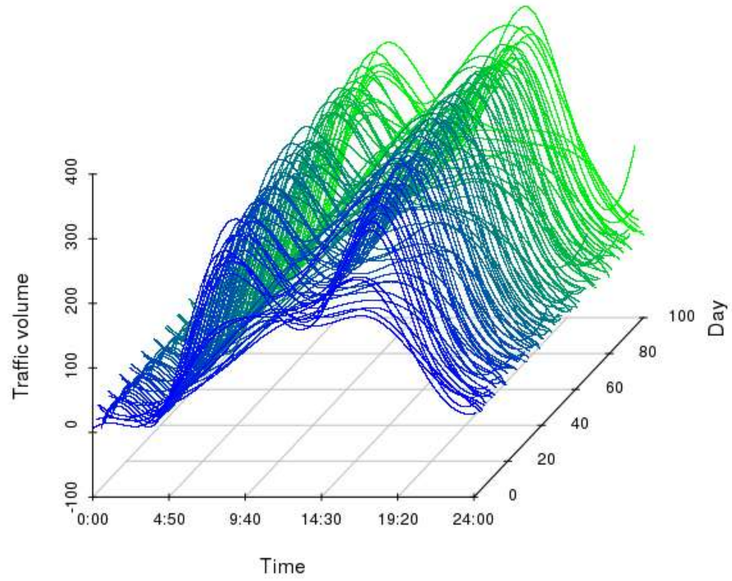
TEMA 5: análisis de datos funcionales



José R. Berrendero

Departamento de Matemáticas, Universidad Autónoma de Madrid

1 / 26

Temas a tratar

El objetivo es dar una idea general de qué es el análisis de datos funcionales

- ¿Qué es el análisis de datos funcionales?
- Modelos probabilísticos y espacio muestral
- Preprocesado:
  - Representación en términos de una base
  - Suavizado
- Estimación de la media y de la función de covarianzas
- Componentes principales funcionales
- Regresión funcional

¿Qué es el análisis de datos funcionales?

El **análisis de datos funcionales (FDA)** tiene por objetivo el tratamiento estadístico de datos que pueden contemplarse como funciones.

Cuevas (2014) resume así la evolución de las técnicas estadísticas según tipo de datos y espacio parámetrico:

Statistical theory	$\mathcal{X}$	Θ	Dating back to
Classical parametric inf.	$\mathbb{R}$	$\Theta\subset\mathbb{R}$	1920s
Multivariate analysis	$\mathbb{R}^d$ $(n >> d)$	$\Theta\subset\mathbb{R}^k$ $(n >> k)$	1940s
Nonparametrics	$\mathbb{R}^d$ $(n >> d)$	A function space	1960s
High dimensional problems	$\mathbb{R}^d$ $(n < d)$	$\Theta\subset\mathbb{R}^k$	2000s
Functional Data Analysis	A function space	$\mathbb{R}^k$, or a function space	1990s

Fundamentos matemáticos del FDA:

- Sobre las propiedades del espacio muestral en el que *viven* los datos: **análisis funcional**, especialmente la teoría de espacios de Hilbert.
- Sobre los modelos probabilísticos: la teoría de **procesos estocásticos**.

Bibliografía básica

Dos artículos históricos sobre inferencia con procesos estocásticos:

- Grenander, U. (1950). Stochastic Processes and Statistical Inference. *Arkiv för Matematik*, **1**, 195–277.
- Parzen, E. (1961). An approach to time series analysis. *The Annals of Mathematical Statistics*, **32**, 951-989.

La primera monografía que se publicó sobre FDA:

- Ramsay and Silverman (2005). *Functional Data Analysis*. Springer. (Primera edición de 1997.)

Libros sobre FDA de orientación más teórica, especialmente el segundo:

- Horváth and Kokoszka (2012). *Inference for Functional Data with Applications*. Springer.
- Hsing and Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Bibliografía básica

Un texto reciente que puede servir como introducción:

- Kokoszka and Reimherr (2017). *Introduction to Functional Data Analysis*. CRC Press.

Sobre series temporales funcionales:

- Bosq (2000). *Linear Processes in Function Spaces*. Springer.

