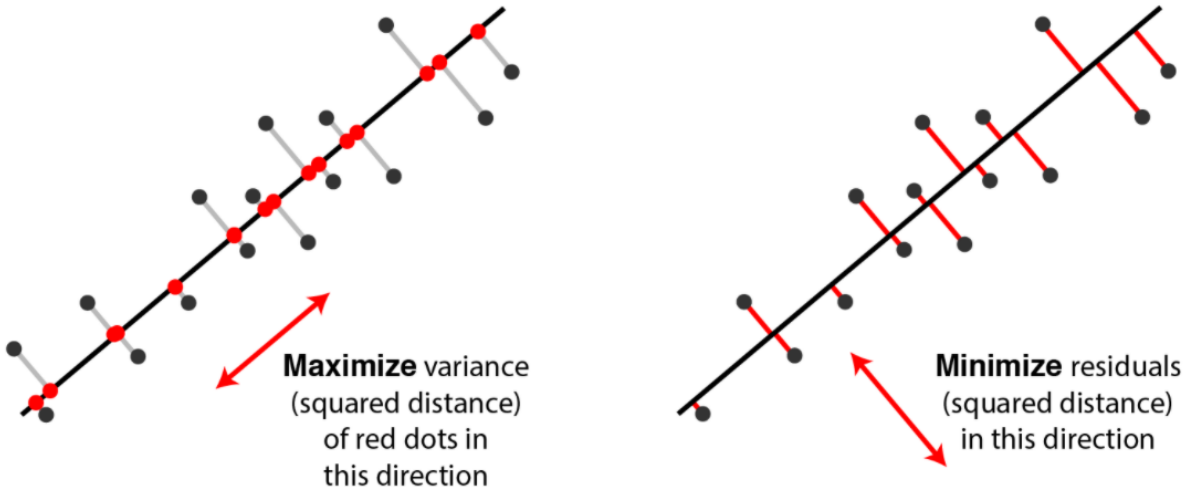
Stat 435 Intro to Statistical Machine Learning

Week 2: PCA and other non-linear methods

Richard Li

May 15, 2017

Motivation of PCA



Two equivalent views of principal component analysis.

Geometric view of PCA

If we think of PCA as maximizing variances,

PCA **rotates** the data set so as to align the directions in which it is spread out the most with the principal axes.

Remember in your previous HW, you found it difficult to visualize your classifier when there are more than two predictors, and you scratched your head trying to find two predictors that give you the best visual separation ~~(or you just randomly plotted something)~~.

A nice 3D visualization of the rotation:
`http://setosa.io/ev/principal-component-analysis/`

Connection of the two views of PCA

Maximize $\text{var}(\{z_{11}, z_{21}, \cdots, z_{n1}\})$ where

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \phi_{31} x_{i3} + \cdots + \phi_{p1} x_{ip}$$

