Lecture 4: Variable Selection and LASSO

Statistical Learning and Data Mining

Xingye Qiao

Department of Mathematical Sciences
Binghamton University

E-mail: qiao@math.binghamton.edu

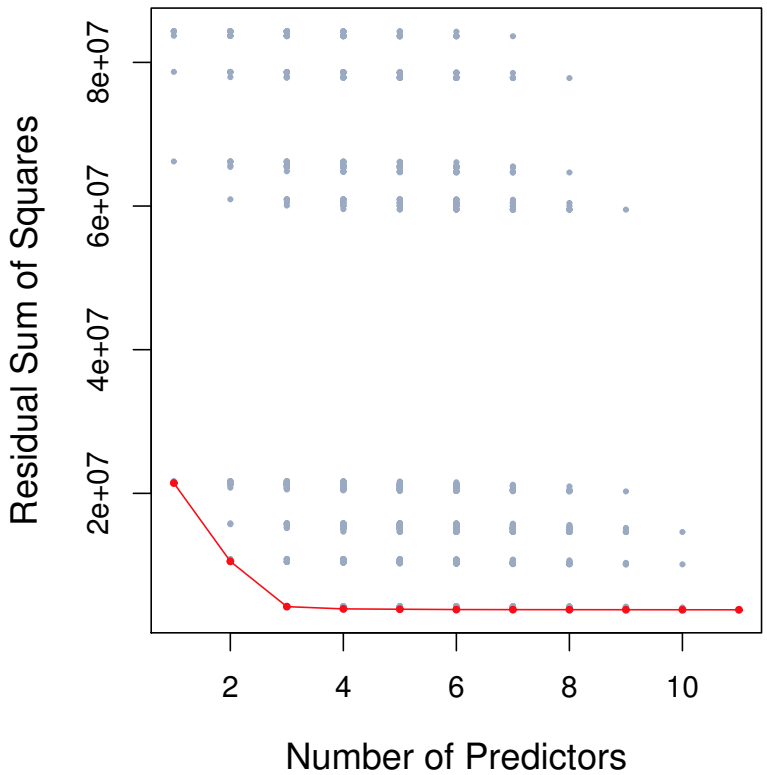
Read: ISR Chs. 6.1–6.2, 6.5–6.6 and ESLII Chs. 3.3–3.4 and SLS Ch. 2

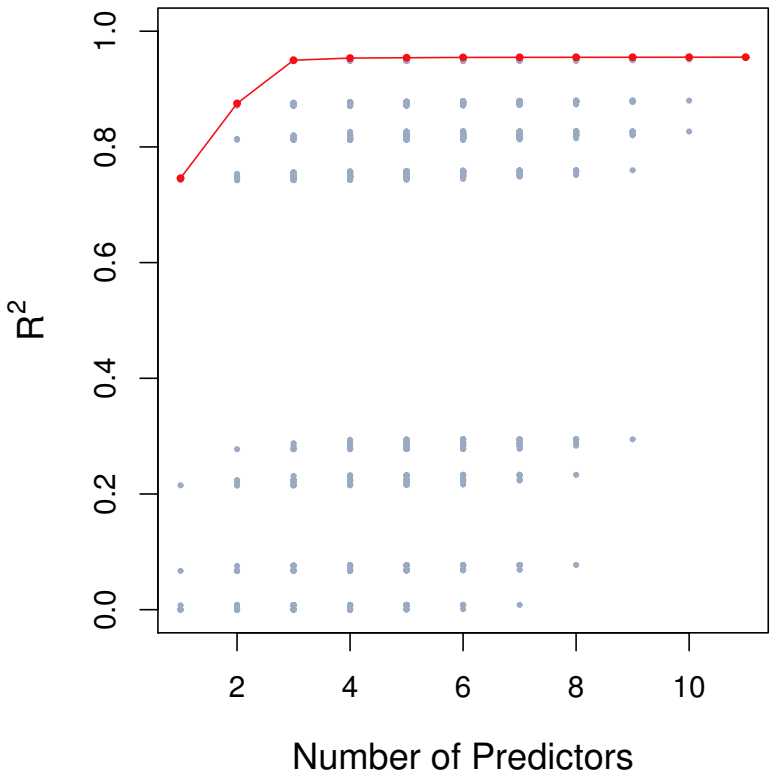
Outline

The next section would be . . . . . .

Combinatorial approaches

- Search through all possible submodels and identify the best one (under certain criterion).
- Among p variables, there are $p(p-1)/2$ potential submodels with 2 variables to inspect, and $2^p - 1$ submodels with different sizes.
- Often conducted in steps:
  - Step 1: For each $k = 1, \ldots, p$, fit all $\begin{bmatrix} p \\ k \end{bmatrix}$ submodels with exactly k variables
  - Step 2: For each k, find the best submodel with the smallest RSS.
  - Step 3: Select a single best submodel from among the p submodels above, using certain criterion.
- Selection Criterion: cross-validated prediction error, adjusted R^2, Mallow's C_p, PRESS, AIC and BIC.
- Very computationally inefficient.


