Applied functional data analysis (Stat 8550)

the sample midterm two

1. What is the main purpose of the principal component analysis (PCA)? write down the expression of the first PC score of he functional PCA in terms of the original sample function?

- Based on the following R codes and output, answer the following questions,

```
> library(fda)
>  X=as.matrix(read.table("CO.txt"))
>  X=scale(X,center=TRUE, scale=FALSE)
>  t=0:23
>  basis.obj=create.bspline.basis(c(0,24), nbasis=50)
>  fdParobj = fdPar(basis.obj, 2, 1e-8)
>  fit.list=smooth.basis(t, t(X), fdParobj)
>  CO.fd=fit.list$fd
> pac.Par=fdPar(basis.obj, 2, 1e-3)
> pca.list=pca.fd(CO.fd, 2, pac.Par)
> str(pca.list)
List of 5
 $ harmonics:
 $ values   :
 $ scores   :
 $ varprop  :
 $ meanfd   :
```

2. How many PC components do the R codes calculate?
3. Write down the optimization problem for the first PC function for the R function "pca.fd"?
4. There are two smoothness tuning parameters 10^{-8} and 10^{-3}. What are the purposes of the two tuning parameters and the corresponding smoothness penalties?
5. What are the meanings of the five components in the output "pca.list".
6. What is the main purpose of the functional LDA? What is the main difference between functional LDA and functional PCA?
7. Write down the definitions of the between-class variance and the within-class variance?

- Consider the fitting.

```
> source("scalar.on.function.R")
> ncurves=1
> fit.cv=cv.sigcomp(t.x.list, X.list, Y, s.n.basis=40)
> Y.pred=pred.sigcomp(fit.cv, X.list)
> residual=Y.pred-Y
> y.bar=apply(Y,2,mean)
> 1-sum((residual[,1])^2)/sum((Y[,1]-y.bar[1])^2)
[1] 0.9693437
> 1-sum((residual[,2])^2)/sum((Y[,2]-y.bar[2])^2)
[1] 0.9087465
 >  cov(Y)
         V1       V2
V1 117.1167  657.636
V2 657.6360 7308.746
```

8. In function-on-scalar regression, what is the difference between functional PC regression and PLS regression?

9. Based on the R code above, what are the R^2 for the two responses, respectively? What is the combined R^2?

10. Write down the function-on-function model?

11. Write down the definitions of MSIPE and MSIEE?

Solution to sample mid 2

1. The main purpose is the dimension reduction. That is, we transfer the original variables which are defined in a high-dimensional or infinite-dimensional space to a new variable defined in a lower dimensional space which can capture most variations in the original data.

$$Y_i = \begin{pmatrix} Y_{i1} = \int_a^b X_i(s)\beta_1(s)\,ds \\ Y_{i2} = \int_a^b X_2(s)\beta_1(s)\,ds \\ \vdots \\ Y_{in} = \int_a^b X_n(s)\beta_1(s)\,ds \end{pmatrix}$$

2. Two components

3. $$\max_{\beta} \frac{\iint \beta(s)\Sigma_X(s,t)\beta(t)\,ds\,dt}{\int \beta^2(s)\,ds + \lambda \int \beta''(s)^2\,ds} \quad \text{subject to } \int \beta^2(s)\,ds = 1$$

where $\Sigma_X(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}[X_i(s)-\bar{X}_\cdot(s)][X_i(t)-\bar{X}_\cdot(t)]$ is the covariance function.

4. $10^{-\delta}$ is the tuning parameter for smoothing $X_i(s)$.

10^{-3} is ---------- for smoothing $\beta(s)$

5. "harmonics" contains the PC functions

"values" – eigenvalues "scores" – PC scores

"varprop" – proportions of ~~variation~~ variations explained by each component.

"meanfd" – functional object of the mean curve.

6. To find new variables to make the between class variance as large as possible and make the within-class variance as small as possible.

FPAC only concerns the variation in $X(s)$

FLDA focus on the relationship between $X(s)$ and the class label Y.

7. $$\Sigma_B = \frac{1}{n-k}\sum_{k=1}^{K} n_k(\overline{y_\bullet^{(k)}} - \overline{y_{\bullet\bullet}})^2$$

$$\Sigma_W = \frac{1}{n-k}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\overline{y_i^{(k)}} - \overline{y_\bullet^{(k)}})^2$$

8. FPCA finds new predictors by maximizing the variace of $\int_a^b X(s)\beta(s)ds$, whereas FPLS finds new predictors by maximizing the covariance

$$Cov\left(Y, \int_a^b X(s)\beta(s)ds\right)$$

9 $R_1^2 = 0.969$, $R_2^2 = 0.908$

$$R^2_{comb} = \frac{0.969 \times 117.11 + 0.9087 \times 7308.7}{117.11 + 7308.7} = 0.9087$$

10. $$Y(t) = \mu(t) + \int_a^b X(s)\beta(s,t)\,ds + \varepsilon(t)$$

11. $$MSIPE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \int_a^b \left(Y_i^{(pred)}(t) - Y_i^{(test)}(t) \right)^2 dt$$

$$MSIEE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \int_a^b \left(Y_i^{(pred)}(t) - Y_i^{(test)}(t) + \varepsilon_i^{(test)}(t) \right)^2 dt$$