High-dimensional variable selection: from association to intervention

Peter Bühlmann

Seminar für Statistik, ETH Zürich

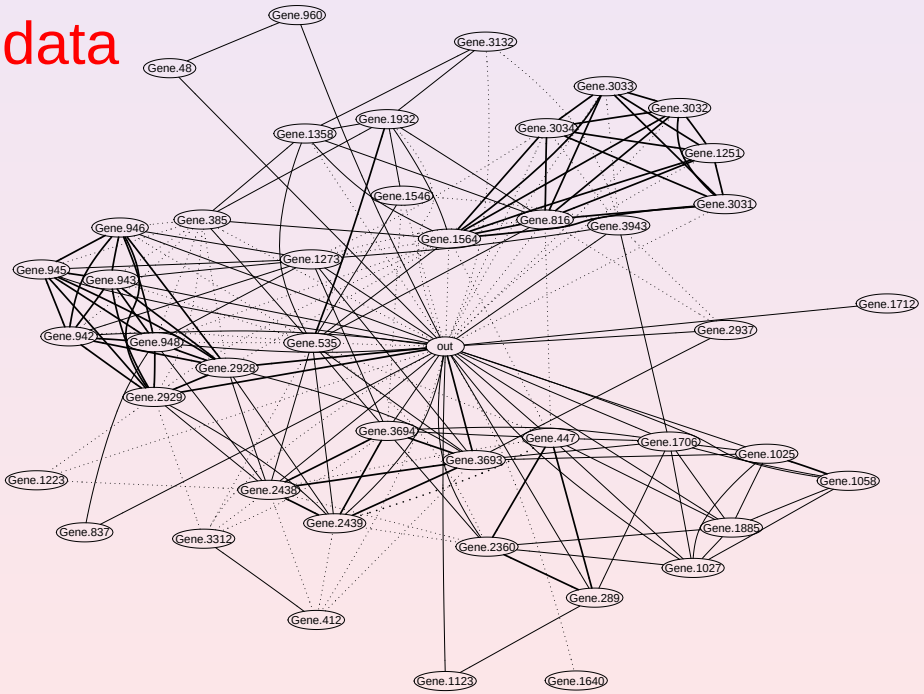
July 2008

An example: Riboflavin production in Bacillus Subtilis

in collaboration with DSM (former Roche Vitamines)

response variables $Y \in \mathbb{R}$: riboflavin production rate
covariates $X \in \mathbb{R}^p$: expressions from $p = 4088$ genes
sample size $n = 71$ from a "homogeneous" population of genetically engineered mutants of Bacillus Subtilis

$p \gg n$ and
high quality data



goal: improve riboflavin production rate of Bacillus Subtilis

more refined question:
what is the effect of knocking-down a single gene on the riboflavin production rate?

$\rightsquigarrow$ this is a question of interventional type; not association

outline:
we will use intervention calculus (e.g. Pearl) $\approx$ causal analysis

- in the high-dimensional framework
- based on observational data only $\rightsquigarrow$ we will infeer minimal bounds for interventional/causal effects

(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 realistsic for our problem
if we change one gene, some others will also change
and these are not (cannot be) kept fixed

Intervention calculus

"dynamic" notion of importance:

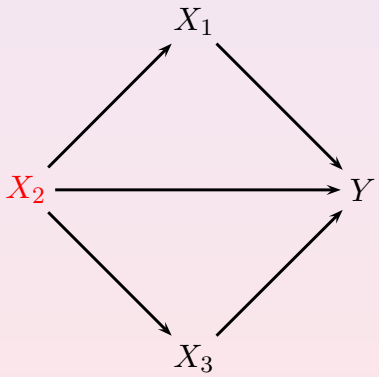
if we set a variable $X^{(j)}$ to a value x (intervention)

$\rightsquigarrow$ some other variables $X^{(k)}$ $(k \neq j)$ and maybe Y will change

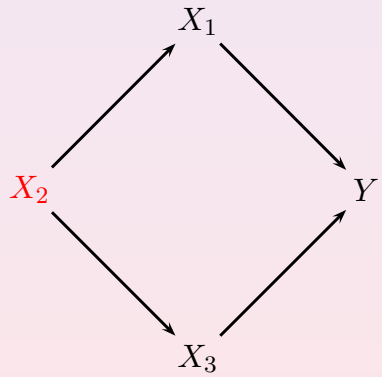
we want to quantify the total effect of
$X^{(j)}$ on Y plus "all changed" $X^{(k)}$ on Y

a graph or influence diagram will be very useful

Example 1



Example 2



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

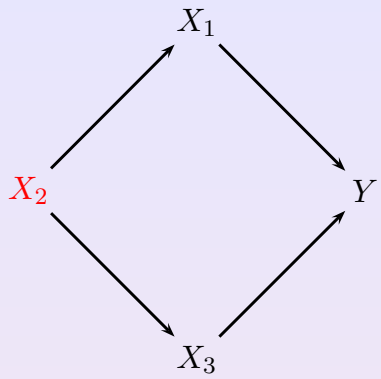
for DAG's: recursive factorization of joint distribution