s.t $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

} 1st principle component

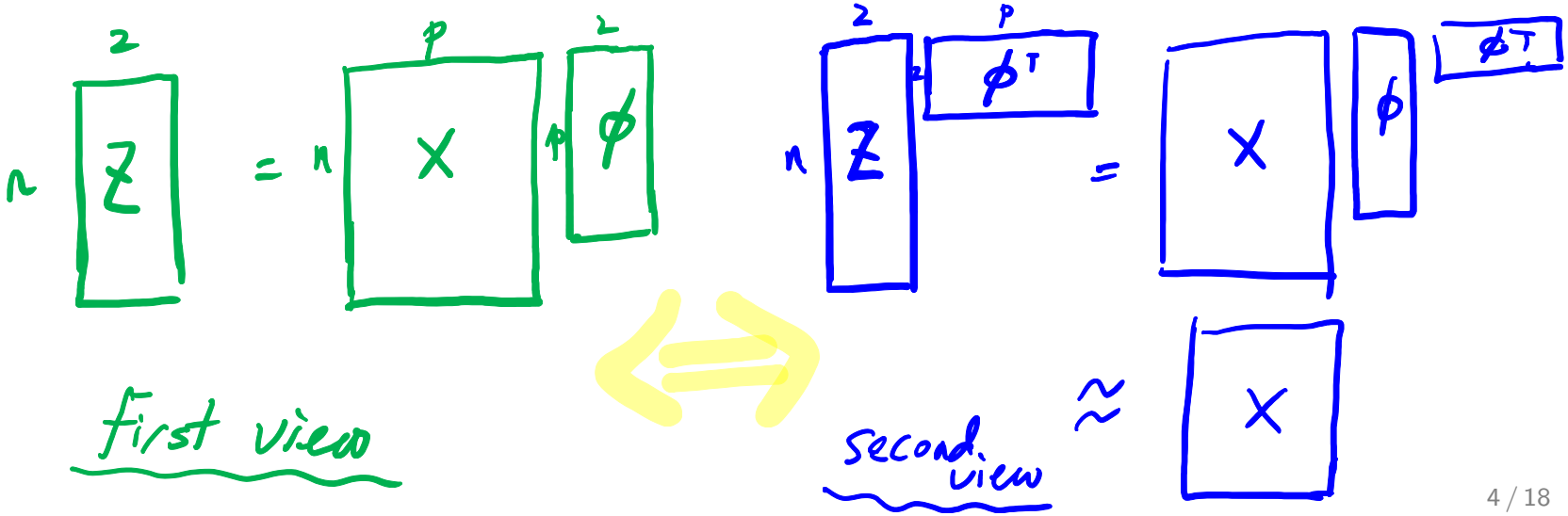
Maximize $\text{var}(\{z_{12}, z_{22}, \cdots, z_{n2}\})$ where

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \phi_{32} x_{i3} + \cdots + \phi_{p2} x_{ip}$$

s.t $\sum_{j=1}^{p} \phi_{j2}^2 = 1$, $\sum_{j=1}^{p} \phi_{j1} \phi_{j2} = 0$

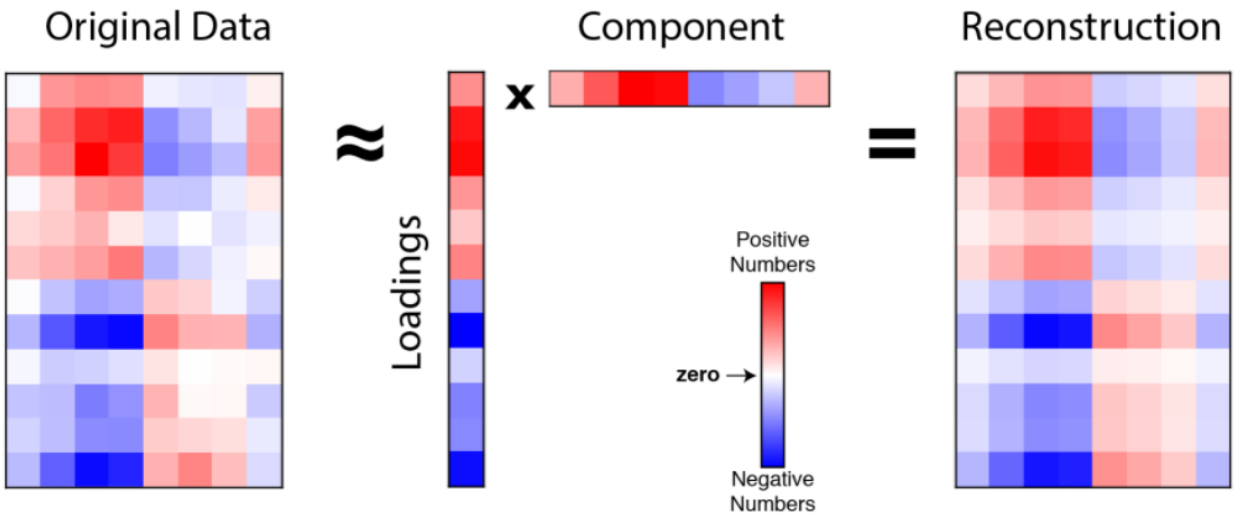
} 2nd principle component

.....

