Modelo Factorial para Datos Funcionales

Workshop on Econometrics and Data Science, 2018
Universidad Autónoma de Nuevo León.

Israel Martinez-Hernandez
Postdoctoral Fellow at KAUST,
israelmat@cimat.mx
israel.martinezhernandez@kaust.edu.sa

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

STATISTICS
@KAUST

Marc G. Genton's Research Group: Spatio-Temporal Statistics and Data Science

https://stsda.kaust.edu.sa/Pages/Home.aspx

Outline

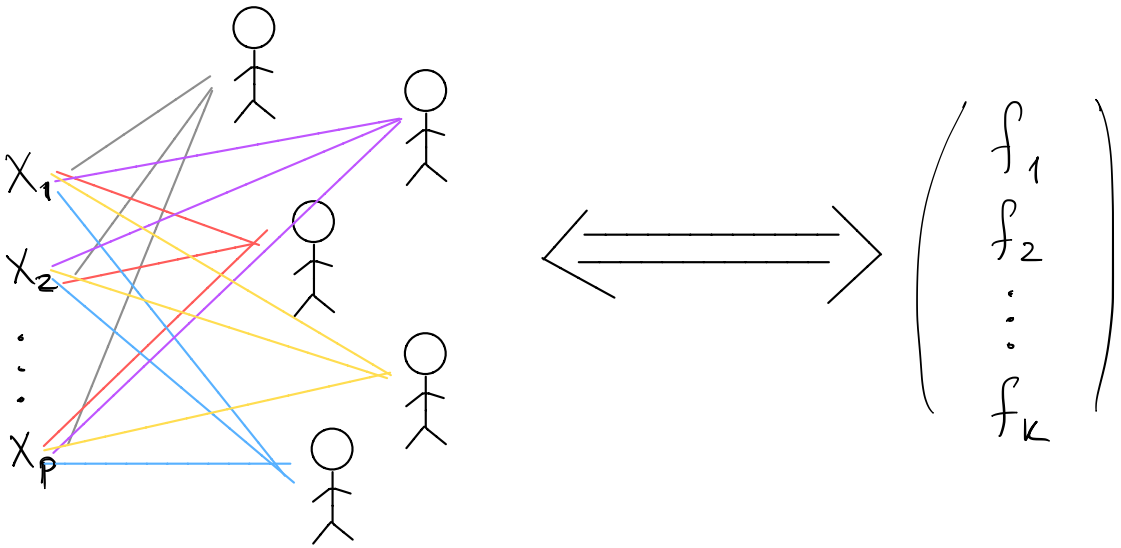
Modelos Factoriales Multivariados

Introduction to Functional Data

Modelos Dinámicos Factoriales para FTS
- Estimación
- Ejemplos

Modelos Factoriales

- Los modelos factoriales describen una colección de variables aleatorias (por ejemplo encuestas) a través de un número pequeño de variables, llamadas factores.



Covarianza tot= Cov compartida + Varianza especifica + Varianza de errores de medición

Modelos Factoriales

Sea $(X_1, X_2, \ldots, X_p)$ un vector de variables aleatorias correlacionadas

$$
\begin{aligned}
X_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + e_1 \\
X_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + e_2 \\
&\vdots \\
X_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pk} f_k + e_p
\end{aligned}
$$

reescribiendo

$$\boldsymbol{X} = \Lambda \boldsymbol{f} + \boldsymbol{e}$$

donde $k << p$, $\text{Cov}(f_i, f_j) = \delta_{\{i=j\}}$ y $\text{Cov}(f_i, e_j) = 0$. Los parámetros λ_{ij} son llamados pesos factoriales o factor loadings, f_i los factores, y $\lambda_i^T f$ es llamado como los componentes comunes.

Modelos Factoriales

Así, la varianza de cada variable aleatoria X_i se puede escribir como

$$\mathrm{Var}(X_i) = \sigma_i^2 = \sum_{i=1}^{k} \lambda_i^2 + \psi_i,$$

donde ψ_i es la vairanza de e_i.

De igual manera, la covarianza entre dos variables leatorias es,

$$\mathrm{Cov}(X_i, X_j) = \sigma_{i,j} = \sum_{i=1}^{k} \lambda_i \lambda_j.$$

Por lo tanto, la matriz de covarianzas Σ del vector de variables aleatorias es

$$\Sigma = \Lambda \Lambda^T + \Psi.$$

Modelos Factoriales

Existen dos tipos de modelos factoriales:

1. Los factores son observables; factores economicos.

   El modelo de Fama & French;
   Indican que el exceso del retorno esperado de una cartera se puede definir por tres factores

$$r = r_f + \beta_1(r_m - r_f) + \beta_2(SMB) + \beta_3(HML) + \varepsilon,$$

   donde $r =$ exceso de retorno esperado
   Los factores son, $r_m - r_f$, SMB y HML.

   En este caso la esitmación puede ser obtenida a través de minimos cuadrados.

   Ver Fama, E. F.; French, K. R. (1993)

