High-dimensional data analysis, final report, Due 6/24(Wed.)

102225014 Ai-Chun Chen

1) Derive B-spline bases with order M=4 under a knot sequence $\xi_0 < \xi_1 < \xi_2$ **with an equally spaced mesh** $\Delta = \xi_1 - \xi_0 = \xi_2 - \xi_1$**.**

NOTE: Result from M=3 cannot be used since the augmented knots are different.

<u>solution</u>:

Augmented knots:

K=1, M=4 (K+2M=9)

$\tau_1 \le \tau_2 \le \ldots \le \tau_9$ such that

i) $\tau_1 = \tau_2 = \tau_3 = \tau_4 = \xi_0$

ii) $\tau_5 = \xi_1$

iii) $\tau_6 = \tau_7 = \tau_8 = \tau_9 = \xi_2$

When m=1:

$\mathrm{B}_{1,1}(x) = 0$; (doesn't exist a x such that $\tau_1 \le x < \tau_2$)

$\mathrm{B}_{2,1}(x) = 0$;

$\mathrm{B}_{3,1}(x) = 0$;

$\mathrm{B}_{4,1}(x) = I(\xi_0 \le x < \xi_1)$;

$\mathrm{B}_{5,1}(x) = I(\xi_1 \le x < \xi_2)$;

$\mathrm{B}_{6,1}(x) = 0$;

$\mathrm{B}_{7,1}(x) = 0$;

$\mathrm{B}_{8,1}(x) = 0$.

When m=2:

$\mathrm{B}_{1,2}(x) = 0$;

$\mathrm{B}_{2,2}(x) = 0$;

$\mathrm{B}_{3,2}(x) = \frac{\xi_1 - x}{\Delta} I(\xi_0 \le x < \xi_1)$;

$\mathrm{B}_{4,2}(x) = \frac{x - \xi_0}{\Delta} I(\xi_0 \le x < \xi_1) + \frac{\xi_2 - x}{\Delta} I(\xi_1 \le x < \xi_2)$;

$\mathrm{B}_{5,2}(x) = \frac{x - \xi_1}{\Delta} I(\xi_1 \le x < \xi_2)$;

$\mathrm{B}_{6,2}(x) = 0$;

$\mathrm{B}_{7,2}(x) = 0$.

When m=3:

$$B_{1,3}(x) = 0;$$

$$B_{2,3}(x) = \left(\frac{\xi_1 - x}{\Delta}\right)^2 I(\xi_0 \le x < \xi_1);$$

$$B_{3,3}(x) = \left\{\frac{(x-\xi_0)(\xi_1 - x)}{\Delta^2} + \frac{(\xi_2 - x)(x-\xi_0)}{2\Delta^2}\right\} I(\xi_0 \le x < \xi_1) + \frac{(\xi_2 - x)^2}{2\Delta^2} I(\xi_1 \le x < \xi_2)$$

$$B_{4,3}(x) = \left\{\frac{(x-\xi_1)(\xi_2 - x)}{\Delta^2} + \frac{(\xi_2 - x)(x-\xi_0)}{2\Delta^2}\right\} I(\xi_1 \le x < \xi_2) + \frac{(x-\xi_0)^2}{2\Delta^2} I(\xi_0 \le x < \xi_1)$$

$$B_{5,3}(x) = \left(\frac{x-\xi_1}{\Delta}\right)^2 I(\xi_1 \le x < \xi_2);$$

$$B_{6,3}(x) = 0.$$

When m=4:

$$B_{1,4}(x) = \left(\frac{\xi_1 - x}{\Delta}\right)^3 I(\xi_0 \le x < \xi_1)$$

$$B_{2,4}(x) = \left\{\frac{(x-\xi_0)(\xi_1 - x)^2}{\Delta^3} + \frac{(\xi_2 - x)(x-\xi_0)(\xi_1 - x)}{2\Delta^3} + \frac{(\xi_2 - x)^2(x-\xi_0)}{4\Delta^3}\right\} I(\xi_0 \le x < \xi_1) + \frac{(\xi_2 - x)^3}{4\Delta^3} I(\xi_1 \le x < \xi_2)$$