Regresión funcional no lineal:

- Ferraty and Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

Dos artículos de revisión, el primero muy citado:

- Cuevas (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, **147**, 1-23.
- Wang et al (2015). Functional Data Analysis. *Annual Review of Statistics and Its Application*, **3**, 257-295.

Implementaciones en R

- Véase esta página para un listado de paquetes con implementaciones de técnicas de FDA.
- Paquetes que vamos a usar en esta presentación:

```
library(refund)
library(fda.usc)
```

- Información sobre `fda.usc`

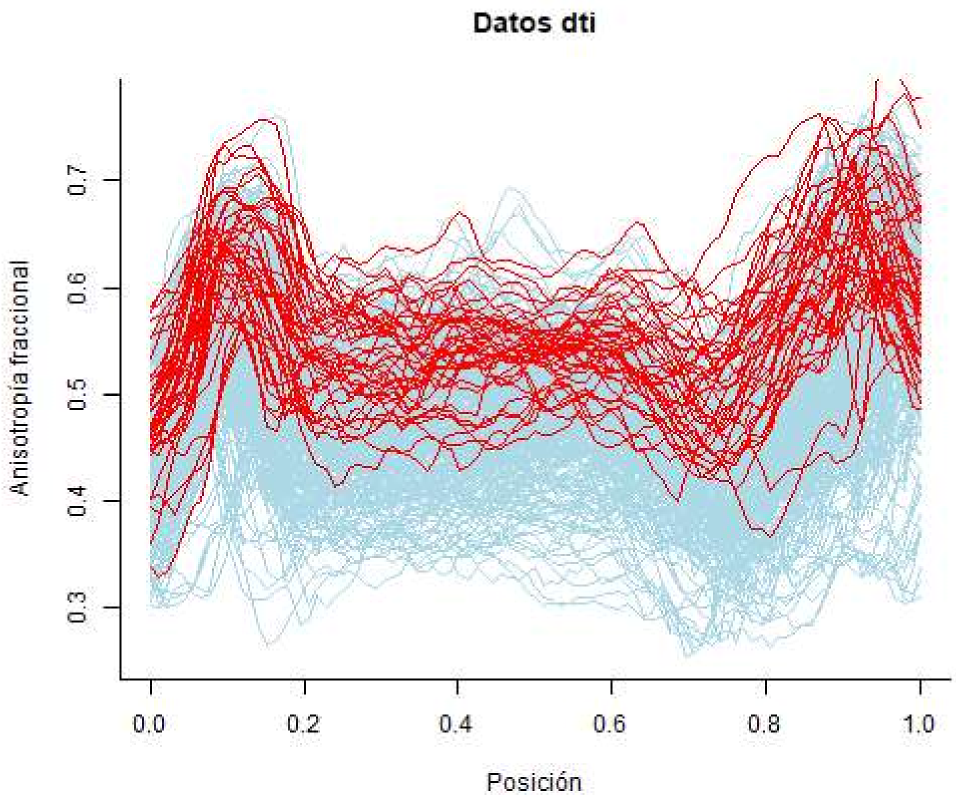
Ejemplo

- Referencia: Kokoszka and Reimherr (2017), pag. 14.
- DTI (*diffusion tensor imaging*) es una técnica de imagen de resonancia magnética basada en la forma en la que el agua se difunde en el cerebro.
- El agua se difunde isotrópicamente (igual en todas direcciones) en el cerebro excepto en la sustancia blanca donde lo hace anisotrópicamente.
- El cuerpo calloso es el haz de fibras nerviosas que sirve como vía de comunicación entre un hemisferio cerebral y otro.
- La anisotropía fraccional es un valor entre 0 y 1 que mide el nivel de anisotropía (y por lo tanto la cantidad de sustancia blanca) en una posición particular del cuerpo calloso.
- En 376 personas (pacientes y controles) se ha medido la anisotropía fraccional en 93 posiciones equiespaciadas.
- Los datos se encuentran en el conjunto de datos `DTI` del paquete `refund`.

