

# Statistical Foundations of Data Science

**Jianqing Fan**

**Princeton** University

<https://fan.princeton.edu>



## 2. Penalized Least-Squares

- 2.1. Classical model selection
- 2.3. Lasso and  $L_1$ -regularization
- 2.5. Concentration inequalities
- 2.7. Variable expansion & g-penalty

- 2.2. Penalized least-squares
- 2.4. Folded concave regularization
- 2.6. Estimation of residual variance

# 2.1 Classical Model Selection

# Best subset selection

Data:  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$  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  $\beta_S$ .

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

# Relation with $L_0$ -penalty

Traditional:  $L_0$ -penalty

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

Computation: Best subset

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

Greedy algorithms: stepwise addition/deletion, stepwise, 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 $\lambda$ ?

# Stein identity and prediction error

**Model:**  $Y_i = \mu(\mathbf{X}_i) + \varepsilon_i$ .

Let  $\hat{\mu}$  be an estimator

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

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

**PE:**  $E \|\mathbf{Y}^{\text{new}} - \hat{\mu}\|^2 = n\sigma^2 + E \|\mu - \hat{\mu}\|^2 = E \{ \|\mathbf{Y} - \hat{\mu}\|^2 \} + 2df_{\hat{\mu}}\sigma^2$ .

**Unbiased estimate of PE.** Let  $df_{\hat{\mu}} = \sigma^{-2} \sum_{i=1}^n \text{cov}(\hat{\mu}_i, Y_i)$

model fixed

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

■ For linear predictor  $\hat{\mu} = \mathbf{S}\mathbf{Y}$ ,  $df_{\hat{\mu}} = \text{tr}(\mathbf{S})$  since  $\text{cov}(\hat{\mu}_i, Y_i) = s_{ii}\sigma^2$ .

# Efforts in model selection: choosing $\lambda$

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

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

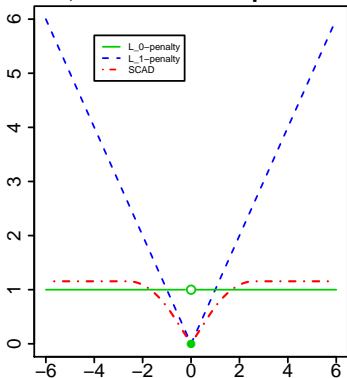
## 2.2 Penalized Least-Squares



# Convex and folded concave relaxations

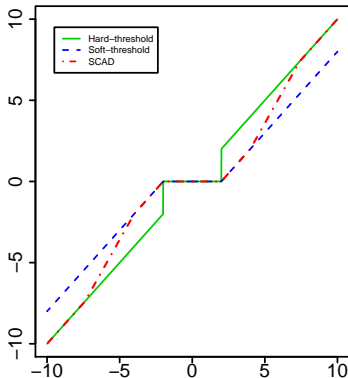
Penalized least-squares:  $Q(\beta) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p p_\lambda(|\beta_j|)$

L<sub>0</sub>, L<sub>1</sub> and SCAD penalties



(a)

Solutions to PLS



(b)

# Convex penalties

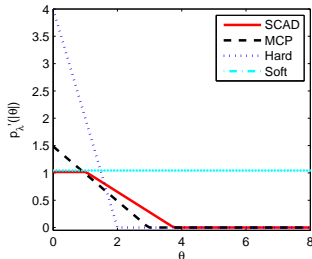
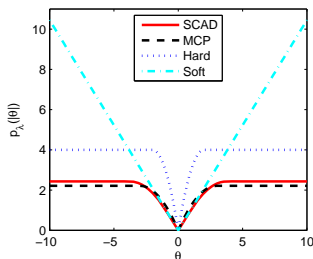
- ★  $L_2$  penalty  $p_\lambda(|\theta|) = \lambda|\theta|^2 \implies$  ridge regression
- ★  $L_q$ -penalty  $p_\lambda(|\theta|) = \lambda|\theta|^q \implies$  Bridge reg (*Frank and Friedman, 93*).  
 $q \geq 1$  is convex,  $q \in (0, 1)$  is folded concave.
- ★  $L_1$  penalty  $p_\lambda(|\theta|) = \lambda|\theta| \implies$  **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)

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

(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).

