Statistics for high-dimensional data: Toward more reliable results

Peter Bühlmann
ETH Zürich

August 2, 2009

High-dimensional data

Riboflavin production with Bacillus Subtilis

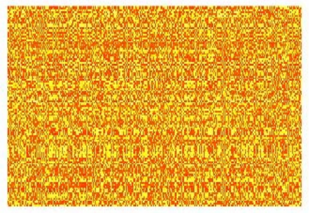
(in collaboration with DSM (Switzerland))

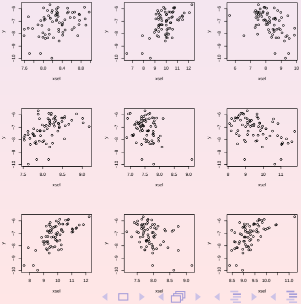
goal: improve riboflavin production rate of Bacillus Subtilis
using clever genetic engineering

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

Y versus 9 “reasonable” genes

gene expression data





general framework:

$$Z_1, \ldots, Z_n \text{ i.i.d. or stationary}$$
$$\dim(Z_i) \gg n$$

for example:
$Z_i = (X_i, Y_i),\ X_i \in \mathbb{R}^p, Y_i \in \mathbb{R}$: regression with $p \gg n$
$Z_i = (X_i, Y_i),\ X_i \in \mathbb{R}^p, Y_i \in \{0, 1\}$: classification with $p \gg n$

numerous applications:
biology, imaging, economy, environmental sciences, ...

High-dimensional linear models

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $Y = X\beta + \epsilon$

goals:

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

High-dimensional linear models

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i,\ i = 1, \ldots, n$$

$p \gg n$

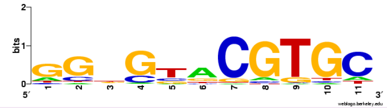
in short: $Y = X\beta + \epsilon$

goals:

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

Motif regression and variable selection

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?

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

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i$$

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

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

High-dimensional linear model

$$Y = X\beta + \epsilon, \quad p \text{ large; or } p \gg n$$

we need to regularize...

and there are many proposals

- Bayesian methods for regularization
- greedy algorithms: aka forward selection or boosting
- preliminary dimension reduction
- ...

e.g. 2'650'000 entries on Google Scholar for
“high dimensional linear model” ...

High-dimensional linear model

$$Y = X\beta + \epsilon, \quad p \text{ large; or } p \gg n$$

we need to regularize...
and there are many proposals

- Bayesian methods for regularization
- greedy algorithms: aka forward selection or boosting
- preliminary dimension reduction
- ...

e.g. 2’650’000 entries on Google Scholar for
“high dimensional linear model” ...

High-dimensional linear model

$$Y = X\beta + \epsilon, \quad p \text{ large; or } p \gg n$$

we need to regularize...
and there are many proposals

- Bayesian methods for regularization
- greedy algorithms: aka forward selection or boosting
- preliminary dimension reduction
- ...

e.g. 2’650’000 entries on Google Scholar for
“high dimensional linear model” ...

Penalty-based methods

if true β_{true} is sparse w.r.t.

- $\|\beta_{\text{true}}\|_0 =$ number of non-zero coefficients
  $\rightsquigarrow$ penalize with the $\|\cdot\|_0$-norm:
  $\text{argmin}_\beta(n^{-1}\|Y - X\beta\|^2 + \lambda\|\beta\|_0)$, e.g. AIC, BIC
  $\rightsquigarrow$ computationally infeasible if p is large (2^p sub-models)
- $\|\beta_{\text{true}}\|_1 = \sum_{j=1}^p |\beta_{\text{true},j}|$
  $\rightsquigarrow$ penalize with the $\|\cdot\|_1$-norm, i.e. Lasso:
  $\text{argmin}_\beta(n^{-1}\|Y - X\beta\|^2 + \lambda\|\beta\|_1)$
  $\rightsquigarrow$ convex optimization:
  computationally feasible and very fast for large p

The Lasso (Tibshirani, 1996)

Lasso for linear models (and analogously for GLM's)

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

$\rightsquigarrow$ 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

Lasso for prediction: $\hat{\beta}(\lambda)^T x_{new}$

Lasso for variable selection:

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

$$\text{for} \qquad \mathcal{S} = \{j; \beta_j \neq 0\}$$

no significance testing involved

it's convex optimization only!

Lasso for prediction: $\hat{\beta}(\lambda)^T x_{new}$

Lasso for variable selection:

$$\begin{aligned} & \hat{\mathcal{S}}(\lambda) = \{j;\ \hat{\beta}_j(\lambda) \neq 0\} \\ \text{for} \quad & \mathcal{S} = \{j; \beta_j \neq 0\} \end{aligned}$$

no significance testing involved

it's convex optimization only!

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 $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i,$

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

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

really?
do we trust our selection algorithm?
how stable are the findings?

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 $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i,$

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

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

really?

do we trust our selection algorithm?

how stable are the findings?

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

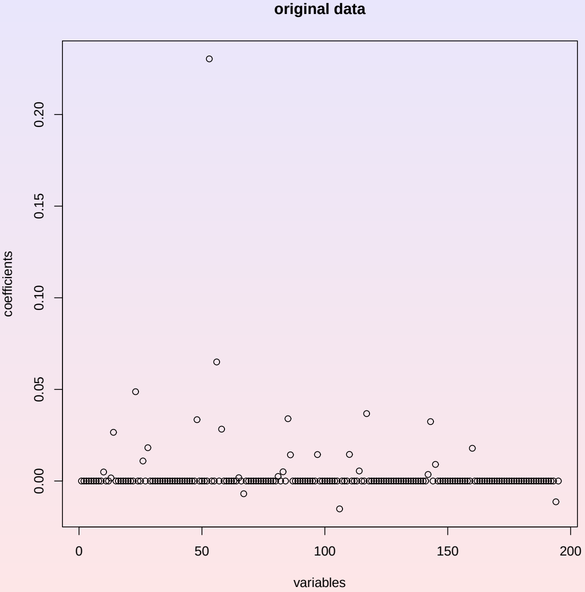
variable selection in linear model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i,$

