Statistical Foundations of Data Science

Jianqing Fan

Princeton University

https://fan.princeton.edu



2. Penalized Least-Squares

2.1. Classical model selection

2.2. Penalized least-squares

2.3. Lasso and L_1-regularization

2.4. Folded concave regularization

2.5. Concentration inequialities

2.6. Estimation of residual variance

2.7. Variable expansion & g-penalty

2.1 Classical Model Selection

Best subset selection

Data: $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ i.i.d. from $Y = \mathbf{X}^T\beta + \varepsilon$, $E(\varepsilon|\mathbf{X}) = 0$.

Matrix notation: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$.

Arts of Modeling: ★categorical (SES, age group, zipcode) ★polynomials ★splines ★additive model ★interaction models ★bivariate tensors ★multivariate tensors ★kernel tricks ★time-series

■Model size $p = \dim(\mathbf{X})$ can easily get very large

Objective: To select active components S and estimate β_S.

Best subset S_m: Minimizes RSS among $\binom{p}{m}$ possible subsets.

Relation with L_0-penalty

Traditional: L_0-penalty

$$n^{-1}\|\mathbf{y}-\mathbf{X}\beta\|^2+\lambda\sum_{j=1}^{p}\|\beta_j\|_0.$$

Computation: Best subset

- Given $\|\beta\|_0 = m$, the solution is the best subset S_m.
- Find m to minimize $n^{-1}\mathbf{RSS}_m+\lambda m$. NP hard, all subets.

Greedy algorithms: stepwise addition/deletion, stepwize, matching pursuit.

Theory: The method works well even for high-dimensional problems.

(*Birgé, Massart, Baron, 99; Shen, Pan, Zhu, 11)* see also **chap 4**

How to choose λ?

Stein identity and prediction error

Model: $Y_i = \mu(\mathbf{X}_i) + \varepsilon_i$. Let $\widehat{\mu}$ be an estimator

From $\|\mu - \widehat{\mu}\|^2 = \|\mathbf{Y} - \widehat{\mu}\|^2 - \|\mathbf{Y} - \mu\|^2 + 2(\widehat{\mu} - \mu)^T(\mathbf{Y} - \mu)$, we have

Stein identity: $\mathrm{E}\,\|\mu - \widehat{\mu}\|^2 = \mathrm{E}\left\{\|\mathbf{Y} - \widehat{\mu}\|^2 - n\sigma^2\right\} + 2\underbrace{\sum_{i=1}^n \mathrm{cov}(\widehat{\mu}_i, \mathbf{Y}_i)}_{\equiv \sigma^2\, df_{\widehat{\mu}}}$.

PE: $\mathrm{E}\,\|\mathbf{Y}^{\mathrm{new}} - \widehat{\mu}\|^2 = n\sigma^2 + \mathrm{E}\,\|\mu - \widehat{\mu}\|^2 = \mathrm{E}\left\{\|\mathbf{Y} - \widehat{\mu}\|^2\right\} + 2df_{\widehat{\mu}}\sigma^2$.

Unbiased estimate of PE. Let $df_{\widehat{\mu}} = \sigma^{-2}\sum_{i=1}^n \mathrm{cov}(\widehat{\mu}_i, \mathbf{Y}_i)$ model fixed

$$C_p(\widehat{\mu}) = \underbrace{\|\mathbf{Y} - \widehat{\mu}\|^2}_{\mathrm{RSS}} + 2\sigma^2\, df_{\widehat{\mu}}.$$

For linear predictor $\widehat{\mu} = \mathbf{SY}$, $df_{\widehat{\mu}} = \mathrm{tr}(\mathbf{S})$ since $\mathrm{cov}(\widehat{\mu}_i, Y_i) = s_{ii}\sigma^2$.

Efforts in model selection: choosing λ

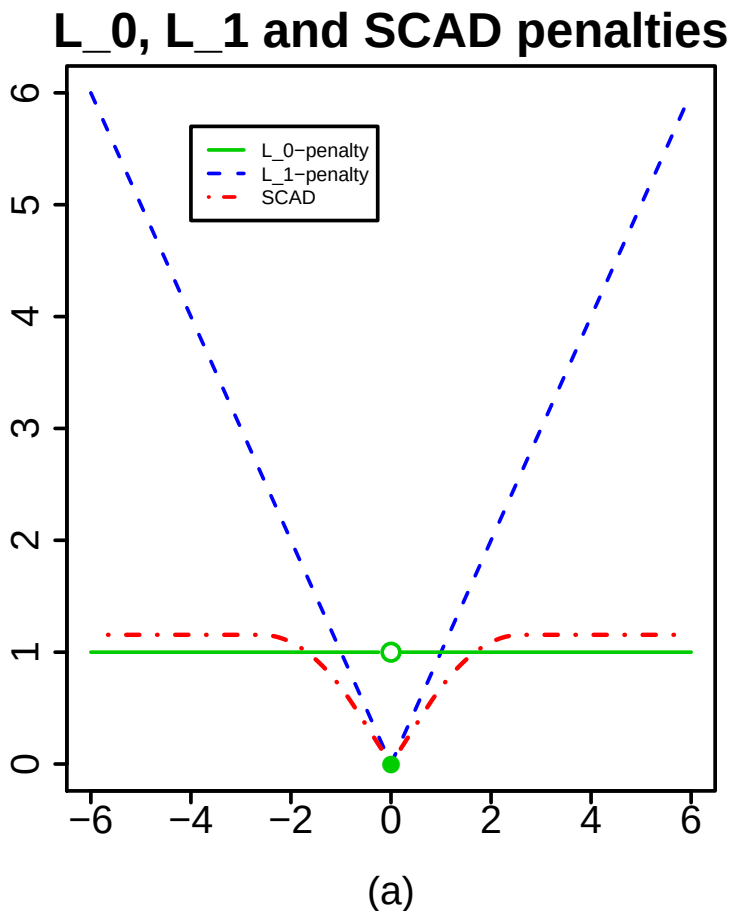
Best subset: Find m to minimize $n^{-1}\mathbf{RSS}_m + \lambda m\sigma^2$

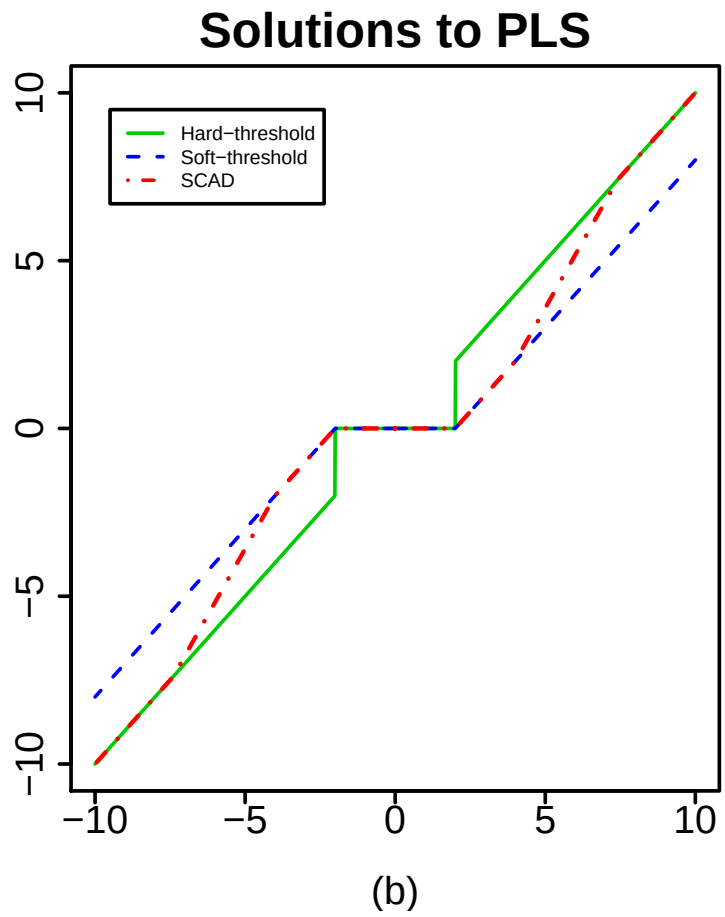
- C_p (*Mallows, 73*) and AIC (*Akaike, 74*): $\lambda = 2$, model $\widehat{\mathcal{M}}$
- $\mathrm{AIC}(\widehat{\mu}) = \log(\|\mathbf{Y} - \widehat{\mu}\|^2/n) + 2\, df_{\widehat{\mu}(\lambda)}/n.$ ★w/o σ^2 approx the same
- BIC (*Schwarz, 1978*): $\lambda = \log(n)$.
- RIC (*Foster & George, 1994*): $\lambda = 2\log(p)$
- Adjusted-R^2: $R_{adj,m} = 1 - \frac{n-1}{\mathbf{n-m}} \frac{\mathbf{RSS_m}}{\mathrm{SD}_y}$.
- Generalized cross-validation (*Craven and Wahba, 1979*): $\mathrm{GCV}(m) = \frac{\mathbf{RSS_m}}{\mathbf{n(1-m/n)^2}}$

