Chapter 6 Linear Model Selection and Regularization

Zhenmiao Zhang

November 18, 2020

Outline

- Overview
  - Motivation
- Subset Selection
  - Best Subset Selection
  - Stepwise Selection
- Shrinkage Methods
  - Ridge Regression
  - Lasso Regression
  - Parameter Selection
- Dimension Reduction Methods
  - The Framework
  - Principle Components Regression
  - Partial Least Squares
- Considerations in High Dimensions

Linear Regression

The standard linear regression model with p predictors:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon.$$

The model is typically fit with least squares. For n observations:

$$\begin{aligned} X &= [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n\times(p+1)}, \\ Y &= [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^{n\times 1}, \\ \beta &= [\beta_0, \beta_1, \ldots, \beta_p]^T \in \mathbb{R}^{(p+1)\times 1}, \end{aligned}$$

the estimated $\hat{\beta}$ should be:

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 \Rightarrow \frac{\partial \|Y - X\beta\|_2^2}{\partial \beta} = 0$$

$$\Rightarrow -X^T(Y - X\hat{\beta}) = 0 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y.$$

Number of observations and predictors

Number of observations: n
Number of predictors: p ($p+1$ coefficients)

- $p \ll n$: the least squares estimates tend to have low variance, and will perform well on test observations.
- $p \approx n$: there can be high variability in the least squares fit, resulting in overfitting and consequently poor predictions on test set.
- $p \geq n$: there is no longer a unique least squares coefficient estimate; the variance is infinite; the MSE on an independent test set becomes extremely large.

We need exclude irrelevant variables from a multiple regression model!

Outline

Overview

- Motivation

Subset Selection

- Best Subset Selection
- Stepwise Selection

Shrinkage Methods

- Ridge Regression
- Lasso Regression
- Parameter Selection

Dimension Reduction Methods

- The Framework
- Principle Components Regression
- Partial Least Squares

Considerations in High Dimensions

Best Subset Selection

Best Subset Selection: fit a separate least squares regression for each possible combination of the p predictors; select the best model from among the 2^p possibilities.

Algorithm 1: Best Sample Selection

```
Let $M_0$ denote the null model with no predictors. It
  simply predicts the sample mean for each observation.
for $k = 1, 2, ...p$ do
    Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
    Pick the best among these $\binom{p}{k}$ models, and call it
      $M_k$. Here best is defined as having the smallest
      $RSS$, or equivalently largest $R^2$.
end
Select a single best model from among $M_0, ..., M_p$ using
  cross-validated prediction error, $C_p(AIC)$, $BIC$, or
  adjusted $R^2$.
```

How to select the best model?

Indirect approach: we can estimate test error by making an adjustment to the training error:

- C_p estimate of test MSE: $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$
- AIC for maximum likelihood: $SIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$
- BIC from Bayesian: $BIC = \frac{1}{n\hat{\sigma}^2}(RSS + log(n)d\hat{\sigma}^2)$
- Adjusted R^2: Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$

Direct approach: we can directly estimate the test error, using either a validation set or a cross-validation method.

Indirect approach

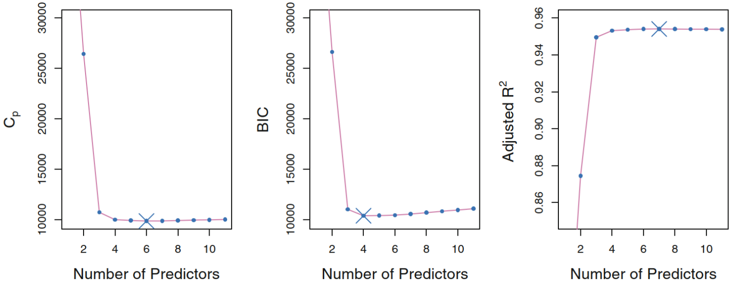


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* **Credit** *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

