High-dimensional graphical modeling and causal inference

Peter Bühlmann

joint work with

Markus Kalisch

Marloes Maathuis

Caroline Uhler

Sara van de Geer

cannot do confirmatory causal inference without randomized intervention experiments...

but we can do better than proceeding naively

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

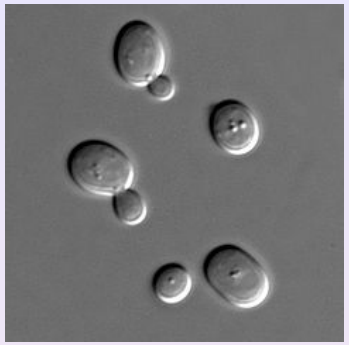
Genomics

1. Flowering of arabidopsis thaliana

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

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

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)

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

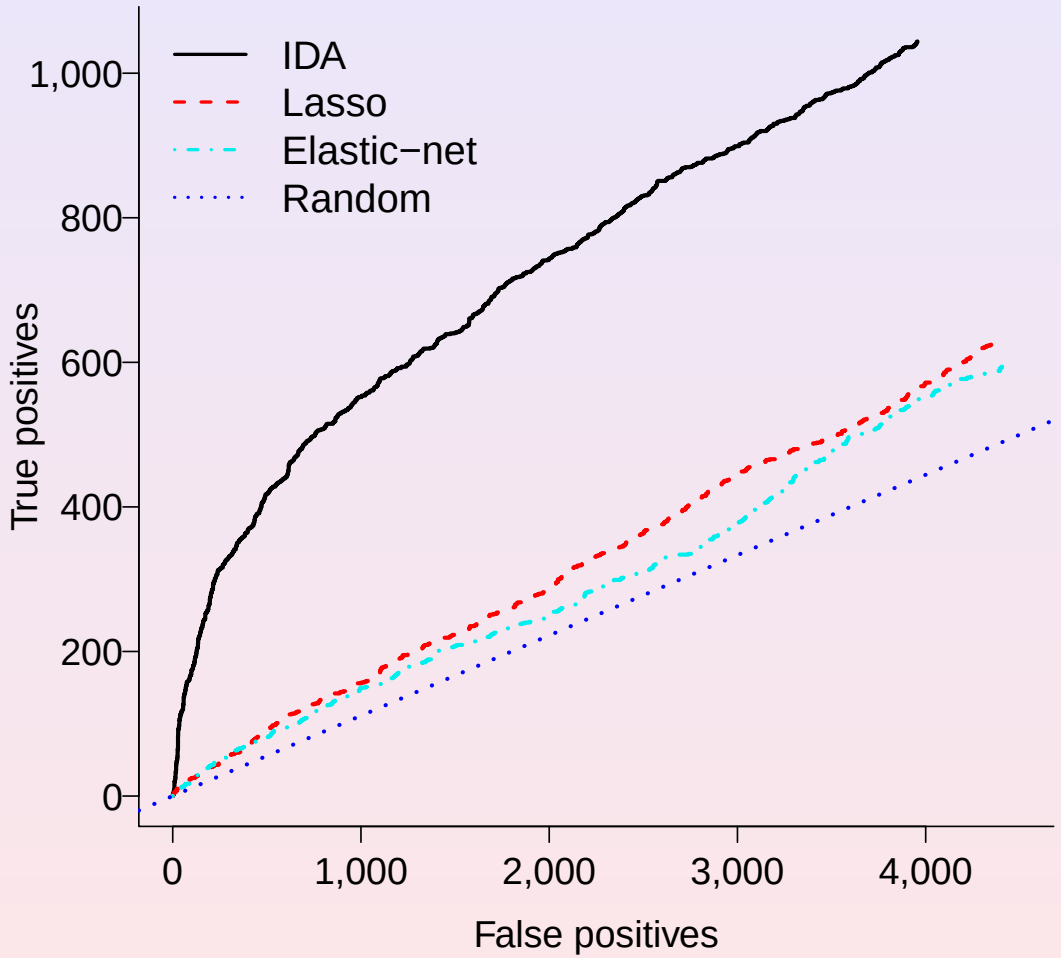
$$\mathrm{Var}(X^{(j)}) \equiv 1 \text{ for all } j$$

$|\beta_j|$ 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

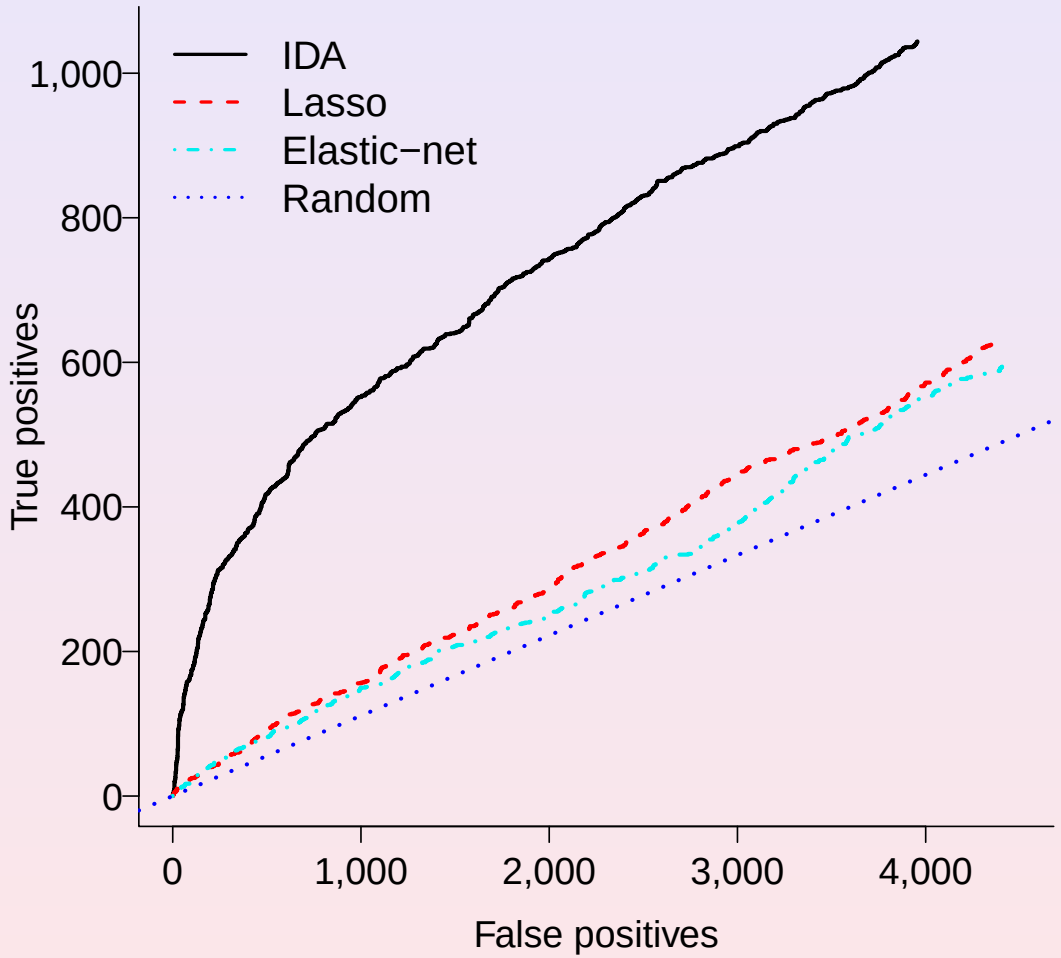
$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these others are not (cannot be) kept fixed