Residual Sum of Squares
2e+07
4e+07
6e+07
8e+07
2
4
6
8
10
Number of Predictors

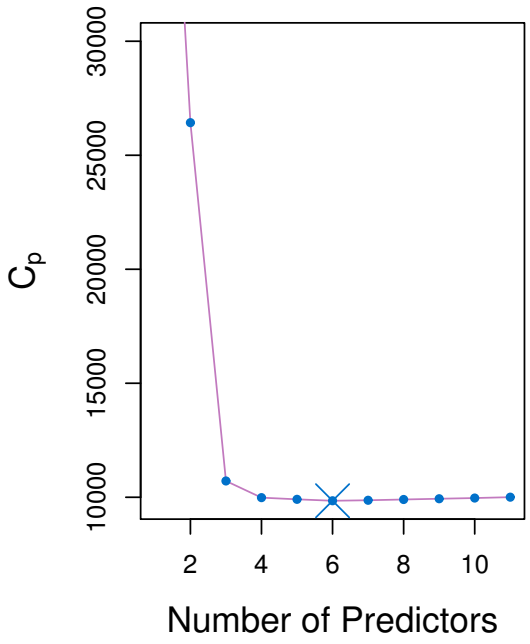

R^2
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
Number of Predictors

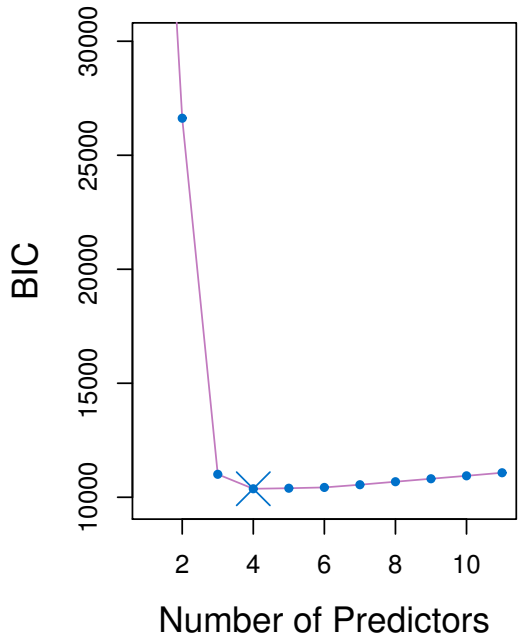
Stepwise Selection

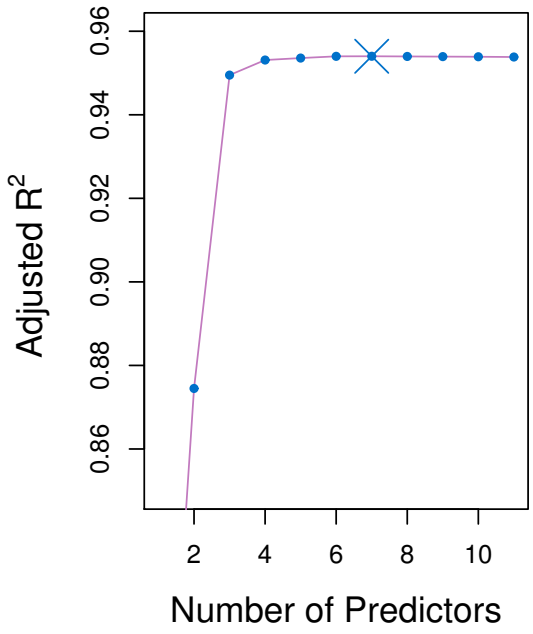
- For computational reasons, best subset selection cannot be applied with very large p
- Stepwise methods are attractive alternatives.
- Forward, backward, hybrids, etc.
- Go through a much smaller collection of submodels in a greedy way
- Select a single best model using certain criterion in the end.

Selection Criteria

- Common idea: indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting.
- Suppose there are d predictors in a submodel
  - $\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2)$
  - $\text{BIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + \log(n)d\hat{\sigma}^2)$
  - Adjusted $R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$
- AIC (C_p) and BIC both have rigorous theoretical justifications; though popular and intuitive, adjusted R^2 is not as well motivated.


C_p
10000
15000
20000
25000
30000
2
4
6
8
10
Number of Predictors


BIC
10000
15000
20000
25000
30000
2
4
6
8
10
Number of Predictors


Adjusted R^2
0.86
0.88
0.90
0.92
0.94
0.96
2
4
6
8
10
Number of Predictors

ℓ_0 regularization

- Recall AIC and BIC
  - AIC $= \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2d\hat{\sigma}^2)$
  - BIC $= \frac{1}{n\hat{\sigma}^2}(\text{RSS} + \log(n)d\hat{\sigma}^2)$
- Both can be written as

$$\propto \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}$$

  with different λ choices ($2\hat{\sigma}^2$ and $\log(n)\hat{\sigma}^2$ resp.)
- $\sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\} = (\sum_{j=1}^{p} |\beta_j|^0)$ the ℓ_0 "norm" of $\boldsymbol{\beta}$
- Note that ℓ_0 "norm" is not actually a proper norm, but this allows a nice connection with the ridge regression (ℓ_2 norm) and lasso (ℓ_1 norm).