2.2 Penalized Least-Squares

Convex and folded concave relaxations

Penalized least-squares: $Q(\beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$





Convex penalties

- ★ L_2 penalty $p_\lambda(|\theta|) = \lambda|\theta|^2 \Longrightarrow$ ridge regression
- ★ L_q-penalty $p_\lambda(|\theta|) = \lambda|\theta|^q \Longrightarrow$ Bridge reg (*Frank and Friedman, 93*).

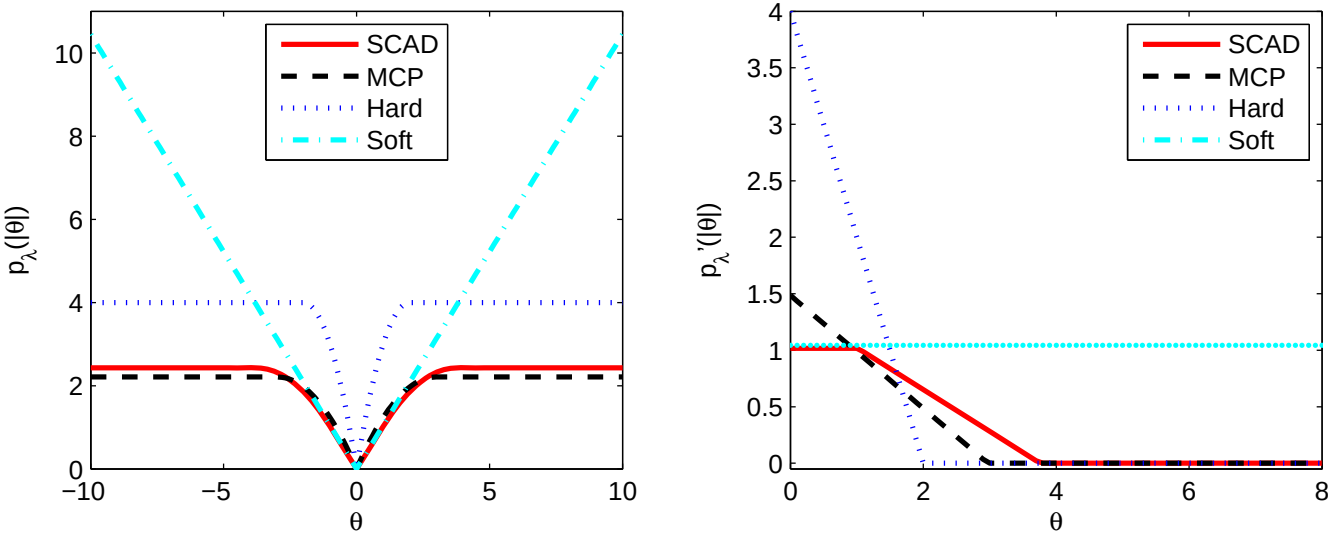
  $q \geq 1$ is convex, $q \in (0,1)$ is folded concave.
- ★ L_1 penalty $p_\lambda(|\theta|) = \lambda|\theta| \Longrightarrow$ **LASSO** (*Tibshirani 1996*).
- ★ Elastic net $p_\lambda(\theta) = \lambda_1|\theta| + \lambda_2\theta^2$ (*Zou & Hastie, 05*)

Folded Concave Penalty

Smoothly Clipped Absolute Deviation(SCAD)

$$p'_\lambda(\theta) = \lambda\left\{ I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}, \quad a > 2$$

(Fan, 1997). Set $a = 3.7$ from Bayesian viewpoint (Fan & Li, 01)



Minimum Concavity P. (MCP): $p'_\lambda(\theta) = (a\lambda - \theta)_+ / a$ *(Zhang,10)*.

$a = 1 \Longrightarrow$ Hard-Thresholding $p_\lambda(\theta) = \lambda^2 - (\lambda - |\theta|)_+^2$.

What is the role of penalty function?

Orthonormal design

LSE: $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}$, when $\mathbf{X}^T\mathbf{X} = I_p$.

$$Q(\beta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \frac{1}{2n}\sum_{j=1}^{p}(\widehat{\beta}_j - \beta_j)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

⟹ **comp** min. (*Antoniadis and Fan, 01; Fan and Li, 01*) solutions drawn on page 9

$$\mathbf{\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|)}$$

1. $p_\lambda(|\theta|) = \frac{\lambda}{2}\theta^2 \Longrightarrow$ **shrinkage**: $\widehat{\theta}_\lambda = (1+\lambda)^{-1}z$.
2. $p_\lambda(|\theta|) = \lambda|\theta| \Longrightarrow$ **soft-threshold**: $\widehat{\theta}_S = \mathrm{sgn}(z)(|z| - \lambda)_+$.
3. $p_\lambda(\theta) = \lambda^2 - (\lambda - |\theta|)_+^2 \Longrightarrow$ **hard-threshold** $\widehat{\theta}_H = zI(|z| > \lambda)$.

Solutions of SCAD and MCP

$$\widehat{\theta}_{\text{SCAD}}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \le 2\lambda; \\ \text{sgn}(z)[(a-1)|z| - a\lambda]/(a-2), & \text{when } 2\lambda < |z| \le a\lambda; \\ z, & \text{when } |z| \ge a\lambda. \end{cases} .$$

- $a = \infty \Longrightarrow$ soft-thresholding

$$\theta_{\text{MCP}}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+/(1 - 1/a), & \text{when } |z| < a\lambda; \\ z, & \text{when } |z| \ge a\lambda. \end{cases}$$

Desired Properties

- **Continuity**: to avoid instability in model prediction.
- **Sparsity**: to reduce model complexity (set small coeff. to 0).
- **Unbiasedness**: to avoid unnecessary modeling bias
  (unbiased when true coefficients are large). (*Fan and Li, 2001*)

Method	continuity	sparsity	unbiasedness
Best subset		$\surd$	$\surd$
Ridge	$\surd$		
LASSO	$\surd$	$\surd$	
SCAD	$\surd$	$\surd$	$\surd$

- **Ideal**: L_0-penalty — hard to compute
- **Popular**: L_1 and Elastic net — large biases, missing big variables
- **Great**: SCAD, MCP — trade-off performance and computing