and indeed:



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

and indeed:



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

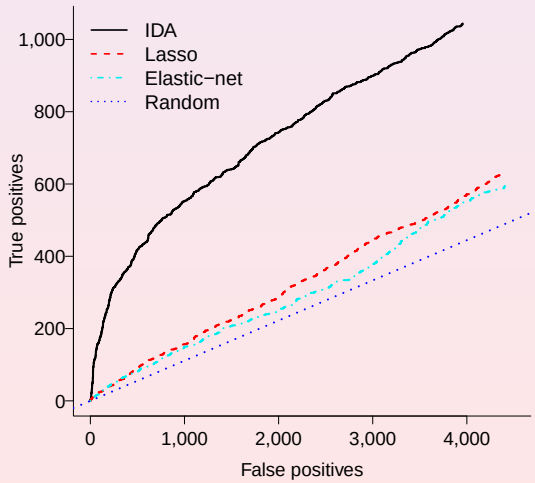
Effects of single gene knock-downs on all other genes (yeast)

(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



DAGs and causal effects

- univariate response Y
- p-dimensional covariate X

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

$$\mathrm{do}(X^{(j)} = x)$$

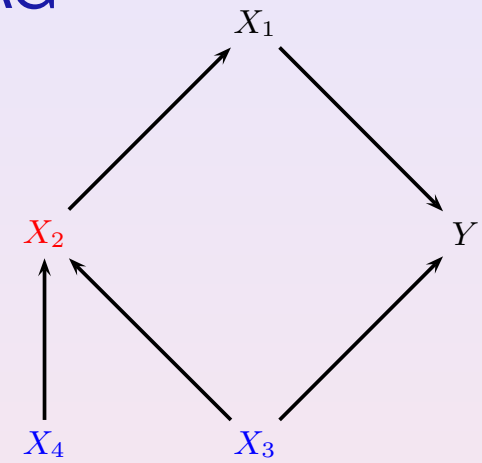
$\rightsquigarrow$ this is a question of intervention type

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,...

... a substantial machinery... (e.g. Pearl)

two main assumptions

- causal influence diagram is a DAG



relaxation allowing for cycles:

e.g. Hyttinen, Eberhardt & Hoyer (2010, 2012)

- there are no hidden (relevant) variables

relaxation including hidden variables:

e.g. Spirtes, Glymour & Scheines (2000); Colombo, Maathuis & Richardson (2012),...

if $Y, X^{(1)}, \ldots, X^{(p)} \sim \mathcal{N}_{p+1}(\mu, \Sigma)$
$\rightsquigarrow$ intervention (or causal) effect is a real-valued parameter

$$\theta_j \equiv \frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)] \quad \text{constant w.r.t. } x$$

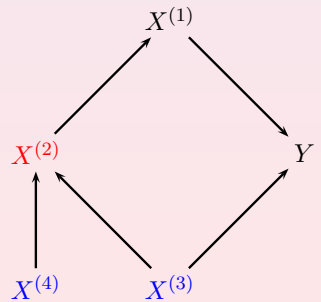
and the intervention effect parameter can be characterized as:

for $Y \notin \mathrm{pa}(j)$: θ_j is the regression parameter in

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

only need parental set and regression

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



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)

Inferring intervention effects from observational distribution

main problem: inferring DAG or parental set(s) 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*

a lot of work about identifiability:
Verma & Pearl (1991); Spirtes, Glymour & Scheines (1993); Tian & Pearl (2000–2002); Lauritzen & Richardson (2002); Shpitser & Pearl (2006–2011); vanderWeele & Robins (2007–2011); Drton, Foygel & Sullivant (2011);...

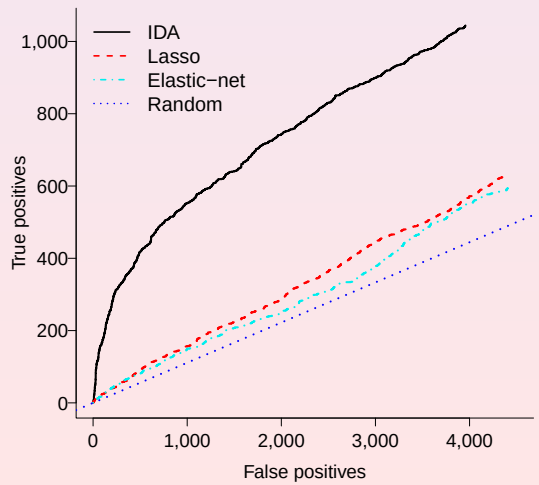
we cannot estimate causal/intervention effects from observational distribution

but can estimate "informative" lower bounds of causal eff. based on Markov equivalence class of DAGs

$$\text{true causal effect } |\theta_j| \geq \underbrace{\alpha_j}_{\text{identifiable}}$$