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

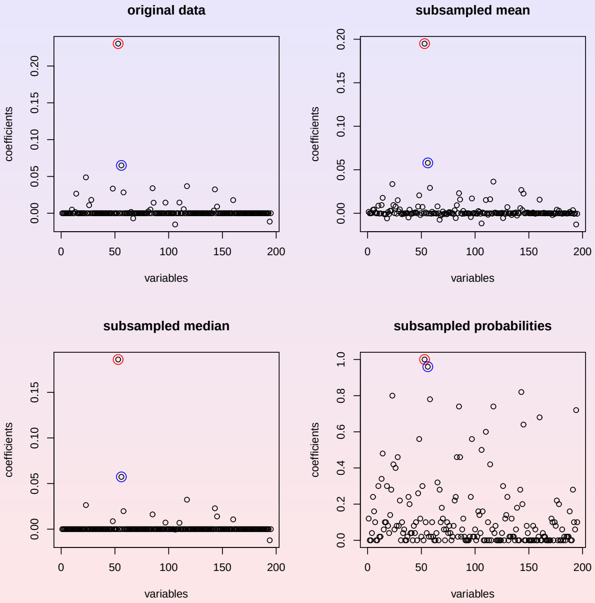
$\rightsquigarrow$ Lasso selects 26 covariates and $R^2 \approx 50\%$
i.e. 26 interesting candidate motifs
and hence report these findings to the biologists...

really?
do we trust our selection algorithm?
how stable are the findings?

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



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



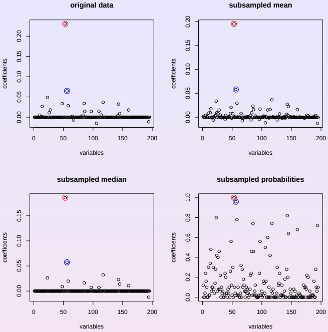
original data

subsampled mean

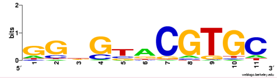
subsampled median

subsampled probabilities

$\leadsto$ 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, ...)
ongoing biological validation with Ricci lab (ETH Zurich)

Further outline of the talk

1. some methodology and theory (mainly) for Lasso
   $\rightsquigarrow$ understand whether the motif regression example is special? Or whether we expect such a behavior?
2. subsampling and stability
3. P-values, FWER and FDR control
4. and more...

High-dimensional linear models and the Lasso

$$Y_i = (\beta_0+) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $Y = X\beta + \epsilon$

goals:

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

High-dimensional linear models and the Lasso

$$Y_i = (\beta_0 +) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \epsilon_i, \; i = 1, \ldots, n$$

$p \gg n$

in short: $Y = X\beta + \epsilon$

goals:

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

Lasso for linear models (Tibshirani, 1996)

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

$\leadsto$ convex optimization problem

Why the Lasso/ℓ_1-penalization hype?

among other things (which will be discussed later)

$$\ell_1\text{-penalty approach approximates } \underbrace{\ell_0\text{-penalty problem}}_{\text{what we usually want}}$$

consider underdetermined system of linear equations:

$$A_{p\times p}\beta_{p\times 1} = b_{p\times 1}, \ \ \mathrm{rank}(A) = m < p$$

ℓ_0-penalty-problem: solve for β which is sparsest w.r.t. $\|\beta\|_0$
i.e. “Occam’s razor”

Donoho & Elad (2002), ...: if A is not too ill-conditioned (in the sense of linear dependence of sub-matrices)

$$\begin{aligned} & \text{sparsest solution } \beta \text{ w.r.t. } \|\cdot\|_0\text{-norm} \\ = \ & \underbrace{\text{sparsest solution } \beta \text{ w.r.t. } \|\cdot\|_1\text{-norm}}_{\text{amounts to a convex optimization}} \end{aligned}$$

Why the Lasso/ℓ_1-penalization hype?

among other things (which will be discussed later)

$$\ell_1\text{-penalty approach approximates} \quad \underbrace{\ell_0\text{-penalty problem}}_{\text{what we usually want}}$$

consider underdetermined system of linear equations:

$$A_{p\times p}\beta_{p\times 1} = b_{p\times 1}, \;\; \mathrm{rank}(A) = m < p$$

ℓ_0-penalty-problem: solve for β which is sparsest w.r.t. $\|\beta\|_0$
i.e. “Occam’s razor”

Donoho & Elad (2002), ...: if A is not too ill-conditioned (in the sense of linear dependence of sub-matrices)

$$\begin{aligned} & \text{sparsest solution } \beta \text{ w.r.t. } \|\cdot\|_0\text{-norm} \\ = \;\; & \underbrace{\text{sparsest solution } \beta \text{ w.r.t. } \|\cdot\|_1\text{-norm}}_{\text{amounts to a convex optimization}} \end{aligned}$$

Prediction (with the Lasso)

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!

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

theory: consistency (Greenshtein & Ritov, 2004) and optimality
Bunea, Tsybakov & Wegkamp (2006, 2007); van de Geer (2008);
Bickel, Ritov & Tsybakov (2009);...

Prediction (with the Lasso)

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!

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

theory: consistency (Greenshtein & Ritov, 2004) and optimality
Bunea, Tsybakov & Wegkamp (2006, 2007); van de Geer (2008);
Bickel, Ritov & Tsybakov (2009);...

