Regularization methods in multiple regression

March 21, 2024

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

$Y = (Y_1, \dots, Y_n)^T$ - wektor of trait values for n individuals

$X_{n \times p}$ - matrix of regressors

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β is not unique

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β is not unique

Ridge regression:

$$\hat{\beta} = argmin_{\beta \in R^p} L(b) \ \ \text{, where} \ \ L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β is not unique

Ridge regression:

$$\hat{\beta} = argmin_{\beta \in R^p} L(b) \quad \text{, where } \; L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

$$\frac{\partial L(b)}{\partial b} = -2X'(Y - Xb) + 2\gamma b = 0$$

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β is not unique

Ridge regression:

$$\hat{\beta} = argmin_{\beta \in R^p} L(b) \quad \text{, where} \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

$$\frac{\partial L(b)}{\partial b} = -2X'(Y - Xb) + 2\gamma b = 0$$

$$-X'Y + (X'X + \gamma I)b = 0 \quad \Leftrightarrow \quad b = (X'X + \gamma I)^{-1} X'Y$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \; M = X(X'X + \gamma I)^{-1} X'$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \; M = X(X'X + \gamma I)^{-1} X'$$

$$Tr[M] = Tr\left[(X'X + \gamma I)^{-1} X'X\right]$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \; M = X(X'X + \gamma I)^{-1} X'$$

$$Tr[M] = Tr\left[(X'X + \gamma I)^{-1} X'X\right]$$

$$Tr[M] = \sum_{i=1}^{n} \lambda_i(M), \quad \text{where} \quad \lambda_1(M), \ldots, \lambda_n(M) \quad \text{are eigenvalues of} \quad M$$

Eigenvalues of M

$$X'Xu = \lambda u$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \quad Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

$$\hat{P}E = RSS + 2\sigma^2 \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

Ridge regression - orthogonal design

$$X'X = I, \;\; \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

$$= \frac{\gamma^2}{(1+\gamma)^2}\beta_i^2 + \frac{\sigma^2}{(1+\gamma)^2}$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

$$= \frac{\gamma^2}{(1+\gamma)^2}\beta_i^2 + \frac{\sigma^2}{(1+\gamma)^2}$$

$$E||\hat{\beta} - \beta||^2 = \frac{\gamma^2}{(1+\gamma)^2}||\beta||^2 + \frac{p\sigma^2}{(1+\gamma)^2}$$

Ridge regression - orthogonal design (2)

When ridge is better than LS ?

Ridge regression - orthogonal design (2)

When ridge is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge regression - orthogonal design (2)

When ridge is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Ridge regression - orthogonal design (2)

When ridge is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Otherwise, when

$$||\beta||^2 < \frac{\gamma+2}{\gamma} p\sigma^2$$

Ridge regression - orthogonal design (2)

When ridge is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Otherwise, when

$$||\beta||^2 < \frac{\gamma + 2}{\gamma} p\sigma^2$$

$$\gamma < \frac{2p\sigma^2}{||\beta||^2 - p\sigma^2}$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \ \text{ then } \ \|\gamma\|_1 > \|\beta\|_1.$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^n |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \ \text{ then } \ \|\gamma\|_1 > \|\beta\|_1.$$

$$k = ||\beta||_0 = \#\{i: \ \beta_i \neq 0\}$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \;\text{ then }\; \|\gamma\|_1 > \|\beta\|_1.$$

$$k = ||\beta||_0 = \#\{i : \; \beta_i \neq 0\}$$

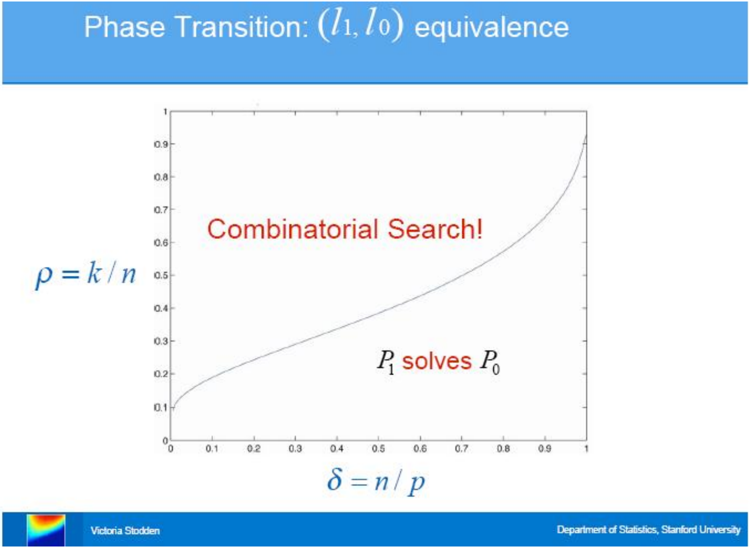
Basis Pursuit can recover β if k is small enough.

Transition curve (Donoho and Tanner, 2005)

Let's assume than $p \to \infty$, $n/p \to \delta$ and $k/n \to \epsilon$.

If X_{ij} are iid $N(0, \tau^2)$ then the probability that BP recovers β converges to 1 if $\epsilon < \rho(\delta)$ and to 0 if $\epsilon > \rho(\delta)$, where $\rho(\delta)$ is the *transition curve*.

