Ch 6.3: Dimension Reduction

Lecture 13 - CMSE 381

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Weds, March 11, 2024

Announcements

Last time:

- Shrinkage: Ridge and Lasso

This lecture:

- PCA / PCR
- PLS

Section 1

Last time

Shrinkage

Find β to minimize:

Least Squares:

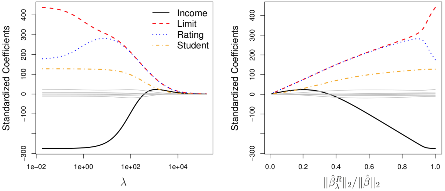
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

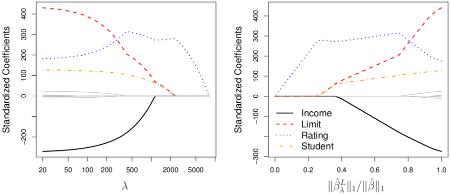
Ridge:

$$RSS + \sum_{j=1}^{p} \beta_j^2$$

The Lasso:

$$RSS + \sum_{j=1}^{p} |\beta_j|$$





Also, find best choice of λ or s using CV

Section 2

Dimension Reduction

Linear transformation of predictors

Original Predictors:

$$X_1, \cdots, X_p$$

New Predictors:

$$Z_1, \cdots, Z_M$$

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

- The goal is for $M \ll p$
- Need to figure out good φ to do this

Two dimensions down to 1.....

- $$Z_1 = X_1 + 3 \cdot X_2$$
- Matrix version:
$$(Z_1) = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$
- Example data:

X_1	X_2	Z_1
0	1	4
3	4	17
1	1	4
-1	0	-1

an example with just three dimensions down to two

- $\varphi = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \\ 0 & 1/2 \end{pmatrix}$

$$\begin{cases} Z_1 = X_1 + 0 \cdot X_2 + 0 \cdot X_3 \\ Z_2 = 0 \cdot X_1 + \frac{1}{2} X_2 + \frac{1}{2} X_3 \end{cases}$$

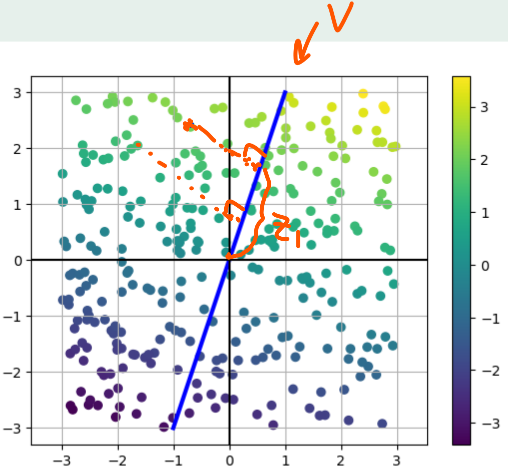
matrix form:

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \equiv z = \overbrace{(X_1, X_2, X_3)}^{X} \varphi$$

$$z_1 = X \varphi_1 \qquad z_2 = X \varphi_2$$

$$\begin{pmatrix}1\\3\end{pmatrix} \xrightarrow{\text{normalize}} \begin{pmatrix}\frac{1}{\sqrt{10}}\\ \frac{3}{\sqrt{10}}\end{pmatrix} \equiv V$$

- Get the φ to unit column norm
- Example: $Z_1 = \frac{1}{\sqrt{10}}X_1 + \frac{3}{\sqrt{10}}X_2$
- The Z value is the distance from the projection of (X_1, X_2) onto the vector $(1, 3)$ to the origin.



