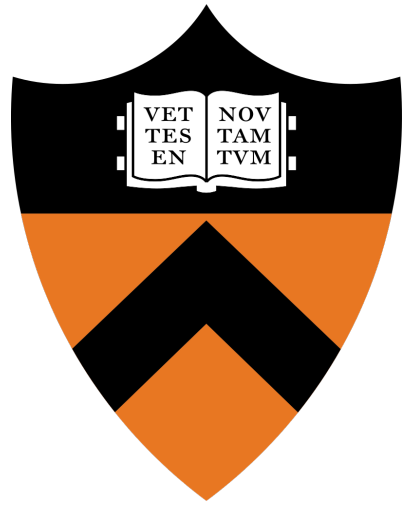
Model Selection and Lasso



Yuxin Chen

Princeton University, Fall 2020

Outline

- Model selection
- Lasso estimator
- Risk inflation
- Minimax risk for sparse vectors

Asymptotic notation

- $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}$$

- $f(n) \gtrsim g(n)$ or $f(n) = \Omega(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \geq \text{const}$$

- $f(n) \asymp g(n)$ or $f(n) = \Theta(g(n))$ means

$$\text{const}_1 \leq \lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \leq \text{const}_2$$

- $f(n) = o(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = 0$$

Model selection

All models are wrong but some are useful.

— George Box

Basic linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta},$$

- $\boldsymbol{y} = [y_1, \cdots, y_n]^\top \in \mathbb{R}^n$: observed data / response variables
- $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]^\top$: design matrix / feature matrix (known)
  - assumed to be full rank
- $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_p]^\top \in \mathbb{R}^p$: unknown signal / regression coefficients
- $\boldsymbol{\eta} = [\eta_1, \cdots, \eta_n]^\top \in \mathbb{R}^n$: noise

Throughout this lecture, we assume Guassian noise

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

Model selection / feature selection

Regression:

- find relationship between response y_i and explanatory variables $\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,p}$
- use the fitted model to make prediction

Question: which $\underbrace{\text{(sub)-set of variables / features}}_{\text{model}}$ should we include?

- Myth: nothing is lost by including every feature / variable available
- Paradoxically, we can often achieve better predictions by discarding a fraction of variables

Tradeoff

- Model too small $\Longrightarrow$ large bias (underfitting)
- Model too large $\Longrightarrow$ large variance and poor prediction (overfitting)

How to achieve a desired tradeoff between predictive accuracy and parsimony (model complexity)?

Underfitting

Recall that the least squares (LS) estimate is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

- Divide the design matrix into 2 parts: $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}]$
- $\tilde{\boldsymbol{x}} = \begin{bmatrix} \tilde{\boldsymbol{x}}^{(1)} \\ \tilde{\boldsymbol{x}}^{(2)} \end{bmatrix}$: new data
- LS estimate based only on $\boldsymbol{X}^{(1)}$:

  $$\hat{\boldsymbol{\beta}}^{(1)} := (\boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(1)})^{-1} \boldsymbol{X}^{(1)\top} \boldsymbol{y}$$

  with prediction at $\tilde{\boldsymbol{x}}$ given by

  $$\hat{y}_{\text{underfit}} = \tilde{\boldsymbol{x}}^{(1)\top} \hat{\boldsymbol{\beta}}^{(1)}$$

- LS estimate based on true model

  $$\hat{\boldsymbol{\beta}} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

  with prediction at $\tilde{\boldsymbol{x}}$ given by

  $$\hat{y}_{\text{true}} = [\tilde{\boldsymbol{x}}^{(1)\top}, \tilde{\boldsymbol{x}}^{(2)\top}] \hat{\boldsymbol{\beta}}$$

Bias due to underfitting

Suppose the ground truth is $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{bmatrix}$, then

$$\begin{aligned} \mathbb{E}\left[\hat{\boldsymbol{\beta}}^{(1)}\right] &= \left(\boldsymbol{X}^{(1)\top}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)\top}\left(\boldsymbol{X}^{(1)}\boldsymbol{\beta}^{(1)} + \boldsymbol{X}^{(2)}\boldsymbol{\beta}^{(2)}\right) \\ &= \boldsymbol{\beta}^{(1)} + \underbrace{\left(\boldsymbol{X}^{(1)\top}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)\top}\boldsymbol{X}^{(2)}\boldsymbol{\beta}^{(2)}}_{\text{bias}} \end{aligned}$$

$\Longrightarrow$ $\hat{\boldsymbol{\beta}}^{(1)}$ is a biased estimate of $\boldsymbol{\beta}^{(1)}$

Prediction variance due to underfitting

Fact 8.1

$$\mathbf{Var}\left[\hat{y}_{\text{true}}\right] \geq \mathbf{Var}\left[\hat{y}_{\text{underfit}}\right]$$

- **Implications:** the "apparent" prediction variance tends to decrease when we adopt small models
- (Exercise): compute the prediction variance under overfitting

Proof of Fact 8.1

Observe that

$$\mathbf{Cov}[\hat{\boldsymbol{\beta}}] = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \mathbf{Cov}\left[\boldsymbol{y}\right] \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$$

$$= \sigma^2 \begin{bmatrix} (\boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(1)})^{-1} + \boldsymbol{L}\boldsymbol{M}\boldsymbol{L}^\top & -\boldsymbol{L}\boldsymbol{M} \\ -\boldsymbol{M}\boldsymbol{L}^\top & \boldsymbol{M} \end{bmatrix} \text{ (matrix inversion identity)}$$

where $\boldsymbol{L} = \left(\boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(1)}\right)^{-1} \boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(2)}$ and

$$\boldsymbol{M} = \left\{ \boldsymbol{X}^{(2)\top} \left(\boldsymbol{I} - \boldsymbol{X}^{(1)} \left(\boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(1)}\right)^{-1} \boldsymbol{X}^{(1)\top} \right) \boldsymbol{X}^{(2)} \right\}^{-1} \succeq \mathbf{0}.$$

This gives

$$\mathbf{Var}\left[\hat{y}_{\text{true}}\right] = \left[\tilde{\boldsymbol{x}}^{(1)\top}, \tilde{\boldsymbol{x}}^{(2)\top}\right] \mathbf{Cov}\left[\hat{\boldsymbol{\beta}}\right] \begin{bmatrix} \tilde{\boldsymbol{x}}^{(1)} \\ \tilde{\boldsymbol{x}}^{(2)} \end{bmatrix}$$

$$= \sigma^2 \tilde{\boldsymbol{x}}^{(1)\top} \left(\boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(1)}\right)^{-1} \tilde{\boldsymbol{x}}^{(1)} + \sigma^2 \left(\boldsymbol{L}^\top \tilde{\boldsymbol{x}}^{(1)} - \tilde{\boldsymbol{x}}^{(2)}\right)^\top \boldsymbol{M} \left(\boldsymbol{L}^\top \tilde{\boldsymbol{x}}^{(1)} - \tilde{\boldsymbol{x}}^{(2)}\right)$$

$$\geq \sigma^2 \tilde{\boldsymbol{x}}^{(1)\top} \left(\boldsymbol{X}^{(1)\top} \boldsymbol{X}^{(1)}\right)^{-1} \tilde{\boldsymbol{x}}^{(1)} = \mathbf{Var}\left[\hat{y}_{\text{underfit}}\right]$$

Model selection criteria

Choosing a subset of explanatory variables might improve prediction

Question: which subset shall we select?

One strategy

(1) pick a criterion that measures how well a model performs

(2) evaluate the criterion for each subset and pick the best

One popular choice: choose a model that predicts well

Prediction error and model error

- training set: $\boldsymbol{y}, \boldsymbol{X}$
- $\hat{\boldsymbol{\beta}}$: an estimate based on training set
- new data: $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\eta}} \in \mathbb{R}^m$, where $\tilde{\boldsymbol{\eta}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$
- Goal: use $\hat{\boldsymbol{\beta}}$ to predict $\tilde{\boldsymbol{y}}$

One may assess the quality of the estimate based on its *prediction error* on $\tilde{\boldsymbol{y}}$, i.e.

$$\begin{aligned}
\text{PE} &:= \mathbb{E}\left[\left\|\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{y}}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\tilde{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\|^2\right] + 2\mathbb{E}\left[(\tilde{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^\top(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})\right] + \mathbb{E}\left[\left\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right\|^2\right] \\
&= \underbrace{\mathbb{E}\left[\left\|\tilde{\boldsymbol{X}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\|^2\right]}_{:=\text{ME (model error)}} + \underbrace{m\sigma^2}_{\text{variability of data}}
\end{aligned}$$

Residual sum of squares (RSS)

We shall set $\tilde{\boldsymbol{X}} = \boldsymbol{X}$ (and hence $m = n$) out of simplicity