$$B_{3,4}(x) = \left\{\frac{(x-\xi_0)^2(\xi_1 - x)}{2\Delta^3} + \frac{(\xi_2 - x)(x-\xi_0)^2}{2\Delta^3}\right\} I(\xi_0 \le x < \xi_1) + \left\{\frac{(x-\xi_1)(\xi_2 - x)^2}{2\Delta^3} + \frac{(\xi_2 - x)^2(x-\xi_0)}{2\Delta^3}\right\} I(\xi_1 \le x < \xi_2)$$

$$B_{4,4}(x) = \left\{\frac{(x-\xi_1)^2(\xi_2 - x)}{\Delta^3} + \frac{(\xi_2 - x)(x-\xi_0)(x-\xi_1)}{2\Delta^3} + \frac{(\xi_2 - x)(x-\xi_0)^2}{4\Delta^3}\right\} I(\xi_1 \le x < \xi_2) + \frac{(x-\xi_0)^3}{4\Delta^3} I(\xi_0 \le x < \xi_1)$$

$$B_{5,4}(x) = \left(\frac{x-\xi_1}{\Delta}\right)^3 I(\xi_1 \le x < \xi_2).$$

2) Express the above B-spline bases in terms of $z_i(t) = \dfrac{t-\xi_i}{\Delta}$ **for** $i = 0, 1, 2.$

<u>solution</u>:

m=1,

$$B_{1,1}(x) = B_{2,1}(x) = B_{3,1}(x) = B_{6,1}(x) = B_{7,1}(x) = B_{8,1}(x) = 0;$$

$$B_{4,1}(x) = I(\xi_0 \le x < \xi_1); \quad B_{5,1}(x) = I(\xi_1 \le x < \xi_2).$$

m=2:

$$B_{1,2}(x) = 0;$$

$$B_{2,2}(x) = 0;$$

$$B_{3,2}(x) = -z_1(x) I(\xi_0 \le x < \xi_1)$$

$$B_{4,2}(x) = z_0(x) I(\xi_0 \le x < \xi_1) - z_2(x) I(\xi_1 \le x < \xi_2)$$

$$\mathrm{B}_{5,2}(x) = z_1(x) I(\xi_1 \le x < \xi_2)$$

$$\mathrm{B}_{6,2}(x) = 0;$$

$$\mathrm{B}_{7,2}(x) = 0.$$

m=3:

$$\mathrm{B}_{1,3}(x) = 0;$$

$$\mathrm{B}_{2,3}(x) = z_1{}^2(x) I(\xi_0 \le x < \xi_1);$$

$$\mathrm{B}_{3,3}(x) = \left\{ -z_0(x) z_1(x) - \frac{1}{2} z_0(x) z_2(x) \right\} I(\xi_0 \le x < \xi_1) + \frac{1}{2} z_2{}^2(x) I(\xi_1 \le x < \xi_2)$$

$$\mathrm{B}_{4,3}(x) = \left\{ -z_1(x) z_2(x) - \frac{1}{2} z_0(x) z_2(x) \right\} I(\xi_1 \le x < \xi_2) + \frac{1}{2} z_0{}^2(x) I(\xi_0 \le x < \xi_1)$$

$$\mathrm{B}_{5,3}(x) = z_1{}^2(x) I(\xi_1 \le x < \xi_2);$$

$$\mathrm{B}_{6,3}(x) = 0.$$

m=4:

$$\mathrm{B}_{1,4}(x) = -z_1{}^3(x) I(\xi_0 \le x < \xi_1)$$

$$\mathrm{B}_{2,4}(x) = \left[z_0(x)\{1 - z_0(x)\}^2 + \frac{1}{2} z_0(x)\{z_0(x) - 1\}\{z_0(x) - 2\} + \frac{1}{4} z_0(x)\{z_0(x) - 2\}^2 \right] I(\xi_0 \le x < \xi_1)$$
$$- \frac{1}{4} z_2{}^3(x) I(\xi_1 \le x < \xi_2)$$

