

PROYECTO FIN DE CARRERA:

ANÁLISIS DE DATOS FUNCIONALES. IMPLEMENTACIÓN Y APLICACIONES.

Alumno: Valentín Navarro Pérez

Director de Proyecto: Pedro Delicado Useros.

Licenciatura en Ciencias y Técnicas Estadísticas.

Facultat de Matemàtiques i Estadística.

Universitat Politècnica de Catalunya

Julio 2004.



DADES DEL PROJECTE:

Nom de l'estudiant: Valentín Navarro Pérez

DNI: 34752340

Títol del Projecte: Análisis de datos financieros,
implementación y aplicación.

Director del Projecte: Pedro Delgado Uceros

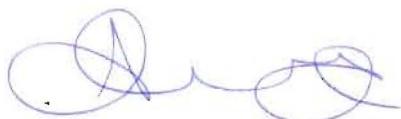
Tutor del Projecte:

QUALIFICACIÓ

10, Matrícula d'Honor

MEMBRES DEL TRIBUNAL (nom i signatura)

President:



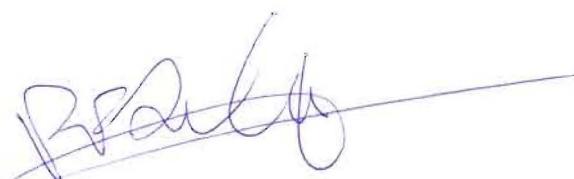
Tomás Aluja

Vocal:



Pere Pascual

Secretari



Pedro Delgado

Data: 7 - 7 - 2004

ÍNDICE:

1.	Introducción.	6
2.	Introducción a la teoría del análisis de datos funcionales.	11
2.1.	Análisis descriptivo funcional.	11
2.1.1.	Producto escalar y definiciones.	11
2.1.2.	Estadísticos descriptivos en notación de producto escalar	13
2.1.3.	Producto escalar. Definiciones para datos funcionales.	15
2.1.4.	Estadísticos descriptivos en análisis de datos funcional.	17
2.1.4.1.	Estadísticos sobre una función.	17
2.1.4.2.	Estadísticos de muestras de una función aleatoria.	22
2.1.4.3.	Estadísticos de muestras de dos o más funciones aleatorias.	23
2.2.	Análisis de componentes principales en espacios con producto escalar.	24
2.2.1.	Proyecciones expresadas como productos escalares	25
2.2.2.	Elementos propios	27
2.2.3.	El problema del análisis de componentes principales mediante producto escalar.	28
2.3.	Análisis de componentes principales funcionales (ACPF).	29
2.3.1.	Introducción y definiciones en el espacio funcional L^2 .	29
2.3.2.	El problema del análisis de componentes principales funcionales.	31
2.3.3.	Interpretación de las ACPFs.	33
2.3.4.	ACPF para funciones representadas a través de funciones base conocidas.	35
3.	Estimación de funciones por splines	37
3.1.	Representación de datos. Suavizado e interpolación	37
3.2.	Teoría de splines cúbicos interpoladores.	38
3.3.	Cálculo computacional de splines cúbicos interpoladores: método de los momentos.	41
3.4.	Regresión no paramétrica con splines.	43
4.	Métodos computacionales y subrutinas en R.	47
4.1.	Operaciones con splines en lenguaje R.	47

4.1.1.	Representación de splines en lenguaje R.	47
4.1.2.	Procedimientos de evaluación de splines.	49
4.1.3.	Suma o resta de splines.	50
4.1.4.	Producto de dos splines y cuadrado de un spline.	52
4.1.5.	Integral definida de un spline.	53
4.2.	Procedimientos de cálculo de R de la descriptiva funcional.	54
4.2.1.	Cálculo de la media de una función.	54
4.2.2.	Cálculo de la varianza de una función.	55
4.2.3.	Cálculo de la covarianza de dos splines registrados.	56
4.2.4.	Cálculo de los momentos no centrados de primer y segundo orden.	58
4.3.	Cálculo de componentes principales funcionales para funciones representadas en splines.	59
4.3.1.	Planteamiento del problema.	59
4.3.2.	Cálculo de vectores y valores propios mediante raíz cuadrada de W .	60
4.3.3.	Cálculo de los coeficientes.	61
4.3.4.	Función en R <i>fpea.sint</i> .	63
5	Etapa de registro.	64
5.1.	Etapa de registro.	64
5.2.	Registro de knots.	64
5.3.	Errores en la etapa de registro.	67
5.4.	Reducción del número de knots.	68
6	Análisis funcional de datos en pirámides de población.	70
6.1.	Aplicación del análisis funcional de datos a pirámides de población.	70
6.1.1.	Idea y definiciones.	70
6.1.2.	Obtención de datos.	72
6.1.3.	Transformación de pirámides en funciones.	73
6.2.	Análisis descriptivo de la serie.	74
6.2.1.	Análisis descriptivo de primer nivel.	74
6.2.2.	Análisis descriptivo de segundo nivel.	80
6.2.3.	Críticas sobre el uso de la descriptiva funcional.	82
6.3.	Análisis de datos funcionales en las pirámides.	83
6.3.1.	Cálculo, interpretación y nivel de representación de las CPFs.	83

6.3.2. Resultados y comentarios.	88
6.4. Críticas al trabajo.	92
7 Análisis funcional de datos en datos farmacocinéticos.	93
7.1. Planteamiento del problema.	93
7.1.1. Ensayos fase I en oncología.	93
7.1.2. Situación y objetivos.	95
7.2. Registro de datos.	96
7.2.1. Problemas en la elección del tipo de spline.	97
7.2.2. Origen del tiempo de las funciones y rango de registro.	98
7.2.3. Transformación de los datos.	100
7.2.4. Registro y reducción de knots de la lista de spline.	101
7.3. Análisis descriptivo.	103
7.3.1. Descriptiva funcional de primer nivel y segundo nivel grupo a grupo.	104
7.3.2. Descriptiva funcional entre grupos.	106
7.4. Componentes principales funcionales.	108
7.4.1. Preparación del análisis.	108
7.4.2. Estudio del Docetaxel.	108
7.4.3. Estudio del Paclitaxel.	112
8. Conclusiones	117
Bibliografía.	118
Apéndice	119
1. Detalles de cálculos	119
1.1. Cálculo de spline por regresión.	119
1.2. Cálculo de la matriz W.	120
2. Sistemas de ecuaciones tridiagonales.	122
2.1. Matrices tridiagonales.	122
2.2. Descomposición LU de una matriz tridiagonal.	123
2.3. Solución del sistema LUx=b	124
2.4. Subrutina en R del cálculo de splines interpoladores.	125
2.5. Subrutina en R del cálculo de splines regresión..	128
3. Otros algoritmos en R.	131
3.1. Algoritmo mutpol.	131

3.2.	Algoritmo int.pl	132
3.3.	Algoritmo med.f.sint.	133
3.4.	Algoritmo var.f.sint.	133
3.5.	Algoritmo covar.f.sint.	134
3.6.	Algoritmo mom.1	135
3.7.	Algoritmo mom.2	135
3.8.	Algoritmo desc.f.spls.	137
3.9.	Algoritmo de cálculo de componentes principales funcionales	142
4.	Otras funciones de soporte.	144
4.1.	Algoritmo buscacoef.	144
4.2.	Algoritmo PL.	145
4.3.	Algoritmo spl.a2.	145

Agradecimientos.

Para Sandra;

Gracias por darme aliento y ánimo tanto durante la carrera como durante la elaboración del proyecto. Gracias por esperar y por no desesperar, en los momentos de trabajo desde casa.

1.-INTRODUCCIÓN.

Supongamos que un investigador quiere probar una hipótesis y para ello plantea un posible estudio a un estadístico. Éste piensa en qué datos se deberían recoger para reflejar fielmente el fenómeno estudiado. Pero simultáneamente también piensa cómo analizarlos: en el modelo que posteriormente va a utilizar para contrastar la hipótesis del investigador. Este hecho induce en ocasiones a que variables con una cierta naturaleza sean recogidas de forma distinta para que el análisis que el estadístico tiene en mente pueda ser aplicado.

Pongamos un ejemplo con análisis de medidas repetidas. Supongamos que queremos analizar el comportamiento a lo largo del tiempo de la concentración en la sangre de un quimioterápico en función de dos tipos de sistemas de tratamiento (o bombas). Un planteamiento posible sería medir la concentración para cada uno de los individuos en unos tiempos predefinidos, asignando al azar el tipo de esquema. Posteriormente realizaríamos un análisis de medidas repetidas y llegaríamos a la conclusión de que; si las medias trazan a lo largo del tiempo una misma línea entonces las bombas de tratamiento con las que son iguales; si las medias de cada bomba trazan líneas paralelas un esquema tiene un efecto mayor constante en el tiempo sobre la otra; y si las medias se cruzan entonces los esquemas producen un efecto distinto en la administración del medicamento. En este ejemplo podríamos escoger evaluar la concentración en intervalos de una centésima de día desde el origen del tiempo (véase figura 1.1). Nótese que al estar el eje transformado comenzaría en -4.6 ($\log(0.01)=-4.6$). Vemos que las funciones tienen un comportamiento muy distinto entre el tiempo inicial y el tiempo $=\exp(-3)$, sin embargo el análisis de medidas repetidas no lo detectaría.

En este ejemplo nos hemos dejado llevar por el tipo de análisis posterior, ya que en realidad la concentración de sangre a lo largo del tiempo no se ha considerado como una función sino como un vector de valores. La consecuencia directa en el ejemplo es la de que no vamos a saber qué valores se alcanzan en los tiempos en que no hemos evaluado la concentración. El planteamiento de medidas repetidas no va a responder bien a nuestra pregunta en aquellos tiempos no evaluados. Esto es así porque desde un inicio no hemos asumido que estábamos

trabajando con funciones. Se hace necesario por tanto realizar un trato estadístico que tenga en cuenta que estamos trabajando con este tipo de datos.

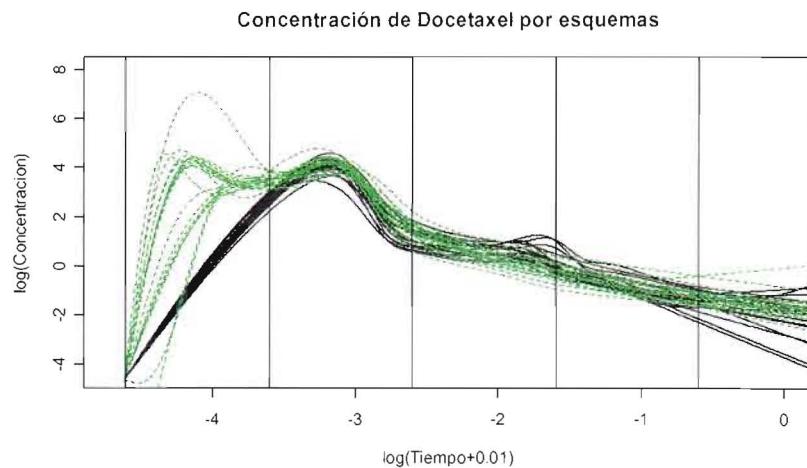


Figura 1.1: Representación de dos muestras de la concentración de docetaxel (en mg/m²) a lo largo del tiempo medido en porción de día y con transformación logarítmica ($\log(t+0.01)$).

El Análisis de datos funcionales

El análisis de datos funcionales es aquella parte de la estadística que trabaja con muestras de funciones aleatorias. Ramsay y Silverman (1997), entre otros, han introducido herramientas de análisis de para este tipo de datos. Las medidas de tendencia central, de dispersión y de relación entre variables aplicadas a muestras de variables aleatorias se pueden definir de manera análoga para muestras de datos funcionales. Sólo hace falta considerar que ahora estaremos trabajando en un espacio vectorial distinto: el espacio L^2 (funciones de cuadrado integrable). En realidad Ramsay y Silverman definen estas medidas manera general para cualquier espacio que tenga definido un producto escalar, y luego deduce de ellas las definiciones para muestras de funciones. En este proyecto se ha hecho hincapié en el análisis descriptivo y en el análisis de componentes principales. En el capítulo 2 haremos referencia a estas medidas.

Splines y etapa de registro

Las definiciones de las medidas que se utilizan en el análisis de datos funcionales necesitan para su cálculo la forma explícita de cada una de las funciones de la muestra, es decir $f_i(x)$ ha de ser

conocida. Sin embargo en la mayoría de las ocasiones no disponemos de su expresión explícita. En realidad la forma de $f(x)$ suele ser desconocida. Lo único que normalmente tenemos son observaciones de $f(x)$ realizadas en algunos puntos del eje de ordenadas. Además, en ocasiones estas evaluaciones suelen estar sometidas a errores de medida. La primera fase del estudio del análisis de datos funcionales en esta situación será, transformar nuestros puntos evaluados en funciones. Esto puede hacerse mediante el uso de splines, en los que haremos referencia en el capítulo 3.

Sin embargo no basta con pasar los datos a splines para poder empezar el análisis. Existen fuentes de “ruido” asociadas tanto a la recogida de datos discretos como a la elección del spline. Este ruido es el que se debe solventar en lo que llamaremos la etapa de registro de la que hablaremos en el capítulo 5.

Funciones en R

Dado que el análisis de datos funcionales es una herramienta novedosa de análisis, se ha elaborado un paquete específico de algoritmos con las que poder realizar los cálculos. Este conjunto de algoritmos ha sido implementado en el lenguaje orientado a objetos R. Se han definido los distintos tipos de objetos, los algoritmos de registro, de cálculo de splines, de cálculo de estadísticos descriptivo funcionales y de cálculo de componentes principales funcionales. En el capítulo 4 se detalla cómo se realizan los cálculos y qué es lo que hace cada uno de los algoritmos.

Aplicaciones de la descriptiva funcional y del análisis de componentes principales funcionales.

Para ilustrar las enormes posibilidades que el análisis de datos funcionales puede revelarnos sobre muestras de funciones se han utilizado dos ejemplos.

El primero, que se recoge en el capítulo 6, se ha escogido el recuento de población de 225 países y/o regiones (para simplificar a partir de ahora los llamaremos siempre países) del registro del 2001 del *Census Bureau* de los Estados Unidos. Las pirámides de población se han transformado para construir para cada uno de los países una función densidad (entre las edades -82.5 y +82.5, considerando las mujeres como edades negativas). Por lo tanto tendremos 225 funciones densidad a las que podremos aplicar tanto una descriptiva como un análisis de

componentes principales funcionales. Se han creado funciones en R específicas para trabajar **con pirámides de población** (véase como ejemplo la figura 1.3). Veremos cómo este tipo de **análisis** nos revela aspectos muy interesantes de las pirámides de población.

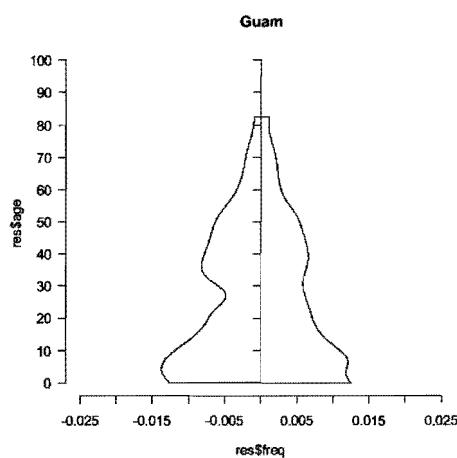


Figura 1.3: Representación funcional de la pirámide de población de Guam, utilizando la función en lenguaje R `piramyd.graph`.

El segundo, que se recoge en el capítulo 7, son datos de dos quimioterápicos que se **administran** a la vez en dos esquemas de tratamiento distintos. Estos datos han presentado **múltiples** problemas de registro y podremos observar cómo se han solucionado. El análisis **descriptivo** nos mostrará las relaciones entre los dos esquemas de tratamiento, y el análisis de **componentes principales funcionales** nos desvelará en qué intervalos de tiempo los esquemas **son equivalentes** y en cuales no.

Objetivo final del proyecto

El objetivo de este proyecto ha sido el de hacer posible un análisis de datos funcionales, **descriptivo** y de componentes principales funcionales. Es decir,

- Comprender la bibliografía y la teoría del análisis de datos funcionales tanto el análisis descriptivo como el análisis de componentes principales;
- Comprender la teoría de splines interpoladores o de regresión que nos permitan transformar los datos discretos en funciones;

- Implementar algoritmos en R que permitan registrar los datos, calcular descriptiva funcional y análisis de componentes principales; e
- Interpretar las medidas descriptivas y las componentes principales funcionales, tanto en general como para los dos ejemplos tratados.

2. – INTRODUCCIÓN A LA TEORÍA DEL ANÁLISIS DE DATOS FUNCIONALES.

2.1. – ANÁLISIS DESCRIPTIVO FUNCIONAL.

2.1.1.-Producto escalar y definiciones

Las medidas del análisis descriptivo pueden ser expresadas con notación algebraica. Ramsay y Silverman las definen así para que luego sean fácilmente adaptables a la naturaleza de nuestros datos: variables, vectores o funciones. La clave de todo esto es cómo definimos para cada elemento (variable, vector o función) el producto escalar.

Empezaremos con la definición de producto escalar y con sus propiedades. Tengamos en cuenta que ahora los elementos pueden ser de cualquier tipo: vectores o funciones.

- DEFINICIÓN 2.1:

El producto escalar euclíadiano de elementos x e y del espacio vectorial E , denotado como $\langle x, y \rangle$, es una aplicación

$$\langle \cdot, \cdot \rangle : E \times E \rightarrow \mathbb{R}$$

$$(x, y) \rightarrow \langle x, y \rangle$$

que satisface las siguientes propiedades

1. Simetría: $\langle x, y \rangle = \langle y, x \rangle$ para todo x e y ,
2. Positividad: $\langle x, x \rangle \geq 0$ para todo x , con $\langle x, x \rangle = 0 \Leftrightarrow x = 0$;
3. Bilinealidad: Para todo $a, b \in \mathbb{R}$, $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$

El producto escalar se puede interpretar como medida de cantidad de asociación entre dos elementos del espacio vectorial en el que trabajamos. También si lo aplicamos a un mismo elemento es una medida de magnitud del mismo elemento.

Podemos ampliar el producto escalar euclíadiano si introducimos un operador W . En este caso denotamos el producto escalar general como $\langle x, y \rangle_w$ y su definición dependerá del elemento del espacio vectorial con el que trabajamos. El producto escalar general será pues la aplicación sobre el espacio E que satisface las propiedades de simetría, positividad y bilinealidad al incorporar W .

Ejemplo 2.1: x e y son vectores en \Re^n y W una matriz definida positiva definimos el producto escalar como:

$$\langle x, y \rangle_W = x^T W y$$

- Y se cumplen las siguientes propiedades:

- (1) $x^T W y = y^T W x$ (simetría).
- (2) $x^T W x \geq 0$ y $x^T W x = 0 \Leftrightarrow x = 0$ (positividad)
- (3) $(ax + by)^T W z = ax^T W z + by^T W z$ (bilinealidad)

Tal como hemos dicho antes se pueden definir medidas de asociación entre elementos o bien de magnitud de un elemento dentro del espacio vectorial:

DEFINICION 2.2:

La norma de un elemento x del espacio, se define como:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

PROPIEDADES:

- (1) $\|x\| \geq 0$ y $\|x\| = 0 \Leftrightarrow x = 0$
- (2) $\|ax\| = |a| \cdot \|x\| \quad \forall a \in \Re$
- (3) $\|x + y\| \leq \|x\| + \|y\|$
- (4) $|\langle x, y \rangle| \leq \|x\| \cdot \|y\| = \sqrt{\langle x, x \rangle \cdot \langle y, y \rangle}$ (desigualdad de Cauchy-Schwartz).

COROLARIO:

De la desigualdad de Cauchy-Schwartz podemos deducir la desigualdad del coseno, que es la siguiente:

$$(5) \quad -1 \leq \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \leq +1$$

DEFINICIÓN 2.3:

Diremos que dos elementos $\langle x, y \rangle$ del espacio son ortogonales si

$$\langle x, y \rangle = 0$$

DEFINICIÓN 2.4:

El ángulo θ entre los elementos x e y los definimos como:

$$\theta = \arccos \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$$

Una posible interpretación de la norma de un elemento x puede ser la de la magnitud del elemento x dentro del espacio vectorial. La desigualdad de Cauchy-Schwartz nos está indicando que el valor absoluto del producto escalar de dos elementos está acotado por el producto de las normas de los elementos. Si el valor absoluto del producto escalar de estos dos elementos se aproxima a la cota entonces los elementos están definidos en direcciones del espacio vectorial semejantes, por lo que podemos decir que el grado de asociación entre ellos es grande. Sin embargo si el producto escalar es próximo a cero el grado de asociación es pequeño y las direcciones que definen los elementos son casi ortogonales. El símbolo negativo o positivo indica que el sentido creciente de la dirección definida por x corresponde a un sentido de decrecimiento en la dirección definida por y . Podemos ligar entonces el producto escalar al concepto de ángulo a través de la propiedad (5).

Ejemplo 2.2: Sea $x=(1,0)$ e $y=(0,-1)$ vectores en \mathbb{R}^2 . Entonces

$$\theta = \arccos \frac{\langle (1,0), (0,-1) \rangle}{\| (1,0) \| \cdot \| (0,-1) \|} = \arccos \frac{(1 \times 0) + (0 \times -1)}{\sqrt{(1 \times 1) + (0 \times 0)} \cdot \sqrt{(0 \times 0) + (-1 \times -1)}} = \arccos \left(\frac{0}{1} \right) = \frac{\pi}{2}$$

$\cos \theta = 0$

Estos dos vectores no tienen ningún grado de asociación.

DEFINICIÓN 2.5:

La distanza entre x e y se define como:

$$d_{x,y} = \| x - y \| = \sqrt{\langle x - y, x - y \rangle}$$

Si las direcciones y los sentidos que definen x e y son muy parecidos y además la norma de los elementos es parecida entonces el tamaño de $x-y$ será pequeño y por tanto la distancia entre x e y será corta. Tanto si las direcciones son perpendiculares como si las direcciones son iguales pero de sentido contrario, las distancias se harán grandes.

Ejemplo 2.3: Sea $x=(1,0)$, $y=(0,-1)$ y $z=(2,0)$ vectores en \mathbb{R}^2 . Entonces

$$(i) d_{x,y} = \| x - y \| = \| (1,0) - (0,-1) \| = \| (1,1) \| = \sqrt{1^2 + 1^2} = \sqrt{2}$$

$$(ii) d_{x,z} = \| x - z \| = \| (1,0) - (2,0) \| = \| (-1,0) \| = \sqrt{1^2 + 0^2} = 1$$

2.1.2. –Estadísticos descriptivos en notación de producto escalar.

Tal como habíamos anunciado en el capítulo 1, es posible generalizar definiciones para los estadísticos básicos de tendencia central (media) y de dispersión (varianza o correlación) si los definimos a través del producto escalar, sin especificar el espacio vectorial en el que estamos trabajando. Veremos las definiciones más importantes:

DEFINICIÓN 2.6:

Definimos la media como la cantidad obtenida al realizar el siguiente producto escalar:

$$\bar{x} = \frac{1}{N} \langle x, \mathbf{1} \rangle,$$

donde $\mathbf{1}$ es elemento unidad y $N = \langle \mathbf{1}, \mathbf{1} \rangle = \|\mathbf{1}\|^2$.

Ejemplo 2.4: Sea el espacio de posibles muestras $X = (X_1, \dots, X_N)$ de tamaño N , para una muestra en particular:

$$\begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}$$

- El elemento unidad será el vector: $\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}$,
- Y la media la definimos como:

$$\bar{x} = \frac{1}{N} \langle x, \mathbf{1} \rangle = \frac{1}{N} (x_1 \quad \dots \quad x_N) \cdot \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{N} (x_1 + x_2 + \dots + x_N)$$

DEFINICIÓN 2.7:

Definimos la varianza como la cantidad obtenida al realizar la siguiente operación:

$$S_x^2 = \frac{1}{N} \cdot \langle x - \bar{x} \cdot \mathbf{1}, x - \bar{x} \cdot \mathbf{1} \rangle = \frac{1}{N} \|x - \bar{x} \cdot \mathbf{1}\|^2$$

donde $\bar{x} \cdot \mathbf{1} = (\bar{x}, \bar{x}, \dots, \bar{x})$

Se puede interpretar esta medida como el módulo de la diferencia entre el elemento x respecto al elemento media.

DEFINICIÓN 2.8:

Definimos la covarianza entre x e y como la cantidad obtenida al realizar el siguiente producto escalar:

$$S_{x,y} = \frac{1}{N} \cdot \langle x - \bar{x} \cdot \mathbf{1}, y - \bar{y} \cdot \mathbf{1} \rangle$$

En primer lugar los elementos x e y los centramos en el elemento cero quedándonos así con el producto escalar entre x e y centrados. Podemos entonces aplicar la idea anteriormente expuesta: el producto escalar es una medida de la relación entre x centrada e y centrada.

DEFINICIÓN 2.9:

Definimos la *correlación entre x e y* como la cantidad obtenida al realizar el siguiente producto escalar:

$$r_{x,y} = \frac{S_{x,y}}{S_x S_y} = \frac{\langle x - \bar{x}\mathbf{1}, y - \bar{y}\mathbf{1} \rangle}{\|x - \bar{x}\mathbf{1}\| \cdot \|y - \bar{y}\mathbf{1}\|} = \cos(x - \bar{x}\mathbf{1}, y - \bar{y}\mathbf{1})$$

La *correlación entre x e y* es la proporción de la magnitud explicada por la relación de direcciones entre x e y respecto al total de magnitud de estos dos elementos ($S_x \cdot S_y$).

Ejemplo 2.5: Sea el espacio de posibles muestras $X = (X_1, \dots, X_N)$ de tamaño N , para una muestra en particular:

$$(1) S_x^2 = \frac{1}{N} (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_N - \bar{x}) \cdot \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_N - \bar{x} \end{pmatrix} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \text{ varianza de una muestra.}$$

$$(2) S_{x,y} = \frac{1}{N} (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_N - \bar{x}) \cdot \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{pmatrix} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \text{ covarianza de una muestra.}$$

una muestra.

$$(3) r_{x,y} = \frac{S_{x,y}}{S_x S_y} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \cdot \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}} \text{ es la correlación entre dos muestras.}$$

2.1.3. -Producto escalar. Definiciones para datos funcionales

Si nos referimos al análisis funcional de datos nuestros elementos x e y son funciones de \mathfrak{R} en \mathfrak{R} , y W es una función de peso definida positiva. En adelante todas las integrales serán definidas sobre \mathfrak{R} , salvo que explícitamente se indique lo contrario. En este caso:

DEFINICIÓN 2.10:

Sea el espacio de funciones de cuadrado integrable,

$$L^2 = \left\{ f : \mathbb{R} \rightarrow \mathbb{R} : \int_{\mathbb{R}} f^2(t) dt < \infty \right\}.$$

Definimos el *producto escalar euclíadiano* en L^2 como:

$$\langle x, y \rangle = x' y = \int x(t) y(t) dt$$

OBSERVACIÓN: El espacio de funciones de cuadrado integrable L^2 es un espacio de Hilbert.

Cuando nos referimos al espacio L^2 podemos imaginar que estamos trabajando con un vector de infinitas componentes, una para cada valor $t \in \mathbb{R}$. El producto escalar entonces será la suma del producto de las infinitas componentes de x e y . Teniendo en cuenta que la distancia entre componente y componente es infinitesimal, el producto escalar se expresa a través de integrales. Observemos que si $x(t)$ crece a la vez que $y(t)$ en los mismos intervalos de t entonces el producto escalar se hará grande; mientras que si cuando $x(t)$ es grande e $y(t)$ es pequeño, el producto no se hace grande. Deducimos del punto anterior que el concepto de producto escalar es el de una magnitud de la relación entre las funciones $x(t)$ e $y(t)$. Veremos que $\langle x, y \rangle$ es, en efecto, un producto escalar.:

(i) Simetría: $\langle x, y \rangle = x' y = \int x(t) y(t) dt = \int y(t) x(t) dt = \langle y, x \rangle$

(ii) Positividad: $\langle x, x \rangle = x' x = \int x(t) x(t) dt \geq 0$ y $\langle x, x \rangle = \int x(t) x(t) dt = 0 \Leftrightarrow x = 0$

(iii) Bilinealidad: Para todo $a, b \in \mathbb{R}$

$$(ax + by)' z = \int [a \cdot x(t) + b \cdot y(t)] z(t) dt = \int [a \cdot x(t) \cdot z(t) + b \cdot y(t) \cdot z(t)] dt =$$

$$\int a \cdot x(t) \cdot z(t) dt + \int b \cdot y(t) \cdot z(t) dt = a \langle x, z \rangle + b \langle y, z \rangle$$

DEFINICIÓN 2.11:

- Definimos el *producto escalar general* en el espacio L^2 como:

$$\langle x, y \rangle_w = x' y = \int w(t) x(t) y(t) dt \text{ donde } w(t) \in L^2$$

El producto escalar general lo que hace es ponderar por una función $w(t)$ para dar más peso a unos intervalos de t y restarselo a otros.

• PROPIEDADES:

(i) Simetría:

$$\langle x, y \rangle_w = \int w(t)x(t)y(t)dt = \int w(t)y(t)x(t)dt = \langle y, x \rangle_w$$

(ii) Positividad:

$$\langle x, x \rangle_w = \int w(t)x(t)x(t)dt \geq 0 \text{ y } \langle x, x \rangle_w = \int w(t)x(t)x(t)dt = 0 \Leftrightarrow x = 0$$

(iii) Bilinealidad: Para todo $a, b \in \mathfrak{R}$

$$\begin{aligned} \langle ax + by, z \rangle_w &= \int w(t)[a \cdot x(t) + b \cdot y(t)]z(t)dt = \int [a \cdot w(t) \cdot x(t) \cdot z(t) + b \cdot w(t) \cdot y(t) \cdot z(t)]dt = \\ &a \int w(t) \cdot x(t) \cdot z(t)dt + b \int w(t) \cdot y(t) \cdot z(t)dt = a \langle x, z \rangle_w + b \langle y, z \rangle_w \end{aligned}$$

DEFINICIÓN 2.12:

La L^2 -norma de $x(t)$ se define como:

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\|x\|^2} = \sqrt{\int x(t)^2 dt}$$

La norma nos da la magnitud del elemento $x(t)$ dentro del espacio. Las propiedades de la norma se cumplen de la misma manera en este espacio. Existe también en el caso de las funciones el concepto de ángulo derivada de la desigualdad de Cauchy-Schwartz:

DEFINICIÓN 2.13:

El ángulo θ entre las funciones $x(t)$ e $y(t)$ los definimos como:

$$\theta = \arccos \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} = \frac{\int x(t)y(t)dt}{\sqrt{\int x(t)^2 dt \cdot \int y(t)^2 dt}}$$

Explica la relación de dependencia entre $x(t)$ e $y(t)$.

DEFINICIÓN 2.14:

Diremos que una función $x(t)$ es ortogonal a $y(t)$ si:

$$\langle x, y \rangle = \int x(t) \cdot y(t)dt = 0$$

2.1.4. -Estadísticos descriptivos en análisis funcional de datos.

2.1.4.1.- Estadísticos sobre una función.

En el análisis funcional de datos hay que hacer distinciones entre los diferentes estadísticos, en función de si se definen para estudiar la tendencia central y la dispersión dentro de un solo

elemento (estadísticos sobre una función), o bien si se definen para estudiar una muestra de funciones (estadísticos de una muestra de una función aleatoria) o bien si se definen para estudiar muestras de dos o más funciones aleatorias. La primera clase de estadísticos pretende resumir la información de toda una función con pocas medidas. La segunda clase de estadísticos intentan caracterizar la función aleatoria mientras que la tercera clase intenta captar la relación entre dos o más funciones aleatorias. En este punto exponemos los primeros.

Hallar la media, la varianza y la covarianza de una función $x(t)$ es simplemente un ejercicio de deducción si tenemos en cuenta las definiciones de estas cantidades en notación algebraica y la definición del producto escalar (2.10). Para simplificar los conceptos supondremos que el intervalo en el que mueve t será $[0, T]$. Por lo tanto:

DEFINICIÓN 2.14:

Definimos la *media de la función $x(t)$* o también *valor medio de la función $x(t)$* como:

$$\bar{x} = \frac{1}{T} \langle x, \mathbf{1} \rangle = \frac{1}{T} \cdot \int_0^T x(t) \cdot \mathbf{1}(t) dt$$

Donde la función $\mathbf{1}(t)=1 \forall t \in \mathbb{R}$.

La media de $x(t)$ (\bar{x}) va a representar la tendencia central y el nivel medio de todos los valores de $x(t)$. No hay que confundir este concepto ($\bar{x} \in \mathbb{R}$) con los conceptos de función media (punto 1.1.4.2) y función valor medio, que veremos a continuación.

DEFINICIÓN 2.15:

La *función valor medio de $x(t)$* , que denotaremos por $\bar{x} \cdot \mathbf{1}(t)$, es aquella cuyo valor $\forall t \in [0, T]$ es el de la media de $x(t)=\bar{x}$.

Esta definición va a tener sentido a la hora de expresar la varianza y la covarianza de una función.

PROPIEDAD:

Sea $x(t)$ una función continua definida en el intervalo $[a, b]$ y sea \bar{x} su valor medio. Definimos ahora la función valor medio $f_{\bar{x}}(t) = \bar{x} \forall t \in [a, b]$. Entonces se cumple que:

$$\int_a^b x(t) dt = \int_a^b f_{\bar{x}}(t) dt$$

Es decir, el área que encierra la función $x(t)$ es la misma que el área que encierra la función de valor constante igual al valor medio de la función en el intervalo $[a, b]$.

PRUEBA:

$$\begin{aligned} \int_a^b x(t) dt - \int_a^b f_{\bar{x}}(t) dt &= \left(\int_a^b f_{\bar{x}}(t) dt \right) \underset{\text{es cte.}}{=} \int_a^b x(t) dt - f_{\bar{x}}(t) \int_a^b dt = \left(\begin{array}{l} \text{definición} \\ \text{de } \bar{x} \end{array} \right) = \\ \int_a^b x(t) dt - \left(\frac{1}{\int_a^b dt} \int_a^b x(t) dt \right) \int_a^b dt &= \left(\int_a^b dt \right) \underset{\text{es cte.}}{=} \int_a^b x(t) dt - \frac{a}{\int_a^b dt} \int_a^b x(t) dt = 0, \end{aligned}$$

como queríamos demostrar.

DEFINICIÓN 2.16:

La varianza de $x(t)$, que denotamos con $S_{x(t)}^2$ se define como:

$$S_{x(t)}^2 = \frac{1}{T} \langle x - \bar{x} \cdot \mathbf{1}, x - \bar{x} \cdot \mathbf{1} \rangle = \frac{1}{T} \cdot \int_0^T (x(t) - \bar{x} \cdot \mathbf{1}(t))^2 dt$$

Esta cantidad representa la variación media (al cuadrado) de todos los valores de la función respecto a su valor medio. Gráficamente una función con mucha variabilidad será aquella que englobe un área grande entre dicha función y la función valor medio.

DEFINICIÓN 2.17:

La covarianza entre dos funciones $x(t)$ e $y(t)$, que denotamos con $S_{x(t),y(t)}$ se define como:

$$S_{x(t),y(t)} = \frac{1}{T} \langle x - \bar{x} \cdot \mathbf{1}, y - \bar{y} \cdot \mathbf{1} \rangle = \frac{1}{T} \int_0^T (x(t) - \bar{x} \cdot \mathbf{1}(t)) \cdot (y(t) - \bar{y} \cdot \mathbf{1}(t)) dt$$

DEFINICIÓN 2.18:

La correlación entre dos funciones $x(t)$ e $y(t)$, que denotamos con $r_{x(t),y(t)}$ se define como:

$$r_{x(t),y(t)} = \frac{S_{x(t),y(t)}}{S_{x(t)}S_{y(t)}} = \frac{\frac{1}{T} \langle x - \bar{x} \cdot \mathbf{1}, y - \bar{y} \cdot \mathbf{1} \rangle}{\frac{1}{T} \langle x - \bar{x} \cdot \mathbf{1}, x - \bar{x} \cdot \mathbf{1} \rangle \frac{1}{T} \langle y - \bar{y} \cdot \mathbf{1}, y - \bar{y} \cdot \mathbf{1} \rangle} =$$

$$= \frac{\frac{1}{T} \cdot \int_0^T (x(t) - \bar{x} \cdot 1(t)) \cdot (y(t) - \bar{y} \cdot 1(t)) dt}{\frac{1}{T} \int_0^T (x(t) - \bar{x} \cdot 1(t))^2 dt \cdot \frac{1}{T} \int_0^T (y(t) - \bar{y} \cdot 1(t))^2 dt}$$

En este caso la función covarianza mide la relación entre dos funciones $x(t)$ e $y(t)$ y la correlación la sitúa entre -1 y $+1$. Digamos que la covarianza es el producto de dos áreas: la comprendida entre una función $x(t)$ y su valor medio (Ax), y la comprendida entre una función $y(t)$ y su valor medio (Ay). Se pueden dar tres situaciones: $R \cong +1$; $R \cong -1$ y $R \cong 0$. Cuando $R \cong +1$ tendremos que en el mismo rango de $t \in \mathfrak{R}$, Ax tiene el mismo signo que Ay . Cuando $R \cong -1$ Ax y Ay tienen signo distinto. Por último cuando $R \cong 0$ tenemos que en unos intervalos de $t \in \mathfrak{R}$ $Ax \cong 0$ y por tanto $Ax \cdot Ay \cong 0$, y en otros intervalos $Ay \cong 0$ y por tanto $Ax \cdot Ay \cong 0$. También entre funciones periódicas si una tiene un periodo múltiplo de dos del otro también su correlación será cero.

El valor medio, la varianza y la correlación nos van a servir para identificar bien el comportamiento particular de cada una de nuestras funciones y sus relaciones dos a dos.

Ejemplo 2.6:

Sean las funciones,

$$f_1(t) = \sin(t \cdot \pi) \quad t \in [0, \beta]; \quad f_2(t) = \sin(t \cdot \pi + \pi) \quad t \in [0, \beta]; \quad f_3(t) = 2 + 2 \sin(t \cdot \pi) \quad t \in [0, \beta];$$

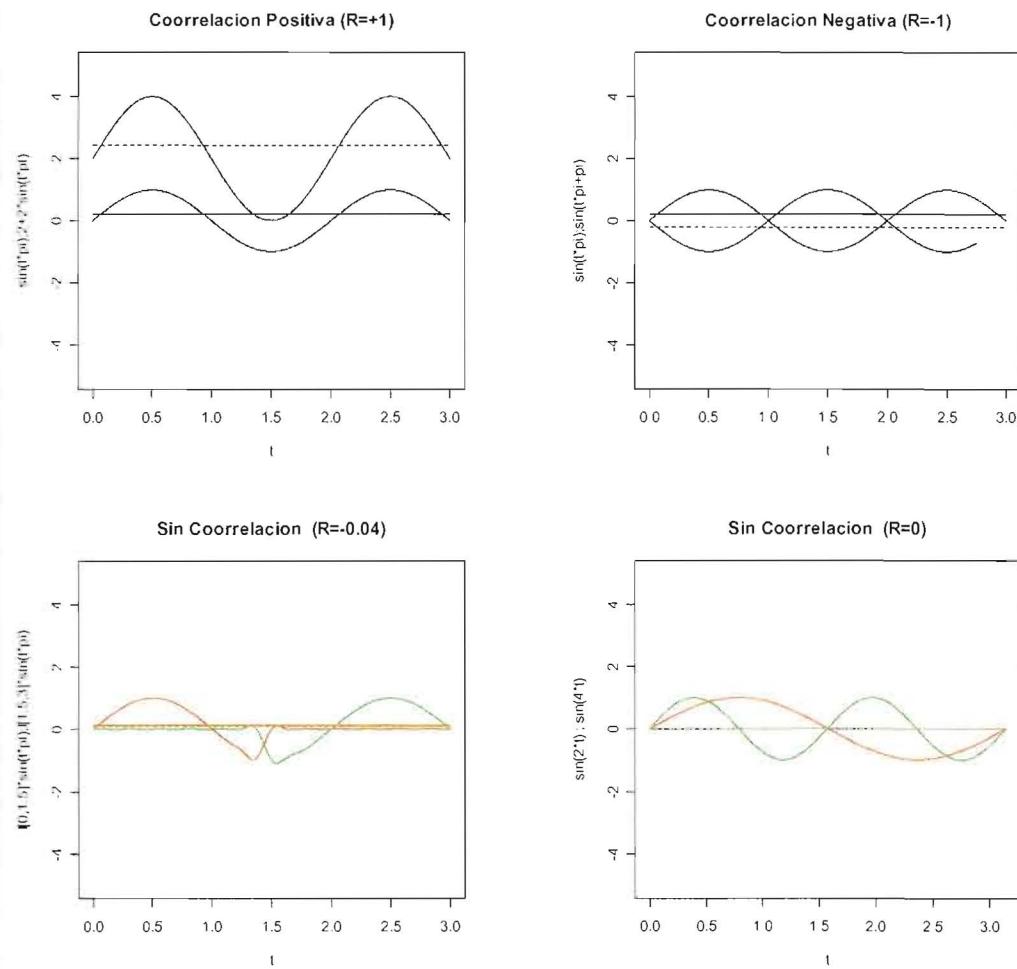
$$f_4(t) = I_{[t < 1, \beta]} \sin(t \cdot \pi) \quad t \in [0, \beta]; \quad f_5(t) = I_{[t > 1, \beta]} \sin(t \cdot \pi) \quad t \in [0, \beta], \quad f_6(t) = \sin(2t \cdot) \quad t \in [0, \pi] \text{ y}$$

$$f_7(t) = \sin(4t \cdot) \quad t \in [0, \pi]$$

Tenemos evaluadas estas funciones en puntos equidistantes de distancia 0.125 en el intervalo $[0,3]$, y hemos calculado para estos puntos su spline interpolador. Hemos calculado su matriz de correlaciones obteniendo lo siguiente:

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.0000000	1.0000000	-0.9989560	0.65189879	0.73199419
[2,]	1.0000000	1.0000000	-0.9989560	0.65189879	0.73199419
[3,]	-0.9989560	-0.9989560	1.0000000	-0.64823756	-0.73390805
[4,]	0.6518988	0.6518988	-0.6482376	1.00000000	-0.03945599
[5,]	0.7319942	0.7319942	-0.7339080	-0.03945599	1.00000000

Vemos que la correlación entre f_1 y f_2 es igual a uno, f_1 y f_3 es prácticamente igual a -1 y f_4 y f_5 es prácticamente cero. Veamos los gráficos:



Figuras 2.1-2.4: Representación gráfica de las funciones f_1, f_2, f_3, f_4, f_5 y f_6 con sus valores medios respectivos. En ellos se puede observar cómo la correlación se relaciona con el comportamiento dos a dos de las funciones.

Obsérvese que en el caso $R \geq +1$ los tramos en que Af_1 es positivo, Af_2 también es positivo y viceversa. En el caso en que $R \leq -1$, los tramos de Af_1 positivos son los tramos de Af_2 negativos y viceversa. En el caso de la izquierda con $R \geq 0$, cuando f_4 tiene área, $Af_5 \geq 0$ y viceversa. En el caso de la derecha con $R=0$ la función f_7 cumple 2 ciclos cuando f_6 sólo cumple uno.

2.1.4.2.- Estadísticos de muestras de una función aleatoria.

Sabemos que, para muestras de valores en \mathbb{R} la media, la varianza y la covarianza son valores también en \mathbb{R} . Cuando trabajamos con muestras cuyos elementos son vectores en \mathbb{R}^n , la naturaleza de los estadísticos varía: la media también es un vector pero luego tenemos expresada de forma compacta tanto la varianza como la covarianza, mediante lo que denominamos matriz de varianzas covarianzas.

En el análisis funcional sucede algo parecido a lo que sucede con el caso de los vectores. Podemos pensar que en el análisis funcional trabajamos con vectores con infinitas componentes, una para cada $t \in \mathbb{R}$. Tendremos así una expresión de la función media y varianza como funciones a lo largo de t (vector con infinitas componentes) y las funciones covarianza y correlación entre diferentes funciones serán funciones que van de \mathbb{R}^2 a \mathbb{R} .

DEFINICIÓN 2.19:

Sea $x_1(t), x_2(t), \dots, x_N(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0, T]$.

Definimos la *función media muestral de $x(t)$* como:

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

Si fijamos el valor de t en un punto concreto y evaluamos ahí todas las funciones entonces obtenemos una muestra de tamaño N de la cual podemos extraer su media. La *función media muestral de $x(t)$* nos va a dar la media de esos valores en ese valor de t de forma explícita en una sola función.

DEFINICIÓN 2.20:

Sea $x_1(t), x_2(t), \dots, x_N(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0, T]$.

Definimos la *función varianza muestral de $x(t)$* como:

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

Y definimos la *función desviación estándar muestral de x(t)* como:

$$Stdev_{x(t)}(t) = \sqrt{Var_{x(t)}(t)}$$

De la misma manera que la función media muestral de $x(t)$ nos indica la tendencia central de las funciones en un t dado, las funciones varianza muestral y desviación estándar muestral de $x(t)$ nos cuantifican el valor medio al cuadrado y el valor medio respectivamente de las desviaciones respecto la media en t .

