# Análisis de datos funcionales - Taller 3

Valeria Bejarano - Camilo Avellaneda

## Contents

(1) Usando el dataset TECATOR:

**Descripción de la base:**

Este trabajo hace uso de la paquetería *fda*, *fda.usc*, *roahd* y *tidyverse* [Ramsay et al., 2020, Febrero-Bande and Oviedo de la Fuente, 2012, Ieva et al., 2019] del software R [R Core Team, 2020]. De manera preliminarse realiza un análisis exploratorio con el fin de visualizar las curvas en el dataset en general y según su contenido de grasa, con el objetivo de determinar qué conjunto de funciones base es el apropiado para realizar el suavizado. Las siguientes figuras muestran las curvas observadas, para todo el conjunto de curvas en el Gráfico 1 y de manera desagregada por su contenido de grasa en el Gráfico 2. Como no se observan comportamientos cíclicos se utiliza un conjunto base de B-splines, con lo que se procede a estimar el valor del parámetro de penalización por curvatura $\lambda$.

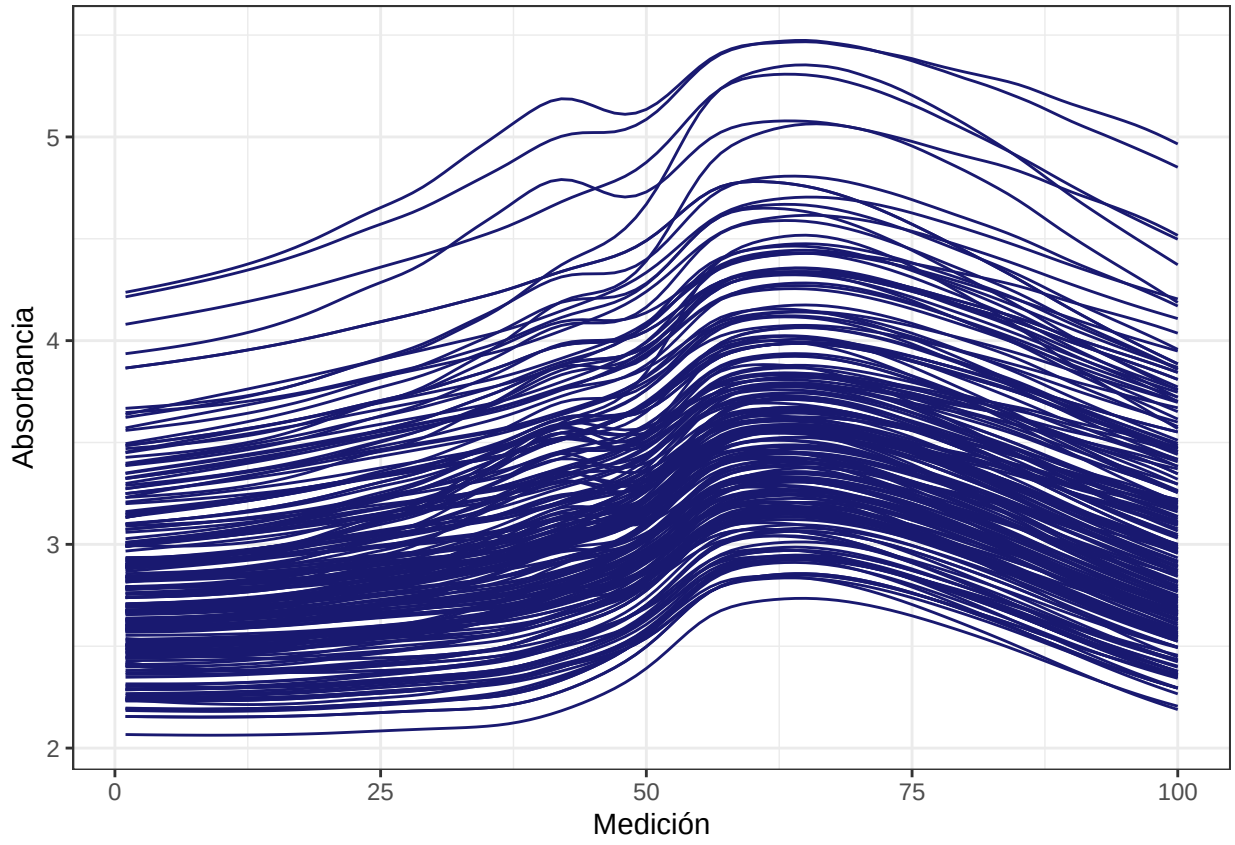


Gráfico 1: Gráfico de las realizaciones correspondientes al proceso de absorbancia.

