Estimación Penalizada con Datos Funcionales (Análisis de Datos Funcionales)

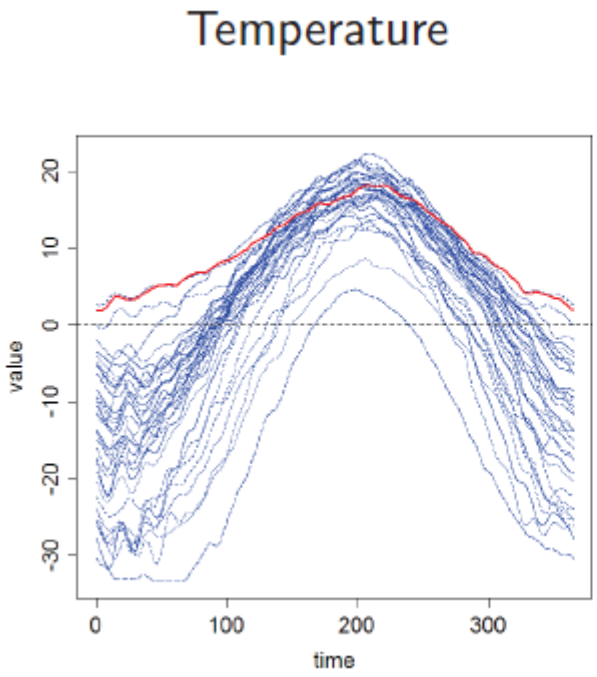
Mat. Juan Carlos García, MSc.
Ing. Mat.(e) Elena Ochoa
Escuela de Ciencias Físicas y Matemáticas
Universidad de las Américas(UDLA)

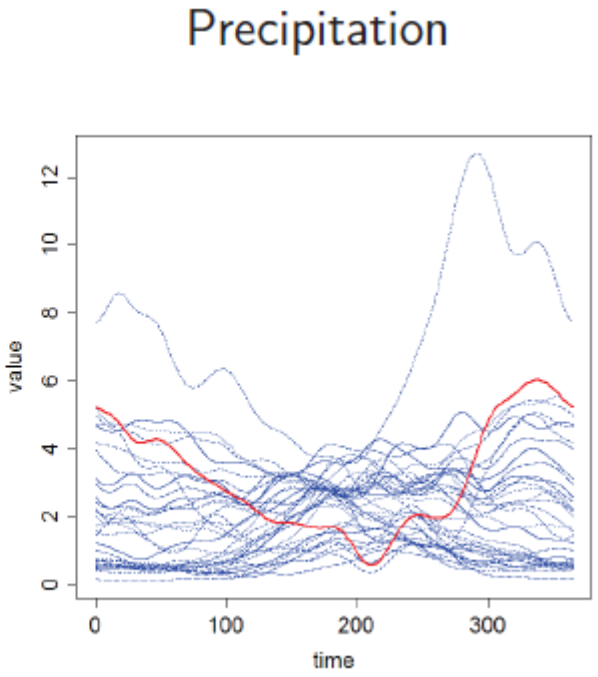
INTRODUCCIÓN

- El análisis de datos multivariantes permite el estudio de observaciones que constituyen un conjunto finito de números; sin embargo, en los casos reales aparecen situaciones donde los datos que se estudian son procesos continuos (temperaturas, cotizaciones bursátiles, etc.) y ante estos nuevos retos surge como respuesta la estadística de datos funcionales .

Qué es el Dato Funcional

Datos Meteorológicos:
Cambio de temperatura en el transcurso de un año, tomados treinta-cinco estaciones meteorológicas en todo Canadá.





Caracterización y clasificación de diferentes tipos de cánceres :
Donde se tiene una muestra aleatoria de 25 muestras tumorales (normales) y 25 muestras tumorales (malignas), donde se han medido los niveles de expresión de 100 genes.

