Extensions on theoretical results for slow convergence

aim: analyze $\|X(\hat{\beta} - \beta^0\|_2/n$ without (major) conditions on the design X
the proof technique decouples into a deterministic and probabilistic part (the set $\mathcal{T}$)

the deterministic part remains the same for other probabilistic structures (other analysis for $\mathbb{P}[\mathcal{T}]$) such as:

- heteroscedastic errors with
  $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 \not\equiv$ const.
- dependent observs. $\rightsquigarrow$ fixed design and dependent errors
- non-Gaussian errors:
  – sub-Gaussian distribution
  – second moments plus bounded X: see Example 14.3 in Bühlmann and van de Geer (2011)
- random design: assume that ε is independent of X
  $\rightsquigarrow$ condition on X: invoke the results for fixed design and integrate out

heteroscedastic errors

$\varepsilon \sim \mathcal{N}_n(0, D)$, where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

assume that: $\sigma_i^2 \leq \underbrace{\sigma^2}_{\text{some pos. const.}} < \infty$

Then, Corollary 6.1 remains true with σ^2 as above

Proof:
exactly as before but exploiting that $V_j \sim \mathcal{N}(0, \tau_j^2)$ with $\tau_j \leq 1$
and using that $\mathbb{P}[|V_j| > c] \leq \mathbb{P}[\underbrace{|Z|}_{\sim|\mathcal{N}(0,1)|} > c]$

Exercise: work out the details.

errors from stationary distribution

$\varepsilon \sim \mathcal{N}_n(0, \Gamma)$, where $\Gamma_{i,j} = R(i-j) = R(j-i)$

assume that: $\sum_{k=-\infty}^{\infty} |R(k)| < \infty$ and $|X_i^{(j)}| \leq K_X < \infty$

Then, Corollary 6.1 remains true with $\sigma^2 = K_X^2 \sum_{k=-\infty}^{\infty} |R(k)|$

Proof:
Exercise. (A bit more tricky...)

Compatibility condition, Restricted eigenvalues, Sparse eigenvalues

if we want to identify and estimate the regression parameter β^0
$\rightsquigarrow$ need necessarily more assumptions

example:
if $X^{(1)} = X^{(2)} \rightsquigarrow$ cannot distinguish between β_1^0 and β_2^0
– in low dimensions: typically assume $\text{rank}(X) = p$ $(\leq n)$
– $\text{rank}(X) \leq \min(n, p) \rightsquigarrow$ for $p > n$: never achieve full rank p

how to measure “degree” of identifiability?

suppose $X\theta = X\beta^0$
then:

$$\begin{aligned} 0 &= \|X(\theta-\beta^0)\|_2^2/n = (\theta-\beta^0)^T\hat{\Sigma}(\theta-\beta^0) \\ &\geq \underbrace{\lambda_{\min}^2(\hat{\Sigma})}_{\text{min. eigenval. of } \hat{\Sigma}} \|\theta-\beta^0\|_2^2 \end{aligned}$$

$\hat{\Sigma} = X^T X/n$

since $\lambda_{\min}^2(\hat{\Sigma}) = \min_{u;u\neq 0} \frac{u^T\hat{\Sigma}u}{/\|u\|_2^2}$

for $p > n$: $\lambda_{\min}^2(\hat{\Sigma}) = 0 \rightsquigarrow$ bound above is “useless”

idea: restrict to small sub-matrices

$\rightsquigarrow$ sparse eigenvalues (Meinshausen & Yu, 2009)

$$\phi^2_{\min}(m) = \min_{S \subseteq \{1,\ldots,p\}} \left(\lambda^2_{min}(\hat{\Sigma}_S); |S| \leq m \right)$$

$$\iff \qquad \phi^2_{\min}(m) = \min_{\beta \neq 0; \|\beta\|_0 \leq m} \frac{\beta^T \hat{\Sigma} \beta}{\|\beta\|_2^2}$$

Denote by $s_\theta = \|\theta\|_0$, $s_0 = \|\beta^0\|_0$

Then: if we require $\phi^2_{\min}(s_\theta + s_0) > 0$:

since $\|\theta - \beta^0\|_0 \leq s_\theta + s_0$ we obtain

$$0 = \|X(\theta - \beta^0)\|_2^2/n \geq \phi^2_{\min}(s_\theta + s_0)\|\theta - \beta^0\|_2^2$$

$$\rightsquigarrow \quad \theta = \beta^0$$

Conclusion:
if we restrict to sparse vectors θ with at most the sparsity of β^0, i.e., $\|\theta\|_0 = s_\theta \leq \|\beta^0\|_0 = s_0$

$\rightsquigarrow$ can identify the regression parameter vector if $\phi^2_{\min}(2s_0) > 0$

(have not shown that identification works with the Lasso...)

