ENFOQUE ESTADÍSTICO DEL APRENDIZAJE

INTRODUCCIÓN AL ANÁLISIS DE DATOS FUNCIONALES

Gabriel Omar Masi
Zonia Morales

Objetivos

- Proporcionar un marco teórico introductorio sobre el análisis de datos funcionales, abordando su representación, el proceso de suavizado y las principales medidas estadísticas aplicables a objetos funcionales.
- Desarrollar un análisis de clusterización de las alturas en función del sexo utilizando un dataset de personas entre 0 y 18 años, con el objetivo de caracterizar el comportamiento de crecimiento según el sexo. Para mejorar la discriminación, se analizarán las derivadas de las funciones que representan las alturas de cada persona y se evaluará si este enfoque permite una clusterización más precisa.

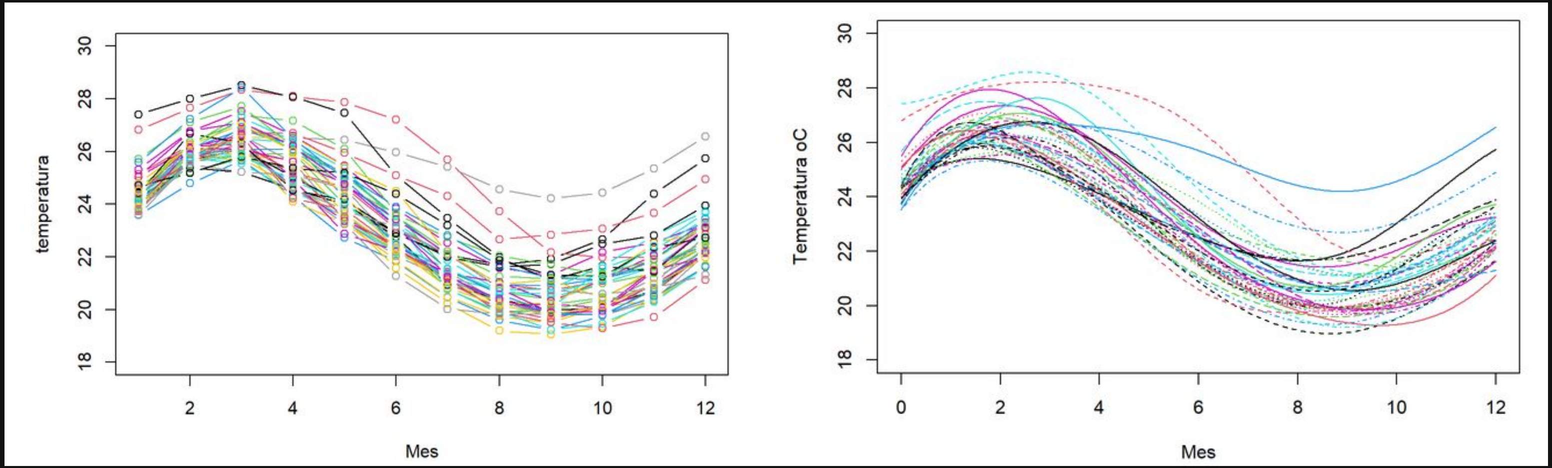
Introducción Conceptual

¿De qué se trata el análisis de datos funcionales?

01 ¿Qué es FDA?

Estudia y analiza datos que pueden representarse como funciones continuas en un dominio (por ejemplo, tiempo, espacio). En lugar de trabajar con puntos discretos, se modelan como curvas o trayectorias completas.

$$\{x_n(t) : t \in [T_1, T_2],\ n = 1, 2, \ldots, N\}.$$



Aristizabal R y Bohorquez C. Curso de Estadística Espacial

02

Motivación

- **Aprovechar las funciones:** ofrecen información más rica que los datos discretos: tendencias globales, derivadas, patrones locales, y más.
- **Capturar relaciones complejas:** Relación entre curvas: Por ejemplo, correlación entre curvas de actividad física y salud.
- **Representar datos complejos** (como funciones) con pocos coeficientes mediante bases funcionales (B-splines, Fourier).

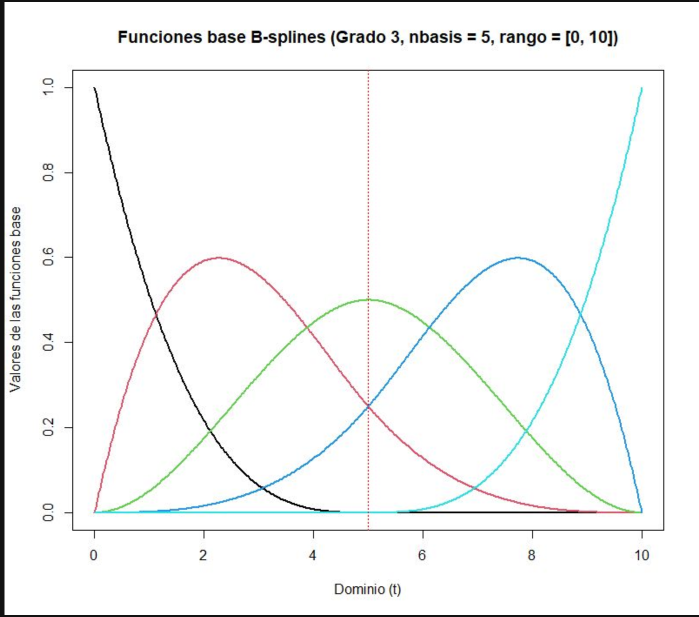
03 Expansión en Bases

Σf η

Una **base funcional** es un conjunto de funciones fundamentales que se combinan linealmente para representar cualquier función en el espacio de interés.

