Review of Chapters 2 and 3

1. Recap

prediction with the Lasso:

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n = O_P(\sqrt{\frac{\log(p)}{n}}\|\beta^0\|_1)\ (n \to \infty)$$

assuming

- fixed design matrix $\mathbf{X}$ (analogous result for random design)
- Gaussian errors (can be relaxed)

2. Variable screening and $\|\hat{\beta} - \beta^0\|_q$-norms

estimation of parameters:

$$\|\hat{\beta} - \beta^0\|_q = o_P(1) \; (n \to \infty)$$

assuming

- compatibility condition on the (fixed) design $\mathbf{X}$
- Gaussian errors (can be relaxed)

more details: for $\lambda \asymp \sqrt{\log(p)/n}$,

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

the latter result needs a slightly stronger condition on the design

Variable screening

active set (of variables): $S_0 = \{j;\ \beta_j^0 \neq\}$
estimated active set: $\hat{S}_0 = \{j;\ \hat{\beta}_j \neq 0\}$

Question: is $\hat{S}_0 = S_0$ with high probability?
$\rightsquigarrow$ often too ambitious goal
problems with small $|\beta_j^0|$'s

denote by $S_0^{\mathrm{relevant}(C)} = \{j;\ |\beta_j^0| \geq C\}$
result: if $\|\hat{\beta} - \beta^0\|_1 \leq a_n$ with high probability, then:
if $C_n > a_n$,

$$\hat{S} \supset S_0^{\mathrm{relevant}(C_n)} \text{ with high probability}$$

Proof is elementary (Problem 2.3)

implication: typically,

$$\|\hat{\beta} - \beta^0\|_1 \le O(s_0\sqrt{\log(p)/n}) \text{ with high prob.}$$

hence, when assuming a

$$\text{"beta-min condition"}: \min_{j \in S_0^c} |\beta_j^0| \gg s_0\sqrt{\log(p)/n}$$

$\leadsto \hat{S} \supset S_0$

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

for this we require

- compatibility condition on the (fixed) design $\mathbf{X}$
- beta-min condition
- Gaussian errors (can be relaxed)

Variable selection

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

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

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

these conditions are often not fulfilled in practice
$\rightsquigarrow$ variable screening is realistic; 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 $\min_{j \in S_0^c} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$:

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

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

these conditions are often not fulfilled in practice
$\leadsto$ variable screening is realistic; variable selection is not very realistic
better "translation":
LASSO = Least Absolute Shrinkage and Screening Operator

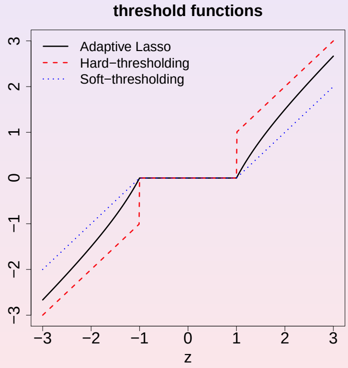
version of Table 2.2 in the book:

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

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

adaptive Lasso works well in practice (more sparse than Lasso) and has better theoretical properties than Lasso for variable screening (and selection)

alternatives: thresholding the Lasso; Relaxed Lasso

Computational algorithm for Lasso

can use a very generic coordinate descent algorithm

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

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

this implies (by setting the sub-differential to zero) the KKT-conditions (Lemma 2.1):

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

Let $\beta^{[0]} \in \mathbb{R}^p$ be an initial parameter vector. Set $m = 0$.

repeat

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 $\mathcal{S}^{[m]}$.

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).

until numerical convergence

for the squared error loss: the up-date in Step 4 is explicit

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

Generalized linear models (GLMs)

univariate response Y, covariate $X \in \mathcal{X} \subseteq \mathbb{R}^p$

GLM: $Y_1, \ldots, Y_n$ independent

$$g(\mathbb{E}[Y_i|X_i = x]) = \underbrace{\mu + \sum_{j=1}^{p} \beta_j x^{(j)}}_{=f(x)=f_{\mu,\beta}(x)}$$

$g(\cdot)$ real-valued, known link function; μ an intercept term

Lasso: defined as ℓ_1-norm penalized negative log-likelihood (μ is not penalized)

Example: logistic (penalized) regression
$Y \in \{0, 1\}$, $g(\pi) = \log(\pi/(1 - \pi))$ ($\pi \in (0, 1)$)