Restricted eigenvalues

instead of sparse eigenvalues:
Lasso identifies β^0 under weaker conditions

idea: can restrict to additional cone condition:
for a subset $S \subseteq \{1, \ldots, p\}$, consider cone

$$\|\beta_{S^c}\|_1 \leq \underbrace{3}_{\text{arbitrary}} \|\beta_S\|_1 \qquad (1)$$

the ℓ_1-norm in the (large) complement S^c of S is small
Restricted eigenvalue (Bickel, Ritov & Tsybakov, 2009):

$$\kappa^2(m, 3) = \min_{S;\ |S| \leq m} \quad \min_{\beta \neq 0,\ \beta \text{ fulfills (1)}} \frac{\beta^T \hat{\Sigma} \beta}{\|\beta_S\|_2^2}$$

Why such a cone?

can show that for Lasso, with high probability (on $\mathcal{T}$):
cone condition is fulfilled for $\hat{\beta} - \beta^0$ for $S = S_0$

$$\text{on } \mathcal{T}: \ \|(\hat{\beta} - \beta^0)_{S_0^c}\|_1 = \|\hat{\beta}_{S_0^c}\|_1 \leq 3\|(\hat{\beta} - \beta^0)_{S_0}\|_1$$

if $\kappa^2(s_0, 3) > 0$: can identify β^0 with the Lasso
Proof: an incomplete version, see visualizer

Compatibility constant

slightly weaker condition the restricted eigenvalue Compatibility constant for the set S_0 (van de Geer, 2007):

$$\phi_0^2 = \min_{\beta \neq 0} \quad \min_{\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1} \frac{s_0 \beta^T \hat{\Sigma} \beta}{\|\beta_{S_0}\|_1^2} \quad (s_0 = |S_0| = \|\beta^0\|_0)$$

see also Bühlmann and van de Geer, p. 106

It holds that

$$\phi_0^2 \geq \kappa^2(s_0, 3)$$

Proof: Cauchy-Schwarz inequality

$$\|\beta_{S_0}\|_1 \leq \sqrt{s_0}\|\beta_{S_0}\|_2$$

In the denominator of $\kappa^2(s_0, 3)$ we have $\|\beta_{S_0}\|_2^2$
$\rightsquigarrow \phi_0^2 \geq \kappa^2(s_0, 3)$. □

compatibility condition holds:
$\phi_0^2 > 0$

and this is still enough to identify the true parameter with Lasso using only $\phi_0^2 > 0$ (instead of the stronger assumption $\kappa^2(s_0, 3) > 0$)

Oracle inequality for the Lasso

$$Y = X\beta^0 + \varepsilon, \quad p \gg n$$

for the Lasso:
Theorem 6.1 in Bühlmann and van de Geer (2011)
assume: compatibility condition holds with compatibility constant ϕ_0^2 ($\geq L > 0$)
Then, on $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$:

$$\|X(\hat{\beta} - \beta^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2$$

recall: $\mathcal{T} = \{2 \max_{j=1,\ldots,p} |\varepsilon^T X^{(j)}/n| \leq \lambda_0\}$

$\mathbb{P}[\mathcal{T}]$ large if $\lambda_0 \asymp \sqrt{\log(p)/n}$

When does the compatibility condition hold?