Transition curve (2)



Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \epsilon$

Or alternatively: $\min_{b\in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda||b||_1$

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \epsilon$

Or alternatively: $\min_{b\in R^p} \frac{1}{2}||y - Xb||_2^2 + \lambda||b||_1$

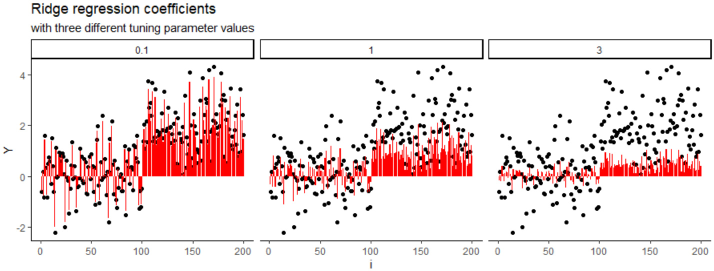
BPDN (Chen and Donoho, 1994) or LASSO (Tibshirani, 1996)

Solution to orthogonal design, ridge

- The Solution for ridge is given by

$$\hat{\beta}_i^{ridge} = \frac{y_i}{1+\lambda} = \frac{\hat{\beta}_i^{LS}}{1+\lambda}.$$

  - Leads to a shrinkage by a factor $\frac{1}{1+\lambda}$ of the coefficients.

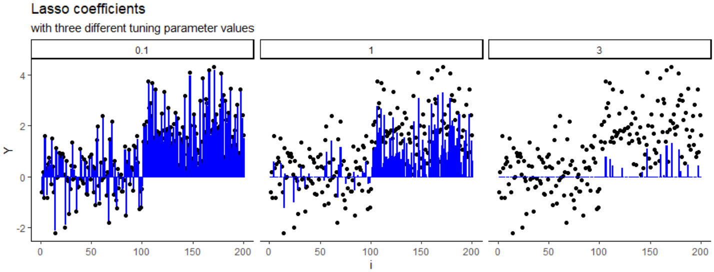


Solution to orthogonal design, lasso

- The Solution for lasso is given by

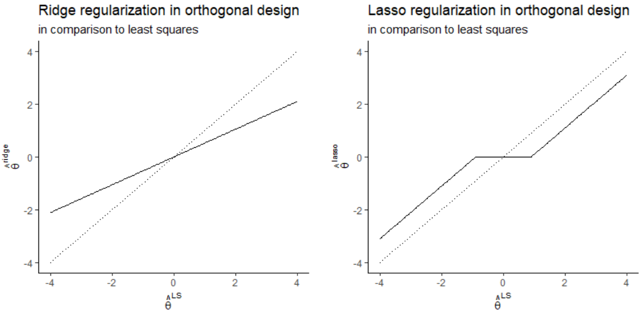
$$\hat{\beta}_i^{lasso} = sign(y_i)\left(|y_i| - \lambda\right)_+ = sign(\hat{\beta}_i^{LS})\left(|\hat{\beta}_i^{LS}| - \lambda\right)_+$$

  - Set small values to exactly zero (sparse solution)
  - Fixed shrinkage of λ for non-zero coefficients.



Solution to orthogonal design, joint

- Common representation: Plot relation between $\hat{\beta}^{LS}$ and $\hat{\beta}^{ridge}$ / $\hat{\beta}^{lasso}$



Selection of the tuning parameter for LASSO

- General rule: the reduction of λ_L results in identification of more elements from the true support (true discoveries) but at the same time it produces more falsely identified variables (false discoveries)
- The choice of λ_L is challenging- e.g. crossvalidation typically leads to many false discoveries
- When $X^T X = I$ Lasso selects X_j iff $|\hat{\beta}_j^{LS}| > \lambda$
- Selection $\lambda = \sigma \Phi^{-1}(1 - \alpha/(2p)) \approx \sigma\sqrt{2 \log p}$ corresponds to Bonferroni correction and controls FWER.

Application

- One of the main advantages of Lasso and ridge is thar they can be use when $p > n$.

Application

- One of the main advantages of Lasso and ridge is thar they can be use when $p > n$.
- A typical application is genetics. Here
  - X_i- samples were scanned with a microarray, that measures the expression of 10000s of genes simultaneously.
  - y_i- time for severe breast cancer to metastasize.
  - The goal is to identify patients with poor prognosis in order to administer more aggressive follow-up treatment for them.

Application

- One of the main advantages of Lasso and ridge is thar they can be use when $p > n$.
- A typical application is genetics. Here
  - $X_i -$ samples were scanned with a microarray, that measures the expression of 10000s of genes simultaneously.
  - $y_i -$ time for severe breast cancer to metastasize.
  - The goal is to identify patients with poor prognosis in order to administer more aggressive follow-up treatment for them.
- Two typical genetic "models"
  - quantitative trait loci (QTL) a single or few important gene.
  - Polygene: many genes with small individual effect.

Which model is lasso which model is ridge?