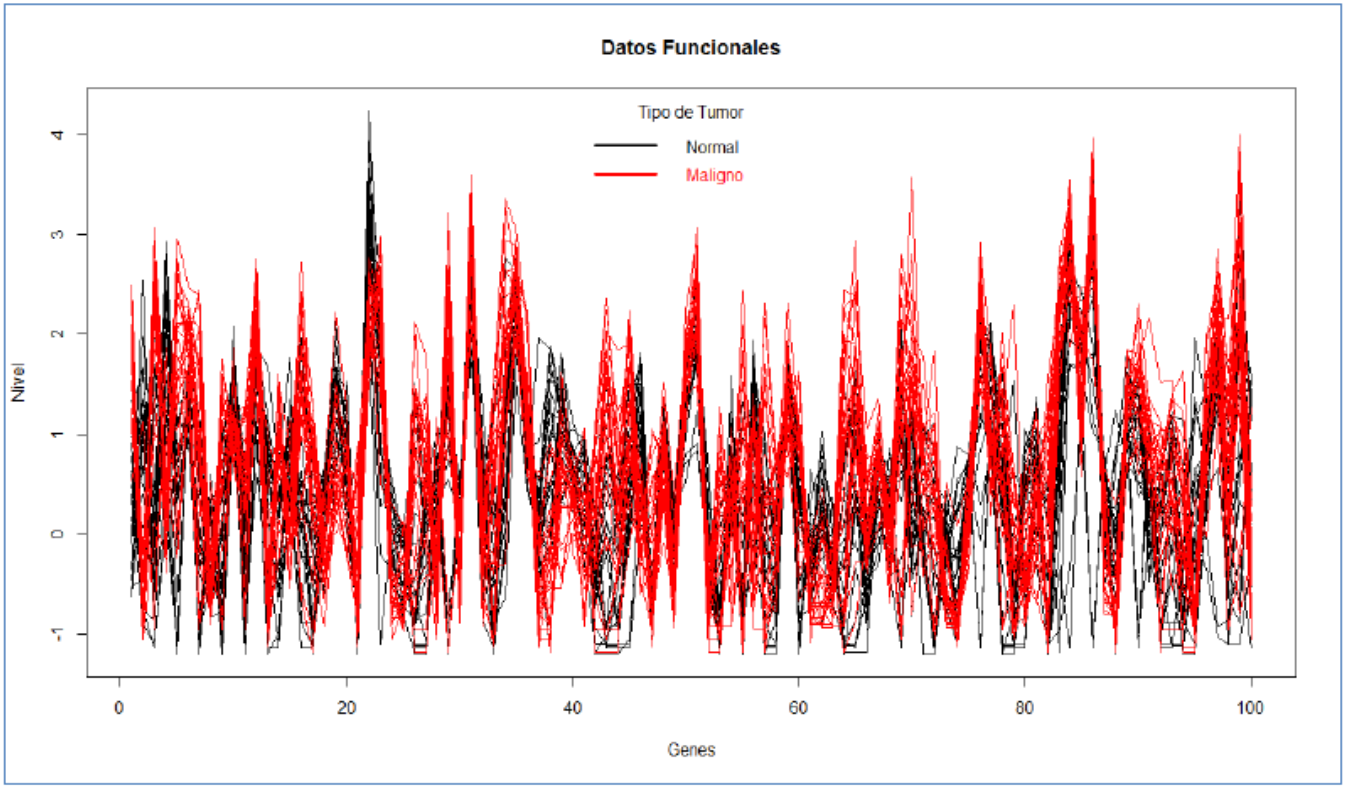
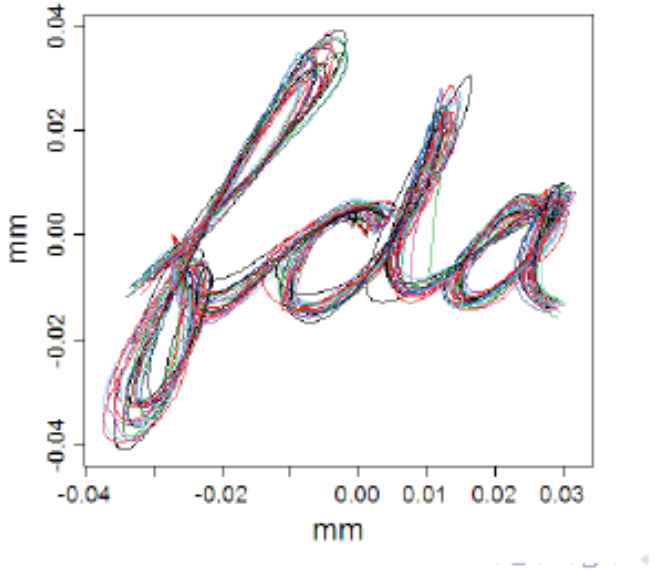


Figura 1: El gráfico de los datos funcionales representados mediante curvas de color negro para los tumores normales y rojo para los malignos.

Datos de escritura a mano:
Las medidas de la posición de la punta de una pluma de escribir "FDA". 20 repeticiones, mediciones tomadas a 200 hertzios.



En el ADF la unidad básica de información es la función completa , mas que un conjunto de valores.

VARIABLE FUNCIONAL

Definición : Una variable aleatoria $\mathcal{X}$ se dice que es una variable funcional si toma valores en un espacio infinito dimensional (espacio funcional) E . Una observación x de $\mathcal{X}$ es llamado un dato funcional.

LOS PRIMEROS PASOS EN AL ANÁLISIS DE DATOS FUNCIONAL

- ¿ Qué espacio funcional describe nuestra situación?

$\mathcal{L}_p([a,b]), p \neq 2$: Espacio de Banach.

