# The de-sparsified or de-biased Lasso

Recap: if $p < n$ and $\text{rank}(X) = p$, then:

$$\hat{\beta}_{\text{OLS},j} = Y^T Z^{(j)} / (X^{(j)})^T Z^{(j)}$$

$$Z^{(j)} = X^{(j)} - X^{(-j)} \hat{\gamma}^{(j)}$$

$$= \text{OLS residuals from } X^{(j)} \text{ vs. } X^{(-j)} = \{X^{(k)};\ k \neq j\}$$

$$\hat{\gamma}^{(j)} = \text{argmin}_{\gamma} \|X^{(j)} - X^{(-j)}\gamma\|_2^2$$

idea for high-dimensional setting:
use the Lasso for the residuals $Z^{(j)}$

## The de-sparsified Lasso

consider

$$
\begin{aligned}
Z^{(j)} &= X^{(j)} - X^{(-j)}\hat{\gamma}^{(j)} \\
&= \text{Lasso residuals from } X^{(j)} \text{ vs. } X^{(-j)} = \{X^{(k)};\ k \neq j\} \\
\hat{\gamma}^{(j)} &= \text{argmin}_{\gamma} \|X^{(j)} - X^{(-j)}\gamma\|_2^2 + \lambda_j \|\gamma\|_1
\end{aligned}
$$

build projection of $Y$ onto $Z^{(j)}$:

$$
\frac{Y^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \underbrace{=}_{Y = X\beta^0 + \varepsilon} \beta_j^0 + \underbrace{\sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \beta_k^0}_{\text{bias}} + \frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}
$$

estimate bias and subtract it:

$$\widehat{\text{bias}} = \sum_{k \neq j} \frac{(X^{(k)})^T X^{(j)}}{(X^{(j)})^T Z^{(j)}} \underbrace{\hat{\beta}_k}_{\text{standard Lasso}}$$

$\leadsto$ de-sparsified Lasso estimator

$$\hat{b}_j = \frac{Y^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} - \sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \hat{\beta}_k \quad (j = 1, \ldots, p)$$

not sparse! Never equal to zero for all $j = 1, \ldots, p$

can also be represented as

$$\hat{b}_j = \underbrace{\hat{\beta}_j}_{\text{standard Lasso}} + \frac{(Y - X\hat{\beta})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \quad \text{“de-biased Lasso”}$$

using that

$$\frac{Y^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} = \beta_j^0 + \sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} \beta_k^0 + \frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}$$

we obtain

$$\sqrt{n}(\hat{b}_j - \beta_j^0) = \underbrace{\sqrt{n} \sum_{k \neq j} \frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} (\beta_k^0 - \hat{\beta}_k)}_{\sqrt{n}\cdot \text{ (bias term of de-biased Lasso)}} + \underbrace{\sqrt{n} \frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}}_{\text{fluctuation term}}$$

so far, this holds for any $Z^{(j)}$

assume fixed design $X$, e.g. condition on $X$
Gaussian error $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$

fluctuation term:

$$\sqrt{n}\frac{\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}} = \frac{n^{-1/2}\varepsilon^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}/n} \sim \mathcal{N}(0, \frac{\sigma^2 \|Z^{(j)}\|_2^2/n}{|(X^{(j)})^T Z^{(j)}/n|^2})$$

bias term of de-biased Lasso: we exploit two things

- $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n})$
- KKT condition for Lasso (on $X^{(j)}$ versus $X^{(-j)}$): $|(X^{(k)})^T Z^{(j)}/n| \le \lambda_j/2$

therefore:

$$\begin{aligned} & \sqrt{n}\sum_{k\neq j}\frac{(X^{(k)})^T Z^{(j)}}{(X^{(j)})^T Z^{(j)}}(\beta_k^0 - \hat{\beta}_k) \\ = & \sqrt{n}\sum_{k\neq j}\frac{(X^{(k)})^T Z^{(j)}/n}{(X^{(j)})^T Z^{(j)}/n}(\beta_k^0 - \hat{\beta}_k) \\ \le & \sqrt{n}\max_{k\neq j}\left|\frac{(X^{(k)})^T Z^{(j)}/n}{(X^{(j)})^T Z^{(j)}/n}\right|\|\hat{\beta} - \beta^0\|_1 \\ \le & \sqrt{n}\frac{\lambda_j/2}{(X^{(j)})^T Z^{(j)}/n}O_P(s_0\sqrt{\log(p)/n}) \\ = & O_P(s_0\log(p)/\sqrt{n}) = o_P(1) \text{ if } s_0 \ll \frac{\sqrt{n}}{\log(p)} \end{aligned}$$

if $\lambda_j \asymp \sqrt{\log(p)/n}$ and $(X^{(j)})^T Z^{(j)}/n \asymp O(1)$

summarizing $\rightsquigarrow$
*Theorem 10.1 in the notes*
assume:

