FSAN-815/ELEG-815: Foundations of Statistical Learning

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware

Fall 2014

Shrinkage Methods

The standard linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon. \quad (1)$$

Two problems typically raise with the least squares estimates.

- *Prediction accuracy*: The least squares estimates often have low bias but large variance.
- *Interpretation*: With a large number of predictors, a smaller subset is desired that exhibit the strongest effects.

Shrinkage Methods

Shrinkage

Fitting a model involving all p predictors. Shrinking coefficients towards zero. This shrinkage has the effect of reducing variance. Shrinkage methods can also perform variable selection.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_{p-1} X_{p-1} + \beta_p X_p + \varepsilon. \quad (2)$$

Example: Credit data set:

- Y =balance
- X=age, education, rating, income, gender, student...

Ridge Regression

Ridge regression shrinkage minimizes a penalized residual sum of squares,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right] \\ &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\left[\underbrace{||\mathbf{y}-\beta_0-\mathbf{X}\boldsymbol{\beta}||_2^2}_{\text{Loss}}+\underbrace{\lambda||\boldsymbol{\beta}||_2^2}_{\text{Penalty}}\right],\end{aligned}$$

where $||\boldsymbol{\beta}||_2$ is the $\ell-2$ norm of $\boldsymbol{\beta}$: $||\boldsymbol{\beta}||_2=\sqrt{\sum_{j=1}^{p}\beta_j^2}$.

Here $\lambda\geq 0$ is a tuning parameter, which controls the strength of the penalty term. Note that:

- When $\lambda=0$, we get the linear regression estimate.
- When $\lambda=\infty$, we get $\boldsymbol{\beta}^{ridge}=0$.
- For λ in between, we are balancing two ideas: fitting a linear model of Y on X, and shrinking the coefficients.

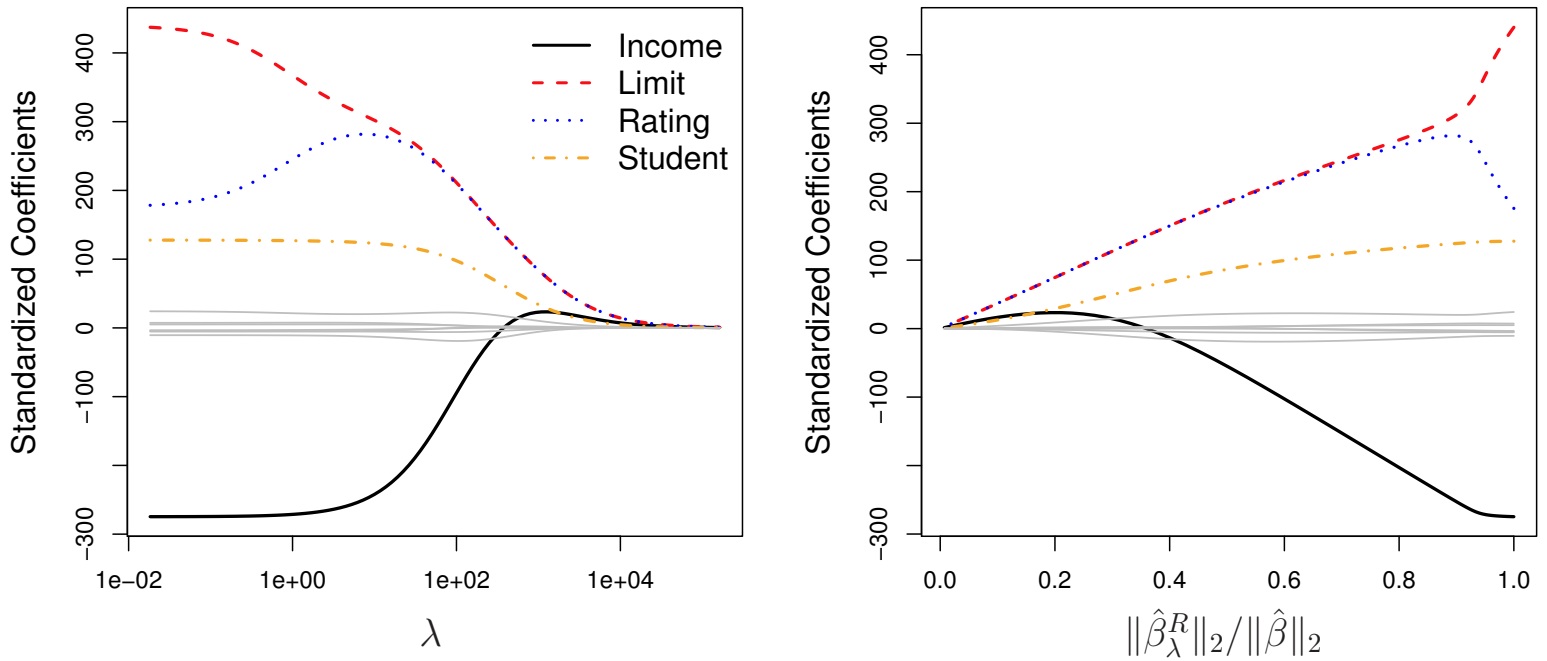
Important Details

- If we center the columns of $\mathbf{X}$, then the intercept estimate ends up just being $\hat{\beta}_0 = \bar{y}$, so we usually assume that $\mathbf{y}$, $\mathbf{X}$ have been centered and don't include an intercept.

- The penalty term $||\boldsymbol{\beta}||_2^2$ is unfair if the predictor variables are not on the same scale. Variables are not measured in the same units, we typically scale the columns of $\mathbf{X}$ (to have sample variance 1), and then we perform ridge regression.

Ridge Regression

Ridge regression estimates for the Credit data set are displayed. Each curve corresponds to estimate for one of the ten variables.



- Left: As $\lambda \uparrow$, the ridge estimates $\hat{\beta}_k \to 0$.
- Right: The ℓ_2 norm, $\| \hat{\boldsymbol{\beta}} \|_2 = (\sum_{j=1}^{p} \beta_j^2)^{1/2}$ measures the distance of β from zero.

Ridge Regression

The penalized residual sum of squares (PRSS):

$$PRSS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2, \tag{3}$$

Differentiating with respect to $\boldsymbol{\beta}$, we obtain,

$$\frac{\partial PRSS}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} \tag{4}$$

PRSS($\boldsymbol{\beta}$) is convex. Set the first derivative to zero,

$$\lambda\boldsymbol{\beta} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{5}$$

The ridge regression solution,

$$\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}. \tag{6}$$

- Solution is indexed by the turning parameter λ.
- Inclusion of λ makes problem non-singular even if $\mathbf{X}^T\mathbf{X}$ is not invertible.