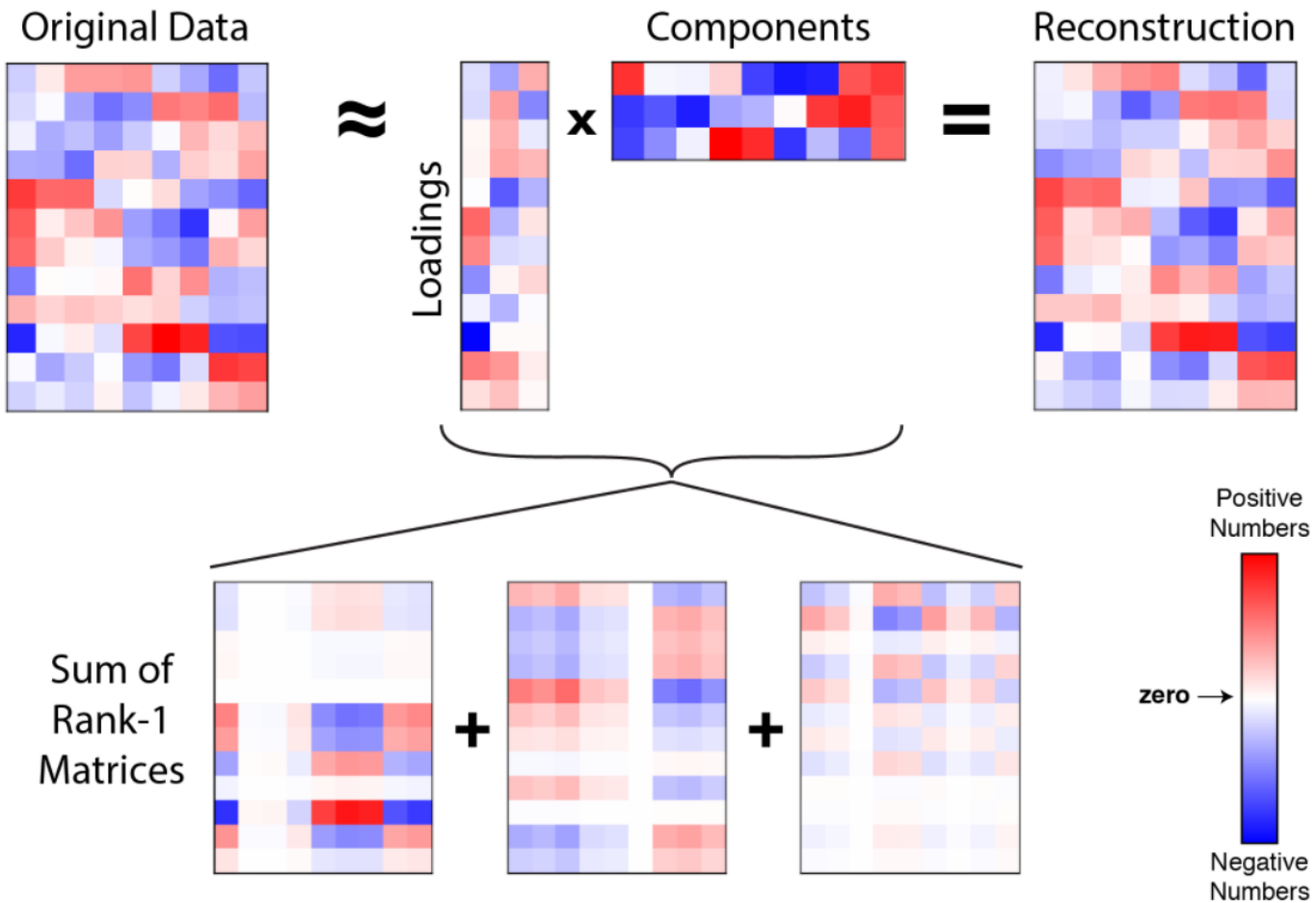


Alternative view of PCA

(10.2.2 of ISLR) If we think of PCA as minimizing residuals,



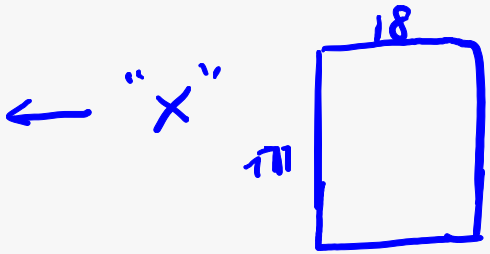
Alternative view of PCA

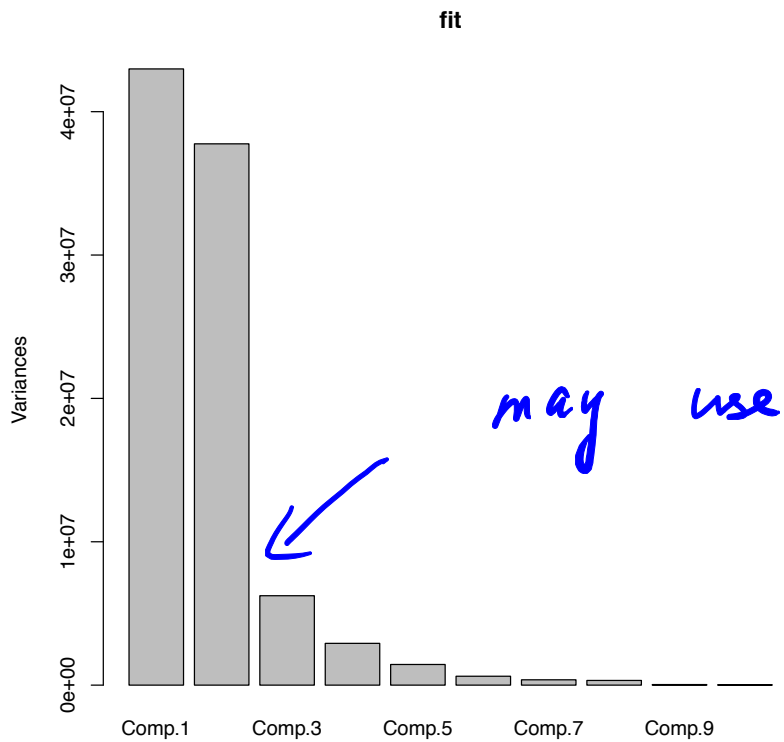


This sequence matrices explain less and less variation in the data.
What happens when there are p of them? Try the codes in 10.4 of ISLR, see if it is true.

PCA using R

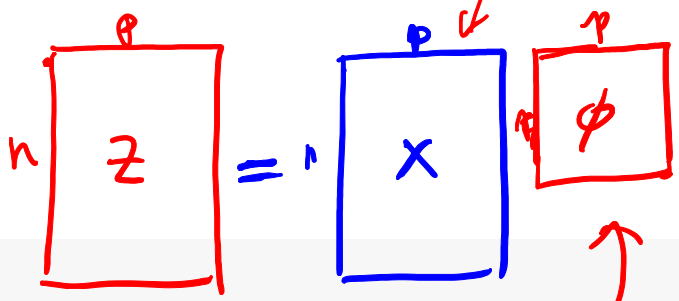
```
library(ISLR)
data(College)
dim(College)

