Square-root lasso:
pivotal recovery of sparse signals via conic programming

REPORTER:Ge,Liu,Luo,Wang

January 3, 2021

Structure

Changes

- Different choice of penalty level λ:

$$\lambda = \sigma c 2 n^{1/2} \Phi^{-1}(1 - \alpha/2p), (lasso)$$
$$\lambda = c n^{1/2} \Phi^{-1}(1 - \alpha/2p), (s - lasso)$$

- Objective function:

$$\hat{Q}(\beta) + \frac{\lambda}{n} \|\beta\|_1, (lasso)$$
$$\hat{Q}(\beta)^{1/2} + \frac{\lambda}{n} \|\beta\|_1, (s - lasso)$$

Contribution

- Achieve the same near-oracle rates of convergence as lasso, without knowing σ.Additionally, we could drop the assumption of noise's normality under specific condition.
- Due to the maintenance of global convexity, square-root lasso could be set as a solution to a conic programming problem.

- Our task is to choose a $\hat{\beta}$, such that[1]

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta)^{1/2} + \frac{\lambda}{n}\|\beta\|_1 \quad (1)$$

- Generally, the optimal β to minimum a differentiable function $f(\beta)$ is a point where the gradient vanishes, i.e.

$$\nabla_\beta f(\beta) = 0$$

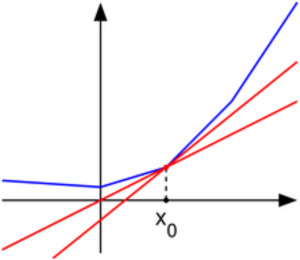
- Unfortunately, the gradient of ℓ_1-norm doesn't exists.

[1] The definitions of $\hat{Q}$: $\hat{Q}(\beta) = n^{-1} \sum_{i=1}^{n} (y_i - x_i' \beta)^2$

- Rigorously, a subgradient (or subderivative in $\mathbb{R}^1 space$) of a **convex function** $f : U \to \mathbb{R}$ at a point x_0 is all vectors v satisfying:

$$f(x) - f(x_0) \geq v \cdot (x - x_0) \qquad (2)$$

for every $x \in U$, where U is a subset of $\mathbb{R}^n$.



Properties of subgradient:

- A convex function is differentiable at a point x_0 if and only if the subgradient is made up of only one vector, which is the gradient of f at x_0.
- A point x_0 is a global minimum of a convex function f if and only if zero is contained in the subdifferential.

Choose of Penalty Level

- Let

$$\tilde{S} = \nabla \hat{Q}^{1/2}(\beta_0) = \frac{\mathbf{E}_n(x\sigma\epsilon)}{\{\mathbf{E}_n(\sigma^2\epsilon^2)\}^{1/2}} = \frac{\mathbf{E}_n(x\epsilon)}{\{\mathbf{E}_n(\epsilon^2)\}^{1/2}} \tag{3}$$

  where β_0 is the real parameter value and $\mathbf{E}_n(f) = \mathbf{E}_n\{f(z)\} = \sum_{i=1}^n f(z_i)/n$.

- Take derivative of each dimensions of β at point β_0, and apply the property of subgradient:

$$-\tilde{S}_j + \lambda/n \geq 0, \tilde{S}_j + \lambda/n \geq 0 \quad j = 1, \cdots, p \tag{4}$$

Choose of Penalty Level

- One should choose a λ such that $\lambda/n \geq \max_{1 \leq j \leq p} |\tilde{S}_j|$,i.e.

$$\lambda \geq \Lambda, \quad \Lambda = n\|\tilde{S}\|_\infty \tag{5}$$

- For reasons of efficiency and regularization, we set λ the smallest level such that

$$\lambda \geq c\Lambda, \quad \Lambda = n\|\tilde{S}\|_\infty \tag{6}$$

with a high probability $1 - \alpha$, where $c > 1$ is a theoretical constant to be stated later.

Choose of Penalty Level

- The rule above is not practical, since we do not observe Λ directly. However, we can proceed as follows:

1. When we know the distribution of errors exactly, e.g., $F_0 = \Phi$, we propose to set λ as c times the $(1 - \alpha)$ quantile of Λ given X. This choice of the penalty level precisely implements (6) and is easy to compute by simulation.
2. When we do not know F_0 exactly, but instead know that F_0 is an element of some family $\mathcal{F}$, we can rely on either finite sample or asymptotic upper bounds on quantiles of Λ given X.