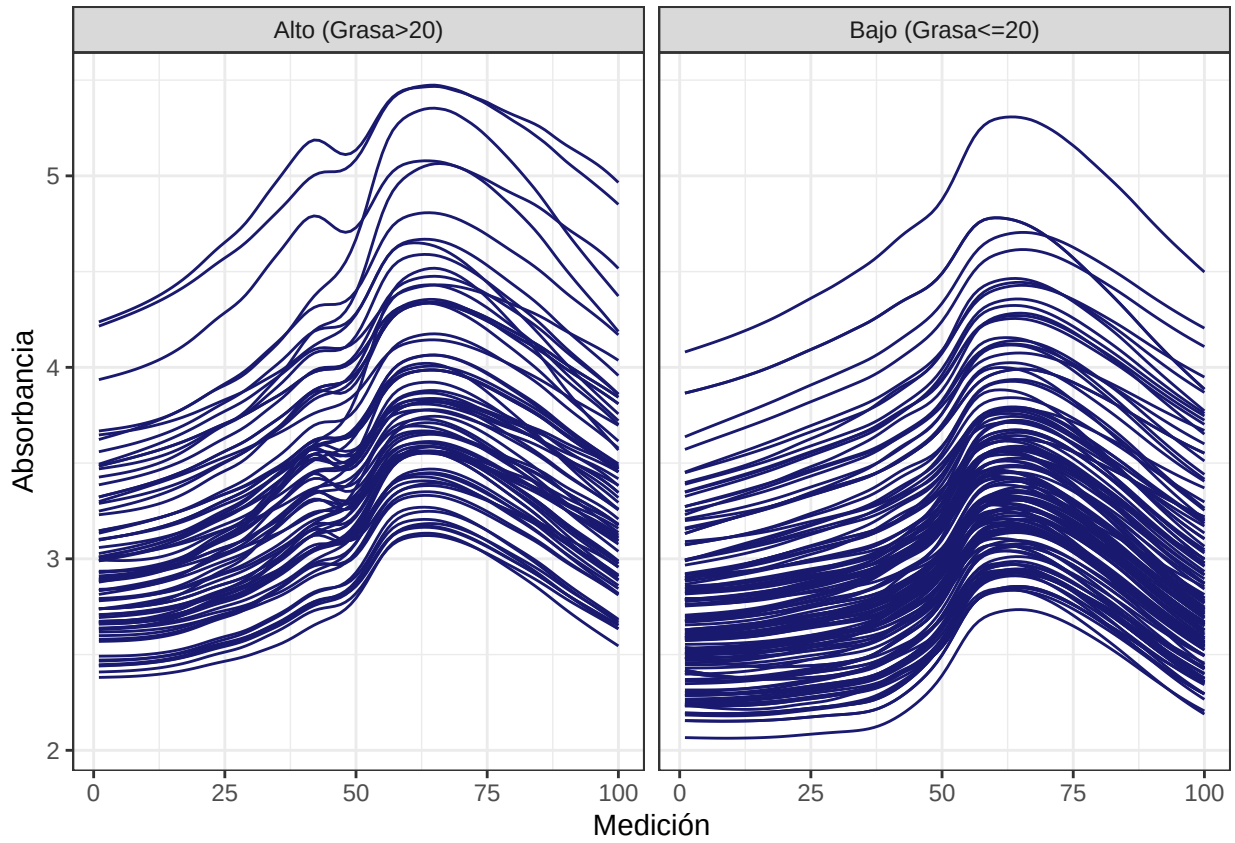


Gráfico 2: Gráfico de las realizaciones correspondientes al proceso de absorbancia, según su contenido de grasa (grupo alto y grupo bajo).

Gráfico 5: Curvas correspondientes al grupo alto en grasa, junto con la curva media y la media global de todas las observaciones en azul.