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

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

Example

Example 2



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

truncated factorization for $\mathrm{do}(X^{(2)} = x)$,
i.e. intervention at $X^{(2)}$ by setting it to the value x:

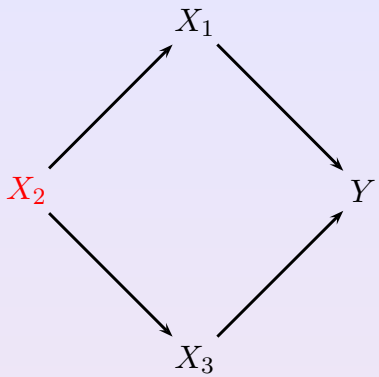
$$
\begin{aligned}
& P(Y, X^{(1)}, X^{(3)}|\mathrm{do}(X^{(2)} = x)) \\
= \quad & P(Y|X^{(1)}, X^{(3)}) P(X^{(1)}|X^{(2)} = x) P(X^{(3)}|X^{(2)} = x) \cdot 1
\end{aligned}
$$

$$
P(Y|\mathrm{do}(X^{(2)} = x)) = \int P(Y, X^{(1)}, X^{(3)}|\mathrm{do}(X^{(2)} = x)) dX^{(1)} dX^{(3)}
$$

$$
= \int P(Y|X^{(1)}, X^{(3)}) P(X^{(1)}|X^{(2)} = x) P(X^{(3)}|X^{(2)} = x) dX^{(1)} dX^{(3)}
$$

Example

Example 2



$$P(Y, X^{(1)}, X^{(2)}, X^{(3)})$$
$$= P(Y|X^{(1)}, X^{(3)})P(X^{(1)}|X^{(2)})P(X^{(3)}|X^{(2)})P(X^{(2)})$$

truncated factorization for $\mathrm{do}(X^{(2)} = x)$,
i.e. intervention at $X^{(2)}$ by setting it to the value x:

$$P(Y, X^{(1)}, X^{(3)}|\mathrm{do}(X^{(2)} = x))$$
$$= P(Y|X^{(1)}, X^{(3)})P(X^{(1)}|X^{(2)} = x)P(X^{(3)}|X^{(2)} = x) \cdot 1$$

$$P(Y|\mathrm{do}(X^{(2)} = x)) = \int P(Y, X^{(1)}, X^{(3)}|\mathrm{do}(X^{(2)} = x))dX^{(1)}dX^{(3)}$$

$$= \int P(Y|X^{(1)}, X^{(3)})P(X^{(1)}|X^{(2)} = x)P(X^{(3)}|X^{(2)} = x)dX^{(1)}dX^{(3)}$$

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

- observational data
  $\rightsquigarrow$ need to estimate conditional distributions
- an influence diagram
  $\rightsquigarrow$ need to estimate structure of a graph/influence diagram

intervention effect: for example

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

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

otherwise: we have an intervention effect “within the system of measured variables” which is better than considering just association

when having no unmeasured confounder (variable):

intervention effect (as defined) = causal effect

otherwise: we have an intervention effect “within the system of measured variables” which is better than considering just association

recap: Gaussian case

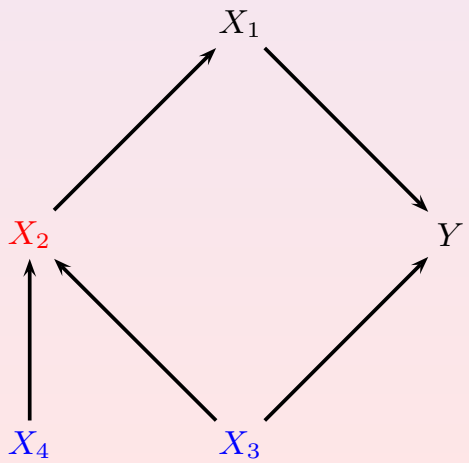
$$\frac{\partial}{\partial x}\mathbb{E}[Y|\mathrm{do}(X^{(j)} = x)] \equiv \theta_j \text{ for all } x$$

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

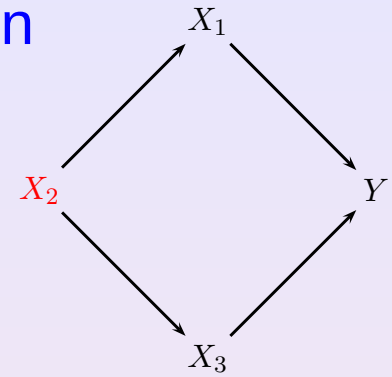
θ_j is the regression parameter in

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

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



Intervention versus association



$$
\begin{aligned}
Y &= X^{(1)} + X^{(3)} + \varepsilon, \\
X^{(2)} &= \varepsilon^{(2)}, \\
X^{(1)} &= 0.8X^{(2)} + \varepsilon^{(1)}, \\
X^{(3)} &= 0.8X^{(2)} + \varepsilon^{(3)}
\end{aligned}
$$

intervention effect: $\theta_2 = 1.6$

$\theta_1 = \theta_3 = 1 \rightsquigarrow X^{(2)}$ has largest interventional importance

regression effect: $\beta_2 = 0$

$\beta_1 = \beta_3 = 1 \rightsquigarrow X^{(2)}$ has smallest (zero) interventional imp.

