Variable Selection in Functional Linear Concurrent Regression

Rahul Ghosal and Arnab Maity

15 September 2019

Introduction

This document presents an illustration of the variable selection method proposed in Ghosal and Maity (2019). The whole process is based on using the following steps.

- Using the `preprocess` function to smooth noisy covariates.
- Finally using the `FLCM.select` function, which performs variable selection from the given input data.

All the functions mentioned above are included in the source file `varselect.R`.

Required libraries

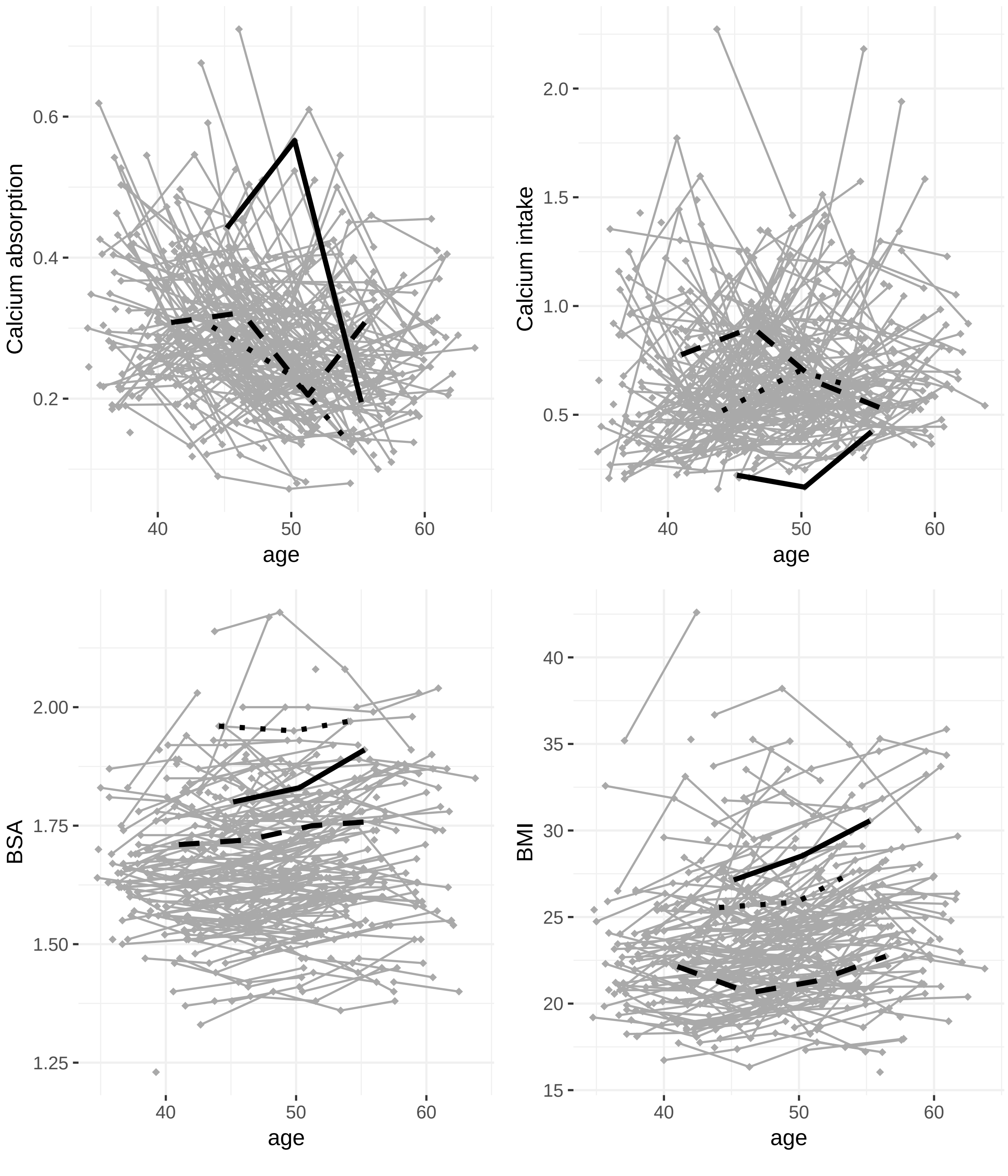
Loading the required libraries

```
library(MASS)
library(mgcv)
library(refund)
library(fda)
library(parallel)
library(grpreg)
```