- the case where the structures of new and old data are the same

Unfortunately, we do not have access to PE (as we don't know $\boldsymbol{\beta}$)

$\Longrightarrow$ need an operational criterion for estimating PE

- One candidate: estimate PE via residual sum of squares

$$\text{RSS} := \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$$

$\Longrightarrow$ training error

Training error underestimates prediction error

Suppose $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\Pi}\boldsymbol{y}$ for some given $\boldsymbol{\Pi}$ with $\mathrm{Tr}(\boldsymbol{\Pi}) > 0$ (e.g. LS), then

$$\mathsf{PE} = \mathbb{E}[\mathsf{RSS}] + 2\sigma^2 \mathrm{Tr}(\boldsymbol{\Pi}) > \mathbb{E}[\mathsf{RSS}] \tag{8.1}$$

Proof:

$$\begin{aligned}
\mathsf{PE} - \mathbb{E}[\mathsf{RSS}] &= \mathbb{E}\left[\|\tilde{\boldsymbol{y}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\right] - \mathbb{E}\left[\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2\right] \\
&= \mathbb{E}\left[\|\tilde{\boldsymbol{y}}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 - 2\langle \tilde{\boldsymbol{y}}, \boldsymbol{X}\hat{\boldsymbol{\beta}}\rangle\right] \\
&\quad - \mathbb{E}\left[\|\boldsymbol{y}\|^2 + \|\boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 - 2\left\langle \boldsymbol{y}, \boldsymbol{X}\hat{\boldsymbol{\beta}}\right\rangle\right] \\
&= 2\mathbb{E}\left[\langle \boldsymbol{y} - \tilde{\boldsymbol{y}}, \boldsymbol{X}\hat{\boldsymbol{\beta}}\rangle\right] = 2\mathbb{E}\left[\langle \boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}, \boldsymbol{\Pi}\boldsymbol{y}\rangle\right] \\
&= 2\mathbb{E}\left[\langle \boldsymbol{\eta}, \boldsymbol{\Pi}\boldsymbol{\eta}\rangle\right] \overset{\text{(a)}}{=} 2\mathrm{Tr}\left(\boldsymbol{\Pi}\mathbb{E}\left[\boldsymbol{\eta}\boldsymbol{\eta}^\top\right]\right) \\
&= 2\sigma^2 \mathrm{Tr}(\boldsymbol{\Pi}),
\end{aligned}$$

where (a) follows from the identity $\mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \mathrm{Tr}(\boldsymbol{B}\boldsymbol{A}^\top)$.

Example: least squares (LS) estimator

The least squares solution is

$$\hat{\boldsymbol{\beta}}^{\mathsf{ls}} := \arg\min_{\hat{\boldsymbol{\beta}}} \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$$

The fitted values $\hat{\boldsymbol{y}}^{\mathsf{ls}}$ is given by

$$\hat{\boldsymbol{y}}^{\mathsf{ls}} = \boldsymbol{\Pi}^{\mathsf{ls}}\boldsymbol{y} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}.$$

$$\implies \quad \mathsf{PE} = \mathbb{E}[\mathsf{RSS}] + 2\sigma^2 \mathrm{Tr}(\boldsymbol{\Pi}^{\mathsf{ls}}) = \mathbb{E}[\mathsf{RSS}] + 2\sigma^2 p$$

LS estimator for a given model

Suppose the model (i.e. support of $\boldsymbol{\beta}$) is $S \subseteq \{1, \cdots, p\}$. The least squares solution given S is

$$\hat{\boldsymbol{\beta}}_S := \arg\min_{\hat{\boldsymbol{\beta}}} \{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 : \hat{\beta}_i = 0 \text{ for all } i \notin S\}$$

The fitted values $\hat{\boldsymbol{y}}$ is then given by

$$\hat{\boldsymbol{y}} = \boldsymbol{\Pi}_S \boldsymbol{y} := \boldsymbol{X}_S (\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1} \boldsymbol{X}_S^\top \boldsymbol{y},$$

where $\boldsymbol{X}_S$ is formed by the columns of $\boldsymbol{X}$ at indices in S.

Mallows' C_p statistic

In view of (8.1),

$$\mathsf{PE}(\hat{\boldsymbol{\beta}}_S) = \mathbb{E}[\mathsf{RSS}(\hat{\boldsymbol{\beta}}_S)] + 2\sigma^2 \mathrm{Tr}(\boldsymbol{\Pi}_S) = \mathbb{E}[\mathsf{RSS}(\hat{\boldsymbol{\beta}}_S)] + 2|S|\sigma^2,$$

since

$$\mathrm{Tr}(\boldsymbol{\Pi}_S) = \mathrm{Tr}(\boldsymbol{X}_S(\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top) = \mathrm{Tr}((\boldsymbol{X}_S^\top \boldsymbol{X}_S)^{-1}\boldsymbol{X}_S^\top \boldsymbol{X}_S) = |S|.$$

Definition 8.2 (C_p statistic, Mallows '73)

$$C_p(S) = \underbrace{\mathsf{RSS}(\hat{\boldsymbol{\beta}}_S)}_{\text{training error}} + \underbrace{2\sigma^2|S|}_{\text{model complexity}}$$

C_p is an unbiased estimate of prediction error

Model selection based on C_p statistic

1. Compute $C_p(S)$ for each model S
2. Choose $S^* = \arg\min_S C_p(S)$

This is essentially an ℓ_0-regularized least-squares problem

$$\text{minimize}_{\hat{\boldsymbol{\beta}}} \quad \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + \underbrace{2\sigma^2\|\hat{\boldsymbol{\beta}}\|_0}_{\text{penalized by model complexity}} \tag{8.2}$$

Example: orthogonal design

Suppose $\boldsymbol{X} = \boldsymbol{I}$, then (8.8) reduces to

$$\text{minimize}_{\hat{\boldsymbol{\beta}}} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\beta}_i)^2 + 2\sigma^2 \mathbf{1}\{\hat{\beta}_i \neq 0\}$$

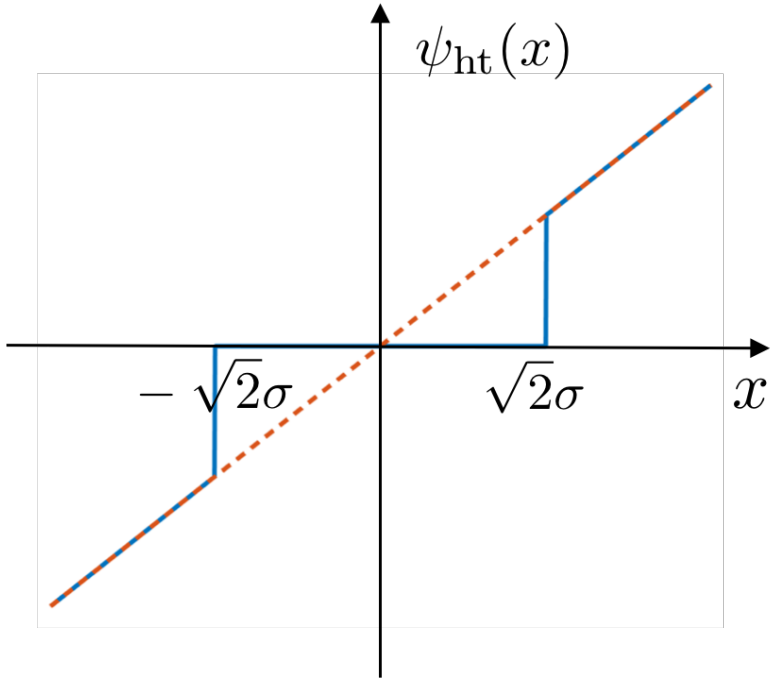
Solving this problem gives

$$\hat{\beta}_i = \begin{cases} 0, & |y_i| \leq \sqrt{2}\sigma \\ y_i, & |y_i| > \sqrt{2}\sigma \end{cases}$$

hard thresholding

- Keep large coefficients; discard small coefficients

Example: orthogonal design



$$\hat{\beta}_i = \psi_{\mathsf{ht}}(y_i; \sqrt{2}\sigma) := \begin{cases} 0, & |y_i| \leq \sqrt{2}\sigma \\ y_i, & |y_i| > \sqrt{2}\sigma \end{cases}$$

hard thresholding

Hard thresholding preserves data outside threshold zone

Lasso estimator

Convex relaxation: Lasso (Tibshirani '96)