Modelos Factoriales

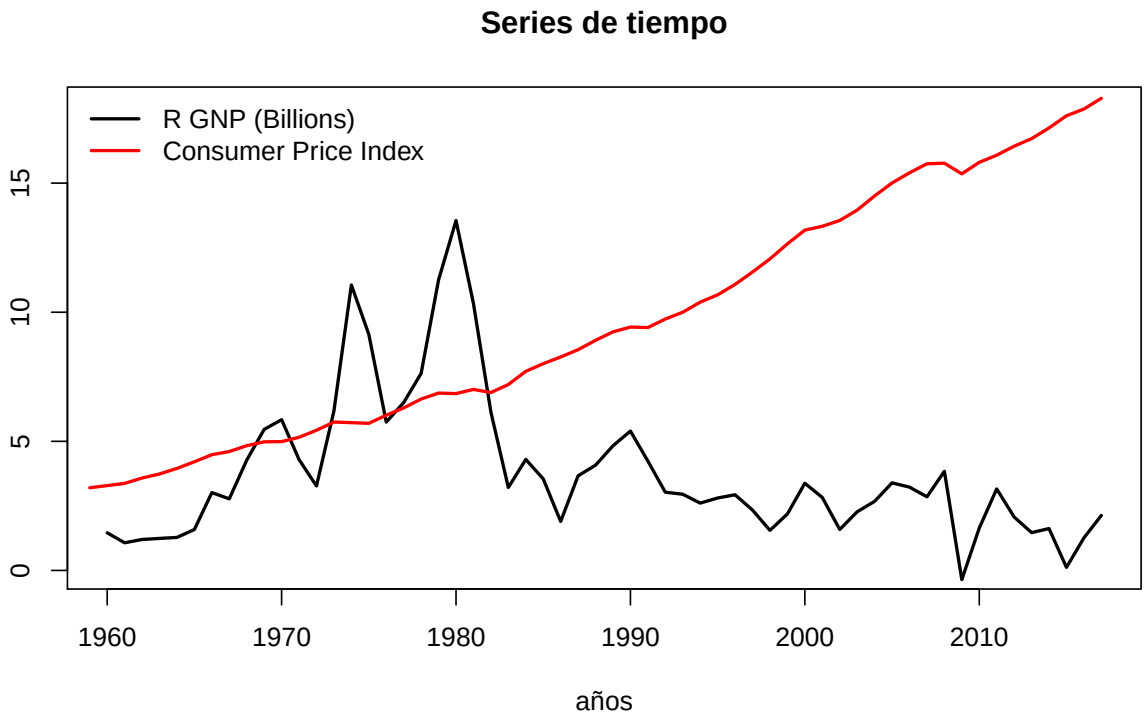
2. Los factores son varaibles aleatorias latentes, i.e., no se observan.

   Para este enfoque, los metodos comunes para extraer la estructura de los factores son:

   - Análisis Factorial,
   - Análisis de Componentes Principales (PCA) o,
   - Máxima verosimilitud bajo supuestos distribucionales.

Modelos Factoriales Dinámicos

Que tal si los datos observados son series de tiempo, i,e., cada componente del vector aleatorio $X_i = \{x_{it}\,;\, t \in \mathbb{N}\}$.



En este caso los factores f_i tambien dependen del tiempo; f_{it} para $i = 1, 2, \cdots, k$. Además estos factores tienen una dinamica a través del tiempo.

Modelos Factoriales Dinámicos

$$
\begin{aligned}
x_{1t} &= \lambda_{11} f_{1t} + \lambda_{12} f_{2t} + \cdots + \lambda_{1k} f_{kt} + e_{1t} \\
&\vdots \\
x_{pt} &= \lambda_{p1} f_{1t} + \lambda_{p2} f_{2t} + \cdots + \lambda_{pk} f_{kt} + e_{pt}
\end{aligned}
$$

reescribiendo para cada i

$$x_{it} = \boldsymbol{\lambda}_i^T \boldsymbol{f}_t + e_{it}, \quad t = 0, 1, \cdots, T.$$

- Los factores $\boldsymbol{f}_t$ pueden ser tales que

  $$A(L)\boldsymbol{f}_t = \boldsymbol{u}_t.$$

  donde $A(L)$ es un polinomio de operadores de retardo. Por ejemplo, $\boldsymbol{f}_t$ puede ser un $\mathrm{VAR}(p)$.
- Los factores $\boldsymbol{f}_t$ pueden ser no estaionarios; con raíz unitaria.

Modelos Factoriales Dinámicos

Sea $X_i = (x_{i1}, x_{i2}, \cdots, , x_{iT})^{'}$, $F = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_T)^{'}$ y $\Lambda = (\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_p)$.

Vamos a escribir como

$$X = (X_1^{'}, X_2^{'}, \cdots, X_p^{'}),$$

la matriz de observaciones. Cada componente de la serie está en columna. Así

$$X = F\Lambda + e,$$

$$(T \times p) = (T \times k)(k \times p) + (T \times p),$$

Los parametros a estimar son Λ y F.
Supuestos:

- $F^{'}F/T = I_k$
- $\Lambda^{'}\Lambda$ es diagonal.

Estimación vía Componentes Principales

Dado k, el método de componentes principales consiste en estimar los parametros, resolviendo el problema

$$\min_{\Lambda^k, F^k} S(k) = \min_{\Lambda^k, F^k} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - (\boldsymbol{\lambda}_i^k)^T \boldsymbol{f}_t^k)^2$$

Sujeto a las normalizaciones anteriores. Obteneindo así

- $\hat{F}^k = \sqrt{T} \times$ matriz de eigenvectores correspondinetes a los primeros k eigenvalores mas grandes de la matrix XX'.
- $\hat{\Lambda}^{k\prime} = \hat{F}^{k\prime} X / T$

Cualquier otra rotación de $\hat{F}^k$ es también una solución

Ver Bai, J. and Ng, S. 2002. Determining the Number of Factors.

Estimación vía Componentes Principales

Y que tal si los facores son no estacionarios, por ejemplo

$$\boldsymbol{f}_t = \boldsymbol{f}_{t-1} + \boldsymbol{\eta}_t,$$

y entonces lo que se observa es

$$x_{it} = D_{it} + \boldsymbol{\lambda}_i^T \boldsymbol{f}_t + \varepsilon_{it},$$

incluyendo una tendencia deterministica D_{it}.

Más aun, se puede tener que algunas componentes de los factores son estacionarias y el resto no estacionarias.

