

Extensions

the proof technique **decouples** into a deterministic and probabilistic 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 $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma_i^2 \neq \text{const.}$
- ▶ dependent observations $\rightsquigarrow$ 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 ε is independent of X
 $\rightsquigarrow$ condition on X : invoke the results for fixed design and integrate out

heteroscedastic errors

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

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

Then, Corollary 6.1 remains true with σ^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 $\mathbb{P}[|V_j| > c] \leq \mathbb{P}[\underbrace{|Z|}_{\sim |\mathcal{N}(0,1)|} > c]$

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...)

Compatibility condition, Restricted eigenvalues, Sparse eigenvalues

if we want to identify and estimate the regression parameter β^0
 $\leadsto$ need necessarily more assumptions

example:

- if $X^{(1)} = X^{(2)} \leadsto$ cannot distinguish between β_1^0 and β_2^0
- in low dimensions: typically assume $\text{rank}(X) = p (\leq n)$
 - $\text{rank}(X) \leq \min(n, p) \leadsto$ for $p > n$: never achieve full rank p

how to measure “degree” of identifiability?

suppose $X\theta = X\beta^0$

then:

$$\begin{aligned} 0 &= \|X(\theta - \beta^0)\|_2^2/n = (\theta - \beta^0)^T \hat{\Sigma} (\theta - \beta^0) \\ &\geq \underbrace{\lambda_{\min}^2(\hat{\Sigma})}_{\text{min. eigenval. of } \hat{\Sigma}} \|\theta - \beta^0\|_2^2 \\ &\quad \hat{\Sigma} = X^T X/n \end{aligned}$$

since $\lambda_{\min}^2(\hat{\Sigma}) = \min_{u; u \neq 0} \frac{u^T \hat{\Sigma} u}{\|u\|_2^2}$

for $p > n$: $\lambda_{\min}^2(\hat{\Sigma}) = 0 \leadsto$ bound above is “useless”

idea: **restrict to small sub-matrices**

$\leadsto$ sparse eigenvalues (Meinshausen & Yu, 2009)

$$\phi_{\min}^2(m) = \min_{S \subseteq \{1, \dots, p\}} \left(\lambda_{\min}^2(\hat{\Sigma}_S); |S| \leq m \right)$$
$$\iff \phi_{\min}^2(m) = \min_{\beta \neq 0; \|\beta\|_0 \leq m} \frac{\beta^T \hat{\Sigma} \beta}{\|\beta\|_2^2}$$

Denote by $s_\theta = \|\theta\|_0$, $s_0 = \|\beta^0\|_0$

Then: if we require $\phi_{\min}^2(s_\theta + s_0) > 0$:

since $\|\theta - \beta^0\|_0 \leq s_\theta + s_0$ we obtain

$$0 = \|X(\theta - \beta^0)\|_2^2/n \geq \phi_{\min}^2(s_\theta + s_0) \|\theta - \beta^0\|_2^2$$
$$\leadsto \theta = \beta^0$$

Conclusion:

if we restrict to **sparse** vectors θ with at most the sparsity of β^0 ,
i.e., $\|\theta\|_0 = s_\theta \leq \|\beta^0\|_0 = s_0$

$\leadsto$ can identify the regression parameter vector if $\phi_{\min}^2(2s_0) > 0$
(have not shown that identification works with the Lasso...)

Restricted eigenvalues

instead of sparse eigenvalues:

Lasso identifies β^0 under weaker conditions

idea: can restrict to additional **cone condition**:

for a subset $S \subseteq \{1, \dots, p\}$, consider cone

$$\|\beta_{S^c}\|_1 \leq \underbrace{3}_{\text{arbitrary}} \|\beta_S\|_1 \quad (1)$$

the ℓ_1 -norm in the (large) complement S^c of S is small
Restricted eigenvalue (Bickel, Ritov & Tsybakov, 2009):

$$\kappa^2(m, 3) = \min_{S: |S| \leq m} \min_{\beta \neq 0, \beta \text{ fulfills (1)}} \frac{\beta^T \hat{\Sigma} \beta}{\|\beta_S\|_2^2}$$

Why such a cone?

can show that for Lasso, with high probability (on $\mathcal{T}$):
cone condition is fulfilled for $\hat{\beta} - \beta^0$ for $S = S_0$

$$\text{on } \mathcal{T} : \|(\hat{\beta} - \beta^0)_{S_0^c}\|_1 = \|\hat{\beta}_{S_0^c}\|_1 \leq 3\|(\hat{\beta} - \beta^0)_{S_0}\|_1$$

if $\kappa^2(s_0, 3) > 0$: can identify β^0 with the Lasso

Proof: an incomplete version, see visualizer

Compatibility constant

slightly weaker condition the restricted eigenvalue
Compatibility constant for the set S_0 (van de Geer, 2007):

$$\phi_0^2 = \min_{\beta \neq 0} \min_{\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1} \frac{\mathbf{s}_0 \beta^T \hat{\Sigma} \beta}{\|\beta_{S_0}\|_1^2} \quad (\mathbf{s}_0 = |S_0| = \|\beta^0\|_0)$$

see also Bühlmann and van de Geer, p. 106

It holds that

$$\phi_0^2 \geq \kappa^2(\mathbf{s}_0, 3)$$

Proof: Cauchy-Schwarz inequality

$$\|\beta_{S_0}\|_1 \leq \sqrt{\mathbf{s}_0} \|\beta_{S_0}\|_2$$

In the denominator of $\kappa^2(\mathbf{s}_0, 3)$ we have $\|\beta_{S_0}\|_2^2$
 $\leadsto \phi_0^2 \geq \kappa^2(\mathbf{s}_0, 3).$

