Linear Methods for Regression: Part II

Peng Zhang

February 9, 2019

Review

Why are we often not satisfied with the least squares estimates?

- Prediction accuracy:especially when $p > n$, to control the variance.
- Interpretation: with large number of predictors, we often like to determine a smaller subset that exhibit the strongest effects.

Three classes of methods

1. Subset Selection: retain only a subset of the variables and eliminate the rest from the model
   - Best Subset Selection
   - Forward- and Backward- Stepwise Selection
   - Forward- Stagewise regression
2. Shrinkage: fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero
   - Ridge Regression
   - The Lasso
   - Least Angle Regression (LAR)
3. Dimension Reduction: project the p predictors into a M-dimensional subspace, where $M < p$
   - Principal Components Regression
   - Partial Least Squares

Best Subset Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, ..., p$:
   - (a) Fit all C_p^k models that contain exactly k predictors
   - (b) Pick the best among these C_p^k models, and call it $\mathcal{M}_k$. Here "best" is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ using cross-validated prediction error, Cp (AIC), BIC, or adjusted R^2.

Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, ..., p-1$:
   - (a) Consider $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor
   - (b) Pick the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here "best" is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ using cross-validated prediction error, Cp (AIC), BIC, or adjusted R^2.

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.
2. For $k = p, p-1, ..., 1$:
   - (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors
   - (b) Pick the best among these k models, and call it $\mathcal{M}_{k-1}$. Here "best" is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_p$ using cross-validated prediction error, Cp (AIC), BIC, or adjusted R^2.

Forward-Stagewise Regression

- 1, Start with the residual $r = y - \bar{y}$, $\beta_1, \beta_2, \cdots, \beta_p = 0$
- 2, Find the predictor x_j most correlated with r. if none of the variables have correlation with the residuals, then break;
- 3, Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j =< r, x_j >$
- 4, set $r \rightarrow r - \delta_j * x_j$ and repeat steps 2 and 3 many times.

Ridge Regression

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{1}$$

equivalently with

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{2}$$

$$\text{subject to: } \sum_{j=1}^{p} \beta_j^2 \leq t$$

The ridge criterion in matrix form

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta \tag{3}$$

The ridge regression solutions are

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

The Lasso

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{5}$$

equivalently with

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{6}$$

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

Least Angle Regression

- 1. Standardize the predictors to have mean zero and unit norm. Start with the residual $r = y - \bar{y}$, $\beta_1, \beta_2, ..., \beta_p = 0$.
- 2. Find the predictor x_j most correlated with r.
- 3. Move β_j from 0 towards its least-squares coefficient $< x_j, r >$, until some other competitor x_k has as much correlation with the current residual as does x_j .
- 4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on (x_j, x_k), until some other competitor x_l has as much correlation with the current residual.
- 5. Continue in this way until all p predictors have been entered. After $\min(N - 1, p)$ steps, we arrive at the full least-squares solution.