Characterization of folded-concave PLS

Concavity:

$$\kappa(p_\lambda;\mathbf{v}) = \max_j[-p''_\lambda(|v|_j)] = \max_j \lim_{\varepsilon\to 0+} \sup_{t_1<t_2\in(|v_j|-\varepsilon,|v_j|+\varepsilon)} -\frac{p'_\lambda(t_2)-p'_\lambda(t_1)}{t_2-t_1}.$$

- ★ For L_1 penalty, $\kappa(p_\lambda;\mathbf{v}) = 0$ for any $\mathbf{v}$.
- ★ For SCAD, $\kappa(p_\lambda;\mathbf{v}) = 0$ if $|\mathbf{v}| \notin [\lambda, a\lambda]$, else $\kappa(p_\lambda;\mathbf{v}) = (a-1)^{-1}\lambda^{-1}$

Theorem 3.1. Necessary and sufficient conditions

Let $S = \mathrm{supp}(\widehat{\beta})$, $\widehat{\beta}_1 = \widehat{\beta}_S$, $\mathbf{X}_1 = \mathbf{X}_S$, $\mathbf{X}_2 = \mathbf{X}_{S^c}$. If $\widehat{\beta}$ is a local min, then

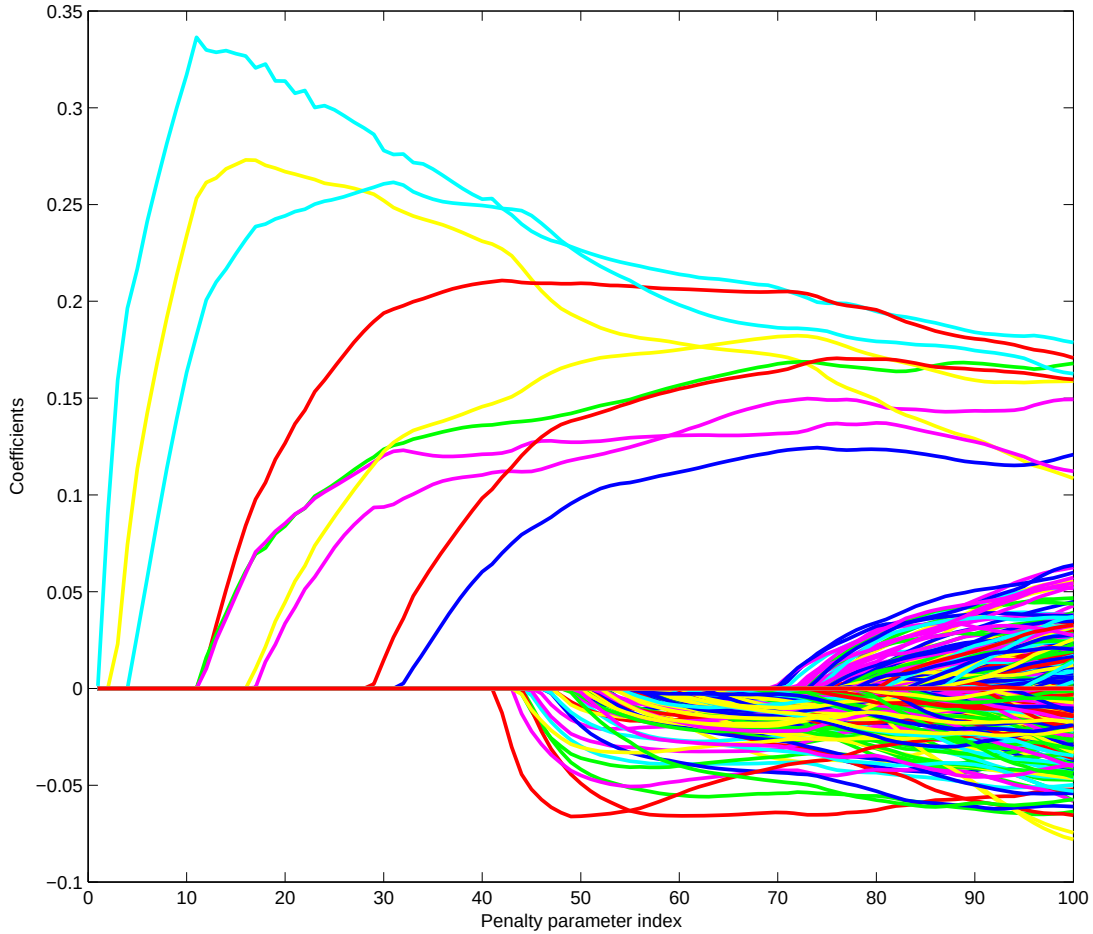
$$n^{-1}\mathbf{X}_1^T(\mathbf{Y}-\mathbf{X}\widehat{\beta}) - p'_\lambda(|\widehat{\beta}_1|)\mathrm{sgn}(\widehat{\beta}_1) = \mathbf{0},$$
$$\|n^{-1}\mathbf{X}_2^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})\|_\infty \le p'_\lambda(0+),$$
$$\lambda_{\min}(n^{-1}\mathbf{X}_1^T\mathbf{X}_1) \ge \kappa(p_\lambda;\widehat{\beta}_1).$$

They are sufficient if inequalities are strict.

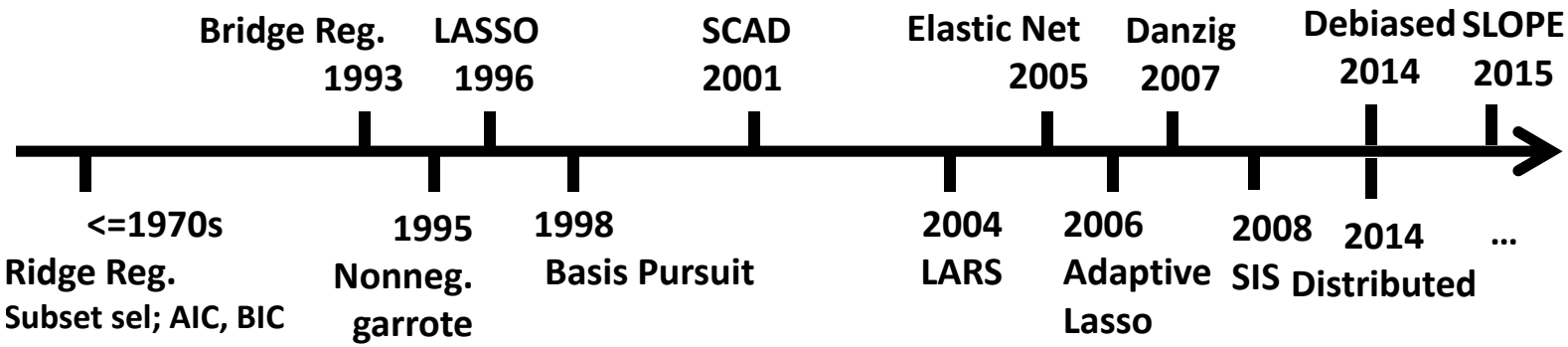
Solution Paths

Target: $Q(\beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$

Solution path: $\widehat{\beta}(\lambda)$ as a function of λ or $\|\widehat{\beta}(\lambda)\|_1$.



