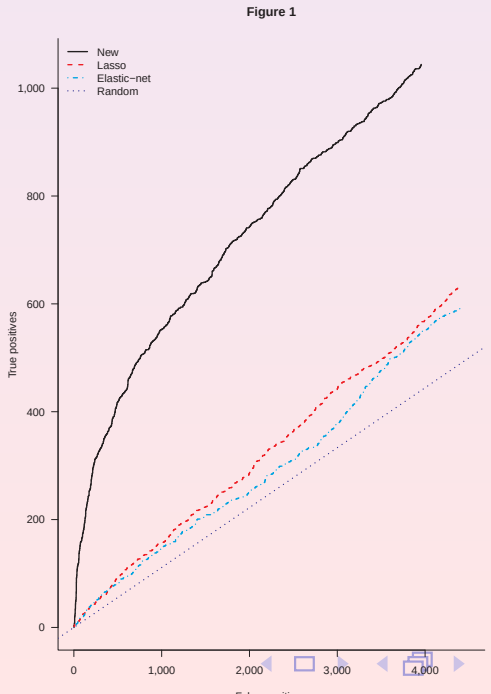
Effects of single gene knock-downs on all other genes (yeast)

(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

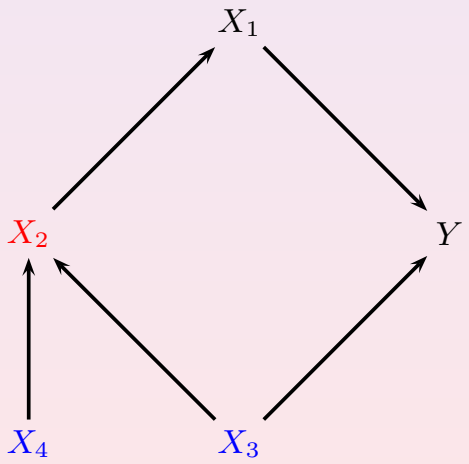
not the effect of $X^{(j)}$ on Y when keeping all other variables fixed (regression effect)

Intervention calculus (a review)

"dynamic" notion of an effect:
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 DAGs (Directed Acyclic Graphs)
random variables are represented as nodes in the DAG

assume a Markov condition, saying that

$X^{(j)}|X^{(\mathrm{pa}(j))}$ cond. independent of its non-descendant variables

$\rightsquigarrow$ 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)

for simplicity: just consider DAGs (Directed Acyclic Graphs)
random variables are represented as nodes in the DAG

assume a Markov condition, saying that

$X^{(j)}|X^{(\mathrm{pa}(j))}$ cond. independent of its non-descendant variables

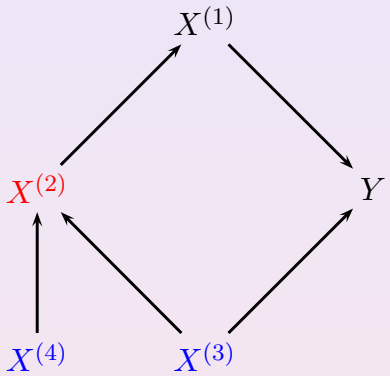
$\rightsquigarrow$ recursive factorization of joint distribution

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

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

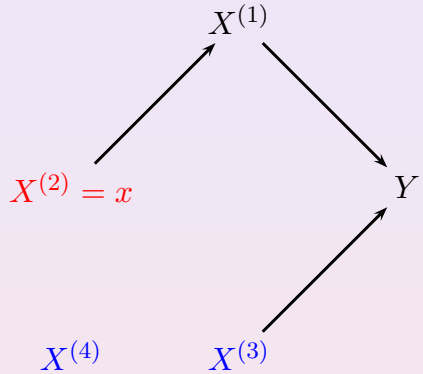
assume Markov property for causal DAG:

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 observational probability distribution P

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

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

recap:
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

recap:
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)

Inferring intervention effects from observational distribution

main problem: inferring DAG from observational data

impossible! can only infer equivalence class of DAGs (several DAGs can encode exactly the same conditional independence relationships)

Example:

X causes *Y*

Y causes *X*

and we cannot estimate causal/intervention effects from observational distribution