Existen resultados en la literatura. Ver la tesis doctoral de Fancisco Corona

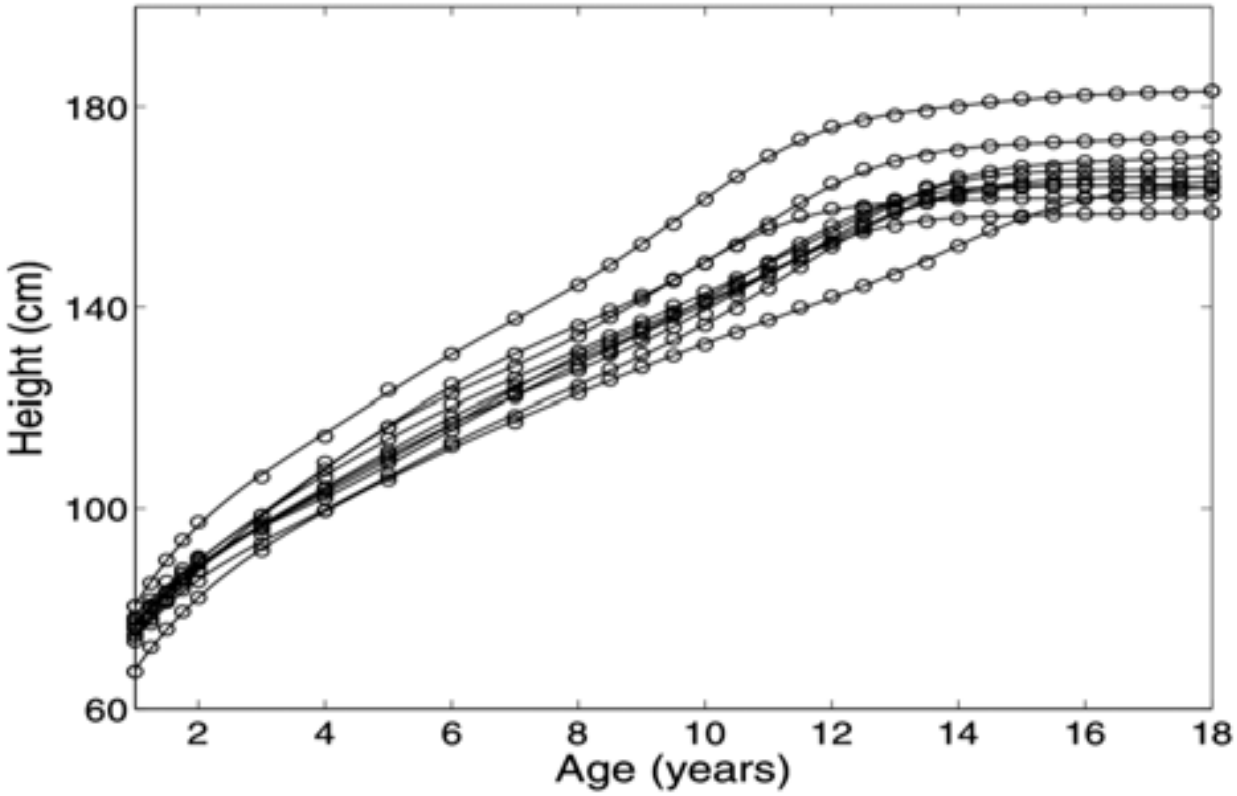
Datos Funcionales

- $\Omega : \; w \rightarrow x_i \in \mathbb{R}$

- $\Omega : \; w \rightarrow x_i(s) \in \mathbb{H}$

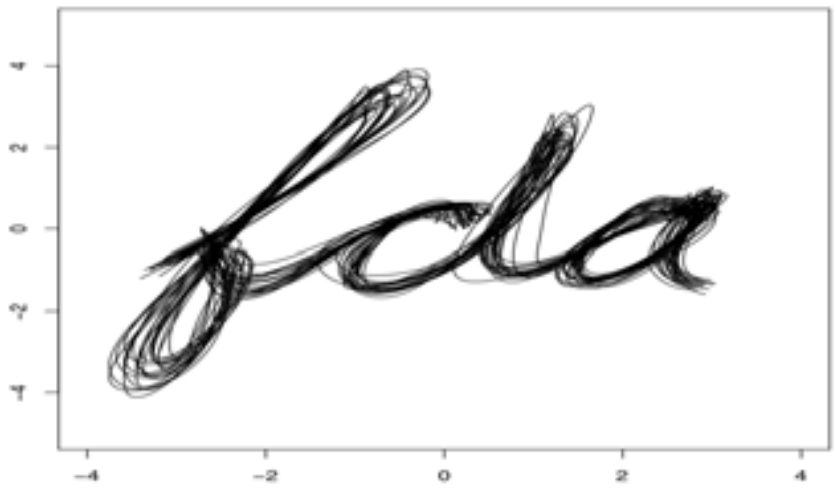
Ejemplos de datos Funcionales

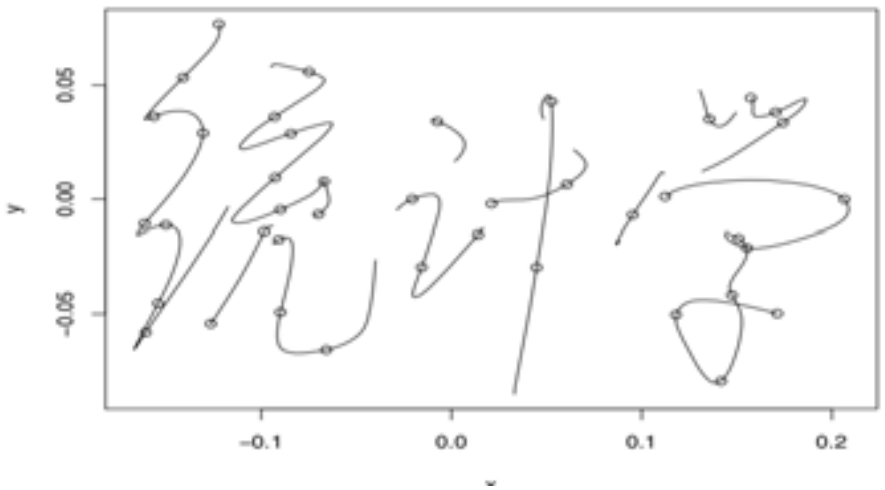
Medición de alturas de persona



Ejemplos de datos Funcionales

Datos de manuscritos





Variable Aleatoria Funcionales

Sea $\mathbb{H}$ el espacio de todas las funciones continuas en $[0,1]$,

$$\mathbb{H} = \{f : [0,1] \to \mathbb{R};\ f \text{ es continua}\}.$$

Definamos el producto punto $\langle \cdot, \cdot \rangle$ de dos funciones f y g en $\mathbb{H}$ como

$$\langle f, g \rangle = \int_0^1 f(s)g(s)ds$$

Una variable aleatoria funcional es una variable aleatoria que toma valores en $\mathbb{H}$,

$$X : \Omega \longrightarrow \mathbb{H}.$$

Si el primer momento existe, la media es definida como el único elemento μ de $\mathbb{H}$ tal que

$$\mathbb{E}(\langle X, z \rangle) = \langle \mu, z \rangle, \text{ for all } z \in \mathbb{H}$$

