Análisis de Datos Funcionales

Ing. Nelson Ariza Morales
Centro de Investigación en Matemáticas A.C.
Unidad Monterrey

June 10, 2024

Índice

Definición

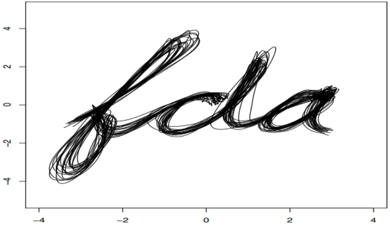


Figure: Medidas de posición de la punta de un bolígrafo con la inscripción “fda”. 20 réplicas.

Análisis de datos funcionales

=

Análisis de datos que son funciones.

¿Cuáles son las características más obvias de estos datos?

Cantidad

Frecuencia

Tendencias similares

Definición

¿De datos a funciones?

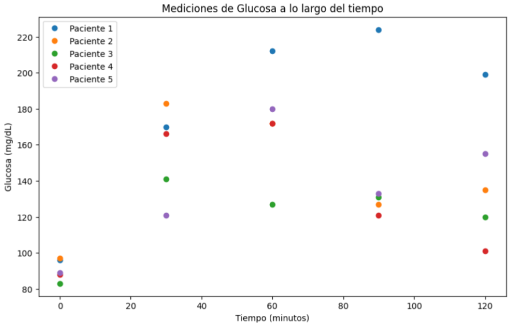


Figure: Mediciones de prueba de tolerancia a glucosa de 5 pacientes

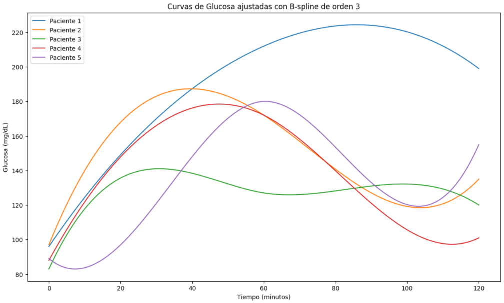


Figure: Mediciones de la glucosa con B -Splines de orden 3

Modelo fundamental

$$x_i = f(t_i) + \epsilon_i$$

- Datos: $x_1, x_2, ..., x_n$
- Errores de medición: $\epsilon_1, \epsilon_2, ..., \epsilon_n$
- Se asume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ y ϵ_i son independientes.
- No tenemos la forma paramétrica de $f(t)$

¿Cómo estimar $f(t)$?

La función $f(t)$ puede expresarse como una combinación lineal de funciones base:

$$f(t) = \sum_{j=1}^{K} c_j \phi_j(t)$$

donde $\phi_1(t), \ldots, \phi_J(t)$ se denominan funciones base, y $c_1, \ldots, c_J$ se refieren a los coeficientes de estas funciones base.

Pregunta

¿Cómo decidir las funciones base?

Funciones base

Una base de funciones es un conjunto de funciones conocidas $\phi_1(t), \ldots, \phi_J(t)$, que son matemáticamente independientes entre sí y que tienen la ventaja de poder describir de forma particular, a partir de la suma ponderada o combinación lineal un determinado número de n funciones *(Curtis, 2014)*.

- Monomiales
- Fourier
- B-Splines

Funciones base: Monomiales

Las bases monomiales utilizan polinomios de la forma:

$$\Phi(t) = (1, t, t^2, \dots, t^k)$$

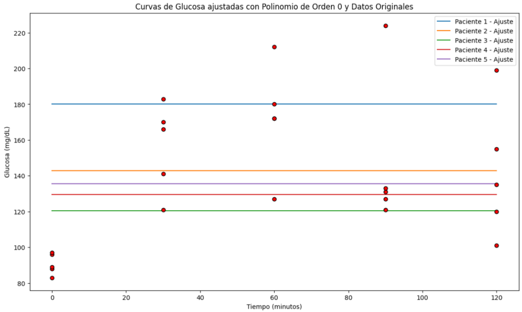


Figure: Monomial de orden 1

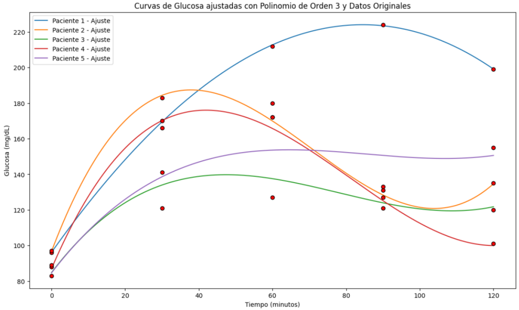


Figure: Monomial de orden 3

Funciones base: Fourier

Ideal para datos periódicos, las bases de Fourier se componen de senos y cosenos de frecuencia creciente:

$$\Phi(t) = (1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \ldots)$$

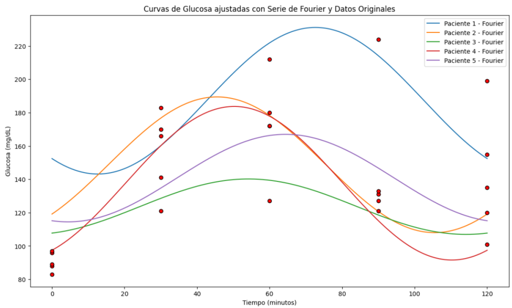


Figure: Fourier orden 2

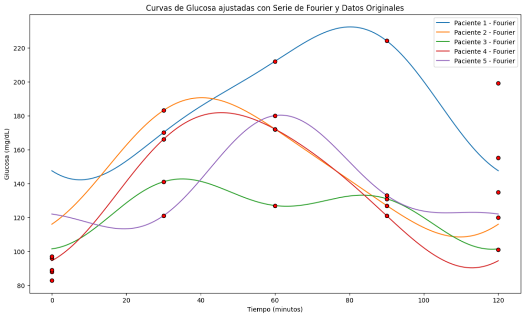


Figure: Fourier de orden 5

Funciones base: B-Splines

Los splines son segmentos polinómicos unidos de un extremo a otro.

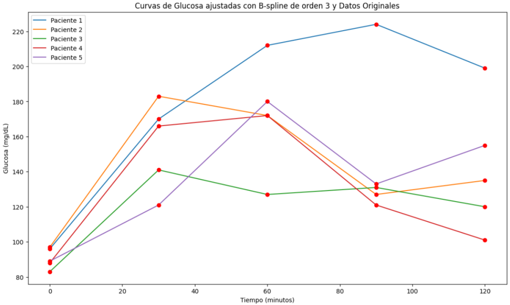


Figure: Splines de orden 1

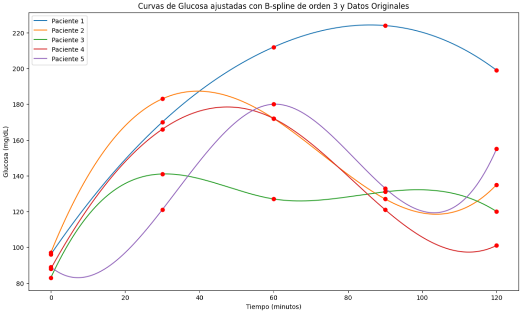


Figure: Spline de orden 3

Estadisticas Descriptivas

La media funcional se define como el promedio de un conjunto de funciones y se calcula de la siguiente manera:

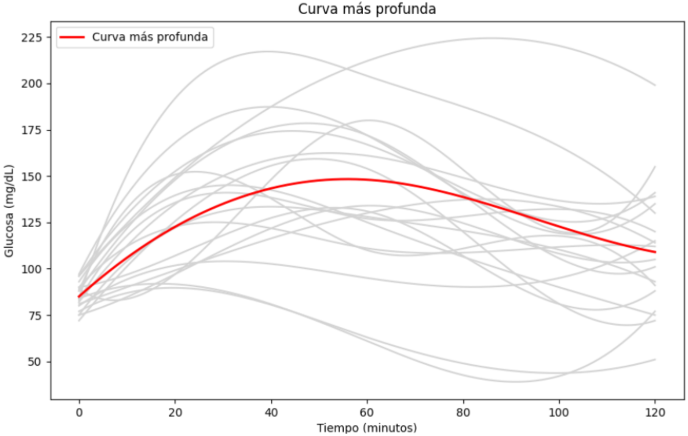
$$\bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t)$$

