BIOS 7720
Homework 3
Due 05/13/2021

For all problems which involve simulation or fitting models to data, code to reproduce results must be included in your solutions.

Students are allowed to work together in groups of up to 3 on this assignment, though all students must submit their own independent write-ups. For coding problems, students may work together on designing and strategizing methods for implementation, but coding up of solutions must be done independently. Please include the name of any other students you worked with on this assignments in your submission.

P1. (40 points) Name three characteristics of a data problem/data set (or a variable contained within a data set) that would suggest FDA methods may be useful and/or appropriate in analyzing the data.

P2. (40 points) This question will have you simulate data for a generalized function on scalar regression model and evaluate model performance in terms of estimating fixed effects (accuracy and inference) as well as subject-level predictions. Specifically, you will simulate data according to the model

$$g(E[y_i(s)|x_i, \boldsymbol{b}_i]) = f_0(s) + f_1(s)x_i + b_i(s) \tag{1}$$

where the outcome are binary data and $g()$ is the logit function. That is, model (1) is a GLMM with a binomial outcome. You will perform a simulation study using $R = 100$ simulated datasets of $N = 200$ participants (200 functions) with data according to model (1) above. Use the following simulation settings for all R simulated datasets:

$$\begin{aligned}
x_i &\sim N(0,1) \\
f_0(s) &= 0.5 + (s - 0.5)^2 \\
f_1(s) &= \sin(2\pi s) \\
b_i(s) &= \sum_{k=1}^{4} \phi_k(s)\xi_{ik} \\
&\quad \phi_1(s) = \sqrt{2}\text{sin}(2\pi s) \\
&\quad \phi_2(s) = \sqrt{2}\text{cos}(2\pi s) \\
&\quad \phi_3(s) = \sqrt{2}\text{sin}(4\pi s) \\
&\quad \phi_4(s) = \sqrt{2}\text{cos}(4\pi s)
\end{aligned}$$

You can simulate one of the R datasets using the code below

```
## general set up
N <- 200               # number of participants (functions)
J <- 50                # number of observations per function
sind <- seq(0,1,len=J) # functional domain of observed functions
## fixed effects
f0   <- function(s) 0.5 + (s-0.5)^2 # f_0(s)
f1   <- function(s) sin(2*pi*s)     # f_1(s)
x  <- rnorm(N)                      # x_i
## random effects
K      <- 4              # number of \phi
lambda <- 0.5^(0:(K-1)) # true eigenvalues (defines variance of \xi_i)
# \phi_k
Phi <- sqrt(2)*cbind(sin(2*pi*sind), cos(2*pi*sind),
                     sin(4*pi*sind), cos(4*pi*sind))
# \xi_{ik}
xi_raw <- matrix(rnorm(N*K), N, K)
xi     <- xi_raw %*% diag(sqrt(lambda))
# simulate b_i(s) as \sum_k \xi_ik \phi_k(t)
bi <- xi %*% t(Phi)

# simulate linear predictor by adding the fixed effects ot the random effects
# note that vapply will return the results in matrix format with one column per function
# instead of the more common 1 row per function
fixed_eff <- vapply(1:N, function(i, sind) f0(sind) + f1(sind)*x[i], numeric(J),
                    sind=sind)
# because the fixed effect matrix is transposed, need to transpose the ranom effect matrix b_i(s)
eta       <- fixed_eff + t(bi)
# simulate outcome by applying g^{-1}
expit     <- function(x) 1/(1+exp(-x))
ptrue     <- expit(eta)
Y <- vapply(ptrue, function(x) rbinom(n=1, size=1, prob=x), numeric(1))
Y <- matrix(Y, N, J, byrow=TRUE)
```

For each simulated dataset you should fit the relevant GFoSR model by creating the appropriate data frame and calling *mgcv::bam()* as follows:

```
library("mgcv")
df_fit <- data.frame(y=as.vector(t(Y)),
                     x=rep(x, each=J),
                     id=factor(rep(1:N, each=J)),
                     sind = rep(sind, N))
fit <- bam(y ~ s(sind,k=10) +
               s(sind, by=x, bs="cr",k=10) +
               ti(id, sind,bs=c("re","cr"), k=c(5,5), mc=c(TRUE, FALSE)),
           data=df_fit, method="fREML", family=binomial, discrete=TRUE)
```

(a) (20 points) For each simulated data set $R = 1, \ldots, 100$ evaluate fixed effects estimation accuracy and inference at each point along the functional domain $(s_1, \ldots, s_J)$ according to the criteria:

- $\text{MSE}_r^f(s) = (\hat{f}_{kr}(s) - f_{kr}(s))^2$
- 95% Coverage Probability$(s) = 1(\hat{f}_{kr}(s) - 2\text{SE}(\hat{f}_{kr}(s)) \leq f_{kr}(s) \leq \hat{f}_{kr}(s) + 2\text{SE}(\hat{f}_{kr}(s)))$

where $\hat{f}_{kr}(s)$ is the k^{th} estimated coefficient $(k = 0, 1)$ for the r^{th} simulated dataset. Summarize your results and support your claims using tables or figures.

(b) (20 points) For each simulated data set evaluate subject prediction accuracy on the response scale at each point along the functional domain $(s_1, \dots, s_J)$ using the criteria

$$\text{MSE}_r^y(s) = \sum_{i=1}^{N} (\hat{y}_{ir}(s) - y_{ir}(s))^2$$

where $y_{ir}(s)$ is the response (0/1) value for subject i at point s on the functional domain in the r^{th} dataset and $\hat{y}_{ir}(s)$ is $\hat{E}[y_i(s)|x_i, \tilde{\boldsymbol{b}}_i]$ (i.e. predicted probability that participant/function i's response at s is 1). Summarize your results and support your claims using tables or figures.