Inferring interventional effects

main problem: inferring DAG from data

$\rightsquigarrow$ impossible: can only infer equivalence class of DAG's

$\rightsquigarrow$ for each variable $X^{(j)}$, can only estimate

set of interventional effects

the population parameters: sets of interventional 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: $\mathcal{G}_1, \ldots, \mathcal{G}_m$
- for every DAG-member $\mathcal{G}_r$, and every variable $X^{(j)}$:
  single interventional effect $\theta_{r,j}$
  summarize them by

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

unique values may occur:

it may happen that for some j: $\theta_{1,j} = \theta_{2,j} = \ldots, \theta_{m,j}$
i.e. the jth interventional effect is unique

but in general, the population parameter is a (multi-) set Θ
it holds that:

for every j: true intervention effect from true DAG $\theta_{\text{true},j} \in \Theta$

typically: $\{\theta_{\text{true},j};\ j = 1, \ldots, p\} \overset{\text{strict}}{\subset} \Theta$

Pearl:

"... a $\underbrace{\text{causal concept}}_{\text{single causal effect}}$ cannot be defined from the distribution alone"

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

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

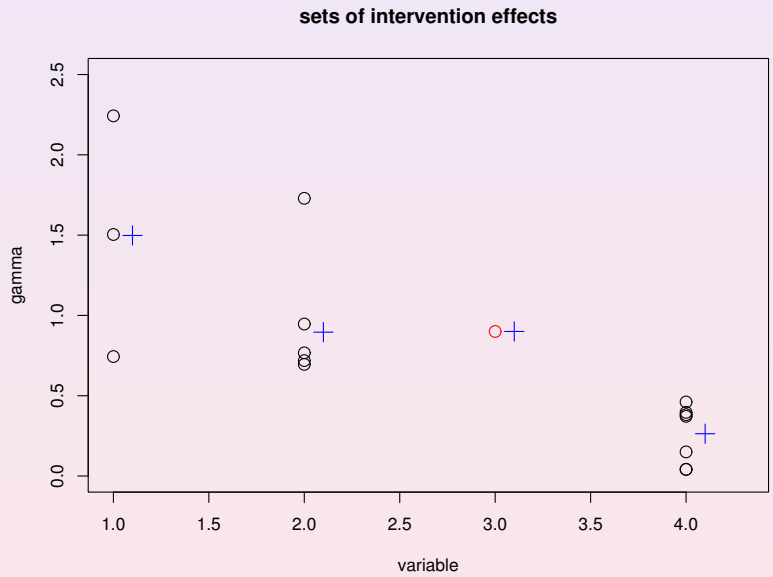
Multiplicities

the assumed values of $\theta_{1,j}, \ldots, \theta_{m,j}$ are

$$\gamma_{1,j}, \ldots, \gamma_{k_j,j} \quad (k_j \leq m)$$

$\gamma_{r,j}$ occurs $n(\gamma_{r,j})$ times

- values $|\gamma_{r,j}|$ $(j = 1, 2, 3, 4, ...)$
- multiplicities: $m = 7$
  unique effect in red
- (weighted) mean:
  $\sum_r n(\gamma_{r,j})|\gamma_{r,j}|/7 =$
  $\sum_{r=1}^{7} |\theta_{r,j}|/7$



using the additional concept of multiplicities
we can define an average interventional effect

Computationally tractable algorithm: population version

conceptually: so far, we described $\Theta_{m \times p}$ by finding/searching for all members (DAG's) within an equivalence class of DAG's

searching all DAG's is computationally infeasible if p is large (we actually can do this up to $p \approx 15$)

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

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

- directed edge in CPDAG: every member (DAG) in equivalence class has this directed edge
- undirected edge in CPDAG: some members (DAG's) have this edge with opposite directions
- no edge in CPDAG: no edge for every member (DAG)

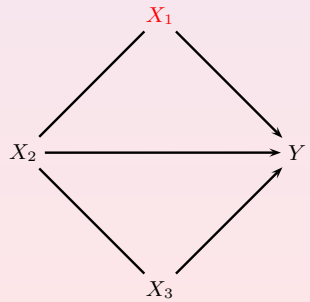
Local algorithm to find all intervention effects $\theta_{r,j}$

(Maathuis, Kalisch & PB, 2008)

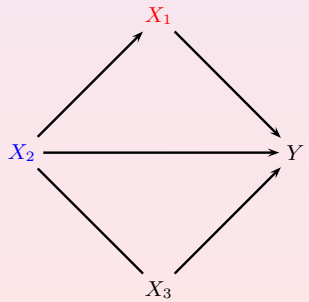
input: CPDAG (true underlying equivalence class of DAG's)