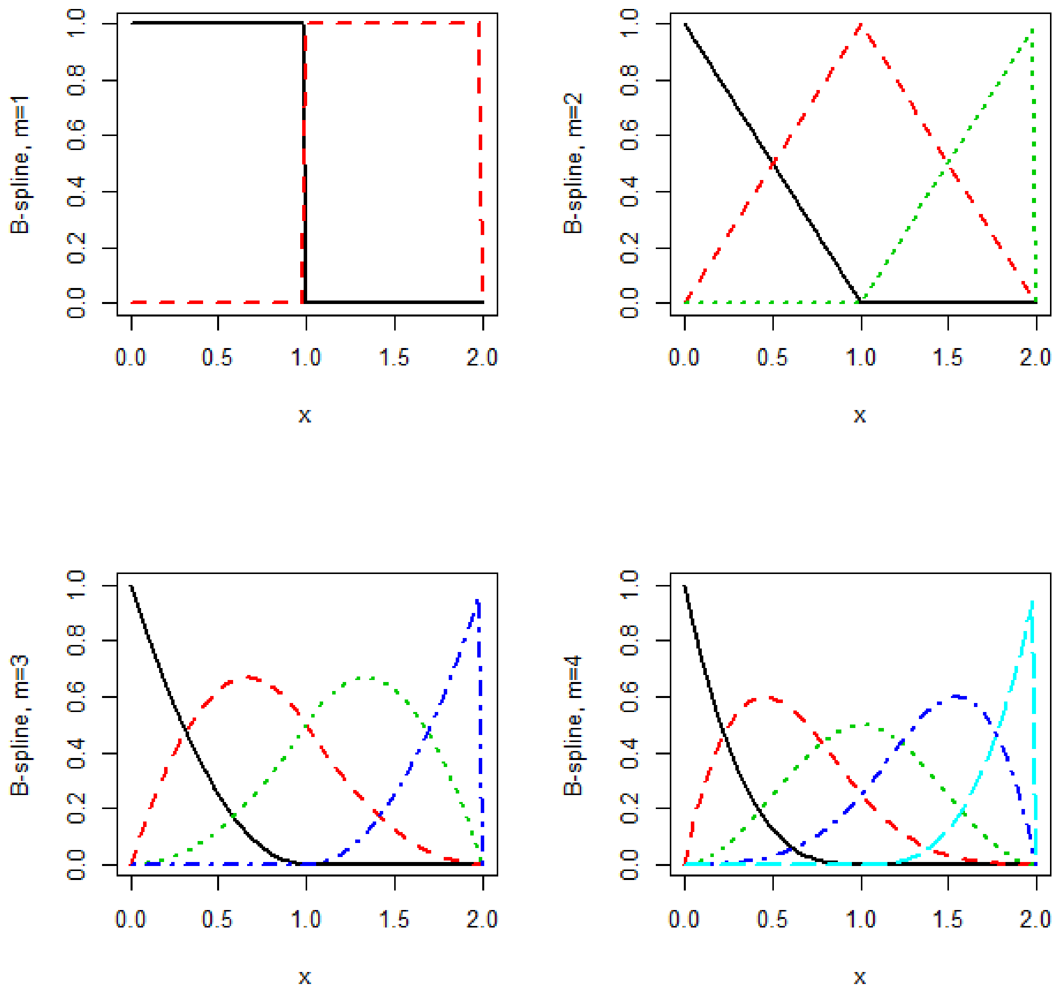
$$\mathrm{B}_{3,4}(x) = \left[\frac{1}{2} z_0{}^2(x)\{1 - z_0(x)\} + \frac{1}{2} z_0{}^2(x)\{2 - z_0(x)\} \right] I(\xi_0 \le x < \xi_1)$$
$$+ \left\{ \frac{1}{2}\{z_1(x) + 1\}\{z_1(x) - 1\}^2 + \frac{1}{2} z_1(x)\{z_1(x) - 1\}^2 \right\} I(\xi_1 \le x < \xi_2)$$