Loading and plotting the dietary calcium absorption data

```
##load calcium data
calcium<-read.csv("calcium.csv")
#####################################
# @ n=188, p=3                      #
# @ calabs = response
# @ caldiet = covariate
# @ bsa = covariate
# @ bmi = covariate
#####################################
attach(calcium)
library(ggplot2)
library(gridExtra)
temp <- data.frame(x = age[1:10], y =calabs[1:10],group=id[1:10] )
temp2 <- data.frame(x = age[1:10], y =caldiet[1:10],group=id[1:10] )
temp3 <- data.frame(x = age[1:10], y =bsa[1:10],group=id[1:10] )
temp4 <- data.frame(x = age[1:10], y =bmi[1:10],group=id[1:10] )
linet<-c()
linet[1:3]<-c("solid")
linet[4:7]<-c("dashed")
linet[8:10]<-c("dotted")
```

```r
par(mfrow=c(2,2))
par(mar=c(5.1,4.1,4.1,2.1))
#setEPS()
#postscript("calciumsel.eps", width=7, height=10)
pp <- ggplot(calcium,aes(x=age,y=calabs,group=id)) +
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major = element_line(colour = "gray94"),
        panel.grid.minor = element_line(colour = "gray94"))+
  xlab("age") +ylab("Calcium absorption")+geom_line(color='dark gray')+
  geom_point(shape=18,color='dark gray')+
  geom_line(data = temp, aes(x = x, y = y,group=group),linetype=linet,size=1.2)
pp2 <- ggplot(calcium,aes(x=age,y=caldiet,group=id)) +
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major = element_line(colour = "gray94"),
        panel.grid.minor = element_line(colour = "gray94"))+
  xlab("age") +ylab("Calcium intake")+geom_line(color='dark gray')+
  geom_point(shape=18,color='dark gray')+
  geom_line(data = temp2, aes(x = x, y = y,group=group),linetype=linet,size=1.2)
pp3 <- ggplot(calcium,aes(x=age,y=bsa,group=id)) +
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major = element_line(colour = "gray94"),
        panel.grid.minor = element_line(colour = "gray94"))+
  xlab("age") +ylab("BSA")+geom_line(color='dark gray')+
  geom_point(shape=18,color='dark gray')+
  geom_line(data = temp3, aes(x = x, y = y,group=group),linetype=linet,size=1.2)
pp4 <- ggplot(calcium,aes(x=age,y=bmi,group=id)) +
  theme(panel.background = element_rect(fill = "white"),
        panel.grid.major = element_line(colour = "gray94"),
        panel.grid.minor = element_line(colour = "gray94"))+
  xlab("age") +ylab("BMI")+geom_line(color='dark gray')+
  geom_point(shape=18,color='dark gray')+
  geom_line(data = temp4, aes(x = x, y = y,group=group),linetype=linet,size=1.2)
grid.arrange(pp, pp2,pp3, pp4, ncol=2)
```



Preprocessing the data and removing noise

```
y<-calabs       ##Response
mydata<-calcium[,c(1,2,4,5,6)] ##data in long format
names(mydata)[2]<-c("time")     ##1st column id, 2nd time, rest covariates
source('varselect.R')
#Preprocessing Covariates
```

```
mydata<- preprocess(mydata)
```

Adding simulated covariates

```
set.seed(1)                # Set seed for reproducibility
p=15                       # adding 15 simulated covariates
n=length(unique(mydata$id))# 188
A<-matrix(0,n,p)
B<-matrix(0,n,p)
for(i in 1:n)
{for(j in 1:p)
{
  A[i,j]<-rnorm(1,0,2)
}
}
for(i in 1:n)
{for(j in 1:p)
{
  B[i,j]<-rnorm(1,0,2)
}
}
X<-function(i,j,t){A[i,j]*sqrt(2)*sin(pi*j*t/200)+B[i,j]*sqrt(2)*cos(pi*j*t/200)}


for(i in 1:527)
  for(j in 6:20)
  {
    {mydata[i,j]<- X(mydata$id[i],(j-5),mydata$time[i])}}
##Final list of Covariates
names(mydata)[6:20]<-paste("pseudo",1:15)
names(mydata)[-c(1:2)]
```

```
##  [1] "caldiet"   "bsa"       "bmi"       "pseudo 1"  "pseudo 2"
##  [6] "pseudo 3"  "pseudo 4"  "pseudo 5"  "pseudo 6"  "pseudo 7"
## [11] "pseudo 8"  "pseudo 9"  "pseudo 10" "pseudo 11" "pseudo 12"
## [16] "pseudo 13" "pseudo 14" "pseudo 15"
```

Performing Variable Selection

```
#######Inputs############
# @ y= response
# @ mydata = a dataframe in long format column 1=id, column 2 = time,
# rest of the column covariates
# @ nbasis1 = Number of basis functions for intercept
# @ nbasis  = Number of basis functions for regression functions
# @ cvl     = Length of crossvalidation grid for parameter \psi
#########################
# cvl set to 10 for illustration purpose,use cvl>=100
FLCM.select(y,mydata,nbasis1=7,nbasis2=15,cvl=10)
```

```
## $scad
## [1] "caldiet"
##
## $mcp
## [1] "caldiet"
```

Both the SCAD and MCP for FLCM selects only caldiet and discards the pseudovariables.