3 Dimension Reduction Methods

So far we have controlled variance in two ways:

(1) Used a subset of original variables

- best subset, forward/backward selection, lasso

(2) shrinking coefficients towards zero

- ridge regression, lasso.

These methods all defined using original predictor variables $x_1, \dots, x_p$.

We now explore a class of approaches that

(1) transform the predictors

(2) then perform least squares using transformed variables.

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

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

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

for constants $\phi_{1m}, \dots, \phi_{pm}$ $\quad m = 1, \dots, M$

(2) Fit the linear regression model using least squares

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

regression coefficients.

If $\{\phi_{jm}\}_{\substack{j=1,\dots,p \\ m=1,\dots,M}}$ 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 $M < p$.

$\beta_0, \beta_1, \dots, \beta_p$

$\theta_0, \theta_1, \dots, \theta_M$

from least squares

have to come from somewhere else (hopefully we've picked well).

NOTE:

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

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

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

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

$\Rightarrow$ special case of original linear regression problem (with β_j constrained)

↳ if can bias coefficient estimate

↳ if $p > n$ (or $p \approx n$) selecting $M \ll p$ can reduce variance.

All dimension reduction methods work in two steps.

1. transformed predictors are obtained (get $\{\phi_{jm}\}_{\substack{j=1,\dots,p \\ m=1,\dots,M}}$)
2. linear Model is fit using M transformed predictors from 1.

selection of $\{\phi_{jm}\}_{\substack{j=1,\dots,p \\ m=1,\dots,M}}$ can be done in multiple ways.

We will talk about 2.

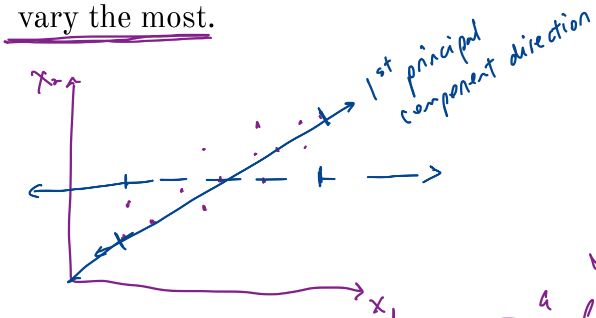
One way to choose $Z_1, \dots, Z_M$

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.

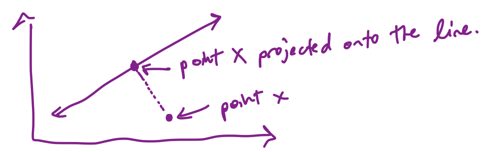
PCA is an unsupervised approach for reducing the dimension of an $n \times p$ data matrix X.

The *first principal component* directions of the data is that along which the obervations vary the most.



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

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



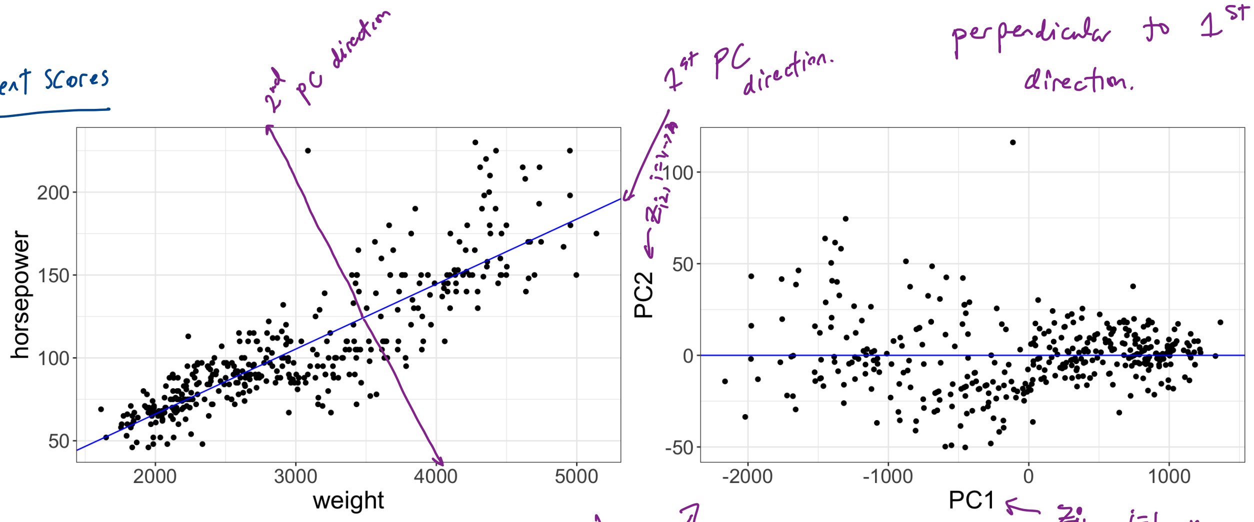
out of every possible linear combination of X_1 and X_2 such that $\phi_{11}^2 + \phi_{21}^2 = 1$, choose ϕ_{i1} s.t. $Var\left(\phi_{11}(X_1 - \bar{X}_1) + \phi_{21}(X_2 - \bar{X}_2)\right)$ is maximized.

$Z_{i1} = \phi_{11}(x_{i1} - \bar{x}_1) + \phi_{21}(x_{i2} - \bar{x}_2)$ for $i = 1, \dots, n$ are called 1st principal component scores

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 to 1st PC direction.

1st PC direction = dimension along which data vary the most.



projected onto principal component directions.

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 $Z_1, \ldots, Z_M$ a choice we are making.
   score vectors.
2. Fit a linear regression model w/ $Z_1, \ldots, Z_M$ as predictors using least squares.