- parents of $X^{(j)}$: $\mathrm{pa}(j)$
- undirected neighbors of $X^{(j)}$: undir-neigh(j)
- consider all subsets $\mathcal{S}$ of undir-neigh(j)
  make $\mathcal{S}$ a set of parental nodes of $j \rightsquigarrow$ new graph $G_{\mathcal{S}\to j}$
  check whether new graph $G_{\mathcal{S}\to j}$ has no new v-structure with collider j: if yes, denote the set by $\mathcal{S}_+$
- for all such $\mathcal{S}_+$ and all j: regression
  $Y = \theta_{\mathcal{S}_+,j} X^{(j)} + \sum_{k\in \mathrm{pa}(j)} \theta_{\mathcal{S}_+,k} X^{(k)} + \sum_{\ell\in\mathcal{S}_+} \theta_{\mathcal{S}_+,\ell} X^{(\ell)} + \text{error}$
- denote by $\Theta_{\text{local}} = \{\theta_{\mathcal{S}_+,j};\ \text{all subsets } \mathcal{S}_+\}$

CPDAG

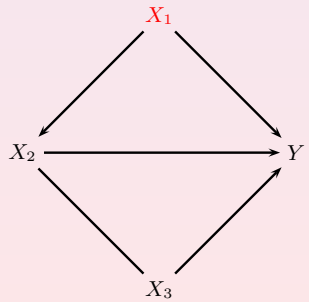


$G^{(1)}_{\mathcal{S}_+\to 1}$



Y *vs.* $X^{(1)} + X^{(2)}$

$G^{(2)}_{\mathcal{S}_+\to 1}$



Y *vs.* $X^{(1)}$

Theorem (Maathuis, Kalisch & PB, 2008)

$\Theta_{\text{local}} = \Theta$,where equality is in terms of sets
but the mutiplicities are not the same

huge computational gain if p is large, e.g. $p \approx 5'000$

Estimation

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

use the PC-algorithm (Spirtes & Glymour, 1991)

underlying 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)

implication of faithfulness:
for the skeleton of the true DAG (directions of edges are removed)

$$\begin{aligned} & \text{edge between } i \text{ and } j \\ \Leftrightarrow \quad & X^{(i)} \text{ dependent of } X^{(j)} \text{ given } X^{(\mathcal{S})}, \text{ for all } \mathcal{S} \subseteq \{1, \ldots, p\} \end{aligned}$$

considering all subsets is only "conceptual" $\rightsquigarrow$ see below (and impossible to compute)

note: for conditional independence graph and regression

$$\begin{aligned} & \text{edge between } i \text{ and } j \\ \Leftrightarrow \quad & X^{(i)} \text{ dependent of } X^{(j)} \text{ given } \{X^{(k)};\ k \neq i, j\} \end{aligned}$$

in the Gaussian case: need to estimate whether

$$\text{Parcor}(X^{(i)}, X^{(j)} | X^{(\mathcal{S})}) = 0 \text{ or } \neq 0$$

conceptually for all subsets $\mathcal{S}$; but in fact, only for "some"

thanks to faithfulness, we can gradually move-up from marginal to higher-order partial correlations

$\rightsquigarrow$ key feature to deal with $p \gg n$

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 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 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

one tuning parameter (cut-off parameter) α for truncation of estimated Z-transofrmed 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

modification of the above algorithm (for estimation of some separating sets) yields an estimate of the CPDAG (equivalence class of DAG's)

Theorem (Kalisch & PB, 2007; Maathuis, Kalisch & PB, 2008)

- $X^{(1)}, \ldots, X^{(p)} \sim \mathcal{N}_p(\mu, \Sigma)$ faithful to a DAG
- $p = p_n = O(n^\alpha)$ $(0 \leq \alpha < \infty)$ (high-dimensional)
- $\max_j |\mathrm{ne}(j)| = o(n)$ (sparsity)
- non-zero (partial) correlations $\gg n^{-1/2}$
  maximal (partial) correlation $\leq C < 1$