$$\mathrm{B}_{4,4}(x) = \left[\frac{1}{2} z_0(x)\{z_0(x) - 1\}\{2 - z_0(x)\} - \frac{1}{4} z_0{}^2(x)\{z_0(x) - 2\} + \{z_0(x) - 1\}^2\{2 - z_0(x)\} \right] I(\xi_1 \le x < \xi_2)$$
$$+ \frac{1}{4}\{z_1(x) + 1\}^3 I(\xi_0 \le x < \xi_1)$$

$$\mathrm{B}_{5,4}(x) = z_1{}^3(x) I(\xi_1 \le x < \xi_2)$$

3) Depicts the B-spline basis function with knots $\xi_0 = 0$, $\xi_1 = 1$ and $\xi_2 = 2$ using R; All the basis functions plotted in one figure. Distinguish them by color.

solution:



4) Give second derivatives of the B-splines

B-spline is $f(x) = \beta_1 \mathrm{B}_{1,4}(x) + \beta_2 \mathrm{B}_{2,4}(x) + \beta_3 \mathrm{B}_{3,4}(x) + \beta_4 \mathrm{B}_{4,4}(x) + \beta_5 \mathrm{B}_{5,4}(x)$

The second derivatives with respect to x is

$$f''(x) = \begin{cases} (-6x+12)\beta_1 + (21x/2-9)\beta_2 + (3-6x)\beta_3 + (3x/2)\beta_4, & x \in [0,1) \\ (3-3x/2)\beta_2 + (6x-9)\beta_3 + (12-21x/2)\beta_4 + (6x-6)\beta_5, & x \in [1,2) \end{cases}$$

5) Express the penalization term explicitly using quadratic form: $\frac{1}{\Delta^5}\mathbf{g}^T\mathbf{A}\mathbf{g}$, where the matrix $\mathbf{A}$ should be calculated.

$$[\,\mathbf{A}\,]_{jk} = \int_{[0,2)} \mathrm{B}''_{j,4}(x)\mathrm{B}''_{k,4}(x)dx$$

We have

$$\mathbf{A} = \begin{bmatrix} 84 & -39 & 3 & 6 & 0 \\ -39 & 24 & -6 & -3 & 1.5 \\ 3 & -6 & 6 & -6 & 3 \\ 6 & -3 & -6 & 24 & -16.5 \\ 0 & 1.5 & 3 & -16.5 & 12 \end{bmatrix}.$$

6) Fit the smoothing spline to any dataset. Demonstrate how the fitted curve changes as the smoothing parameter changes as $\lambda \to \infty$.

I choose the data set from Portland cement data sheet, it can be found in R package "AICcmodavg" and open the data cement by "data(cement)".
I display B-spline by using the second column of data cement

```
> x2
 [1] 26 29 56 31 52 55 71 31 54 47 40 66 68
```