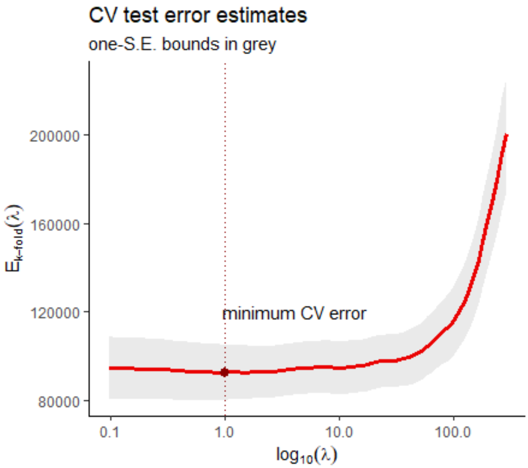
$p >> n$

https://www.mv.helsinki.fi/home/mjxpirin/HDS$_{c}ourse/material/vantveer.txt$
D = read.table("vantveer.txt", header = T)
print(dim(D))
X = as.matrix(D[,2:ncol(D)])
y = D$Months

```
## [1]    98 24189
```

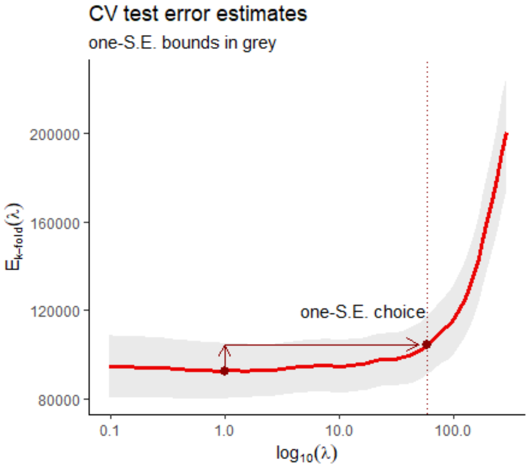
Choosing the regularization parameter

- Cross-validation for choosing λ.
- Often the cross validated error is often flat
- Prediction favours more parameters



Choosing the regularization parameter II

- A solution is instead choose a more regularized solution that have similar error.
- "similar" = within one S.E.

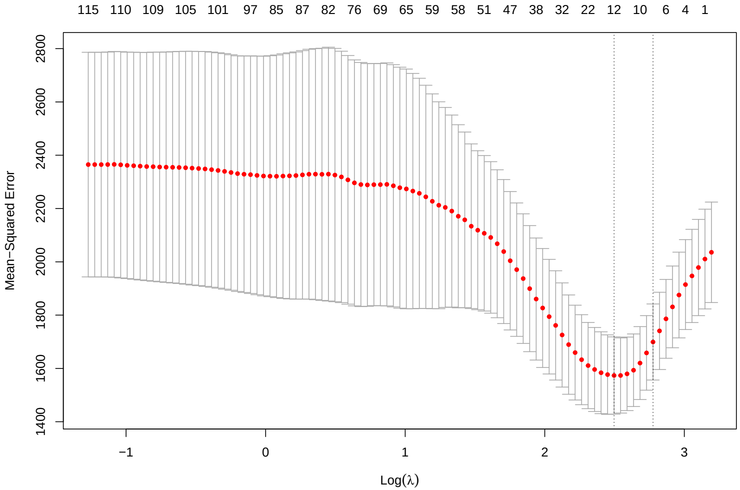


Choosing the regularization parameter (R)

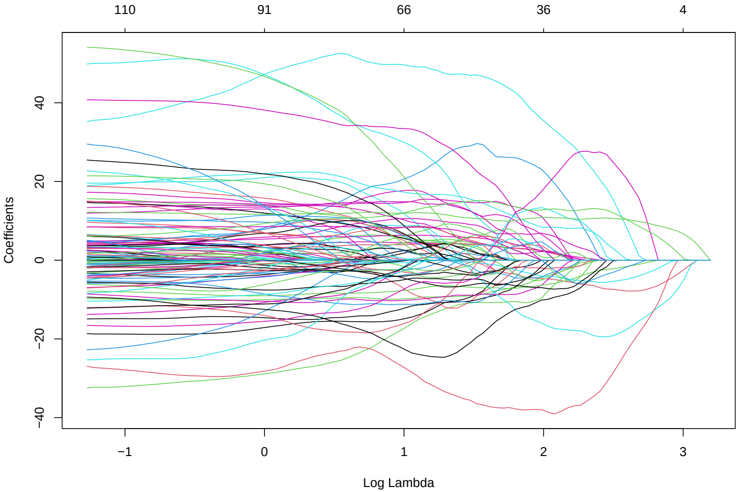
- Go to package in R **glmnet** for both lasso and ridge.
- Need to find λ, use k-fold cross-validation.

[] library(glmnet) cvfit <- cv.glmnet(X, y, alpha=1, nfolds = 10, intercept = T, standardize = T)
plot(cvfit)

fit λ II (lasso)

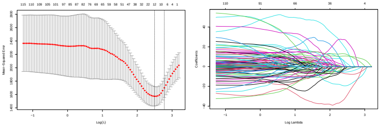


fit λ III (lasso)



Under the hood

- For each cross validation sample on solves an entire path $\lambda \in \{100, 10, 1, 0.1..., \}$
- Recall that $p = 24189$, thus for each $\lambda \in \{100, 10, 1, 0.1..., \}$ and each cross-validation set we need to estimate 24189 parameters to optimum.