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

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

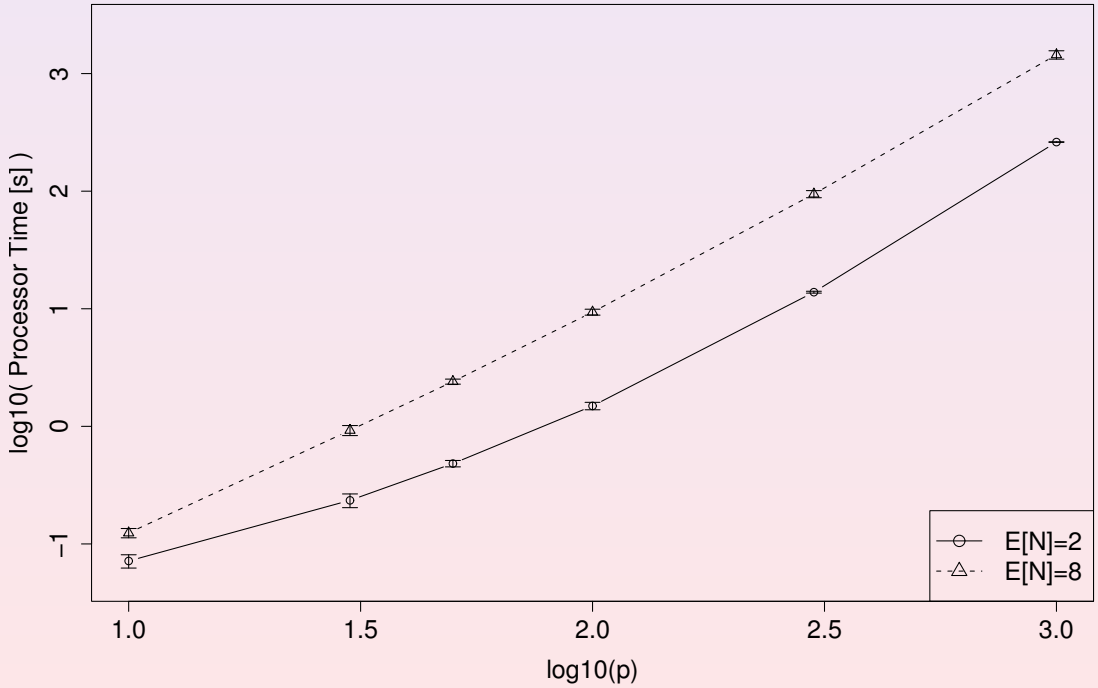
$$\mathbb{P}[\widehat{\text{CPDAG}}(\alpha) = \text{ true CPDAG}] = 1 - O(\exp(Cn^{1-\delta}))$$

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

computational complexity:

crudely bounded to be polynomial in p
sparser underlying structure $\rightsquigarrow$ faster algorithm

we can easily do the computations for
sparse cases with $p \approx 10^4$ $\approx$ 2-5 hrs CPU time



How well can we do?

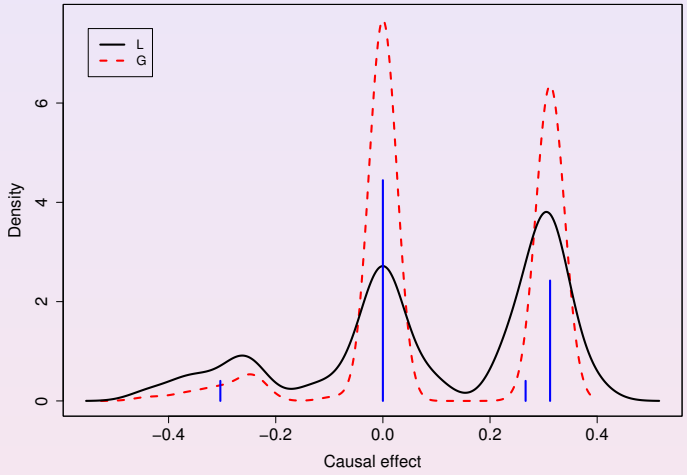
two methods:

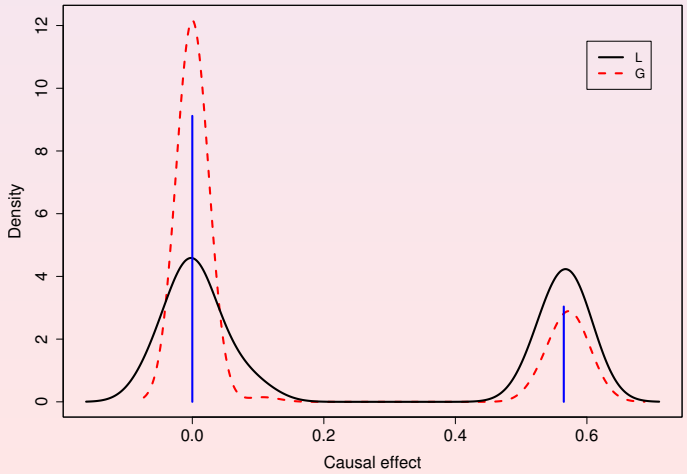
- local algorithm: as described
- global algorithm:
  searching for all DAG's within an equivalence class
  and computing intervention effects from all these DAG's

for our simulation models:
could compute with global method up to $p = 14$
whereas local algorithm can handle large $p \approx 10^4$