Variable selection (with the Lasso)

we aim for increased understanding
but we cannot easily evaluate the selection method

$\rightsquigarrow$ it is highly desirable to
assess uncertainty, assign relevance or significance

motif regression
$n = 287$ samples, $p = 195$ variables (candidate motifs)

use Lasso as variable selection method:

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

Lasso selects 26 variables (motifs)
when choosing $\lambda = \hat{\lambda}_{CV}$ via cross-validation

and we have seen problems when trusting it blindly!
(also with other methods than Lasso)

Variable selection (with the Lasso)

we aim for increased understanding
but we cannot easily evaluate the selection method

$\rightsquigarrow$ it is highly desirable to
assess uncertainty, assign relevance or significance

motif regression
$n = 287$ samples, $p = 195$ variables (candidate motifs)

use Lasso as variable selection method:

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

Lasso selects 26 variables (motifs)
when choosing $\lambda = \hat{\lambda}_{CV}$ via cross-validation

and we have seen problems when trusting it blindly!
(also with other methods than Lasso)

theory for variable selection with Lasso: is it misleading?

Theorem (Meinshausen & PB, 2004 (publ: 2006))

- sufficient and necessary neighborhood stability condition on the design X; see also Zhao & Yu (2006)
- $p = p_n$ is growing with n
  - $p_n = O(n^{\alpha})$ for some $0 < \alpha < \infty$ (high-dimensionality)
  - $|\mathcal{S}_{true,n}| = O(n^{\kappa})$ for some $0 < \kappa < 1$ (sparsity)
  - the non-zero β_j's are outside the $n^{-1/2}$-range
  - $Y,\ X^{(j)}$'s Gaussian (not crucial)

Then: if $\lambda = \lambda_n \sim const.n^{-1/2-\delta/2}$ $(0 < \delta < 1/2)$,

$$\begin{aligned} \mathbb{P}[\hat{\mathcal{S}}(\lambda) = \mathcal{S}_{true}] &= 1 - O(\exp(-Cn^{1-\delta}))\ \ (n \to \infty) \\ &\approx 1 \text{ even for relatively small } n \end{aligned}$$

Problem 1:

Neighborhood stability condition is restrictive

sufficient and necessary for consistent model selection with Lasso

it fails to hold if design matrix exhibits
“strong linear dependence” (in terms of sub-matrices)

if it fails and because of necessity of the condition
$\Rightarrow$ Lasso is not consistent for selecting the relevant variables

neighborhood stability condition $\Leftrightarrow$ irrepresentable condition (Zhao & Yu, 2006)

$$n^{-1}X^TX \to \Sigma$$

active set $\mathcal{S} = \{j;\ \beta_j \neq 0\} = \{1,\ldots,p_{eff}\}$ consists of the first p_{eff} variables; partition

$$\Sigma = \left(\begin{array}{cc} \Sigma_{\mathcal{S},\mathcal{S}} & \Sigma_{\mathcal{S},\mathcal{S}^c} \\ \Sigma_{\mathcal{S}^c,\mathcal{S}} & \Sigma_{\mathcal{S}^c,\mathcal{S}^c} \end{array} \right)$$

irrep. condition : $|\Sigma_{\mathcal{S}^c,\mathcal{S}}\Sigma_{\mathcal{S},\mathcal{S}}^{-1}\mathrm{sign}(\beta_1,\ldots,\beta_{p_{eff}})| < 1$

a nice formulation, but:
no way to check this assumption in practice
(and the condition is restrictive)

Problem 2: Choice of λ

for prediction oracle solution

$$\lambda_{\text{opt}} = \text{argmin}_{\lambda} \mathbb{E}[(Y - \sum_{j=1}^{p} \hat{\beta}_j(\lambda) X^{(j)})^2]$$

$\mathbb{P}[\hat{\mathcal{S}}(\lambda_{\text{opt}}) = \mathcal{S}_{true}] < 1$ ($n \to \infty$) (or $= 0$ if $p_n \to \infty$ ($n \to \infty$))

asymptotically: prediction optimality yields too large models (Meinshausen & PB, 2004; related example by Leng et al., 2006)

“Problem 3”: small non-zero regression coefficients (i.e. high noise level)

we cannot reliably detect variables with small non-zero coefficients

but (under some conditions)
we can still detect the variables with large regression effects

If neighborhood stability condition fails to hold (problem 1)

under sparse eigenvalue assumptions for $n^{-1}X^TX$
"typically" much weaker assumptions than neighborhood stability

van de Geer (2008); Zhang & Huang (2008); Meinshausen & Yu (2000); Bickel, Ritov & Tsybakov (2009); van de Geer & PB (20??):
for suitable $\lambda = \lambda_n$ and with large probability

$$\|\hat{\beta} - \beta\|_1 = \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j| \le \underbrace{C}_{\text{depending on } X, \sigma^2} \sqrt{\log(p) p_{eff}/n}$$

hence: $\max_j |\hat{\beta}_j - \beta_j| \le \|\hat{\beta} - \beta\|_1 \le C\sqrt{\log(p) p_{eff}/n}$

and if $\min_j \{|\beta_j|;\ \beta_j \neq 0\} > C\sqrt{\log(p) p_{eff}/n}$

then $\hat{\beta}_j \neq 0$ for all $j \in \mathcal{S}$, i.e. $\hat{\mathcal{S}} \supseteq \mathcal{S}$

with large probability

$$\hat{\mathcal{S}} \supseteq \mathcal{S}$$

$$|\hat{\mathcal{S}}| \le O(min(n,p)) \underbrace{=}_{\text{if } p \gg n} O(n)$$

i.e. a huge dimensionality reduction in the original covariates!

furthermore: "typically", for prediction-optimal λ_{opt}

$$\hat{\mathcal{S}}(\lambda_{\mathrm{opt}}) \supseteq \mathcal{S}$$

$\leadsto$ Lasso as an
excellent screening procedure

i.e. true active set is contained in estimated active set from Lasso

with large probability

$$\hat{\mathcal{S}} \supseteq \mathcal{S}$$

$$|\hat{\mathcal{S}}| \le O(min(n,p)) \underbrace{=}_{\text{if } p \gg n} O(n)$$

i.e. a huge dimensionality reduction in the original covariates!

