Statistical Foundations of Data Science

Jianqing Fan

Princeton University

https://fan.princeton.edu

ZOOM ID Lectures: **970 4936 8998** Office Hours: **996 4030 7631**

Annotated Lecture Notes: web view

2. Penalized Least-Squares

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) ★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, 12)

see also **chap 4**

Estimating prediction error

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

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

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

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

■ 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

Approx same: $\mathrm{AIC}(\widehat{\mu}) = \log(\|\mathbf{Y} - \widehat{\mu}\|^2/n) + 2\, df_{\widehat{\mu}(\lambda)}/n.$ ★w/o σ^2

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

- C_p *(Mallows, 73)* and AIC *(Akaike, 74)*: $\lambda = 2\sigma^2$, model $\widehat{\mathcal{M}}$
- BIC *(Schwarz, 1978)*: $\lambda = \log(n)\sigma^2$.
- RIC *(Foster & George, 1994)*: $\lambda = 2\log(p)\sigma^2$
- Adjusted-R^2: $R_{adj,m} = 1 - \frac{n-1}{\mathbf{n-m}}\frac{\mathrm{RSS}_\mathbf{m}}{\mathrm{SD}_y}$.
- Generalized cross-validation *(Craven and Wahba, 1979)*: $\mathrm{GCV}(m) = \frac{\mathrm{RSS}_\mathbf{m}}{\mathbf{n(1-m/n)^2}}$, equivalent to PLS with $\lambda = 2\sigma^2$.

Cross-validation

■ Model-free or nonparametric approach to PE *(Allen, 74; Stone, 74)*

Multiple fold CV. Random partition data into equal size subsamples $\{\mathcal{S}_j\}_{j=1}^k$.

Testing and training set: data in $\mathcal{S}_k$ and $\{\mathcal{S}_j\}_{j\neq k}$.

Estimating of PE: $\mathbf{CV}_k(m) = n^{-1}\sum_j\left\{\sum_{i\in\mathcal{S}_j}(Y_i - \widehat{\beta}_{m,-\mathcal{S}_j}^T \mathbf{X}_{i,\mathcal{M}_m})^2\right\}$

■ $\widehat{\beta}_{m,-\mathcal{S}_j}$ = fitted coef of model $\mathcal{M}_m$ w/o using data in $\mathcal{S}_j$.

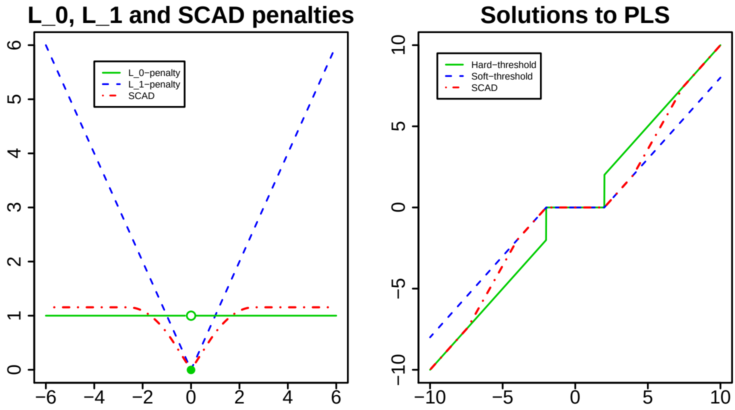
Choice of k: $k = n$ *(best, but expensive; leave-one out)*, 10 or 5 (5-fold).

Why CV: $E(Y_i - \mathbf{X}_i^T\widehat{\beta}_{-i}^{\lambda})^2 = E\varepsilon_i^2 + E\|\underbrace{\mathbf{X}_i^T(\widehat{\beta}_{-i}^{\lambda} - \beta^*)}_{\text{indep}}\|^2$.

2.2 Penalized Least-Squares

Convex and folded concave relaxations

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



(a) (b)

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)*
- ★ Bayesian variable selection: **posterior mode**. *(Mitchell, 88; George & McCulloch 93, Berger and Pericchi, 96 JASA)*.