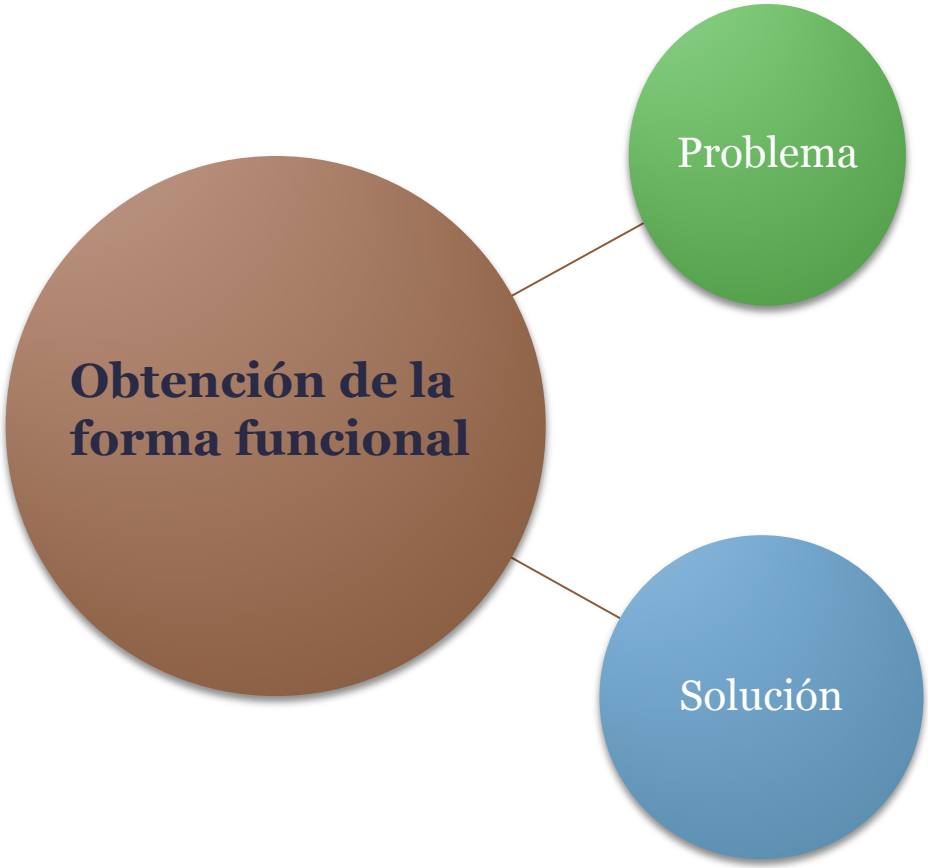
$$d(\mathcal{X}_i, \mathcal{X}_j) = ||\mathcal{X}_i - \mathcal{X}_j||_p$$

$\mathcal{L}_2([a,b])$: Espacio de Hilbert separable.

$$\langle f, g \rangle = \int f(x)g(x)d\mu$$

Base ortonormal: $\mathcal{X} = \sum_{i=1}^{\infty} \langle \mathcal{X}, e_i \rangle e_i$

$$d(\mathcal{X}_i, \mathcal{X}_j) = ||\mathcal{X}_i - \mathcal{X}_j|| = \sqrt{\langle \mathcal{X}_i - \mathcal{X}_j, \mathcal{X}_i - \mathcal{X}_j \rangle}$$



- Dispone de datos funcionales x_{ij} (t_j) discretizados en un conjunto de puntos

$$\{t_j\}_{j=1}^{N} \in [a, b]$$

- Representación básica de:

$$y_{ij} = \chi_i(t_j) + \varepsilon_j$$

Con t_j continua y $\chi_i(t_j)$ suave.

Base: $\{\phi_1(t), \ldots, \phi_p(t)\}$

$$\chi_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t) \quad i=1,\ldots,N$$

Representación de datos funcionales

- En representación no paramétrica de las funciones.

Método Basis-expansion

$$\chi_i(t_j) = \sum_{j=1}^{p} a_{ij}\phi_j(t) \quad i=1,\ldots,N$$

Sistema base: Bases de B-splines y Sistemas base de Fourier

Coeficientes básicos

Bases de Fourier

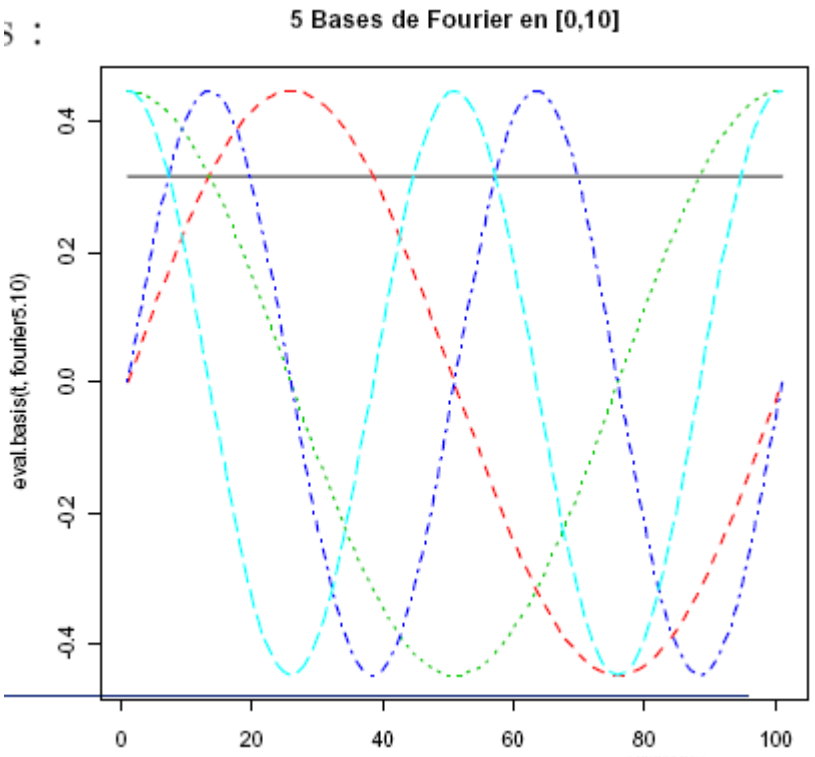
Los datos deben mostrar una cierta periodicidad ,como por ejemplo variaciones de temperatura.