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

**What is the role of penalty function?**

# Orthonormal design

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

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



**comp** min. (Antoniadis and Fan, 01; Fan and Li, 01)

solutions drawn on page 9

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

- 1  $\rho_\lambda(|\theta|) = \frac{\lambda}{2}\theta^2 \implies$  **shrinkage**:  $\hat{\theta}_\lambda = (1 + \lambda)^{-1}z$ .
- 2  $\rho_\lambda(|\theta|) = \lambda|\theta| \implies$  **soft-threshold**:  $\hat{\theta}_S = \text{sgn}(z)(|z| - \lambda)_+$ .
- 3  $\rho_\lambda(\theta) = \lambda^2 - (\lambda - |\theta|)_+^2 \implies$  **hard-threshold**  $\hat{\theta}_H = zI(|z| > \lambda)$ .

# Solutions of SCAD and MCP

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

■  $a = \infty \implies$  soft-thresholding

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

# 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		✓	✓
Ridge	✓		
LASSO	✓	✓	
SCAD	✓	✓	✓

- **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:

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

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

$$n^{-1} \mathbf{X}_1^T (\mathbf{Y} - \mathbf{X} \hat{\beta}) - p_\lambda'(|\hat{\beta}_1|) \text{sgn}(\hat{\beta}_1) = \mathbf{0},$$

$$\|n^{-1} \mathbf{X}_2^T (\mathbf{Y} - \mathbf{X} \hat{\beta})\|_\infty \leq p_\lambda'(0+),$$

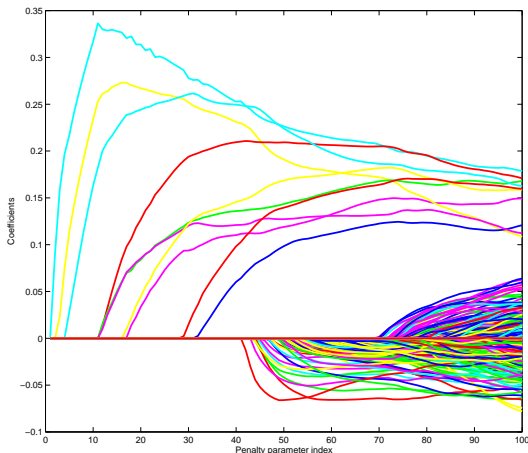
$$\lambda_{\min}(n^{-1} \mathbf{X}_1^T \mathbf{X}_1) \geq \kappa(p_\lambda; \hat{\beta}_1).$$

They are sufficient if inequalities are strict.

# Solution Paths

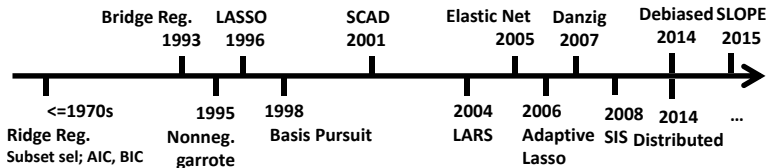
Target:  $Q(\beta) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p \rho_\lambda(|\beta_j|)$

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





## 2.3 LASSO and $L_1$ Regularization



Necessary condition for Lasso:

$$n^{-1} \mathbf{X}_1^T (\mathbf{Y} - \mathbf{X}_1 \hat{\beta}_1) - \lambda \operatorname{sgn}(\hat{\beta}_1) = \mathbf{0}, \quad \|n^{-1} \mathbf{X}_2^T (\mathbf{Y} - \mathbf{X}_1 \hat{\beta}_1)\|_\infty \leq \lambda,$$

$$\rightarrow \|n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X} \hat{\beta})\|_\infty \leq \lambda$$

$$\rightarrow \hat{\beta} = 0 \text{ when } \lambda > \|n^{-1} \mathbf{X}^T \mathbf{Y}\|_\infty.$$

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

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

# Model Selection Consistency for LASSO

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

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

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

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

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

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

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

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

# Remarks on Irrepresentable Condition

- ★  $(\mathbf{X}_1 \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2$  reg coef of 'unimportant'  $X_j$  ( $j \notin S_0$ ) on  $\mathbf{X}_{S_0}$ .
- ★ Irrepresentable cond requires sum of (the signed reg coefs of each  $X_j$  on  $\mathbf{X}_{S_0}$ )  $\leq 1$ . The bigger  $S_0$ , the harder the condition
- ★ Irrepresentable condition is restrictive, leading to false negatives, compensated by many false positives.

# Risks of Lasso (I)

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

$$\gamma = \begin{pmatrix} -1 \\ \beta \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} Y \\ \mathbf{x} \end{pmatrix}, \quad \Sigma^* = E(\mathbf{z}\mathbf{z}^T) \quad \mathbf{s}_n^* = n^{-1} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T.$$

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:  $\hat{\beta} = \operatorname{argmin}_{\|\beta\|_1 \leq c} \|\mathbf{Y} - \mathbf{X}\beta\|^2$ .