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 DAGs (CPDAG)
- find all DAG-members of true equivalence class (CPDAG):
  $D_1, \ldots, D_m$
- for every DAG-member D_r, and every variable $X^{(j)}$:
  single intervention effect $\theta_{r,j}$
  summarize them by

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

and we cannot estimate causal/intervention effects from observational distribution

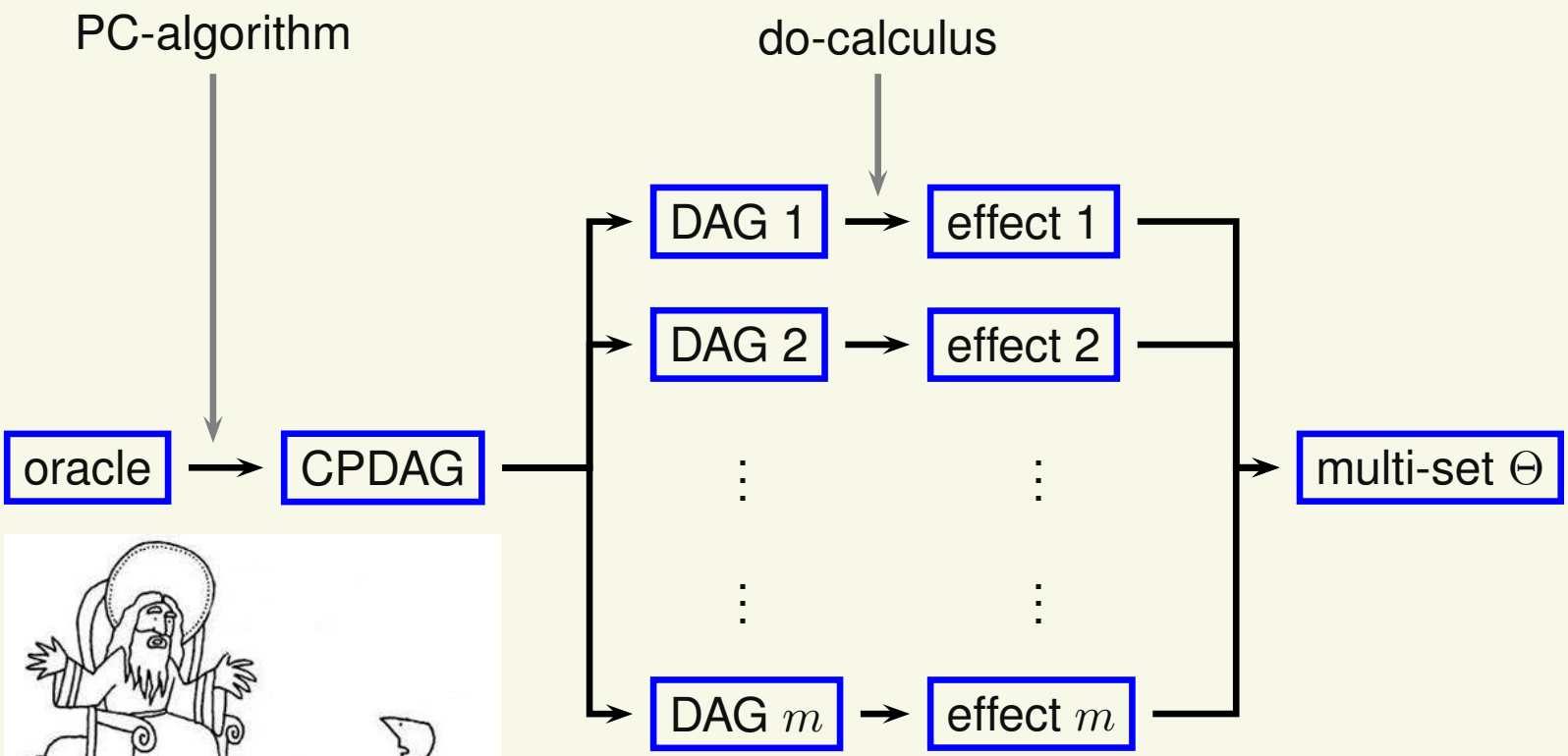
but we will be able to estimate lower bounds of causal effects

conceptual "procedure":