(Maathuis, Kalisch & PB, 2009)

R-package: `pcalg`
what we used in the yeast example to score importance of genes according to size of $\hat{\alpha}_j$



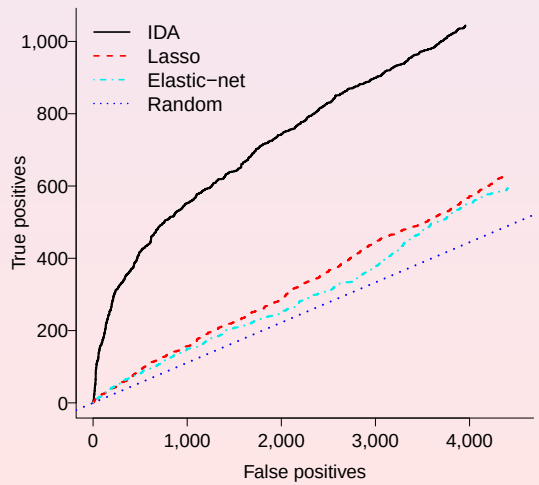
we cannot estimate causal/intervention effects from observational distribution

but can estimate "informative" lower bounds of causal eff. based on Markov equivalence class of DAGs

$$\text{true causal effect } |\theta_j| \geq \underbrace{\alpha_j}_{\text{identifiable}}$$

(Maathuis, Kalisch & PB, 2009)

`R`-package: `pcalg`
what we used in the yeast example to score importance of genes according to size of $\hat{\alpha}_j$



Estimation of the Markov equivalence class of DAGs

notation: drop the Y-notation ($Y = X^{(1)}, X^{(2)}, \ldots, X^{(p)}$)

goal: infer CPDAG (Markov equivalence class of DAGs)
$\rightsquigarrow$ "structure learning"

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

two main approaches:

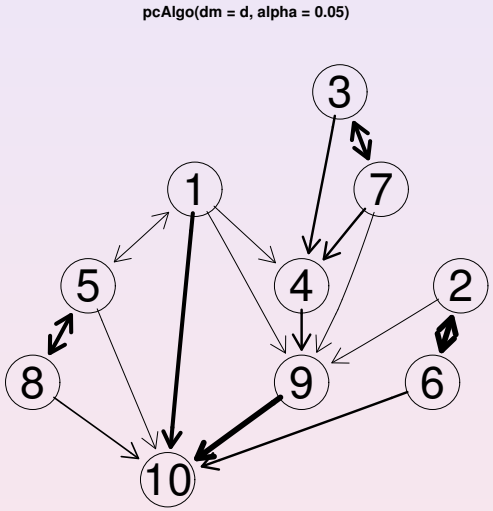
- multiple testing of conditional (in-)dependences
- score-based methods: MLE as prime example

Estimation of the Markov equivalence class of DAGs

notation: drop the Y-notation ($Y = X^{(1)}, X^{(2)}, \ldots, X^{(p)}$)

goal: infer CPDAG (Markov equivalence class of DAGs)
$\rightsquigarrow$ "structure learning"

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



two main approaches:

- multiple testing of conditional (in-)dependences
- score-based methods: MLE as prime example

Faithfulness assumption

for inferring CPDAG via conditional (in-)dependences ("essentially" necessary for conditional dependence testing approaches)

a distribution P is called faithful to a DAG D if all conditional independences can be inferred from the graph

(can infer some conditional independences from a Markov assumption; but we require here "all" conditional independences)

assuming faithfulness: $\rightsquigarrow$ can infer the CPDAG from a list of conditional (in-)dependence relations

Faithfulness assumption

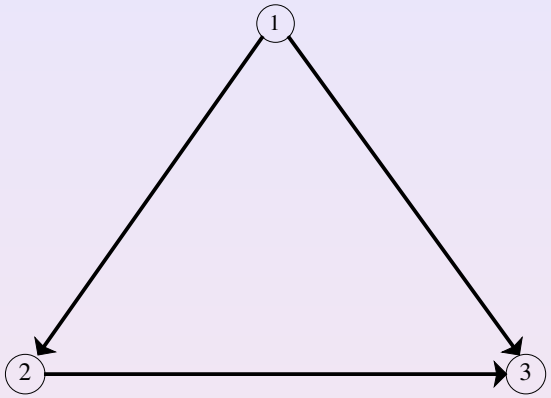
for inferring CPDAG via conditional (in-)dependences ("essentially" necessary for conditional dependence testing approaches)

a distribution P is called faithful to a DAG D if all conditional independences can be inferred from the graph

(can infer some conditional independences from a Markov assumption; but we require here "all" conditional independences)

assuming faithfulness: $\rightsquigarrow$ can infer the CPDAG from a list of conditional (in-)dependence relations

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. $\alpha = \beta = 1,\ \gamma = -1$

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

failure of faithfulness due to cancellation of coefficients

failure of exact faithfulness is “rare” (Lebesgue measure zero)

but for statistical estimation (in the Gaussian case):
“often” require strong faithfulness (Robins, Scheines, Spirtes & Wasserman, 2003):

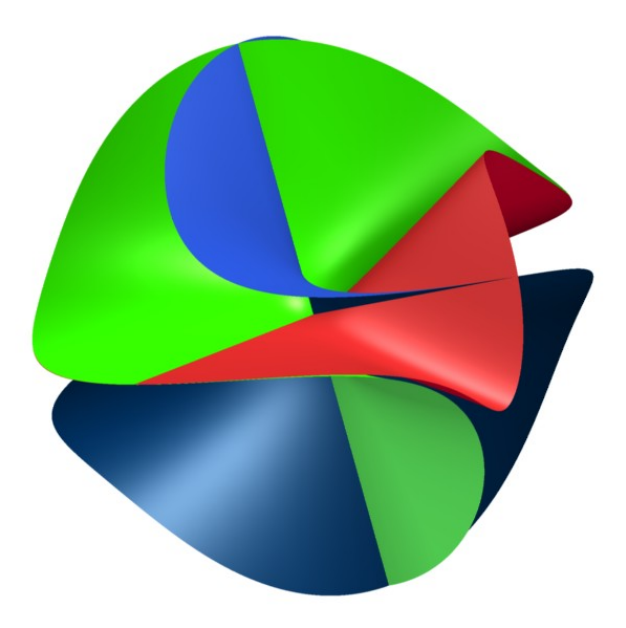
faithfulness &

$$\min\left\{|\rho(i,j|S)|;\ \rho(i,j|S)\neq 0,\ i\neq j,\ |S|\leq d\right\}\geq\tau,$$