The next section would be . . . . . .

ℓ_1: best convex relaxation of ℓ_0

- Best subset, forward stepwise, and backward stepwise selection...: select a subset of the variable, achieving variable selection. However, **computationally infeasible** for large p
- Ridge regression is computational attractive, but **does not really select variables**.
- What we need: something that both selects variable, and is easy to compute.
- One bottleneck of ℓ_0 regularization (as AIC, BIC) is that it is not convex, and hence we cannot borrow the strength of many efficient optimization tool.
- ℓ_1 norm ($\sum_{j=1}^{p} |\beta_j|^1$) is convex, and fairly close to the ℓ_0 norm ($\sum_{j=1}^{p} |\beta_j|^0$).

Linear regression with ℓ_1 regularization

- Relatively recent alternative to ridge regression.
- Lasso: least absolute shrinkage and selection operator

Alamy/Lisa Dearing

Linear regression with ℓ_1 regularization

- Compared to ridge regression, β_j^2 term replaced by $|\beta_j|$
- Lagrangian form:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \; RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Recall ridge regression:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \; RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Best subset selection w/ AIC (and $\lambda = 2\hat{\sigma}^2$)

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \; RSS + \lambda \sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\}$$

Constrained form

- Lasso has the following constrained form:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\operatorname{argmin}} \; RSS$$

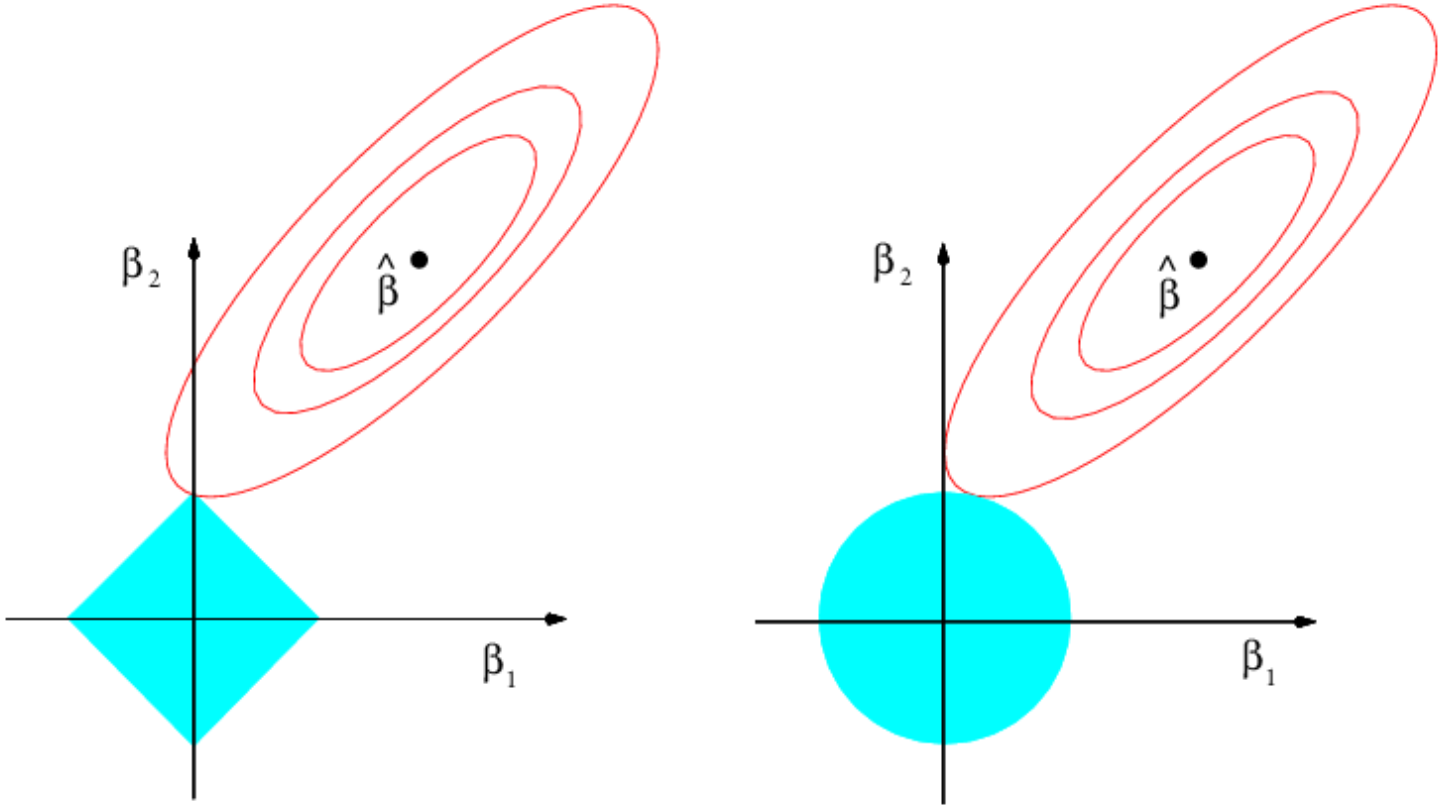
$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s.$$

