BIOS 835: Shrinkage Methods

Alexander McLain

September 7, 2023

"There's two buttons I never like to hit: that's panic and snooze"

Introduction

- Today, we'll talk about shrinkage methods
- All shrinkage methods are biased (by design) even when the model is correct.
- The will, however, result in lower variance than the full LR model.
- Some shrinkage methods we'll discuss:
  - Ridge Regression
  - Lasso
  - Elastic Net
  - Least Angle Regression (LAR)

Instability of LS Estimators

- Recall that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ can be unstable
- Further, when $rank(\boldsymbol{X}) < p$ $\boldsymbol{X}'\boldsymbol{X}$ is singular
  - $\boldsymbol{X}$ is ill-conditioned
  - columns of $\boldsymbol{X}$ are collinear
  - $p > n$
- We can measure ill-conditioning by $\kappa = d_1/d_p$ where d_1 and d_p are the largest and smallest eigenvalues, respectively, of $\boldsymbol{X}$.
- We can measure collinearity for a predictor X_k via $VIF_k = (1 - R_k^2)^{-1}$ where R_k^2 is the R^2 for all other covariates regressed on X_k

Biased Regression Methods

- The instability of the LS estimates is due to $\boldsymbol{X}'\boldsymbol{X}$ being singular or nearly singular.
- **Solution** allow $\hat{\boldsymbol{\beta}}$ to be biased.
- We'll assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ have been centered (no β_0).
- First, we discuss Principal Component Analysis (PCA) to get to PC Regression (PCR), which can be viewed as a shrinkage method.

Introduction

- Principal components explain the variance-covariance structure of a set of variables through some linear combinations.
- $\boldsymbol{X}$: variables which we want to explain the variance-covariance matrix $\Sigma = Cov(\boldsymbol{X})$
- $\boldsymbol{Z}$: a linear combination of the X's (this is not the response)

$$\begin{aligned}
\boldsymbol{Z}_1 &= \boldsymbol{X}\boldsymbol{v}_1 = v_{11}\boldsymbol{X}_1 + v_{12}\boldsymbol{X}_2 + \ldots + v_{1p}\boldsymbol{X}_p \\
\boldsymbol{Z}_2 &= \boldsymbol{X}\boldsymbol{v}_1 = v_{21}\boldsymbol{X}_1 + v_{22}\boldsymbol{X}_2 + \ldots + v_{2p}\boldsymbol{X}_p \\
\vdots \quad & \quad \vdots \quad \vdots \\
\boldsymbol{Z}_p &= \boldsymbol{X}\boldsymbol{v}_1 = v_{p1}\boldsymbol{X}_1 + v_{p2}\boldsymbol{X}_2 + \ldots + v_{pp}\boldsymbol{X}_p
\end{aligned}$$

PC goal

- Using standard results

$$Var(\boldsymbol{Z}_i) = \boldsymbol{v}_i'\Sigma\boldsymbol{v}_i \quad \text{and} \quad Cov(\boldsymbol{Z}_i, \boldsymbol{Z}_k) = \boldsymbol{v}_i'\Sigma\boldsymbol{v}_k \text{ for } i, k = 1, 2, \ldots, p$$

Main idea of Principal Components

- Set $\boldsymbol{Z}_1 = \boldsymbol{X}\boldsymbol{v}_1$ such that $\boldsymbol{v}_1'\Sigma\boldsymbol{v}_1$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$.
- Set $\boldsymbol{Z}_2 = \boldsymbol{X}\boldsymbol{v}_2$ such that $\boldsymbol{v}_2'\Sigma\boldsymbol{v}_2$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = 0$.

⋮

- Set $\boldsymbol{Z}_i = \boldsymbol{X}\boldsymbol{v}_i$ such that $\boldsymbol{v}_i'\Sigma\boldsymbol{v}_i$ is maximized over all $\boldsymbol{v}$ such that $\boldsymbol{v}'\boldsymbol{v} = 1$ and $Cov(\boldsymbol{Z}_k, \boldsymbol{Z}_i) = 0$ for all $k = 1, 2, \ldots, i = 1$.

