Shrinkage Methods II

Statistical Data Mining I

Fall 2017

Rachael Hageman Blair

Shrinkage Methods

- We will tackle a number of shrinkage methods:
  - Ridge Regression
  - The LASSO
  - Least Angle Regression (LAR)
  - Principal Components Regression
  - Partial Least Squares

Shrinkage Methods

- We will tackle a number of shrinkage methods:
  - ~~Ridge Regression~~
  - The LASSO
  - Least Angle Regression (LAR)
  - Principal Components Regression
  - Partial Least Squares

J. R. Statist. Soc. **B** (1996)
58, *No.* 1, *pp.* 267–288

Regression Shrinkage and Selection via the Lasso

By ROBERT TIBSHIRANI†

University of Toronto, Canada

[Received January 1994. Revised January 1995]

The Annals of Statistics
2004, Vol. 32, No. 2, 407–499

LEAST ANGLE REGRESSION

BY BRADLEY EFRON,[1] TREVOR HASTIE,[2] IAIN JOHNSTONE[3]
AND ROBERT TIBSHIRANI[4]

Stanford University

Shrinkage Methods: The Lasso

- The lasso estimate is defined as:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to:

$$\sum_{j=1}^{p} |\beta_j| \le t.$$

L1 lasso penalty

- Equivalently:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- No closed form solution.
- Quadratic programming problem.
- The nature of the shrinkage is not obvious.

Shrinkage Methods: The Lasso

Setting the penalty:

- Setting t sufficiently small will cause some coefficients to be exactly zero. (enforces sparsity ~ like a continuous subset selection).

- If t is set to be larger than:
$$t_0 = \sum_{j=1}^{p} \left|\hat{\beta}_j\right|$$
then you are not going to shrink anything. The LS coefficients will be as is.

- If t is set to: $t_0 = \frac{1}{2}\sum_{j=1}^{p} \left|\hat{\beta}_j\right|$
then the LS coefficients will be shrunk by about 50% on average.

- In general, t should be chosen to minimize the EPE.