Choose of Penalty Level

- In order to describe our choice of λ formally, define for $0 < \alpha < 1$

$$\Lambda_F(1-\alpha|X) = (1-\alpha) - \text{quantile of } \Lambda_F|X \tag{7}$$

$$\Lambda(1-\alpha) = n^{1/2}\Phi^{-1}(1-\alpha/2p) \leq \{2n\log(2p/\alpha)\} \tag{8}$$

where $\Lambda_F = n\|\mathbf{E}_n(x\xi)\|_\infty / \{\mathbf{E}_n(\xi^2)\}^{1/2}$, with i.i.d $\xi_i(i=1, \cdots, n)$ having law F.

- In the normal case, $F_0 = \Phi$,λ can be either of

$$\lambda = c\Lambda_\Phi(1-\alpha|X), \lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi(1-\alpha/2p) \tag{9}$$

where we call the exact and asymptotic options respectively.

Choose of Penalty Level

Some mild conditions needed for following result:

- *Condition* 1.

$$\log^2(p/\alpha)\log(1/\alpha) = o(n) \text{ and } p/\alpha \to \infty \text{ as } n \to \infty \quad (10)$$

- *Condition* 2.
  There exists a finite constant $q > 2$ such that the law F_0 is an element of the family $\mathcal{F}$ such that $\sup_{n\geq 1}\sup_{F\in\mathcal{F}} \mathbf{E}_F(|\epsilon|^q) < \infty$;the design X obeys:

$$\sup_{n\geq 1, 1\leq j\leq p} \mathbf{E}_n(|x_j|^q < \infty)$$

  .

- *Condition* 3.
  As $n \to \infty$,$p \leq \alpha n^{\eta(q-2)/2}/2$ for some constant $0 < \eta < 1$, and $\alpha^{-1} = o[n^{\{(q/2-1)\vee(q/4)\}\wedge()q/2-2}/(\log n)^{q/2}]$, where $q > 2$ is defined in Condition 2.

Lemma1

Suppose that $F_0 = \Phi$.
(i)Assume $p/\alpha > 8$.For any $1 < \ell < \{n/\log(1/\alpha)\}^{1/2}$, the asymptotic option in (9) implements $\lambda \geq c\Lambda$ with probability at least $1 - \alpha\tau$,where

$$\tau = \{1 + \frac{1}{\log(p/\alpha)}\}\frac{\exp[2\log(2p/\alpha)\ell\{log(1-\alpha)/n\}^{1/2}]}{1 - \ell\{\log(1/\alpha)/n\}^{1/2}} - \alpha^{\ell^2/4-1},$$

when under Condition 1,$\tau = 1 + o(1)$ by setting $\ell \to \infty$, $\ell = o[n^{1/2}/\{\log(p/\alpha)\log^{1/2}(1/\alpha)\}]$ as $n \to \infty$.
(ii)Assume $p/\alpha > 8$ and $n > 4\log(2/\alpha)$. Then

$$\Lambda_\Phi(1-\alpha|X) \leq \nu\{2n\log(2p/\alpha)\}, \nu = \frac{\{1 + 2/\log(2p/\alpha)\}^{1/2}}{1 - 2\{\log(2/\alpha)/n\}^{1/2}}$$

where under Condition 1, $\nu = 1 + o(1)$ as $n \to \infty$.

Choose of Penalty Level

- In the nonnormal case, the semi-exact option of λ is:

$$\lambda = c \max_{F \in \mathcal{F}} \Lambda_F(1 - \alpha | X) \tag{11}$$

lemma2

(i) The exact option implements $\lambda \geq c\Lambda$ with probability at least $1 - \alpha$, if $F_0 = F$.
(ii) The semi-exact option implements $\lambda \geq c\Lambda$ with probability at least $1 - \alpha$, if either $F_0 \in \mathcal{F}$ or $\Lambda_F(1 - \alpha|X) \geq \Lambda_{F_0}(1 - \alpha|X)$ for some $F \in \mathcal{F}$.
Suppose further that Condition 2 and 3 hold. Then :
(iii) the asymptotic option implements $\lambda \geq c\Lambda$ with probability at least $1 - \alpha - o(\alpha)$,
(iv) the magnitude of the penalty level of the exact and semi-exact options satisfies the inequality

$$\max_{F \in \mathcal{F}} \Lambda_F(1 - \alpha|X) \leq \{2n \log(2p/\alpha)\}^{1/2}\{1 + o(1)\}, n \to \infty \tag{12}$$

Notes

- The proof of LEMMA 1 and LEMMA 2 is cumbersome, including the use of **Chernoff tail bound**, **Rosenthal's inequality** and **Vonbahr-Esseen's inequalities.**
- The Conditions 2 and 3 are only one possible set of sufficient conditions That guarantees the Gaussian-like conclusions of LEMMA 2, using moderate deviation theory of Slastnikov(1982).