Direct approach

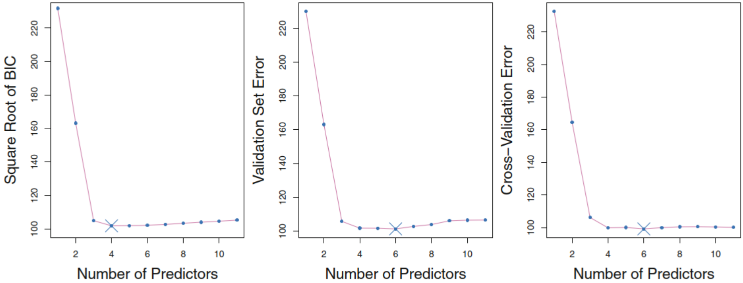


FIGURE 6.3. *For the* Credit *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC*. Center: *Validation set errors.* Right: *Cross-validation errors.*

Drawbacks of best subset selection

Best subset selection has two main drawbacks:

- For computational reasons, best subset selection cannot be applied with very large p ($p \geq 40$).
- An enormous search space can lead to overfitting and high variance of the coefficient estimates. **Guided search** is necessary.

Forward Stepwise Selection

Alternative - Forward Stepwise Selection: select the best model only from $1 + p(p+1)/2$ models.

Algorithm 2: Forward Stepwise Selection

```
Let $M_0$ denote the null model with no predictors.
for $k = 0, 1, ...p-1$ do
    Consider all $p-k$ models that augment the
      predictors in $M_k$ with one additional predictor.
    Choose the best among these $p-k$ models, and call
      it $M_{k+1}$. The best is defined as having smallest
      $RSS$, or largest $R^2$.
end
Select a single best model from among $M_0, ..., M_p$ using
  cross-validated prediction error, $C_p(AIC)$, $BIC$, or
  adjusted $R^2$.
```

Backward Stepwise Selection

Alternative - Backward Stepwise Selection: begins with the full least squares model.

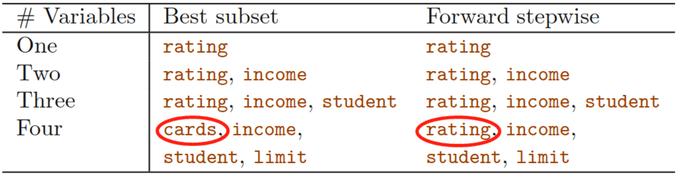
Algorithm 3: Backward Stepwise Selection

```
Let $M_p$ denote the full model, which contains all $p$ predictors.
for $k = p, p - 1, ...1$ do
    Consider all $k$ models that that contain all but one of the predictors in $M_k$, for a total of k $-$ 1 predictors.
    Choose the best among these $k$ models, and call it $M_{k-1}$. The best is defined as having smallest $RSS$, or largest $R^2$.
end
Select a single best model from among $M_0, ..., M_p$ using cross-validated prediction error, $C_p(AIC), BIC$, or adjusted $R^2$.
```

Hybrid Approach

- Similarly to forward selection, variables are added to the model sequentially.
- After adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.
- Better model space exploration while retaining computational advantages of both stepwise selections.

Comparison of the stepwise selection approaches



# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income, student, limit	rating, income, student, limit

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* Credit *data set. The first three models are identical but the fourth models differ.*

- Forward stepwise, backward stepwise and hybrid approach are not guaranteed to yield the best model containing a subset of the p predictors.
- Forward stepwise and hybrid approach can be used even when $n \leq p$. In this case, it is possible to construct sub-models $M_0, ..., M_{n-1}$ only.

Outline

Overview

- Motivation

Subset Selection

- Best Subset Selection
- Stepwise Selection

Shrinkage Methods

- Ridge Regression
- Lasso Regression
- Parameter Selection

Dimension Reduction Methods

- The Framework
- Principle Components Regression
- Partial Least Squares

Considerations in High Dimensions

Ridge Regression

$$\hat{\beta}^R = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Centered trick for bias term

Transform the training data to centered form:

$$
\begin{aligned}
X &= [x_1 - \bar{x}, ..., x_n - \bar{x}]^T \\
Y &= [y_1 - \bar{y}, ..., y_n - \bar{y}]^T \\
\beta &= [\beta_1, ..., \beta_p]^T, \beta_0 = \bar{y} - \beta^T \bar{x}
\end{aligned}
$$

Ridge Regression (centered):

$$\hat{\beta}^R = \arg\min_{\beta} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_2^2 \tag{1}$$

Another formulation (centered):

$$\hat{\beta}^R = \arg\min_{\beta} \|X\beta - Y\|_2^2, \ s.t. \ \|\beta\|_2^2 \leq s \tag{2}$$