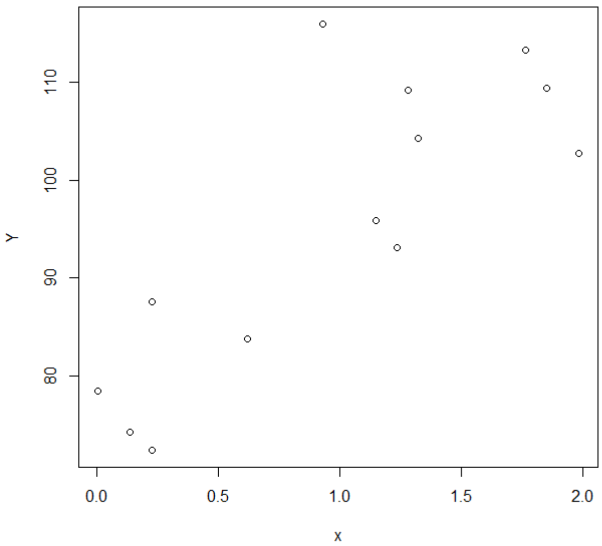
and the response values $\mathbf{y}$

```
> Y
 [1]  78.5  74.3 104.3  87.6  95.9 109.2 102.7  72.5  93.1 115.9  83.8 113.3
[13] 109.4
```

And transform the range of $\mathbf{x}_2$ to $[0, 2]$:

```
> x
 [1] 0.004395604 0.136263736 1.323076923 0.224175824 1.147252747 1.279120879
 [7] 1.982417582 0.224175824 1.235164835 0.927472527 0.619780220 1.762637363
[13] 1.850549451
```

The scatter plot of $\mathbf{x}_2$ to $\mathbf{y}$ is provided



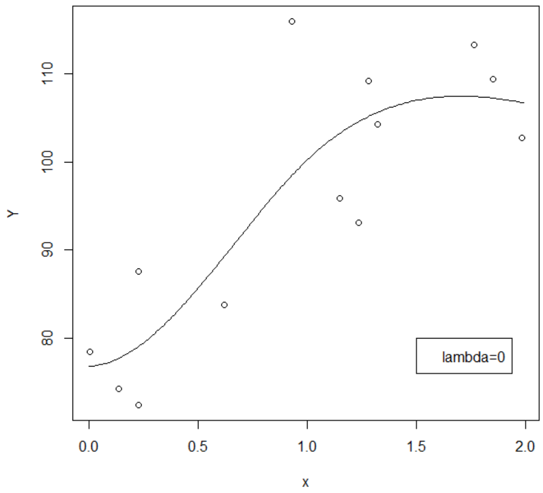
We use $\hat{f}(x) = \sum_{j=1}^{5} \hat{\beta}_j \mathrm{B}_{j,4}(x)$ to fit this data, where $\hat{\beta}_j$ is the j-th component of

$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{A})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$. The design matrix $\mathbf{X}$ is followed

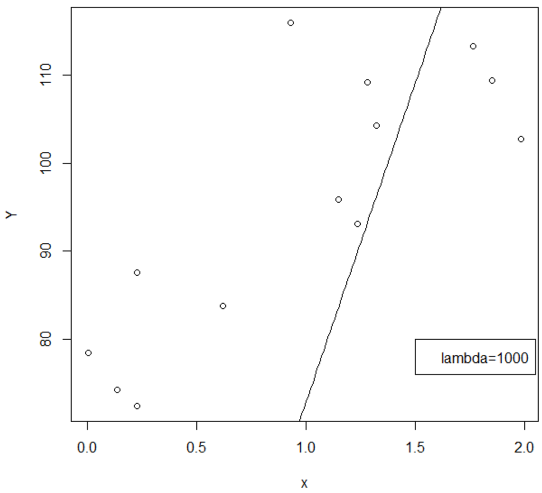
$$\mathbf{X} = \begin{bmatrix} \mathrm{B}_{1,4}(x_1) & \cdots & \mathrm{B}_{5,4}(x_1) \\ \\ \mathrm{B}_{1,4}(x_n) & \cdots & \mathrm{B}_{5,4}(x_n) \end{bmatrix}.$$