Variable Aleatoria Funcionales

Sea X una variable aleatoria funcional con segundo momento finito y media cero. El operador de covarianza de X es definido como

$$C : \mathbb{H} \longrightarrow \mathbb{H},$$

tal que

$$\begin{aligned} C(z) &= \mathbb{E}\left(\langle X, z\rangle X\right), \text{ for all } z \in \mathcal{H} \\ &= \int \sigma(u, \cdot) z(u) \mathrm{d}u, \end{aligned}$$

donde $\sigma(v, u) = \mathrm{Cov}(X(v), X(u))$ (*es una superficie*).

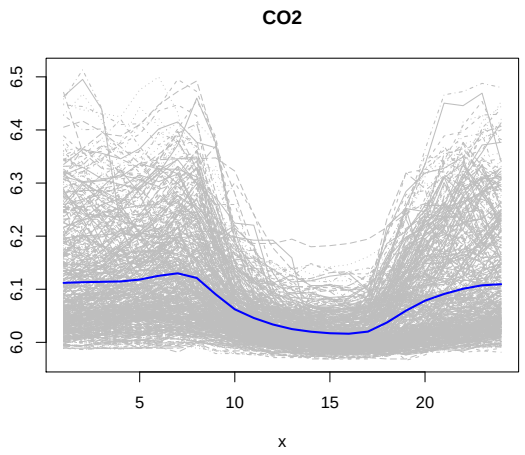
Las funciones $v(s)$ que cumplen

$$\int \sigma(u, s) v(u) \mathrm{d}u = \lambda v(s)$$

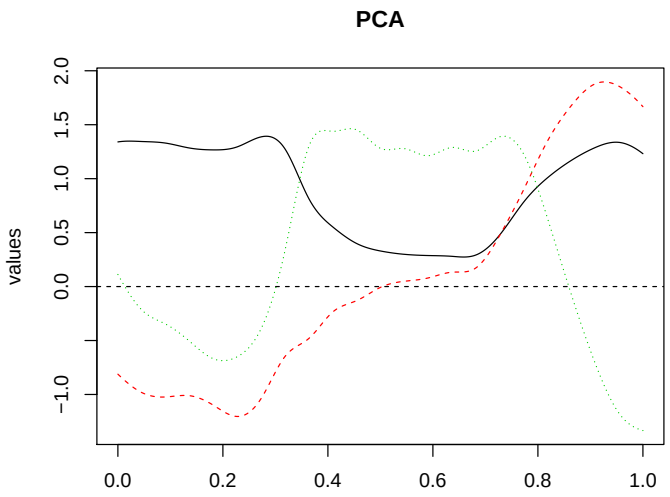
son llamadas eigenfunciones y λ eigenvalores

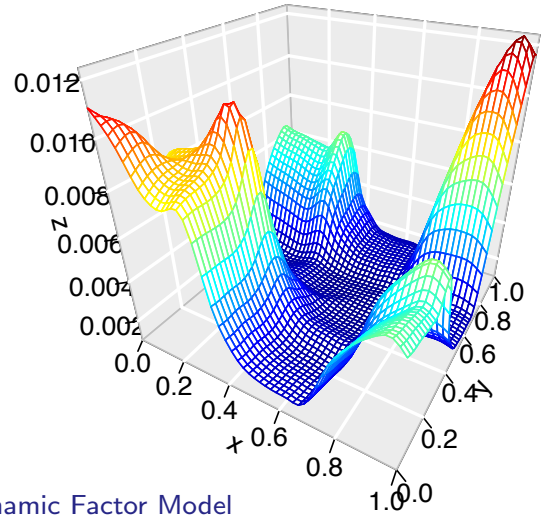
Ejemplos

La curva azul representa la curve media.