furthermore: “typically”, for prediction-optimal λ_{opt}

$$\hat{\mathcal{S}}(\lambda_{\mathrm{opt}}) \supseteq \mathcal{S}$$

$\leadsto$ Lasso as an
excellent screening procedure

i.e. true active set is contained in estimated active set from Lasso

with large probability

$$\hat{\mathcal{S}} \supseteq \mathcal{S}$$

$$|\hat{\mathcal{S}}| \le O(min(n,p)) \underbrace{=}_{\text{if } p \gg n} O(n)$$

i.e. a huge dimensionality reduction in the original covariates!

furthermore: “typically”, for prediction-optimal λ_{opt}

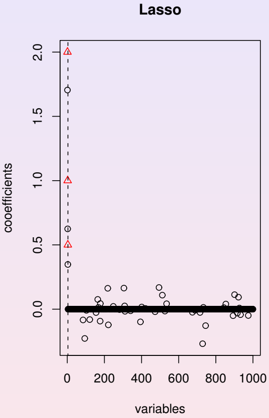
$$\hat{\mathcal{S}}(\lambda_{\mathrm{opt}}) \supseteq \mathcal{S}$$

$\leadsto$ Lasso as an
excellent screening procedure

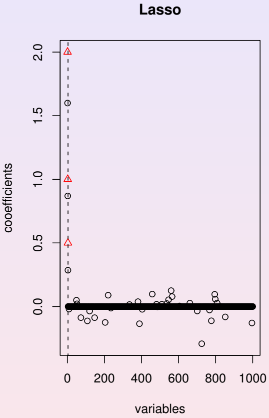
i.e. true active set is contained in estimated active set from Lasso

Lasso screening is $\underbrace{\text{easy to use,}}_{\text{prediction optimal tuning}}$ $\underbrace{\text{computationally efficient,}}_{O(np\min(n,p))}$ and statistically accurate

$p_{eff} = 3, \; p = 1'000, \; n = 50$; 2 independent realizations



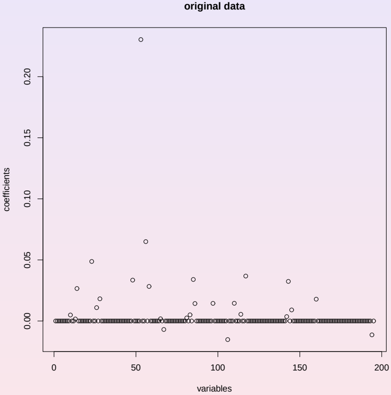
44 selected variables



36 selected variables

Motif regression ($p = 195,\ n = 287$)

26 selected covariates when using $\hat{\lambda}_{CV}$



presumably: the truly relevant variables are among the 26 selected covariates

First conclusion

Lasso is a good screening method: with high probability

$$\hat{\mathcal{S}} \supseteq \mathcal{S}$$

and two or multi-stage methods can be used
$\leadsto$ re-estimation on much smaller model with variables from $\hat{\mathcal{S}}$

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

but still:	often unstable selections
and	no measure of significance

First conclusion

Lasso is a good screening method: with high probability

$$\hat{\mathcal{S}} \supseteq \mathcal{S}$$

and two or multi-stage methods can be used
$\leadsto$ re-estimation on much smaller model with variables from $\hat{\mathcal{S}}$

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

but still: often unstable selections
and no measure of significance

similar “picture” for other screening procedures

- (gradient-type) boosting (Friedman, 2001; PB & Yu, 2003)
- Sure Independence Screening (SIS) (Fan & Lv, 2008)
- forward selection (orthogonal matching pursuit) (Tropp, 2004)

under suitable conditions on the design X: $\rightsquigarrow$ $\hat{S} \supseteq S$
(and $\hat{S} = S$ is much harder in high-dimensional case)

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

but still: often unstable selections
and no measure of significance

Stability Selection (Meinshausen & PB, 2008)

using subsampling (or bootstrapping)

another motif regression example
$Y_i \in \mathbb{R}$: univariate response measuring expression of gene i
$X_i = (X_i^{(1)}, \ldots, X_i^{(p)}) \in \mathbb{R}^p$:
$X_i^{(j)}$ = abundance score of candidate motif j in DNA segment around gene i (using sequence data and computational biology algorithms, e.g. MDSCAN)

linear regression model with $n = 1'200,\ p = 660$

$$Y = X\beta + \varepsilon$$

and the goal is selection of the relevant variables

Stability Selection (Meinshausen & PB, 2008)

using subsampling (or bootstrapping)

another motif regression example
$Y_i \in \mathbb{R}$: univariate response measuring expression of gene *i*
$X_i = (X_i^{(1)}, \ldots, X_i^{(p)}) \in \mathbb{R}^p$:
$X_i^{(j)}$ = abundance score of candidate motif *j* in DNA segment around gene *i* (using sequence data and computational biology algorithms, e.g. MDSCAN)

linear regression model with $n = 1'200,\ p = 660$

$$Y = X\beta + \varepsilon$$

and the goal is selection of the relevant variables

Using the Lasso...

the 9 most promising motifs, in descending order of $|\hat{\beta}_j(\hat{\lambda}_{CV})|$

motif j	41	29	635	19	34	603	618	596	30
$\lvert\hat{\beta}_j\rvert$	1.42	1.27	0.81	0.61	0.57	0.49	0.33	0.3	0.3

in total, 20 motifs have a non-zero regression coefficient

report motifs in this order?
how many? all 20?

Using the Lasso...

the 9 most promising motifs, in descending order of $|\hat{\beta}_j(\hat{\lambda}_{CV})|$

motif j	41	29	635	19	34	603	618	596	30
$\|\hat{\beta}_j\|$	1.42	1.27	0.81	0.61	0.57	0.49	0.33	0.3	0.3

in total, 20 motifs have a non-zero regression coefficient

report motifs in this order?
how many? all 20?