- Compared to $\sum_{j=1}^{p} \beta_j^2 \leq s$ (ridge regression) or $\sum_{j=1}^{p} \mathbb{1}\{\beta_j \neq 0\} \leq s$ (best subset)

Remarks

- Solution not linear in $\boldsymbol{y}$.
- No closed form solution (except in special cases)
- Important sparsity (variable selection) property: unlike ridge regression, results in coefficient estimates that are exactly equal to zero.
- Note for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have

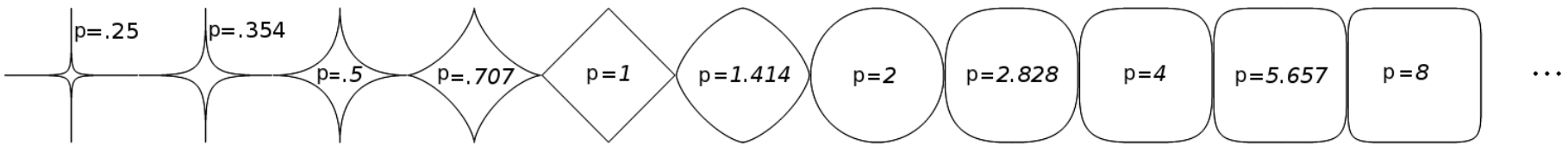
$$
\begin{aligned}
&(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\
=&(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})^T(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}) + (\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} - \mathbf{X}\boldsymbol{\beta}) \\
=&\underbrace{(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})^T(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})}_{\text{independent of } \boldsymbol{\beta}} + (\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})
\end{aligned}
$$



- Each point = a possible estimate.
- Ellipsoid centered at the OLS estimate.
- Points on the same ellipsoid have same RSS.
- Cyan diamond and circle = ℓ_1 and ℓ_2 norm constraints.
- ℓ_1 constraint lead to zero estimate more easily.

ℓ_q norm balls

One may consider more general ℓ_q norm as the penalty function.



- $q = 0$: subset selection. computationally infeasible.
- $q = 1$: Lasso. Sparsity property.
- $q = 2$: Ridge regression. No sparsity. But shrink correlated variables together.

A special case: orthonormal input $\mathbf{X}$

Let us suppose that $\mathbf{X}$ is orthonormal: columns are orthogonal, and each has unit norm. Then $\mathbf{X}^T\mathbf{X} = \mathbb{I}$ and $\hat{\beta}_{OLS} = \mathbf{X}^T\boldsymbol{y}$.

In this case, the three procedures (subset selection, ridge regression and lasso) have explicit solutions.

Orthonormal input: subset selection

Define $|\hat{\beta}|_{(k)}$ as the kth largest number among all $|\hat{\beta}_j^{OLS}|$'s ($j = 1, \ldots, p$). It can be shown that the subset selection estimator for β_j is

$$\hat{\beta}_j^{OLS} \mathbb{1}\{\hat{\beta}_j^{OLS} \geq |\hat{\beta}|_{(M)}\}$$

if we are to select M variables. This is named **hard thresholding.**

Note that

$$(\boldsymbol{y} - \mathbf{X}_{(-k)}\hat{\boldsymbol{\beta}}_{(-k)}^{OLS})^T(\boldsymbol{y} - \mathbf{X}_{(-k)}\hat{\boldsymbol{\beta}}_{(-k)}^{OLS}) = \text{RSS} + \mathbf{X}_k^T\mathbf{X}_k\hat{\beta}_{OLS,k}^2 = \text{RSS} + \hat{\beta}_{OLS,k}^2$$

Orthonormal input: ridge regression

Easy to see that the ridge regression estimator for β_j is

$$\hat{\beta}_j^{OLS}/(1+\lambda)$$

Orthonormal input: Lasso

We first define Soft-Thresholding operator:

$$S_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$$

where $(a)_+$ means $\max(a, 0)$. In addition, we suppose that $\boldsymbol{y}$ has been standardized (so that intercept is not considered).
The Lasso estimator for β_j is