Folded Concave Penalty

Smoothly Clipped Absolute Deviation (SCAD)

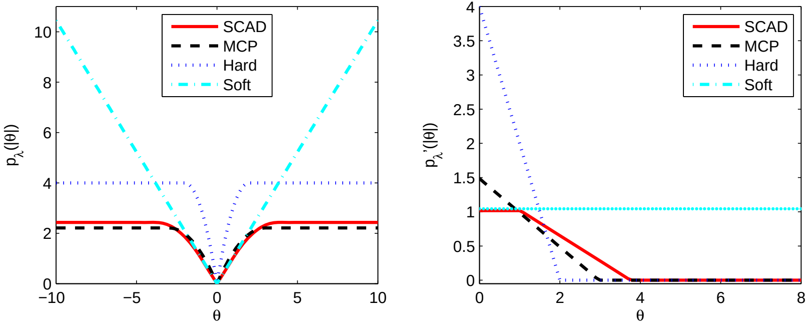
$$p'_\lambda(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\}, \quad a > 2$$

(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 $p_\lambda(\theta) = \lambda^2 - (\lambda - |\theta|)_+^2$.

Folded Concave Penalties and Their Derivaves



- Associated risks computed in the figure on page 17.

Orthonormal design

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

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

$\Longrightarrow$ **comp** min. *(Antoniadis and Fan, 01; Fan and Li, 01)* solutions drawn on page 10

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

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

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

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

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

Desired Properties

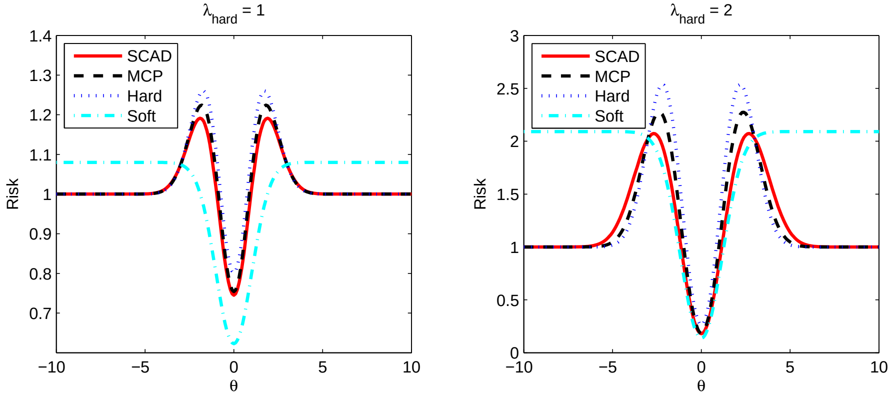
- **Continuity**: to avoid instability in model prediction.
- **Sparsity**: to reduce model complexity (set small coeff. to 0).
- **Unbiasedness**: to avoid unnecessary modeling bias (unbiased when true coefficients are large). (*Fan and Li, 2001*)

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

Performance Comparisons

$Z \sim N(\theta, 1)$ ■ $R(\widehat{\theta}, \theta) = E_\theta(\widehat{\theta} - \theta)^2$ ■ same risk at $\theta = 3$



Characterization of folded-concave PLS

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

- ★ 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 2.1. 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

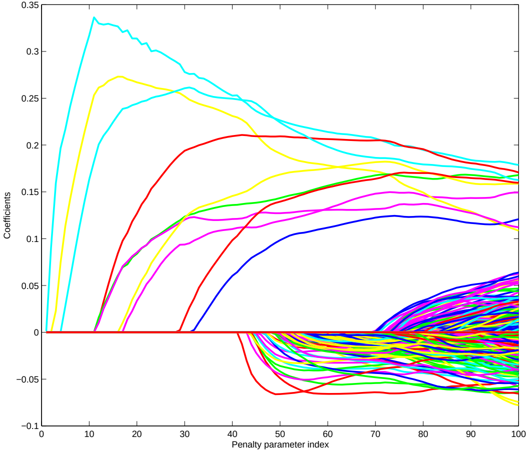
$$
\begin{aligned}
&n^{-1}\mathbf{X}_1^T(\mathbf{Y}-\mathbf{X}\widehat{\beta}) - p'_\lambda(|\widehat{\beta}_1|)\,\mathrm{sgn}(\widehat{\beta}_1) = \mathbf{0},\\
&\|n^{-1}\mathbf{X}_2^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})\|_\infty \le p'_\lambda(0+),\\
&\lambda_{\min}(n^{-1}\mathbf{X}_1^T\mathbf{X}_1)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

They are sufficient if inequalities are strict.