- probability distribution P from a DAG, generating the data
  $\rightsquigarrow$ true underlying equivalence class of DAGs (CPDAG)
- find all DAG-members of true equivalence class (CPDAG):
  $D_1, \ldots, D_m$
- for every DAG-member D_r, and every variable $X^{(j)}$:
  single intervention effect $\theta_{r,j}$
  summarize them by

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

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

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

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

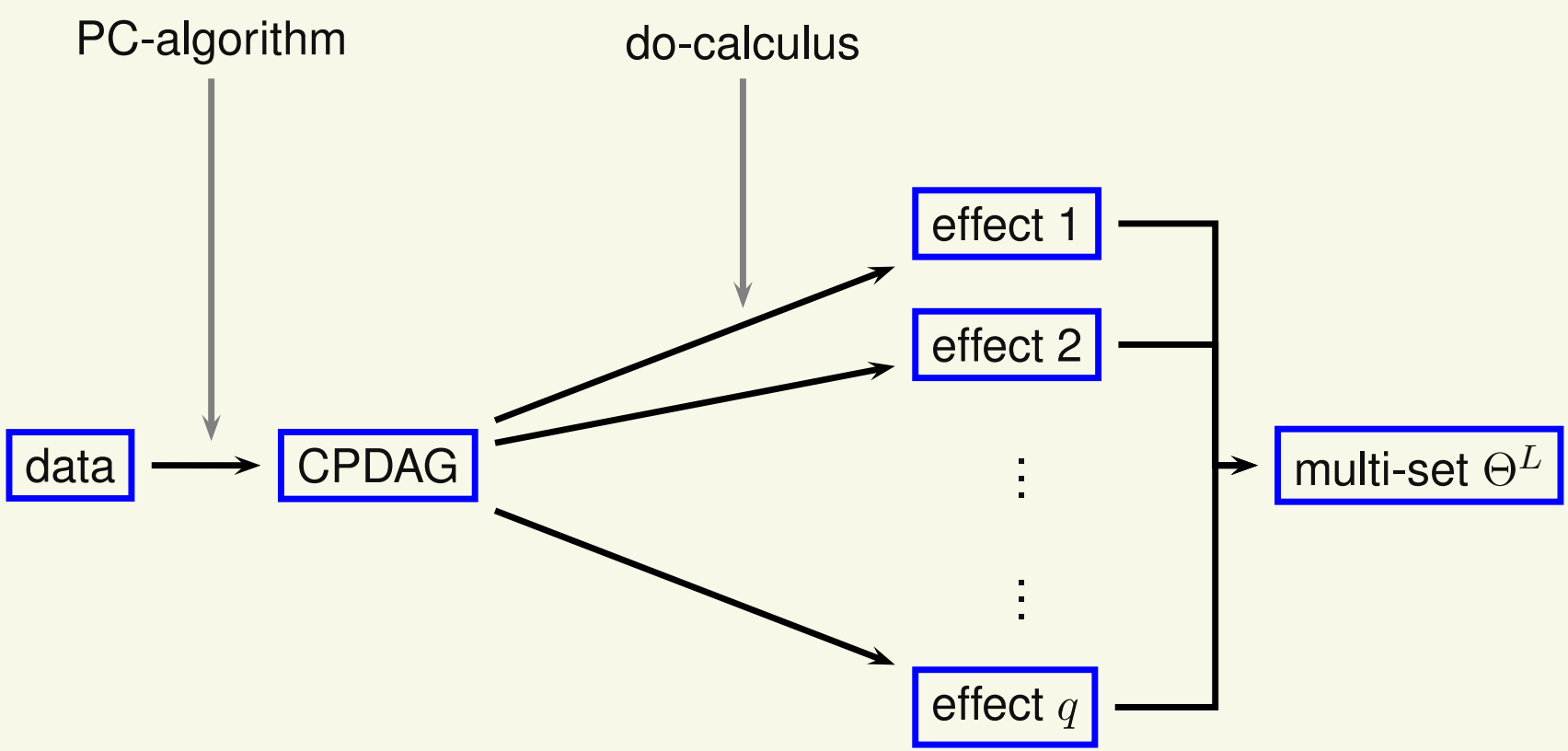
minimal absolute effect α_j is a lower bound for true absolute intervention effect

"Optimization" I: $\exists$ 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