$$S_{\lambda/2}(\hat{\beta}_j^{OLS})$$

$$
\begin{aligned}
&\text{RSS} \\
=&(\boldsymbol{y} - \mathbf{X}_k\beta_k - \mathbf{X}_{(-k)}\boldsymbol{\beta}_{(-k)})^T(\boldsymbol{y} - \mathbf{X}_k\beta_k - \mathbf{X}_{(-k)}\boldsymbol{\beta}_{(-k)}) \\
=&(\boldsymbol{y} - \mathbf{X}_k\beta_k)^T(\boldsymbol{y} - \mathbf{X}_k\beta_k) + \boldsymbol{\beta}_{(-k)}^T\mathbf{X}_{(-k)}^T\mathbf{X}_{(-k)}\boldsymbol{\beta}_{(-k)} \\
&\quad - 2(\boldsymbol{y} - \mathbf{X}_k\beta_k)^T\mathbf{X}_{(-k)}\boldsymbol{\beta}_{(-k)} \\
=&(\boldsymbol{y} - \mathbf{X}_k\beta_k)^T(\boldsymbol{y} - \mathbf{X}_k\beta_k) + \underbrace{\boldsymbol{\beta}_{(-k)}^T\mathbf{X}_{(-k)}^T\mathbf{X}_{(-k)}\boldsymbol{\beta}_{(-k)} - 2\boldsymbol{y}^T\mathbf{X}_{(-k)}\boldsymbol{\beta}_{(-k)}}_{\text{independent of } \beta_k}
\end{aligned}
$$

Hence minimizing RSS with respect to β_k can be done by minimizing $(\boldsymbol{y} - \mathbf{X}_k\beta_k)^T(\boldsymbol{y} - \mathbf{X}_k\beta_k)$ alone with no regard to the other variables. Hence in orthonormal design, only need to consider a single variable regression problem when try to estimate β_k.

Lasso for single variable

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta|$$

- Standard approach? Take gradient. Solve gradient equation.
- However, $|\beta|$ does not have derivative at 0.
- If $|\hat{\beta}_{OLS}| \geq \frac{\lambda}{2} > 0$, then the gradient equation is

$$\begin{aligned} 0 &= -\boldsymbol{x}^T(\boldsymbol{y} - \boldsymbol{x}\beta) + \frac{\lambda}{2}\text{sign}(\beta) \\ &= -\hat{\beta}_{OLS} + \beta + \frac{\lambda}{2}\text{sign}(\beta) \\ \Rightarrow \beta &= \hat{\beta}_{OLS} - \frac{\lambda}{2}\text{sign}(\beta) \end{aligned}$$

- If $0 \leq \hat{\beta}_{OLS} < \frac{\lambda}{2}$, then the gradient $-\hat{\beta}_{OLS} + \beta + \frac{\lambda}{2}\text{sign}(\beta)$ is

$$\begin{cases} < 0, \text{ as } \beta < 0 \\ > 0, \text{ as } \beta > 0 \end{cases} \Rightarrow \beta = 0 \text{ is at the minimal.}$$

Lasso solution in orthonormal design

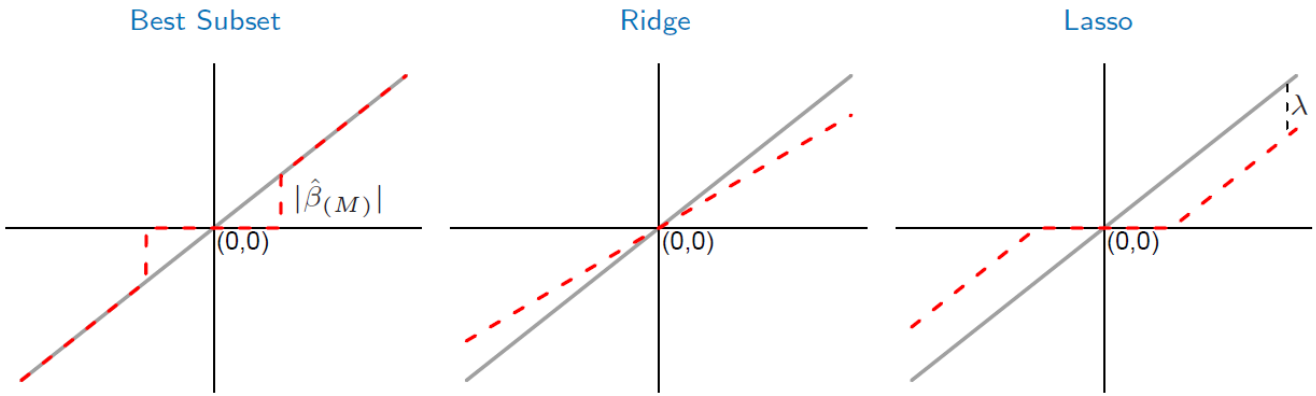
- In summary, the lasso solution for single variable regression is

$$S_{\frac{\lambda}{2}}(\hat{\beta}_{OLS})$$

- Hence, the Lasso estimator for β_j in orthonormal design is

$$S_{\frac{\lambda}{2}}(\hat{\beta}_j^{OLS})$$

Comparison between subset selection, ridge regression and Lasso



- Ridge regression does a proportional shrinkage.
- Lasso translates each coefficient by a constant factor, truncating at zero. (Soft-thresholding)
- Best-subset selection drops all variables with coefficients smaller than the Mth largest (Hard-thresholding)

Interpreting Coefficients.