Tuning Parameter λ

The ridge regression solution,

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

- Solution is indexed by the parameter λ
- So for each λ, we have a solution.
- The λs trace out a path of solutions.
- λ is the shrinkage parameter
- λ controls the size of the coefficients.
- λ controls amount of regularization.
- As $\lambda \to 0$, we obtain the LS solutions.
- As $\lambda \to \infty$, we have $\hat{\boldsymbol{\beta}}^{\text{ridge}}_{\lambda=\infty} = 0$.

Proving that $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ is biased

Let $\mathbf{R} = \mathbf{X}^T\mathbf{X}$.
Then:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}} &= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y} \\
&= (\mathbf{R} + \lambda \mathbf{I}_p)^{-1}\mathbf{R}(\mathbf{R}^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{R}(\mathbf{I}_p + \lambda \mathbf{R}^{-1}))^{-1}\mathbf{R}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{I}_p + \lambda \mathbf{R}^{-1})^{-1}\mathbf{R}^{-1}\mathbf{R}\hat{\boldsymbol{\beta}}^{ls} \\
&= (\mathbf{I}_p + \lambda \mathbf{R}^{-1})^{-1}\hat{\boldsymbol{\beta}}^{ls}
\end{aligned}$$

So,

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}}) &= E((\mathbf{I}_p + \lambda \mathbf{R}^{-1})^{-1}\hat{\boldsymbol{\beta}}^{ls}) \\
&= (\mathbf{I}_p + \lambda \mathbf{R}^{-1})^{-1}\boldsymbol{\beta} \\
&\neq \boldsymbol{\beta}.
\end{aligned}$$

Data Augmentation Approach

- The ℓ_2 PRSS can be written as:

$$\begin{aligned} \text{PRSS}(\boldsymbol{\beta})_{\ell_2} &= \sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\mathbf{x}_j^T\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \\ &= \sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\mathbf{x}_j^T\beta_j)^2 + \sum_{j=1}^{p}(0 - \sqrt{\lambda}\beta_j)^2 \end{aligned}$$

- Hence, the ℓ_2 criterion can be recast as another LS problem for another data set.

Data Augmentation Approach

The ℓ_2 criterion is the RSS for the augmented data set:

$$\mathbf{X}_\lambda = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{Np} \\ \sqrt{\lambda} & 0 & \dots & 0 \\ 0 & \sqrt{\lambda} & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & \sqrt{\lambda} \end{bmatrix}; \ \mathbf{y}_\lambda = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{7}$$

So,

$$\mathbf{X}_\lambda = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_p \end{bmatrix}; \ \mathbf{y}_\lambda = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix}. \tag{8}$$

Solving The Augmented Data Set

So the 'least squares' solution for the augmented data set is:

$$(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T\mathbf{y}_\lambda = \left((\mathbf{X}^T, \sqrt{\lambda}\mathbf{I}_p)\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{I}_p\end{pmatrix}\right)^{-1}(\mathbf{X}^T, \sqrt{\lambda}\mathbf{I}_p)\begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix} \quad (9)$$

$$= (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}, \quad (10)$$

which is simply the ridge solution.

Orthonormal $\mathbf{X}$ In Ridge Regression

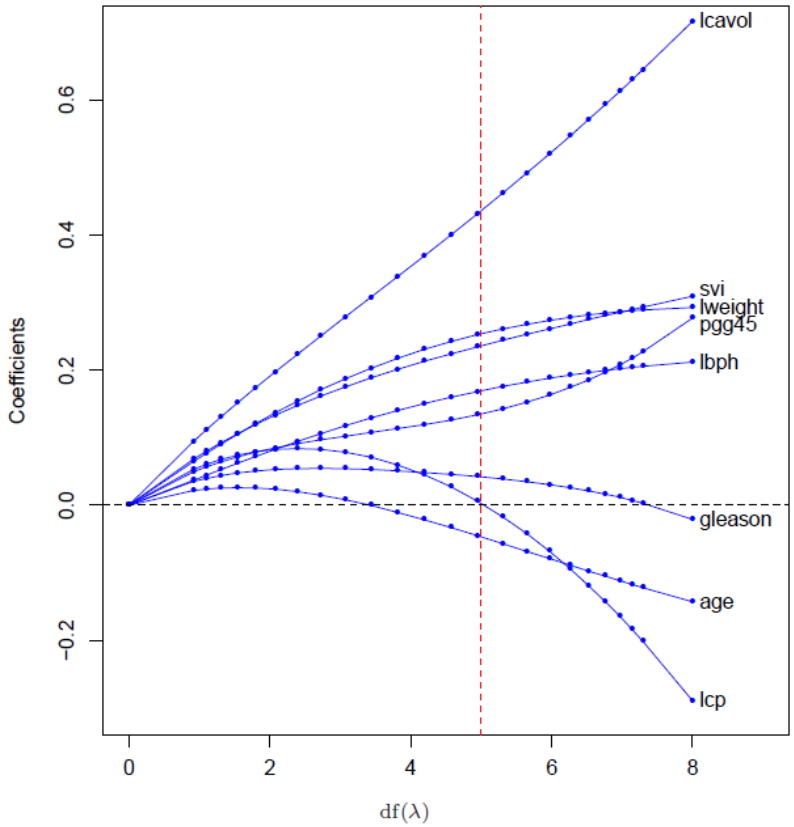
- If $\mathbf{X}$ is orthonormal, then $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, then a couple of closed form properties exist.

- Let $\hat{\boldsymbol{\beta}}^{ls}$ denote the LS solution for our orthonormal $\mathbf{X}$; then

$$\hat{\boldsymbol{\beta}}_{\lambda}^{ridge} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{ls}. \tag{11}$$

Ridge Regression

In case of orthonomal inputs, $\mathbf{X}^T = \mathbf{X}^{-1}$, we have $\hat{\boldsymbol{\beta}}^{ridge} = \hat{\boldsymbol{\beta}}/(1+\lambda)$



Profiles of ridge coefficients for the prostate cancer example, plotted versus $df(\lambda)$, the effective degrees of freedom.