$$\text{periodo } \frac{2\pi}{w}$$

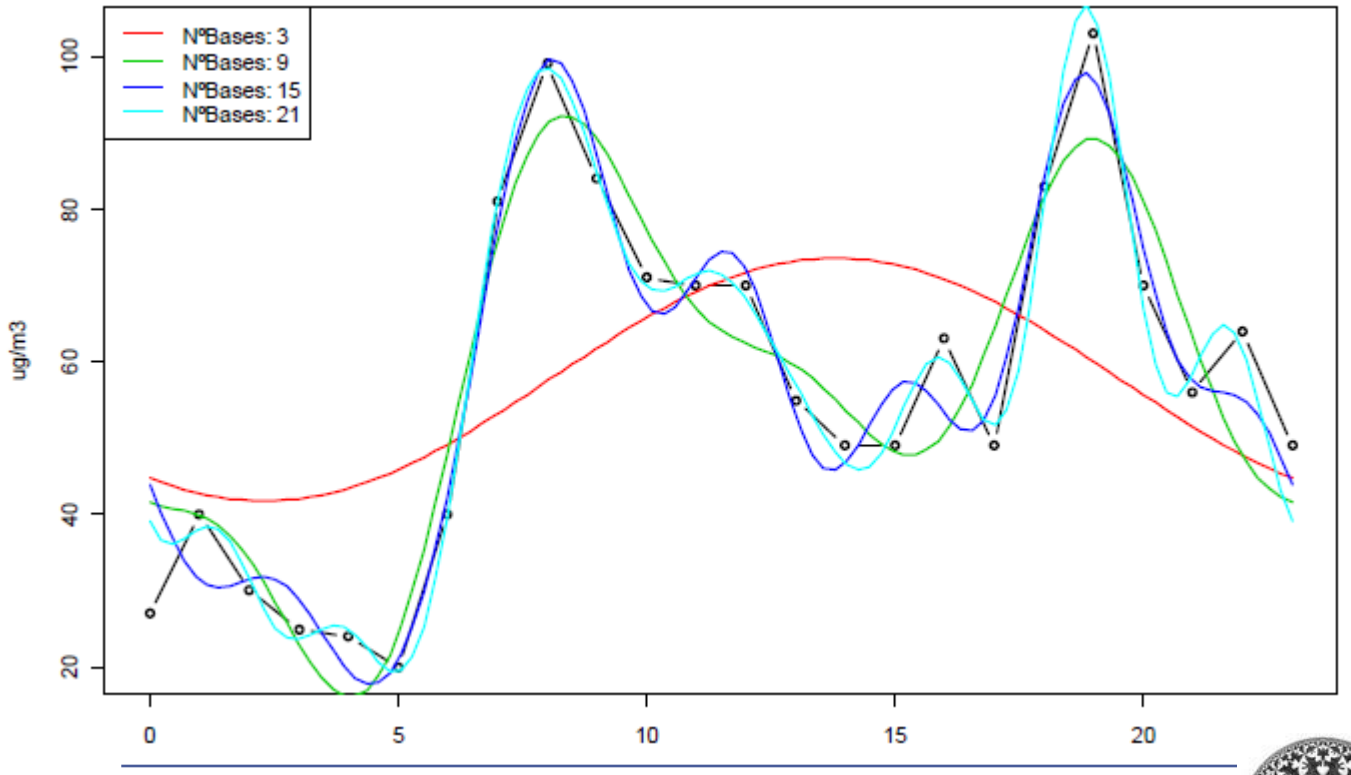
$$T = [0, T] \text{ y } w = \frac{2\pi}{T}$$

Funciones otonormales:

$$\phi_0(t) = \frac{1}{\sqrt{T}},$$
$$\phi_{2r-1}(t) = \frac{\sin(rwt)}{\sqrt{T/2}}, \text{ y}$$
$$\phi_{2r}(t) = \frac{\cos(rwt)}{\sqrt{T/2}}.$$



