Supplementary material to
Modelling a response as a function of high-frequency count data: the association between physical activity and fat mass

Nicole H. Augustin,1 Calum Mattocks,2 Julian J. Faraway,1 Sonja Greven3 and Andy R. Ness4
1 Department of Mathematical Sciences, University of Bath, Bath, UK;
2 Centre for Exercise, Nutrition and Health Sciences, School for Policy Studies, 8 Priory Road, University of Bristol, Bristol, UK;
3 Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstr. 33, 80539 Munich, Germany;
4 School of Oral and Dental Science and School of Social and Community Medicine, Bristol Dental School, Lower Maudlin Street, Bristol, UK

May 8, 2015

1 Processing of activity profiles

The accelerometers were set to an epoch time of 1 minute and hence for each child there was a time series of minute by minute accelerometer measurements (counts per minute) over seven days available at the ages 12, 14 and 16. For preprocessing of the accelerometer data we followed Mattocks et al. [3]. Any sequence with more than 10 zeros was replaced by missing values, since these periods were regarded as periods were the monitor was not worn; days with a mean count less than 150 or a mean count of three standard deviations above the overall mean (prior to exclusions) were invalid; days were only included if the monitor was worn at least 600mins (10h); weekly profiles were invalid if less than 3 valid days were observed.

Activity counts greater than 15000 cpm were set to missing, as it is unclear what activity would result in such high counts, and misrecording of activity was most likely. This is in line with two independent validation studies using MTI Actigraph AM7164 [2, 1] where intensity for running or jogging was recorded at a maximum value of around 12500 cpm.
### Example R code for model selection

```r
### attach library
library(mgcv)
### reading in data containing variables, including individual histograms of accelerometer
### data; everything is measured at the same age, say at 12 years
load(file=paste(pathr, "mdatb80noTrans.RData", sep=""))
### variable description: Z are the relative frequencies of histogram with mid-points P;
### Zweekday and Zweekend are the histograms for weekday and weekend respectively;
### Z2D is the two dimensional histogram of counts per minute by hour of the day
### create an index of observations taking account the fact that
### for the week-end/week-day histograms there are missing values
missindex<-!is.na(apply(mdat$sm$Zweekend,1,sum))
set.seed(5)
nosets<-2
### leave out 1/4 of data for validation
nleave<-round(1/4*length(moddat$fat1[missindex]))
sid<-sample(moddat$id,nleave,replace=False)
subs!<-(moddat$id %in% sid) & missindex
kPA<-10
### function for model selection criteria
printresults<-function(gobject,moddat=moddat,sid=sid,y=moddat$fat1){
edf<-sum(gobject$edf)
nobs<-length(gobject$y)
f<-predict(gobject,newdata=moddat)
SE<-(y-exp(f))^2
SE<-SE[moddat$id %in% sid]
PE<mean(SE,na.rm=TRUE)
results<-cbind(nobs(gobject), sum(gobject$edf), summary(gobject)$r.sq,
AIC(gobject), BIC(gobject), PE)
names(results)<-c("N","edf","adj.R-sq","AIC","BIC", "PE")
results
}
### list of models considered
fo<-list(log(fat) ~ sex + mnum_obese+weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad") + s(height,bs="tp",k=40),
log(fat) ~ mnum_obese+weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad") + s(height,bs="tp",k=40),
log(fat) ~ sex + weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad") + s(height,bs="tp",k=40),
log(fat) ~ sex + mnum_obese+weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad") + s(height,bs="tp",k=40),
log(fat) ~ sex + mnum_obese+weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad"))
b <- gam(fo[[1]],data=moddat,method="REML",subset=subs) # log-normal model
### look at residual plots regarding choice of distribution
gam.check(b) # fit ok
### try other options regarding distribution
b.2 <- gam(fat~sex + mnum_obese+weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad") + s(height,bs="tp",k=40),
data=moddat,method="REML",subset=subs, family=gaussian(link=log))
# compare with gaussian but using log link (instead of log-normal)
gam.check(b.2) # fit not so good
b.3 <- gam(fat~sex + mnum_obese+weartime + s(P[-1],k=kPA,by=Z[-1],bs="ad") + s(height,bs="tp",k=40),
data=moddat,method="REML",subset=subs, family=Gamma(link=\log))
# compare with gamma but using log link
gam.check(b.3) # fit better than b.2 but not as good as b

### STEP 1: Check out baseline model 3.1
result<-matrix(NA,nrow=6, ncol=7)
dimnames(result)<-list(c("model 1",dimnames(anova(b)$pTerms.table)[[1]]),
dimnames(anova(b)$pTerms.table)[[1]],c("N", "edf", "adj.R-sq", "AIC", "BIC", "PE", "p-val"))
results[1,1:6]<-printresults(b,moddat,sid,y=moddat$fat)
results[1,7]<-(anova(b)$pTerms.pr,anova(b)$pTerms.table[4])
### these terms are all significant, now check their relative importance
for (i in 2:length(fo)){
text<-gam(fo[[i]],data=moddat,method="REML",subset=subs)
results[1,7]<-printresults(text,moddat,sid,y=moddat$fat)
}
results

### STEP 2 model selection - can we make it more complicated?

cpm estimated from histogram data:
moddat$cpmtest <-moddat$Z %*% moddat$P[1,]
```
### separate histograms for male and females:

```r
moddat$Z1 <- as.numeric(moddat$sex=="female")*moddat$sm$Z
moddat$Z2 <- as.numeric(moddat$sex=="male")*moddat$sm$Z
```

```r
resultsB<-matrix(NA,nrow=6, ncol=6)
dimnames(resultsB)<-list(c("base","+ hist (1)","+ cpm","hist by gender","2Dhist","hist by ME"),
c("N","edf","R.sq","AIC","BIC","PE"))
```

```r
moddat$Z2D<-sweep(moddat$Z2D,1,rowSums(moddat$Z2D),FUN="/")
```

```r
# setting up index so that first bins relating to the lowest intensity bin
# are left out.
ind2D<seq(1,ncol(moddat$P1),80)
moddat$P1<-moddat$P1[-,ind2D]
moddat$P2<-moddat$P2[-,ind2D]
moddat$Z2D<-moddat$Z2D[-,ind2D]
```