La varianza funcional mide la dispersión de las funciones alrededor de la media funcional y se define como:

$$Var_x(t) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i(t) - \bar{x}(t))^2$$

Profundidad de Banda

Permite ordenar curvas de muestra en un espacio funcional según su centralidad o atipicidad. Este enfoque fue significativamente avanzado por López-Pintado y Romo (2009) y Sun y Genton (2011)



Metodologías Profundidad de Banda

Sun y Genton.
Definen la profundidad de una curva $x_i(t)$ como la frecuencia con la que $x_i(t)$ queda dentro de las bandas formadas por otras curvas en la muestra:

$$BD_J(x_i, P) = \sum_{j=2}^{J} P\{G(x_i) \subset B(X_1, \ldots, X_j)\}$$

donde $B(X_1, \ldots, X_j)$ es una banda delimitada por j curvas aleatorias y $G(x_i)$ es el gráfico de la función $x_i(t)$. La gráfica de una función $x(t)$ es el subconjunto del plano

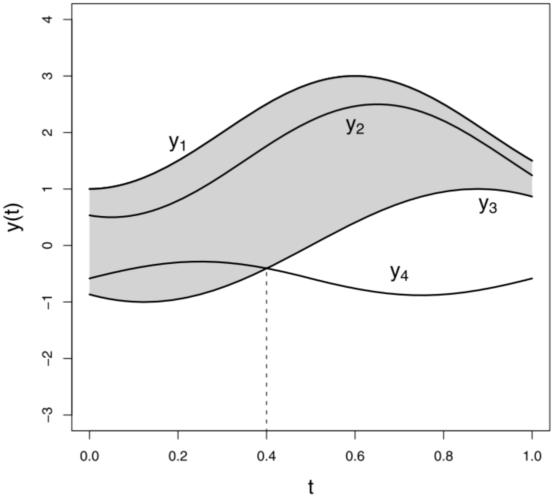
$$G(x) = \{(t, x(t)) : t \in T\}.$$

Lopez y Pintado
Proponen una variante más flexible, la profundidad de banda modificada (MBD), que considera la proporción del dominio en el que una curva está dentro de la banda formada por otras curvas:

$$MBD_n^{(j)} = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \lambda_r \{A(x; x_{i_1}, \ldots, x_{i_j})\}$$

donde λ representa la medida de Lebesgue sobre el intervalo I, y la suma se extiende sobre todos los subconjuntos de j curvas.

Sun&Genton vs Pintado&Romo



Boxplot Funcional

La banda en $\mathbb{R}^2$ delimitada por las curvas $x_1, \ldots, x_k$ es

$$B(x_1, \ldots, x_k) = \{(t, x(t)) : t \in T, \min_{r=1,\ldots,k} x_r \leq x(t) \leq \max_{r=1,\ldots,k} x_r\}.$$

Consideremos la banda subyacente de las 50% curvas más profundas:

$$C_{0.5} = \left\{(t, x(t)) : \min_{r=1,\ldots,\lfloor n/2 \rfloor} X_r(t) \leq x(t) \leq \max_{r=1,\ldots,\lfloor n/2 \rfloor} X_r(t)\right\}.$$

Los valores atípicos se identifican utilizando una extensión del criterio IQR, similar a los boxplots clásicos:

$$Bigotes = C_{0.5} \pm 1.5 \times \mathrm{IQR}_{\text{funcional}}$$

K-Means

En datos funcionales, las métricas de distancia deben capturar diferencias en el comportamiento global de funciones completas en lugar de puntos discretos.
La distancia L^p es comúnmente utilizada para este fin:

$$d(f,g) = \left(\int (f(t) - g(t))^p \, dt \right)^{1/p}$$

Para $p = 2$, esta distancia es una generalización de la distancia Euclídea al espacio de funciones

Implementación

Este estudio utiliza una base de datos que recoge resultados de la prueba de tolerancia oral a la glucosa (OGTT) realizada a pacientes durante 120 minutos.