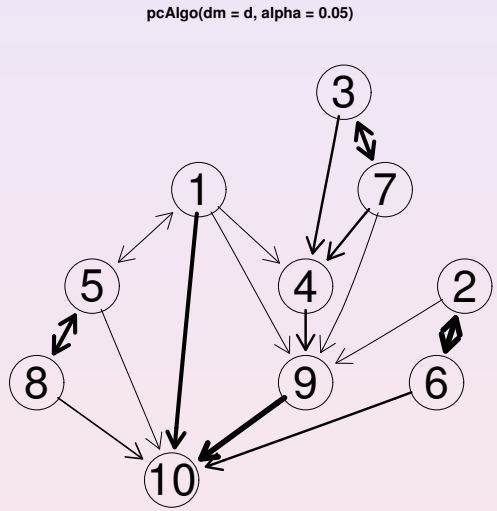
the local $\Theta^L = \Theta$ up to multiplicities

(Maathuis, Kalisch & PB, 2009)

Estimation from finitely many observational data

difficult part: estimation of CPDAG (equivalence class of DAGs)
$\rightsquigarrow$ estimation of structure

$$P \Rightarrow \underbrace{\text{CPDAG}}_{\text{equiv. class of DAGs}}$$



Two different approaches

- multiple statistical testing for conditional independencies
  PC-algorithm (Spirtes et al., 2000)
- score-based methods for penalized maximum likelihood estimator
  $\rightsquigarrow$ challenging issues in optimization

from now on:

absorb Y notationally into X (e.g. $Y = X_1$)

Two different approaches

- multiple statistical testing for conditional independencies
  PC-algorithm (Spirtes et al., 2000)
- score-based methods for penalized maximum likelihood estimator
  $\rightsquigarrow$ challenging issues in optimization

from now on:

absorb Y notationally into X (e.g. $Y = X_1$)

Statistical theory (Kalisch & PB, 2007; Maathuis, Kalisch & PB, 2009)

n i.i.d. observational data points; p variables
high-dimensional setting where $p \gg n$

assumptions:

- $X^{(1)}, \ldots, X^{(p)} \sim \mathcal{N}_p(0, \Sigma)$ Markov and faithful to true DAG
- high-dimensionality: $\log(p) \ll n$
- sparsity: maximal degree $d = \max_j |\mathrm{ne}(j)| \ll n$
- signal strength: non-zero (partial) correlations suff. large
  $\min\{|\rho_{i,j|S}|; \rho_{i,j|S} \neq 0,\ i \neq j, |S| \leq d\} \gg \sqrt{d \log(p)/n}$
- "coherence": maximal (partial) correlations $\leq C < 1$
  $\max\{|\rho_{i,j|S}|;\ i \neq j, |S| \leq d\} \leq C < 1$

Then:

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

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

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

The role of “sparsity” in causal inference

as usual: sparsity is necessary for accurate estimation in presence of noise

but here: “sparsity” (so-called protectedness) is crucial for identifiability as well

X causes Y Y causes X

cannot tell from observational data the direction of the arrow

the same situation arises with a full graph with more than 2 nodes

$\rightsquigarrow$

causal identification really needs “sparsity”
the better the “sparsity” the tighter the bounds for causal effects

Penalized maximum likelihood estimator and Optimization II

why another approach than multiple testing?
$\rightsquigarrow$ can be used for more general problems of inferring causal effects based on observational and ("a few") interventional data

n i.i.d. observational data points from $\mathcal{N}_p(0, \Sigma)$ which is Markov w.r.t. DAG D
$\rightsquigarrow$ write down the negative log-likelihood

$$-\ell(\Sigma, D; \text{data}) = \dots$$

unknown quantities are Σ and D

Penalized maximum likelihood estimator and Optimization II

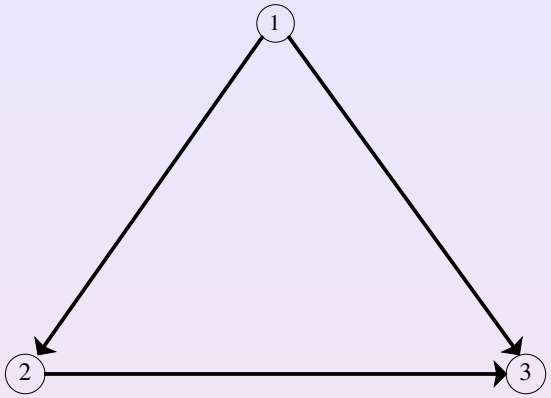
why another approach than multiple testing?
$\rightsquigarrow$ can be used for more general problems of inferring causal effects based on observational and ("a few") interventional data

