The de-biased Lasso: its Gaussian limiting distribution

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

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

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

all is very easy!

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 σ

Choice of tuning parameters

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

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

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

error due to bias estimation is bounded by:

$$|\ldots| \le \sqrt{n} \frac{\lambda_j/2}{|(X^{(j)})^T Z^{(j)}/n|} \|\hat{\beta} - \beta^0\|_1 \propto \frac{\lambda_j}{\|Z^{(j)}\|_2^2/n}$$

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

$\leadsto$ 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!

Why the $1/\sqrt{n}$ convergence rate?

de-biased/de-sparsified Lasso is considering

- low-dimensional components $\{\beta^0_j;\, j \in A\}$ with $|A|$ small

  $$\sum_{j \in A} c_j \sqrt{n}(\hat{b}_j - \beta^0_j) \Rightarrow \mathcal{N}(0, \sum_{j,j' \in A} c_j c_{j'} V_{j,j'}),\ V = \lim_n n\mathrm{Cov}(\hat{b})$$

  for large $|A|$: the sum would blow up the variance and the scaling with $\sqrt{n}$ is not correct

- high-dimensional β^0 and ℓ_∞-norm:

  $$\sqrt{n}\|\hat{b} - \beta^0\|_\infty \sim \underbrace{\text{maximum of } p \text{ dependent Gaussian r.v.'s}}_{\sim \underbrace{C\sqrt{\log(p)}}_{\text{under independence/weak dependence}}}$$

  $\rightsquigarrow \sqrt{\log(p)/n}$ convergence rate

Multiple testing adjustment

if we test all hypotheses, for all $j = 1, \ldots, p$:

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

we have to adjust/correct for multiple testing
different type I error measures:
for multiple tests, one can control for:

$$\text{FamilyWise Error Rate:} \qquad \text{FWER} = P[V > 0],$$
$$\text{False Discovery Rate:} \qquad \text{FDR} = \mathbb{E}[V/R],$$
$$V = \text{ number of } \underbrace{\text{false positives}}_{\text{null-hyp. rejected although it is true}}, \ R = \text{ number of rejections}$$

other measures exist: but these are the two most common ones

- input: raw p-values p_j for jth hypothesis test (e.g. from de-biased Lasso)
- output: corrected p-values $p_{\mathrm{corr},j}$

  reject $H_{0,j} \iff p_{\mathrm{corr},j} \le \alpha$: then,

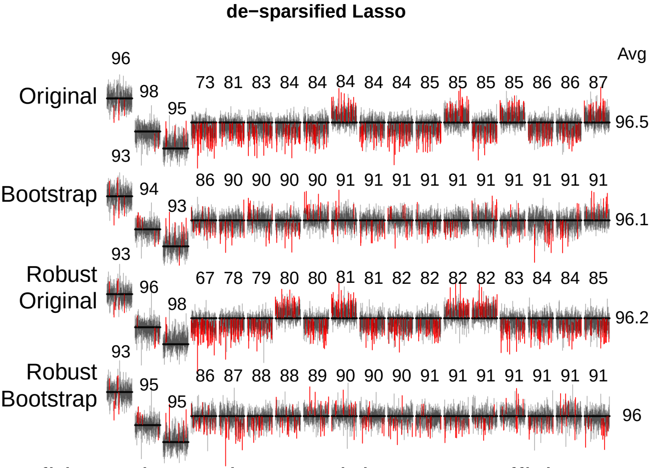
  $$\text{FWER} \le \alpha \text{ or FDR} \le \alpha$$

  depending on the adjustment method
- for controlling FWER: Bonferroni-Holm procedure
  for controlling FDR: Benjamini-Hochberg procedure (which is only proven to be correct for independent hypotheses)

`R`-software: `p.adjust` or also package `hdi` has some more clever adjustment for dependent p-values from de-biased Lasso

Empirical results

R-software `hdi` (Meier et al.)



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 $\hat{S}_\lambda$?

prime example: $\hat{S}_\lambda$ from Lasso in linear model with p covariates

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:

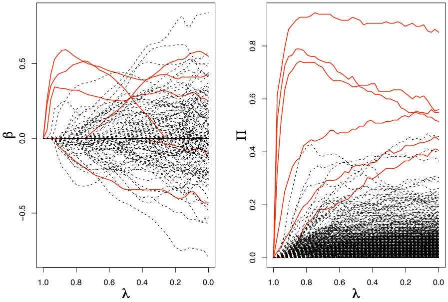
$$\hat{\Pi}_K(\lambda) = \mathbb{P}^*[K \subseteq \hat{S}_\lambda(I^*)] \approx B^{-1} \sum_{b=1}^{B} I(K \subseteq \hat{S}_\lambda(I^{*b}))$$

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 (≈ 500) 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

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

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):

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

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
$\hat{S}_\Lambda = \cup_{\lambda \in \Lambda} \hat{S}\lambda$
$q_\Lambda = \mathbb{E}[\hat{S}_\Lambda(\underbrace{\quad I \quad}_{\text{random subsample}})]$

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

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

suppose we know q_Λ (see later)
strategy: specify $\mathbb{E}[V] = v_0$ (e.g. $= 5$)
$\rightsquigarrow$ for $\pi_{\mathrm{thr}} := \frac{1}{2} + \frac{q_\Lambda^2}{2pv_0}$: $\mathbb{E}[V] \leq v_0$

example: regression model with $p = 1000$ variables

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

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

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

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

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

error control becomes better (for the same threshold)

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

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

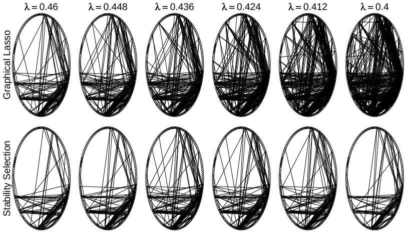
- distribution of $X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all π (because of equi-correlation)
- distribution of $Y | X^{(S_0)}, \{X^{(\pi(j))};\ j \in S_0^c\}$ is the same for all π (because it depends only on $X^{(S_0)}$)
- therefore: distribution of $Y, X^{(S_0)}, \{X^{(\pi(j))};\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

An illustration for graphical modeling

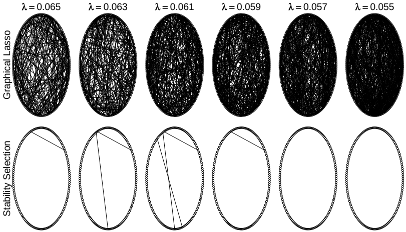
$p = 160$ gene expressions, $n = 115$

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