Lasso (Least absolute shrinkage and selection operator)

$$\text{minimize}_{\hat{\boldsymbol{\beta}}} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\hat{\boldsymbol{\beta}}\|_1 \tag{8.3}$$

for some regularization parameter $\lambda > 0$

- It is equivalent to

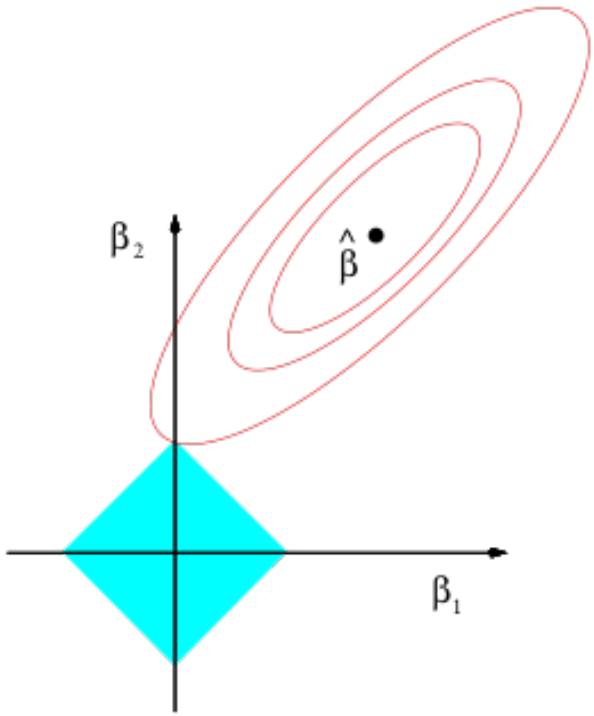
$$\begin{aligned} \text{minimize}_{\hat{\boldsymbol{\beta}}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 \\ \text{s.t.} \quad & \|\hat{\boldsymbol{\beta}}\| \leq t \end{aligned}$$

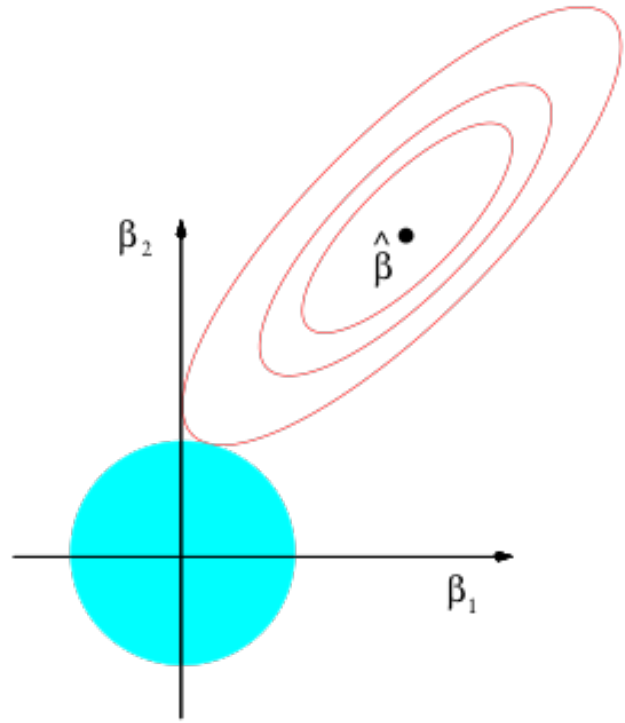
  for some t that depends on λ
  - a quadratic program (QP) with convex constraints

- λ controls model complexity: larger λ restricts the parameters more; smaller λ frees up more parameters

Lasso vs. MMSE (or ridge regression)

$\hat{\boldsymbol{\beta}}$: least squares solution





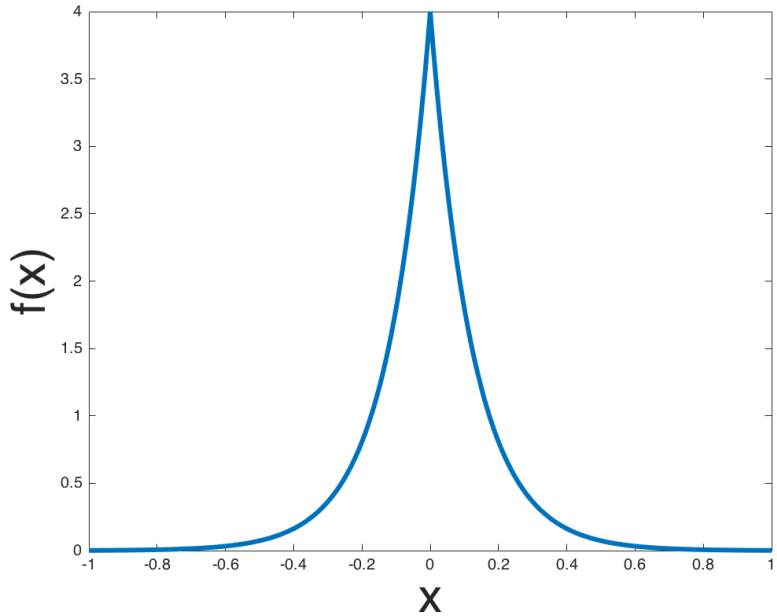
$$\begin{aligned} \text{minimize}_{\boldsymbol{\beta}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_1 \leq t \end{aligned}$$

$$\begin{aligned} \text{minimize}_{\boldsymbol{\beta}} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_2 \leq t \end{aligned}$$

Fig. credit: Hastie, Tibshirani, & Wainwright

A Bayesian interpretation

Orthogonal design: $\boldsymbol{y} = \boldsymbol{\beta} + \boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$.



Impose an i.i.d. prior on β_i to encourage sparsity *(Gaussian is not a good choice)*:

(Laplacian prior) $$\mathbb{P}(\beta_i = z) = \frac{\lambda}{2} e^{-\lambda |z|}$$

A Bayesian interpretation of Lasso

Posterior of $\boldsymbol{\beta}$:

$$\mathbb{P}\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) \quad \propto \quad \mathbb{P}(\boldsymbol{y}|\boldsymbol{\beta})\mathbb{P}(\boldsymbol{\beta}) \quad \propto \quad \prod_{i=1}^{n} e^{-\frac{(y_i-\beta_i)^2}{2\sigma^2}} \frac{\lambda}{2} e^{-\lambda|\beta_i|}$$

$$\propto \quad \prod_{i=1}^{n} \exp\left\{-\frac{(y_i-\beta_i)^2}{2\sigma^2} - \lambda|\beta_i|\right\}$$

$\Longrightarrow$ maximum *a posteriori* (MAP) estimator:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{\frac{(y_i-\beta_i)^2}{2\sigma^2} + \lambda|\beta_i|\right\} \quad \text{(Lasso)}$$

Implication: Lasso is MAP estimator under Laplacian prior

Example: orthogonal design

Suppose $\boldsymbol{X} = \boldsymbol{I}$, then Lasso reduces to

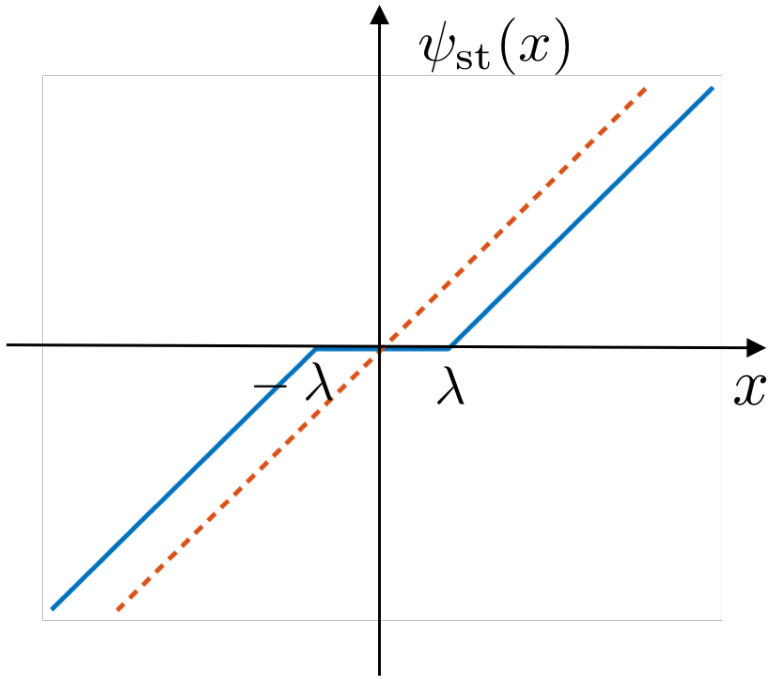
$$\text{minimize}_{\hat{\boldsymbol{\beta}}} \quad \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\beta}_i)^2 + \lambda|\hat{\beta}_i|$$

