High-dimensional statistics: from assocation to causal inference

Peter Bühlmann
ETH Zürich

December 2010

Sara van de Geer
ETH Zurich

Nicolai Meinshausen
Oxford University

Marloes Maathuis
ETH Zurich

Markus Kalisch
ETH Zurich

High-dimensional data

1. Classification of tumor samples based on gene expression microarray data, e.g. $p = 7130, n = 49$ ($p \gg n$)

Table 1. Test set error rates based on leave one out cross validation for leukemia, colon, estrogen, nodal, lymphoma and NCI data with gene subsets from feature selection ranging between 10 to all genes for several classifiers. LogitBoost error rates are reported with optimal stopping (minimum cross-validated error across iterations), after a fixed number of 100 iterations as well as with the estimated stopping parameter. The cross validation with estimated stopping parameters for the lymphoma and NCI data with all genes was not feasible

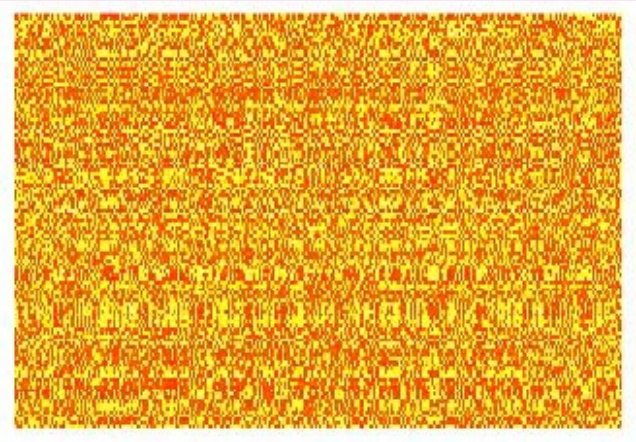
Leukemia	10	25	50	75	100	200	3571
LogitBoost, optimal	4.17%	2.78%	4.17%	2.78%	2.78%	2.78%	2.78%
LogitBoost, estimated	6.94%	5.56%	5.56%	4.17%	4.17%	5.56%	5.56%
LogitBoost, 100 iterations	5.56%	2.78%	4.17%	2.78%	2.78%	2.78%	2.78%
AdaBoost, 100 iterations	4.17%	4.17%	4.17%	4.17%	4.17%	2.78%	4.17%
1-nearest-neighbor	4.17%	1.39%	4.17%	5.56%	4.17%	2.78%	1.39%
Classification tree	22.22%	22.22%	22.22%	22.22%	22.22%	22.22%	23.61%
Colon	10	25	50	75	100	200	2000
LogitBoost, optimal	14.52%	16.13%	16.13%	16.13%	16.13%	14.52%	12.90%
LogitBoost, estimated	22.58%	19.35%	22.58%	20.97%	22.58%	19.35%	19.35%
LogitBoost, 100 iterations	14.52%	22.58%	22.58%	19.35%	17.74%	16.13%	16.13%
AdaBoost, 100 iterations	16.13%	24.19%	24.19%	17.74%	20.97%	17.74%	17.74%
1-nearest-neighbor	17.74%	14.52%	14.52%	20.97%	19.35%	17.74%	25.81%
Classification tree	19.35%	22.58%	29.03%	32.26%	27.42%	14.52%	16.13%
Estrogen	10	25	50	75	100	200	7129
LogitBoost, optimal	4.08%	4.08%	2.04%	2.04%	2.04%	4.08%	2.04%
LogitBoost, estimated	6.12%	6.12%	6.12%	6.12%	6.12%	6.12%	6.12%
LogitBoost, 100 iterations	8.16%	6.12%	6.12%	4.08%	4.08%	8.16%	6.12%
AdaBoost, 100 iterations	8.16%	8.16%	2.04%	2.04%	6.12%	4.08%	4.08%
1-nearest-neighbor	4.08%	8.16%	18.37%	12.24%	14.29%	14.29%	16.33%
Classification tree	4.08%	4.08%	4.08%	4.08%	4.08%	4.08%	4.08%
Nodal	10	25	50	75	100	200	7129
LogitBoost, optimal	16.33%	18.37%	22.45%	22.45%	22.45%	18.37%	20.41%
LogitBoost, estimated	22.45%	30.61%	30.61%	34.69%	28.57%	26.53%	24.49%
LogitBoost, 100 iterations	18.37%	20.41%	26.53%	42.86%	42.86%	18.37%	22.45%
AdaBoost, 100 iterations	18.37%	16.33%	28.57%	40.82%	36.73%	22.45%	28.57%
1-nearest-neighbor	18.37%	30.61%	30.61%	42.86%	36.73%	36.73%	48.98%
Classification tree	22.45%	20.41%	20.41%	20.41%	20.41%	20.41%	20.41%
Lymphoma	10	25	50	75	100	200	4026
LogitBoost, optimal	1.61%	3.23%	1.61%	1.61%	1.61%	3.23%	8.06%
LogitBoost, estimated	3.23%	3.23%	3.23%	1.61%	3.23%	3.23%	-%
LogitBoost, 100 iterations	1.61%	3.23%	1.61%	1.61%	1.61%	3.23%	8.06%
AdaBoost, 100 iterations	4.84%	3.23%	1.61%	1.61%	1.61%	1.61%	3.23%
Nearest neighbor	1.61%	0.00%	0.00%	0.00%	0.00%	1.61%	1.61%
Classification tree	22.58%	22.58%	22.58%	22.58%	22.58%	22.58%	25.81%
NCI	10	25	50	75	100	200	5244
LogitBoost, optimal	32.79%	31.15%	27.87%	22.95%	26.23%	24.59%	31.15%
LogitBoost, estimated	36.07%	44.26%	36.07%	39.34%	44.26%	47.54%	-%
LogitBoost, 100 iterations	37.70%	44.26%	34.43%	29.51%	26.23%	24.59%	36.07%
AdaBoost, 100 iterations	50.82%	37.70%	34.43%	29.51%	32.79%	29.51%	36.07%
Nearest neighbor	36.07%	29.51%	27.87%	24.59%	22.95%	22.95%	27.87%
Classification tree	70.49%	68.85%	65.57%	65.57%	60.66%	62.30%	62.30%

2. Riboflavin production with Bacillus Subtilis
(in collaboration with DSM (Switzerland))

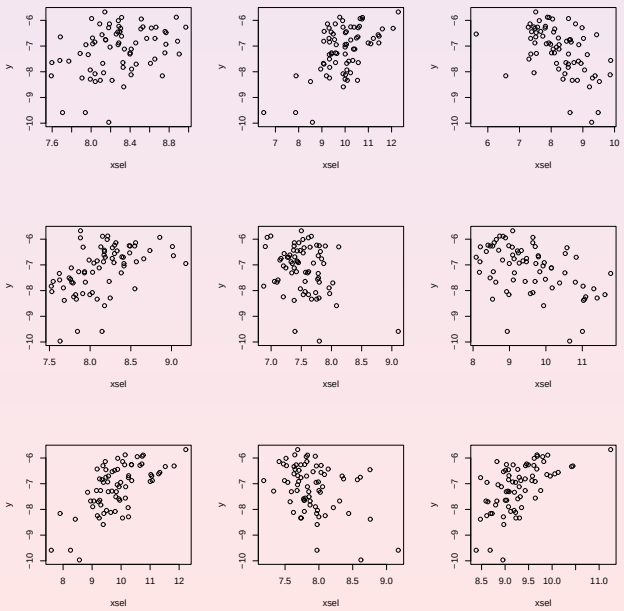
goal: improve riboflavin production rate of Bacillus Subtilis using clever genetic engineering

response variables $Y \in \mathbb{R}$: riboflavin (log-) production rate
covariates $X \in \mathbb{R}^p$: expressions from $p = 4088$ genes
sample size $n = 115$, $p \gg n$

gene expression data



Y versus 9 “reasonable” genes



High-dimensional linear models

$$Y_i = (\mu +) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$

goals:

- prediction, e.g. w.r.t. squared prediction error
- estimation of parameter β
- variable selection
  i.e. estimating the effective variables
  (having corresponding coefficient $\neq 0$)

Exemplifying the outline

binary lymph node classification using gene expressions

a high noise problem: $n = 49$ samples, $p = 7130$ gene expr.

despite that it is classification:
$p(x) = \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$
$\rightsquigarrow$ $\hat{p}(x)$ via linear model; can then do classification

cross-validated misclassification error (2/3 training; 1/3 test)

Lasso	L_2Boosting	FPLR	Pelora	1-NN	DLDA	SVM
21.1%	17.7%	35.25%	27.8%	43.25%	36.12%	36.88%

with variable selection

best 200 genes (Wilcoxon test)
no additional variable selection

from a practical perspective:
if you trust in cross-validation: can "validate" how good we are
i.e. prediction may be a black box, but we can "evaluate" it

Exemplifying the outline

binary lymph node classification using gene expressions

a high noise problem: $n = 49$ samples, $p = 7130$ gene expr.

despite that it is classification:
$p(x) = \mathbb{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$
$\rightsquigarrow$ $\hat{p}(x)$ via linear model; can then do classification

cross-validated misclassification error (2/3 training; 1/3 test)

Lasso	L_2Boosting	FPLR	Pelora	1-NN	DLDA	SVM
21.1%	17.7%	35.25%	27.8%	43.25%	36.12%	36.88%

with variable selection

best 200 genes (Wilcoxon test)
no additional variable selection

from a practical perspective:
if you trust in cross-validation: can "validate" how good we are
i.e. prediction may be a black box, but we can "evaluate" it

"however"

- cross-validation has large variability...
  still want to know whether a method is good or optimal for prediction
- if concerned about $\|\hat{\beta} - \beta^0\|$ (estimation error)
  $\rightsquigarrow$ no easy (cross-) validation available
- if concerned about the active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$ and variable selection
  $\rightsquigarrow$ no easy (cross-) validation available

and this is the outline:

- prediction, estimation, variable selection in regression/classification
- and then graphical modeling and intervention/causal analysis

The Lasso (Tibshirani, 1996)

Lasso for linear models

$$\hat{\beta}(\lambda) = \text{argmin}_\beta(n^{-1}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \underbrace{\lambda}_{\geq 0} \underbrace{\|\beta\|_1}_{\sum_{j=1}^p |\beta_j|})$$

$\rightsquigarrow$ convex optimization problem

- Lasso does variable selection
  some of the $\hat{\beta}_j(\lambda) = 0$
  (because of "ℓ_1-geometry")
- $\hat{\beta}(\lambda)$ is a shrunken LS-estimate

more about "ℓ_1-geometry"

equivalence to primal problem

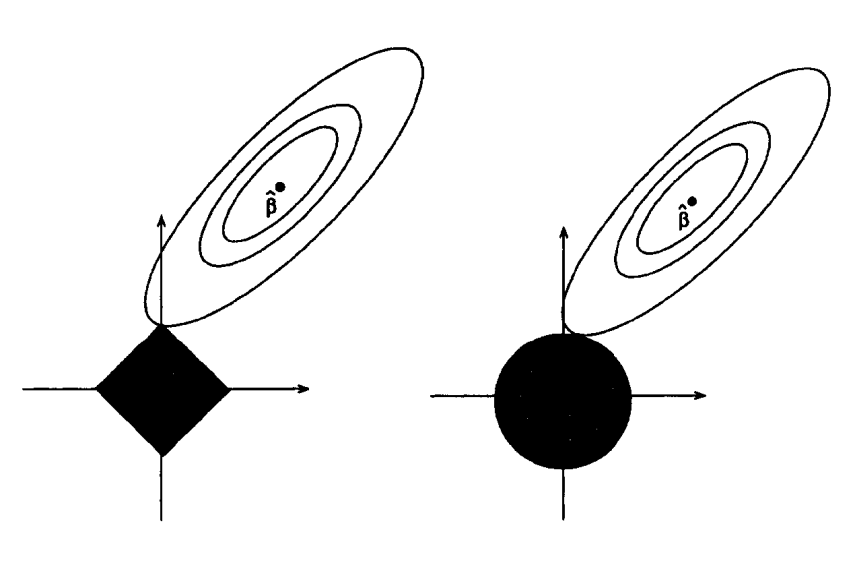
$$\hat{\beta}_{\text{primal}}(R) = \text{argmin}_{\beta;\|\beta\|_1 \le R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n,$$

with a one-to-one correspondence between λ and R which depends on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$
[such an equivalence holds since

- $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ is convex in β
- convex constraint $\|\beta\|_1 \le R$

see e.g. Bertsekas (1995)]

p=2



left: ℓ_1-"world"
residual sum of squares reaches a minimal value (for certain constellations of the data) if its contour lines hit the ℓ_1-ball in its corner
$\rightsquigarrow \ \hat{\beta}_1 = 0$