$X = \begin{pmatrix}x_1\\x_2\end{pmatrix}$, $\quad P_V X = (X \cdot V) \cdot V$

↓

length of projection

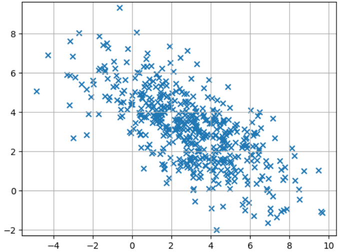
||

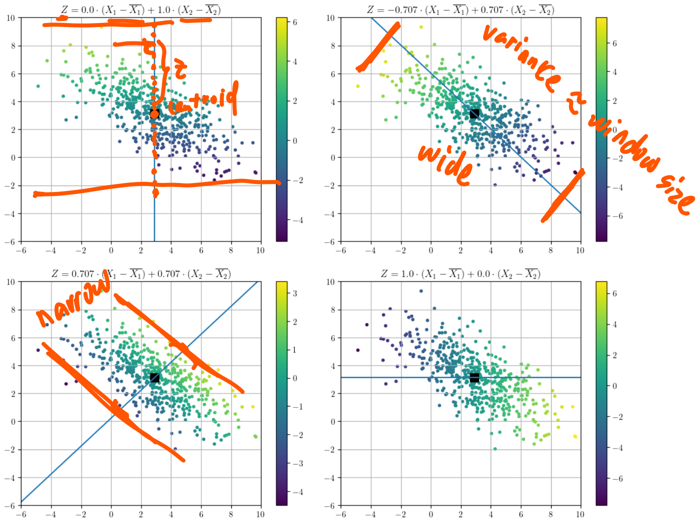
Z_1

Projection onto a line

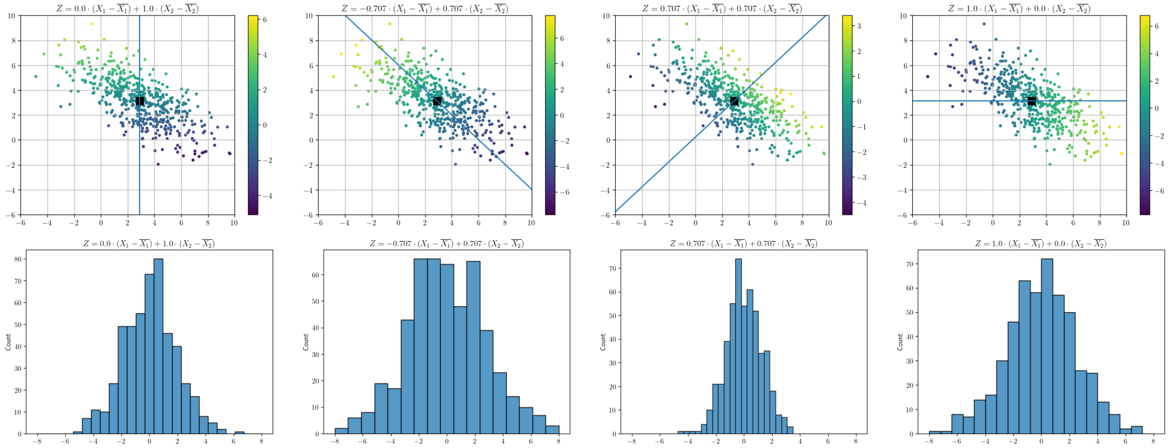
https://www.desmos.com/calculator/cih7wy8oyg

Different projections





Histograms of Z values



The goal

- Find good φ's for some $M \ll p$
- Fit regression model on Z_i's using least squares

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i$$

- in this notation, θs replace βs

- Dimension reduction comes from the fact that we're fitting models on a smaller number of variables
- Next two subsection: ways to find φs: PCA and partial least squares

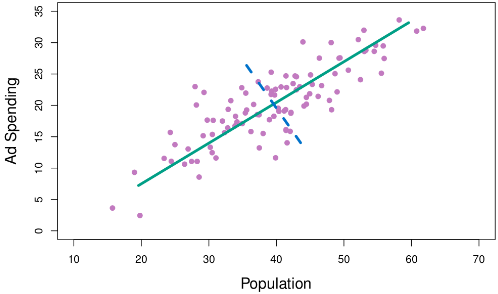
$M \ll p$

$$y = \theta_0 + \sum_{m=1}^{p} \theta_m X_m + \varepsilon \Rightarrow y = \theta_0 + \sum_{m=1}^{M} \theta_m Z_m + \varepsilon$$