DEFINICIÓN 2.21:

Sea $x_1(t), x_2(t), \dots, x_N(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0, T]$.

La *función covarianza muestral de x(t) entre t_1 y t_2* será:

$$Cov_{x(t)}(t_1, t_2) = \frac{1}{N-1} \sum_{i=1}^N (x_i(t_1) - \bar{x}(t_1)) \cdot (x_i(t_2) - \bar{x}(t_2))$$

DEFINICIÓN 2.22:

Sea $x_1(t), x_2(t), \dots, x_N(t)$ una muestra de funciones de una función aleatoria $x(t)$, definida en $[0, T]$.

La *función correlación muestral de x(t) entre t_1 y t_2* será:

$$Corr_{x(t)}(t_1, t_2) = \frac{Cov_{x(t)}(t_1, t_2)}{\sqrt{Var_{x(t)}(t_1) \cdot Var_{x(t)}(t_2)}}$$

Estas dos medidas nos van a indicar la magnitud de la relación entre el comportamiento de la función $x(t)$ en el valor t_1 y el comportamiento de $x(t)$ en el valor t_2 . En realidad si fijamos t_1 y t_2 podríamos pensar que tenemos dos muestras del mismo tamaño: $x_1(t_1), x_2(t_1), \dots, x_N(t_1)$ y $x_1(t_2), x_2(t_2), \dots, x_N(t_2)$ y podríamos calcular la covarianza o el coeficiente de correlación entre las dos muestras. Con esto resumimos la dependencia de los registros a través de los distintos t .

2.1.4.3.- Estadísticos de muestras de dos o más funciones aleatorias.

Supongamos que ahora tenemos muestras de tamaño N de dos funciones aleatorias $x(t)$ e $y(t)$ y queremos saber qué relación tiene una con otra. Queremos saber la magnitud de la dependencia entre una y otra función para t_1 y t_2 fijados.

DEFINICIÓN 2.23:

Sean dos funciones aleatorias $x(t)$ e $y(t)$ y sean:

- $x_1(t), x_2(t), \dots, x_N(t)$ muestra de $x(t)$
- $y_1(t), y_2(t), \dots, y_N(t)$ muestra de $y(t)$

entonces la función de covarianza cruzada de $x(t)$ e $y(t)$ en t_1, t_2 se define como:

$$Cov_{x(t),y(t)}(t_1, t_2) = \frac{1}{N-1} \sum_{i=1}^N [x_i(t_1) - \bar{x}(t)] \cdot [y_i(t_2) - \bar{y}(t_2)]$$

DEFINICIÓN 2.24:

Sea la situación de la definición 2.23. Entonces la función de correlación cruzada de $x(t)$ e $y(t)$

en t_1, t_2 se define como:

$$Corr_{x(t),y(t)}(t_1, t_2) = \frac{Cov_{x(t),y(t)}(t_1, t_2)}{\sqrt{Var_{x(t)}(t_1) \cdot Var_{y(t)}(t_2)}}$$

Observemos que si fijamos t_1 y t_2 lo único que estamos haciendo es calcular la el coeficiente de correlación entre la muestra $x_1(t_1), x_2(t_1), \dots, x_N(t_1)$ y la muestra $y_1(t_2), y_2(t_2), \dots, y_N(t_2)$.

2.2. – ANÁLISIS DE COMPONENTES PRINCIPALES EN ESPACIOS CON PRODUCTO ESCALAR.

El análisis de componentes principales en espacios con producto escalar (ACPPE) definido considera que todos los elementos del espacio en el que trabajamos (sea \mathbb{R}^n o sea L^2) se pueden expresar como combinación lineal de unos pocos elementos. Es decir,

$$\mathbf{x}_i = \sum_{k=1}^K f_{i,k} \cdot \xi_k \text{ donde } K \text{ es menor que la dimensión del espacio.}$$

El objetivo del ACPPE es el de condensar la máxima información dentro de la combinación lineal con el menor número de elementos posible. En el ACPPE, por tanto, buscamos en primer lugar una nueva base de elementos $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ que contenga la máxima variabilidad posible en su primera componente. De la variabilidad restante, es decir, del espacio ortogonal al generado por el elemento ξ_1 , buscamos que la segunda componente contenga la máxima posible, y así sucesivamente hasta formar una base con los elementos ortogonales generadores

de todo el espacio. La transformación con la que conseguimos una base con las propiedades anteriormente descritas es la resultante de obtener los elementos propios. El concepto de elemento propio no es más que la generalización del concepto de vector propio pero definido para cualquier espacio a través de productos escalares

Resuelto su cálculo a este nivel general posteriormente se trataría simplemente de adaptar la notación a la naturaleza del espacio con el que trabajaremos. En primer lugar introduciremos unas definiciones, luego veremos cómo se resuelve el problema de elementos propios y en el punto siguiente nos plantearemos el problema de buscar funciones propias como caso particular del problema que planteamos ahora.

2.2.1.- Proyecciones expresadas como productos escalares.

Sean v_1, v_2, \dots, v_n elementos del espacio E . Entonces podemos considerar el conjunto de elementos posibles de E que se pueden obtener mediante combinaciones lineales de estos elementos generadores, que llamaremos V .

$x \in V$ si $x = v'\alpha$ donde $v = (v_1, v_2, \dots, v_n)$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$. ($\forall i: \alpha_i \in \mathbb{R}$)

Este conjunto de elementos V será un subespacio vectorial, pues cumple las siguientes propiedades:

- (i) El subespacio ha de contener el elemento nulo 0.
- (ii) Sea $x, y \in V$ entonces $x+y \in V$
- (iii) Sea $\lambda \in \mathbb{R}$ y $x \in V$, entonces $\lambda x \in V$.

DEFINICIÓN 2.25:

Sea P un operador y v_1, v_2, \dots, v_n elementos del espacio vectorial E que definen un subespacio vectorial V . Entonces P es un operador de proyección ortogonal hacia V si cumple que:

(1) $\forall z \in E, Pz \in V$. Es decir, para todo elemento z del espacio E , su proyección Pz es uno de los elementos del subespacio V :

$$Pz = v'c, \text{ para algún } c \in \mathbb{R}^n$$

(2) El proyector P aplicado a un elemento del subespacio vectorial da como resultado el mismo elemento y :

$$\forall y \in V, Py = y$$

(3) $\forall z$, el residuo $z - Pz$ es ortogonal a todo elemento $v \in V$:

$$\langle v' \alpha, \tilde{z} - P_{\tilde{z}} \rangle = 0, \forall \alpha \in \mathfrak{R}''$$

En resumen, con $P_{\tilde{z}}$ obtenemos el elemento de V que está más cerca de \tilde{z} .

PROPOSICIÓN:

Sea

\tilde{z} : Elemento del espacio vectorial E

$\underline{v} = (v_1, v_2, \dots, v_n)$ vector que contiene los elementos generadores del subespacio vectorial de V .

Entonces la proyección ortogonal de \tilde{z} sobre el subespacio V se halla de la siguiente manera:

$$P_{\tilde{z}} = v' K^+ \langle v, \tilde{z} \rangle,$$

donde:

$K = v v' = (k_{ij})_{ij} = (\langle v_i, v_j \rangle)_{ij}$ matriz nxn de productos escalares de elementos que generan V ,

$\langle v, \tilde{z} \rangle$ es el producto escalar entre elementos del espacio E .

K^+ = Inversa generalizada de K ($K K^+ K = K$)

DEMOSTRACIÓN:

- En primer lugar demostraremos que si y es un elemento de V , entonces $Py = y$.

Si y es un elemento de V entonces $y = \underline{v}' \underline{c}$, donde \underline{c} es un vector con componentes en \mathfrak{R} .

Por tanto:

$$Py = \underline{v}' K^+ \langle v, \underline{v} c \rangle = \underline{v}' K^+ K c$$

$$y - Py = \underline{v}' c - \underline{v}' K^+ K c = \underline{v}' (I - K^+ K) c = \underline{v}' d, \text{ donde } I \text{ es la matriz identidad.}$$

Si $\|y - Py\|^2 = 0$ entonces querrá decir que $y - Py$ es el elemento cero.

$$\|y - Py\|^2 = d' \underline{v} \underline{v}' d = d' K d = d' (K - K K^+ K) c = 0$$

- A continuación demostraremos que el residuo $z - Pz$ es ortogonal a V , o lo que es lo mismo, a todo elemento $w \in V$. Esto es equivalente a probar que $\langle P_{\tilde{z}} - w, \tilde{z} - P_{\tilde{z}} \rangle = 0$.

$$\langle P_{\tilde{z}} - w, \tilde{z} - P_{\tilde{z}} \rangle = \begin{pmatrix} \text{propiedad del} \\ \text{producto escalar} \end{pmatrix} = \langle P_{\tilde{z}} - w, \tilde{z} \rangle - \langle P_{\tilde{z}} - w, P_{\tilde{z}} \rangle.$$

$$\langle P_{\tilde{z}} - w, P_{\tilde{z}} \rangle = (Pw = w) = \langle P_{\tilde{z}} - Pw, P_{\tilde{z}} \rangle = \langle P(\tilde{z} - w), P_{\tilde{z}} \rangle = \begin{pmatrix} P_{\tilde{z}} = v' K^+ \langle v, \tilde{z} \rangle \\ P(\tilde{z} - w) = v' K^+ \langle v, \tilde{z} - w \rangle \end{pmatrix} =$$

$$\langle \tilde{z} - w, v' \rangle K^+ v v' K^+ \langle v, \tilde{z} \rangle = \begin{pmatrix} v v' = K \\ K^+ K K^+ = K^+ \end{pmatrix} = \langle \tilde{z} - w, v' \rangle K^+ \langle v, \tilde{z} \rangle = \langle P_{\tilde{z}} - w, \tilde{z} \rangle$$

y por tanto $\langle P_{\tilde{z}} - w, \tilde{z} \rangle - \langle P_{\tilde{z}} - w, P_{\tilde{z}} \rangle = 0$,

como queríamos demostrar.

2.2.2.- Elementos propios.

DEFINICIÓN 2.26:

Sea Π un operador lineal definido en un espacio E con producto escalar (un endomorfismo),

$$\Pi: E \rightarrow E$$

$$e \rightarrow g = \Pi e,$$

que cumple:

Π es un operador definido positivo $\langle e, \Pi e \rangle \geq 0$ y,

Π es un operador simétrico $\langle e, \Pi f \rangle = \langle \Pi e, f \rangle = \langle f, \Pi e \rangle$.

Diremos que u es un *elemento propio de valor propio* λ respecto de Π si se cumple que:

$$\Pi u = \lambda u.$$

Ejemplo 2.7: Espacio vectorial \mathbb{R}^n .

Definimos el endomorfismo A , que es una matriz $n \times n$

$$A: \mathbb{R}^n \rightarrow \mathbb{R}^n$$

$$x \rightarrow y = Ax,$$

entonces un vector propio v con valor propio λ respecto a A es un vector que cumple:

$$Av = \lambda v$$

Un vector propio v respecto de A es aquel vector cuya dirección no cambia cuando se somete a una aplicación A , aunque sí cambie su norma en λ unidades. El conjunto de todos los vectores propios respecto a A forman una base de \mathbb{R}^n que posee las propiedades que nos interesan para el análisis de componentes principales.

PROPOSICIÓN (sin demostración):

Sea Π un operador lineal endomórfico y simétrico y sea x un elemento del espacio E . Sea el problema de maximización con restricción siguiente:

$$\text{Max } \langle x, \Pi x \rangle \text{ sujeto a } \|x\| = 1,$$

Entonces, la solución del problema es el elemento propio u de Π con mayor valor propio λ .

2.2.3.- El problema del Análisis de Componentes Principales mediante producto escalar.

Sea la representación

$$\hat{x}_i = \sum_{k=1}^K f_{i,k} \cdot \xi_k .$$

Queremos aproximar el valor del elemento x_i a través de la combinación lineal anterior. Es decir, que la diferencia entre el elemento y su aproximación sea mínima. Supongamos que $k=1$. Veremos entonces que se trata de resolver un problema de maximización de la covarianza sujeto a unas restricciones:

$$\begin{aligned} & \min_{\xi \in E} \left\{ \min_{f_E \in \mathcal{R}} \sum_{i=1}^N \langle (x_i - f_i \xi), (x_i - f_i \xi) \rangle \right\} = \min_{\xi \in E} \left\{ \min_{f_E \in \mathcal{R}} \sum_{i=1}^N \langle x_i, x_i \rangle + f_i^2 \langle \xi, \xi \rangle - 2 \cdot f_i \langle x_i, \xi \rangle \right\} = \\ & = (\text{derivando e igualando a cero: } \langle \xi, \xi \rangle = 1 \text{ y } 2f_i = 2 \cdot \langle x_i, \xi \rangle \Rightarrow f_i = \langle x_i, \xi \rangle) = \\ & = \min_{\xi \in E} \left\{ \sum_{i=1}^N \langle x_i, x_i \rangle + \langle x_i, \xi \rangle^2 - 2 \cdot \langle x_i, \xi \rangle^2 \right\} = \min_{\xi \in E} \left\{ \sum_{i=1}^N \langle x_i, x_i \rangle - \langle x_i, \xi \rangle^2 \right\} \Leftrightarrow \\ & \Leftrightarrow \max_{\xi \in E} \sum_{i=1}^N \langle x_i, \xi \rangle^2 \end{aligned}$$

Así pues, el problema del ACPPE se reduce a un problema de maximización de la función covarianza, sujeto a restricciones de norma igual a uno y de ortogonalidad cuya solución es el cálculo de elementos propios.

PROPOSICIÓN (sin demostración)

Sea Π un operador lineal endomórfico y simétrico y sea x un elemento del espacio E . Sean los problemas sucesivos siguientes:

(1) Sea U_1 el espacio generado por el elemento u_1 , solución del problema:

Max $\langle x, \Pi x \rangle$ sujeto a $\|x\|=1$

(2) Sea U_2 el espacio generado por el elemento u_2 , solución del problema:

Max $\langle x, \Pi x \rangle$ sujeto a $\|x\|=1$ y $\langle x, u_1 \rangle = 0$

...

(i) Sea U_i el espacio generado por el elemento u_i solución del problema:

$$\text{Max } \langle x, \prod_{j=1}^i u_j \rangle \text{ sujeto a } \|x\| = 1 \text{ y } \langle x, u_j \rangle = 0 \text{ para } j < i.$$

Entonces la solución $u_1, \dots, u_p, \dots, u_n$ de los problemas sucesivos anteriores corresponde a los elementos propios del operador lineal.

COROLARIO:

La solución de los problemas sucesivos anteriores utilizando como operador lineal el operador covarianza corresponden a los elementos propios de dicho operador, y por tanto solucionan el problema ACPPE.

2.3.- ANÁLISIS DE COMPONENTES PRINCIPALES FUNCIONAL (ACPF).

Una vez tenemos resuelto el problema del ACPPE, hemos de adaptar la notación al espacio vectorial L^2 . Para ello definiremos en primer lugar algunos conceptos y posteriormente resolveremos el problema del análisis de componentes principales funcionales.

El ACPF pretende explicar el conjunto de funciones de la muestra a partir de unas pocas funciones:

$$\hat{x}_i(t) = \sum_{k=1}^K f_{i,k} \cdot \xi_k(t)$$

Pero, ¿cuáles son estas funciones?. Tal como se deduce del apartado anterior se puede caracterizar cualquier función como combinación lineal de las funciones propias. El problema de calcular valores y vectores propios, o bien valores y funciones propias, no es más que la aplicación del problema general de buscar elementos propios en un espacio donde se trabaja con vectores en \mathbb{R}^n o con funciones L^2 respectivamente.

2.3.1.-Introducción y definiciones en el espacio funcional L^2 .

Recordemos el planteamiento en el análisis de componentes principales. Sea $X_{n \times k}$ una matriz que representa una muestra o una población de tamaño n representada en vectores de dimensión k . Queremos realizar un cambio de base que nos permita explicar en pocas componentes la máxima información (variabilidad) existente.

$A: \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$x \rightarrow y = Ax$$

Lo que busca el ACP clásico es buscar un cambio de base de manera “astuta”. Ese cambio de base pretende que la proyección de los puntos en ese nuevo eje contenga la máxima varianza posible. Posteriormente vuelve a buscar lo mismo para el espacio ortogonal al primer eje, y así sucesivamente.

El análisis de componentes principales funcional (ACPF) pretende lo mismo: buscar un cambio de base de manera que en la primera componente se explique la máxima variabilidad, en la segunda la máxima variabilidad del espacio ortogonal a la primera, etc. La única novedad es que ahora nuestra base actual contiene elementos que son funciones, en vez de vectores.

DEFINICIÓN 2.27:

Sea $\Pi : \mathbb{R}^2 \rightarrow \mathbb{R}$. Dada $f \in L^2$, la función Π permite definir otra función de L^2 , a la que llamaremos

$$\Pi(f):$$

$$\Pi(f) : \mathbb{R}^2 \rightarrow \mathbb{R}$$

$$s \mapsto \Pi(f)(s) = \int_{\mathbb{R}} \Pi(s, t) \cdot f(t) dt$$

Así, Π define un endomorfismo en L^2 , al que también llamaremos Π .

$$\Pi : L^2 \rightarrow L^2$$

$$f \mapsto \Pi(f)$$

PROPOSICIÓN:

El anterior endomorfismo es lineal.

DEMOSTRACIÓN:

Sean $\alpha, \beta \in \mathbb{R}$, f y $g \in L^2$. Sea $s \in \mathbb{R}$. Entonces,

$$\begin{aligned} \Pi(\alpha f + \beta g)(s) &= \int_{\mathbb{R}} \Pi(s, t) (\alpha f(t) + \beta g(t)) dt = \alpha \int_{\mathbb{R}} \Pi(s, t) f(t) dt + \beta \int_{\mathbb{R}} \Pi(s, t) g(t) dt = \\ &= \alpha \Pi(f)(s) + \beta \Pi(g)(s) \end{aligned}$$

como queríamos demostrar.

OBSERVACIONES:

- Podríamos considerar que Π es una matriz de infinitas filas e infinitas columnas (una para cada real).
- La función covarianza muestral permite definir es un operador lineal endomórfico en L^2 .

DEFINICIÓN 2.28:

Sea Π un operador lineal definido en el espacio L^2 (un endomorfismo),

$$\Pi: L^2 \rightarrow L^2$$

$$f \rightarrow g = \Pi f$$

Entonces diremos que Π es un operador definido positivo si

$$\langle f, \Pi f \rangle \geq 0 \quad \forall f \quad y \quad \langle f, \Pi f \rangle = 0 \Leftrightarrow f = 0.$$

DEFINICIÓN 2.29:

Sea Π un operador lineal definido positivo en el espacio L^2 (un endomorfismo), entonces Π es simétrico si:

$$\langle f, \Pi g \rangle = \langle \Pi f, g \rangle = \langle g, \Pi f \rangle$$

DEFINICIÓN 2.30:

Sea M el operador lineal endomórfico:

$$M: L^2 \rightarrow L^2$$

$$f \rightarrow g = Mf$$

Entonces diremos que f es una función propia sobre M si $\exists \lambda \in \mathbb{R}$ tal que:

$$Mf = \lambda f$$

2.3.2.-El problema del análisis de componentes principales funcionales(ACPF).

La mejor manera para interpretar qué es lo que hace el ACPF es la de explicarlo como una expansión del análisis de componentes principales (ACP) con datos multivariantes.

- En el caso del ACP tenemos que:

$\underline{X} = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n)$ es una muestra de vectores en \mathbb{R}^p centrados. Es decir, que el vector media de toda la matriz X resulta ser el vector nulo. Con el ACP buscamos expresar las funciones como combinación lineal de unas variables llamadas *componentes principales*. Es por eso que queremos buscar los coeficientes de cada x_i ,

$$f_{i,j} = \sum_{j=1}^p \xi_{j,i} x_{i,j} = \langle \xi_i, \underline{x}_i \rangle \quad i=1, \dots, N$$

tal que $\frac{1}{N} \sum_{i=1}^N f_{i,i}^2 = \frac{1}{N} \sum_{i=1}^N \langle \xi_i, x_i \rangle^2$ sea máximo sujeto a la restricción de que la norma

del vector ξ_i sea igual a uno. Es decir que buscamos una dirección ξ_i de la nube de puntos tal que la varianza de los puntos proyectados sobre ésta sea la máxima entre todas las direcciones posibles. Del espacio ortogonal restante podemos volver a realizar la misma operación y así sucesivamente hasta obtener la base $\xi_1, \xi_2, \dots, \xi_p$

- Sean $x_1(t), x_2(t), \dots, x_n(t)$ funciones centradas (su integral en \mathfrak{R} es cero) y definidas en L^2 . Estas funciones definen un espacio que tiene dimensión n . Con el ACPF buscamos para cada una de las funciones un valor resultado de aplicar el producto escalar general al espacio L^2 : haremos de adaptar a la naturaleza del espacio la definición de producto escalar. Es por eso que si antes x_i e ξ_i eran vectores y el producto escalar se expresaba como un sumatorio, ahora x_i e ξ_i serán funciones y el producto escalar se expresará como una integral:

$$f_{i,i} = \langle \xi_i, x_i \rangle = \int_{-\infty}^{+\infty} \xi_i(s) x_i(s) ds$$

Y buscaremos la $\xi_{i,1}$ tal que $\frac{1}{N} \sum_{i=1}^N f_{i,i}^2 = \frac{1}{N} \sum_{i=1}^N \langle \xi_i, x_i \rangle^2$ sea máximo sujeto a que

$$\|\xi_i\| = \int_{-\infty}^{+\infty} \xi_i(s) ds = 1.$$

Del espacio ortogonal restante podemos volver a realizar la misma operación y así sucesivamente hasta obtener la base $\xi_1, \xi_2, \dots, \xi_p$. Formalizando, lo que queremos resolver es esta sucesión de problemas de maximización continua con restricciones (problema ACPF):

- problema (1) $\begin{cases} \text{Max}_{\xi_1 \in L^2} \sum_{i=1}^N \langle \xi_1, x_i \rangle^2 \\ \text{sujeto a } \|\xi_1\| = 1 \end{cases}$
- problema (2) $\begin{cases} \text{Max}_{\xi_2 \in L^2} \sum_{i=1}^N \langle \xi_2, x_i \rangle^2 \\ \text{sujeto a } \|\xi_2\| = 1 \text{ y } \langle \xi_1, \xi_2 \rangle = 0 \end{cases}$
- problema (j) $\begin{cases} \text{Max}_{\xi_j \in L^2} \sum_{i=1}^N \langle \xi_j, x_i \rangle^2 \\ \text{sujeto a } \|\xi_j\| = 1 \text{ y } \langle \xi_k, \xi_j \rangle = 0 \text{ para } k < j \end{cases}$

PROPOSICIÓN:

Sea el problema ACPF y sean $x_1(t), x_2(t), \dots, x_N(t)$ funciones definidas en \mathfrak{R} .

Sea además

$$v(s, t) = \frac{1}{N} \sum_{i=1}^N x_i(s) x_i(t), \text{ donde}$$

$$\begin{aligned} V : L^2(\mathfrak{R}, \mathfrak{R}) &\rightarrow \mathfrak{R} \\ (s, t) &\mapsto v(s, t) \end{aligned}$$

y donde,

$$\begin{aligned} v(s, \cdot) : \mathfrak{R} &\rightarrow \mathfrak{R} \\ s &\mapsto \int_{\mathfrak{R}} v(s, t) y(t) dt, \text{ que es función de } s. \end{aligned}$$

Entonces la solución del problema ACPF es la solución de la ecuación propia siguiente:

$$\int_{\mathfrak{R}} v(s, t) \xi(t) dt = \langle v(s, \cdot), \xi \rangle = \lambda \cdot \xi(s)$$

que corresponde al vector de las funciones propias de $x_1(t), x_2(t), \dots, x_N(t)$ sobre el operador covarianza $v(s, t)$.

DEMOSTRACIÓN:

El problema que nos planteamos es el siguiente:

$$\begin{aligned} \max_{\xi \in L^2} \frac{1}{N} \sum_{i=1}^N \langle x_i, \xi \rangle^2 &= \max_{\xi \in L^2} \frac{1}{N} \sum_{i=1}^N \left(\int_{\mathfrak{R}} x_i(t) \xi(t) dt \right)^2 = \max_{\xi \in L^2} \frac{1}{N} \sum_{i=1}^N \int_{\mathfrak{R}} x_i(t) \xi(t) dt \int_{\mathfrak{R}} x_i(s) \xi(s) ds = \\ &= \max_{\xi \in L^2} \frac{1}{N} \sum_{i=1}^N \int_{\mathfrak{R}} \int_{\mathfrak{R}} \xi(t) x_i(t) x_i(s) \xi(s) ds dt = \max_{\xi \in L^2} \int_{\mathfrak{R}} \xi(t) \left(\int_{\mathfrak{R}} \left(\frac{1}{N} \sum_{i=1}^N x_i(t) x_i(s) \right) \xi(s) ds \right) dt = \\ &= \max_{\xi \in L^2} \int_{\mathfrak{R}} \xi(t) V \xi dt = \max_{\xi \in L^2} \langle \xi, V \xi \rangle \end{aligned}$$

Y basándonos en lo dicho en el punto 2.2.3, la solución es buscar las funciones propias de $v(s, t)$.

2.3.3.-Interpretación de las ACPFS.

Como resultado del problema de maximización anteriormente planteado obtenemos la ecuación

$$\hat{x}_i(t) = \sum_{k=1}^N f_{i,k} \cdot \xi_k(t),$$

donde los coeficientes son $f_{i,k}$ y las funciones componentes principales son $\xi_k(t)$. Con estas últimas podremos explicar resumidamente cada x_i .

Como hemos dicho antes queremos representar x_i a partir de las mínimas funciones posibles. Para saber cuál es el porcentaje de representación con un número de componentes en concreto utilizamos los valores propios.

- En ACP: La suma de los valores propios corresponde a la traza de la matriz de varianzas-covarianzas, que es la varianza total.

$$\text{traza}(V) = \sum_{j=1}^p V(x_j) = \sum_{j=1}^p \sigma_j^2 = \sum_{j=1}^p \frac{1}{N} \sum_i^N (x_{i,j} - \bar{x}_j)^2 = \frac{1}{N} \sum_i \sum_j (x_{i,j} - \bar{x}_j)^2 = \sum_{j=1}^p \lambda_j$$

- En ACPF: De la misma manera que en ACP, la suma de los valores propios corresponde a la varianza total.

$$\int_R v(t, t) dt = \frac{1}{N} \sum_{i=1}^N \left(\int_R x_i(t)^2 dt \right) = \sum_{j=1}^N \lambda_j$$

DEFINICIÓN 2.31:

Sean l el número de primeras componentes con las que queremos representar x_i , entonces el *porcentaje de representación* será el siguiente:

$$\% \text{ representación} = 100 \times \frac{\sum_{i=1}^l \lambda_i}{\sum_{i=1}^N \lambda_i}$$

Uno de los objetivos del ACPF es el de poder representar como puntos en el espacio de dos o tres dimensiones la totalidad de la muestra. Las coordenadas de los puntos serían los coeficientes $(f_{i,1}, f_{i,2}, f_{i,3})$ y los ejes representan las direcciones de crecimiento de las funciones propias $\xi_1(t)$, $\xi_2(t)$ y $\xi_3(t)$. Para hacer una interpretación completa que sitúe en el espacio cada una de las x_i 's hemos de interpretar a qué pauta de comportamiento corresponde cada una de las componentes principales $\xi_i(t)$. En este caso no podemos utilizar el mismo método que en ACP ya que no tenemos representación de las componentes en función de las p variables, que

en este caso sería $p=\infty$. Lo que hacemos en este caso para interpretar las componentes principales funcionales es representar la función media juntamente con la función media $\pm C^* \xi(t)$, donde C es una constante que facilita la interpretación del gráfico.

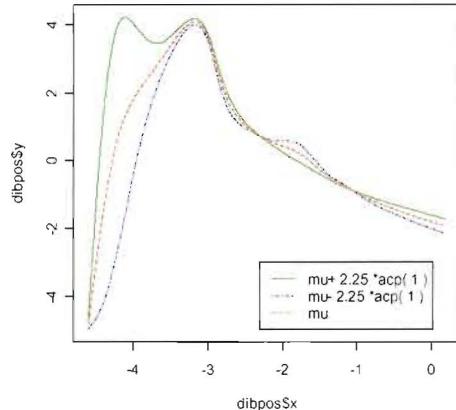


Figura 2.5: Representación de la función media $\pm C^* \xi(t)$ para datos de concentración en la sangre de Docetaxel de ambos esquemas con transformación logarítmica en la concentración y en el tiempo.

2.3.4.- ACPF para funciones representadas a través de funciones base conocidas.

En muchas ocasiones nos vamos a encontrar con que nuestras funciones están expresadas como una combinación lineal de una base de funciones. Para este caso en particular tenemos una adaptación del ACPF.:

Sea $\phi_1(t), \phi_2(t), \dots, \phi_k(t)$ una base de funciones, entonces tendremos representadas nuestras funciones de la siguiente manera:

$$x_i(t) = \sum_{i=1}^k c_{i,k} \phi_k(t)$$

En forma matricial estaríamos diciendo lo siguiente: $\underline{x} = C \underline{\phi}$ donde \underline{x} y $\underline{\phi}$ son vectores de funciones y C es una matriz $N \times K$ que contiene los coeficientes de la combinación lineal.

La función de covarianza teniendo en cuenta esta representación sería:

$$\nu(s, t) = \frac{1}{N} \phi(s)' C' C \phi(t)$$

Recordemos que la ecuación propia a resolver era:

$$\int v(s,t) \xi(t) dt = \langle v(s,\cdot), \xi \rangle = \lambda \cdot \xi(s)$$

Supondremos también que las funciones propias pueden ser expresadas en la base anterior. Es decir,

$$\xi_i(s) = \sum_{k=1}^K b_k \phi_k(s)$$

Considerando la expresión de la varianza actual:

$$\int_R v(s,t) \xi(s) dt = \int_R \frac{1}{N} \phi(s)' C' C \phi(t) \phi(t)' b dt = \phi(s)' N^{-1} C' C W b$$

Con lo que la ecuación propia que queda por resolver es:

$$\phi(s)' N^{-1} C' C W b = \lambda b$$

$$\text{Donde } W = \left(w_{k_1, k_2} \right)_{k_1, k_2} = \left(\int_R \phi_{k_1}(t) \phi_{k_2}(t) dt \right)_{k_1, k_2}$$

Como esto ha de ser para todo $s \in \mathcal{R}$ entonces tendremos que todo se reduce a calcular los vectores propios de una matriz A:

$$N' C' C W b = \lambda b \Leftrightarrow A b = \lambda b$$

3. – ESTIMACIÓN DE FUNCIONES POR SPLINES.

3.1.- REPRESENTACIÓN DE DATOS. SUAVIZADO E INTERPOLACIÓN.

Para el análisis de datos funcionales teórico es preciso conocer la forma explícita de cada una de las funciones. Sin embargo en la mayoría de las situaciones reales no conoceremos esta forma. Es más, como resultado de nuestro experimento o de la observación vamos a obtener simplemente los datos observados de la función a lo largo de su eje de ordenadas. Podemos decir que en general esta será nuestra situación:

SITUACIÓN:

- Sea $f_1(x), f_2(x), \dots, f_N(x)$ una muestra de funciones aleatorias.
- Para cada $f_i(x)$ habremos evaluado la función obteniendo $y_{i,1}, y_{i,2}, \dots, y_{i,n_i}$ como imagen de la misma evaluada en la serie de valores $x_{i,1}, x_{i,2}, \dots, x_{i,n_i}$.

Ante esta situación nos tenemos que plantear cómo transformar estos datos en una buena estimación de la función observada. Desgraciadamente a la hora de tomar estos valores también podemos cometer errores: error observacional. Por lo tanto habremos de considerar para la estimación de la función si el error observacional puede ser o no ser obviado.

Existen varios métodos de conseguir funciones a partir de esta situación: el polinomio interpolador de lagrange, splines interpoladores (Bonet et al.) o regresión no paramétrica (Lange K.), entre otros.

En todos ellos el planteamiento de la situación será el siguiente:

PROBLEMA:

Convertir los valores $y_{i,1}, y_{i,2}, \dots, y_{i,n_i}$ de alguna manera eficiente en una función $f_i(x)$ tal que

$$y = f_i(x_{i,j}) + \varepsilon_{i,j},$$

donde $\varepsilon_{i,j}$ es un error observacional aleatorio con $E[\varepsilon_{i,j}] = 0$

Para la situación en que sabemos que el error observacional es despreciable usaremos el spline interpolador, por las buenas propiedades que tiene. Mientras que en el caso en que tengamos que tener en cuenta el error observacional o de medida usaremos regresión no paramétrica por splines.

Los splines se pueden definir de forma genérica y luego en particular se puede definir una clase de splines. La definición genérica es la siguiente:

DEFINICIÓN 3.1:

Sea:

- $S(x)$ una función definida en $[a,b]$,
- $a=x_0 < x_1 < \dots < x_n = b$ conjunto de puntos ordenados.

Entonces se dice que $S(x)$ es un *spline de grado p y nodos* $x_0 < x_1 < \dots < x_n$ si se verifica que:

- a) $S(x)$ es un polinomio de grado menor o igual a p en cada intervalo $[x_i, x_{i+1}]$;
- b) La función $S(x)$ tiene derivadas hasta de orden $(p-1)$ continuas en $[a,b]$.

3.2.- TEORÍA DE SPLINES CÚBICOS INTERPOLADORES.

Cuando tenemos un gran número de evaluaciones de la función, herramientas como el polinomio interpolador suelen dar problemas, ya que acaban dando como solución polinomios de muy alto grado, que suelen ser muy poco “suaves” (concepto que más adelante explicaremos). Entre todas las posibilidades (Polinomios cúbicos de Lagrange, Interpolación de Hermite, etc..) nos dedicaremos al uso de splines, ya que son la función interpoladora que minimiza la integral de la segunda derivada al cuadrado entre todas las funciones de L^2 , tal como describen P.J Green y B.W. Silverman.

Es por la propiedad anterior por la que los splines interpoladores dan como resultado estimaciones de la función más “suave”, y no precisa de polinomios de alto grado. En particular son los splines de tercer grado los que consideraremos por sus buenas propiedades. Veamos la definición:

DEFINICIÓN 3.2:

Sea

- $x_0 < x_1 < \dots < x_n$ conjunto de puntos ordenados (llamados también knots).
- f_0, f_1, \dots, f_n valores de la función evaluada en los puntos anteriores: $f_i = f(x_i)$, $i=0, \dots, n$.

Entonces el *spline cúbico interpolador* $S(x)$ es una función definida en el intervalo $[x_0, x_n]$ con las siguientes propiedades:

- a) $S(x)$ es un polinomio cúbico en cada intervalo $[x_i, x_{i+1}]$;
- b) $S(x_i) = f_i$ en cada nodo x_i .
- c) La segunda derivada $S''(x)$ existe y es continua a lo largo del intervalo $[x_0, x_n]$.
- d) En los nodos extremos $S''(x_0) = S''(x_n) = 0$.

PROPOSICIÓN (p3.1):

Existe exactamente un único spline en $[x_0, x_n]$ que satisfaga las propiedades anteriores.

DEMOSTRACIÓN (intuitiva):

De la definición anterior podemos deducir que un spline cúbico es la solución de un sistema de ecuaciones. Si el sistema de ecuaciones es compatible determinado entonces habremos demostrado que es único.

Tenemos que $S(x)$ es:

$$s_0(x) = s_{0,1} + s_{0,2}x + s_{0,3}x^2 + s_{0,4}x^3 \text{ para el intervalo } [x_0, x_1]$$

$$s_1(x) = s_{1,1} + s_{1,2}x + s_{1,3}x^2 + s_{1,4}x^3 \text{ para el intervalo } [x_1, x_2]$$

...

$$s_{n-1}(x) = s_{n-1,1} + s_{n-1,2}x + s_{n-1,3}x^2 + s_{n-1,4}x^3 \text{ para el intervalo } [x_{n-1}, x_n]$$

Por lo que tenemos que calcular $4n$ coeficientes.

Sobre la base de la definición encontramos que tiene que haber una serie de restricciones.

- Restricciones de interpolación: El spline ha de pasar por los puntos (x_i, f_i)

$$s_0(x_0) = f_0, s_1(x_1) = f_1, \dots, s_{n-1}(x_n) = f_n, (n+1 \text{ restricciones})$$

- Restricciones de continuidad:

$$s_0(x_1) = s_1(x_1), s_1(x_2) = s_2(x_2), \dots, s_{n-2}(x_{n-1}) = s_{n-1}(x_n) \quad (n-1 \text{ restricciones})$$

- Restricciones de derivada continua:

$$s'_0(x_1) = s'_1(x_1), s'_1(x_2) = s'_2(x_2), \dots, s'_{n-2}(x_{n-1}) = s'_{n-1}(x_n) \quad (n-1 \text{ restricciones})$$

- Restricciones de segunda derivada continua:

$$s''_0(x_1) = s''_1(x_1), s''_1(x_2) = s''_2(x_2), \dots, s''_{n-2}(x_{n-1}) = s''_{n-1}(x_n) \quad (n-1 \text{ restricciones})$$

- Restricciones de punto extremo (el spline se convierte en rectas fuera de $[x_0, x_n]$):

$$s''_0(x_0) = 0 \text{ y } s''_{n-1}(x_n) = 0 \quad (2 \text{ restricciones})$$

Por lo tanto tenemos un sistema de ecuaciones con $4n$ incógnitas y $4n$ ecuaciones, lo que lo convierte en un sistema compatible determinado. Por tanto tiene una única solución, como queríamos demostrar.

Otra de las ventajas de los splines es la de que pertenece al espacio L^2 . Además entre todas las funciones de ese espacio el spline es el que minimiza la siguiente cantidad:

$$\int (f''(x))^2 dx,$$

que es una medida global de la curvatura de la función. Las funciones que oscilan mucho entre los puntos a interpolar tienen pendientes muy cambiantes y segundas derivadas muy altas. Esto se traduce en lo que gráficamente llamamos ser una función “poco suave” (con altibajos bruscos).

Sea $s(x)$ el spline interpolador de una función $f(x)$ en los nodos $x_0 < x_1 < \dots < x_n$. Si $g(x)$ es cualquier otra función dos veces diferenciable continua e interpoladora de $f(x)$ en ese nodo, entonces:

$$\int_{x_0}^{x_n} g''(x)^2 dx \geq \int_{x_0}^{x_n} s''(x)^2 dx$$

con igualdad si y sólo si $g(x) = f(x)$

DEMOSTRACIÓN:

- Cuando $\int_{x_0}^{x_n} g''(x)^2 dx = \infty$ entonces no hay nada que probar.

- Observar que:

$$\begin{aligned} \int_{x_0}^{x_n} g''(x)^2 dx &= \int_{x_0}^{x_n} [(g''(x) - s''(x)) + s''(x)]^2 dx = \\ &= \int_{x_0}^{x_n} (g''(x) - s''(x))^2 dx + \int_{x_0}^{x_n} s''(x)^2 dx + 2 \int_{x_0}^{x_n} s''(x)(g''(x) - s''(x)) dx \end{aligned}$$

- ¿Cuánto vale $\int_{x_0}^{x_n} s''(x)(g''(x) - s''(x)) dx$?

$$\int_{x_0}^{x_n} [g''(x) - s''(x)] s''(x) dx = \left(\begin{array}{c} \text{int egrando} \\ \text{por intervalos} \end{array} \right) = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} [g''(x) - s''(x)] s''(x) dx =$$

$$\left(\begin{array}{c} \text{int egrando} \\ \text{por partes} \end{array} \right) = \sum_{i=0}^{n-1} \left\{ [g'(x) - s'(x)] s''(x) \Big|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} [g'(x) - s'(x)] s'''(x) dx \right\} = \sum (d - e)$$

- $e = (s'''(x) = \text{constante}) = \alpha_i [g(x) - s(x)] \Big|_{x_i}^{x_{i+1}} = 0$ por ser interpoladores en x_i y en x_{i+1} .

$$\begin{aligned} \sum d &= \left(\begin{array}{c} \text{evaluando} \\ \text{en } x_i \text{ e } x_{i+1} \end{array} \right) = \sum_{i=1}^{n-1} \{ s''(x_{i+1}) [g'(x_{i+1}) - s'(x_{i+1})] - s''(x_i) [g'(x_i) - s'(x_i)] \} = \\ &= \{ s''(x_1) [g'(x_1) - s'(x_1)] - s''(x_0) [g'(x_0) - s'(x_0)] \} + \{ s''(x_2) [g'(x_2) - s'(x_2)] - s''(x_1) [g'(x_1) - s'(x_1)] \} + \\ &+ \{ s''(x_3) [g'(x_3) - s'(x_3)] - s''(x_2) [g'(x_2) - s'(x_2)] \} + \dots + \{ s''(x_n) [g'(x_n) - s'(x_n)] - s''(x_{n-1}) [g'(x_{n-1}) - s'(x_{n-1})] \} = \\ &= \{ s''(x_n) [g'(x_n) - s'(x_n)] - s''(x_0) [g'(x_0) - s'(x_0)] \} = \left(\begin{array}{c} \text{Por definición} \\ s''(x_0) = s''(x_n) = 0 \end{array} \right) = 0 \end{aligned}$$

Por lo tanto:

$$\int_{x_0}^{x_n} g''(x)^2 dx = \int_{x_0}^{x_n} [g''(x) - s''(x)]^2 dx + \int_{x_0}^{x_n} s''(x)^2 dx \geq \int_{x_0}^{x_n} s''(x)^2 dx, \text{ como queríamos demostrar.}$$

Esto quiere decir que la inecuación de la proposición p3.2 será una igualdad cuando $g''(x) = s''(x)$.

Cuando esto no se produzca, como integramos algo que es siempre estrictamente no negativo la desigualdad será en el sentido en que marca la proposición.

Como conclusión a este apartado podemos decir que el spline interpolador es la función ideal para utilizar tanto por pertenecer al espacio L^2 como por poseer de entre todas las funciones que pertenecen a L^2 , la propiedad de ser la más “suave”.

3.3.- CÁLCULO COMPUTACIONAL DE SPLINES CÚBICOS INTERPOLADORES: MÉTODO DE LOS MOMENTOS.

Una manera intuitiva de construir un spline cúbico interpolador consiste simplemente en montar el sistema de ecuaciones de la proposición p3.1 y resolverlo. El problema es que la matriz asociada a este sistema de ecuaciones es demasiado grande, tiene muchos ceros y no es simétrica. Existe una manera alternativa de simplificar enormemente los cálculos: el método de los momentos.

NOTACIÓN:

Sean $b_i = x_{i+1} - x_i$, sea $M_i = s''(x_i)$ también llamados *momentos* de $s(x)$, y sea $B_i = s'(x_i)$, para $i=0, \dots, n$.