Shrinkage Methods: The Lasso

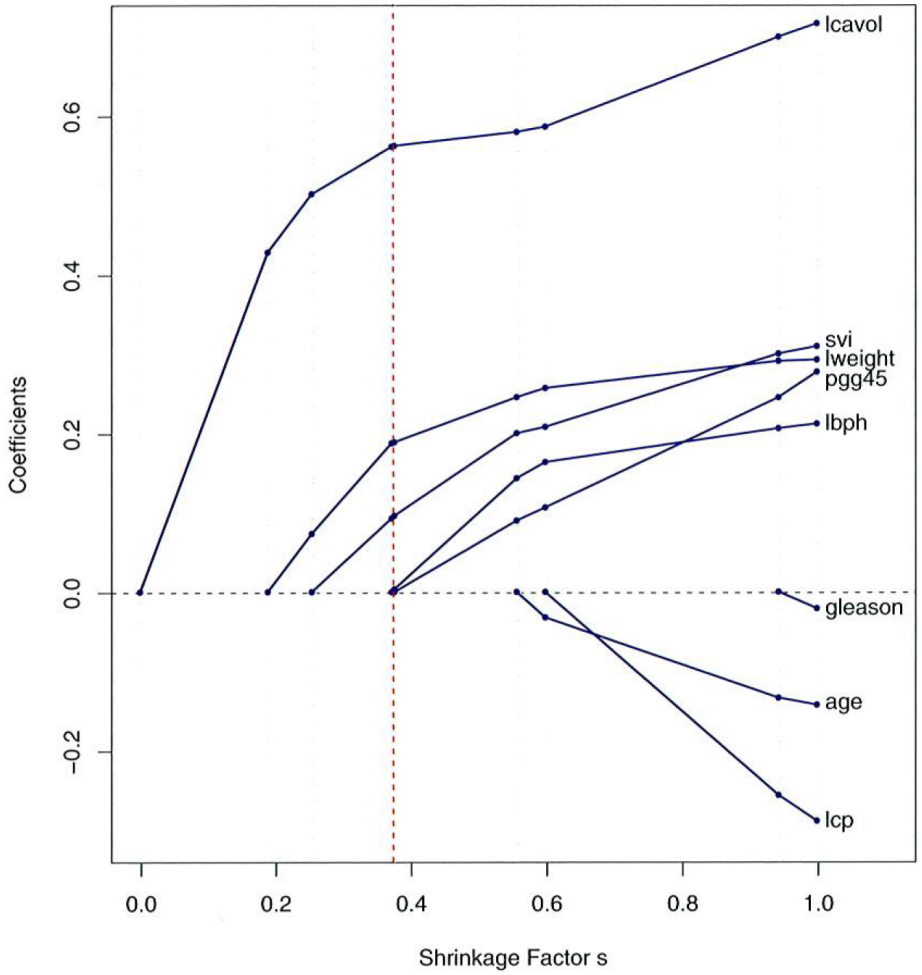


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Ridge Solution

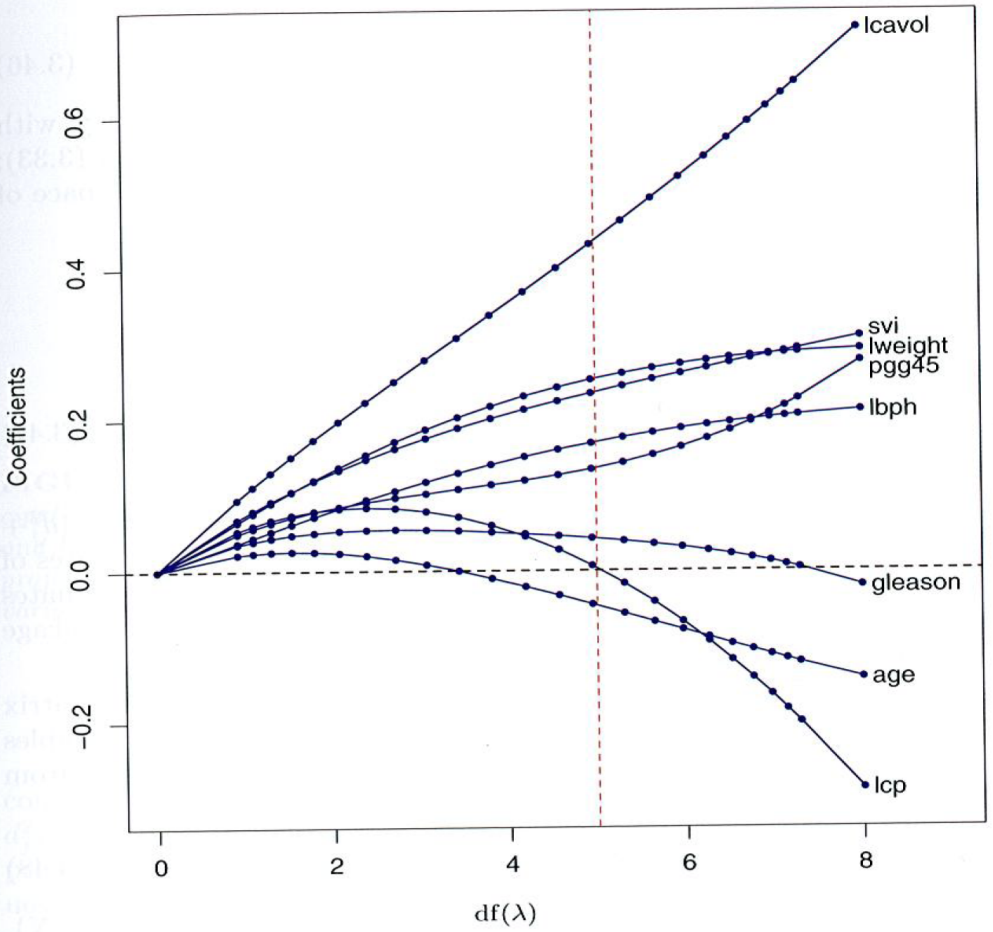