⋮

PC solution

Theorem

Let $\Sigma = Cov(\boldsymbol{X})$ *where* $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p\}$. *Let* Σ *have eigenvalue-eigenvector pairs* $(d_1, \boldsymbol{v}_1)$, $(d_2, \boldsymbol{v}_2)$, ..., $(d_p, \boldsymbol{v}_p)$ *where* $d_1 \geq d_2 \geq \ldots \geq d_p$. *Then the* i*th principal component is*

$$\boldsymbol{Z}_i = \boldsymbol{X}\boldsymbol{v}_i = v_{i1}\boldsymbol{X}_1 + v_{i2}\boldsymbol{X}_2 + \ldots + v_{ip}\boldsymbol{X}_p$$

for $i = 1, 2, \ldots, p$ *where*

$$Var(\boldsymbol{Z}_i) = \boldsymbol{v}_i'\Sigma\boldsymbol{v}_i = d_i \quad \text{and} \quad Cov(\boldsymbol{Z}_i, \boldsymbol{Z}_k) = \boldsymbol{v}_i'\Sigma\boldsymbol{v}_k = 0 \text{ for } i, k = 1, 2, \ldots, p$$

Properties

- Let σ_{ii} be the (i, i)th component of Σ.

$$\sum_{i=1}^{p} Var(\boldsymbol{Z}_i) = \sum_{i=1}^{p} Var(\boldsymbol{X}_i) = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} d_i = tr(\Sigma) = tr(\Lambda)$$

  where Λ is a diagonal matrix of with ith element d_i, the eigenvalues.
- Recall $d_1 \geq d_2 \geq \ldots \geq d_p$ where $Var(\boldsymbol{Z}_i) = d_i$.
- The correlation between Z_i and X_k is

$$\rho_{Z_i, X_k} = \frac{v_{ik}\sqrt{d_i}}{\sqrt{\sigma_{kk}}}$$

  (which can also be measured by v_{ik}).

Properties

- The total proportion variability that can be attributed to Z_k is

$$\frac{d_k}{\sum_{i=1}^{p} d_i}$$

- Suppose

$$\frac{d_1 + d_2 + d_3}{\sum_{i=1}^{p} d_i} = 0.97.$$

what would this imply about $Z_4, \ldots, Z_p$?

How many PC's to use

- If $\frac{\sum_{i=1}^{r} d_i}{\sum_{i=1}^{P} d_i}$ is "close to one," then only the first r PC's are needed.
- Scree plot: a plot of d_i used for selecting r
- *Kaiser's Rule* for the PC of a correlation matrix cutoff at 1.
- PC's are not invariant to scaling
- PC on standardized data = PC on a correlation matrix
- PC's are unique up to a rotation (varimax criteria).

Connection to SVD

Definition

The *singular value decomposition* (SVD) of an $N \times p$ centered input matrix $\boldsymbol{X}$ has the form

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}',$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $N \times p$ and $p \times p$ **orthogonal matrices**, and $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \ldots \geq d_p$. (starting to sound familiar?)

- the columns of $\boldsymbol{U}$ are orthogonal unit vectors of length N called the *left singular vectors* of $\boldsymbol{X}$
- the columns of $\boldsymbol{V}$ are orthogonal unit vectors of length p called the *right singular vectors* of $\boldsymbol{X}$
- The d_i's are called the *singular values* of $\boldsymbol{X}$

Connection to SVD

- The sample covariance matrix of $\boldsymbol{X}$ is $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}/N$, so using the definition of SVD we have
$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}'$$
which is called the *eigen decomposition* $\boldsymbol{X}'\boldsymbol{X}$.
- The columns of $\boldsymbol{V}$, $\boldsymbol{v}_i$, are equal to the eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$
- The singular values of $\boldsymbol{X}$, d_i, are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and the diagonal values of $\boldsymbol{D}^2$.
- Note that $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D}$ where $\boldsymbol{Z}$ is an $N \times p$ matrix of *PC scores*.
- Efficient algorithms exist to calculate the SVD of $\boldsymbol{X}$, so this is how PC's are estimated.

Principal components regression

- Principal components regression (PCR) forms the linear combination $\boldsymbol{z}_m = \boldsymbol{X}\boldsymbol{v}_m$ and then regresses $\boldsymbol{y}$ on $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_M$ for some $M \leq p$.
- Since the $\boldsymbol{z}_j$'s are orthogonal, this is just the sum of univariate regressions

$$\hat{\boldsymbol{y}}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \boldsymbol{z}_m$$

  where $\hat{\theta}_m = \langle \boldsymbol{z}_m, \boldsymbol{y} \rangle / \langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle$.
- The $\boldsymbol{z}_m$'s are linear combinations of the $\boldsymbol{x}$'s
- The coefficients for $\boldsymbol{x}_j$ from PCR can be expressed as

$$\beta^{\text{pcr}}_{j(M)} = \sum_{m=1}^{M} \hat{\theta}_m v_{jm}$$

Ridge Regression

- Potential instability in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ could be solved by adding a constant λ to $\boldsymbol{X}'\boldsymbol{X}$.
- Adding a constant to the diagonal of $\boldsymbol{X}'\boldsymbol{X}$ will bring it closer to a matrix with 0's on the off diagonal.
- This results in the ridge regression estimate of $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix.