MÉTODO DE LOS MOMENTOS: (Bonet et al)

Se trata de expresar el sistema de ecuaciones que resuelve el problema de cálculo del spline interpolador, en función de los valores de la segunda derivada (M_i) y de los valores ya dados (x_i, f_i) , aprovechando que la segunda derivada es una recta. En el tramo $[x_i, x_{i+1}]$ tendremos la siguiente recta:

$$s_i''(x) = M_i + \frac{M_{i+1} - M_i}{b_i} (x - x_i) \text{ para } i=0, \dots, n-1$$

- Podemos obtener tanto $s'_i(x)$ como $s_i(x)$ integrando dos veces respecto x :

$$\begin{aligned} s'_i(x) &= \int_{x_i}^x s''_i(t) dt = \int_{x_i}^x M_i dt + \int_{x_i}^x \frac{M_{i+1} - M_i}{b_i} (t - x_i) dt = B_i + M_i (x - x_i) + \frac{M_{i+1} - M_i}{b_i} \frac{(x - x_i)^2}{2} \\ s_i(x) &= \int_{x_i}^x s'_i(t) dt = \int_{x_i}^x B_i + M_i (t - x_i) + \frac{M_{i+1} - M_i}{b_i} \frac{(t - x_i)^2}{2} dt = \\ &= A_i + B_i (x - x_i) + M_i \frac{(x - x_i)^2}{2} + \frac{M_{i+1} - M_i}{b_i} \frac{(x - x_i)^3}{6} \end{aligned}$$

- Podemos expresar ambas ecuaciones en función únicamente de M_i , eliminando A_i y B_i . Para ello aprovecharemos que $s_i(x_i)$ es interpolador y que tanto la primera como la segunda derivadas son continuas: $s'_i(x_i) = s'_{i+1}(x_i)$ y $s''_i(x_i) = s''_{i+1}(x_i)$.

$$(1) \quad s_i(x_i) = f_i \quad y \quad s_i(x_i) = A_i \Rightarrow A_i = f_i$$

$$(2) \quad s_i(x_i) = f_i \quad y \quad s_i(x_{i+1}) = f_{i+1} = \begin{cases} \text{expresión} \\ \text{en } M_i \end{cases} = f_i + B_i b_i + M_i \frac{b_i^2}{2} + \frac{M_{i+1} - M_i}{b_i} \frac{b_i^3}{6} \quad \text{y por lo}$$

$$\text{tanto tenemos que } B_i = \frac{f_{i+1} - f_i}{b_i} - (M_{i+1} - M_i) \frac{b_i}{6} - M_i \frac{b_i}{2} \quad i=0, \dots, n-1$$

(3) Utilizamos $s_i(x_{i+1}) = s_{i+1}(x_{i+1})$ para montar el nuevo sistema de ecuaciones:

$$\begin{aligned} s_i(x_{i+1}) = s_{i+1}(x_{i+1}) &\Leftrightarrow B_i + M_i (x_{i+1} - x_i) + \frac{(M_{i+1} - M_i)(x_{i+1} - x_i)^2}{b_i} = B_{i+1} \Leftrightarrow \begin{cases} \text{sustituyendo } b_i \\ \text{y } B_i \end{cases} \\ &\Leftrightarrow M_i b_i + (M_{i+1} - M_i) \frac{b_i}{2} + \frac{f_{i+1} - f_i}{b_i} - (M_{i+1} - M_i) \frac{b_i}{6} - M_i \frac{b_i}{2} = \\ &\frac{(f_{i+2} - f_{i+1})}{b_{i+1}} - (M_{i+2} - M_{i+1}) \frac{b_{i+1}}{6} - M_{i+1} \frac{b_{i+1}}{2} \end{aligned}$$

Y aislando las incógnitas obtenemos el sistema de ecuaciones siguiente:

$$b_i \cdot M_i + 2(b_i - b_{i+1})M_{i+1} + b_{i+1} \cdot M_{i+2} = 6 \left(\frac{f_{i+2} - f_{i+1}}{b_{i+1}} - \frac{f_{i+1} - f_i}{b_i} \right) \quad (\text{notación } d_{i+1}) \quad i=0, \dots, n-2$$

Expresado matricialmente:

$$\begin{pmatrix} 2(b_0 + b_1) & b_1 & & & & \\ b_1 & 2(b_1 + b_2) & b_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & b_{n-3} & 2(b_{n-3} + b_{n-2}) & b_{n-2} & \\ & & & b_{n-2} & 2(b_{n-2} + b_{n-1}) & M_{n-1} \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \end{pmatrix} \Leftrightarrow Rm = d$$

El sistema de ecuaciones anterior tiene asociada una matriz de sistema de tipo tridiagonal. Estas matrices permiten llegar a la solución de manera eficiente desde el punto de vista computacional. En el apéndice se facilita una variante de un método de solución de ecuaciones tridiagonales.

También se puede expresar d matricialmente en función del vector $f = (f_0, f_1, \dots, f_n)$:

$$\begin{pmatrix} \frac{1}{b_0} & -\left(\frac{1}{b_0} + \frac{1}{b_1}\right) & & & & \\ & \frac{1}{b_1} & -\left(\frac{1}{b_1} + \frac{1}{b_2}\right) & & & \\ & & \ddots & \ddots & & \\ & & & b_{n-3} & -\left(\frac{1}{b_{n-3}} + \frac{1}{b_{n-2}}\right) & \\ & & & & b_{n-2} & 0 \\ & & & & & -\left(\frac{1}{b_{n-2}} + \frac{1}{b_{n-1}}\right) & b_{n-1} \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} = d \Leftrightarrow Qf = d$$

Por lo que podemos expresar el problema como solucionar la ecuación matricial:

$$R\mathbf{m} = \mathcal{Q}\mathbf{f}$$

Una vez solucionado el sistema tendremos que calcular, en primer lugar los valores de la primera derivada y luego los coeficientes de cada uno de los polinomios que forman el spline. La expresión de la primera derivada es la que ha aparecido anteriormente:

$$B_i = \frac{f_{i+1} - f_i}{h_i} - (M_{i+1} - M_i) \frac{h_i}{6} - M_i \frac{h_i}{2} \quad i=0, \dots, n-1$$

Y los coeficientes de cada uno de los polinomios del spline se hallan de la siguiente manera:

$$\begin{aligned} S_i &= A_i + B_i(x - x_i) + M_i \frac{(x - x_i)^2}{2} + \frac{M_{i+1} - M_i}{6 h_i} \frac{(x - x_i)^3}{6} = \begin{pmatrix} \text{aislando} \\ \text{por coeficientes} \end{pmatrix} = \\ A_i + B_i(x - x_i) + \frac{M_i}{2} (x^2 - 2x x_i + x_i^2) + \frac{M_{i+1} - M_i}{6 h_i} (x^3 - 3x^2 x_i + 3x x_i^2 - x_i^3) &= \\ = \left\{ \frac{M_i}{2} x_i^2 - x_i^3 \left(\frac{M_{i+1} - M_i}{6 h_i} \right) - B_i x_i + A_i \right\} + x \left\{ -M_i x_i + 3x_i^2 \frac{M_{i+1} - M_i}{6 h_i} + B_i \right\} + \\ + x^2 \left\{ \frac{M_i}{2} - 3x_i \frac{M_{i+1} - M_i}{6 h_i} \right\} + x^3 \left\{ \frac{M_{i+1} - M_i}{6 h_i} \right\} \end{aligned}$$

Donde forzamos que $A_i = f_i$ ya que consideramos que no hay error en nuestros datos.

3.4- REGRESIÓN NO PARAMÉTRICA CON SPLINES.

En muchas ocasiones los datos se obtienen a través de métodos en los cuales cabe el error de medida. Por ejemplo en el medidor de concentración de un medicamento en la sangre va a haber un error de medida o bien un error de precisión de la máquina. En este caso la interpolación no va a ser el objetivo principal pues sabemos que los valores obtenidos pueden no ajustarse a la realidad. En estos casos nos va a interesar que la función obtenida se acerque lo más posible a la función real, y también que la función obtenida sea suave, al igual que le pedíamos al spline interpolador. Un criterio propuesto por K. Lange, que se ajusta a nuestros propósitos es considerar que la función g que queremos encontrar tiene que minimizar la siguiente expresión:

$$J_\alpha(g) = \alpha \sum_{i=0}^n w_i [y_i - g(x_i)]^2 + (1 - \alpha) \int_{x_0}^{x_n} g''(x)^2 dx \quad \text{con } 0 \leq \alpha \leq 1$$

Donde α es el peso relativo que le daremos al criterio de ajuste de mínimos cuadrados respecto al peso de la penalización por falta de suavidad. Cabe notar que si $\alpha=1$ el polinomio interpolador cumple con que $J_\alpha(g) \geq J_\alpha(s)$. Buscaremos la función que minimice la expresión anterior.

PROPOSICIÓN (p3.3)

Sea $n \geq 3$ y sean los puntos x_1, \dots, x_n tales que $a < x_1 < \dots < x_n < b$.

Sea $W_2^1([a, b]) = \left\{ g : [a, b] \rightarrow \mathbb{R} / \exists g^{(1)}(x) / \int_a^b (g^{(1)}(x))^2 dx < \infty \right\}$.

Dadas las observaciones y_1, \dots, y_n y un parámetro de suavizado $\lambda > 0$, la solución del problema

$$[p] \min_{g \in W_2^1[a, b]} \Psi(g) = \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int_a^b (g''(x))^2 dx,$$

es un spline cúbico natural con nodos en x_1, \dots, x_n .

DEMOSTRACIÓN:

Sea $g \in W_2^1([a, b])$ y que no es un spline con nodos t_1, \dots, t_r . Sea \hat{g} el spline cúbico natural con nodos en x_1, \dots, x_n que interpola $(x_1, g(x_1)), \dots, (x_n, g(x_n))$. Así $\hat{g}(x_i) = g(x_i), i = 1, \dots, n$ y

$$\sum_{i=1}^n (y_i - g(x_i))^2 = \sum_{i=1}^n (y_i - \hat{g}(x_i))^2.$$

Por otra parte, $\int_a^b (\hat{g}''(x))^2 dx < \int_a^b (g''(x))^2 dx$ por la proposición p3.2. Así,

$\Psi(\hat{g}) < \Psi(g)$ y por lo tanto el óptimo del problema [p] hay que buscarlo entre los splines cúbicos con nodos en x_1, \dots, x_n como queríamos demostrar.

Existe la posibilidad de calcular el spline de regresión utilizando una variación del método de los momentos.

PROPOSICIÓN (p3.4):

Sean

- $m = (M_1, \dots, M_{n-1})$ vector de momentos.
- $f = (f_0, \dots, f_n) = (f(x_0), \dots, f(x_n))$ vector de valores del spline en los knots.
- $y = (y_0, \dots, y_n)$ vector de valores reales de la función en los knots.
- R y Q las matrices descritas en el punto 3.1.2.
- W matriz de pesos para cada uno de los valores.

Entonces el spline s que minimiza:

$$J_\alpha(s) = \alpha \sum_{i=0}^n w_i [y_i - g(x_i)]^2 + (1-\alpha) \int_0^n g''(x)^2 dx = \alpha A + (1-\alpha)B$$

es el que tiene como momentos:

$$\hat{m} = [\alpha R + (1-\alpha) Q W^{-1} Q']' \alpha Q y$$

Donde A es la parte de mínimos cuadrados y B es la penalización a funciones no suaves.

Además los valores estimados de la y tienen la siguiente expresión:

$$\hat{y} = y - \left(\frac{1-\alpha}{\alpha} \right) W^{-1} Q' m$$

La anterior proposición nos proporciona en primer lugar los momentos

DEMOSTRACIÓN:

- Nótese que:

$$\begin{aligned} s''_i(x) &= M_i + \frac{M_{i+1} - M_i}{h_i} (x - x_i) = \begin{pmatrix} \text{expresado} \\ \text{tambien} \end{pmatrix} = \frac{1}{h_i} \{ b_i M_i + M_{i+1}(x - x_i) - M_i(x - x_i) \} = (b_i = x_{i+1} - x_i) \\ &= \frac{1}{h_i} \{ M_i(x_{i+1} - x_i - x + x_i) + M_{i+1}(x - x_i) \} = \frac{1}{h_i} \{ M_i(x_{i+1} - x) + M_{i+1}(x_i - x) \} \end{aligned}$$

- Podemos expresar A a través de vectores y matrices

$$A = \sum_{i=0}^n w_i [y_i - f(x_i)]^2 = (y - f)' W (y - f)$$

- Calculamos cuánto vale B:

$$\begin{aligned} B &= \int_{x_i}^{x_{i+1}} s''(x)^2 dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} s''(x)^2 dx = \begin{pmatrix} \text{sustituimos} \\ s''(x) \end{pmatrix} = \sum_{i=0}^{n-1} \frac{1}{h_i} \int_{x_i}^{x_{i+1}} [M_i(x_{i+1} - x) + M_{i+1}(x_i - x)]^2 dx = \\ &= (\text{apendice 1.1.a}) = \sum_{i=0}^{n-1} \frac{1}{3} h_i (M_i^2 + M_{i+1}^2 + M_i M_{i+1}) = (\text{apendice 1.1.b}) = \\ &= \frac{1}{6} \sum_{i=1}^{n-1} h_{i-1} M_{i-1} M_i + 2 M_i^2 (h_{i-1} + h_i) + h_i M_i M_{i+1} = m' R m \end{aligned}$$

- Aprovechando el hecho de que R es simétrica e invertible, la función a minimizar expresada matricialmente es ahora:

$$\begin{aligned} J_\alpha(g) &= \alpha (y - f)' W (y - f) + (1-\alpha) m' R m = \begin{pmatrix} R m = Q f \\ m = R^{-1} Q f \\ m' = f' Q' R^{-1} \end{pmatrix} = \\ &= \alpha (y - f)' W (y - f) + (1-\alpha) f' Q' R^{-1} Q f \end{aligned}$$

- Podemos minimizar $J_\alpha(s)$ respecto de f derivando e igualando al vector cero:

$$\frac{\partial J_\alpha(s)}{\partial f} = 0 \Leftrightarrow -2\alpha W(y - f) + 2(1-\alpha)Q'R^{-1}Qf = 0 \Leftrightarrow \\ \hat{f} = \hat{y} = y - \left(\frac{1-\alpha}{\alpha}\right)W^{-1}Q'R^{-1}Qf = (m = R^{-1}Qf) = y - \left(\frac{1-\alpha}{\alpha}\right)W^{-1}Q'm$$

Y con esto ya hemos obtenido la estimación de las y 's.

- Pero también interesa despejar m :

$$\frac{\partial J_\alpha(s)}{\partial f} = 0 \Leftrightarrow -2\alpha W(y - f) + 2(1-\alpha)Q'R^{-1}Qf = 0 \Leftrightarrow -2\alpha W(y - f) + 2(1-\alpha)Q'm = 0 \Leftrightarrow \\ \begin{pmatrix} \text{multiplicando} \\ \text{por } W^{-1}Q \end{pmatrix} \Leftrightarrow -2\alpha QW^{-1}Wy - 2\alpha QW^{-1}Wf + 2(1-\alpha)QW^{-1}Q'm = 0 \Leftrightarrow \\ (Qf = R) \Leftrightarrow -2\alpha Qy + 2\alpha Rm + 2(1-\alpha)QW^{-1}Q'm = 0 \Leftrightarrow \\ m = [\alpha R + (1-\alpha)QW^{-1}Q']' \alpha Qy$$

como queríamos demostrar.

Una vez calculados tanto los momentos como las estimaciones de las y 's, el proceso de estimación de los coeficientes de los splines es idéntico pero considerando ahora los nuevos momentos y también que $A_i = \hat{y}_i$. Es decir:

$$B_i = \frac{\hat{y}_{i+1} - \hat{y}_i}{h_i} - (M_{i+1} - M_i) \frac{b_i}{6} - M_i \frac{b_i}{2} \quad i=0, \dots, n-1$$

Y los coeficientes de cada uno de los polinomios del spline se hallan de la siguiente manera:

$$S_i = \left\{ \frac{M_i}{2} x_i^2 - x_i^3 \left(\frac{M_{i+1} - M_i}{6h_i} \right) - B_i x_i + \hat{y}_i \right\} + x \left\{ -M_i x_i + 3x_i^2 \frac{M_{i+1} - M_i}{6h_i} + B_i \right\} + \\ + x^2 \left\{ \frac{M_i}{2} - 3x_i \frac{M_{i+1} - M_i}{6h_i} \right\} + x^3 \left\{ \frac{M_{i+1} - M_i}{6h_i} \right\}$$

4. – MÉTODOS COMPUTACIONALES Y SUBRUTINAS EN R.

Una vez conocidas las herramientas teóricas necesitaremos procedimientos que nos faciliten los cálculos. Para ello se ha decidido implementar los comandos en lenguaje R (proyecto CRAN). Este software es gratuito y posee unas propiedades que lo hacen especialmente apropiado para el proyecto. R posee tanto comandos propios de un lenguaje de programación como comandos estadísticos. Esto permitirá crear una serie de objetos y de comandos con los que los cálculos serán más sencillos.

4.1.- OPERACIONES CON SPLINES EN LENGUAJE R.

Tal como se ha comentado previamente, en el análisis de datos funcionales se opera con la forma explícita de las funciones. Esto en principio sobre una hoja en blanco resulta fácil pero en el ordenador sólo es posible con software específico. Por suerte vamos a trabajar con splines, que son fácilmente implementables. En esta sección explicaremos detalles de cómo se suman, restan y multiplican splines y explicaremos los algoritmos implementados en R que realizan estas operaciones.

4.1.1.- Representación de splines en lenguaje R.

Recordemos que en la definición general de spline (3.1) $S(x)$ es una función definida a trozos en donde en cada intervalo es un polinomio de grado p .

$$S(x) = \sum_{i=0}^{n-1} \xi_i(x) \theta_i(x) \text{ donde}$$

$$\theta_i(x) \text{ es un polinomio de grado } p \text{ y } \xi_i(x) = \begin{cases} 1 & \text{si } x \in [x_i, x_{i+1}] \\ 0 & \text{si } x \notin [x_i, x_{i+1}] \end{cases}$$

En lenguaje R podemos definir de manera fácil tipos sencillos de objetos tales como variables, vectores, matrices, etc. Un tipo muy útil de objeto la lista. Una lista es un conjunto de elementos sencillos. Hemos escogido para un objeto de tipo spline una lista con el siguiente formato:

Spline \equiv lista(x,s,r), donde

`x=Spline$x`: Vector ordenado de knots $x_0 < x_1 < \dots < x_n$. (n: número de subintervalos)

`S=Spline$s`: Matriz de coeficientes de los polinomios. La fila i-ésima corresponde al polinomio efectivo en el intervalo $[x_{i-2}, x_{i-1}]$. Nótese que los intervalos $(-\infty, x_0]$ y $(x_n, +\infty)$ también tienen un polinomio asociado (fila 1 y n+2 respectivamente)

`r=Spline$r`: Valor lógico. "F" =spline no registrado. "T"=spline registrado

Nótese que tanto los splines interpoladores como los splines por regresión no paramétrica van a tener la misma forma. Esto resulta muy ventajoso ya que tanto los procedimientos de análisis descriptivo como el cálculo del análisis de componentes principales funcional trabaja con el mismo formato de memoria, lo que reduce el número de subrutinas a programar. El significado específico de la componente *r* quedará aclarado en el capítulo 5. De momento podemos trabajar con la idea de que dos splines registrados tienen los mismos nodos.

EJEMPLO:

Supongamos que tenemos los siguientes datos:

```
> x  
[1] 0 1 2 3 5  
> y  
[1] 2 3 4 3 2
```

El comando `splinecalc(x,y)` calcula el spline interpolador para estos datos:

```
spl1<-splinecalc(x,y)  
> spl1  
$x  
[1] 0 1 2 3 5  
$s  
[,1] [,2] [,3] [,4]  
[1,] 2.000000 0.8546512 0.000000 0.00000000  
[2,] 2.000000 0.8546512 0.000000 0.14534884  
[3,] 2.872093 -1.7616279 2.616279 -0.72674419  
[4,] -9.034884 16.0988372 -6.313953 0.76162791  
[5,] 13.962209 -6.8982558 1.351744 -0.09011628  
[6,] 2.697674 -0.1395349 0.000000 0.00000000  
$r  
[1] "f"
```

Por lo tanto lo que hemos obtenido es lo siguiente:

$$S(x) = \begin{cases} 2 + 0.85x & \text{si } x \in (-\infty, 0) \\ 2 + 0.85x + 0.15x^3 & \text{si } x \in [0, 1] \\ 2.87 - 1.76x + 2.62x^2 - 0.73x^3 & \text{si } x \in [1, 2) \\ -9.03 + 16.10x - 6.31x^2 + 0.76x^3 & \text{si } x \in [2, 3) \\ 13.96 - 6.90x + 1.35x^2 - 0.09x^3 & \text{si } x \in [3, 5] \\ 2.69 - 0.13x & \text{si } x \in (5, +\infty) \end{cases}$$

Representadolo gráficamente:

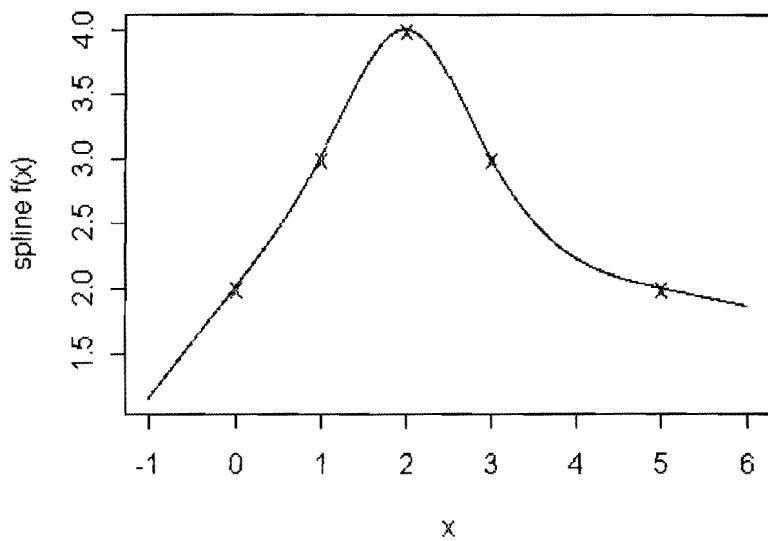


Figura 4.1: Representación gráfica del spline interpolador que pasa por los puntos $\{(0,2), (1,3), (2,4), (3,3), (5,2)\}$, calculado y dibujado con los comandos `splinecalc` y `dibusspline` respectivamente.

4.1.2.- Procedimientos de evaluación de splines.

Supongamos que ya tenemos calculado nuestro spline y queremos saber el valor de $S(x)$. Como $S(x)$ es una función definida a trozos averiguaremos primero en qué intervalo está situado x , y luego evaluaremos el polinomio que corresponde a ese intervalo en ese valor.

(A) Función buscacoeff.

Entrada: Spline a evaluar, x_s .

Salida: Coeficientes del polinomio, número de intervalo $I_i \in x_s$.

Descripción: Inicialmente comprueba que x_s se halla en el intervalo de los knots. Si está dentro entonces realiza un recorrido por los knots de los splines hasta encontrar el intervalo $I_i \in x_s$, dando de salida los coeficientes del polinomio correspondiente. Si está por debajo del mínimo de los knots la salida son los primeros coeficientes y si está por encima del máximo la salida son los últimos coeficientes. Cuando la evaluación se tiene que hacer para un vector de valores ordenados de tamaño n es muy útil guardar la posición “ i ” en que se ha quedado la última búsqueda, ya que así nos ahorraremos “ i ” evaluaciones y hallar los n polinomios representará $O(n)$ evaluaciones. Para detalles del código fuente consultar apéndice (4.1).

Una vez averiguados cuáles son los coeficientes del polinomio $S_i(x)$ únicamente hay que evaluarlo en x_s . Esto lo realizamos con la función PL.

(B) Función PL

Entrada: Coeficientes del polinomio (pol) y valor x_s ($equis$).

Salida: $pol(equis)$.

Descripción: Evalua el polinomio $pol(a_0 + a_1 * x + a_2 * x^2 + \dots + a_n * x^{n-1})$ en el valor $equis$. pol es un vector formado por los coeficientes en este orden $(a_0, a_1, a_2, \dots, a_n)$.

4.1.3.- Suma o resta de splines.

Las operaciones implicadas en una suma o resta de splines son sumas o restas de polinomios por intervalos. Nótese que:

$$\begin{aligned} p^3(x) + q^3(x) &= (a_0 + b_1 x + c_1 x^2 + d_1 x^3) + (a_2 + b_2 x + c_2 x^2 + d_2 x^3) = \\ &= (a_0 + a_2) + (b_1 + b_2)x + (c_1 + c_2)x^2 + (d_1 + d_2)x^3 \end{aligned}$$

Sumar y restar splines es entonces proceso muy sencillo si los splines que intervienen en las operaciones ya han pasado por el proceso de *registro* (véase capítulo 5). En este caso no es preciso ningún procedimiento en R sino simplemente sumar o restar las matrices de coeficientes de cada uno de los splines. Veamos un ejemplo:

Supongamos que calculamos dos splines con diferentes knots:

<pre>> datos1 [[1]] [1] 0 1 2 3 5 [[2]] [1] 2 3 4 3 2</pre>	<pre>> datos2 [[1]] [1] 2 3 4 6 [[2]] [1] 1 2 3 2 > spl1<-splinecalc(datos1[[1]],datos1[[2]]) > spl2<-splinecalc(datos2[[1]],datos2[[2]])</pre>
---	--

El resultado será el siguiente:

<pre>> spl1 \$x [1] 0 1 2 3 5 \$s [,1] [,2] [,3] [,4] [1,] 2.000000 0.8546512 0.000000 0.00000000 [2,] 2.000000 0.8546512 0.000000 0.14534884 [3,] 2.872093 -1.7616279 2.616279 -0.72674419 [4,] -9.034884 16.0988372 -6.313953 0.76162791 [5,] 13.962209 -6.8982558 1.351744 -0.09011628 [6,] 2.697674 -0.1395349 0.000000 0.00000000 \$r [1] "f"</pre>	<pre>> spl2 \$x [1] 2 3 4 6 \$s [,1] [,2] [,3] [,4] [1,] -0.8695652 0.9347826 0.0000000 0.00000000 [2,] -1.3913043 1.7173913 -0.3913043 0.06521739 [3,] 9.1739130 -8.8478261 3.1304348 -0.32608696 [4,] -20.0434783 13.0652174 -2.3478261 0.13043478 [5,] 8.1304348 -1.0217391 0.0000000 0.00000000 \$r [1] "f"</pre>
---	--

Observamos que la suma de las dos funciones si no las registramos sería: En $(-\infty, 0)$ $\text{spl}_1 + \text{spl}_2$; en $[0, 1)$ $\text{spl}_1 + \text{spl}_2$; en $[1, 2)$ $\text{spl}_1 + \text{spl}_2$; en $[2, 3)$ $\text{spl}_1 + \text{spl}_2$ en $[3, 4)$ $\text{spl}_1 + \text{spl}_2$; en $[4, 5)$ $\text{spl}_1 + \text{spl}_2$ en $[5, 6)$ $\text{spl}_1 + \text{spl}_2$ y en $[6, +\infty)$ $\text{spl}_1 + \text{spl}_2$

Pero si pasamos previamente la fase de registro:

```

> listaejemplo<-regspl(list(sp11,sp12))
> listaejemplo
$ x
[1] 0 1 2 3 4 5 6
$s
$s[[1]]
[,1]      [,2]      [,3]      [,4]
[1,]  2.000000  0.8546512  0.000000  0.00000000
[2,]  2.000000  0.8546512  0.000000  0.14534884
[3,]  2.872093 -1.7616279  2.616279 -0.72674419
[4,] -9.034884 16.0988372 -6.313953  0.76162791
[5,] 13.962209 -6.8982558  1.351744 -0.09011628
[6,] 13.962209 -6.8982558  1.351744 -0.09011628
[7,]  2.697674 -0.1395349  0.000000  0.00000000
[8,]  2.697674 -0.1395349  0.000000  0.00000000

```

\$s[[2]]				
	[,1]	[,2]	[,3]	[,4]
[1,]	-0.8695652	0.9347826	0.0000000	0.00000000
[2,]	-0.8695652	0.9347826	0.0000000	0.00000000
[3,]	-0.8695652	0.9347826	0.0000000	0.00000000
[4,]	-1.3913043	1.7173913	-0.3913043	0.06521739
[5,]	9.1739130	-8.8478261	3.1304348	-0.32608696
[6,]	-20.0434783	13.0652174	-2.3478261	0.13043478
[7,]	-20.0434783	13.0652174	-2.3478261	0.13043478
[8,]	8.1304348	-1.0217391	0.0000000	0.00000000

\$r
[1] "t"

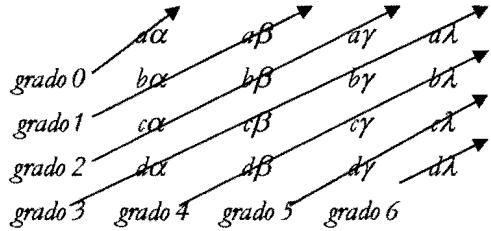
Al estar registrados los intervalos en los knots son los mismos para los dos splines y por lo tanto el problema se reduce a una suma de matrices: $listaejemplo\$s[[1]] + listaejemplo\$s[[2]]$ en el caso de una suma y $listaejemplo\$s[[1]] - listaejemplo\$s[[2]]$ en el de una resta.

4.1.4.- Producto de dos splines y cuadrado de un spline.

Al igual que en la suma y resta de polinomios, para calcular el producto de dos splines vamos a exigir que los splines implicados en las operaciones estén registrados. Por lo tanto el producto de dos splines registrados será un spline registrado para los mismos intervalos, donde el polinomio de I_i será el producto de los dos polinomios correspondientes a I_i de cada spline. Lo que sí habremos de implementar es una pequeña subrutina que nos calcule el producto de dos polinomios de mismo grado o el cuadrado de un polinomio. Si bien este algoritmo está preparado para polinomios de cualquier grado explicaremos el algoritmo para el caso de polinomios de grado 3.

$$p(x) \times q(x) = (a + bx + cx^2 + dx^3) \times (e + fx + gx^2 + hx^3) = ae + (af + be)x + (ag + bf + ce)x^2 + \\ + (ah + bg + cf + de)x^3 + (bh + cg + dh)x^4 + (ch)x^5 + dhx^6$$

Si representamos matricialmente este producto,



observamos que la suma de los elementos de las semidiagonales inversas corresponde al coeficiente de cada uno de los grados del polinomio producto. La idea del siguiente algoritmo consiste en calcular los coeficientes del polinomio producto haciendo el recorrido de la matriz anterior. Los índices del algoritmo dentro de cada grado evolucionan restando filas y sumando columnas. Lo único que cambia es que hasta el grado 3 empezamos desde la columna 1 y a partir del grado 4 hasta el final empezamos en la columna 2,3,4, etc. El código del algoritmo se encuentra en el apéndice.

El algoritmo de calcular el cuadrado del polinomio es análogo al anterior.

EJEMPLO:

$$p(x) \times q(x) = (1 + 2x + 3x^2 + 4x^3) \times (2 + 3x + 4x^2 + 5x^3) = 1 \cdot 2 + (1 \cdot 3 + 2 \cdot 2)x + (1 \cdot 4 + 2 \cdot 3 + 3 \cdot 2)x^2 + \\ + (1 \cdot 5 + 2 \cdot 4 + 3 \cdot 3 + 4 \cdot 2)x^3 + (2 \cdot 5 + 3 \cdot 4 + 4 \cdot 3)x^4 + (3 \cdot 5 + 4 \cdot 4)x^5 + 4 \cdot 5x^6 = \\ = 2 + 7x + 16x^2 + 30x^3 + 34x^4 + 31x^5 + 20x^6$$

```
>> mutpol(c(1,2,3,4),c(2,3,4,5))
[1] 2 7 16 30 34 31 20
```

4.1.5.- Integral definida de un spline.

La integral de un spline, al ser éste función definida a trozos, es una suma de integrales definidas para cada uno de los intervalos que definen los knots. Es decir:

$$\int_a^b S(x)dx = \sum_{i=-1}^n \xi_i S_i(x)dx = \int_{-\infty}^a S_{-1}(x)I_{[a,b]}(x)dx + \int_a^0 S_0(x)I_{[a,b]}(x)dx + \\ \dots + \int_{x_i}^{x_{i+1}} S_i(x)I_{[a,b]}(x)dx + \dots + \int_{x_{n-1}}^{\infty} S_{n-1}(x)I_{[a,b]}(x)dx + \int_{x_n}^{\infty} S_n(x)I_{[a,b]}(x)dx$$

Cada una de las integrales es de fácil cálculo al ser la integral en polinomio. Se ha creado una función que evalúa el polinomio integrado llamada *int.pl*. El cálculo de la sub-integral siguiente se reduce a aplicar la regla de Barrow de la siguiente manera:

$$\int_{x_i}^{x_{i+1}} S_i(x) dx = \text{int.pl}(S_i(x); x_{i+1}) - \text{int.pl}(S_i(x); x_i)$$

Calcular la integral completa se reduce a sumar todas las sub-integrales. El procedimiento *int.pl* se encuentra en el apéndice.

4.2.- PROCEDIMIENTOS DE CÁLCULO EN R DE LA DESCRIPTIVA FUNCIONAL.

4.2.1.- Cálculo de la media de una función.

La fórmula de cálculo de la media de una función es la siguiente:

$$\bar{x}_{x(t)} = \frac{1}{\int_{x_0}^x I(t) dt} \int_{x_0}^x x(t) I(t) dt = \frac{1}{\frac{x_n - x_0}{n} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} S_i(t) dt} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} S_i(t) dt = \begin{cases} \text{considerando} \\ \text{la función} \\ \text{int.pl} \end{cases}$$

$$= \frac{1}{\frac{x_n - x_0}{n}} \sum_{i=0}^{n-1} (\text{int.pl}(S_i(t); x_{i+1}) - \text{int.pl}(S_i(t); x_i))$$

El procedimiento de cálculo de la media de una función se llama *med.f.sint* y se ayuda del procedimiento *int.pl* (véase punto 4.1.5) y de otro de verificación de si se trata o no de un spline (*esspline*). El código principal se encuentra en el apéndice.

EJEMPLO: Supongamos que tenemos los siguientes datos:

> datos1	\$s
[1]	
[1] 0 1 2 3 5	[,1] [,2] [,3] [,4]
	[1,] 2.000000 0.8546512 0.000000 0.00000000
	[2,] 2.000000 0.8546512 0.000000 0.14534884
[{2}]	[3,] 2.872093 -1.7616279 2.616279 -0.72674419
[1] 2 3 4 3 2	[4,] -9.034884 16.0988372 -6.313953 0.76162791
	[5,] 13.962209 -6.8982558 1.351744 -0.09011628
> spl1	[6,] 2.697674 -0.1395349 0.000000 0.00000000
\$x	
[1] 0 1 2 3 5	\$r
	[1] "f"
	> med.f.sint(spl1)
	[1] 2.8625

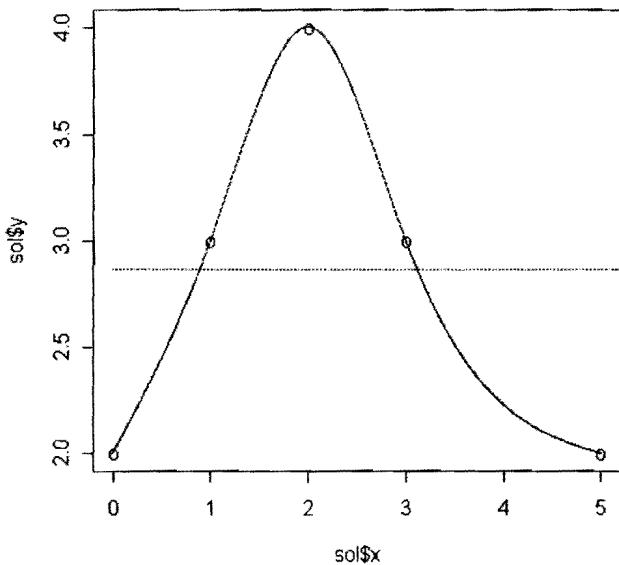


Figura 4.2: Representación gráfica del spline interpolador que pasa por los puntos $\{(0,2), (1,3), (2,4), (3,3), (5,2)\}$, y de la función constante valor medio de dicho spline.

Obsérvese que el área bajo la línea roja tiene el mismo valor que el área bajo la función.

4.2.2.- Cálculo de la varianza de una función.

La fórmula de cálculo de la varianza de una función es la siguiente:

$$\sigma_{x(t)}^2 = \frac{1}{\int_{x_0}^{x_n} f(t) dt} \int_{x_0}^{x_n} (x(t) - \bar{x}_{x(t)})^2 dt = \frac{1}{x_n - x_0} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} (f_i(t) - \bar{x}_{x(t)})^2 dt$$

Fijémonos que la expresión que hay que elevar al cuadrado es un polinomio menos un valor constante. Esto equivale a restarle la media de la función al término de grado cero del polinomio. Esto es lo que hace la función *constpl*. Además en esta función se utilizan las funciones *esspline*, *int.pl* y *polquad* (calcula el cuadrado de un polinomio y funciona como *mutpol* del punto 4.1.4). Por lo tanto:

$$\sigma_{x(t)} = \frac{1}{x_n - x_0} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \text{polquad}\left(\text{consPL}(S_i(t); \bar{x}_{x(t)}), dt\right) = \begin{pmatrix} \text{polaux} = \\ \text{polquad}\left(\text{consPL}(S_i(t); \bar{x}_{x(t)}), dt\right) \end{pmatrix} =$$

$$= \frac{1}{x_n - x_0} \sum_{i=0}^{n-1} (\text{int.pl}(\text{polaux}; x_{i+1}) - \text{int.pl}(\text{polaux}; x_i))$$

EJEMPLO:

Utilizando el mismo spline que en el punto 4.2.1 obtenemos lo siguiente:

```
> var.f.sint(spl1)
[1] "es spline interpolador"
[1] 0.4570101
```

OBSERVACIÓN:

Tanto la media como la varianza de una función se ven afectados por la operación de registro debido a que los knots de inicio y de fin, que aparecen en la fórmula de cálculo, pueden ser distintos a cuando el spline no estaba registrado.

4.2.3.- Cálculo de la covarianza de dos splines registrados.

El cálculo de la covarianza entre dos funciones es parecido al de la varianza. La única diferencia a nivel de programación está en que usaremos la función *mutpol* en vez de *polquad*. En los cálculos de la media y la varianza de una función no era necesario que los splines estuviesen registrados. En el cálculo de la covarianza sí vamos a exigir que lo estén, para evitar los problemas que ya comentábamos en el punto 4.1.3. Es decir:

$$\sigma_{x(t), y(t)} = \frac{1}{\int_{x_0}^x j(t) dt} \int_{x_0}^x (x(t) - \bar{x}_{x(t)}) (y(t) - \bar{y}_{y(t)}) dt = \begin{pmatrix} \text{polauxxx} = \\ \left(\text{consPL}(S_{x(t), y(t)}(t); \bar{x}_{x(t)}), dt\right) \end{pmatrix}$$

$$= \frac{1}{x_n - x_0} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \text{mutpol}(\text{polauxxx}; \text{polauxxy}) dt = \begin{pmatrix} \text{polaux} = \\ \text{mutpol}(\text{polauxxx}; \text{polauxxy}) \end{pmatrix} =$$

$$= \frac{1}{x_n - x_0} \sum_{i=0}^{n-1} (\text{int.pl}(\text{polaux}; x_{i+1}) - \text{int.pl}(\text{polaux}; x_i))$$

La función *covar.f.sint* es la que realiza el cálculo de la covarianza. Ésta se ayuda de otros procedimientos ya comentados: *mutpol*, *consPL*, *int.pl* y *med.f.sint*. La función en R se encuentra en el apéndice.

EJEMPLO:

> datos1 [[1]] [1] 0 1 2 3 5 [[2]] [1] 2 3 4 3 2 > datos2 [[1]] [1] 2 3 4 6 [[2]] [1] 1 2 3 2	> spl1<-splinecalc(datos1[[1]],datos1[[2]]) > spl2<-splinecalc(datos2[[1]],datos2[[2]]) listaspls<-regsp1(list(spl1,spl2)) > covar.f.sint(listaspls,1,2) [1] -0.3805915 > covar.f.sint(listaspls,2,1) [1] -0.3805915
--	--

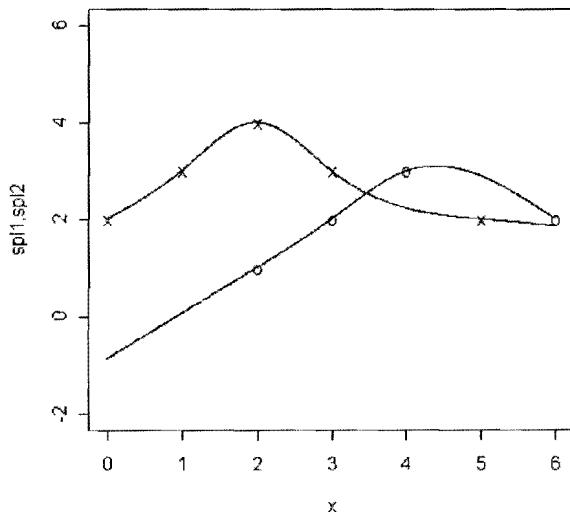


Figura 4.3: Representación gráfica de un spline interpolador que pasa por los puntos $\{(0,2), (1,3), (2,4), (3,3), (5,2)\}$, y un spline interpolador generado con los puntos $\{(2,1), (3,2), (4,3), (6,2)\}$. Estas dos funciones tienen una covarianza de -0.38 .

4.2.4.- Cálculo de los momentos no centrados de primer y segundo orden.

No existe ninguna función que calcule directamente la función media ni la función varianza, aunque la ejecución del procedimiento *desc.f.spls* nos proporciona una lista que los contiene. La función media se puede calcular a partir de la función *mom.1*(a partir de ahora también $m_1(t)$), que calcula el momento no centrado de una lista de splines registrada, multiplicada por N. Es decir, calcula únicamente el sumatorio. La varianza se puede calcular también mediante los momentos de primer y segundo orden. El sumatorio de las funciones al cuadrado se calcula con la función *mom.2* ($m_2(t)$)

La función media consiste simplemente en calcular *mom.1* y dividirla entre el tamaño de muestra.

$$\bar{x}(t) = \frac{1}{N} \sum_{i=1}^N x_i(t) = \frac{1}{N} m_1(t)$$

De igual manera la función varianza se puede obtener a través de los momentos de primer y segundo orden:

$$\begin{aligned} V(t) &= \frac{1}{N} \sum_{i=1}^N (x_i(t) - \bar{x}(t))^2 = \frac{1}{N} \sum_{i=1}^N x_i(t)^2 + \frac{1}{N} \sum_{i=1}^N \bar{x}(t)^2 - \frac{2}{N} \sum_{i=1}^N x_i(t) \bar{x}(t) = \\ &= \frac{1}{N} m_2(t) + \bar{x}(t)^2 - \frac{2}{N} \bar{x}(t) \cdot N \cdot \bar{x}(t) = \frac{1}{N} m_2(t) - \bar{x}(t)^2 = \frac{1}{N} m_2(t) - \left(\frac{1}{N} m_1(t) \right)^2 \end{aligned}$$

Recordemos que tanto $m_1(t)$ como $m_2(t)$ se ejecutan si los splines están registrados según los mismos knots que la lista de splines de origen. Calcular los $m_1(t)$ y $m_2(t)$ es fácil, si nos basamos en el siguiente hecho:

$$\begin{aligned} m_1(t) &= \sum_{i=1}^N S_i(t) = \begin{pmatrix} \text{Definicion} \\ \text{de spline} \end{pmatrix} = \sum_{i=1}^N \left(\sum_{j=-1}^{n_i} \xi_{i,j} \theta_{i,j} \right) = \begin{pmatrix} \text{Splines} \\ \text{registrados} \end{pmatrix} = \sum_{j=-1}^n \xi_j \left(\sum_{i=1}^N \theta_{i,j} \right) = \\ &= \sum_{j=-1}^n \xi_j \theta_j \\ m_2(t) &= \sum_{i=1}^N S_i^2(t) = \begin{pmatrix} \text{Definicion} \\ \text{de spline} \end{pmatrix} = \sum_{i=1}^N \left(\sum_{j=-1}^{n_i} \xi_{i,j} \theta_{i,j}^2 \right) = \begin{pmatrix} \text{Splines} \\ \text{registrados} \end{pmatrix} = \sum_{j=-1}^n \xi_j \left(\sum_{i=1}^N \theta_{i,j}^2 \right) = \\ &= \sum_{j=-1}^n \xi_j \theta_j^2 \end{aligned}$$

Hemos de tener en cuenta que los coeficientes de los polinomios en R son una matriz de coeficientes, donde la i-esima fila contiene los coeficientes del polinomio para el intervalo

$[x_{i,2}, x_{i,l}]$. El momento de primer orden multiplicado por N se calculará simplemente sumando las matrices de coeficientes. El momento de segundo orden multiplicado por N consiste en calcular para cada spline una matriz con los coeficientes de los polinomios al cuadrado y sumarlos. El detalle de las funciones *mom.1*, *mom.2* y *descf.spls* se pueden encontrar en el apéndice.

4.3.- CÁLCULO DE COMPONENTES PRINCIPALES FUNCIONALES PARA FUNCIONES REPRESENTADAS EN SPLINES.

4.3.1.- Planteamiento del problema

Supongamos que tenemos una muestra de funciones aleatorias debidamente registradas y expresadas como splines:

$$\hat{X}_1(t), \hat{X}_2(t), \dots, \hat{X}_N(t)$$

Donde

$$\hat{X}_i(t) = S_i(t) = \sum_{j=0}^{n-1} \xi_j(x) \phi_{i,j}(t) \text{ donde}$$

$$\phi_{i,j}(t) \text{ es un polinomio de grado } p \text{ y } \xi_j(t) = I_{[x_j, x_{j+1}]}(t) = \begin{cases} 1 & \text{si } t \in [x_j, x_{j+1}] \\ 0 & \text{si } t \notin [x_j, x_{j+1}] \end{cases}$$

y knots x_0, x_1, \dots, x_n

Si queremos representar el spline cúbico (interpolador o de regresión) en función de los coeficientes de cada uno de los polinomios que definen el spline obtendremos lo siguiente:

$$\begin{aligned} \hat{X}_i(t) &= \sum_{j=0}^{n-1} I_{[x_j, x_{j+1}]}(t) \phi_{i,j}(t) = \\ &= I_{[x_0, x_1]}(t) (c_{i,1} + c_{i,2}t + c_{i,3}t^2 + c_{i,4}t^3) + I_{[x_1, x_2]}(t) (c_{i,5} + c_{i,6}t + c_{i,7}t^2 + c_{i,8}t^3) + \dots = \\ &= \sum_{j=0}^{n-1} I_{[x_j, x_{j+1}]}(t) \left(\sum_{k=0}^{p-1} c_{i,j(p+1)+(k+1)} t^k \right) \end{aligned}$$

donde $p=3$, N es el tamaño de la muestra y n es el número de knots.