Prediction Error And The Bias-Variance Tradeoff

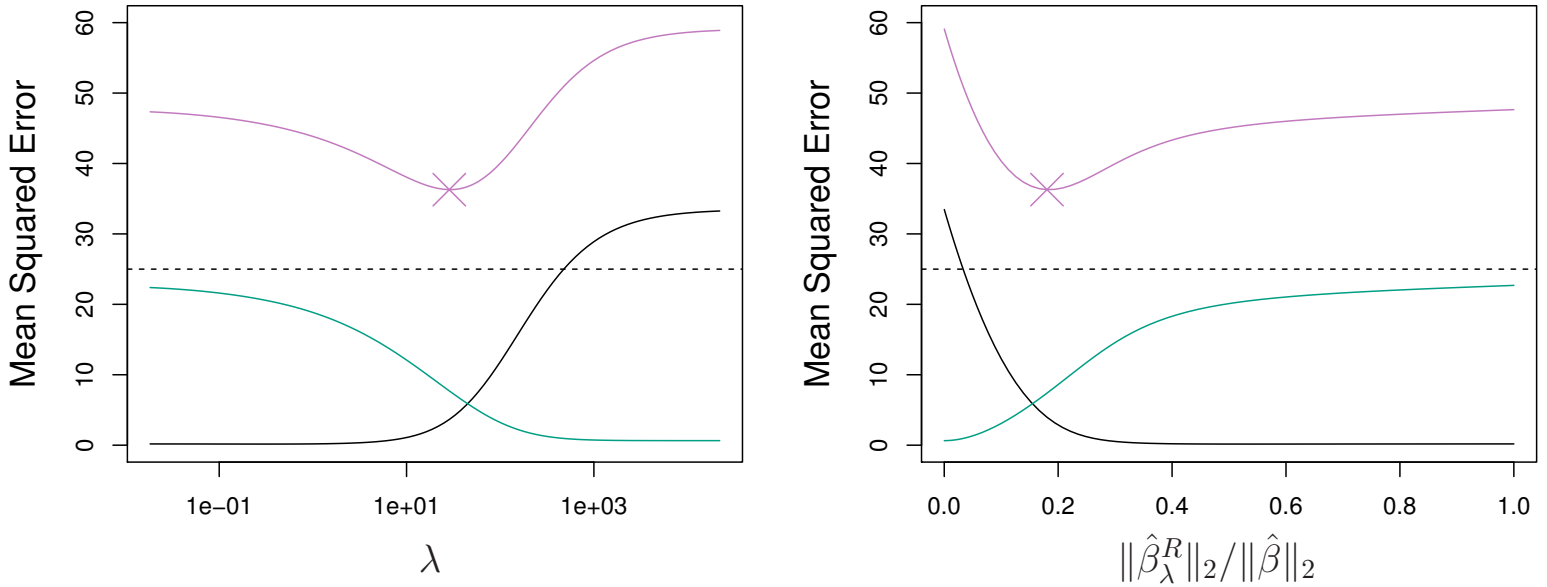
- So good estimators should, on average have, small prediction errors.
- Let's consider the PE at a particular target points $\mathbf{x}_0$:

$$\mathrm{PE}(\mathbf{x}_0) = \sigma_\varepsilon^2 + \mathrm{Bias}^2(\mathrm{f}(\mathbf{x}_0)) + \mathrm{Var}(\mathrm{f}(\mathbf{x}_0)). \tag{12}$$

- Such a decomposition is known as the bias-variance tradeoff.
  - As model becomes more complex (more terms included), local structure/curvature can be picked up.
  - But coefficient estimates suffer from high variance as more terms are included in the model.
- So introducing a little bias in our estimate for β might lead to a substantial decrease in variance, and hence to a substantial decrease in PE.

Ridge Regression

Bias-variance trade-off.



Squared bias (black), variance (green), and test mean squared error (purple).

- $\lambda = 0$, the variance is high but there is no bias.
- As λ increases, the variance decreases, at the expense of bias.

Lasso Regression

- Tibshirani introduced the LASSO: least absolute shrinkage and selection operator.
- LASSO coefficients are the solutions to the ℓ_1 optimization problem: The lasso estimate is defined as

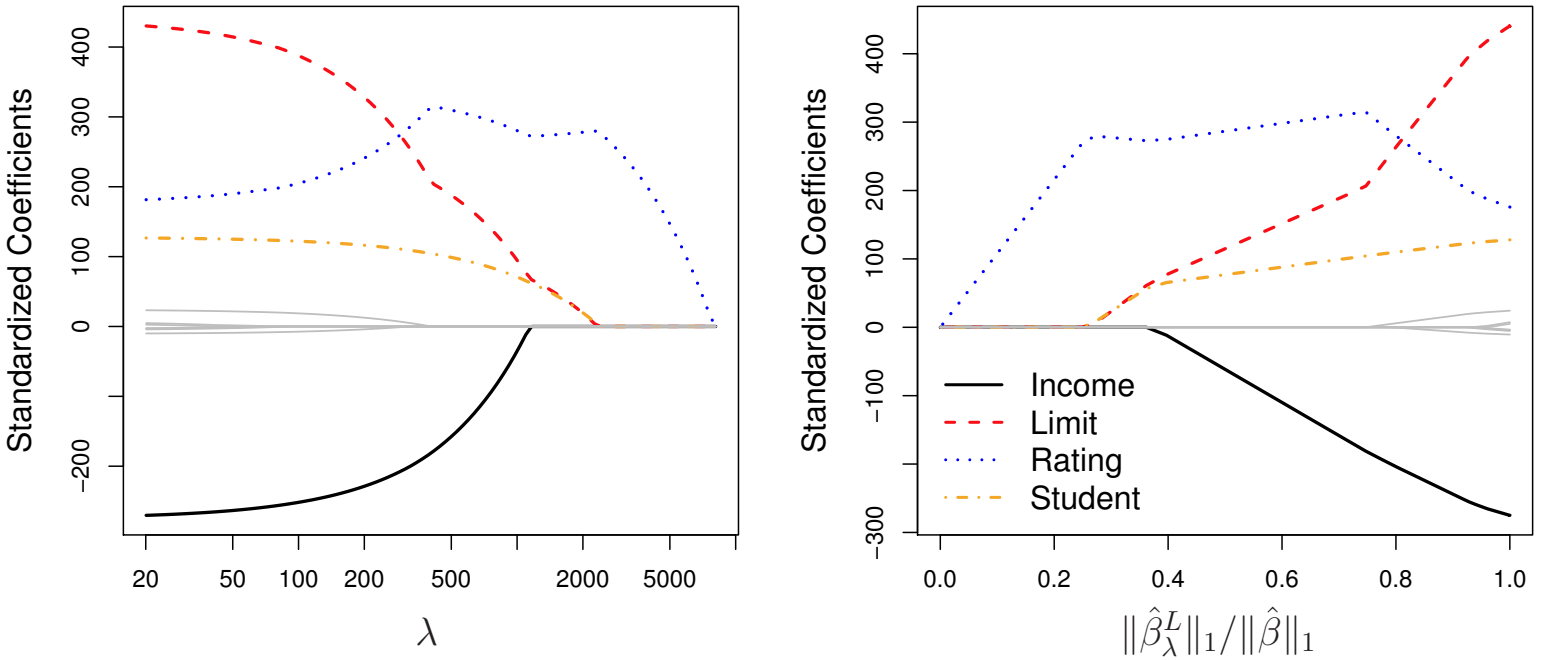
$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{lasso} &= \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{N} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda ||\beta_j||_1 \right].
\end{aligned}$$

