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

*Corollary 6.1*

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

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 orcale on the variables from $S_0 = \{j;\ \beta_j^0 \neq 0\}$

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

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

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P(\log(p)\frac{s_0}{n})$$

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

# Extensions

the proof technique decouples into a deterministic and probablistic part (the set $\mathcal{T}$)

the deterministic part remains the same for other probabilistic structures (other analysis for $\mathbb{P}[\mathcal{T}]$) such as:

- heteroscedastic errors with
  $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 \not\equiv$ const.
- dependent observations $\leadsto$ for fixed design, dependent errors
- non-Gaussian errors
  sub-Gaussian distribution
  second moments plus bounded $X$: see Example 14.3 in Bühlmann and van de Geer (2011)
- random design: assume that $\varepsilon$ is independent of $X$
  $\leadsto$ condition on $X$: invoke the results for fixed design and integrate out

heteroscedastic errors

$\varepsilon \sim \mathcal{N}_n(0, D)$, where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

assume that: $\sigma_i^2 \leq \underbrace{\sigma^2}_{\text{some pos. const.}} < \infty$

Then, Coroallry 6.1 remains true with $\sigma^2$ as above

Proof:

exactly as before but exploiting that $V_j \sim \mathcal{N}(0, \tau_j^2)$ with $\tau_j \leq 1$ and using that $\mathbb{P}[V_j > c] \leq \mathbb{P}[\underbrace{Z}_{\sim \mathcal{N}(0,1)} \leq c]$

Exercise: work out the details.

errors from stationary distribution

$\varepsilon \sim \mathcal{N}_n(0, \Gamma)$, where $\Gamma_{i,j} = R(i-j) = R(j-i)$
assume that: $\sum_{k=-\infty}^{\infty} |R(k)| < \infty$ and $|X_i^{(j)}| \leq K_X < \infty$

Then, Corollary 6.1 remains true with $\sigma^2 = K_X^2 \sum_{k=-\infty}^{\infty} |R(k)|$

Proof:
Exercise. (A bit more tricky...)

# Oracle inequality

aim: what can we say about

- $\|\hat{\beta} - \beta^0\|_q$ for $q \in \{1, 2\}$
- fast convergence rate for $\|X(\hat{\beta} - \beta^0)\|_2^2/n$

consider again

$$\mathcal{T} = \{ \max_{j \in \{1,\ldots,p\}} 2|\varepsilon^T X^{(j)}|/n \leq \lambda_0 \}$$

*Theorem 6.1 in Bühlmann and van de Geer (2011)*
assume: compatibility condition holds with compatibility constant $\phi_0^2 \geq L > 0$
Then, on $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$:

$$\|X(\hat{\beta} - \beta^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2$$

$$\|X(\hat{\beta} - \beta^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2$$

$\leadsto$

$$\|X(\hat{\beta} - \beta^0\|_2^2/n \leq 4\lambda^2 s_0/\phi_0^2 \asymp \log(p)s_0/n \quad \text{fast converg. rate}$$

$$\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda s_0/\phi_0^2 \asymp s_0\sqrt{\log(p)/n} \quad \text{estimation error. for par.}$$

for oracle inequality (and estimation error): we cleary need some assumptions on $X$

for $p > n$ and $\mathrm{rank}(X) = n$, the null-space of $X$ is not only the zero vector:

$$X\xi = 0 \text{ for inifintely many } \xi \neq 0$$

$\leadsto$ $X\beta^0 = X\theta$ for $\theta = \beta^0 + \xi$ with any $\xi$ such that $X\xi = 0$.

we cannot identify the true parameter $\beta^0$ from (inifintely many) data
$\leadsto$ we have to make an assumption on $X$

# Compatibility condition

the compatibility condition holds for the true active set $S_0$ with compatibility constant $\phi_0^2 > 0$ if:

$$\forall \beta \text{ satisfying } \|\beta_{S_0^C}\|_1 \leq 3\|\beta_{S_0}\|_1 :$$
$$\|\beta_{S_0}\|_1^2 \leq (\beta^T \hat{\Sigma} \beta) s_0 / \phi_0^2$$

see p. 106 in Bühlmann and van de Geer (2011)

$\rightsquigarrow$

*Theorem 6.1 in Bühlmann and van de Geer (2011)*
assume: compatibility condition holds with compatibility constant $\phi_0^2$ $(\geq L) > 0$
Then, on $\mathcal{T}$ and for $\lambda \geq 2\lambda_0$:

$$\|X(\hat{\beta} - \beta^0\|_2^2/n + \lambda\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2$$

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

estimation of parameters: thanks to the oracle inequality

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

assuming

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

$\leadsto$ convergence to zero if sparsity $s_0 = o(\sqrt{n/\log(p)})$

under restricted eigenvalue assumption (slightly stronger than compatbility condition): one can show that

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

$\leadsto$ convergence to zero under weaker ass. $s_0 = o(n/\log(p))$

# Variable screening

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

Question 1: is $\hat{S}_0 = S_0$ with high probability?
$\leadsto$ often too ambitious goal
problems with small $|\beta_j^0|$'s

Question 2: can we do variable screening $\hat{S} \supseteq S_0$ with high probability?
still very relevant in practice: dimensionality reduction!