- Unlike ridge regression, the lasso performs variable selection, and hence results in models that are easier to interpret.
- Zero coefficient $=$ variable does not contribute to $\boldsymbol{y}$
- Remember that lasso is an operator – meaning that it can be manipulated.
- If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.

The next section would be . . . . . .

Solution Paths

- A Solution Path is a path of the coefficient estimate for all different tuning parameter values.
- Lasso has piece-wise linear solution paths.
- Lasso has various different implementations, ranging from very slow to very fast.

Quadratic program with linear constraints.

- The lasso problem is essentially a quadratic programming problem with $2p$ linear constraint.
- Recast β_j as $\beta_j = \beta_j^+ - \beta_j^-$ with constraints that $\beta_j^+ > 0$ and $\beta_j^- > 0$ (hence $2p$ constraints)
- Replace $\|\boldsymbol{\beta}\|_1$ by $\sum_{j=1}^{p}(\beta_j^+ + \beta_j^-)$
- Very inefficient for large p.
- Must train $\boldsymbol{\beta}$ for each fixed λ
- Want to have a method to calculate the full solution path corresponding to a full span of λ values.

Least angle regression (LARS)

- Forward stepwise regression: an ancient idea. To add one variable at a time. At each step, identify the best variable to include in the model, and then update the least square fit sequentially.
- Least angle regression (LARS): a similar idea. But instead of exploiting the current variable at much as possible, LARS only fits the current variable to a certain level.

LARS–1

- At first step, identify the variable most correlated with the response.
- "Slowly" increase the coefficient for this variable (along the direction of the LS fit)
  - This causes the correlation between this variable and the residual of the fit to decrease (as the coefficient slowly increases).
  - . . . , until another variable has the same absolute correlation (with the residual) as the current one. At this point, both variables are in the *active* set.

LARS–2

- At each point when a new variable enters the active set (denoted as $\mathcal{A}$), "slowly" increase the coefficient of the variables in $\mathcal{A}$ along the direction of the LS fit of the current residual on $\mathcal{A}$, i.e.

$$\boldsymbol{\beta}_{\mathcal{A}}(\alpha) = \boldsymbol{\beta}_{\mathcal{A}}(0) + \alpha\boldsymbol{\delta}$$

  where $\boldsymbol{\delta} = (\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T\boldsymbol{r}$ and $\boldsymbol{r} = \boldsymbol{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}(0)$. Note that the variable which just entered the active set had coefficient 0.
  - An important feature of this process is that the correlations between the variables in $\mathcal{A}$ and the residual $\boldsymbol{r}(\alpha) := \boldsymbol{y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}}(\alpha)$ are all equal (to each other) and they decrease at the same time as $\alpha \uparrow$
- . . . , until there is another variable which has as much correlation with the residual $\boldsymbol{r}(\alpha)$ as the ones in the active set $\mathcal{A}$, at which point, the new variable enters the active set and the iterations restart.

- At each time point, there is a value for $\boldsymbol{\beta}$. This corresponds to a value of $t = \sum_{j=1}^{p} |\beta_j|$
- At the very beginning, there was no variable in the active set. This corresponds to $t = 0$, or $\lambda = \infty$
- In the very end, all variables are in the active set. This corresponds to $t = \sum_{j=1}^{p} |\hat{\beta}_j^{OLS}|$, or $\lambda = 0$