- Notice the similarity to the ridge regression problem:
The ℓ_2 norm *ridge penalty* $\| \boldsymbol{\beta} \|_2$ is replaced by the ℓ_1 norm, *lasso penalty* $\| \boldsymbol{\beta} \|_1 = \sum_{j=1}^{p} |\beta_j|$.

Lasso Regression

- As with ridge regression, the lasso shrinks the coefficient estimates.
- The ℓ_1 norm penalty forces some of the coefficient estimates to zero when λ is large. Lasso performs **variable selection**.
- Lasso yields **sparse** models, models with only a subset of the variables.
- Unlike ridge regression, $\hat{\boldsymbol{\beta}}_{\lambda}^{lasso}$ has no closed form.
- Original implementation involves quadratic programming techniques from convex optimization.
- lars package in R implements the LASSO.

Lasso Regression



Lasso regression estimates for the Credit data set.

The Variable Selection Property of the Lasso

One can show that the Ridge and Lasso regression coefficient estimates solve the problems

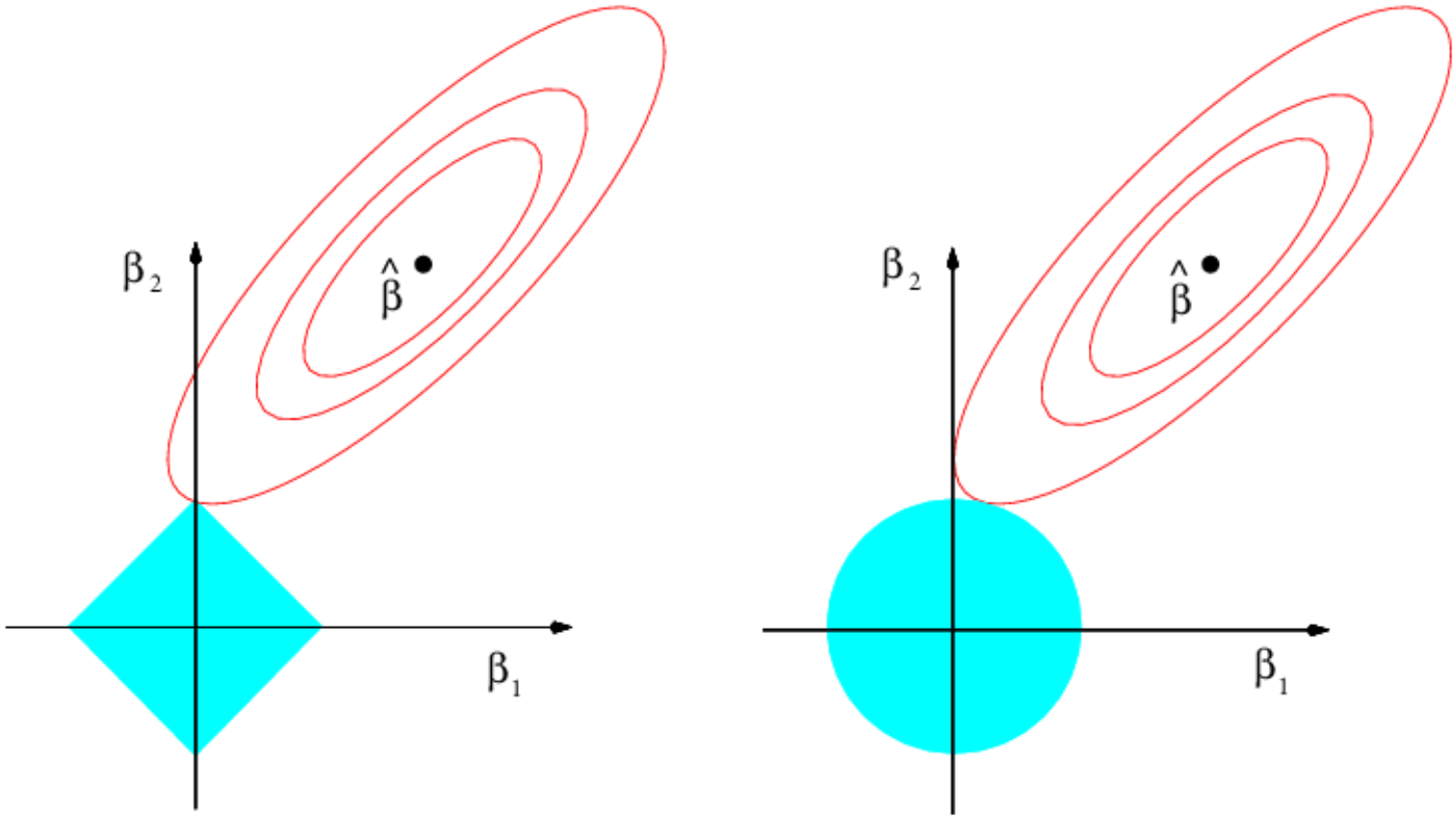
$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{argmin}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} \tag{13}$$

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

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{argmin}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} \tag{14}$$

