Lecture 16: Dimensionality reduction

Reading: Sections 6.3, 6.4

STATS 202: Data mining and analysis

Sergio Bacallado
October 31, 2013

Regularization methods

- 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 $\phi_{11}, \ldots, \phi_{p1}$ for the first principal component define the directions of greatest variance in the space of variables.

Example. `VSArrests` dataset.

Variable	`VrbanPop`	`Murder`	`Assault`	`Rape`
Loading	$\phi_{11} = -0.15$	$\phi_{21} = 0.7$	$\phi_{31} = 0.6$	$\phi_{41} = 0.6$

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

Principal Components Regression

Recall: The scores $z_{11}, \ldots, z_{n1}$ for the first principal component define the deviation of the samples along this direction.

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

Example. `VSArrests` dataset.

Sample	`Alabama`	`Arkansas`	...	`Wyoming`
Score	$z_{11} = 0.3$	$z_{21} = -0.5$	...	$z_{n1} = -1$

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

The model is:

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

Principal Components Regression

Idea:

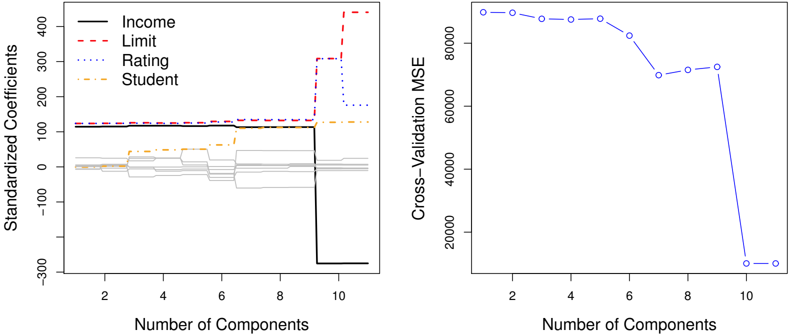
- Replace the original predictors, $X_1, X_2, \ldots, X_p$, with the first M score vectors $Z_1, Z_2, \ldots, Z_M$.
- Perform least squares regression, to obtain coefficients $\theta_0, \theta_1, \ldots, \theta_M$.

Equivalent to a linear regression onto $X_1, \ldots, X_p$, with coefficients:

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

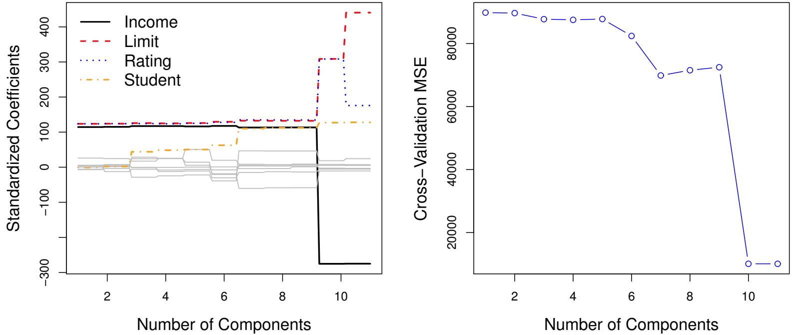
This constraint in the form of β_j introduces *bias*, but it can lower the *variance* of the model.

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 the coefficients β_j estimated for each M.

The coefficients shrink as we decrease M!

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: $\mathbf{X} = UD^{1/2}V^T$, where $D^{1/2} = \text{diag}(\sqrt{d_1}, \ldots, \sqrt{d_p})$; then

$$(\mathbf{X}^T \mathbf{X})^{-1} = VD^{-1}V^T$$

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

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: $\mathbf{X} = UD^{1/2}V^T$, where $D^{1/2} = \text{diag}(\sqrt{d_1}, \ldots, \sqrt{d_p})$; then

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

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

Relationship between PCR and Ridge regression

Predictions of least squares regression:

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

Predictions of Ridge regression:

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

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

Predictions of PCR:

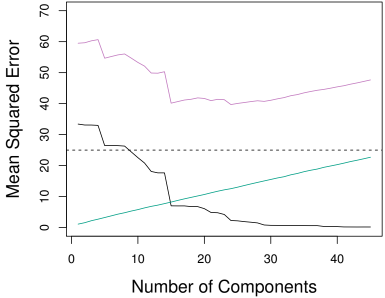
$$\hat{y} = \mathbf{X}\hat{\beta}^{\mathsf{PC}} = \sum_{j=1}^{p} u_j \mathbf{1}(j \leq M) u_j^T y$$

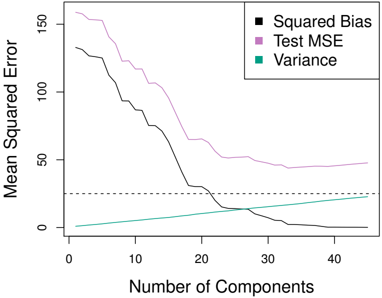
The projections onto small principal components are shrunk to zero.

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).

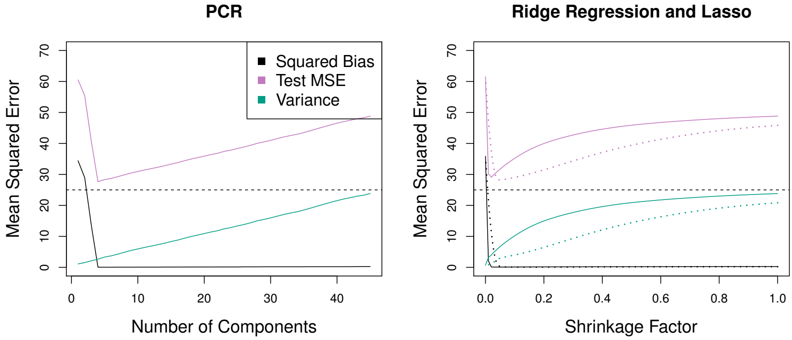




Simulated example

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

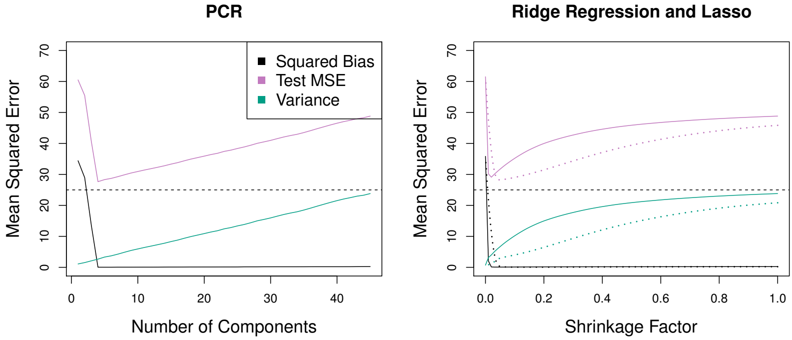
The response is a function of the first 5 principal components.



Simulated example

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

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



Partial least squares

- Principal components regression derives $Z_1, \ldots, Z_M$ using *only* the predictors $X_1, \ldots, X_p$.
- 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

- At each step, we try to find the linear combination of predictors that is most highly correlated to the response.
- After each step, we transform the predictors such that they are *independent* from the linear combination chosen.
- Compared to PCR, partial least squares has less bias and more variance (a stronger tendency to overfit).