Proof of equivalence

Proof
Both (1) and (2) are convex programming, thus the K-K-T are sufficient and necessary conditions to optimal points. The Lagrange function of (2):

$$L_2(\beta; \lambda_{L_2}) = \|X\beta - Y\|_2^2 + \lambda_{L_2}(\|\beta\|_2^2 - s).$$

K-K-T conditions:

$$\begin{aligned} &\frac{\partial L_2(\beta; \lambda_{L_2})}{\partial \beta} = 0, \\ &\|\beta\|_2^2 - s \leq 0, \\ &\lambda_{L_2} \geq 0, \\ &\lambda_{L_2}(\|\beta\|_2^2 - s) = 0. \end{aligned}$$

Proof of equivalence

The Lagrange function of (1):

$$L_1(\beta) = \|X\beta - Y\|_2^2 + \lambda\|\beta\|_2^2.$$

K-K-T conditions:

$$\frac{\partial L_1(\beta)}{\partial \beta} = 0.$$

□

Let $\lambda_{L_2} = \lambda \geq 0$, we have $\frac{\partial L_2(\beta;\lambda_{L_2})}{\partial \beta} = \frac{\partial L_1(\beta)}{\partial \beta}$. Considering K-K-T are sufficient and necessary conditions, all the optimal points of (1) are optimal points of (2), and vice versa (let $s = \|\hat{\beta}\|_2^2$ when $\lambda \neq 0$).

Solution of ridge regression

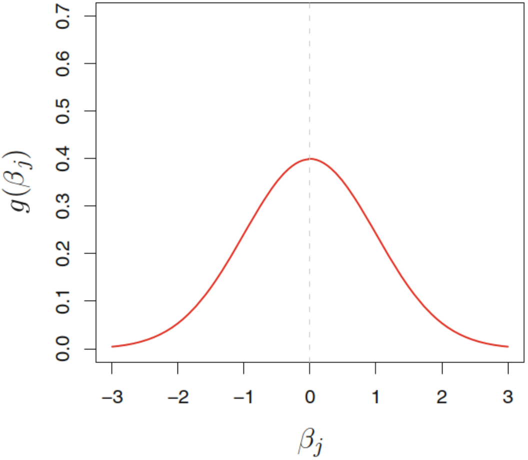
From the K-K-T conditions of (1), we can directly give the estimate of β:

$$\frac{\partial L_1(\beta)}{\partial\beta} = 0 \Rightarrow X^T(X\hat{\beta}^R - Y) + \lambda\hat{\beta}^R = 0$$
$$\Rightarrow \hat{\beta}^R = (X^TX + \lambda I)^{-1}X^TY$$

If $\lambda \neq 0$, we have $s = \|\hat{\beta}^R\|_2^2 = \|(X^TX + \lambda I)^{-1}X^TY\|_2^2$.
If $\lambda = 0$, $\hat{\beta}^R = (X^TX)^{-1}X^TY = \hat{\beta}$ (least square).

Bayesian interpretation of ridge regression

Ridge regression is maximum a posteriori estimation (MAP) that assumes β has prior distribution of $\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I)$.



Bayesian interpretation of ridge regression

Proof

The model underlying linear regression is $Y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In Bayesian statistics, β is a random vector that has a prior distribution. We can get:

$$\beta \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda} I),$$

$$i.e., f_\beta(\beta) = (2\pi)^{-\frac{p}{2}} p^{-\frac{1}{2}} \exp(-\frac{1}{2}\beta^T \frac{\lambda}{\sigma^2}\beta)$$

$$Y|X, \beta \sim \mathcal{N}(X\beta, \sigma^2 I),$$

$$i.e., f_{Y|X,\beta}(Y|X, \beta) = (2\pi)^{-\frac{n}{2}} n^{-\frac{1}{2}} \exp(-\frac{1}{2}(Y - X\beta)^T \frac{1}{\sigma^2}(Y - X\beta))$$

Bayesian interpretation of ridge regression

The mode of posterior distribution (conditional distribution) of β (assume X is fixed):