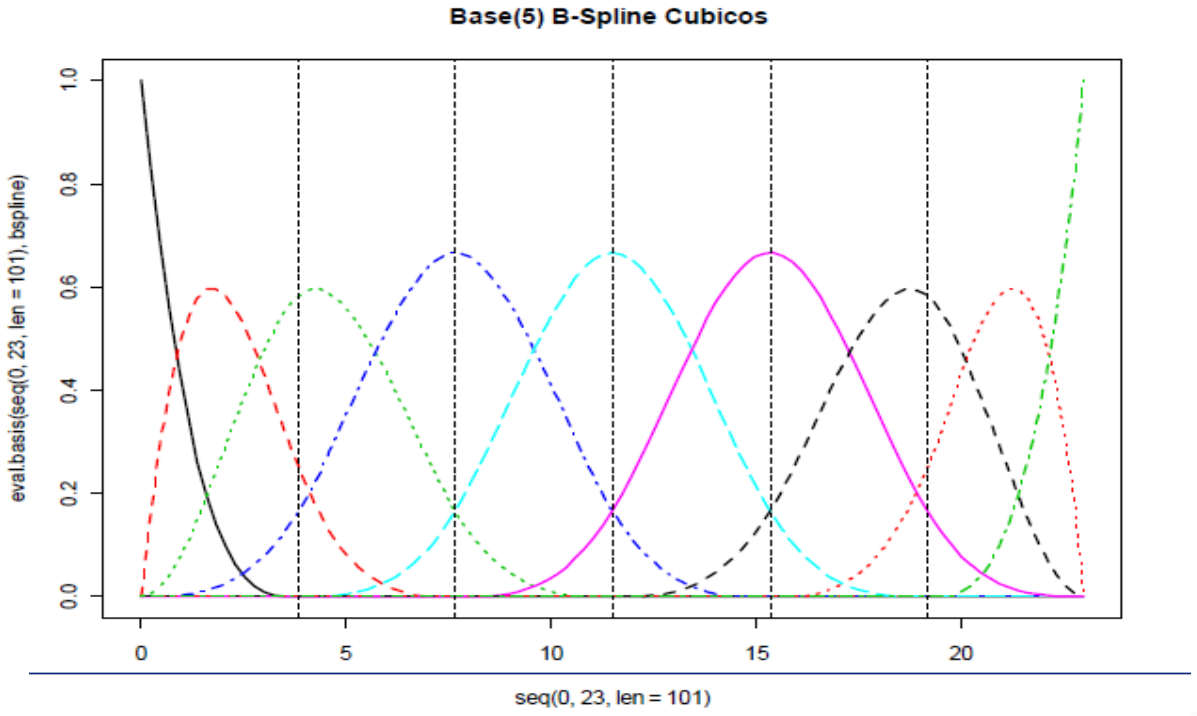
Representación con Fourier



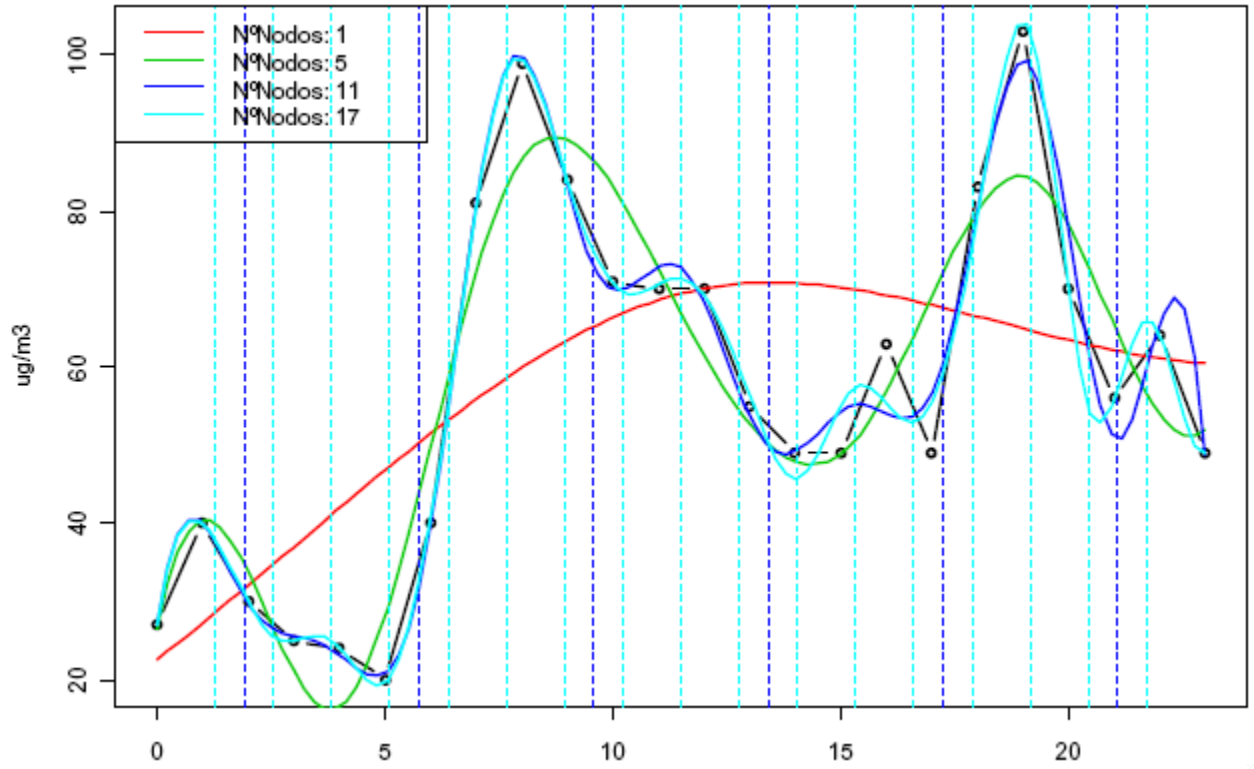
BASES DE B-SPLINES

- Splines son segmentos de polinomio se unieron de extremo a extremo.
- Los puntos en los que los segmentos se unen se denominan nodos.
- Sistema definido por:
  - El orden m (orden = grado +1) del polinomio
  - La ubicación de los nodos.
- Número de funciones bases:
  - Orden + número de nodos interiores

- Entre mayor sea el número de nodos la curva presentara mayor flexibilidad
- Los B-splines se calculan con el algoritmo de Boor, entre ellos los más usados son los cúbicos.