$$\tau\asymp\sqrt{\log(p)/n}$$

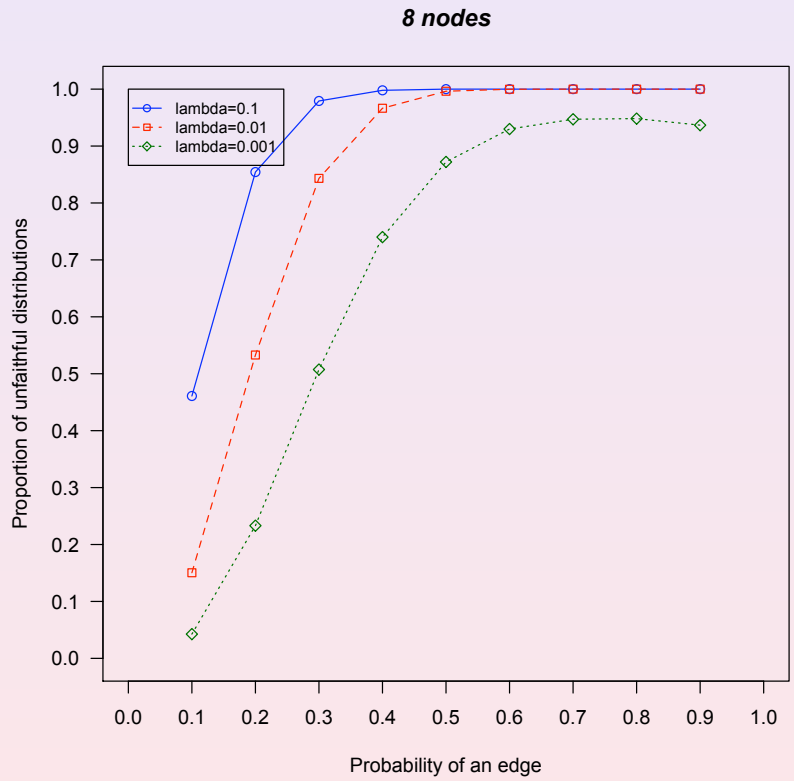
... strong faithfulness can be rather severe
(Uhler, Raskutii, PB & Yu, 2012)

3 nodes, full graph



unfaithful distributions
due to exact cancellation

8 nodes, varying sparsity



$\mathbb{P}$[not strongly faithful]

Estimating the CPDAG: the PC-algorithm (Spirtes & Glymour, 1991)

- crucial assumption:
  distribution P (strongly) faithful to the true underlying DAG
- less crucial but convenient:
  Gaussian assumption for $X^{(1)}, \ldots, X^{(p)} \rightsquigarrow$ can work with partial correlations

- input: $\hat{\Sigma}_{\mathrm{MLE}}$
  but we only need to consider many small sub-matrices of it (assuming sparsity of the graph)
- output: based on a clever data-dependent (random) sequence of multiple tests
  estimated CPDAG

Statistical theory for PC-algorithm
(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 $d = \max_j |\mathrm{ne}(j)| \ll n$
- "coherence": maximal (partial) correlations $\leq C < 1$
  $\max\{|\rho_{i,j|S}|;\ i \neq j, |S| \leq d\} \leq C < 1$
- signal strength/strong faithfulness:
  $\min\{|\rho_{i,j|S}|; \rho_{i,j|S} \neq 0,\ i \neq j, |S| \leq d\} \gg \sqrt{d \log(p)/n}$

Then, for some suitable tuning param. and $0 < \delta < 1$:

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

(Restricted) strong-faithfulness (Uhler, Raskutti, PB & Yu, 2012)

strong-faithfulness: faithfulness &

$$\min\left\{|\rho(i,j|S)|;\ \rho(i,j|S)\neq 0,\ i\neq j,\ |S|\leq d\right\}\geq\tau$$

$\tau \asymp \sqrt{d\log(p)/n}$, $d =$ max. degree of DAG

sufficient and necessary for PC-/conservative PC-algorithm:
restricted strong-faithfulness

1. adjacency strong-faithfulness

$$\min\left\{|\rho(i,j|S)|;\ \rho(i,j|S)\neq 0,\ (i,j)\in E,\ |S|\leq d\right\}\geq\tau$$

2. orientation strong-faithfulness

$$\min\left\{|\rho(i,j|S)|;\ (i,j,S)\in \mathrm{neigh}\right\}\geq\tau$$

$\mathrm{neigh} = \{(i,j,S);\ i,j$ not adjacent, (i,j,k) unshielded triple with i,j not d-separated by $S\}$

goal: understand

$$p(\tau) = \mathbb{P}[\text{failure of } \tau \text{ restricted strong-faithfulness}]$$

when edge weights β_{jk} (for edge $j \to k$) i.i.d. $\mathrm{Uniform}([-1,1])$

results (Uhler, Raskutti, PB & Yu, 2012):

- upper bound:

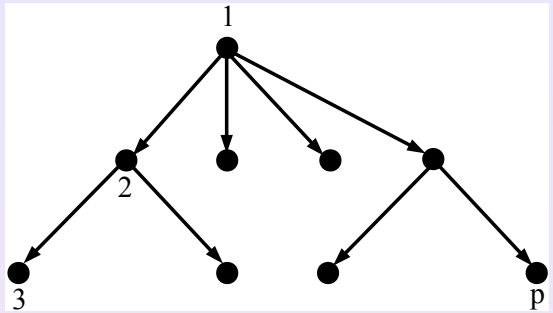
$$\begin{aligned} & p(\tau) \\ \leq \quad & C_1 C_2(|E|) \underbrace{\kappa^k}_{\max_{i,j,S} \mathrm{Var}(X^{(i)}|X^{(S)})} \tau^k \underbrace{\sum_{i,j,S} \deg(\mathrm{Cov}(\mathrm{X}^{(\mathrm{i})}, \mathrm{X}^{(\mathrm{j})}|\mathrm{X}^{(\mathrm{S})}))}_{\text{often large}} \end{aligned}$$

  k depends on polynomials character. strict unfaithfulness

- lower bounds
  - for trees
  - for cycles
  - for bi-partite graphs

Trees:

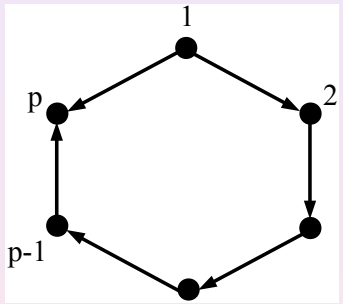
$$p(\tau) \geq 1 - (1 - \tau)^{p-1}$$



Cycles:

$$p(\tau) \geq 1 - (1 - \tau)^{3p-2}$$

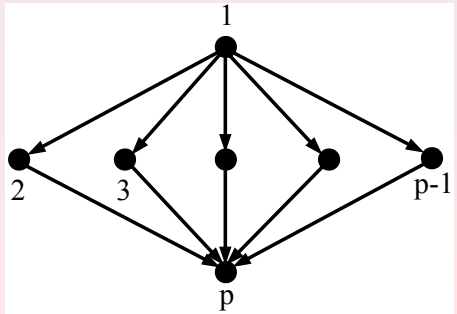
$\rightsquigarrow$ similar regime as trees



Bipartite graphs:

$$p(\tau) \geq 1 - (1 - \tau)^{(p-2)(2^{p-3}+1)}$$

$\rightsquigarrow$ a “disaster”...!



"most favorable" case: trees

$$p(\tau) \geq 1 - (1 - \tau)^{p-1}$$

with $\tau = \sqrt{\log(p)/n}$ (for bounded degree trees) $\rightsquigarrow$

$\mathbb{P}[\tau$ restricted strong-faithfulness holds$] \to 1 \Rightarrow p = o(\sqrt{n})$

- due to necessity of restricted strong-faithfulness
- assuming framework with i.i.d. sampling of edge weights (Uniform, Gaussian, Laplace,...)

$\Rightarrow$ cannot achieve high-dimensional consistency of PC-algorithm (conditional independence testing approaches) without further conditions

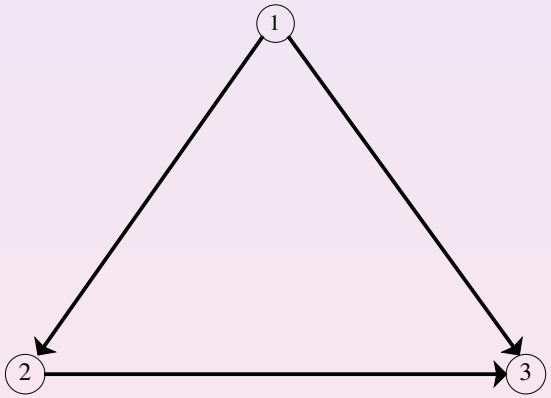
(e.g. saying that non-zero edge weights are "very" large)

Maximum likelihood estimation

without requiring strong faithfulness!


R.A. Fisher

Gaussian DAG is Gaussian linear structural equation model:



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

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

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

$$X = BX + \varepsilon$$

non-zeroes of B $\Rightarrow$ knowledge of the corresponding DAG

if we would know the order of the variables
$\rightsquigarrow$ (high-dimensional) multivariate regression

but we don't know the order of the variables:

- can only identify equivalence class of B's $\rightarrow$ "`obvious`"
- neg. log-likelihood is non-convex fct.(B) $\rightarrow$ `next slides`
- learning of ordering has large complexity (in general p!)

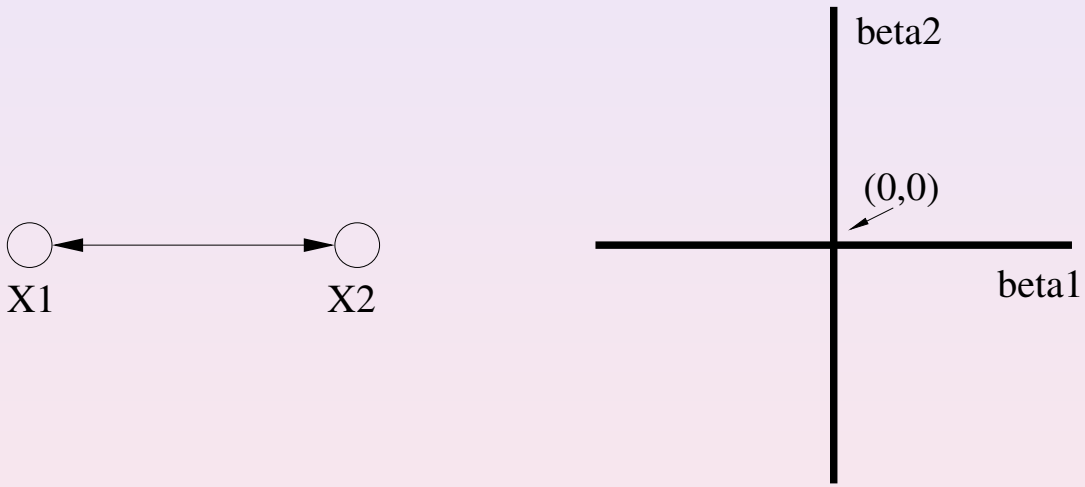
ℓ_0-penalized MLE

$$\hat{B}, \{\hat{\sigma}_j^2\} = \text{argmin}_{B;\ \{\sigma_j^2\}} - \ell(B, \{\sigma_j^2\};\ \text{data}) + \lambda \underbrace{\|B\|_0}_{\sum_{jk} I(B_{jk} \neq 0)}$$

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

Toy-example

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



non-convex parameter space!
(convex relaxation? $\rightarrow$ `see discussion`)

Why ℓ_0-penalty?