“It could have been different” (Tukey)

$\rightsquigarrow$ subsampling with sample size $n = \lfloor n/2 \rfloor$
“selection probability” for each motif: $\Pi_j = P^*(\hat{\beta}_j^* \neq 0)$

motif j	41	29	635	19	34	603	618	596	30
$\lvert\hat{\beta}_j\rvert$	1.42	1.27	0.81	0.61	0.57	0.49	0.33	0.3	0.3
$\hat{\Pi}_j$	100%	100%	100%	74%	98%	32%	81%	80%	97%

rather report motif 603 or 30 ?

motif j	41	29	635	19	34	603	618	596	30
$\lvert\hat{\beta}_j\rvert$	1.42	1.27	0.81	0.61	0.57	0.49	0.33	0.3	0.3
Π_j	100%	100%	100%	74%	98%	32%	81%	80%	97%

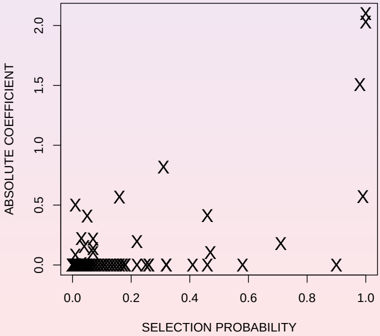
(and not very different results when using a two-stage procedure,
as e.g. the Adaptive Lasso)

“Semi-”Synthetic data

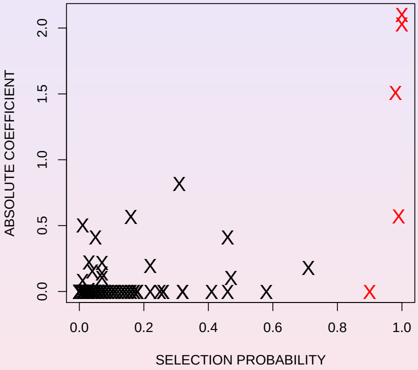
select 5 motifs $m_1, \ldots, m_5$ at random among all $p = 660$ motifs and set

$$Y = \sum_{j=1}^{5} \underbrace{X^{(m_j)}}_{\text{real}} \underbrace{\beta_{m_j}}_{\text{synthetic}} + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (n = 1'200,\ p = 660)$$

σ^2, β chosen to achieve very low SNR=0.1



now we know the “ground-truth”



red: motifs with $\beta_j \neq 0$ black: motifs with $\beta_k = 0$

Stability selection

consider (first) linear model setting

$$Y_i = (\beta_0) + \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \; i = 1, \ldots, n \; (\ll p)$$

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

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 $I^* \subseteq \{1, \dots, n\}$, $|I^*| = \lfloor n/2 \rfloor$
- run the selection algorithm $\hat{S}^{\lambda}(I^*)$ on I^*
- do these steps many times and compute the relative selection frequencies

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

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

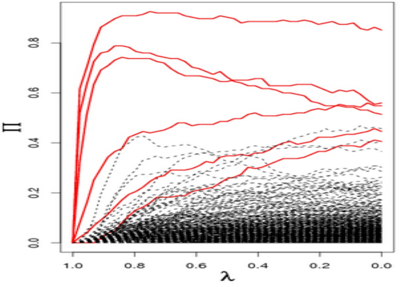
choice of $\pi_{\text{thr}} \rightsquigarrow$ see later

note: some vague relations to
the “problem of regions” (Efron & Tibshirani, 1998)

if we consider many regularization parameters:

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

Λ can be discrete, a singleton or continuous

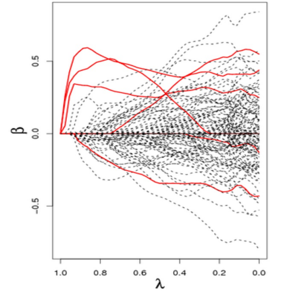


$$\hat{S}^{\text{stable}} = \{j;\ \max_{\lambda \in \Lambda} \hat{\Pi}_j^{\lambda} \geq \pi_{\text{thr}}\}$$

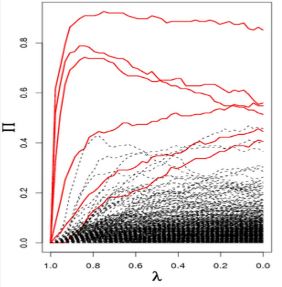
see also Bach (2009) for a related proposal

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

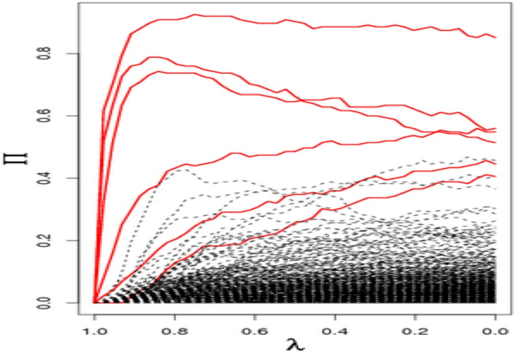


Stability selection

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 $V = |S^C \cap \hat{S}^{\text{stable}}| =$ number of false positives

Theorem (Meinshausen & PB, 2008)
main assumption: exchangeability condition
in addition: $\hat{S}$ has to be better than "random guessing"
Then:

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

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

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 $\{I_{\{j \in \hat{S}^{\lambda}\}};\ j \in S^{C}\}$ is exchangeable
note: only some requirement for noise variables

for specific problems, we can prove error control under weaker assumptions...

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)
  $\leadsto$ variable selection, graphical modeling, clustering, ...

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

for specific problems, we can prove error control under weaker assumptions...