Para determinar el valor de $\lambda$ se construyó una función que utiliza una grilla de valores y a partir de cada uno de ellos se calcula el coeficiente de validación cruzada generalizada y se selecciona el $\lambda$ que minimiza dicho criterio, como se muestra en los siguientes comandos.

```
df_gcv <- map_dfr(seq(-5,5,by=0.01),~Select_lambda(x= 1:100 , Y_mat= base::t(absorp), lambda=.x))
lambda_optimo <- df_gcv[which.min(df_gcv$gcv),"lambda"]
```

Una vez se suavizan las curvas, se procede a validar los supuestos que son requeridos para llevar a cabo las pruebas de hipótesis de manera óptima. Para esto, se tienen en cuenta tres aspectos, que son:

1. Exclusión de curvas atipicas,
2. Independencia entre las curvas y
3. Verificación de la existencia de puntos de cambio.

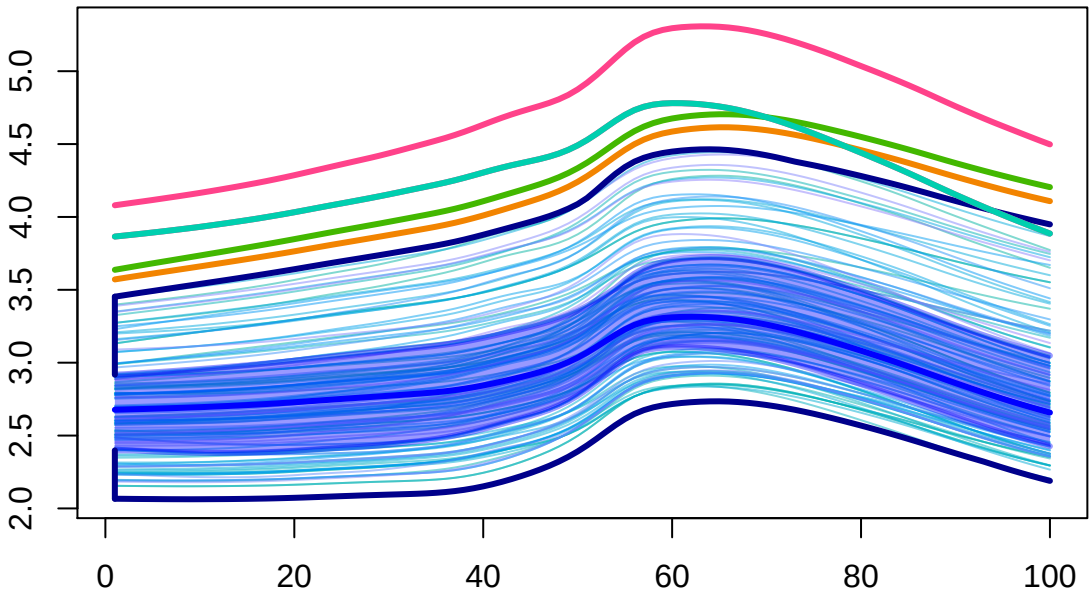


Gráfico 3: Boxplot funcional para el grupo de curvas bajo en grasa.