Esto se puede expresar matricialmente como $\mathbf{x} = \mathbf{C}\boldsymbol{\phi}$, y considerando que se trata de splines cúbicos y de que $(n-1)(p+1) + (p+1) = n(p+1)$ entonces tenemos que:

$$\begin{pmatrix} \dot{X}_1(t), \dot{X}_2(t), \dots, \dot{X}_N(t) \end{pmatrix} =$$

$$\begin{pmatrix} c_{1,1} & c_{1,2} & c_{1,3} & c_{1,4} & \cdots & c_{1,(n-1)(p+1)+1} & c_{1,(n-1)(p+1)+2} & c_{1,(n-1)(p+1)+3} & c_{1,n(p+1)} \\ c_{2,1} & c_{2,2} & c_{2,3} & c_{2,4} & \cdots & c_{2,(n-1)(p+1)+1} & c_{2,(n-1)(p+1)+2} & c_{2,(n-1)(p+1)+3} & c_{2,n(p+1)} \\ c_{3,1} & c_{3,2} & c_{3,3} & c_{3,4} & \cdots & c_{3,(n-1)(p+1)+1} & c_{3,(n-1)(p+1)+2} & c_{3,(n-1)(p+1)+3} & c_{3,n(p+1)} \\ \vdots & & & & & & & & \\ c_{N,1} & c_{N,2} & c_{N,3} & c_{N,4} & \cdots & c_{N,(n-1)(p+1)+1} & c_{N,(n-1)(p+1)+2} & c_{N,(n-1)(p+1)+3} & c_{N,n(p+1)} \end{pmatrix} \begin{pmatrix} I_{[x_0, x_1]} \\ t \cdot I_{[x_0, x_1]} \\ t^2 I_{[x_0, x_1]} \\ t^3 I_{[x_0, x_1]} \\ \vdots \\ I_{[x_{n-1}, x_n]} \\ t \cdot I_{[x_{n-1}, x_n]} \\ t^2 I_{[x_{n-1}, x_n]} \\ t^3 I_{[x_{n-1}, x_n]} \end{pmatrix}$$

Nótese que en nuestro formato de spline la primera y la última fila de coeficientes corresponden a los polinomios que actúan en los intervalos $(-\infty, x_0)$ y $(x_n, +\infty)$ respectivamente. Pero en este caso estos dos coeficientes no se tienen en cuenta en el cálculo de las componentes principales. La subrutina en R que transforma una lista de splines registrados en la matriz C anterior se llama *monta.c* y se puede ver el código fuente en el apéndice (3,7)

Recordemos que en el punto 2.3.4 detallábamos que todo se reduce a calcular los vectores propios b y valores propios λ de una matriz A :

$$N' C' C W b = \lambda b \Leftrightarrow A b = \lambda b$$

Donde $W = (w_{k_1, k_2})_{k_1, k_2} = \left(\int_{\Omega} \phi_{k_1}(t) \phi_{k_2}(t) dt \right)_{k_1, k_2}$

Nótese que la matriz W es la matriz de covarianzas $\text{cov}[\phi, \phi]$. El detalle del cálculo de la matriz W puede verse en el punto 1.2. del apéndice.

4.3.2.- Cálculo de vectores y valores propios mediante raíz cuadrada de W .

En principio mediante la función *eigen* del lenguaje R podríamos calcular los vectores y valores propios de A . El problema es que este algoritmo cuando A no es simétrica, en ocasiones da una solución en números complejos. Existe una manera para realizar el cálculo de vectores y valores propios sobre una matriz simétrica:

$$\begin{aligned} \frac{1}{N} C' CW b = \lambda b &\Leftrightarrow \left(\begin{array}{l} \text{premultipliando por } W^{1/2} \\ \text{y considerando } W = W^{1/2} W^{1/2} \end{array} \right) \Leftrightarrow \frac{1}{N} W^{1/2} C' CW^{1/2} W^{1/2} b = \lambda W^{1/2} b \Leftrightarrow \\ &\Leftrightarrow \left(\begin{array}{l} \text{consideramos} \\ u = W^{1/2} b \end{array} \right) \Leftrightarrow \frac{1}{N} W^{1/2} C' CW^{1/2} u = \lambda u \end{aligned}$$

Fijémonos que ahora la matriz a diagonalizar va a ser simétrica, los valores propios λ son los mismos que los del problema original, y que los vectores propios del problema original b se obtienen fácilmente a través de la expresión: $b = W^{1/2} u$. Todo el problema se reduce entonces a calcular $W^{1/2}$.

Sea W una matriz simétrica. Entonces sabemos que

$$W^{1/2} = U D^{1/2} U'$$

Donde U y D provienen de la descomposición svd en vectores y valores propios de W .

Este resultado es bien conocido y no se precisa demostración.

4.3.3.- Cálculo de los coeficientes.

Una vez calculadas las funciones componentes principales nos interesa saber cuáles son los coeficientes de los individuos en las componentes $f_{i,k}$ tales que:

$$\hat{x}_i(t) = \sum_{k=1}^K f_{i,k} \cdot \xi_k(t)$$

Si $x(t)$ y $\xi(t)$ estuviesen expresados en una misma base $\phi(t)$, entonces

$$\underline{x}(t) = C\phi(t) \text{ y } \underline{\xi}(t) = B\phi(t) \Rightarrow \underline{x}(t) = CB^{-1}\underline{\xi}(t) = F\underline{\xi}(t)$$

y F serían los coeficientes de $x(t)$ en las CPFs. Pero en nuestra implementación no trabajamos con una base sino con un sistema de generadores que tiene dimensión $4n$, cuando la dimensión del espacio de splines es n . Por lo tanto el cálculo de las CPFs no es un cambio de base, y la matriz B que relaciona $\xi(t)$ con el sistema de generadores no es invertible. Así pues no podemos calcular los coeficientes F tales que $x(t) = F\xi(t)$.

Pero necesitamos una manera de relacionar $x_i(t)$ y $\xi_k(t)$ en función de su forma. Lo verdaderamente interesante va a ser relacionar cada una de las funciones muestra con las funciones principales mediante un solo coeficiente. Para ello necesitaremos una medida de relación entre $x_i(t)$ y $\xi_k(t)$. Podemos considerar scores a la covarianza y la correlación entre función principal y función de la muestra.

DEFINICIÓN 4.1

Diremos que el *score* de $x_i(t)$ en $\xi_j(t)$ a:

$$\alpha_{i,j} = \text{Corr}[x_i, \xi_j]$$

PROPOSICION (p.4.2)

Sea x el vector de funciones aleatorias, ξ el vector de funciones propias, W la matriz de varianzas-covarianzas $\text{cov}[\phi, \phi]$ y B la matriz de vectores propios. Entonces la covarianza y la correlación entre $x_i(t)$ y $\xi_j(t)$ tienen la siguiente expresión:

$$\text{Cov}[x, \xi] = CWB$$

$$\text{Corr}[x, \xi] = \text{Diag}(V[x])^{-\frac{1}{2}} \text{Cov}[x, \xi]$$

DEMOSTRACIÓN:

Para una función de la muestra en particular: $x_i(t) = c'_i \phi(t)$

Para una función propia: $\xi_j(t) = b'_j \phi(t)$

$$\text{Cov}[x_i, \xi_j] = c'_i \text{Cov}[\phi, \phi] b_j = c'_i W b_j$$

Y ampliándolo para matrices:

$$\text{Cov}[x, \xi] = CWB$$

En el caso de la correlación:

$$\text{Corr}[x_i, \xi_j] = \frac{\text{Cov}[x_i, \xi_j]}{\sqrt{\text{Var}[x_i] \text{Var}[\xi_j]}} = \begin{pmatrix} \text{al ser } \|\xi_j\| = 1 \\ y \|\xi_j\| = \text{Var}[\xi_j] \end{pmatrix} = \frac{\text{Cov}[x_i, \xi_j]}{\sqrt{\text{Var}[x_i]}}$$

Como sabemos que: $V[x] = CWC'$, si nos quedamos con una matriz con la diagonal de $V[x]$, elevada a $(-1/2)$ y el resto ceros, premultiplicando obtendremos, para cada uno de los elementos de la matriz resultante, la expresión de la correlación:

$$\text{Corr}[x, \xi] = \text{Diag}(V[x])^{-\frac{1}{2}} \text{Cov}[x, \xi],$$

donde

$$Diag(V[x])^{-\frac{1}{2}} = \begin{pmatrix} \sqrt{V[x_1]} & 0 & \cdots & 0 \\ 0 & \sqrt{V[x_2]} & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{V[x_n]} \end{pmatrix}$$

como queríamos demostrar.

4.3.4.- Función en R fPCA.sint.

La función R fPCA.sint se encarga de realizar las operaciones basándose en los resultados anteriores.

Entrada: Lista de splines registrados, interpoladores o de regresión.

Salida: Lista con vectores propios B, valores propios b, splines componentes principales ξ , scores correlación y dimensiones de las matrices.

Descripción: La función fPCA.sint en primer lugar calcula las matrices C y W, basándose en la teoría del punto 4.3.1 y del apéndice 1.2. Posteriormente calcula la matriz raíz cuadrada de W para crear la matriz simétrica que se diagonaliza (punto 4.3.2). Tras diagonalizar, deshace el cambio de variable y calcula los splines componentes principales ayudándose de la función sint.pca. Y por último calcula los scores covarianza y correlación, basándose en la idea del punto 4.3.3.

5. – ETAPA DE REGISTRO.

5.1.- ETAPA DE REGISTRO.

Definimos la etapa de registro como aquella etapa en la que transformamos los datos discretos en datos funcionales listos para el cálculo del análisis descriptivo. En esta etapa decidimos:

- Tipo de spline que utilizaremos y parámetro de suavizado en caso de spline regresión.
- Origen del tiempo de las funciones y rango de registro.
- Registro de knots sobre una lista de splines.
- Transformaciones.

Esta etapa es clave ya que con un mal registro podemos obtener malas estimaciones y tener problemas en los algoritmos de cálculo.

El primer punto hace referencia al capítulo anterior y básicamente consiste en conseguir splines que representen bien la función teórica de la que se extrae la muestra. En ocasiones el polinomio interpolador puede reproducir bien la función teórica. Pero en los casos en que no, habrá que recurrir al spline regresión y habrá que encontrar un parámetro de suavizado que se adapte bien a la función. Cuando tenemos poca información sobre la función observada, pero sabemos que nuestras mediciones están sujetas a error, el spline regresión suaviza el ruido.

En muchas ocasiones el origen del tiempo y el rango de observación teóricos está bien definidos. Sin embargo los datos discretos pueden haber sido mal introducidos y pueden aparecer tiempos fuera de rango. Es importante tener en cuenta esto a la hora de registrar los knots de una lista.

Las transformaciones nos ayudarán a destacar zonas de variabilidad de las funciones.

En este capítulo se detallará el registro de knots, los problemas de la etapa de registro y el caso de reducción de knots.

5.2.- REGISTRO DE KNOTS.

Una manera simplificar el cálculo de estadísticos descriptivos de una lista de splines es aplicar el registro de knots. Esta proceso permite que las operaciones básicas de suma y producto de splines se reduzcan a sumas y productos de matrices. El planteamiento es el siguiente:

- Un spline es una terna compuesta por los siguientes elementos $S_i(x) = \{K_i, \xi_i, \theta_i\}$ tales que :

$K_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n_i} \mid x_{i,1} < x_{i,2} < \dots < x_{i,n_i}\}$ conjunto ordenado de knots,

$\xi_i = \{\xi_{i,-1}, \xi_{i,0}, \dots, \xi_{i,n_i}\}$ funciones indicatrices con,

$$\xi_{i,j} = I_{[x_j, x_{j+1}]}, \xi_{i,-1} = I_{[-\infty, x_0]} \text{ y } \xi_{i,n_i} = I_{[x_{n_i}, +\infty]}$$

$\theta_i = \{\theta_{i,-1}, \theta_{i,0}, \dots, \theta_{i,n_i}\}$ conjunto de polinomios de grado 3,

que satisfacen las propiedades vistas en el capítulo 4.

Partiendo de esta base consideremos la siguiente definición

DEFINICIÓN 5.1:

Llamaremos *registro de knots* al proceso consistente en expresar todos los splines de una lista dada, de manera que compartan tanto el conjunto de knots como el de las funciones indicatrices. Es decir:

- Sea una lista de splines $S_1(x), S_2(x), \dots, S_N(x)$.
- Registrar los knots de esta lista de splines será expresar cada uno de ellos de la siguiente manera:

$S_i^R(x) = \{K^R, \xi^R, \theta_i^R\}$ tal que :

$$K^R = \bigcup_{i=1}^N K_i = \{x_o^R, x_1^R, \dots, x_n^R \mid x_o^R < x_1^R < \dots < x_n^R\}, \text{ que será un conjunto de tamaño } n+1,$$

$\xi^R = \{\xi_{i,-1}^R, \xi_i^R, \dots, \xi_{i,n_i}^R\}$ definidos análogamente en función de K^R y

$\theta_i^R = \{\theta_{i,-1}^R, \theta_{i,0}^R, \dots, \theta_{i,n_i}^R\}$ donde

$$\theta_{i,j}^R = \theta_{i,k} \text{ cuando } [x_{i,j}^R, x_{i,j+1}^R] \subset [x_{i,k}^R, x_{i,k+1}^R] \text{ con } n \geq n_i$$

Observese de la definición que:

- Para todos los splines el conjunto de knots y el de funciones indicatrices es el mismo
- El nuevo conjunto de polinomios θ_i^R es de mayor o igual dimensión pero contiene los mismos polinomios del spline sin registrar ya que algunos de ellos estarán repetidos.

EJEMPLO 5.1:

Sean los puntos de $f_1 = \{(x,y): (0,1), (1,2), (3,4), (4,2)\}$ y $f_2 = \{(x,y): (1,2), (2,4), (4,3)\}$

Calculamos los splines y nos queda lo siguiente:

Spline de f_1 :

$$\begin{array}{ll}
 1+0.81x & (-\infty, 0) \\
 1+0.81x+0.19x^3 & (0, 1) \\
 1.56-0.87x+1.68x^2-0.37x^3 & (1, 3) \\
 -23.75+24.44x-6.75x^2+0.56x^3 & (3, 4) \\
 12.25 -2.56x & (4, + \infty)
 \end{array}$$

Spline de f_2 :

$$\begin{array}{ll}
 -0.42+2.42x & (-\infty, 1) \\
 0+1.17x+1.25x^2 -0.42x^3 & (1, 2) \\
 -5+8.67x-2.50x^2+ 0.21x^3 & (2, 4) \\
 8.33-1.33x & (4, + \infty)
 \end{array}$$

En la etapa de registro vamos a considerar los mismos knots y los mismos intervalos tanto para uno como para otro spline. Por lo tanto los expresaremos así:

Spline de f_1 :

$$\begin{array}{ll}
 1+0.81x & (-\infty, 0) \\
 1+0.81x+0.19x^3 & (0, 1) \\
 1.56-0.87x+1.68x^2-0.37x^3 & (1, 2) \\
 1.56-0.87x+1.68x^2-0.37x^3 & (2, 3) \\
 -23.75+24.44x-6.75x^2+0.56x^3 & (3, 4) \\
 12.25 -2.56x & (4, + \infty)
 \end{array}$$

Spline de f_2 :

$$\begin{array}{ll}
 -0.42+2.42x & (-\infty, 0) \\
 -0.42+2.42x & (0, 1) \\
 0+1.17x+1.25x^2 -0.42x^3 & (1, 2) \\
 -5+8.67x-2.50x^2+ 0.21x^3 & (2, 3) \\
 -5+8.67x-2.50x^2+ 0.21x^3 & (3, 4) \\
 8.33-1.33x & (4, + \infty)
 \end{array}$$

Fijémonos que lo único que hemos hecho es reproducir el conjunto de intervalos que se deriva de la unión de los knots de un spline y otro, y a continuación hemos añadido el polinomio correspondiente a ese intervalo. Por eso nos quedan filas repetidas.

Obsérvese también que si queremos estimar la suma $f_1 + f_2$ mediante la suma de splines, $S_1(x) + S_2(x)$, es muy sencillo. Si definimos una matriz con los coeficientes de los polinomios de cada uno de los splines, la matriz suma contendrá los coeficientes de los polinomios del spline suma. El producto de dos splines será una matriz cuyas filas contendrán los coeficientes del polinomio resultante de multiplicar los dos polinomios asociados de las filas de los dos splines.

5.2. ERRORES EN LA ETAPA DE REGISTRO.

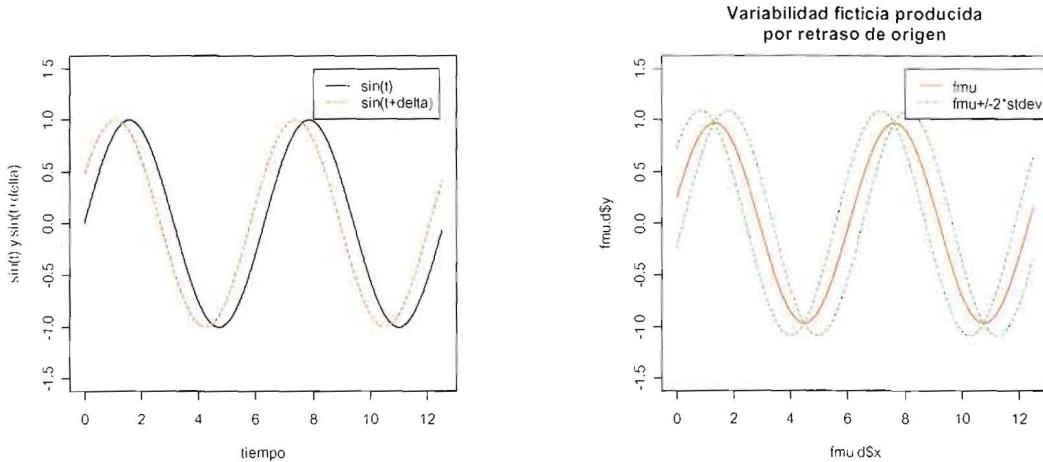
5.2. ERRORES EN LA ETAPA DE REGISTRO.

La etapa de registro es clave ya que existen errores asociados por realizar de manera incorrecta los cuatro puntos comentados anteriormente:

- Un error en el tipo de spline puede dar lugar a valores poco realistas o bien a valores extremos tanto de mínimos como de máximos. El spline interpolador puede dar formas muy irregulares y pocos suaves si tenemos muchos puntos en un rango pequeño.
- Un error en el origen del tiempo puede sesgar la estimación del valor medio y crear variabilidades ficticias debidas al “retraso” que puede tener una función respecto otra. Esto puede ser producido simplemente por haber sido sometida al estudio en un instante de tiempo distinto (Δ).

Dos funciones idénticas $f_i(x)$ tienen una función varianza igual a cero en todo \mathfrak{R} , sin embargo $f_i(x)$ y $f_i(x+\Delta)$ tienen función varianza distinta a cero, simplemente por tener un origen distinto.

Corrigiendo el origen de las funciones se evita inflar la función varianza de manera artificial.



Figuras 5.1 y 5.2: Representación de $\sin(t)$ y de $\sin(t+\Delta)$. Si no se ajusta este retraso Δ , al hacer la descriptiva funcional la función varianza será distinta de cero, aun tratándose de la misma función.

- Considerar un rango más grande de datos respecto los valores en que queremos evaluar la función puede dar a lugar a sesgo del valor medio y a variabilidades ficticias: Recordemos que los polinomios asignados a los extremos derecho e izquierdo de los splines son rectas. Si consideramos estos extremos de las funciones splines, estamos sumando áreas de rectas que van o a $-\infty$ o a $+\infty$, y estas áreas estarán contribuyendo al cálculo de estadísticos de primer y segundo nivel.
- Registrar una lista de splines puede crear un número elevado de knots, lo que puede producir inefficiencia en los algoritmos de cálculo, e incluso en el cálculo de componentes principales

puede producir matrices singulares. En el siguiente punto veremos cómo podemos solucionar este problema.

5.3.- REDUCCIÓN DEL NÚMERO DE KNOTS.

Tal y como acabamos de comentar la etapa de registro de knots puede acarrear problemas en el rendimiento de los algoritmos. En ocasiones puede ocurrir que en el procedimiento de componentes principales funcionales la diagonalización de la matriz asociada no pueda realizarse, a causa de que elevado número de knots que causan determinantes próximos a cero.

Sin embargo en una lista con splines registrados, si visualizamos las funciones vemos que con pocos knots podríamos obtener splines que explicasen formas casi idénticas, que nos conduzcan a soluciones eficientes de los algoritmos.

El propósito de este punto es el de conseguir, a partir de una lista de splines con knots registrados, una nueva lista con muchos menos knots de manera que no se pierda demasiada información. Esta pérdida de información la definimos a continuación.

DEFINICIÓN 5.2.

Diremos que la *pérdida asociada a la reducción de knots* de un spline extenso $S(x)$ respecto de su spline reducido $S_{rd}(x)$ en un intervalo $[a,b]$ es:

$$L(S(x), S_{rd}(x)) = \int_a^b (S(x) - S_{rd}(x))^2 dx$$

OBSERVACIONES:

- La pérdida va a depender de cómo escojamos el intervalo $[a,b]$.
- También dependerá de la manera en que escojamos el nuevo conjunto reducido de knots. Más que la cantidad de knots, va a ser más interesante escoger una combinación de knots cuyas posiciones en el eje de ordenadas permitan que el spline reducido se asemeje mucho al spline extenso.

Para conseguir una lista de splines reducida el algoritmo es el siguiente:

ALGORITMO DE REDUCCIÓN DE KNOTS:

Entrada: Lista de splines extensos (LSe) y conjunto reducido de knots (Kr).

Salida: Lista de splines reducidos (LSr), pérdida para cada uno de los splines (PSi), pérdida total (Pt).

PASOS:

- Ordenamos de menor a mayor Kr.
- Creamos (LSr) con knots Kr y coeficientes de los polinomios cero y Psi=0 para todo i.
- Para cada spline i de LSe (LSe i):

Evaluamos en el spline extenso LSe i en los puntos de Kr y los guardamos en el vector Y.

Calculamos un nuevo spline interpolador con los puntos Kr e Y.

La matriz de coeficientes asociada al nuevo spline será la matriz de coeficientes de LSri

Calculamos PSi:

Realizamos una sublistas registrada con LSe i y con LSri.

Calculamos un subspline $(LSe_i - LSri)^2$, e integramos en el rango de Kr.

- La suma del vector PSi será Pt.
- FIN.

En el capítulo 6 veremos que esta operación no es necesaria ya que tenemos pocos knots y una muestra grande, pero en el capítulo 7 veremos que en el ejemplo de datos de farmacocinética necesitaremos esta operación para poder calcular las componentes principales funcionales.

6. – ANÁLISIS FUNCIONAL DE DATOS EN PIRÁMIDES DE POBLACIÓN.

Para ilustrar el uso tanto del análisis funcional descriptivo (gráfico y numérico) como del ACPF hemos escogido dos parrillas de datos bien distintas. La primera parrilla son datos de pirámides de población (recuentos de población por edad y sexo). La segunda corresponde a concentraciones de quimioterápicos en la sangre a lo largo del tiempo desde su administración.

6.1. - APLICACIÓN DEL ANÁLISIS FUNCIONAL DE DATOS A PIRÁMIDES DE POBLACIÓN.

6.1.1. - Idea y definiciones

En demografía la manera más extendida de representar gráficamente la estructura por sexos y edad de una población en demografía es mediante una pirámide de población.

DEFINICIÓN 6.1.

Una *pirámide de población* es un gráfico compuesto por dos histogramas en el que la edad, usualmente en grupos de 5 años, se representa en el eje vertical, y las frecuencias (absolutas o relativas) crecen, desde el cero hacia la derecha en los hombres, y desde el cero hacia la izquierda en mujeres. Las pirámides han de cumplir:

- Para pirámides de frecuencia absoluta el área del total de las barras ha de corresponder al total de la población.
- Para pirámides de frecuencia relativa el área del total de las barras ha de ser igual a uno.

CONSECUENCIA:

La altura de cada una de las barras en el caso de edades agrupadas en grupos de cinco años es:

- Para pirámide de frecuencias absolutas:

$$h_{i-i+4,g} = \frac{1}{5} (p_{i,g} + p_{i+1,g} + p_{i+2,g} + p_{i+3,g} + p_{i+4,g})$$

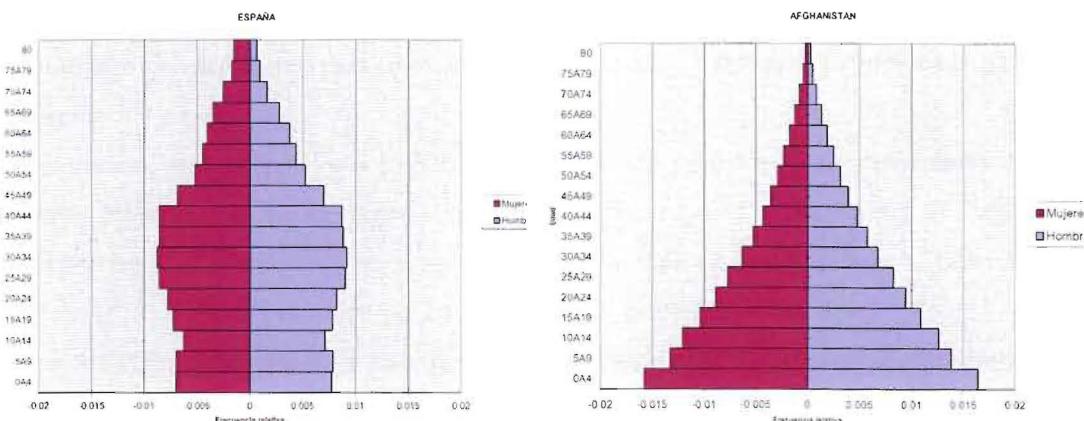
- Para pirámide de frecuencias relativas:

$$h_{i-i+4,g} = \frac{1}{5} \frac{p_{i,g} + p_{i+1,g} + p_{i+2,g} + p_{i+3,g} + p_{i+4,g}}{N}$$

donde

- p_{ig} = Total de población de edad i para el género g (hombres, mujeres).

- $N = \sum_{g=\{hombres, mujeres\}} \left(\sum_{j=1}^{\max\{edad\}} p_{i,g} \right)$



Figuras 6.1 y 6.2: Pirámides de población de España y Afganistán.

Lo interesante de las pirámides de población es que nos permiten ver simultáneamente la distribución de la edad de la población tanto en hombres como en mujeres. En el ejemplo vemos cómo Afganistán tiene mayor proporción de población en edades bajas y de manera aproximadamente lineal desciende conforme crece la edad. Además, observando para todos los grupos de edad, la frecuencia en los hombres es ligeramente superior a la de las mujeres. Sin embargo la pirámide de España tiene forma de ánfora, en la que la mayor frecuencia se centra entre los 25 y 40 años y a partir de los 55 años las mujeres representan más población que los hombres. Por tanto, la forma de las pirámides explica mucho sobre la distribución de las poblaciones.

El objetivo de este apartado será el de intentar resumir el comportamiento de las pirámides de población en unas pocas formas “principales” o también funciones de pautas de variabilidad. Interpretaremos las pirámides como funciones de frecuencias a lo largo del eje de la edad y luego calcularemos las ACPF's de estas funciones.

Vamos a suponer que todas las funciones se pueden explicar como una combinación lineal de K funciones básicas,

$$\hat{x}_i = \sum_{k=1}^K f_{i,k} \cdot \xi_k, \text{ donde } K \text{ es menor que la dimensión del espacio.}$$

6.1.2. - Obtención de datos

Los datos de pirámides de población han sido obtenidos a través de la Web de la *International Data Base (IDB)*, del *International Programs Center, U.S. Census Bureau*, con datos de 2000. En esta Web podemos encontrar, entre otros muchos datos, el recuento de población por grupos quinquenales de edad y sexo para cada uno de los 223 países o regiones, listados en la tabla de las páginas 6-7 y 6-8.

Los datos han sido trasladados a *EXCEL* y mediante una macro se han transformado. En la primera columna se ha traspasado el intervalo de edad. Los primeros 17 registros corresponden a hombres y los siguientes 17 a mujeres. Las siguientes columnas representan cada una el recuento de población para cada uno de los grupos de edad y sexo.

Una vez ejecutada la macro se han exportado los datos en formato csv (comma separate values) y se ha importado al R mediante la función *prop()*. Esta función a su vez calcula las frecuencias relativas de cada uno de los grupos de edad-sexo usando la fórmula de pirámides de frecuencia relativa indicada en el punto 6.1.1. La salida es un objeto en R llamado *datos* con dos atributos: *data* e *índice*. De esta forma tenemos acceso a las frecuencias de cada uno de los países de manera rápida. El índice está calculado para poder representar la pirámide abatida, es decir, la edad de los hombres es la marca de clase del grupo de edad en positivo y la de la mujer en negativo.

Marca	Frec.Relat	Marca	Frec.Relat	Marca	Frec.Relat	Marca	Frec.Relat
-82.5	0.002682684	-37.5	0.008102872	2.5	0.004462603	47.5	0.006517787
-77.5	0.003033323	-32.5	0.008402554	7.5	0.004658441	52.5	0.006135334
-72.5	0.004131856	-27.5	0.008594614	12.5	0.004914393	57.5	0.005766256
-67.5	0.004736286	-22.5	0.007549443	17.5	0.005682587	62.5	0.004727523
-62.5	0.004348084	-17.5	0.005980113	22.5	0.007209823	67.5	0.005493312
-57.5	0.005503198	-12.5	0.005214808	27.5	0.008286283	72.5	0.005163782
-52.5	0.006012251	-7.5	0.004967031	32.5	0.008127295	77.5	0.004287671
-47.5	0.006452897	-2.5	0.004758099	37.5	0.007968606	82.5	0.005297734
-42.5	0.007419115			42.5	0.007411345		

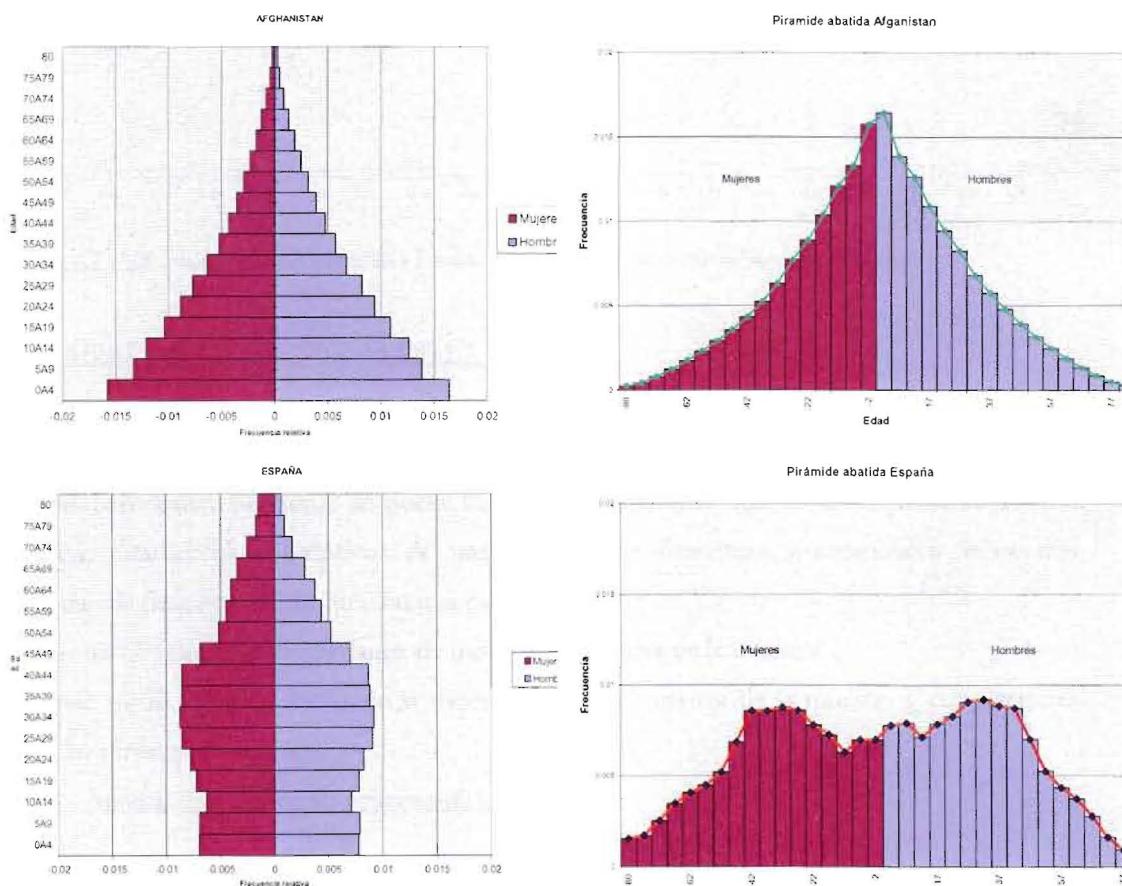
Tabla 6.1: Frecuencias relativas por edades (edades negativas hombres y positivas mujeres) de España.

6.1.3. - Transformación de pirámides en funciones.

Teniendo en cuenta que las pirámides son dos histogramas solapados y que la herramienta que utilizaremos es el cálculo de ACPF's, lo primero que haremos será transformar las pirámides en funciones mediante el abatimiento. La idea es la siguiente:

- Consideraremos las edades de las mujeres como negativas y la de los hombres como positivas. Representaremos en el eje horizontal la edad y en el eje vertical la frecuencia o densidad (abatimiento de la pirámide)
- La nueva función será: $f(\text{edad con signo}) = \text{frecuencia}$.

Gráficamente:

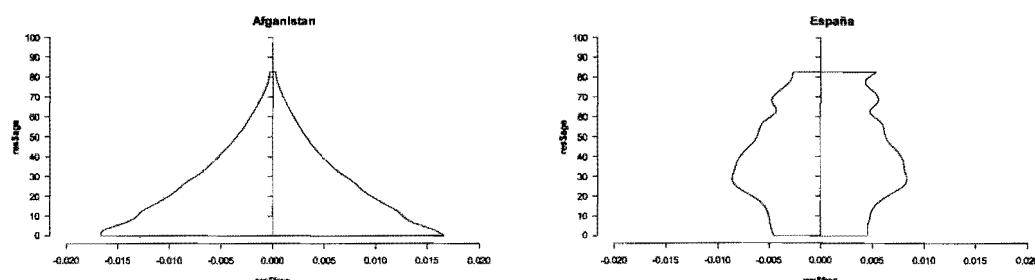


Figuras 6.3, 6.4, 6.5 y 6.6: Pirámides (fig.6.1 y 6.3) y pirámides abatidas (fig. 6.2 y 6.4) de Afganistán y España respectivamente. Obsérvese que en las pirámides abatidas el rango de edad va desde el -87.5 al 87.5.

Fijémonos que ahora estamos considerando que la edad con signo es una variable aleatoria continua que tiene definida una función de densidad distinta según sea el país. Cada una de las frecuencias relativas correspondiente a cada intervalo de edad será entonces la estimación de la función densidad en el valor de la edad igual a la marca de la clase del intervalo.

A partir de ahora trabajaremos con dichas funciones de densidad. Por tanto tenemos ahora 223 funciones, una para cada uno de los países. Nuestro objetivo va a ser el intentar describir las pautas de comportamiento de estas funciones de manera simple.

Se han creado dos funciones para representar en R las pirámides ya con formato de función:



Figuras 6.7 y 6.8: Pirámides de Afganistán y España representadas mediante splines interpoladores.

6.2. - ANÁLISIS DESCRIPTIVO DE LA SERIE.

El análisis descriptivo de la serie se calcula mediante la función `desc.f.spls`. Recordemos que el análisis descriptivo funcional se podía hacer en tres niveles: análisis descriptivo sobre una función, estadísticos descriptivos de una muestra de funciones y estadísticos sobre dos muestras de funciones. Esta función nos calcula:

Primer nivel: valor medio y varianza de todas las funciones de la muestra.

Segundo nivel: Función media de la muestra, función varianza de la muestra y correlaciones entre las distintas funciones.

6.2.1. - Análisis descriptivo de primer nivel.

Observando el resultado en la tabla 6.1 para nuestra serie de datos se constata que el valor medio de cada uno de los países muestra valores muy parecidos entre sí. Este hecho tiene su explicación en la interpretación geométrica del valor medio de una función: El valor medio de una función continua, finita definida en un intervalo $[a,b]$ es la altura que ha de tener una recta

paralela al eje horizontal para que contenga el mismo área que la propia función (véase propiedad de la definición 2.14). Por tanto si pensamos en una función de valor constante definida entre -82.5 y 82.5 , para que ésta sea una densidad (es decir, de área igual a uno), la altura ha de ser de $0,006060$, que es precisamente el valor medio al que se aproxima todas las funciones. Vemos entonces que, para que el área valga uno, los valores por encima de $0,006060$ en un grupo de edad han de ser compensados con valores por debajo de $0,006060$ en otros grupos de edad. Las discrepancias entre el valor medio de cada una de las funciones y la cifra $0,06060$ pueden ser atribuidas al error de la estimación por splines.

Otra consecuencia de la interpretación geométrica de la media es la de que la varianza será igual a cero si la distribución de la población es uniforme, es decir, si la función abatida es constante de valor $0,006060$. Por lo tanto la varianza va a ser un indicador de semejanza o distancia respecto a una pirámide ideal uniforme.

En las tablas siguientes destacan por su poca variabilidad países como Suecia y Bélgica ($0,13 \times 10^5$) y Grecia ($0,15 \times 10^5$), y destacan por su alta variabilidad países como Islas Marshall ($3,38 \times 10^5$), Franja de Gaza ($3,33 \times 10^5$) y Congo Kinshasa ($3,18 \times 10^5$).

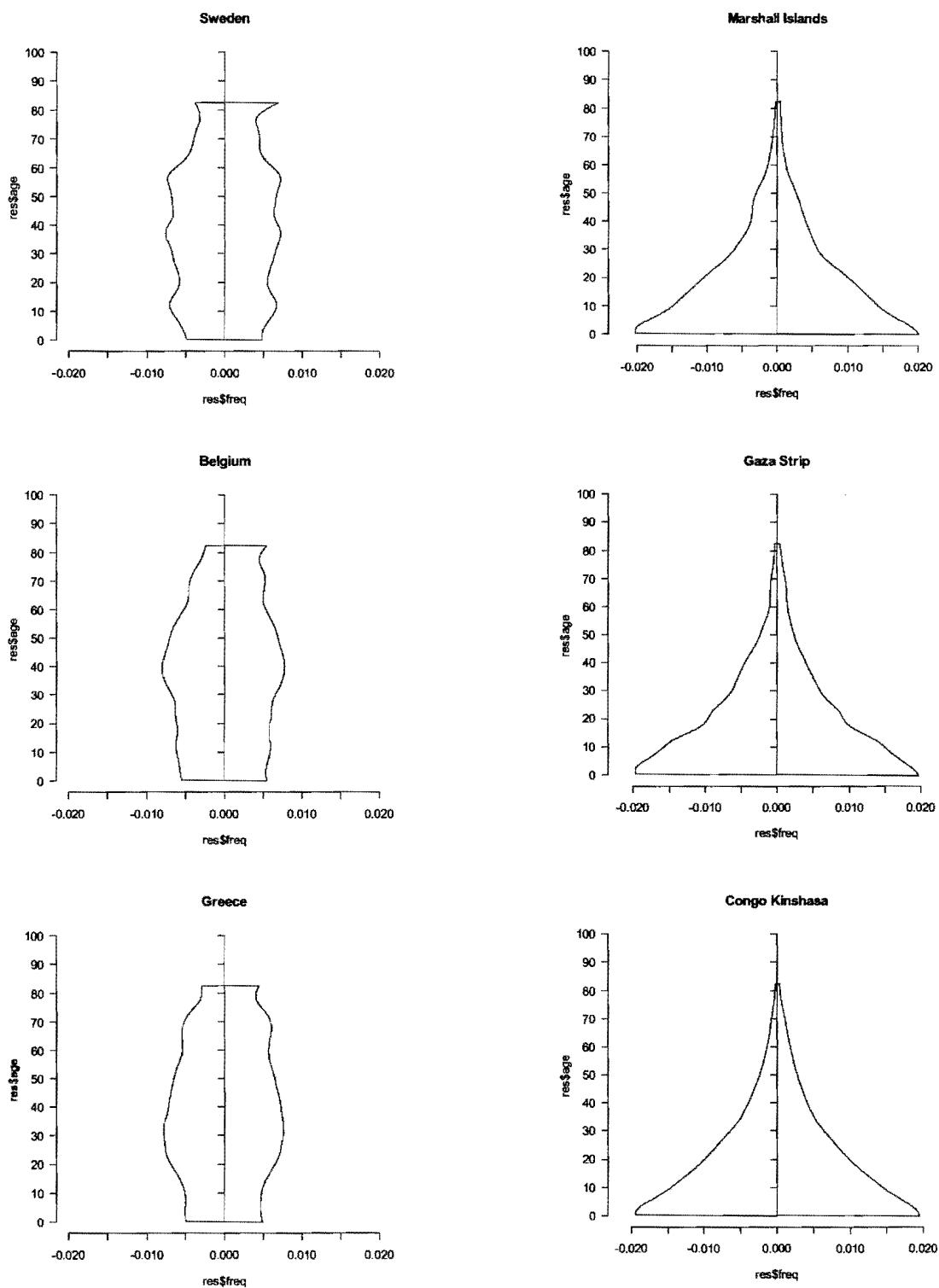
Estos resultados parecen indicar que los patrones de población en países con un comportamiento típico europeo, de América del norte o algunas islas del Caribe o del Océano pacífico, parecen acercarse a la distribución uniforme y por eso su variabilidad no supera el 1×10^5 . Países con variabilidad entre 1 y 2, a grandes rasgos, parecen corresponder a países árabes, de Asia central y de Latinoamérica desarrollados. Por último variabilidades superiores a dos corresponderían mayoritariamente a países africanos y a países poco desarrollados árabes o latinoamericanos.

Por lo tanto podemos decir que en la serie de pirámides de población el valor medio nos va a servir para comprobar la efectividad del spline, y la varianza nos dará una medida de lo uniforme que va a ser cada una de las funciones de la muestra. En las figuras siguientes podemos observar el contraste entre los dos extremos: por una parte países muy próximos a una distribución uniforme (Suecia, Bélgica y Grecia) y por otra a países con mayor variabilidad (Isla Marshall, Franja de Gaza y Congo Kinshasa).

Pais	Media (x10 ⁻⁴)	Varianza (x10 ⁻⁵)	Pais	Media (x10 ⁻⁴)	Varianza (x10 ⁻⁵)	Pais	Media (x10 ⁻⁴)	Varianza (x10 ⁻⁵)
Afghanistan	60.54	2.29	Central.African.Republic	60.39	2.44	Greece	59.462	0.148
Albania	60.26	0.91	Chad	60.52	3.08	Greenland	60.447	0.909
Algeria	60.44	1.70	Chile	60.21	0.70	Grenada	60.475	2.260
American.Samoa	60.41	1.53	China	60.32	0.77	Guadeloupe	59.951	0.722
Andorra	59.60	0.53	Colombia	60.43	1.20	Guam	60.311	1.192
Angola	60.56	2.43	Comoros	60.51	2.45	Guatemala	60.453	2.261
Anguilla	59.96	0.90	Congo..Brazzaville.	60.50	2.41	Guernsey	58.964	0.201
Antigua.and.Barbuda	60.33	1.35	Congo.Kinshasa.	60.53	3.18	Guinea	60.551	2.375
Argentina	59.86	0.55	Costa.Rica	60.31	1.16	Guinea.Bissau	60.523	2.310
Armenia	60.28	0.70	Cote.d.Ivoire	60.54	2.91	Guyana	60.317	1.218
Aruba	59.94	0.44	Croatia	59.85	0.19	Haiti	60.399	2.143
Australia	59.60	0.29	Cuba	59.72	0.58	Honduras	60.421	2.304
Austria	59.43	0.24	Cyprus	59.82	0.39	Hong.Kong.S.A.R.	59.839	0.607
Azerbaijan	60.33	1.02	Czech.Republic	59.80	0.28	Hungary	59.742	0.254
Bahamas..The	60.13	0.96	Denmark	59.29	0.20	Iceland	59.670	0.365
Bahrain	60.50	1.41	Djibouti	60.55	2.25	India	60.420	1.294
Bangladesh	60.46	1.83	Dominica	60.05	1.07	Indonesia	60.463	1.167
Barbados	59.90	0.63	Dominican.Republic	60.41	1.35	Iran	60.443	1.706
Belarus	59.94	0.34	Ecuador	60.33	1.58	Iraq	60.491	2.430
Belgium	59.40	0.15	Egypt	60.49	1.48	Ireland	59.773	0.380
Belize	60.44	2.28	El.Salvador	60.31	1.75	Man..Isle.of	58.912	0.146
Benin	60.53	3.06	Equatorial.Guinea	60.50	2.27	Israel	59.821	0.629
Bermuda	59.99	0.42	Eritrea	60.48	2.49	Italy	59.213	0.186
Bhutan	60.50	1.89	Estonia	59.80	0.27	Jamaica	60.133	1.119
Bolivia	60.31	1.81	Etiopia	60.52	2.97	Japan	59.350	0.164
Bosnia.and.Herzegovina	60.35	0.51	Faroe.Islands	59.41	0.27	Jersey	59.321	0.345
Botswana	60.35	2.27	Fiji	60.53	1.36	Jordan	60.483	1.945
Brazil	60.34	1.05	Finland	59.48	0.19	Kazakhstan	60.308	0.818
Brunei	60.49	1.38	France	59.29	0.16	Kenya	60.495	2.564
Bulgaria	59.90	0.19	French.Guiana	60.27	0.94	Kiribati	60.550	1.987
Burkina.Faso	60.52	3.06	French.Polynesia	60.39	1.07	Kuwait	60.543	1.785
Burma	60.43	1.27	Gabon	60.45	1.10	Kyrgyzstan	60.347	1.444
Burundi	60.46	2.94	Gambia..The	60.52	2.63	Laos	60.470	2.376
Cambodia	60.49	2.23	Gaza.Strip	60.50	3.33	Latvia	59.692	0.272
Cameroon	60.48	2.30	Georgia	60.14	0.36	Lebanon	60.307	1.274
Canada	59.53	0.32	Germany	59.40	0.22	Lesotho	60.405	1.901
Cape.Verde	60.15	2.20	Ghana	60.47	2.14	Liberia	60.414	2.481
Cayman.Islands	60.08	0.62	Gibraltar	59.32	0.19	Libya	60.452	1.842

Tabla 6.2: Valor medio y valor varianza de los 223 países.