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

The Variable Selection Property of the Lasso



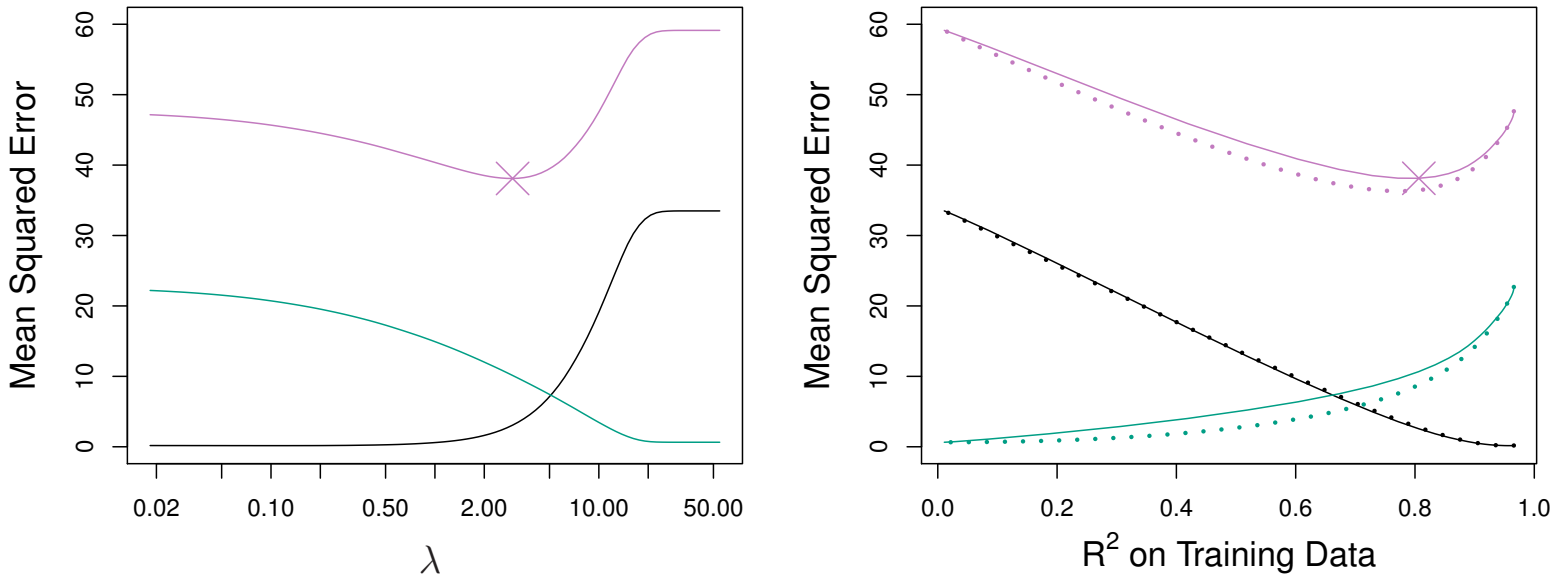
- *RSS* has elliptical contours, centered at the *LS* estimate.
- Constraint regions, $\beta_1^2 + \beta_2^2 \leq t$, and $|\beta_1| + |\beta_2| \leq t$.

Comparing the Lasso and Ridge Regression

Lasso has a major advantage over ridge regression. It produces simpler models that involve only a subset of the predictors. Which method leads to better prediction accuracy?

- Lasso performs better when a small number of predictors have substantial coefficients, and the remaining predictors are small or zero.
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.

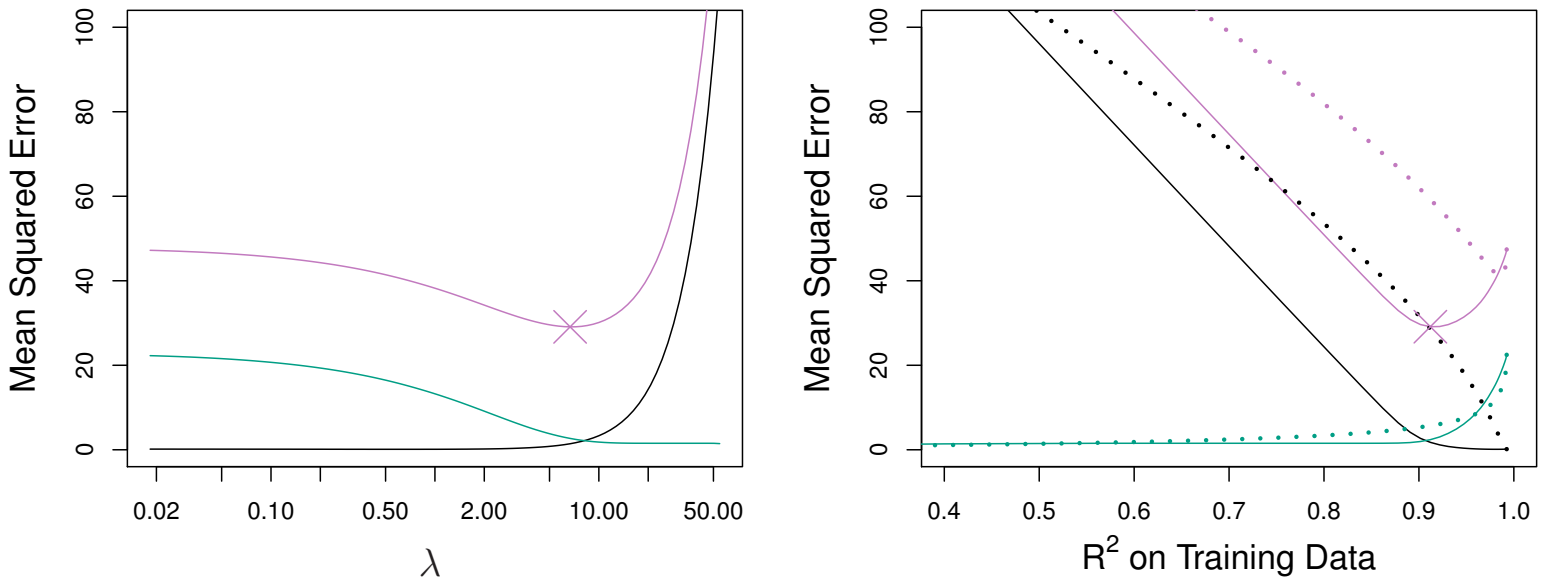
Comparing the Lasso and Ridge Regression



In this example all 45 predictors were related to the response.

- Left: Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso.
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed).