Under the hood

- For each cross validation sample on solves an entire path $\lambda \in \{100, 10, 1, 0.1..., \}$
- Recall that $p = 24189$, thus for each $\lambda \in \{100, 10, 1, 0.1..., \}$ and each cross-validation set we need to estimate 24189 parameters to optimum.

The Hessian Screening Rule

Johan Larsson
Department of Statistics
Lund University
johan.larsson@stat.lu.se

Jonas Wallin
Department of Statistics
Lund University
jonas.wallin@stat.lu.se

The Strong Screening Rule for SLOPE

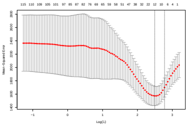
Johan Larsson
Dept. of Statistics, Lund University
johan.larsson@stat.lu.se

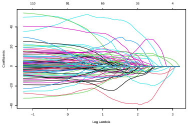
Małgorzata Bogdan
Dept. of Mathematics, University of Wroclaw
Dept. of Statistics, Lund University
malgorzata.bogdan@uwr.edu.pl

Jonas Wallin
Dept. of Statistics, Lund University
jonas.wallin@stat.lu.se

-

Abstract



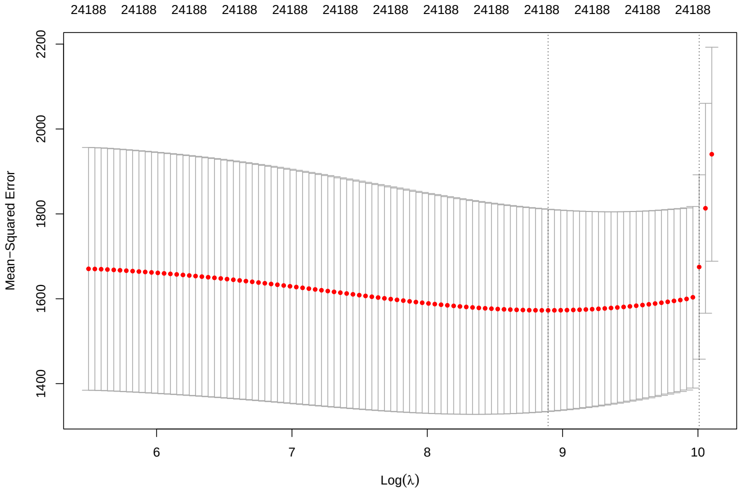


ridge

- Then ridge ($\alpha = 0$), fitting λ the same way

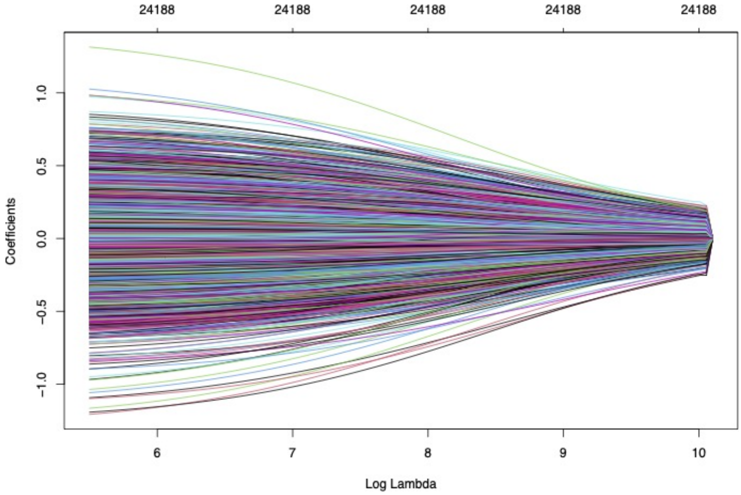
[] cvfit <- cv.glmnet(X, y, alpha=0, nfolds = 10, intercept = T, standardize = T) plot(cvfit)

ridge II



24188 24188 24188 24188 24188 24188 24188 24188 24188 24188 24188 24188 24188
Mean−Squared Error
2200
2000
1800
1600
1400
6
7
8
9
10
Log(λ)

ridge III



24188
24188
24188
24188
24188
Coefficients
1.0
0.5
0.0
−0.5
−1.0
6
7
8
9
10
Log Lambda

LASSO Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = 1_{x>0} - 1_{x<0}$

LASSO Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \dots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = 1_{x>0} - 1_{x<0}$

Let $I := \{i \in \{1, \dots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

LASSO Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = 1_{x>0} - 1_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentable condition:

$$\|X_{\bar{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

LASSO Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = 1_{x>0} - 1_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentable condition:

$$\|X_{\bar{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

When

$$\|X_{\bar{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty > 1$$

then probability of the support recovery by LASSO is smaller than 0.5 (Wainwright, 2009).

