Recap

Oracle inequality 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 $\lambda_0 \asymp \sqrt{\log(p)/n}$

implications:

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

these are the (minimax) optimal rates:

no other method can do better

Variable Screening

assume compatibility condition and (e.g.) Gaussian errors
in addition, require beta-min condition:

$$\min_{j \in S_0} |\beta_j^0| \gg s_0 \sqrt{\log(p)/n}$$

$$\Longrightarrow \mathbb{P}[\hat{S} \supseteq S_0] \to 1 \ (p \geq n \to \infty)$$

with high probab: Lasso selects a superset of the active set S_0
$\leadsto$ Lasso does not miss an important active variable!

in practice: $\lambda = \lambda_{CV} \leadsto$ leads “typically” to a too large model

LASSO: Least Absolute Shrinkage and Screening Operator

Variable Selection

obtaining

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

necessarily requires restrictive condition on X, the so-called irrepresentability condition (= neighborhood stability condition)

Adaptive Lasso (Zou, 2006)

two-stage procedure:

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

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

at least as sparse (typically more sparse) than Lasso

$\leadsto$ good/"better" for very sparse underlying mechanisms/models