Section 3

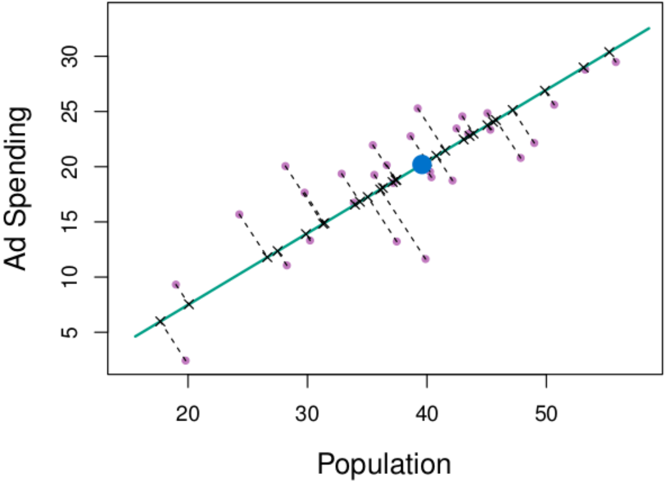
Review of PCA

An example dataset



- Population size in tens of thousands of people
- Ad spending for a company in thousands of dollars
- 100 data points
- By eye: green line is the direction of most variability, calle the principal direction
- Meaning if we project observations onto the line, then the projected observations would have the largest possible variance
-

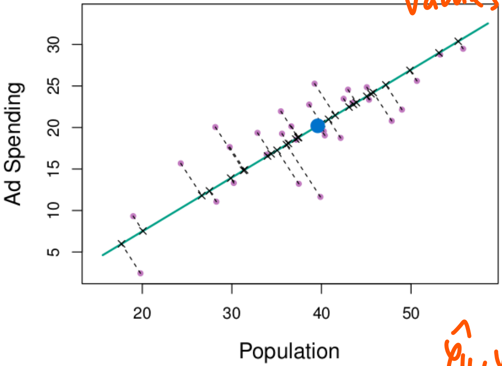
Projection onto first PC



$$Z_1 = 0.839 \cdot (\texttt{pop} - \overline{\texttt{pop}}) + 0.544 \cdot (\texttt{ad} - \overline{\texttt{ad}})$$

$v_1 \cdots v_m$: PC

values in $z_1 \cdots z_m$: PC scores



- Of every linear combo of pop and ad, this one has the highest variance

PC 1

$v_1 = (\varphi_{1,1}, \varphi_{2,1})$

- Maximizes $\mathrm{Var}(\varphi_{1,1}(pop - \overline{pop}) + \varphi_{2,1}(ad - \overline{ad}))$
- Require $\varphi_{11}^2 + \varphi_{21}^2 = 1$, otherwise, just make φ super big to blow up variance

$$\hat\varphi_{11}, \hat\varphi_{21} = \arg\max_{\varphi_{11}, \varphi_{21}} \mathrm{Var}\Big(\underbrace{\varphi_{11}(X_1 - \bar X_1) + \varphi_{21}(X_2 - \bar X_2)}_{z_1}\Big) \quad \text{s.t. } \varphi_{11}^2 + \varphi_{21}^2 = 1$$

$$\Downarrow$$

$$\to \hat z_1 = \hat\varphi_{11}(X_1 - \bar X_1) + \hat\varphi_{21}(X_2 - \bar X_2)$$

The other principal components

- Z_2 is linear combo of variables uncorrelated to Z_1
- Find the one that explains the most variance
- Result: requires Z_2 to be the direction that explains the most variance among all directions that are uncorrelated with Z_1
- Requires Z_k to be the direction that explains the most variance among all directions that are orthogonal or perpendicular to $Z_1, Z_2, \ldots, Z_{k-1}$

$$PC2: V_2 = (\varphi_{12}, \varphi_{22})$$

$$\hat{z}_2 \leftarrow \max \operatorname{Var}\left(\varphi_{12}(X_1 - \bar{X}_1) + \varphi_{22}(X_2 - \bar{X}_2)\right)$$

