

Function approximation using basis functions

- Regression. Suppose that we observe (X_i, Y_i) : $1 \leq i \leq n$, where

$$Y_i = f(X_i) + \varepsilon_i,$$

and ε_i 's are IID errors with mean zero and variance σ^2 . The problem of interest in regression is to estimate f based on (X_i, Y_i) 's.

- Estimation approach. Choose a set of functions $\{B_j\}_{j=1}^J$ so that f can be approximated well using $\sum_{j=1}^J a_j B_j$. Then the coefficients a_j 's can be estimated using least squares method. That is, a_k 's are chosen so that

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^J a_j B_j(X_i) \right)^2 \quad (1)$$

is minimized. Let $\hat{a}_1, \dots, \hat{a}_J$ be the solution to the minimization problem in (1). Let

$$\hat{f} = \sum_{j=1}^J \hat{a}_j B_j$$

Then \hat{f} is the estimator of f based on basis approximation and least square estimation using basis functions B_1, \dots, B_J .

- Some choices for the B_j 's are
 - Trigonometric functions.
 - Polynomials.
- Given Y_i : $1 \leq i \leq n$ and $Z_{i,j}$: $1 \leq i \leq n$ and $1 \leq j \leq J$, the vector $(a_1, \dots, a_J)^T$ that minimizes

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^J a_j Z_{i,j} \right)^2$$

is given by $(Z^T Z)^{-1} Z^T Y$, where Z is the $n \times J$ matrix $(Z_{i,j})$ and Y is the $n \times 1$ vector (Y_i) . Let $\hat{a} = (Z^T Z)^{-1} Z^T Y$. \hat{a} can be computed in R using the `lm` function

```
lm(Y~Z-1)$coef
```

or using the `solve` function to compute $(Z^T Z)^{-1}$.

```
solve(t(Z) %*% Z) %*% t(Z) %*% Y
```

- We can also use `ginv` to compute the generalized inverse of $Z^T Z$. Theoretically, the generalized inverse of $Z^T Z$ is the same as $(Z^T Z)^{-1}$ when $(Z^T Z)^{-1}$ exists. However, `ginv(t(Z) %*% Z) %*% t(Z) %*% Y` may differ from `solve(t(Z) %*% Z) %*% t(Z) %*% Y` due to computational error.

- Example 1. Let $f(x) = x \sin(20x)$ for $x \in [0, 1]$. Suppose that $n = 1000$, $(X_1, \dots, X_n) = \text{seq}(0, 1, \text{length}=n)$, and $Y_i = f(X_i)$ for $i = 1, \dots, n$. Find the estimator of f based on best linear approximation using 11 basis functions $1, \cos(2\pi kx), \sin(2\pi kx)$: $k = 1, \dots, 5$.

(a) Find $(a_0, a_1, \dots, a_5, b_1, \dots, b_5)$ that minimizes

$$\sum_{i=1}^n \left(Y_i - a_0 - \sum_{k=1}^5 (a_k \cos(2\pi k X_i) + b_k \sin(2\pi k X_i)) \right)^2.$$

(b) Let $(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_5, \hat{b}_1, \dots, \hat{b}_5)$ be the solution to the above minimization problem. Let

$$\hat{f}(x) = \hat{a}_0 + \sum_{k=1}^5 (\hat{a}_k \cos(2\pi kx) + \hat{b}_k \sin(2\pi kx)) \text{ for } x \in [0, 1].$$

Then \hat{f} is the estimator of f based on best linear approximation using 11 basis functions $1, \cos(2\pi kx), \sin(2\pi kx)$: $k = 1, \dots, 5$. Plot \hat{f} on $[0, 1]$.

(c) Find the ISE $\int_0^1 (\hat{f}(x) - f(x))^2 dx$.

```
#(a)
n <- 1000
x <- seq(0,1,length=n)
f <- function(x){ x*sin(20*x) }
y <- f(x)
m <- 5
Z <- matrix(0, n, 2*m)
for (k in 1:m){
  Z[,k] <- cos(2*pi*k*x)
  Z[,k+m] <- sin(2*pi*k*x)
}
Z <- cbind(rep(1,n), Z)
a1=(ginv(t(Z) %*% Z) %*% t(Z) %*% y)[,1]
a2=lm(y~Z-1)$coef
a1-a2
```

#estimated coefficients
#estimated coefficients

```
#(b)
fhat <- function(x){
  m <- 5
  n <- length(x)
  Z <- matrix(0, n, 2*m)
  for (k in 1:m){
    Z[,k] <- cos(2*pi*k*x)
    Z[,k+m] <- sin(2*pi*k*x)
  }
  Z <- cbind(rep(1,n), Z)
  ans <- Z %*% a2
  return(ans[,1])
}
```

```

}
curve(fhat,0,1)
curve(f,0,1, add=T, col=2)

#(c)
g <- function(u){ return((fhat(u)-f(u))^2) }
integrate(g,0,1)$value #ISE 0.01065894

```

- Exercise 1. Let $f(x) = \sin(20x)$ for $x \in [0, 1]$. Suppose that $n = 1000$, $(X_1, \dots, X_n) = \text{seq}(0, 1, \text{length}=n)$, and $Y_i = f(X_i)$ for $i = 1, \dots, n$. Let \hat{f} be the estimator of f based on best linear approximation using basis functions $1, x, \dots, x^m$.
 - (a) Find the ISE of estimating f based on $(X_1, Y_1), \dots, (X_n, Y_n)$ by approximating f using the best linear combination of $1, x, \dots, x^m$, where $m = 10$.
 - (b) Compute the ISE in Part (a) with m replaced by 11, 12, and 13. Does the ISE decrease as m increases?
- Exercise 2. In Exercise 1, replace Y_i with $f(X_i) + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 2$. Compute the approximate IMSE based on 100 trials for $m = 10, 11, 12, 13$. Does the IMSE decrease as m increases?
- The leave-one-out cross-validation function. Consider estimating f based on best linear approximation using basis functions B_1, \dots, B_J .
 - Let \hat{f} denote the estimator of f using the full data (X_i, Y_i) : $1 \leq i \leq n$.
 - Let \hat{f}_{-i} denote the estimator of f with (Y_i, X_i) removed from the data.

Then

$$RSSCV = \sum_{i=1}^n \left(Y_i - \hat{f}_{-i}(X_i) \right)^2.$$

It can be shown that $RSSCV$ can be computed using the formula

$$\sum_{i=1}^n \frac{(Y_i - \hat{f}(X_i))^2}{(1 - h_{ii})^2},$$

where h_{ii} is the i -th diagonal element of the hat matrix $Z(Z^T Z)^{-1} Z^T$ and Z is the $n \times J$ matrix whose j -th column is $(B_j(X_1), \dots, B_j(X_n))^T$ for $j = 1, \dots, J$.

- Example 2. Compute the RSSCV for the data in Example 1.

```

n <- 1000
x <- seq(0,1,length=n)
f <- function(x){ x*sin(20*x) }
y <- f(x)
m <- 5
Z <- matrix(0, n, 2*m)

```

```

for (k in 1:m){
  Z[ ,k] <- cos(2*pi*k*x)
  Z[ ,k+m] <- sin(2*pi*k*x)
}
Z <- cbind(rep(1,n), Z)
mod <- lm(y~Z-1)
hii.v <- lm.influence(mod)$hat
rsscv <- sum( mod$resid^2/(1-hii.v)^2 )

```

- Exercise 3. For the data in Example 1, compute $RSSCV = \sum_{i=1}^n \left(Y_i - \hat{f}_{-i}(X_i) \right)^2$ directly and compare it with the result in Example 2.