Ridge Regression Properties

- Another way to view the ridge regression model is that

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \text{argmin}_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||^2 \right\},$$

  where $||\boldsymbol{\beta}||^2 = \sum_{j=1}^{p} \beta_j^2$ (note: the λ here is the same as on the previous slide)

- (Yet) Another way to view the ridge regression model is that $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ minimizes

$$ESS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) \quad \text{such that} \quad ||\boldsymbol{\beta}||^2 \leq t,$$

  which makes explicit the size constraint on the parameters.

- The relationship between t and λ is made through Lagrange multipliers.

Ridge Regression and Bayes

The ridge regression estimate naturally comes up in Bayesian regression analysis.

- If we were to estimate the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ in a Bayesian framework with
  - $\boldsymbol{e} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, and
  - $\boldsymbol{\beta} \sim MVN(\boldsymbol{0}, \sigma^2_{\beta} \boldsymbol{I})$ (prior distribution).
- Then the mean (or mode) of the posterior distribution of $\boldsymbol{\beta}|\boldsymbol{Y}$ is $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ with $\lambda = \sigma^2/\sigma^2_{\beta}$.

Ridge Regression and PCR

If $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{X}$ and $\boldsymbol{\alpha} = \boldsymbol{V}'\boldsymbol{\beta}$ (thus $\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{\alpha}$)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{1}$$

- If we were to apply a ridge regression PCR model, the jth coefficient is

$$\hat{\alpha}_j^{\text{ridge}} = \frac{d_j^2}{d_j^2 + \lambda}\hat{\alpha}_j$$

where $\hat{\boldsymbol{\alpha}}$ is the LS estimate of (1), and d_j is the jth eigenvalue.

Ridge Regression and PCR

If $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{X}$ and $\boldsymbol{\alpha} = \boldsymbol{V}'\boldsymbol{\beta}$ (thus $\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{\alpha}$)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{1}$$

- If we were to apply a ridge regression PCR model, the jth coefficient is

$$\hat{\alpha}_j^{\text{ridge}} = \frac{d_j^2}{d_j^2 + \lambda}\hat{\alpha}_j$$

  where $\hat{\boldsymbol{\alpha}}$ is the LS estimate of (1), and d_j is the jth eigenvalue.
- Then $\hat{\boldsymbol{\beta}}^{\text{ridge}} = \boldsymbol{V}\hat{\boldsymbol{\alpha}}^{\text{ridge}}$.
- Thus, the shrinkage of $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ compared to $\hat{\boldsymbol{\beta}}$ is a function of the singular values.

Ridge Regression Degrees of Freedom

- Degrees of freedom are a good measure of the flexibility of the model.
- The *effective degrees of freedom* for the ridge model are

$$\begin{aligned} df(\lambda) &= \text{tr}\{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\}, \\ &= \text{tr}(\boldsymbol{H}_\lambda) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \end{aligned}$$

- Note that when $\lambda = 0$ we have $df(\lambda) = p$, and $df(\lambda) \to 0$ as $\lambda \to \infty$.
- All ridge coefficients will be non-zero.

Ridge Regression Properties

- The bias and variance of Ridge coefficients have nice closed forms

$$Var_\lambda(\hat{\beta}^{\text{ridge}}) = \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2}$$

$$Bias_\lambda^2(\hat{\beta}^{\text{ridge}}) = \lambda^2 \sum_{j=1}^{p} \frac{(\hat{\alpha}_j^{\text{ridge}})^2}{(d_j^2+\lambda)^2}$$

thus,

$$MSE_\lambda(\hat{\beta}^{\text{ridge}}) = \sigma^2 \sum_{j=1}^{p} \frac{\sigma^2 d_j^2 + \lambda(\hat{\alpha}_j^{\text{ridge}})^2}{(d_j^2+\lambda)^2}$$

Lasso regression model

- The Lasso regression model, results in solution $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ minimizes

$$ESS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) \quad \text{such that} \quad \sum_{i=1}^{p} |\beta_j| \leq t,$$

  which makes explicit the size constraint on the parameters.

