Práctica 2

1. Muestre que

   (a) La primer dirección principal maximiza $\text{VAR}(\langle \alpha, X \rangle)$ sobre $\mathcal{S}_1 = \{\alpha : \|\alpha\| = 1\}$.

   (b) La $(k+1)-$ésima dirección principal maximiza $\text{VAR}(\langle \alpha, X \rangle)$ sobre $\mathcal{S}_1 = \{\alpha : \|\alpha\| = 1\}$ sujeto a $\text{COV}(\langle \alpha, X \rangle, \langle \phi_j, X \rangle) = 0$, para $1 \leq j \leq k$.

   (c) Deduzca que como en el caso multivariado, las componentes principales tienen varianza λ_j y son no correlacionadas.

2. Considere el conjunto de datos `CanadianWeather` de la librería `fda` y los promedios mensuales de las 35 estaciones, `CanadianWeather$monthlyTemp`.

   (a) A partir del núcleo de covarianza empírico, $\widehat{\gamma}(t,s)$, obtenga estimadores de las 4 primeras direcciones principales. Qué porcentaje de la variabilidad total explican?

   (b) Para cada $j = 1, \ldots, 4$, haga un plot con los datos en gris, la media $\widehat{\mu}(t)$ en negro y las curvas $\widehat{\mu}(t) + \gamma\,\widehat{\phi}_j(t)$ y $\widehat{\mu}(t) - \gamma\,\widehat{\phi}_j(t)$ en rojo, tomando $\gamma = 5$ y 10. Qué observa?

   (c) Haga un gráfico de los escores $\widehat{\xi}_{i,j}$ versus $\widehat{\xi}_{i,k}$, $1 \leq i \leq n$, $1 \leq k < j \leq 4$, donde $\widehat{\xi}_{i,\ell} = \langle X_i - \widehat{\mu}, \widehat{\phi}_\ell \rangle$

   (d) Repita el análisis después de haber eliminado los datos detectados como atípicos por el boxplot funcional. Qué observa?

3. En el Rdata `lipdata.Rdata` se dan los datos correspondientes al movimiento del labio inferior al pronunciar la palabra *bob*. Los registros corresponden a 32 repeticiones de dicha palabra y las 501 mediciones se tomaron en el intervalo $[0, 0.69]$.

   (a) Grafique las curvas junto con el estimador $\widehat{\mu}$ de su media.

   (b) Realice el boxplot funcional de los datos e identifique los datos atípicos si los hubiere.

   (c) Grafique la superficie $\widehat{\gamma}(t,s)$.

   (d) A partir de la covarianza empírica $\widehat{\gamma}(t,s)$ obtenga estimadores de las 5 primeras direcciones principales. Qué porcentaje de la variabilidad total explican?

   (e) Para cada $j = 1, \ldots, 4$, haga un plot con los datos en gris, la media $\widehat{\mu}(t)$ en negro y las curvas $\widehat{\mu}(t) + \gamma\,\widehat{\phi}_j(t)$ y $\widehat{\mu}(t) - \gamma\,\widehat{\phi}_j(t)$ en rojo, tomando $\gamma = 0.5$ y 1. Qué observa?

   (f) Haga un gráfico de los escores $\widehat{\xi}_{i,j}$ versus $\widehat{\xi}_{i,k}$, $1 \leq i \leq n$, $1 \leq k < j \leq 4$, donde $\widehat{\xi}_{i,\ell} = \langle X_i - \widehat{\mu}, \widehat{\phi}_\ell \rangle$.

   (g) Repita el análisis después de haber eliminado las observaciones 24,25 y 27. Qué observa?

4. Considere un proceso $X \in L^2([-1,1])$ con un desarrollo finito de Karhunen–Loève,

$$X = Z_1\phi_1 + Z_2\phi_2 + Z_3\phi_3 \tag{1}$$

donde $\phi_1(t) = \sin(4\pi t)$, $\phi_2(t) = \cos(7\pi t)$ y $\phi_3(t) = \cos(15\pi t)$, $t \in [-1, 1]$. Supongamos que $Z_j \sim N(0, \sigma_j^2)$ donde $\sigma_1 = 4$, $\sigma_2 = 2$ and $\sigma_3 = 1$, Z_j independientes entre sí.

(a) Calcule $\mathbb{E}X$ y Γ el operador de covarianza de X. Muestre que Γ tiene rango finito. Cuanto valen sus autovalores y sus autofunciones? Podría haberlo deducido directamente de (1)?