2.3 LASSO and L_1 Regularization



LASSO

Necessary condition for Lasso:

$$n^{-1}\mathbf{X}_1^T(\mathbf{Y}-\mathbf{X}_1\widehat{\beta}_1)-\lambda \mathrm{sgn}(\widehat{\beta}_1)=\mathbf{0}, \qquad \|n^{-1}\mathbf{X}_2^T(\mathbf{Y}-\mathbf{X}_1\widehat{\beta}_1)\|_\infty \le \lambda,$$

$$\Rightarrow \|n^{-1}\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})\|_\infty \le \lambda \qquad \Rightarrow \widehat{\beta}=0 \text{ when } \lambda > \|n^{-1}\mathbf{X}^T\mathbf{Y}\|_\infty.$$

Solving $\widehat{\beta}_1$ and substituting in, we have

$$\|(n\lambda)^{-1}\mathbf{X}_2^T(\mathbf{I}_n-\mathbf{P}_{\mathbf{X}_1})\mathbf{Y}-\mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathrm{sgn}(\widehat{\beta}_1)\|_\infty \le 1.$$

Model Selection Consistency for LASSO

True Model: $\mathbf{Y} = \mathbf{X}_{\mathcal{S}_0}\beta_0 + \varepsilon$

Selection Consistency: $\text{supp}(\widehat{\beta}) = \mathcal{S}_0 \Longrightarrow \mathbf{X}_{\mathcal{S}_0} = \mathbf{X}_1$

Necessary Condition: Using true model and $(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\mathbf{X}_{\mathcal{S}_0} = 0$

$$\|(n\lambda)^{-1}\mathbf{X}_2^T(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_1})\varepsilon - \mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1\mathbf{X}_1)^{-1}\,\text{sgn}(\widehat{\beta}_1)\|_\infty \le 1.$$

First term negligible: $\|\mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\,\text{sgn}(\widehat{\beta}_1)\|_\infty \le 1.$

Irrepresentable Cond: If sign consistency $\text{sgn}(\widehat{\beta}) = \text{sgn}(\beta_0)$,

$$\|\mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\,\text{sgn}(\beta_{\mathcal{S}_0})\|_\infty \le 1, \qquad \textit{(Zhao and Yu, 06)},$$

necessary for sel. consistency; sufficient if 1 replaced by $1 - \eta$.

Remarks on Irrepresentabe Condition

- ★ $(\mathbf{X}_1\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$ reg coef of 'unimportant' X_j ($j \notin \mathcal{S}_0$) on $\mathbf{X}_{\mathcal{S}_0}$.
- ★ Irrepresentable cond requires sum of (the signed reg coefs of each X_j on $\mathbf{X}_{\mathcal{S}_0}$) ≤ 1. The bigger $\mathcal{S}_0$, the harder the condition
- ★ Irrespresentable condition is restrictive, leading to false negatives, compensated by many false positives.

Risks of Lasso (I)

Risk: $R(\beta) = E(Y - \mathbf{X}^T\beta)^2 = E(\gamma^T\mathbf{Z})^2 = \gamma^T\mathbf{\Sigma}^*\gamma,$

$$\gamma = \begin{pmatrix} -1 \\ \beta \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix}, \qquad \mathbf{\Sigma}^* = E(\mathbf{Z}\mathbf{Z}^T) \qquad \mathbf{S}_n^* = n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T.$$

Empirical risk: $R_n(\beta) = n^{-1}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\beta)^2 = \gamma^T\mathbf{S}_n^*\gamma$ **cov-learning**

Dual problem: $\widehat{\beta} = \text{argmin}_{\|\beta\|_1 \le c} \|\mathbf{Y} - \mathbf{X}\beta\|^2.$

$$\begin{aligned} & |R(\beta) - R_n(\beta)| = |\gamma^T(\mathbf{\Sigma}^* - \mathbf{S}_n^*)\gamma| \\ \le \quad & \|\mathbf{\Sigma}^* - \mathbf{S}_n^*\|_{\max}\|\gamma\|_1^2 = (1 + \|\beta\|_1)^2\|\mathbf{\Sigma}^* - \mathbf{S}_n^*\|_{\max}. \\ \le \quad & (1+c)^2\|\mathbf{\Sigma}^* - \mathbf{S}_n^*\|_{\max}. \end{aligned}$$

Risks of Lasso (II)

If $\|\beta_0\|_1 \le c$, then $R_n(\widehat{\beta}) - R_n(\beta_0) \le 0$ and

$$\begin{aligned} 0 \le R(\widehat{\beta}) - R(\beta_0) &\le \{R(\widehat{\beta}) - R_n(\widehat{\beta})\} + \{R_n(\beta_0) - R(\beta_0)\} \\ &\le 2(1+c)^2 \|\mathbf{\Sigma}^* - \mathbf{S}_n^*\|_{\max} \end{aligned}$$

Risk consistency: $R(\widehat{\beta}) - R(\beta_0) \to 0$ *(Greenshtein and Ritov, 04)*

- ★ Sample cov has rate $O(\sqrt{(\log p)/n})$ for subGaussian data.
- ★ Consistency requires $\|\beta_0\|_1 \le c = o((n/\log p)^{1/4})$
- ★ Need uniform convergence rate of $\mathbf{S}_n^*$.

Accuracy of Sample Covariance matrix

Result: If $\max_{i,j} P\{\sqrt{n}|\sigma_{ij} - \widehat{\sigma}_{ij}| > x\} < \exp(-\mathbf{C}x^{\mathbf{a}})$, for big x,

$$\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_{\max} = O_P\left(\frac{(\log \mathbf{p})^{\mathbf{1/a}}}{\sqrt{\mathbf{n}}}\right), \qquad \text{for } p \times p \text{ matrix } \mathbf{\Sigma}, \text{ as } p \to \infty.$$

- Impact of dim is limited.
- Req exponential tail (sub-Weibull).

Proof:

$$\begin{aligned} P\{\sqrt{n}\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} > b_n\} &\leq \sum_{i,j} P\{\sqrt{n}|\widehat{\sigma}_{ij} - \sigma_{ij}| > b_n\} \\ &\leq p^2 \exp(-Cb_n^a) = 1/p^8 \end{aligned}$$

by taking $b_n = (10C^{-1}\log p)^{1/a}$. Conclusion follows.

Remarks

- ★ It requires only marginal behavior and uses the union bounds.
- ★ The inequality is referred to as concentration inequality.
- ★ Sample covariance matrix is basically the sample mean of EX_iX_j; needed concentration inequality for sample mean or its robust estimation.
- ★ Sub-Weibull tail: $a = 2$ sub-Gaussian and $a = 1$ sub-exponential.

Sparsity of Lasso Solution

- Let $\widehat{\boldsymbol{\Delta}} = \widehat{\beta} - \beta_0$, where β_0 is true parameter. Then, $F(\widehat{\boldsymbol{\Delta}}) \leq 0$.
- $F(\boldsymbol{\Delta}) = \underbrace{R_n(\boldsymbol{\Delta}+\beta_0)/2 - R_n(\beta_0)/2}_{(1)} + \lambda(\underbrace{\|\boldsymbol{\Delta}+\beta_0\|_1 - \|\beta_0\|_1}_{(2)})$

By convexity, if $\boxed{\lambda \geq 2n^{-1}\|\mathbf{X}^T\varepsilon\|_\infty}$, then

$$(1) \quad \geq \quad -|\frac{1}{2}R_n'(\beta_0)^T\boldsymbol{\Delta}| \geq -\frac{1}{n}\|\mathbf{X}^T\varepsilon\|_\infty\|\boldsymbol{\Delta}\|_1 \geq -\lambda\|\boldsymbol{\Delta}\|_1/2$$

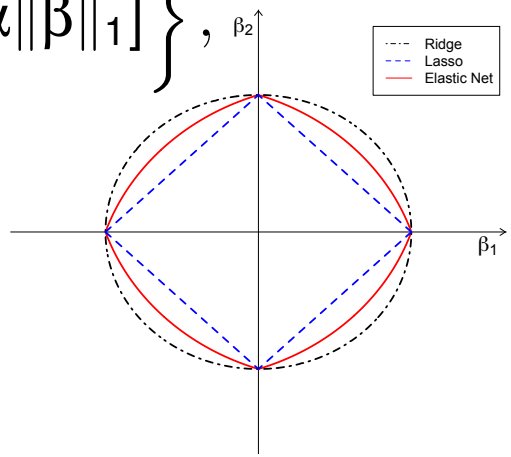
$$\text{set } \beta^* = \beta_0$$

$$(2) \quad = \quad \|\boldsymbol{\Delta}_{S_0} + \beta^*_{S_0}\|_1 + \|\boldsymbol{\Delta}_{S_0^c}\|_1 - \|\beta^*_{S_0}\|_1 \geq -\|\boldsymbol{\Delta}_{S_0}\|_1 + \|\boldsymbol{\Delta}_{S_0^c}\|_1$$

Combining these and $F(\widehat{\boldsymbol{\Delta}}) \leq 0$, we have $\|\widehat{\boldsymbol{\Delta}}_{S_0^c}\|_1 \leq 3\|\widehat{\boldsymbol{\Delta}}_{S_0}\|_1$.