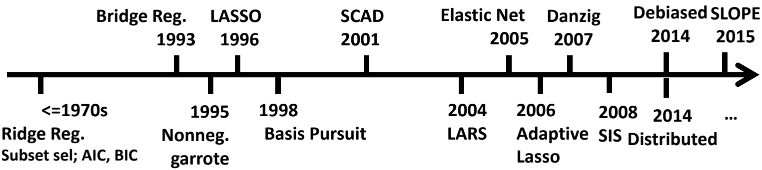
Solution Paths

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:

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

$\Longrightarrow$ $\|n^{-1}\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})\|_\infty\le\lambda$ $\qquad$ $\Longrightarrow$ $\widehat{\beta}=0$ 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: $\mathrm{supp}(\widehat{\beta}) = \mathcal{S}_0$ ➡ $\mathbf{X}_{\mathcal{S}_0} = \mathbf{X}_1$

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

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

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

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

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

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 cond.

★ Example: $X_j = \rho s^{-1/2}\sum_{k\in\mathcal{S}_0} \mathrm{sgn}(\beta_k)X_k + \sqrt{1-\rho^2}\varepsilon_j$, where $s = |\mathcal{S}_0|$. Sum of reg coef is $\sqrt{s}|\rho|$, can exceed 1!

★ Irrespresentable condition is restrictive, leading to false positives, 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: $\widehat{\beta} = \mathrm{argmin}_{\|\beta\|_1 \le c} \|\mathbf{Y} - \mathbf{X}\beta\|^2.$

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

Risks of Lasso (II)

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

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

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

- ★ Sample cov has rate $O(\sqrt{(\log p)/n})$ for subGaussian data.
- ★ Persistency 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 $\max_{i,j} P\{\sqrt{n}|\sigma_{ij} - \widehat{\sigma}_{ij}| > x\} < \exp(-\mathbf{C}\mathbf{x}^{1/a})$ for big x,

$$\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\max} = O_P\left(\frac{(\log \mathbf{p})^{\mathbf{a}}}{\sqrt{\mathbf{n}}}\right).$$

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

Proof:

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

by taking $b_n = (10C^{-1}\log p)^a$. 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.
- ★ Other robust estimators can also be in the regression.
- ★ Sub-Weibull tail: $a = 1/2$ sub-Gaussian and $a = 1$ sub-exponential.

Sparsity of Lasso Solution

- Let $\widehat{\mathbf{\Delta}} = \widehat{\beta} - \beta_0$, $\beta^* = \beta_0$ and $S = \mathcal{S}_0$. Then, $F(\widehat{\mathbf{\Delta}}) \le 0$.
- $F(\mathbf{\Delta}) = \underbrace{R_n(\mathbf{\Delta}+\beta_0)/2 - R_n(\beta_0)/2}_{(1)} + \lambda(\underbrace{\|\mathbf{\Delta}+\beta_0\|_1 - \|\beta_0\|_1}_{(2)})$

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

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

$$(2) \quad = \quad \|\mathbf{\Delta}_S + \beta_S^*\|_1 + \|\mathbf{\Delta}_{S^c}\|_1 - \|\beta_S^*\|_1 \ge -\|\mathbf{\Delta}_S\|_1 + \|\mathbf{\Delta}_{S^c}\|_1$$

Combining these and $F(\widehat{\mathbf{\Delta}}) \le 0$, we have $\|\widehat{\mathbf{\Delta}}_{S^c}\|_1 \le 3\|\widehat{\mathbf{\Delta}}_S\|_1$.

Weighted and adaptive lasso and SLOPE

Lasso: Creating biases for large coefficients.

Weighted lasso: $\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\beta\|^2+\lambda\sum_{j=1}^{p} w_j|\beta_j|$.

★solved by Lasso via rescaling: $\alpha=\beta\cdot\mathbf{w}$, $\mathbf{X}^*=\mathbf{X}/\mathbf{w}$.