$n = 1000, p = 8, \mathbb{E}[\text{neighborhood} - \text{size}] = 3$

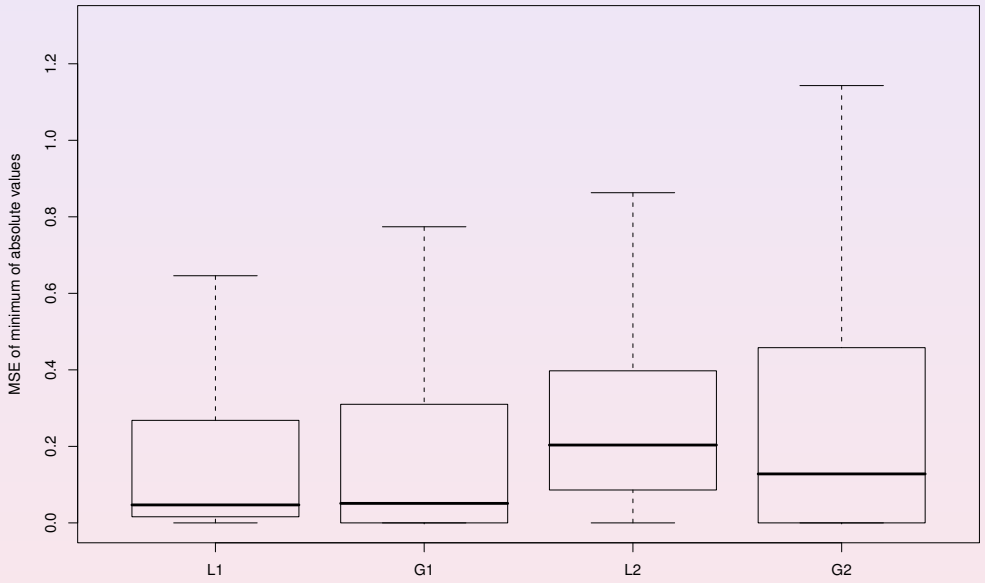
densities of intervention effects $\hat{\theta}$, including multiplicities
local (black) and global (red) method; true values (blue)





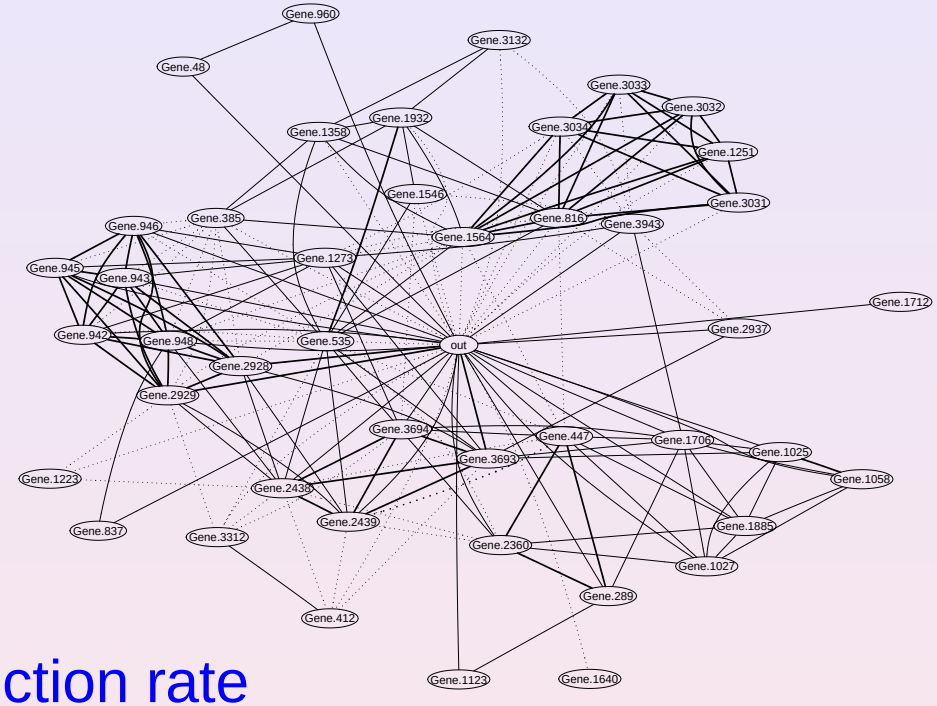
$n = 2000$ (left); $n = 20$ (right); $p = 10$;
$\mathbb{E}[\text{neighborhood} - \text{size}] = 4$

MSE for lower bound: $\mathbb{E}[(\min_r |\hat{\theta}_{r,j}| - \min_r |\theta_{r,j}|)^2]$



$\rightsquigarrow$ for small n: global algorithm slightly better
but computationally infeasible for $p > 15$