Comparing the Lasso and Ridge Regression



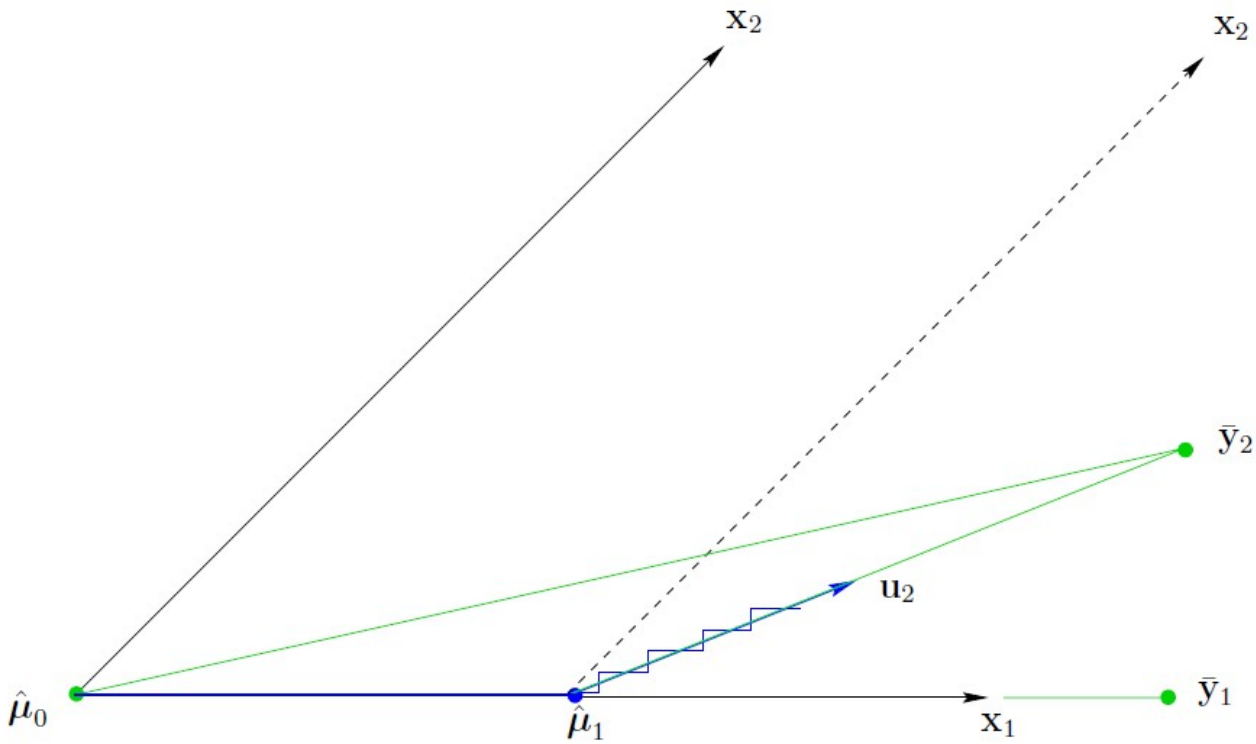
Here the the response is a function of only 2 out of 45 predictors.

- Left: Squared bias (black), variance (green), and test MSE (purple) for the lasso.
- Right: Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed).

Least Angle Regression

1. Standardize the predictors to have mean zero and unit norm. Start with all coefficients equal to zero and find the predictor most correlated with the response, say x_{j_1}.
2. Take the largest step possible in the direction of this predictor until some other predictor, say x_{j_2}, has much correlation with the current residual.
3. Proceed in a direction equiangular between x_{j_1} and x_{j_2} until a third variable x_{j_3} earns its way into the " most correlated " set.
4. LAR then proceeds equiangular between x_{j_1}, x_{j_2} and x_{j_3} until a fourth variable enters, and so on.

Illustration of 2 covariates



The LAR algorithm in the case of $p = 2$ covariates; $\bar{y}_2$ is the projection of $\mathbf{y}$ in $\text{span}\{x_1, x_2\}$.

Algorithm

LAR builds up estimates

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{15}$$

Correlation between $\mathbf{X}$ and residual

$$\mathbf{c}(\hat{\boldsymbol{\mu}}) = \mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{X}^T(\bar{\mathbf{y}} - \hat{\boldsymbol{\mu}}) \tag{16}$$

For the example shown in last page, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$.

- Algorithm begins at $\hat{\boldsymbol{\mu}}_0 = 0$, the residual vector $\bar{y}_2 - \hat{\boldsymbol{\mu}}_0$ has greater correlation with $\mathbf{x}_1$.
- The next LAR estimate is $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_0 + \hat{\gamma}_1 \mathbf{x}_1$, where $\hat{\gamma}_1$ is chosen such that $\bar{y}_2 - \hat{\boldsymbol{\mu}}_1$ bisects the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$.
- Then $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1 + \hat{\gamma}_1 \mathbf{u}_2$, where $\mathbf{u}_2$ is the unit bisector.

Mathematical formulation of LAR

Assumption: The covariate vectors $\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_p$ are *linearly independent*. $\mathscr{A}$ is a subset of the indices $1, 2, ..., p$

- $$\begin{aligned} \mathbf{X}_{\mathscr{A}} &= (...s_j \mathbf{x}_j ...)_{j \in \mathscr{A}} \\ \mathbf{G}_{\mathscr{A}} &= \mathbf{X}_{\mathscr{A}}^T \mathbf{X}_{\mathscr{A}} \\ A_{\mathscr{A}} &= (\mathbf{1}_{\mathscr{A}}^T \mathbf{G}_{\mathscr{A}}^{-1} \mathbf{1}_{\mathscr{A}})^{-1/2} \end{aligned} \tag{17}$$

- *Equiangular vector*

$$\mathbf{u}_{\mathscr{A}} = \mathbf{X}_{\mathscr{A}} \omega_{\mathscr{A}}, \ \textit{where } \omega_{\mathscr{A}} = A_{\mathscr{A}} G_{\mathscr{A}}^{-1} \mathbf{1}_{\mathscr{A}} \tag{18}$$

is the unit vector making equal angles, with the columns of $\mathbf{X}_{\mathscr{A}}$

$$\mathbf{X}_{\mathscr{A}}^T \mathbf{u}_{\mathscr{A}} = A_{\mathscr{A}} \mathbf{1}_{\mathscr{A}} \ \textit{and} \ \| \mathbf{u}_{\mathscr{A}} \|^2 = 1 \tag{19}$$

Algorithm

We begin at $\hat{\boldsymbol{\mu}}_0 = 0$ and build up $\boldsymbol{\mu}$ by steps. Suppose $\hat{\boldsymbol{\mu}}_{\mathscr{A}}$ be the current LAR estimate.

