

How good is the de-biased Lasso?

asymptotic efficiency:

for the de-biased Lasso to “work” we require

- ▶ sparsity: $s_0 = o(\sqrt{n}/\log(p))$
this cannot be beaten in a minimax sense
- ▶ compatibility condition for X

for optimality in terms of the lowest possible asymptotic variance achieving the “Cramer-Rao” lower bound:

- ▶ require **in addition** that $X^{(j)}$ versus $X^{(-j)}$ is sparse:
 $s_j \ll n/\log(p)$

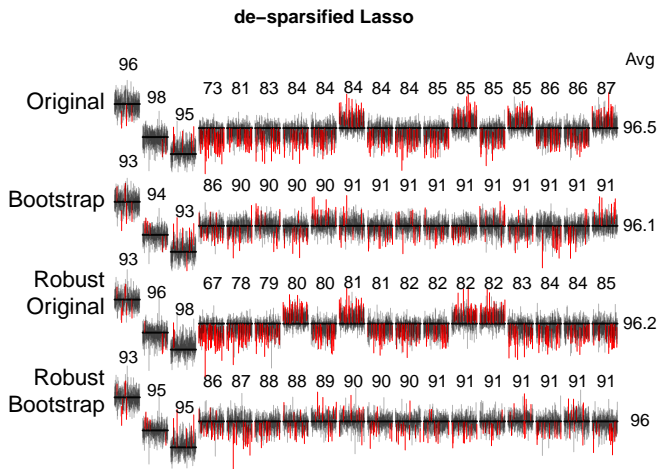
then... skipping details, the de-biased Lasso achieves (see Theorem 10.2):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0) \implies \mathcal{N}\left(0, \underbrace{\sigma^2 \Theta_{jj}}_{\text{Cramer-Rao lower bound}}\right)$$

$\Theta = \Sigma_X^{-1} = \text{Cov}(X)^{-1} \rightsquigarrow$ as for OLS in low dimensions!

Empirical results

R-software hdi




black: confidence interval covered the true coefficient
red: confidence interval failed to cover

Stability Selection (Ch. 10 in Bühlmann and van de Geer (2011))

Journal of the
Royal Statistical Society

J. R. Statist. Soc. B (2010)
72, Part 4, pp. 417–473

SERIES B
Statistical
Methodology 

Stability selection

Nicolai Meinshausen

University of Oxford, UK

and Peter Bühlmann

Eidgenössische Technische Hochschule Zürich, Switzerland

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, February 3rd, 2010, Professor D. M. Titterton in the Chair]

has been developed before one knew about the
de-biased/de-sparsified Lasso

even with new tools such as the de-biased/de-sparsified Lasso:
estimation of discrete structures (“relevant” variables in a
generalized linear model; edges in a graphical model) is
notoriously difficult

e.g. choice of tuning parameters...?

The generic setup

i.i.d. data Z_1, \dots, Z_n

main example: $Z_i = (X_i, Y_i)$ from regression or classification

\hat{S}_λ is a “feature selection” method/algorithm among $\{1, \dots, p\}$ features

can we assign “relevance” to the selected features in \hat{S}_λ ?

a “natural” approach: resampling!

here: use subsampling:

- ▶ I^* random sub-sample of size $\lfloor n/2 \rfloor$ of $\{1, \dots, n\}$
- ▶ compute $\hat{S}_\lambda(I^*)$
- ▶ repeat B times to obtain $\hat{S}_\lambda(I^{*1}), \dots, \hat{S}_\lambda(I^{*B})$
- ▶ consider the “overlap” among $\hat{S}_\lambda(I^{*1}), \dots, \hat{S}_\lambda(I^{*B})$

regarding the latter, for example:

$$\hat{\Pi}_K(\lambda) = \mathbb{P}^*[K \subseteq \hat{S}_\lambda(I^*)] \approx B^{-1} \sum_{b=1}^B I(K \subseteq \hat{S}_\lambda(I^{*b}))$$

e.g. $\hat{\Pi}_j(\lambda)$ ($j \in \{1, \dots, p\}$)

the probability \mathbb{P}^* is with respect to subsampling: a sum over $\binom{n}{m}$ terms, $m = \lfloor n/2 \rfloor$, i.e., all possible subsampling combinations

↪ it is approximated by B (≈ 100) times random subsampling

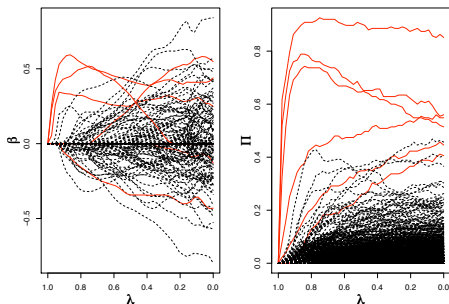
The stability regularization path

Riboflavin data: $n = 115$, $p = 4088$

Y : log-production rat of riboflavin by bacillus subtilis

X : gene expressions of bacillus subtilis

all X -variables permuted except 6 “a-priori relevant” genes



left: Lasso regularization path (red: the 6 non-permuted “relevant” genes)

right: Stability path with $\hat{\Pi}_j$ on y -axis (red: the 6 non-permuted “relevant” variables stick out much more clearly from the noise covariates)