#### list of models considered

```r
fo2 <list(
  log(fat) ~ sex * as.factor(mum_obese) + s(height,bs="tp",k=40) + weartime,
  log(fat) ~ sex * as.factor(mum_obese) + s(P[,,-1],k=kPA,by=Z[,,-1],bs="ad")
    + s(height,bs="tp",k=40),
  log(fat) ~ sex * as.factor(mum_obese) + s(P[,,-1],k=kPA,by=Z[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=22[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=21[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=Zweekday[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=Zweekend[,,-1],bs="ad")
    + s(height,bs="tp",k=40) + weartime, ## functions of histogram by sex
  log(fat) ~ sex * as.factor(mum_obese) + te(P1,P2,k=(kPA,10),by=Z2D,
    bs=c("cr","cr")) + s(height,bs="tp",k=40) + weartime, ## 2Dhist
  log(fat) ~ sex * as.factor(mum_obese) + s(P[,,-1],k=kPA,by=22[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=21[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=20[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=19[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=18[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=17[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=16[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=15[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=14[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=13[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=12[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=11[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=10[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=9[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=8[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=7[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=6[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=5[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=4[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=3[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=2[,,-1],bs="ad")
    + s(P[,,-1],k=kPA,by=1[,,-1],bs="ad")
    + s(height,bs="tp",k=40) + weartime, ## functions of histogram by sex)
)
```

```r
for (i in 1:length(fo2)){
test <- gam(fo2[[i]],data=moddat,method="REML",subset=subs)
resultsB[i,]<-printresults(test,moddat,sid,y=moddat$fat)
}
```

```r
resultsB
```

### 3. checking whether non-linear part of smooth is significant using
### thin plate regression splines where it is possible to remove the linear part
### # from the smooth

```r
ba <- gam(log(fat) ~ sex * as.factor(mum_obese) + s(P[,,-1],k=kPA,by=22[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=21[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=20[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=19[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=18[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=17[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=16[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=15[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=14[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=13[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=12[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=11[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=10[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=9[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=8[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=7[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=6[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=5[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=4[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=3[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=2[,,-1],bs="ad")
  + s(P[,,-1],k=kPA,by=1[,,-1],bs="ad")
  + s(height,bs="tp",k=40) + weartime, data=moddat,method="REML",subset=subs)
summary(ba)
anova(ba)
```

### References