- In order to figure out the asymptotic bounds on the estimation error $\hat{\delta} = \hat{\beta} - \beta_0$ in the Euclidean norm $\|\hat{\delta}\|_2 = (\delta'\delta)^{1/2}$, we try to estimate $\|\delta\|_{2,n}$ under the restricted eigenvalues condition,where:

$$\|\delta\|_{2,n} = \left[E_n\left\{\left(x'\delta\right)^2\right\}\right]^{1/2} = \left\{\delta' E_n\left(xx'\right)\delta\right\}^{1/2}$$

- And the restricted set $\Delta_{\bar{c}}$ can be derived from the $\lambda \geqslant a\Lambda$

$$\Delta_{\bar{c}} = \{\delta \in \mathbb{R}^p : \|\delta_{T^c}\|_1 \leqslant \bar{c}\,\|\delta_T\|_1\,, \delta \neq 0\}\,, \quad \bar{c} = \frac{c+1}{c-1}$$

Asymptotic Bounds on Estimation Error

- To connect $\|\delta\|_{2,n}$ and $\|\hat{\delta}\|$,we define the following restricted eigenvalues of the Gram matrix $E_n(xx')$:

$$\kappa_{\bar{c}} = \min_{\delta \in \Delta_{\bar{c}}} \frac{s^{1/2}\|\delta\|_{2,n}}{\|\delta_T\|_1}, \quad \tilde{\kappa}_{\bar{c}} = \min_{\delta \in \Delta_{\bar{c}}} \frac{\|\delta\|_{2,n}}{\|\delta\|_2}$$

- **Condition 4**[2] There exist finite constants $n_0 > 0$ and $\kappa > 0$, such that the restricted eigenvalues obey $\kappa_{\bar{c}} \geqslant \kappa$ and $\tilde{\kappa}_{\bar{c}} \geqslant \kappa$ for all $n > n_0$
- Moreover, let $m = s \log n$, there exist n', s.t $\forall n > n'$:

$$0 < k \leqslant \min_{\|\delta_{T^c}\|_0 \leqslant m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|_2^2} \leqslant \max_{\|\delta_{T^c}\|_0 \leqslant m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|_2^2} \leqslant k' < \infty \tag{13}$$

[2] The Condition 4 and sufficiency of (13) follows from Bickel et al. (2009):

- Consider the basic model described as below,we have the following theorem:
  - $y_i = x_i'\beta_0 + \sigma\epsilon_i \quad (i = 1, \ldots, n)$
  - $E_{F_0}(\epsilon_i) = 0, \quad E_{F_0}(\epsilon_i^2) = 1$
  - $T = \operatorname{supp}(\beta_0)$ has $s < n$ elements
  - $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1 \quad (j = 1, \ldots, p)$

Theorem

Let $c > 1, \bar{c} = (c+1)/(c-1)$, and suppose that λ obeys the growth restriction $\lambda s^{1/2} \leqslant n\kappa_{\bar{c}}\rho$, for some $\rho < 1$. If $\lambda \geqslant c\Lambda$, then

$$\left\|\hat{\beta} - \beta_0\right\|_{2,n} \leqslant A_n \sigma \left\{E_n\left(\epsilon^2\right)\right\}^{1/2} \frac{\lambda s^{1/2}}{n}, \qquad \textit{where } A_n = \frac{2(1+1/c)}{\kappa_{\bar{c}}\left(1-\rho^2\right)}$$

Target: $\|\hat{\beta} - \beta\|_2 \lesssim \sigma\{s\log(2p/\alpha)/n\}^{1/2}$

Several Key steps in theorem proof:

- $\{\hat{Q}(\hat{\beta})\}^{1/2} - \left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2} \leqslant \frac{\lambda}{n}\left\|\beta_0\right\|_1 - \frac{\lambda}{n}\|\hat{\beta}\|_1 \leqslant \frac{\lambda}{n}\left(\left\|\hat{\delta}_T\right\|_1 - \left\|\hat{\delta}_{T^c}\right\|_1\right)$
- $\{\hat{Q}(\hat{\beta})\}^{1/2} - \left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2} \geqslant -\|\tilde{S}\|_\infty\|\hat{\delta}\|_1 \geqslant -\frac{\lambda}{cn}\left(\left\|\hat{\delta}_T\right\|_1 + \left\|\hat{\delta}_T\right\|_1\right)$
- $\hat{Q}(\hat{\beta}) - \hat{Q}\left(\beta_0\right) = \|\hat{\delta}\|_{2,n}^2 - 2E_n\left(\sigma\epsilon x'\hat{\delta}\right),$

  $2\left|E_n\left(\sigma\epsilon x'\hat{\delta}\right)\right| \leqslant 2\left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2}\|\tilde{S}\|_\infty\|\hat{\delta}\|_1$
- $\hat{Q}(\hat{\beta}) - \hat{Q}\left(\beta_0\right) = \left[\{\hat{Q}(\hat{\beta})\}^{1/2} + \left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2}\right]\left[\{\hat{Q}(\hat{\beta})\}^{1/2} - \left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2}\right]$
- $\{\hat{Q}(\hat{\beta})\}^{1/2} \leqslant \left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2} + \frac{\lambda}{n}\left(\frac{s^{1/2}\|\hat{\delta}\|_{2,n}}{\kappa_{\bar{c}}}\right)$

Finally,we can derive the theorem:

$$\left\{1 - \left(\frac{\lambda s^{1/2}}{n\kappa_{\bar{c}}}\right)^2\right\}\|\hat{\delta}\|_{2,n}^2 \leqslant 2\left(\frac{1}{c} + 1\right)\left\{\hat{Q}\left(\beta_0\right)\right\}^{1/2}\frac{\lambda s^{1/2}}{n\kappa_{\bar{c}}}\|\hat{\delta}\|_{2,n}$$

Estimation Error with Related Corollary

- Based on the Theorem, we have the following corollary for different cases:[3]

Corollary

Consider the model described in $(1)-(4)$. *Suppose further that* $F_0 = \Phi, \lambda$ *is chosen according to the exact option:*
$\lambda = c\Lambda_\Phi(1-\alpha \mid X), \lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi(1-\alpha/2p)$
and the related condition are satisfied, then with probability at least $1-\alpha-\gamma$

$$\tilde{\kappa}_{\bar{c}} \left\| \hat{\beta} - \beta_0 \right\|_2 \leqslant \left\| \hat{\beta} - \beta_0 \right\|_{2,n} \leqslant B_n \sigma \left\{ \frac{2s \log(2p/\alpha)}{n} \right\}^{1/2}$$

where $B_n = \frac{2(1+c)v\omega}{\kappa_{\bar{c}}(1-\rho^2)}$

[3] Recall Lemma: $\Lambda_\Phi(1-\alpha \mid X) \leq \nu\{2n\log(2p/\alpha)\}$

Corollary

Consider the model described in $(1)-(4)$. *Suppose further that* $F_0 = \Phi$,*Conditions 4 and 1 hold, and* $(s/n)\log(p/\alpha) \to 0$, *as* $n \to \infty$,*There is an o(1) term such that with probability at least* $1-\alpha-o(1)$

$$\kappa \left\| \hat{\beta} - \beta_0 \right\|_2 \leqslant \left\| \hat{\beta} - \beta_0 \right\|_{2,n} \leqslant C_n \sigma \left\{ \frac{2s\log(2p/\alpha)}{n} \right\}^{1/2}$$

, *where* $C_n = \frac{2(1+c)}{\kappa\{1-o(1)\}}$

Corollary

Consider the model described in $(1)-(4)$. *Let* λ *be specified according to the asymptotic, exact or semi-exact option as following:*[a]

$\lambda = c\Lambda_F(1-\alpha \mid X), \quad \lambda = c \max_{F\in\mathcal{F}} \Lambda_F(1-\alpha \mid X)$

$\lambda = c\Lambda(1-\alpha) = cn^{1/2}\Phi^{-1}(1-\alpha/2p)$,*There is an* $o(1)$ *term such that with probability at least* $1-\alpha-o(1)$

$$\kappa \left\|\hat\beta-\beta_0\right\|_2 \leqslant \left\|\hat\beta-\beta_0\right\|_{2,n} \leqslant C_n\sigma\left\{\frac{2s\log(2p/\alpha)}{n}\right\}^{1/2}$$