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta} f_{\beta|X,Y}(\beta|X,Y) \\ &= \arg\max_{\beta} \frac{f_{X,\beta}(X,\beta)f_{Y|X,\beta}(Y|X,\beta)}{f_{X,Y}(X,Y)} \\ &= \arg\max_{\beta} \frac{f_X(X)f_{\beta|X}(\beta|X)f_{Y|X,\beta}(Y|X,\beta)}{f_{X,Y}(X,Y)} \\ &= \arg\max_{\beta} f_{\beta}(\beta)f_{Y|X,\beta}(Y|X,\beta)\end{aligned}$$

Bayesian interpretation of ridge regression

$$= \arg\max_{\beta} \exp\Big(-\frac{1}{2}\beta^T\frac{\lambda}{\sigma^2}\beta - \frac{1}{2}(Y - X\beta)^T\frac{1}{\sigma^2}(Y - X\beta)\Big)$$

$$= \arg\min_{\beta} \frac{1}{2}\beta^T\frac{\lambda}{\sigma^2}\beta + \frac{1}{2}(Y - X\beta)^T\frac{1}{\sigma^2}(Y - X\beta)$$

$$= \arg\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

$$= \arg\min_{\beta} \|X\beta - Y\|_2^2 + \lambda\|\beta\|_2^2$$

□

Standardizing the predictors

For ordinary least squares, $X_j\hat{\beta}_j$ depends on the scaling of X_j. But for ridge regression, $X_j\hat{\beta}_{j,\lambda}$ will depend not only on the value of λ, but also on the scaling of the j^{th} predictor. It may even depend on the scaling of other predictors. Standardize the predictors before ridge regression:

$$\tilde{x_{ij}} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}}$$

An application

Ridge regression more or less shrinks every dimension of the data by the same proportion.

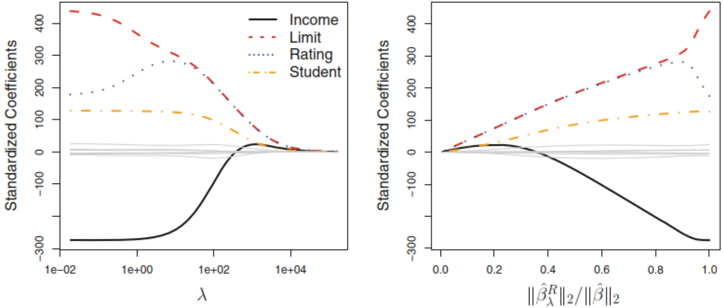


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* Credit *data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

An application

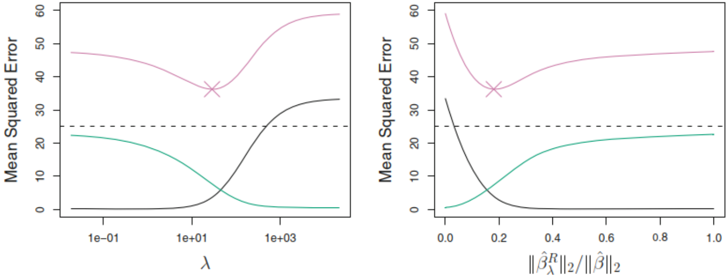


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Lasso Regression

$$\hat{\beta}^R = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Centered form

Transform the training data to centered form:

$$X = [x_1 - \bar{x}, ..., x_n - \bar{x}]^T$$
$$Y = [y_1 - \bar{y}, ..., y_n - \bar{y}]^T$$
$$\beta = [\beta_1, ..., \beta_p]^T, \beta_0 = \bar{y} - \beta^T \bar{x}$$

Lasso Regression (centered):

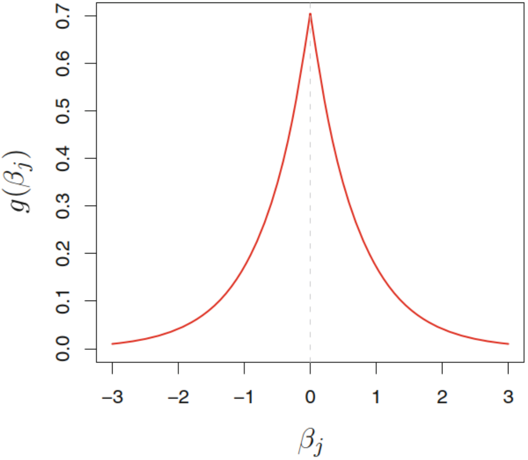
$$\hat{\beta}^L = \arg\min_{\beta} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_1$$

Another formulation (centered):

$$\hat{\beta}^L = \arg\min_{\beta} \|X\beta - Y\|_2^2, \ s.t. \ \|\beta\|_1 \leq t$$

Bayesian interpretation of lasso regression

Lasso regression is maximum a posteriori estimation (MAP) that assumes β has prior distribution of $\beta \sim \mathit{Laplace}(0, \frac{2\sigma^2}{\lambda} I)$.