ℓ_2-"world" is different

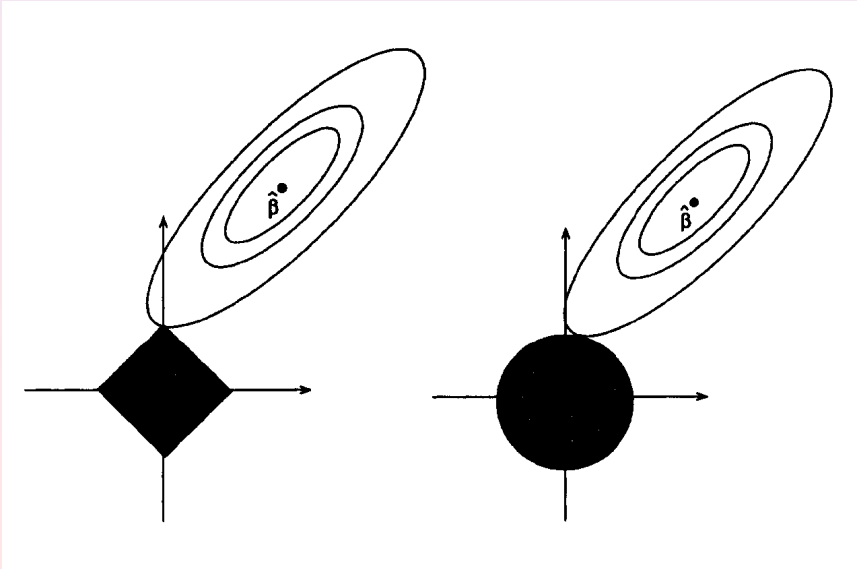
Ridge regression,

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \text{argmin}_\beta \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda\|\beta\|_2^2 \right),$$

equivalent primal equivalent solution

$$\hat{\beta}_{\text{Ridge;primal}}(R) = \text{argmin}_{\beta;\|\beta\|_2 \le R} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n,$$

with a one-to-one correspondence between λ and R



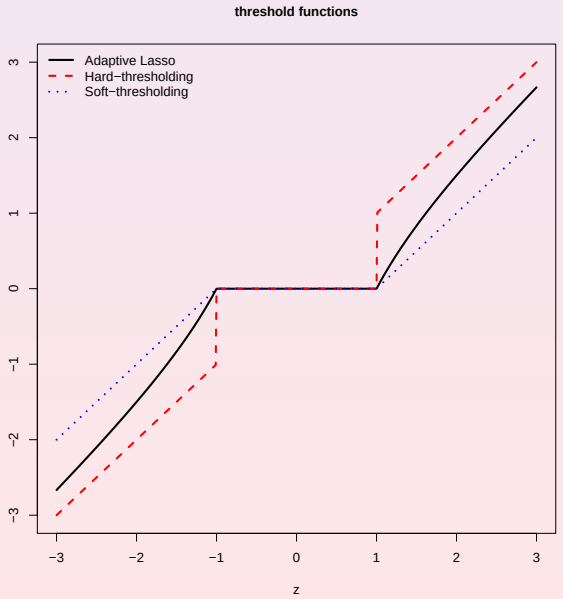
Orthonormal design

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad n^{-1}\mathbf{X}^T\mathbf{X} = I$$

Lasso = soft-thresholding estimator

$$\hat{\beta}_j(\lambda) = \mathrm{sign}(Z_j)(|Z_j| - \lambda/2)_+, \quad \underbrace{Z_j}_{=\mathrm{OLS}} = (n^{-1}\mathbf{X}^T\mathbf{Y})_j,$$

$$\hat{\beta}_j(\lambda) = g_{\mathrm{soft}}(Z_j),$$



Using the Lasso...

in practice: choose λ via cross-validation (e.g. 10-fold)

use cross-validation again to validate the procedure
(need double cross-validation)

binary lymph node classification using gene expressions:
a high noise problem
$n = 49$ samples, $p = 7130$ gene expressions

cross-validated misclassification error (2/3 training; 1/3 test)

Lasso	L_2Boosting	FPLR	Pelora	1-NN	DLDA	SVM
21.1%	17.7%	35.25%	27.8%	43.25%	36.12%	36.88%

with variable selection

best 200 genes (Wilcoxon test)
no additional variable selection

and Lasso selects on CV-average 13.12 out of $p = 7130$ genes

Theory for the Lasso: Prediction and estimation

fixed design linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$

Basic inequality

$$n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq 2n^{-1}\varepsilon^T\mathbf{X}(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1$$

Proof:

$$n^{-1}\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq n^{-1}\|\mathbf{Y} - \mathbf{X}\beta^0\|_2^2 + \lambda\|\beta^0\|_1$$

$n^{-1}\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2 = n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 + n^{-1}\|\varepsilon\|_2^2 - 2n^{-1}\varepsilon^T\mathbf{X}(\hat{\beta} - \beta^0)$
$n^{-1}\|\mathbf{Y} - \mathbf{X}\beta^0\|_2^2 = n^{-1}\|\varepsilon\|_2^2$
$\rightsquigarrow$ statement above □

need a bound for $2n^{-1}\varepsilon^T\mathbf{X}(\hat{\beta} - \beta^0)$

$$2n^{-1}\varepsilon^T \mathbf{X}(\hat{\beta} - \beta^0) \quad \leq \quad 2 \max_{j=1,\ldots,p} |n^{-1} \sum_{i=1}^{n} \varepsilon_i X_i^{(j)}| \|\hat{\beta} - \beta^0\|_1$$

consider

$$\mathcal{T} = \mathcal{T}(\lambda_0) = \{2 \max_j |n^{-1} \sum_{i=1}^{n} \varepsilon_i X_i^{(j)}| \leq \lambda_0\}$$

the probabilistic part of the problem

$$\text{on } \mathcal{T}: \quad 2n^{-1}\varepsilon^T \mathbf{X}(\hat{\beta} - \beta^0) \quad \leq \quad \lambda_0 \|\hat{\beta} - \beta^0\|_1 \leq \lambda_0 \|\hat{\beta}\|_1 + \lambda_0 \|\beta^0\|_1$$

and hence using the Basic inequality

$$\text{on } \mathcal{T}: \quad n^{-1} \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 + (\lambda - \lambda_0)\|\hat{\beta}\|_1 \leq (\lambda_0 + \lambda)\|\beta^0\|_1$$

for $\lambda \geq 2\lambda_0$:

$$\text{on } \mathcal{T} = \mathcal{T}(\lambda_0): \quad 2n^{-1} \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 + \lambda \|\hat{\beta}\|_1 \leq 3\lambda \|\beta^0\|_1$$

choice of λ and probability of the set $\mathcal{T}$

λ as small as possible such that $\lambda \geq 2\lambda_0$ (see above)
λ_0 such that $\tau = \tau(\lambda_0)$ has large probability

$$V_j = n^{-1/2}\sigma^{-1}\sum_{i=1}^{n}\varepsilon_i X_i^{(j)} \quad \rightsquigarrow \mathcal{T}(\lambda_0) = \{2\max_{j=1,\ldots,p}|V_j| \leq \lambda_0 n^{1/2}\sigma^{-1}\}$$

Example:
Gaussian errors $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$
and scaled covariates $n^{-1}\|\mathbf{X}^{(j)}\|_2^2 \equiv 1$
then: $V_j \sim \mathcal{N}(0,1) \quad \rightsquigarrow$

$$\lambda_0 = 2\sigma\sqrt{\frac{u^2 + 2\log(p)}{n}} \Rightarrow \mathbb{P}[\mathcal{T}(\lambda_0)] \geq 1 - 2\exp(u^2/2)$$

can generalize to non-Gaussian errors (sub-Gaussian distr., higher moments), to dependent errors, ...

for prediction with high-dimensional ℓ_1-penalization:

$$\lambda \asymp \lambda_0 \asymp \sqrt{\log(p)/n}$$

unless the variables are very correlated
$\rightsquigarrow$ would relax the $\log(p)$ factor a bit

recall for $\lambda \geq 2\lambda_0$:

$$\text{on } \mathcal{T}(\lambda_0): \quad 2n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq 3\lambda\|\beta^0\|_1$$

and hence: for λ (and λ_0) $\asymp \sqrt{\log(p)/n}$,

$$n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 = \|\beta^0\|_1 O_P(\sqrt{\log(p)/n})$$

$\rightsquigarrow$

- consistency for prediction if $\|\beta^0\|_1 = o(\sqrt{n/\log(p)})$
  essentially recovering Greenshtein & Ritov (2004)
  with a simple structure how to generalize to other settings
- convergence rate $O_P(\sqrt{\log(p)/n})$ is "far from optimal"
- no assumptions on the (fixed) design matrix

aim: $n^{-1}\|\mathbf{X}(\hat{\beta}-\beta^0)\|_2^2 = s_0 O_P(\log(p)/n)$, $s_0 = |S_0| = |\{j;\ \beta_j^0 \neq 0\}|$

unfortunately, for the Lasso and other computationally feasible methods: need conditions on $\mathbf{X}$

idea: recall the basic inequality

$$n^{-1}\|\mathbf{X}(\hat{\beta}-\beta^0)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq 2n^{-1}\varepsilon^T\mathbf{X}(\hat{\beta}-\beta^0) + \lambda\|\beta^0\|_1$$

simple re-writing (triangle inequality) on $\mathcal{T}(\lambda_0)$, with $\lambda \geq 2\lambda_0$,

$$2\|(\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0)\|_2^2 + \lambda\|\hat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1$$

where $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$

relate $\|\hat{\beta}_{S_0} - \beta^0_{S_0}\|_1$ to (with $\leq$ relation) $(\hat{\beta}-\beta^0)\hat{\Sigma}(\hat{\beta}-\beta^0)$
(and bring it to the left hand side)
this is a kind of restricted ℓ_1-eigenvalue problem

reminder:
$\|\beta\|_2^2 \leq \frac{\beta^T \hat{\Sigma} \beta}{\Lambda_{min}^2}$ where Λ_{min}^2 is the smallest eigenvalue of $\hat{\Sigma}$

here: Compatibility condition (van de Geer, 2007)
smallest restricted ℓ_1-eigenvalue:
active set S_0 with $s_0 = |S_0|$
compatibility constant $\phi_0^2 > 0$ such that for all β satisfying $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$, it holds that

$$\|\beta_{S_0}\|_1^2 \leq \frac{(\beta^T \hat{\Sigma} \beta) s_0}{\phi_0^2}$$

(appearance of s_0 due to $\|\beta_{S_0}\|_1^2 \leq s_0 \|\beta_{S_0}\|_2^2$)

oracle inequality

for $\lambda \geq 2\lambda_0$:

on $\mathcal{T}(\lambda_0)$: $n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2$

asymptotics: $\lambda \asymp \sqrt{\log(p)/n}$,

$$n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 \leq \frac{s_0}{\phi_0^2} O_P(\log(p)/n),$$

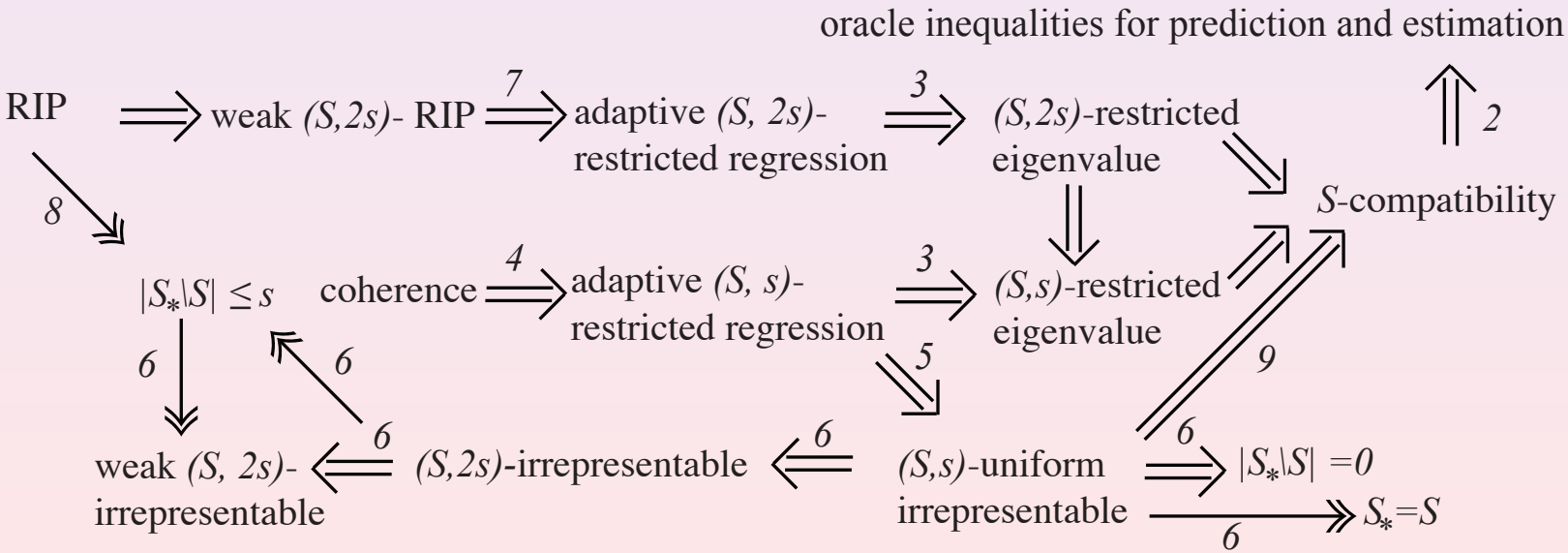
$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{s_0}{\phi_0^2} O_P(\sqrt{\log(p)/n})$$

just make the appropriate assumptions to prove what you like...

real question:
how restrictive is compatibility condition (smallest restricted ℓ_1-eigenvalue)?

it is (slightly) weaker than the restricted eigenvalue assumption (Bickel, Ritov & Tsybakov, 2009)

more generally: (van de Geer & PB, 2009)

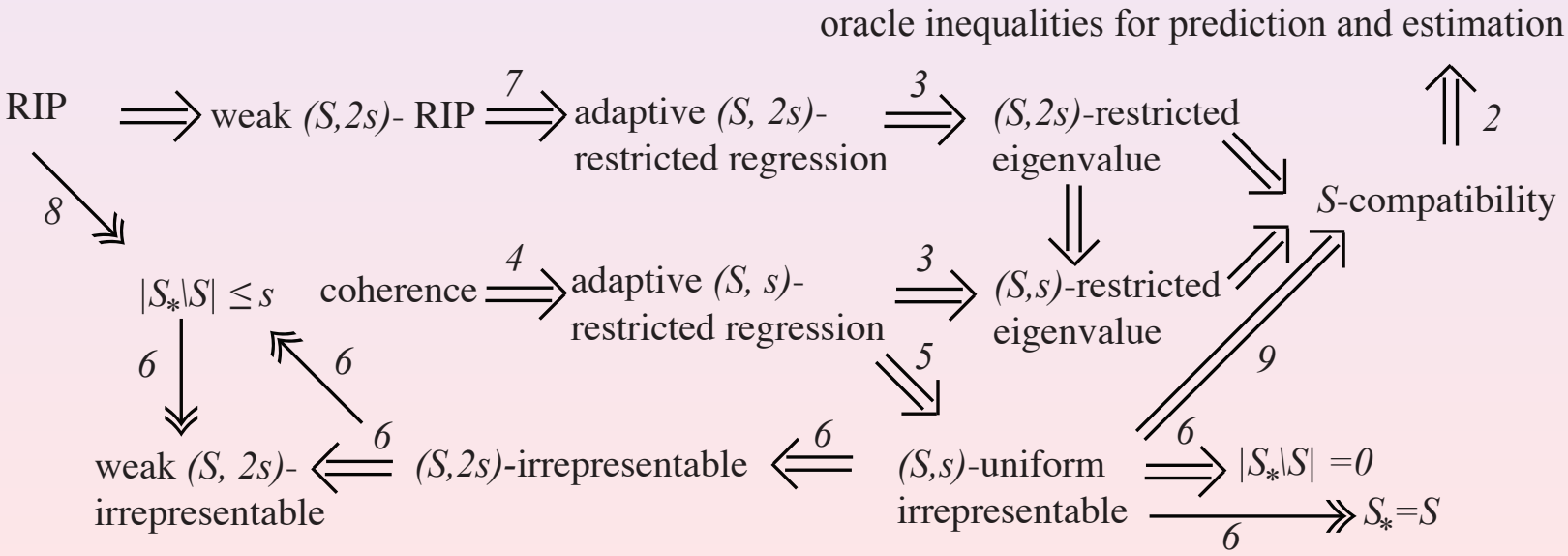


just make the appropriate assumptions to prove what you like...