Preparación y representación gráfica

```
# lee y prepara los datos
data(refund::DTI)

Y <- DTI$cca
Y <- DTI$cca[-c(126,130,131,125,319,321),] # datos perdidos
case <- as.factor(DTI$case[-c(126,130,131,125,319,321)]) # 1, paciente; y 0, con
case <- recode(case, "0" = "control", "1" = "paciente")
m <- dim(Y)[2]

# Crea el objeto de la clase fdata para datos funcionales
dti <- fdata(Y,
             argvals = seq(0, 1, length=m),
             names = list(main = "Datos dti",
                          xlab = "Posición",
                          ylab = "Anisotropía fraccional"))

# Representación gráfica
plot(dti[case == "paciente"], col = "lightblue", bty = "l")
lines(dti[case == "control"], col = "red")
```

Preparación y representación gráfica



9 / 26

Dos situaciones diferentes

En la práctica se observa cada función x_i discretizada (en un conjunto finito $t_{i,1}, \ldots, t_{i,m_i}$).

Se suelen considerar dos casos:

- **Funciones densamente observadas**: los puntos $t_{i,j}, \quad j = 1, \ldots, m_i$, se pueden elegir arbitrariamente próximos.
- **Datos funcionales sparse**: hay limitaciones en cuanto a los puntos en los que las funciones se pueden observar. Este tipo de datos también se llaman *longitudinales*.

Los valores de t suelen corresponder a instantes en el tiempo, pero no necesariamente. En el ejemplo, indican posiciones en el espacio.

Aquí vamos a suponer que todas las funciones se observan en los mismos puntos $t_1, \ldots, t_m$.

Modelos probabilísticos

La muestra observada $x_1, \ldots, x_n$ corresponde a realizaciones independientes de un **proceso estocástico** $X \equiv \{X(t) : t \in T\}$. Fijamos $T = [0, 1]$, sin que ello suponga mucha pérdida de generalidad (en el contexto de FDA).

Esto significa:

- Para cada $t \in [0, 1]$, $X(t)$ es una variable aleatoria. Desde este punto de vista, un proceso es una familia de variables aleatorias *indexada* por T. Para cada $t \in T$, $X_1(t), \ldots, X_n(t)$ son vaiid.

- Un proceso también se puede entender como una función aleatoria seleccionada mediante un mecanismo aleatorio bien definido sobre un cierto espacio (el espacio muestral). La muestra es un conjunto de estas funciones.

Los datos funcionales son **gaussianos** si las distribuciones finito dimensionales del proceso subyacente X son normales multivariantes, es decir, si para todo p y cualesquiera $t_1, \ldots, t_p \in T$, el vector $(X(t_1), \ldots, X(t_p))$ tiene distribución normal p-dimensional.

¿Cuál es el espacio muestral?

La gran mayoría de técnicas disponibles en FDA suponen que los datos corresponden a procesos cuyas trayectorias pertenecen a un **espacio de Hilbert H**.

Tenemos a nuestra disposición varias estructuras en H:

- Las funciones son vectores. Podemos sumarlas y multiplicarlas por escalares con las propiedades esperadas.
- Hay un producto escalar $\langle f, g \rangle$ que permite hablar de ortogonalidad: $f \perp g$ si $\langle f, g \rangle = 0$. Ventajas:
    - Teorema de Pitágoras, que interviene en diversas descomposiciones de interés.
    - La idea fundamental de proyección sobre un conjunto (esencial para el método de mínimos cuadrados).
- El producto escalar induce una norma $\|f\| = \langle f, f \rangle^{1/2}$, que permite hablar de distancia entre datos funcionales $d(f, g) = \|f - g\|$.

¿Cuál es el espacio muestral?

- La distancia incorpora las ideas de convergencia y continuidad: $\lim_{n\to\infty} f_n = g$ si $\|f_n - g\| \to 0$, cuando $n \to \infty$. Gracias a ello, es posible estudiar la consistencia de los estimadores.
- Un espacio de Hilbert es completo: las sucesiones de Cauchy son convergentes. Esta propiedad es fundamental para probar la existencia de proyecciones, por ejemplo.

El espacio más utilizado

Es usual suponer $H = L^2[0,1]$: funciones $f : [0,1] \to \mathbb{R}$ de cuadrado integrable $\int_0^1 |f(t)|^2 dt < \infty$ dotado del producto escalar

$$\langle f, g \rangle := \int_0^1 f(t)g(t)dt.$$

Esta es la generalización más directa de la norma y el producto escalar euclídeos para vectores de dimensión finita.

Representación en términos de una base

Suponemos que los datos pertenecen a $H = L^2[0,1]$.

