Advanced Methods Homework 3

Bohao Tang

1 Variable Selection

1. For Linear model with normal error

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon} \qquad \vec{\varepsilon} \sim N(0, \sigma^2 I_n),\ \vec{\beta} \text{ of shape } p\times 1$$

Then the likelihood is:

$$P(Y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}^{\,n}} e^{-\frac{(\vec{Y}-X\vec{\beta})'(\vec{Y}-X\vec{\beta})}{2\sigma^2}}$$

The MLE $\vec{\beta}$, $\hat{\sigma}$ for $P(Y|\theta)$ is $\hat{\beta} = (X'X)^{-1}X'\vec{Y}$

$$\hat{\sigma}^2 = \frac{\vec{Y}'(I-H)\vec{Y}}{n} \quad \text{where } H = X(X'X)^{-1}X'$$

So $$AIC = 2\left[\frac{n}{2}\log 2\pi + \frac{n}{2}\log\frac{\vec{Y}'(I-H)\vec{Y}}{n} + \frac{n}{2}\right] + 2(p+1)$$

$$= n+2+2p + n\log\left\{\frac{2\pi}{n}\left[\vec{Y}'(I-H)\vec{Y}\right]\right\}$$

$$= n\log\left(\frac{SSE}{n}\right) + 2p + \text{constant without } p.$$

2 Ridge Regression

1. Suppose the SVD decomposition of design matrix X is $X = UDV$

Then $var(\hat{\beta}_{ridge}) = \sigma^2 V(D^2+\lambda I)^{-1}D^2(D^2+\lambda I)^{-1}V'$

$Var(\hat{\beta}_{LS}) = \sigma^2 V(D^2)^{-1}V'$ suppose $D = diag\{\cdots d_i \cdots\}$

$$var(\hat{\beta}_{LS}) - var(\hat{\beta}_{ridge}) = \sigma^2 V\, diag\left\{\cdots, -\frac{d_i^2}{(d_i^2+\lambda)^2} + \frac{1}{d_i^2}\cdots\right\}V'$$

since for all i and d_i, $\frac{1}{d_i^2} - \frac{d_i^2}{(d_i^2+\lambda)^2} \geq 0$, so $var(\hat{\beta}_{LS}) - var(\hat{\beta}_{ridge})$ is semi positive definite $\Rightarrow var(\hat{\beta}_{LS}) \geq var(\hat{\beta}_{ridge})$

2: The goal is to minimize $(\vec{Y}-X\vec{\beta})'(\vec{Y}-X\vec{\beta})+\lambda\beta'\beta$

use derivative we get that $\hat{\beta}_{ridge}=(X'X+\lambda I)^{-1}X'Y$

So the hat matrix $H_\lambda = X(X'X+\lambda I)^{-1}X'$

Suppose the SVD of X is $X=UDV$ $\quad D=\begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix}$

Then $H_\lambda = UD(D^2+\lambda I)^{-1}DU'$

$\Rightarrow tr(H_\lambda) = tr(D(D^2+\lambda I)^{-1}D) = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda}$

3 Principal Components

1. Suppose $X'X = U\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix}U'$

where $U=[\vec{p}_1,\vec{p}_2,\cdots\vec{p}_p]$ are orthogonal

and $\lambda_1\geq\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_i\geq\cdots\geq\lambda_p\geq 0$

Then it is to prove $\quad \max\limits_{\substack{\vec{a}\perp span\{\vec{p}_1,\vec{p}_2\cdots\vec{p}_i\} \\ \|\vec{a}\|=1}} \vec{a}'\cdot X'X\cdot\vec{a} = \lambda_{i+1}$

Proof: Since $\vec{a}\perp\{\vec{p}_1,\vec{p}_2\cdots\vec{p}_i\}$ ~~we~~ and $\vec{p}_1\cdots\vec{p}_p$ is a base

we have that $\vec{a}=a_{i+1}\vec{p}_{i+1}+\cdots+a_p\vec{p}_p$ and since $\|\vec{a}\|=1$, $\sum_{l=i+1}^{p}a_l^2=1$

Then $\vec{a}'X'X\vec{a} = \lambda_{i+1}a_{i+1}^2+\lambda_{i+2}a_{i+2}^2+\cdots+\lambda_p a_p^2$

$\leq \lambda_{i+1}(a_{i+1}^2+a_{i+2}^2+\cdots a_p^2) = \lambda_{i+1}$

and the equality can be satisfied ~~only~~ when simply let $a_{i+1}=1$ and others $=0$

So the $i+1$ th principal components explain the maximum variability in direction orthogonal to the first i th.