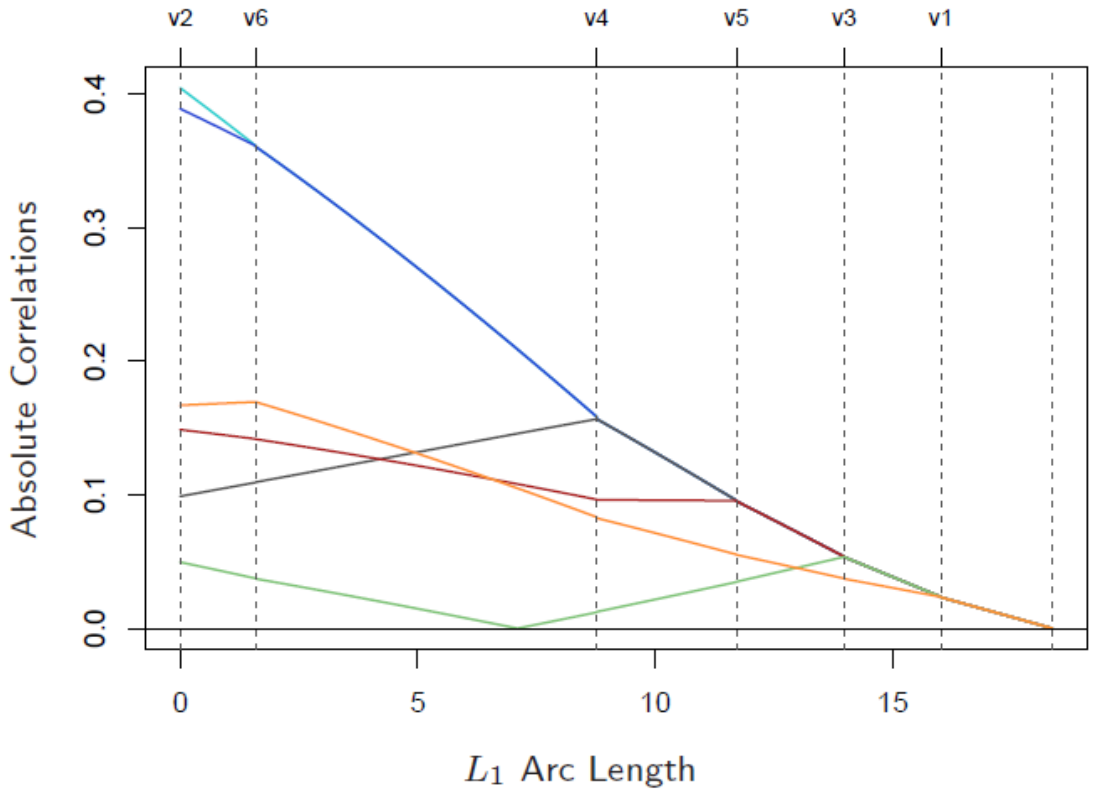


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

Principal Components Regression

1. Standardize each x_j to have mean zero and variance one.
2. Compute the matrix $\mathbf{X}^T\mathbf{X}$,and find the eigendecomposition of $\mathbf{X}^T\mathbf{X}$ as $\mathbf{X}^T\mathbf{X} = VD^2V^T$. The columns of V are denoted v_m,which is principal component direction and the diagonal elements of D are denoted d_m. $d1 \geqslant d2 \geqslant \cdots \geqslant d_p \geqslant 0$
3. For $m = 1, 2, \cdots, p$
   - $z_m = Xv_m$, mth principal component of X.
   - $\hat{\theta}_m =< z_m, y > / < z_m, z_m >=< z_m, y > /d_m^2$.
4. Given a value $M \leq p$, the estimate of y is $\hat{y}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m$.
5. $\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$

The mth principal component direction v_m solves:

$$\max_{\alpha} \text{Var}(\mathbf{X}\alpha) \tag{7}$$
$$\text{subject to:} ||\alpha|| = 1, \alpha^T \mathbf{S} v_t = 0, t = 1, \cdots, m-1.$$

where $\mathbf{S} = \mathbf{X}^T\mathbf{X}$ is the sample covariance matrix of the x_j.

$$\text{Var}(\mathbf{X}\alpha) = \alpha^T \mathbf{X}^T \mathbf{X}\alpha = \alpha^T \mathbf{S}\alpha \tag{8}$$

v_1 maximizing $\alpha^T\mathbf{S}\alpha$, is the normalized eigenvector of $\mathbf{S}$ with largest eigenvalue. v_2 is the normalized eigenvector of $\mathbf{S}$ with second largest eigenvalue.