- ensures the same score for Markov-equivalent structures (this would not be true when using ℓ_1-norm penalty)
- ℓ_0-penalty leads to decomposable score

$$\text{score}(D, \mathbf{X}) = \sum_{j=1}^{p} g_j(\mathbf{X}^{(j)}, \mathbf{X}^{(\text{pa}_D(j))})$$

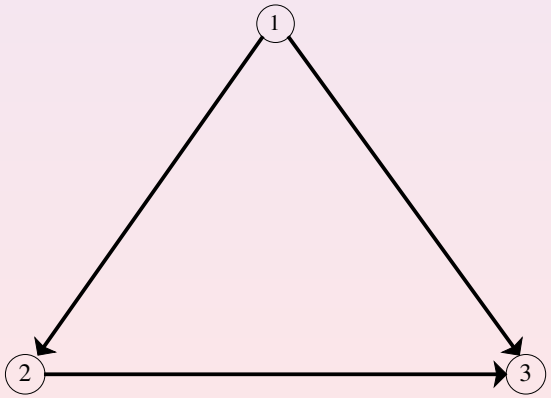
$\rightsquigarrow$ dynamic programming for computation if $p \approx 20 - 30$
(not easily possible with ℓ_1-norm penalization)
recall that the estimation problem is non-convex...

Statistical properties for ℓ_0-penalized MLE (van de Geer & PB, 2012)

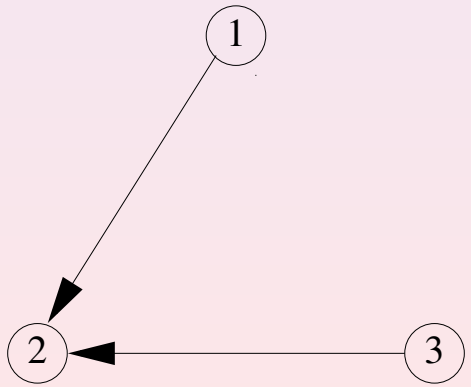
the target:

ℓ_0-penalized MLE estimates a DAG with fewest edges which represents the true distribution: minimal edges I-MAP

in the Gaussian linear structural eqn. model:
$\mathrm{Cov}(X) = \Sigma = (I - B)^{-1}\Omega(I - B)^{-T}$, $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$
true Σ^0 of data-generating distribution
minimal edges I-MAP: a DAG and corresponding B^0, Ω^0 such that $\Sigma^0 = (I - B^0)^{-1}\Omega^0(I - B^0)^{-T}$
not unique (in general)



non-faithful distribution where $\mathrm{Cov}(X^{(1)}, X^{(3)}) = 0$



minimal edges I-MAP

no faithfulness required for inferring minimal edges I-MAP
$\rightsquigarrow$ no strong-faithfulness required either

and when assuming faithfulness:

$$\begin{aligned} &\text{equivalence class of minimal edges I-MAP} \\ =\ &\text{(usual) Markov equivalence class} \end{aligned}$$

without requiring strong-faithfulness!

main condition required for ℓ_0-penalized MLE:
permutation beta-min condition
for an ordering of the variables
i.e. permutation $\pi : \{1, \ldots, p\} \to \{1, \ldots, p\}$
consider regressions of

$$X_{\pi(j)} \text{ versus } X_{\pi(j-1)}, \ldots, X_{\pi(1)} \quad \text{(Gram-Schmidt)}$$

$\rightsquigarrow$ coefficients $B^0(\pi)$
(for a true ordering π^0: $B^0(\pi^0)$ is most ℓ_0-sparse)

permutation beta-min condition:
for any π, "most of non-zero" $|B^0_{jk}(\pi)|$ are sufficiently large

technically: for any π
$(1 - \eta)$ $\underbrace{s(\pi)}$ edges (j, k) with
no. of edges in $B^0(\pi)$

$$|B^0_{jk}(\pi)| > \sqrt{\log(p)/n}\underbrace{(\sqrt{p/s_0} \vee 1)}_{\text{typically} \asymp 1}/\eta_0$$

example:
AR(1) (and AR(k) with fixed k) model satisfy the permutation beta-min condition

AR(1) model is a chain, i.e., a tree with maximal degree = 2
$\rightsquigarrow$ still "bad" in terms of strong faithfulness

Main results (van de Geer & PB, 2012)

assume permutation beta-min condition (and other "mild conditions")
then:

- with high probability: for $\lambda^2 \asymp \log(p)/n$

$$\|\hat{B} - B^0(\hat{\pi})\|_F^2 + \|\hat{\Omega} - \Omega^0(\hat{\pi})\|_F^2 = O(\lambda^2 s_0)$$

$$s_0 = \text{ no. of edges in minimal edges I-MAP}$$

- number of estimated edges is in the correct order of magnitude

$$\hat{s} \asymp s_0$$

- exact edge recovery of minimal edges I-MAP: our result "essentially requires" $p = o(\sqrt{n/\log(n)})$
  (which is the best case regime for strong faithfulness condition)

no strong faithfulness condition!

Main results (van de Geer & PB, 2012)

assume permutation beta-min condition (and other "mild conditions")
then:

- with high probability: for $\lambda^2 \asymp \log(p)/n$

$$\|\hat{B} - B^0(\hat{\pi})\|_F^2 + \|\hat{\Omega} - \Omega^0(\hat{\pi})\|_F^2 = O(\lambda^2 s_0)$$

$$s_0 = \text{ no. of edges in minimal edges I-MAP}$$

- number of estimated edges is in the correct order of magnitude

$$\hat{s} \asymp s_0$$

- exact edge recovery of minimal edges I-MAP: our result "essentially requires" $p = o(\sqrt{n/\log(n)})$
  (which is the best case regime for strong faithfulness condition)