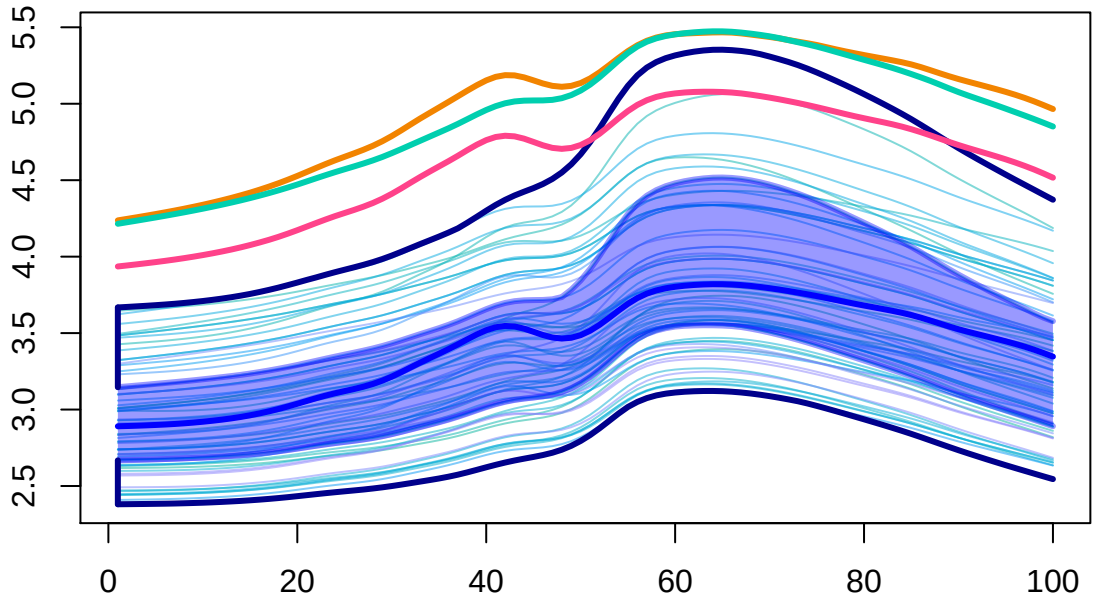


Gráfico 3: Boxplot funcional para el grupo de curvas bajo en grasa.

Para el primer caso, se elaboraron gráficos boxplot para cada grupo, los cuales se muestran en los gráficos 3 y 4. Las curvas de colores diferentes al azul representan curvas atipicas, por lo cual son excluidas para análisis posteriores.

En los procedimientos que se describen a continuación, cuando se requiere el cálculo de la norma $\langle \cdot, \cdot \rangle$ correspondiente a funciones en el espacio $\mathcal{L}^2$ se utilizó la función *inprod* del paquete $fda$. Por otro lado, para verificar el supuesto de independencia, se desea probar

$$H_0 : \text{Las observaciones son independientes vs}$$
$$H_a : \text{Las observaciones no son independientes,}$$

para lo que se realiza el cálculo de la estadística de prueba que se muestra en la ecuación (1), donde $\hat{C}_h$ es la matriz de covarianzas entre scores con un rezago h, mientras que $\hat{C}_0$ es la matriz de varianzas y covarianzas entre scores, que es igual a la matriz diagonal con los valores propios del operador de covarianza. De manera adicional, a partir de ejercicios previos, se observó que es suficiente trabajar con una función propia, ya que ésta captura un gran porcentaje de variabilidad.

$$Q = \sum_{k=1}^{N-1} tr[\hat{C}_h' \hat{C}_0^{-1} \hat{C}_h \hat{C}_0^{-1}]. \tag{1}$$

```
Test_altos <- Independence_Q(x = 1:100, Y_mat = Matrix_alto_without,
lambda = lambda_optimo, n_pca = 1)
Test_altos
```

```
## [1] 0.2519135
```

```
Test_bajos <- Independence_Q(x = 1:100, Y_mat = Matrix_bajo_without,
lambda = lambda_optimo, n_pca = 1)
Test_bajos
```

```
## [1] 0.5771531
```

Para la determinación del valor $p$ asociado a cada prueba, se hace uso de la distribución $\chi^2$ con $pH$ grados de libertad, donde $H$ es el número de curvas menos 1. De esta manera, los valores p son 1 y 1 para el grupo de altos y bajos, respectivamente. De esta manera, se observa que no se rechaza la hipótesis de independencia en ambos casos.

Por último, para verificar la existencia de puntos de cambio en las observaciones funcionales, se desea probar

$$H0 : E[X_k] = E[X_m] \ \forall k, m = 1, ..., N \text{ vs}$$
$$Ha : \text{Por lo menos } E[X_k] \neq E[X_m] \text{ para } k, m = 1, ..., N,$$

para lo que se usa la estadística de prueba que se muestra en la ecuación (2), donde $\langle X_i(t), \eta_l(t) \rangle$ y $\lambda_l$ es el $l$-ésimo valor propio asociado al operador de covarianza en cuestión. Para la determinación de valores $p$ en esta prueba se uso métodos de remuestreo y permutaciones entre las agrupaciones conformadas, ya que bajo hipótesis nula, se supone que deberían ser en promedio iguales.

$$\frac{1}{N} \sum_{l=1}^{d} \frac{\sum_{i=1}^{K} \hat{\epsilon}_{li} - \frac{K}{N} \sum_{j=1}^{N} \hat{\epsilon}_{lj}}{\lambda_l} \tag{2}$$