, *where* $C_n = \frac{2(1+c)}{\kappa\{1-o(1)\}}$

[a] Recall Lemma2
$\max_{F\in\mathcal{F}} \Lambda_F(1-\alpha \mid X) \leq \{2n\log(2p/\alpha)\}^{1/2}\{1+o(1)\}, n\to\infty$

Computational Properties

- The original object function is:

$$\underset{\beta \in \mathbb{R}^p}{Min} \{\hat{Q}(\beta)^{1/2} + \frac{\lambda}{n} \|\beta\|_1\}, \hat{Q}(\beta) = \frac{\Sigma_{i=1}^{n} (y_i - x_i^{'} \beta)^2}{n} \tag{14}$$

- We have:

$$\beta_j^+ = max(\beta_j, 0), \quad \beta_j^- = -min(\beta_j, 0)$$
$$\beta = \beta^+ - \beta^-, \quad ||\beta||_1 = \Sigma_{j=1}^{p} (\beta_j^+ + \beta_j^-)$$
$$v_i = y_i - x_i^{'} \beta^+ + x_i^{'} \beta^-, \quad \hat{Q}(\beta)^{1/2} = \frac{||v||}{n^{1/2}}$$
$$Q^{n+1} = \{(v, t) \in R^n \times R : t \geq ||v||\}$$

Computational Properties

- Thus we can rewrite the object function(14) as

$$\min_{t,v,\beta^+,\beta^-} \frac{t}{n^{1/2}} + \frac{\lambda}{n}\Sigma_{i=1}^{p}(\beta_j^+ + \beta_j^-) \tag{15}$$
$$=(\frac{1}{n^{1/2}}, \frac{\lambda}{n}, ..., \frac{\lambda}{n})(t, \beta_1^+, ..., \beta_p^+, \beta_1^-, ..., \beta_p^-)'$$

- The standard Conic Programming Problem:

$\min_u c^{'} u$ subject to $Au = b \quad u \in C$ where C is a Cone.

- The second order conic programming problem:

$$\min_u c^{'} u, ||Au + b||_2 \leq a^{'} u + d$$

Computational Properties

- It is easy to transform (15) into a second order conic programming problem form.In fact this is the method to solve square root Lasso.
- Furthermore, Conic Programming has a tractable dual form and we write the dual problem of (15) below

$$\max_{a \in R^n} \frac{1}{n} \Sigma_{i=1}^n y_i a_i, |\Sigma_{i=1}^n x_{ij} a_i / n| \leq \lambda, ||a|| \leq n^{1/2}. \tag{16}$$

and the optimal $\hat{a}_i$ equal the residuals $y_i - x_i^{'} \hat{\beta}$ up to a renormalization factor.

Theorem 2

The square-root lasso problem in (14)(with solution $\hat{\beta}$) is equivalent to the conic programming problem(15)(with solution $\hat{\beta^+}$,$\hat{\beta^-}$,$\hat{t}$),which admits the strongly dual problem in (16)(with solution $\hat{a}$).
if $y = X\hat{\beta} \neq 0$ we have:$\hat{\beta} = \hat{\beta^+} - \hat{\beta^-}$,$\hat{v}_i = y_i - x_i'\hat{\beta}$ and $\hat{a} = n^{1/2}\hat{v}/||\hat{v}||$

Experiments:square-root lasso performance

Now we focus on the performance of square-root lasso on a test we set the parameters as follow

- $1 - \alpha = 0.95$
- c=1.1
- n=100,p=500
- $\beta_0 = (1, 1, 1, 1, 1, 0, 0..)$
- $x_i \sim N(0, \Sigma)$ with the Toeplitz correlation $\Sigma_{jk} = (1/2)^{|j-k|}$
- $X = (x_1^{'}, ...x_n^{'})^{'}$
- $y_i = x_i^{'}\beta_0 + \sigma\epsilon_i$
- $\epsilon_i \sim F_0$

Some indicator to evaluate the square-root Lasso performance:

- **Relative empirical risk**: $\frac{E(||\hat{\beta}-\beta_0||_{2,n})}{E(||\beta^*-\beta_0||_{2,n})}$ where β^* is the oracle estimator with known true support of β_0 (just use OLS)
- **The average number of regressors selected outside the true model**:$E\{|supp(\hat{\beta}) \setminus supp(\beta_0)|\}$
- **The average number of regressors missed from the true model**: $E|supp(\beta_0) \setminus supp(\hat{\beta})|$