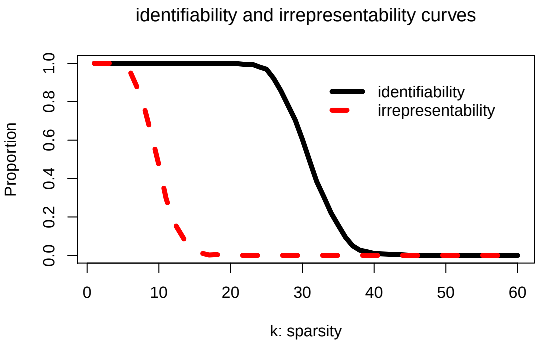
Separation of true and false predictors

Theorem (Tardivel, Bogdan, 2019)

For any $\lambda > 0$ LASSO can separate well the causal and null features if and only if vector β is identifiable with respect to l_1 norm and $min_{i \in I}|\beta_i|$ is sufficiently large.

Irrepresentability and identifiability curves

n=100, p=300, elements of X were generated as iid N(0,1)



Modifications of LASSO

Corollary

Appropriately thresholded LASSO can properly identify the sign of sufficiently large β if and only if β is identifiable with respect to l_1 norm.

Conjecture

Adaptive (reweighted) LASSO can properly identify the sign of sufficiently large β if and only if β is identifiable with respect to l_1 norm.

Adaptive LASSO

Adaptive LASSO [Zou, *JASA* 2006], [Candès, Wakin and Boyd, *J. Fourier Anal. Appl.* 2008]

$$\beta_{aL} = argmin_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda \sum_{i=1}^{p} w_i |b|_i \right\}, \tag{1}$$

where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.

Adaptive LASSO

Adaptive LASSO [Zou, *JASA* 2006], [Candès, Wakin and Boyd, *J. Fourier Anal. Appl.* 2008]

$$\beta_{aL} = argmin_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda \sum_{i=1}^{p} w_i |b|_i \right\}, \qquad (1)$$

where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.
Reduces bias and improves model selection properties

Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))

Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))

Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}

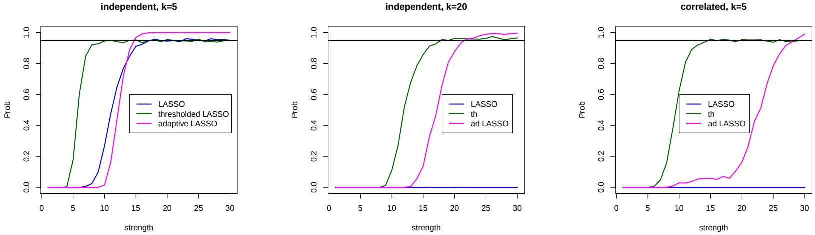
Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}
4. For adaptive LASSO - weights based on LASSO estimator with λ as in 2 and 3, selection based on LASSO with λ as in 1

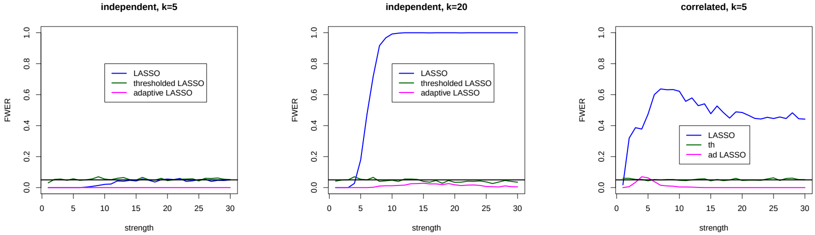
Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}
4. For adaptive LASSO - weights based on LASSO estimator with λ as in 2 and 3, selection based on LASSO with λ as in 1
5. Threshold selected by using knockoff control variables (Foygel-Barber and Candès, 2015; Candès, Fan, Janson, Lv, 2016)

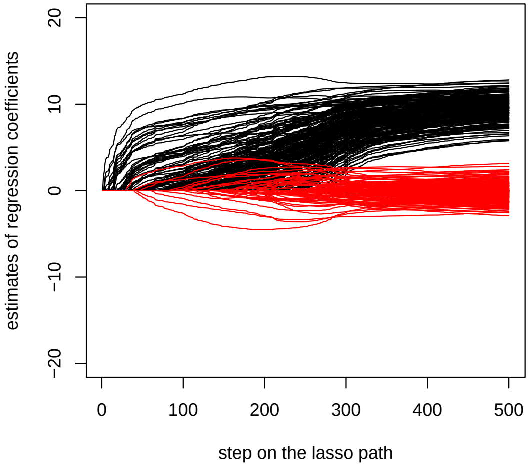
Probability of the sign recovery



Family Wise Error Rate



Thresholded LASSO (1)



Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

$\hat{\beta}(\lambda)$ - vector of $2p$ estimates of regression coefficients by LASSO applied on the augmented design matrix $X_{aug} = [X, \tilde{X}]$

Knockoffs and LCD statistics

Foygel-Barber and Candés (Ann. Stat. 2015), Candès, Fan, Janson and Lv (JRSSB, 2017) - augment X with the matrix $\tilde{X}$ of specifically constructed control variables

Necessary requirement:

$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.