Adaptive lasso: $w_j=|\beta_j|^{-\gamma}$ (e.g, $\gamma=0.5,1,2$) *(Zou, 06)*

★One-step implementation of folded-concave penalty $p_\lambda(\cdot)$:

$$w_j=p_\lambda'(|\widehat{\beta}_j^{\text{lasso}}|)/\lambda.$$

SLOPE: $\operatorname{argmin}_\beta\left\{\frac{1}{n}\|\mathbf{y}-\mathbf{X}\beta\|^2+\sum_{j=1}^{p}\lambda_j|\beta|_{(j)}\right\}$,

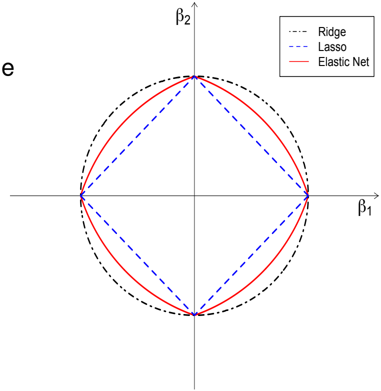
★$\lambda_j=\Phi^{-1}(1-jq/2p)\sigma/\sqrt{n}$, related to FDR q.

Elastic Net

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

★Mitigate collinearity

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

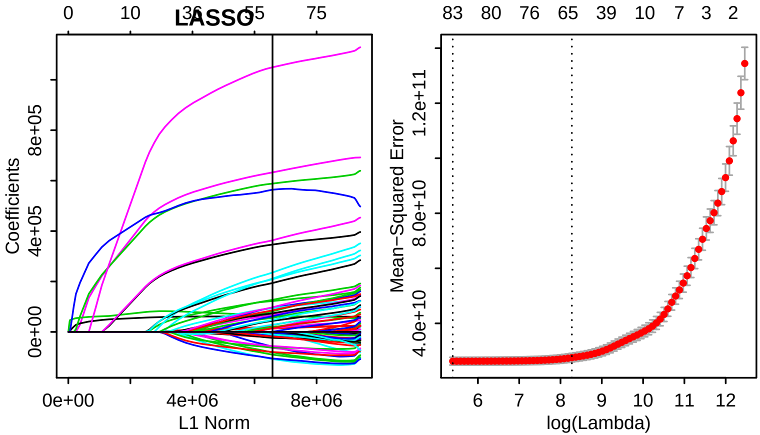


Housing Prediction: Zillow data: House price + house attributes (bedrooms, bathrooms, sqft.living sqft.lot, ...) in 70 zip codes, sold in 2014 & 15. ($n_{train} = 15,129$, $n_{test} = 6,484$).