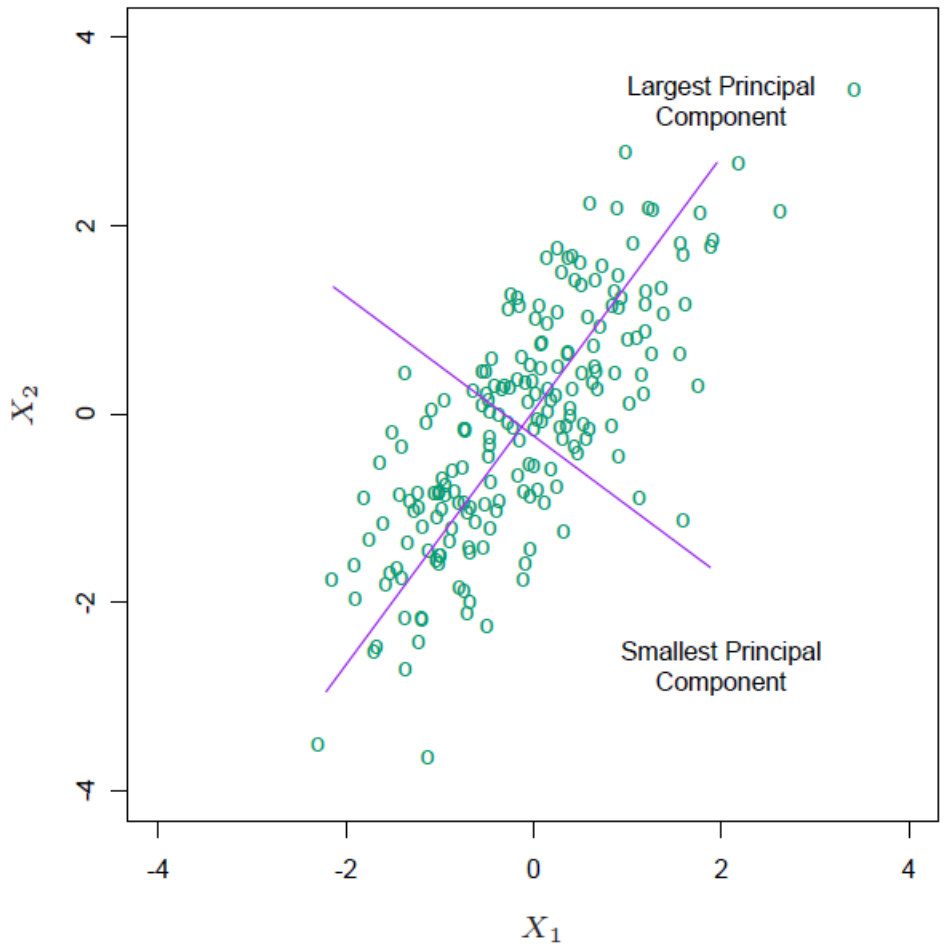


FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

Partial Least Squares

1. Standardize each x_j to have mean zero and variance one. Set $\hat{y}(0) = \bar{y}\mathbf{1}$, and $x_j^{(0)} = x_j, j = 1, \cdots, p$.
2. For $m = 1, 2, \cdots, p$
   - $z_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} x_j^{(m-1)}$, where $\hat{\varphi}_{mj} = < x_j^{(m-1)}, y >$.
   - $\hat{\theta}_m = < z_m, y > / < z_m, z_m >$.
   - $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$.
   - Orthogonalize each $x_j^{(m1)}$ with respect to z_m: $x_j^{(m)} = x_j^{(m-1)} - [< z_m, x_j^{(m-1)} > / < z_m, z_m >] z_m, j = 1, 2, \cdots, p$.
3. Output the sequence of fitted vectors $\{\hat{y}^{(m)}\}_1^p$. Since the $\{z_t\}_1^m$ are linear in the original x_j , so is $\hat{y}^{(m)} = X\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