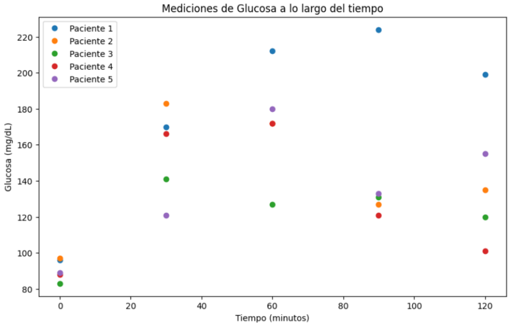


Figure: Mediciones de prueba de tolerancia a glucosa de 5 pacientes

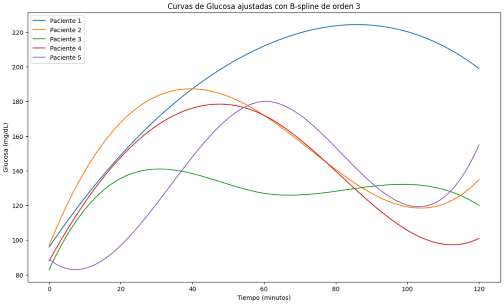
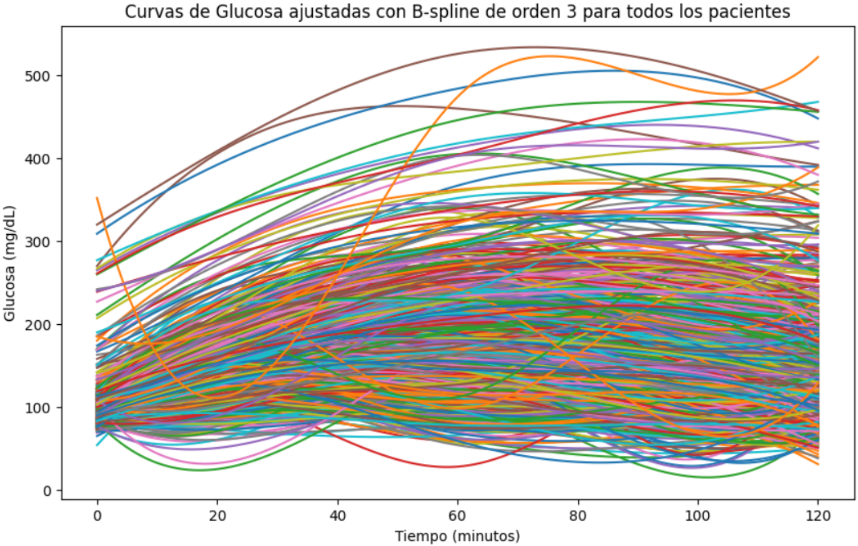
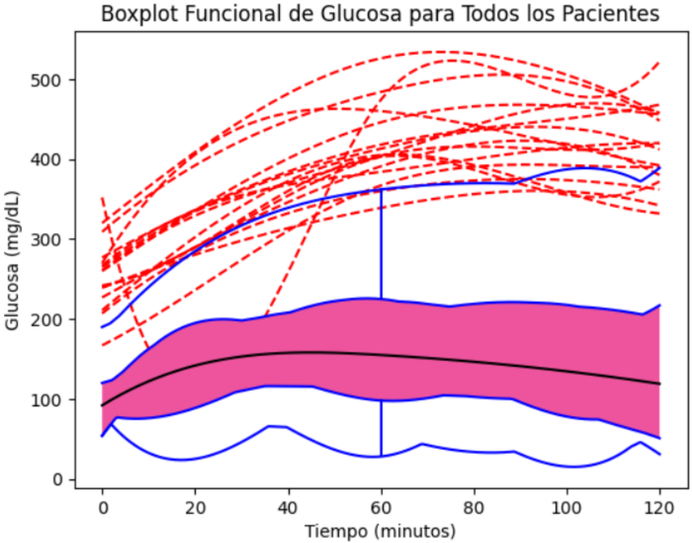


Figure: Mediciones de la glucosa con B -Splines de orden 3

Visualicemos



Baseline


Boxplot Funcional de Glucosa para Todos los Pacientes
Glucosa (mg/dL)
0
100
200
300
400
500
0
20
40
60
80
100
120
Tiempo (minutos)