2: First since $U = XV'D^{-1}$ and $V'D^{-1}$ is invertiable

the column space of U is the same as column space of X

Proof: every vector in column space of X can be writen as $X\vec{\beta}$ for some $\vec{\beta}$,

so, since $X\vec{\beta} = XV'D^{-1}(DV\vec{\beta}) = U(DV\vec{\beta})$

and $U\vec{y} = X(V'D^{-1}\vec{y})$

we have $col(X) = col(U)$

Second U is orthogonal, so the column vectors of U forms an orthonormal basis for $col(U)$

Combine this two argument, U results in an orthonormal baris for $col(X)$.

Then $\hat{y} = \sum_{j=1}^{p} \vec{u}_i \langle \vec{u}_i, \vec{y} \rangle$ is indeed the project of y into $col(X)$.

4. Time Series Analysis

1: MA(1) is ~~X(t)=~~ $X_t = Z_t + \theta Z_{t-1}$ for Z_t i.i.d $\sim N(0, \sigma^2)$

Then $cov(X_t, X_{t+h}) = cov(Z_t + \theta Z_{t-1}, Z_{t+h} + \theta Z_{t+h-1})$

$$= \begin{cases} (1+\theta^2)\sigma^2 & h = 0 \\ \theta\sigma^2 & h = \pm 1 \\ 0 & h \text{ otherwise} \end{cases}$$

So the auto correlation function $\rho(h)$ is

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \theta/(1+\theta^2) & h = \pm 1 \\ 0 & |h| > 1 \end{cases}$$

2: For AR(p) model

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t \quad Z_t \overset{i.i.d.}{\sim} N(0,\sigma^2)$$

Then we have that

$$\gamma(1) = E[X_t X_{t-1}] = \phi_1 E X_{t-1}^2 + \phi_2 E X_{t-2} X_{t-1} + \cdots \phi_p E[X_{t-1} X_{t-p}] + E[X_{t-1} Z_t]$$
$$= \phi_1 \gamma(0) + \phi_2 \gamma(1) + \cdots + \phi_p \gamma(p-1)$$

$$\gamma(2) = E[X_t X_{t-2}] = \phi_1 \gamma(1) + \phi_2 \gamma(0) + \cdots + \phi_p \gamma(p-2)$$

$$\gamma(3) = E[X_t X_{t-3}] = \phi_1 \gamma(2) + \phi_2 \gamma(1) + \phi_3 \gamma(0) + \cdots + \phi_p \gamma(p-3)$$

$$\vdots$$

$$\gamma(p) = E[X_t X_{t-p}] = \phi_1 \gamma(p-1) + \phi_2 \gamma(p-2) + \phi_3 \gamma(p-3) + \cdots \phi_p \cdot \gamma(0)$$

$$\Rightarrow \Gamma \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} \gamma(1) \\ \vdots \\ \gamma(p) \end{pmatrix} \quad \text{where } \Gamma = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \gamma(1) & \cdots & \gamma(p-2) \\ \gamma(2) & \gamma(1) & \gamma(0) & \gamma(1) \cdots & \gamma(p-3) \\ \vdots & & & & \\ \gamma(p-1) & \gamma(p-2) & \cdots & & \gamma(0) \end{pmatrix}.$$

$$\Rightarrow \begin{pmatrix} \hat\phi_1 \\ \vdots \\ \hat\phi_p \end{pmatrix} = \begin{pmatrix} 1, & \hat\rho(1), & \hat\rho(2) & \cdots & \hat\rho(p-1) \\ \hat\rho(1), & 1, & \hat\rho(1) & \cdots & \hat\rho(p-2) \\ \hat\rho(2), & \hat\rho(1), & 1 & \cdots & \hat\rho(p-3) \\ \vdots & & & & \vdots \\ \hat\rho(p-1), & \cdots & & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat\rho(1) \\ \hat\rho(2) \\ \vdots \\ \hat\rho(p) \end{pmatrix}$$

also $\gamma(0) = E[X_t X_t] = \phi_1 \gamma(1) + \phi_2 \gamma(2) + \cdots \phi_p \gamma(p) + \sigma^2$

