

Lecture 15: Dimensionality reduction

Reading: Sections 6.3

STATS 202: Data mining and analysis

October 28, 2019

Shrinkage methods

Ridge regression:

$$\min_{\beta} \text{RSS}(\beta) + \lambda \sum_{j=1}^p \beta_j^2$$

The Lasso:

$$\min_{\beta} \text{RSS}(\beta) + \lambda \sum_{j=1}^p |\beta_j|$$

As we increase λ we increase bias, but reduce variance.

A very special case

Suppose $n = p$ and our matrix of predictors is $\mathbf{X} = I$.

Then, the objective function in Ridge regression can be simplified:

$$\sum_{j=1}^p (y_j - \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda \beta_j^2.$$

It is easy to show that

$$\hat{\beta}_j^R = \frac{y_j}{1 + \lambda}.$$

A very special case

Similar story for the Lasso; the objective function is:

$$\sum_{j=1}^p (y_j - \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

and we can minimize the terms that involve each β_j separately:

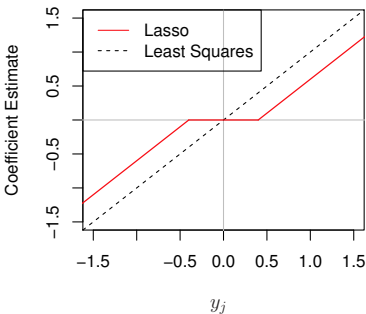
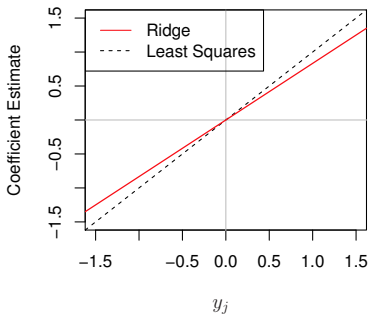
$$(y_j - \beta_j)^2 + \lambda |\beta_j|.$$

It is easy to show that

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| < \lambda/2. \end{cases}$$

Lasso and Ridge coefficients as a function of λ

Special case $\mathbf{X} = I$. Each coefficient $\hat{\beta}_j^R, \hat{\beta}_j^L$ depends only on y_j .



Regularization methods: strategies for trading off bias and variance by restricting model flexibility

- ▶ Variable selection:
 - ▶ Best subset selection
 - ▶ Forward and backward stepwise selection
- ▶ Shrinkage
 - ▶ Ridge regression
 - ▶ The Lasso (a form of variable selection)
- ▶ Dimensionality reduction:
 - ▶ **Idea:** Define a small set of M predictors which *summarize* the information in all p predictors.

Principal Components Regression

Recall: The loadings $\phi_{11}, \dots, \phi_{p1}$ for the first principal component define the directions of greatest variance in the space of variables.

Example. USArrests dataset.

Variable	UrbanPop	Murder	Assault	Rape
Loading	$\phi_{11} = 0.28$	$\phi_{21} = 0.54$	$\phi_{31} = 0.59$	$\phi_{41} = 0.54$

Interpretation: The first principal component measures the overall rate of crime.

Principal Components Regression

The loadings also define a new variable derived from the inputs:

$$Z_1 = \sum_{j=1}^p \phi_{j1} X_j$$

The first principal component scores z_{11}, \dots, z_{n1} are the values of this variable for each datapoint:

$$z_{i1} = \sum_{j=1}^p \phi_{j1} x_{ij}$$

Example. USArrests dataset.

Sample	Alabama	Alaska	...	Wyoming
Score	$z_{11} = 172$	$z_{21} = 196$...	$z_{n1} = 122$

Interpretation: The scores for the first principal component measure the overall rate of crime in each state.

Principal Components Regression

Idea:

- ▶ Replace the original predictors, X_1, X_2, \dots, X_p , with the first M score vectors Z_1, Z_2, \dots, Z_M .
- ▶ Perform least squares regression, to obtain coefficients $\theta_0, \theta_1, \dots, \theta_M$.