When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

$\hat{\beta}(\lambda)$ - vector of $2p$ estimates of regression coefficients by LASSO applied on the augmented design matrix $X_{aug} = [X, \tilde{X}]$

$$W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{p+j}|$$

Knockoff filter

Define a random threshold as

$$\hat{t}(\lambda) = \min\left\{ t > 0 : \frac{1 + \#\{j : W_j(\lambda) \leq -t\}}{\#\{j : W_j(\lambda) \geq t\}} \leq q \right\}$$

and select

$$\widehat{\mathcal{S}(\lambda)} = \{j : W_j(\lambda) \geq \hat{t}(\lambda)\},$$

Knockoff filter

Define a random threshold as

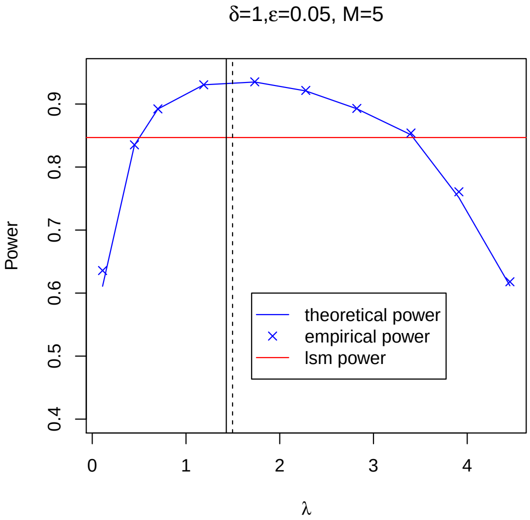
$$\hat{t}(\lambda) = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j(\lambda) \leq -t\}}{\#\{j : W_j(\lambda) \geq t\}} \leq q \right\}$$

and select

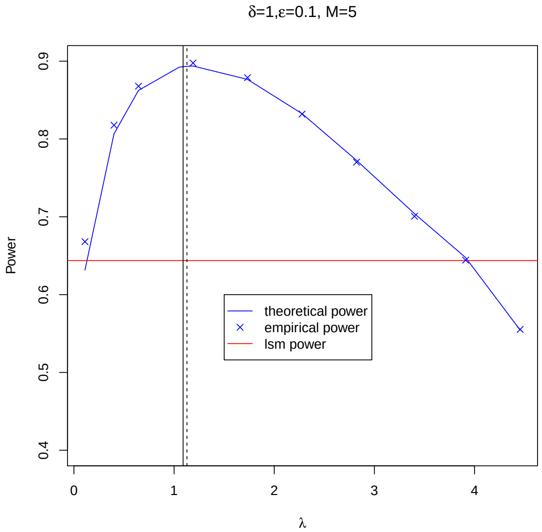
$$\widehat{\mathcal{S}(\lambda)} = \{j : W_j(\lambda) \geq \hat{t}(\lambda)\},$$

Foygel-Barber and Candès (2015), Candès, Fan, Janson and Lv (2017) - The above knockoff procedure $KN(\lambda, q)$ controls FDR at the level q.

Gain in power over LSM



Gain in power over LSM



Theoretical results using the mean field asymptotics

Su, Bogdan, Candès, Ann. Stat. 2017 - FDR-Power Tradeoff Diagram for LASSO

Theoretical results using the mean field asymptotics

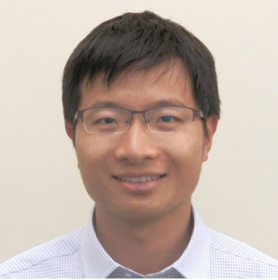
Su, Bogdan, Candès, Ann. Stat. 2017 - FDR-Power Tradeoff Diagram for LASSO

Weinstein, Su, Bogdan, Barber, Candés, Ann. Stat. 2023 - Breaking the tradoff diagram with thresholded LASSO

Sorted L-One Penalized Estimation

M.B., E.van den Berg, C.Sabatti, W.Su, E.J.Candès, AOAS 2015






Sorted L-One Penalized Estimation

$$\hat{\beta} = argmin_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

where $\left|b\right|_{(1)} \geq \ldots \geq \left|b\right|_{(p)}$ are ordered magnitudes of coefficients of b and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ is the sequence of tuning parameters.

Sorted L-One Penalized Estimation

$$\hat{\beta} = argmin_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

where $\left|b\right|_{(1)} \geq \ldots \geq \left|b\right|_{(p)}$ are ordered magnitudes of coefficients of b and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ is the sequence of tuning parameters.

The above optimization problem is convex and can be efficiently solved even for large design matrices.

Sorted L-One Penalized Estimation

$$\hat{\beta} = argmin_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}.$$

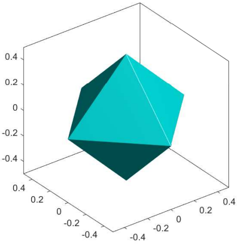
where $\left|b\right|_{(1)} \geq \ldots \geq \left|b\right|_{(p)}$ are ordered magnitudes of coefficients of b and $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ is the sequence of tuning parameters.

The above optimization problem is convex and can be efficiently solved even for large design matrices.

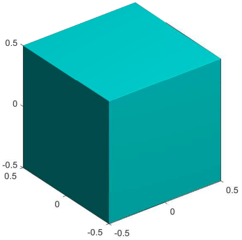
Sorted L-One Norm: $J_\lambda(b) = \sum_{i=1}^{p} \lambda_i \left|b\right|_{(i)}$ reduces to $||b||_1$ if $\lambda_1 = \ldots = \lambda_p$ and to $||b||_\infty$ if $\lambda_1 > \lambda_2 = \ldots = \lambda_p = 0$.

