Linear Model Selection and Regularization II: Dimension Reduction

Rebecca C. Steorts, Duke University

STA 325, Chapter 6 ISL

Agenda

Dimension Reduction Techniques

Dimension Reduction Techniques

The methods so far in Chapter 6 have controlled variance in two different ways, either by using a subset of the original variables, or by shrinking their coefficients toward zero.

All of these methods are defined using the original predictors $X_1, \ldots, X_p$.

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 Techniques

Let $Z_1, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors.

That is,

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

for some constants $\psi_{1m}, \ldots, \psi_{pm}$, $m = 1, \ldots, M$.

We can then fit the regression model

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

using least squares.

Dimension Reduction

In equation 2, the regression coefficients are given be $\theta_1, \ldots, \theta_M$.

If the constants $\psi_{1m}, \ldots, \psi_{pm}$ are chosen wisely, then such dimension reduction approaches can often outperform least squares regression.

Dimension Reduction

The term dimension reduction comes from the fact that this approach reduces the problem of estimating the $p + 1$ coefficents

$$\beta_0, \beta_1, \ldots, \beta_p$$

to the simpler problem of estimating the $M + 1$ coefficients

$$\theta_1, \ldots, \theta_M$$

, where $M < p$.

The dimension problem has been reduced from $p + 1$ to $M + 1$.

Dimension Reduction

From equation 1, note:

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \psi_{jm} x_{ij} \tag{3}$$

$$= \sum_{m=1}^{M} \sum_{j=1}^{p} \theta_m \psi_{jm} x_{ij} \tag{4}$$

$$= \sum_{j=1}^{p} \beta_j x_{ij} \tag{5}$$

where

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

Equation 3 can be thought of as a special case of the original linear regression model.

Dimension Reduction

Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form 3.

This constraint on the form of the coefficients has the potential to bias the coefficient estimates.

However, in situations where p is large relative to n, selecting a value of $M << p$ can significantly reduce the variance of the fitted coefficients.

If $M = p$, and all the Z_m are linearly independent, then 3 poses no constraints. In this case, no dimension reduction occurs, and so fitting 2 is equivalent to performing least squares on the original p predictors.

Dimension Reduction Algorithm

All dimension reduction methods work in two steps.

1. The transformed predictors $Z_1, \ldots, Z_M$ are obtained.
2. The model is fit using these M predictors. The choice $Z_1, \ldots, Z_M$ can be acheived in different ways.

We consider two approaches: **principal components** and **partial least squares**.

Principal Components Regression

Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables.

PCA is discussed in greater detail as a tool for unsupervised learning in Chapter 10.

Here we describe its use as a dimension reduction technique for regression.

An Overview of Principal Components Analysis

PCA is a technique for reducing the dimension of a $n \times p$ data matrix X.

The first principal component direction of the data is that along which the observations vary the most.

For instance, consider Figure 6.14, which shows population size (pop) in tens of thousands of people, and ad spending for a particular company (ad) in thousands of dollars, for 100 cities. The green solid line represents the first principal component direction of the data. We can see by eye that this is the direction along which there is the greatest variability in the data. That is, if we projected the 100 observations onto this line (as shown in the left-hand panel of Figure 6.15), then the resulting projected observations would have the largest possible variance; projecting the observations onto any other line would yield projected observations with lower variance. Projecting a point onto a line simply involves finding the location on the line which is closest to the point.