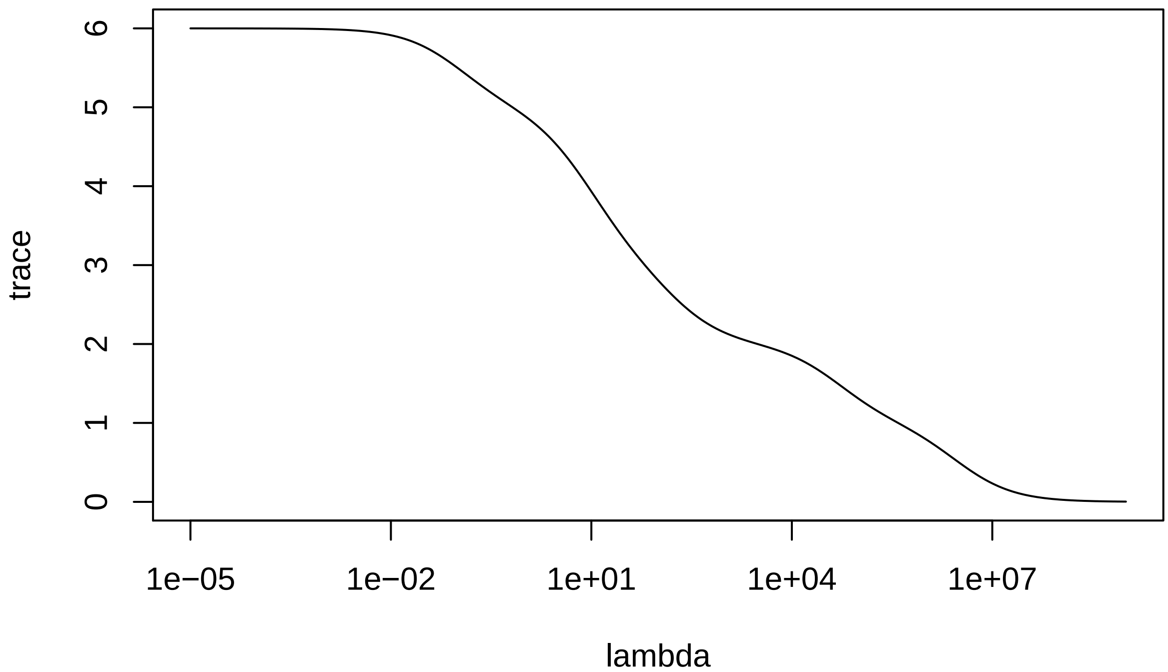
$$\Rightarrow \hat\sigma^2 = \hat\gamma(0)\left(1 - \hat\phi_1 \hat\rho(1) - \hat\phi_2 \hat\rho(2) \cdots - \hat\phi_p \hat\rho(p)\right)$$

where $\hat\gamma(h) = \frac{1}{n-h} \sum_{j=1}^{n-h} (X_{j+h} - \bar X)(X_j - \bar X)$ and $\hat\rho(h) = \frac{\hat\gamma(h)}{\hat\gamma(0)}$

Coding and data analysis exercises

1.

```r
require(stats)

myridge <- function(X, y, lambda){

    Design = cbind(1, X)
    s = svd(Design)
    D = s$d
    U = s$u
    V = s$v

    trace = c()
    beta = c()
    for(l in lambda){
     trace = c(trace, sum(D^2 / (D^2 + l)))
     beta = cbind(beta, V %*% diag((D^2 + l)^-1) %*% diag(D) %*% t(U) %*% y)
    }

    plot(lambda, trace, type = "l", log = "x")

    return(beta)
}

mtcars_selected = as.matrix(mtcars[c("mpg","cyl","disp","hp","drat","wt")])
X = mtcars_selected[,2:6]
y = mtcars_selected[,1]
lambda = 10 ^ seq(-5,9,length.out = 1000)
dull = myridge(X, y, lambda)
```



2.

This function `mypcr` use principle components regression and return with two lists of parameter `gamma` and fitted values `fiited_values` of every choice of largest several components. Along with a plot.

```
require(stats)

mypcr <- function(X, y){

    Design = cbind(1, X)
    s = svd(Design)
    D = s$d
    U = s$u
    V = s$v
    Z = U %*% diag(D)

    scores = c()
    score_ratio_of_first_m = c()
    gamma = list()
    fitted_values = list()
    num_of_components = 1:length(D)
    for(d in D){
     scores = c(scores, d)
     m = length(scores)
     score_ratio_of_first_m = c(score_ratio_of_first_m, sum(scores)/sum(D))
```

```
    if(m == 1){
      g = matrix(1/D[1:m]) %*% t(U[,1:m]) %*% y
      gamma[[m]] = g
    }
    else{
      g = diag(1/D[1:m]) %*% t(U[,1:m]) %*% y
      gamma[[m]] = g
    }
    f = U[,1:m] %*% t(U[,1:m]) %*% y
    fitted_values[[m]] = f
  }

  plot(num_of_components, score_ratio_of_first_m, type = "l")

  return(list("components" = Z,
              "gamma" = gamma,
              "fitted_values" = fitted_values,
              "ratios" = score_ratio_of_first_m))
}

mtcars_selected = as.matrix(mtcars[c("mpg","cyl","disp","hp","drat","wt")])
X = mtcars_selected[,2:6]
y = mtcars_selected[,1]
dull = mypcr(X, y)
```