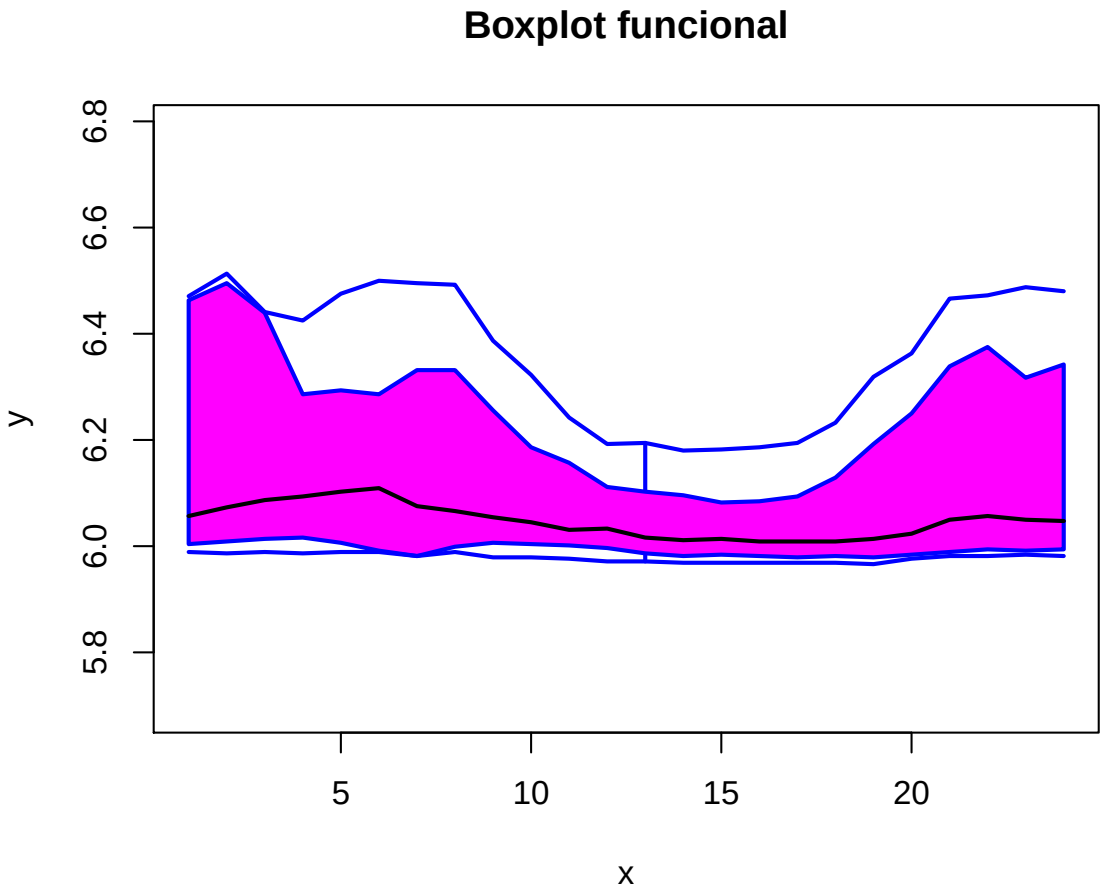


Lado izquierdo:eigenfunciones estimadas, lado derecho: la superfice $\hat{\sigma}(t, s)$.



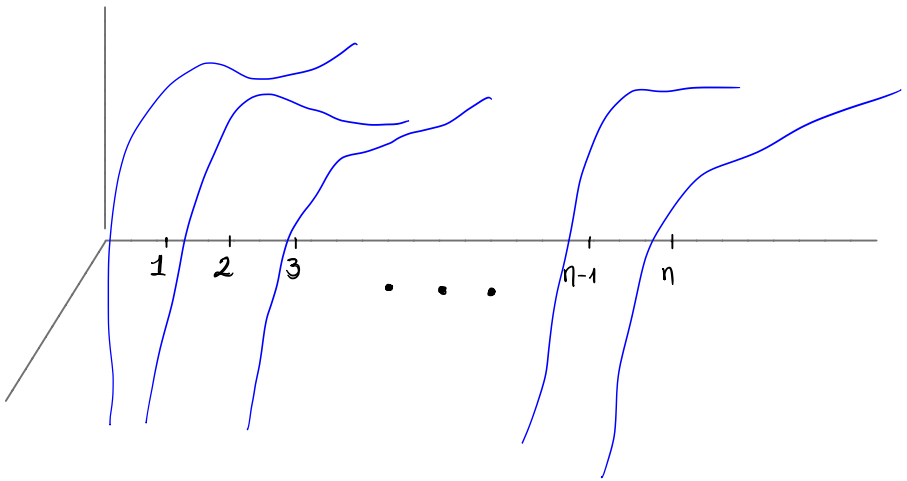


Ejemplos



Serie de Tiempo Funcional

Una serie de tiempo funcional es una sucesión de variables aleatorias funcionales $\{X_n, -\infty < n < \infty\}$

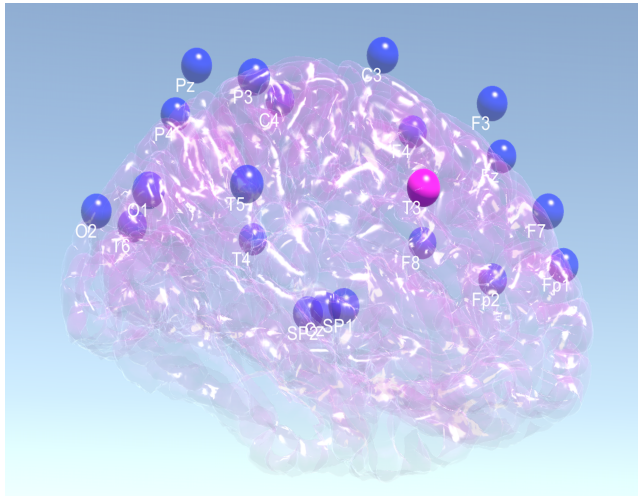


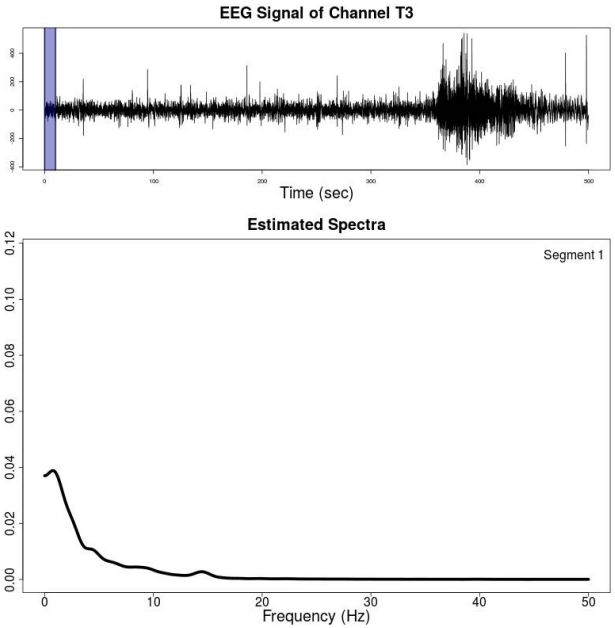
En este caso, tenemos que el operador de covarianza con lag h es (asumiendo que la media es cero)

$$C_{X_n,X_{n+h}}(z) = \mathbb{E}\left(\langle X_n, z\rangle X_{n+h}\right), \text{ for all } z \in \mathcal{H}$$
$$= \int \sigma_h(u,\cdot)z(u)\mathrm{d}u$$