What is a good truncation value (for $\hat{\Pi}$)?

aim: choose π_{thr} such that

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

has not too many false positives

Λ can be a singleton or a range of values

as a measure for type I error control (against false positives):

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

where S_0 is the set of the true relevant features, e.g.:

- active variables in regression
- true edges in a graphical model

“the miracle”:

a simple formula connecting π_{thr} with $\mathbb{E}[V]$

consider a setting with p possible features

$\hat{S}(\lambda)$ is a feature selection algorithm

$$\hat{S}_\Lambda = \cup_{\lambda \in \Lambda} \hat{S}(\lambda)$$

$$q_\Lambda = \mathbb{E}[|\hat{S}_\Lambda(\underbrace{\quad}_I \quad)|]$$

random subsample

Theorem 10.1

Assume:

- ▶ exchangeability condition:
 $\{1(j \in \hat{S}(\lambda)), j \in S_0^c\}$ is exchangeable for all $\lambda \in \Lambda$
- ▶ \hat{S} is not worse than random guessing

$$\frac{\mathbb{E}|S_0 \cap \hat{S}_\Lambda|}{\mathbb{E}(|S_0^c \cap \hat{S}_\Lambda|)} \geq \frac{|S_0|}{|S_0^c|}.$$

Then, for $\pi_{\text{thr}} \in (1/2, 1)$:

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q_\Lambda^2}{p}.$$

suppose we know q_Λ (see later)

strategy: specify $\mathbb{E}[V] = v_0$ (e.g. = 5)

\leadsto for $\pi_{\text{thr}} := \frac{1}{2} + \frac{q_\Lambda^2}{2pv_0}$: $\mathbb{E}[V] \leq v_0$

example: regression model with $p = 1000$ variables

\hat{S}_λ = the top 10 variables from Lasso (e.g. the different λ from Lasso by CV and choose the top 10 variables with the largest absolute values of the corresponding estimated coefficients; if less than 10 variables are selected, take the selected variables) the value λ corresponds to the “top 10”; Λ is a singleton

we then know that $q_\Lambda = \mathbb{E}[|\hat{S}_\lambda(I)|] \leq 10$

For $\mathbb{E}[V] = v_0 := 5$ we then obtain

$$\pi_{\text{thr}} = \frac{1}{2} + \frac{q_\Lambda^2}{2pv_0} = 0.5 + \frac{10^2}{2 * 1000 * 5} = 0.51$$

there is room to play around
recommendation: take $|\hat{S}(\lambda)|$ rather large and stability selection
will reduce again to reasonable size

when taking the “top 30”, the threshold becomes

$$\pi_{\text{thr}} = \frac{1}{2} + \frac{q_{\lambda}^2}{2p\nu_0} = 0.5 + \frac{30^2}{2 * 1000 * 5} = 0.59$$

adding noise...

can always add (e.g. independent $\mathcal{N}(0, 1)$) noise covariates
enlarged dimension ρ_{enlarged}

error control becomes better (for the same threshold)

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{\text{thr}} - 1} \frac{q_{\lambda}^2}{\rho_{\text{enlarged}}}$$

this sometimes helps indeed in practice – at the cost of loss in power

The assumptions for mathematical guarantees

not worse than random guessing

$$\frac{\mathbb{E}(|S_0 \cap \hat{S}_\lambda|)}{\mathbb{E}(|S_0^c \cap \hat{S}_\lambda|)} \geq \frac{|S_0|}{|S_0^c|}$$

perhaps hard to check but very reasonable...

for Lasso in linear models it holds assuming the variable screening property

asymptotically: if beta-min and compatibility condition hold

exchangeability condition $\{1(j \in \hat{S}(\lambda)), j \in S_0^c\}$ is
exchangeable for all $\lambda \in \Lambda$

a restrictive assumption
but the theorem is very general, for any algorithm \hat{S}

a very special case where exchangeability condition holds:
random equi-correlation design linear model

$$Y = X\beta^0 + \varepsilon, \text{Cov}(X)_{i,j} \equiv \rho (i \neq j), \text{Var}(X_j) \equiv 1 \forall j$$

distributions of $(Y, X^{(S_0)}, \{X^{(j)}; j \in S_0^c\})$ and of $(Y, X^{(S_0)}, \{X^{(\pi(j))}; j \in S_0^c\})$ are the same for any permutation $\pi : S_0^c \rightarrow S_0^c$