Consideramos una base ortonormal $\{e_\ell : \ell = 1, 2, \ldots\}$ de H:

$$\|e_\ell\| = 1, \ \text{ para todo } \ell \geq 1; \text{ y } \langle e_\ell, e_k \rangle = 0, \text{ si } \ \ell \neq k$$

Hay varias bases que se pueden usar: trigonométricas, splines, etc.

Toda función $x \in H$ admite la representación:

$$x(t) = \sum_{\ell=1}^{\infty} \langle x, e_\ell \rangle e_\ell(t)$$

Una técnica básica en FDA consiste en representar cada dato mediante la serie anterior truncada en un número de sumandos conveniente N:

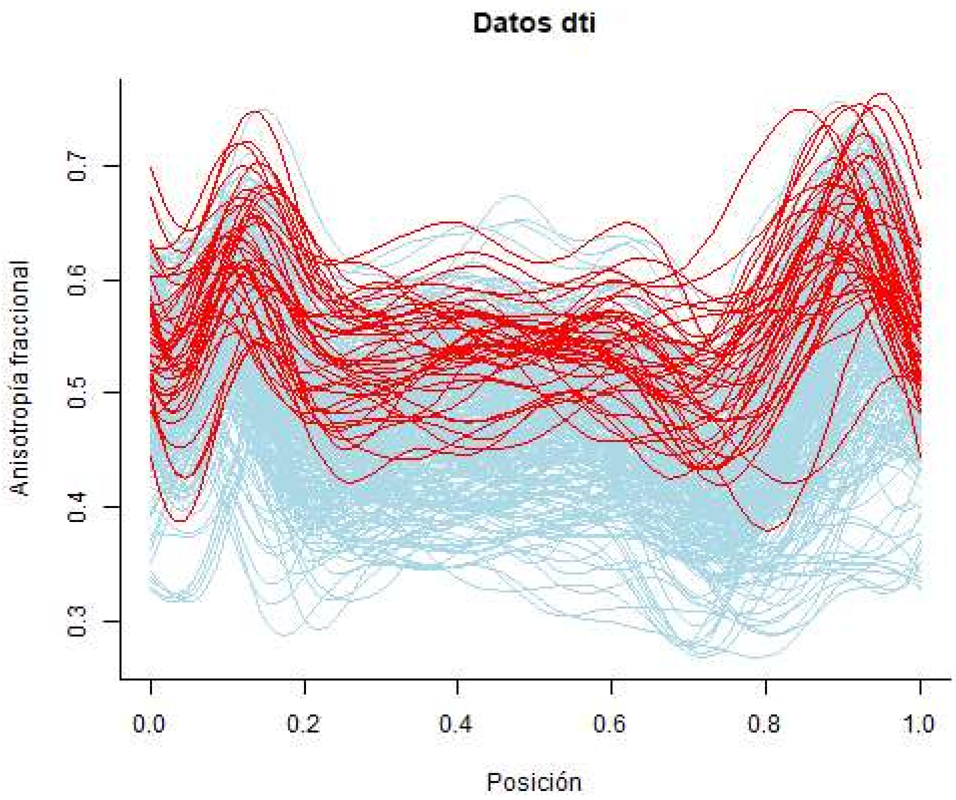
$$x_i(t_j) \approx \sum_{\ell=1}^{N} \langle x_i, e_\ell \rangle e_\ell(t_j), \quad i = 1, \ldots, n; \ \ j = 1, \ldots, m.$$

Ejemplo

- Usamos `fdata2fd` para obtener la representación y `fdata` para pasarla a un objeto de la clase `fdata`.
- `type.basis` permite elegir entre diferentes bases.
- `nbasis` determina el número de elementos de la suma N.
- `nderiv` es un argumento opcional que permite trabajar con las derivadas en lugar de con las funciones originales (una de las ventajas de usar la representación).