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 alpha = 0.5?
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$glmnet.fit); title('LASSO')
abline(v=sum(abs(beta.cvglm1[-1])))  #Lasso solution path
plot(fit.cvglm1)                      #Estimated MSE
  ##### compute testing errors  ############
X_test <- model.matrix(~., data = test_data[,5:22])  ##test data model
pred.glm1<- predict(fit.cvglm1, newx = X_test, s = "lambda.min")   ## computed predicted value
mse.pred.glm1 <- sum((test_data$price - pred.glm1)^2)      #MSE
1 - mse.pred.glm1/ sum((test_data$price - mean(Y))^2)      #out-sample-R^2
```

Housing Prediction: Zillow data



Danzig Selector

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

high confident set

■ $\widehat{\beta}_{\mathrm{DZ}} =$ solution; ■ require $\lambda \geq \|n^{-1}\mathbf{X}^T\underbrace{(\mathbf{Y} - \mathbf{X}\beta_0)}_{\varepsilon}\|_\infty$

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

(*Candés & Tao, 07*)

- ★ Let $\widehat{\boldsymbol{\Delta}} = \widehat{\beta}_{DZ} - \beta_0$. Then, $\|\beta_0\|_1 \geq \|\widehat{\beta}_{\mathrm{DZ}}\|_1 = \|\beta_0 + \widehat{\boldsymbol{\Delta}}\|_1$.
- ★ As in (2), $\|\beta_0 + \widehat{\boldsymbol{\Delta}}\|_1 \geq \|\beta_0\|_1 - \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1 + \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0^c}\|_1$.
- ★ They imply $\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1 \geq \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0^c}\|_1$ —sparsity

Rate of Convergence

★ $\|\widehat{\boldsymbol{\Delta}}\|_2 \geq \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_2 \geq \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1/\sqrt{s} \geq \|\widehat{\boldsymbol{\Delta}}\|_1/(2\sqrt{s})$

★ $\|\mathbf{X}\widehat{\boldsymbol{\Delta}}\|_2^2 \leq \|(\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\Delta}})\|_\infty \|\widehat{\boldsymbol{\Delta}}\|_1 \leq 2n\lambda\|\widehat{\boldsymbol{\Delta}}\|_1.$

Combine last two, we have

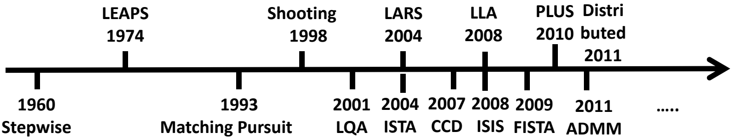
$$\|\mathbf{X}\widehat{\boldsymbol{\Delta}}\|_2^2 \leq 4n\lambda\sqrt{s}\|\widehat{\boldsymbol{\Delta}}\|_2.$$

Restricted eigenvalue condition: *(Bickel, et al, 09)*

$$\min_{\|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0}\|_1 \geq \|\widehat{\boldsymbol{\Delta}}_{\mathcal{S}_0^c}\|_1} n^{-1}\|\mathbf{X}\widehat{\boldsymbol{\Delta}}\|_2^2/\|\widehat{\boldsymbol{\Delta}}\|_2^2 \geq a$$

L_2 consistency: $\|\widehat{\boldsymbol{\Delta}}\|_2^2 \leq 16a^{-2}\lambda^2 s$

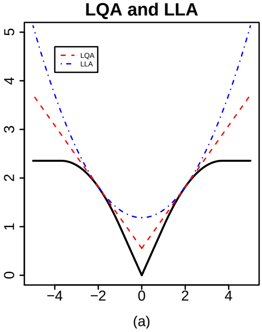
3.4 Algorithms for Folded-concave Regularization



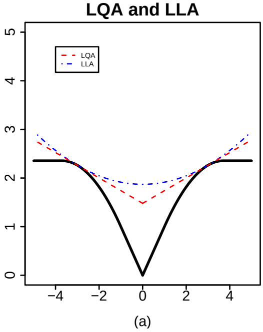
$$Q(\beta) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

Local quadratic approximations

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



(a)



(a)

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

$$\beta^{(k+1)} = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\beta^{(k)})\}^{-1}\mathbf{X}'\mathbf{y}, \quad \Sigma_\lambda(\beta^{(k)}) = \mathrm{diag}\{p'_{\lambda_1}(|\beta^{(k)}|)/|\beta_{(k)}|\}$$

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

Local linear approximations and one-step estimator

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

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

- ★ Fan et al (14) shows that if initial estimator is good, one-step procedure obtains the **oracle** estimator, and is a fixed point.
- ★ Later, algorithmic construction of **oracle** estimator.
- ★ Optimality demonstrated in Chap 4 of book.

Remarks

- LLQ computes explicitly the updates whereas LLA requires PLS.
- LLA gives better approximation, particularly around the origin.
- With LLA, PLS is iter. reweighted LASSO, w/ weights given by $p'_\lambda(\cdot)$.
- 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:

$$Q(\beta) \le Q^{\text{approx}}(\beta), \qquad Q(\beta^*) = Q^{\text{approx}}(\beta^*)$$

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

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

Coordinate descent algorithms

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

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

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

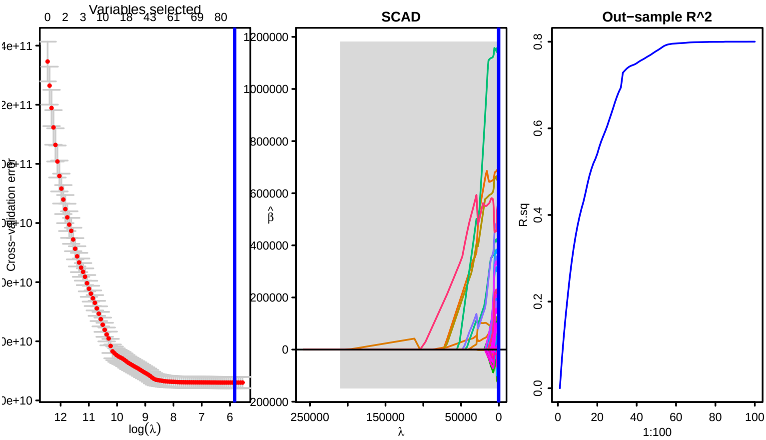
$$\begin{aligned} Q_j(\beta_j) &\equiv Q(\beta_{10}, \cdots, \beta_{j-1,0}, \beta_j, \beta_{j+1,0}, \cdots, \beta_{p,0}) \\ &= \frac{1}{2n}\|\mathbf{R}_j - \mathbf{X}_j\beta_j\|^2 + p_\lambda(|\beta_j|) + c, \end{aligned}$$

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

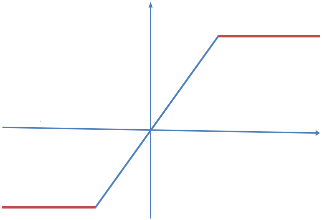
Housing Prediction by SCAD: Zillow data: $R^2 = 0.80$

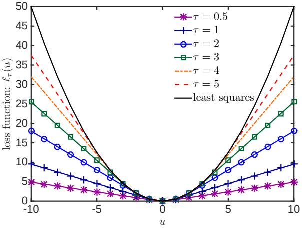
```
library('ncvreg')                  #loading the library for use
X <- model.matrix(~.,data = train_data[,5:22])
Y <- train_data$price              #data model for training
cvfit.SCAD <- cv.ncvreg(X, Y, penalty="SCAD")   #SCAD using CV