Bayesian interpretation of lasso regression

Proof

The model underlying linear regression is $Y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In Bayesian statistics, β is a random vector that has a prior distribution. We can get:

$$\beta \sim Laplace(0, \frac{2\sigma^2}{\lambda} I),$$

$$i.e., f_\beta(\beta) = \frac{\lambda}{4\sigma^2} \exp(-\frac{\lambda \|\beta\|_1}{2\sigma^2})$$

$$Y|X, \beta \sim \mathcal{N}(X\beta, \sigma^2 I),$$

$$i.e., f_{Y|X,\beta}(Y|X, \beta) = (2\pi)^{-\frac{n}{2}} n^{-\frac{1}{2}} \exp(-\frac{1}{2}(Y - X\beta)^T \frac{1}{\sigma^2}(Y - X\beta))$$

Bayesian interpretation of lasso regression

The mode of posterior distribution (conditional distribution) of β (assume X is fixed):

$$\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} f_{\beta|X,Y}(\beta|X,Y) \\
&= \arg\max_{\beta} \frac{f_{X,\beta}(X,\beta) f_{Y|X,\beta}(Y|X,\beta)}{f_{X,Y}(X,Y)} \\
&= \arg\max_{\beta} \frac{f_X(X) f_{\beta|X}(\beta|X) f_{Y|X,\beta}(Y|X,\beta)}{f_{X,Y}(X,Y)} \\
&= \arg\max_{\beta} f_{\beta}(\beta) f_{Y|X,\beta}(Y|X,\beta)
\end{aligned}$$

Bayesian interpretation of lasso regression

$$= \arg\max_{\beta} \exp\Big(-\frac{\lambda\|\beta\|_1}{2\sigma^2} - \frac{1}{2}(Y - X\beta)^T \frac{1}{\sigma^2}(Y - X\beta)\Big)$$

$$= \arg\min_{\beta} \frac{\lambda\|\beta\|_1}{2\sigma^2} + \frac{1}{2}(Y - X\beta)^T \frac{1}{\sigma^2}(Y - X\beta)$$

$$= \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) + \lambda\|\beta\|_1$$

$$= \arg\min_{\beta} \|X\beta - Y\|_2^2 + \lambda\|\beta\|_1$$

□

An application

The lasso more or less shrinks all coefficients toward zero by a similar amount, and sufficiently small coefficients are shrunken all the way to zero.

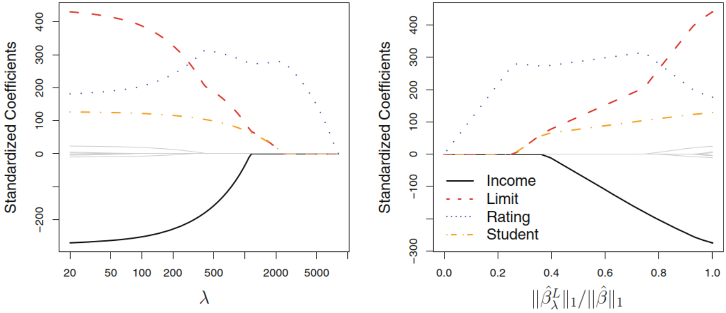


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Lasso estimates can shrink to 0 but ridge cannot

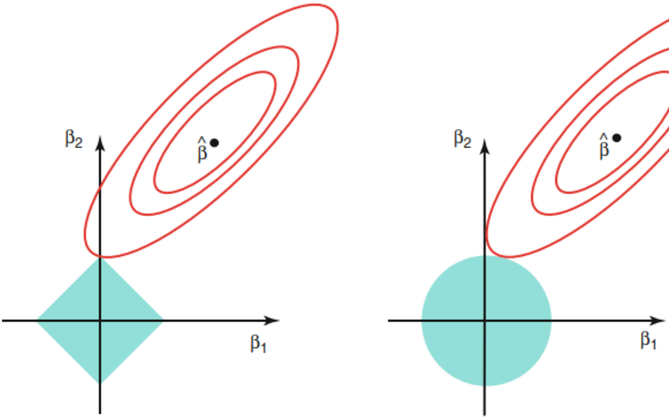


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \le s$ and $\beta_1^2 + \beta_2^2 \le s$, while the red ellipses are the contours of the RSS.*

Reference *link* of proof.