Key idea: Often a small # of PC suffice to explain most of the variability in the data, as well as the relationship w/ predictor.

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 model on $X_1, \ldots, X_p$ because we can mitigate overfitting.

How to choose M, the number of components?

M can be thought of as a tuning parameter
$\Rightarrow$ use CV method to choose!

as $M \uparrow p$, PCR $\rightarrow$ least squares. $\Rightarrow$ bias $\downarrow$ but variance $\uparrow$, will see bias-variance tradeoff in the form of a U-shape in the test MSE.

Note: PCR is not feature selection!

each of the M principal components used in the linear regression is a linear combination of all p of the original predictors!

$\Rightarrow$ while PCR works well to reduce variance, it doesn't give us a sparse model.

PCR more like ridge regression than the lasso. (not going to help w/ interpretation)

NOTE: recommended standardizing predictors $X_1, \ldots, X_p$ to each have st. dev. $= 1$ before getting the PCs.

3.2 Partial Least Squares

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

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.

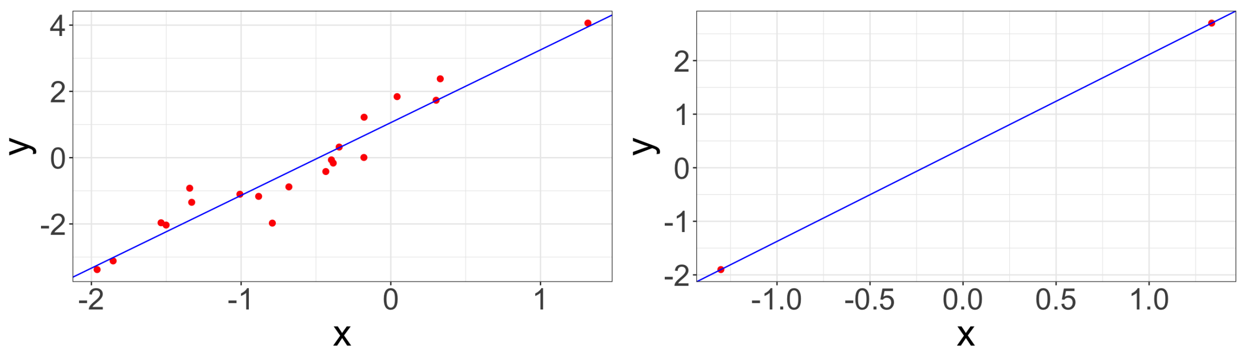
4 Considerations in High Dimensions

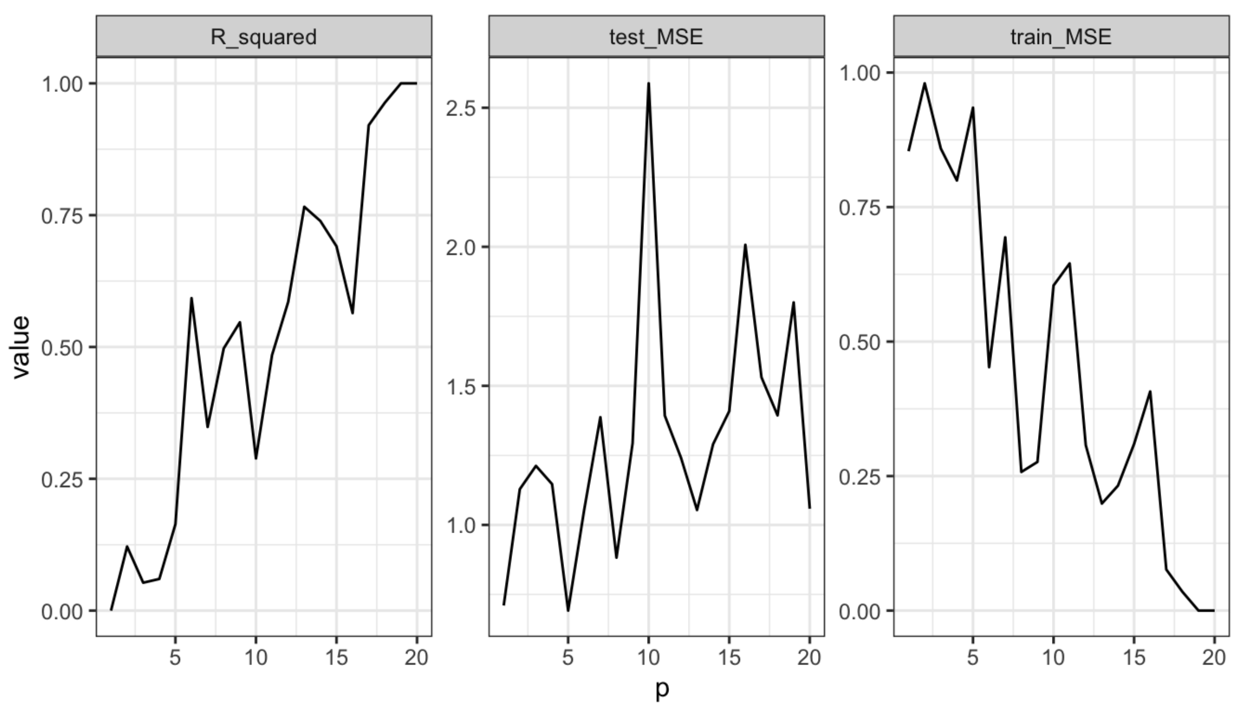
Most traditional statistical techniques for regression and classification are intendend for the low-dimensional setting.

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.