n i.i.d. observational data points from $\mathcal{N}_p(0, \Sigma)$ which is Markov w.r.t. DAG D
$\rightsquigarrow$ write down the negative log-likelihood

$$-\ell(\Sigma, D; \text{data}) = \dots$$

unknown quantities are Σ and D

Gaussian DAG is Gaussian linear structural equation model:



$$X^{(1)} \leftarrow \varepsilon^{(1)}$$
$$X^{(2)} \leftarrow \beta_{21} X^{(1)} + \varepsilon^{(2)}$$
$$X^{(3)} \leftarrow \beta_{31} X^{(1)} + \beta_{32} X^{(2)} + \varepsilon^{(3)}$$

in general:

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

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

$\rightsquigarrow$ reparametrization

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

$\rightsquigarrow$ explicit form of likelihood

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

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

Challenges in optimization

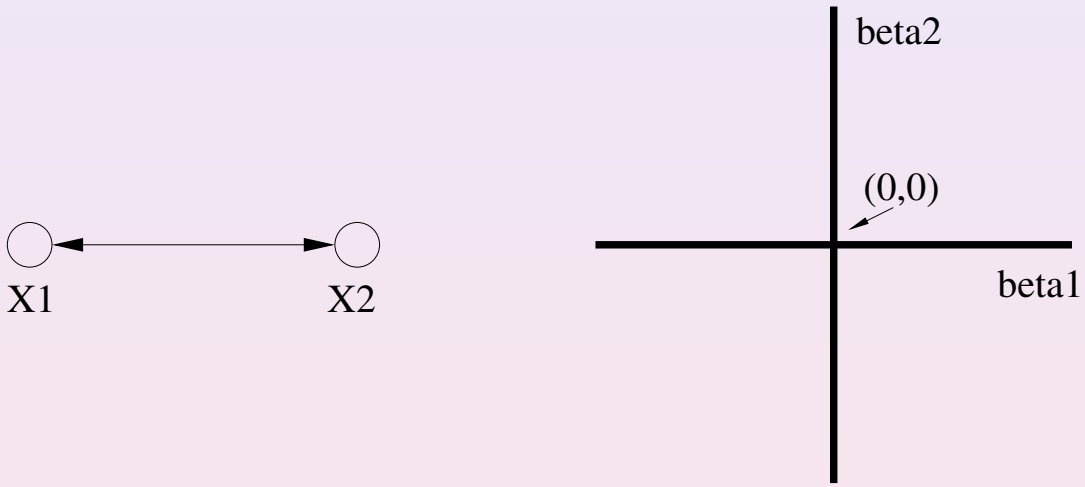
$$\begin{aligned}\hat{\Sigma},\ \hat{D} &= \text{argmin}_{\Sigma;D \text{ a DAG}} - \ell(\Sigma, D;\ \text{data}) + \lambda|D| \\ &= \text{argmin}_{B;\ \{\sigma_j^2;j\}} - \ell(B, \{\sigma_j^2;\ j\};\ \text{data}) + \lambda \underbrace{\|B\|_0}_{\sum_{ij} I(B_{ij}\neq 0)}\end{aligned}$$

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)

Toy-example

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



non-convex parameter space!
(no straightforward way to do convex relaxation, etc.)