Riboflavin production with Bacillus Subtilis



Y: riboflavin production rate

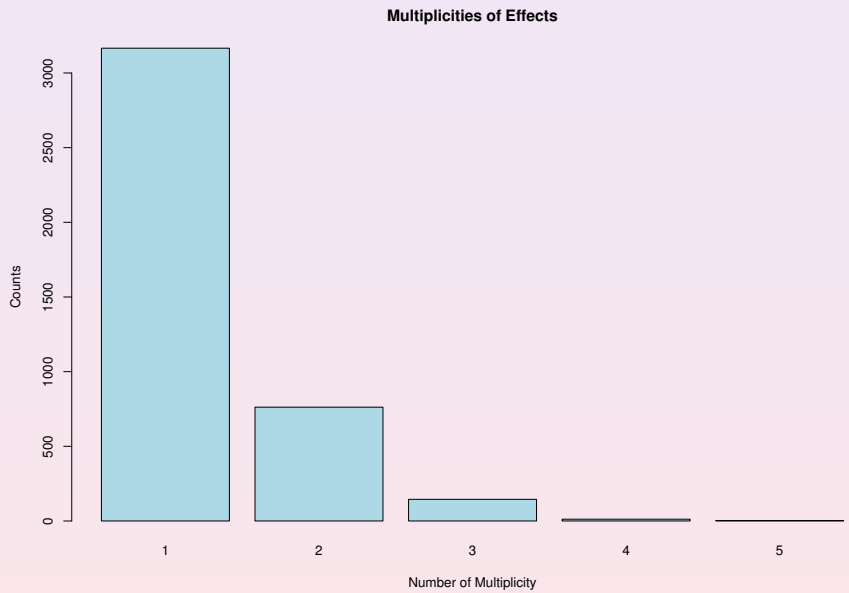
covariates $X \in \mathbb{R}^p$: expressions from $p = 4088$ genes

sample size $n = 71$ from a "homogeneous" population of genetically engineered mutants of Bacillus Subtilis

goal: estimate intervention effects of the $p = 4088$ genes

we use regularization parameter $\alpha = 0.01$
(a more principled choice via sub-sampling/bootstraping is possible)

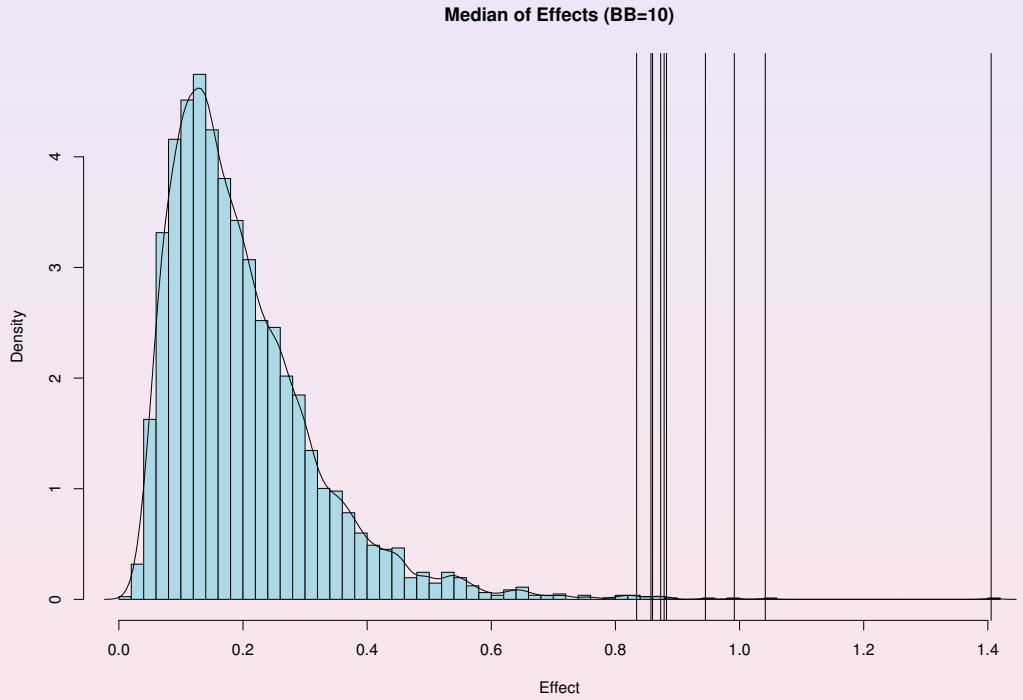
multiplicities of $\hat{\theta}$:



degree of uniqueness is high $\rightsquigarrow$ $\min_r |\hat{\theta}_{r,j}|$ is "tight" lower bound

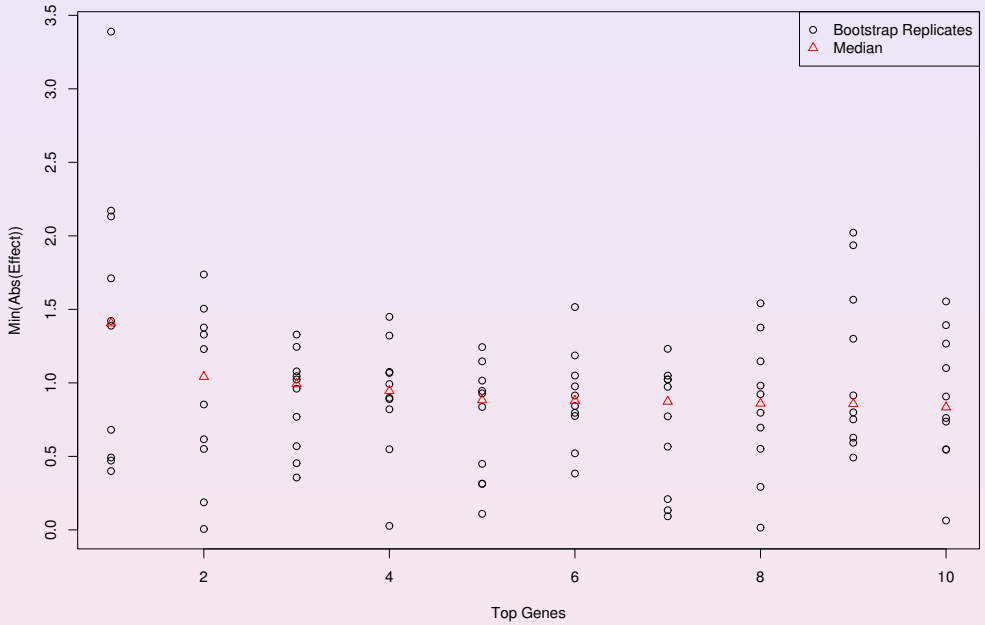
bootstrap analysis (10 replicates)