## [1] 777  18

College[, 1] <- as.numeric(College[, 1] == "Yes")
fit <- princomp(College)
plot(fit)
```





may use only the first 2 components

Confirm the reconstruction



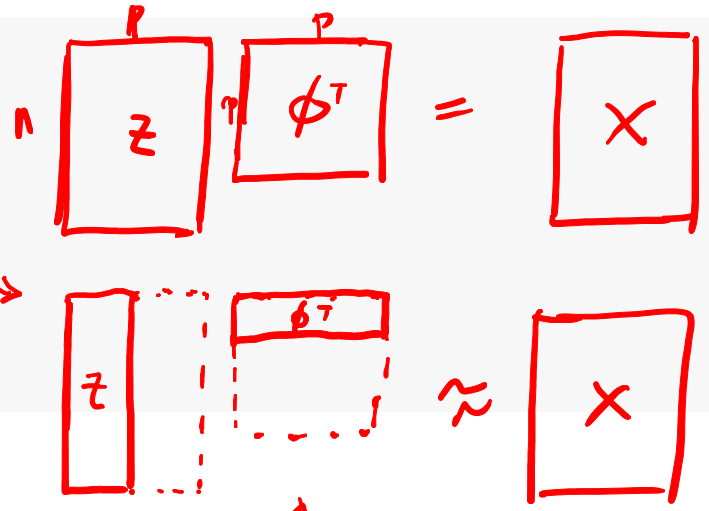

```
dim(fit$loading)

## [1] 18 18

dim(fit$scores)

## [1] 777  18
```

center the data

```
College.center <- as.matrix(College)
College.center <- apply(College.center, 2, function(x){x - mean(x)})
transform <- College.center %*% fit$loading
mean(abs(fit$scores - transform))

## [1] 1.052119e-14
```