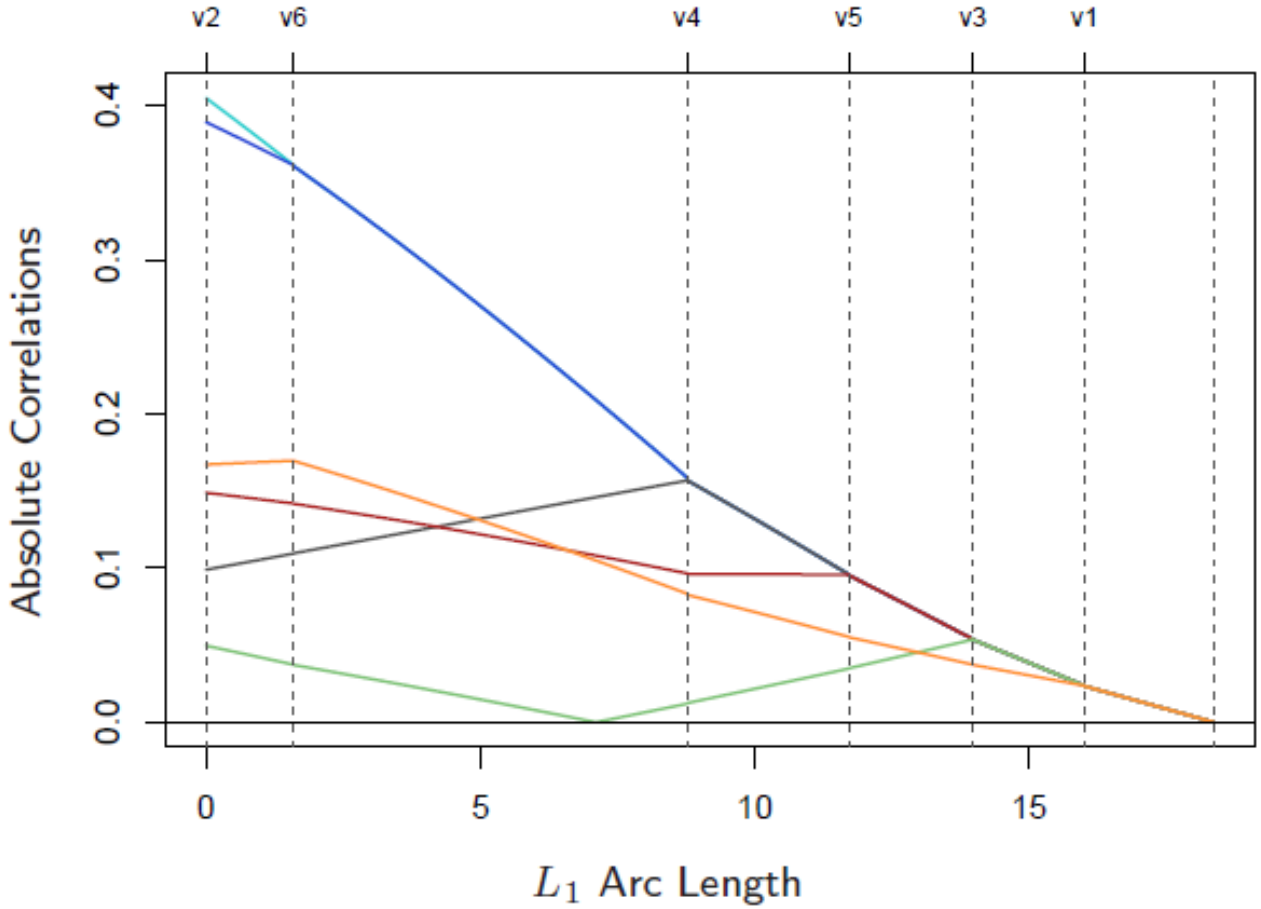
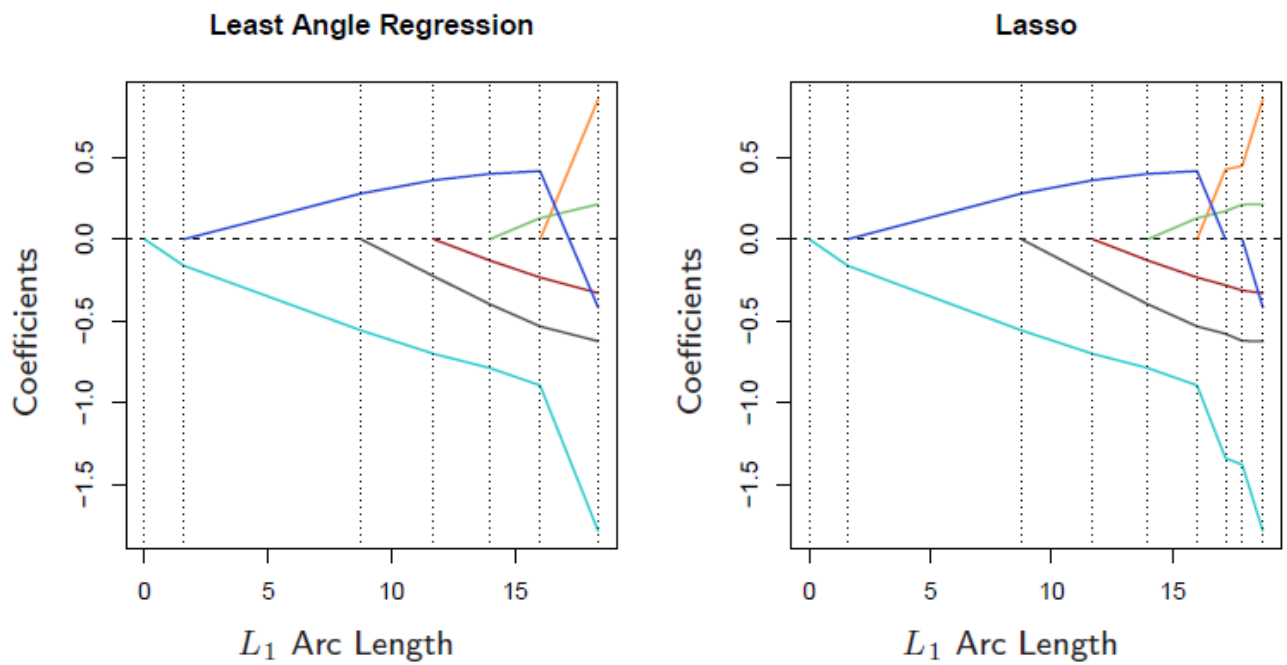