Análisis por genero

Mujeres

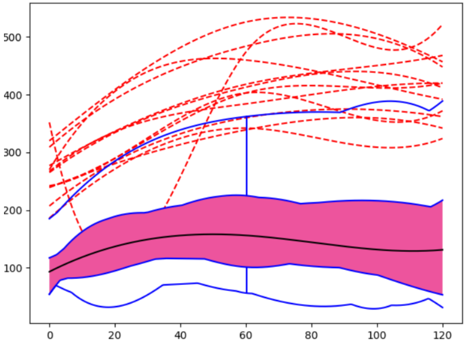


Figure: Boxplot para los datos de hombres

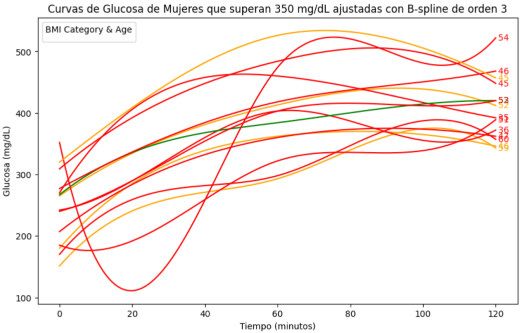


Figure: Pacientes que superan 370 mg/dL de glucosa

Análisis por genero

Hombres

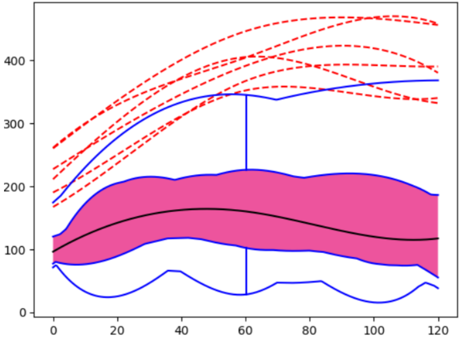


Figure: Boxplot para los datos de hombres

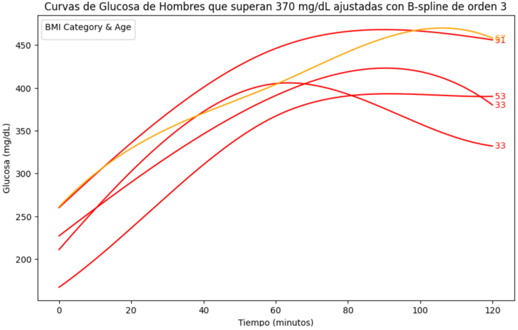


Figure: Pacientes que superan 370 mg/dL de glucosa

Análisis por condición médica

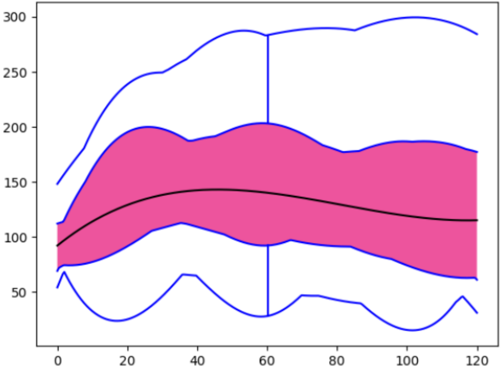


Figure: Boxplot para los datos de pacientes normales

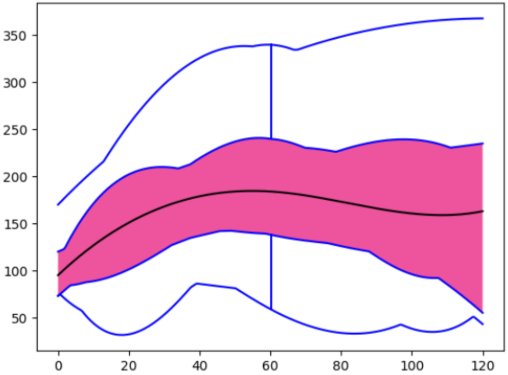


Figure: Boxplot para los datos de pacientes prediabeticos

DIABETES TIPO 2

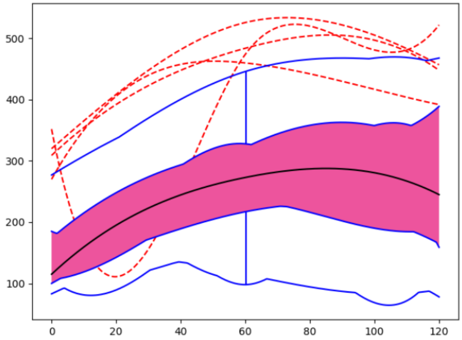


Figure: Boxplot para los datos de pacientes con diabetes tipo 2

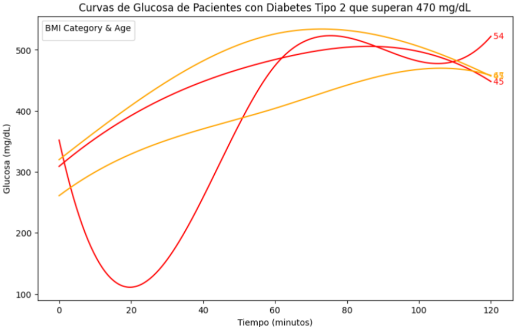
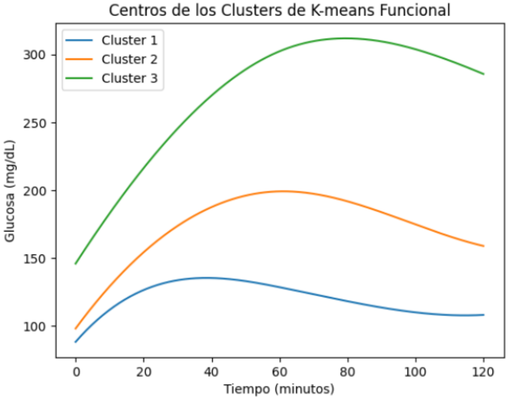
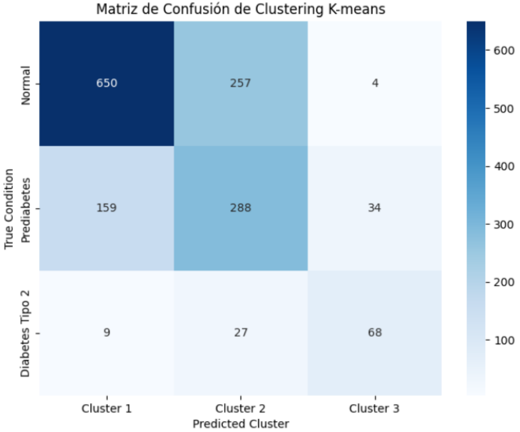


Figure: Pacientes con diabetes tipo 2 que superan 470 mg/dL

K MEANS


Centros de los Clusters de K-means Funcional
Cluster 1
Cluster 2
Cluster 3
300
250
200
150
100
0
20
40
60
80
100
120
Tiempo (minutos)
Glucosa (mg/dL)


Matriz de Confusión de Clustering K-means