Comparison between Lasso and Ridge Regression

All 45 predictors were related to the response.

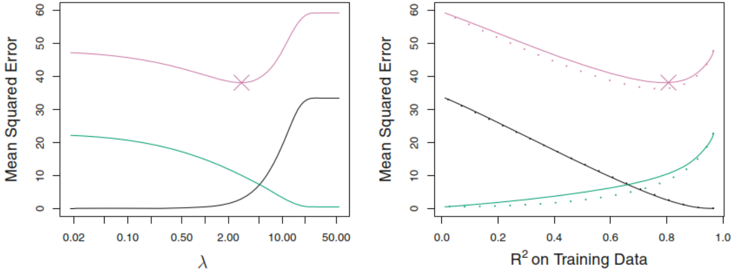


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Comparison between Lasso and Ridge Regression

Only two predictors are related to the response.

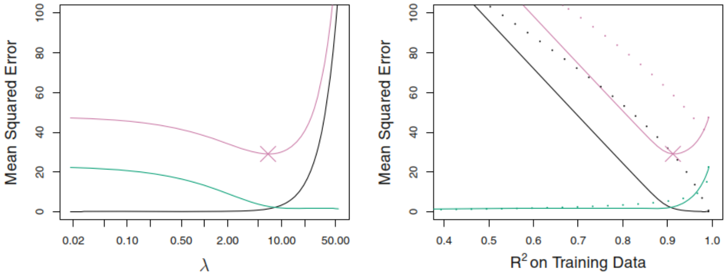


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

An example of cross-validation

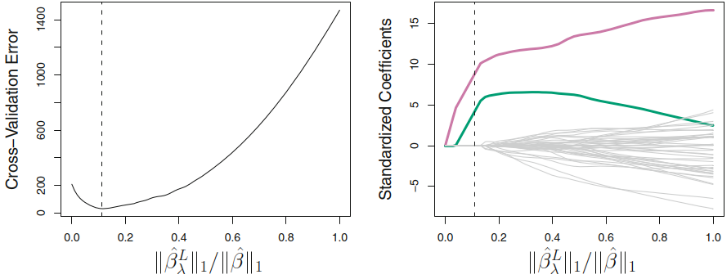


FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Outline

Overview

- Motivation

Subset Selection

- Best Subset Selection
- Stepwise Selection

Shrinkage Methods

- Ridge Regression
- Lasso Regression
- Parameter Selection

Dimension Reduction Methods

- The Framework
- Principle Components Regression
- Partial Least Squares

Considerations in High Dimensions

The framework

Let $Z_1, Z_2, ...Z_m$ represent $M < p$ linear combinations of original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for some constants $\phi_{1m}, \phi_{2m}, ..., \phi_{pm}$. We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, 2, ..., n,$$

using least squares.

What is PCA?

The target: find the low-dimension representation of X that explains the most variance of the covariance matrix of X.

The solution: eigenvalue decomposition or singular value decomposition.

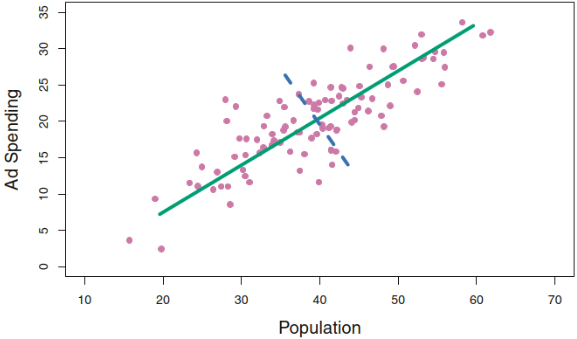


FIGURE 6.14. *The population size* (pop) *and ad spending* (ad) *for* 100 *different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Eigenvalues decomposition

Theorem

We can decompose any square, symmetric $n \times n$ matrix S as

$$S = U \Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T,$$

where $\Lambda = diag(\lambda_1, ..., \lambda_n)$, with $\lambda_1 \geq ... \geq \lambda_n$ the eigenvalues. $U = [u_1, ..., u_n]$ is a $n \times n$ orthonormal matrix ($U^T U = I$) that contains the eigenvectors.