real question:
how restrictive is compatibility condition (smallest restricted ℓ_1-eigenvalue)?

it is (slightly) weaker than the restricted eigenvalue assumption (Bickel, Ritov & Tsybakov, 2009)

more generally: (van de Geer & PB, 2009)



Does compatibility condition hold in practice?

it is non-checkable (in contrast to checkable but restrictive conditions Juditsky & Nemirovski (2008) ... which presumably would often fail in e.g. genomic data-sets)

assume that $X_1, \ldots, X_n$ i.i.d. with $\mathbb{E}[X] = 0,\ \mathrm{Cov}(X) = \Sigma$

- compatibility constant $\phi^2_{0,\Sigma}$ for Σ is bounded away from zero
  (maybe even the smallest eigenvalue of Σ is bounded away from zero)
- moment conditions for X (including e.g. Gaussian case)
- sparsity $s_0 = O(\sqrt{n/\log(p)})$

$\rightsquigarrow$

$$\phi^2_{0,\hat{\Sigma}} \geq \phi^2_{0,\Sigma}/2 \quad \text{with high probability}$$

(van de Geer & PB, 2009)
for sparse problems, the compatibility condition is "likely to hold"

Summary I (for Lasso)

for fixed design linear models:

fact 1:
no design conditions and mild assumption on error distribution:

- "slow" rate $n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 = \|\beta^0\|_1 O_P(\sqrt{\log(p)/n})$
  consistency for prediction if $\|\beta^0\|_1 = o(\sqrt{n/\log(p)})$

fact 2:
compatibility condition (or restricted eigenvalue condition) and mild assumption on error distribution:

- fast rate $n^{-1}\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2 \leq \frac{s_0}{\phi_0^2} O_P(\log(p)/n)$
- $\|\hat{\beta} - \beta^0\|_1 \leq \frac{s_0}{\phi_0^2} O_P(\sqrt{\log(p)/n})$

"myth": design assumptions for Lasso (in fact 2) are restrictive
"not really true" in the regime $s_0 = O(\sqrt{n/\log(p)})$

Remark:
fast convergence rate for prediction is possible without design conditions using

- ℓ_0-penalization (Barron, Birgé & Massart, 1999)
  computationally infeasible
- exponential weighting (Dalalyan & Tsybakov, 2008)
  computationally "cumbersome"

theory and methodology generalizes to
non-convex loss functions (GLMs),
additive models (Group Lasso), multitask models, ...
and "similar findings" with Dantzig selector, orthogonal matching pursuit, boosting,...

Variable selection

Example: Motif regression
for finding HIF1α transcription factor binding sites in DNA seq.
Müller, Meier, PB & Ricci

$Y_i \in \mathbb{R}$: univariate response measuring binding intensity of HIF1α on coarse DNA segment i (from CHIP-chip experiments)
$X_i = (X_i^{(1)}, \ldots, X_i^{(p)}) \in \mathbb{R}^p$:
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i (using sequence data and computational biology algorithms, e.g. MDSCAN)

question: relation between the binding intensity Y and the abundance of short candidate motifs?

$\rightsquigarrow$ linear model is often reasonable
"motif regression" (Conlon, X.S. Liu, Lieb & J.S. Liu, 2003)

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \; n = 287, \; p = 195$$

goal: variable selection
$\rightsquigarrow$ find the relevant motifs among the $p = 195$ candidates

Lasso for variable selection

$$\hat{S}(\lambda) = \{j;\ \hat{\beta}_j(\lambda) \neq 0\}$$

for $$S_0 = \{j; \beta_j^0 \neq 0\}$$

no significance testing involved

it's convex optimization only!

(and that can be a problem... see later)

Motif regression

for finding HIF1α transcription factor binding sites in DNA seq.

$Y_i \in \mathbb{R}$: univariate response measuring binding intensity on coarse DNA segment i (from CHIP-chip experiments)
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i

variable selection in linear model $Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i$,

$i = 1, \ldots, n = 287,\ p = 195$

$\rightsquigarrow$ Lasso selects 26 covariates and $R^2 \approx 50\%$
i.e. 26 interesting candidate motifs

Theory for the Lasso: Part II (variable selection)

for (fixed design) linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$ with
active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$
two key assumptions

1. neighborhood stability condition for design $\mathbf{X}$
   $\Leftrightarrow$ irrepresentable condition for design $\mathbf{X}$
2. beta-min condition

$$\min_{j \in S_0} |\beta_j^0| \geq C\sqrt{\log(p)/n}, \quad C \text{ suitably large}$$

both conditions are sufficient and "essentially" necessary for

$$\hat{S}(\lambda) = S_0 \text{ with high probability}, \quad \lambda \gg \underbrace{\sqrt{\log(p)/n}}_{\text{larger than for pred.}}$$

already proved in Meinshausen & PB, 2004 (publ: 2006)
and both assumptions are restrictive!

Theory for the Lasso: Part II (variable selection)

for (fixed design) linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$ with
active set $S_0 = \{j;\ \beta_j^0 \neq 0\}$
two key assumptions

1. neighborhood stability condition for design $\mathbf{X}$
   $\Leftrightarrow$ irrepresentable condition for design $\mathbf{X}$
2. beta-min condition

$$\min_{j \in S_0} |\beta_j^0| \geq C\sqrt{\log(p)/n}, \quad C \text{ suitably large}$$

both conditions are sufficient and "essentially" necessary for

$$\hat{S}(\lambda) = S_0 \text{ with high probability}, \quad \lambda \gg \underbrace{\sqrt{\log(p)/n}}_{\text{larger than for pred.}}$$

already proved in Meinshausen & PB, 2004 (publ: 2006)
and both assumptions are restrictive!

neighborhood stability condition $\Leftrightarrow$ irrepresentable condition

(Zhao & Yu, 2006)

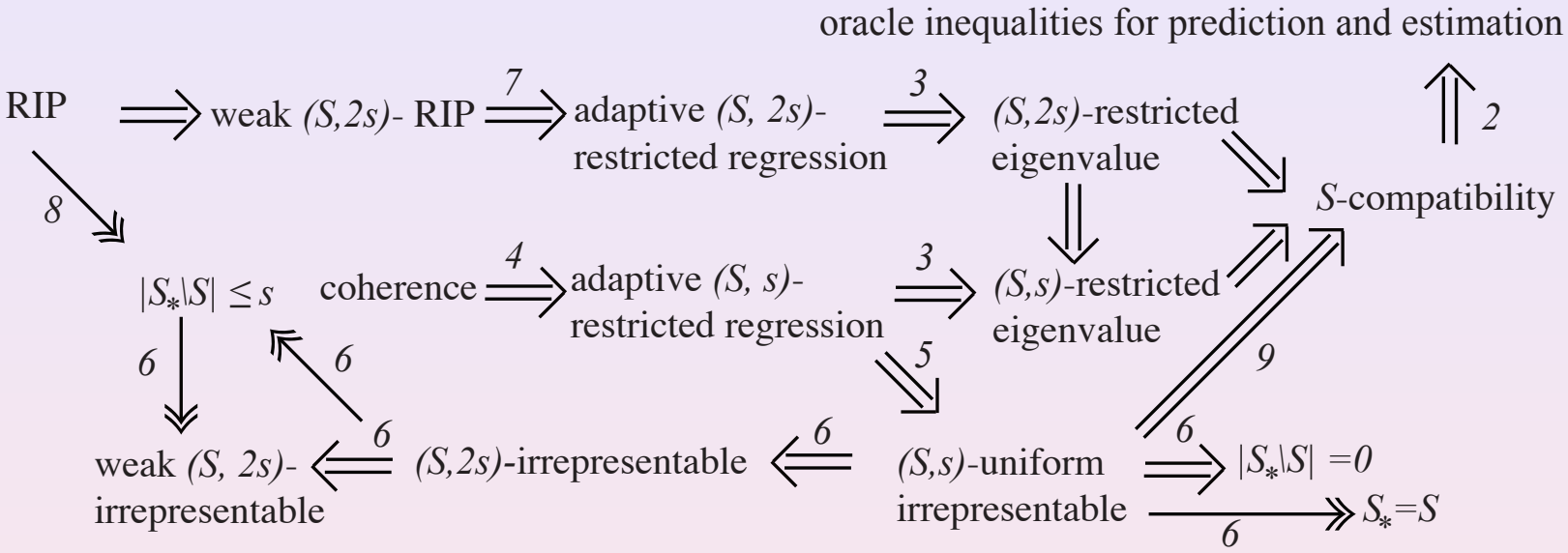
$$n^{-1}\mathbf{X}^T\mathbf{X} = \hat{\Sigma}$$

active set $S_0 = \{j;\ \beta_j \neq 0\} = \{1, \ldots, s_0\}$ consists of the first s_0 variables; partition

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{S_0,S_0} & \hat{\Sigma}_{S_0,S_0^c} \\ \hat{\Sigma}_{S_0^c,S_0} & \hat{\Sigma}_{S_0^c,S_0^c} \end{pmatrix}$$

irrep. condition : $\|\hat{\Sigma}_{S_0^c,S_0}\hat{\Sigma}_{S_0,S_0}^{-1}\operatorname{sign}(\beta_1^0, \ldots, \beta_{s_0}^0)^T\|_\infty < 1$

various design conditions (van de Geer & PB, 2009)



irrepresentable condition is (much more) restrictive than the compatibility condition
(and irrepresentable condition is necessary for recovery of S_0 with the Lasso)

not very realistic assumptions... what can we expect?

recall: under compatibility condition and mild assumption on error distribution

$$\|\hat{\beta} - \beta^0\|_1 \leq C \frac{s_0}{\phi_0^2} \sqrt{\log(p)/n}$$

consider the relevant active variables

$$S_{\text{relev}} = \{j;\ |\beta_j^0| > C \frac{s_0}{\phi_0^2} \sqrt{\log(p)/n}\}$$

then, clearly,

$$\hat{S} \supseteq S_{\text{relev}} \text{ with high probability}$$

screening for detecting the relevant variables is possible!
without beta-min condition and assuming compatibility condition only

in addition: assuming beta-min condition

$$\min_{j \in S_0} |\beta_j^0| > C \frac{s_0}{\phi_0^2} \sqrt{\log(p)/n}$$

$$\hat{S} \supseteq S_0 \text{ with high probability}$$

screening for detecting the true variables

Tibshirani (1996):
LASSO = Least Absolute Shrinkage and Selection Operator

new translation PB (2010):
LASSO = Least Absolute Shrinkage and Screening Operator

Practical perspective

choice of λ: $\hat{\lambda}_{CV}$ from cross-validation
empirical and theoretical indications (Meinshausen & PB, 2006) that

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

moreover

$$|\hat{S}(\hat{\lambda}_{CV})| \leq \min(n, p) (= n \text{ if } p \gg n)$$

$\rightsquigarrow$ huge dimensionality reduction (in the original covariates)

recall:

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

and we would then use a second-stage to reduce the number of false positive selections

$\rightsquigarrow$ re-estimation on much smaller model with variables from $\hat{S}$

- OLS on $\hat{S}$ with e.g. BIC variable selection
- thresholding coefficients and OLS re-estimation (Zhou, 2009)
- adaptive Lasso (Zou, 2006)
- ...

recall:

$$\hat{S}(\hat{\lambda}_{CV}) \supseteq S_0 \quad (\text{or } S_{\text{relev}})$$

and we would then use a second-stage to reduce the number of false positive selections

$\rightsquigarrow$ re-estimation on much smaller model with variables from $\hat{S}$

- OLS on $\hat{S}$ with e.g. BIC variable selection
- thresholding coefficients and OLS re-estimation (Zhou, 2009)
- adaptive Lasso (Zou, 2006)
- ...