Se realizó la programación tanto de la estadística de prueba como el método de remuestreo asociado, de tal manera que la función indica el número donde los grupos de funciones no son iguales, en caso de que se rechace la hipótesis nula. Las salidas que se muestran a continuación presentan los valores que se utilizaron para segmentar los dos grupos denotados por $K$ y los respectivos valores $p$.

```
Check_bajos <- Check_differences(x = 1:100, Y_mat = Matrix_bajo_without,
               n_pca = 1, lambda = lambda_optimo,
               n_times = 500)
Check_bajos$K
```

```
##  [1]  10  15  20  25  30  35  40  45  50  55  60  65  70  75  80  85  90  95 100
## [20] 105 110 115 120
```

```
Check_bajos$p_values
```

```
##  [1] 0.112 0.122 0.122 0.392 0.724 0.934 0.674 0.554 0.300 0.156 0.698 0.828
## [13] 0.936 0.756 0.564 0.858 0.938 0.912 0.854 0.692 0.880 0.580 0.178
```

```
Check_altos <- Check_differences(x = 1:100, Y_mat = Matrix_alto_without,
               n_pca = 1, lambda = lambda_optimo,
               n_times = 500)
Check_altos$K
```

```
##  [1] 10 15 20 25 30 35 40 45 50 55 60
```

```
Check_altos$p_values
```

```
##  [1] 0.376 0.244 0.484 0.950 0.250 0.514 0.722 0.868 0.218 0.468 0.194
```

Se puede observar que los valores $p$ asociados a la prueba de punto de cambio son mayores a 0.05, por lo cual no se rechaza la hipótesis nula en los casos considerados.

- Formule y verifique la prueba de hipótesis adecuada para la carne con alto contenido de grasa. Verifique los supuestos asociados a la prueba de hipótesis.

Suponiendo que se tiene una secuencia de $n$ funciones aleatorias $X_1(t), \ldots, X_n(t)$, con $X_i(t) \in \mathcal{L}^2$ y se desea llevar a cabo la prueba de hipótesis relacionada con una función promedio proveniente de dicho proceso estocástico funcional, inicialmente se tienen dos perspectivas. La primera de ellas se basa en la norma, mientras que la segunda se basa principalmente en los componentes principales funcionales. Partiendo de que se desea probar

$$H_0 : \mu(t) = \mu_0(t) \text{ vs } H_a : \mu(t) \neq \mu_0(t),$$

donde $\mu(t)$ representa la función promedio poblacional y $\mu_0(t)$ es una función de interés. Las dos perspectivas mencionadas anteriormente se describen a continuación:

- Prueba basada en la norma:

En este caso la estadística de prueba se muestra en la ecuación (3)

$$\begin{aligned} T_{NORM} &= n||\bar{X}(t) - \mu(t)||^2 \\ &= n\sum_{k=1}^{\infty}\langle \bar{X}(t) - \mu(t), \eta_k(t)\rangle \\ &\approx n\sum_{k=1}^{q}\langle \bar{X}(t) - \mu(t), \eta_k(t)\rangle, \end{aligned} \tag{3}$$

donde $\eta_k(t)$ la $k$-ésima función propia del operador de covarianza $\mathcal{C}_X$, que corresponde al conjunto de realizaciones funcionales $X_i(t)$ y $q$ es un número de funciones propias definido de acuerdo a un porcentaje de variabilidad.

- Prueba basada en los componentes principales funcionales:

Esta perspectiva se basa en la estadística $T^2$ de Hotelling considera la estadística de prueba que se muestra en la ecuación (4)

$$\begin{aligned} T^2_{FPCA} &= n\sum_{k=1}^{\infty}\frac{\epsilon_k^2}{\lambda_k} \\ &= n\sum_{k=1}^{\infty}\frac{\langle \bar{X}(t) - \mu(t), \eta_k(t)\rangle^2}{\lambda_k} \\ &\approx n\sum_{k=1}^{q}\frac{\langle \bar{X}(t) - \mu(t), \eta_k(t)\rangle^2}{\lambda_k} \end{aligned} \tag{4}$$

Para fines prácticos en este ejercicio se plantea el objetivo de determinar si la función promedio en cada uno de los grupos es constante o no. Es decir, que se va a probar

$$\begin{aligned} H_0 &: \mu(t) = \mu_0, \\ H_a &: \mu(t) \neq \mu_0. \end{aligned}$$