$$\text{s.t. } \varphi_{12}^2 + \varphi_{22}^2 = 1$$

$$\text{and } \varphi_{11}\varphi_{12} + \varphi_{21}\cdot\varphi_{22} = 0$$

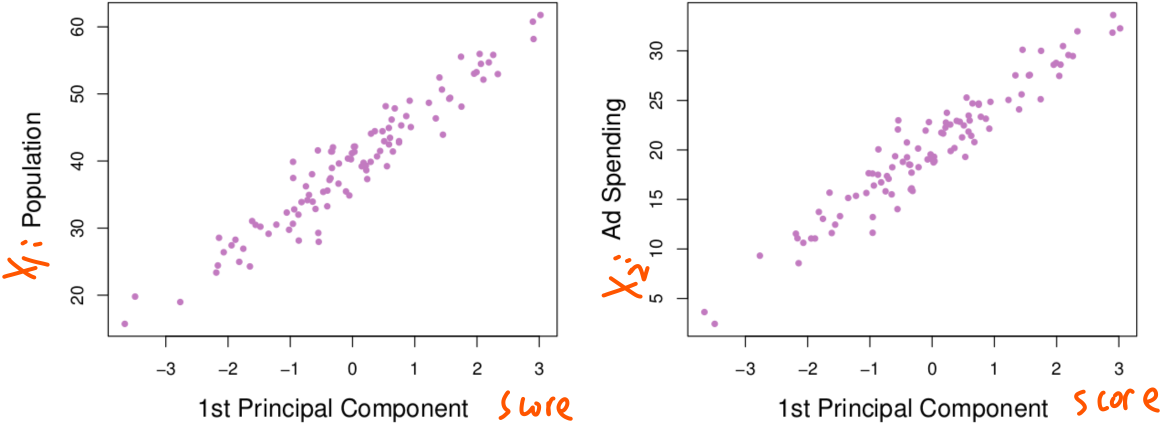
$$\vdots$$

$$\hat{z}_k \leftarrow \max \operatorname{Var}(z_k)$$

$$\text{s.t. } \|V_k\|^2 = 1$$

$$V_k \perp V_j,\ j = 1, \ldots k-1$$

Principal component scores



$$z_{i1} = 0.839 \cdot (\mathtt{pop}_i - \overline{\mathtt{pop}}) + 0.544 \cdot (\mathtt{ad}_i - \overline{\mathtt{ad}})$$

Computing all PC directions together

Let X be the data matrix whose $(i, k)th$ entry is the kth predictor's value for observation i. We do the following to find the PCs.

- Centralize the data $X^c = X - \frac{1}{n}X11^T$ $\leftarrow$ X − X.mean( )
- Compute the SVD of X^c, $X^c = U\Sigma V^T$
- the kth PC is v_k which is the kth column of V.
- the PC scores of Z_k is $X^c v_k$.
- the PC score z_{ik} is $X_i^c v_k$, where X_i^c is the ith row of X^c.

Section 4

Principal Components Regression

So you've found your PCA coefficients

Now what?

- Do linear regression on $Z_1, \cdots, Z_M$
- Book calls this "principal components regression "
- Linear model is $y = \theta_0 + \theta_1 Z_1 + \cdots + \theta_M Z_M$

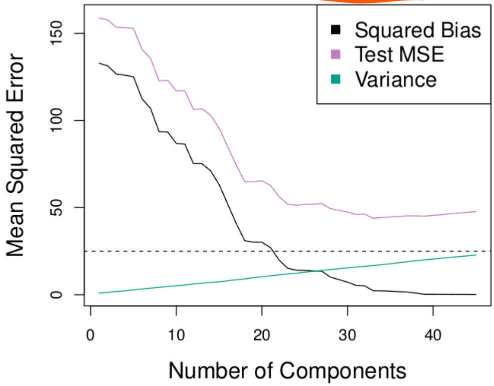
What are we assuming?