Elastic Net and Danzig Selector

$$\arg\min_{\beta}\left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\beta\|^2+\lambda[(1-\alpha)\|\beta\|^2+\alpha\|\beta\|_1]\right\},$$



★Mitigate collinearity

★$\alpha=1 \Longrightarrow$ Lasso; $\alpha=0 \Longrightarrow$ Ridge

Danzig selector: $\min_{\beta\in R^p}\|\beta\|_1,$ $\quad s.t. \quad \|n^{-1}\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta)\|_\infty \le \lambda.$

high confident set

■require $\lambda \ge \|n^{-1}\mathbf{X}^T(\underbrace{\mathbf{Y}-\mathbf{X}\beta_0}_{\varepsilon})\|_\infty$ ★max spur corr. *(Candés & Tao, 07)*

Linear program: $\min_{\mathbf{u}}\sum_{i=1}^p u_i, \quad \mathbf{u}\ge 0, \quad -\mathbf{u}\le\beta\le\mathbf{u}, \quad -\lambda\mathbf{1}\le n^{-1}\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\beta)\le\lambda\mathbf{1}$

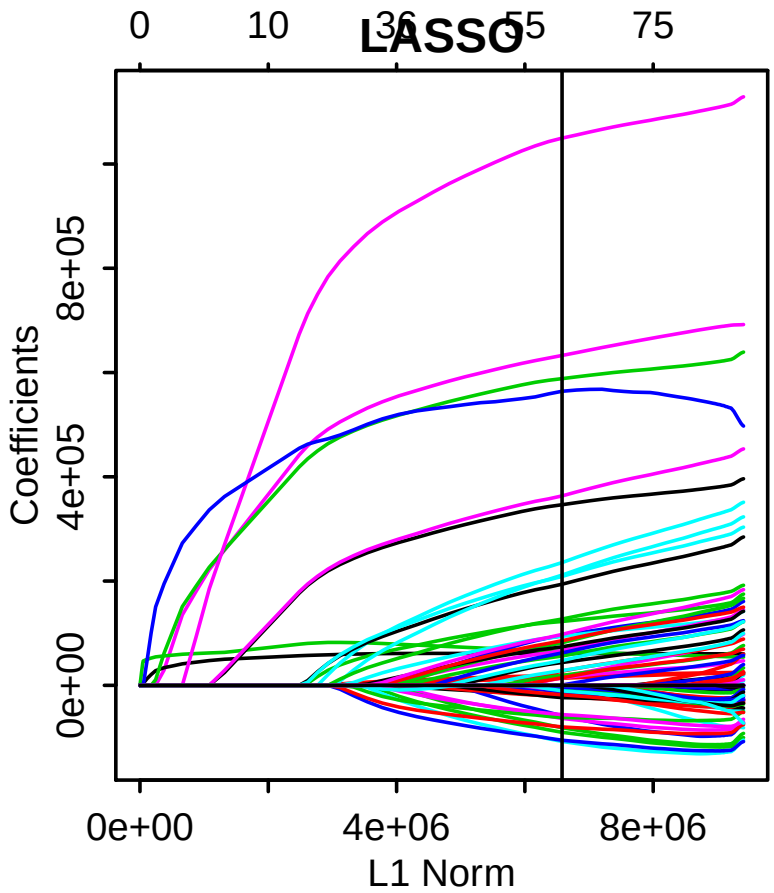
Housing Prediction: Zillow data: House price + house attributes (bedrooms, bathrooms, sqft.living sqft.lot, ...) in 70 zip codes, sold in 2014 & 15.

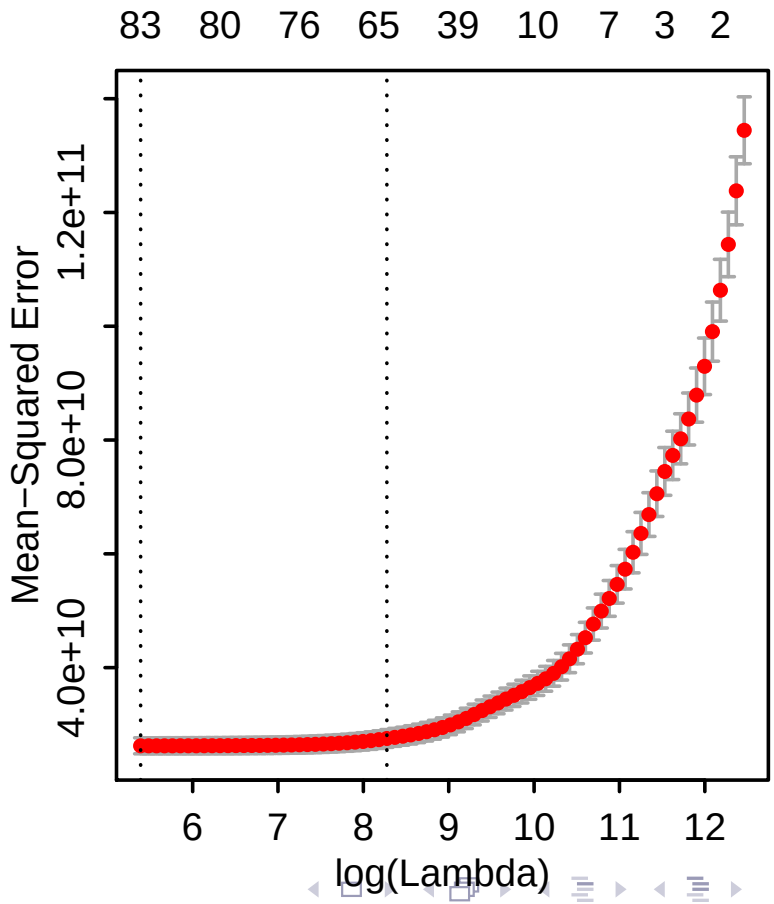
($n_{train} = 15,129$, $n_{test} = 6,484$).