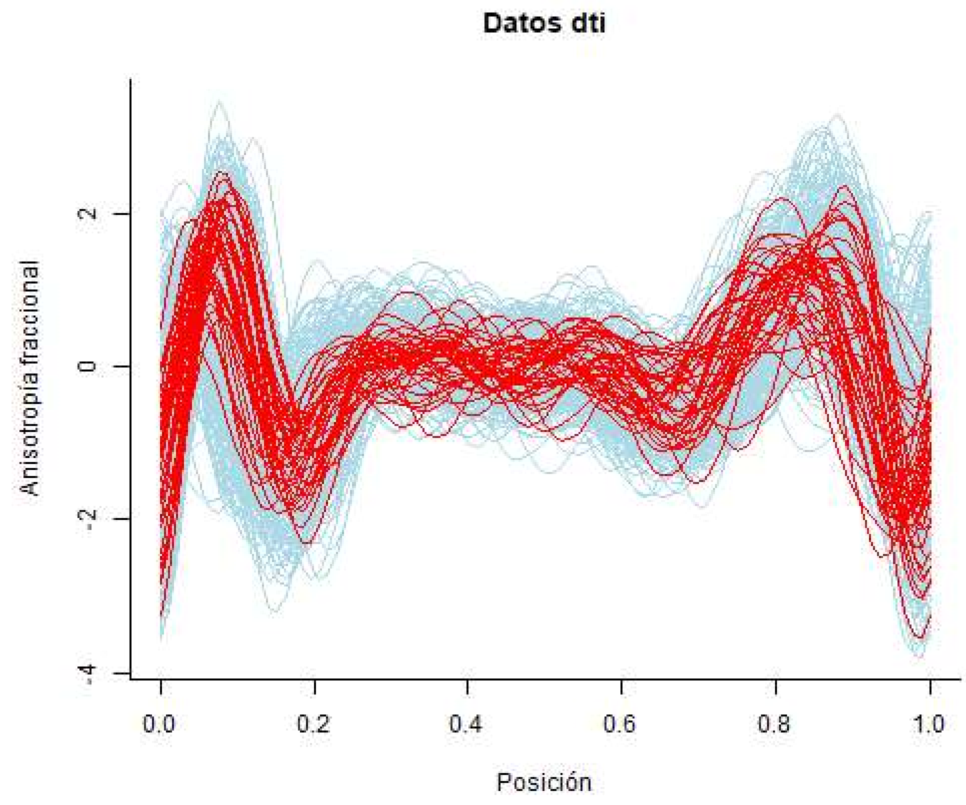
```
dti_fourier <- fdata2fd(dti, type.basis= "fourier", nbasis= 50)
dti_fourier <- fdata(dti_fourier, argvals = seq(0, 1, length=m),
                     names = list(main = "Datos dti",
                                  xlab = "Posición",
                                  ylab = "Anisotropía fraccional"))

plot(dti_fourier[case == "paciente"], col = "lightblue", bty = "l")
lines(dti_fourier[case == "control"], col = "red")
```

Ejemplo



16 / 26

Ejemplo (primeras derivadas)



Suavizado

En algunos casos puede ser conveniente obtener versiones más suaves de los datos.

Una posibilidad es usar el método del núcleo con pesos de tipo Nadaraya-Watson:

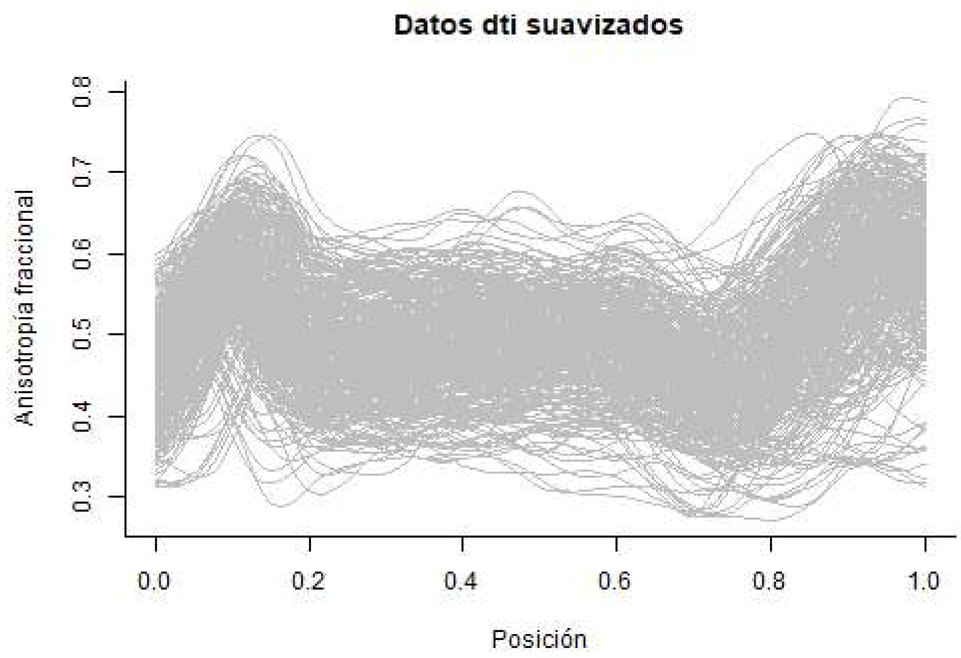
$$\tilde{x}_i(t) = \sum_{j=1}^{m} s_j(t) x_i(t_j),$$