Summary II (for Lasso)

variable selection: estimation of $S_0 = \{j;\ \beta_j^0 \neq 0\}$ requires (necessarily)

- irrepresentable condition for design
- beta-min condition on the coefficients

both of them are restrictive

but variable Screening is more realistic
assuming compatibility condition on the design (smallest restricted ℓ_1-eigenvalue)

$$\hat{S}(\lambda) \supseteq S_{\text{relev}},$$

assuming beta-min cond.: $\hat{S}(\lambda) \supseteq S_0$

also here: mainly focused on the Lasso in linear models

many extensions have been worked out:
Group Lasso, Fused Lasso, sparsity-smoothness penalty, Dantzig-selector,...
concave penalties: SCAD, MC+, and related adaptive Lasso,...
Orthogonal matching pursuit, boosting,...
marginal screening (sure independence screening),...

empirical and theoretical results are “similar”

- prediction is “easy”
- estimation of parameters and variable screening is often “reasonably accurate”
- variable selection is “hard”

Gaussian graphical models

$$X_1, \ldots, X_n \text{ i.i.d. } \sim \mathcal{N}_p(0, \Sigma)$$

goal: infer zeroes of Σ^{-1}:

$$\Sigma^{-1}_{jk} \neq 0 \quad \Leftrightarrow \quad X^{(j)} \not\perp X^{(k)} | X^{(\{1,\ldots,p\}\setminus\{j,k\})} \quad \Leftrightarrow \quad \text{edge } j-k$$

nodewise regression can do the job:

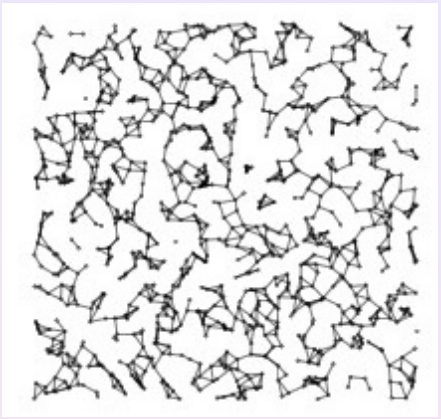
$$X^{(j)} = \sum_{k \neq j} \beta_k^{(j)} X^{(k)} + \varepsilon^{(j)}, \quad j = 1, \ldots, p$$

$\rightsquigarrow$

$$\beta_k^{(j)} \neq 0, \ \beta_j^{(k)} \neq 0 \Leftrightarrow \Sigma^{-1}_{jk} \neq 0 \Leftrightarrow \text{edge } j-k$$

Meinshausen & PB (2006):
p Lasso regressions $\rightsquigarrow \hat{\beta}^{(j)}$

$$\begin{aligned} & \text{estimate edge } j - k \\ \Leftrightarrow \quad & \hat{\beta}_k^{(j)} \neq 0 \text{ and/or } \hat{\beta}_j^{(k)} \neq 0 \end{aligned}$$



does not use the constraint of positive definiteness for Σ

but for inferring edge set (support estimation):
uncoupled nodewise regression requires substantially weaker irrepresentable condition than simultaneous GLasso approach based on multivariate Gaussian likelihood
(Friedman et al., 2007; Banerjee et al., 2008)
see Meinshausen (2004; publ. 2008)

Back to variable selection in regression

Motif regression for finding HIF1α transcription factor binding sites in DNA sequences

$Y_i \in \mathbb{R}$: univariate response measuring binding intensity on coarse DNA segment i (from CHIP-chip experiments)
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i

variable selection in linear model $Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i$,

$i = 1, \ldots, n = 287,\ p = 195$

$\rightsquigarrow$ Lasso selects 26 covariates and $R^2 \approx 50\%$
i.e. 26 interesting candidate motifs
and hence report these findings to the biologists...

really?
how stable are the findings?

Back to variable selection in regression

Motif regression for finding HIF1α transcription factor binding sites in DNA sequences

$Y_i \in \mathbb{R}$: univariate response measuring binding intensity on coarse DNA segment i (from CHIP-chip experiments)
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i

$$\text{variable selection in linear model } Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i,$$

$$i = 1, \ldots, n = 287, \; p = 195$$

$\rightsquigarrow$ Lasso selects 26 covariates and $R^2 \approx 50\%$
i.e. 26 interesting candidate motifs
and hence report these findings to the biologists...

really?

how stable are the findings?

Back to variable selection in regression

Motif regression for finding HIF1α transcription factor binding sites in DNA sequences

$Y_i \in \mathbb{R}$: univariate response measuring binding intensity on coarse DNA segment i (from CHIP-chip experiments)
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment i

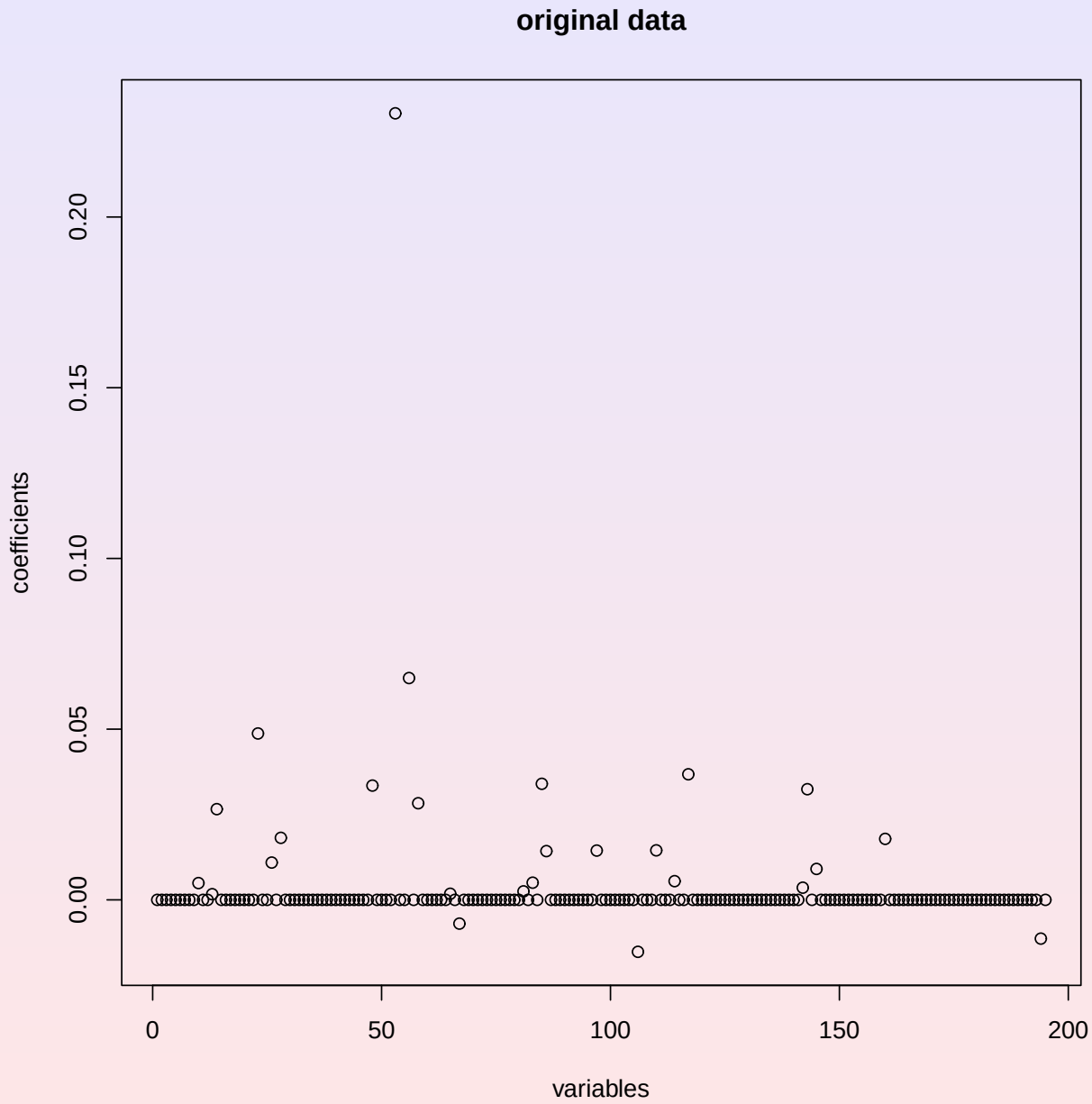
variable selection in linear model $Y_i = \mu + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i$,

$i = 1, \ldots, n = 287,\ p = 195$

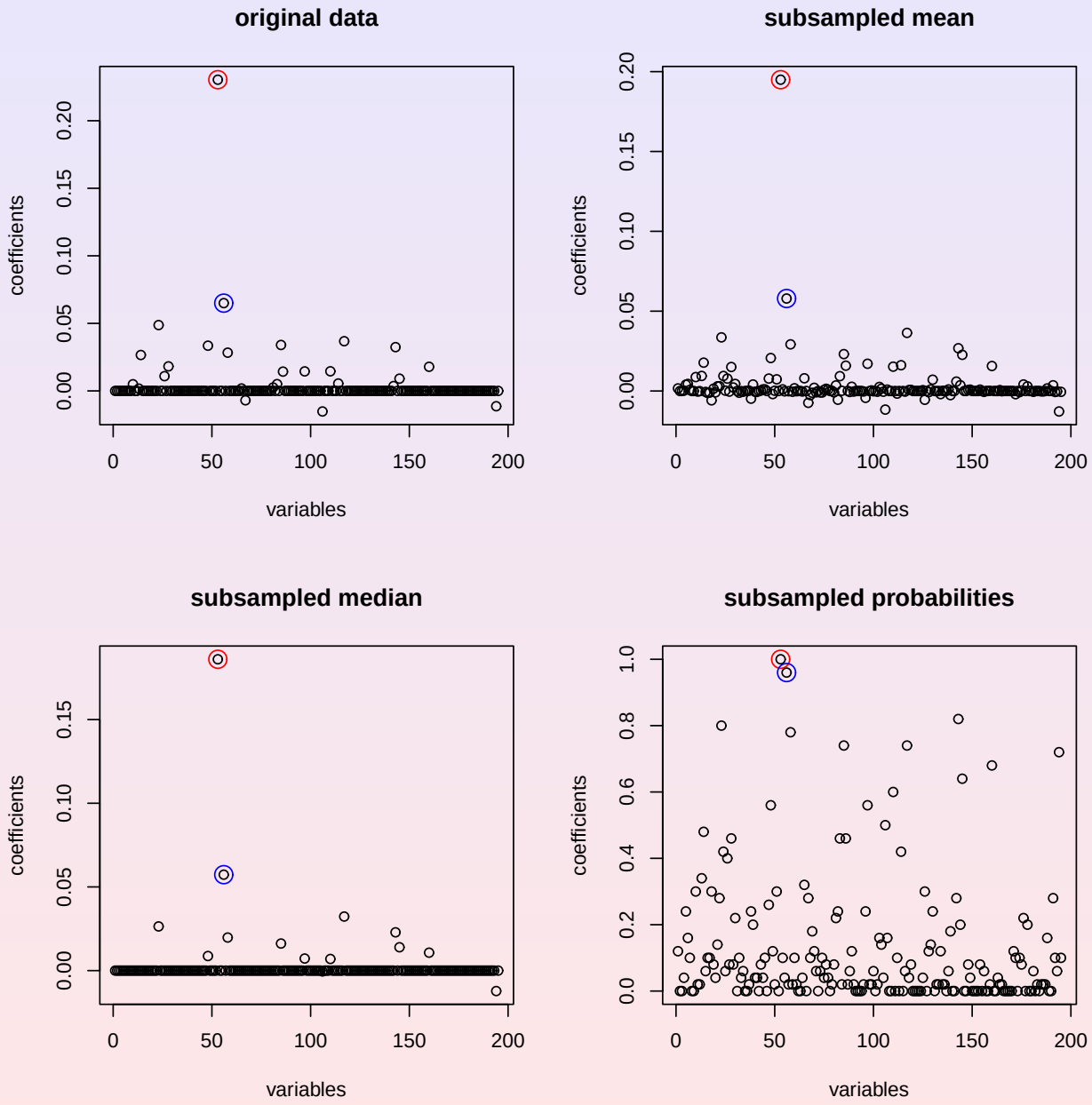
$\rightsquigarrow$ Lasso selects 26 covariates and $R^2 \approx 50\%$
i.e. 26 interesting candidate motifs
and hence report these findings to the biologists...

really?
how stable are the findings?

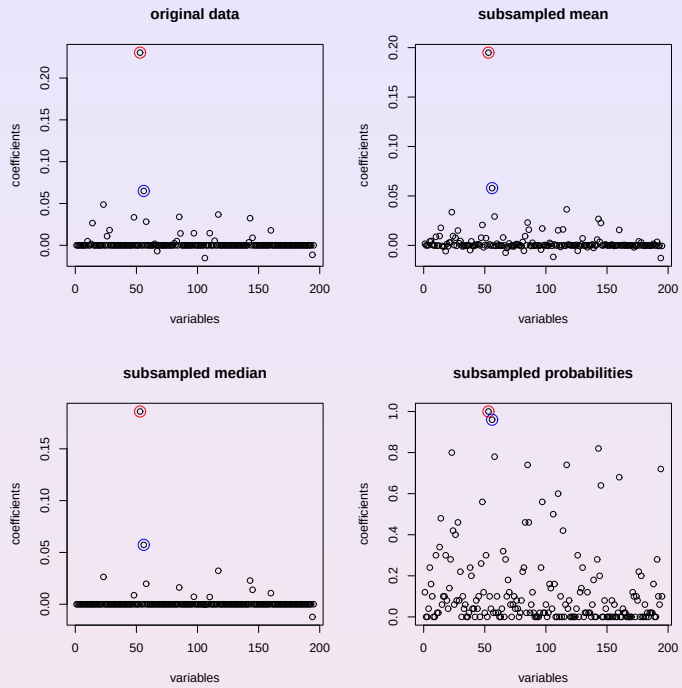
estimated coefficients $\hat{\beta}(\hat{\lambda}_{\mathrm{CV}})$



stability check: subsampling with subsample size $\lfloor n/2 \rfloor$



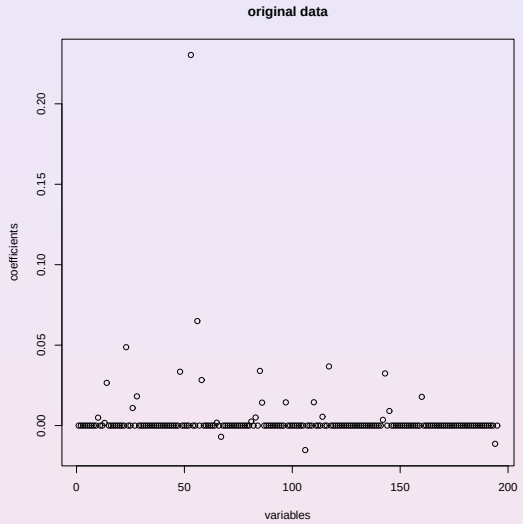
$\rightsquigarrow$ only 2 “stable” findings
($\neq$ 26)



one variable (○):
corresponds to true, known motif
other variable (○): good additional support for relevance
(nearness to transcriptional start-site of important genes, ...)