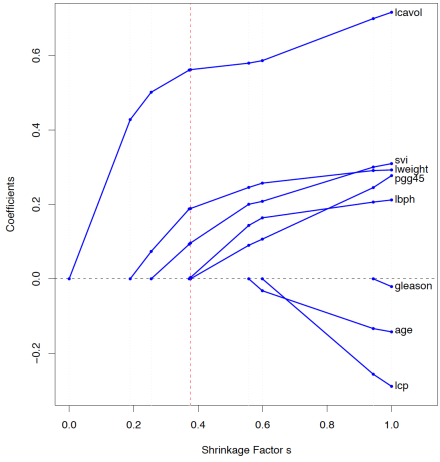
- The solution to this is such that

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \text{argmin}_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_j| \right\},$$

- The relationship between these forms is made through Lagrange multipliers.

Lasso properties

- A major difference between the Lasso and Ridge models is that making t sufficiently small (or λ large) will result in some β coefficients being exactly zero.
- Let $t_0 = \sum_{i=1}^{p} |\hat{\beta}_j^{ls}|$ then if $t = t_0/2$ the Lasso will result in coefficients that are the least squares coefficients shrunk by about 50% on average.
- There is no closed-form solution for Lasso coefficients (they are nonlinear in $\boldsymbol{y}$).



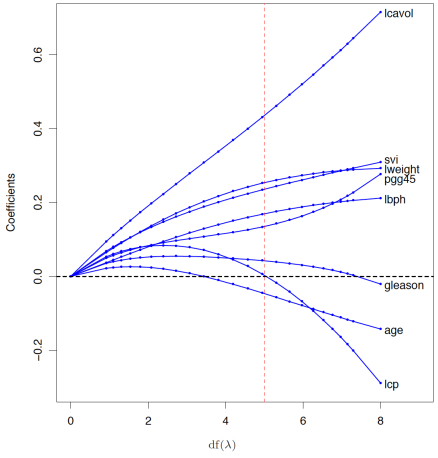


Figure: From ESL, where $s = t/\sum_1^p |\hat{\beta}_j^{ls}|$

Ridge vs Lasso vs Elastic Net

- The Lasso regression model,

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \text{argmin}_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} |\beta_j| \right\},$$

- The ridge regression model,

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \text{argmin}_{\beta} \left\{ (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} \beta_j^2 \right\},$$

- The elastic net regression model (Zou and Hastie, 2005),

$$\hat{\boldsymbol{\beta}}^{\text{en}} = \text{argmin}_{\beta} \left[(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda \sum_{i=1}^{p} \left\{ \alpha\beta_j^2 + (1-\alpha)|\beta_j| \right\} \right],$$

Ridge vs Lasso vs Best Subset

- Suppose $\boldsymbol{X}$ is a standardized orthonormal input matrix
- The estimators for Ridge, Lasso, and best subset, denoted by $\hat{\beta}^{\text{ridge}}$, $\hat{\beta}^{\text{lasso}}$ and $\hat{\beta}^{\text{bss}}$ respectively, are

$$\begin{aligned}
\hat{\beta}_j^{\text{bss}} &= \hat{\beta}_j^{ls} I\left(|\hat{\beta}_j| \geq |\hat{\beta}_{(M)}|\right) \text{ of size M} \\
\hat{\beta}_j^{\text{ridge}} &= \frac{\hat{\beta}_j^{ls}}{1+\lambda} \\
\hat{\beta}_j^{\text{lasso}} &= \text{sign}\left(\hat{\beta}_j^{ls}\right)\left(|\hat{\beta}_j^{ls}| - \lambda\right)_+
\end{aligned}$$

where $\hat{\beta}_{(M)}$ is the Mth largest regression coefficient.

Least Angle Regression

- Least Angle Regression (LAR) was developed in Efron *et al.* (2004).[1]
- LAR starts by entering the variable most closely correlated with the response. LAR moves the coefficient (continuously) towards the LS estimate until another variable has an equal correlation with the residuals of the model.
- Then, both variables are moved toward their LS estimates (while keeping their correlation tied and decreasing) until another variable has equal correlation with the residuals. Etc.
- LAR is similar to the forward stagewise regression in that only "some" of a variable's influence on the outcome is entered into the model.

[1] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R., 2004. Least angle regression. *The Annals of statistics*, 32(2), pp.407-499.