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

LASSO Solution

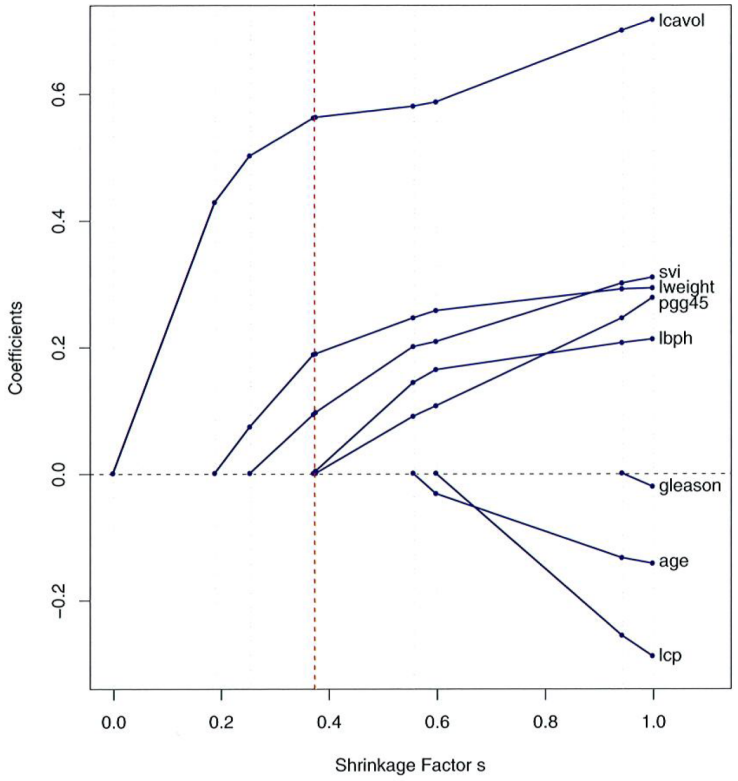
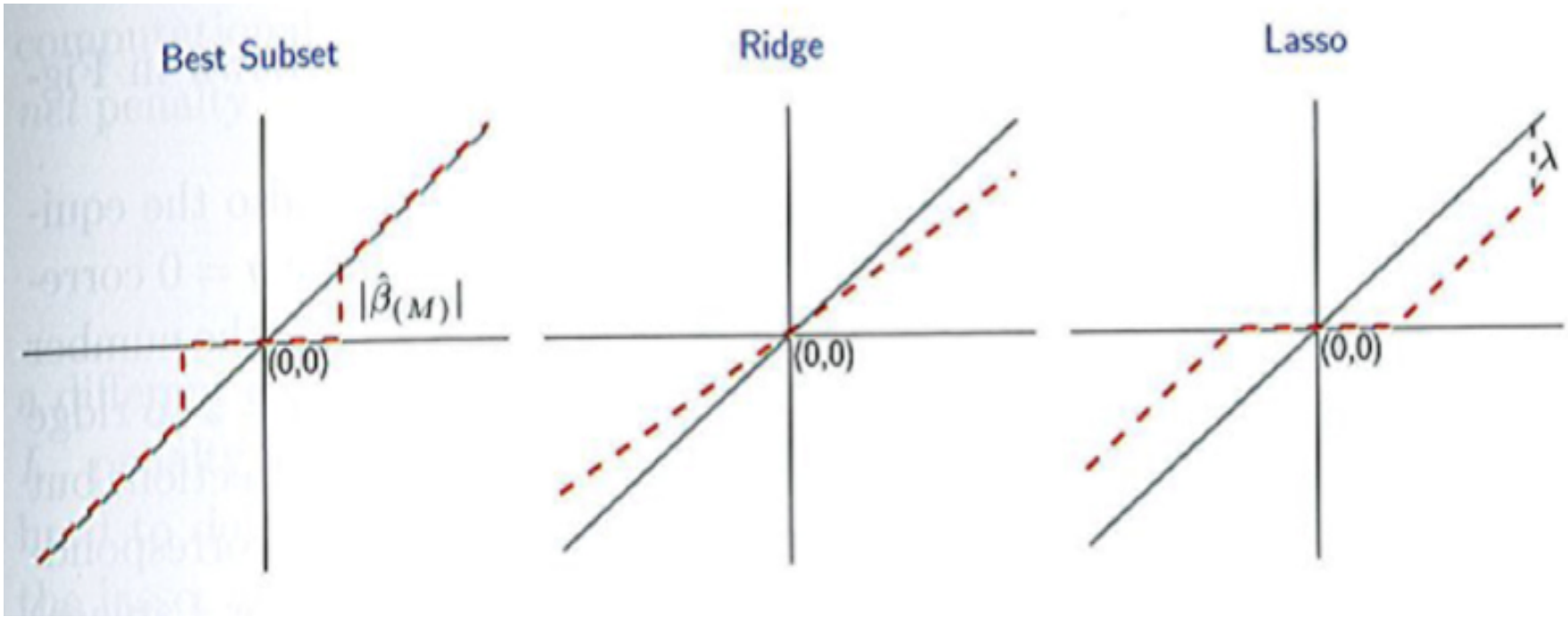


