# The Lasso (Tibshirani, 1996)

Lasso for linear models

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

$\rightsquigarrow$ convex optimization problem

- Lasso does variable selection
  some of the $\hat{\beta}_j(\lambda) = 0$
  (because of "$\ell_1$-geometry")
- $\hat{\beta}(\lambda)$ is a shrunken LS-estimate

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

Lasso for variable selection:

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

for $\quad S_0 = \{j; \beta_j^0 \neq 0\}$

no significance testing involved
it's convex optimization only!

(and that can be a problem... see later)

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

Lasso for variable selection:

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

for $$S_0 = \{j; \beta_j^0 \neq 0\}$$

no significance testing involved
it's convex optimization only!

(and that can be a problem... see later)

# Some results from asymptotic theory

triangular array of observations:

$$Y_{n;i} = \sum_{j=1}^{p_n} \beta_{n;j} X_{n;i}^{(j)} + \varepsilon_{n;i}, \; i = 1, \ldots, n; \quad n = 1, 2, \ldots$$

consistency:

$$(\hat{\beta}(\lambda) - \beta_0)^T \Sigma_X (\hat{\beta}(\lambda) - \beta_0) = o_P(1) \; (n \to \infty),$$

$\Sigma_X = n^{-1}\mathbf{X}^T\mathbf{X}$ in case of a fixed design
$\Sigma_X$ equals covariance of the covariate $X$ in case of a random design

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n \text{ for fixed design,}$$
$$\mathbb{E}[(X_{new}(\hat{\beta}(\lambda) - \beta^0))^2] \text{ for random design,}$$

consistency holds under the main assumption:

$$\|\beta\|_1 = O\left(\sqrt{\frac{n}{\log(p)}}\right)$$

when choosing $\lambda$ in a suitable range.

optimal prediction:

$$\mathbb{E}[\|\mathbf{X}(\hat{\beta}(\lambda) - \beta^0)\|_2/n] = O\left(\frac{s_0 \log(p)}{n}\right),$$

$s_0 = \mathrm{card}(S_0)$
if one would knew a-priori the $s_0$ relevant covariates
use OLS which yields

$$\mathbb{E}[\|\mathbf{X}(\hat{\beta}_{OLS} - \beta^0)\|_2/n] = \frac{s_0}{n}$$

for optimal prediction: we need additional assumptions on the design $\mathbf{X}$

# Variable screening

under some additional assumptions on the design:

for suitable $\lambda = \lambda_n$ and with large probability

$$\|\hat{\beta} - \beta\|_1 = \sum_{j=1}^{p} |\hat{\beta}_j - \beta_j| \leq \underbrace{C}_{\text{depending on } \mathbf{X}, \sigma^2} \sqrt{\log(p) s_0 / n}$$

hence: $\max_j |\hat{\beta}_j - \beta_j| \leq \|\hat{\beta} - \beta\|_1 \leq C\sqrt{\log(p) s_0/n}$

and if $\min_j \{|\beta_j|;\ \beta_j \neq 0\} > C\sqrt{\log(p) s_0/n}$

then $\hat{\beta}_j \neq 0$ for all $j \in S_0$, i.e. $\hat{S} \supseteq S_0$

with large probability

$$\hat{S} \supseteq S_0$$

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

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

furthermore: "typically", for prediction-optimal $\lambda_{\mathrm{opt}}$

$$\hat{S}(\lambda_{\mathrm{opt}}) \supseteq S_0$$

$\leadsto$ Lasso as an
excellent screening procedure

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

with large probability

$$\hat{S} \supseteq S_0$$

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

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

furthermore: "typically", for prediction-optimal $\lambda_{\text{opt}}$

$$\hat{S}(\lambda_{\text{opt}}) \supseteq S_0$$

$\rightsquigarrow$ Lasso as an
excellent screening procedure

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

with large probability

$$\hat{S} \supseteq S_0$$

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

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

furthermore: "typically", for prediction-optimal $\lambda_{\mathrm{opt}}$

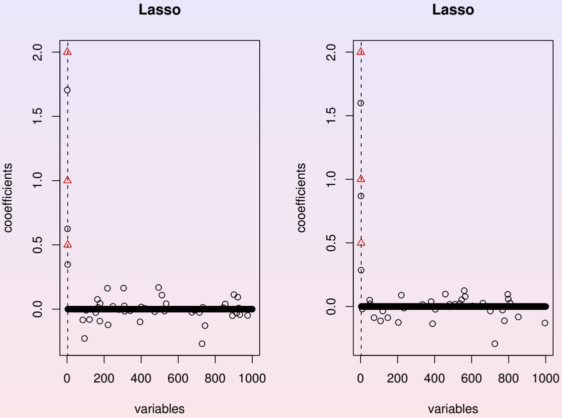
$$\hat{S}(\lambda_{\mathrm{opt}}) \supseteq S_0$$

$\leadsto$ 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

## $s_0 = 3,\ p = 1'000,\ n = 50$; 2 independent realizations

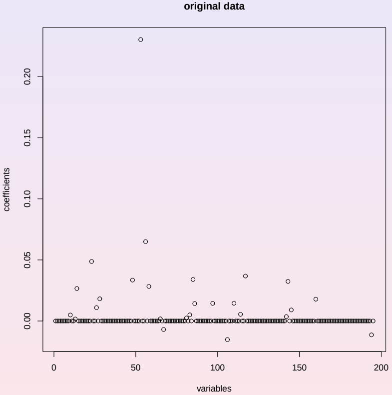



44 selected variables

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

# Variable selection with Lasso

an older formulation:

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)
    - $|S_{true,n}| = |S_{0,n}| = O(n^{\kappa})$ for some $0 < \kappa < 1$ (sparsity)
    - the non-zero $\beta_j$'s are outside the $n^{-1/2}$-range
    - $Y,\ X^{(j)}$'s Gaussian (not crucial)

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

$$\begin{aligned} \mathbb{P}[\hat{S}(\lambda) = S_0] &= 1 - O(\exp(-Cn^{1-\delta}))\ \ (n \to \infty) \\ &\approx 1 \text{ even for relatively small } n \end{aligned}$$

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)

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

active set $S_0 = \{j;\ \beta_j \neq 0\} = \{1, \ldots, s_0\}$ consists of the first $s_0$ variables; partition

$$\Sigma = \begin{pmatrix} \Sigma_{S_0,S_0} & \Sigma_{S_0,S_0^c} \\ \Sigma_{S_0^c,S_0} & \Sigma_{S_0^c,S_0^c} \end{pmatrix}$$

$$\text{irrep. condition}: \quad |\Sigma_{S_0^c,S_0} \Sigma_{S_0,S_0}^{-1} \mathrm{sign}(\beta_1, \ldots, \beta_{s_0})| < 1$$

not easy to get insights when it holds...

## Problem 2: Choice of $\lambda$

for prediction oracle solution

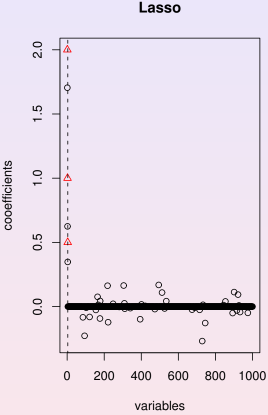
$$\lambda_{\mathrm{opt}} = \mathrm{argmin}_{\lambda} \mathbb{E}[(Y - \sum_{j=1}^{p} \hat{\beta}_j(\lambda) X^{(j)})^2]$$

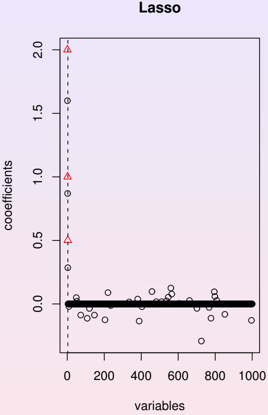
$\mathbb{P}[\hat{S}(\lambda_{\mathrm{opt}}) = S_0] < 1 \ (n \to \infty) \quad (\text{or} = 0 \text{ if } p_n \to \infty \ (n \to \infty))$

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

# recap: variable screening

$s_0 = 3,\ p = 1'000,\ n = 50$; 2 independent realizations




44 selected variables 36 selected variables

$\rightsquigarrow$ want to get rid of the variables with small estimated coefficients

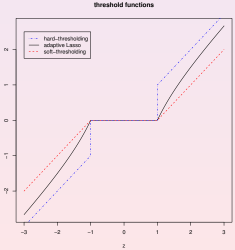
# Adaptive Lasso (Zou, 2006)

re-weighting the penalty function

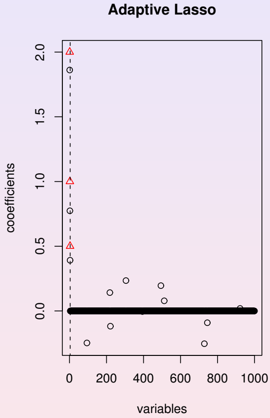
$$\hat{\beta} = \text{argmin}_{\beta}(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{init,j}|}),$$

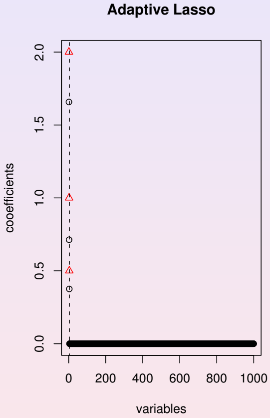
$\hat{\beta}_{init,j}$ from Lasso in first stage $\underbrace{\text{(or OLS if } p < n)}_{\text{Zou (2006)}}$

for orthogonal design,
if $\hat{\beta}_{init}$ = OLS:
Adaptive Lasso = NN-garrote
$\leadsto$ less bias than Lasso



## $s_0 = 3,\ p = 1'000,\ n = 50$
## same 2 independent realizations from before

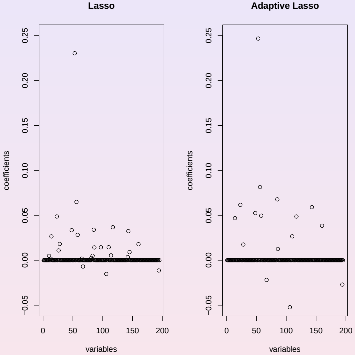



13 selected variables
(44 with Lasso)

3 selected variables
(36 with Lasso)

## Motif regression: $n = 287, \; p = 195$




26 selected variables

16 selected variables

trivial property

$$\hat{\beta}_{init,j} = 0 \Rightarrow \hat{\beta}_j = 0$$

since

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

another motif regression (linear model): $n = 2587$, $p = 666$

| | Lasso | 1-Step | 2-Step |
|---|---|---|---|
| test set squared prediction error | 0.6193 | 0.6230 | 0.6226 |
| number of selected variables | 91 | 42 | 28 |

$\leadsto$ substantially sparser model fit with
twice-iterated adaptive Lasso (three-stage procedure)

Relaxed Lasso (Meinshausen, 2007)

similar in spirit to the adaptive Lasso; and similar in performance