Unit balls for different SLOPE sequences by D.Brzyski

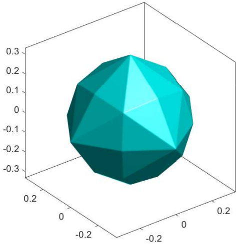
[(2,2,2)]



[(2,0,0)]



[(3,2,1)]



FDR control with SLOPE

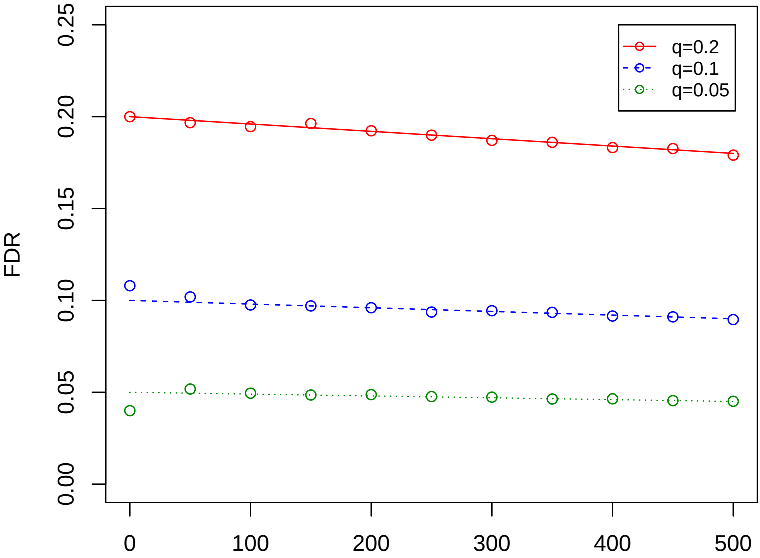
Theorem (B,van den Berg, Sabatti, Su and Candès (2015))

When $X^T X = I$ SLOPE with

$$\lambda_i := \sigma \Phi^{-1}\left(1 - i \cdot \frac{q}{2p}\right)$$

controls FDR at the level $q\frac{p_0}{p}$.

Orthogonal design, $n = p = 5000$



Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (2018, AOS):

SLOPE with the BH related sequence of tuning parameters attains minimax rate for the estimation error $||\hat{\beta} - \beta||^2$.

Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (2018, AOS):

SLOPE with the BH related sequence of tuning parameters attains minimax rate for the estimation error $||\hat{\beta} - \beta||^2$.

SLOPE rate of the estimation error - $k \log(p/k)$

LASSO rate of the estimation error - $k \log p$

Asymptotic optimality of SLOPE

Let $k = ||\beta||_0$ and consider the setup where $k/p \to 0$ and $\frac{k \log p}{n} \to 0$.

X is standardized so that each column has a unit L_2 norm.

Su and Candès (Annals of Statistics, 2016),

Bellec, Lecué, Tsybakov (2018, AOS):

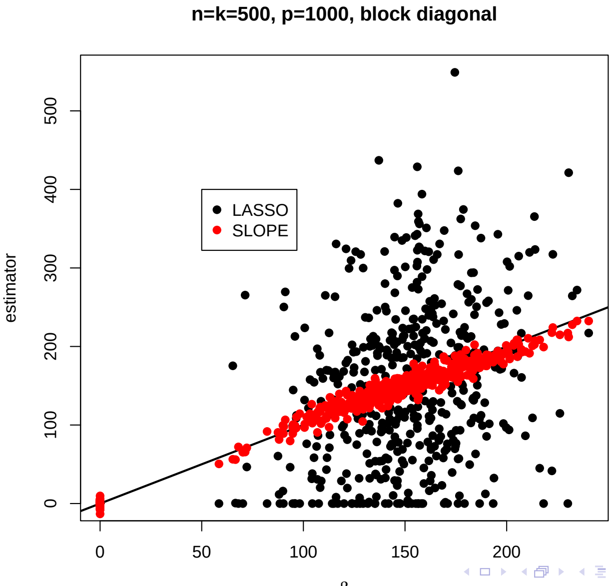
SLOPE with the BH related sequence of tuning parameters attains minimax rate for the estimation error $||\hat{\beta} - \beta||^2$.

SLOPE rate of the estimation error - $k \log(p/k)$

LASSO rate of the estimation error - $k \log p$

Extension to logistic regression by Abramovich and Grinshtein (2018, IEEE Trans. Inf. Theory)