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

## Risks of Lasso (II)

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

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

Risk consistency:  $R(\hat{\beta}) - R(\beta_0) \rightarrow 0$  (Greenshtein and Ritov, 04)

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

# Accuracy of Sample Covariance matrix

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

$$\|\Sigma - \hat{\Sigma}\|_{\max} = O_P\left(\frac{(\log p)^{1/a}}{\sqrt{n}}\right).$$

■ Impact of dim is limited.

■ Req exponential tail (sub-Weibull).

**Proof:**

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

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  $\hat{\Delta} = \hat{\beta} - \beta_0$ , where  $\beta_0$  is true parameter. Then,  $F(\hat{\Delta}) \leq 0$ .

$$\text{■ } F(\Delta) = \underbrace{R_n(\Delta + \beta_0)/2 - R_n(\beta_0)/2}_{(1)} + \lambda \underbrace{(\|\Delta + \beta_0\|_1 - \|\beta_0\|_1)}_{(2)}$$

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

$$(1) \geq -\left| \frac{1}{2} R'_n(\beta_0)^T \Delta \right| \geq -\frac{1}{n} \|\mathbf{X}^T \varepsilon\|_\infty \|\Delta\|_1 \geq -\lambda \|\Delta\|_1 / 2$$

set  $\beta^* = \beta_0$

$$(2) = \|\Delta_{S_0} + \beta_{S_0}^*\|_1 + \|\Delta_{S_0^c}\|_1 - \|\beta_{S_0}^*\|_1 \geq -\|\Delta_{S_0}\|_1 + \|\Delta_{S_0^c}\|_1$$

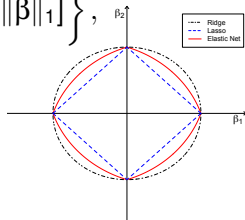
Combining these and  $F(\hat{\Delta}) \leq 0$ , we have  $\|\hat{\Delta}_{S_0^c}\|_1 \leq 3\|\hat{\Delta}_{S_0}\|_1$ .

# Elastic Net and Danzig Selector

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

★ Mitigate collinearity

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



Danzig selector:  $\min_{\beta \in \mathbb{R}^p} \|\beta\|_1,$

$$s.t. \quad \|n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\beta)\|_{\infty} \leq \lambda.$$

high confident set

■ require  $\lambda \geq \|n^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\beta_0)\|_{\infty}$

★ max spur corr.

(Candés & Tao, 07)

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

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")          #

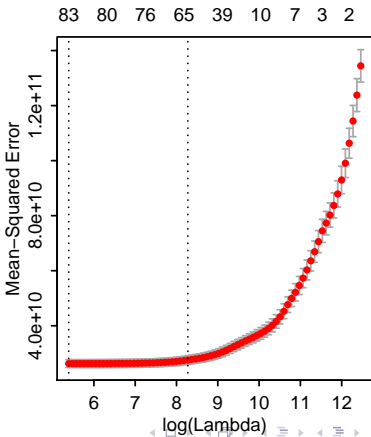
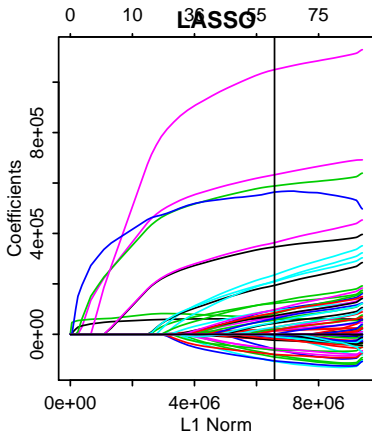
fit.cvglm1 <- cv.glmnet(X,Y,nfolds = 5, alpha = 1);      #cross validation, what is
fit.cvglm1$lambda.min
beta.cvglm1 <- coef(fit.cvglm1, s=fit.cvglm1$lambda.1se) ###coef at 1se

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
```

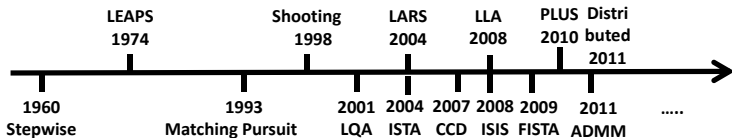
```