□

compatibility condition holds:

$$\phi_0^2 > 0$$

and this is still enough to identify the true parameter with Lasso using only $\phi_0^2 > 0$ (instead of the stronger assumption $\kappa^2(s_0, 3) > 0$)

Oracle inequality for the Lasso

$$Y = X\beta^0 + \varepsilon, \quad p \gg n$$

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 $\lambda \geq 2\lambda_0$:

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

$$\text{recall:} \quad \mathcal{T} = \left\{ 2 \max_{j=1, \dots, p} |\varepsilon^T X^{(j)}|/n \leq \lambda_0 \right\}$$

$$\mathbb{P}[\mathcal{T}] \text{ large if } \lambda_0 \asymp \sqrt{\log(p)/n}$$

When does the compatibility condition hold?

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 Σ . Assume that

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

Then, for some $C > 0$:

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

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

Implications of oracle inequality

assume that $\phi_0^2 \geq L > 0$

$$\text{on } \mathcal{T}: \|X(\hat{\beta} - \beta^0)\|_2^2/n + \lambda \|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0 / \phi_0^2$$

if $\lambda (= 2\lambda_0) \asymp \sqrt{\log(p)/n}$: as $p \geq n \rightarrow \infty$,

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

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

thus, when compatibility condition holds:

1. fast rate of convergence for prediction

if S_0 with $s_0 = o(n)$ would be known:

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

$\rightsquigarrow$ factor $\log(p)$ is the (small!) price for not knowing S_0

2. estimation error for β^0 in terms of the ℓ_1 -norm

Variable screening

active set (of variables): $S_0 = \{j; \beta_j^0 \neq 0\}$

estimated active set: $\hat{S}_0 = \{j; \hat{\beta}_j \neq 0\}$

make an assumption that true regression coefficients are not too small

$$\text{"beta-min condition"} : \min_{j \in S_0} |\beta_j^0| > \underbrace{4\lambda s_0 / \phi_0^2}_{\text{bound for } \|\hat{\beta} - \beta^0\|_1}$$

$$\implies \mathbb{P}[\hat{S} \supseteq S_0] \geq \mathbb{P}[\mathcal{T}] = \text{"large"}$$

with high probab: Lasso selects a superset of the active set S_0

$\leadsto$ Lasso does not miss an important active variable!

$$\mathbb{P}[\hat{S} \supseteq S_0] \geq \mathbb{P}[\mathcal{T}] = \text{“large”}$$

Proof:

Suppose that $\hat{S} \not\supseteq S_0$: $\leadsto$ there exists $j^* \in S_0$ with $\hat{\beta}_{j^*} = 0$

But then, on $\mathcal{T}$:

$$\|\hat{\beta} - \beta^0\|_1 \geq \|\hat{\beta}_{j^*} - \beta_{j^*}^0\| = |\beta_{j^*}^0| > 4\lambda s_0 / \phi_0^2$$

which is a contradiction to the oracle inequality

$$(\text{for } \|\hat{\beta} - \beta^0\|_1 \leq \lambda 4s_0 / \phi_0^2)$$

□

Theory versus Practice

theory:

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

if the following hold:

- ▶ compatibility condition for the (fixed) design X
- ▶ beta-min condition
- ▶ i.i.d. 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 so large...
even if one chooses λ very small which results in a typically larger set $\hat{S}$...

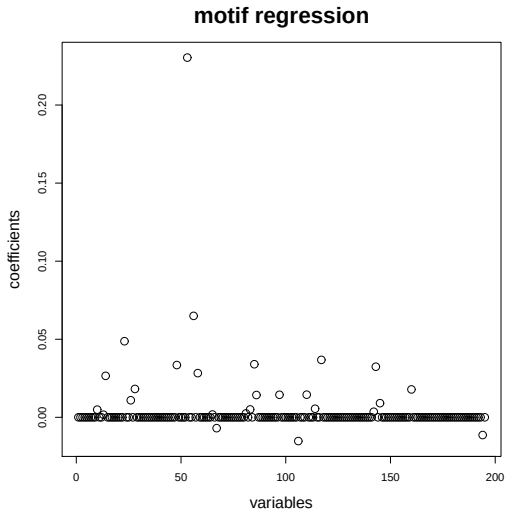
possible reasons to explain with theory:

- ▶ compatibility constant ϕ_0^2 might be very small (due to highly correlated columns in X or near linear dependence among a few columns of X)
 $\leadsto \|\hat{\beta} - \beta^0\|_1 \leq 4\lambda s_0 / \phi_0^2$
 $\leadsto$ 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 $S_{\text{substantial}(C)} = \{j; |\beta_j^0| \geq \underbrace{C}_{\text{large}}\}$

The Lasso workhorse



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

Variable selection

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

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

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

this condition is often not fulfilled in practice
(and choosing the correct λ would be difficult as well)

↪ variable screening is realistic (“choose λ by CV”)
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 $\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n}$:

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

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

this condition is often not fulfilled in practice
(and choosing the correct λ would be difficult as well)

↪ variable screening is realistic (“choose λ by CV”)

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