“learning” from the example:

- using $\hat{S}(\lambda) = \{j;\ \hat{\beta}_j(\lambda) \neq 0\}$ for S_0 is questionable



- from theoretical point of view, many things can go wrong

as I explained for Lasso (and also true for many other methods)

- assigning uncertainty is completely missing

Stability Selection (Meinshausen & PB, 2010)

using subsampling (or bootstrapping)

consider (first) linear model setting

$$Y_i = (\mu +) \sum_{j=1}^{p} \beta_j^0 X_i^{(j)} + \varepsilon_i, \ i = 1, \ldots, n \ (\ll p)$$

set of active variables: $S_0 = \{j; \ \beta_j^0 \neq 0\}$

variable selection procedure:

$$\hat{S}^\lambda \subseteq \{1, \ldots, p\},$$

λ a tuning parameter

prime example: Lasso (Tibshirani, 1996)

subsampling:

- draw sub-sample of size $\lfloor n/2 \rfloor$ without replacement, denoted by $I^* \subseteq \{1, \ldots, n\}$, $|I^*| = \lfloor n/2 \rfloor$
- run the selection algorithm $\hat{S}^\lambda(I^*)$ on I^*
- do these steps many times and compute the relative selection frequencies

$$\hat{\Pi}_j^\lambda = P^*(j \in \hat{S}^\lambda(I^*)),\ j = 1, \ldots, p$$

P^* is w.r.t. sub-sampling (and maybe other sources of randomness if a randomized selection algorithm is invoked)

could also use bootstrap sampling with replacement...

subsampling:

- draw sub-sample of size $\lfloor n/2 \rfloor$ without replacement, denoted by $I^* \subseteq \{1, \ldots, n\}$, $|I^*| = \lfloor n/2 \rfloor$
- run the selection algorithm $\hat{S}^\lambda(I^*)$ on I^*
- do these steps many times and compute the relative selection frequencies

$$\hat{\Pi}_j^\lambda = P^*(j \in \hat{S}^\lambda(I^*)),\ j = 1, \ldots, p$$

P^* is w.r.t. sub-sampling (and maybe other sources of randomness if a randomized selection algorithm is invoked)

could also use bootstrap sampling with replacement...

Stability selection

$$\hat{S}^{\text{stable}} = \{j;\ \hat{\Pi}_j^{\lambda} \geq \pi_{\text{thr}}\}$$

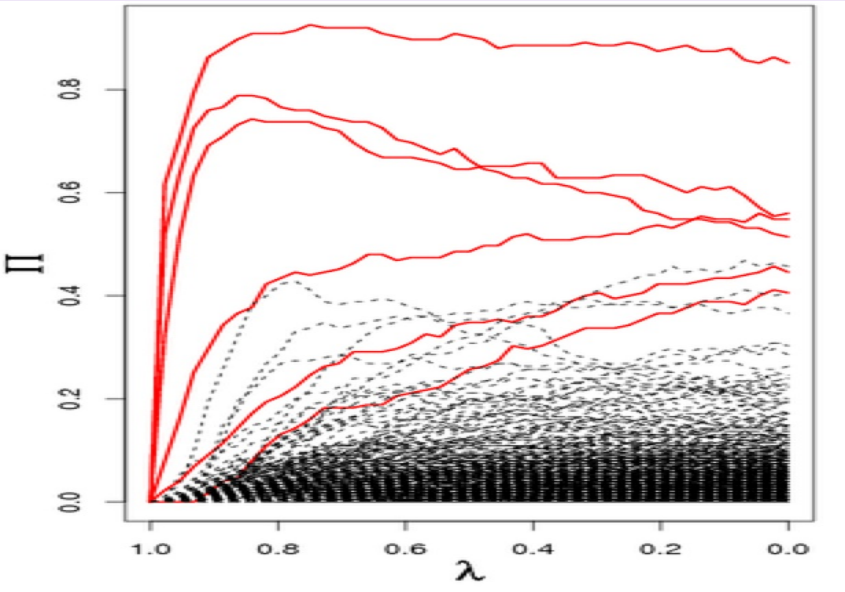
depends on λ via $\hat{\Pi}_j^{\lambda} = P^*(j \in \hat{S}^{\lambda}(I^*))$

choice of $\pi_{\text{thr}} \rightsquigarrow$ see later

if we consider many regularization parameters:

$$\{\hat{S}^{\lambda};\ \lambda \in \Lambda\}$$

Λ can be discrete, a singleton or continuous



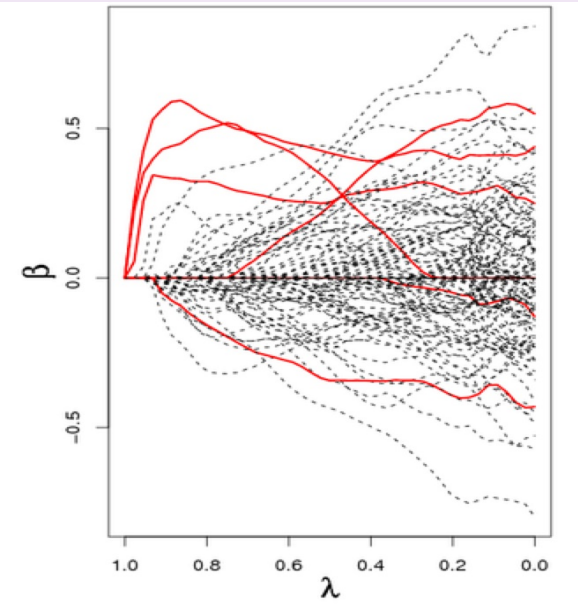
$$\hat{S}^{\text{stable}} = \{j;\ \max_{\lambda \in \Lambda} \hat{\Pi}_j^{\lambda} \geq \pi_{\text{thr}}\}$$

see also Bach (2009) for a related proposal

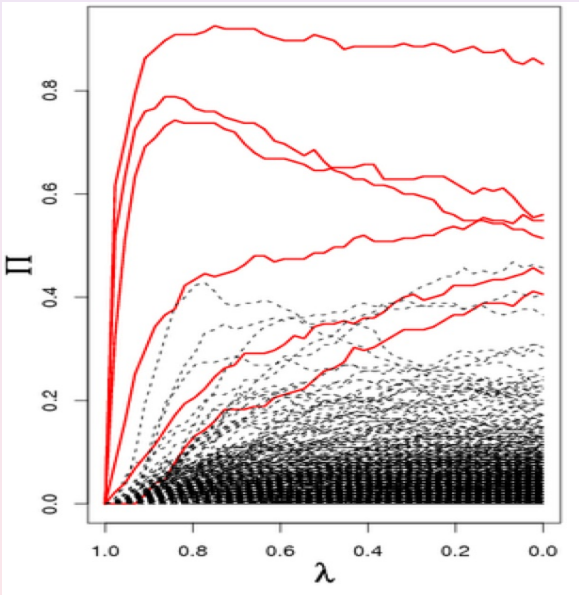
The Lasso and its corresponding stability path

$Y =$ riboflavin production rate in Bacillus Subtilis (log-scale)
X: $p = 4088$ gene expressions (log-scale),
sparsity p_{eff} "=" 6 (6 "relevant" genes;
all other variables permuted)
sample size $n = 115$

Lasso



Stability selection

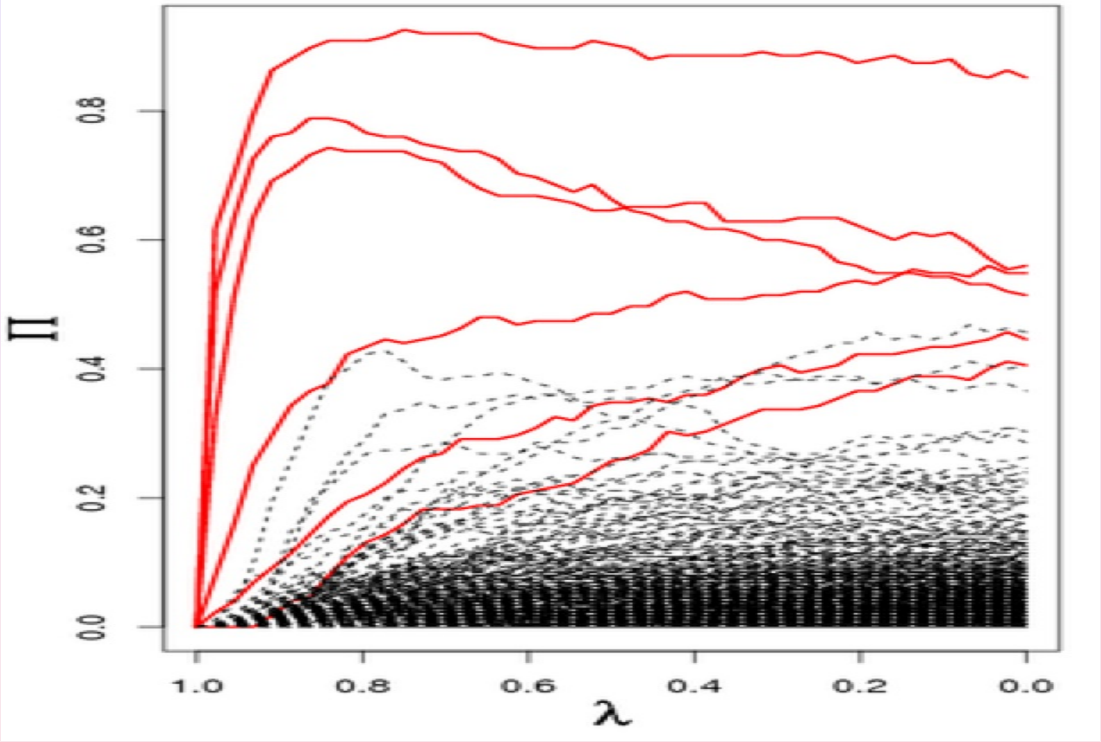


with stability selection: the 4-6 "true" variables are sticking out much more clearly from noise covariates

stability selection cannot be reproduced by simply selecting the right penalty with Lasso

stability selection provides a fundamentally new solution

Choice of threshold $\pi_{\text{thr}} \in (0, 1)$?



How to choose the threshold π_{thr}?

consider a selection procedure which selects q variables
(e.g. top 50 variables when running Lasso over many λ's)

denote by $V = |S_0^c \cap \hat{S}^{\text{stable}}| =$ number of false positives

Theorem (Meinshausen & PB, 2010)
main assumption: exchangeability condition
in addition: $\hat{S}$ has to be better than "random guessing"
Then:

$$E(V) \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q^2}{p}$$

i.e. finite sample control, even if $p \gg n$
$\rightsquigarrow$ choose threshold π_{thr} to control e.g. $E[V] \leq 1$ or
$P[V > 0] \leq E[V] \leq \alpha$

note the generality of the Theorem...

- it works for any method which is better than “random guessing”
- it works not only for regression but also for “any” discrete structure estimation problem (whenever there is a include/exclude decision)
  $\rightsquigarrow$ variable selection, graphical modeling, clustering, ...

and hence there must be a fairly strong condition...
Exchangeability condition:
the distribution of $\{I_{\{j \in \hat{S}^\lambda\}};\ j \in S_0^c\}$ is exchangeable
note: only some requirement for noise variables

note the generality of the Theorem...

- it works for any method which is better than “random guessing”
- it works not only for regression but also for “any” discrete structure estimation problem (whenever there is a include/exclude decision)
  $\rightsquigarrow$ variable selection, graphical modeling, clustering, ...

and hence there must be a fairly strong condition...
Exchangeability condition:
the distribution of $\{I_{\{j \in \hat{S}^\lambda\}};\ j \in S_0^c\}$ is exchangeable
note: only some requirement for noise variables

Discussion of the conditions in case of
random design linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$:

- no beta-min condition
  (but the Theorem is only about false positives)

- exchangeability condition is restrictive:
  example where it holds: $\Sigma = \mathrm{Cov}(X)$ from equicorrelation

the theory is (as of now) too rough and does not indicate better theoretical behavior for variable selection than for adaptive Lasso (or thresholded Lasso)

Discussion of the conditions in case of
random design linear model $\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon$:

- no beta-min condition
  (but the Theorem is only about false positives)

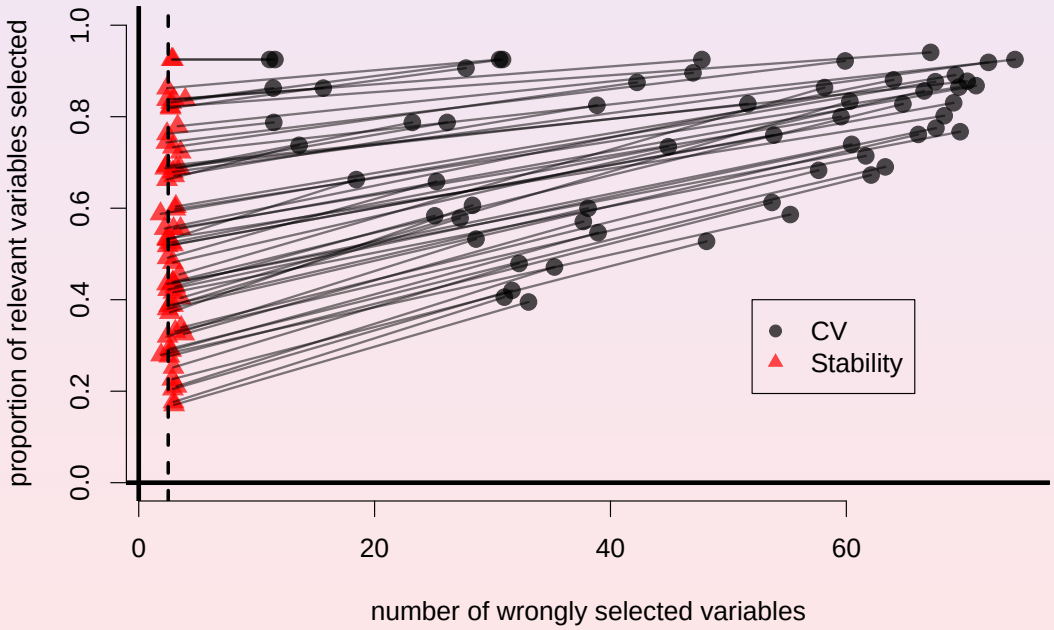
- exchangeability condition is restrictive:
  example where it holds: $\Sigma = \mathrm{Cov}(X)$ from equicorrelation

the theory is (as of now) too rough and does not indicate better theoretical behavior for variable selection than for adaptive Lasso (or thresholded Lasso)