Out-of-sample R^2: 0.80

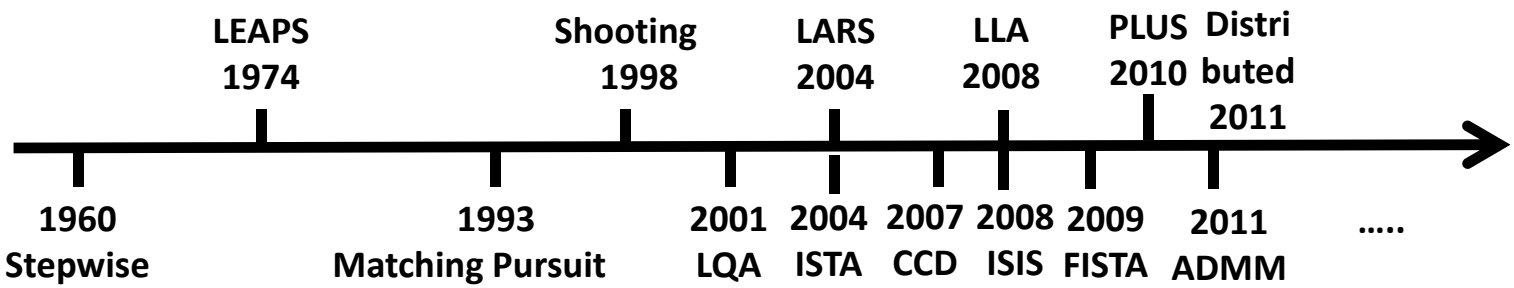
```
library('glmnet')              #use the package
X <- model.matrix(~.,data = train_data[,5:22])
Y <- train_data$price

fit.glm1 <- glmnet(X, Y, alpha=1)      #Lasso fit, solution path
plot(fit.glm1, main="Lasso")           #

fit.cvglm1 <- cv.glmnet(X,Y,nfolds = 5, alpha = 1);      #cross validation, what is
fit.cvglm1$lambda.min
beta.cvglm1 <- coef(fit.cvglm1, s=fit.cvglm1$lambda.1se) ###coef at 1se

pdf("Zillow1.pdf", width=4.6, height=2.6, pointsize=8)
par(mfrow = c(1,2), mar=c(5,5,3,1)+0.1, mex=0.5)
plot(fit.cvglm1$glmnet.fit); title('LASSO')
abline(v=sum(abs(beta.cvglm1[-1])))  #Lasso solution path
```

```
plot(fit.cvglm1)                              #Estimated MSE
  ##### compute testing errors  ############
X_test <- model.matrix(~., data = test_data[,5:22])  ##test data model
pred.glm1<- predict(fit.cvglm1, newx = X_test, s = "lambda.min")   ## computed pred
mse.pred.glm1 <- sum((test_data$price - pred.glm1)^2)       #MSE
1 - mse.pred.glm1/ sum((test_data$price - mean(Y))^2)       #out-sample-R^2
```





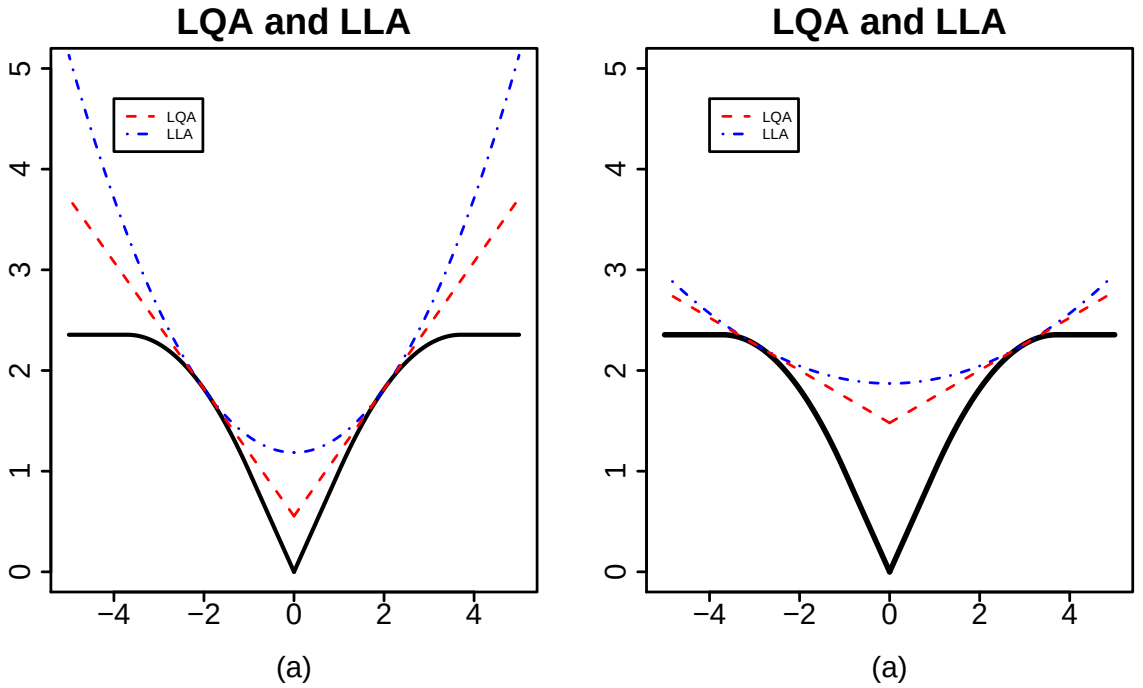
3.4 Algorithms for Folded-concave Regularization



$$Q(\beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

Local quadratic approximations

Target: $Q^{\text{approx}}(\beta) = \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\beta\|^2 + \sum_{j=1}^{p}\left\{p_\lambda(|\beta_j^{(k)}|) + \frac{p'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|}(\beta_j^2 - |\beta_j^{(k)}|^2)\right\}$



Iterative formulas for LQA: (*ridge reg, Fan & Li, 01*)

$$\beta^{(k+1)} = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\beta^{(k)})\}^{-1}\mathbf{X}'\mathbf{y}, \quad \Sigma_\lambda(\beta^{(k)}) = \text{diag}\{p'_\lambda(|\beta^{(k)}|)/|\beta_{(k)}|\}$$

- Delete X_j in the iteration, if $|\beta_j^{(k+1)}| \le \eta$.

Local linear approximations and one-step estimator

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(k)}|) + \mathbf{p'_\lambda(|\beta_j^{(k)}|)}(|\beta_j| - |\beta_j^{(k)}|)$$

Target: $Q^{\text{approx}}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p w_j|\beta_j|, \qquad w_j = p'_\lambda(|\beta_j^{(k)}|)$

One-step estimator: $\widehat{\beta}^{(0)} = \mathbf{0} \Longrightarrow$ **Lasso** $\Longrightarrow$ Adaptive Lasso

- ★ With LLA, PLS is iter. reweighted LASSO, w/ weights given by $p'_\lambda(\cdot)$.
- ★ Iterations reduces biases of previous estimates.
- ★ SCAD = Iteratively Reweighted LASSO and is an Oracle estimator.

Remarks

- ★ LLQ computes explicitly the updates whereas LLA requires PLS.
- ★ LLA gives better approximation, particularly around the origin.
- ★ Adaptive lasso: $p'_\lambda(|\beta_j^*|) = \lambda|\beta_j^*|^{-\gamma}$ $(\gamma > 0)$. **drawback**: Zero is an absorbing state.
- ★ Both LQA and LLA are a specific member of MM algorithm:

$$Q(\beta) \le Q^{\text{approx}}(\beta|\beta_{init}), \qquad Q(\beta_{init}) = Q^{\text{approx}}(\beta_{init}|\beta_{init})$$

- ★ **Convergence of MM** (Majorization Minimization):

$$Q(\beta^{(k)}) \underbrace{=}_{cond} Q^{\text{approx}}(\beta^{(k)}|\beta^{(k)}) \underbrace{\ge}_{min} Q^{\text{approx}}(\beta^{(k+1)}|\beta^{(k)}) \underbrace{\ge}_{major} Q(\beta^{(k+1)}).$$

Coordinate descent algorithms

★ Optimize one variable at a time: Given current value β_0, update

$$\widehat{\beta}_j = \text{argmin}_{\beta_j} Q(\beta_{1,0}, \cdots, \beta_{j-1,0}, \beta_j, \beta_{j+1,0}, \cdots, \beta_{p,0}),$$

★ For PLS, $\mathbf{R}_j = \mathbf{Y} - \mathbf{X}_{-j}\widehat{\beta}_{-j,0}$ without j^{th} variable. Then