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Shrinkage Methods: The Ridge and Lasso



Shrinkage Methods: Bayesian Interpretation

Generalization can be viewed as **Bayesian Estimates**:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}.$$

Prior density

Bayes Estimates with different priors:

q = 0: Subset Selection

q = 1: LASSO

q = 2: Ridge Regression

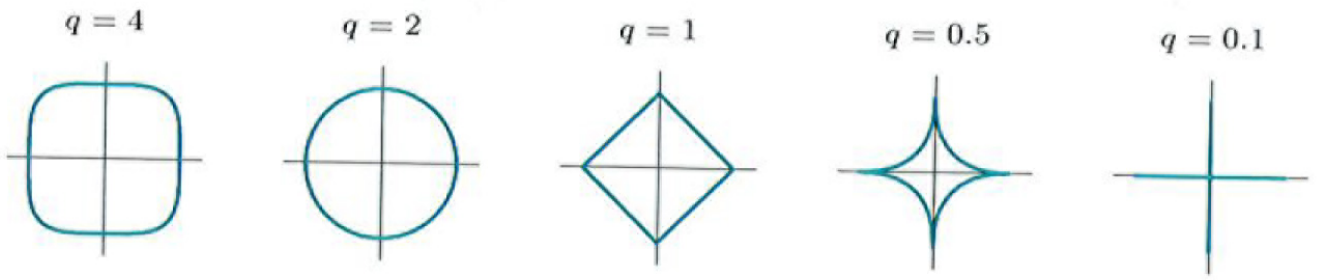


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Shrinkage Methods: Least Angle Regression

- Similar to forward stepwise regression.

- Instead of including variables at each step, the estimated parameters are moved toward the least squares fit continuously according to the correlation with residual.

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.

2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.

3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.

4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ 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.

Shrinkage Methods: Least Angle Regression

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.

2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.

3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.

4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ 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.

Shrinkage Methods: Least Angle Regression

Recall – Forward Stagewise Regression:

- Start with all coefficients β_j equal to zero.
- Find the predictor x_j most correlated with y, and add it into the model. Take residuals $r = y - \hat{y}$.
- Continue, at each stage adding to the model the predictor most correlated with r.
- Continue until all predictors are in the model.