The Lasso estimate $\hat{\boldsymbol{\beta}}$ is then given by

$$\hat{\beta}_i = \begin{cases} y_i - \lambda, & y_i \geq \lambda \\ y_i + \lambda, & y_i \leq -\lambda \\ 0, & \text{else} \end{cases}$$

soft thresholding

Example: orthogonal design



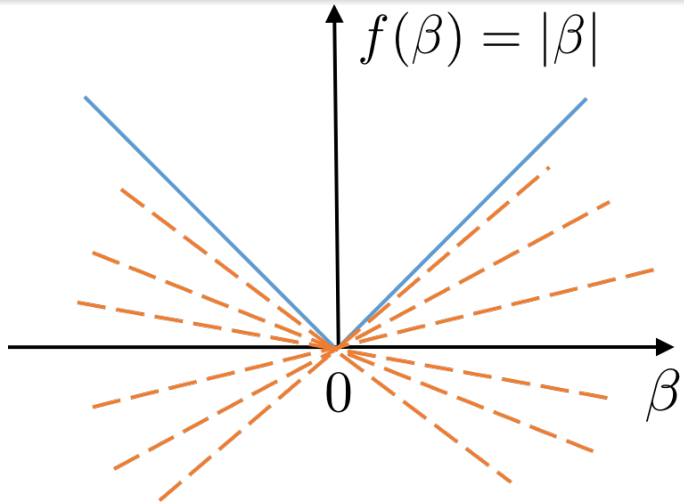
$$\hat{\beta}_i = \psi_{\text{st}}(y_i; \lambda) = \begin{cases} y_i - \lambda, & y_i \geq \lambda \\ y_i + \lambda, & y_i \leq -\lambda \\ 0, & \text{else} \end{cases}$$

soft thresholding

Soft thresholding shrinks data towards 0 outside threshold zone

Optimality condition for convex functions

For any convex function $f(\boldsymbol{\beta})$, $\boldsymbol{\beta}^*$ is an optimal solution iff $\mathbf{0} \in \partial f(\boldsymbol{\beta}^*)$, where $\partial f(\boldsymbol{\beta})$ is the set of all subgradients at $\boldsymbol{\beta}$



- s is a subgradient of $f(\beta) = |\beta|$ if

$$\begin{cases} s = \text{sign}(\beta), & \text{if } \beta \neq 0 \\ s \in [-1, 1], & \text{if } \beta = 0 \end{cases} \tag{8.4}$$

Optimality condition for convex functions

For any convex function $f(\boldsymbol{\beta})$, $\boldsymbol{\beta}^*$ is an optimal solution iff $\mathbf{0} \in \partial f(\boldsymbol{\beta}^*)$, where $\partial f(\boldsymbol{\beta})$ is the set of all subgradients at $\boldsymbol{\beta}$

- The subgradient of $f(\beta) = \frac{1}{2}(y - \beta)^2 + \lambda|\beta|$ can be written as

$$g = \beta - y + \lambda s \quad \text{with } s \text{ defined in } (8.4)$$

- We see that $\hat{\beta} = \psi_{\text{st}}(y; \lambda)$ by checking optimality conditions for two cases:
  - If $|y| \leq \lambda$, taking $\beta = 0$ and $s = y/\lambda$ gives $g = 0$
  - If $|y| > \lambda$, taking $\beta = y - \text{sign}(y)\lambda$ gives $g = 0$

Single-parameter setup

Consider the case where there is only a single parameter $\hat{\beta} \in \mathbb{R}$:

$$\text{minimize}_{\hat{\beta} \in \mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{y} - \hat{\beta}\boldsymbol{z}\|_2^2 + \lambda|\hat{\beta}|.$$

Then one can verify that (homework)

$$\hat{\beta} = \psi_{\text{st}}\left(\frac{\boldsymbol{z}^\top \boldsymbol{y}}{\|\boldsymbol{z}\|_2^2}; \frac{\lambda}{\|\boldsymbol{z}\|_2^2}\right) = \begin{cases} \frac{\boldsymbol{z}^\top \boldsymbol{y}}{\|\boldsymbol{z}\|_2^2} - \frac{\lambda}{\|\boldsymbol{z}\|_2^2}, & \text{if } \boldsymbol{z}^\top \boldsymbol{y} > \lambda \\ 0, & \text{if } |\boldsymbol{z}^\top \boldsymbol{y}| \leq \lambda \\ \frac{\boldsymbol{z}^\top \boldsymbol{y}}{\|\boldsymbol{z}\|_2^2} + \frac{\lambda}{\|\boldsymbol{z}\|_2^2}, & \text{else} \end{cases}$$

Algorithm: coordinate descent

Idea: repeatedly cycle through the variables and, in each step, optimize only a single variable

- When updating $\hat{\beta}_j$, we solve

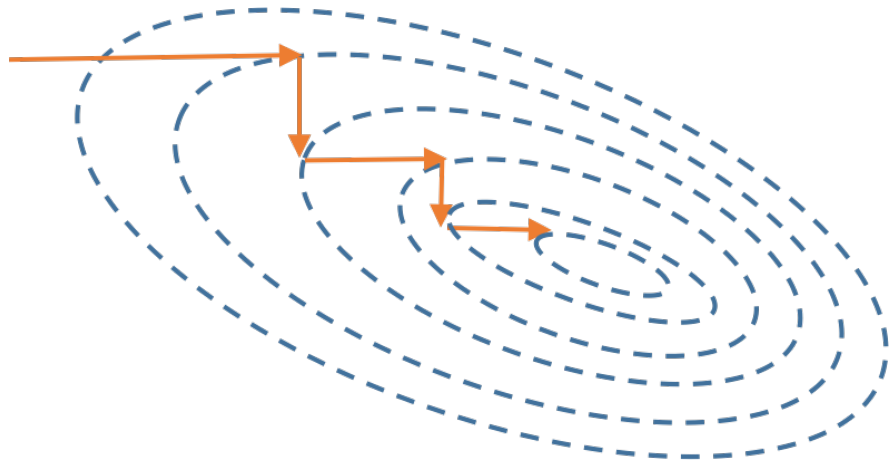
$$\text{minimize}_{\hat{\beta}_j \in \mathbb{R}} \ \frac{1}{2}\Big\|\boldsymbol{y} - \sum_{i:i\neq j} \boldsymbol{X}_{:i}\hat{\beta}_i - \boldsymbol{X}_{:,j}\hat{\beta}_j\Big\|_2^2 + \lambda|\hat{\beta}_j| + \lambda \sum_{i:i\neq j} |\hat{\beta}_i|$$

  where $\boldsymbol{X}_{:,j}$ is jth column of $\boldsymbol{X}$

- This is exactly the single-parameter setting, and hence

$$\hat{\beta}_j \ \leftarrow \ \psi_{\text{st}}\left(\frac{\boldsymbol{X}_{:,j}^\top\left(\boldsymbol{y} - \sum_{i:i\neq j} \boldsymbol{X}_{:i}\hat{\beta}_i\right)}{\|\boldsymbol{X}_{:j}\|^2}; \ \frac{\lambda}{\|\boldsymbol{X}_{:j}\|^2}\right)$$

Algorithm: coordinate descent



Algorithm 8.1 Coordinate descent for Lasso

Repeat until convergence

for $j = 1, \cdots, n$**:**

$$\hat{\beta}_j \leftarrow \psi_{\text{st}}\left(\frac{\boldsymbol{X}_{:,j}^{\top}\left(\boldsymbol{y} - \sum_{i:i\neq j} \boldsymbol{X}_{:i}\hat{\beta}_i\right)}{\|\boldsymbol{X}_{:j}\|_2^2}; \frac{\lambda}{\|\boldsymbol{X}_{:j}\|_2^2}\right) \tag{8.5}$$

Risk inflation

Ideal risk: orthogonal design

$$y_i = \beta_i + \eta_i, \qquad i = 1, \cdots, n$$

Let's first select / fix a model and then estimate: for a fixed model $S \subseteq \{1, \cdots, n\}$, the LS estimate $\hat{\boldsymbol{\beta}}_S$ is

$$(\hat{\boldsymbol{\beta}}_S)_i = \begin{cases} y_i, & \text{if } i \in S \\ 0, & \text{else} \end{cases}$$

- **Mean square estimation error for a fixed model S:**

$$\mathsf{MSE}(\hat{\boldsymbol{\beta}}_S, \boldsymbol{\beta}) := \mathbb{E}[\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}\|^2] = \sum_{i \in S} \mathbb{E}[(y_i - \beta_i)^2] + \sum_{i \notin S} \beta_i^2$$

$$= \underbrace{|S|\sigma^2}_{\text{variance due to noise}} + \underbrace{\sum_{i \notin S} \beta_i^2}_{\text{bias (since we don't estimate all coefficients)}}$$

Ideal risk: orthogonal design

- **Smallest MSE for a fixed model size:** If we fix the model size $|S| = k$, then the best model achieves