Para desarrollar la prueba se utilizaron las ecuaciones (3) y (4), mientras que para la determinación de valores críticos se usaron métodos de remuestreo. Para generar las nuevas observaciones producto del método bootstrap, se hizo uso de la metodología descrita por Paparoditis and Sapatinas [2016]. La idea descrita allí se hace para dos poblaciones y se resume en generar observaciones como se describe en la ecuación (5), donde $\epsilon_{ij}^+ = X_{ij}(t) - \bar{X}_i.(t)$, con $\bar{X}_{i.}(t) = 1/N\sum_{j=1}^{N} X_{ij}(t)$. De esta forma, se observa que el remuestreo se realiza sobre los $\epsilon_{ij}^+$ calculados a partir de la realización original. Esta idea se aplica en el siguiente punto, cuyo objetivo es realizar una prueba a partir de dos poblaciones. En esta parte, que corresponde a la media de una

población, se hace uso de una idea análoga, es decir que las observaciones funcionales simuladas $X_i(t)^+$ se obtuvieron mediante la formula que se encuentra en la ecuación (6).

$$X_{ij}(t)^+ = \bar{X}_{N+M} + \epsilon_{ij}^+ \tag{5}$$

$$X_i(t)^+ = \bar{X}_N + \epsilon_i^+ \tag{6}$$

A continuación se presentan los valores p correspondientes a las estadísticas de prueba basados en la norma y en los componentes principales funcionales, a partir del método bootstrap para el grupo alto en grasa. Allí se observa que estos valores son mayores a 0.05, por lo cual se concluye que no hay suficiente evidencia para determinar que la media es diferente a una constante. Considerando que esto es cuestionable, se elaboró el Gráfico 5, donde se visualiza la función promedio en color rojo y la media global utilizada en la prueba en color azul. Allí se nota que son bastante similares, lo cual justifica la decisión obtenida en la prueba.

```
Test_one_mean_fda(x = 1:100, Y_mat = Matrix_alto_without,
                  lambda = lambda_optimo,
                  n_pca = 1, mean = mean(Matrix_alto_without), n_times = 500)
```

```
##   p_value_norm p_value_fpca
## 1        0.138        0.138
```