Some numerical experiments

Variable selection in linear models using Lasso

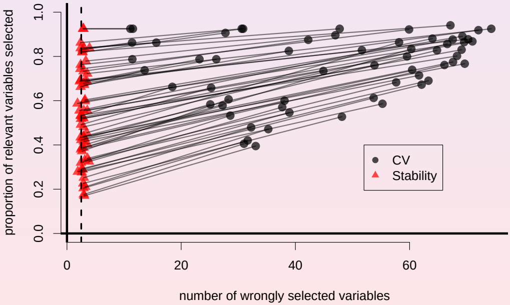
a range of scenarios:

$p = 660$ with design from a real data set about motif regression

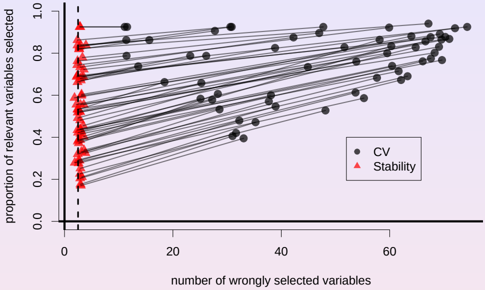
$n \in \{450, 750\}$, sparsity $p_{eff} \in \{4, 8, \ldots, 40\}$ (using artificial β)

signal to noise ratio $\in \{0.25, 1, 4\}$

control for $E[V] \leq 2.5$



control for $E[V] \leq 2.5$



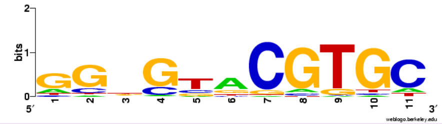
stability selection yields:

- 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

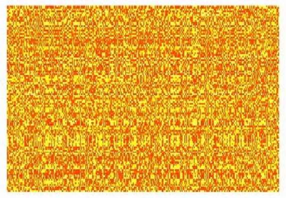


Graphical modeling using GLasso

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

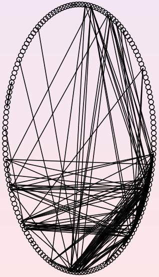
infer conditional independence graph using ℓ_1-penalization
i.e. infer zeroes of Σ^{-1} from $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}_p(0, \Sigma)$

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



gene expr. data

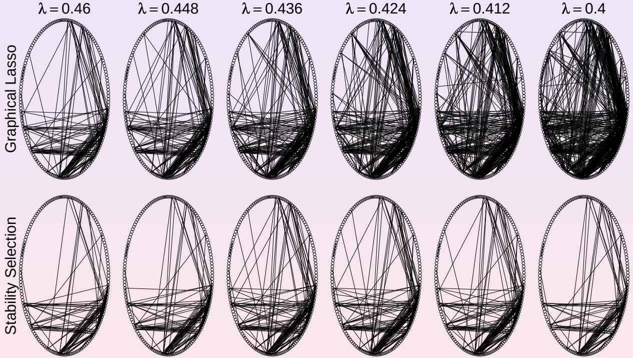
$$\overset{?}{\Longrightarrow}$$



zero-pattern of Σ^{-1}

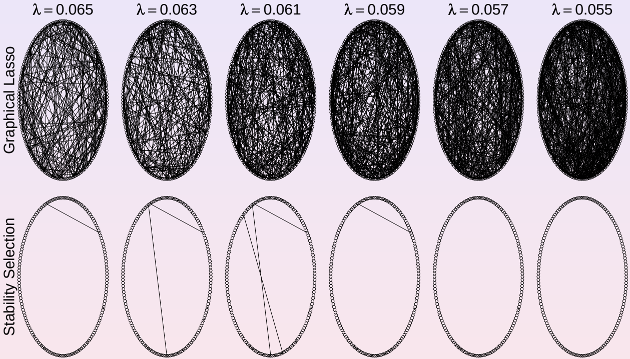
sub-problem of riboflavin production with bacillus subtilis
$p = 160,\ n = 115$
stability selection with $E[V] \leq 5$

varying the regularization parameter λ in ℓ_1-penalization



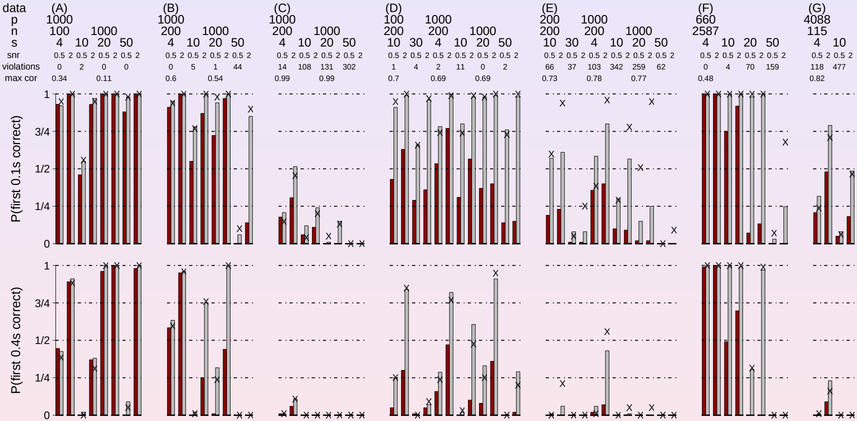
with stability selection: choice of initial λ-tuning parameter does not matter much (as proved by our theory)
just need to fix the finite-sample control

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)

probabilities: selected variables include no noise variable and at least 10% or 40% of the correct var.