The mth PLS component direction ϕ_m solves the squared sample covariance:

$$\max_{\alpha} \text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha) \tag{9}$$
$$\text{subject to:} ||\alpha|| = 1, \alpha^T \mathbf{S}\phi_t = 0, t = 1, \cdots, m-1.$$

the objective function

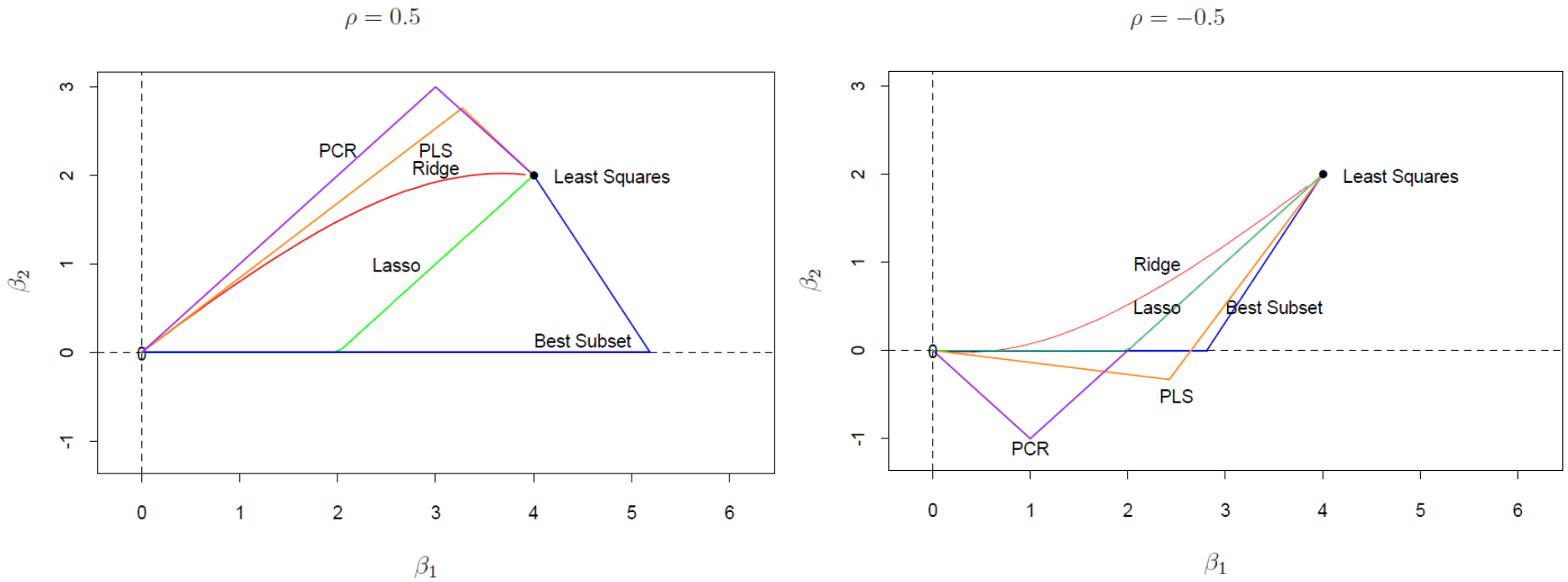
$$\text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha) \propto (\mathbf{y}^T\mathbf{X}\alpha)^2 \tag{10}$$

Suppose $\mathbf{W} = \mathbf{X}^T\mathbf{y}$, then $(\mathbf{y}^T\mathbf{X}\alpha)^2 = (\mathbf{W}^T\alpha)^2$, This gives what we may call the first 'canonical covariance' variable with $\phi_1 = \frac{\mathbf{W}}{||\mathbf{W}||}$, then the second canonical covariance variable ϕ_2 has to maximize the expression (9), $\phi_2 \propto w - \frac{\mathbf{W}^T\mathbf{S}\mathbf{W}}{\mathbf{W}^T\mathbf{S}^2\mathbf{W}}\mathbf{S}\mathbf{W}$

- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $z_1, z_2, \cdots, z_M$ that are linear combinations of the original features, and then fits a linear model via OLS using these M new features.
- But unlike PCR, PLS identifies these new features in a supervised way – that is, it makes use of the response y in order to identify new features that not only approximate the old features well, but also that are related to the response.
- Roughly speaking, the PLS approach attempts to find directions that have high variance and have high correlation with the response, in contrast to PCR which keys only on high variance.