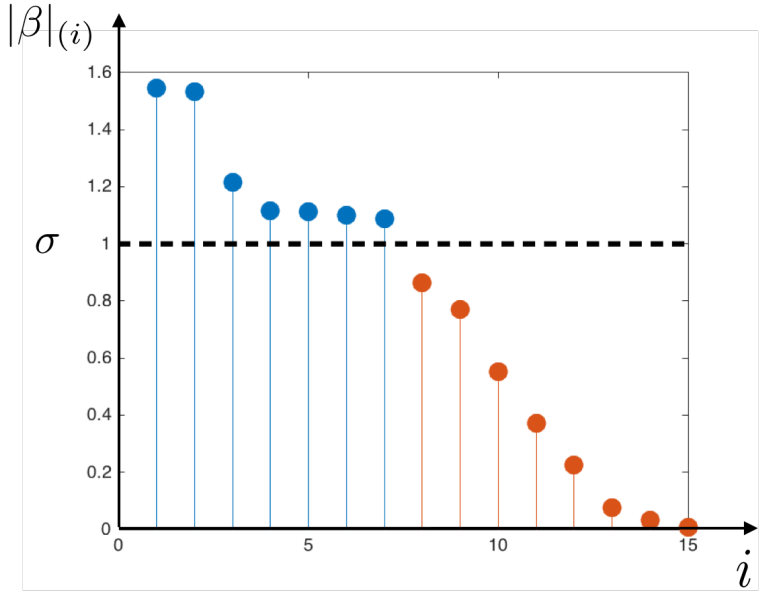
$$\mathsf{MSE}_k(\boldsymbol{\beta}) := \min_{S:|S|=k} \mathsf{MSE}(\hat{\boldsymbol{\beta}}_S, \boldsymbol{\beta}) = k\sigma^2 + \min_{S:|S|=k} \sum_{i \notin S} \beta_i^2$$

$$= k\sigma^2 + \sum_{i=k+1}^{n} |\beta|_{(i)}^2$$

where $|\beta|_{(1)} \geq |\beta|_{(2)} \geq \cdots \geq |\beta|_{(n)}$ are order statistics of $\{|\beta_i|\}$

Implication: good estimation is possible when $\boldsymbol{\beta}$ compresses well

Optimizing the risk over all possible models



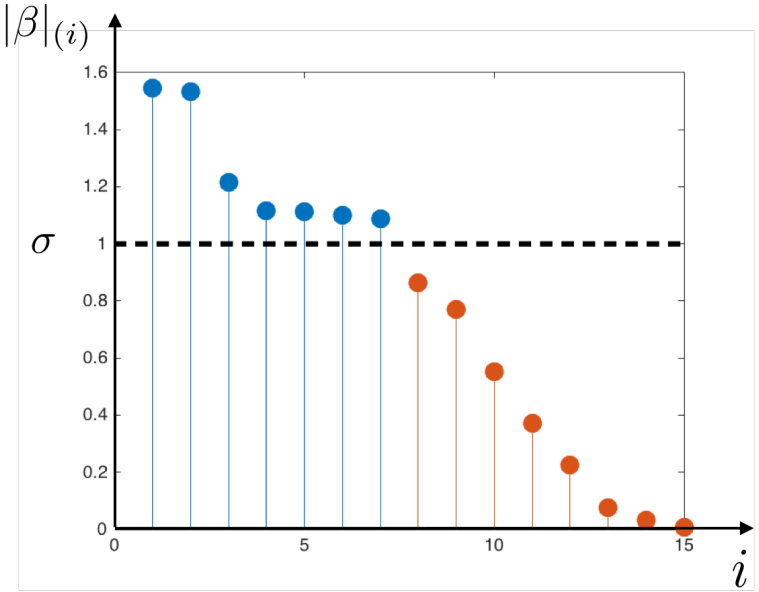
- **Ideal risk (smallest MSE over all models):** minimizing over all possible model size k gives

$$\mathsf{MSE}^{\mathsf{ideal}}(\boldsymbol{\beta}) := \min_k \min_{S:|S|=k} \mathsf{MSE}(\hat{\boldsymbol{\beta}}_S, \boldsymbol{\beta}) = \min_k \left\{ k\sigma^2 + \sum_{i=k+1}^{n} |\beta|^2_{(i)} \right\}$$

$$= \sum_{i=1}^{n} \min\{\sigma^2, \beta_i^2\}$$

Oracle lower bound

$$\mathrm{MSE}^{\mathrm{ideal}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \min\{\sigma^2, \beta_i^2\}$$



- β_i is worth estimating iff $|\beta_i| > \sigma^2$
- $\mathrm{MSE}^{\mathrm{ideal}}$ is the optimal risk *if an oracle reveals which variables are worth estimating and which can be safely ignored*
- With the oracle information, one can achieve $\mathrm{MSE}^{\mathrm{ideal}}$ via

$$\hat{\beta}_i^{\mathrm{ideal}} = \begin{cases} y_i, & \text{if } |\beta_i| > \sigma, \\ 0 & \text{else} \end{cases}$$

(eliminate irrelevant variables)

Risk inflation

- **Problem:** unfortunately, we do NOT know which model S is the best and hence cannot attain $\text{MSE}^{\text{ideal}}$ ...

- Instead, we shall treat it as an oracle lower bound, and consider the increase in estimation error due to selecting rather than knowing the correct model

Risk inflation

Definition 8.3 (risk inflation, Foster & George '94)

The risk inflation of an estimator $\hat{\boldsymbol{\beta}}$ is

$$\mathsf{RI}(\hat{\boldsymbol{\beta}}) = \sup_{\boldsymbol{\beta}} \frac{\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})}{\mathsf{MSE}^{\mathsf{ideal}}(\boldsymbol{\beta})},$$

where $\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) := \mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2]$.

- Idea: calibrate the actual risk against the ideal risk for each $\boldsymbol{\beta}$ to better reflect the potential gains / loss
- Suggestion: *find a procedure that achieves low risk inflation!*

Risk inflation by soft / hard thresholding

Consider identity design $\boldsymbol{X} = \boldsymbol{I}$, and $\hat{\beta}_i = \psi_{\mathsf{ht}}(y_i; \lambda)$ or $\hat{\beta}_i = \psi_{\mathsf{st}}(y_i; \lambda)$ with threshold zone $[-\lambda, \lambda]$

- For the extreme case where $\boldsymbol{\beta} = \mathbf{0}$,

$$\mathsf{MSE}^{\mathsf{ideal}}(\boldsymbol{\beta}) = \sum\nolimits_{i=1}^{p} \min\{\sigma^2, \beta_i^2\} = 0$$

- In order to control risk inflation, λ needs to be sufficiently large so as to ensure $\hat{\beta}_i \approx 0$ for all i. In particular,

$$\begin{aligned} \max\nolimits_{1 \le i \le p} |y_i| &= \max_{1 \le i \le p} |\eta_i| \approx \sigma\sqrt{2 \log p} \quad \text{(exercise)} \\ \Longrightarrow \quad & \lambda \ge \sigma\sqrt{2 \log p} \end{aligned}$$

Risk inflation by soft / hard thresholding

Theorem 8.4 (Foster & George '94, Johnstone, Candes)

Let $\hat{\boldsymbol{\beta}}$ be either a soft or hard thresholding procedure with threshold $\lambda = \sigma\sqrt{2\log p}$. Then

$$\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq (2\log p + c)\left(\sigma^2 + \mathsf{MSE}^{\mathsf{ideal}}(\boldsymbol{\beta})\right)$$

where $c = 1$ for soft thresholding and $c = 1.2$ for hard thresholding.

For large p, one typically has $\mathsf{MSE}^{\mathsf{ideal}}(\boldsymbol{\beta}) \gg \sigma^2$. Then Theorem 8.4 implies

$$\mathsf{RI}(\hat{\boldsymbol{\beta}}) \approx 2\log p$$

Proof of Theorem 8.4 for soft thresholding

WLOG, assume that $\sigma = 1$. The risk of soft thresholding for a single coordinate is

$$r_{\mathrm{st}}(\lambda, \beta_i) := \mathbb{E}[(\psi_{\mathrm{st}}(y_i; \lambda) - \beta_i)^2]$$

where $y_i \sim \mathcal{N}(\beta_i, 1)$.

1. There are 2 very special points that we shall single out: $\beta_i = 0$ and $\beta_i = \infty$. We start by connecting $r_{\mathrm{st}}(\lambda, \beta_i)$ with $r_{\mathrm{st}}(\lambda, 0)$ and $r_{\mathrm{st}}(\lambda, \infty)$.

