# Recap

## Lasso for orthonormal design: explicit solution

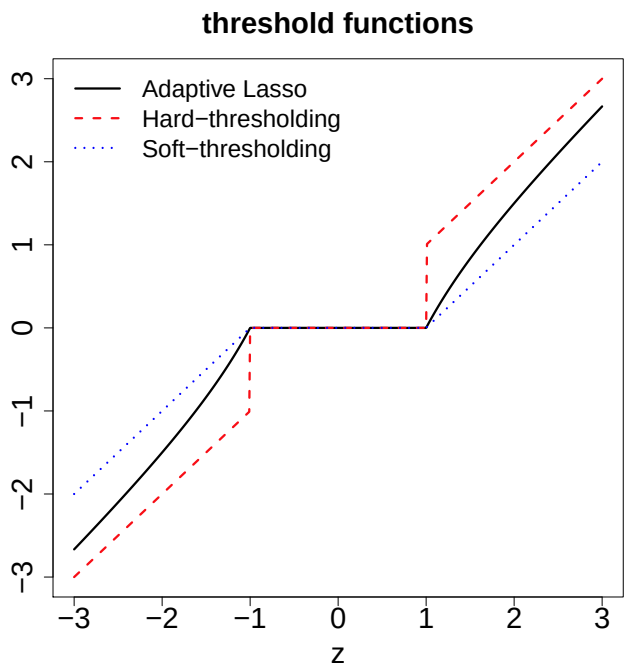
$$X^T X/n = I_{p\times p} \quad \text{(implying that } p \le n\text{)}$$

Lasso = soft-thresholding of ordinary least squares

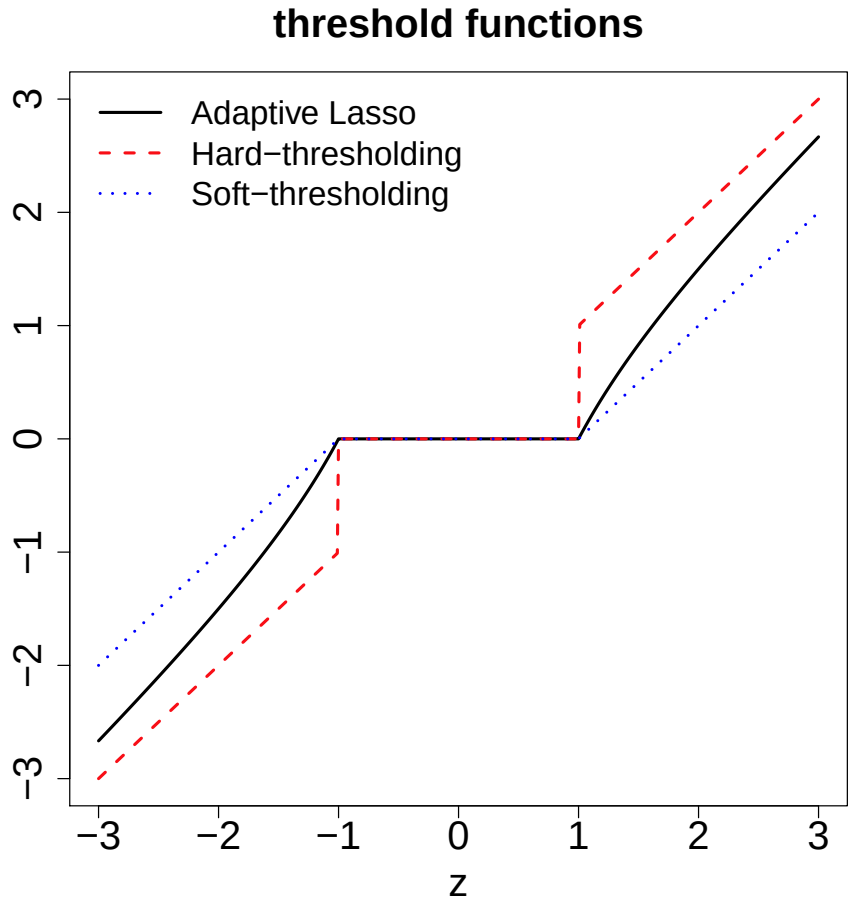
*Proposition 1*

Assume orthogonal design $X^T X/n = I$. Then,

$$\hat{\beta}_j(\lambda) = g_{\lambda/2}(Z_j), \quad Z_j = (X^T Y)_j/n = \hat{\beta}_{\mathrm{OLS},j},$$
$$g_\lambda(z) = \mathrm{sign}(z)(|z| - \lambda)_+$$

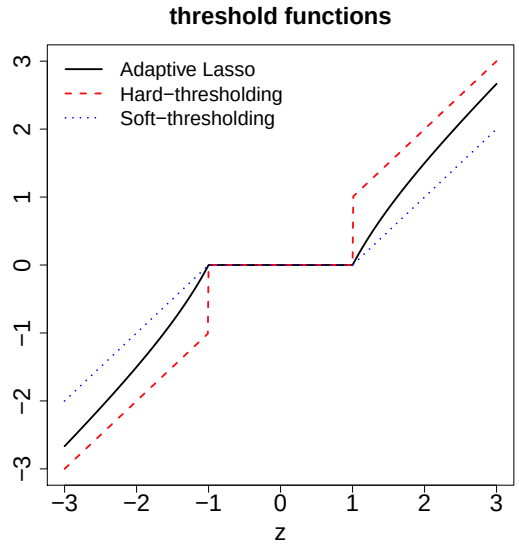


# Lasso for orthonormal design: explicit solution




$\rightsquigarrow$ Lasso (blue dashed line) exhibits bias!