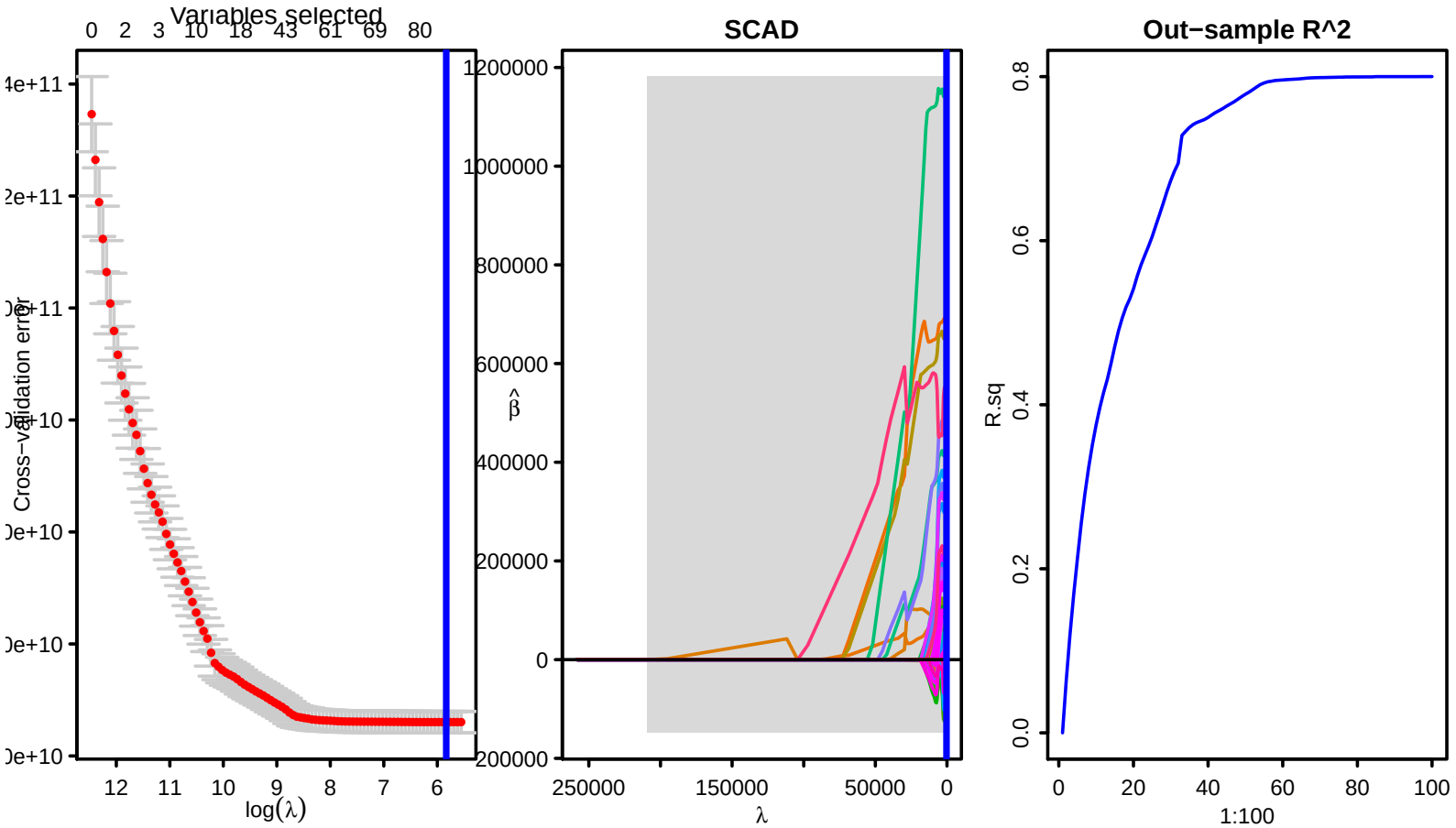
$$\begin{aligned} Q_j(\beta_j) &\equiv Q(\beta_{10}, \cdots, \beta_{j-1,0}, \beta_j, \beta_{j+1,0}, \cdots, \beta_{p,0}) \\ &= \frac{1}{2n}\|\mathbf{R}_j - \mathbf{X}_j\beta_j\|^2 + p_\lambda(|\beta_j|) + c, \end{aligned}$$

★ For Lasso, SCAD and MCP, $\widehat{\beta}_j$ admits analytic solution.

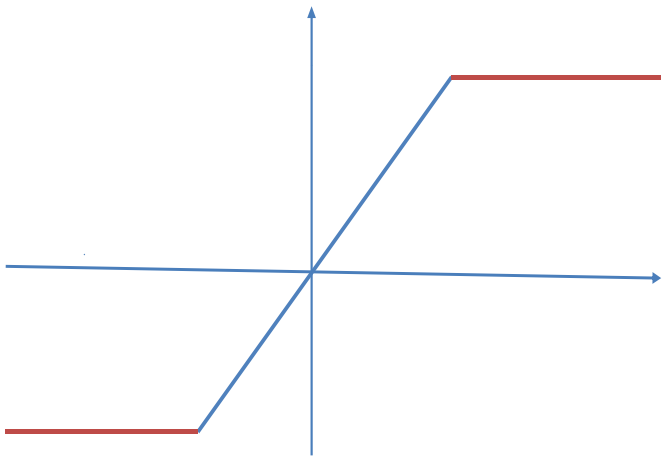
Housing Prediction by SCAD: Zillow data: $R^2 = 0.80$

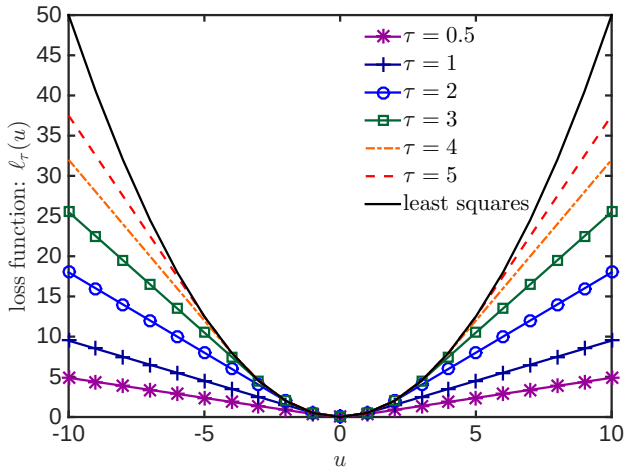
```
library('ncvreg')                  #loading the library for use
X <- model.matrix(~.,data = train_data[,5:22])
Y <- train_data$price              #data model for training
cvfit.SCAD <- cv.ncvreg(X, Y, penalty="SCAD")   #SCAD using CV

### prediction of the test data and out-sample R^2
predict.SCAD = predict(cvfit.SCAD, X=X_test)
mse.pred.scad <- sum((test_data$price - predict.SCAD)^2)   #MSE
1 - mse.pred.scad/ sum((test_data$price - mean(Y))^2)  #out-sample-R^2

### change smoothing parameter
fit.SCAD <- ncvreg(X, Y, penalty="SCAD")  ## give a family of fits for diff lambda
predict.SCAD = predict(fit.SCAD, X=X_test); #a family of predict for diff lambda
R.sq = NULL
for(j in 1:100){
mse.pred.scad <- sum((test_data$price - predict.SCAD[,j])^2) #MSE at jth
R.sq = c(R.sq, 1 - mse.pred.scad/ sum((test_data$price - mean(Y))^2))}

pdf("Zillow2.pdf", width=4.6, height=2.6, pointsize=8)
par(mfrow = c(1,3), mar=c(5,5,3,1)+0.1, mex=0.5)
plot(cvfit.SCAD)
abline(v=log(cvfit.SCAD$lambda.min),lwd=2,col=4)
plot(fit.SCAD, main="SCAD")
abline(v=cvfit.SCAD$lambda.min,lwd=2,col=4)
plot(1:100,R.sq, main="Out-sample R^2", type="l", col=4)
```