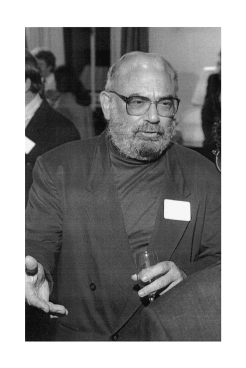
red: Lasso

grey: stability selection with Lasso

grey cross ×: additional randomization on covariates

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

Leo Breiman



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

P-values (Meinshausen, Meier & PB, 2008)

for more specific problems assuming weaker assumptions (no exchangeability condition)

for simplicity: focus on P-values for regression coefficients
$H_0^{(j)}: \ \beta_j = 0$

$$Y_i = (\beta_0 +) \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ \ (i = 1, \ldots, n), \ \ p \gg n$$

A first idea: sample splitting with sub-samples of sizes $\lfloor n/2 \rfloor$

related to subsampling with sub-sample size $\lfloor n/2 \rfloor$

- select variables on first half of the sample $\leadsto$ $\hat{\mathcal{S}}$
- compute OLS for variables in $\hat{\mathcal{S}}$ on second half of the sample
  $\leadsto$ P-values $P^{(j)}$ based on Gaussian linear model

$$\begin{aligned} &\text{if } j \in \hat{\mathcal{S}}: \quad P^{(j)} \text{ from } t\text{-statistics} \\ &\text{if } j \notin \hat{\mathcal{S}}: \quad P^{(j)} = 1 \ \ (\text{i.e. if } \hat{\beta}^{(j)} = 0) \end{aligned}$$

Bonferroni-corrected P-values:

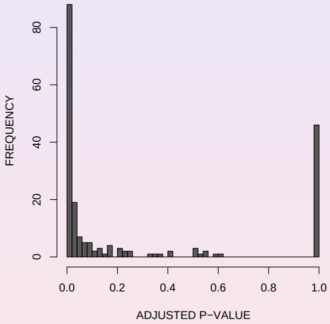
$$P_{\mathrm{corr}}^{(j)} = \min(P^{(j)} \cdot |\hat{\mathcal{S}}|, 1)$$

$\leadsto$ (conserv.) familywise error control with $P_{\mathrm{corr}}^{(j)}$ $(j = 1, \ldots, p)$

(Wasserman & Roeder, 2008)

this is a “P-value lottery”
motif regression example: $p = 195, \; n = 287$

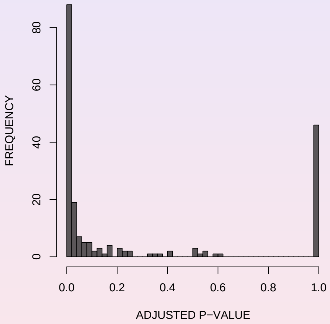
adjusted P-values for same important variable
over different random sample-splits



in addition: bad “efficiency”
$\leadsto$ improve by aggregating over many sample-splits

this is a "P-value lottery"
motif regression example: $p = 195,\ n = 287$

adjusted P-values for same important variable
over different random sample-splits



in addition: bad "efficiency"
$\leadsto$ improve by aggregating over many sample-splits

Multi sample-split P-values and aggregation

run the sample-splitting procedure B times:

$$\text{P-values: } P^{(j)}_{\mathrm{corr},1}, \ldots, P^{(j)}_{\mathrm{corr},B}$$

(assuming a Gaussian linear model with fixed design)

goal:
aggregation of $P^{(j)}_{\mathrm{corr},1}, \ldots, P^{(j)}_{\mathrm{corr},B}$ to a single P-value $P^{(j)}_{\mathrm{final}}$
problem: dependence among $P^{(j)}_{\mathrm{corr},1}, \ldots, P^{(j)}_{\mathrm{corr},B}$

define

$$Q^{(j)}(\gamma) = \underbrace{q_\gamma}_{\text{emp. } \gamma\text{-quantile fct.}} (P^{(j)}_{\mathrm{corr},b}/\gamma;\ b = 1, \ldots B)$$

e.g: $\gamma = 1/2$, aggregation with the median
$\leadsto$ (conserv.) familywise error control for any fixed value of γ

what is the best γ?   it really matters
$\leadsto$ can "search" for it an correct with an additional factor

“adaptively” aggregated P-value:

$$P^{(j)}_{\text{final}} = (1 - \log(\gamma_{\min})) \cdot \inf_{\gamma \in (\gamma_{\min}, 1)} Q^{(j)}(\gamma)$$

$$Q^{(j)}(\gamma) = q_\gamma(P^{(j)}_{\text{corr},b}/\gamma;\ b = 1, \ldots B)$$

$$\rightsquigarrow \text{reject } H_0^{(j)} :\ \beta_j = 0 \iff P^{(j)}_{\text{final}} \le \alpha$$

$P^{(j)}_{\text{final}}$ equals roughly a raw P-value based on sample size $\lfloor n/2 \rfloor$, multiplied by

a factor $\approx$ $(5 - 10) \cdot |\hat{\mathcal{S}}|$

(which is to be compared with p)

for familywise error rate (FWER) =
$\mathbb{P}$[at least one false positive selection]

Theorem (Meinshausen, Meier & PB, 2008)
assumptions: Gaussian linear model (with fixed design) and

- $\lim_{n\to\infty}\mathbb{P}[\hat{\mathcal{S}} \supseteq \mathcal{S}] = 1$ screening property
- $|\hat{\mathcal{S}}| < \lfloor n/2 \rfloor$ sparsity property

Then:

$P_{\mathrm{final}}^{(j)}$'s yield asymptotic FWER control

$$\limsup_{n\to\infty}\mathbb{P}(\min_{j\in\mathcal{S}^c} P_{\mathrm{final}}^{(j)} \le \alpha) \le \alpha$$

i.e. (conservative) familywise error control

False discovery rate (FDR) (Benjamini & Hochberg, 1995)

based on ordered $P_{\mathrm{final}}^{(j)}$'s from before
$\leadsto$ control of FDR for multiple testing of regression coefficients with $p \gg n$
(Meinshausen, Meier & PB, 2008)

assumptions for selector $\hat{\mathcal{S}}$:
are satisfied for

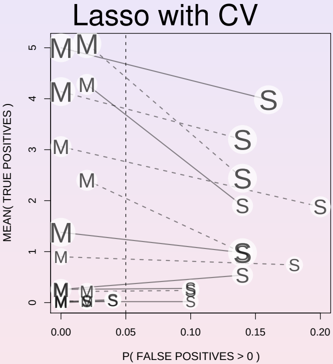
- Lasso
  - assuming restricted eigenvalue conditions on the design (Bickel, Ritov & Tsybakov, 2009)
    or even weaker conditions (van de Geer & PB, 20??)
  - assuming sparsity of true regression coefficients

