BIOS 7720
Homework 2
Due 04/26/2021

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.

Throughout this homework we'll be using a few packages for data processing/analysis. In particular, we'll be using functions from the tidyverse to do data tidying/manipulation, and functions from the *mgcv* and *refund* packages for data analysis. I load them below.

```
library("tidyverse")
library("mgcv")
library("refund")
```

P1. (20 points) This question relates to estimating scores from fPCA. Consider the fPCA model:

$$\begin{aligned} y_i(s) &= f_0(s) + b_i(s) + \epsilon_i(s) \\ &= f_0(s) + \sum_{k=1}^{K} \xi_{ik}\phi_k(s) + \epsilon_i(s) \\ \boldsymbol{\xi}_{K\times 1} &\sim N(\mathbf{0}_{K\times 1}, \boldsymbol{\Sigma}_b) \\ \epsilon_i(s) &\overset{\text{iid}}{\sim} N(0, \sigma^2_\epsilon) \end{aligned}$$

We will take advantage of the fact that the residuals and $\boldsymbol{\xi}$ are normal to derive the conditional distribution of $\boldsymbol{x}_i$ (the random effects) given the observed $\boldsymbol{y}_i$.

Now, suppose we observe y_i on a grid of $s_1, \ldots, s_J$. Denote $\boldsymbol{y}_i = [y_i(s_1) \cdots y_i(s_J)]^t$. With slight abuse of notation let $f_0(\boldsymbol{s}) = [f_0(s_1) \cdots f_0(s_J)]^t$. In matrix notation we have

$$\begin{bmatrix} \boldsymbol{y}_{J\times 1} \\ \boldsymbol{\xi}_{K\times 1} \end{bmatrix} = \begin{bmatrix} f_0(\boldsymbol{s}) \\ \mathbf{0}_{K\times 1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Phi}_{J\times K} & \boldsymbol{I}_{J\times J} \\ \boldsymbol{I}_{K\times K} & \mathbf{0}_{K\times J} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_{K\times 1} \\ \boldsymbol{\epsilon}_{J\times 1} \end{bmatrix}$$

Note I dropped the subscript i on both $\boldsymbol{y}_i$ and $\boldsymbol{\xi}_i$ above to provide explicit dimensions for your convenience. Then, taking the variance of the quantity above,

it is easy to see that

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{\xi}_i \end{bmatrix} \sim N\left(\begin{bmatrix} f_0(\boldsymbol{s}) \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Phi}\boldsymbol{\Sigma}_b\boldsymbol{\Phi}^t + \sigma_\epsilon^2\boldsymbol{I} & \boldsymbol{\Phi}\boldsymbol{\Sigma}_b \\ \boldsymbol{\Sigma}_b\boldsymbol{\Phi}^t & \sigma_\epsilon^2\boldsymbol{I} \end{bmatrix} \right)$$

Denote $\boldsymbol{\Omega} = \boldsymbol{\Phi}\boldsymbol{\Sigma}_b\boldsymbol{\Phi}^t + \sigma_\epsilon^2\boldsymbol{I}$. Using the properties of the multivariate normal, it follows that

$$E[\boldsymbol{\xi}_i|\boldsymbol{y}_i] = \boldsymbol{\Sigma}_b\boldsymbol{\Phi}^t\boldsymbol{\Omega}^{-1}(\boldsymbol{y} - f_0(\boldsymbol{s})) \tag{1}$$

```
set.seed(19840)
# simulation settings
N  <- 100  # number of functions to simulate
ns <- 50   # number of observations per function
sind <- seq(0,1,len=ns) # functional domain of observed functions
K  <- 4    # number of true eigenfunctions
lambda <- 0.5^(0:(K-1))   # true egenfunctions
sig2 <- 2  # error variance
# set up true eigenfunctions
Phi <- sqrt(2)*cbind(sin(2*pi*sind), cos(2*pi*sind),
                     sin(4*pi*sind), cos(4*pi*sind))
# simulate coefficients
# first, simulate standard normals, then multiply by the
# standard deviation to get correct variance
xi_raw <- matrix(rnorm(N*K), N, K)
xi     <- xi_raw %*% diag(sqrt(lambda))
# simulate functional responses as \sum_k \xi_ik \phi_k(t)
x <- xi %*% t(Phi)
y <- x + matrix(rnorm(N*ns, mean=0, sd=sqrt(sig2)), N, ns)
```

(a) (5 points) Fit fPCA to the simulated data above using *refund::fpca.face()* to the simulated data above (y in the code).

(b) (5 points) Using Equation (1) above, manually obtain estimated scores from the fPCA fit via *refund::fpca.face()* simulated data. You may plug in the estimated quantities $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{b}}$, $\hat{\sigma}_\epsilon^2$, $\hat{f}_0(s)$ obtained from your fPCA fit. Note that by default *refund::fpca.face()* does not return $\hat{\sigma}_\epsilon^2$, but it will if you specify the argument var=TRUE to your function call.

(c) (5 points) Obtain estimated scores using the numeric integration approach discussed in class.

(d) (5 points) Compare the true scores to those returned from *refund::fpca.face()* and those you manually derived in parts (b) and (c).