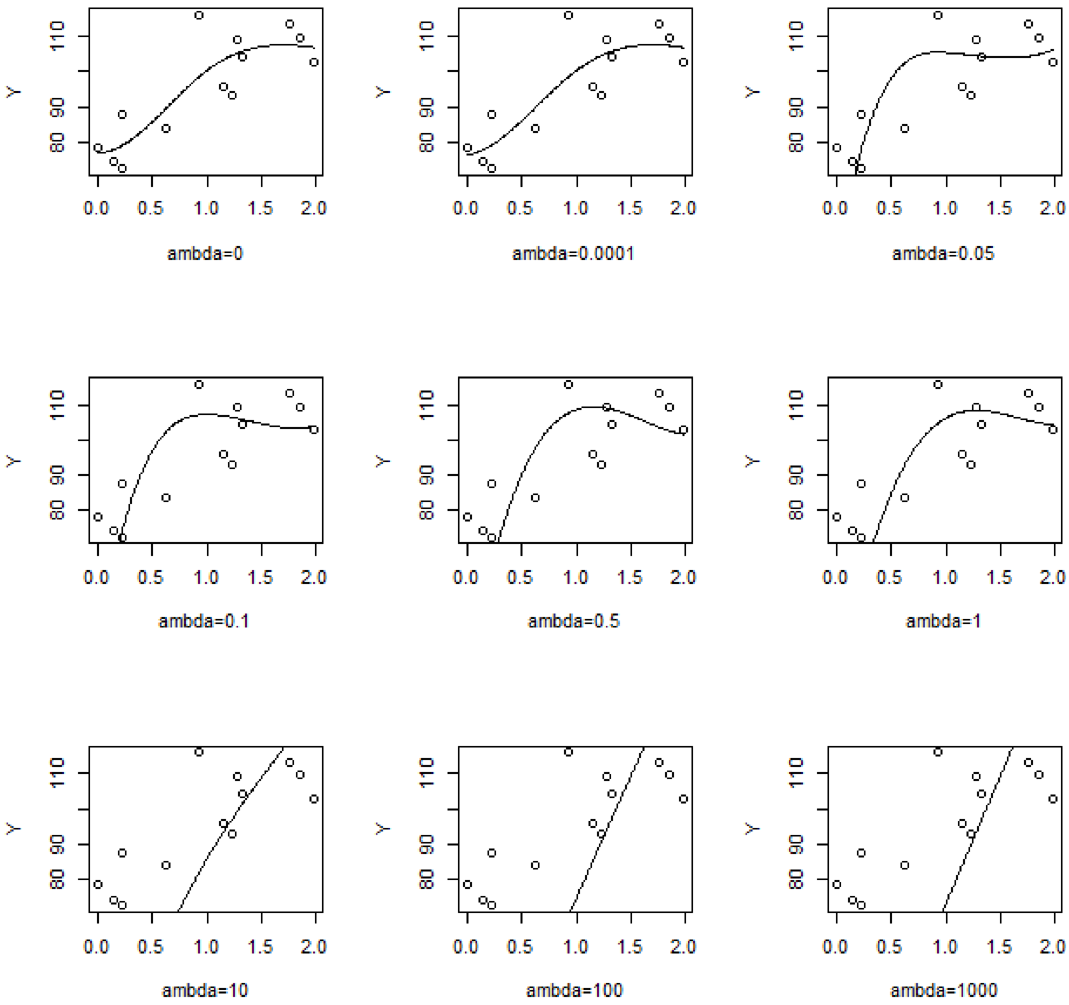
The follow-up figures show how the fitted curve change when the smoothing parameter λ grows.

$\lambda = 0$



$\lambda = 1000$





Conclusion:

When λ is small, the fitted curve tend to be more bending, maybe overfit from data. This phenomenon disappear when λ grows up. When λ is 100, the fitted curve is nearly a straight line.