### 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$lambda.min),lwd=2,col=4)
fit.SCAD <- ncvreg(X, Y, penalty="SCAD")
plot(fit.SCAD, main="SCAD")
abline(v=cvfit.SCAD$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 estimation errors: e.g.

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

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

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

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

Concentration inequalities

Theorem 2.2 $Y_1, \cdots, Y_n$ 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

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

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

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

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

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

Concentration inequalities for robust mean

Robust means. Y_i are i.i.d. with mean μ and variance σ^2.

d) **Truncated loss** (Adaptive Huber estimator): Let

$$\widehat{\mu}_\tau = \operatorname{argmin} \sum_{i=1}^n \rho_\tau(Y_i - \mu), \qquad \rho_\tau(x) = \begin{cases} x^2, & \text{if } |x| \le \tau \\ \tau(2|x| - \tau), & \text{if } |x| > \tau \end{cases}$$

Then, for $\tau = \sqrt{n}c/t$ with $c \ge \mathrm{SD}(Y)$, we have *(Fan, Li, Wang, 17)*

$$P(|\widehat{\mu}_\tau - \mu| \ge t\frac{c}{\sqrt{n}}) \quad \le \quad 2\exp(-t^2/16), \quad \forall t \le \sqrt{n/8}.$$

e) **Truncated mean**: Set $\widetilde{Y}_i = \operatorname{sgn}(Y_i)\min(|Y_i|, \tau)$.

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i - \mu\Big| \ge t\frac{\sigma}{\sqrt{n}}\Big) \le 2\exp\Big(-ct^2\Big)$$

for some universal constant c. *(Fan, Wang, Zhu, 21)*.

- Sample mean has Cauchy tail: $P(|\bar{Y} - \mu| > t\frac{\sigma}{\sqrt{n}}) \le 1/t^2$.

Remarks

There are concentration inequalities for

- ★ sum of sub-exponential distributions
- ★ 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 $|g(\mathbf{x}_1, \cdots, \mathbf{x}_n) - g(\mathbf{x}_1, \cdots, \mathbf{x}_{i-1}, \mathbf{x}_i', \cdots, \mathbf{x}_n)| \leq c_i$ 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

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

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 persistency.

★ Slow rate and require $\|\beta_0\|_1 \leq c = o((n/\log p)^{1/4})$.

★ 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!

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

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

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

Final estimate: $\widehat{\sigma}_{\mathbf{RCV}}^2 = (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)/2$ 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. B-spline). Additive model becomes an extended 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

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

Full Nonparametrics and Kernel Tricks

Saturated nonparametric model: $Y = f(\mathbf{X}) + \varepsilon$. **curse of dimensionality**

Kernel Tricks: $Y_i = \sum_{j=1}^n \beta_j \underbrace{K(\mathbf{X}_i/\lambda, \mathbf{X}_j/\lambda)}_{X_{ij}^* \ \textit{prototype}} + \varepsilon$ $\quad f(\mathbf{x}) \approx \sum_{j=1}^n \beta_j K(\mathbf{x}/\lambda, \mathbf{X}_j/\lambda)$

★Polynomial kernel $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^d$ for a positive integer d

★Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/2)$

★Coefficients β_j can be regularized ($p = n$): $\|\mathbf{y} - \mathbf{K}\beta\|^2 + \lambda^* \beta^T \mathbf{K} \beta$

■Also applicable to structured nonparametrics.