Our computation: Greedy Interventional Equivalence Search

(Hauser & PB, 2011)

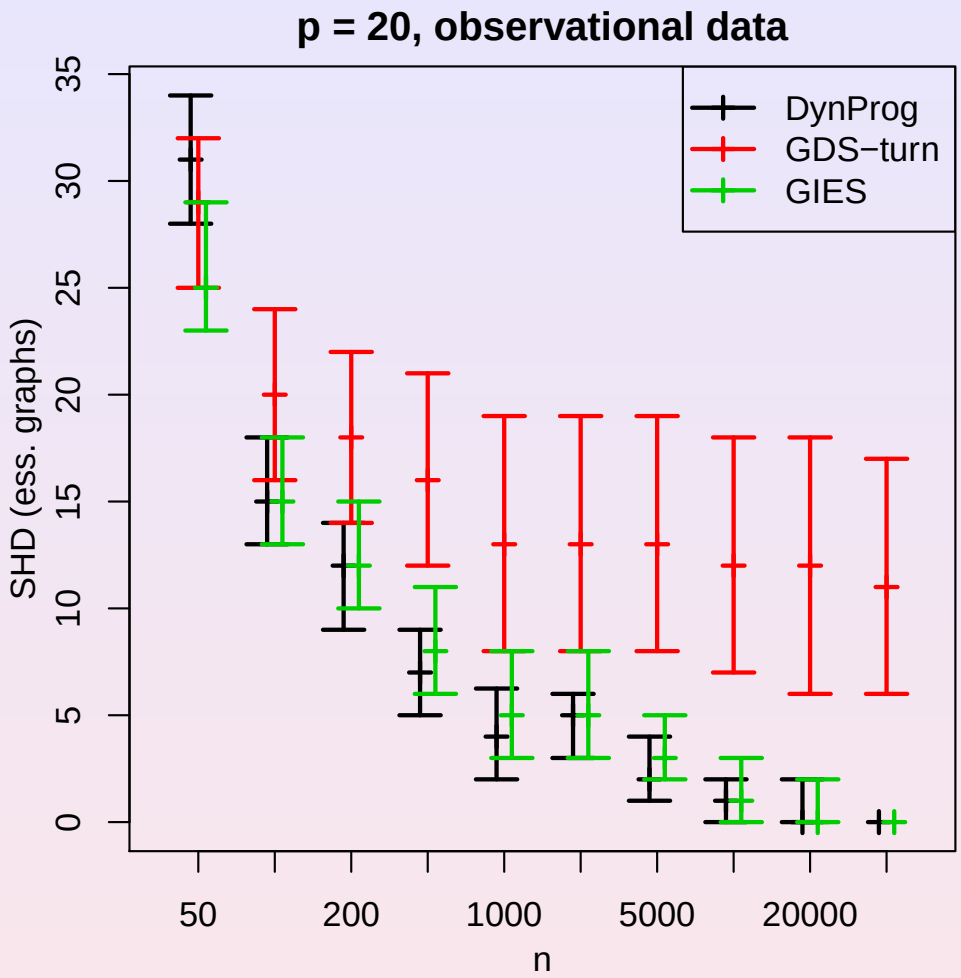
do greedy search over equivalent classes (cf. Chickering, 2002)
forward and backward and turning phase

forward:

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

this can be done efficiently without enumerating all members in the equivalence classes (Hauser & PB, 2011) – but it's non-trivial

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



greedy equivalent (class) search is

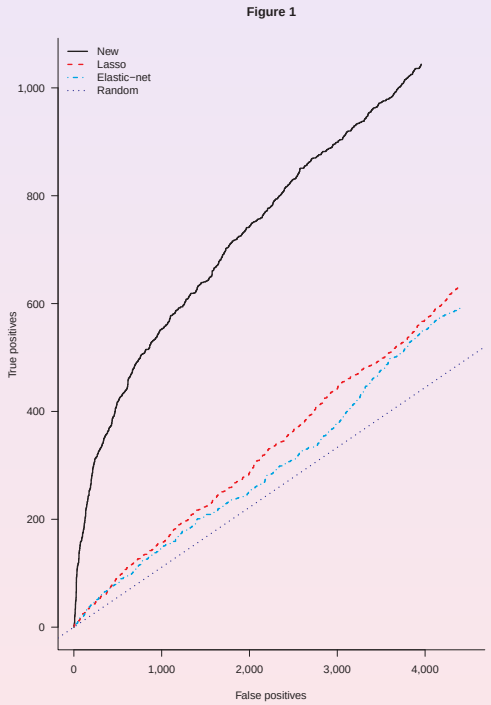
- much better than greedy search (over DAGs)
- and for small dimension as good as exhaustive search

Successes in biology

Effects of single gene knock-downs on all other genes in yeast (Maathuis, Colombo, Kalisch & PB, 2010)