Housing Prediction by SCAD: Zillow data



2.5 Concentration Inequalities





A motivating example

- fundamental tools for controlling max spurious correlation errors: e.g.

$$P\Big\{\|n^{-1}\mathbf{X}^T\varepsilon\|_\infty > t\Big\} \le \sum_{j=1}^{p} P\Big\{|\underbrace{n^{-1}\textstyle\sum_{i=1}^{n} X_{ij}\varepsilon_i}_{Z_j}| > t\Big\}.$$

If $n^{-1}\|\mathbf{X}_j\|^2 = 1$ and $\varepsilon_i \sim_{i.i.d.} N(0,\sigma^2)$, then $Z_j \sim N(0,\sigma^2/n)$ and

$$P\Big\{|Z_j| \ge t\frac{\sigma}{\sqrt{n}}\Big\} \le \frac{2}{\sqrt{2\pi}}\int_t^\infty \frac{\mathbf{x}}{\mathbf{t}}\exp(-x^2/2)dx = \frac{2}{\sqrt{2\pi}}\exp(-t^2/2)/t$$

By taking $t = \sqrt{2(1+\delta)\log p}$, with prob $\ge 1 - o(p^{-\delta})$,

$$\|n^{-1}\mathbf{X}^T\varepsilon\|_\infty \le \sqrt{2(1+\delta)}\sigma\sqrt{\frac{\log p}{n}}.$$

- Tail probability of sum of independent random variables, however dependence $\{X_j\}_{j=1}^p$ is.

Concentration inequalities

Theorem 3.2 $Y_1, \cdots, Y_n$ are independent r.v. w/ mean 0.

Let $S_n = \sum_{i=1}^n Y_i$ be the sum of the random variables.

a) **Hoeffding inequality**: If $Y_i \in [a_i, b_i]$, then

$$P(|S_n| \geq t) \leq 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

b) **Berstein's inequality**. If $E|Y_i|^m \leq m! M^{m-2} v_i/2$, then

$$P(|S_n| \geq t) \leq 2\exp\Big(-\frac{t^2}{2(v_1 + \cdots + v_n + Mt)}\Big).$$

c) **Sub-Gaussian**: If $E\exp(aY_i) \leq \exp(v_i a^2/2)$, then

$$P(|S_n| \geq t) \leq 2\exp\Big(-\frac{t^2}{2(v_1 + \cdots + v_n)}\Big).$$

Remarks

There are concentration inequalities for

- ★ sum of sub-exponential distributions
- ★ Robust mean estimators: Winsorization + Huber estimator
- ★ empirical processes
- ★ random matrices
- ★ MLE
- ★ Self-normalized average
- ★ Markovian chains and mixing processes

Bounded diff ineq: If $Z_n = g(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ with $|g(\mathbf{x}_1, \cdots, \mathbf{x}_n) - g(\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_i', \cdots, \mathbf{x}_n)| \le c_i$ for all data $\{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{x}_i'$ (changing only one data point from $\mathbf{x}_i$ to $\mathbf{x}_i'$) and $\{\mathbf{X}_i\}$ indep, then

$$P(|Z_n - EZ_n| > t) \le 2\exp\left(-\frac{2t^2}{c_1^2 + \cdots + c_n^2}\right).$$

2.6 Estimation of Residual Variance

Residual Variance

- ★ Important for stat inference and model selection. It is benchmark for optimal prediction.

- ★ $\sigma^2 = R(\beta_0)$, consistently estimated by LASSO residual variance $R_n(\widehat{\beta}) = n^{-1}\|\mathbf{Y} - \mathbf{X}\widehat{\beta}\|^2$, when risk consistency.

- ★ This requires $\|\beta_0\|_1 \leq c = o((n/\log p)^{1/4})$ and has slow rates.

- ★ See Sec 1.3.3. for spurious corr and underestimation of σ^2.

Refitted Cross-validation

- ★ Randomly split the data;
- ★ Select variables using the first half data;
- ★ Refit the selected model using the second half to get $\widehat{\sigma}^2$;
- ★ and vice verse; take the average of the two estimators.

Key difference: Refitting eliminates spurious correlation! (*Fan, Guo, Hao, 12)*

RCV

Random Division of Data: $(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})$ and $(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})$

Refitted residual variances: With selected model $\widehat{M}_1$, compute

$$\widehat{\sigma}_1^2 = \frac{(\mathbf{y}^{(2)})^T(\mathbf{I}_{n/2} - \mathbf{P}^{(2)}_{\widehat{M}_1})\mathbf{y}^{(2)}}{n/2 - |\widehat{M}_1|},$$

where $\mathbf{P}^{(2)}_{\widehat{M}_1} = \mathbf{X}^{(2)}_{\widehat{M}_1}(\mathbf{X}^{(2)T}_{\widehat{M}_1}\mathbf{X}^{(2)}_{\widehat{M}_1})^{-1}\mathbf{X}^{(2)T}_{\widehat{M}_1}$ and $\widehat{\sigma}_2^2$

Final estimate: $\widehat{\sigma}^2_{\mathbf{RCV}} = (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)/2$ or its weighted average.

Advantages: ■Require only **sure screening**

■Reduce influence of spurious correlation.

2.7 Variable Expansions and Group Penalty

Structured Nonparametric Models

★ **Additive model** (*Stone, 85, Hastie and Tibshirani, 90)*

$$Y = f_1(X_1) + \cdots + f_p(X_p) + \varepsilon$$

★ **Two-d nonparametric interactions**:

$$Y = \sum_{i=1}^{p} f_i(X_i) + \sum_{i<j} f_{i,j}(X_i, X_j) + \varepsilon$$

★ **Varying coefficient model**:

$$Y = \beta_0(U) + \beta_1(U)X_1 + \cdots + \beta_p(U)X_p + \varepsilon.$$

Expanded Linear Models

Ex 1: **Additive model** $Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon$ *(Stone, 85, Hastie and Tibshirani, 90)*

Approx $f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} B_{jk}(x)$ using basis functions $\{B_{jk}(x)\}_{k=1}^{K_j}$ (e.g. spline basis). Additive model becomes an expanded linear model

$$Y = \sum_{j=1}^{p} \Big\{ \sum_{k=1}^{K_j} \beta_{jk} B_{jk}(X_j) \Big\} + \varepsilon.$$

Ex 2: **Bivar interaction model**: $Y = \sum_{j=1}^{p} f_j(X_j) + \sum_{i<j} f_{i,j}(X_i, X_j) + \varepsilon$ can be approximated as an expanded linear model

$$Y = \sum_{j=1}^{p} \Big\{ \underbrace{\sum_{k=1}^{K_j} \beta_{jk} B_{jk}(X_j)}_{\beta_j^T \mathbf{x}_j} \Big\} + \sum_{i<j} \Big\{ \underbrace{\sum_{k=1}^{K_i} \sum_{l=1}^{K_j} \gamma_{ijkl} B_{ik}(X_i) B_{jl}(X_j)}_{\gamma_{ij}^T \mathbf{x}_{ij}} \Big\} + \varepsilon.$$

Group penalty

- Covariate divided into groups, e.g. additive model:
- Data: $\mathbf{Y} = \sum_{j=1}^{p} \underbrace{\mathbf{X}_j}_{n \times K_j} \beta_j + \varepsilon$
- Group PLS: $\frac{1}{2n} \|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j \beta_j\|^2 + \sum_{j=1}^{p} p_\lambda(\|\beta_j\|)$
- Select or kill a group of variables. *(Lin & Yuan, 06)*
- Appeared already in Antoniadis & Fan (JASA, 01) for selecting blocks of wavelet coefficients.