Corollary 6.8 from Bühlmann and van de Geer (2011) – modified form

Assume that the row vectors of X are i.i.d. sampled from a sub-Gaussian distribution with mean zero and covariance matrix Σ. Assume that

- $\lambda^2_{min}(\Sigma) > 0$
- $s_0 = |S_0| = O(\sqrt{n/\log(p)})$

Then, for some $C > 0$:

$$\phi_0^2 \geq C\lambda^2_{\min}(\Sigma) > 0 \text{ with probability } \to 1 \ (n \to \infty)$$

Example: Toeplitz matrix $\Sigma_{ij} = \rho^{|i-j|}$ $(0 \leq \rho < 1)$:
$\lambda^2_{min}(\Sigma) \geq L_\rho > 0$ where L_ρ is independent of p

Implications of oracle inequality

assume that $\phi_0^2 \geq L > 0$

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

if $\lambda(= 2\lambda_0) \asymp \sqrt{\log/n}$: as $p \geq n \to \infty$,

$$\|X(\hat{\beta} - \beta^0\|_2^2/n = O_P(s_0 \log(p)/n)$$
$$\|\hat{\beta} - \beta^0\|_1 = O_P(s_o\sqrt{\log(p)/n})$$

thus, when compatibility condition holds:

1. fast rate of convergence for prediction
   if S_0 with $s_0 = o(n)$ would be known:
   $\|X\hat{\beta}_{\text{OLS}} - \beta^0\|_2^2/n = O_P(s_0/n)$
   $\rightsquigarrow$ factor $\log(p)$ is the (small!) price for not knowing S_0
2. estimation error for β^0 in terms of the ℓ_1-norm

Variable screening

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

estimated active set: $\hat{S}_0 = \{j;\ \hat{\beta}_j \neq 0\}$

make an assumption that true regression coefficients are not too small

"beta-min condition" : $\min_{j \in S_0} |\beta_j^0| > \underbrace{4\lambda s_0/\phi_0^2}_{\text{bound for } \|\hat{\beta} - \beta^0\|_1}$

$$\Longrightarrow\ \mathbb{P}[\hat{S} \supseteq S_0] \geq \mathbb{P}[\mathcal{T}]\ = \text{"large"}$$

with high probab: Lasso selects a superset of the active set S_0

$\rightsquigarrow$ Lasso does not miss an important active variable!

$$\mathbb{P}[\hat{S} \supseteq S_0] \geq \mathbb{P}[\mathcal{T}] \ = \text{“large”}$$

Proof:
Suppose that $\hat{S} \not\supseteq S_0$: $\rightsquigarrow$ there exists $j^* \in S_0$ with $\hat{\beta}_{j^*} = 0$
But then, on $\mathcal{T}$:
$\|\hat{\beta} - \beta^0\|_1 \geq \|\hat{\beta}_{j^*} - \beta^0_{j^*}| = |\beta^0_{j^*}| > 4\lambda s_0/\phi_0^2$
which is a contradiction to the oracle inequality
(for $\|\hat{\beta} - \beta^0\|_1 \leq \lambda 4 s_0/\phi_0^2$)
□

Theory versus Practice

theory:

$$\mathbb{P}[\hat{S} \supseteq S_0] \to 1$$

if the following hold:

- compatibility condition for the (fixed) design X
- beta-min condition
- i.i.d. Gaussian errors (can be relaxed)

in addition: $|\hat{S}| \leq \min(n, p)$
hence: huge dimensionality reduction if $p \gg n$