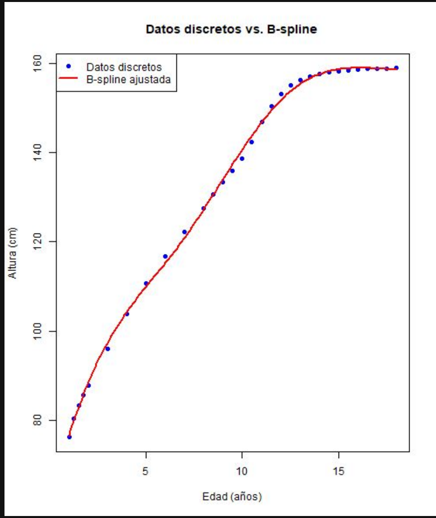
Transforman funciones que pertenecen a espacios funcionales de dimensión infinita en representaciones finitas y manejables.

$$f(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$$



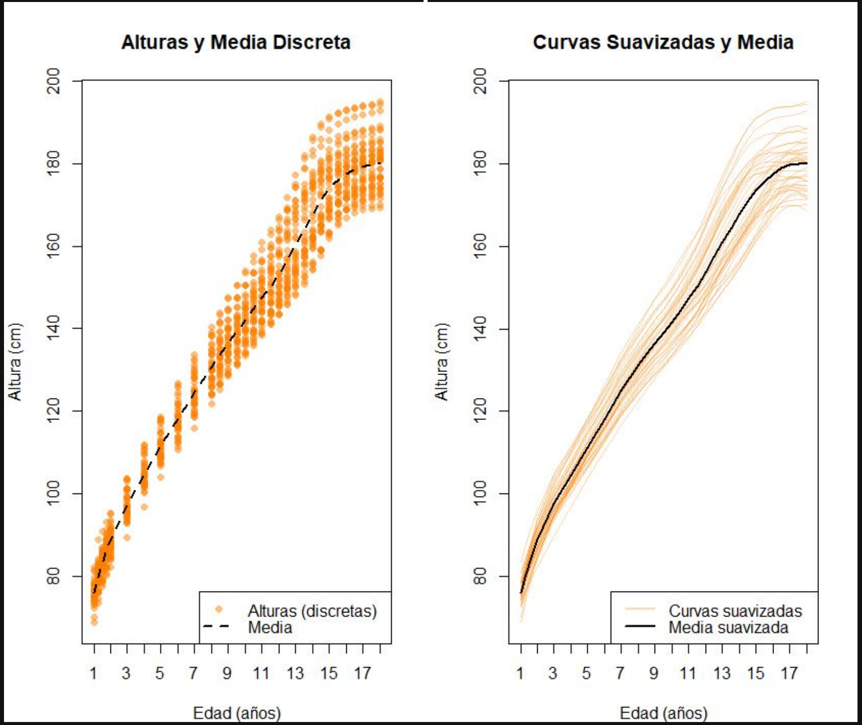

```
spline.basis=create.bspline.basis(rangeval=c(0,10), nbasis=5)
plot(spline.basis, lty=1, lwd=2)
```

03 Expansión en Bases, b-splines




```r
# Cargar datos de crecimiento
data(growth)
# Definir edad y altura de la primera niña
age <- growth$age; height <- growth$hgtf[,1]
# Crear 5 funciones base B-splines
basis <- create.bspline.basis(rangeval = range(age), nbasis = 5)
# Ajustar las B-splines a los datos
altura_fd <- smooth.basis(age, height, basis)$fd
# Graficar la curva ajustada
plot(altura_fd, lwd = 2, main = "Curva ajustada con 5 B-splines")
# Superponer los puntos originales
points(age, height, col = "red", pch = 16)
```

04 Estimadores de medidas Descriptivas en Datos Funcionales: Media



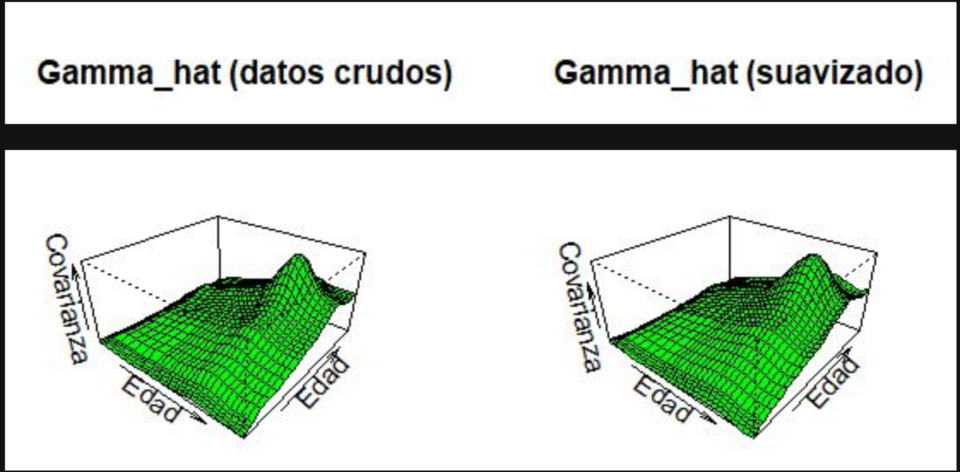
En $\mathcal{H} = L^2(I)$, el espacio de funciones cuadrado integrable:

$$\mu(t) = \mathbb{E}[X(t)] \text{ para todo } t \in I.$$

La media funcional se estima como:

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t) = \bar{X}_n(t)$$

04 Estimadores de medidas Descriptivas en Datos Funcionales: Covarianza y varianza



En $\mathcal{H} = L^2(I)$, el espacio de funciones cuadrado integrable:

$$(\Gamma u)(t) = \int_I \gamma(s,t) f(t) ds.$$

$$\gamma(s,t) = \operatorname{Cov}(X(s), X(t)).$$

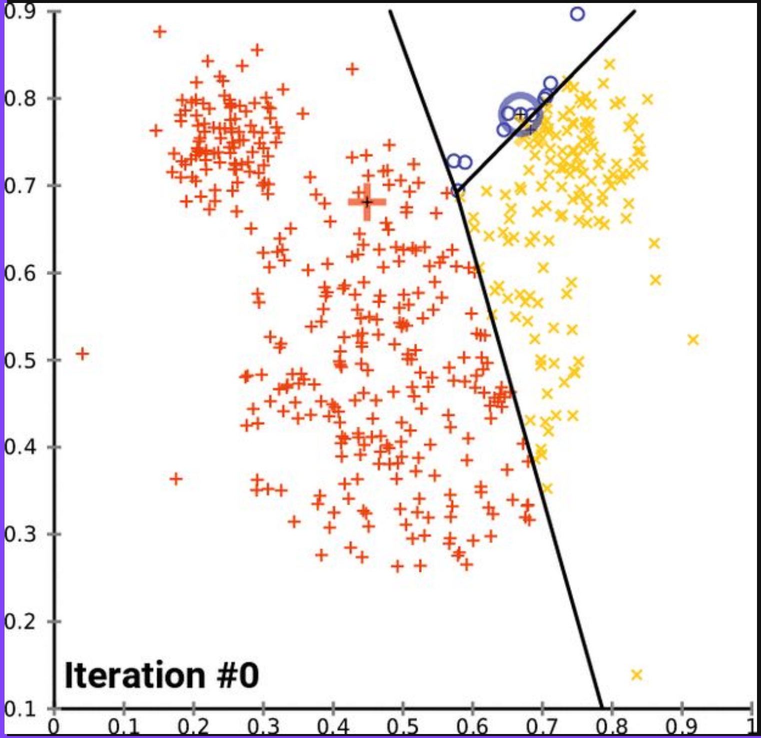
El núcleo se estima como:

$$\hat{\gamma}(t,s) = \frac{1}{n} \sum_{i=1}^{n} \left[X_i(t) - \bar{X}_n(t)\right] \left[X_i(s) - \bar{X}_n(s)\right]$$

05

Clustering

05 Clustering K Means



Inicialización:

- Se seleccionan k centroides iniciales al azar, donde k es el número de grupos deseados.

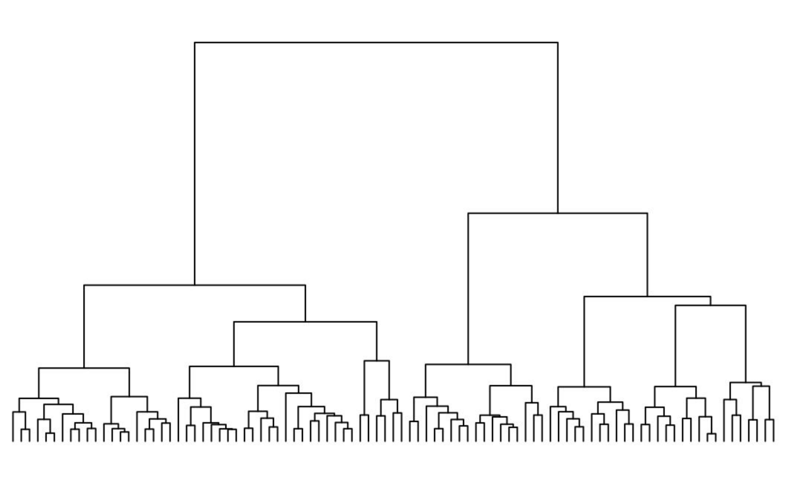
Asignación de Clusters:

- Para cada punto en el conjunto de datos, se calcula la distancia (usualmente euclidiana) a cada uno de los k centroides.
- Cada punto se asigna al cluster con el centroide más cercano.

Recalcular Centroides:

- Para cada cluster, se calcula el nuevo centroide como el promedio (media) de los puntos asignados a ese cluster.