1. Current correlations: $\hat{\mathbf{c}} = \mathbf{X}^T(\mathbf{y} - \hat{\boldsymbol{\mu}}_{\mathscr{A}})$.
2. The greatest absolute current correlations $\hat{C} = \max_j\{|\hat{c}_j|\}$ and the corresponding indices $\mathscr{A} = \{j : |\hat{c}_j| = \hat{C}\}$; $s_j = sign\{\hat{c}_j\}$ for $j \in \mathscr{A}$.
3. Compute $\mathbf{X}_{\mathscr{A}}, A_{\mathscr{A}}, \mathbf{u}_{\mathscr{A}}$ and inner product $\mathbf{a} \equiv \mathbf{X}^T\mathbf{u}_{\mathscr{A}}$.
4. Update $\hat{\boldsymbol{\mu}}_{\mathscr{A}}$: $\hat{\boldsymbol{\mu}}_{\mathscr{A}+} = \hat{\boldsymbol{\mu}}_{\mathscr{A}} + \hat{\gamma}\mathbf{u}_{\mathscr{A}}$
   where

$$\hat{\gamma} = min^+_{j \in \mathscr{A}^c}\{\frac{\hat{C} - \hat{c}_j}{A_{\mathscr{A}} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_{\mathscr{A}} + a_j}\}, \tag{20}$$

   where "min^+" indicates that the minimum is taken over only positive components within each choice of j.

Algorithm

To explain the equation, define

$$\hat{\boldsymbol{\mu}}(\gamma) = \hat{\boldsymbol{\mu}}_{\mathscr{A}} + \gamma \mathbf{u}_{\mathscr{A}} \tag{21}$$

for $\gamma > 0$, the current correlation is

$$c_j(\gamma) = \mathbf{x}_j^T(y - \hat{\boldsymbol{\mu}}(\gamma)) = \hat{c}_j - \gamma a_j \tag{22}$$

For $j \in \mathscr{A}$, we have

$$|c_j(\gamma)| = \hat{C} - \gamma A_{\mathscr{A}} \tag{23}$$

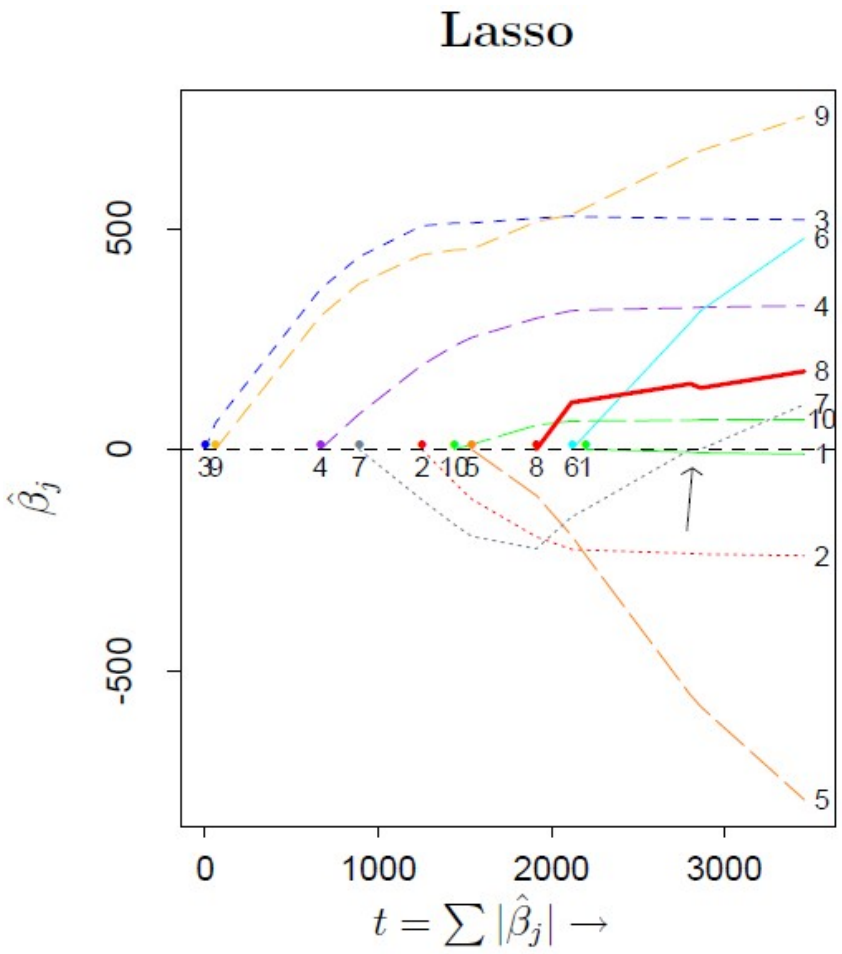
- All of the maximal absolute current correlations decline equally.
- $\hat{\gamma}$ is the smallest positive value such that some new index $\hat{j}$ joins the active set.