- The directions in which $X_1, \cdots, X_p$ have the most variation are the directions that are associated with Y
- If assumption holds, then fitting on Z_i better than fitting on X_i since fewer variables lessens chance of overfitting

Doing better

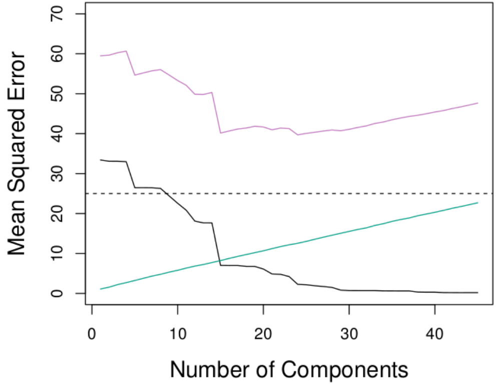
Example with simulated data: $n = 50$ observations of $p = 45$ predictors
Y is a function of 2 predictors

- Right side is usual least squares estimate
- Typical U shape for Test MSE
- PCR with good choice of number of components improvement over plain least squares



Doing better

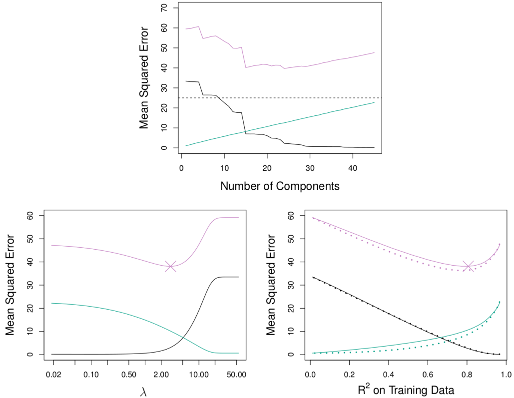
Example with simulated data: $n = 50$ observations of $p = 45$ predictors
Y is a function of all predictors



- Right side is usual least squares estimate
- Even better improvement using PCR than previous
- note the change in y-axis values

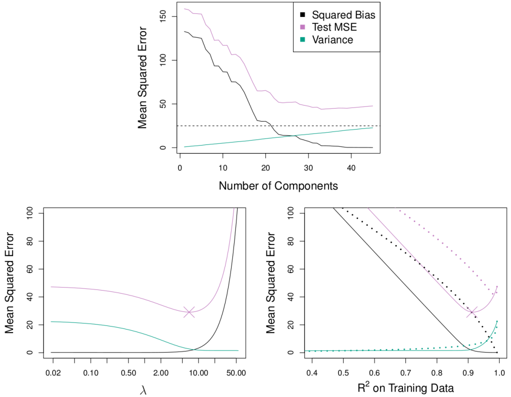
Comparison to results on shrinkage

Y is a function of all predictors



Shrinkage does better

Y is a function of 2 predictors

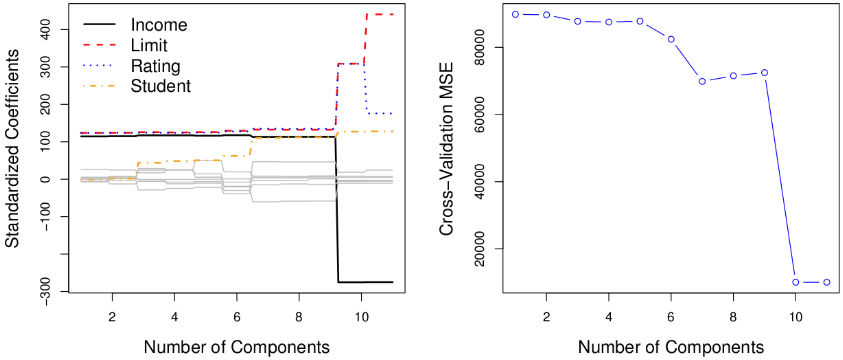


What is missing?

Why the PCR result is worse than Lasso on the above example?