$n = 63$

observational data



Arabidopsis thaliana (Stekhoven, Maathuis, Hennig & PB, 2011)

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

observational data with $n = 47$ and $p = 21'326$
$\rightsquigarrow$ lower bound estimates $\hat{\alpha}_j$ for causal effect of every gene/variable on Y (using the PC-algorithm)

apply stability selection (Meinshausen & PB, 2010)
$\rightsquigarrow$ assigning uncertainties via control of PCER (per comparison error rate)

Causal gene ranking

	Gene	summary rank	median effect	expression	error (PCER)	name
1	AT2G45660	1	0.60	5.07	0.0017	AGL20 (SOC1)
2	AT4G24010	2	0.61	5.69	0.0021	ATCSLG1
3	AT1G15520	2	0.58	5.42	0.0017	PDR12
4	AT3G02920	5	0.58	7.44	0.0024	replication protein-related
5	AT5G43610	5	0.41	4.98	0.0101	ATSUC6
6	AT4G00650	7	0.48	5.56	0.0020	FRI
7	AT1G24070	8	0.57	6.13	0.0026	ATCSLA10
8	AT1G19940	9	0.53	5.13	0.0019	AtGH9B5
9	AT3G61170	9	0.51	5.12	0.0034	protein coding
10	AT1G32375	10	0.54	5.21	0.0031	protein coding
11	AT2G15320	10	0.50	5.57	0.0027	protein coding
12	AT2G28120	10	0.49	6.45	0.0026	protein coding
13	AT2G16510	13	0.50	10.7	0.0023	AVAP5
14	AT3G14630	13	0.48	4.87	0.0039	CYP72A9
15	AT1G11800	15	0.51	6.97	0.0028	protein coding
16	AT5G44800	16	0.32	6.55	0.0704	CHR4
17	AT3G50660	17	0.40	7.60	0.0059	DWF4
18	AT5G10140	19	0.30	10.3	0.0064	FLC
19	AT1G24110	20	0.49	4.66	0.0059	peroxidase, putative
20	AT1G27030	20	0.45	10.1	0.0059	unknown protein

- biological validation by gene knockout experiments in progress.

red: biologically known genes responsible for flowering

in collaboration with Hennig and Gruissem lab, ETH Zurich:
performed validation experiment with mutants corresponding to these top 20 - 3 = 17 genes

- 14 mutants easily available $\rightsquigarrow$ only test for 14 genes
- more than usual: mutants showed low germination or survival...
- 9 among the 14 mutants survived (sufficiently strongly), i.e. 9 mutants for which we have an outcome
- 3 among the 9 mutants (genes) showed a significant effect for Y relative to the wildtype (non-mutated plant)

$\rightsquigarrow$ that is: besides the three known genes, we find three additional genes which exhibit a significant difference in terms of “time to flowering”

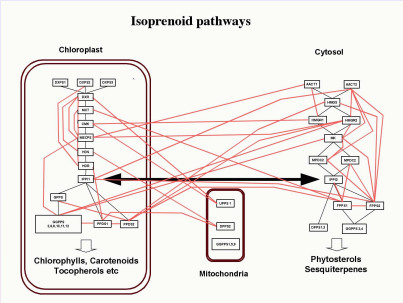
in short:

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

Conclusions

1. 1. Beware of over-interpretation!



so far, based on current data:
we can not reliably infer a causal network

despite theorems...
(perturbation of the data yields unstable networks)

2. Causal inference relies on subtle uncheckable(!) assumptions
$\rightsquigarrow$ experimental validations are important (simple organisms in biology are great for pursuing this!)

3. many technical issues in identifiability, high-dimensional statistical inference and optimization

4. but there is a clear potential:

for stable ranking/prediction of intervention/causal effects

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

Editorial in Nature Methods (April 2010)

this is extremely useful in computational biology

and in this sense:
“causal inference from observational data is much further developed than 30 years ago when it was thought to be impossible”

Thank you!

`R`-package: `pcalg`
(Kalisch, Mächler, Colombo, Maathuis & PB, 2010)

References:

- Hauser, A. and Bühlmann, P. (2011). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. Preprint arXiv:1104.2808v1.
- 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.