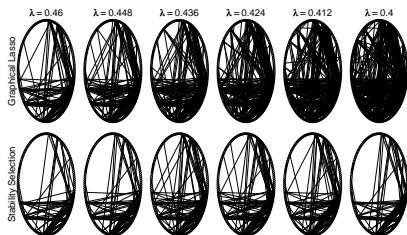
- ▶ distribution of $X^{(S_0)}, \{X^{(\pi(j))}; j \in S_0^c\}$ is the same for all π (because of equi-correlation)
- ▶ distribution of $Y|X^{(S_0)}, \{X^{(\pi(j))}; j \in S_0^c\}$ is the same for all π (because it depends only on $X^{(S_0)}$)
- ▶ therefore: distribution of $Y, X^{(S_0)}, \{X^{(\pi(j))}; j \in S_0^c\}$ is the same for all π
and hence exchangeability condition holds for any (measurable) function $\hat{S}(\lambda)$

An illustration for graphical modeling

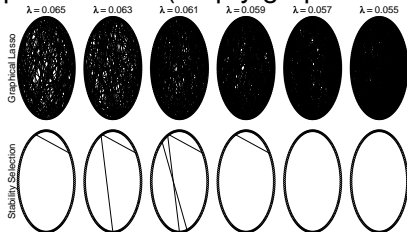
$p = 160$ gene expressions, $n = 115$

GLasso estimator, selecting among the $\binom{p}{2} = 12'720$ features

stability selection with $\mathbb{E}[V] \leq v_0 = 30$



with permutation (empty graph is correct)



Stability Selection is extremely easy to use
and super-generic

the sufficient assumptions (far from necessary) for
mathematical guarantees are restrictive
but the method seems to work very well in practice

P-values based on multi sample splitting

(Ch. 11 in Bühlmann and van de Geer (2011))

Stability Selection

- ▶ uses subsampling many times – a good thing!
- ▶ provides control of the expected number of false positives rather than e.g. the familywise error rate \leadsto we will “address” this with multi sample splitting and aggregation of P-values

familywise error rate (FWER):

$$\text{FWER} = \mathbb{P}[V > 0], \quad V \text{ number of false positives}$$

Fixed design linear model

$$Y = X\beta^0 + \varepsilon$$

instead of de-biased/de-sparsified method, consider the “older” technique (which is not statistically optimal but more generic and more in the spirit of stability selection)

split the sample into two parts I_1 and I_2 of equal size $\lfloor n/2 \rfloor$

- ▶ use (e.g.) Lasso to select variables based on I_1 : $\hat{S}(I_1)$
- ▶ perform low-dimensional statistical inference on I_2 based on data $(x_{I_2}^{(\hat{S}(I_1))}, Y_{I_2})$;

for example using the t -test for single coefficients β_j^0

(if $j \notin \hat{S}(I_1)$, assign the p-value 1 to the hypothesis

$$H_{0,j} : \beta_j^0 = 0,$$

due to independence of I_1 and I_2 , this is a “valid” strategy (see later)

validity of the (single) data splitting procedure
consider testing $H_{0,j} : \beta_j^0 = 0$ versus $H_{A,j} : \beta_j^0 \neq 0$
assume Gaussian errors for the fixed design linear model :
thus, use the t -test on the second half of the sample I_2 to get a
p-value

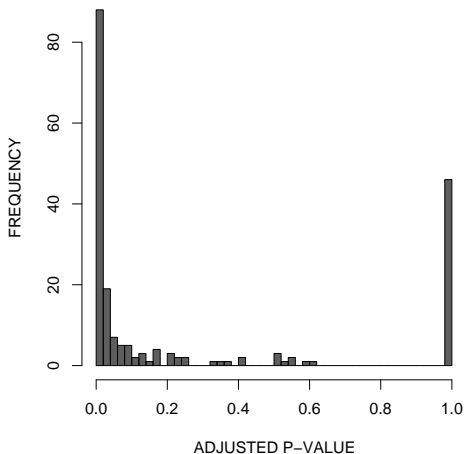
$$P_{\text{raw},j} \text{ from } t\text{-test based on } X_{I_2, \hat{S}(I_1)}, Y_{I_2}$$

$P_{\text{raw},j}$ is a valid p-value (controlling type I error) for testing $H_{0,j}$
if $\hat{S}(I_1) \supseteq S_0$ (i.e., the screening property holds)

if the screening property does not hold: $P_{\text{raw},j}$ is still valid for
 $H_{0,j}(M) : \beta_j(M) = 0$ where $M = \hat{S}(I_1)$ is a selected sub-model
and $\beta(M) = (X_M^T X_M)^{-1} X_M^T \mathbb{E}[Y]$

a p-value lottery depending on **the random split** of the data

motif regression $n = 287$, $p = 195$



~> should aggregate/average over multiple splits!