05 Clustering Jerárquico

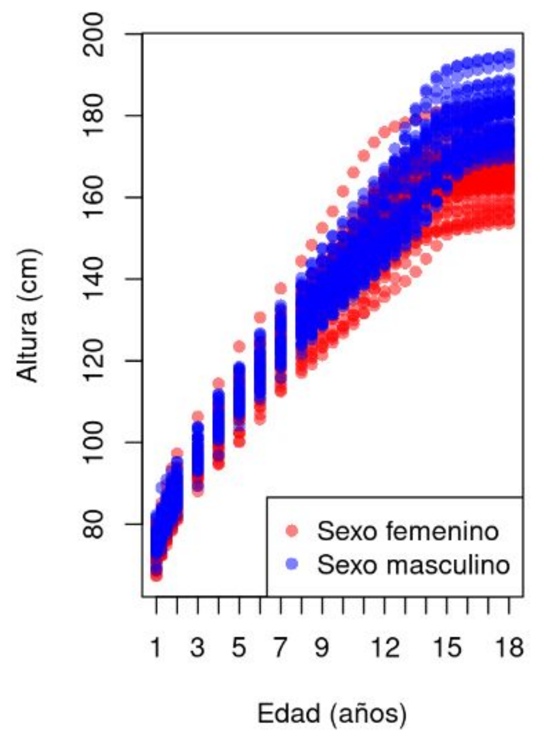


1. **Construcción de la Matriz de Distancias**
   - Mide la similaridad entre cada par de elementos
2. **Algoritmo de Agrupamiento**:
   - Se agrupan los datos iterativamente, comenzando con cada registro como un grupo independiente, luego combinándolos gradualmente según su similitud.
3. **Generación del Dendrograma**:
   - Un dendrograma es un árbol que muestra cómo se agrupan los puntos en diferentes niveles de similitud.
   - Cada nivel representa un paso en el proceso de agrupamiento.
4. **Corte del Dendrograma**:
   - Una vez construido el dendrograma, se puede "cortar" a un nivel deseado para definir un número específico de clústeres (por ejemplo, 2 grupos).

Aplicación de FDA y Clustering

Dataset crecimiento niños, niñas y adolescentes

Aplicación



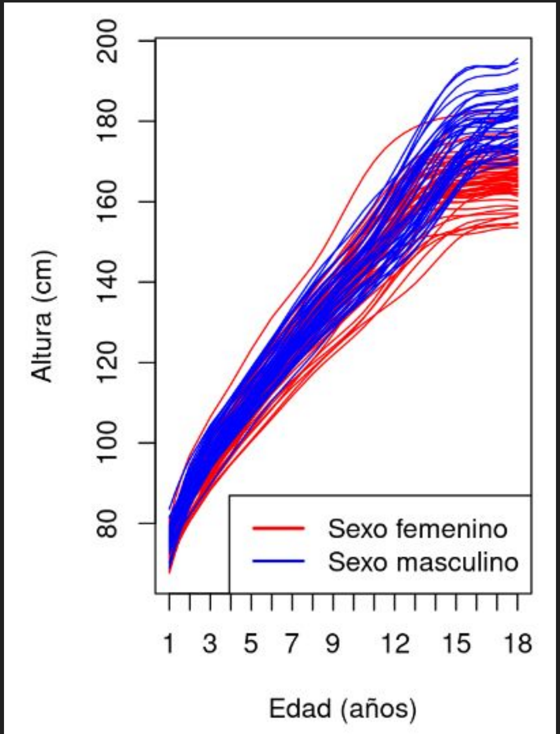
- Partimos de 93 registros
- Cada registro cuenta con 31 medidas de altura (registradas a lo largo del tiempo)
- Se intentará agrupar los individuos por sexo

Aplicación

- Para obtener curvas suaves, que representan a nuestros datos debemos:

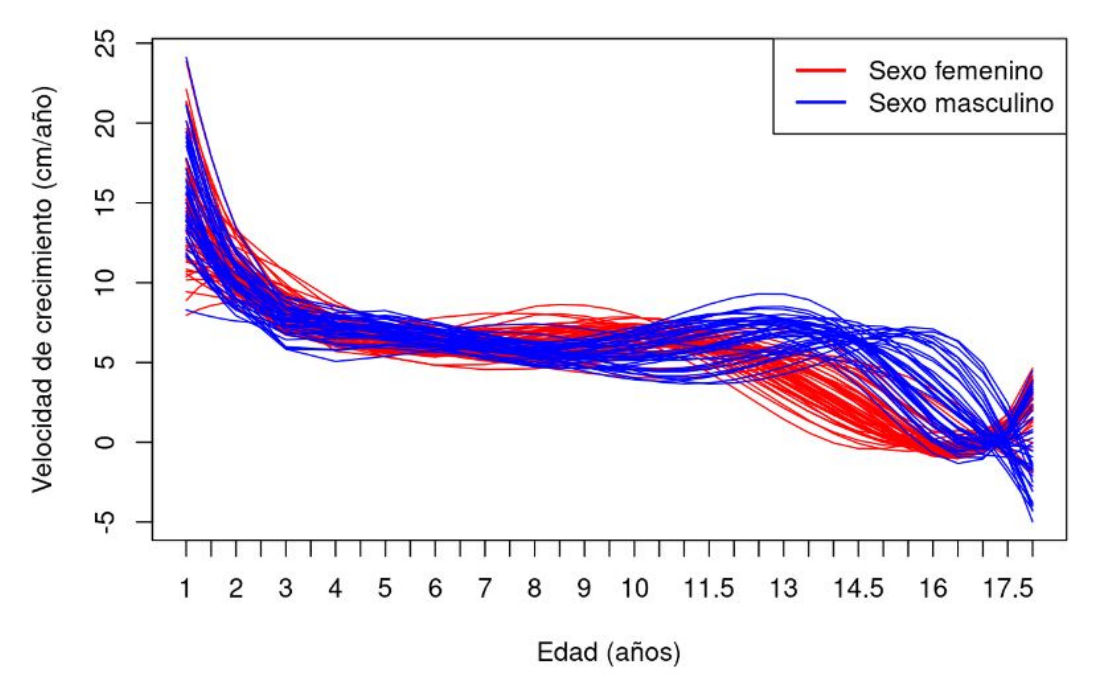
- Primero procedemos a generar una base utilizando B-Spline con N = 8