donde

$$s_j(t) = \frac{K\left(\frac{t-t_j}{h}\right)}{\sum_{\ell=1}^{m} K\left(\frac{t-t_\ell}{h}\right)}$$

Ejemplo

```
# Parámetro de suavizado h calculado previamente por VC
# Si no se fija h, se calcula usando validación cruzada
suavizado <- optim.np(dti, h = 0.0167, type.S = S.NW)
dti_suavizado <- suavizado$fdata.est
plot(dti_suavizado, col = "gray", bty = "l", main = "Datos dti suavizados")
```



Estimación de media y covarianza

Dos funciones importantes a estimar en relación con el proceso X son

- La función media $\mu(t) = \mathrm{E}(X(t))$
- La función de covarianzas $K(s,t) = \mathrm{E}[(X(t)-\mu(t))(X(s)-\mu(s))]$

Las versiones muestrales de estas funciones son:

- La función media muestral: $\bar{x}(t) = n^{-1}\sum_{i=1}^n x_i(t)$
- La función de covarianzas muestral:

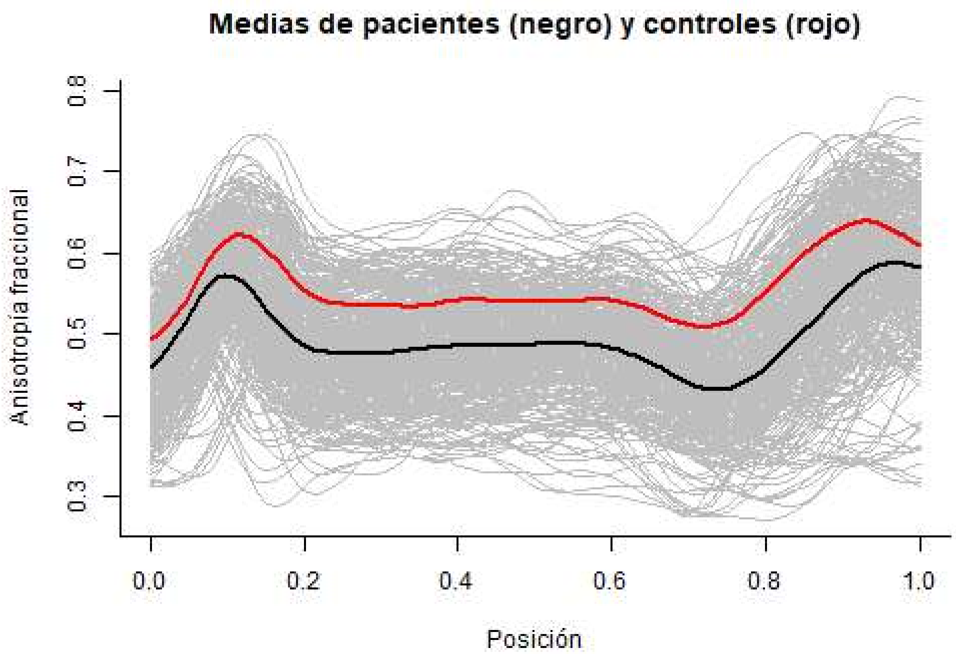
$$\hat{K}(s,t) = n^{-1}\sum_{i=1}^n (x_i(t)-\bar{x}(t))(x_i(s)-\bar{x}(s))$$

Si queremos estimadores para todo $t \in [0,1]$ será necesario aplicar algún método de suavizado a los estimadores anteriores.