Variance maximization problem

Let data matrix $X = [x_1, ..., x_n] \in \mathbb{R}^{p \times n}$, and assume X is centered, which is $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 0$. Then the sample covariance matrix is

$$Var(X) = C = \frac{1}{n-1}XX^T.$$

The projection of X on a direction vector u ($\|u\|_2^2 = 1$) can be represented as uu^TX, with u^TX the transformed coordinates of X on basis u.

Then the **variance maximization problem** is

$$\hat{u} = \arg\max_u Var(u^TX) = \arg\max_u u^TCu.$$

Variance maximization problem

Solution is $\hat{u} = u_1$, where u_1 is the first eigenvector of C.

Proof

Eigenvalue decomposition of C:

$$C = U \Lambda U^T.$$

Let $v = U^T u$, we have $\|v\|_2^2 = v^T v = u^T U U^T u = u^T u = 1$,

$$\max u^T C u = \max u^T U \Lambda U^T u = \max v^T \Lambda v$$
$$= \max \lambda_1 v_1^2 + ... + \lambda_p v_p^2, \ \ s.t. \|v\|_2^2 = 1.$$

Because $\lambda_1 \geq ... \geq \lambda_p$, we get $v^* = [1, 0, ..., 0]^T$, and corresponding $u^* = U v^* = [u_1, ..., u_p] v^* = u_1$.

□

An application of u_1

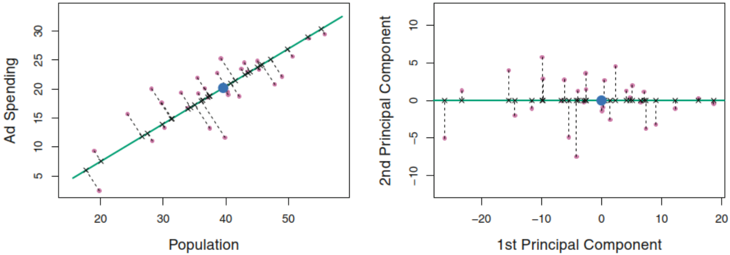


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\text{pop}}, \overline{\text{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

Iteration method to find k principle components

We say u_1 is the first principle component of X. The projection of X on u_1 is $u_1 u_1^T X$. The projection of X on a hyperplane orthogonal to u_1 (deflation, the unexplained information by u_1) is

$$X^{(2)} := X - u_1 u_1^T X = (I - u_1 u_1^T)X.$$

We get the second principle components by solving variance maximization problem on $X^{(2)}$:

$$\begin{aligned} C^{(2)} &= (I - u_1 u_1^T)C(I - u_1 u_1^T) \\ &= (I - u_1 u_1^T)\Big(\sum_{i=1}^{n} \lambda_i u_i u_i^T\Big)(I - u_1 u_1^T) \\ &= \sum_{i=2}^{n} \lambda_i u_u u_i^T. \end{aligned}$$

Iteration method to find k principle components

$$C^{(2)} = \sum_{i=2}^{n} \lambda_i u_u u_i^T$$

- The largest eigenvalue of $C^{(2)}$ is λ_2.
- The second principle component is u_2.

After k iterations, we get the k directions (principle components) of largest variance, which is

$$U^{(k)} = [u_1, u_2, ..., u_k].$$

Then the k-dimension representation of X is $U^{(k)T}X$, and the rank-k approximation of X is

$$U^{(k)}U^{(k)T}X.$$

Theorems of PCA

Theorem $\hat{X}^{(k)} = U^{(k)} U^{(k)T} X$ solves the problem

$$\arg\min_{\hat{X}^{(k)}} \|\hat{X}^{(k)} - X\|_F, \ s.t. Rank(\hat{X}^{(k)}) \leq k,$$

and we have

$$\begin{aligned}
&\frac{1}{n-1}\|\hat{X}^{(k)} - X\|_F^2 = \frac{1}{n-1}\|U^{(k)} U^{(k)T} X - X\|_F^2 \\
&= \frac{1}{n-1}\|(I - U^{(k)} U^{(k)T})X\|_F^2 \\
&= Trace((I - U^{(k)} U^{(k)T})C(I - U^{(k)} U^{(k)T})) \\
&= \lambda_{k+1} + ... + \lambda_p.
\end{aligned}$$

Principle components regression and an example

Settings in PCR replace X with $U^{(k)T}X$.