- $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
- $\lambda_j = C_j \sqrt{\log(p)/n}$ and $\|Z^{(j)}\|_2^2/n \geq L > 0$
- $s_0 = o(\sqrt{n}/\log(p))$ (a bit more sparse than “usual”)
- $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0 \sqrt{\log(p)/n})$
  (i.e., compatibility constant $\phi_o^2$ bounded away from zero)

Then:

$$\sigma^{-1}\sqrt{n}\frac{(X^{(j)})^T Z^{(j)}/n}{\|Z^{(j)}\|_2/\sqrt{n}}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}(0,1) \quad (j = 1, \ldots, p)$$

more precisely:

$$\sigma^{-1}\sqrt{n}\frac{(X^{(j)})^T Z^{(j)}/n}{\|Z^{(j)}\|_2/\sqrt{n}}(\hat{b}_j - \beta_j^0) = W_j + \Delta_j$$

$$(W_1, \ldots, W_p)^T \sim \mathcal{N}_p(0, \Omega),\ \Omega_{jj} \equiv 1\ \forall j,\ \max_{j=1,\ldots,p} |\Delta_j| = o_P(1)$$

confidence intervals for $\beta_j^0$:

$$\hat{b}_j \pm \hat{\sigma} n^{-1/2} \frac{\|Z^{(j)}\|_2/\sqrt{n}}{|(X^{(j)})^T Z^{(j)}/n} \Phi^{-1}(1 - \alpha/2)$$

$\hat{\sigma}^2 = \|Y - X\hat{\beta}\|_2^2/n$ or $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|_2^2/(n - \|\hat{\beta}\|_0^0)$

can also test

$$H_{0,j}: \ \beta_j^0 = 0 \text{ versus } H_{A,j}: \ \beta_j^0 \neq 0$$

can also test group hypothesis: for $G \subseteq \{1, \ldots, p\}$

$$H_{0,G}: \ \beta_j^0 \equiv 0 \forall j \in G$$
$$H_{A,G}: \exists j \in G \text{ such that } \beta_j^0 \neq 0$$

under $H_{0,G}$:

$$\max_{j \in G} \sigma^{-1} \sqrt{n} \frac{|(X^{(j)})^T Z^{(j)}/n|}{\|Z^{(j)}\|_2/\sqrt{n}} |\hat{b}_j| = \max_{j \in G} |W_j + \Delta_j| \asymp \underbrace{\max_{j \in G} |W_j|}_{\text{distr. simulated}}$$

and plug-in $\hat{\sigma}$ for $\sigma$

# Choice of tuning parameters

as usual: $\hat{\beta} = \hat{\beta}(\hat{\lambda}_{\mathrm{CV}})$; what is the role of $\lambda_j$?

$$\text{variance} = \sigma^2 n^{-1} \frac{\|Z^{(j)}\|_2^2/n}{|(X^{(j)})^T Z^{(j)}/n|^2} \asymp \sigma^2/\|Z^{(j)}\|_2^2$$

if $\lambda_j \searrow$ then $\|Z^{(j)}\|_2^2 \searrow$, i.e. large variance

error due to bias estimation is bounded by:

$$|\ldots| \le \sqrt{n} \frac{\lambda_j/2}{|(X^{(j)})^T Z^{(j)}/n|} \|\hat{\beta} - \beta^0\|_1 \propto \lambda_j$$

assuming $\lambda_j$ is not too small
if $\lambda_j \searrow$ (but not too small) then bias estimation error $\searrow$

$\rightsquigarrow$ inflate the variance a bit to have low error due to bias estimation: control type I error at the price of slightly decreasing power

## 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: $s_j \ll n/\log(p)$

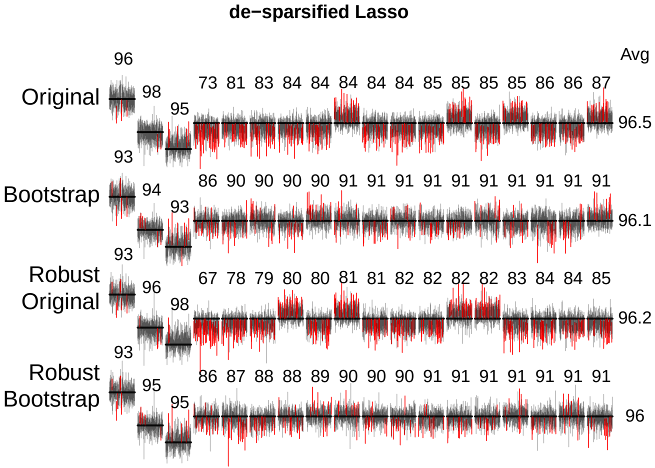
then... skipping details, the de-biased Lasso achieves (see Theorem 10.2):