Ejemplos de Series de Tiempo Funcional

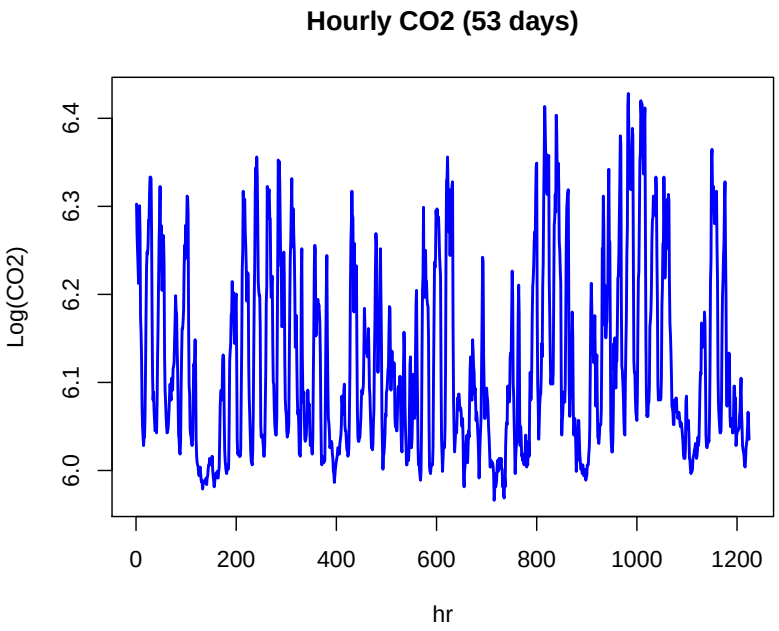
Análisis espectral de electroencefalogramas

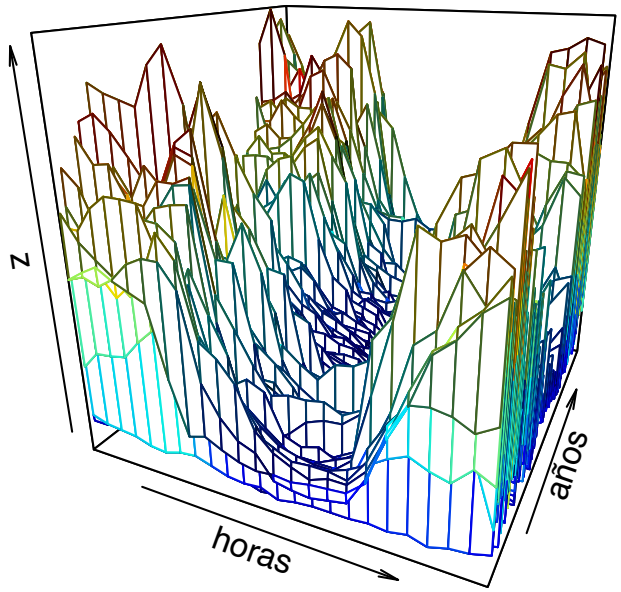




Ejemplos de Series de Tiempo Funcional

A partir de un proceso estocastico continuo





Datos Funcionales

Los datos funcionales se relacionan con:

- Altas dimensiones.
- Estrutura temporal, estructura temporal.
- Alta frecuencia.
- Big Data.

Concepto de Estacionariedad

Una serie de tiempo funcional, $\{X_n, -\infty < n < \infty\}$ se dice que es estacionaria si

- $\mathbb{E}(\|X_n\|^2) < \infty$
- $\mathbb{E}(X_n) = \mu$, es decir, no depende de n
- $C_{X_{n+h},X_{m+h}}(z) = C_{X_n,X_m}(z)$, $z \in \mathcal{H}$ para toda $m, n \in \mathbb{Z}$ y $h \geq 0$, es decir, solo depende de h.

Si $\{X_n\}$ es una serie de tiempo funcional estacionaria entonces, para cualquier $v \in \mathbb{H}$ diferente de cero se tiene que

$$\{\langle X_n, v\rangle\}_n,$$

es una serie de tiempo univariada (escalar) y es estacionaria.

Estimadores

supongamos que observamos una serie de tiempo funcional estacionaria, $\{X_1, \cdots, X_n\}$. El estimador de la media es definida como

$$\hat{\mu}(s) = \frac{1}{N}\sum_{i=1}^{n} X_i(s),$$

y el estimador del operador de covarianza como

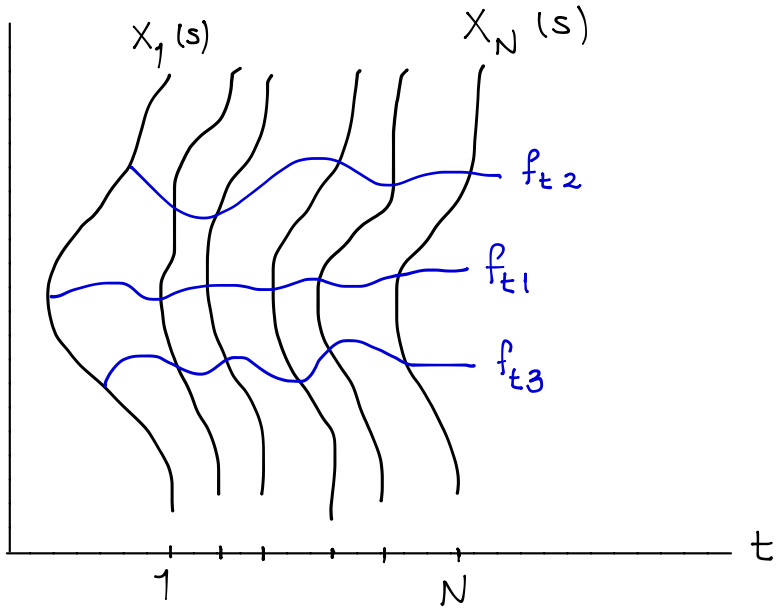
$$\hat{C}_h(z) = \frac{1}{N-h}\sum_{i=1}^{n-h}\langle X_i, z\rangle X_{i+h}.$$

Consecuentemente los estimadores correspondientes a las eigenfunciones,

$$\hat{C}_0(\hat{v}_j)(s) = \lambda_j \hat{v}_j(s).$$

Modelos Dinámicos Factoriales para FTS

La idea de un modelo factorial para datos funcionales es representar a los datos a través un número finito de variables aleatorias,



Se quiere describir la dinámica con series de tiempo univariadas y se suponen que estos no son observados.