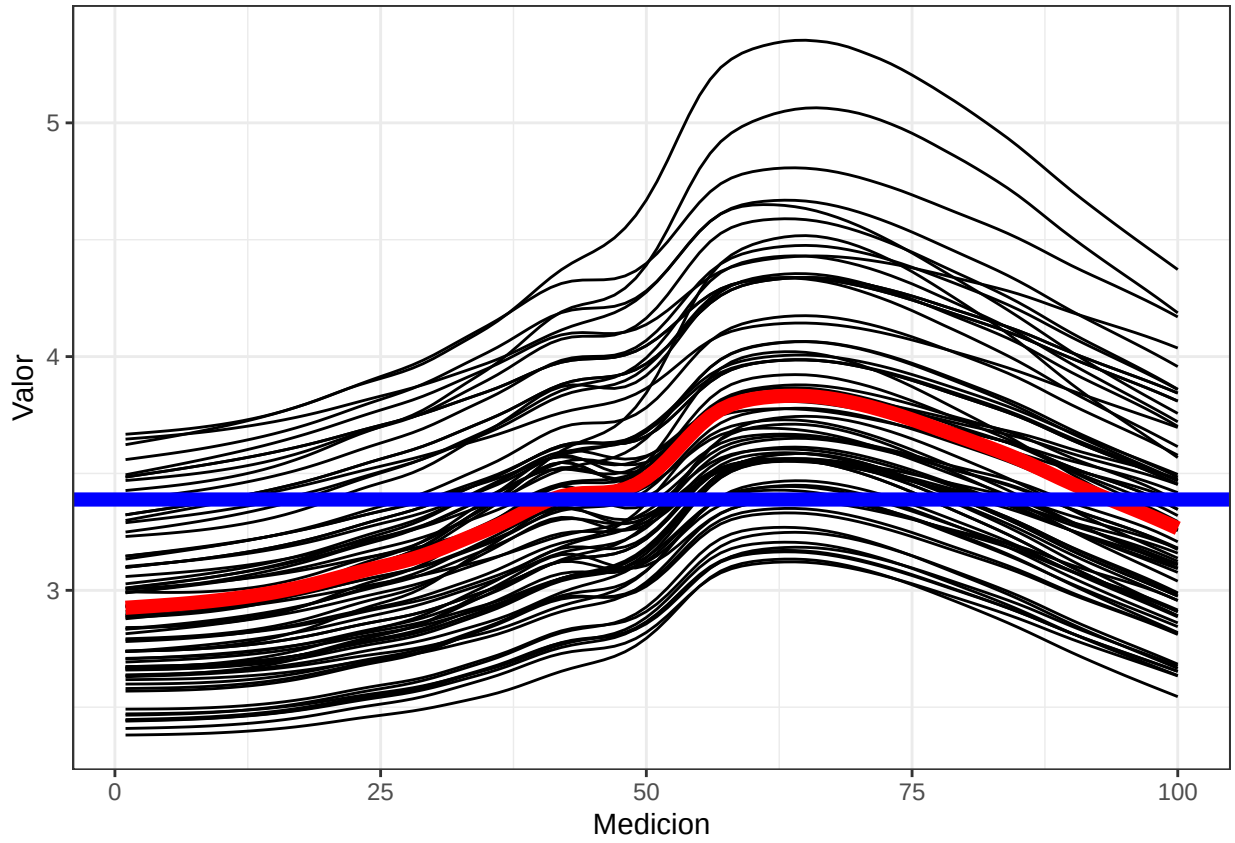


Gráfico 5: Curvas correspondientes al grupo alto en grasa, junto con la curva media y la media global de todas las observaciones en azul.

- Verifique si las funciones medias de la carne con alto contenido de grasa y de la carne con bajo nivel de grasa, son iguales. Verifique los supuestos asociados a la prueba de hipótesis.

En este caso, se tienen dos agrupaciones de curvas, en adelante denotadas por $X_1, \ldots, X_N$ y $Y_1, \ldots, Y_M$, respectivamente, con $X_k, Y_i \in \mathcal{L}^2$. Se desea probar:

$$H_0: \ \mu_1(t) = \mu_2(t) \ H_a: \ \mu_1(t) \neq \mu_2(t)$$

Para ello, nuevamente se tienen dos perspectivas. La primera de ellas se basa en la norma, mientras que la segunda se basa en los componentes principales funcionales.

- Prueba basada en la norma:

Se utiliza la estadística de prueba que se muestra en la ecuación (7), donde $\varphi$ representa un proceso gaussiano, $Z_k$ son variables aleatorias con distribución normal y $\tau_k$ son los valores propios correspondientes al operador de covarianza $\mathcal{C}_\varphi = (1-\theta)\mathcal{C}_X + \theta\mathcal{C}_Y$.

$$\begin{aligned} U &= \frac{NM}{N+M}||\bar{X}(t) - \bar{Y}(t)|| \\ &= \int \varphi^2(t)dt \\ &= \sum_{k=1}^{\infty} \tau_k Z_k^2 \\ &\approx \sum_{k=1}^{q} \hat{\tau}_k \hat{Z}_k^2. \end{aligned} \tag{7}$$

- Prueba basada en los componentes principales funcionales:

En esta segunda perspectiva, se utilizaron los valores y funciones propias de $\mathcal{C}_\varphi$. La estadística de prueba se muestra en la ecuación (8), donde $\lambda_k$ es el $k$-ésimo valor propio de $\mathcal{C}_\varphi$ y $\epsilon_k = \langle \bar{X} - \bar{Y}, \eta_k \rangle$ con $\eta_k$ la $k$-ésima función propia correspondiente.

$$U_{FPCA} = \sum_{k=1}^{q} \frac{\hat{\epsilon}_k^2}{\hat{\lambda}_k}. \tag{8}$$

La determinación de los valores y funciones propias de $\mathcal{C}_\varphi$ se realizó mediante la matriz que se detalla en Horváth and Kokoszka [2012], la cual se muestra en la ecuación (9).

$$\begin{aligned} \hat{z}_{MN}(t,s) = &\frac{N}{N+M}\frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \bar{X}_N(t))(X_i(s) - \bar{X}_N(s)) \\ &+ \frac{M}{N+M}\frac{1}{M}\sum_{i=1}^{N}(X_i(t) - \bar{X}_M(t))(X_i(s) - \bar{X}_M(s)). \end{aligned} \tag{9}$$

La determinación de valores $p$ se realizó de acuerdo a la metodología dada en Paparoditis and Sapatinas [2016], que se describió previamente. Los resultados se presenta a continuación, donde se observa que los valores $p$ son menores a 0.05, por lo cual se concluye que hay suficiente evidencia para rechazar la hipótesis de igualdad de funciones promedio poblacionales.

```
##   p_value_norm p_value_fpca
## 1            0            0
```

- Verifique si el operador de covarianza de la carne con alto contenido de grasa es igual al operador de covarianza con bajo nivel de grasa. Verifique los supuestos asociados a la prueba de hipótesis.