- Luego, le aplicamos la transformación Data2fd que retorna las curvas



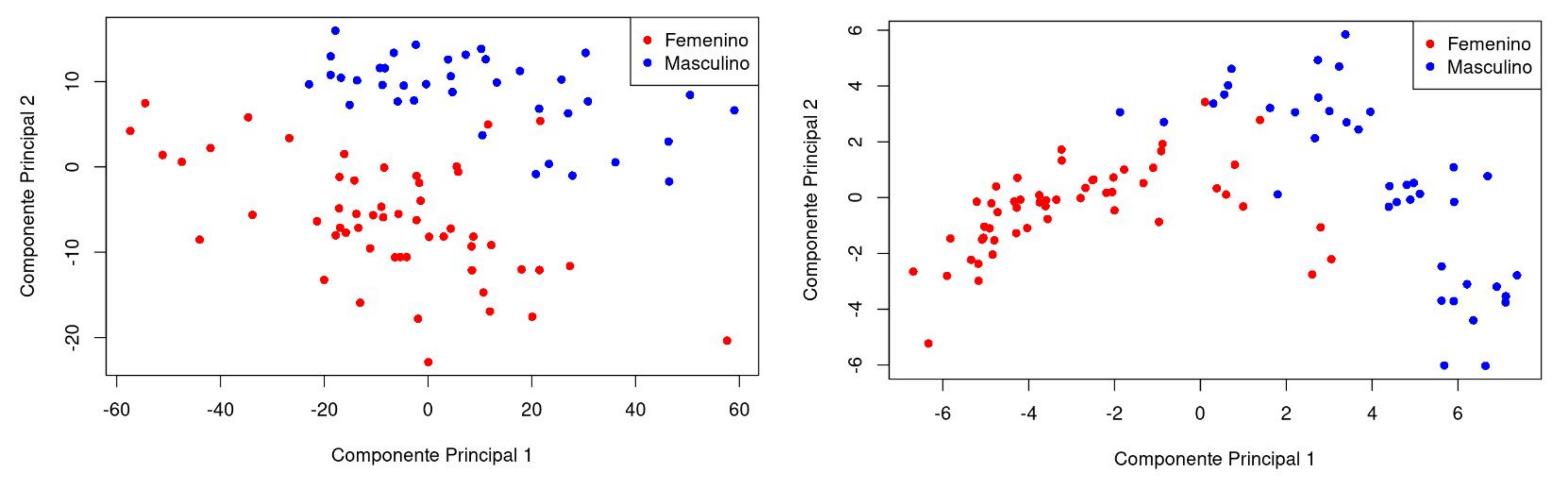
Aplicación

Las curvas resultantes pueden derivarse y obtenemos lo siguiente



Aplicación

Aplicar PCA a las curvas para poder graficarlas de manera puntual y/o optimizar la performance de los algoritmos de clustering