need to make an assumption that true regression coefficients are not too small

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

$$\Longrightarrow \mathbb{P}[\hat{S} \supseteq S_0] \to 1 \text{ if } \|\hat{\beta} - \beta^0\|_1 = O_P(s_0 \sqrt{\log(p)/n})$$

Proof: suppose that $j^* \in S_0$ but $j^* \notin \hat{S}$

$$\|\hat{\beta} - \beta^0\|_1 \geq |\hat{\beta}_{j^*} - \beta_{j^*}^0| = |\beta_{j^*}^0| \gg s_0 \sqrt{\log(p)/n}$$

which is a contradiction $\square$

analogously: if

- beta-min condition $\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$
- $\|\hat{\beta} - \beta^0\|_2 = O_P(\sqrt{s_0 \log(p)/n})$

$\rightsquigarrow \mathbb{P}[\hat{S} \supseteq S_0] \to 1$

## Theory versus Practice

theory:

$$\mathbb{P}[\hat{S} \supseteq S_0] \to 1$$

if the following hold:

- compatibility condition for the (fixed) design $X$
- beta-min condition
- Gaussian errors (can be relaxed)

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

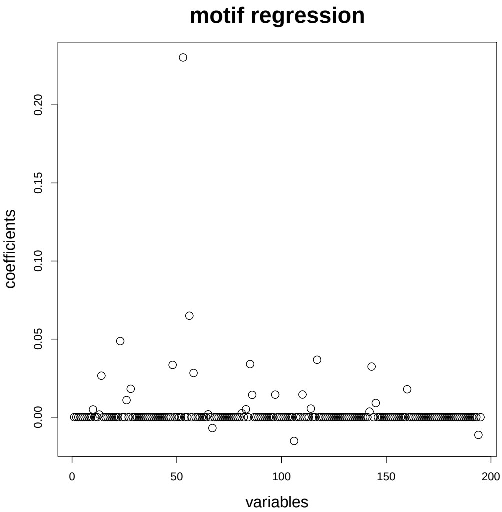
in practice: $\mathbb{P}[\hat{S} \supseteq S_0]$ may not be soo large...
even if one chooses $\lambda$ very small which results in a typically larger set $\hat{S}$...

possible reasons to explain with theory:

- compatibility constant $\phi_0^2$ might be very small (due to highly correlated columns in $X$ or near linear dependence among a few columns of $X$)
  $\leadsto \|\hat{\beta} - \beta^0\|_1 \le 4\lambda s_0/\phi_0^2$
  $\leadsto$ requires a stronger beta-min condition!
- errors are non-Gaussian (heavy tailed)

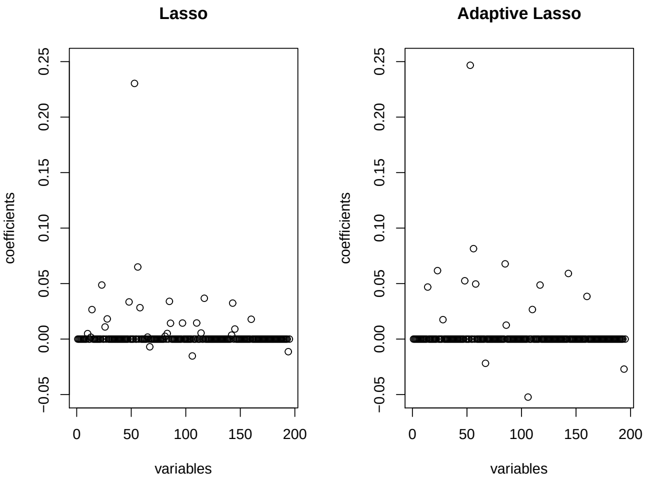
it is "empirically evident" though: $\mathbb{P}[\hat{S} \supseteq S_{\text{substantial}(C)}]$ large
where $S_{\text{substantial}(C)} = \{j;\ |\beta_j^0| \ge \underbrace{C}_{\text{large}}\}$

# The Lasso workhorse




$$p = 195, n = 143, |\hat{S}(\lambda_{CV})| = 26$$

# The (adaptive) Lasso workhorse




$$p = 195, n = 143, |\hat{S}_{\mathrm{Lasso}}(\lambda_{CV})| = 26$$

# When does the compatibility condition hold?

have seen that the compatibility condition plays a major role for estimating $\beta^0$ and for fast convergence rate for prediction

*Corollary 6.8 from Bühlmann and van de Geer (2011) – modified form*
Assume that the row vectors of $X$ are i.i.d. sampled from a sub-Gaussian distribution with mean zero and covariance matrix $\Sigma$. Assume that

- $\lambda_{min}^2(\Sigma) > 0$
- $s_0 = |S_0| = O(\sqrt{n/\log(p)})$

Then: $\phi_0^2 \geq \lambda_{\min}^2(\Sigma) > 0$ with probability $\rightarrow 1$ $(n \rightarrow \infty)$

Example: Toeplitz matrix $\Sigma_{ij} = \rho^{|i-j|}$ $(0 \leq \rho < 1)$:
$\lambda_{min}^2(\Sigma) \geq L > 0$ where $L$ is independent of $p$