Least Angle Regression Algorithm:

- Start with all coefficients β_j equal to zero.
- Find the predictor x_j most correlated with y.
- Increase the coefficient β_j in the direction of the sign of its correlation with y. Take the residuals along the way, and stop when some other predictor x_k has as much correlation with the residual r.
- Increase (β_j, β_k) in their joint least squares direction, until some other predictor has as much correlation with the residual r.
- Continue until all predictors are in the model.

Shrinkage Methods: Least Angle Regression

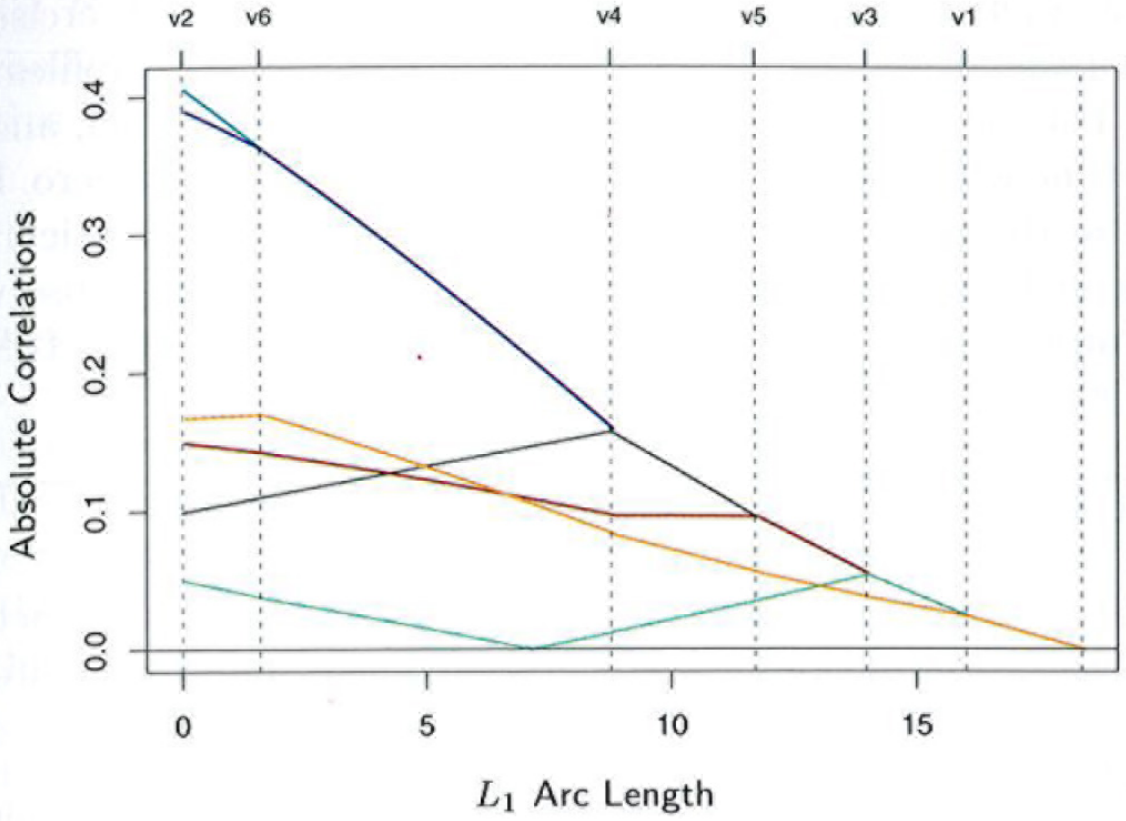


FIGURE 3.14. *Progression of the absolute correlations during each step of the LAR procedure, using a simulated data set with six predictors. The labels at the top of the plot indicate which variables enter the active set at each step. The step length are measured in units of L_1 arc length.*