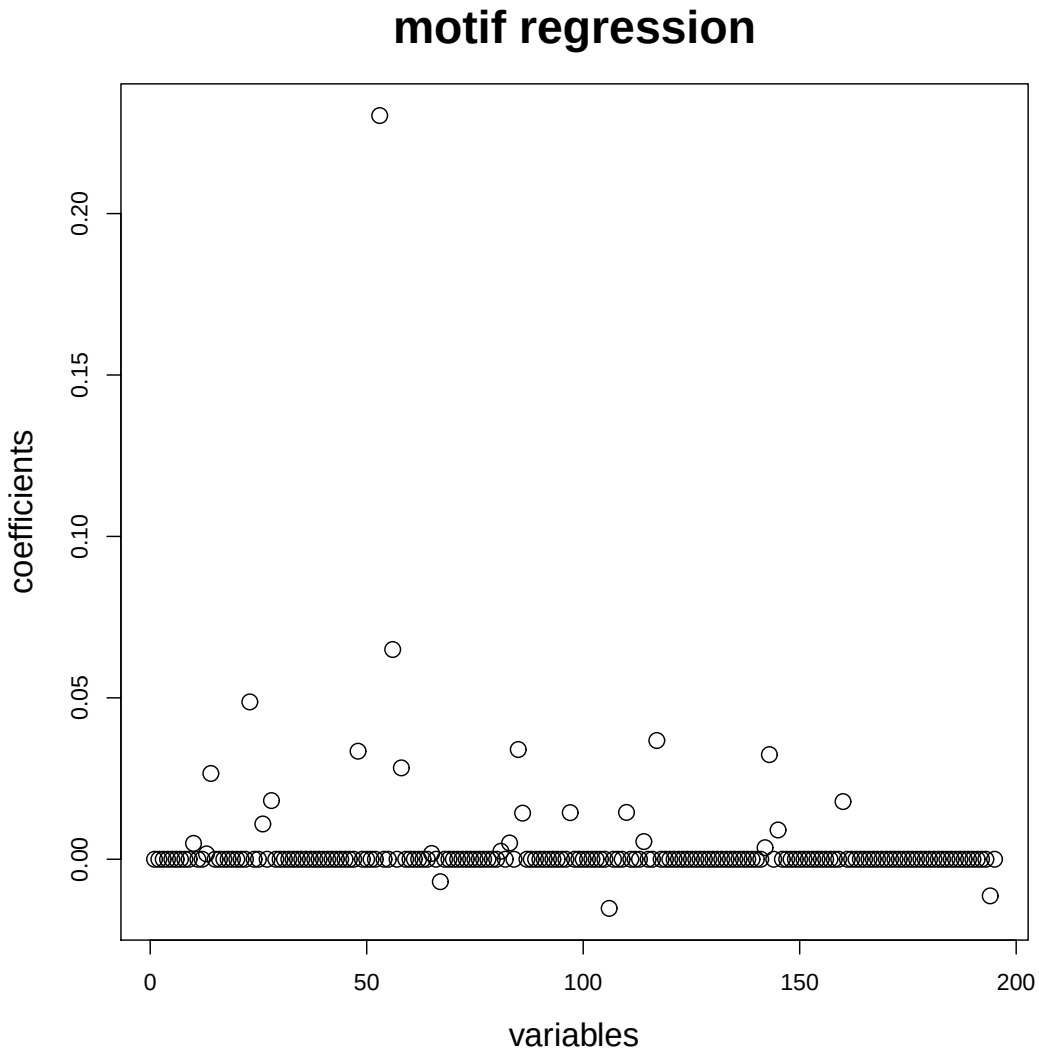
in practice: $\mathbb{P}[\hat{S} \supseteq S_0]$ may not be so large...
even if one chooses λ very small which results in a typically larger set $\hat{S}$...

possible reasons to explain with theory:

- compatibility constant ϕ_0^2 might be very small (due to highly correlated columns in X or near linear dependence among a few columns of X)
  $\rightsquigarrow \|\hat{\beta} - \beta^0\|_1 \leq 4\lambda s_0/\phi_0^2$
  $\rightsquigarrow$ requires a stronger beta-min condition!
- errors are non-Gaussian (heavy tailed)

it is "empirically evident" though: $\mathbb{P}[\hat{S} \supseteq S_{\text{substantial}(C)}]$ large
where $S_{\text{substantial}(C)} = \{j;\ |\beta_j^0| \geq \underbrace{C}_{\text{large}}\}$

The Lasso workhorse



$$p = 195, n = 143, |\hat{S}(\lambda_{CV})| = 26$$

Variable selection

under more restrictive irrepresentable condition or
neighborhood stability condition on the design X
and assuming beta-min condition $\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$:

$$\mathbb{P}[\hat{S} = S_0] \to 1 \ (n \to \infty)$$

the irrepresentable condition is sufficient and essentially
necessary for consistent variable selection

this condition is often not fulfilled in practice
(and choosing the correct λ would be difficult as well)

$\rightsquigarrow$ variable screening is realistic ("choose λ by CV")
variable selection is not very realistic

better "translation":
LASSO = Least Absolute Shrinkage and Screening Operator

Variable selection

under more restrictive irrepresentable condition or
neighborhood stability condition on the design X
and assuming beta-min condition $\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$:

$$\mathbb{P}[\hat{S} = S_0] \to 1 \ (n \to \infty)$$

the irrepresentable condition is sufficient and essentially
necessary for consistent variable selection

this condition is often not fulfilled in practice
(and choosing the correct λ would be difficult as well)

$\rightsquigarrow$ variable screening is realistic ("choose λ by CV")
variable selection is not very realistic
better "translation":
LASSO = Least Absolute Shrinkage and Screening Operator