En este caso, se tienen dos agrupaciones de curvas, en adelante denotadas por $X_1, \ldots, X_N$ y $Y_1, \ldots, Y_M$, respectivamente, con $X_k, Y_i \in \mathcal{L}^2$. Se desea probar

$$H_0: \ C_x = C_y$$
$$H_a: \ C_x \neq C_y,$$

lo cual se a llevar a cabo mediante la estadística de prueba descrita en la ecuación (10), que se presenta a continuación:

$$\hat{T} = \frac{N+M}{2}\theta(1-\theta)\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{\langle \hat{C}_x(\eta_i) - \hat{C}_y(\eta_i), \eta_j\rangle^2}{(\theta\lambda_{yj} + (1-\theta)\lambda_{xi})(\theta\lambda_{xj} + (1-\theta)\lambda_{yi})}, \tag{10}$$

donde

$$\hat{\lambda}_{xk} = \frac{1}{N}\sum_{n=1}^{N}\langle X_n, \hat{\eta}_k\rangle^2 \text{ y } \hat{\lambda}_{yk} = \frac{1}{M}\sum_{m=1}^{M}\langle Y_m, \hat{\eta}_k\rangle^2. \tag{11}$$

Mientras que por otro lado,

$$\hat{C}_x(\eta_i) = \frac{1}{N}\sum_{k=1}^{N}\langle X_k(t), \eta(t)\rangle X_k(t) \text{ y } \hat{C}_y(\eta_i) = \frac{1}{M}\sum_{k=1}^{M}\langle Y_k(t), \eta(t)\rangle Y_k(t). \tag{12}$$

La determinación del valor $p$ en este caso, se realiza a partir de la distribución $\chi^2$ con $\frac{p(p+1)}{2}$ grados de libertad. Se observa que el valor $p$, utilizando la distribución $\chi^2$, es mayor a 0.05, por lo cual los operadores son estadísticamente diferentes. Por otra parte, mediante el método bootstrap el valor $p$ es menor a 0, por lo cual en ese caso se rechaza la hipótesis de igualdad entre operadores de covarianza. Los resultados mencionados se presentan a continuación:

```
Test_original <- Cov_operators_test_stat(x = 1:100 , Y_mat_1 = Matrix_bajo_without,
                                         Y_mat_2 = Matrix_alto_without,
                                         pca_overall = pca_overall,
                                         lambda = lambda_optimo)

p_value <- pchisq(Test_original, n_pca*(n_pca+1)/2, lower.tail = FALSE)
p_value
```

```
##           [,1]
## [1,] 0.2371937
```

```
mean(Results_boots$Test_stat_boots > as.double(Test_original))
```

```
## [1] 0
```

Se incluyen en los procedimientos de prueba las propuestas vistas en clase y las presentadas en los papers Paparoditis and Sapatinas [2016] y Lopez-Pintado and Qian [2021].

(2) Utilice el dataset sobre las curvas espectrométricas de la producción de azúcar a partir de la remolacha y aplique la propuesta presentada por Qiu et al. [2021].

# References


Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. URL http://www.jstatsoft.org/v51/i04/.

Lajos Horváth and Piotr Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

Francesca Ieva, Anna Maria Paganoni, Juan Romo, and Nicholas Tarabelloni. roahd Package: Robust Analysis of High Dimensional Data. *The R Journal*, 11(2):291–307, 2019. doi: 10.32614/RJ-2019-032. URL https://doi.org/10.32614/RJ-2019-032.

Sara Lopez-Pintado and Kun Qian. A depth-based global envelope test for comparing two groups of functions with applications to biomedical data. *Statistics in Medicine*, 2021.

E Paparoditis and T Sapatinas. Bootstrap-based testing of equality of mean functions or equality of covariance operators for functional data. *Biometrika*, 103(3):727–733, 2016.

Zhiping Qiu, Jianwei Chen, and Jin-Ting Zhang. Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis*, 157:107160, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

J. O. Ramsay, Spencer Graves, and Giles Hooker. *fda: Functional Data Analysis*, 2020. URL https://CRAN.R-project.org/package=fda. R package version 5.1.9.