Shrinkage Methods: Least Angle Regression

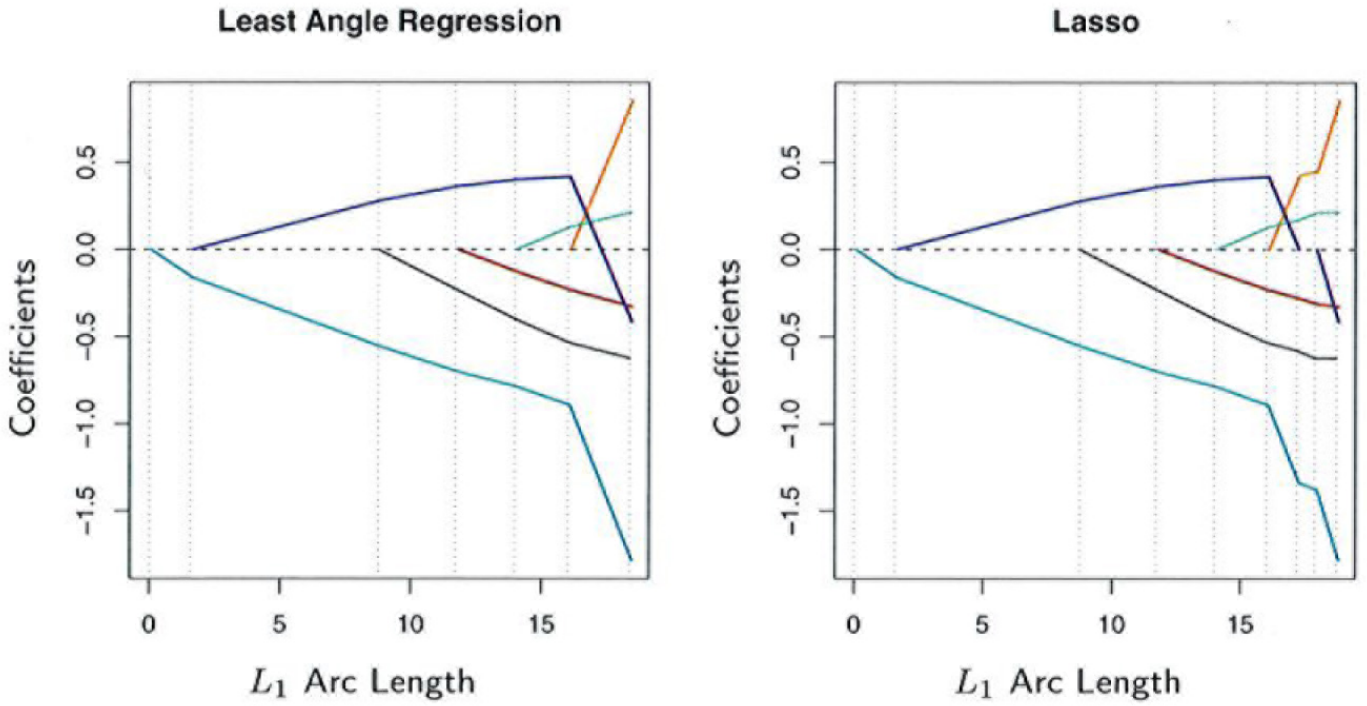


FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about* 18.

Shrinkage Methods: Least Angle Regression

Least Angle Regression Algorithm:

- Start with all coefficients β_j equal to zero. y
- Find the predictor x_j most correlated with y .
- Increase the coefficient β_j in the direction of the sign of its correlation with y. Take the residuals along the way, and stop when some other predictor x_k has as much correlation with the residual r.
- Increase (β_j, β_k) in their joint least squares direction, until some other predictor has as much correlation with the residual r.
- Continue until all predictors are in the model.

Note with one modification – this solution gives the entire solution path. That modification: when a coefficient hits zero – remove it from the active set of predictors and re-compute the current joint least squares direction.