Lemma 8.5

$$\begin{aligned} r_{\mathrm{st}}(\lambda, \beta) &\leq r_{\mathrm{st}}(\lambda, 0) + \beta^2 \quad \textit{(quadratic upper bound)} \\ r_{\mathrm{st}}(\lambda, \beta) &\leq r_{\mathrm{st}}(\lambda, \infty) = 1 + \lambda^2 \end{aligned}$$

Proof of Theorem 8.4 for soft thresholding

2. The next step is to control $r_{\mathrm{st}}(\lambda, 0)$

Lemma 8.6

$$r_{\mathrm{st}}(\lambda, 0) \leq 2\phi(\lambda)/\lambda \overset{\lambda=\sqrt{2\log p}}{\ll} 1/p \qquad (\textit{very small})$$

where $\phi(z) := \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$.

3. With these lemmas in mind, we are ready to prove Theorem 8.4

$$\begin{aligned}
\sum_{i=1}^{p} \mathbb{E}[(\beta_i - \hat{\beta}_i)^2] &\leq p r_{\mathrm{st}}(\lambda, 0) + \sum_{i=1}^{p} \min\left\{\beta_i^2, \lambda^2 + 1\right\} \\
&< 1 + \sum_{i=1}^{p} \min\left\{\beta_i^2, 2\log p + 1\right\} \\
&\leq (2\log p + 1)\left[1 + \sum_{i=1}^{p} \min\left\{\beta_i^2, 1\right\}\right] \\
&= (2\log p + 1)\left(1 + \mathsf{MSE}^{\mathsf{ideal}}\right)
\end{aligned}$$

Proof of Lemma 8.5

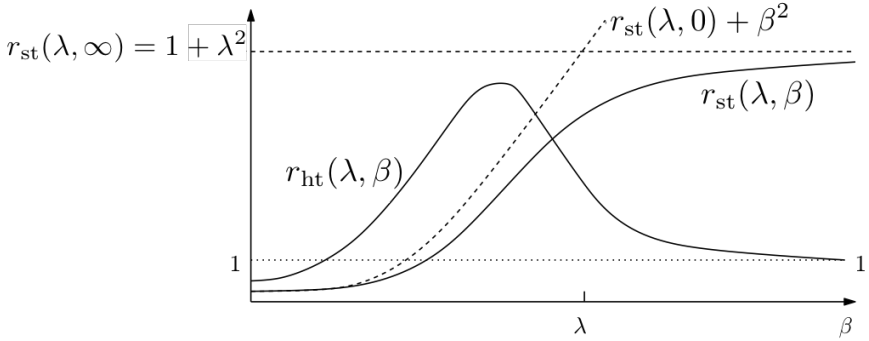


Figure adapted from Johnstone '15

WLOG, assume $\beta \geq 0$.

(1) To prove $r_{\mathrm{st}}(\lambda,\beta) \leq r_{\mathrm{st}}(\lambda,0) + \beta^2$, it suffices to show $\frac{\partial r_{\mathrm{st}}}{\partial \beta} \leq 2\beta$, as

$$r_{\mathrm{st}}(\lambda,\beta) - r_{\mathrm{st}}(\lambda,0) = \int_0^\beta \frac{\partial r_{\mathrm{st}}(\lambda,\beta)}{\partial \beta} \mathrm{d}\beta \leq \int_0^\beta 2\beta \mathrm{d}\beta = \beta^2.$$

This follows since (exercise)

$$\frac{\partial r_{\mathrm{st}}(\lambda,\beta)}{\partial \beta} = 2\beta \mathbb{P}\left(Z \in [-\lambda-\beta, \lambda-\beta]\right) \in [0, 2\beta], \tag{8.6}$$

with $Z \sim \mathcal{N}(0,1)$

Proof of Lemma 8.5

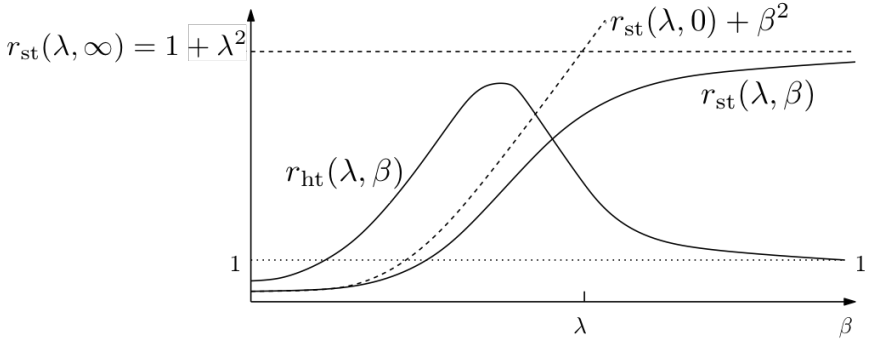


Figure adapted from Johnstone '15

WLOG, assume $\beta \geq 0$.

(2) The identity (8.6) also shows $r_{\rm st}$ is increasing in $\beta > 0$, and hence

$$r_{\rm st}(\lambda,\beta) \leq r_{\rm st}(\lambda,\infty) = \mathbb{E}[((\beta+z-\lambda)-\beta)^2] = 1+\lambda^2 \qquad (8.7)$$

Proof of Lemma 8.6 (Candes)

$$
\begin{aligned}
\frac{r_{\mathrm{st}}(\lambda,0)}{2} &= \int_{\lambda}^{\infty}(y-\lambda)^2\phi(y)\mathrm{d}y \\
&= \int_{\lambda}^{\infty}(y-2\lambda)\,y\phi(y)\mathrm{d}y + \lambda^2\int_{\lambda}^{\infty}\phi(y)\mathrm{d}y \\
&\overset{\text{(a)}}{=} -\int_{\lambda}^{\infty}(y-2\lambda)\,\phi'(y)\mathrm{d}y + \lambda^2\mathbb{P}\{Z>\lambda\} \\
&\overset{\text{(b)}}{=} -(y-2\lambda)\,\phi(y)\Big|_{\lambda}^{\infty} + \int_{\lambda}^{\infty}\phi(y)\mathrm{d}y + \lambda^2\mathbb{P}\{Z>\lambda\} \\
&= -\lambda\phi(\lambda) + \left(1+\lambda^2\right)\mathbb{P}\{Z>\lambda\} \\
&\overset{\text{(c)}}{\leq} -\lambda\phi(\lambda) + \left(1+\lambda^2\right)\frac{\phi(\lambda)}{\lambda} = \frac{\phi(\lambda)}{\lambda},
\end{aligned}
$$

where (a) follows since $\phi'(y) = -y\phi(y)$, (b) follows from integration by parts, and (c) holds since $\mathbb{P}\{Z>\lambda\} \leq \frac{\phi(\lambda)}{\lambda}$.

Optimality

Theorem 8.7 (Foster & George '94, Jonestone)

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \frac{\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})}{\sigma^2 + \mathsf{MSE}^{\mathsf{ideal}}(\boldsymbol{\beta})} \geq (1 + o(1)) 2 \log p$$

- Soft and hard thresholding rules—depending only on available data without access to an oracle—can achieve the ideal risk up to the multiplicative factor $(1 + o(1)) 2 \log p$
- This $2 \log p$ factor is asymptotically optimal for unrestricted $\boldsymbol{\beta}$

Comparison with canonical selection procedure

1. Minimax-optimal procedure w.r.t. risk inflation

$$\hat{\beta}_i = \psi_{\mathsf{ht}}\left(y_i;\ \sigma\sqrt{2\log p}\right)$$

2. Canonical selection based on C_p statistics

$$\hat{\beta}_i = \psi_{\mathsf{ht}}\left(y_i;\ \sqrt{2}\sigma\right)$$

   - Optimal procedure employs a much larger threshold zone
   - **Reason:** $\min_S C_p(S)$ underestimates $\min_S \mathbb{E}[\mathsf{PE}(S)]$ since

$$\mathbb{E}\left[\min_S C_p(S)\right] \underbrace{\leq}_{\text{sometimes } \ll} \min_S \mathbb{E}\big[C_p(S)\big] = \min_S \mathbb{E}\big[\mathsf{PE}(S)\big]$$

   - e.g. when $\boldsymbol{\beta} = \mathbf{0}$, $\|\psi_{\mathsf{ht}}\left(\boldsymbol{y}; \sqrt{2}\sigma\right) - \boldsymbol{\beta}\|^2 \asymp n \gg 0$ with high prob.

More general models

Let's turn to a general design matrix $\boldsymbol{X}$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} \quad \text{where } \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

One can take the ideal risk to be

$$\mathsf{MSE}^{\mathsf{ideal}} := \min_S \mathsf{PE}(S) = \min_S \Big\{ \underbrace{\mathbb{E}[\|\boldsymbol{X}_S \hat{\boldsymbol{\beta}}_S - \boldsymbol{X}\boldsymbol{\beta}\|_2^2]}_{\text{model error}} + |S|\sigma^2 \Big\}$$

More general models

Consider the ℓ_0-penalized selection procedure

$$\text{minimize}_{\hat{\boldsymbol{\beta}}} \quad \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda^2\sigma^2\|\hat{\boldsymbol{\beta}}\|_0 \tag{8.8}$$

for some $\lambda \asymp \sqrt{\log p}$

Theorem 8.8 (Foster & George '94, Birge & Massart '01, Jonestone, Candes)

$$(\textit{achievability}) \qquad \mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \lesssim (\log p)\left\{\sigma^2 + \mathsf{MSE}^{\text{ideal}}(\boldsymbol{\beta})\right\}$$

$$(\textit{minimax lower bound}) \qquad \inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}} \frac{\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})}{\sigma^2 + \mathsf{MSE}^{\text{ideal}}(\boldsymbol{\beta})} \gtrsim \log p$$

(8.8) is nearly minimax optimal for arbitrary designs!