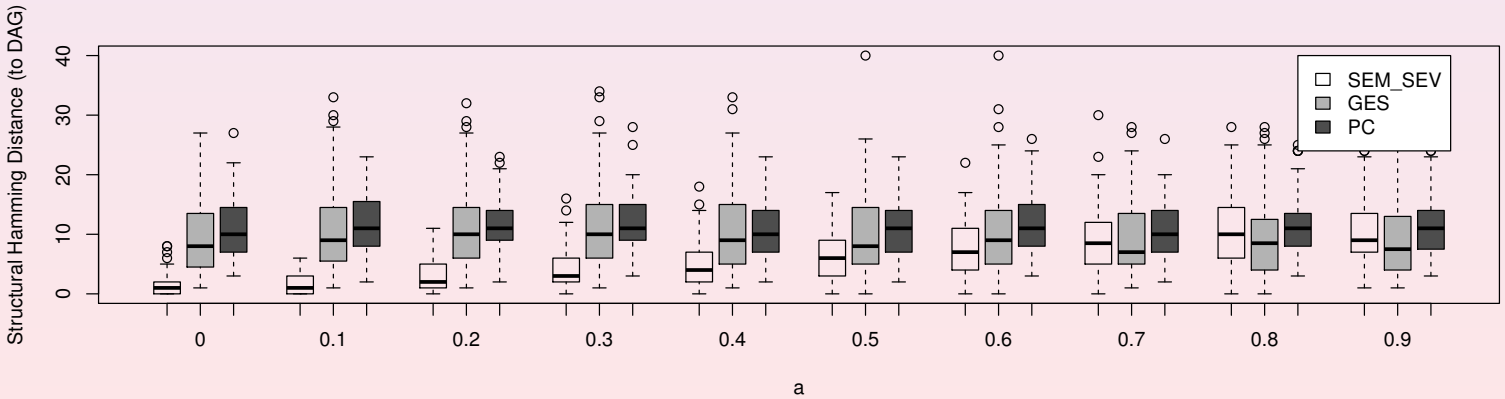
no strong faithfulness condition!

improvement for linear structural equation model with same error variances, for the regime $p = o(n/\log(n))$

$$X^{(j)} \leftarrow \sum_{k \in \mathrm{pa}(j)} B_{jk} + \varepsilon^{(j)}, \ \mathrm{Var}(\varepsilon^{(j)}) \equiv \omega^2 \ \ \text{(i.e. } \Omega = \omega^2 I\text{)}$$

only "standard" beta-min condition instead of permutation beta-min condition:
"most of non-zero" $|B^0_{jk}(\pi^0)|$ are sufficiently large
instead for all $\pi \rightsquigarrow$ only for true ordering π^0

and we have supporting empirical results to quantify the improvement if error variances are "approximately the same"

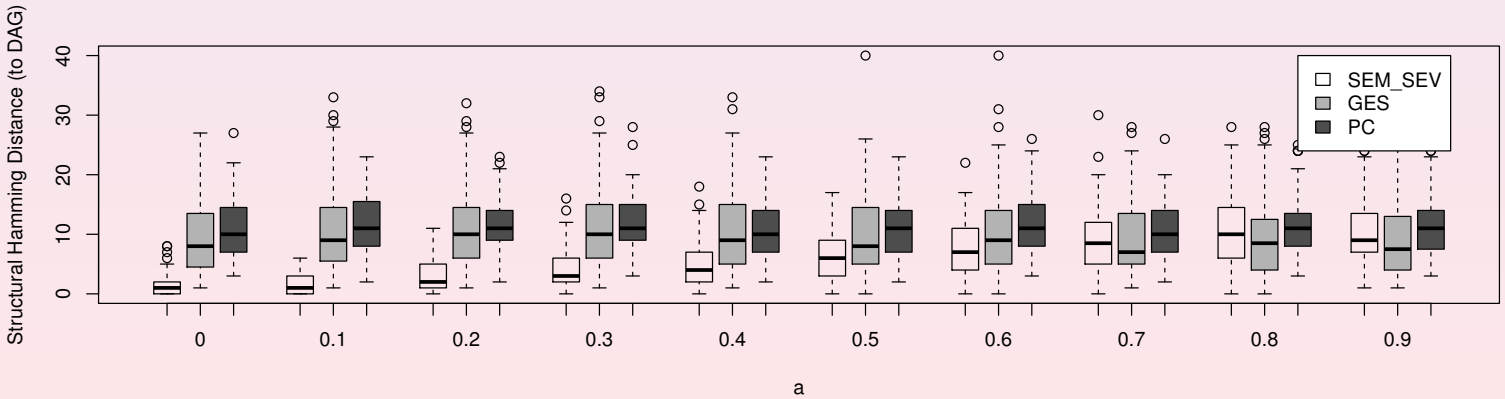


improvement for linear structural equation model with same error variances, for the regime $p = o(n/\log(n))$

$$X^{(j)} \leftarrow \sum_{k \in \mathrm{pa}(j)} B_{jk} + \varepsilon^{(j)}, \ \mathrm{Var}(\varepsilon^{(j)}) \equiv \omega^2 \ \ \text{(i.e. } \Omega = \omega^2 I\text{)}$$

only "standard" beta-min condition instead of permutation beta-min condition:
"most of non-zero" $|B^0_{jk}(\pi^0)|$ are sufficiently large
instead for all $\pi \rightsquigarrow$ only for true ordering π^0

and we have supporting empirical results to quantify the improvement if error variances are "approximately the same"



Route via structural equation models: many interesting 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 (2012)
- 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)

  e.g. $X^{(j)} \leftarrow \sum_{k \in \mathrm{pa}(j)} f_k(X^{(k)}) + \varepsilon^{(j)}$ (additive strctl. eqns.)
  Nowzohour & PB (in progress)
- linear structural eqns. with non-Gaussian errors:
  $\ldots$, at least one $\varepsilon^{(j)}$ non-Gaussian
  Shimizu (2006)

Route via structural equation models: many interesting 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 (2012)
- 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)