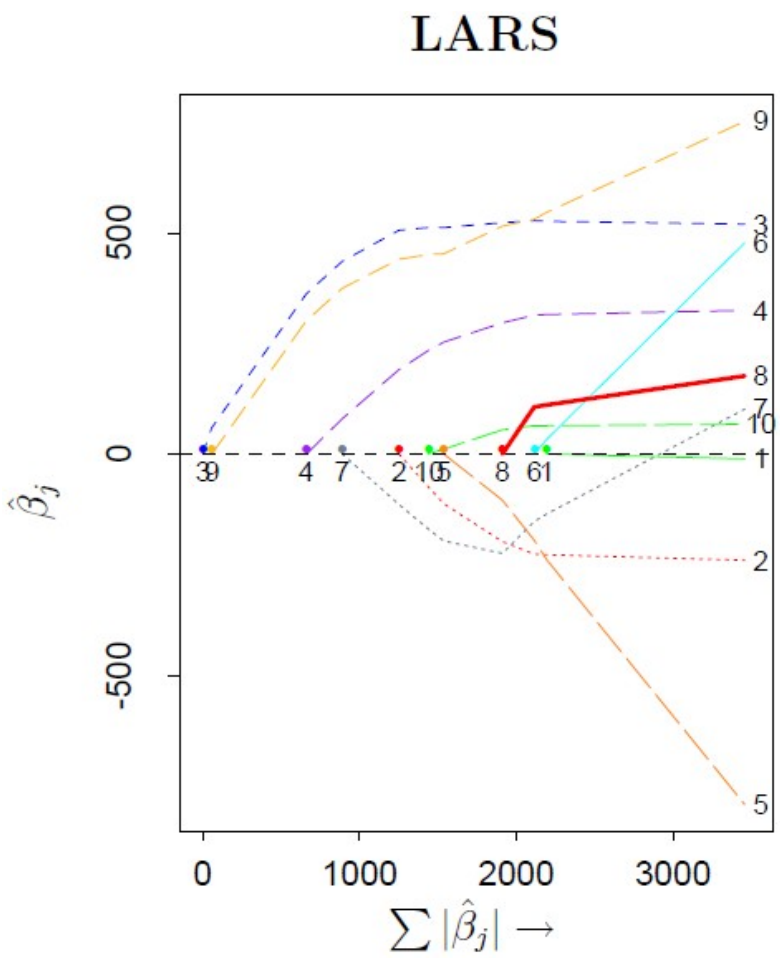
Example

- **Diabetes study:** 442 diabetes patients measured on 10 baseline variables: Age, sex, body mass index, average blood pressure and six blood serum measurements.
- **Response variable:** A measure of disease progression one year after baseline.

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

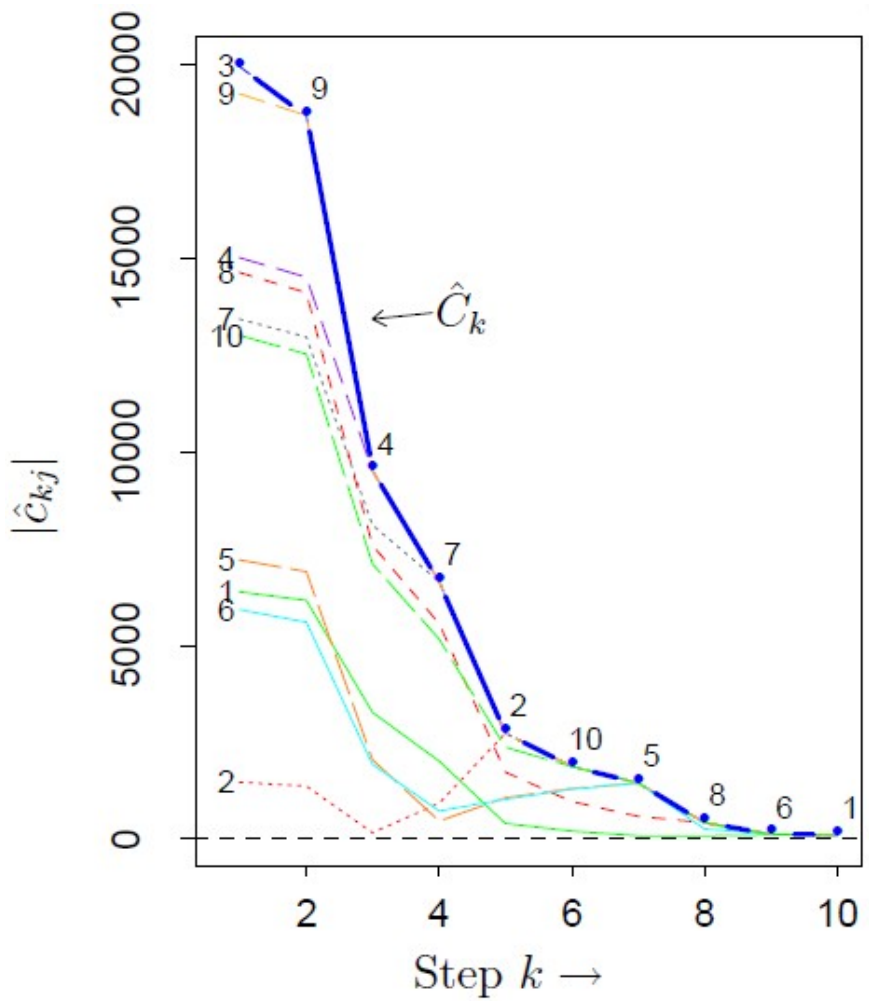
Example





The variables join the active set $\mathscr{A}$ in the same order as for the Lasso: $3, 9, 4, 7, \ldots, 1$.

Example



- Variables enter active set $\mathscr{A}$ in order $3, 9, 4, 7, ..., 1$.
- Maximum current correlation $\hat{C}_k$ declining with k.

The LARS/Lasso Relationship

Suppose $\mathscr{A}$ is the active set of variables at some stage in the algorithm, we have

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\beta) = |\hat{c}_j| s_j, \ \forall j \in \mathscr{A} \tag{24}$$

Consider the lassocriterion

$$R(\boldsymbol{\beta}) = \frac{1}{2} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel_2^2 + \lambda \parallel \beta \parallel_1 \tag{25}$$

$\mathscr{B}$ is active set of variables in the solution for a given λ.

$$\frac{\partial R(\boldsymbol{\beta})}{\partial \beta_j} = -\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \times sign(\beta_j) = 0. \tag{26}$$

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \lambda \times sign(\beta_j), \ \forall j \in \mathscr{B} \tag{27}$$

LAR and Lasso are identical only if the sign of β_j matches the sign of the inner product.

Multiple Outputs

To predict multiple outputs $Y_1, Y_2, ..., Y_K$ from inputs $X_0, X_1, ..., X_p$, assume a linear model for each output

$$\begin{aligned} Y_k &= \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \varepsilon_k \\ &= f_k(X) + \varepsilon_k \end{aligned} \tag{28}$$

With N training cases,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{29}$$

where $\mathbf{Y}$ is the $N \times K$ response matrix, $\mathbf{X}$ is the $N \times (p+1)$ input matrix, $\mathbf{B}$ is the $(p+1) \times K$ matrix of parameters, $\mathbf{E}$ is the $N \times K$ matrix of errors.

Multiple Outputs

$$RSS(\mathbf{B}) = \sum_{k=1}^{K}\sum_{i=1}^{N}(y_{ik} - f_k(x_i))^2 \qquad (30)$$
$$= tr[(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})]$$

The least squares estimates is

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad (31)$$

- Coefficients for the kth outcome are LS estimates in the regression of y_k on $x_0, x_1, \ldots x_p$.

Multiple Outputs

To apply shrinkage methods in the multiple output case, one could apply a univariate technique individually to each outcome or simultaneously to all outcomes.
For example, in ridge regression

- Apply ridge regression to each of the K columns of the outcome matrix $\mathbf{Y}$, using possibly different parameters λ.
- Apply ridge regression to all columns using the same value of λ.