Lasso is suboptimal for coherent design

$$\boldsymbol{X} = \begin{bmatrix} 1 & & & \epsilon \\ & 1 & & \vdots \\ & & \ddots & \epsilon \\ & & 1 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -1/\epsilon \\ 1/\epsilon \end{bmatrix}, \quad \text{and} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix}$$

When $\epsilon \to 0$, solution to Lasso is

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix}$$

far from the truth

- Issue: the last 2 columns of $\boldsymbol{X}$ are too similar / correlated

Minimax risk for sparse vectors

Asymptotic minimax risk for sparse vectors

So far we've considered risk without any restriction on $\boldsymbol{\beta}$. Practically, prior knowledge (like sparsity of $\boldsymbol{\beta}$) might be exploited to yield more accurate estimates.

Theorem 8.9

Suppose $\boldsymbol{X} = \boldsymbol{I}$. For any k-sparse $\boldsymbol{\beta}$ with $k \ll p$, the asymptotic minimax risk is

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_0 \le k} \mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (1 + o(1)) 2\sigma^2 k \log(p/k)$$

Minimaxity of soft thresholding estimator

Consider $\hat{\beta}_i = \psi_{\mathrm{st}}(y_i; \lambda)$ with $\lambda = \sigma\sqrt{2\log p}$ as before

- If $\boldsymbol{\beta} = \mathbf{0}$, one has $\hat{\boldsymbol{\beta}} \approx \mathbf{0}$ as discussed before
- If $\beta_1 \gg \beta_2 \gg \cdots \gg \beta_k \gg \sigma$ and $\beta_{k+1} = \cdots = \beta_p = 0$, then

$$\hat{\beta}_i \approx \begin{cases} y_i - \lambda, & \text{if } i \le k \\ 0, & \text{else} \end{cases}$$

$$\begin{aligned} \Longrightarrow \quad \mathrm{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &\approx \sum_{i=1}^{k} \mathbb{E}\left[(y_i - \beta_i - \lambda)^2\right] = k(\sigma^2 + \lambda^2) \\ &= k\sigma^2(2\log p + 1) > \underbrace{2k\sigma^2 \log(p/k)}_{\text{minimax risk}} \end{aligned}$$

  - Need to pick a smaller threshold λ

Minimaxity of soft thresholding estimator

Theorem 8.10

Suppose $\boldsymbol{X} = \boldsymbol{I}$. For any k-sparse $\boldsymbol{\beta}$ with $k \ll p$, the soft thresholding estimator $\hat{\beta}_i = \psi_{\mathrm{st}}(y_i; \lambda)$ with $\lambda = \sigma\sqrt{2\log(p/k)}$ obeys

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \leq (1 + o(1))2\sigma^2 k \log(p/k)$$

- Threshold λ determined by sparsity level

Sanity check

If $\beta_1 \gg \cdots \gg \beta_k \gg \sigma$ and $\beta_{k+1} = \cdots = \beta_p = 0$, then

$$\hat{\beta}_i \approx \begin{cases} y_i - \lambda, & \text{if } i \leq k \\ 0, & \text{else} \end{cases}$$

$$\begin{aligned} \implies \quad \mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &\approx k(\sigma^2 + \lambda^2) \quad \text{(as shown before)} \\ &= k\sigma^2(2\log(p/k) + 1) \\ &\approx \underbrace{2k\sigma^2 \log(p/k)}_{\text{minimax risk}} \end{aligned}$$

Proof of Theorem 8.11

WLOG, suppose $\sigma = 1$. Under the sparsity constraint,

$$\begin{aligned}
\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) &= \sum_{i=1}^{p} r_{\mathrm{st}}(\lambda, \beta_i) = \sum_{i:\beta_i \neq 0} r_{\mathrm{st}}(\lambda, \beta_i) + (p-k)\, r_{\mathrm{st}}(\lambda, 0) \\
&\leq k r_{\mathrm{st}}(\lambda, \infty) + (p-k)\, r_{\mathrm{st}}(\lambda, 0) \qquad (8.9) \\
&\leq k\left(1+\lambda^2\right) + 2p\frac{\phi(\lambda)}{\lambda} \qquad (8.10) \\
&= (1+o(1))2k\log(p/k) + \frac{k}{\sqrt{\pi \log(p/k)}} \\
&= (1+o(1))2k\log(p/k),
\end{aligned}$$

where (8.9) follows since $r_{\mathrm{st}}(\lambda, \beta)$ is increasing in β, and (8.10) comes from (8.7) and Lemma 8.6

Adaptivity to unknown sparsity

- **Problem of "optimal" soft thresholding:** knowing the sparsity level *a priori* is often unrealistic
- **Question:** can we develop an estimator that is adaptive to unknown sparsity?

Adaptivity cannot be achieved via soft thresholding with fixed thresholds, but what if we adopt data-dependent thresholds?

Data-dependent thresholds

Let $|y|_{(1)} \geq \cdots \geq |y|_{(p)}$ be the order statistics of $|y_1|, \cdots, |y_p|$

Key idea: use a different threshold for y_i based on its rank

$$\hat{\beta}_i = \psi_{\mathrm{st}}(y_i; \lambda_j) \quad \text{if } |y_i| = |y|_{(j)} \tag{8.11}$$

- originally due to Benjamini & Hochberg '95 for controlling false discovery rate

How to set thresholds? (non-rigorous)

Consider k-sparse vectors, and WLOG suppose $\sigma = 1$. Recall that when we use soft thresholding with $\lambda = \sqrt{2\log(p/k)}$, the least favorable signal is

$$\beta_1 \gg \cdots \gg \beta_k \gg \sigma \quad \text{and} \quad \beta_{k+1} = \cdots = \beta_p = 0 \tag{8.12}$$

If we use data-dependent thresholds and if $\{\lambda_i\}_{i>k}$ are sufficiently large, then

$$\hat{\beta}_i \approx \begin{cases} y_i - \lambda_i, & \text{if } i \le k \\ 0, & \text{else} \end{cases}$$

$$\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \approx \sum_{i=1}^{k} \mathbb{E}[(y_i - \lambda_i - \beta_i)^2] = \sum_{i=1}^{k} \left(1 + \lambda_i^2\right)$$

How to set thresholds? (non-rigorous)

If the estimator is minimax for each k and if the worst-case $\boldsymbol{\beta}$ for each k is given by (8.12), then