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

LARS and LASSO

- The solutions to LARS and LASSO are very similar. With a small modification, the LARS algorithm produces the whole solution path of LASSO.
- Modification: if one variables in the active set has a correlation (with the residual) 0, then drop the variable out of the active set, recalculate the direction of increment $\boldsymbol{\delta} = (\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}}^T \boldsymbol{r}$



Properties of LARS

- We can exactly calculate the needed step size at the beginning of each step, and do not need to carefully take many tiny steps and recheck the correlation for many times.
- The solution path is piecewise linear (with respect to t). Only need to work out the values at turning points.
- Extremely fast.

Coordinate descent

Coordinate descent algorithm

1. Fix all the other variables, and vary the coefficient for one variable β_j to achieve the minimal of $Q(\beta_j|\boldsymbol{\beta}_{-j})$
2. Do the same for all the variables.
3. Iterate until convergence

This is actually pretty easy and fast.

$$Q(\beta) = \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

$$Q(\beta_j; \boldsymbol{\beta}_{-j}) = \frac{1}{2}\|\boldsymbol{Y} - \mathbf{X}_{(-j)}\boldsymbol{\beta}_{-j} - \mathbf{X}_{(j)}\beta_j\|^2 + \lambda\|\boldsymbol{\beta}_{-j}\|_1 + \lambda|\beta_j|$$

Here $\boldsymbol{Y} - \mathbf{X}_{(-j)}\boldsymbol{\beta}_{-j}$ is the new vector of response, and β_j is the only unknown variable. This is a one-dimensional simple linear regression model with the residual on the jth variable $\mathbf{X}_{(j)}$.

If each variable has been pre-normalized, then it is just as simple as

$$\beta_j(\lambda) = S_\lambda\left[\mathbf{X}_{(j)}^T\{\boldsymbol{Y} - \mathbf{X}_{(-j)}\boldsymbol{\beta}_{-j}\}\right]$$

Warm start

- For each j, the update can be done in just one line of command. We can quickly go over all the variables (call this one round), and we can do many rounds until the estimates do not change between two rounds.
- Often we want to calculate the solution for a range of λ, not just one λ
- Warm start: provide a reasonable guess for the correct solution at each stage.
  - Start with setting λ to be a large value, so that $\widehat{\boldsymbol{\beta}}_{lasso}(\lambda) = \mathbf{0}$ (we may want to choose the smallest such λ).
  - Decrease the λ by a little, and use the solution from the previous λ as the initial values (called "warm start").

LARS vs Coordinate descent

- LARS provides exact piecewise linear solution to LASSO, while the solution from Coordinate descent is on a grid of many discrete points.
- For large problem, Coordinate descent is faster.
- Piecewise linear path algorithm can be found useful for other problems, as long as the loss function is piecewise linear or quadratic in β, and the penalty function is piecewise linear in β
- Coordinate descent turns out to be useful for a broader problems in statistics and machine learning.

Gradient descent

The gradient descent algorithm takes the following update iteratively to minimize convex and differentiable $f(\boldsymbol{\omega})$:

$$\boldsymbol{\omega}^{(k+1)} \leftarrow \boldsymbol{\omega}^{(k)} - \gamma f'(\boldsymbol{\omega}_{(k)})$$

where $0 < \gamma \leq 1$ is step size (or learning rate.)

However, in Lasso the objective is not differentiable. In particular,

$$f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta})$$

where g is the RSS which is convex and differentiable, but $h = \lambda \|\boldsymbol{\beta}\|_1$ is convex but not differentiable.

Proximal gradient descent for Lasso

Instead of the update

$$\boldsymbol{\beta}^{(k+1)} \leftarrow \boldsymbol{\beta}^{(k)} - \gamma f'(\boldsymbol{\omega}_{(k)})$$

use

$$\boldsymbol{\beta}^{(k+1)} \leftarrow S_{\gamma\lambda}(\boldsymbol{\beta}^{(k)} - \gamma g'(\boldsymbol{\omega}_{(k)}))$$

where $S_a : \mathbb{R}^p \mapsto \mathbb{R}^p$ is element-wise soft-thresholding defined in the obvious way.

Gradient descent update according to the differentiable part (g), but followed by soft-thresholding to correct for the regularization. See SLS pp.107.

Effective degrees of freedom

- Formal definition

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}(\hat{y}_i, y_i)$$

- Closed form for linear regression and ridge regression.
- No closed form for lasso (since nonlinear)
- **The number of nonzero coefficients is an unbiased (and consistent) estimate for the degrees of freedom of the lasso.**
- With this result, easy to use AIC or BIC type of selection procedure.

Lecture 4 R code

- Subset selection
- OLS
- Lasso
- Solution path