Comparison of the Selection and Shrinkage Methods

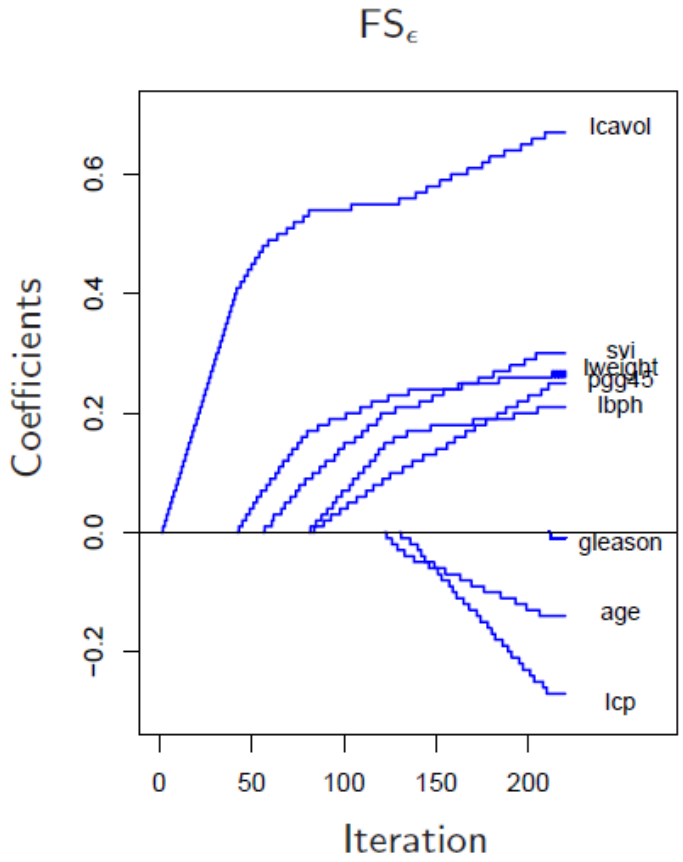
Two correlated inputs X_1,X_2 with correlation ρ. Assume the true regression coefficients are $\beta_1 = 4$,$\beta_2 = 2$.