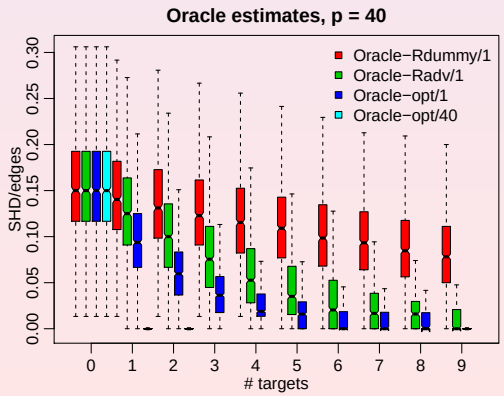
  e.g. $X^{(j)} \leftarrow \sum_{k \in \mathrm{pa}(j)} f_k(X^{(k)}) + \varepsilon^{(j)}$ (additive strctl. eqns.)
  Nowzohour & PB (in progress)
- linear structural eqns. with non-Gaussian errors:
  $\ldots,$ at least one $\varepsilon^{(j)}$ non-Gaussian
  Shimizu (2006)

Observational-interventional data

increase identifiability from (randomized) interventional data
combination of observational-interventional data is common in e.g. biology

yeast example:
63 observational and 231 interventional data

- MLE for Gaussian observational-interventional data (Hauser & PB, 2012a)
- active learning by choosing sequentially the next best intervention for identifying the true DAG
  (and solving the Eberhardt conjecture) (Hauser & PB, 2012b)



Concluding discussion

1. we have achieved some success in biology applications
(simple organisms: yeast and arabidopsis thaliana)
$\rightsquigarrow$ but there seems ample room for improvement

2. methods based on inferring conditional independences necessarily require version of strong faithfulness
(e.g. PC-algorithm)
$\rightsquigarrow$ restrictive in term of dimensionality

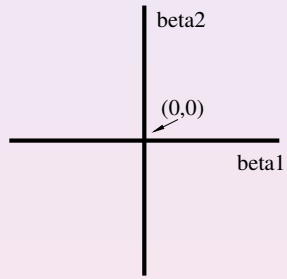
3. route via structural equation models does not require strong faithfulness;
and "natural restrictions" lead to full identifiability!

the price to pay with MLE (and other estimators?) for structural equation models: computation!

re-consider (penalized) MLE for linear Gaussian case:

$$\begin{aligned} \text{model:} \quad & X = BX + \varepsilon \\ \text{penalized MLE:} \quad & \hat{B}, \hat{\Omega} = \text{argmin}_{B,\Omega} \ell(B, \Omega; \text{data}) + \lambda \text{pen}(B) \end{aligned}$$

under non-convex constraint
of no directed cycles



can we do efficient convex relaxation for

$$\mathcal{S} = \{\Pi(I - B)^{-1}\Omega(I - B)^{-T}\Pi^T; \Pi \text{ perm.}, \ B \text{ lower triang.}, \ \Omega\}$$?

so far, our solution:

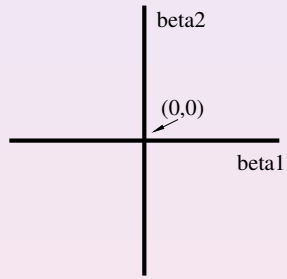
- dynamic programming if $p \approx 20 - 30$
- greedy equivalence class search if p is large
  (only ad-hoc... but reasonable results)

the price to pay with MLE (and other estimators?) for structural equation models: computation!

re-consider (penalized) MLE for linear Gaussian case:

$$\text{model:} \quad X = BX + \varepsilon$$
$$\text{penalized MLE:} \quad \hat{B}, \hat{\Omega} = \text{argmin}_{B,\Omega} \ell(B, \Omega; \text{data}) + \lambda \text{pen}(B)$$

under non-convex constraint
of no directed cycles



can we do efficient convex relaxation for

$$\mathcal{S} = \{\Pi(I - B)^{-1}\Omega(I - B)^{-T}\Pi^T; \Pi \text{ perm.}, \ B \text{ lower triang.}, \ \Omega\} \ ?$$

so far, our solution:

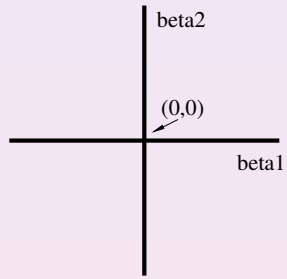
- dynamic programming if $p \approx 20 - 30$
- greedy equivalence class search if p is large
  (only ad-hoc... but reasonable results)

the price to pay with MLE (and other estimators?) for structural equation models: computation!

re-consider (penalized) MLE for linear Gaussian case:

$$\begin{aligned} \text{model:} \quad & X = BX + \varepsilon \\ \text{penalized MLE:} \quad & \hat{B}, \hat{\Omega} = \text{argmin}_{B,\Omega} \ell(B, \Omega; \text{data}) + \lambda \text{pen}(B) \end{aligned}$$

under non-convex constraint
of no directed cycles



can we do efficient convex relaxation for

$$\mathcal{S} = \{\Pi(I - B)^{-1}\Omega(I - B)^{-T}\Pi^T; \Pi \text{ perm.}, \ B \text{ lower triang.}, \ \Omega\} \ ?$$

so far, our solution:

- dynamic programming if $p \approx 20 - 30$
- greedy equivalence class search if p is large (only ad-hoc... but reasonable results)

Thank you!

`R`-package: `pcalg`
(Kalisch, Mächler, Colombo, Maathuis & PB, 2012)

References:

- van de Geer, S. and Bühlmann, P. (2012). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. arXiv:1205.5473v1
- Peters, J. and Bühlmann, P. (2012). Identifiability of Gaussian structural equation models with same error variances. arXiv:1205.2536v1
- Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B. (2012). Geometry of faithfulness assumption in causal inference. arXiv:1207.0547v2 (To appear in the Annals of Statistics).
- Kalisch, M., Mächler, M., Colombo, D., Maathuis, M.H. and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. Journal of Statistical Software 47 (11), 1-26.
- Stekhoven, D.J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M.H. and Bühlmann, P. (2011). Causal stability ranking. Bioinformatics 28, 2819-2823.
- Hauser, A. and Bühlmann, P. (2012). Two optimal strategies for active learning of causal models from interventions. Proc. of the 6th European Workshop on Probabilistic Graphical Models (PGM 2012), pp. 123-130, 2012.
- 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.
- 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.

Non-equal error variances

SEV-method $\rightsquigarrow \hat{D}$; completion to Markov-equivalence class $\rightsquigarrow \mathcal{E}(\hat{D})$

performance of $\mathcal{E}(\hat{D})$ for true CPDAG