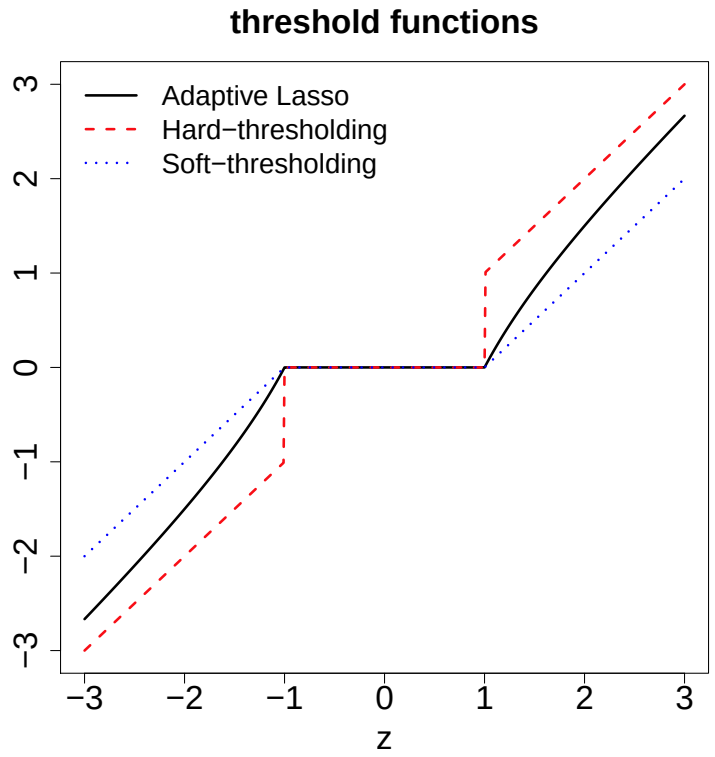
version of Table 2.2 in the book:

property	design condition	size of non-zero coeff.
slow prediction conv. rate	no requirement	no requirement
fast prediction conv. rate	compatibility	no requirement
estimation error bound $\|\hat{\beta} - \beta^0\|_1$	compatibility	no requirement
variable screening	compatibility or restricted eigenvalue	beta-min condition weaker beta-min cond.
variable selection	neighborhood stability $\Leftrightarrow$ irrepresentable cond.	beta-min condition

Adaptive Lasso

is a good way to address the bias problems of the Lasso

for orthonormal design



two-stage procedure:

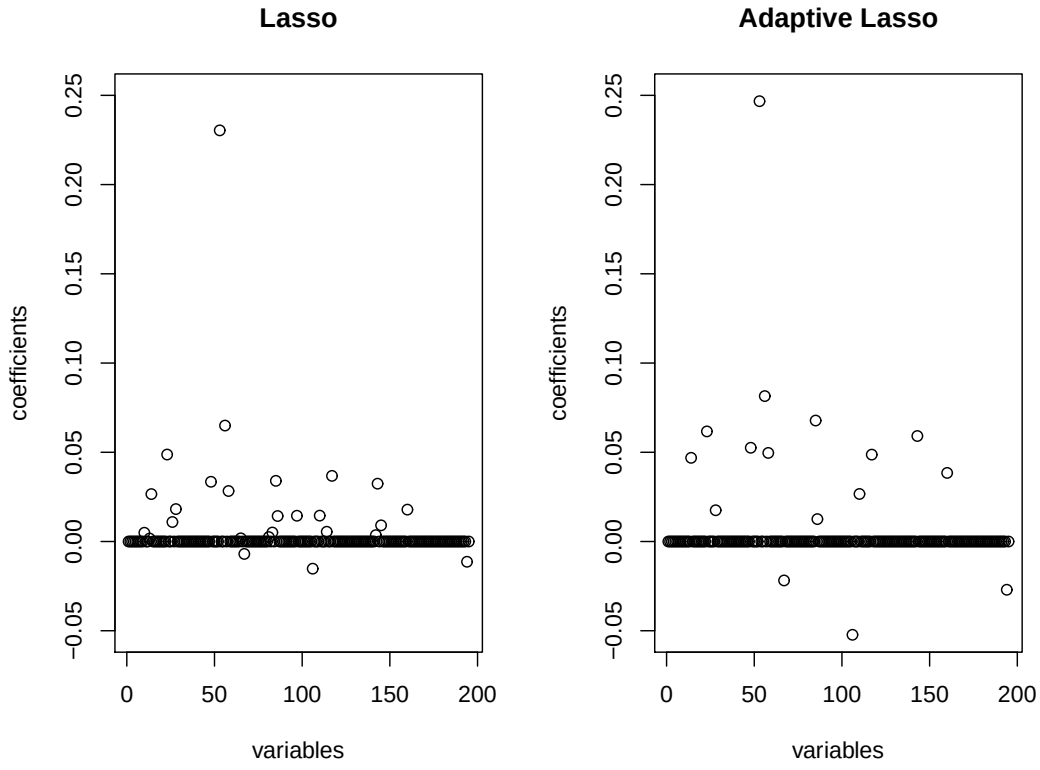
- initial estimator $\hat{\beta}_{\text{init}}$, e.g., the Lasso
- re-weighted ℓ_1-penalty