Modelos Dinámicos Factoriales para FTS

Sea $\mathbb{X} = \{X_n(s);\ s \in [0,1],\ n = 1, 2, \ldots, N\}$ una serie de tiempo funcional.

$\mathbb{X}$ sigue el modelo functional factorial (FDF) si

$$X_n(s) = \sum_{i=1}^{K} \lambda_i(s) f_{n,i} + \varepsilon_n(s), \tag{1}$$

- tal que para cada $i = 1, \ldots, K$, $\mathbf{f}_i = \{f_{n,i}\ ,\ n = 1, \ldots, N\}$ es una serie de tiempo univariada
- $\mathrm{Cov}(\mathbf{f}_i, \mathbf{f}_j) = 0\ i \neq j$,
- $$\int \lambda_i(t)\lambda_j(t)dt = \delta_{\{j=i\}}.$$
- Los errores $\varepsilon_n(s)$ se asume que están definidos en el espacio ortogonal a $\mathrm{span}\{\lambda_1, \cdots, \lambda_K\}$.

Algunas Propiedades

Si $\mathbb{X}$ sigue el modelo FDF, entonces

- La media del datos funcionales cumple

$$\mathbb{E}\{X_n(s)\} = \sum_{i=1}^{K} \lambda_i(s)\mathbb{E}(f_{n,i}) + \mathbb{E}\{\varepsilon_n(s)\}$$

- El operador de covarianza cumple

$$C_{X_n} = \sum_{i=1}^{K} \mathbb{E}(f_{n,i}^2)\lambda_i \otimes \lambda_i + C_\varepsilon.$$

  Así, está bien definida.

- Para $h \geq 1$,

$$C_h = \sum_{i=1}^{K} \mathrm{Cov}(f_{n,i}, f_{n+h,i})\lambda_i \otimes \lambda_i = \sum_{i=1}^{K} \gamma_i(h)\lambda_i \otimes \lambda_i.$$

Estimadores

Sea $X(t)$ una variable aleatoria funcional. La solución $\{v_j\}_{j=1}^{K}$ del problema de optimización

$$\operatorname{argmin} \mathbb{E}\left(\|X - \sum_{i=1}^{K} \langle X, z_i \rangle z_i\|_2^2 \right),$$

son las eigenfunciones del operador de covarianza C, i.e.,

$$X(s) = \sum_{i=1}^{K} \langle X, v_i \rangle v_i(s) + \epsilon(s).$$

- La expansión de *Karhunen-Loève*,

$$X(t) = \sum_{i=1}^{\infty} \langle X, \lambda_i \rangle v_i(t).$$

Ejemplos

Así, un estimador para las funciones $\lambda_i(s)$ (loading factors), puede ser definido como

$$\hat{\lambda}_i = \hat{v}_i.$$

o alguna rotación de estas. La estimación de las trayectorias de los factores se definen como

$$\hat{\mathbf{f}}_i = \{\langle X_1, \hat{v}_i \rangle, \langle X_2, \hat{v}_i \rangle, \cdots, \langle X_N, \hat{v}_i \rangle\}.$$

Luego se puede ajustar un modelo ARMA(p,q) para cada $\hat{\mathbf{f}}_i$ y a partir.
El numero de factores K puede ser estimado usando por ejemplo el criterio de infromación de akaike (AIC). Sea $\hat{K}$ el estimador que minimiza AIC.

Ejemplo vía Simulación

Simulamos el modelo

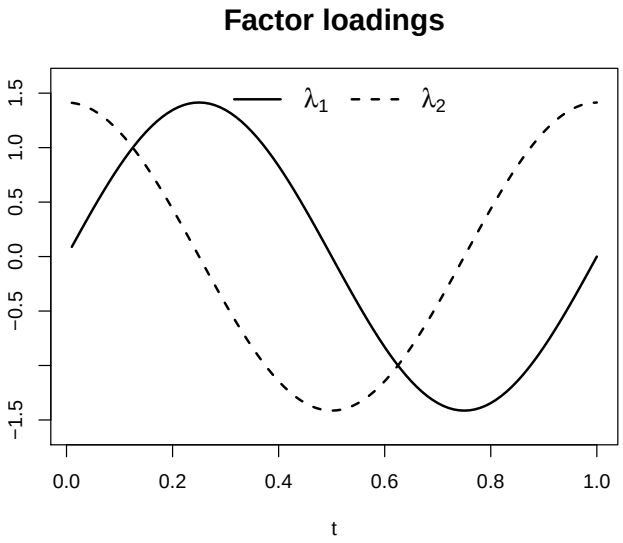
$$X_n(s) = \lambda_1(s) f_{1n} + \lambda_2(s) f_{2n} + \varepsilon_n(s),$$