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

```



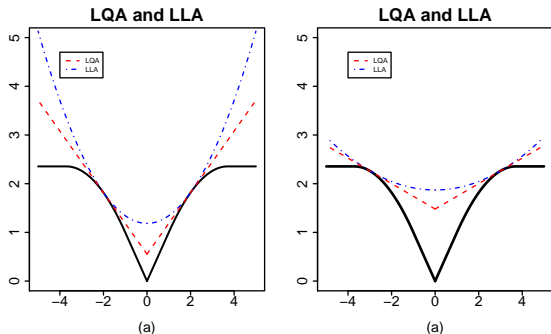
# 3.4 Algorithms for Folded-concave Regularization



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

# 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)

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

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

# Local linear approximations and one-step estimator

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

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

One-step estimator:  $\hat{\beta}^{(0)} = \mathbf{0} \implies \text{Lasso} \implies \text{Adaptive Lasso}$

- ★ 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:

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

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

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



# Coordinate descent algorithms

- ★ Optimize one variable at a time: Given current value  $\beta_0$ , update

$$\hat{\beta}_j = \operatorname{argmin}_{\beta_j} Q(\beta_{1,0}, \dots, \beta_{j-1,0}, \beta_j, \beta_{j+1,0}, \dots, \beta_{p,0}),$$

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

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

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

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

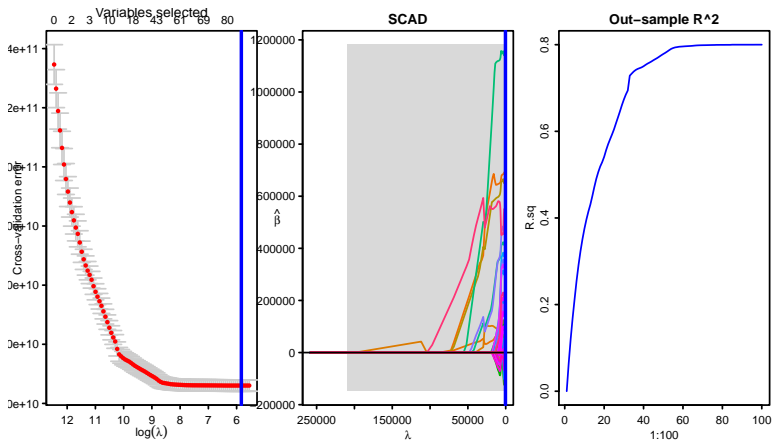
```
library('ncvreg')           #loading the library for use
X <- model.matrix(~.,data = train_data[,5:22])
Y <- train_data$price       #data model for training
cvfit.SCAD <- cv.ncvreg(X, Y, penalty="SCAD")  #SCAD using CV

### 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- $R^2$ 

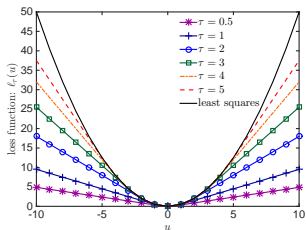
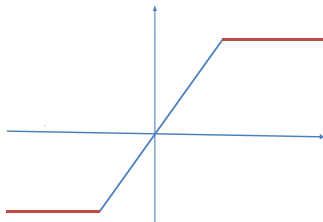
### change smoothing parameter
fit.SCAD <- ncvreg(X, Y, penalty="SCAD")  ## give a family of fits for diff lambda
predict.SCAD = predict(fit.SCAD, X=X_test); #a family of predict for diff lambda
R.sq = NULL
for(j in 1:100){
mse.pred.scad <- sum((test_data$price - predict.SCAD[,j])^2) #MSE at jth
R.sq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

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

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

$$P\left\{|Z_j| \geq t\frac{\sigma}{\sqrt{n}}\right\} \leq \frac{2}{\sqrt{2\pi}} \int_t^{\infty} \frac{\mathbf{x}}{\mathbf{t}} \exp(-x^2/2) dx = \frac{2}{\sqrt{2\pi}} \exp(-t^2/2)/t$$

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

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

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

# Concentration inequalities

**Theorem 3.2**  $Y_1, \dots, Y_n$  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

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

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

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

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

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

# 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  $Z_n = g(\mathbf{X}_1, \dots, \mathbf{X}_n)$  with  $|g(\mathbf{x}_1, \dots, \mathbf{x}_n) - g(\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}'_i, \dots, \mathbf{x}_n)| \leq c_i$  for all data  $\{\mathbf{x}_j\}_{j=1}^n$  and  $\mathbf{x}'_i$  (changing only one data point from  $\mathbf{x}_i$  to  $\mathbf{x}'_i$ ) and  $\{\mathbf{X}_j\}$  indep, then

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

## 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  $R_n(\hat{\beta}) = n^{-1} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$ , 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  $\sigma^2$ .

# Refitted Cross-validation

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

Key difference: Refitting eliminates spurious correlation!

(*Fan, Guo, Hao, 12*)

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

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

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

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

Final estimate:  $\hat{\sigma}_{\text{RCV}}^2 = (\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$  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

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

**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

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

# 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.