median of bootstrapped $\min_r |\hat{\theta}^*_{r,j}|$, for all $j = 1, 2, \ldots, p$



top 10 genes (variables) indicated by vertical line

bootstrapped $\min_r |\hat{\theta}^*_{r,j}|$ (top 10),
ranked by median of bootstrapped values

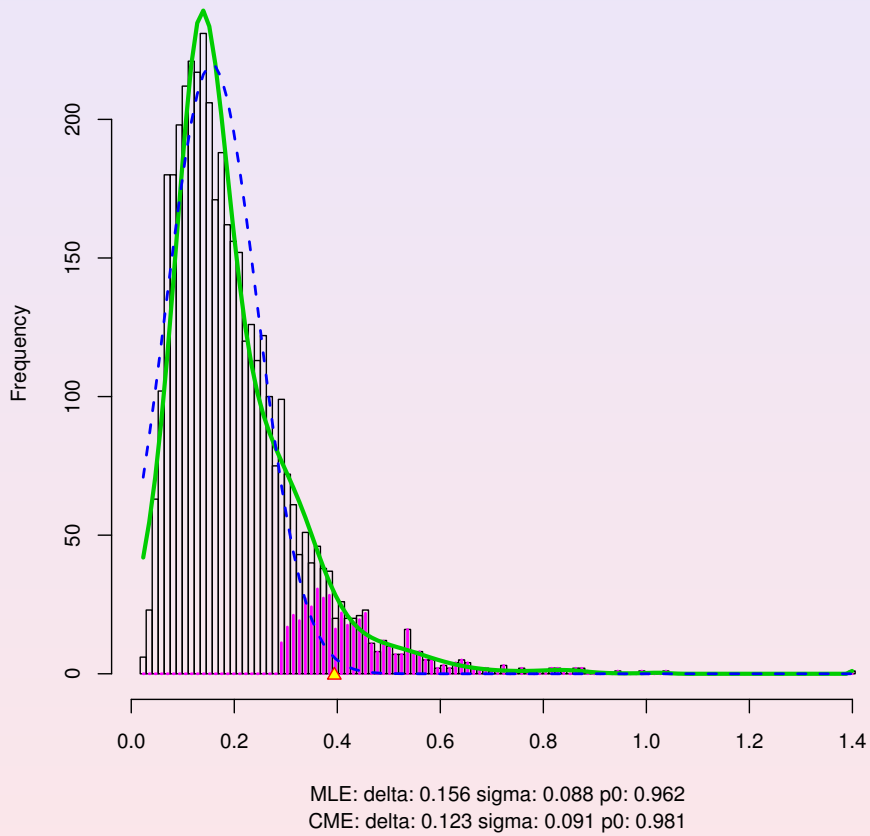


median of bootstrapped minimal intervention effects ≥ 0.85
$\rightsquigarrow$ log-productivity of riboflavin changes by ≥ 0.85 under 1-fold change of gene expression

one interesting gene aomng the “top10” which

- is biologically “plausible”
- has not been modified so far – but DSM plans to do so

if you do not trust asymptotics...
we have a scoring wich is built upon intervention calculus
$\rightsquigarrow$ use local FDR (Efron, 2001-2006) to quantify “high”



cut-off at 0.4 yields local FDR ≤ 0.2
(for bootstrapped median of $\min_r |\hat{\theta}_{r,j}|$)

Conclusions

- intervention analysis using observational data only:
  in absence of an influence diagram (graph)
  $\rightsquigarrow$ can infer bounds on intervention/causal effects
  the bounds are tight (for some variables) if multiplicity of Θ is 1 (for some variables)
- even in the sparse high-dimensional context:
  intervention analysis is computationally feasible and statistically "reasonable" and consistent
- variability and uncertainty:
  in absence of anything better so far, we use the bootstrap...
- unmeasured confounders:
  conceptually, we can make use of
  ancestral graphs (Drton & Richardson) and
  e.g. the FCI algorithm (Spirtes, Glymour & Scheines, 2000)