Lecture 16: Linear Model Selection and Regularization (Textbook 6.3 and 6.4)

Dimension Reduction Methods

- The methods that we have discussed so far in this chapter have involved fitting linear regression models, via least squares or a shrunken approach, using the original predictors, $X_1, X_2, ..., X_p$.
- We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as **dimension reduction methods**.

Dimension Reduction Methods

- Let $Z_1, Z_2, ..., Z_M$ represent $M < p$ linear combinations of our original p predictors,
$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j,$$
for some constants $\phi_{1m}, ..., \phi_{pm}$, and $m = 1, ..., M$.

- We then fit the linear regression
$$Y = \theta_0 + \sum_{m=1}^{M} \theta_m Z_m + \epsilon$$
using the ordinary least squares.

- In the previous model, the regression coefficients are given by $\theta_0, ..., \theta_M$. If the constants $\phi_{1m}, ..., \phi_{pm}$ are chosen wisely, then such dimension reduction approaches can often outperform OLS regression.

- The term dimension reduction comes from the fact that the dimension of the problem has been reduced from $p + 1$ to $M + 1$.

Dimension Reduction Methods

All dimension reduction methods work in two steps.

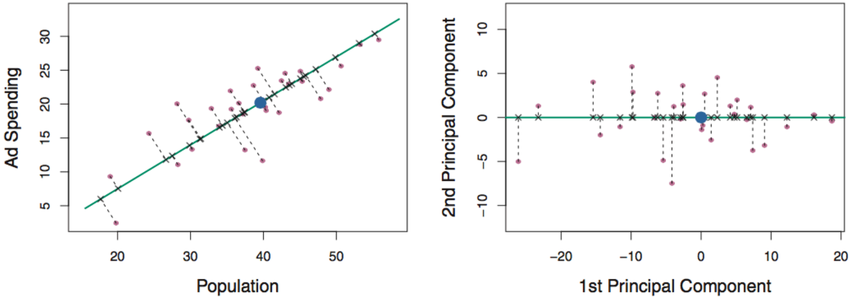
- First, the transformed predictors $Z_1, Z_2, ..., Z_M$ are obtained.
- Second, the model is fit using these M predictors.
- However, the construction of $Z_1, Z_2, ..., Z_M$ can be achieved in different ways.
- We will consider the **principal components** method.

Principal Components Regression

- The idea is to apply principal components analysis (PCA) to the $n \times p$ features matrix $\mathbf{X}$. This has nothing to do with the outcome!
- The first principal component is that (normalized) linear combination of the variables with the largest variance.
- The second principal component has largest variance, subject to being uncorrelated with the first.
- And so on. (We will discuss more details in Chapter 10).
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

Advertising Data

Consider two features: population size (pop) and ad spending for a particular company (ad).



The first principal component direction is shown in green. It is the dimension along which the data vary the most. We get

$$z_{i1} = 0.839 \times (pop_i - \bar{pop}) + 0.544 \times (ad_i - \bar{ad}).$$

The principal components regression (PCR) just fits a simple linear model for y_i versus z_{i1} using least squares.

Principal Components Regression

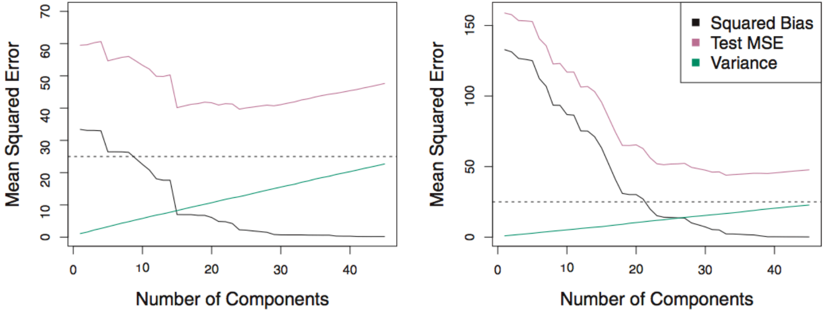
- The **principal components regression** (PCR) approach involves constructing the first M principal components, $Z_1, ..., Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares.

- We hope that a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.

- As we explained before, it is one of the dimension reduction method, which reduces fitting a linear model with $p + 1$ predictors to $M + 1$ predictors.

- This attains better bias-variance trade-off.

Two Simulated Examples



- How to select the number of components M? Cross-validation!
- PCR does not perform feature selection.
- Similar to the ridge and lasso regression, we generally recommend standardizing each predictor.

High-Dimensional Data

- **Definition**: High-dimensional data occurs when the number of features p exceeds the number of observations n.
- **Challenges**: Traditional models (e.g., linear regression) may overfit and perform poorly due to increased variance.
- **Examples**: Genomics (e.g., SNP data) and marketing (e.g., search terms).

Regularization Techniques Overview

- **Purpose**: Regularization reduces model flexibility, managing overfitting.
- **Techniques**:
  - **Ridge Regression**: Adds a penalty for large coefficients.
  - **Lasso Regression**: Encourages sparsity by setting some coefficients to zero.
  - **Principal Component Regression (PCR)**: Reduces dimensionality by using principal components.

Ridge Regression

- **Concept**: Adds an L^2 penalty to the regression to shrink coefficients.
- **Equation**: $\text{RSS} + \lambda \sum \beta_j^2$
- **Pros**: Useful when predictors are highly correlated.
- **Limitations**: Does not produce sparse models.

Lasso Regression

- **Concept**: Adds an L^1 penalty, setting some coefficients to zero.
- **Equation**: $\text{RSS} + \lambda \sum |\beta_j|$
- **Pros**: Produces sparse models, good for feature selection.

Principal Component Regression (PCR)

- **Concept**: Reduces predictors by using principal components.
- **Advantages**: Reduces dimensionality and mitigates multicollinearity.
- **Trade-Off**: Not all components correlate with the response.

Model Evaluation in High Dimensions

- **Traditional Metrics**: R^2, adjusted R^2, and p-values are unreliable.
- **Why ?**: Measures of model fit on the training data
- **Better Metrics**: MSE or R^2 on an independent test set