Existen LFGN y TCL para procesos que garantizan la consistencia de estos estimadores y determinan su distribución asintótica.

Ejemplo

```
dti_pacientes <- dti_suavizado[case=="paciente"]
dti_controles <- dti_suavizado[case=="control"]
plot(dti_suavizado, col = "gray", bty = "l",
     main = "Medias de pacientes (negro) y controles (rojo)")
lines(func.mean(dti_pacientes), lwd = 2)
lines(func.mean(dti_controles), lwd = 2, col = 'red')
```



Matriz y operador de covarianzas

Finito dimensional	**Funcional**
Aplicación lineal	Operador lineal
$x \mapsto Kx$	$x \mapsto Kx(\cdot) = \int_0^1 K(\cdot, s)x(s)ds$
Autovalores y autovectores	Autovalores y autofunciones
$Ku = \lambda v$	$Kv(\cdot) = \lambda v(\cdot)$
Diagonalización	Teorema de Mercer
$K = V\Lambda V' = \sum_{\ell=1}^{p} \lambda_\ell v_\ell v_\ell'$	$K(s,t) = \sum_{\ell=1}^{\infty} \lambda_\ell v_\ell(s) v_\ell(t)$
Coordenadas en la base de autovectores	Desarrollo de Karhunen-Loève
$x - \mu = \sum_{\ell=1}^{p} ((x-\mu)' v_\ell) v_\ell$	$x(t) - \mu(t) = \sum_{\ell=1}^{\infty} \langle x - \mu, v_\ell \rangle v_\ell(t)$
Descomposición espectral	Descomposición espectral
$Kx = \sum_{\ell=1}^{p} \lambda_\ell (x' v_\ell) v_\ell$	$Kx(t) = \sum_{\ell=1}^{\infty} \lambda_\ell \langle x, v_\ell \rangle v_\ell(t)$

Componentes principales funcionales

Cada dato funcional se puede representar usando la base de autofunciones del operador de covarianzas (Karhunen-Loève):

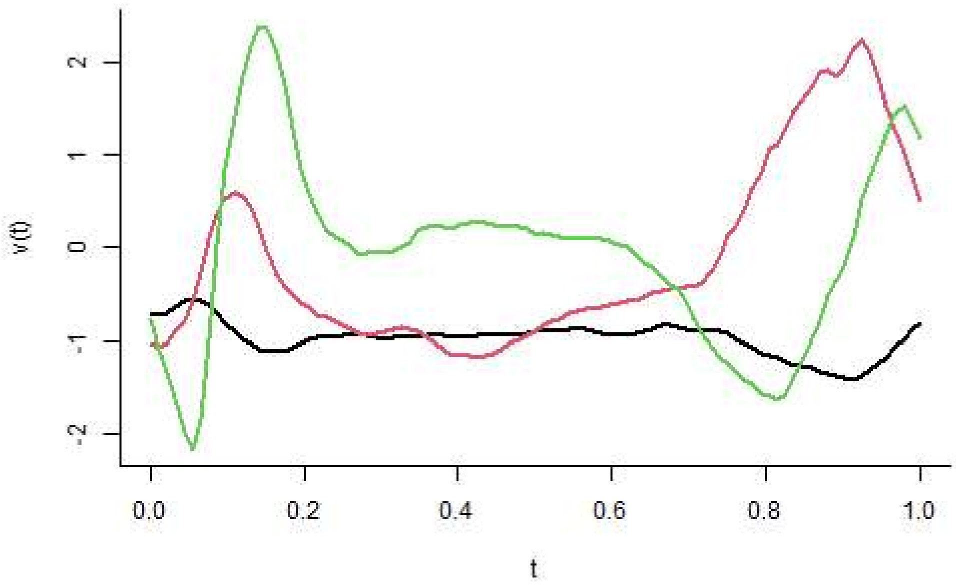
$$X_i(t) - \mu(t) = \sum_{\ell=1}^{\infty} \langle X_i - \mu, v_\ell \rangle v_\ell(t) := \sum_{\ell=1}^{\infty} Z_{i,\ell} v_\ell(t).$$

Las v.a. $Z_{i,1}, Z_{i,2}, \dots$ son las componentes principales funcionales de X_i.