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}(0, \underbrace{\sigma^2 \Theta_{jj}}_{\text{Cramer-Rao lower bound}})$$

$\Theta = \Sigma_X^{-1} = \mathrm{Cov}(X)^{-1} \rightsquigarrow$ 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
SERIES B Statistical Methodology

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

**Stability selection**

Nicolai Meinshausen
*University of Oxford, UK*

and Peter Bühlmann
*Eidgenössiche 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. Titterington *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 $Z_1, \ldots, Z_n$

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

$\hat{S}_\lambda$ is a "feature selection" method/algorithm among $\{1, \ldots, p\}$ features

can we assign "relevance" to the selected features in $\hat{S}_\lambda$?

a "natural" approach: resampling!
here: use subsampling:

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

regarding the latter, for example:

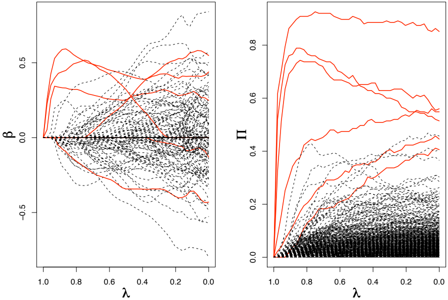
$$\hat{\Pi}_K(\lambda) = \mathbb{P}^*[K \subseteq \hat{S}_\lambda(I^*)] \approx B^{-1} \sum_{b=1}^{B} I(K \subseteq \hat{S}_\lambda(I^{*b}))$$

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

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
$\leadsto$ it is approximated by $B$ ($\approx 100$) 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 $\pi_{\text{thr}}$ 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
$\Lambda$ 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 $\pi_{\text{thr}}$ with $\mathbb{E}[V]$

consider a setting with $p$ possible features
$\hat{S}(\lambda)$ is a feature selection algorithm
$\hat{S}_\Lambda = \cup_{\lambda \in \Lambda} \hat{S}(\lambda)$
$q_\Lambda = \mathbb{E}[\hat{S}_\Lambda(\underbrace{I}_{\text{random subsample}})]$

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

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

suppose we know $q_\Lambda$ (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

$\hat{S}_\lambda =$ the top 10 variables from Lasso (e.g. the different $\lambda$ 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 $\lambda$ corresponds to the "top 10"; $\Lambda$ 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

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

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

$$\pi_{\text{thr}} = \frac{1}{2} + \frac{q_\Lambda^2}{2pv_0} = 0.5 + \frac{30^2}{2 * 1000 * 5} = 0.59$$

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

error control becomes better (for the same threshold)

$$\mathbb{E}[V] \leq \frac{1}{2\pi_{\mathrm{thr}} - 1} \frac{q_\Lambda^2}{p_{\mathrm{enlarged}}}$$

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

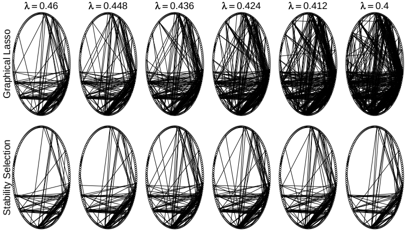
$$Y = X\beta^0 + \varepsilon,\ \mathrm{Cov}(X)_{i,j} \equiv \rho\ (i \neq j),\ \mathrm{Var}(X_j) \equiv 1 \forall j$$

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 $\pi : S_0^c \to S_0^c$

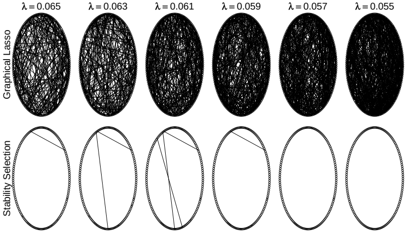
- distribution of $X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all $\pi$ (because of equi-correlation)
- distribution of $Y|X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all $\pi$ (because it depends only on $X^{(S_0)}$)
- therefore: distribution of $Y, X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all $\pi$
  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