LAR Algorithm

LAR algorithm

1. Standardize predictors. Start with residual $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ and $\boldsymbol{\beta} = 0$.
2. Find the predictor $\boldsymbol{x}_j$ most correlated with $\boldsymbol{r}$.
3. Move β_j from $0 \rightarrow \langle \boldsymbol{x}_j, \boldsymbol{r} \rangle$ until another variable $\boldsymbol{x}_k$ has equal absolute correlation with the residual.
4. Move (β_j, β_k) towards their joint least squares estimates of the current $\boldsymbol{r}$ on $(\boldsymbol{x}_j, \boldsymbol{x}_k)$, until another variable $\boldsymbol{x}_l$ has equal absolute correlation with the residual.
5. Continue until all p predictors have been entered.

*LAR can be used when $p > n$

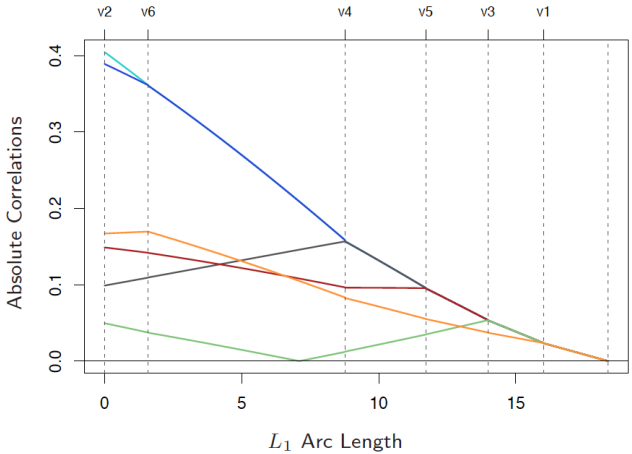


Figure: From ESL (online version)

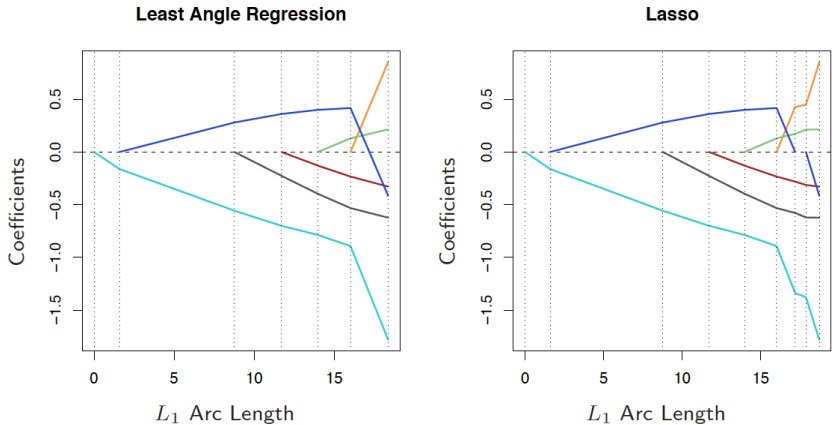


Figure: From ESL (online version), note the similarities.

Modified LAR Algorithm

- A Modified LAR algorithm is done by adding
  4a. If a non-zero coefficient hits zero, drop its variable from the active set and recompute the current joint least squares direction
- The Modified LAR algorithm is an extremely efficient way of estimating the LASSO model (recall the need for quadratic programming with LASSO).
- The current joint least squares direction and the point where another variable will have equal absolute correlation can be calculated (the variables don't actually have to be moved continuously).

Partial Least Squares

- PCR was performed by making a linear combination of the $\boldsymbol{X}$'s, then performing LR on the derived linear combinations.
- Partial Least Squares (PLS), like PCR, also constructs a linear combination of the $\boldsymbol{X}$'s and performs LR on the derived linear combinations.
- Note that when performing PCA (for PCR), the outcome ($\boldsymbol{y}$) is never used.
- PLS does use the outcome to form the linear combinations.