The model is:

$$\begin{aligned}y_i &= \theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2} + \dots + \theta_M z_{iM} \\&= \theta_0 + \theta_1 \sum_{j=1}^p \phi_{j1} x_{ij} + \theta_2 \sum_{j=1}^p \phi_{j2} x_{ij} + \dots + \theta_M \sum_{j=1}^p \phi_{jM} x_{ij} \\&= \theta_0 + \left[\sum_{m=1}^M \theta_m \phi_{1m} \right] x_{i1} + \dots + \left[\sum_{m=1}^M \theta_m \phi_{pm} \right] x_{ip}\end{aligned}$$

Principal Components Regression

Idea:

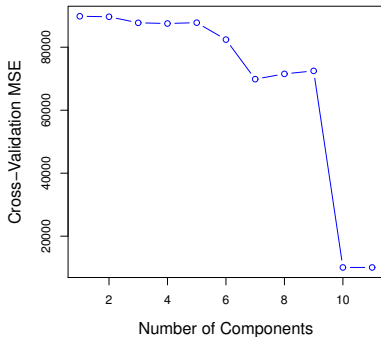
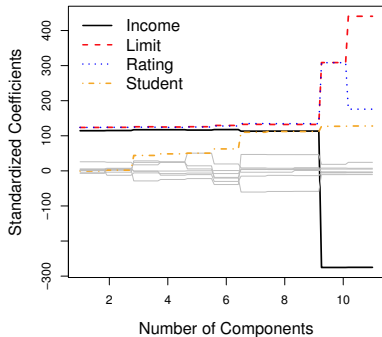
- ▶ Replace the original predictors, X_1, X_2, \dots, X_p , with the first M score vectors Z_1, Z_2, \dots, Z_M .
- ▶ Perform least squares regression, to obtain coefficients $\theta_0, \theta_1, \dots, \theta_M$.

Equivalent to a linear regression onto X_1, \dots, X_p , with coefficients constrained to be of the form

$$\beta_j = \sum_{m=1}^M \theta_m \phi_{jm}$$

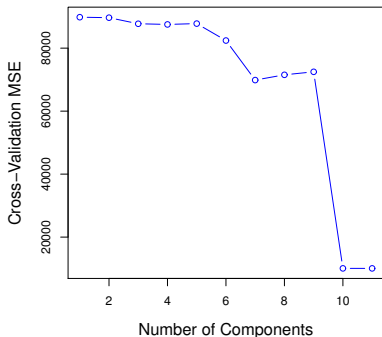
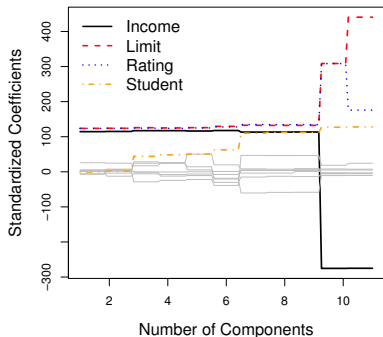
This constraint in the form of β_j introduces *bias*, but it can lower the *variance* of the model.

Application to the Credit dataset



- ▶ A model with 11 components is equivalent to least-squares regression
- ▶ Best error is achieved with 10 components (almost no dimensionality reduction)

Application to the Credit dataset



The left panel shows, for each M , the coefficients $\hat{\beta}_j^{PCR} = \sum_{m=1}^M \hat{\theta}_m \phi_{jm}$ for each *initial input variable* X_j .

Relationship between PCR and Ridge regression

Least squares regression: want to minimize

$$RSS = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T (y - \mathbf{X}\beta) = 0$$

$$\implies \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

Solve the singular value decomposition: $\mathbf{X} = U D^{1/2} \Phi^T$, where $D^{1/2} = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_p})$; then

$$(\mathbf{X}^T \mathbf{X})^{-1} = \Phi D^{-1} \Phi^T$$

where $D^{-1} = \text{diag}(1/d_1, 1/d_2, \dots, 1/d_p)$.

Relationship between PCR and Ridge regression

Ridge regression: want to minimize

$$RSS + \lambda \|\beta\|_2^2 = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta) + \lambda \beta^T \beta$$

$$\frac{\partial(RSS + \lambda \|\beta\|_2^2)}{\partial \beta} = -2\mathbf{X}^T(y - \mathbf{X}\beta) + 2\lambda\beta = 0$$

$$\implies \hat{\beta}_\lambda^R = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T y$$

Solve the singular value decomposition: $\mathbf{X} = U D^{1/2} \Phi^T$, where $D^{1/2} = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_p})$; then

$$(\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} = \Phi D_\lambda^{-1} \Phi^T$$

where $D_\lambda^{-1} = \text{diag}(1/(d_1 + \lambda), 1/(d_2 + \lambda), \dots, 1/(d_p + \lambda))$.