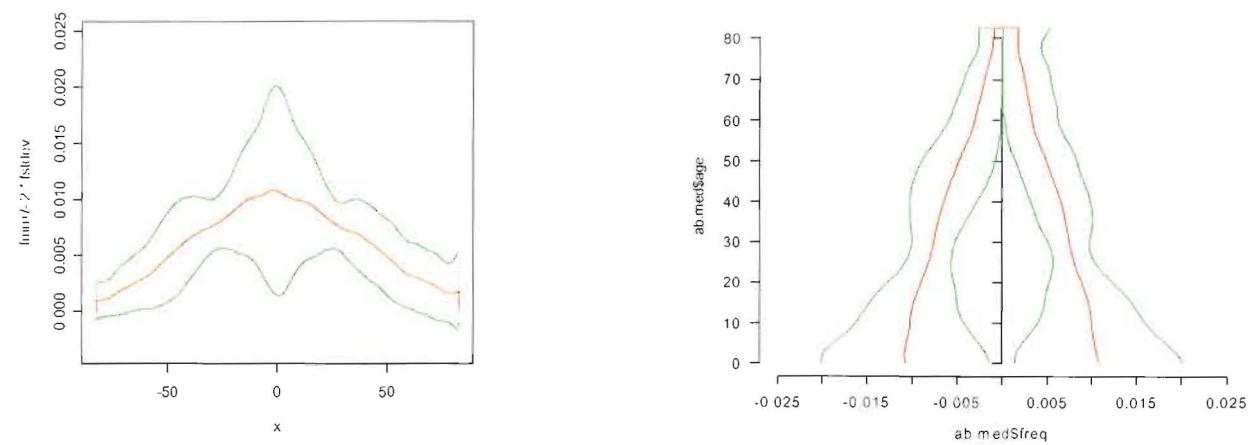
Pais	Media (x10 ⁻⁴)	Varianza (x10 ⁻⁵)	Pais	Media (x10 ⁻⁴)	Varianza (x10 ⁻⁵)	Pais	Media (x10 ⁻⁴)	Varianza (x10 ⁻⁵)
Liechtenstein	59.82	0.38	Palau	60.33	1.09	Sweden	58.88	0.13
Lithuania	59.88	0.30	Panama	60.20	1.05	Switzerland	59.27	0.25
Luxembourg	59.63	0.27	Papua.New.Guinea	60.52	1.89	Syria	60.49	2.23
Macau.S.A.R.	60.14	0.89	Paraguay	60.38	1.73	Taiwan	60.14	0.58
Macedonia..The.Former.Yugo. .Rep.of	60.23	0.44	Peru	60.38	1.40	Tajikistan	60.37	2.14
Madagascar	60.49	2.66	Philippines	60.45	1.69	Tanzania	60.48	2.72
Malawi	60.53	2.77	Poland	59.94	0.37	Thailand	60.32	0.76
Malaysia	60.44	1.37	Portugal	59.68	0.19	Togo	60.53	2.76
Maldives	60.54	2.67	Puerto.Rico	59.73	0.45	Tonga	60.51	2.15
Mali	60.53	3.00	Qatar	60.55	1.77	Trinidad.and.Tobago	60.21	0.86
Malta	59.80	0.27	Reunion	60.35	1.15	Tunisia	60.36	1.12
Marshall.Islands	60.52	3.38	Romania	60.04	0.31	Turkey	60.31	1.02
Martinique	59.82	0.63	Russia	60.02	0.40	Turkmenistan	60.45	1.81
Mauritania	60.57	2.85	Rwanda	60.51	2.57	Turks.and.Caicos.Islands	60.41	1.30
Mauritius	60.36	0.81	Saint.Helena	60.02	0.57	Tuvalu	60.47	1.31
Mayotte	60.55	2.94	Saint.Kitts.and.Nevis	59.85	0.93	Uganda	60.58	3.63
México	60.41	1.42	Saint.Lucia	60.23	1.52	Ukraine	59.96	0.28
Moldova	60.19	0.54	Saint.Pierre.and.Miquelon	59.91	0.50	United.Arab.Emirates	60.55	1.70
Monaco	58.49	0.13	Saint.Vincent.&..Grenadines	60.16	1.26	United.Kingdom	59.25	0.18
Mongolia	60.42	1.65	Samoa	60.38	1.86	United.States	59.49	0.32
Montenegro	59.89	0.35	San.Marino	59.30	0.25	Uruguay	59.71	0.35
Montserrat	59.76	0.88	Sao.Tome.and.Principe	60.43	3.08	Uzbekistan	60.40	1.68
Morocco	60.40	1.52	Saudi.Arabia	60.54	2.21	Vanuatu	60.49	1.75
Mozambique	60.53	2.45	Senegal	60.49	2.54	Venezuela	60.31	1.26
Namibia	60.40	2.47	Serbia	59.93	0.19	Vietnam	60.33	1.37
Nauru	60.59	2.11	Seychelles	60.18	1.27	Virgin.Islands	60.07	0.55
Nepal	60.51	2.06	Sierra.Leone	60.53	2.56	Virgin.Islands..British	60.31	0.96
Netherlands	59.55	0.26	Singapore	60.12	0.99	West.Bank	60.43	2.58
Netherlands.Artilles	60.20	0.60	Slovakia	59.95	0.40	Yemen	60.44	3.09
New.Caledonia	60.38	0.97	Slovenia	59.76	0.28	Zambia	60.51	3.32
New.Zealand	59.65	0.39	Solomon.Islands	60.48	2.55	Zimbabwe	60.44	2.41
Nicaragua	60.51	2.10	Somalia	60.50	2.71			
Niger	60.55	3.11	South.Africa	60.35	1.32			
Nigeria	60.54	2.51	Korea..South	60.28	0.66			
Korea..North	60.37	0.74	Spain	59.36	0.23			
Northern.Mariana.Islands	60.54	1.83	Sri.Lanka	60.28	0.81			
Norway	59.13	0.21	Sudan	60.55	2.63			
Oman	60.49	2.36	Suriname	60.38	1.23			
Pakistan	60.44	2.00	Swaziland	60.53	2.80			



Figuras 6.9 – 6.14: Representación de pirámides con varianza baja (a la izquierda) y alta (a la derecha).

6.2.2. - Análisis descriptivo de segundo nivel.

El análisis de segundo nivel corresponde básicamente a representar la función media, la función varianza (o desviación estándar) y el valor de las correlaciones. Abajo podemos ver representadas gráficamente estas funciones tanto abatidas como en formato pirámide.



Figuras 6.15 y 6.16: Representación de las funciones media y varianza a la izquierda y de las mismas funciones en forma de pirámide (a la derecha).

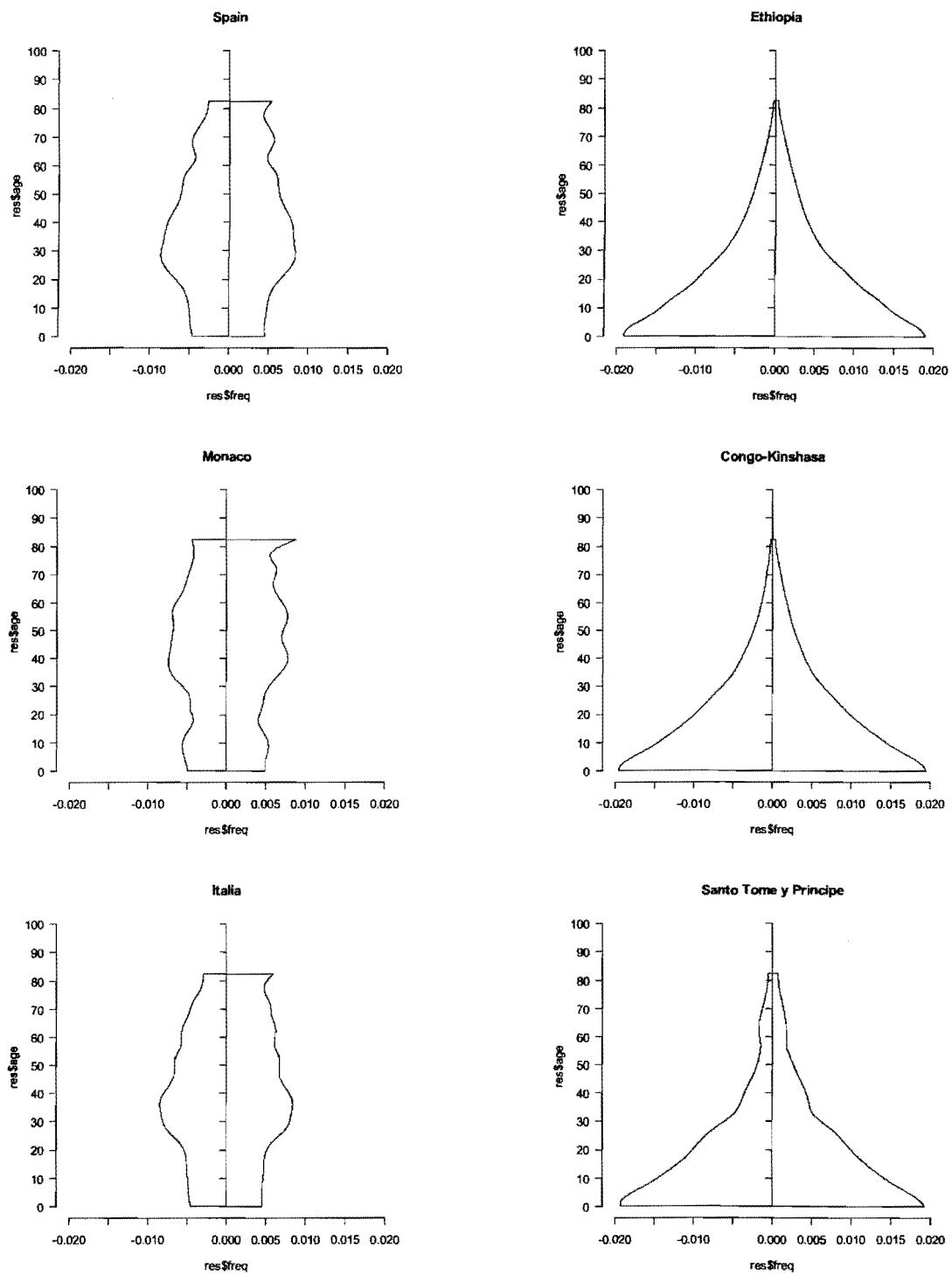
Podemos observar que la función media corresponde a una pirámide de base ancha y que, conforme avanza la edad, la densidad es menor. La variabilidad de la función media es 0.954×10^5 , semejante a la de países como Groelandia (0.909×10^5), Guayana Francesa (0.94×10^5) o Singapur (0.99×10^5).

La desviación estándar de las frecuencias va a ser distinta en función de la edad. Vemos en los gráficos anteriores que la variabilidad entre países es mayor en edades inferiores. Conforme aumentamos en edad, ésta disminuye hasta aproximadamente los treinta años; edad en que la variabilidad es menor. Luego vuelve a crecer hasta los 50 años de edad en las mujeres y los 40 en los hombres. A continuación se mantiene constante en hombres y sin embargo decrece ligeramente en las mujeres hasta las edades más avanzadas. Por lo tanto, las edades que hacen distinguir más unos países de otros son las inferiores a 20 y la franja entre 40 y 50 años, distinguiendo entre mujeres (izquierda del gráfico) y hombres (derecha del gráfico).

Otro análisis de segundo nivel es el de las covarianzas o correlaciones dos a dos. En este caso no podemos representar en una tabla dichas correlaciones ya que en total son 24753 cifras. Sin embargo podemos destacar algunas parejas de países según su valor de correlación:

Países	R	Países	R
España – Etiopía	-2.22×10^{-4}	Italy - Sao.Tome.and.Príncipe	-0.255
Alemania - Tajikistán	8.18×10^{-5}	Madagascar - Mauritania	0.9998
Mónaco - Sao.Tome.and.Príncipe	-0.471	Congo Kinshasa - Etiopía	0.9996

Tabla 6.3: Ejemplo de correlaciones entre parejas de países.

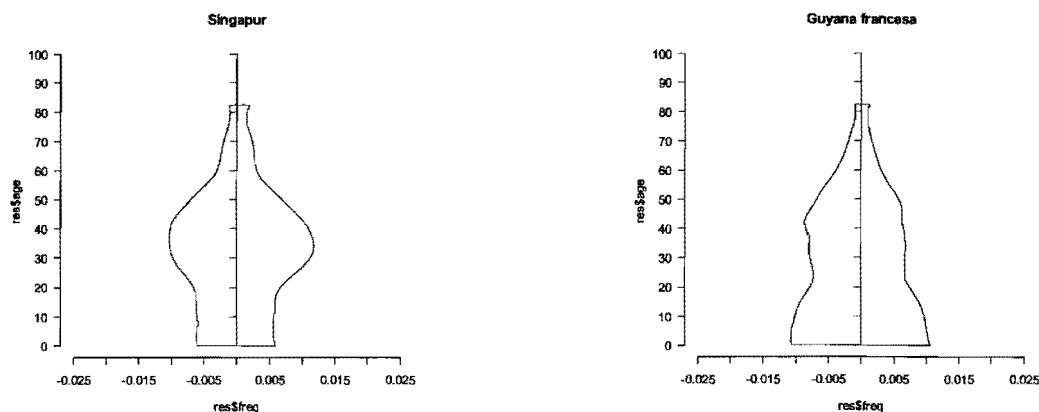


Figuras 6.17-6.22: Pirámides de algunas de las parejas referidas en la tabla 6.3.

En general países de las mismas regiones o de semejantes condiciones socioeconómicas se correlacionan positivamente. Destaca Mónaco por tener correlación negativa con muchos países africanos y de América Latina, posiblemente debido a que es una población con una frecuencia relativa muy grande en edades altas. España y Etiopía tienen una correlación prácticamente cero debido a que España tiene poca área entre su función y la función uniforme (o también varianza), y además en las edades donde España tiene algo de área es donde Etiopía tiene frecuencias relativas próximas a la uniforme. Las pirámides de Congo-Kinshasa y de Etiopía son prácticamente idénticas y por eso su correlación es igual a uno. Mónaco y Santo Tomé y Príncipe destacan por su correlación negativa ya que el primero tiene poca población en edades bajas y mucha en altas, que es precisamente lo contrario a lo que pasa en el segundo país.

6.2.3. - Críticas sobre el uso de la descriptiva funcional.

Podemos considerar que la descriptiva funcional nos permite evaluar con un solo valor (su varianza) cual es el comportamiento global de la pirámide, y en cuánto se acercan a la pirámide uniforme (valor medio). Los resultados nos sugieren una clasificación o ranking de países según su varianza. Además la correlación nos permite ver en cuánto se parecen unas funciones dos a dos, en caso de tener valores o cerca de cero o cerca de uno. En principio nos podría parecer que las formas de las pirámides para países con varianzas semejantes habrían de ser semejantes y que la correlación nos indicaría en qué se diferencian. Sin embargo esto no es del todo cierto ya que tanto la varianza como la correlación no tienen en cuenta el otro eje, la edad. Nos podemos encontrar con países de muy parecida varianza y tener una estructura por edades desigual. En todo caso la varianza nos indicará que la magnitud de la desigualdad de la forma de un país es igual a la magnitud de la desigualdad del otro. No podemos decir nada más. Esto queda de relieve en las siguientes pirámides.



Figuras 6.23 y 6.24: Pirámides de países con valor varianza muy parecido pero de formas diferentes.

La varianza de Singapur y la Guyana francesa es de 0.99×10^{-5} y 0.94×10^{-5} respectivamente, y su correlación es de 0.668. Sin embargo vemos que sus formas no son del todo similares.

En conclusión, si bien el análisis descriptivo nos ayuda a dar un resumen global de la forma de las funciones, si queremos tener en cuenta el eje de las ordenadas necesitaremos otra herramienta. La solución nos la ofrece el análisis de componentes principales funcionales (ACPF).

6.3. - ANÁLISIS DE DATOS FUNCIONALES DE LAS PIRÁMIDES.

La motivación que nos conduce a aplicar el análisis de componentes principales funcionales es la de buscar unas pautas de comportamiento de las frecuencias por edades de las pirámides de población. Es decir, primeramente buscaremos unas funciones (componentes principales funcionales o CPF's) que resuman el comportamiento de la variabilidad de las funciones. Una vez encontradas calcularemos los coeficientes que nos indicarán si la forma de la pirámide de un país en concreto se explica más por una o por otra pauta de variabilidad.

6.3.1. – Cálculo, interpretación y nivel de representación de las CPF's.

La función en R que calculará las componentes principales funcionales para splines (tanto para spline de tipo interpolador o de regresión) es *fpcasint()*. Antes de aplicar esta función habremos registrado los splines y posteriormente los habremos centrado (es decir, a todas las funciones le restamos la función media). El resultado de este algoritmo es una lista de vectores propios, valores propios, matriz de covarianzas y matriz de correlaciones entre las funciones de la muestra y las funciones componentes principales (véase punto 4.3.1.). Nos va a interesar interpretar las CPF's.

Recordando el esquema de las componentes principales clásicas,

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \text{ matriz } n \times p \text{ y } (CP_1, CP_2, \dots, CP_p) = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,p} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,p} \\ \vdots & \vdots & & \vdots \\ \alpha_{p,1} & \alpha_{p,2} & \dots & \alpha_{p,p} \end{pmatrix} (X_1, X_2, \dots, X_p),$$

y por lo tanto podíamos expresar la componente principal i-ésima en función de las variables;

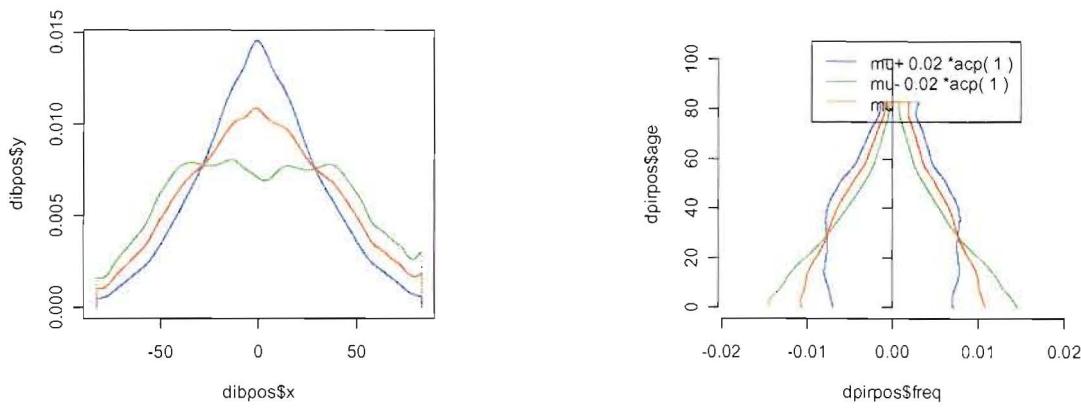
$$CP_i = \alpha_{i,1}X_1 + \alpha_{i,2}X_2 + \dots + \alpha_{i,p}X_p.$$

Podríamos decir que en ACPF X es una matriz de infinitas componentes y por extensión A también lo es, teniendo así un real para cada valor de X_i . En este caso X_1, \dots, X_n pasaría de ser sucesión a ser función. Por

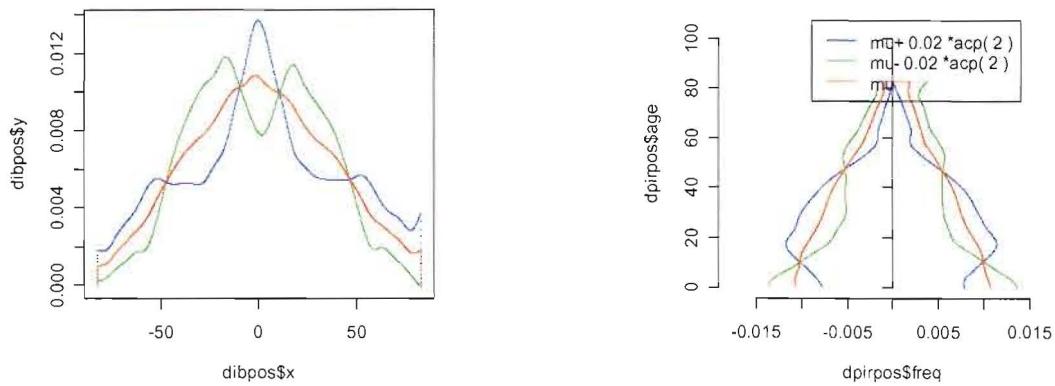
tanto a diferencia de las componentes principales clásicas las CPF's no son vectores sino que ahora son funciones. Numéricamente las CPF's tienen representación de spline. Por lo tanto lo mejor para interpretar cuál es su significado es representarla gráficamente. El representar gráficamente una CPF nos da una pista poco útil. Sin embargo representar la función media +/- la CPF nos va a aportar la información decisiva, ya que veremos cómo afecta sumar o restar la CPF a la media.

La subrutina en R que nos permite visualizar este gráfico es `fpa.fmuplot()`. Tendrá como parámetros de entrada el objeto salida de `fPCA.sint()`, el spline función media (registrado) y otros parámetros. Esta subrutina también tiene incorporado una opción de representación en formato pirámide. Si lo aplicamos a nuestras pirámides de población obtenemos lo siguiente:

- CPF(1)



- CPF(2)



Figuras 6.25 –6.28: Representación de la función media +/- CPF's primera y segunda como función (izquierda) y como pirámide (derecha). Esta representación nos permite interpretar las pautas de variabilidad explicadas por las CPF's.

Estudiemos el caso de país con coeficiente grande y positivo en la primera CPF: respecto la pirámide media en edades jóvenes tendrá frecuencias inferiores, a los 30 años tendrá la misma frecuencia y en edades superiores tendrá más frecuencia respecto la pirámide media. Es decir, tendrá forma de ánfora.

Un país con coeficiente grande y negativo en la primera CPF: respecto la pirámide media, en edades jóvenes tendrá más frecuencia y en edades más allá de los treinta la frecuencia será menor. Podemos decir que su forma es de embudo.

Un país con coeficiente grande y positivo en la segunda CPF: Respecto la pirámide media, las frecuencias bajas en edades entre cero y diez tendrá menor frecuencia que la media, entre diez y cincuenta tendrá más y de cincuenta en adelante tendrá menos frecuencia.

Acabamos de interpretar las dos primeras componentes principales: Pero, podemos explicar todas las formas de pirámides sólo con estas dos componentes? Necesitamos saber el nivel de representatividad de las dos primeras componentes principales. Esto lo extraemos de los valores propios. Al igual que en las componentes principales clásicas la varianza total es igual a la suma de valores propios. Por lo tanto si sumamos los dos primeros valores propios, y los dividimos entre la suma total obtendremos la proporción de representatividad de las dos primeras componentes principales. A través de la función `fPCA.sint.plot()` además obtenemos un gráfico.

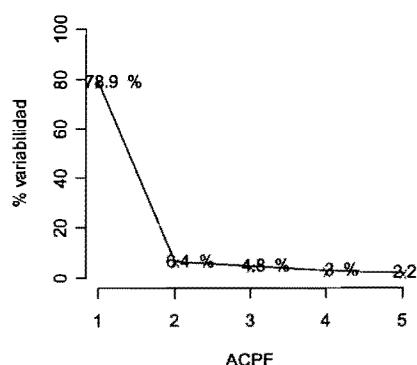


Figura 6.29: Gráfico de representatividad de las primeras 5 CPFs. Vemos que con las dos primeras CPFs obtenemos 85.3 %, cifra más que aceptable.

PAIS	CPF(1)	CPF(2)	PAIS	CPF(1)	CPF(2)	PAIS	CPF(1)	CPF(2)
Afghanistan	-0.983	-0.118	Cayman.Islands	0.852	-0.184	Ghana	-0.961	0.112
Albania	0.326	0.298	Central.African.Republic	-0.986	-0.098	Gibraltar	0.935	-0.256
Algeria	-0.696	0.635	Chad	-0.942	-0.297	Greece	0.955	-0.145
American.Samoa	-0.738	-0.287	Chile	0.928	0.127	Greenland	0.651	0.008
Andorra	0.947	-0.026	China	0.788	0.327	Grenada	-0.711	0.580
Angola	-0.958	-0.206	Colombia	-0.244	0.528	Guadeloupe	0.779	0.265
Anguilla	0.724	0.578	Comoros	-0.957	-0.139	Guam	-0.209	-0.612
Antigua.and.Barbuda	0.333	0.409	Congo.Brazzaville.	-0.989	-0.070	Guatemala	-0.990	-0.109
Argentina	0.859	-0.186	Congo.Kinshasa.	-0.969	-0.233	Guernsey	0.959	-0.207
Armenia	0.730	0.261	Costa.Rica	-0.172	0.883	Guinea	-0.983	-0.148
Aruba	0.923	-0.170	Cote.d.Ivoire	-0.988	-0.117	Guinea.Bissau	-0.968	-0.139
Australia	0.980	-0.155	Croatia	0.931	-0.277	Guyana	0.128	0.875
Austria	0.969	-0.153	Cuba	0.894	0.141	Haiti	-0.919	0.083
Azerbaijan	0.376	0.656	Cyprus	0.944	-0.102	Honduras	-0.988	-0.033
Bahamas.The	0.544	0.612	Czech.Republic	0.943	-0.096	Hong.Kong.S.A.R.	0.928	0.043
Bahrain	0.283	0.249	Denmark	0.934	-0.310	Hungary	0.941	-0.124
Bangladesh	-0.649	0.539	Djibouti	-0.908	-0.334	Iceland	0.965	-0.162
Barbados	0.928	0.221	Dominica	0.308	0.549	India	-0.737	0.518
Belarus	0.947	0.034	Dominican.Republic	-0.902	0.343	Indonesia	-0.046	0.786
Belgium	0.959	-0.249	Ecuador	-0.934	0.283	Iran	-0.395	0.707
Belize	-0.989	0.006	Egypt	-0.839	0.452	Iraq	-0.978	0.088
Benin	-0.977	-0.185	El.Salvador	-0.951	0.002	Ireland	0.964	-0.016
Bermuda	0.933	-0.191	Equatorial.Guinea	-0.973	-0.178	Man..Isle.of	0.938	-0.265
Bhutan	-0.962	-0.195	Eritrea	-0.936	-0.203	Israel	0.856	-0.173
Bolivia	-0.949	0.084	Estonia	0.941	-0.047	Italy	0.959	-0.180
Bosnia.and.Herzegovina	0.948	0.081	Ethiopia	-0.964	-0.253	Jamaica	0.090	0.839
Botswana	-0.931	0.136	Faroe.Islands	0.875	-0.271	Japan	0.921	-0.262
Brazil	0.338	0.840	Fiji	-0.632	0.500	Jersey	0.927	-0.249
Brunei	0.076	0.629	Finland	0.937	-0.301	Jordan	-0.816	0.403
Bulgaria	0.940	-0.141	France	0.955	-0.244	Kazakhstan	0.661	0.456
Burkina.Faso	-0.973	-0.208	French.Guiana	0.409	-0.215	Kenya	-0.936	0.200
Burma	-0.024	0.934	French.Polynesia	0.287	0.840	Kiribati	-0.955	-0.092
Burundi	-0.977	-0.153	Gabon	-0.399	-0.482	Kuwait	0.033	0.495
Cambodia	-0.946	-0.032	Gambia.The	-0.961	-0.245	Kyrgyzstan	-0.792	0.309
Cameroon	-0.991	-0.104	Gaza.Strip	-0.962	-0.236	Laos	-0.987	-0.116
Canada	0.980	-0.121	Georgia	0.930	0.080	Latvia	0.936	-0.059
Cape.Verde	-0.853	-0.054	Germany	0.956	-0.201	Lebanon	0.045	0.580

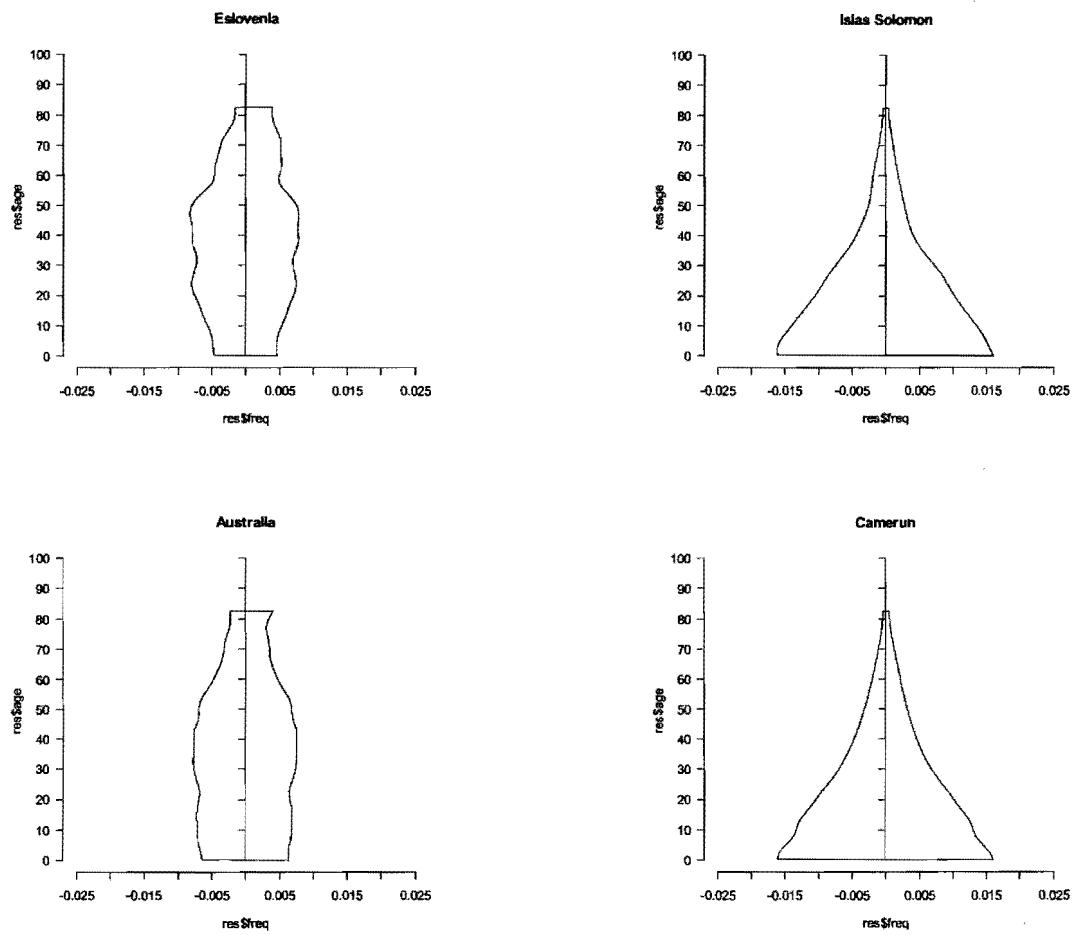
Tabla 6.4: Scores de la primera y segunda CPFs para cada uno de los países

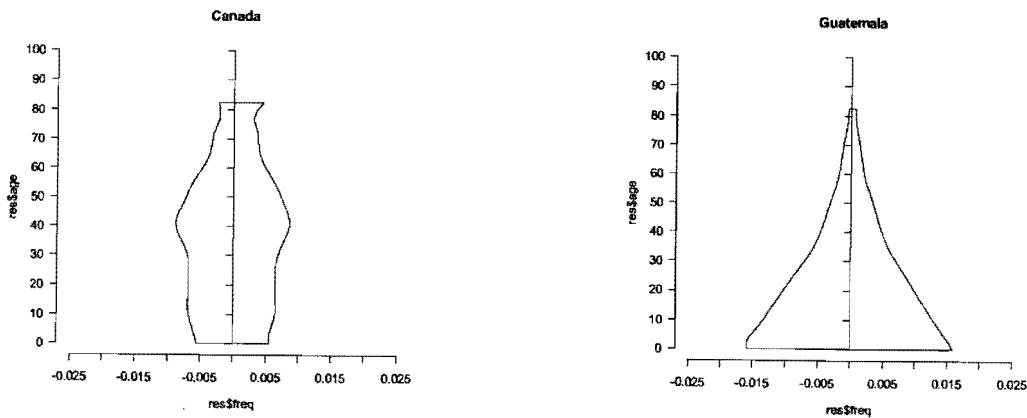
PAIS	CPF(1)	CPF(2)	PAIS	CPF(1)	CPF(2)	PAIS	CPF(1)	CPF(2)
Lesotho	-0.980	-0.021	Northern.Mariana.Islands	0.410	0.476	Spain	0.964	-0.054
Liberia	-0.921	-0.296	Norway	0.941	-0.279	Sri.Lanka	0.767	0.506
Libya	-0.785	0.484	Oman	-0.715	-0.243	Sudan	-0.986	-0.121
Liechtenstein	0.960	-0.166	Pakistan	-0.984	-0.002	Suriname	-0.162	0.703
Lithuania	0.957	-0.010	Palau	0.550	0.209	Swaziland	-0.986	-0.124
Luxembourg	0.961	-0.214	Panama	0.163	0.877	Sweden	0.925	-0.291
Macau.S.A.R.	0.782	0.311	Papua.New.Guinea	-0.965	-0.021	Switzerland	0.965	-0.215
Macedonia.The.Former.Yugo.	0.963	0.044	Paraguay	-0.964	-0.207	Syria	-0.955	0.229
Madagascar	-0.962	-0.232	Peru	-0.872	0.410	Taiwan	0.924	0.213
Malawi	-0.975	0.007	Philippines	-0.966	0.168	Tajikistan	-0.962	0.069
Malaysia	-0.867	0.180	Poland	0.932	0.010	Tanzania	-0.985	-0.125
Maldives	-0.970	-0.158	Portugal	0.954	-0.125	Thailand	0.830	0.384
Mali	-0.950	-0.286	Puerto.Rico	0.929	-0.049	Togo	-0.988	-0.083
Malta	0.926	-0.234	Qatar	0.380	0.126	Tonga	-0.857	0.232
Marshall.Islands	-0.967	-0.227	Reunion	-0.062	0.287	Trinidad.and.Tobago	0.684	0.497
Martinique	0.829	0.166	Romania	0.927	0.015	Tunisia	0.166	0.909
Mauritania	-0.967	-0.214	Russia	0.931	0.025	Turkey	0.359	0.884
Mauritius	0.801	0.306	Rwanda	-0.950	0.134	Turkmenistan	-0.932	0.235
Mayotte	-0.901	-0.211	Saint.Helena	0.902	0.108	Turks.and.Caicos.Isla.	0.006	-0.067
Mexico	-0.674	0.602	Saint.Kitts.and.Nevis	0.360	0.296	Tuvalu	-0.210	0.271
Moldova	0.857	0.145	Saint.Lucia	-0.460	0.758	Uganda	-0.969	-0.228
Monaco	0.887	-0.415	Saint.Pierre.and.Miquelon	0.831	-0.246	Ukraine	0.937	-0.063
Mongolia	-0.411	0.842	St.Vincent.and.Grenadines	-0.096	0.873	United.Arab.Emirates	0.248	-0.031
Montenegro	0.957	-0.058	Samoa	-0.131	0.496	United.Kingdom	0.955	-0.242
Montserrat	0.373	0.513	San.Marino	0.951	-0.200	United.States	0.958	-0.194
Morocco	-0.799	0.550	Sao.Tome.and.Principe	-0.959	-0.255	Uruguay	0.873	-0.287
Mozambique	-0.987	-0.029	Saudi.Arabia	-0.777	-0.353	Uzbekistan	-0.814	0.436
Namibia	-0.986	-0.030	Senegal	-0.986	-0.107	Vanuatu	-0.863	0.439
Nauru	-0.889	0.088	Serbia	0.914	-0.194	Venezuela	-0.401	0.817
Nepal	-0.986	-0.046	Seychelles	0.169	0.808	Vietnam	-0.432	0.809
Netherlands	0.967	-0.216	Sierra.Leone	-0.937	-0.287	Virgin.Islands	0.639	-0.379
Netherlands.Antilles	0.905	-0.054	Singapore	0.829	0.302	Virgin.Islands..British	0.747	0.505
New.Caledonia	0.319	0.303	Slovakia	0.952	0.083	West.Bank	-0.970	-0.150
New.Zealand	0.943	-0.066	Slovenia	0.982	-0.050	Yemen	-0.970	-0.190
Nicaragua	-0.945	0.274	Solomon.Islands	-0.991	-0.053	Zambia	-0.986	-0.047
Niger	-0.948	-0.278	Somalia	-0.931	-0.165	Zimbabwe	-0.776	0.388
Nigeria	-0.980	-0.167	South.Africa	-0.436	0.754			
Korea..North	0.764	0.168	Korea..South	0.905	0.208			

6.3.2 - Resultados y comentarios.

El último paso consiste en calcular los coeficientes para cada uno de los países con el fin de entender la forma de los mismos y relacionarlos con la forma de países semejantes. La tabla 6.4 muestra los coeficientes de los dos CPF's para cada uno de los países. Gráficamente se representan los dos primeros scores en un plano, como se muestra en la figura 6.41.

Si nos fijamos en la tabla observamos que los países con la CPF(1) mayor son Eslovenia (0,982), Australia y Canadá (ambos 0.980). Mientras que los que tienen una CPF(1) menor son las Islas Solomon, Camerún(ambos con -0.991) y Guatemala (-0.990).

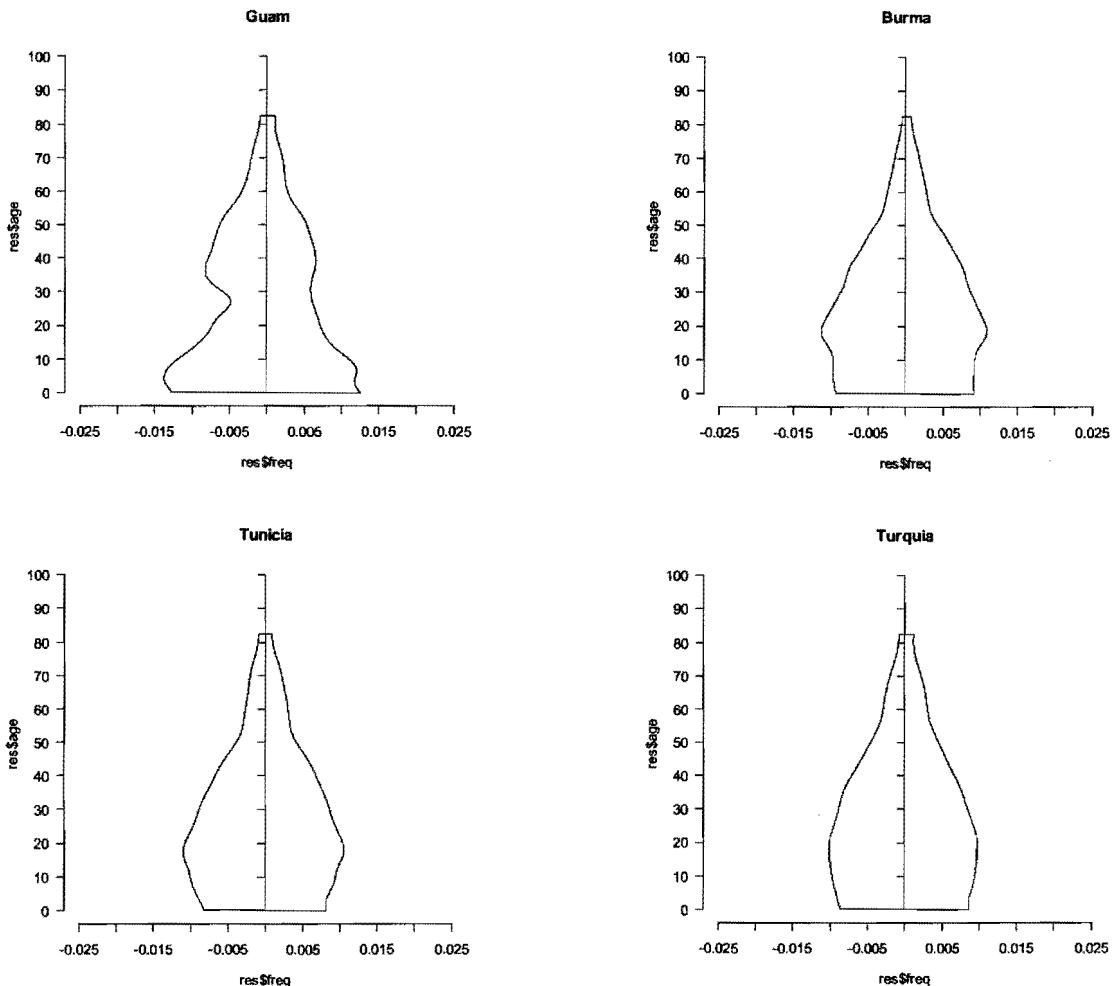




Figuras 6.30-6.35: Representación de pirámides de países con un score de la primera CPF próximo a +1 (a la izquierda) o a -1 (a la derecha).

A la vista de estos gráficos, y teniendo en cuenta la figura 6.26 de la función media +/- $\text{CPF}(1)$, observamos que los tres gráficos de la izquierda se asemejan mucho a la línea de función media más $\text{CPF}(1)$, mientras que los tres gráficos de la derecha se asemejan mucho a la que resta $\text{CPF}(1)$. Existen aproximadamente 60 países con un coeficiente $\text{CPF}(1)$ de 0.90 o superior, lo que quiere decir que estos 60 países tienen como característica principal la forma de ánfora. De manera general encontramos países europeos (Reino Unido, Francia, España, Italia, Noruega, Suecia, Hungría, Bulgaria, etc), norteamericanos (Estados Unidos y Canadá) y países oceánicos (Australia y Nueva Zelanda). Existen también 70 países con coeficiente $\text{CPF}(1)$ inferior a -0.90, lo que significa que su característica principal va a ser la forma de embudo. A grandes rasgos esta forma es característica del centro y sur de África (los Congos, Namibia, Mozambique, etc...), algunos países de Centroamérica (Honduras, El Salvador, Nicaragua, República Dominicana, Haití, etc..) o Sudamérica (Ecuador, Bolivia, Paraguay), y algunos países del Asia central (Nepal, Pakistán, Afganistán, Irak, Yemen, Filipinas), entre otros.

En cuanto a los coeficientes de los países respecto la $\text{CPF}(2)$, observamos que no hay muchos países que estén fuertemente correlacionados. Destaca Guam (-0.612) y con correlaciones positivas destacan Birmania (Burma) (0.934) (con una excesiva mortalidad según publica la CIA en www.cia.gov), Túnez (0.909) y Turquía (0.884).



Figuras 6.36-6.39: Pirámides de población con la segunda CPF próxima a +1 (Birmania, Túnez y Turquía) o negativa (Guam).

La figura 6.40 nos muestra que la nube de puntos de los 223 países tiene forma de media luna. Empezando por la derecha (primera CPF próxima a +1) encontramos países como Suecia, Serbia, Canadá, Australia, España, etc de pirámides tipo ánfora. Conforme seguimos la media luna (incremento de la CPF(2) y disminución de la CPF(1)) nos encontramos a China y Palau. Los países con mayor CPF(2) (Brazil, Turquía, Túnez y Birmania) suelen tener poca correlación con la CPF(1). Avanzando más (CPF(1) de -0.6 y CPF(2) de 0.5) nos encontramos países como India, Perú o Bangladesh. Finalmente en el extremo izquierdo nos encontramos muchos países con CPF(1) cercana a -1 de claro patrón con forma de embudo como Yemen, Camerún, Guatemala, Congo-Kinshasa e Islas Salomón.