SLOPE vs LASSO



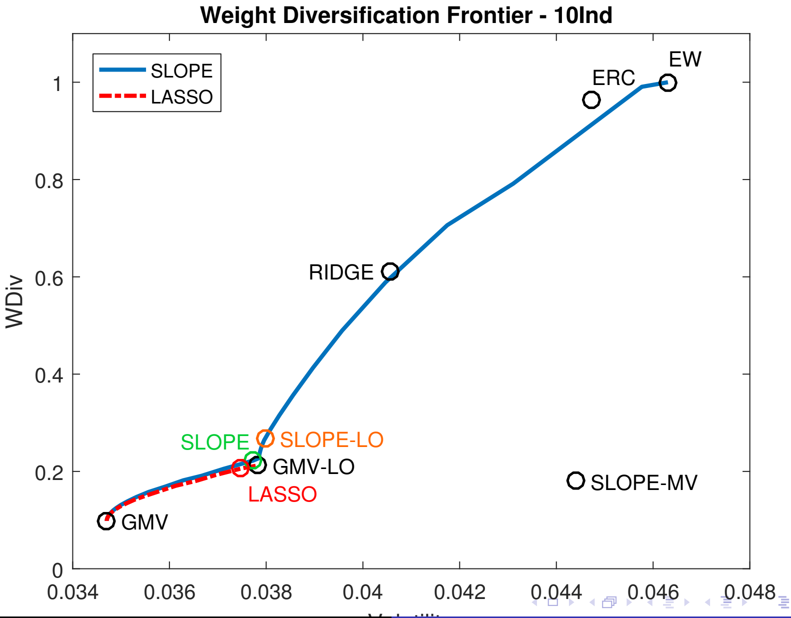
Portfolio Optimization, (P. Kremmer, S. Lee, M. Bogdan, S. Paterlini, JBF 2020)

$R_{t\times k} = (R_1, \dots, R_k)$ - asset returns,
Σ - the covariance matrix of R

$$\min_{w\in\mathbb{R}^k} \frac{\phi}{2} w'\Sigma w + J_\lambda(w) \tag{2}$$

$$\text{s.t.} \sum_{i=1}^{k} w_i = 1 \tag{3}$$

Evolution of Portfolio



Extensions

- Elastic net

$$ {}_{\beta}||y - X\beta||_2^2 + \lambda \left(\alpha|\beta|_1 + (1-\alpha)\, ||\beta||_2^2 \right) $$

where $w_j = \frac{1}{\hat{\beta}^{LS}}$

- Adaptive lasso:

$$ {}_{\beta}||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j| $$

where $w_j = \frac{1}{|\hat{\beta}^{LS}|^{\gamma}}$ or $w_j = \frac{1}{|\hat{\beta}^{ridge}|^{\gamma}}$

  - Sparser then lasso,
  - less regularization on parameters.

- Group lasso:

$$ {}_{\beta}||y - X\beta||_2^2 + \lambda \sum_{j=1}^{J} ||\beta_j||_2 $$

- The respective versions of SLOPE

The Bayesian connection I:ridge

- Ridge regression solution:

$$\hat{\beta}^{ridge} = argmin_{\beta} \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \frac{1}{2} \lambda \sum_{j=1}^{p} \beta_j^2 \quad (3.41)$$

The Bayesian connection I:ridge

- Ridge regression solution:

$$\hat{\beta}^{ridge} = argmin_{\beta} \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \frac{1}{2} \lambda \sum_{j=1}^{p} \beta_j^2 \quad (3.41)$$

- A Probabilistic interpretation of the regularization is

$$\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2 = \frac{\lambda}{2} \beta^T \beta = \frac{\lambda}{2} (\beta - 0)^T (\beta - 0).$$

Thus the ridge penalty can be considered a prior

$$\beta \sim \mathcal{N}\left(\beta; 0, \frac{1}{\lambda} I\right)$$

The Bayesian connection I:ridge

- Ridge regression solution:

$$\hat{\beta}^{ridge} = argmin_{\beta} \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \frac{1}{2} \lambda \sum_{j=1}^{p} \beta_j^2 \quad (3.41)$$

- A Probabilistic interpretation of the regularization is

$$\frac{\lambda}{2} \sum_{j=1}^{p} \beta_j^2 = \frac{\lambda}{2} \beta^T \beta = \frac{\lambda}{2} (\beta - 0)^T (\beta - 0) .$$

  Thus the ridge penalty can be considered a prior

$$\beta \sim \mathcal{N} \left(\beta; 0, \frac{1}{\lambda} I \right)$$

- And the ridge solution is thus the MAP (maximum a posteriori) estimate of:

$$\pi(\beta | y, \lambda) \propto \mathcal{N}(y; X\beta, I) \mathcal{N} \left(\beta; 0, \frac{1}{\lambda} I \right)$$

The Bayesian connection II:lasso

- Lasso:

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} \frac{1}{2} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3.52)$$

The Bayesian connection II:lasso

- Lasso:

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} \frac{1}{2} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3.52)$$

- A Probabilistic interpretation of the regularization

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

is that is log density of p independent variables with Laplace distributions

$$\beta \sim \prod_{i=1}^{p} Laplace \left(0, \frac{1}{\lambda} \right)$$

The Bayesian connection II:lasso

- Lasso:

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} \frac{1}{2} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (3.52)$$

- A Probabilistic interpretation of the regularization

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

is that is log density of p independent variables with Laplace distributions

$$\beta \sim \prod_{i=1}^{p} Laplace\left(0, \frac{1}{\lambda}\right)$$

- And the lasso is thus the MAP estimate of:

$$\pi(\beta|y, \lambda) \propto \mathcal{N}(y; X\beta, I) \prod_{i=1}^{p} Laplace\left(0, \frac{1}{\lambda}\right)$$