Illustration of the reconstruction

```
Z1 <- College.center %*% fit$loading[, 1]
approx1 <- Z1 %*% t(fit$loading[, 1])
mean(abs(approx1 - College.center))

## [1] 627.799

plot(approx1, College.center)
abline(0, 1, col="red")
```



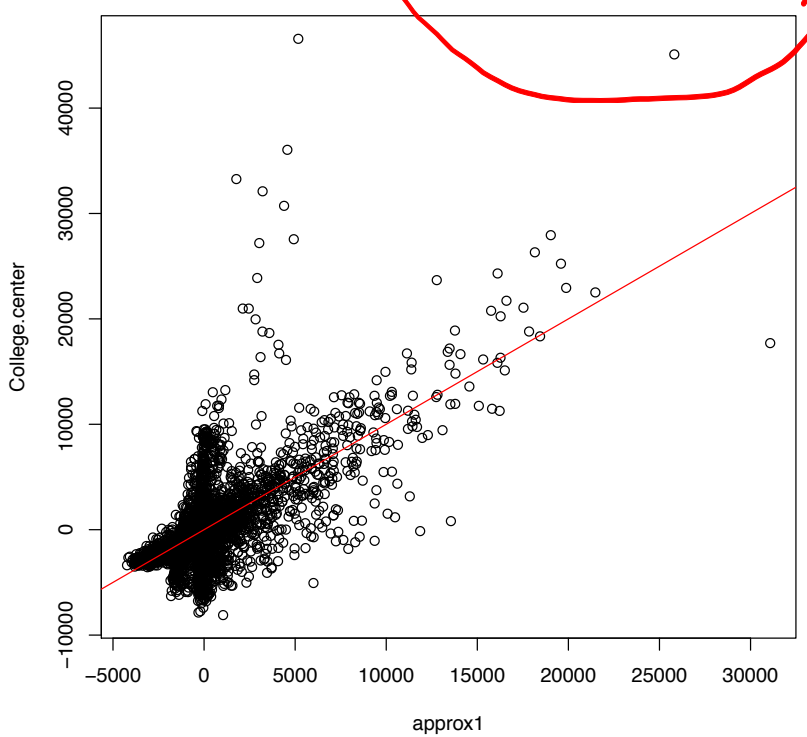
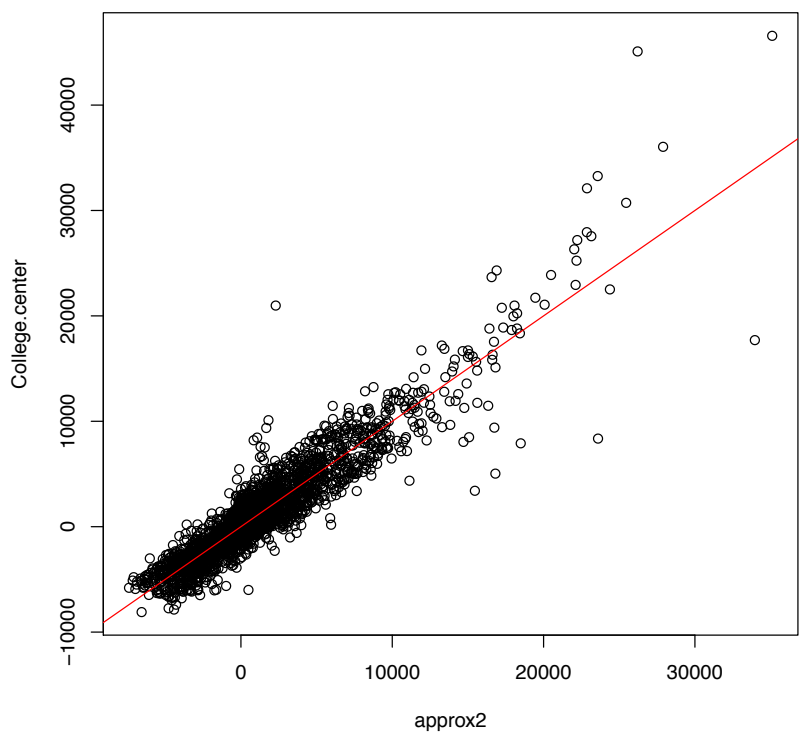