Representación de los scores de las dos primeras CPFs

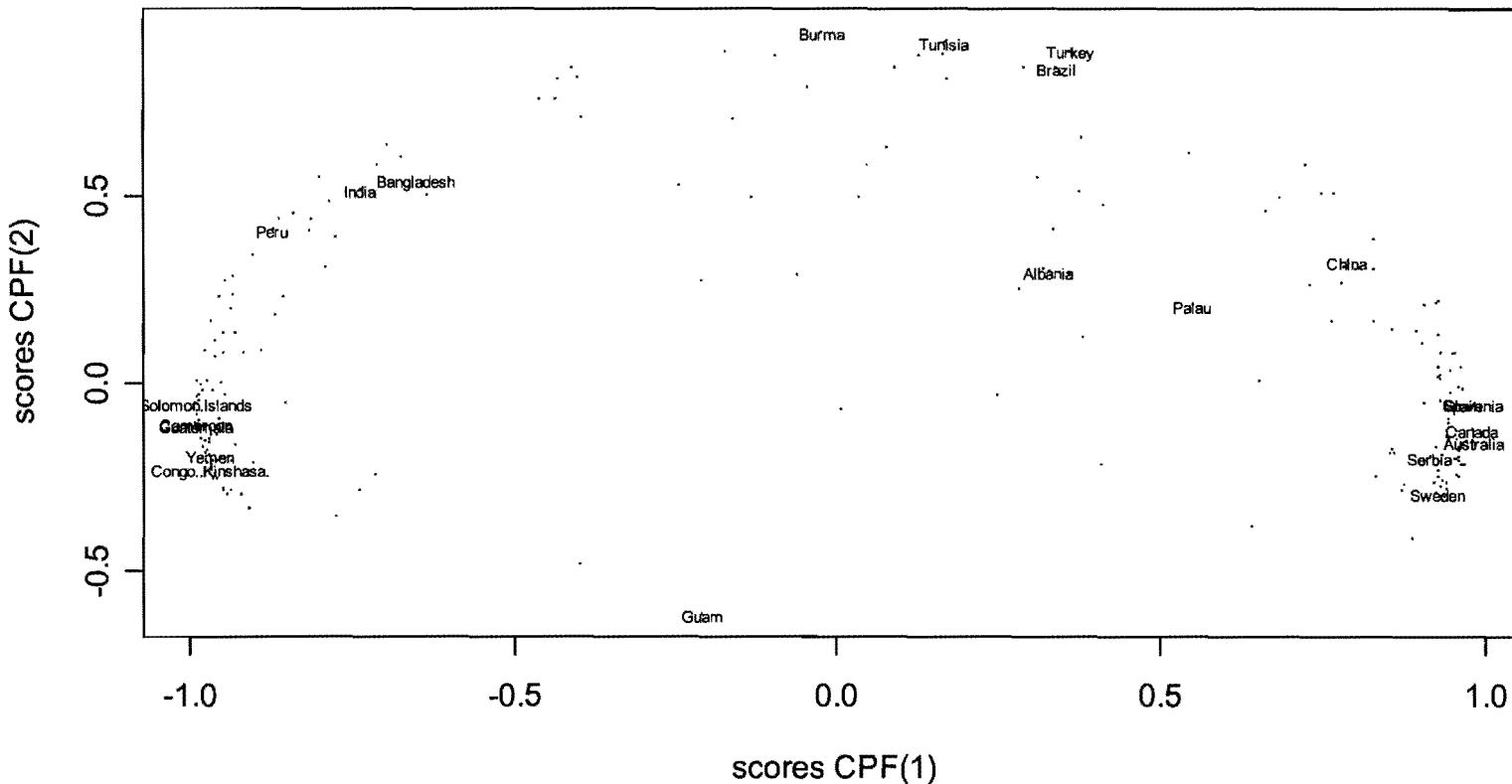


Figura 6.40: Representación de los scores de las dos primeras CPFs.

6.4.- Críticas al trabajo.

- El análisis de datos funcionales permite tratar las pirámides como un objeto compacto, hecho que permite comparar países de manera simple. Además nos permite ver países que están próximos en su forma de una manera gráfica e intuitiva. Este tipo de análisis aporta una gran luz a la descripción general de las formas de las pirámides ya que compara todas las formas entre sí, resumiéndolas de manera sencilla, o bien situándolas como un punto en el espacio, o bien con medidas de varianza y correlación con un sólo valor.
- El análisis descriptivo funcional nos aporta mucha información, aunque cabe advertir que no considera el eje de ordenadas. A diferencia del análisis descriptivo, con el ACPF sí que detectamos las diferentes pautas de comportamiento en función del eje de ordenadas, que en este caso es la edad. El ACPF es un buen complemento a la información que ya nos ofrece la descriptiva funcional.
- Como crítica negativa, en este trabajo se ha considerado que la pirámide abatida es una función densidad continua. Esto no es del todo cierto ya que en el cero seguramente tenemos problemas de discontinuidad. De todas maneras considerar la edad como variable continua es desde un punto de vista teórico más preciso.
- También se ha utilizado en las categorías *80 años y más* la marca de clase 82,5, hecho el cual es discutible. Esto se ha hecho así porque disponíamos de una información limitada.

7. - ANÁLISIS FUNCIONAL DE DATOS EN DATOS FARMACOCINÉTICOS.

7.1. – PLANTEAMIENTO DEL PROBLEMA.

7.1.1.- Ensayos fase I en oncología.

Los ensayos clínicos fase I en oncología tienen como objetivo encontrar una dosis máxima tolerada en medicamentos o combinaciones de medicamentos aún no probados en humanos. A diferencia del resto de fármacos, en quimioterapia se parte de la idea de que cuanta más dosis se pueda administrar, mayor es la eficacia del tratamiento. Pero en todo medicamento existe un límite de tolerancia que no se puede sobreponer. Ese límite lo delimitan los efectos secundarios no deseados del fármaco: la toxicidad. Para estos ensayos se definen dos conceptos. El primer concepto es el de *toxicidad limitante de dosis* (TLD), que es aquella inaceptable o excesiva, dependiendo del fármaco y del estudio. El segundo concepto es el de *dosis máxima tolerada* (DMT), que es aquella dosis la cual no se puede superar sin que exista una probabilidad indeseablemente alta de que se produzca una toxicidad limitante de dosis. Existen varios diseños que buscan encontrar esta dosis. El más simple es el algoritmo siguiente:

Se busca una escalada de dosis por el método de fibonacci (basada en la sucesión de Fibonacci y partiendo desde una dosis inicial una décima parte de la DMT encontrada en animales): D_1 , D_2, \dots, D_n tal que $D_1 < D_2 < \dots < D_n$, $D_{i+2} = D_{i+1} + D_i$

Iniciar:

$$D_i = D_1$$

Repetir:

- Se incluyen 3 pacientes de manera secuencial y se observan con D_i ;
- Se calcula la proporción de TLD en cada inclusión.
- Si la proporción de TLD es igual a cero al fin de las tres inclusiones, se pasa a la dosis superior: $D_i = D_{i+1}$.
- Si la proporción de TLD es igual a 33 % se observan 3 pacientes más en la misma dosis.

- Si la proporción de TLD es mayor al 33% con 3 pacientes o bien es mayor o igual a 33% con 6 pacientes D_i será dosis no tolerada.

hasta D_i no tolerada.

FIN.

D_{t-1} será la dosis máxima tolerada recomendada para estudios más avanzados (fase II).

Otro objetivo del ensayo es calcular parámetros farmacocinéticos del medicamento, que miden cómo varía la concentración del medicamento en la sangre a lo largo del tiempo. Esto es importante ya que normalmente, tanto el nivel de concentración máximo como el área bajo la curva, suelen estar relacionados tanto con la eficacia del tratamiento como con la toxicidad, y en especial con la toxicidad limitante de dosis.

Actualmente el método más utilizado para calcular estos parámetros es el denominado método del trapecio. Este método no es más que unir los puntos de concentraciones observados a lo largo del tiempo. El máximo de concentración será el máximo observado, y el área encerrada por la poligonal será la concentración total. Para el punto de cruce de la curva con el eje de las ordenadas se utiliza una extrapolación lineal hasta encontrarse con el eje de ordenadas.

EJEMPLO 6.1:

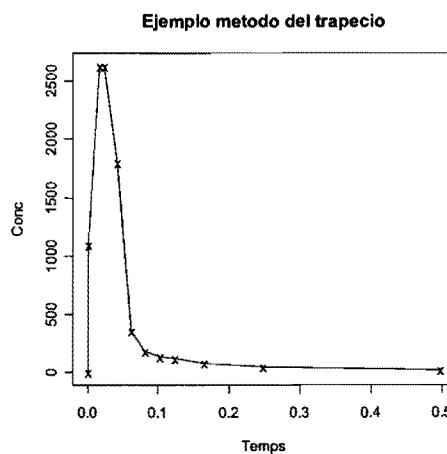


Figura 7.1: Concentración de docetaxel por el método del trapecio del paciente 1

Este método es discutible por dos motivos: En primer lugar el valor máximo de concentración va a depender de los tiempos que se haya decidido para las tomas de muestra y estará infraestimado en caso de no obtener la muestra justo cuando se obtenga el máximo real. En

segundo lugar el área real va a ser estimada con sesgo por suponer que la función entre punto y punto es recta y no curva.

Dado que la concentración es una función del tiempo parece apropiado aplicar el planteamiento del análisis de datos funcionales. También parece más apropiado aplicar splines interpoladores o de regresión como consecuencia de lo comentado en el punto 3.2. Sabemos que los datos pueden estar sujetos a error, por lo que es preferible utilizar splines de regresión. Veremos un ejemplo de las consecuencias de aplicar el spline interpolador en vez de splines de regresión.

7.1.2.- Situación y objetivos.

En este ensayo se incluyeron 20 pacientes de los cuales sólo fueron evaluables 19 debido a que de uno no se tienen recogidos los datos suficientes para el análisis. El tratamiento consiste en administrar vía intravenosa (una infusión), una combinación de fármacos de manera secuencial. El tratamiento total son 6 infusiones administradas una vez cada 21 días (duración de un ciclo), aunque para el estudio de farmacocinética sólo estudiaremos las dos primeras. La particularidad es que en el primer ciclo a unos pacientes se les administra inicialmente el esquema PD (Paclitaxel seguido de Docetaxel) y a otros el esquema DP (la secuencia inversa). En el segundo ciclo, a los que en el primero se les había practicado la secuencia PD se les administra el tratamiento con la secuencia DP y viceversa. Es decir:

	PRIMER CICLO		SEGUNDO CICLO	
	Grupo	Tratamiento	Grupo	Tratamiento
Individuo que inicia con esquema PD	PD	Paclitaxel en 3h seguido de docetaxel en 1h.	DP	Docetaxel en 1h seguido de Paclitaxel en 3h
Individuo que inicia con esquema DP	DP	Docetaxel en 1h seguido de Paclitaxel en 3h.	PD	Paclitaxel en 3h seguido de docetaxel en 1h.

Tabla 7.1: Esquema de tratamientos.

Vamos a suponer que las variaciones de concentración no van a depender de si en el primer ciclo inicia con PD o con DP. Es decir, esperaremos que el comportamiento de las

concentraciones de un esquema PD o DP es el mismo se produzca en el primer ciclo o en el segundo.

Nuestro objetivo será inicialmente describir cómo varían las concentraciones de docetaxel o paclitaxel según sea el esquema PD o DP. Para esto calcularemos las funciones media y varianza para:

Docetaxel en esquema PD Paclitaxel en esquema PD

Docetaxel en esquema DP Paclitaxel en esquema DP.

Posteriormente se unificarán las muestras de Docetaxel en esquema PD y DP y se realizará un ACPF para ver si existen diferentes pautas de comportamiento. De manera análoga se hará con Paclitaxel.

Antes de calcular las funciones de concentraciones hemos de tener en cuenta de que la dosis es distinta para cada paciente, ya que éstos han entrado en niveles de dosis distintos. El cálculo de la dosis administrada depende de la dosis del nivel, expresada en mg/m², y de la superficie corporal, que depende del peso y la altura. Por lo tanto previamente habremos de calcular la concentración ajustada por la dosis. Es decir:

$$C = \frac{C_{\text{sin ajustar}}}{\text{Dosis} / S.C}$$

NOTACIÓN:

C_{ijk} = Concentración ajustada del fármaco i en el grupo de tratamiento j e individuo k,

Donde i={P, D} (P =Paclitaxel, D =Docetaxel),

j={PD, DP} (PD =Paclitaxel antes de Docetaxel, DP =Docetaxel antes de Paclitaxel)

k={1,..,19} individuos de la muestra.

7.2.- REGISTRO DE DATOS.

Aprovecharemos los datos de concentraciones de Docetaxel y Paclitaxel para ilustrar los pasos y los problemas que plantea realizar una mala etapa de registro. Posteriormente realizaremos el análisis de manera correcta.

Por último, transformar los datos puede resaltar aquellos intervalos en donde una función tiene más variabilidad, y que tal vez por escala no se aprecian en representaciones sin transformar.

7.2.1. – Problemas en la elección del tipo de spline.

En primer lugar hemos de decidir qué tipo de spline utilizamos. Recordemos que inicialmente considerábamos aplicar splines interpoladores cuando las medidas a lo largo del tiempo no estaban sujetas a error. En caso de posibles errores de medida tenía más sentido el spline regresión (véase punto 3.4). El criterio con el que decidiremos qué tipo de spline es mejor será el gráfico: representaremos el spline junto con los puntos reales y veremos si el comportamiento del spline se ajusta a la imagen intuitiva que deberíamos obtener. Diremos que intuitivamente se ajusta si el spline no tiene comportamientos extraños y si el máximo del spline está a una distancia no excesiva del máximo observado en los datos. La estimación del spline interpolador no comporta problemas de cálculo, pues es único. Sin embargo el spline regresión depende del parámetro de suavizado α que debemos encontrar.

Tras probar varios valores de α se ha escogido $\alpha=0,999999995$ porque con $\alpha=0.999999990$ el spline no se acercaba al máximo de la muestra y $\alpha=0,999999999$ no era diferente a la estimación del spline interpolador. Valores inferiores a $\alpha=0.90$ dan rectas de regresión.

Se han representado gráficamente para las 76 funciones, tanto la representación en spline interpolador como la del spline regresión. En todas ellas se observa que el spline regresión suaviza alguno de los comportamientos irregulares del spline interpolador. A veces, valores cercanos en el tiempo pero de muy distinto valor, producen en el spline interpolador unas oscilaciones con intervalos fuera del rango de valores empíricamente posible. Además parece que el spline interpolador sobreestima el valor del máximo de la función. Sin embargo el spline regresión evita estas oscilaciones porque no tiene la restricción de pasar por el punto observado. Un ejemplo representativo es el ejemplo 7.2., donde vemos que el spline interpolador sobreestima el máximo y estima concentraciones negativas en el intervalo entre 0,1 y 0,2.

EJEMPLO 7.2:

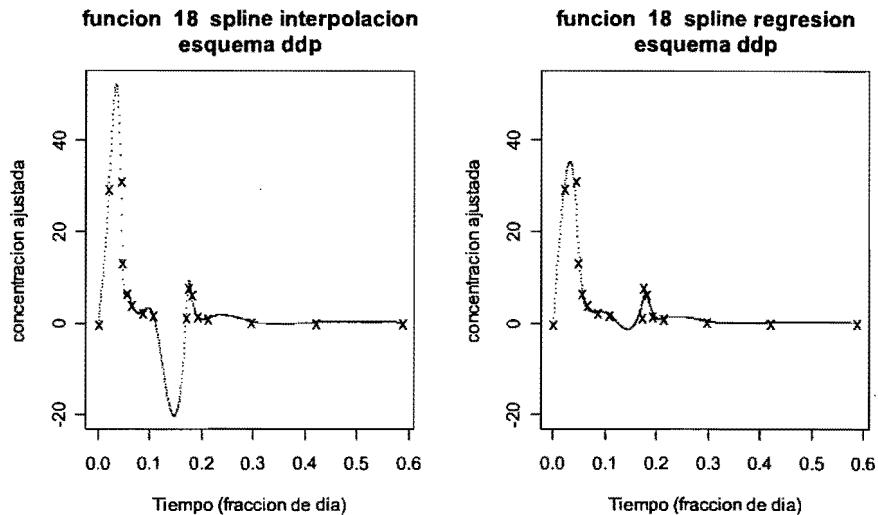


Figura 7.2 y 7.3: Estimación por spline interpolador (izquierda) y de regresión (a la derecha) de la concentración de Docetaxel en grupos DP y PD en el paciente 18.

Cabe destacar que el utilizar un tipo de spline u otro conduce a resultados distintos en la descriptiva de primer y de segundo nivel. Es por eso que decidir el tipo de spline más apropiado es básico dentro del análisis.

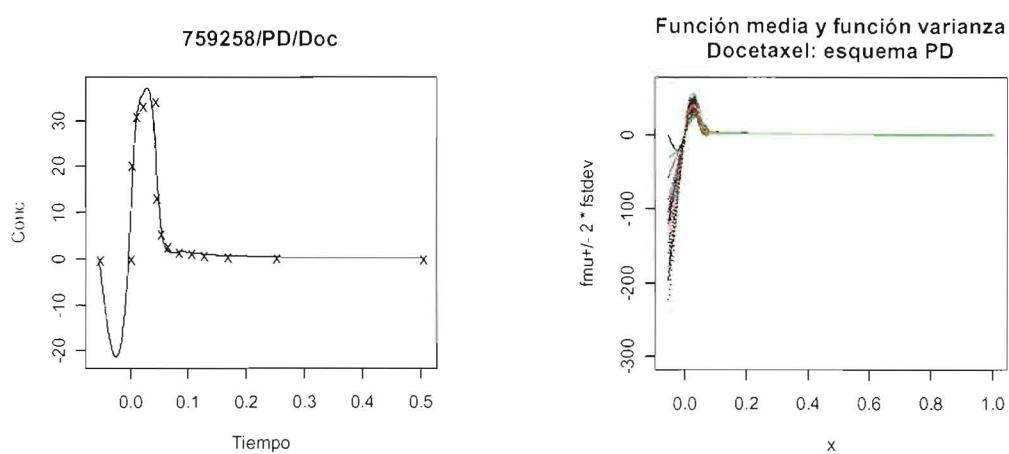
7.2.2. – Origen del tiempo de las funciones y rango de registro.

El rango de registro y el origen del tiempo es importante. Existen pacientes que a tiempo cero ya tenían concentraciones diferentes de cero y que además tenían concentraciones igual a cero en un origen de tiempo negativo (por ejemplo $t=-0.05$). El hecho de tener tiempos negativos es una señal de que el origen del tiempo ha sido mal tomado. Esto tiene dos repercusiones: La primera es que la banda de varianza será mayor si el origen real de la función está alejado del origen teórico (véase capítulo 5). La segunda consiste en que al hacer la operación de registro de knots `regspl()` estaremos forzando a que todos los splines sean evaluados también en esos valores negativos de tiempo. La función en tiempos negativos también contará a la hora de posteriormente calcular el valor medio y la varianza de cada una de las funciones. En nuestro ejemplo los splines tienden a menos infinito muy deprisa cuando las ordenadas van de valores

positivos a negativos. Por eso estaríamos considerando áreas que sesgan hacia valores negativos el valor medio, y sobreestiman la varianza de las funciones.

La función `regspl()` registra los knots una lista de splines. Es decir, hace una lista de splines cuyos knots son el conjunto de todos los knots posibles, y calcula para cada uno de los splines los coeficientes de los polinomios correspondientes a cada intervalo posible. En este caso el paciente 2 produce que el registro de knots empieza en el tiempo -0.05. Si calculamos la descriptiva funcional de primer nivel también estaremos considerando los tiempos entre -0.05 y 0. Como inicialmente la pendiente es positiva y empieza en el cero, en el intervalo anterior los splines tendrán valor negativo, con lo que tanto el valor medio como la varianza se verán afectadas. Como ejemplo veremos que la descriptiva de primer y de segundo nivel, si no restringimos los tiempos, se ve afectada de manera considerable.

EJEMPLO 7.3.



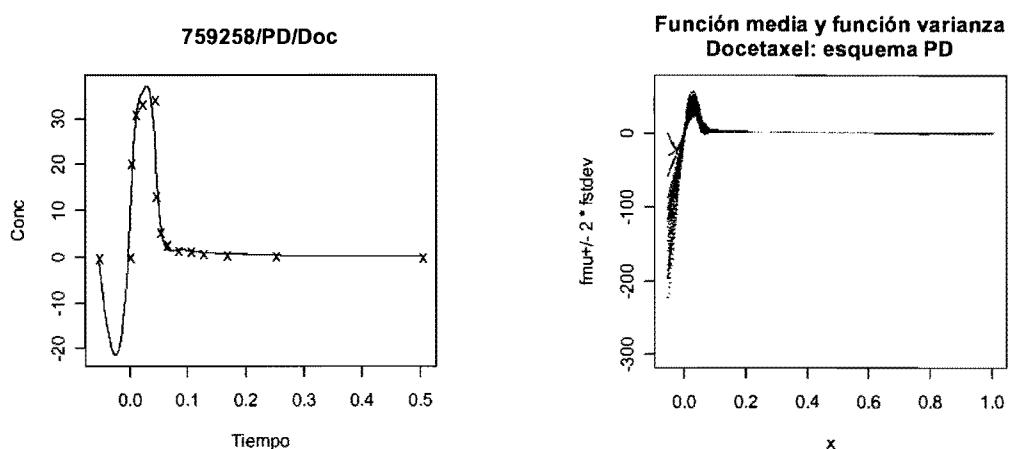
Figuras 7.4 y 7.5: Ejemplo de función con tiempo negativo (a la izquierda) problemática para registro. Al considerar este tiempo negativo dentro del registro, la función media y función varianza (a la derecha) también serán evaluadas en ese tiempo.

Observamos cómo un mal registro sesga y magnifica el valor medio y la varianza respectivamente de cada una de las funciones.

positivos a negativos. Por eso estaríamos considerando áreas que sesgan hacia valores negativos el valor medio, y sobreestiman la varianza de las funciones.

La función `regspl()` registra los knots una lista de splines. Es decir, hace una lista de splines cuyos knots son el conjunto de todos los knots posibles, y calcula para cada uno de los splines los coeficientes de los polinomios correspondientes a cada intervalo posible. En este caso el paciente 2 produce que el registro de knots empieza en el tiempo -0.05. Si calculamos la descriptiva funcional de primer nivel también estaremos considerando los tiempos entre -0.05 y 0. Como inicialmente la pendiente es positiva y empieza en el cero, en el intervalo anterior los splines tendrán valor negativo, con lo que tanto el valor medio como la varianza se verán afectadas. Como ejemplo veremos que la descriptiva de primer y de segundo nivel, si no restringimos los tiempos, se ve afectada de manera considerable.

EJEMPLO 7.3.



Figuras 7.4 y 7.5: Ejemplo de función con tiempo negativo (a la izquierda) problemática para registro. Al considerar este tiempo negativo dentro del registro, la función media y función varianza (a la derecha) también serán evaluadas en ese tiempo.

Observamos cómo un mal registro sesga y magnifica el valor medio y la varianza respectivamente de cada una de las funciones.

	REGISTRO INCORRECTO	REGISTRO CORRECTO		REGISTRO INCORRECTO	REGISTRO CORRECTO	
	media	varianza	media	varianza	media	varianza
f1	-2.41453961	619.97730	2.253368	59.31510	f11	-1.66224612
f2	-3.02815570	873.16943	2.555408	72.57059	f12	-2.82562872
f3	-0.54617800	307.96599	2.638428	62.98715	f13	-1.04686417
f4	-2.56318181	555.68912	1.864712	49.74582	f14	-2.09396513
f5	-1.26973703	396.49835	2.248074	63.47533	f15	-1.07131119
f6	-2.17010447	675.39623	2.706709	65.97350	f16	0.09701882
f7	1.18055401	53.11172	1.940109	41.66075	f17	-1.18836786
f8	-2.40176989	686.70662	2.497289	63.67218	f18	-1.12421574
f9	-2.17097894	563.49742	2.242821	59.02299	f19	-0.50146714
f10	-2.09092514	458.92606	1.916779	46.93005		

Tabla 7.2. Descriptiva de primer y de segundo nivel para docetaxel en esquema PD con spline regresión.

En la tabla de la descriptiva de primer nivel vemos que muchos de los valores medios son negativos, algo que es experimentalmente imposible. En la figura 7.5 se observa cómo todos los splines tienen valores negativos en la parte negativa del eje de ordenadas. Esto explica que, a causa de que existe un área negativa que se resta del área entre cero y uno, los valores medios sean inferiores a cero. Es por eso que es muy importante que en la operación de registro se limite el rango de valores. Esto se consigue mediante este comando:

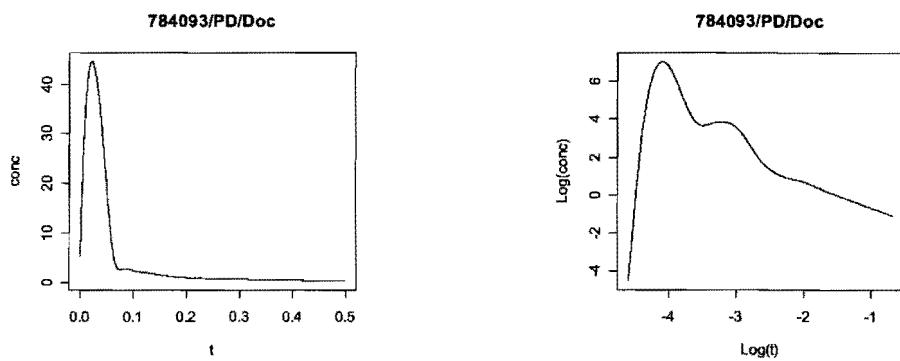
```
listarregddp<-regspl(listaregddp,xlims=c(0,1.3))
listarregdpd<-regspl(listaregdpd,xlims=c(0,1.3))
listarregpdp<-regspl(listaregpdp,xlims=c(0,1.3))
listarregppd<-regspl(listaregppd,xlims=c(0,1.3))
```

Finalmente para solventar estos problemas se ha corregido cada una de las funciones situando el origen en $t=0$ y se han registrado las listas de splines entre 0 y 1,3. También se han eliminado 6 funciones porque no tenían evaluadas las concentraciones en puntos clave de tiempo, y los splines resultantes no han sido capaces de estimar la forma esperada. Además éstos afectaban a la estimación de la función media y de las bandas de varianza. También se han realizado en 14 funciones ajuste de origen.

7.2.3.- Transformación de los datos.

En ocasiones puede ser interesante transformar los datos para conseguir una visión más detallada de las variaciones se sufren las funciones a lo largo del eje de ordenadas. En la

representación obtenida sin transformar vemos que los tiempos en donde se observa mayor variabilidad es en los cercanos a cero, mientras que los cercanos a uno las funciones son prácticamente plana. Si aplicamos el logaritmo en los tiempos se ensanchará la escala de tiempos cercanos a cero y se estrechará la escala de tiempos hacia el uno. Si aplicamos el logaritmo a las concentraciones veremos más matrices de la variabilidad de las concentraciones próximas a cero, mientras que veremos el pico de máxima concentración no tan acentuado. Por lo tanto parece interesante aplicar esta transformación.



Figuras 7.6 y 7.7: Representación de una función sin transformación (izquierda) y con transformación logarítmica (derecha). A la derecha se aprecia que la bajada de concentración no es lineal, tal como parece a la izquierda.

7.2.4. – Registro y reducción de knots de la lista de spline.

En el ejemplo del capítulo 6 de pirámides de población los problemas de registro de knots no se producían, gracias a que partíamos de histogramas con las mismas marcas de clase para una edad y sexo concretas. Sin embargo en datos farmacocinéticos tenemos que en muy pocas ocasiones coinciden exactamente los tiempos de medición entre una farmacocinética y otra. Esto da como resultado una lista registrada de splines con muchos knots (125) y unas matrices de coeficientes con muchas filas repetidas. Cuando ejecutamos los procedimientos de descriptiva funcional este efecto no tiene importancia, pero sí la tiene a la hora de diagonalizar la matriz asociada al cálculo de CPF'S, ya que la matriz resultante es singular.

```

> res.2<-fpca.sint(listamalcen)
Error in eigen(A, symmetric = T) : NA/NaN/Inf in foreign function call (arg 3)
In addition: Warning message:
NaNs produced in: sqrt(Wvp$values)

```

La solución a esta situación consiste en reducir los knots de la lista de splines. El comando de reducción se ejecuta de la siguiente manera:

```

> lredddp<-f.regred(listarregddp,knots=seq(-4.61,0.27,0.2),tipo="int",grdif=F)
> lreddpd<-f.regred(listarregdpd,knots=seq(-4.61,0.27,0.2),tipo="int",grdif=F)
> lredpdp<-f.regred(listarregpdp,knots=seq(-4.61,0.27,0.2),tipo="int",grdif=F)
> lredppd<-f.regred(listarregppd,knots=seq(-4.61,0.27,0.2),tipo="int",grdif=F)

```

Este comando necesita la lista de splines que se reducirá, junto con la lista de nuevos knots (*knots=...*) En este caso hemos utilizado, para reducir el número de knots, el spline interpolador (*tipo="int"*) en el cálculo del spline reducido.(véase punto 5.3). A continuación puede verse una tabla con las pérdidas de cada una de las funciones y con las pérdidas totales.

NHC	DOCTAXEL DP				DOCTAXEL PD				PACLTAXEL DP				PACLTAXEL PD			
	Area ²	Redx 10 ⁻⁴	10 ^{5*} Red/A ²	Area ²	Redx 10 ⁻⁴	10 ^{5*} Red/A ²	Area ²	Redx 10 ⁻⁴	10 ^{5*} Red/A ²	Area ²	Redx 10 ⁻⁴	10 ^{5*} Red/A ²	Area ²	Redx 10 ⁻⁴	10 ^{5*} Red/A ²	
pac160	---	---	---	22.26	1.39	0.62	---	---	---	16.03	7.34	4.58				
pac170	21.97	5.13	2.34	23.70	14.70	6.20	20.02	5.68	2.84	16.59	18.30	11.03				
pac178	25.07	3.79	1.51	20.84	4.75	2.28	20.55	11.90	5.79	15.51	2.22	1.43				
pac330	20.85	2.21	1.06	25.78	5.33	2.07	16.53	9.81	5.94	18.85	5.41	2.87				
pac369	19.63	24.31	12.38	---	---	---	30.34	3.54	1.17	17.11	8.54	5.00				
pac492	15.98	2.43	1.52	24.65	4.73	1.92	17.45	9.45	5.42	19.60	9.61	4.90				
pac507	18.76	1.48	0.79	26.51	7.98	3.01	16.96	7.84	4.62	19.51	2.78	1.43				
pac510	19.15	5.96	3.11	24.11	5.90	2.45	19.27	10.66	5.53	19.48	6.60	3.39				
pac542	19.15	3.78	1.97	26.02	2.77	1.07	15.35	13.18	8.58	18.32	12.12	6.61				
pac759	25.66	6.38	2.49	24.93	5.74	2.30	16.50	18.80	11.39	18.73	5.78	3.09				
pac784	20.40	4.06	1.99	42.10	0.60	0.14	18.65	11.59	6.22	20.07	5.96	2.97				
pac830	17.74	30.97	17.46	19.26	2.12	1.10	27.81	12.01	4.32	16.62	2.83	1.70				
pac833	19.88	9.39	4.72	---	---	---	20.34	16.17	7.95	14.80	0.07	0.05				
pac849	19.63	41.12	20.95	22.94	2.84	1.24	21.65	12.43	5.74	16.86	14.89	8.83				
pac868	18.46	1.90	1.03	27.12	4.79	1.77	19.16	13.04	6.81	20.46	7.82	3.82				
pac871	23.03	0.71	0.31	26.97	12.55	4.65	17.39	17.02	9.78	19.42	6.27	3.23				
pac878	20.76	4.88	2.35	26.07	5.55	2.13	34.71	13.67	3.94	17.45	1.10	0.63				
pac883	---	---	---	37.31	9.11	2.44	---	---	---	19.79	24.08	12.17				
pac894	17.02	5.35	3.14	19.60	9.28	4.74	20.98	18.53	8.83	16.89	9.10	5.39				
pac954	17.06	5.26	3.08	29.24	91.54	31.31	16.97	8.00	4.71	19.70	1.48	0.75				

Tabla 7.3: Función de pérdida (área entre curvas) de reducción de registro de knots para cada una de las funciones. Las funciones en donde no hay valor son aquellas que han sido eliminadas del análisis.

Hemos calculado el cociente entre el área de discrepancia entre función y función reducida y el área al cuadrado de cada una de las funciones. Como mucho la tasa de reducción ha sido del el tanto por ciento de discrepancia ha sido del 31×10^{-5} aproximadamente. La reducción de la lista de knots puede tener repercusiones en los estadísticos descriptivos. Veamos un ejemplo para docetaxel, esquema DP, en donde vemos que existen ligeras discrepancias en los estadísticos de primer nivel:

	valor medio		Varianza	
	extensa	reducida	extensa	reducida
Pac1	0.33	0.36	4.41	4.41
Pac2	-0.33	-0.28	5.05	4.92
Pac3	-0.04	0.01	4.28	4.21
Pac4	0.57	0.59	3.71	3.73
Pac5	0.19	0.20	3.25	3.30
Pac6	0.07	0.10	3.85	3.87
Pac7	0.34	0.36	3.82	3.87
Pac8	0.05	0.08	3.93	3.93
Pac9	-0.45	-0.40	5.07	4.95
pac10	0.47	0.50	3.97	3.98
pac11	0.30	0.32	3.56	3.56
pac12	0.13	0.15	4.07	4.09
pac13	0.42	0.44	3.86	3.88
pac14	0.23	0.24	3.74	3.80
pac15	-0.14	-0.10	4.71	4.67
pac16	0.04	0.07	4.27	4.25
pac17	-0.14	-0.11	3.48	3.47
Pac18	-0.08	-0.06	3.50	3.52

Tabla 7.4: Descriptiva funcional numérica de primer nivel calculada para la lista de splines registrada original (extensa) y para la lista de splines con reducción de knots.

Vemos que las discrepancias entre unas y otras medidas se mueven entre las dos y las seis centésimas, por lo que nos podemos dar por satisfechos.

7.3.- ANÁLISIS DESCRIPTIVO.

El objetivo de este punto va a ser caracterizar la variación a lo largo del tiempo de la concentración de medicamentos en función del grupo y del fármaco. Interesa sobre todo observar el efecto de la secuencia (PD o DP) en la concentración del medicamento. Para ello inicialmente realizaremos una descripción de la concentración para cada una de las

posibilidades y posteriormente calcularemos las correlaciones entre las medias de cada grupo. Utilizaremos directamente la lista de splines con reducción de knots, ya que será la que también deberemos utilizar para el análisis de componentes principales funcionales.

7.3.1.- Descriptiva funcional de primer nivel y segundo nivel grupo a grupo.

En primer lugar calcularemos el valor medio y la varianza de cada función para cada grupo y para cada fármaco (para todo $C_{ij,k}$), utilizando spline de regresión, por las razones comentadas anteriormente. En las figuras 7.8-7.11 se representarán gráficamente la función media y las bandas alrededor de la media basadas en la desviación típica puntual. A continuación comentaremos las diferencias entre distintos grupos de concentración con el mismo fármaco.

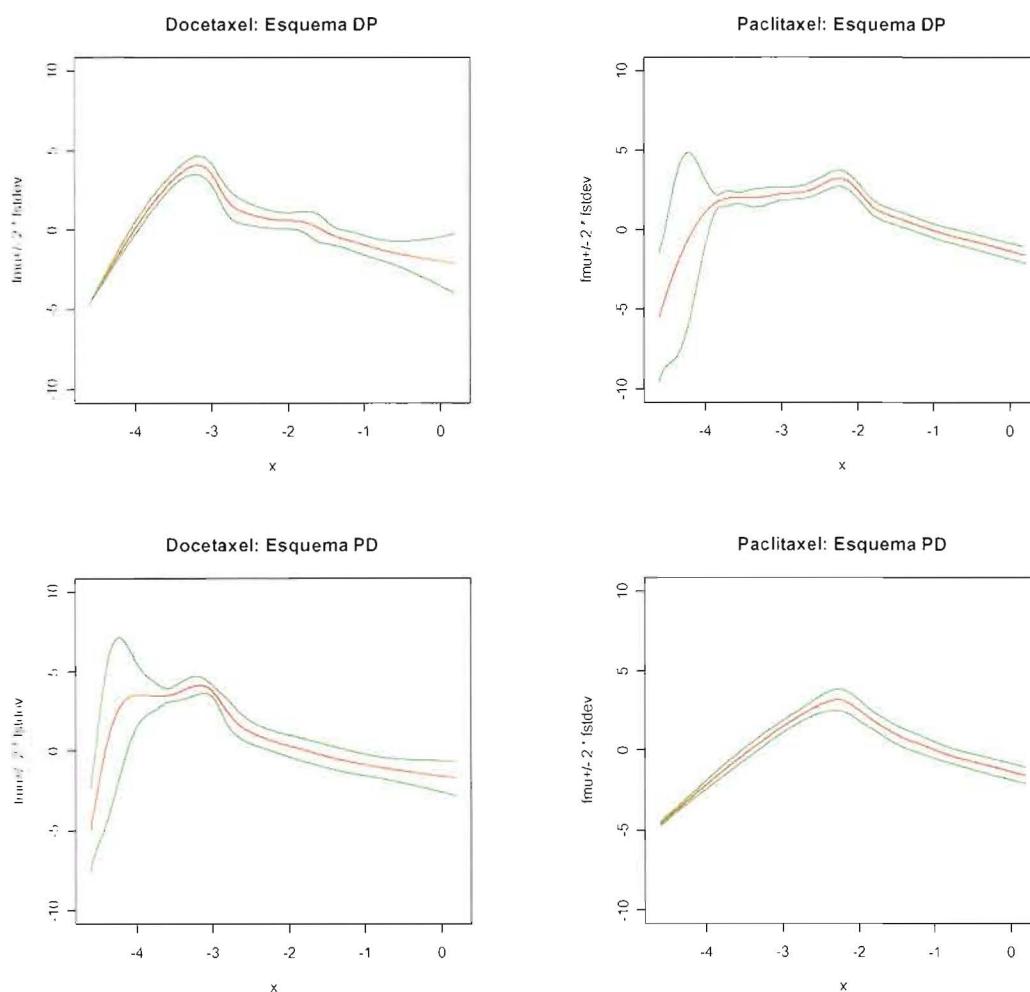
N. PAC	DOCETAXEL						PACLITAXEL					
	GRUPO DP		GRUPO PD		Diferencia		GRUPO DP		GRUPO PD		Diferencia	
	media	varianza	media	varianza	media	varianza	media	varianza	Media	varianza	media	varianza
16094/	--	--	0.59	4.42	--	--	--	--	-0.16	3.37	--	--
170678	0.36	4.41	0.21	4.99	-0.15	0.58	0.24	4.22	-0.33	3.40	-0.57	-0.82
178839	-0.28	4.92	0.63	4.09	0.91	-0.83	0.29	4.29	-0.07	3.27	-0.37	-1.02
330188	0.01	4.21	0.99	4.48	0.99	0.27	0.99	2.54	0.16	3.94	-0.83	1.40
369196	0.59	3.73	---	---	---	---	0.38	6.32	0.04	3.60	-0.34	-2.72
492055	0.20	3.30	0.77	4.70	0.56	1.40	1.22	2.20	0.19	4.08	-1.03	1.88
507823	0.10	3.87	0.94	4.73	0.85	0.87	0.99	2.65	0.08	4.10	-0.91	1.45
51031/	0.36	3.87	1.14	3.73	0.78	-0.13	1.16	2.75	0.13	4.09	-1.02	1.33
542033	0.08	3.93	0.57	5.27	0.49	1.34	0.73	2.76	-0.13	3.85	-0.86	1.10
759258	-0.40	4.95	0.68	4.86	1.08	-0.09	0.91	2.70	0.11	3.92	-0.80	1.22
784093	0.50	3.98	1.39	7.02	0.89	3.04	1.27	2.33	0.31	4.12	-0.96	1.79
830798	0.32	3.56	0.97	3.12	0.65	-0.45	0.24	5.88	0.10	3.48	-0.14	-2.40
833520	0.15	4.09	---	---	---	---	0.23	4.28	-0.14	3.11	-0.37	-1.17
849575	0.44	3.88	0.93	4.05	0.49	0.17	0.49	4.32	0.02	3.55	-0.48	-0.77
868583	0.24	3.80	1.07	4.61	0.83	0.81	0.90	3.23	0.14	4.28	-0.76	1.05
871614	-0.10	4.67	0.67	5.34	0.77	0.68	0.92	2.90	-0.07	4.09	-0.99	1.20
878675	0.07	4.25	1.01	4.53	0.94	0.28	0.49	7.15	0.17	3.65	-0.32	-3.50
883975	---	---	0.34	7.87	---	---	---	---	0.16	4.14	---	---
894472	-0.11	3.47	0.21	4.29	0.32	0.83	0.19	4.43	-0.23	3.51	-0.42	-0.92
954735	-0.06	3.52	0.87	5.56	0.93	2.04	1.04	2.52	0.07	4.14	-0.97	1.62

Tabla 7.5: Valor medio y varianza de cada una de las funciones de la muestra.

En el ejemplo que nos ocupa, el valor medio por su definición, puede ser interpretado como la concentración acumulada recibida de fármaco. La varianza nos va a indicar si la función fluctúa mucho o poco respecto a la función constante valor medio.

Observando las diferencias entre esquemas de tratamiento en Docetaxel, en casi todos los pacientes se absorbe más fármaco con el esquema de tratamiento PD. Además parece observarse más variabilidad en este esquema. De hecho la experiencia de investigadores y enfermeras indica que este esquema de tratamiento (PD) ha presentado más toxicidades. Es posible que tanto la media como la varianza como medidas se pueda relacionar con la toxicidad.

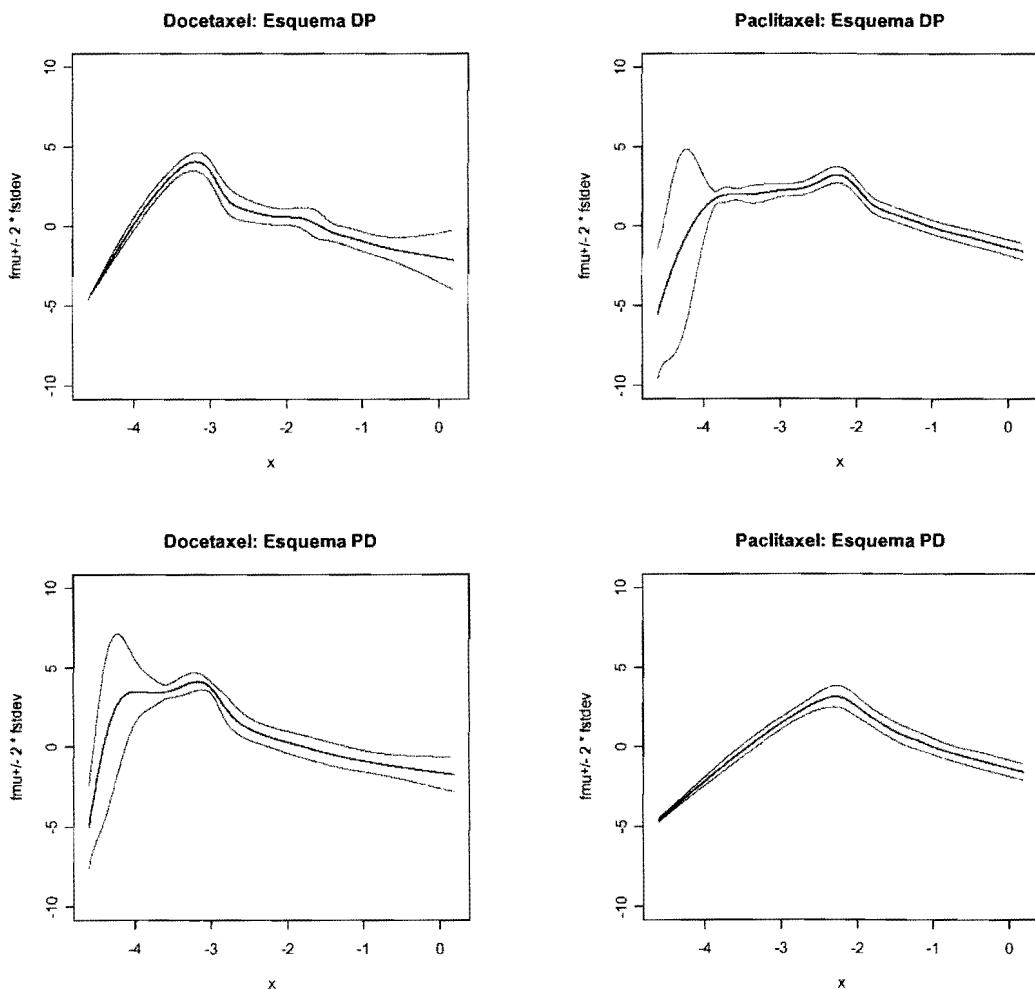
Analizando el comportamiento de Paclitaxel, se observa que los pacientes han absorbido más fármaco en el esquema DP. En cuanto a la variabilidad se puede decir que un esquema u otro no presentan diferencias que puedan ser relevantes, ya que el número de incrementos positivos y negativos son parecidos.