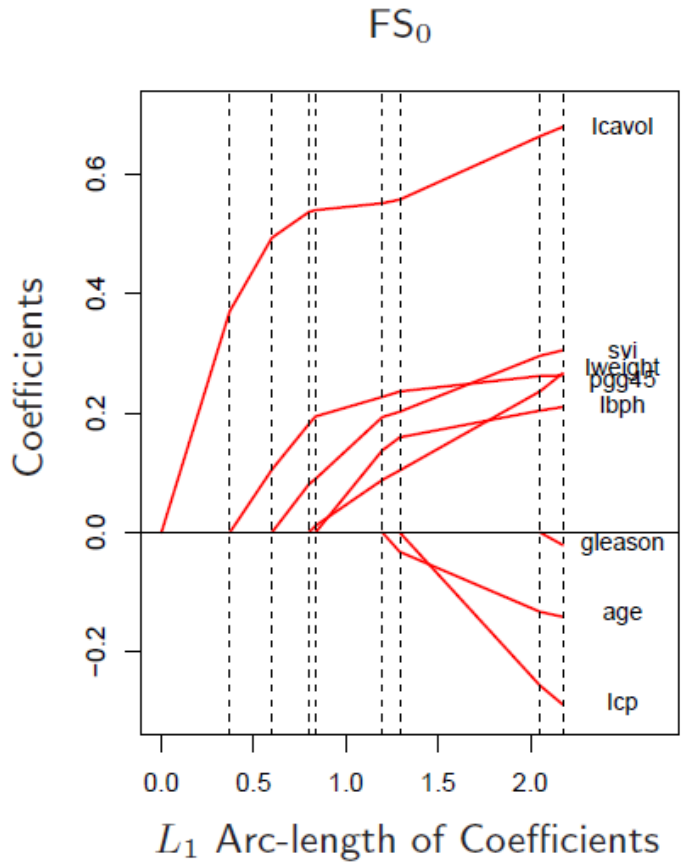


- Ridge regression shrinks all directions, but shrinks low-variance directions more.
- Principal component regression leaves M high-variance directions alone, and discards the rest.
- Partial least squares also tends to shrink the low-variance directions, but can actually inflate some of the higher variance directions. (unstable)
- PLS, PCR and ridge regression tend to behave similarly.
- Ridge regression may be preferred because it shrinks smoothly, rather than in discrete steps.
- Lasso falls somewhere between ridge regression and best subset regression, and enjoys some of the properties of each.

Incremental Forward Stagewise Regression–FS_ϵ

- Start with the residual r equal to y and $\beta_1, \beta_2, \cdots, \beta_p = 0$. All the predictors are standardized to have mean zero and unit norm.
- Find the predictor x_j most correlated with r.
- Update where $\delta_j = \epsilon \cdot \text{sign}[< x_j, r >]$ and $\epsilon > 0$ is a small step size, and set $r \leftarrow r - \delta_j x_j$.
- Repeat steps 2 and 3 many times, until the residuals are uncorrelated with all the predictors.


FS$_\epsilon$
Coefficients
0.6
0.4
0.2
0.0
−0.2
lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp
0
50
100
150
200
Iteration


FS$_0$
Coefficients
0.6
0.4
0.2
0.0
−0.2
lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp
0.0
0.5
1.0
1.5
2.0
L_1 Arc-length of Coefficients

- It generates a coefficient profile by repeatedly updating (by a small amount ϵ) the coefficient of the variable most correlated with the current residuals.
- Letting $\epsilon \to 0$ gives the right panel, which in this case is identical to the lasso path. This limiting procedure infinitesimal forward stagewise regression (FS_0), which plays an important role in non-linear, adaptive methods like boosting.

LAR modification

Algorithm 3.2b *Least Angle Regression: FS_0 Modification.*

4. Find the new direction by solving the constrained least squares problem

$$\min_b ||\mathbf{r} - \mathbf{X}_{\mathcal{A}} b||_2^2 \text{ subject to } b_j s_j \geq 0,\ j \in \mathcal{A},$$

where s_j is the sign of $\langle \mathbf{x}_j, \mathbf{r} \rangle$.

- The modification amouns to a non-negative least squares fit, keeping the signs of the coefficients the same as those of the correlations.
- Like lasso, the entire FS_0 path can be computed very efficiently via the LAR algorithm.
- if the LAR profiles are monotone non-increasing or non-decreasing, then LAR,lasso, FS_0 give identical profiles.

Piecewise Linear Path Algorithms

$$\hat{\beta}(\lambda) = \arg\min_{\beta}[R(\beta) + \lambda J(\beta)] \tag{11}$$

$$R(\beta) = \sum_{i=1}^{N} L(y_i, \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j) \tag{12}$$

where both loss function L and the penalty function J are convex. Sufficient conditions for the solution path $\hat{\beta}$ to be piecewise linear

- R is quadratic or piecewise-quadratic as a function of β.
- J is piecewise linear in β.

Loss function examples: squared loss, absolute-error loss. Penalty function example, L_1, L_∞ on β.

The Dantzig Selector

$$\min_{\beta} ||\beta||_1 \text{ subject to } ||\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)||_\infty \leq s \tag{13}$$

Equivalently as

$$\min_{\beta} ||\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)||_\infty \text{ subject to } ||\beta||_1 \leq t \tag{14}$$

- if $N < p$, as t gets large, DS and lasso yield the least squares solution.
- the operating properties of the DS method are somewhat unsatisfactory.