Illustration of the reconstruction

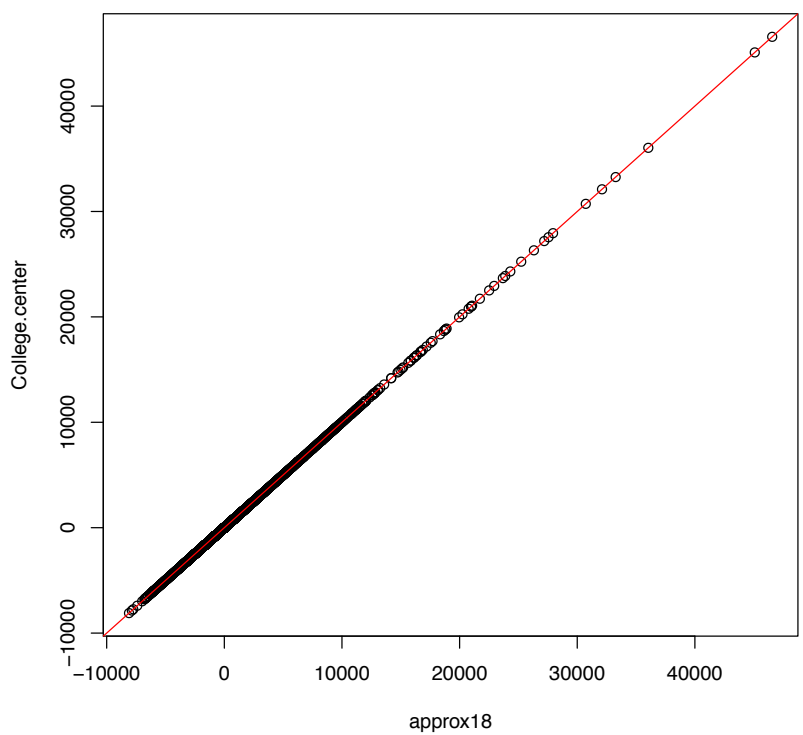
```
Z2 <- College.center %*% fit$loading[, 1:2]
approx2 <- Z2 %*% t(fit$loading[, 1:2])
mean(abs(approx2 - College.center))

## [1] 343.3329

plot(approx2, College.center)
abline(0, 1, col="red")
```



Very reasonable!

Illustration of the reconstruction

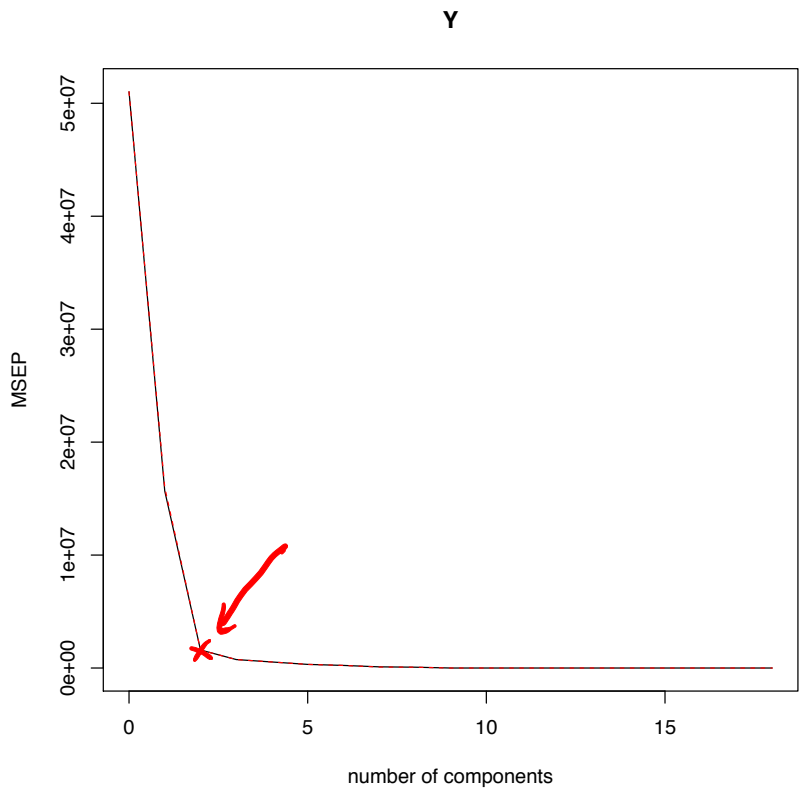
```
Z18 <- College.center %*% fit$loading[, 1:18]
approx18 <- Z18 %*% t(fit$loading[, 1:18])
mean(abs(approx18 - College.center))

## [1] 2.468929e-12

plot(approx18, College.center)
abline(0, 1, col="red")
```