Relationship between PCR and Ridge regression

Predictions of least squares regression:

$$\hat{y} = \mathbf{X}\hat{\beta} = \sum_{j=1}^p u_j u_j^T y, \quad u_j \text{ is the } j\text{th column of } U$$

Predictions of Ridge regression:

$$\hat{y} = \mathbf{X}\hat{\beta}_{\lambda}^R = \sum_{j=1}^p u_j \frac{d_j}{d_j + \lambda} u_j^T y$$

The projection of y onto a principal component is shrunk toward zero. The smaller the principal component, the larger the shrinkage.

Predictions of PCR:

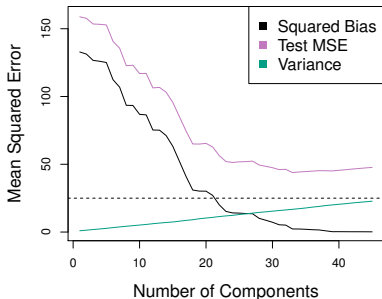
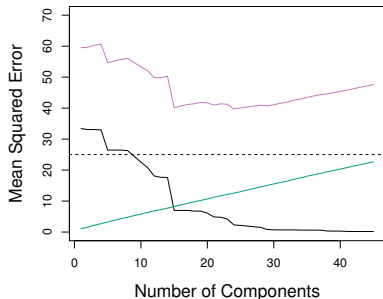
$$\hat{y} = \mathbf{X}\hat{\beta}^{\text{PC}} = \sum_{j=1}^p u_j \mathbf{1}(j \leq M) u_j^T y$$

The projections onto small principal components are shrunk to zero.

Simulated example

In each case $n = 50$, $p = 45$.

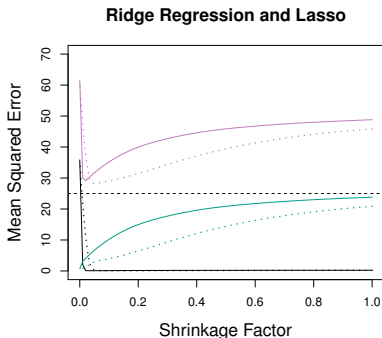
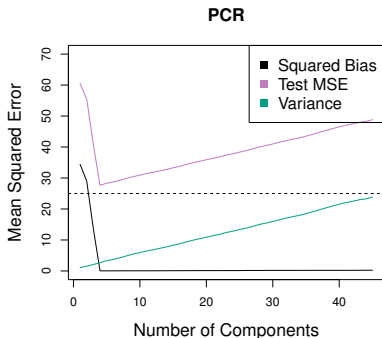
- ▶ **Left:** Response is a function of all the predictors.
- ▶ **Right:** Response is a function of 2 predictors (good for Lasso).



Simulated example

Again, $n = 50$, $p = 45$.

The response is a function of the first 5 principal components.



(Dotted line is ridge regression)

Partial least squares

- ▶ Principal components regression derives Z_1, \dots, Z_M using *only* the predictors X_1, \dots, X_p .
- ▶ In partial least squares, we will use the response Y as well.

Algorithm:

1. Define $Z_1 = \sum_{j=1}^p \phi_{j1} X_j$, where ϕ_{j1} is the coefficient of a simple linear regression of Y onto X_j .
2. Let $X_j^{(2)}$ be the residual of regressing X_j onto Z_1 .
3. Define $Z_2 = \sum_{j=1}^p \phi_{j2} X_j^{(2)}$, where ϕ_{j2} is the coefficient of a simple linear regression of Y onto $X_j^{(2)}$.
4. Let $X_j^{(3)}$ be the residual of regressing $X_j^{(2)}$ onto Z_2 .
5. ...

Partial least squares

- ▶ To form each component, PLS give more weight to the variables / residuals that are more correlated with the response.
- ▶ To form each residual, PLS removes the information explained by the previous components.
- ▶ Compared to PCR, partial least squares has less bias and more variance (a stronger tendency to overfit).