REPRESENTACIÓN CON B-SPLINES



Estimación aproximada de los coeficientes básicos $\{a_{ij}\}$

- Existen 2 formas de aproximar los coeficientes básicos :
  - Si el predictor funcional es observado con error se usa una aproximación suave, como la aproximación de mínimos cuadrados, eligiendo las bases adecuadas.

$$y_{ij} = \chi_i(t_j) + \varepsilon_j \qquad j=1,\ldots,N$$

  - Si son observadas sin error se usan métodos de interpolación , como la interpolación Spline Cúbica.

$$y_{ij} = \chi_i(t_j) \qquad j=1,\ldots,N$$

Aproximación de mínimos cuadrados

Ajustar una curva a las observaciones discretas y_{ij}

Con i=1,...,n y j=0,...,m_i usando el modelo $y_{ij} = \chi_i(t_j) + \varepsilon_j$

y una expansión en términos de funciones básicas para

$$\chi_i(t_j) = \sum_{j=1}^{p} a_{ij}\phi_j(t)$$

Y los valores estimados por el modelo en los nodos de observación serán de la forma

$$\chi_i = \Phi a_i$$

Aproximación de mínimos cuadrados

Los coeficientes de la expansión básica, a_{ij} , *se determinan por el criterio* de mínimos cuadrados, y en forma matricial se tiene que

$$ECM\left(\chi_i \mid a_i\right) = \left(\chi_i - \Phi_i a_i\right)' \left(\chi_i - \Phi_i a_i\right)$$

el estimador de a_i *que minimiza el error de mínimos cuadra*dos es

$$\hat{a}_i = \left(\Phi_i' \Phi_i\right)^{-1} \Phi_i' \chi_i .$$

Los valores ajustados en los nodos de observación son

$$\hat{\chi}_i = \Phi_i \hat{a}_i = \Phi_i \left(\Phi_i' \Phi_i\right)^{-1} \Phi_i' \chi_i$$

La aproximación por mínimos cuadrados es adecuada cuando se asume que los residuos sobre la verdadera curva son independientes e igualmente distribuidos con media 0 y varianza constante.

ESTIMACIÓN POR MÍNIMOS CUADRADOS PENALIZADOS

En un ajuste de curvas con mínimos cuadrados no es fácil controlar el grado de suavidad de la curva ajustada.

Estimación por Splines

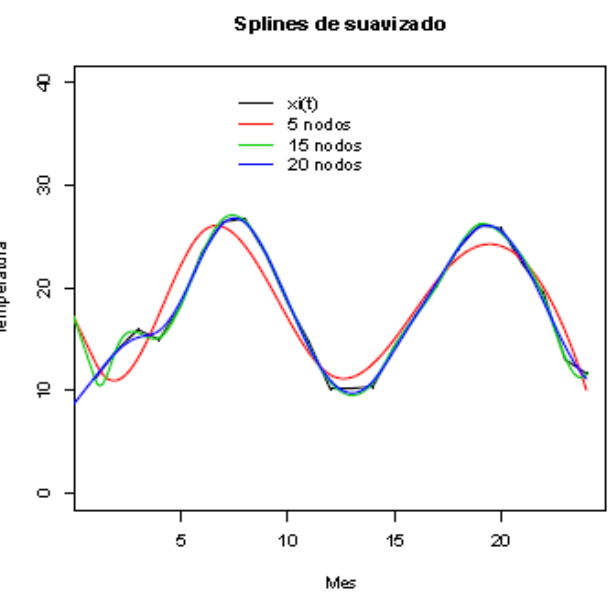
- En el suavizado con splines para el ajuste de curvas:

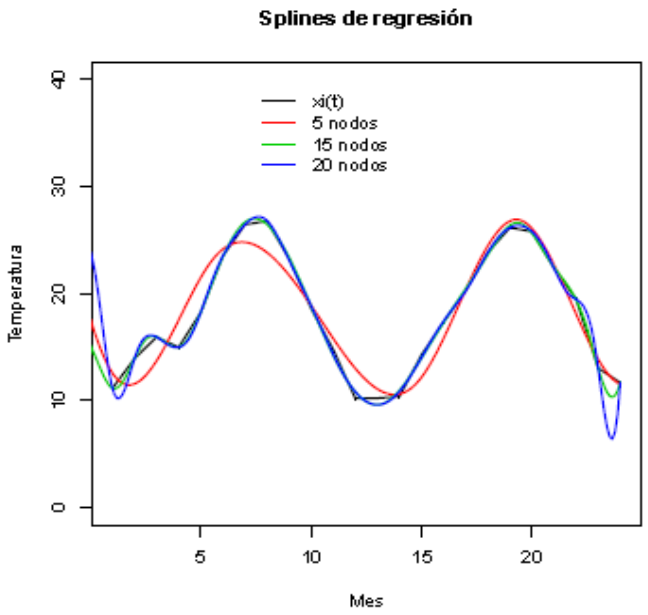
1. Splines de suavizado(smoothing splines)
2. Splines de regresión(regression splines)

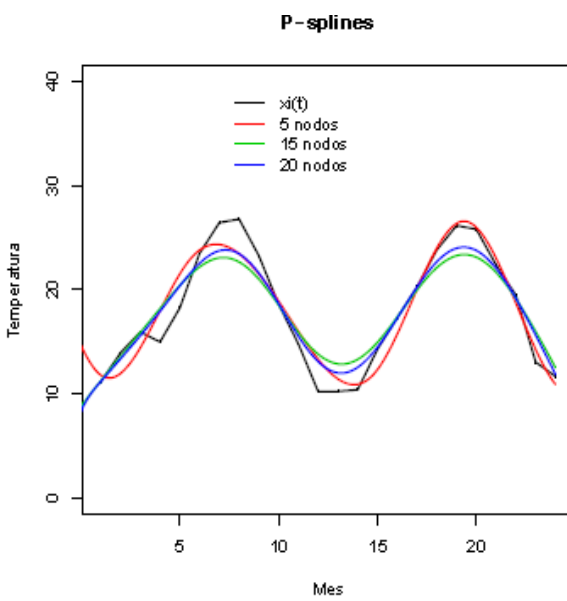
ESCUELA DE CIENCIAS
FÍSICAS Y MATEMÁTICAS