True Condition
Normal
Prediabetes
Diabetes Tipo 2
650
257
4
159
288
34
9
27
68
Cluster 1
Cluster 2
Cluster 3
Predicted Cluster
600
500
400
300
200
100

Ventajas

- Nombrado por primera vez en Dalzell & Ramsay, 1991.
- Poca penetración en campos aplicados .
- Varios metodologías en competencia.
- Recursos/software públicos limitados.
- Análisis de datos más que inferencia.
- Los datos necesitan preprocesamiento.

Bibliografía I

- López-Pintado, S. y Romo, J. (2009). Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 53(4), 1562-1576.
- Sun, Y. y Genton, M. G. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics*, 20(2), 316-334.
- Ramsay, J.O., Silverman, B.W. (2005). Functional Data Analysis (2nd ed.). Springer Series in Statistics. Springer.
- Hernando Bernabé, A. (2017). Development of a Python package for Functional Data Analysis. Universidad Autónoma de Madrid.
- Murillo González, L. (2021). Estudio de la evolución mundial del Covid-19 mediante análisis de datos funcionales. Trabajo de Fin de Máster, Universidad de Granada, Máster en Estadística Aplicada.

¡Gracias por tu atención!

Ing. Nelson Ariza Morales
nelson.ariza@cimat.mx