(b) Se consideran ahora observaciones X_i, $1 \leq i \leq n = 50$ del proceso X dado en (1), es decir,

$$X_i = Z_{i1}\phi_1 + Z_{i2}\phi_2 + Z_{i3}\phi_3$$

donde $Z_{ij} \sim N(0, \sigma_j^2)$ son independientes para $1 \leq i \leq n$ y $1 \leq j \leq 3$. Fijando la semilla en 1223, genere las observaciones sobre una grilla $\{t_j\}$ de puntos equiespaciados de longitud 1000 y grafiquelas.

(c) Realice el boxplot funcional de los datos e identifique los datos atípicos si los hubiere.

(d) Grafique la superficie $\widehat{\gamma}(t, s)$.

(e) A partir de la covarianza empírica $\widehat{\gamma}(t, s)$ obtenga estimadores de las primeras direcciones principales. Cuántas tiene sentido tomar? Cuántas tomaría para explicar un 95% de la variabilidad total?

(f) Para cada j, grafique la $j-$ésima dirección principal real y estimada en un mismo gráfico. Asegurese que signo$(\langle\widehat{\phi}_j, \phi_j\rangle) = 1$.

(g) Para evaluar el efecto de datos atípicos en los estimadores de las direcciones principales, considere las siguientes contaminaciones

C_1: Sean $B_i \sim Bi(1, 0.1)$, $1 \leq i \leq n$ independientes. Defina

$$X_i^{(1)} = \begin{cases} X_i & \text{si } B_i = 0 \\ X_i + 12 & \text{si } B_i = 1 \end{cases}$$

corresponde a sumar a un 10% de las trayectorias un factor 12.

C_2: Reemplace $X_2(t)$ por $X_2(t) + 25$ cuando $-0.4 < t < -0.36$. Llamaremos $X_i^{(2)}$ a las trayectorias obtenidas

C_3: Defina

$$X_i^{(3)} = Z_{i1}^{(3)}\phi_1 + Z_{i2}^{(3)}\phi_2 + Z_{i3}^{(3)}\phi_3$$

donde $Z_{i1}^{(3)} \sim N(0, \sigma_1^2)$,

$$\begin{pmatrix} Z_{i2}^{(3)} \\ Z_{i3}^{(3)} \end{pmatrix} \sim (1-\epsilon)\, N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \text{diag}\left(\sigma_2^2, \sigma_3^2\right)\right) + \epsilon\, N\left(\begin{pmatrix} 4 \\ 4 \end{pmatrix}, \text{diag}\left(0.01, 0.01\right)\right)$$

con $\epsilon = 0.1$. Es decir, se generan $B_i \sim Bi(1, 0.1)$, $1 \leq i \leq n$ independientes e independientes de

$$W_i = (W_{i1}, W_{i2})^{\mathrm{T}} \sim N\left(\begin{pmatrix} 4 \\ 4 \end{pmatrix}, \text{diag}\left(0.01, 0.01\right)\right),$$

entonces $(Z_{i2}^{(3)}, Z_{i3}^{(3)})$ se define como

$$\left(Z_{i2}^{(3)}, Z_{i3}^{(3)}\right)^{\mathrm{T}} = \begin{cases} (Z_{i2}, Z_{i3})^{\mathrm{T}} & \text{si } B_i = 0 \\ (W_{i2}, W_{i3})^{\mathrm{T}} & \text{si } B_i = 1 \end{cases}$$

i. Grafique las trayectorias obtenidas en cada una de las contaminaciones, indicando en rojo las trayectorias contaminadas.

ii. En que caso, es más dificil distinguir los datos atípicos generados de los datos originales?

iii. Haga un boxplot de los escores

$$\xi_{i,j} = \langle X_i, \phi_j \rangle$$

para $1 \leq j \leq 3$, $1 \leq i \leq n$ en cada caso. Si el boxplot identifica outliers a quién corresponden? Interprete.

iv. Para cada $\ell = 1, 2, 3$ llamemos $\widehat{\phi}_j^{(\ell)}$ las direcciones principales estimadas obtenidas con la muestra $\{X_i^{(\ell)}\}_{i=1}^n$. Para cada $\ell = 1, 2, 3$ y $1 \leq j \leq 3$, grafique $\widehat{\phi}_j^{(\ell)}$ y $\widehat{\phi}_j$ en un mismo grafico, asegurese que signo$(\langle \widehat{\phi}_j^{(\ell)}, \widehat{\phi}_j \rangle) = 1$. Qué observa? Es razonable? Explique.

v. Para entender los resultados, calcule $\mathbb{E}\left(X_1^{(\ell)}\right)$ y $\textsc{Var}\left(\langle X_1^{(\ell)}, \phi_j \rangle\right)$ para las contaminaciones C_1 y C_3 y $j = 1, 2, 3$.