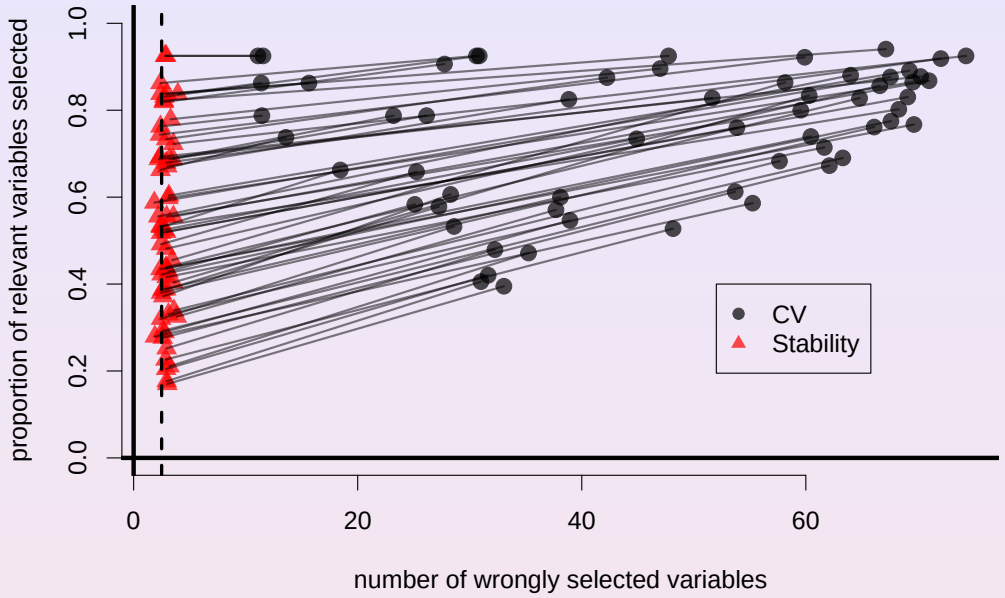
Some numerical experiments

Variable selection in linear models using Lasso
a range of scenarios:
$p = 660$ with design from a real data set about motif regression
$n \in \{450, 750\}$, sparsity $p_{eff} \in \{4, 8, \ldots, 40\}$ (using artificial β)
signal to noise ratio $\in \{0.25, 1, 4\}$

control for $E[V] \leq 2.5$



control for $E[V] \leq 2.5$



stability selection yields:

- accurate control (as proved in theory)
- drastic reduction of false positives in comparison to CV-tuned solution
- not much loss in terms of power (true positives)

Motif regression

stability selection with $\mathbb{E}[V] \leq 1$
$\rightsquigarrow$ two stably selected variables/motifs

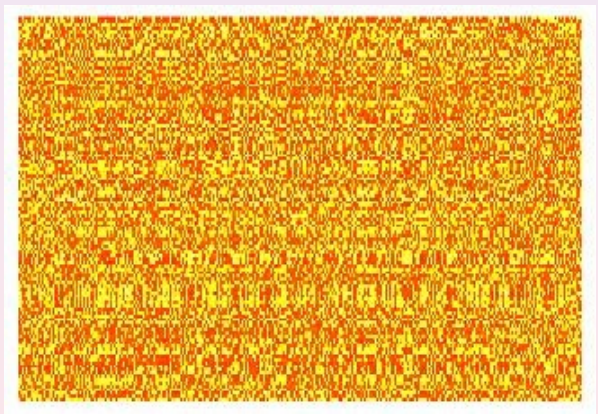
one of them is a known binding site

Graphical modeling using GLasso

(Rothman, Bickel, Levina & Zhu, 2008; Friedman, Hastie & Tibshirani, 2008)

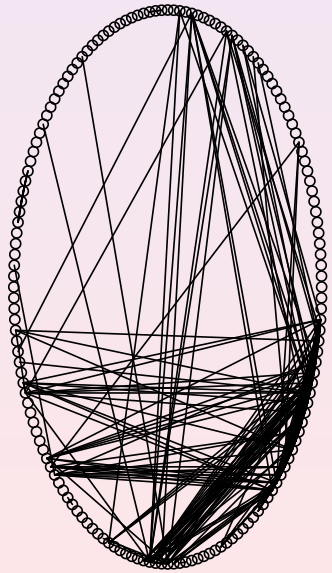
infer conditional independence graph using ℓ_1-penalization
i.e. infer zeroes of Σ^{-1} from $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma)$

$$\Sigma^{-1}_{jk} \neq 0 \quad \Leftrightarrow \quad X^{(j)} \not\perp X^{(k)} | X^{(\{1,\ldots,p\} \setminus \{j,k\})} \quad \Leftrightarrow \quad \text{edge } j - k$$



gene expr. data

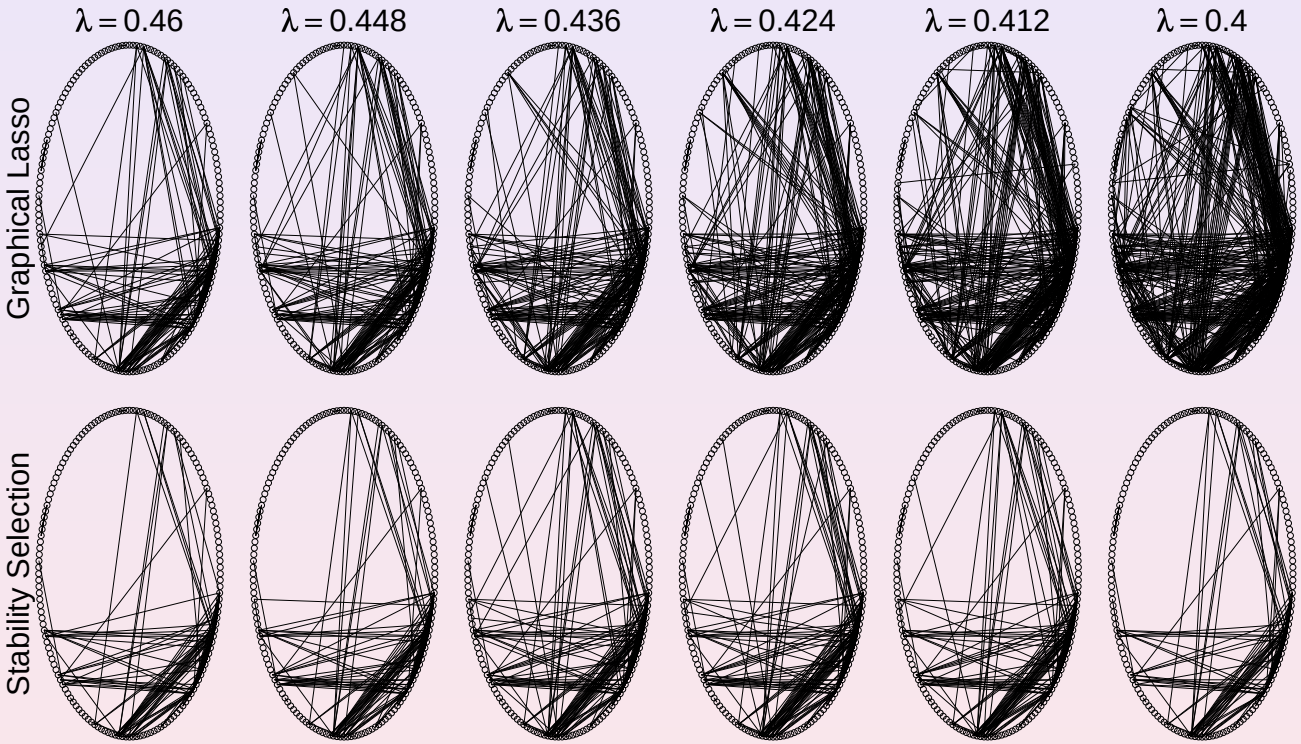
$\overset{?}{\Longrightarrow}$



zero-pattern of Σ^{-1}

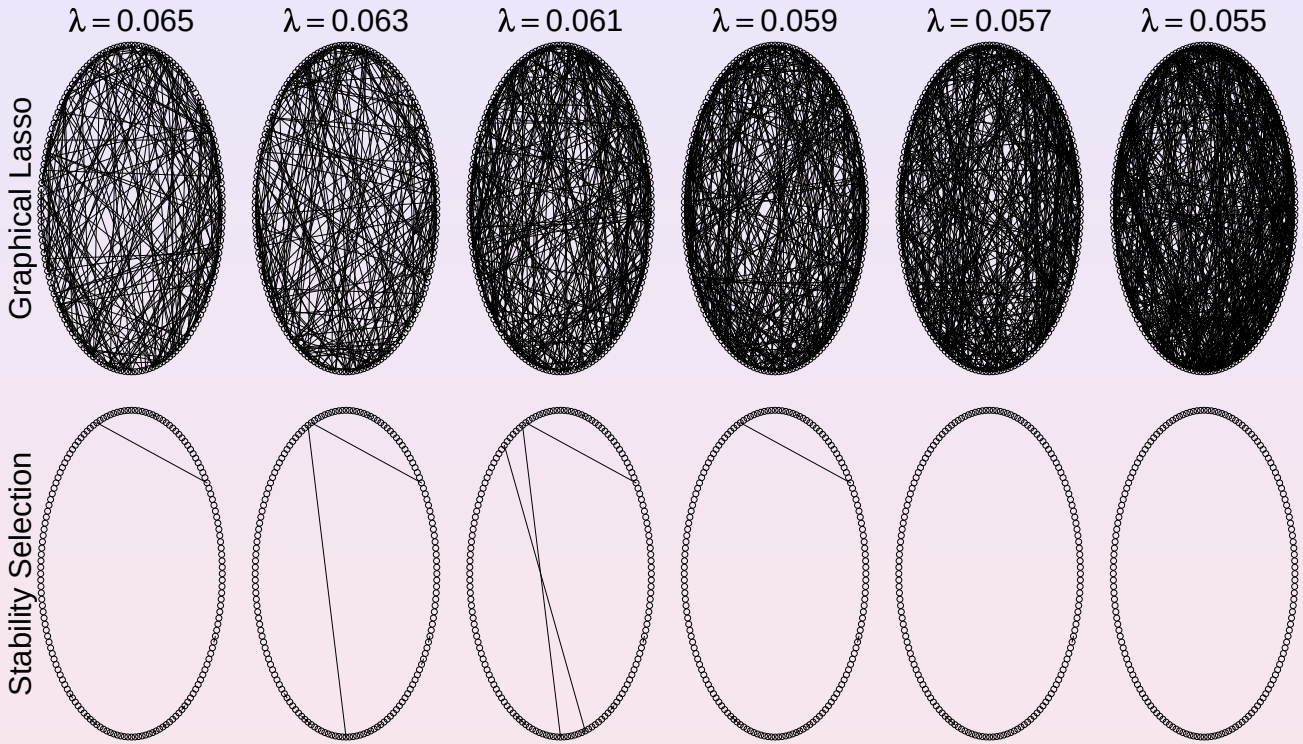
sub-problem of riboflavin production with bacillus subtilis
$p = 160,\ n = 115$
stability selection with $E[V] \leq 5$

varying the regularization parameter λ in ℓ_1-penalization



with stability selection: choice of initial λ-tuning parameter does not matter much (as proved by our theory)
just need to fix the finite-sample control

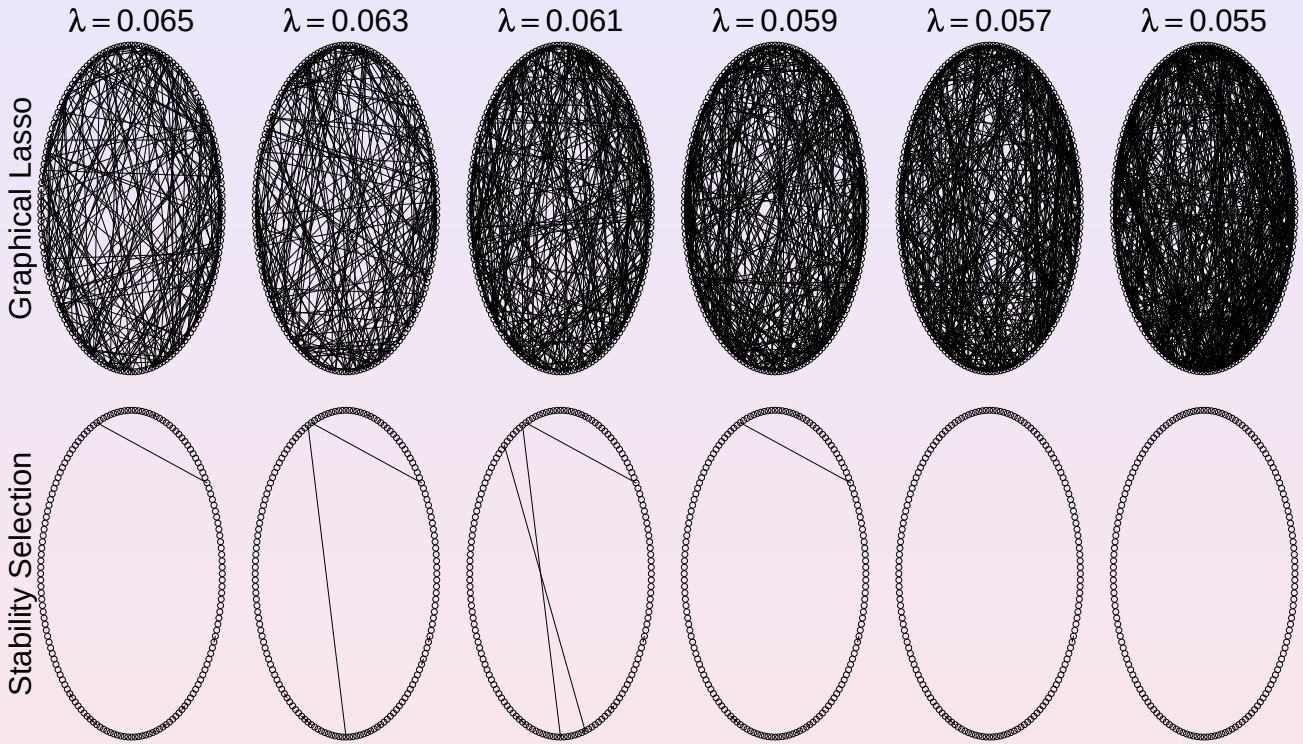
permutation of variables
varying the regularization parameter for the null-case



with stability selection: the number of false positives is indeed controlled (as proved by our theory)

and here: exchangeability condition holds

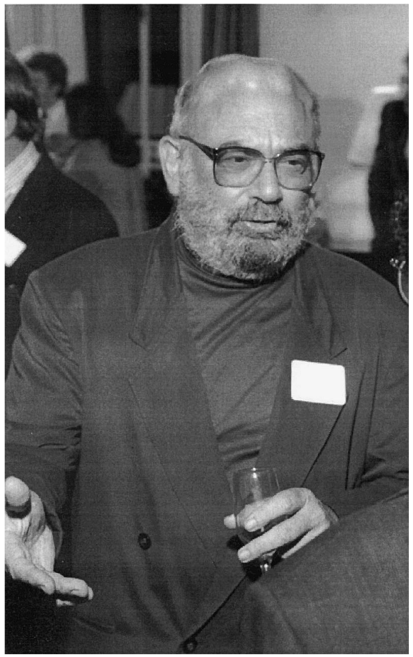
permutation of variables
varying the regularization parameter for the null-case



with stability selection: the number of false positives is indeed controlled (as proved by our theory)
and here: exchangeability condition holds

stability selection is
Bagging the selection outcomes (instead of prediction)