$$\hat{\beta}_{\text{adapt}}(\lambda) = \text{argmin}_\beta \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{\text{init},j}|} \right)$$

adaptive Lasso often works well in practice (more sparse than Lasso) and has better theoretical properties than Lasso for variable screening (and selection) if the truth is assumed to be sparse

alternatives: thresholding the Lasso; Relaxed Lasso

The adaptive Lasso workhorse



$$p = 195, n = 143, |\hat{S}_{\text{ada}-\text{Lasso}}(\lambda_{CV})| = 16$$

we will discuss later in the course the issue of assigning "significance of selected variables"

should we always use the adaptive Lasso?

- it's slightly more complicated – need two Lasso fits
- the differences in large-scale data are perhaps not so large
- I tend to say:
  "Yes, often the adaptive Lasso is perhaps a bit better"

Computational algorithm for Lasso

can use a very generic coordinate descent algorithm (not gradient descent)

motivation of the algorithm:
consider the objective function and the corresponding Karush-Kuhn-Tucker (KKT) conditions by taking the sub-differential:

$$
\begin{aligned}
& \frac{\partial}{\partial j}(\|Y - X\beta\|_2^2/n + \lambda\|\beta\|_1) \\
= \quad & G_j(\beta) + \lambda e_j, \\
& G(\beta) = -2X^T(Y - X\beta)/n, \\
& e_j = \text{sign}(\beta_j) \text{ if } \beta_j \neq 0, \quad e_j \in [-1, 1] \text{ if } \beta_j = 0
\end{aligned}
$$

this implies (by setting the sub-differential to zero) the KKT-conditions (Lemma 2.1, Bühlmann and van de Geer (2011):

$$G_j(\hat{\beta}) = -\text{sign}(\hat{\beta}_j)\lambda \text{ if } \hat{\beta}_j \neq 0,$$
$$|G_j(\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0.$$

an interesting characterization of the Lasso solution!

in abbreviated form:

1: Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. For $m = 1, 2, \ldots$

2: **repeat**

3: Proceed componentwise $j = 1, 2, \ldots, p, 1, 2, \ldots p, 1, 2, \ldots$

update:

if $|G_j(\underbrace{\beta_{-j}^{[m-1]}}_{\text{prev. parameter with } j\text{th comp}=0})| \leq \lambda$: set $\beta_j^{[m]} = 0$,

otherwise: $\beta_j^{[m]}$ is the minimizer of the objective function with respect to the jth component but keeping all others fixed

4: **until** numerical convergence

1: Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. Set $m = 0$.
2: **repeat**
3: Increase m by one: $m \leftarrow m + 1$.
Denote by $\mathcal{S}^{[m]}$ the index cycling through the coordinates $\{1, \ldots, p\}$:
$\mathcal{S}^{[m]} = \mathcal{S}^{[m-1]} + 1 \bmod p$. Abbreviate by $j = \mathcal{S}^{[m]}$ the value of $\mathcal{S}^{[m]}$.
4: if $|G_j(\beta_{-j}^{[m-1]})| \leq \lambda$: set $\beta_j^{[m]} = 0$,
otherwise: $\beta_j^{[m]} = \text{argmin}_{\beta_j} Q_\lambda(\beta_{+j}^{[m-1]})$,
where $\beta_{-j}^{[m-1]}$ is the parameter vector where the jth component is set to zero and $\beta_{+j}^{[m-1]}$ is the parameter vector which equals $\beta^{[m-1]}$ except for the jth component where it is equal to β_j (i.e. the argument we minimize over).
5: **until** numerical convergence

for the squared error loss: the update in Step 4 is explicit (a soft-thresholding operation)