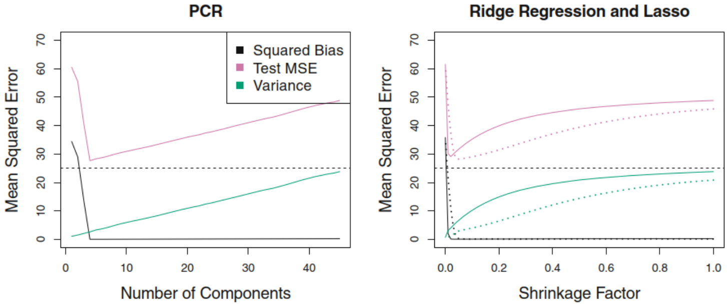


FIGURE 6.19. *PCR, ridge regression, and the lasso were applied to a simulated data set in which the first five principal components of X contain all the information about the response Y. In each panel, the irreducible error $Var(\epsilon)$ is shown as a horizontal dashed line.* Left: *Results for PCR.* Right: *Results for lasso (solid) and ridge regression (dotted). The x-axis displays the shrinkage factor of the coefficient estimates, defined as the ℓ_2 norm of the shrunken coefficient estimates divided by the ℓ_2 norm of the least squares estimate.*

Partial Least Squares

Partial least squares (PLS) is a supervised alternative to PCR.

- PLS computes the first direction Z_1 by setting each ϕ_{j1} equal to the coefficient from the simple linear regression of Y onto X_j.
- PLS then projects X on a hyperplane orthogonal to Z_1, which can be interpreted as the remaining information that has not been explained by the first PLS direction.
- PLS computes the second direction Z_2 using this orthogonalized data in exactly the same fashion as Z_1 was computed based on the original data.
- Iterate.

In practice it often performs no better than ridge regression or PCR (reduce bias but also increase variance).

Outline

Overview

- Motivation

Subset Selection

- Best Subset Selection
- Stepwise Selection

Shrinkage Methods

- Ridge Regression
- Lasso Regression
- Parameter Selection

Dimension Reduction Methods

- The Framework
- Principle Components Regression
- Partial Least Squares

Considerations in High Dimensions

Curse of dimension

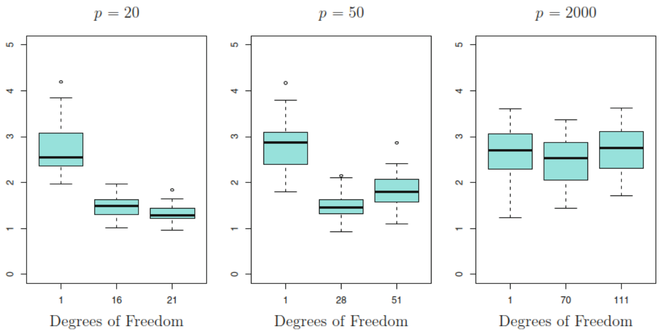


FIGURE 6.24. *The lasso was performed with $n = 100$ observations and three values of p, the number of features. Of the p features, 20 were associated with the response. The boxplots show the test MSEs that result using three different values of the tuning parameter λ in (6.7). For ease of interpretation, rather than reporting λ, the* degrees of freedom *are reported; for the lasso this turns out to be simply the number of estimated non-zero coefficients. When $p = 20$, the lowest test MSE was obtained with the smallest amount of regularization. When $p = 50$, the lowest test MSE was achieved when there is a substantial amount of regularization. When $p = 2{,}000$ the lasso performed poorly regardless of the amount of regularization, due to the fact that only 20 of the 2,000 features truly are associated with the outcome.*

Considerations

- Adding noise features that are not truly associated with the response will lead to a deterioration in the model.
- In the high-dimensional setting, the multicollinearity problem is extreme. The model we have identified is simply one of many possible models for predicting, and it must be validated on independent data sets.