Partial Least Squares Algorithm

PLS Algorithm

1. Standardize each $\boldsymbol{x}_j$ and set $\hat{\boldsymbol{y}}^{(0)} = \bar{\boldsymbol{y}}\mathbf{1}$, and $\hat{\boldsymbol{x}}_j^{(0)} = \boldsymbol{x}_j$, $j = 1, 2, \ldots, p$.
2. For $m = 1, 2, \ldots, M \leq p$
   - 2.1 $\boldsymbol{z}_m = \sum_{j=1}^{p} \hat{\psi}_{mj} \boldsymbol{x}_j^{(m-1)}$, where $\hat{\psi}_{mj} = \langle \boldsymbol{x}_j^{(m-1)}, \boldsymbol{y} \rangle$.
   - 2.2 $\hat{\theta}_m = \langle \boldsymbol{z}_m, \boldsymbol{y} \rangle / \langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle$.
   - 2.3 $\hat{\boldsymbol{y}}^{(m)} = \hat{\boldsymbol{y}}^{(m-1)} + \hat{\theta}_m \boldsymbol{z}_m$.
   - 2.4 Orthogonalize each $\boldsymbol{x}_j^{(m-1)}$ with respect to $\boldsymbol{z}_m$. That is, set $\boldsymbol{x}_j^{(m)} = \boldsymbol{x}_j^{(m-1)} - (\langle \boldsymbol{z}_m, \boldsymbol{x}_j^{(m-1)} \rangle / \langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle) \boldsymbol{z}_m$, for $j = 1, 2, \ldots, p$.
3. Output $\{\hat{\boldsymbol{y}}^{(m)}\}_1^M$. Since the $\{\boldsymbol{z}_l\}_1^m$ are linear in $\boldsymbol{x}_j$, so is $\hat{\boldsymbol{y}}^{(m)} = \boldsymbol{X}\hat{\beta}^{\text{pls}}(m)$.

Partial Least Squares Algorithm (part 2)

1. Standardize each $\boldsymbol{x}_j$ and set $\hat{\boldsymbol{y}}^{(0)} = \boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}$, and $\hat{\boldsymbol{x}}_j^{(0)} = \boldsymbol{x}_j$, $j = 1, 2, \ldots, p$.
2. For $m = 1, 2, \ldots, M \leq p$
   - 2.1 Compute $\hat{\psi}_{mj} = \langle \boldsymbol{x}_j^{(m-1)}, \boldsymbol{y}^{(m-1)} \rangle$ and set $\boldsymbol{z}_m = \sum_{j=1}^{p} \hat{\psi}_{mj} \boldsymbol{x}_j^{(m-1)}$.
   - 2.2 Compute $\hat{\theta}_m = \langle \boldsymbol{z}_m, \boldsymbol{y}^{(m-1)} \rangle / \langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle$ and set $\hat{\boldsymbol{y}}^{(m)} = \hat{\boldsymbol{y}}^{(m-1)} - \hat{\theta}_m \boldsymbol{z}_m$.
   - 2.3 Regress $\boldsymbol{x}_j^{(m-1)}$ with respect to $\boldsymbol{z}_m$ for all j. That is, set

$$\alpha_{mj} = \frac{\langle \boldsymbol{z}_m, \boldsymbol{x}_j^{(m-1)} \rangle}{\langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle}$$

   and then $\boldsymbol{x}_j^{(m)} = \boldsymbol{x}_j^{(m-1)} - \alpha_{mj} \boldsymbol{z}_m$, for $j = 1, 2, \ldots, p$.
3. Then the PLSR model with m components is

$$\hat{\boldsymbol{y}}(k) = \bar{\boldsymbol{y}}\mathbf{1} + \sum_{m=1}^{k} \hat{\theta}_m \boldsymbol{z}_m.$$

Grouped Lasso regression model

- In some problems, the predictors belong to pre-defined groups (e.g., genes belonging to some biological pathway or dummy-coded categorical variables).
- In this situation it may be desirable to shrink and select the members of a group together.
- Suppose that the p predictors are divided into L groups, with p_ℓ being the number of variables in group ℓ
- The Grouped Lasso regression model minimizes

$$(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} ||\boldsymbol{\beta}_\ell||_2,$$

where $\boldsymbol{\beta}_\ell$ are the regression coefficients for group ℓ, and $||\cdot||_2$ is the Euclidean norm.

Fused Lasso regression model

- Along with grouped methods there is the idea of fusion or merging.
- In this situation it may be desirable to shrink and select the members of a group together.
- Consider the categorical covariates; by assuming the coefficient for one group is zero, you fuse that group with the referent group.
- The idea of fusion is to fuse groups together and not have the fusion be subject to the reference group.
- The Fused Lasso regression proposed by Tibshirani et al (2005) minimizes

$$\hat{\boldsymbol{\beta}} = arg\ \min \left\{ (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}) \right\}$$

such that

$$\sum_{j=1}^{p} |\beta_j| \leq s_1 \quad \text{and} \quad \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq s_2.$$