P2. (20 points) In this problem we will apply fPCA to a dataset on Sequential Organ Failure Assessment (SOFA) scores contained in the *refund* dataset. First, load the sofa dataset using the *R* code below.

```r
data(sofa)
str(sofa)
```

```
## 'data.frame': 520 obs. of  7 variables:
##  $ death   : logi  FALSE FALSE TRUE TRUE FALSE FALSE ...
##  $ SOFA    : int [1:520, 1:173] 9 11 14 4 7 13 4 12 11 14 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:520] "1" "2" "3" "4" ...
##   .. ..$ : NULL
##  $ SOFA_raw: int [1:520, 1:173] 9 11 14 4 7 13 4 12 11 14 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:520] "1" "2" "3" "4" ...
##   .. ..$ : NULL
##  $ los     : int  30 16 157 15 9 22 16 8 4 8 ...
##  $ age     : int  59 49 72 52 80 43 43 38 48 68 ...
##  $ male    : logi  FALSE FALSE TRUE TRUE TRUE FALSE ...
##  $ Charlson: int  4 3 2 8 8 8 0 1 6 2 ...
```

The dataset contains SOFA scores on patients hospitalized in the ICU with acute lung injury. Individuals are given one SOFA score per day for each day they're in the ICU. Individuals data records end when they are either discharged from the ICU or experience mortality. Higher SOFA scores indicate a worse health state. In addition to SOFA scores, the dataset contains information on individuals' age, gender, and Charlson co-morbidity index (a measure of health status with higher being worse) information at baseline.

In the following questions you will investigate the association between SOFA score trajectories and mortality. Because of the survival/censoring processes involved, analyzing the full trajectories can be challenging and is beyond the scope of this course. Here, we will focus on analyzing 30-day SOFA trajectories among those individuals who were alive in the ICU for at least 30 days following admission. There is some missingness in the raw data ("SOFA_raw"). These missing data are imputed using last observation carry forward (LOCF) in the "SOFA" element of the sofa dataset.

Below, I provide the code to create this truncated 30-day sample which is entered into the dataframe as "SOFA_trunc".

```r
ndays <- 30
df_P2 <-
   sofa %>%
   filter(los > ndays)
df_P2[["SOFA_trunc"]] <- df_P2$SOFA[,1:ndays]
```

(a) (5 points) Visualize the 30-day SOFA score trajectories separately by mortality status ("death" variable in the dataset where TRUE indicates mortality and FALSE indicates discharge).

(b) (5 points) Apply fPCA to the 30-day data using your preferred method (from class or otherwise). For the first 4 estimated eigenfunctions, plot $\hat{\mu}(s) + / -$

$2\phi_k(s)\text{SD}(\xi_k)$ separately for $k = 1, \ldots, 4$ for day $s = 1, \ldots, 30$. Based on these plots, provide an interpretation of the first four PCs.

(c) (5 points) Fit a logistic regression of eventual mortality on the first four PC scores. Use your results from the previous sub-question to interpret your results. Then, estimate the discriminitive performance of your model using AUC. Compare the in-sample AUC to 10-fold cross-validated AUC. Comment on your results.

(d) (5 points) Use the connection between regression on the PC scores and the generalized functional linear model to plot the estimated $\hat{\gamma}(s)$ plus/minus two (pointwise) standard errors implied by the fitted model. Interpret your results.

P3. (20 points) This problem also makes use of the SOFA dataset. Here we will consider the SOFA score as our outcome and explore the association between baseline covariates and SOFA trajectories.

(a) (5 points) Set up the dataframe in long format for fitting function-on-scalar regression. Include covariate data on age, gender, and charlson score.

(b) (5 points) Fit the "naive" varying coefficient model

$$\begin{aligned} \text{SOFA}_i(s) &= f_0(s) + f_1(s)\text{Age}_i + f_2(s)\text{Charlson}_i + \epsilon_i(s) \\ \epsilon_i(s) &\overset{\text{iid}}{\sim} N(0, \sigma^2_\epsilon) \end{aligned}$$

and plot the estimated $\hat{f}_k(s) + / - 2\text{SE}(\hat{f}_k(s))$ for $k = 0, 1, 2$

(c) (5 points) Does the independent residual assumption seem reasonable here? Justify your claim using relevant numeric summaries and/or figures.

(d) (5 points) Estimate the functional random intercept model

$$\begin{aligned} \text{SOFA}_i(s) &= f_0(s) + f_1(s)\text{Age}_i + f_2(s)\text{Charlson}_i + b_i(s) + \epsilon_i(s) \\ b_i &\sim \text{GP}(0, \Sigma_b) \\ \epsilon_i(s) &\overset{\text{iid}}{\sim} N(0, \sigma^2_\epsilon) \end{aligned}$$

using the iterative approach described in class. Note, you do not need to actually iterate the procedure, you just need to use the method described for estimating $\boldsymbol{\phi}$ from the residuals of the naive model. Plot the estimated $\hat{f}_k(s) + / - 2\text{SE}(\hat{f}_k(s))$ for $k = 0, 1, 2$. Comment on any difference you see from the naive model.

P4. (20 points) When applying FDA we always want to keep in mind simpler models when/where appropriate. Propose one simpler model for each of problems 2 and 3 which still includes all the covariates in some form (e.g. don't drop age from

the model in problem 3). Compare your proposed simpler model to the more complicated model using the model performance measures: AUC (problem 2) and MSE (problem 3). Comment on your findings. You may use either in-sample or cross-validated predictive performance.