Figuras 7.8-7.11: Función media para cada fármaco y grupo. Las bandas alrededor de la media se basan en la raíz cuadrada de la función varianza.

Observando las diferencias entre esquemas de tratamiento en Docetaxel, en casi todos los pacientes se absorbe más fármaco con el esquema de tratamiento PD. Además parece observarse más variabilidad en este esquema. De hecho la experiencia de investigadores y enfermeras indica que este esquema de tratamiento (PD) ha presentado más toxicidades. Es posible que tanto la media como la varianza como medidas se pueda relacionar con la toxicidad.

Analizando el comportamiento de Paclitaxel, se observa que los pacientes han absorbido más fármaco en el esquema DP. En cuanto a la variabilidad se puede decir que un esquema u otro no presentan diferencias que puedan ser relevantes, ya que el número de incrementos positivos y negativos son parecidos.



Figuras 7.8-7.11: Función media para cada fármaco y grupo. Las bandas alrededor de la media se basan en la raíz cuadrada de la función varianza.

En el análisis descriptivo de segundo nivel, por lo que vemos las figuras 7.8-7.11 parece ser que tanto en un esquema como en el otro, los fármacos que entran después sufren una variabilidad al inicio de su administración (ver Docetaxel esquema PD y Paclitaxel esquema DP). También se observa que en Docetaxel la concentración máxima se produce antes en el tiempo y con una pendiente más pronunciada que en Paclitaxel. Si bien en la función media la concentración máxima de Docetaxel es mayor en el esquema DP, ahí la desviación estándar máxima alcanza valores muy altos al inicio respecto el otro esquema. Sin embargo en Paclitaxel los máximos de las desviaciones estándar en uno y otro esquema no están tan distanciados.

7.3.2.- Descriptiva funcional entre grupos.

Una manera de relacionar los grupos es ver cuál es la correlación entre la función media de cada uno de los grupos. Compararemos entonces las correlaciones siguientes:

$$\text{CORR}[\bar{C}_{D,PD}, \bar{C}_{D,DP}] \text{ y } \text{CORR}[\bar{C}_{P,PD}, \bar{C}_{P,DP}]$$

Hasta el momento teníamos cuatro parrillas de datos; una para cada grupo y fármaco. Por lo tanto las medias no compartirán los mismos knots. Es por eso que debemos registrar las cuatro funciones media en una nueva lista. Una vez registrados calcularemos la descriptiva numérica de primer nivel y las correlaciones de las medias dos a dos:

Función Media	media	varianza
Docetaxel, PD	0.78	4.35
Docetaxel, DP	0.14	3.95
Paclitaxel, PD	0.03	3.75
Paclitaxel, DP	0.71	3.11

Tabla 7.6.: Valor medio y varianza de las funciones medias.

El análisis de primer nivel nos sigue indicando que en el $C_{D,PD}$, sigue destacando por su alta variabilidad.

Correlación fina	Docetaxel, PD	Docetaxel, DP	Paclitaxel, PD	Paclitaxel, DP
Docetaxel, PD	1	0.82	0.13	0.65
Docetaxel, DP	0.82	1	0.56	0.85
Paclitaxel, PD	0.13	0.56	1	0.81
Paclitaxel, DP	0.65	0.85	0.81	1

Tabla 7.7.: Correlaciones entre funciones media de los quimioterápicos por esquemas.

Las correlaciones entre las funciones media nos indican que para el esquema DP la correlación entre Docetaxel y Paclitaxel es de 0.85. Esto quiere decir que en muchos intervalos del tiempo, donde la concentración de Docetaxel es alta la de Paclitaxel también lo es. Sin embargo en el esquema PD la correlación es de 0.13. Es decir, en muchos intervalos cuando la concentración es alta en un fármaco, en el otro no lo es tanto. Las correlaciones entre Docetaxel en un esquema y otro es de 0.82, y en Paclitaxel es de 0.82. Esto indica que la forma de unos y otros fármacos tienen bastante parecido aunque no del todo. A la vista del gráfico siguiente vemos que estas discrepancias pueden ser debidas al distinto comportamiento en tiempos iniciales.

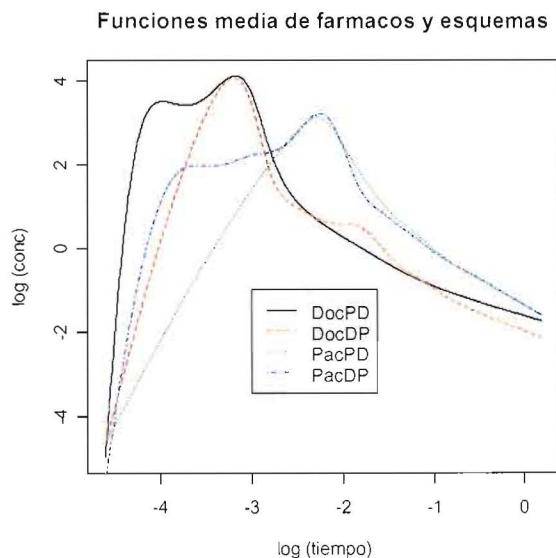


Figura 7.12: Representación gráfica de las funciones media de fármaco-esquema de tratamiento.

El ACPF puede ofrecernos información sobre en qué intervalos de tiempo se produce variabilidad.

Correlación fmú	Docetaxel, PD	Docetaxel, DP	Paclitaxel, PD	Paclitaxel, DP
Docetaxel, PD	1	0.82	0.13	0.65
Docetaxel, DP	0.82	1	0.56	0.85
Paclitaxel, PD	0.13	0.56	1	0.81
Paclitaxel, DP	0.65	0.85	0.81	1

Tabla 7.7.: Correlaciones entre funciones media de los quimioterápicos por esquemas.

Las correlaciones entre las funciones media nos indican que para el esquema DP la correlación entre Docetaxel y Paclitaxel es de 0.85. Esto quiere decir que en muchos intervalos del tiempo, donde la concentración de Docetaxel es alta la de Paclitaxel también lo es. Sin embargo en el esquema PD la correlación es de 0.13. Es decir, en muchos intervalos cuando la concentración es alta en un fármaco, en el otro no lo es tanto. Las correlaciones entre Docetaxel en un esquema y otro es de 0.82, y en Paclitaxel es de 0.82. Esto indica que la forma de unos y otros fármacos tienen bastante parecido aunque no del todo. A la vista del gráfico siguiente vemos que estas discrepancias pueden ser debidas al distinto comportamiento en tiempos iniciales.

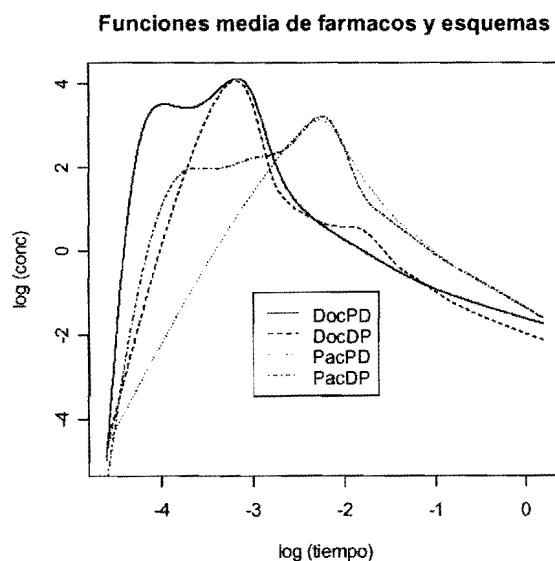


Figura 7.12: Representación gráfica de las funciones media de fármaco-esquema de tratamiento.

El ACPF puede ofrecernos información sobre en qué intervalos de tiempo se produce variabilidad.

7.4.- COMPONENTES PRINCIPALES FUNCIONALES.

7.4.1.- Preparación del análisis.

Como nos interesa encontrar pautas de variabilidad que distingan a un grupo del otro inicialmente crearemos dos listas, una para cada medicamento. Aplicaremos componentes principales funcionales (CPF's) y veremos si logra discriminar un grupo del otro.

7.4.2.- Estudio del Docetaxel.

La lista de splines con reducción de knots previamente ha sido centrada, restando a todos los splines de la muestra la función media. A continuación aplicamos componentes principales funcionales.

En primer lugar haremos el gráfico de representatividad de las CPF's:

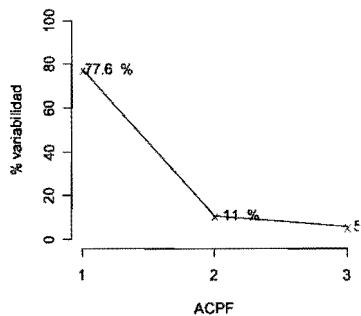


Figura 7:13: Gráfico de representación de cada una de las componentes principales.

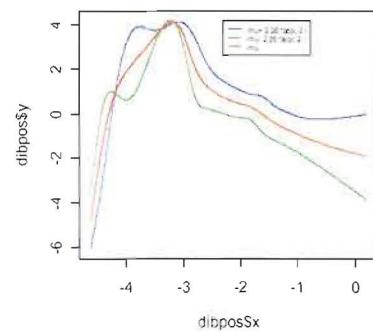
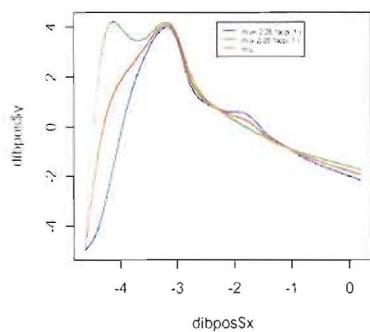
Vemos que con las dos primeras CPF's explicaremos el 88,6 % de la variabilidad total y por tanto la representación con ellas será buena. Comprobemos ahora si las CPF's explicarán las diferencias en esquema de tratamiento.

La tabla de correlaciones (tabla 7.8) muestran que las curvas de concentración en los pacientes tratados con esquema DP se correlacionan positivamente con la primera CPF, mientras que los tratados con esquema PD se correlacionan negativamente. Esto denota que esta componente es la que explica la variabilidad debida a la existencia de dos esquemas. Para poder seguir interpretando esta tabla es necesario hacer las dos representaciones gráficas: la representación

gráfica de estos coeficientes en un plano y el gráfico de cómo distorsiona la función media las CPF's.

ESQUEMA DP			ESQUEMA PD		
PACIENTE	CPF.1	CPF.2	PACIENTE	CPF.1	CPF.2
170DP	-0.92	0.17	170PD	-0.59	0.39
160DP	---	---	160PD	0.73	0.28
178DP	-0.79	-0.54	178PD	0.95	0.09
330DP	-0.91	-0.31	330PD	0.99	0.00
369DP	-0.77	0.45	369PD	---	---
492DP	-0.83	0.09	492PD	0.99	-0.10
507DP	-0.94	-0.12	507PD	0.99	0.05
510DP	-0.91	0.23	510PD	0.87	0.30
542DP	-0.98	-0.18	542PD	0.90	-0.39
759DP	-0.74	-0.61	759PD	0.97	-0.16
784DP	-0.85	0.33	784PD	0.97	0.09
830DP	-0.91	0.17	830PD	0.28	0.87
833DP	-0.94	-0.08	833PD	---	---
849DP	-0.87	0.29	849PD	0.22	0.84
868DP	-0.92	0.12	868PD	0.98	0.15
871DP	-0.88	-0.44	871PD	0.95	-0.28
878DP	-0.95	-0.20	878PD	0.98	0.11
894DP	-0.83	-0.38	894PD	0.20	-0.49
954DP	-0.93	-0.22	954PD	0.89	-0.33

Tabla 7.8: Correlación entre las funciones de cada paciente y las dos primeras CPF's en docetaxel.

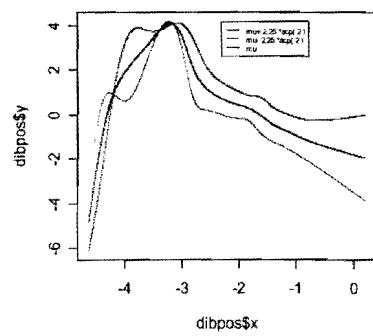
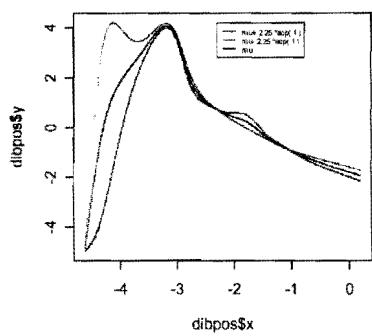


Figuras 7.14 y 7.15: Representación de las CPF's en función de cómo afectan a la media para Docetaxel.

gráfica de estos coeficientes en un plano y el gráfico de cómo distorsiona la función media las CPF'S.

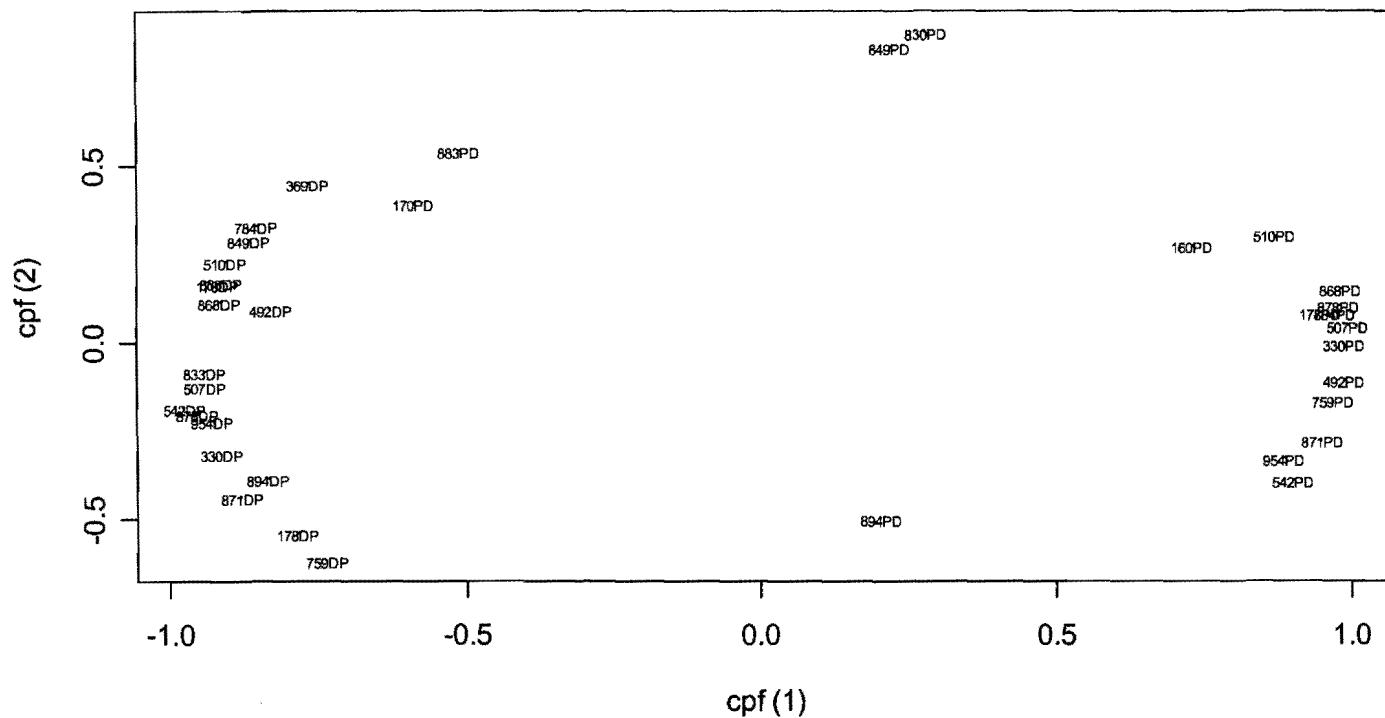
ESQUEMA DP			ESQUEMA PD		
PACIENTE	CPF.1	CPF.2	PACIENTE	CPF.1	CPF.2
170DP	-0.92	0.17	170PD	-0.59	0.39
160DP	---	---	160PD	0.73	0.28
178DP	-0.79	-0.54	178PD	0.95	0.09
330DP	-0.91	-0.31	330PD	0.99	0.00
369DP	-0.77	0.45	369PD	---	---
492DP	-0.83	0.09	492PD	0.99	-0.10
507DP	-0.94	-0.12	507PD	0.99	0.05
510DP	-0.91	0.23	510PD	0.87	0.30
542DP	-0.98	-0.18	542PD	0.90	-0.39
759DP	-0.74	-0.61	759PD	0.97	-0.16
784DP	-0.85	0.33	784PD	0.97	0.09
830DP	-0.91	0.17	830PD	0.28	0.87
833DP	-0.94	-0.08	833PD	---	---
849DP	-0.87	0.29	849PD	0.22	0.84
868DP	-0.92	0.12	868PD	0.98	0.15
871DP	-0.88	-0.44	871PD	0.95	-0.28
878DP	-0.95	-0.20	878PD	0.98	0.11
894DP	-0.83	-0.38	894PD	0.20	-0.49
954DP	-0.93	-0.22	954PD	0.89	-0.33

Tabla 7.8: Correlación entre las funciones de cada paciente y las dos primeras CPF's en docetaxel.



Figuras 7.14 y 7.15: Representación de las CPF's en función de cómo afectan a la media para Docetaxel.

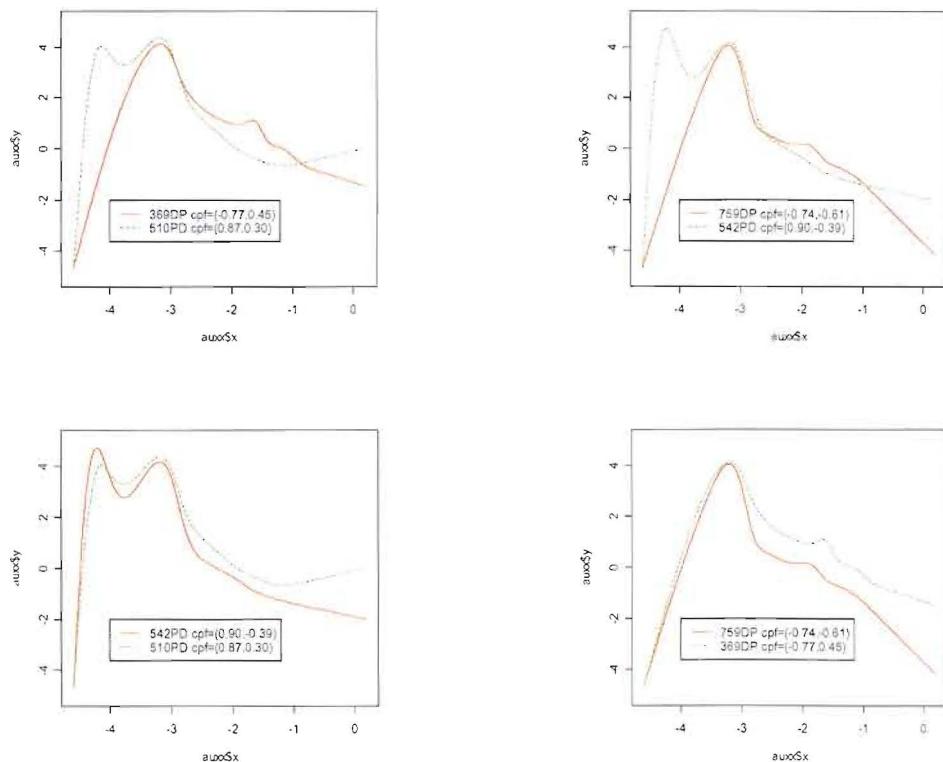
Correlaciones entre CPFs en Docetaxel



Figuras 7.16: Representación de las correlaciones en el planode las CPF's para Docetaxel

De las figuras anteriores se deduce que la primera CPF consigue discriminar entre uno y otro esquema porque explica las discrepancias en el inicio de la infusión al paciente. El pico de concentración máxima común está situado a tiempo $\exp(-3.5)$. Sin embargo aquellos que tienen una CPF más alta tienen otro pico previo de concentración a tiempo aproximado de $\exp(-4.2)$. Este nuevo pico es el que explica la primera CPF y es el rasgo que nos permite diferenciar los grupos de tratamiento en Docetaxel.

Si fijamos la primera CPF y recorremos de negativo a positivo la segunda CPF, veremos que ésta discrimina entre funciones que casi uniformemente tienen mayor concentración. Es decir, fijado si tiene uno o dos picos, las que tengan una segunda CPF mayor estarán por encima de las que tengan una segunda CPF menor. A continuación mostramos unos ejemplos en donde podemos ver representados mejor estos aspectos:

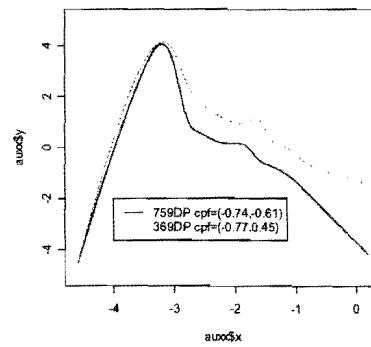
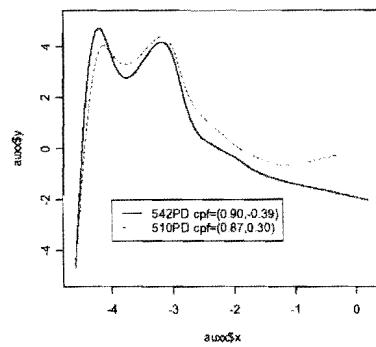
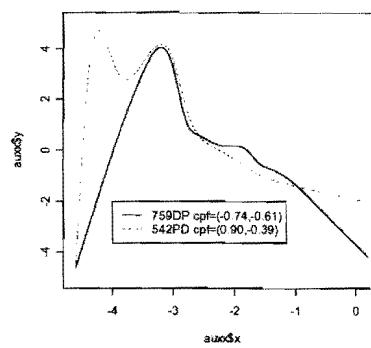
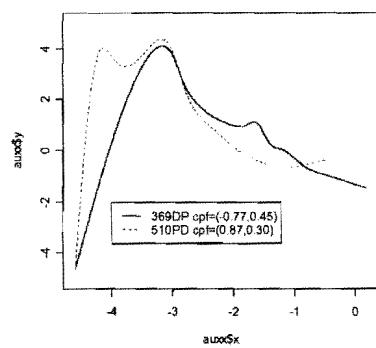


Figuras 7.17-7.20: Comparación de funciones con diferentes correlaciones respecto las CPFs de Docetaxel

Las dos primeras figuras son ejemplos de comparación entre valores extremos de la primera CPF fijada la segunda. Vemos que la primera CPF explica si aparece o no un pico previo de concentración. Las dos figuras de abajo comparan valores extremos de la segunda CPF, fijada

De las figuras anteriores se deduce que la primera CPF consigue discriminar entre uno y otro esquema porque explica las discrepancias en el inicio de la infusión al paciente. El pico de concentración máxima común está situado a tiempo $\exp(-3.5)$. Sin embargo aquellos que tienen una CPF más alta tienen otro pico previo de concentración a tiempo aproximado de $\exp(-4.2)$. Este nuevo pico es el que explica la primera CPF y es el rasgo que nos permite diferenciar los grupos de tratamiento en Docetaxel.

Si fijamos la primera CPF y recorremos de negativo a positivo la segunda CPF, veremos que ésta discrimina entre funciones que casi uniformemente tienen mayor concentración. Es decir, fijado si tiene uno o dos picos, las que tengan una segunda CPF mayor estarán por encima de las que tengan una segunda CPF menor. A continuación mostramos unos ejemplos en donde podemos ver representados mejor estos aspectos:



Figuras 7.17-7.20: Comparación de funciones con diferentes correlaciones respecto las CPFs de Docetaxel

Las dos primeras figuras son ejemplos de comparación entre valores extremos de la primera CPF fijada la segunda. Vemos que la primera CPF explica si aparece o no un pico previo de concentración. Las dos figuras de abajo comparan valores extremos de la segunda CPF, fijada

la primera. Se observa que la segunda CPF está discriminando entre funciones con más o menos concentración a lo largo del tiempo.

7.4.3.- Estudio del Paclitaxel.

Al igual que en el estudio de Docetaxel, hemos unido los dos grupos de Paclitaxel en una sola lista y hemos calculado las CPF's. En el gráfico de abajo vemos que la representatividad de las dos primeras CPF's es del 92,5%.

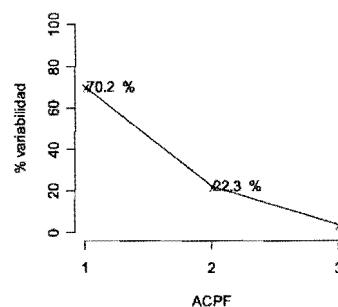


Figura 7.21: Representatividad de las tres primeras CPFs.

ESQUEMA DP			ESQUEMA PD		
PACIENTE	CPF.1	CPF.2	PACIENTE	CPF.1	CPF.2
160DP	---	---	160PD	0.87	-0.33
170DP	0.02	0.88	170PD	0.84	-0.35
178DP	0.08	0.90	178PD	0.87	-0.31
330DP	-0.97	-0.19	330PD	0.90	-0.36
369DP	0.07	0.94	369PD	0.93	-0.36
492DP	-0.95	-0.29	492PD	0.86	-0.37
507DP	-0.97	-0.21	507PD	0.89	-0.35
510DP	-0.98	-0.20	510PD	0.89	-0.37
542DP	-0.98	0.09	542PD	0.87	-0.34
759DP	-0.99	-0.06	759PD	0.89	-0.37
784DP	-0.96	-0.16	784PD	0.79	-0.34
830DP	0.13	0.93	830PD	0.86	-0.36
833DP	0.27	0.85	833PD	0.80	-0.27
849DP	-0.06	0.91	849PD	0.91	-0.37

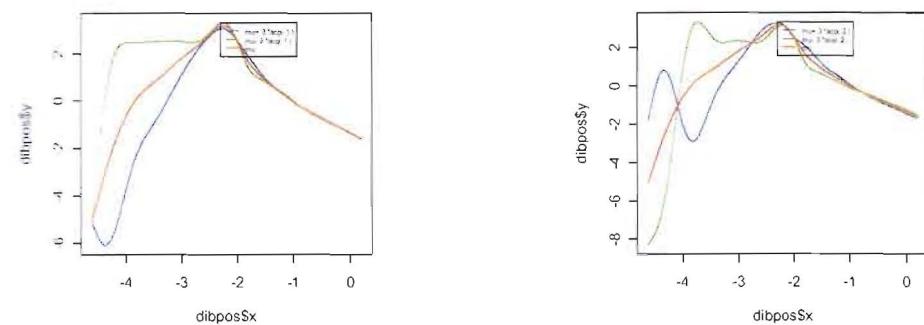
Tabla 7.9: Scores de las funciones para Paclitaxel.

ESQUEMA DP			ESQUEMA PD		
PACIENTE	CPF.1	CPF.2	PACIENTE	CPF.1	CPF.2
868DP	-0.86	0.39	868PD	0.88	-0.38
871DP	-0.96	-0.23	871PD	0.88	-0.36
878DP	-0.02	0.94	878PD	0.91	-0.33
883DP	---	---	883PD	0.88	-0.38
894DP	-0.01	0.86	894PD	0.89	-0.37
954DP	-0.98	-0.08	954PD	0.94	-0.28

Tabla 7.9: (cont) Scores de las funciones para Paclitaxel.

En la tabla anterior (tabla 7.9) vemos cómo las funciones del esquema PD se asocian positivamente a la primera CPF y tiene valores semejantes en la segunda CPF.

El esquema DP en ocasiones se asocia negativamente y en ocasiones esta correlación es nula. Observando la muestra vemos que cuando la primera CPF está cerca de -1, la segunda tiene valores cercanos a cero, mientras que cuando la primera CPF tiene valores cercanos a cero, la segunda tiene valores próximos a +1. Por lo tanto parece haber dos grupos dentro del esquema, hecho que se aprecia mejor en el gráfico de la siguiente página. Con los gráficos de las CPFs en función de cómo afectan a la media podremos interpretar las características de cada subgrupo:



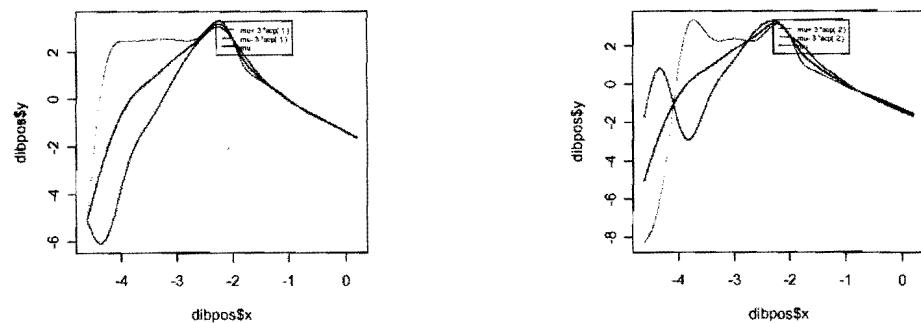
Figuras 7.22 y 7.23: Representación de las CPF's en función de cómo afectan a la media par a Paclitaxel.

ESQUEMA DP			ESQUEMA PD		
PACIENTE	CPF.1	CPF.2	PACIENTE	CPF.1	CPF.2
868DP	-0.86	0.39	868PD	0.88	-0.38
871DP	-0.96	-0.23	871PD	0.88	-0.36
878DP	-0.02	0.94	878PD	0.91	-0.33
883DP	---	---	883PD	0.88	-0.38
894DP	-0.01	0.86	894PD	0.89	-0.37
954DP	-0.98	-0.08	954PD	0.94	-0.28

Tabla 7.9: (cont) Scores de las funciones para Paclitaxel.

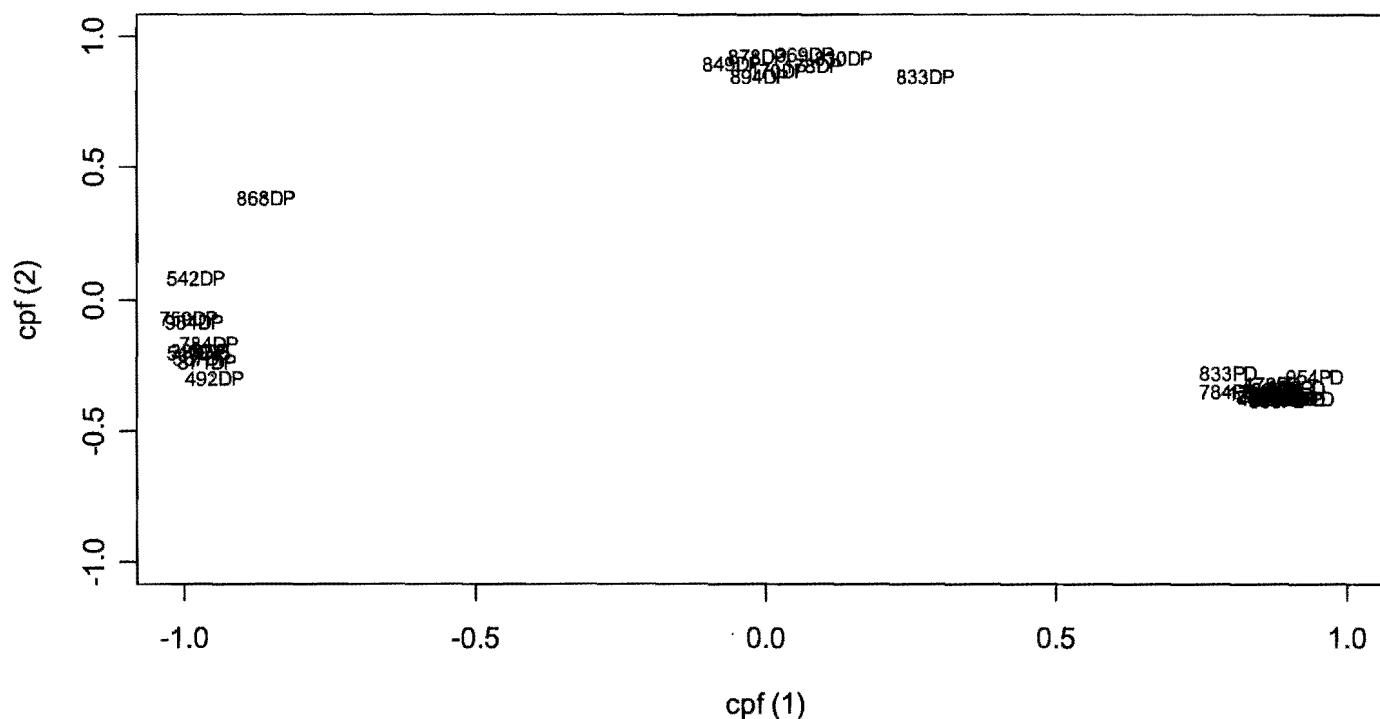
En la tabla anterior (tabla 7.9) vemos cómo las funciones del esquema PD se asocian positivamente a la primera CPF y tiene valores semejantes en la segunda CPF.

El esquema DP en ocasiones se asocia negativamente y en ocasiones esta correlación es nula. Observando la muestra vemos que cuando la primera CPF está cerca de -1, la segunda tiene valores cercanos a cero, mientras que cuando la primera CPF tiene valores cercanos a cero, la segunda tiene valores próximos a +1. Por lo tanto parece haber dos grupos dentro del esquema, hecho que se aprecia mejor en el gráfico de la siguiente página. Con los gráficos de las CPFs en función de cómo afectan a la media podremos interpretar las características de cada subgrupo:



Figuras 7.22 y 7.23: Representación de las CPF's en función de cómo afectan a la media par a Paclitaxel.

Correlaciones entre CPFs y muestra con Paclitaxel

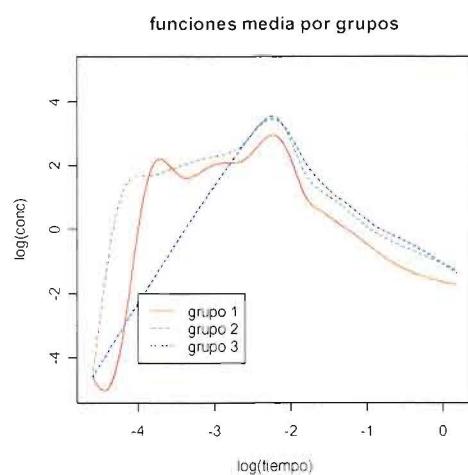


Figuras 7.24: Representación de las correlaciones en el planode las CPF's para Paclitaxel

En las figuras 7.12, vemos que en general las CPFs explican la variabilidad desde el tiempo inicial hasta tiempo -2.25 , ya que no hay gran variabilidad a partir de entonces. Las funciones con la primera CPF positiva tendrán en ese intervalo una concentración inferior respecto de la función media, yéndose al máximo (a tiempo $\exp(-2.25)$ aprox.) de manera rectilínea. Sin embargo las funciones con la primera CPF negativa tendrán una concentración superior a la media, por lo que tendrán un pico de concentración mucho antes del máximo. Las funciones con una primera CPF cercana a cero se parecerán a la media en ese tramo. La segunda CPF cuando es negativa nos está indicando que se produce un pico antes de tiempo $\exp(-4)$, mientras que cuando es negativa ese pico se produce tras el tiempo $\exp(-4)$.

Si interpretamos ahora el gráfico de la nube de puntos, vemos que hay tres grupos.

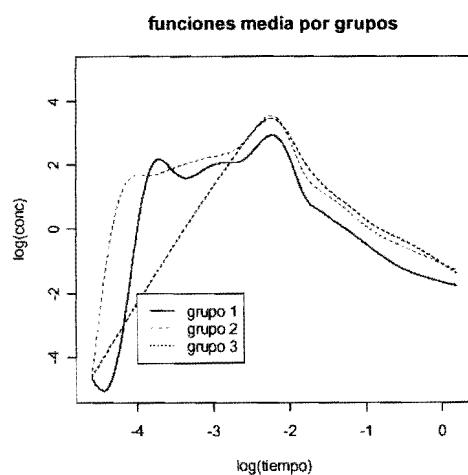
- Grupo 1: Esquema PD, con valores de la primera CPF positivos y valores en la segunda CPF con aproximadamente -0.3 . Serán funciones con una subida rectilínea y sin apenas picos hasta el máximo.
- Grupo 2: Esquema DP, con valores de la primera CPF cercanos a -1 y valores en la segunda CPF con aproximadamente cero. Serán funciones que suben rápidamente antes de tiempo -4 y luego se mantiene constante hasta el tiempo -2.25 .
- Grupo 3: Esquema DP con valores de la primera CPF cercanos a cero y valores en la segunda CPF cercanos a uno. Serán funciones que suben como la media hasta el -4 , y a partir de ahí hacen un pico previo al tiempo -2.25 .



En las figuras 7.12, vemos que en general las CPFs explican la variabilidad desde el tiempo inicial hasta tiempo -2.25 , ya que no hay gran variabilidad a partir de entonces. Las funciones con la primera CPF positiva tendrán en ese intervalo una concentración inferior respecto de la función media, yéndose al máximo (a tiempo $\exp(-2.25)$ aprox.) de manera rectilínea. Sin embargo las funciones con la primera CPF negativa tendrán una concentración superior a la media, por lo que tendrán un pico de concentración mucho antes del máximo. Las funciones con una primera CPF cercana a cero se parecerán a la media en ese tramo. La segunda CPF cuando es negativa nos está indicando que se produce un pico antes de tiempo $\exp(-4)$, mientras que cuando es negativa ese pico se produce tras el tiempo $\exp(-4)$.

Si interpretamos ahora el gráfico de la nube de puntos, vemos que hay tres grupos.

- Grupo 1: Esquema PD, con valores de la primera CPF positivos y valores en la segunda CPF con aproximadamente -0.3 . Serán funciones con una subida rectilínea y sin apenas picos hasta el máximo.
- Grupo 2: Esquema DP, con valores de la primera CPF cercanos a -1 y valores en la segunda CPF con aproximadamente cero. Serán funciones que suben rápidamente antes de tiempo -4 y luego se mantiene constante hasta el tiempo -2.25 .
- Grupo 3: Esquema DP con valores de la primera CPF cercanos a cero y valores en la segunda CPF cercanos a uno. Serán funciones que suben como la media hasta el -4 , y a partir de ahí hacen un pico previo al tiempo -2.25 .



8. – CONCLUSIONES DEL PROYECTO.

Las conclusiones de este proyecto son las siguientes:

- Este análisis es aplicable a todo objeto de estudio que percibamos como funciones. El análisis de datos funcionales consigue tratar como funciones aquellos datos que, tras la observación o experimentación, se perciben con forma de función.
- Estas técnicas sintetizan mucho más la información que si escogemos la estrategia de tratar las funciones de manera discreta. Los estadísticos descriptivos resumen mejor las funciones que si las tratamos de manera discreta (ejemplo de tasas por edad y sexo en pirámides de población) donde en ocasiones se necesitan índices de síntesis.
- El análisis de componentes principales funcionales es un complemento necesario al análisis descriptivo funcional, ya que tiene en cuenta la forma de las funciones y sus diferencias entre intervalos de tiempo.
- La etapa de registro es un punto débil ya que si ésta se lleva a cabo de manera incorrecta, los estadísticos descriptivos pueden llegar a variar mucho, sobre todo los de primer nivel.

Posibles mejoras:

- El hecho de trabajar con un espacio generador para splines y no con una base, puede ser uno de los motivos por los cuales nos encontramos con problemas de singularidad a la hora de calcular componentes principales. Trabajar con B-splines podría ser una solución.
- La creación de una librería con estos procedimientos podría facilitar el entorno de trabajo.

Posibles trabajos futuros:

- Se puede seguir profundizando en cada una de las aplicaciones que se han mostrado en estos los anteriores capítulos.

BIBLIOGRAFÍA

1. Ramsay, J.O. & Silverman, B.W. (1997). *Functional Data Analysis*. Springer. New York.
2. Green, P.J & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, London
3. Lange, K.(1998). *Numerical Analysis for Statisticians*. Springer. New York.
4. Bonet, C.; Jorba, A.; Martínez-Seara, M.T.; Masdemont, J.; Ollé M.; Susin A.; Valencia M. (1994) *Càlcul numèric*. Edicions UPC. Barcelona.
5. R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL <http://www.R-project.org>.
6. Kneip, A. & Utikal K.J. (June 2001) *Inference for Density Families Using Functional Principal Component Analysis*. Journal of the American Statistical Association, Vol. 96, Nº.454.
7. Locantore, N. Marron, J.S. Simpson, D.G. Tripoli, N. Zhang, J.T. Cohen, K.L. (1999). *Robust principal component análisis for functional data*. Test-Sociedad de Estadística e Investigación Operativa. Vol 8, Nº. 1, pp 1-73.

APÉNDICE

1.- DETALLES DE CÁLCULOS.

1.1.- Cálculo de spline por regresión.

a) En el apartado 3.2. se hace referencia a que:

$$\int_{x_i}^{x_{i+1}} s''(x)^2 dx = \frac{1}{b_i} \int_{x_i}^{x_{i+1}} [M_i(x_{i+1} - x) + M_{i+1}(x - x_i)]^2 dx = \frac{b_i}{3} (M_i^2 + M_i M_{i+1} + M_{i+1}^2)$$

• Veamos que realmente es así:

$$\begin{aligned} & \frac{1}{b_i} \int_{x_i}^{x_{i+1}} [M_i(x_{i+1} - x) + M_{i+1}(x - x_i)]^2 dx = \\ &= \frac{1}{b_i} \int_{x_i}^{x_{i+1}} \{M_i^2(x_{i+1} - x)^2 + M_{i+1}^2(x_i - x)^2 + 2M_i(x_{i+1} - x)M_{i+1}(x_i - x)\} dx = \\ &= \frac{M_i^2}{b_i^2} \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2 dx + \frac{M_{i+1}^2}{b_i^2} \int_{x_i}^{x_{i+1}} (x_i - x)^2 dx - \frac{2M_i M_{i+1}}{b_i^2} \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x_i - x) dx = \\ &= \frac{M_i^2}{b_i^2} a + \frac{M_{i+1}^2}{b_i^2} b + \frac{2M_i M_{i+1}}{b_i^2} c \end{aligned}$$

• Calcularemos por separado a, b y c :

$$\begin{aligned} a &= -\frac{(x_{i+1} - x)^3}{3} \Big|_{x_i}^{x_{i+1}} = \frac{(x_{i+1} - x_i)^3}{3} = \frac{b_i^3}{3} \\ b &= \frac{(x_i - x)^3}{3} \Big|_{x_i}^{x_{i+1}} = \frac{(x_{i+1} - x_i)^3}{3} = \frac{b_i^3}{3} \\ c &= \int_{x_i}^{x_{i+1}} (x_{i+1}x_i - xx_{i+1} - xx_i + x^2) dx = x_{i+1}x_i(x_{i+1} - x_i) - x_{i+1}\frac{x^2}{2} \Big|_{x_i}^{x_{i+1}} - x_i\frac{x^2}{2} \Big|_{x_i}^{x_{i+1}} + \frac{x^3}{3} \Big|_{x_i}^{x_{i+1}} = \\ &= x_{i+1}^2x_i - x_{i+1}x_i^2 - \frac{x_{i+1}}{2}[x_{i+1}^2 - x_i^2] - \frac{x_i}{2}[x_{i+1}^2 - x_i^2] + \frac{x_{i+1}^3}{3} - \frac{x_i^3}{3} = \\ &= x_{i+1}^3 \left(\frac{1}{3} - \frac{1}{2}\right) + x_{i+1}^2x_i \left(1 - \frac{1}{2}\right) + x_{i+1}x_i^2 \left(-1 + \frac{1}{2}\right) + x_i^3 \left(-\frac{1}{3} + \frac{1}{2}\right) = \\ &= x_{i+1}^3 \left(-\frac{1}{6}\right) + x_{i+1}^2x_i \left(\frac{1}{2}\right) + x_{i+1}x_i^2 \left(-\frac{1}{2}\right) + x_i^3 \left(\frac{1}{6}\right) = \left(-\frac{1}{6}\right)(x_{i+1} - x_i)^3 = -\frac{1}{6}b_i^3 \end{aligned}$$

- Por lo tanto tenemos que

$$\int_{x_i}^{x_{i+1}} s''(x)^2 dx = \frac{M_i^2}{b_i^2} \cdot \frac{b_i^3}{3} + \frac{M_{i+1}^2}{b_i^2} \cdot \frac{b_i^3}{3} - 2M_i M_{i+1} \left(-\frac{1}{6} \right) b_i^3 = \frac{b_i}{3} (M_i^2 + M_{i+1}^2 + M_i M_{i+1})$$

b) En el mismo apartado se hace referencia a la siguiente igualdad:

$$\sum_{i=0}^{n-1} \frac{1}{3} b_i (M_i^2 + M_{i+1}^2 + M_i M_{i+1}) = \frac{1}{6} \sum_{i=1}^{n-1} b_{i-1} M_{i-1} M_i + 2M_i^2 (b_{i-1} + b_i) + b_i M_i M_{i+1}$$

- Esto se comprueba desarrollando el sumatorio:

$$\begin{aligned} \sum_{i=0}^{n-1} \frac{1}{3} b_i (M_i^2 + M_{i+1}^2 + M_i M_{i+1}) &= \frac{1}{6} \sum_{i=0}^{n-1} 2b_i (M_i^2 + M_{i+1}^2 + M_i M_{i+1}) = \\ &= \frac{1}{6} [2b_0 M_0^2 + 2b_0 M_1^2 + b_0 M_0 M_1 + b_0 M_0 M_1 + 2b_1 M_1^2 + 2b_1 M_2^2 + b_1 M_1 M_2 + b_1 M_1 M_2 + \\ &+ 2b_2 M_2^2 + 2b_2 M_3^2 + b_2 M_2 M_3 + b_2 M_2 M_3 + \dots] = \begin{pmatrix} \text{recolocamos} \\ \text{términos} \end{pmatrix} = \\ &= \frac{1}{6} [(2b_0 M_0^2 + b_0 M_0 M_1) + (b_0 M_0 M_1 + 2M_1^2 \{b_0 + b_1\} + b_1 M_1 M_2) + (b_1 M_1 M_2 + 2M_2^2 \{b_1 + b_2\} + b_2 M_2 M_3) + \dots] \\ &\left(\begin{array}{l} \text{considerando} \\ M_0 = 0 \text{ y } M_n = 0 \end{array} \right) = \frac{1}{6} \sum_{i=1}^{n-1} b_{i-1} M_{i-1} M_i + 2M_i^2 (b_{i-1} + b_i) + b_i M_i M_{i+1} \end{aligned}$$

1.2.- Cálculo de la matriz W.