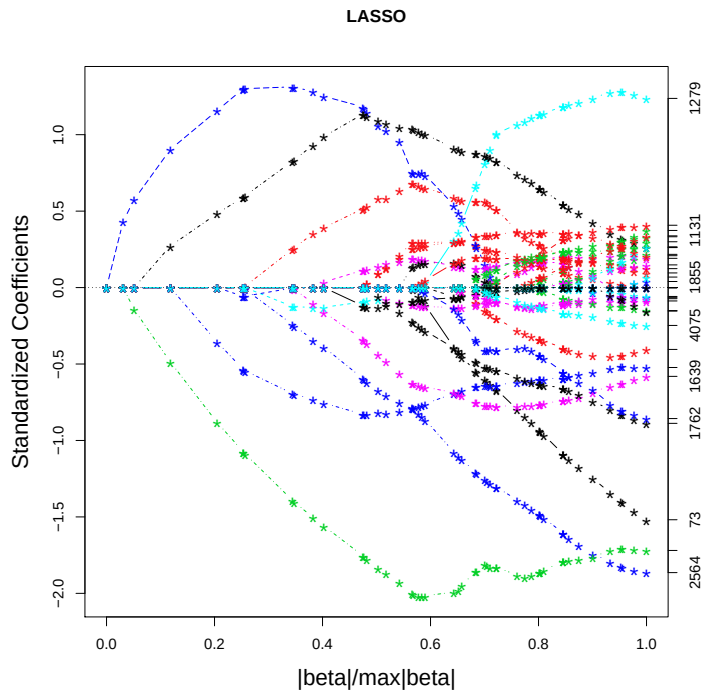
active set strategy can speed up the algorithm for sparse cases: mainly work on the non-zero coordinates and up-date all coordinates e.g. every 20th times

`R`-package `glmnet`

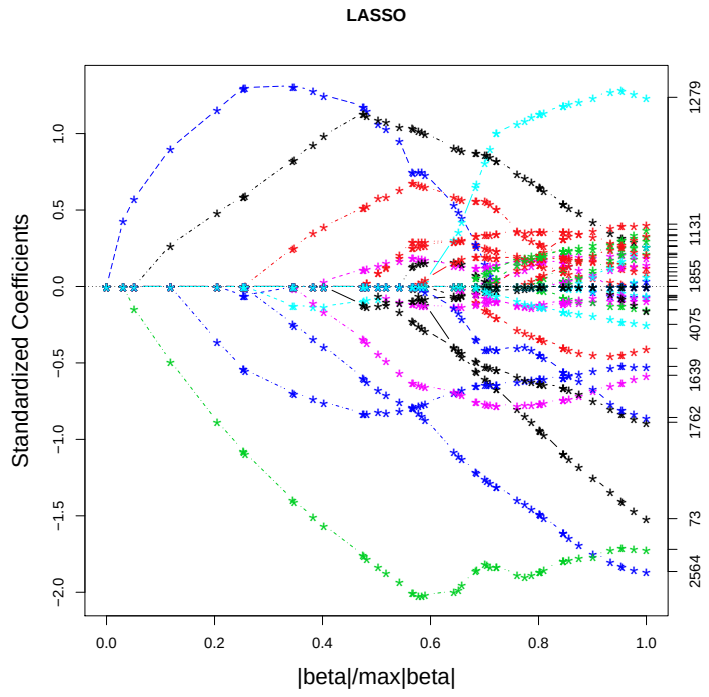
The Lasso regularization path

compute $\hat{\beta}(\lambda)$ over "all" λ

- just a grid of λ-values and interpolate linearly (the true solution path over all λ is piecewise linear)
- for $\lambda_{\max} = |2X^T Y/n|$: $\hat{\beta}(\lambda_{\max}) = 0$
  (because of KKT conditions!)



plot against $\|\hat{\beta}(\lambda)\|_1 / \max_\lambda \|\hat{\beta}(\lambda)\|_1$ (λ small is to the right)



regularization path: in general, “not monotone in the non-zeros”
it can happen in general that e.g.

$$\hat{\beta}_j(\lambda) \neq 0,\ \hat{\beta}_j(\lambda') = 0 \text{ for } \lambda' < \lambda$$