Multiple testing and aggregation of p-values

the issue of multiple testing:

$$\tilde{P}_j = \begin{cases} P_{\text{raw},j} \text{ based on } \mathbf{Y}_{I_2}, \mathbf{X}_{I_2, \hat{S}(I_1)} & , \text{ if } j \in \hat{S}(I_1), \\ 1 & , \text{ if } j \notin \hat{S}(I_1) \end{cases}$$

thus, we can have at most $|\hat{S}(I_1)|$ false positives

\leadsto can correct with Bonferroni with factor $|\hat{S}(I_1)|$ (instead of factor p) to control the familywise error rate

$$\tilde{P}_{\text{corr},j} = \min(\tilde{P}_j \cdot |\hat{S}(I_1)|, 1) \quad (j = 1, \dots, p)$$

decision rule: reject $H_{0,j}$ if and only if $\tilde{P}_{\text{corr},j} \leq \alpha$

$\leadsto \text{FWER} \leq \alpha$

the issue with P-value aggregation:

if we run sample splitting B times, we obtain P-values

$$\tilde{p}_{\text{corr},j}^{[1]}, \dots, \tilde{p}_{\text{corr},j}^{[B]}$$

how to aggregate these dependent p-values to a single one?

for $\gamma \in (0, 1)$ define

$$Q_j(\gamma) = \min \left\{ q_\gamma(\{\tilde{p}_{\text{corr},j}^{[b]}/\gamma; b = 1, \dots, B\}), 1 \right\},$$

where $q_\gamma(\cdot)$ is the (empirical) γ -quantile function

Proposition 11.1 (Bühlmann and van de Geer, 2011)

For any $\gamma \in (0, 1)$, $Q_j(\gamma)$ are P-values which control the FWER

example: $\gamma = 1/2$

aggregate the p-values with the sample median and multiply by the factor 2

avoid choosing γ :

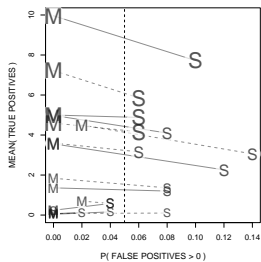
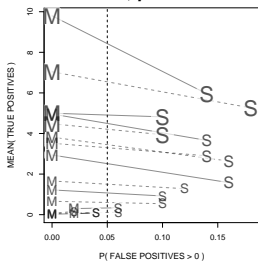
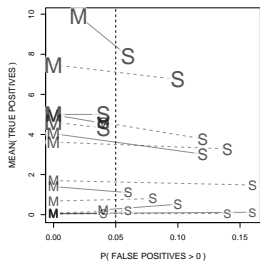
$$P_j = \min \left\{ \underbrace{(1 - \log \gamma_{\min})}_{\text{price to optimize over } \gamma} \inf_{\gamma \in (\gamma_{\min}, 1)} Q_j(\gamma), 1 \right\} \quad (j = 1, \dots, p).$$

Theorem 11.1 (Bühlmann and van de Geer (2011))

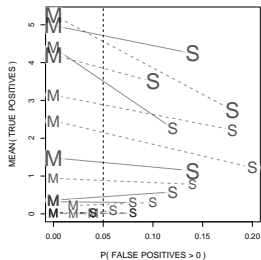
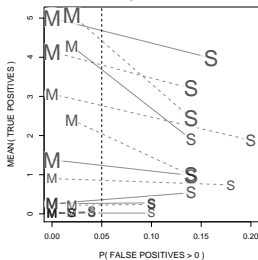
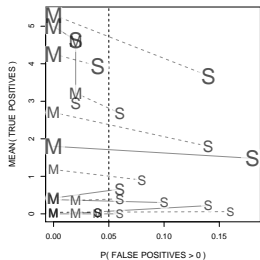
For any $\gamma_{\min} \in (0, 1)$, P_j are P-values which control the FWER

the entire framework for p-value aggregation holds whenever the single p-values are valid ($\mathbb{P}[P_{\text{raw},j} \leq \alpha] \leq \alpha$ under $H_{0,j}$)
has nothing to do with high-dimensional regression and sample splitting

$n = 100, p = 100$



$n = 100, p = 1000$



one can also adapt the method to control the False Discovery Rate (FDR)

multi sample splitting and p-value construction:

- ▶ is very generic, also for “any other” model class
- ▶ is powerful in terms of multiple testing correction: we only correct for multiplicity from $|\hat{S}(I_1)|$ variables
- ▶ it relies in theory on the screening property of the selector in practice: it is a quite competitive method!
- ▶ **Schultheiss et al. (2021)**: can improve multi sample splitting by multi carve methods, based on “technology” from selected inference