ESTIMACIÓN POR MÍNIMOS CUADRADOS PENALIZADOS

Splines de suavizado + Splines de regresión = Splines con penalización o P-splines

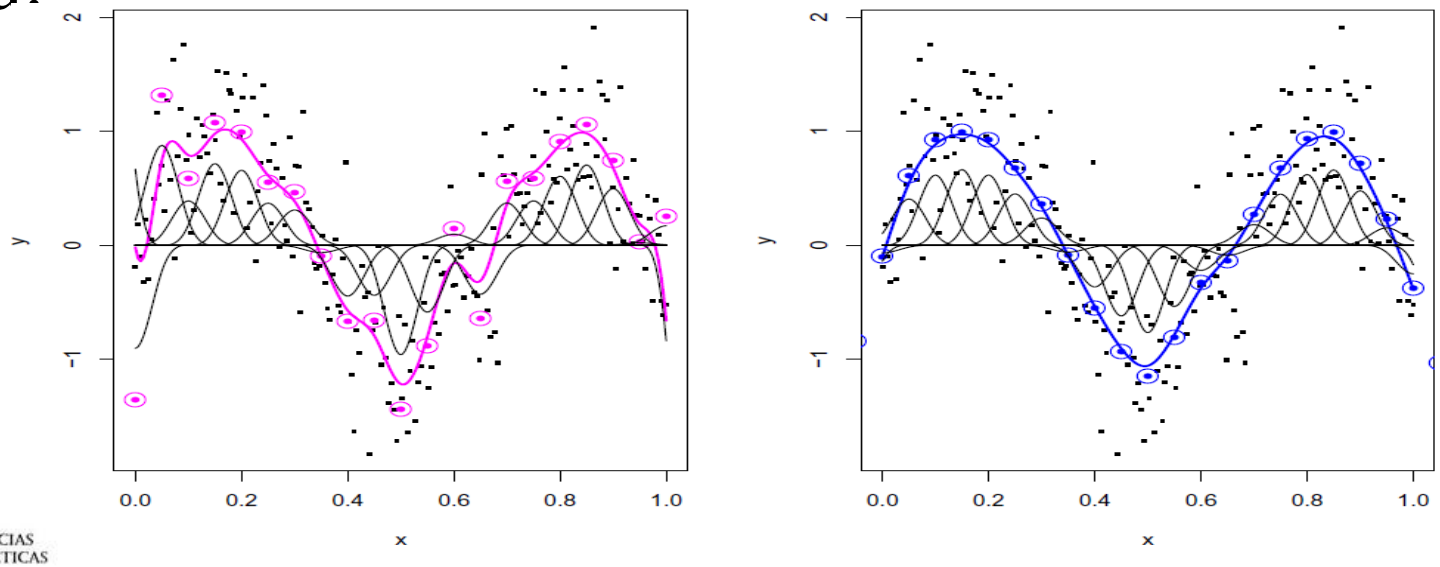






ESTIMACIÓN POR MÍNIMOS CUADRADOS PENALIZADOS

- Con los P-splines la penalización es discreta y que se penalizan los coeficientes básicos de las curvas directamente en lugar de penalizar la curva.



ESCUELA DE CIENCIAS FÍSICAS Y MATEMÁTICAS

Figura 6: Curva estimada con 20 nodos, sin penalizar los coeficientes (izquierda) y penalizando los coeficientes (derecha).

Penalización de la suavidad(P-Splines)

- La integral de la segunda derivada de la curva ajustada $\chi_i(t_j) = a_i'\phi(t)$ al cuadrado en un instante t, se le considera como la curvatura de dicha función en t, y con la finalidad de cuantificar la suavidad de cada una de las curvas $\chi_i(t_j)$, se define la función:

$$PEN_d(\chi_i) = \int [D^d\, a_i'\phi(s)]^2\, ds$$

$$= a_i' \left[\int D^d\phi(s)\, D^d\phi'(s) ds \right] a_i$$

Penalización de la suavidad(P-Splines)

Los P-splines: es una buena aproximación discreta de la integral de la d-ésima derivada al cuadrado

$$PEN_d(\chi_i) = \int [D^d\, a_i' \phi(s)]^2\, ds$$

La penalización se añade a la función de mínimos cuadrados, dando lugar a la función de mínimos cuadrados penalizados.

$$ECMPEN_2(\chi_i, \lambda_i | a_i) = (\chi_i - \Phi_i a_i)'(\chi_i - \Phi_i a_i) + \lambda_i\, a_i' \left[\int D^d \phi(s)\, D^d \phi'(s) ds\right] a_i$$

cuya solución de los coeficientes estimados es:

$$\hat{a}_i = (\Phi_i' \Phi_i + \lambda_i P_d)^{-1} \Phi_i' \chi_i,$$

Criterio de Validación

La dimensión infinita de los datos funcionales

El criterio de validación proyecta los datos funcionales en un espacio K- dimensional

Es importante el número de parámetros de la base y la elegir la base más apropiada para los datos

Entre los diferentes criterios de selección para seleccionar el parámetro $\nu = (K, \lambda)$, hemos puesto en marcha los dos siguientes:

.Criterio de validación (CV),Criterio de Validación Cruzada (GCV) y el Criterio de información de Akaike

Criterio de información de Akaike, consiste en elegir para

cada curva el valor de λ_i que minimice la expresión para los P-splines

SELECCIÓN DEL PARÁMETRO DE SUAVIZADO

- El parámetro de suavizado en los P-splines controlar la suavidad de la curva.
- Los P-splines penalizan los coeficientes que están muy separados entre sí.
- Cuanto de modo que $\lambda_i \to \infty$ obtenemos un ajuste polinómico.
- Cuando $\lambda_i \to 0$ estaremos utilizando mínimos cuadrados ordinarios y por tanto nos aproximamos a un ajuste lineal.