Leo Breiman



and providing error control
in terms of $E[V]$ ($\rightsquigarrow$ conservative FWER control)

Conclusions Part I

for the Lasso (and other computationally feasible methods)
in linear models (and other models):

property	design condition	size of non-zero coeff.
slow converg. rate	no requirement	no requirement
fast converg. rate	restricted eigenvalue	no requirement
variable screening	restricted eigenvalue	beta-min condition
variable selection	neighborhood stability $\Leftrightarrow$ irrepresentable cond.	beta-min condition

for more reliable results in practice, in particular for
variable/feature selection: need something on top of it
$\rightsquigarrow$ e.g. stability selection

Variable selection for causal target

regression is for quantifying association

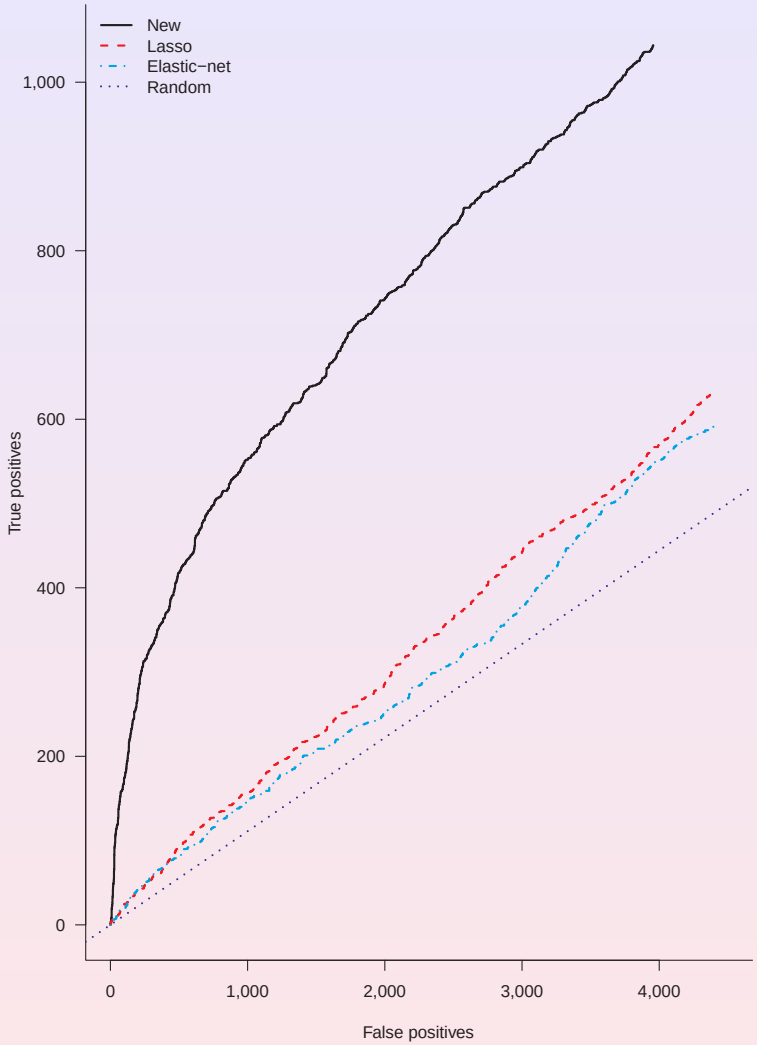
for some applications we need something else

Gene knock-downs in yeast

$p = 5360$ genes
question:
if we would knock-down a single gene, what would be its effect on all other genes?

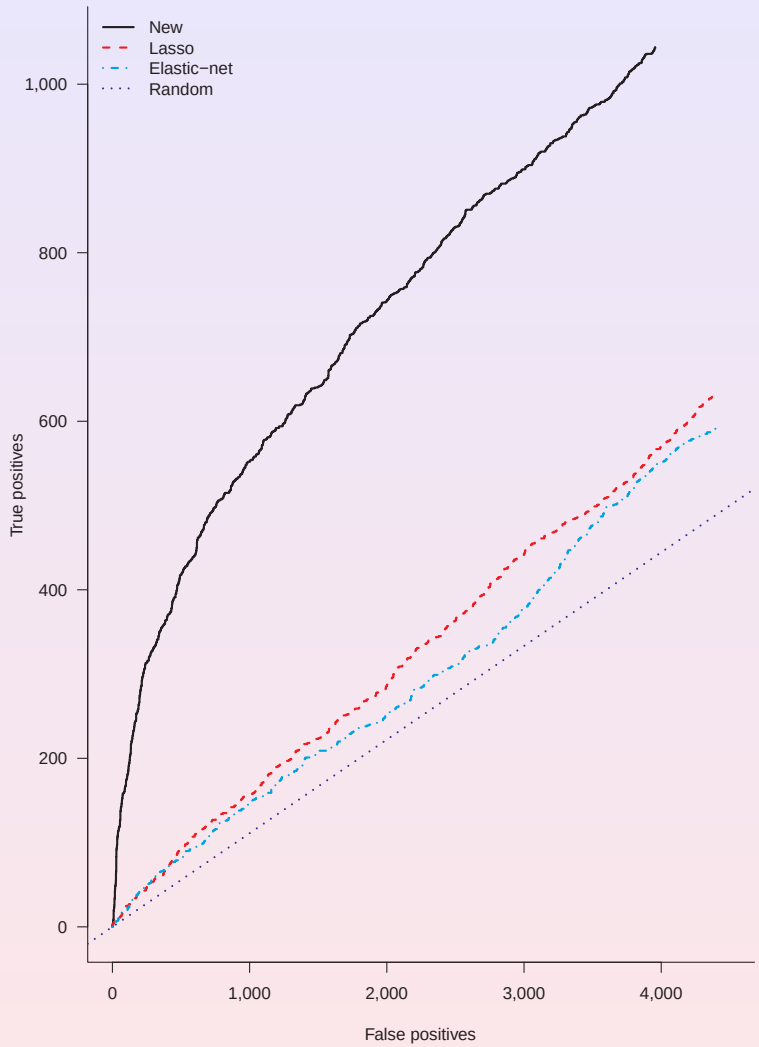
goal:
want to infer/predict such effects without actually doing the intervention
i.e. from observational data

Figure 1



↝ look beyond penalized regression/classification!

Figure 1



$\rightsquigarrow$ look beyond penalized regression/classification!

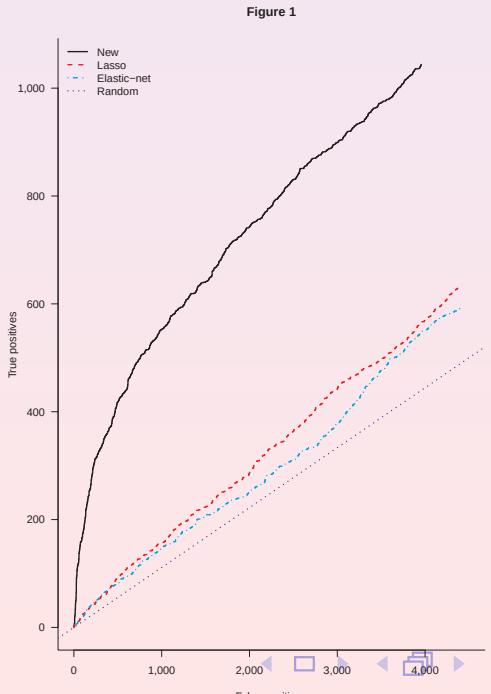
Effects of single gene knock-downs on all other genes (yeast)

(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)

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

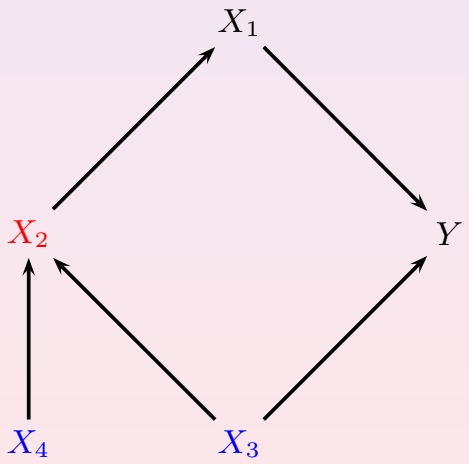
$\rightsquigarrow$ not very realistic for intervention problem
if we change e.g. one gene, some others will also change
and these are not (cannot be) kept fixed

Intervention calculus

"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



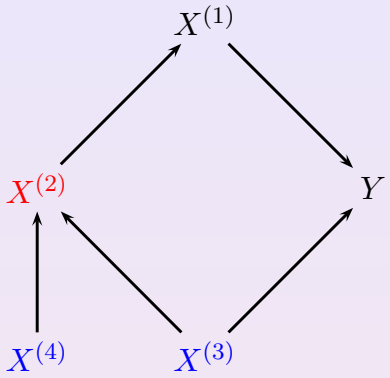
for simplicity: just consider DAGs
(ancestral graphs with hidden variables: more involved)

for DAGs: 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)

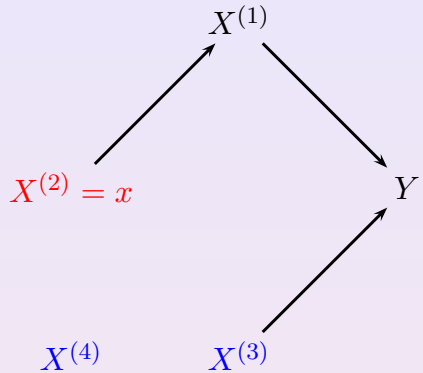
non-intervention



independent errors &
autonom strcl. eqns.
$\Leftrightarrow$ Markov assump:

$$P(Y, X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}) =$$
$$P(Y|X^{(1)}, X^{(3)}) \times$$
$$P(X^{(1)}|X^{(2)}) \times$$
$$P(X^{(2)}|X^{(3)}, X^{(4)}) \times$$
$$P^{(3)} \times$$
$$P^{(4)}$$

intervention at $X^{(2)}$



independent errors &
autonom strcl. eqns:

$$P(Y, X^{(1)}, X^{(3)}, X^{(4)}|\mathrm{do}(X^{(2)} = x)) =$$
$$P(Y|X^{(1)}, X^{(3)}) \times$$
$$P(X^{(1)}|X^{(2)} = x) \times$$
$$P^{(3)} \times$$
$$P^{(4)}$$

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

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

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

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
  $\rightsquigarrow$ need to estimate conditional distributions
- an influence diagram (causal DAG)
  $\rightsquigarrow$ 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

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

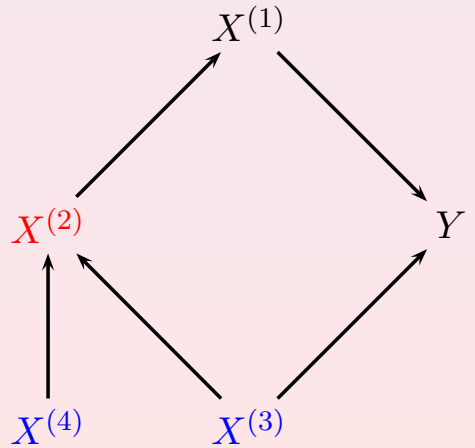
for $Y \notin \mathrm{pa}(j)$:

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 observational data

main problem: inferring DAG from observational data
$\rightsquigarrow$ impossible: can only infer equivalence class of DAGs
(several DAGs can encode exactly the same conditional independence relationships)

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

here:

$$P \text{ and graph } \mathcal{G} \Rightarrow \text{ parameter } \theta(P, \mathcal{G})$$

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: $\mathcal{G}_1, \ldots, \mathcal{G}_m$
- for every DAG-member $\mathcal{G}_r$, 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

