3 Dimension Reduction Methods

So far we have controlled variance (of estimates) in two ways:

(1) Use a subset of original variable
- best subset, forward/backward selection, lasso

(2) shrinking coefficients towards zero
- ridge regression, lasso

These methods all defined using original predictor variables $X_1, \ldots, X_p$.

We now explore a class of approaches that

(1) transform predictors

(2) then fit least squares regression model using transformed variables

We refer to these techniques as *dimension reduction* methods.

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

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

for constants ϕ_{jm}, $m = 1, \ldots, M$.

(2) Fit linear regression model using least squares

$$y_i = \theta_0 + \boxed{\sum_{m=1}^{M} \theta_m z_{im}} + \varepsilon_i, \quad i = 1, \ldots, n$$

regression coefficients.

If ϕ_{jm} chosen well, this can outperform least squares.

The term *dimension reduction* comes from the fact that this approach reduces the problem of estimating $p + 1$ coefficients to the problem of estimating $M + 1$ coefficients where $M < p$.

$\beta_0, \beta_1, \ldots, \beta_p$ — $\theta_0, \theta_1, \ldots, \theta_M$

z = linear combination of x's

Note: $$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \left[\sum_{m=1}^{M} \theta_m \phi_{jm} \right] x_{ij}$$

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

Dimension reduction serves to constrain β_j, since now they must take a particular form.

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

⇒ special case of original linear regression problem

with β_j constrained → can bias our coefficient estimates

→ if $p > n$ (or $p \approx n$), selecting $M << p$ can reduce variance.

All dimension reduction methods work in two steps.

1. transformed predictors are obtained.
2. model is fit using M transformed predictors.

The selection of ϕ_{jm}'s can be done in multiple ways.

↳ We will talk about 2 ways.

First way to choose ϕ_{jm}'s $\Rightarrow$ $Z_1, \dots, Z_M$.

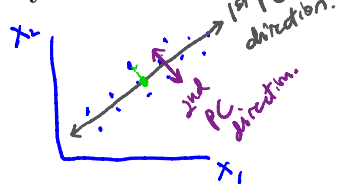
3.1 Principle Component Regression

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

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

① Finding PC directions
② projecting data onto PC directions

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



$p=2$

The 1st principal components are obtained by projecting the data onto the 1st PC direction.

↳ A point is projected onto a line by finding the point on the line closest to the original point.

1st PC direction = direction along which data varies most = line closest to all observations (least squares line).

out of every possible linear combination of X_1 and X_2 such that $\phi_{11}^2 + \phi_{21}^2 = 1$,

choose linear combinations such that

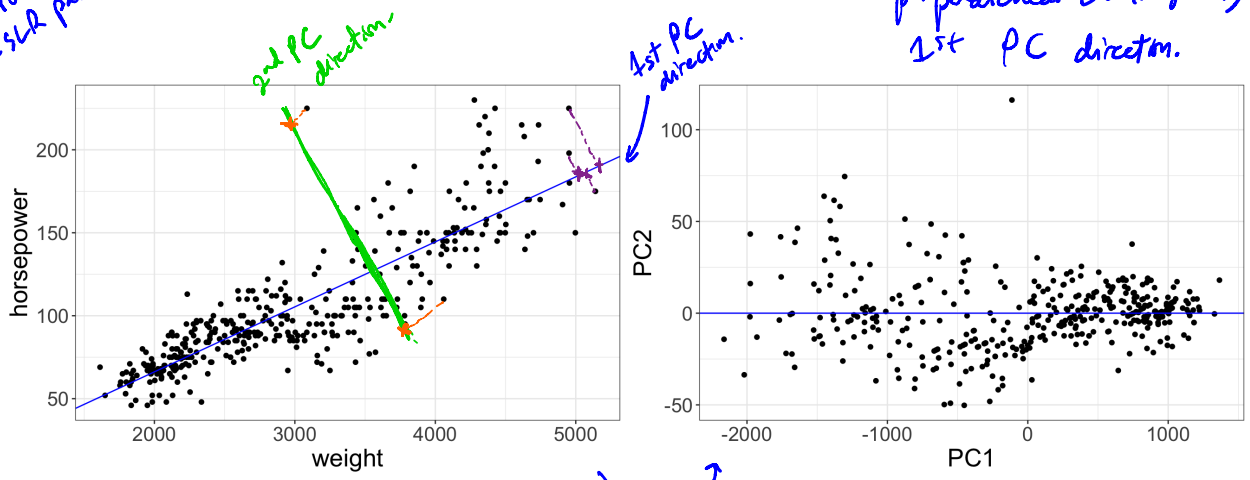
$$\text{Var}\left[\phi_{11}(X_1 - \bar{X}_1) + \phi_{21}(X_2 - \bar{X}_2)\right] \text{ is maximized.}$$

$$\Rightarrow z_{i1} = \phi_{11}(x_{1i} - \bar{x}_1) + \phi_{21}(x_{2i} - \bar{x}_2) \text{ for } i = 1, \dots, n$$ are "principal component scores" (1st.)

We can construct up to p principal components, where the 2nd principal component is a linear combination of the variables that are uncorrelated to the first principal component and has the largest variance subject to this constraint.

$\Rightarrow$ 2nd PC direction is perpendicular (orthogonal) to 1st PC direction.

From Auto data in ISLR package.



The 1st PC contains the most information $\rightarrow$ p^{th} PC contains the least.

The Principal Components Regression approach (PCR) involves

1. Construct first M principal components $Z_1, \ldots, Z_M$
2. fit linear regression model predicting Y using $Z_1, \ldots, Z_M$ by least squares.

Key idea: often a small # of principal components will suffice to explain most of variability in the data X, as well as the relationship with the response.