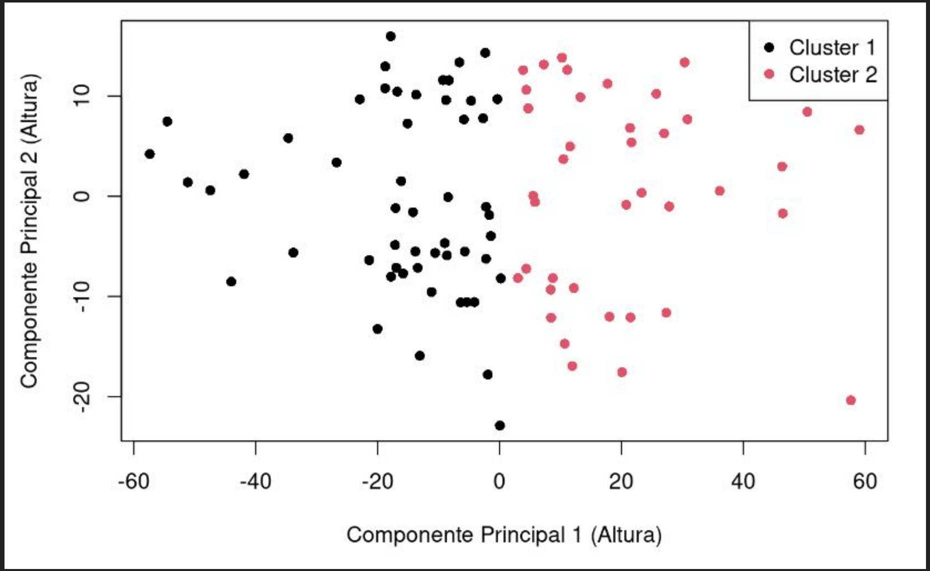


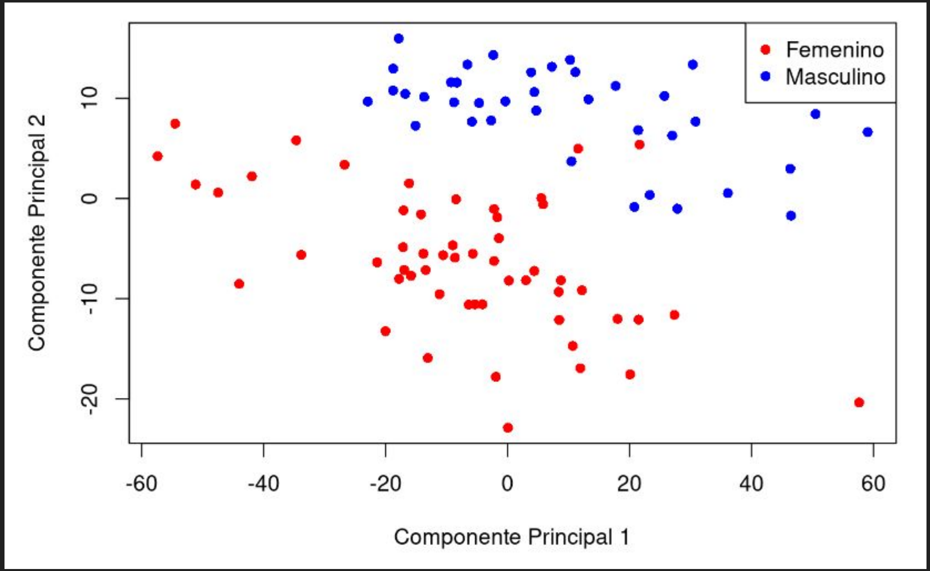
K Means

Al aplicar K Means a las componentes FPCA de las curvas sin derivar, obtenemos el siguiente resultado