CRITERIO DE INFORMACIÓN DE AKAIKE

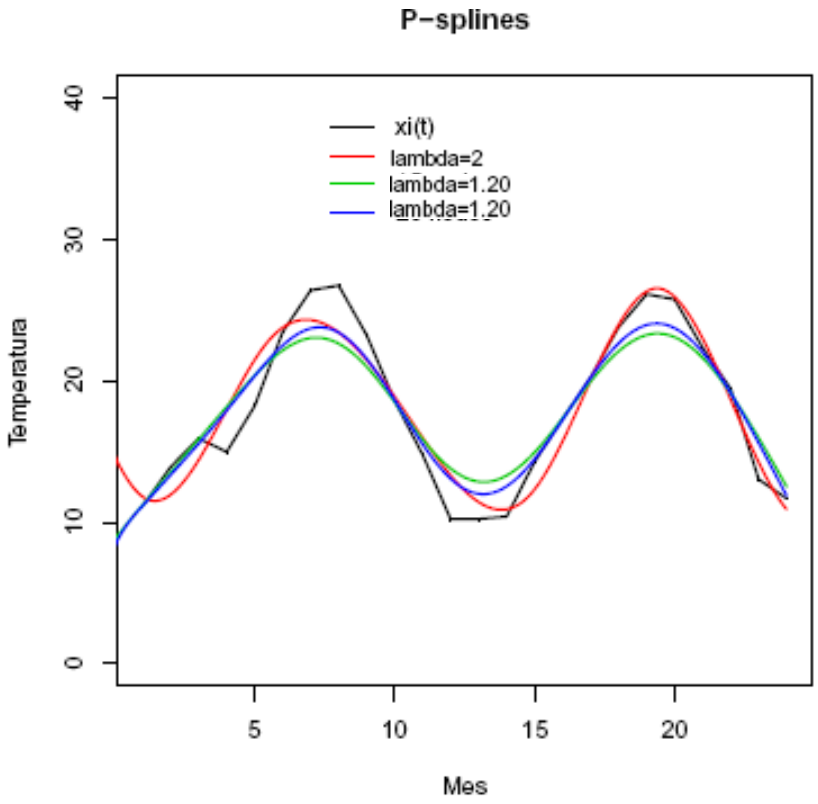
- Consiste en elegir para cada curva el valor de λ_i que minimice la siguiente expresión para los P-splines.

$$AIC = 2log\left(\sum_{k=0}^{m_i}\left(x_{ik}-\hat{x}_{ik}\right)^2\right) - 2\log\left(m_i\right) + 2\log\left(traza\left(H_i\right)\right)$$

donde

$$H_i = \Phi_i\left(\Phi_i'\Phi_i + \lambda_i P_d\right)^{-1}\Phi_i', \qquad \Phi_i = \left(\phi_j\left(t_{ik}\right)\right)_{m_i \times p}$$

$$df(\lambda_i) = traza(H_i)$$



LA CURVA AJUSTADA ES: $\hat{\chi}_i(t) = \hat{a}_i' \phi(t)$.

Análisis descriptivo de Datos Funcionales

- Media: $\bar{\chi}(t) = \frac{\sum_{i=1}^{n} \chi_i(t)}{n}$
- Varianza: $Var(\chi(t)) = \frac{\sum_{i=1}^{n} (\chi_i(t) - \bar{\chi}(t))^2}{n-1}$
- Desviación estándar: $\sqrt{Var(\chi(t))}$
- Covarianza: $Cov(\chi(t_1), \chi(t_2)) = \frac{\sum_{i=1}^{n} (\chi_i(t_1) - \bar{\chi}(t_1))(\chi_i(t_2) - \bar{\chi}(t_2))}{n-1}$

Así, se puede concluir que las estadísticas descriptivas univariadas y bivariadas clásicas se aplican igualmente cuando se tienen datos funcionales. Sin embargo se resalta que en este caso, los objetos calculados corresponden a curvas.

Bibliografía:

[Ferraty and Vieu (2006)] Ferraty F, Vieu P (2006). *Nonparametric Functional Data Analysis.* Springer Series in Statistics. Springer-Velag, New York. Theory and practice.

[Ramsay and Silverman (2005)] *Functional Data Analysis.* Springer Series in Statistics, second edition. Springer-Velag, New York.

[Ramsay *et al.* (2010)] Ramsay JO, Wickham H GS, Hooker G (2010). ***fda:*** *Functional Data Analysis.* **R** package version 2.2.6., http://cran.r-project.org/package=fda.

De Boor, C. (1977). Package for calculating with B-splines. Journal of Numerical Analysis, 14, 441-472.

Eilers, P. y Marx, B. (1996). Flexible smoothing with B-splines and penalties. Statistical Science, 11, 89-121.

http://eio.usc.es/pub/MAESFE/ (librería fda.usc)

http://www.functionaldata.org (librería fda)

http://www.r-project.org/