Exact reconstruction

Principle component regression

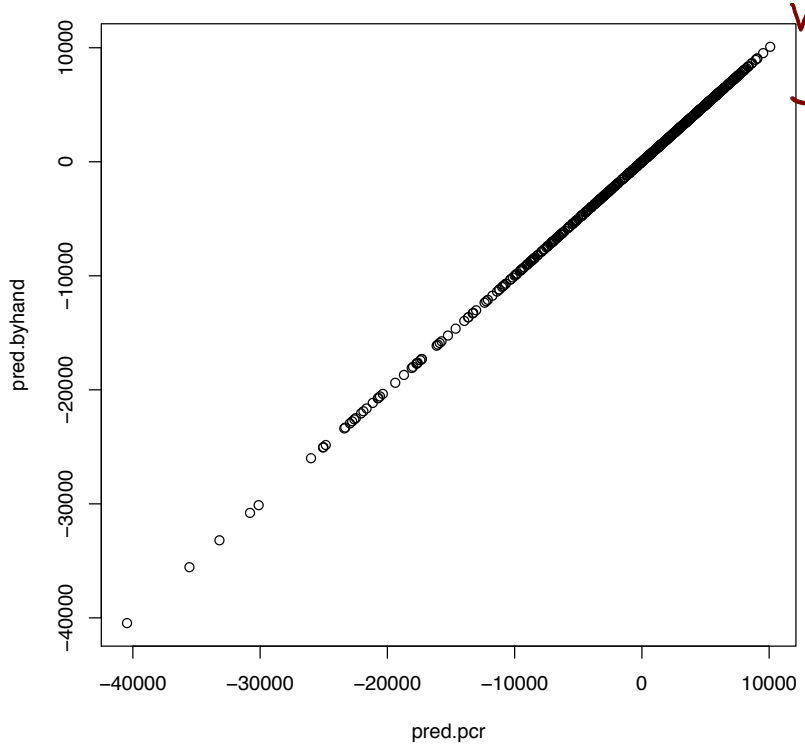
```
library(pls)
Y <- 2 + College.center %*% rnorm(18)
fit.pcr <- pcr(Y ~ College.center, validation ="CV")
validationplot(fit.pcr,val.type="MSEP")
pred.pcr <- predict(fit.pcr, College.center, ncomp = 2)
```



Principle component regression by hand

```
# reproduce this with PCA
Z2 <- College.center %*% fit$loading[, 1:2]
fit.byhand <- lm(Y ~ Z2)
pred.byhand <- cbind(1, Z2) %*% coef(fit.byhand)
plot(pred.pcr, pred.byhand)
```

→ or Z2 = fit$scores[, 1:2]



replicate pcr by hand !

Regression with correlated predictors

It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose $n = p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$, and both predictors and the response are centered to 0, so we do not have to estimate the intercept.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \underbrace{\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}}_{\begin{matrix} = x_1 \\ = x_2 \end{matrix}} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

(a) Write out the ridge regression optimization problem in this setting.

$$\min \left\{ (y_1 - x_1\beta_1 - x_1\beta_2)^2 + (y_2 - x_2\beta_1 - x_2\beta_2)^2 + \lambda\beta_1^2 + \lambda\beta_2^2 \right\}$$