- L_2Boosting, Sure Independence Screening, PC-algorithm,...
  - assuming reasonable conditions on the design
  - assuming sparsity of true regression coefficients

no exchangeability condition is required here

Simulations for FWER: $p = 1000,\ n = 100$

design matrix from multivariate Gaussian with $\Sigma_{j,k} = 0.5^{|j-k|}$
signal to noise ratio $\in \{0.25, 1, 4, 16\}$

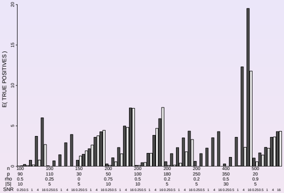


multi sample-split method (M) has

- much better error control than single sample-split method
- (slightly) more power than single split method

Simulations for FDR

for a whole variety of settings



multi sample-split FDR control holds up well (conservative)

if $p < n$: even a bit better than standard FDR if

- p close to n
- strong dependence between the tests

Motif regression

$p = 195, \; n = 287$
for $\alpha = 0.05$, only one variable/motif $\tilde{j}$ remains

$$P^{(\tilde{j})}_{\text{final}} = 0.0059 \;\; (= 0.59\%)$$

and also with FDR control: only this one variable

in this application:
we are rather concerned about false positive findings
$\leadsto$ (conservative) P-values are very useful

Motif regression

$p = 195, \; n = 287$
for $\alpha = 0.05$, only one variable/motif $\tilde{j}$ remains

$$P^{(\tilde{j})}_{\mathrm{final}} = 0.0059 \;\; (= 0.59\%)$$

and also with FDR control: only this one variable

in this application:
we are rather concerned about false positive findings
$\leadsto$ (conservative) P-values are very useful

Where are we?

- sub-sampling for stability selection
- sample-splitting for P-values

are very easy to implement and rather generic
and computationally feasible since convex optimization is fast

(Bayesian approaches offer a “natural alternative” to address the issue of stability and significance)

Convex optimization for sparse problems is fast

can easily deal with $p \approx 10^6$ ("p in the Mega's")

using block gradient descent methods
based on developments and theory of Tseng et al., 2000–2008

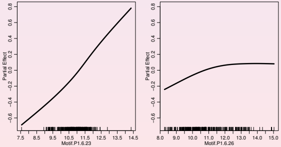
logistic regression case and "Group Lasso"
$p = 10^6,\ p_{eff} = 40$ non-zero parameters, $n = 100$
for 10 different λ-values
CPU using `grplasso` in `R`: 203.16 seconds $\approx$ 3.5 minutes
Meier, van de Geer & PB (2008)

even faster with `glmnet` in R for a plain Lasso problem
Friedman, Hastie & Tibshirani (2008)

I haven't talked about...

- Generalized linear models $(\surd)$
  very similar methodology and theory as for linear models
- Group structure and Group Lasso (Yuan & Lin, 2006) $(\surd)$
  for achieving sparsity in pre-defined groups
- Additive modeling $(\surd)$ (but no simple P-values)
  we should penalize for sparsity and smoothness
  (Ravikumar, Liu, Lafferty & Wasserman, 2007;
  Meier, van de Geer & PB, 2008)

motif additive regression ($p = 195,\ n = 287$)



as before: two stable motifs

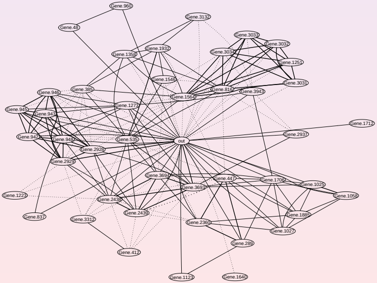
Intervention effects and Causality

back to first example:

Riboflavin production with Bacillus Subtilis

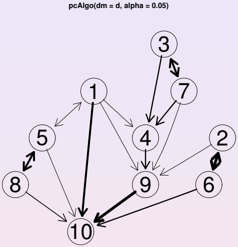
what is the effect of knocking-down a single gene on the riboflavin production rate?

$\rightsquigarrow$ this is a question of intervention type ($\neq$ association)
i.e. of causal type



program to be carried out (Maathuis, Kalisch & PB, 2008)

1. infer graph from data
   (can only infer equivalence class of graphs)



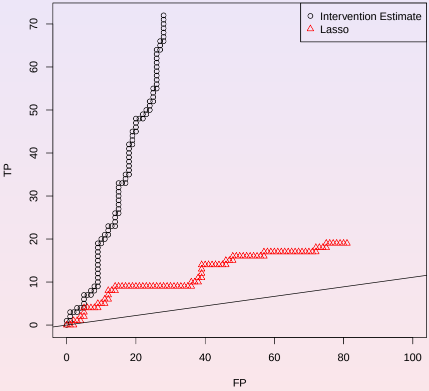
2. run fairly low-dimensional regressions using the structure of the equivalence class of graphs
3. $\rightsquigarrow$ estimates of bounds of causal effects

stability selection is tremendously useful here as well!

single strain interventions in yeast

$n = 63, \; p = 5361$ observational (non-interventional) data

231 intervention experiments for validation



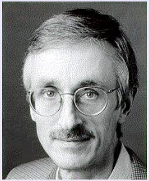
better prediction of intervention/causal effects than Lasso regression for association effects (wrong concept)

Conclusions

in particular for structure estimation:
high-dimensional inference is often unreliable

subsampling, bootstrapping and sample-splitting can be used for stable selection and for assigning error rates

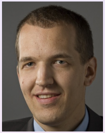
Thank you!


Hans R. Künsch

Peter Bickel

Bin Yu

Nicolai Meinshausen


Lukas Meier

Markus Kalisch

Marloes Maathuis

Sara van de Geer