In other words, we assume that the directions in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with Y.

This is not guaranteed to be true, but often works well in practice.

If this assumption holds, fitting PCR will lead to better results than fitting least squares on $X_1, \ldots, X_p$.

(We can mitigate overfitting).

How to choose M, the number of components?

Note: PCR is not feature selection!

3.2 Partial Least Squares

The PCR approach involved identifying linear combinations that best represent the predictors $X_1, \ldots, X_p$.

Consequently, PCR suffers from a drawback

Alternatively, *partial least squares (PLS)* is a supervised version.

Roughly speaking, the PLS approach attempts to find directions that help explain both the reponse and the predictors.

The first PLS direction is computed,

To identify the second PLS direction,

As with PCR, the number of partial least squares directions is chosen as a tuning parameter.

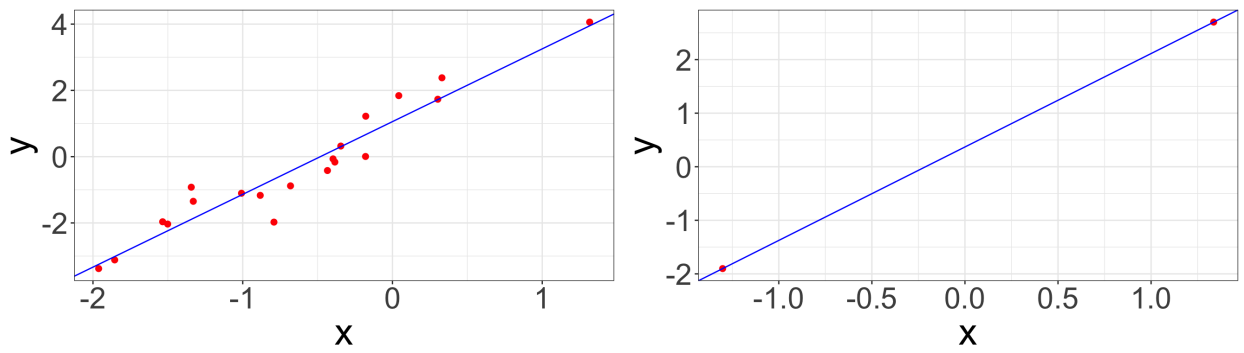
4 Considerations in High Dimensions

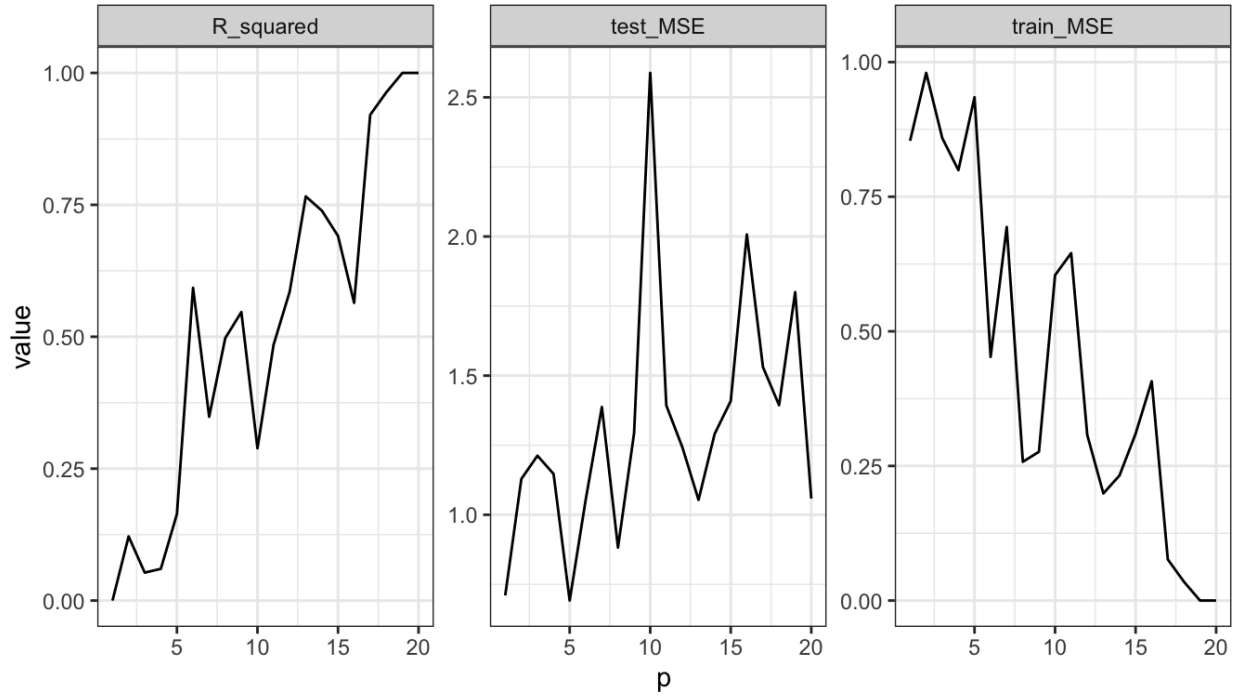
Most traditional statistical techniques for regression and classification are intendend for the low-dimensional setting.

In the past 25 years, new technologies have changed the way that data are collected in many fields. It is not commonplace to collect an almost unlimited number of feature measurements.

Data sets containing more features than observations are often referred to as *high-dimensional*.

What can go wrong in high dimensions?





Many of the methds that we've seen for fitting *less flexible* models work well in the high-dimension setting.

1.

2.

3.

When we perform the lasso, ridge regression, or other regression procedures in the high-dimensional setting, we must be careful how we report our results.