Matriz de Confusión

37	17
16	23



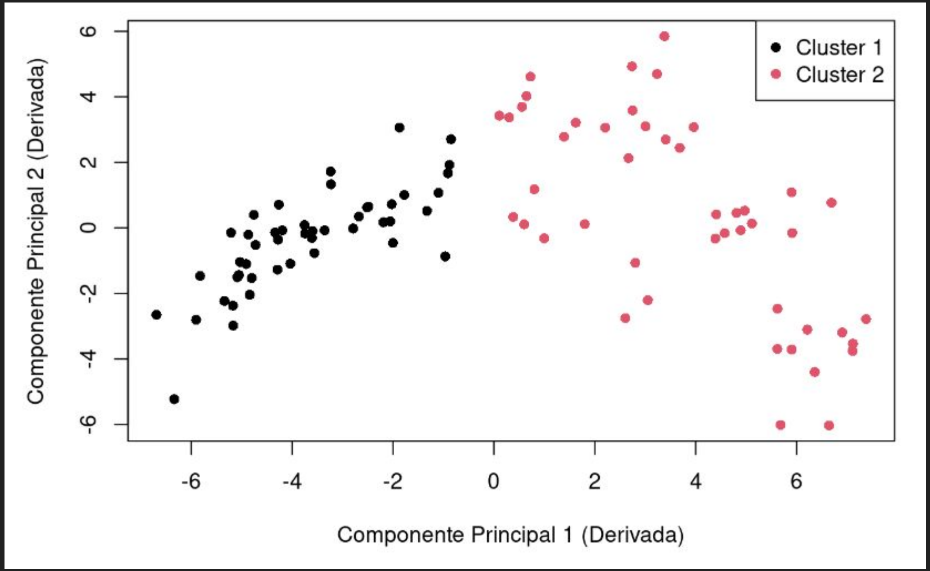


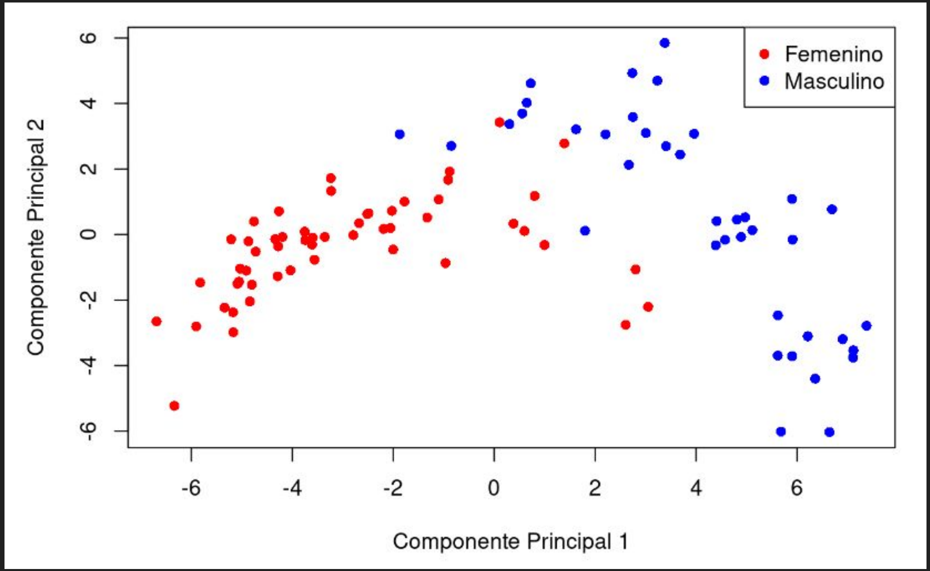
K Means

Al aplicar K Means a las componentes FPCA de las curvas derivadas, obtenemos el siguiente resultado

Matriz de Confusión

45	9
2	37



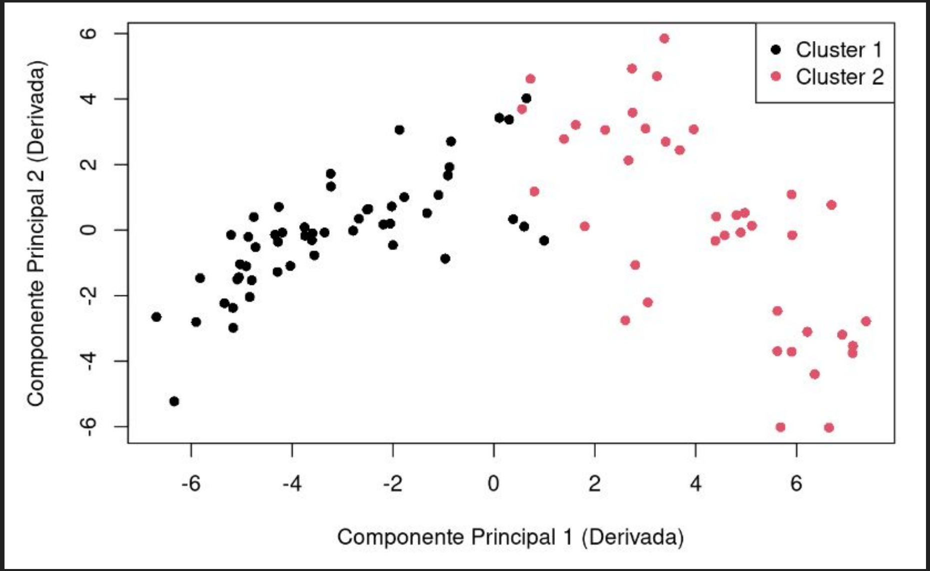


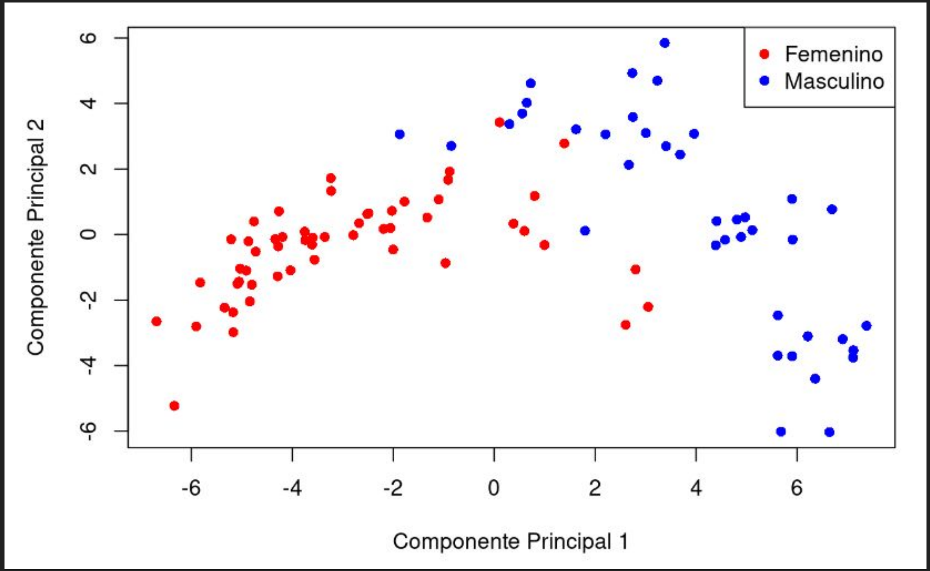
K Means

Si no reducimos la dimensionalidad con FPCA, y aplicamos directamente K Means a las derivadas

