Lasso involves convex optimization

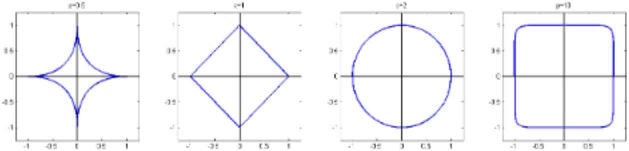


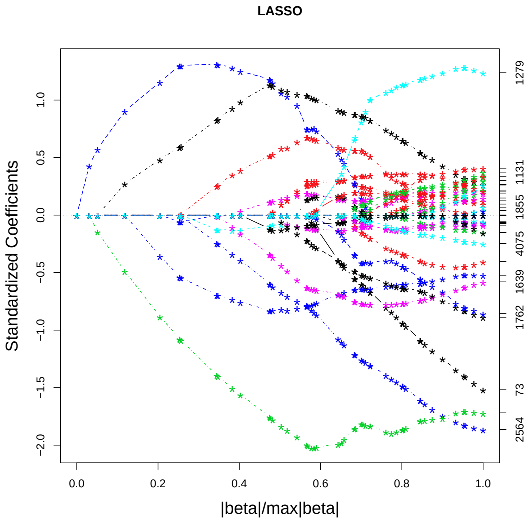
Figure 1: Unit circles for several Minkowski-p-norms $\|\mathbf{x}\|_p$: from left to right $p = 0.5$, $p = 1$ (Manhatten), $p = 2$ (Euclidean), $p = 10$.

from Lange, Zühkle, Holz, Villmann (2014)

- convex: ℓ_p-norm with $p \geq 1$
- sparse: ℓ_p-norm with $p \leq 1$ (need “edges in the ball)

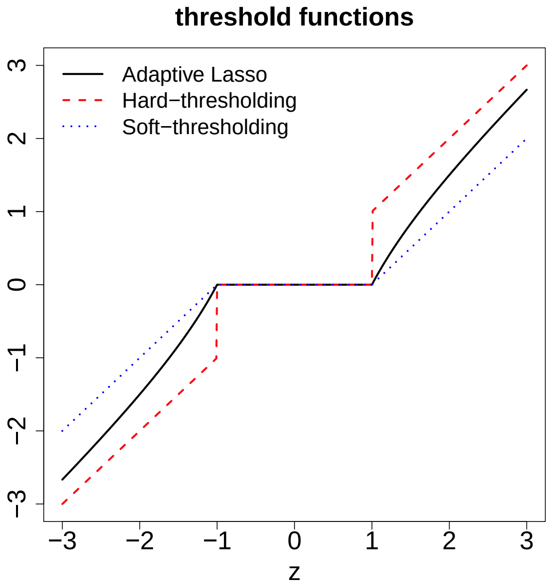
$\rightsquigarrow$ $p = 1$, that is Lasso, for sparse and convex estimator

Lasso regularization path for Riboflavin data



certainly sparse, even when $\lambda \to 0$ (x-axis is like inverse of λ)

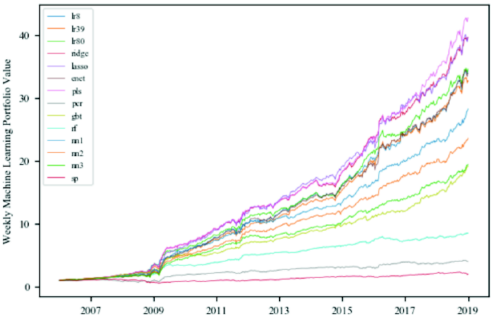
Orthonormal design: threshold functions



Lasso as a prediction machine/method

Fig. 6. Cumulative Performance for Weekly Machine Learning Portfolios

This figure depicts the cumulative weekly returns for equal-weighted long-short machine learning portfolios for 13 machine learning models. ‘lr80’ denotes the linear regression model with all 80 variables. ‘lr39’ denotes the linear regression model with variables that have strong predictability for weekly returns. ‘lr8’ denotes the linear regression model with the eight strongest predictor variables. ‘ridge’ denotes the ridge regression model. ‘lasso’ denotes the lasso regression model. ‘enet’ denotes the elastic net regression model. ‘pls’ denotes the partial least squares regression model. ‘pcr’ denotes the principal component regression model. ‘gbt’ denotes the gradient boosting regression tree model. ‘rf’ denotes the random forest regression model. ‘nn1’ denotes the neural network model with one hidden layer. ‘sp’ denotes the benchmark S&P 500 index. The accumulation period is from January 2006 to December 2018, and the initial investment is set as 1 at the start of the accumulation period.



taken from MSc thesis Jiawen Le (September 2019)

II.4.2 Some results from asymptotic theory

asymptotics when $p = p_n$ (typically $\gg n$, $\to \infty$) and $n \to \infty$

triangular array asymptotics:

$$(X, Y)_{n;1}, \ldots, (X, Y)_{n;n}$$
$$(X, Y)_{n+1;1}, \ldots, (X, Y)_{n+1;n}, (X, Y)_{n+1;n+1}$$
$$(X, Y)_{n+2;1}, \ldots, (X, Y)_{n+2;n}, (X, Y)_{n+2;n+1}, (X, Y)_{n+2;n+2}$$

...

$$Y_{n,i} = \sum_{j=1}^{p_n} \beta^0_{n;j} X^{(j)}_{n;i} + \varepsilon_{n;i}. \quad i = 1, \ldots, n, \; n = 1, 2, \ldots$$

$\mathbb{E}[\varepsilon_{n;i}] = 0$ and usually fixed design X

but we usually do not emphasize the dependence on n

announcement of a few results:

1. for fixed design: if $\|\beta^0\|_1 = o(\sqrt{\frac{n}{\log(p)}})$, then

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = o_P(1)$$

   slow rate, just consistency

2. for fixed design which satisfies a "compatibility condition" (restricted eigenvalue condition) with constant $\phi_0^2 > 0$:

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P\left(\frac{s_0 \log(p)}{n} \frac{1}{\phi_0^2}\right)$$

$$\|\hat{\beta} - \beta^0\|_1 = O_P\left(s_0 \sqrt{\frac{\log(p)}{n}} \frac{1}{\phi_0^2}\right)$$

$$s_0 = |S_0| = |\{j;\ \beta_j^0 \neq 0\}|$$

   ϕ_0^2 close to zero means "badly conditioned (highly correlated)" columns of X

therefore:

- if $s_0 = o\left(\frac{n}{\log(p)}\phi_0^2\right)$:

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P\left(\frac{s_0 \log(p)}{n}\frac{1}{\phi_0^2}\right) \to 0 \ (p, n \to \infty)$$

- if $s_0 = o\left(\sqrt{\frac{n}{\log(p)}}\phi_0^2\right)$:

$$\|\hat{\beta} - \beta^0\|_1 = O_P\left(s_0\sqrt{\frac{\log(p)}{n}}\frac{1}{\phi_0^2}\right) \to 0 \ (p, n \to \infty)$$

Developing the theory for announced results

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 σ. 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 and Asymptotic viewpoint

the proper $\lambda \asymp \sqrt{\log(p)/n}$ (take e.g. $t^2 \asymp \log(p)$)

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 oracle 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

$$\|X(\hat{\beta} - \beta^0)\|_2^2/n = O_P(\log(p)\frac{s_0}{n}) \quad \text{(if } \phi_0^2 \text{ bounded away from zero)}$$

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

Proof of such results: see visualizer

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 $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 \not\equiv$ 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 $D = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

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

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...)