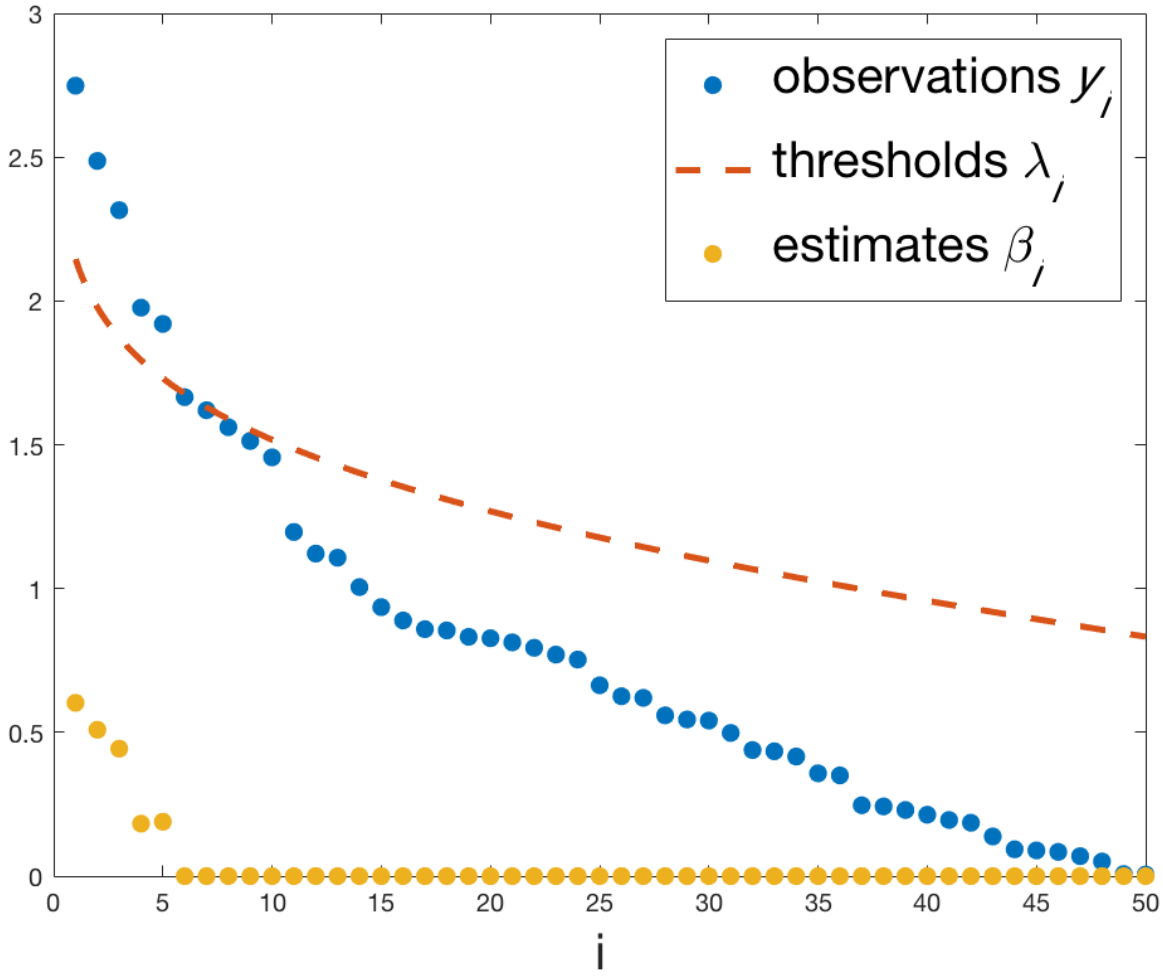
$$\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) \approx 2k \log(p/k), \qquad k = 1, \cdots, p$$

$$\Longrightarrow \quad \sum_{i=1}^{k} \lambda_i^2 \approx 2k \log(p/k), \qquad k = 1, \cdots, p$$

This suggests a choice (think of λ_i^2 as the derivative of $g(x) := 2x \log(p/x)$)

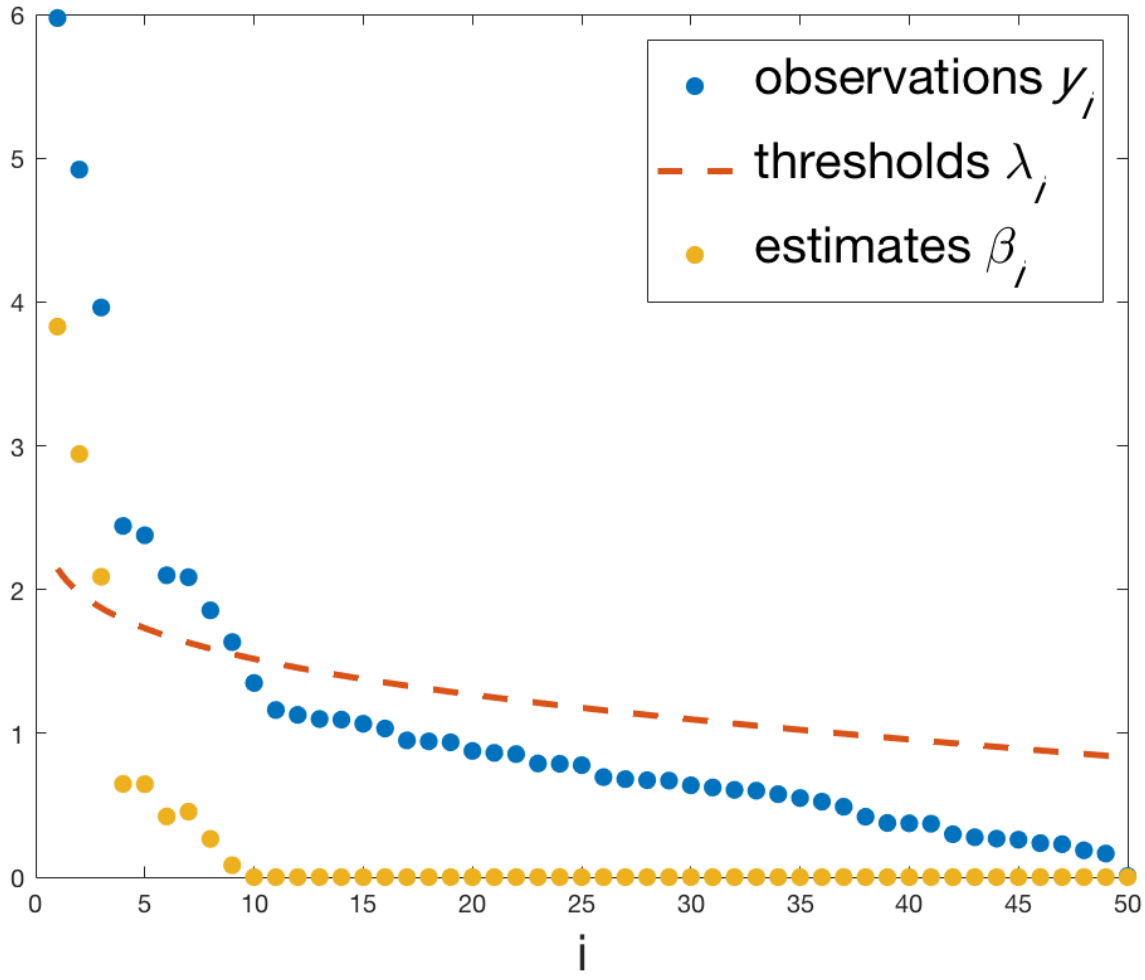
$$\lambda_i^2 \approx 2 \log(p/i) - 2 \approx 2 \log(p/i)$$

How to set thresholds? (non-rigorous)



When $\beta_1 = \cdots = \beta_{50} = 0, \sigma = 1$

How to set thresholds? (non-rigorous)



When $\beta_1 = \cdots = \beta_5 = 3, \beta_6 = \cdots = \beta_{50} = 0, \sigma = 1$

Minimaxity

Theorem 8.11 (Abramovich '06, Su & Candes '16)

Suppose $\boldsymbol{X} = \boldsymbol{I}$, *and* $k \ll p$. *The estimator (8.11) with* $\lambda_i = \sigma\sqrt{2\log(p/i)}$ *is minimax, i.e.*

$$\mathsf{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = (1 + o(1))2\sigma^2 k \log(p/k)$$

- Adaptive to unknown sparsity

Generalization to arbitrary design: SLOPE

SLOPE (Sorted L-One Penalized Estimation): a generalization of LASSO

$$\text{minimize}_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} \quad \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2 + \lambda_1 |\hat{\beta}|_{(1)} + \lambda_2 |\hat{\beta}|_{(2)} + \cdots + \lambda_p |\hat{\beta}|_{(p)}$$

where $\lambda_i = \sigma \Phi^{-1}(1 - iq/(2p)) \approx \sigma\sqrt{2\log(p/i)}$, $0 < q < 1$ is constant, and Φ is CDF of $\mathcal{N}(0,1)$

- This is a convex program if $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ (homework)
- This can be computed efficiently via proximal methods
- SLOPE is minimax and adaptive to unknown sparsity under i.i.d. Gaussian design $\boldsymbol{X}$

Reference

- "*Lecture notes, Theory of Statistics (Stats 300C)*," E. Candes.
- "*Statistical machine learning for high-dimensional data*," J. Fan, R. Li, C. Zhang, H. Zou, 2018.
- "*Linear regression analysis*," G. Seber and A. Lee, *Wiley*, 2003.
- "*Gaussian estimation: sequence and wavelet models*," I. Johnstone, 2015.
- "*Some comments on C_p*," C. Mallows, *Technometrics*, 1973.
- "*The risk inflation criterion for multiple regression*," D. Foster and E. George, *Annals of Statistics*, 1994.
- "*Regression Shrinkage and Selection via the lasso*," R. Tibshirani, *Journal of the Royal Statistical Society*, 1996.
- "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.

Reference

- "*Gaussian model selection*," L. Birge and P. Massart, *Journal of the European Mathematical Society*, 2011.
- "*Controlling the false discovery rate: a practical and powerful approach to multiple testing*," Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society*, 1995
- "*Adapting to unknown sparsity by controlling the false discovery rate*," F. Abramovich, Y. Benjamini, D. Donoho and I. Johnstone, *Annals of Statistics*, 2006.
- "*SLOPE is adaptive to unknown sparsity and asymptotically minimax*," W. Su and E. Candes, *Annals of Statistics*, 2016.