- $Z_{i,1}, Z_{i,2}, \dots$ son v.a. incorreladas con $\mathrm{E}(Z_{i,\ell}) = 0$ y $\mathrm{Var}(Z_{i,\ell}) = \lambda_i$.
- Podemos reemplazar la trayectoria $X_i(t)$ por sus N primeras componentes $Z_i = (Z_{i,1}, \dots, Z_{i,N})$ sin perder mucha información. Después se aplican las técnicas habituales de análisis multivariante a las componentes.
- En la práctica $\mu(t)$ y $K(t, s)$ se estima en un grid finito y el cálculo de los autovalores y autofunciones se reduce a una descomposición espectral matricial.

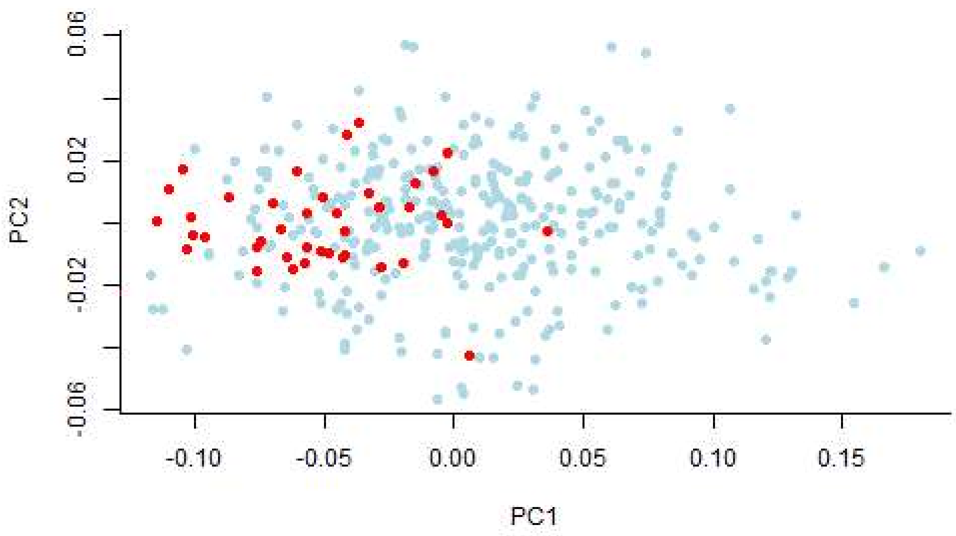
Ejemplo

```
# Estima y representa las tres primeras autofunciones
# PC1 (negro), PC2 (rojo), PC3 (verde)

dti_pc <- fdata2pc(dti, ncomp=3)
plot(dti_pc$rotation, main="", lwd=2, ylab = "v(t)", bty = 'l')
```



Ejemplo

```
# Estima y representa los coeficientes para las dos primeras
# componentes de todos los datos pacientes (azul), controles (rojo)
# % var. explicada = sum(dti_pc$d[1:2]^2) / sum(dti_pc$d^2) = 71% aprox

plot(dti_pc$x[case == "paciente", 1], dti_pc$x[case == "paciente", 2],
     col = 'lightblue', pch = 16, xlab = 'PC1', ylab = 'PC2', bty = 'l')
points(dti_pc$x[case == "control", 1], dti_pc$x[case == "control", 2], col = 're
```



Regresión funcional

Modelo de respuesta escalar y regresor funcional

$$Y_i = \beta_0 + \int_0^1 \beta(t) X_i(t) dt + \epsilon_i, \quad i = 1, \ldots, n.$$

Enfoque 1: regresión de componentes principales funcionales

Si $\beta_\ell := \langle \beta, v_\ell \rangle$, $\int_0^1 \beta(t) X_i(t) dt = \sum_{j=1}^\infty \beta_j Z_{i,\ell} \approx \sum_{j=1}^N \beta_j Z_{i,\ell}$.

Regresión múltiple, usando las variables regresoras $Z_{i,1}, \ldots, Z_{i,N}$.

Enfoque 2: método de penalización

Variantes del problema

$$\arg\min_{\beta \in \mathcal{F}} \sum_{i=1}^n \left(Y_i - \int_0^1 \beta(t) X_i(t) dt \right)^2 + \lambda \int_0^1 [\beta''(t)]^2 dt$$

para diversos espacios de funciones $\mathcal{F}$ y penalizaciones.