Regression with correlated predictors

Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$

$$\frac{\partial l}{\partial \beta_1} = 2x_1^2\beta_1 - 2x_1(y_1 - x_1\beta_2) + 2x_2^2\beta_1 - 2x_2(y_2 - x_2\beta_2) + 2\lambda\beta_1$$

$$= 2\beta_1(x_1^2 + x_2^2 + \lambda) - 2(x_1y_1 + x_2y_2 - \beta_2x_1^2 - \beta_2x_2^2)$$

set this to 0

$$\Rightarrow 2\hat{\beta}_1(x_1^2 + x_2^2 + \lambda) + 2\hat{\beta}_2(x_1^2 + x_2^2) = 2(x_1y_1 + x_2y_2)$$

$$A\hat{\beta}_1 + B\hat{\beta}_2 = C \qquad (*)$$

Regression with correlated predictors

It turns out (or you can observe the symmetry here)

$$\frac{\partial l}{\partial \beta_2} = 0 \Rightarrow A\hat{\beta}_2 + B\hat{\beta}_1 = C \qquad (**)$$

From (*) and (**)

$$\Rightarrow \hat{\beta}_1 = \hat{\beta}_2 = \frac{C}{A+B}$$

Conclusion:

$\hat{\beta}_{ridge}$ "split" the effects between the two identical predictors.

Regression with correlated predictors

Argue that in this setting, the lasso coefficient estimates are not unique

Using the constrained optimization form:

Lasso: $\min \left\{ (y_1 - \beta_1 x_1 - \beta_2 x_1)^2 + (y_2 - \beta_1 x_2 - \beta_2 x_2)^2 \right\}$ (***)

s.t $|\hat{\beta}_1| + |\hat{\beta}_2| \le s$

Suffice to consider the case when data/response are centerred: $x_1 = -x_2$, $y_1 = -y_2$

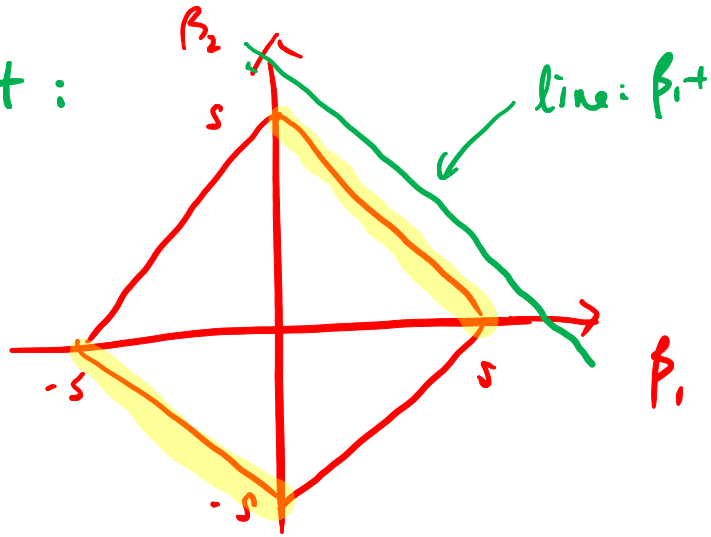
then (***) $= 2\left(y_1 - (\beta_1 + \beta_2) x_1\right)^2$

Regression with correlated predictors

This leads to a simple solution

$$\hat{\beta}_1 + \hat{\beta}_2 = \frac{y_1}{x_1}$$

constraint :



line: $\beta_1 + \beta_2 = \gamma$

$\Rightarrow$ the line is parallel to the boundary thus the minimization occurs only at one of the boundaries (i.e. areas)

Conclusion:

Exist S s.t. all $\hat{\beta}_1 + \hat{\beta}_2 = S$ yield the same lasso objective function

i.e. any "split" of effects is possible! (compare with ridge)

Summary of basis function approaches

- Polynomial regression

  $y \sim \alpha + x^2 + x^3 + \cdots$

- Step function

  $y \sim I(x < c_1) + I(c_1 < x < c_2) + \cdots$

- Regression spline
  - Piecewise polynomial

    $y \sim x I(x < c_1) + x^2 I(x < c_1) + \cdots + x I(c_1 < x < c_2) + \cdots$

  - Cubic spline

    + continuous, first & second derivative continuous

  - Natural cubic spline

    + linear outside of boundary.

- Smoothing spline

  $\sum (y_i - g(x_i))^2 + \lambda \int g''(t)^2 \, dt$

a special case but shrunk version