- A critical assumption for PCR to work is that, the directions in which $X_1, \cdots, X_p$ have the most variation are the directions that are associated with Y
- If assumption holds, then fitting on Z_i better than fitting on X_i since fewer variables lessens chance of overfitting

Picking M



- Credit data set
- Major drop in MSE at 10 compoentns, so barely a savings with the original 11 variables

Properties of PCR

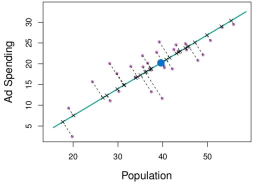
- Note better in previous example because lots of PCs needed
- It does better when info contained in the first few PCs
- Number of PCs to use can be picked with CV
- Standardization needed to make sure big data values don't skew the results

- Not a feature selection method because we don't get a subset of features, we get a collection of linear combos of features

Summary

PCA

- Unsupervised dimensionality reduction
- Choose component Z_1 in the direction of most variance using only X_i's information
- Choose Z_2 and beyond by the same method after "getting rid" of info in the directions already explained



PCR

- Do PCA on input data
- Do Linear Regression on chosen number of PCs.
- Warning: Lose interpretability of the coefficients.

Section 5

Partial Least Squares (PLS)

Supervised alternative

PCR: dimension reduction is Non-supervised

- No input from the Y values before learning the PCs.
- No guarantee that directions that explain Xs help to predict the Ys

Partial Least Squares (PLS):

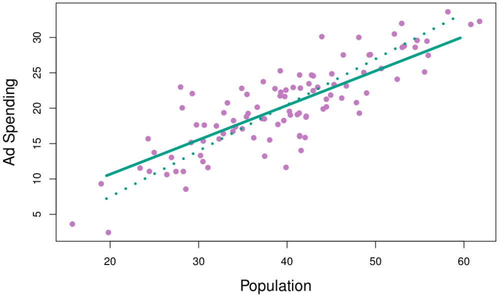
- Identify new features $Z_1, \cdots, Z_M$ linear combos of original where quality measure involves Y
- Fit linear model using least squares on these M features

First direction Z_1 for Partial Least Squares (PLS)

- Set φ_{j1} equal to the coefficient from simple linear regression of Y onto X_j
  This means we do p linear regressions, not that we take the multivariable linear regression coefficients
- The first direction is

$$Z_1 = \sum_{j=1}^{p} \underline{\varphi_{j1}} X_j$$

- PLS places highest weight on variables most strongly related to the response
- In example, PLS chooses direction less change in ad direction than pop relative to PCA



Ex. Prediction of Y =Sales (not shown) on X_1 =Population and X_2 =Ad Spending

- Solid green: First PLS direction
- Dashed: First PC direction

Second (and more) PLS directions

Let's use $Coef(y \sim X_i)$ to denote the $\hat{\beta}_1$ obtained from doing Linear reg $y \sim X_i$
and use $Res(y \sim X_i)$ to denote the residue $y - \hat{y}$ obtained from the same Linear reg.

- Regress each variable on Z_1 and take residuals *This is the remaining info not explained by first PLS direction*
- Compute Z_2 using *orthogonalized data* same as for Z_1
- Number M can be picked using CV after standardizing predictors

PLS Algorithm:

Iter 1. $Z_1 = \sum_{m=1}^{p} \varphi_{m1} X_m$ where $\varphi_{m1} = Coef(y \sim X_m)$

Iter 2. $Z_2 = \sum_{m=1}^{p} \varphi_{m2} X_m^{(1)}$ where $\varphi_{m2} = Coef(y \sim X_m^{(1)})$ and $X_m^{(1)} = Res(X_m \sim Z_1)$

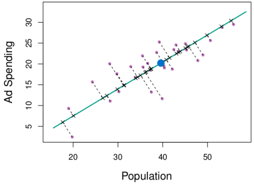
⋮

Iter K. $Z_K = \sum_{m=1}^{p} \varphi_{mK} X_m^{(K-1)}$, where $\varphi_{mK} = Coef(y \sim X_m^{(K-1)})$, $X_m^{K-1} = Res(X_m^{(K-2)} \sim Z_{K-1})$

PCR vs PLS

PCR

- Unsupervised dimensionality reduction + linear regression
- Choose component Z_1 in the direction of most variance using only X_i's information
- Choose Z_2 and beyond by the same method after "getting rid" of info in the directions already explained



PLS

- Supervised dimensionality reduction
- Choose component Z_1 by using simple regression coefficients of each X_i onto Y
- Choose Z_2 and beyond by the same method after "getting rid" of info in the directions already explained
- better suited than PCR for prediction tasks.