Appendix – R code

```
#high-dimensional data final
#Q3

#m=1
b11=0
b21=0
b31=0
b41=function(x){(x>=0 & x<1)}
b51=function(x){(x>=1 & x<2)}
b61=0
b71=0
b81=0

#m=2
b12=0
b22=0
b32=function(x){(1-x)*(x>=0 & x<1)}
b42=function(x){x*(x>=0 & x<1)-(x-2)*(x>=1 & x<2)}
b52=function(x){(x-1)*(x>=1 & x<2)}
b62=0
b72=0

#m=3
b13=0
b23=function(x){(x-1)^2*(x>=0 & x<1)}
b33=function(x){((-x)*(x-1)-1/2*x*(x-2))*(x>=0 & x<1)+1/2*(x-2)^2*(x>=1 &
x<2)}
b43=function(x){(-(x-1)*(x-2)-1/2*x*(x-2))*(x>=1 & x<2)+1/2*x^2*(x>=0 &
x<1)}
b53=function(x){(x-1)^2*(x>=1 & x<2)}
b63=0
```

```
#m=4
b14=function(x){-(x-1)^3*(x>=0 & x<1)}
b24=function(x){ (x*(1-x)^2+1/2*x*(x-1)*(x-2)+1/4*x*(x-2)^2)*(x>=0 & x<1)
- 1/4*(x-2)^3*(x>=1 & x<2)}
b34=function(x){ (1/2*x^2*(1-x)+1/2*x^2*(2-x))*(x>=0 & x<1) +
(1/2*x*(x-2)^2+1/2*(x-1)*(x-2)^2)*(x>=1 & x<2)}
b44=function(x){ (1/2*x*(x-1)*(2-x)-1/4*x^2*(x-2)+(x-1)^2*(2-x))*(x>=1 &
x<2) + 1/4*x^3*(x>=0 & x<1)}
b54=function(x){ (x-1)^3*(x>=1 & x<2)}
par(mfrow=c(2,2))
#m=1
curve(b41,from=0, to=2, col=1, lty=1, ylab="B-spline, m=1",lwd=2)
curve(b51,from=0, to=2, col=2, lty=2,lwd=2, add=TRUE)
#legend(x=0,y=.9,legend = c("b41","b51"), lty=c(1:2), col=c(1:2))
#m=2
curve(b32,from=0, to=2, col=1, lty=1, ylab="B-spline, m=2",lwd=2)
curve(b42,from=0, to=2, col=2, lty=2,lwd=2, add=TRUE)
curve(b52,from=0, to=2, col=3, lty=3,lwd=2, add=TRUE)
#legend(x=.25,y=1,legend = c("b32","b42","b52"), lty=c(1:3), col=c(1:3))
#m=3
curve(b23,from=0, to=2, col=1, lty=1, ylab="B-spline, m=3",lwd=2)
curve(b33,from=0, to=2, col=2, lty=2,lwd=2, add=TRUE)
curve(b43,from=0, to=2, col=3, lty=3,lwd=2, add=TRUE)
curve(b53,from=0, to=2, col=4, lty=4,lwd=2, add=TRUE)
#legend(x=.25,y=1,legend = c("b23","b33","b43","b53"), lty=c(1:4), col=c(1:4))
#m=4
curve(b14,from=0, to=2, col=1, lty=1, ylab="B-spline, m=4",lwd=2)
curve(b24,from=0, to=2, col=2, lty=2,lwd=2, add=TRUE)
curve(b34,from=0, to=2, col=3, lty=3,lwd=2, add=TRUE)
curve(b44,from=0, to=2, col=4, lty=4,lwd=2, add=TRUE)
curve(b54,from=0, to=2, col=5, lty=5,lwd=2, add=TRUE)
#legend(x=.25,y=1,legend = c("b14","b24","b34","b44","b54"), lty=c(1:5),
col=c(1:5))



#Q5
db1 = function(x) (12-6*x)*(x>=0 & x<1)
```

```
db2 = function(x) (21/2*x-9)*(x>=0 & x<1)+(3-3/2*x)*(x>=1 & x<2)
db3 = function(x) (3-6*x)*(x>=0 & x<1)+(6*x-9)*(x>=1 & x<2)