Matriz de Confusión

4	35
49	5



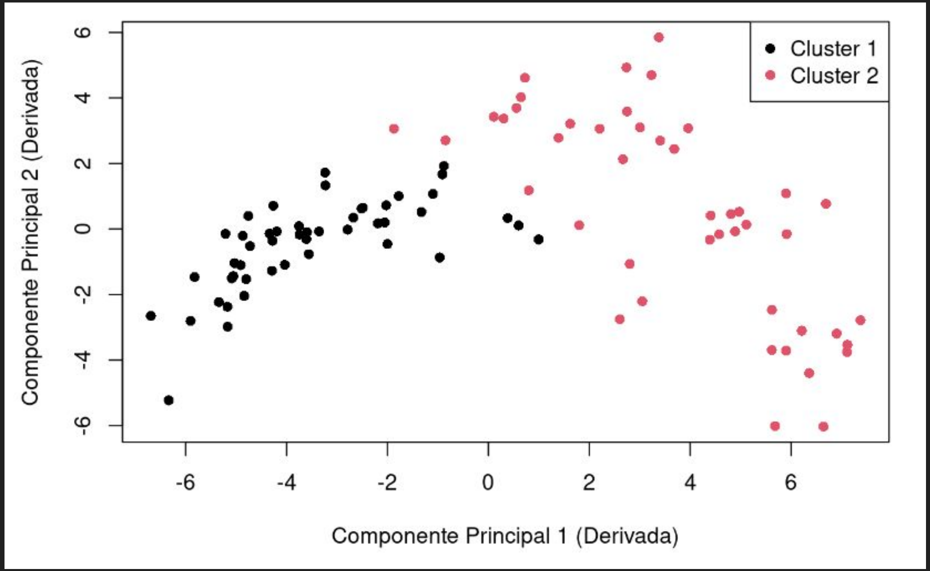


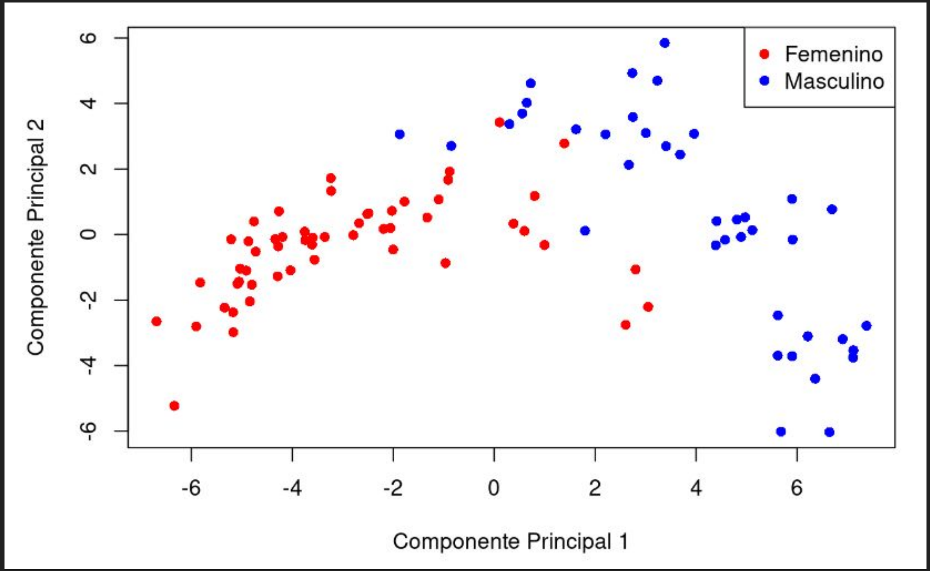
Cluster Jerárquico

Al aplicar un clustering jerárquico a las componentes FPCA de las curvas derivadas, obtenemos

Matriz de Confusión

48	6
0	39





Conclusiones

- FDA es una herramienta muy poderosa para el estudio de datos longitudinales o discretos que evolucionan en un dominio continuo, como el tiempo.

- Las propiedades de las curvas obtenidas nos permiten analizar patrones inherentes en los datos, como tasas de cambio mediante el cálculo de derivadas funcionales.

- Al transformar los datos, podemos mejorar la performance de ciertos algoritmos, por ejemplo, K-Means o Clustering Jerárquico.

- Para el caso en estudio, se encuentra una mayor performance utilizando Clustering Jerárquico, habiendo transformado los datos con B-Splines y luego con un FPCA.

Referencias

- Kokoszka, P., Reimherr, M. (2017). Introduction to Functional Data Analysis. CRC Press. https://www.taylorfrancis.com/books/mono/10.1201/9781315117416/introduction-functional-data-analysis-piotr-kokoszka-matthew-reimherr

- Wu, R., Wang, B., Xu, A. (2021). Functional data clustering using principal curve methods. Communications in Statistics - Theory and Methods, 50(21), 5087-5101. https://www.tandfonline.com/doi/full/10.1080/03610926.2021.1872636#d1e3041

- Parada, D. (2024, septiembre 18). Introducción a los Datos Funcionales y al Análisis de Componentes Principales Funcionales [Video]. En Andres Farral (Canal). YouTube. https://www.youtube.com/watch?v=rKLJq-rKCn0

- Parada, D. (2024, septiembre 25). Introducción a los Datos Funcionales: Las Autofunciones [Video]. En Andres Farral (Canal). YouTube. https://www.youtube.com/watch?v=al7qb7JhvKM