Grouped Lasso

$$\min_{\beta \in R^p} \{||\mathbf{y} - \beta_0 \mathbf{1} - \sum_{t=1}^{L} \mathbf{X}_t \beta_t||_2^2 + \lambda \sum_{t=1}^{L} \sqrt{p_t} ||\beta_t||_2\} \tag{15}$$

- The predictors belong to pre-defined groups, it will shrink and select the members of a group together.
- p predictors are divided into L groups, with p_t the number in group t.
- $\mathbf{X}_t$ is the matrix of predictors corresponding to the tth group,with corresponding coefficient vector β_t.
- This procedure encourages sparsity at both the group and individual levels, that is for some value of λ, an entire group of predictors may drop out of the model.

Further Properties of the Lasso

1. Lasso shrinkage causes the estimates of the non-zero coefficients to be biased towards zero.
2. For reducing this bias
   - One way:Run the lasso to identify the set of non-zero coefficients, then fit an unrestricted linear model to the selected set of features.
   - Relaxed lasso: use the lasso to select the set of non-zero predictors, then apply the lasso again, but using only the selected predictors from the first step.

- Modify the lasso penalty function so that larger coefficients are shrunken less severely, the penalty replaces $\lambda|\beta|$ by $J_a(\beta, \lambda)$, where

$$\frac{dJ_a(\beta,\lambda)}{d\beta} = \lambda \text{sign}(\beta)[I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)] \tag{16}$$

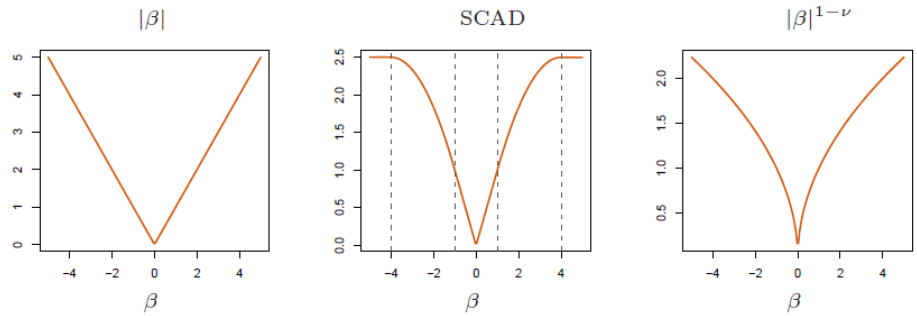


FIGURE 3.20. *The lasso and two alternative non-convex penalties designed to penalize large coefficients less. For SCAD we use $\lambda = 1$ and $a = 4$, and $\nu = \frac{1}{2}$ in the last panel.*

- The adaptive lasso uses a weighted penalty of the form $\sum_{j=1}^{p} w_j|\beta_j|$, here $w_j = 1/|\hat{\beta}_j|^\nu$, $\hat{\beta}_j$ is the ordinary least squares estimate.

Pathwise Coordinate Optimization

- Using simple coordinate descent to compute the lasso solution.
- Fix the penalty parameter λ and optimize successively over each parameter, holding the other parameters fixed at their current values.
- Denote $\tilde{\beta}_k(\lambda)$ the current estimate for β_k at penalty parameter λ.

$$R(\tilde{\beta}(\lambda), \beta_j) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k(\lambda) - x_{ij}\beta_j)^2 + \lambda\sum_{k\neq j}|\tilde{\beta}_k(\lambda)| + \lambda|\beta_j| \quad (17)$$

- It is a univariate lasso problem with response variable the partial residual $y_i - \tilde{y}_i^{(j)} = y_i - \sum_{k\neq j} x_{ik}\tilde{\beta}_k(\lambda)$.
- the explicit solution is

$$\tilde{\beta}_j(\lambda) \leftarrow S(\sum_{i=1}^{N} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda) \quad (18)$$

here $S(t, \lambda) = \text{sign}(t)(|t| - \lambda)_+$