db4 = function(x) (3/2*x)*(x>=0 & x<1)+(12-21/2*x)*(x>=1 & x<2)
db5 = function(x) (6*x-6)*(x>=1 & x<2)

db11 = function(x) db1(x)*db1(x)
db12 = function(x) db1(x)*db2(x)
db13 = function(x) db1(x)*db3(x)
db14 = function(x) db1(x)*db4(x)
db15 = function(x) db1(x)*db5(x)
db22 = function(x) db2(x)*db2(x)
db23 = function(x) db2(x)*db3(x)
db24 = function(x) db2(x)*db4(x)
db25 = function(x) db2(x)*db5(x)
db33 = function(x) db3(x)*db3(x)
db34 = function(x) db3(x)*db4(x)
db35 = function(x) db3(x)*db5(x)
db44 = function(x) db4(x)*db4(x)
db45 = function(x) db4(x)*db5(x)
db55 = function(x) db5(x)*db5(x)

A = matrix(0,5,5)
A[1,1] = integrate(db11,0,2)$value
A[1,2] = integrate(db12,0,2)$value
A[1,3] = integrate(db13,0,2)$value
A[1,4] = integrate(db14,0,2)$value
A[1,5] = integrate(db15,0,2)$value
A[2,2] = integrate(db22,0,2)$value
A[2,3] = integrate(db23,0,2)$value
A[2,4] = integrate(db24,0,2)$value
A[2,5] = integrate(db25,0,2)$value
A[3,3] = integrate(db33,0,2)$value
A[3,4] = integrate(db34,0,2)$value
A[3,5] = integrate(db35,0,2)$value
A[4,4] = integrate(db44,0,2)$value
A[4,5] = integrate(db45,0,2)$value
A[5,5] = integrate(db55,0,2)$value
A[2:5,1] = A[1,2:5]
```

```
A[3:5,2] = A[2,3:5]
A[4:5,3] = A[3,4:5]
A[5,4] = A[4,5]
A

#Q6
library(AICcmodavg)
data(cement)
cement
attach(cement)
x2
x = (x2 - min(x2)+0.1)/(max(x2 - min(x2))+0.5)*2
Y = y
detach(cement)
plot(x,Y)

#Design matrix
X = matrix(0,13,5)
for(i in 1:13){
  X[i,] = c(b14(x[i]),b24(x[i]),b34(x[i]),b44(x[i]),b54(x[i]))
}
beta_r = function(lamb){
  solve(t(X)%*%X+lamb*A)%*%t(X)%*%Y
}

fnc = function(xob,lamb){

beta_r(lamb)[1]*b14(xob)+beta_r(lamb)[2]*b24(xob)+beta_r(lamb)[3]*b34(xob
)+beta_r(lamb)[4]*b44(xob)+beta_r(lamb)[5]*b54(xob)
}
seq = seq(0,1.99,by=0.01)
#lines(seq,fnc(seq,1000))
#legend(1.5,80,legend="lambda=1000")

par(mfrow=c(3,3))
plot(x,Y,xlab="ambda=0")
lines(seq,fnc(seq,0))
plot(x,Y,xlab="ambda=0.0001")
```

```
	lines(seq,fnc(seq,0.0001))
	plot(x,Y,xlab="ambda=0.05")
	lines(seq,fnc(seq,0.05))
	plot(x,Y,xlab="ambda=0.1")
	lines(seq,fnc(seq,0.1))
	plot(x,Y,xlab="ambda=0.5")
	lines(seq,fnc(seq,0.5))
	plot(x,Y,xlab="ambda=1")
	lines(seq,fnc(seq,1))
	plot(x,Y,xlab="ambda=10")
	lines(seq,fnc(seq,10))
	plot(x,Y,xlab="ambda=100")
	lines(seq,fnc(seq,100))
	plot(x,Y,xlab="ambda=1000")
lines(seq,fnc(seq,1000))
```