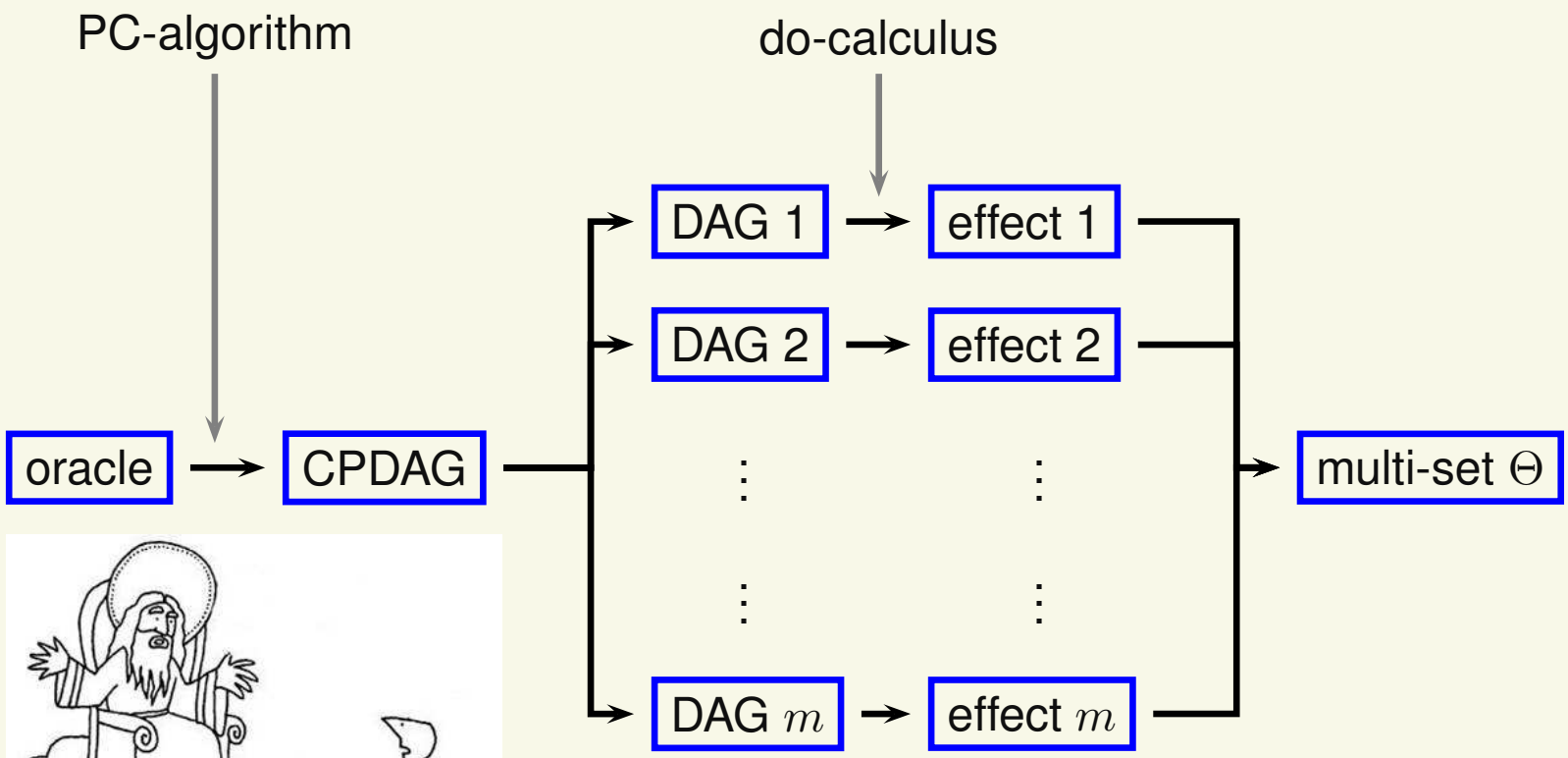
but we will be able to estimate lower bounds of causal effects

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: $\mathcal{G}_1, \ldots, \mathcal{G}_m$
- for every DAG-member $\mathcal{G}_r$, 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

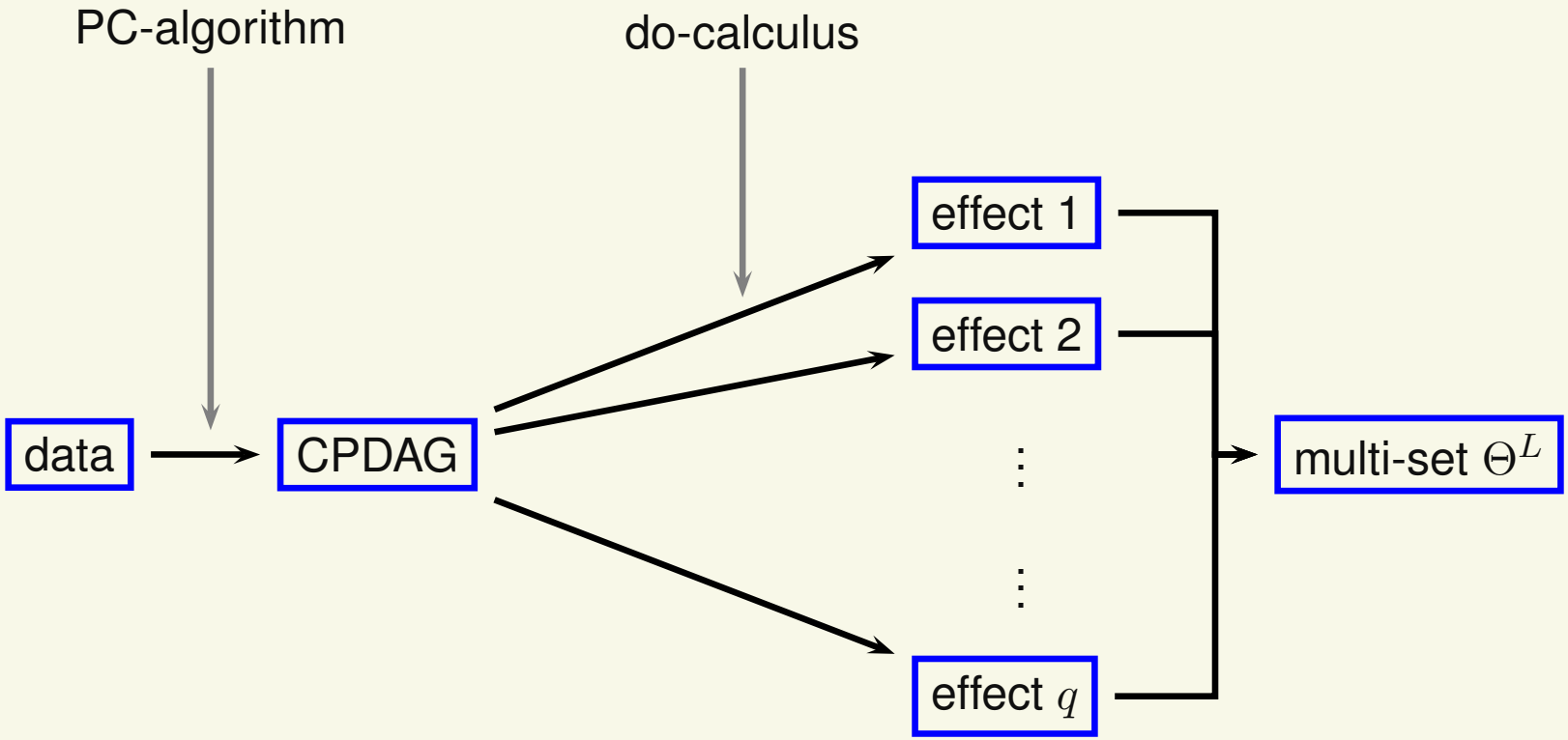
searching for all DAGs is computationally infeasible if p is large (we actually can do this up to $p \approx 15$)

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

Maathuis, Kalisch & PB (2009):

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

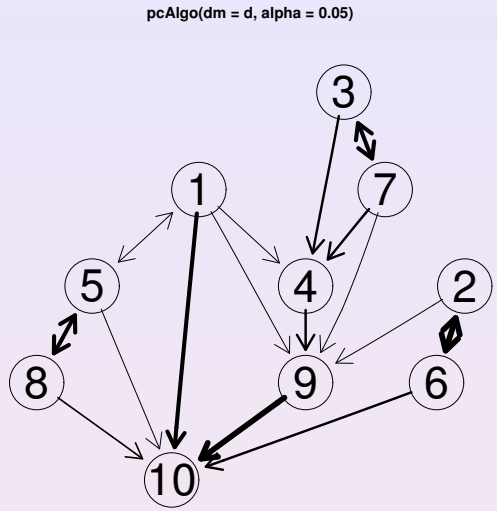
IDA (local sample version)



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

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

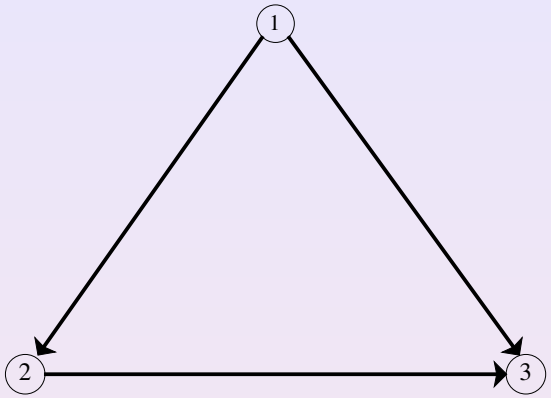
so-called faithfulness assumption allows to reduce to "some subsets S"

Faithfulness assumption

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

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:
   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)

4. move up to partial correlations of order 2:
   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)
5. until removal of edges is not possible anymore

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

the trick is:

Local computations on graphs

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 $\gg n^{-1/2}$ ("signal strength")
  $\min\{|\rho_{n;i,j|S}|; \rho_{n;i,j|S} \neq 0,\ i \neq j,\ |S| \leq d_n\} \gg n^{-1/2}$
- maximal (partial) correlations $\leq C < 1$ ("coherence")
  $\max_{i \neq j; |S| \leq d_n} |\rho_{n;i,j|S}| \leq C < 1$

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)

Criticisms

two main conditions:

- faithfulness assumption
  is it restrictive?
- non-zero partial correlations sufficiently large
  this is the analogue of the beta-min condition in regression

The role of sparsity in causal inference

as usual, sparsity is useful/necessary for estimation in presence of noise

but here: sparsity is crucial for identifiability as well

X causes Y

Y causes X

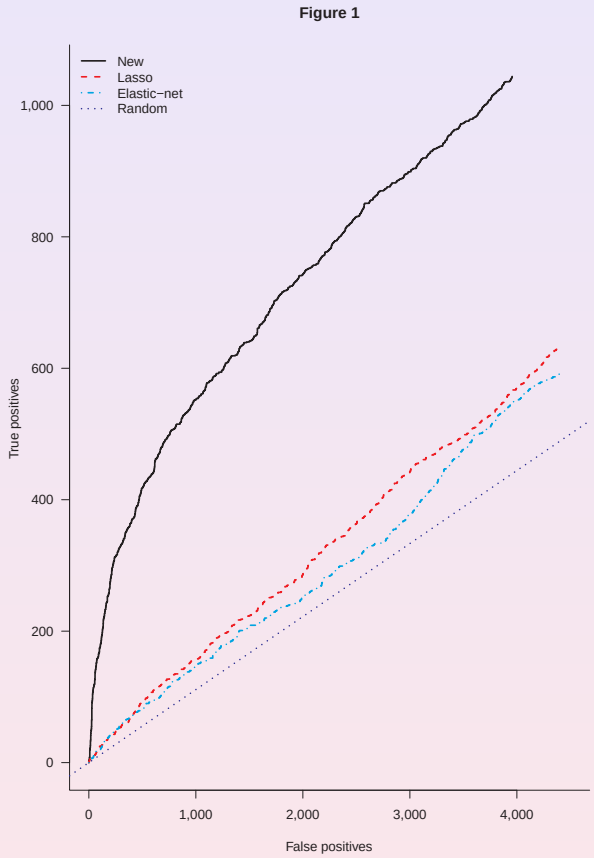
cannot tell from observational data the direction of the arrow

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

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 $p = 21'326$ A. thaliana ecotypes (D. Weigel, Tübingen) and L. Hennig/W. Gruissem (ETH Zürich)

lower bound estimate $\hat{\alpha}_j$ for causal effect of gene j on Y
apply stability selection for lower bounds $\hat{\alpha}_j$'s

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

we 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 on Y relative to the wildtype (non-mutated plant)

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

Beware of over-interpretation!

so far, based on current data:

- we can not reliably infer the causal network
  despite theory... and because of theory
  stability selection yields rather unstable networks
- but we often(?) can do better ranking/prediction for intervention/causal effects
  than sophisticated but conceptually wrong regression methods

intervention/perturbation experiments can be very informative
in progress: combined estimation for observational and interventional data (Hauser & PB, in progress)

Conclusions

high-dimensional statistics: possibilities/limitations if

$$s_0 \sqrt{\log(p)/n} \text{ small/large}; \quad (\text{or } s_0 \log(p)/n \text{ small/large})$$

often subtle conditions on the "design" and "signal strength": they matter in practice!

- prediction is “relatively easy”
- variable selection or structure estimation is much harder
  top priority: efficiently guard against false positives
  (age-old problem in statistics!)
  stability selection, p-values based on sample splitting,...
- trick of convex relaxation (e.g. convex loss function and convex penalty) is beautiful and powerful
  - linear models, GLMs,...
  - not (easily) possible for many models
    e.g. mixture models, mixed-effects models,...,
    DAGs and causal inference
- particularly challenging but important for many scientific problems: causal inference
  - severe identifiability issues
  - nonconvex optimization
    but fairly efficient local computations on graphs

Thank you!

References:

- Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methodology, Theory and Applications. Springer (forthcoming)

- Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). Journal of the Royal Statistical Society: Series B, 72, 417-473
- Maathuis, M.H., Kalisch, M. and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. Annals of Statistics 37, 3133-3164.
- 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.

Convex relaxation?

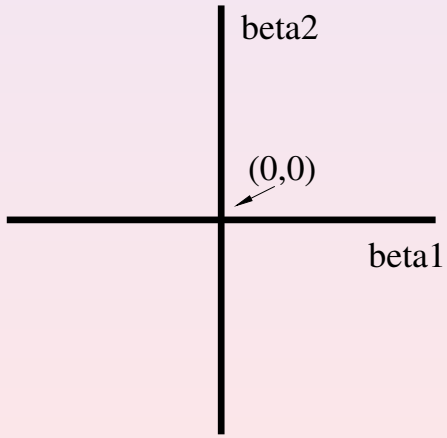
I don't know the answer... but

when parameterizing the (CP)DAG via structural equation models

$\rightsquigarrow$ corresponding parameter space is non-convex!

Example:

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



and hence: no straightforward way to do convex relaxation