Experiments $\epsilon_i \sim N(0,1)$

- Choose λ for square-root lasso:

$$\lambda_{sl} = cn^{1/2}\Phi^{-1}(1-\alpha/2p)$$

- Choose λ for Lasso (the penalty level in the package glmnet):

$$\lambda_l = \frac{c}{n^{1/2}}\sigma\Phi^{-1}(1-\alpha/2p)$$

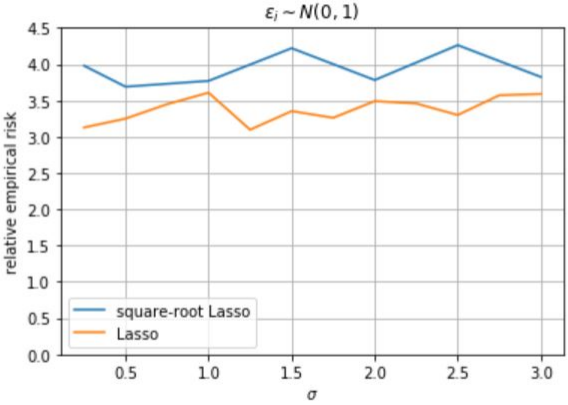
Experiments $\epsilon_i \sim N(0, 1)$

- To do Square-root regression we use package *picos* in python,transforming the regression problem into a second order conic programming problem.
- Use *glmnet* package in R to do Lasso regression

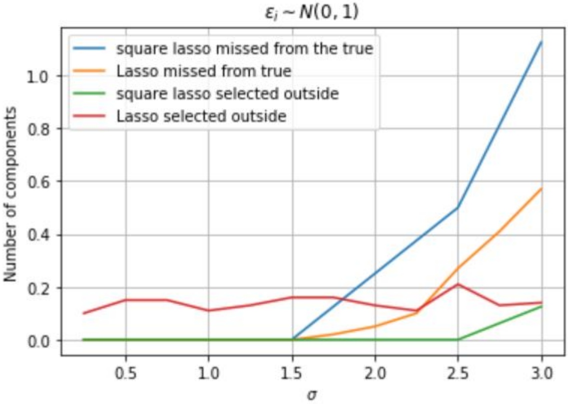
```
Second Order Cone Program
  minimize t/10 + ⟨Lambda/n,  belta+⟩ + ⟨Lambda/n,  belta-⟩
  over
    1×1 real variable t
    100×1 real variable V
    500×1 real variable belta+,  belta-
  subject to
    y[i] - x[i].T • belta+ + x[i].T • belta- = V[i] ∀ i ∈ [0…99]
    ‖V‖  ⩽ t
    belta+ ⩾ 0
    belta- ⩾ 0
```

Figure: Process of SOCP

Calculate **the relative empirical risk**: $\frac{E(||\hat{\beta}-\beta_0||_{2,n})}{E(||\beta^*-\beta_0||_{2,n})}$



Calculate **the average number of regressors missed from the true model** and **the average number of regressors selected outside the true model**:



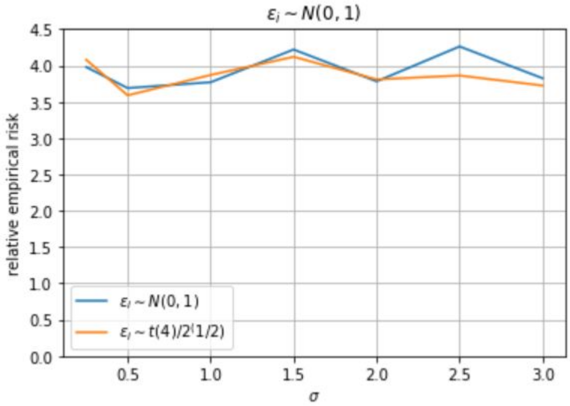
Choose penalty level λ for square-root Lasso,this time the distribution of ϵ_i is not N(0,1) use the formula:

$$\lambda = c \bigwedge_F (1 - \alpha | X),$$

$\bigwedge_F$ is a random variable related to ϵ . We can do simulation to calculate λ

Experiments $\epsilon_i \sim t(4)/2^{(1/2)}$

Calculate **the relative empirical risk**: $\frac{E(||\hat{\beta}-\beta_0||_{2,n})}{E(||\beta^*-\beta_0||_{2,n})}$



Experiments $\epsilon_i \sim t(4)/2^{(1/2)}$

Calculate **the average number of regressors missed from the true model** and **the average number of regressors selected outside the true model** :