Recordemos que $W = (w_{k_1, k_2})_{k_1, k_2} = \left(\int_R \phi_{k_1}(t) \psi_{k_2}(t) dt \right)_{k_1, k_2}$

El cálculo de W en base de spline puede expresarse como la integral de cada uno de los elementos de la matriz resultante del producto de vectores siguiente:

$$W = \int \phi \phi' = \left(\int_a^b \phi_i(t) \phi_j(t) dt \right)_{i,j}$$

Pero en realidad no hace falta calcular $n(p+1) \times n(p+1)$ elementos de la matriz ya que sabemos que es simétrica y que existen muchos productos de funciones $\phi_p(t) \phi_j(t) = 0$. En particular:

- $w_{i,j} = w_{j,i} = \int_a^b \phi_i(t) \phi_j(t) dt \text{ para } \forall i, j \in \{0, 1, \dots, n\}$

- $w_{i(i-1)+f, i(i-1)+c} = \int_{x_{i-1}}^{x_i} t^{f-i} I_{[x_{i-1}, x_i]}(t) t^{c-i} I_{[x_{i-1}, x_i]}(t) dt = \frac{t^{(f-i)(c-i)+1}}{(f-i)(c-i)+1} \Big|_{x_{i-1}}^{x_i} = \frac{x_i^{(f-i)(c-i)+1} - x_{i-1}^{(f-i)(c-i)+1}}{(f-i)(c-i)+1}$

donde $i \in \{1, \dots, n\}$, $f = \{1, 2, 3, 4\}$ y $0 \leq c \leq f$

- $w_{ij}=0$ en otros casos.

Observamos entonces que \mathcal{W} es una matriz formada por submatrices en la diagonal:

$$\left(\begin{array}{cccc} w_{1,1} & w_{1,2} & w_{1,3} & w_{1,4} \\ w_{2,1} & w_{2,2} & w_{2,3} & w_{2,4} \\ w_{3,1} & w_{3,2} & w_{3,3} & w_{3,4} \\ w_{4,1} & w_{4,2} & w_{4,3} & w_{4,4} \\ & \ddots & & \\ & w_{k,k} & w_{k,k+1} & w_{k,k+2} & w_{k,k+3} \\ & w_{k+1,k} & w_{k+1,k+1} & w_{k+1,k+2} & w_{k+1,k+3} \\ & w_{k+2,k} & w_{k+2,k+1} & w_{k+2,k+2} & w_{k+2,k+3} \\ & w_{k+3,k} & w_{k+3,k+1} & w_{k+3,k+2} & w_{k+3,k+3} \\ & \ddots & & \\ & w_{4n-3,4n-3} & w_{4n-3,4n-2} & w_{4n-3,4n-1} & w_{4n-3,4n} \\ & w_{4n-2,4n-3} & w_{4n-2,4n-2} & w_{4n-2,4n-1} & w_{4n-2,4n} \\ & w_{4n-1,4n-3} & w_{4n-1,4n-2} & w_{4n-1,4n-1} & w_{4n-1,4n} \\ & w_{4n,4n-3} & w_{4n,4n-2} & w_{4n,4n-1} & w_{4n,4n} \end{array} \right)$$

Y cada submatriz al integrar tiene la siguiente expresión ($k=4(i-1)+l$):

$$\left(\begin{array}{cccc} w_{k,k} & w_{k,k+1} & w_{k,k+2} & w_{k,k+3} \\ w_{k+1,k} & w_{k+1,k+1} & w_{k+1,k+2} & w_{k+1,k+3} \\ w_{k+2,k} & w_{k+2,k+1} & w_{k+2,k+2} & w_{k+2,k+3} \\ w_{k+3,k} & w_{k+3,k+1} & w_{k+3,k+2} & w_{k+3,k+3} \end{array} \right) = \left(\begin{array}{cccc} x_i - x_{i-1} & \frac{x_{i-1}^2 - x_i^2}{2} & \frac{x_{i-1}^3 - x_i^3}{3} & \frac{x_{i-1}^4 - x_i^4}{4} \\ \frac{x_{i-1}^2 - x_i^2}{2} & x_{i-1}^3 - x_i^3 & x_{i-1}^4 - x_i^4 & x_{i-1}^5 - x_i^5 \\ \frac{x_{i-1}^3 - x_i^3}{3} & \frac{x_{i-1}^4 - x_i^4}{4} & x_{i-1}^5 - x_i^5 & x_{i-1}^6 - x_i^6 \\ \frac{x_{i-1}^4 - x_i^4}{4} & \frac{x_{i-1}^5 - x_i^5}{5} & \frac{x_{i-1}^6 - x_i^6}{6} & \frac{x_{i-1}^7 - x_i^7}{7} \end{array} \right)$$

para todo $i \in \{1, \dots, n\}$. Programar el cálculo de \mathcal{W} para splines cúbicos es entonces fácil ya que se limita a programar simplemente estas $n-1$ submatrices por separado. Además nótese que aunque la submatriz se compone de 16 elementos sólo 7 de ellos son distintos. La matriz anterior si calculamos estos 7 elementos a parte sería:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 7 \end{pmatrix}$$

La subrutina `monta.w` calculará cada una de las $n-1$ submatrices aprovechándose de que sólo hay que calcular estos 7 elementos para cada una de ellas:

```
# Funcion monta.w
monta.w<-function(x,n,N)
{
  W<-matrix(0,nrow=4*(n-1),ncol=4*(n-1))
  for (i in 2:n)
  {
    aux<-x[i]-x[i-1]
    for (k in 2:7)
    {
      aux<-c(aux, ((x[i]**k-x[i-1]**k))/k)
    }
    W[4*(i-2)+1,(4*(i-2)+1):(4*(i-2)+4)]<-aux[1:4]
    W[4*(i-2)+2,(4*(i-2)+1):(4*(i-2)+4)]<-aux[2:5]
    W[4*(i-2)+3,(4*(i-2)+1):(4*(i-2)+4)]<-aux[3:6]
    W[4*(i-2)+4,(4*(i-2)+1):(4*(i-2)+4)]<-aux[4:7]
  }
  W
}

# FIN Funcion monta.w
```

2.- SISTEMAS DE ECUACIONES TRIDIAGONALES.

En este apartado se detalla cómo se soluciona un sistema de ecuaciones en el que la matriz asociada es una matriz tridiagonal. Esto nos será muy útil a la hora de calcular los momentos del spline interpolador (apartado 3.3) de manera eficiente. Empezaremos definiendo qué es una matriz tridiagonal:

2.1.- Matrices tridiagonales.

DEFINICION A.2.1.

Diremos que una matriz $(a_{ij})_{ij}$ de dimensiones $n \times n$ es *tridiagonal* si cumple que $a_{ij}=0$ para toda i,j tal que $|i-j| > 1$. Es decir:

$$\begin{pmatrix} a_{1,1} & a_{1,2} & 0 & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & a_{2,3} & 0 & \cdots & 0 \\ 0 & a_{3,2} & a_{3,3} & a_{3,4} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & a_{n,n-1} & a_{n,n} \end{pmatrix}$$

En un lenguaje de programación el almacenamiento de la matriz tridiagonal se reduce a un vector de n componentes en donde se guardar la diagonal (b) y dos vectores de $n-1$ de longitud para guardar las subdiagonales superior (c) e inferior (a). Es decir:

$$\begin{pmatrix} b_1 & c_1 & 0 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & \cdots & 0 \\ 0 & a_3 & b_3 & c_3 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & \cdots & 0 & a_n & b_n \end{pmatrix}$$

2.2.- Descomposición LU de una matriz tridiagonal.

La descomposición LU de una matriz permite simplificar los cálculos a la hora de encontrar la solución de cualquier sistema de ecuaciones. La idea consiste en descomponer la matriz de sistema en dos matrices de la siguiente forma:

$$A = LU = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_2 & 1 & 0 & \cdots & 0 \\ 0 & a_3 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_n & 1 \end{pmatrix} \begin{pmatrix} \beta_1 & c_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & c_2 & \cdots & 0 \\ 0 & 0 & \beta_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \beta_n \end{pmatrix}$$

Así hemos transformado un sistema de ecuaciones complejo en dos sistemas de ecuaciones muy fáciles de resolver, tal como se verá en el punto 2.3. Hallar las matrices LU es también relativamente sencillo ya que en el caso de A tridiagonal sólo se precisa calcular los coeficientes $\alpha_2, \dots, \alpha_n$ y β_1, \dots, β_n . Si multiplicamos LU obtendremos que:

$$A = LU = \begin{pmatrix} \beta_1 & c_1 & 0 & \cdots & 0 \\ \alpha_2 \beta_1 & \alpha_2 c_1 + \beta_2 & c_2 & \cdots & 0 \\ 0 & \alpha_3 \beta_2 & \alpha_3 c_2 + \beta_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & c_{n-1} \\ 0 & 0 & \cdots & \alpha_n \beta_{n-1} & \alpha_n c_{n-1} + \beta_n \end{pmatrix} = \begin{pmatrix} b_1 & c_1 & 0 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & 0 & \cdots & 0 \\ 0 & a_3 & b_3 & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & 0 & \cdots & 0 & a_n & b_n \end{pmatrix}$$

Podemos hallar α_i y β_i de manera recursiva si resolvemos las ecuaciones desde el primer elemento de LU y vamos bajando en escalera hasta el último. Es decir:

- $\beta_1 = b_1$ y $\alpha_2 \beta_1 = a_2 \Rightarrow \alpha_2 = \frac{a_2}{\beta_1}$
- $\alpha_2 c_1 + \beta_2 = b_2 \Rightarrow \beta_2 = b_2 - \alpha_2 c_1$ y $\alpha_3 \beta_2 = a_3 \Rightarrow \alpha_3 = \frac{a_3}{\beta_2}$
- Y por tanto para $i=2,\dots,n$:
 - (a) $\beta_i = b_i - \alpha_i c_{i-1}$
 - (b) $\alpha_i = \frac{a_i}{\beta_{i-1}}$

2.3.- Solución del sistema $LUX=b$

Partiendo de que ahora el sistema es $LUX=b$

1. Hallamos el vector g tal que: $Lg=b$.
2. Fijémonos que $Ux=g$. Por lo tanto resolvemos ahora el sistema $Ux=g$.

x será la solución de nuestro sistema de ecuaciones.

Paso 1.

Fijémonos el la forma de la matriz L (punto 2.2) y resolvamos el sistema $Lg=b$.

- la primera ecuación se resuelve directamente: $g_1=b_1$,
- La segunda ecuación es: $\alpha_2 g_1 + g_2 = b_2 \Rightarrow g_2 = b_2 - \alpha_2 g_1$,
- Por lo tanto la i -ésima ecuación ($i=2,\dots,n$) es: $\alpha_i g_{i-1} + g_i = b_i \Rightarrow g_i = b_i - \alpha_i g_{i-1}$

Paso 2.

Fijémonos el la forma de la matriz U (punto 2.2) y resolvamos el sistema $Ux=g$.

- La última ecuación se resuelve directamente: $\beta_n x_n = g_n \Rightarrow x_n = \frac{g_n}{\beta_n}$
- La penúltima ecuación es: $\beta_{n-1} x_{n-1} + c_{n-1} x_n = g_{n-1} \Rightarrow x_{n-1} = \frac{g_{n-1} - c_{n-1} x_n}{\beta_{n-1}}$
- Por lo tanto la i-ésima ecuación ($i=n-1, \dots, 1$) es: $\beta_i x_i + c_i x_{i+1} = g_i \Rightarrow x_i = \frac{g_i - c_i x_{i+1}}{\beta_i}$

Y con esto hemos resuelto el sistema de ecuaciones tridiagonal.

2.4.- Subrutina en R del cálculo de splines interpoladores.

Recordemos que para calcular el spline interpolador lo hacíamos mediante el método de los momentos. El sistema de ecuaciones que hay que resolver para calcular los momentos m_i tiene una matriz asociada del sistema de tipo tridiagonal. Por eso vamos a explotar los algoritmos explicados anteriormente para calcular estos m_i 's. Recordemos el sistema:

$$\begin{pmatrix} 2(b_0 + b_1) & b_1 & & & \\ b_1 & 2(b_1 + b_2) & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & b_{n-3} & 2(b_{n-3} + b_{n-2}) & b_{n-2} \\ & & & b_{n-2} & 2(b_{n-2} + b_{n-1}) \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \end{pmatrix} \Leftrightarrow Rm = d$$

donde $d_{i+1} = 6 \left(\frac{f_{i+2} - f_{i+1}}{h_{i+1}} - \frac{f_{i+1} - f_i}{h_i} \right)$

Función `splinecalc`:

Entrada: Vector de x's (ordenadas) e y's (abcisas).

Salida: Objeto de tipo spline

Los pasos del algoritmo son los siguientes:

- Definición de variables
- Ordenación del vector de ordenadas x's.
- Cálculo de los vectores h_i , b_i (en el algoritmo `ad[i]`), a_i y c_i (en el algoritmo como la matriz es simétrica $a_{i+1}=c_i$ i por eso sólo consideramos c_i llamado `adno[i]`) y d_i . Observese que todos los cálculos se realizan con una sola operación iterativa (for).
- Descomposición LU de la matriz: cálculo de alfa y beta, según el algoritmo del punto 2.2.
- Solución del sistema $Lg=d$ según el algoritmo del punto 2.3.

- Solución del sistema $Ux=g$ según el algoritmo del punto 2.4: Vector de momentos m_i .
- Cálculo de las derivadas primeras $b[i]$.(Cap 3.3)
- Cálculo de los coeficientes de los polinomios que forman el spline (Cap 3.3).

```
splinecalc<-function(x,y)
{
  # Calcula el spline cúbico natural mediante el método de los momentos
  # Definicion de variables

  # n:    Número de puntos
  # k:    Dimensión de la matriz del sistema de ecuaciones (A)
  # d:    Vector de valores independientes AM=d
  # h:    Incrementos de valor entre Xi y Xi+1
  # ad:   Elementos de la diagonal de A.
  # adno: Elementos de la subdiagonal y la superdiagonal de (A)
  # alfa: Elementos de la subdiagonal de L en la descomposicion A=LU
  # beta: Elementos de la diagonal de U en la descomposición A=LU
  # g:    Elementos de la superdiagonal de U en la descomposición A=LU
  # m:    Solución del sistema AM=d i valor de la 2a derivada
  # b:    Valor de la la derivada

  n<-length(x)
  k<-n-2
  d<-vector(mode="double", length=k)
  h<-vector(mode="double", length=k)
  ad<-vector(mode="double", length=k)
  adno<-vector(mode="double", length=k-1)
  alfa<-vector(mode="double", length=k)
  beta<-vector(mode="double", length=k)
  g<-vector(mode="double", length=n)
  m<-vector(mode="double", length=k)
  b<-vector(mode="double", length=n)

  ### Primero ordenamos los datos de menor a mayor x

  orden<-order(x)
  x<-x[orden]
  y<-y[orden]
  print(x);print(y)

  ### Calculo de las hi's, de las di's y de la matriz tridiagonal
  ### Como la matriz tridiagonal es simétrica los vectores
  ### diagonal superior y diagonal inferior son iguales

  ### Ojo: h[1]= h(i=0) y h[k] = h(i=k-1)
```

```

####          y[1]= y(i=0)  y  y[k+1]= y(i=k)
###          x[1]= x(i=0)  y  x[k+1]= x(i=k)
###          d[1]= d(i=0)  y  d[k+1]= d(i=k)
### Las dimensiones de la matriz del sistema son n-2 x n-2

h[1]<-x[2]-x[1]
for (i in 1:k)
{
  h[i+1]<-x[i+2]-x[i+1]
  ad[i]<-2*(h[i]+h[i+1])
  if (i<k) {adno[i]<-h[i+1]}
  d[i]<-((y[i+2]-y[i+1])/h[i+1]) - ((y[i+1]-y[i])/h[i])
}
d<-d*6

if(n>3)
{
  ### Descomposición LU para matrices tridiagonales simétricas

beta[1]<-ad[1]
for (i in 2:k)
{
  alfa[i]<-adno[i-1]/beta[i-1]
  beta[i]<-ad[i]-(alfa[i]*adno[i-1])
}

### Resolución Lg=d

g[1]<-d[1]
for (i in 2:k)
{
  g[i]<-d[i]-(alfa[i]*g[i-1])
}

### Resolución Ux=g

m[k]<-g[k]/beta[k]
for (i in 2:k-1)
{
  m[k-i]<-(g[k-i]-(adno[k-i]*m[k-(i-1)]))/beta[k-i]
}
if (n==3){m<-d/ad}
### Resolución de las Bi's

m<-c(0,m,0)
for (i in 1:n)
{

```

```

b[i]<-((y[i+1]-y[i])/h[i])-(m[i+1]-m[i])*(h[i]/6)-m[i]*(h[i]/2)
}

### Solucion del spline

s<-array(NA,dim=c(n-1,4))
for (i in 0:(n-2))
{
  s[i+1,]<-c((m[i+1]/2)*(x[i+1]^2)-((x[i+1]^3)*((m[i+2]-m[i+1])/(6*h[i+1]))-
b[i+1]*x[i+1]+y[i+1],
  (-x[i+1]*m[i+1])+((x[i+1]^2)/2)*(-(m[i+2]-m[i+1])/h[i+1]))+b[i+1]),
  (m[i+1]/2)-((3*x[i+1]*(m[i+2]-m[i+1]))/(6*h[i+1])),
  (m[i+2]-m[i+1])/(6*h[i+1]))
}
b[n]<-s[n-1,2]+2*s[n-1,3]*x[n]+3*s[n-1,4]*x[n]**2
s<-rbind(c(y[1]-(b[1]*x[1]),b[1],0,0),s,c(y[n]-(b[n]*x[n]),b[n],0,0))
r<-"f"
val<-list(x,s,r)
names(val)<-c("x","s","r")
val
}

```

2.4.- Subrutina en R del cálculo de splines regresión.

Se basa en la idea del capítulo 3.4.

```

splint.reg<-function(x,y,w=c(0),alfa=0.5)
{
# Calcula el spline no que minimiza la funcion
# J_alfa_(g)=suma de errores al cuadrado + integral segunda
# derivada de g
### FUNCIONES ADICIONALES

# Funcion errordatos

errordatos<-function(x,y,w,alfa)
#Comprueba que los datos sean apropiados
{
err<-0
if (length(x)!=length(y)|length(x)!=length(w)|length(y)!=length(w))
{
  print("longitud de los datos no corresponden")
  err<-err+1
}
if (alfa>1|alfa<0)
{
  err<-err+1
}

```

```

print("Alfa ha de estar entre 0 y 1")
;
err
;
# Funcion calc.hi
calc.hi<-function(equis,k)
{
hi<-vector(mode="double",length=k+1)
hi[1]<-equis[2]-equis[1]
for (i in 1:k)
{
  hi[i+1]<-equis[i+2]-equis[i+1]
}
hi
}
# Fin funcion calc.hi

# Funcion calc.ri
calc.ri<-function(equis,ene,hi)
{
rij<-matrix(0,nrow=ene-1,ncol=(ene-1))
for(i in 1:(ene-1))
{
  rij[i,i]<-(1/6)*2*(hi[i]+hi[i+1])
  if(i>1){rij[i,i-1]<-(1/6)*hi[i]}
  if(i<(ene-1)){rij[i,i+1]<-(1/6)*hi[i+1]}
}
rij
}
# Fin funcion calc.ri
# Funcion calc.qiant
calc.qiant<-function(equis,ene,hi)
{
qij<-matrix(0,nrow=ene-1,ncol=(ene+1))
for (i in 1:(ene-1))
{
  qij[i,i]<- (-1)*( (1/hi[i])+(1/hi[i+1]) )
  if(i>1){qij[i,i-1]<-(1/hi[i])}
  if(i<(ene-1)){qij[i,i+1]<-(1/hi[i+1])}
}
qij[ene-1,ene]<-(1/hi[ene-1])
qij
}
# Fin Funcion calc.qiant
3
# Funcion calc.qi
calc.qi<-function(equis,ene,hi)
{

```

```

qij<-matrix(0,nrow=ene-1,ncol=(ene+1))
for (i in 1:(ene-1))
{
  qij[i,i]<-(1/hi[i])
  qij[i,i+1]<- (-1)*(-(1/hi[i])+(1/hi[i+1])) 
  qij[i,i+2]<-(1/hi[i+1])
}
qij
}

# Fin Funcion calc.qi

### FIN FUNCIONES ADICIONALES

# PROGRAMA PRINCIPAL
#####
##### DEFINICION DE VARIABLES

n<-length(x)
k<-n-2
h<-vector(mode="double",length=k)
R<-matrix(NA,nrow=,ncol=((n-1)*4))
b<-vector(mode="double",length=n)

##### fin DEFINICION DE VARIABLES

if(w==c(0)){w<-rep((1/length(x)),length(x))}

nota<-errordatos(x,y,w,alfa)
if (nota==0)
{
  h<-calc.hi(x,k)
  R<-calc.ri(x,n-1,h)
  Q<-calc.qiant(x,n-1,h)

  Q<-calc.qi(x,n-1,h)

  SG<-solve( (alfa*R)+(-(1-alfa)*Q) %*% solve(diag(w)) %*% t(Q) )
  m<-SG %*% (alfa*Q) %*% y

  if (alfa==0)
  {
    yhat<-lm(y~x)$fitted.values
  }
  else
  {
    yhat<-y-((1-alfa)/alfa)*solve(diag(w)) %*% t(Q) %*% m
  }
}

```

```

#yhat<-alfa*solve(alfa*diag(w)+(1-alfa)*t(Q)%%solve(R)%%Q%%diag(w))%%y

### Resolución de las Bi's

m<-c(0,m,0)
for (i in 1:n)
{
  b[i]<-((yhat[i+1]-yhat[i])/h[i])-(m[i+1]-m[i])*(h[i]/6)-m[i]*(h[i]/2)
}

### Solucion del spline

s<-array(NA,dim=c(n-1,4))
for (i in 0:(n-2))
{
  s[i+1,]<-c((m[i+1]/2)*(x[i+1]^2)-((x[i+1]^3)*((m[i+2]-m[i+1])/(6*h[i+1]))-
  b[i+1]*x[i+1]+yhat[i+1],
    (-x[i+1]*m[i+1])+((x[i+1]^2)/2)*((m[i+2]-m[i+1])/h[i+1])+b[i+1]),
  (m[i+1]/2)-((3*x[i+1]*(m[i+2]-m[i+1]))/(6*h[i+1])),
  ((m[i+2]-m[i+1])/(6*h[i+1]))
)
b[n]<-s[n-1,2]+2*s[n-1,3]*x[n]+3*s[n-1,4]*x[n]**2
s<-rbind(c(yhat[1]-(b[1]*x[1]),b[1],0,0),s,c(yhat[n]-(b[n]*x[n]),b[n],0,0))
x<-"f"
val<-list(x,s,r)
names(val)<-c("x","s","r")
val

}
}

```

3 - OTROS ALGORITMOS EN R.

3.1.- Algoritmo mutpol.

```

Mutpol<-function(plx,ply)
{
  # Calculo de x(t)*y(t)
  pol2<-vector(mode="double",length=((2*length(plx))-1))
  if(length(plx)==length(ply))
  {
    lg<-length(plx)
    indf<-1;indc<-0
    while (indf<=lg)
    {

```

```

fila<-indf
columna<-1
while (fila>=1)
{
  pol2[indf+indc]<-pol2[indf+indc]+(plxfila]*ply[columna])
  fila<-fila-1
  columna<-columna+1
}
indf<-indf+1
}
indf<-indf-1
indc<-2
while(indc<=lg)
{
  fila<-lg
  columna<-indc
  while(columna<=lg)
  {
    pol2[indf+indc-1]<-pol2[indf+indc-1]+(plxfila]*ply[columna])
    fila<-fila-1
    columna<-columna+1
  }
  indc<-indc+1
}
}
else
{
  print("grado polinomios no coincidentes")
}
pol2
}

```

3.2.- Algoritmo int.pl.

```

### FUNCION intPL
int.PL<-function(pol,equis)
{
# Integra el polinomio a0+a1*x+a2*x^2+...+an*x^(n) y lo evalua en x
# Los coeficientes del polinomio son: c(a0,a1,...,an)
  res<-pol[1]*equis
  for (i in 1:(length(pol)-1))
  {
    res<-res+((equis^(i+1))*pol[i+1])/(i+1)
  }
  res
}

```

```
*** FIN FUNCION intPL
```

3.3.- Algoritmo med.f.sint.

```
med.f.sint<-function(spl)
{
  # Calcula el valor medio de una funcion
  # INPUT: spl-> Objeto de tipo spline (registrado o no)
  # OUTPUT: xmed-> media de la funcion spl.
  # PARAMETROS: n->numero de knots
  ### FUNCIONES ADICIONALES esspline; int.pl:
  ### FUNCION PRINCIPAL

  if (!esspline(spl) !=0)
  {
    print('Este objeto no es un spline interpolador')
    res<-0
  }
  else
  {
    n<-length(spl$x)
    res<-0
    for (i in 2:n)
    {
      res<-res+(int.pl(spl$s[i],spl$x[i])-
                 int.pl(spl$s[i],spl$x[i-1]))
    }
    res<-(1/(spl$x[n]-spl$x[1]))*res
  }
  res
}
```

3.4.- Algoritmo var.f.sint

```
var.f.sint<-function(sint)
{
  # Calcula la varianza de la funcion que expresamos como
  # spline interpolador

  ### FUNCION consPL
  consPL<-function(pol,cte)
  {
    val<-pol
    val[1]<-val[1]-cte
```

```

val
;

*** FIN FUNCION consPL
# Otras funciones: esspline, int.pl, polquad
# PROGRAMA PRINCIPAL
if esspline(sint)==0
{
    print("es spline interpolador")
    n<-length(sint$x)
    zfun<-med.f.sint(sint)
    res<-0
    for (i in 2:n)
    {
        polaux<-consPL(sint$s[i,],mfun)
        polaux<-polquad(polaux)
        res<-res+(int.pl(polaux,sint$x[i])-int.pl(polaux,sint$x[i-1]))
    }
    res<-res/(sint$x[n]-sint$x[1])
}
else
{
    print("no es spline interpolador")
}
res
;

```

3.5.- Algoritmo covar.f.sint

```

Covar.f.sint<-function(lisspl,spx,spy)
{
*** PROGRAMA PRINCIPAL
splx<-list(lisspl$x,lisspl$s[[spx]],'t')
names(splx)<-c('x','s','r')
sply<-list(lisspl$x,lisspl$s[[spy]],'t')
names(sply)<-c('x','s','r')
mfunx<-med.f.sint(splx)
mfuny<-med.f.sint(sply)
n<-length(splx$x)
res<-0
for (i in 2:n)
{
    polauxx<-consPL(splx$s[i,],mfunx)
    polauxy<-consPL(sply$s[i,],mfuny)

```

```

polaux<-mutpol(polauxx,polauxy)
res<-res+(int.pl(polaux,splx$x[i])-int.pl(polaux,splx$x[i-1]))
}
res<-res/(splx$x[n]-splx$x[1])
res
}
### FIN FUNCION
;

```

3.6.- Algoritmo mom.1.

```

mom.1<-function(spllisreg)
{
### Calcula el spline del momento 1 de una
### lista de splines ya registrados.
if (spllisreg$r=="t")
{
  N<-length(spllisreg$s)
  aux<-spllisreg$s[[1]]
  for (i in 2:N)
  {
    aux<-aux+spllisreg$s[[i]]
  }
  val<-list(spllisreg$x,aux,"t")
  names(val)<-c('x','s','r')
}
else
{
  val<-"LISTA DE SPLINES NO REGISTRADA"
}
val
;

```

3.7.- Algoritmo mom.2.

```

mom.2<-function(spllisreg)
{
### Calcula el spline del momento 2 de una
### lista de splines ya registrados.

### FUNCIONES ADICIONALES

# FUNCION MUTPOL

mutpol<-function(plx,ply)
{

```

```

# Calculo de x(t)*y(t)
pol2<-vector(mode="double",length=((2*length(plx))-1))
if(length(plx)==length(ply))
{
  lg<-length(plx)
  indf<-1;indc<-0
  while (indf<=lg)
  {
    fila<-indf
    columnna<-1
    while (fila>=1)
    {
      pol2[indf+indc]<-pol2[indf+indc]+(plx[fila]*ply[columnna])
      fila<-fila-1
      columnna<-columnna+1
    }
    indf<-indf+1
  }
  indf<-indf-1
  indc<-2
  while(indc<=lg)
  {
    fila<-lg
    columnna<-indc
    while(columnna<=lg)
    {
      pol2[indf+indc-1]<-pol2[indf+indc-1]+(plx[fila]*ply[columnna])
      fila<-fila-1
      columnna<-columnna+1
    }
    indc<-indc+1
  }
}
;
else
{
  print("grado polinomios no coincidentes")
}
pol2
;

# FIN FUNCION MUTPOL

# FUNCION SPLQUAD
#     Calcula el cuadrado de un spline
#     INPUT: Coeficientes de un spline
#     OUTPUT: Coeficientes del cuadrado del spline input
#             El spline de salida es de grado 6

```

```

splquad<-function(splcoef)
{
  res<-array(NA,dim=c(nrow(splcoef),2*ncol(splcoef)-1))
  for(i in 1:nrow(splcoef))
  {
    res[i,]<-mutpol(splcoef[i,],splcoef[i,])
  }
  print(dim(splcoef))
  print(res)
}

# FIN FUNCION SPLQUAD

# FIN FUNCIONES ADICIONALES

# PROGRAMA PRINCIPAL

if (spllisreg$r=="t")
{
  N<-length(spllisreg$s)
  aux<-splquad(spllisreg$s[[1]])
  for (i in 2:N)
  {
    aux<-aux+splquad(spllisreg$s[[i]])
  }
  val<-list(spllisreg$x,aux,"t")
  names(val)<-c('x','s','r')
}
else
{
  val<-"LISTA DE SPLINES NO REGISTRADA"
}
val
}

```

3.8.- Algoritmo desc.f.spls.

```

desc.f.spls<-function(spllis,k=2,grnp=750)
{
##### FUNCION desc.f.spl

##### Calcula e imprime por pantalla:
#   1. Para cada funcion su valor medio y varianza.
#   2. La covarianza entre cada una de las funciones.
#   3. La funcion media de la muestra (fmu)
#   4. La funcion varianza de la muestra (fsigma)

```

```

#      5. Representa graficamente fm+/- k*fstdev donde
#          k es un parámetro de la función. Por defecto es 2.

##### Esta función precisa de que estén en la misma carpeta
#      las funciones med.f.sint, var.f.sint, covar.f.sint,mom.1 y mom.2

##### FUNCIONES ADICIONALES

##### FUNCION eslistaspl
eslistaspl<-function(lispos)
{
  # Comprueba que el argumento de la función
  # sea una lista de splines
  res<-T
  noms<-c('x','s','r')
  if(is.null(names(lispos)))
  {
    res<-F
    print('no es spline: ')
  }
  else
  {
    if(length(names(lispos))!=length(noms)){res<-F}
    else
    {
      if(sum(ifelse(names(lispos)==noms,1,0))!=length(noms)){res<-F}
    }
  }
  res
}

#####
##### FIN

# FUNCION MUTPOL

mutpol<-function(plx,ply)
{
  # Calculo de x(t)*y(t)
  pol2<-vector(mode="double",length=((2*length(plx))-1))
  if(length(plx)==length(ply))
  {
    lg<-length(plx)
    indf<-1;indc<-0
    while (indf<=lg)
    {
      fila<-indf
      columna<-1
      while (fila>=1)

```

```

{
  pol2[indf+indc]<-pol2[indf+indc]+(plx[fila]*ply[columna])
  fila<-fila-1
  columna<-columna+1
}
indf<-indf+1
}
indf<-indf-1
indc<-2
while(indc<=lg)
{
  fila<-lg
  columna<-indc
  while(columna<=lg)
  {
    pol2[indf+indc-1]<-pol2[indf+indc-1]+(plx[fila]*ply[columna])
    fila<-fila-1
    columna<-columna+1
  }
  indc<-indc+1
}
;
else
{
  print("grado polinomios no coincidentes")
}
pol2
;

# FIN FUNCION MUTPOL

# FUNCION SPLQUAD
#   Calcula el cuadrado de un spline
#   INPUT: Coeficientes de un spline
#   OUTPUT: Coeficientes del cuadrado del spline input
#           El spline de salida es de grado 6

splquad<-function(splcoef)
{
  res<-array(NA,dim=c(nrow(splcoef),2*ncol(splcoef)-1))
  for(i in 1:nrow(splcoef))
  {
    res[i,]<-mutpol(splcoef[i,],splcoef[i,])
  }
  print(dim(splcoef))
  print(res)
}

```

```

# FIN FUNCION SPLQUAD

#### FIN FUNCIONES ADICIONALES

#### PROGRAMA PRINCIPAL

### Definicion de variables

N<-length(spllis$s)
dfun<-array(NA,dim=c(N, 2))
dcofun<-array(0, dim=c(N,N))

### Fin definicion de variables

if (eslistaspl(spllis))
{
  # Calculo de estadisticos para cada una de las muestras

  colnames(dfun)<-c('media', 'varianza')
  rownames(dfun)<-rep('f', N)
  colnames(dcofun)<-rep('f', N)
  rownames(dcofun)<-rep('f', N)
  for (i in 1:N)
  {
    spaux<-list(spllis$x, spllis$s[[i]], spllis$r)
    names(spaux)<-c('x', 's', 'r')
    dfun[i,1]<-med.f.sint(spaux)
    dfun[i,2]<-var.f.sint(spaux)
    rownames(dfun)[i]<-paste('f', i, sep="")
    rownames(dcofun)[i]<-paste('f', i, sep="")
    colnames(dcofun)[i]<-paste('f', i, sep="")
    for (j in 1:i)
    {
      dcofun[i,j]<-covar.f.sint(spllis, i, j)
      dcofun[j,i]<-covar.f.sint(spllis, i, j)
    }
  }

  # Calculo de la funcion media

  fmu<-mom.l(spllis)
  fmu$s<-fmu$s/N

  # Calculo de la funcion varianza

```

```

m2<-mom.2(spllis)
fmu2<-fmu
fmu2$s<-splquad(fmu2$s)
fvar<-fmu2
fvar$s<-(l/N)*(m2$s-(N*fmu2$s))

# Imprimir resultados

puntsfmu<-dibuspline(fmu,gr=F,npunts=grnp)
puntsvar<-dibuspline(fvar,gr=F,npunts=grnp)
puntstdev<-list(puntsvar$x,sqrt(puntsvar$y))
names(puntstdev)<-c('x','y')

grlim<-c(min(puntsfmu$y-3*puntstdev$y),
          max(puntsfmu$y+3*puntstdev$y))

plot(puntsfmu$x,puntsfmu$y,xlim=range(fmu$x),col=2,
      ylim=grlim,pch='.',type='p',xlab='x',
      ylab=paste('fmu+/-',k,'* fstdev'))

points(puntsfmu$x,puntsfmu$y-k*puntstdev$y,pch='.',col=3)
points(puntsfmu$x,puntsfmu$y+k*puntstdev$y,pch='.',col=3)

for (i in 1:N)
{
aux3<-list(spllis$x,spllis$s[[i]],spllis$r)
names(aux3)<-c('x','s','r')
aux4<-dibuspline(aux3,gr=F,npunts=grnp)
points(aux4$x,aux4$y,pch='.')
}
points(puntsfmu$x,puntsfmu$y,pch='.',col=2)
points(puntsfmu$x,puntsfmu$y-k*puntstdev$y,pch='.',col=3)
points(puntsfmu$x,puntsfmu$y+k*puntstdev$y,pch='.',col=3)

# Salida de resultados

res<-list(dfun,dcofun,fmu,fvar)
names(res)<-c('dfun','dcofun','fmu','fvar')
}
else {
  print('no es una lista de splines registrados')
  res<-NULL
}
res
}

```

3.9.- Algoritmo de cálculo de componentes principales funcionales

Código fuente:

```
fPCA.sint<-function(lreg)
{
  ### Calcula las componentes principales funcionales
  #   donde cada una de las funciones ha sido representadas
  #   como un spline interpolador o regresion. Lo realiza diagonalizando
  #   A=(1/N)*sqrtW*C'C*sqrtW (Solucion de Ab=lb)
  #   que es equivalente al problema real A=(1/N)*C'*C*W*b=lb

  #   VARIABLES INPUT
  #   lreg: lista de splines centrados y
  #         registrados previamente.

  #   VARIABLES DE TRABAJO

  #   n<-numero de knots
  #   N<-numero de splines (tamaño de la muestra)
  #   C<-matriz de coeficientes
  #   W<-matriz de producto escalar de polinomios integrados

  ### FUNCIONES ADICIONALES

  # Funcion sint.pca
  sint.pca<-function(veps,x,N,n)
  {
    # Estructura los coeficientes de las fPCA en forma
    # de spline interpoladores
    # Input : veps: (vectores propios con los coeficientes)
    #          x : (knots)
    # Output: Lista de splines registrados correspondientes
    #          a los coeficientes.

    dimat<-4*(n-1)
    listapca<-list(x,vector(mode='list',length=dimat),'t')
    names(listapca)<-c('x','s','r')
    for (i in 1:dimat)
    {
      listapca$s[[i]]<-array(0,dim=c((n+1),4))
      # Calculamos la pendiente en el knot[1] (m en la ec. y=a+mx)
      listapca$s[[i]][1,2]<-(veps[2,i])+(2*veps[3,i]*x[1])+
```

```

(3*veps[4,i]*(x[1]**2))

# Calculamos (a=y-mx) en el knot[1] (m en la ec. y=a+mx)
listapca$s[[i]][1,1]<-{veps[1,i]+veps[2,i]*x[1]+
                         veps[3,i]*(x[1]**2)+veps[4,i]*(x[1]**3))-+
                         listapca$s[[i]][1,2]*x[1]

for (j in 1:(n-1))
{
  listapca$s[[i]][j+1,]<-veps[((4*j)-3):(4*j),i]
}

# Calculamos la pendiente en el knot[n] (m en la ec. y=a+mx)
listapca$s[[i]][n+1,2]<-(veps[dimat-2,i])+
                         (2*veps[dimat-1,i]*x[n])+ 
                         (3*veps[dimat,i]*(x[n]**2))

# Calculamos (a=y-mx) en el knot[n] (m en la ec. y=a+mx)

listapca$s[[i]][n+1,1]<-(veps[dimat-3,i]+veps[dimat-2,i]*x[n]+
                         veps[dimat-1,i]*(x[n]**2)+ 
                         veps[dimat,i]*(x[n]**3))-+
                         listapca$s[[i]][n+1,2]*x[n]

}

listapca
}

# FIN Funcion sint.pca
### FIN FUNCIONES ADICIONALES

### PROGRAMA

n<-length(lreg$x)
N<-length(lreg$s)
C<-monta.c(lreg,n,N)
W<-monta.w(lreg$x,n,N)

# Realizamos un cambio de variable para hacer
# que la matriz a diagonalizar sea simetrica a traves
# de la matriz raiz cuadrada de W

Wvp<-eigen(W)
sqrtW<-Wvp$vectors%*%diag(sqrt(Wvp$values))%*%t(Wvp$vectors)

A<-(1/N)*sqrtW%*%t(C)%*%C%*%sqrtW

vp<-eigen(A,symmetric=T)
vaps<-vp$values
veps<-vp$vectors

# Deshacemos el cambio de variable

```

```

newveps<-solve(sqrtW) %*% veps

# Montamos las ACPFs como splines interpoladores

splpca<-sint.pca(newveps,lreg$x,N,n)

# Calculamos las correlaciones entre cada funcion muestra
# y cada componente principal

cofun<-C%*%W%*%newveps
dcofun<-dim(cofun)
corrfun<- (diag(diag(C%*%W%*%t(C)) ^(-1/2))) %*%cofun
prueba<-t(newveps)%*%W%*%newveps

val<-list(newveps,vaps,splpca,cofun,corrfun,dim(C),dim(W),dim(newveps))
names(val)<-c('veps','vaps','splpca','cofun','corrfun','dC','dW','dB')
val

### FIN PROGRAMA
}

```

4.- Otras funciones de soporte.

4.1.- Algoritmo Buscacoef.

<pre> ### BUSCACOEF buscacoef<-function(s,eval,ini=1) { i<-ini if (eval>max(s\$x)) { i<-length(s\$x)+1 } else { if (eval<min(s\$x)) { i<-1 } } }</pre>	<pre> else { while (s\$x[i]<eval) { i<-i+1 } } val<-list(s\$s[i],i) names(val)<-c('s','i') val } ### FIN BUSCACOEF </pre>
--	--

4.2.- Algoritmo PL.

```
### FUNCION PL
pl<-function(pol,equis)
{
  res<-pol[1]
  for (i in 1:(length(pol)-1))
  {
    res<-((equis^i)*pol[i+1])+res
  }
}
### FIN FUNCION PL
```

4.3.- Algoritmo spl.a2.

Utiliza funciones ya expuestas anteriormente tales como *intPL*, *mutpol*, *splquad*, y utiliza una nueva llamada *intspl*.

```
Spl.a2<-function(spllis)
{
  ### Calcula para una lista de splines registrados otra
  ### lista con sus splines al cuadrado. Posteriormente
  ### calcula el area de los splines al cuadrado, que es
  ### lo mismo que calcular el area al cuadrado de cada
  ### spline de la lista input.

  ### FUNCIONES ADICIONALES
  ### FUNCION intPL
  # FUNCION MUTPOL
  # FUNCION SPLQUAD
  ### FUNCION intspl
  int.spl<-function(spl)
  {
    # Calcula la integral de un spline no registrado.
    # Input: spline no registrado
    #   (spl$x=knots, spl$s=matriz polinomios)
    # Output: valor de la integral al cuadrado.
    area<-as.double(0)
    long.spl<-length(spl$x)
    for (i in 1:(long.spl-1))
    {
      area<-area+int.pl(spl$s[(i+1),],spl$x[i+1])-
        int.pl(spl$s[(i+1),],spl$x[i])
    }
  }
}
```

```

return(area)
}

# PROGRAMA PRINCIPAL
# Creacion variables
area2<-rep(as.double(0),length=length(spllis$s))
# algoritmo
for (i in 1:length(spllis$s))
{
  aux<-splquad(spllis$s[[i]])
  aux<-list(spllis$x,aux)
  names(aux)<-c('x','s')
  area2[i]<-int.spl(aux)
}
print(area2)
return(area2)

# FIN
}

```

Agradecimientos.

En primer lugar agradezco al profesor Pedro Delicado el tiempo dedicado, las lecciones, los consejos y el interés que ha puesto en la consecución de este proyecto. También merece especial atención la Dra. Margarita García y del Dr. Pontón por ceder los datos del estudio de farmacocinética, y a Xavi Pérez y Maite Encuentra por sus ideas y sugerencias de gran ayuda. También quiero agradecer por igual a los compañeros de la licenciatura los muchos consejos y ánimos que me han ofrecido: Ramón Clèries, Juan Ramón González y Xavi Puig. No puedo olvidarme ni de Víctor Moreno ni de José Miguel Martínez, que me han brindado la oportunidad de exponer este trabajo en las sesiones del SERC y del IMIM respectivamente. Los ánimos de la gente de la UIC, del SERC y del Servicio de Oncología Radioterápica del ICO me han sido de gran apoyo.

Muchas gracias a todos.