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: $\frac{1}{2n}\|\mathbf{Y} - \sum_{j=1}^{p} \mathbf{X}_j\beta_j\|^2 + \sum_{j=1}^{p} p_\lambda(\|\beta_j\|_{W_j})$

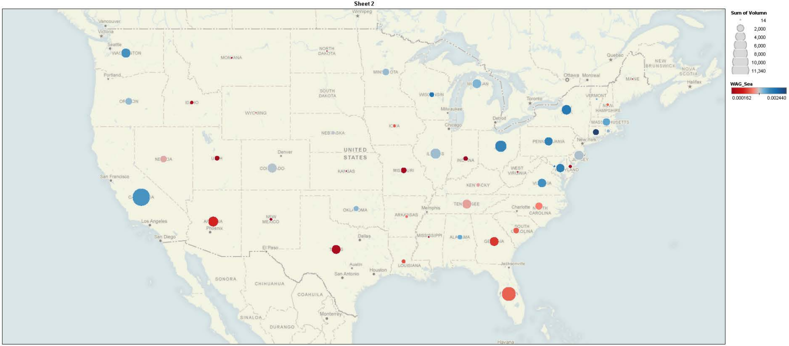
★ Select or kill a group of variables. *(Lin & Yuan, 06)*

★ Appeared in Antoniadis & Fan (01) for selecting blocks of wavelet coefficients.

2.8 Prediction of HPI

Prediction of Home Price Appreciation

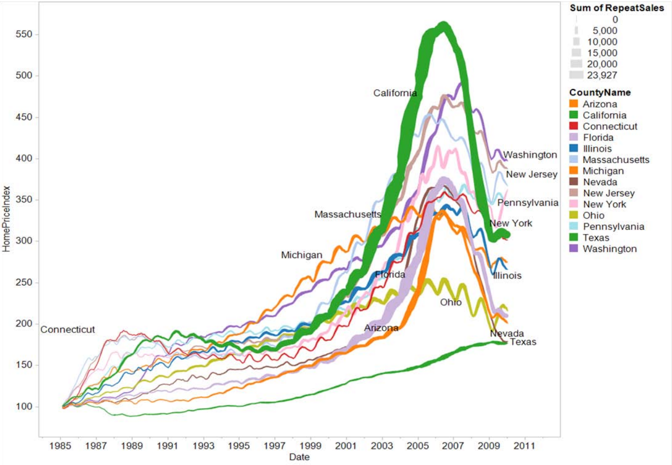
Data: HPA collected at "≈" 1000 Core Based Statistical Areas.



Map based on Longitude (generated) and Latitude (generated). Color shows WAG_Sea. Size shows sum of Volumn. Details are shown for State.

Objective: To project HPA over 30-40 years for approx 1000 CBSAs based on national assumption.

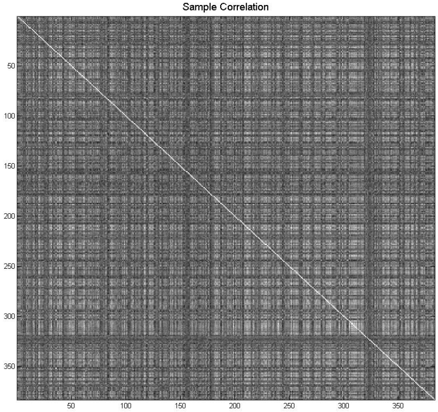
HPI in several states



Date vs. HomePriceIndex. Color shows details about CountyName. Size shows sum of RepeatSales. The view is filtered on CountyName and Date. The CountyName filter keeps 14 members. The Date filter ranges from 1/22/1985 to 1/1/2010. The marks are labeled by CountyName.

Growth of dimensionality

■ Local correlation makes dimensionality growths quickly (1000 neighborhoods requires 1 m parameters).



Conditional sparsity

Model: Y_{t+1} is the HPA in one CBSA:

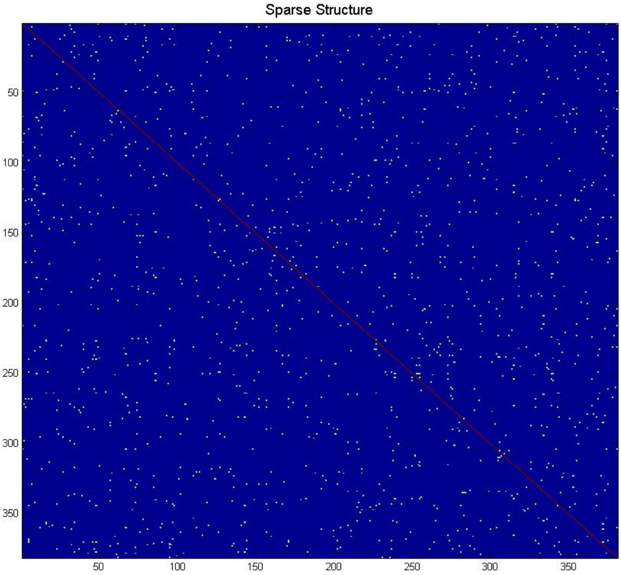
$$Y_{t+1} = \beta_0 + \beta_1 \mathbf{X_{N,t}} + \sum_{j=2}^{382} \beta_j X_{t,j} + \varepsilon_t$$

★Sparsity of $\{\beta_j\}_{j=2}^{382}$ are explored by SCAD

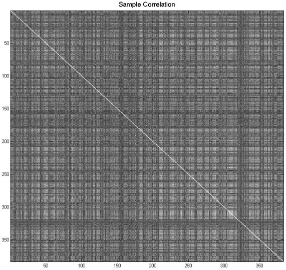
Results: **30% more accurate** than the simple modeling

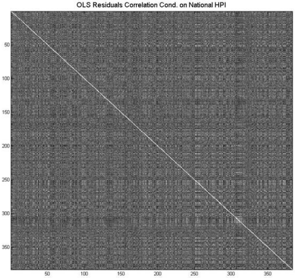
What is benchmark?

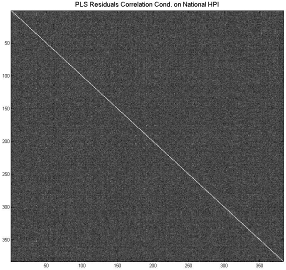
Local neighborhood selection



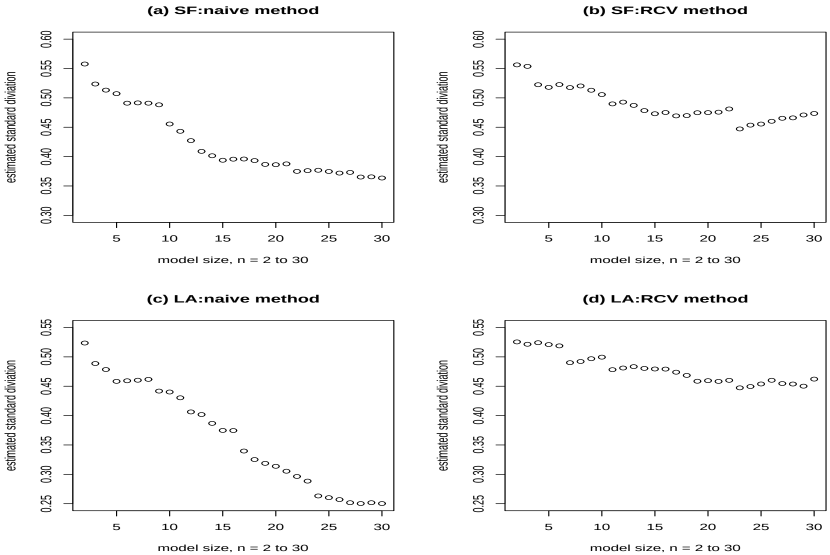
Effectiveness of sparse modeling







Estimated benchmark with different model size



Estimating of prediction errors

San Francisco

Model size	2	3	5	10	15	20	30
N-LASSO	0.558	0.524	0.507	0.456	0.394	0.386	0.364
RCV-LASSO	0.556	0.554	0.518	0.506	0.473	0.475	0.474
R^2 (in %)	76.92	79.83	81.40	85.67	89.79	90.66	92.58

Los Angeles

Model size	2	3	5	10	15	20	30
N-LASSO	0.524	0.489	0.458	0.440	0.375	0.314	0.250
RCV-LASSO	0.526	0.521	0.521	0.499	0.479	0.460	0.462
R^2 (in %)	88.68	90.23	91.56	92.56	94.86	96.57	98.05

Notes

Data period: **Training**: 01/01 – 12/05 **Testing** 01/06 – 12/09

SD of monthly variation: SF = 1.08% LA = 1.69%

Benchmark: 0.53%.

One-month PE in 2006: SF = .67% LA = .86%