(Note that OLS is unbiased)




Hard-thresholding:

*Proposition 2*

Assume orthonormal design $X^T X/n = I$. Then,

$$\hat{\beta}_{\ell_0}(\lambda) = \text{argmin}_\beta \left( \|Y - X\beta\|_2^2/n + \lambda \|\beta\|_0 \right)$$

equals hard-thresholding with threshold value $\sqrt{\lambda}$, that is

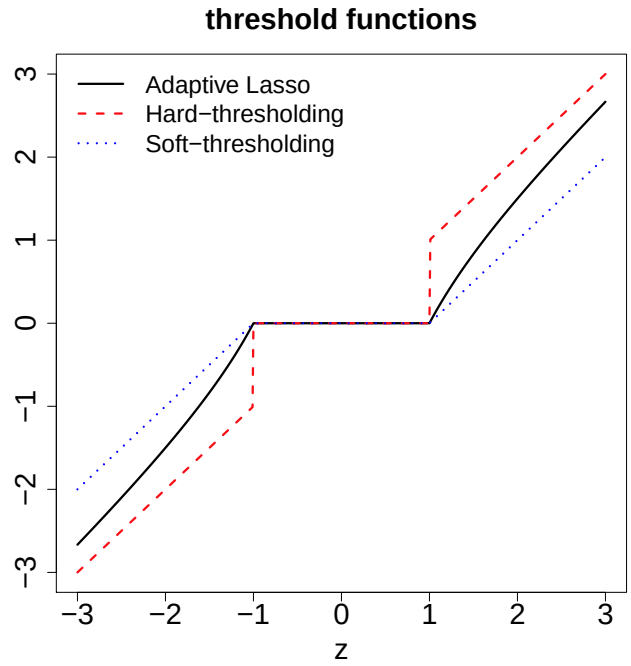
$$\hat{\beta}_{\ell_0;j}(\lambda) = Z_j I(|Z_j| > \sqrt{\lambda}), \; Z = X^T Y/n$$

for example: AIC, BIC are $\ell_0$-norm penalized regression

$\rightsquigarrow$ non-convex difficult optimization (but can use recent progress on mixed-integer programming to deal with $p \leq 500$)

- hard-thresholding exhibits less bias than Lasso (but still $\mathbb{E}[\hat{\beta}_{\ell_0}] \neq \beta^0$)

  but it is hard to compute
- adaptive Lasso (black line in the plot) has also less bias than Lasso but can be computed with two Lasso fits, see later



# II.4. Prediction with the Lasso

goal: estimation of the regression function

$$f(x) = \mathbb{E}[Y|X = x] = \sum_{j=1}^{p} \beta_j^0 x_j = (\beta^0)^T x$$

## II.4.1. Practical aspects

use $\hat{f}(x) = \hat{\beta}(\lambda)^T x$
and choose $\lambda$ via cross-validation

also e.g. in connection with deep neural networks: the prediction from the last layer to $Y$ is based on regularized linear models

- last layer features are $\phi(X_i) \in \mathbb{R}^d$
- Lasso:

$$\hat{\beta}(\lambda) = \text{argmin}_\beta \| Y - (\phi(X_1), \ldots, \phi(X_n))^T \beta \|_2^2 / n + \lambda \|\beta\|_1$$

- prediction $\hat{f}(x) = \hat{\beta}(\lambda)^T \phi(x)$

# How to measure prediction quality?

CV test error

from a theory point of view, we also look at

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n$$

for fixed design:

$$\mathbb{E}[\|X(\hat{\beta} - \beta^0)\|_2^2/n] = \mathbb{E}\|Y_{\text{new}}(X) - X\hat{\beta}\|_2^2/n - \sigma_\varepsilon^2$$
$$Y_{\text{new}} = X\beta^0 + \varepsilon_{\text{new}}$$

that is: expected value of in-sample point prediction accuracy

# Announcement of a few asymptotic results

...

Developing the theory for announced results

Corollary 6.1. in Bühlmann and van de Geer (2011)

assume:

- $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$
- scaled columns $\hat{\sigma}_j^2 \equiv 1 \; \forall \; j$

For

$$\lambda = 4\hat{\sigma}\sqrt{\frac{t^2 + 2\log(p)}{n}}$$

where $\hat{\sigma}$ is an estimator for $\sigma$. Then, with probability at least $1 - \alpha$ where

$$\alpha = 2\exp(-t^2/2) + \mathbb{P}[\hat{\sigma} < \sigma]$$

we have that

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n \leq \frac{3}{2}\lambda\|\beta^0\|_1$$

## Implications and Asymptotic viewpoint

the proper $\lambda \asymp \sqrt{\log(p)/n}$ (take e.g. $t^2 \asymp \log(p)$)

Corollary 6.1 implies:

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P(\underbrace{\lambda}_{\asymp\sqrt{\log(p)/n}} \|\beta^0\|_1) = O_P(\sqrt{\log(p)/n}\|\beta^0\|_1)$$

even for very sparse case with $\|\beta^0\|_1 = O(1)$:
slow convergence rate of order $O_P(\sqrt{\log(p)/n})$

benchmark: OLS oracle on the variables from $S_0 = \{j;\ \beta_j^0 \neq 0\}$

$$\|X(\hat{\beta}_{\mathrm{OLS-oracle}} - \beta^0)\|_2^2/n = O_P(s_0/n), \quad s_0 = |S_0|$$

we will later derive for the Lasso, under additional assumptions on $X$: fast convergence rate

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P(\log(p)\frac{s_0}{n}) \quad \text{(if } \phi_0^2 \text{ bounded away from zero)}$$

for slow rate: no assumptions on $X$ (could have perfectly correlated columns)