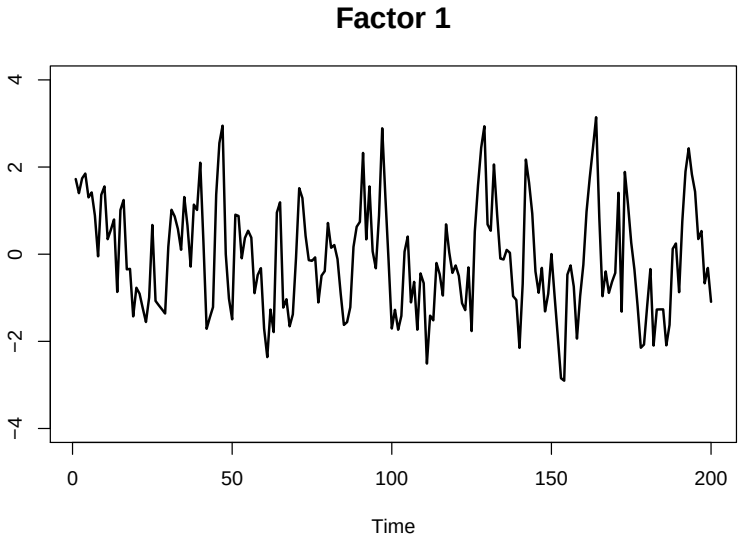
donde

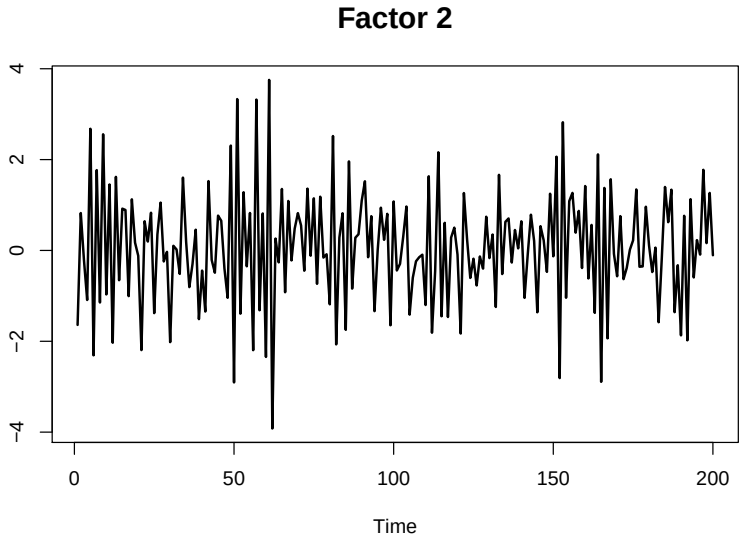
- $\lambda_1(s) = \sin(2\pi s)$ y $\lambda_2(s) = \cos(2\pi s)$, con $s \in [0, 1]$.
- $f_{1n} = 0.6 f_{1,n-1} + \eta_{1n}$ y $f_{2n} = -0.6 f_{2,n-1} + \eta_{2n}$
- $\varepsilon_n(s)$ está generado en términos de la base de Fourier exepto las tres primeras componentes de la base.

El tamaño de muestra considerado es $N = 200$, es decir se simulan $n = 1, 2, \cdots, 200$.

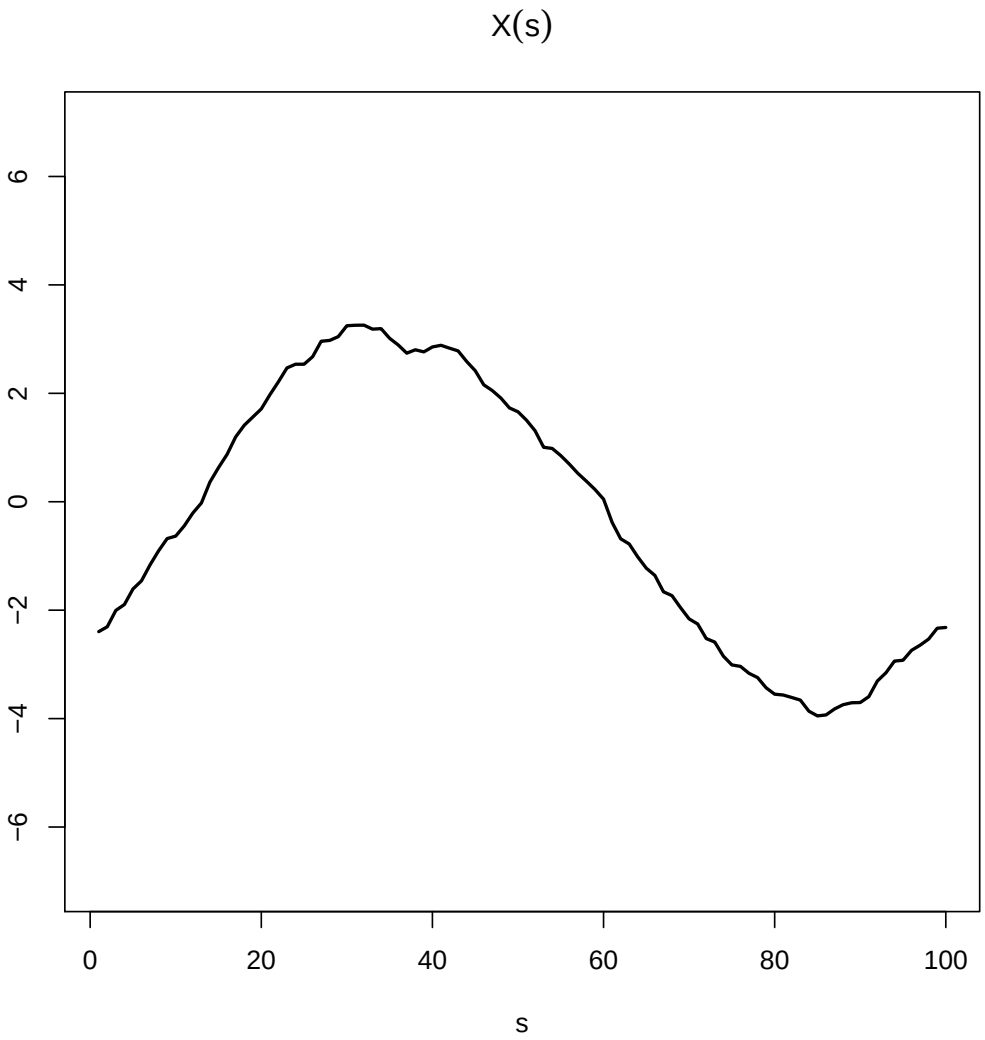
Ejemplo vía Simulación



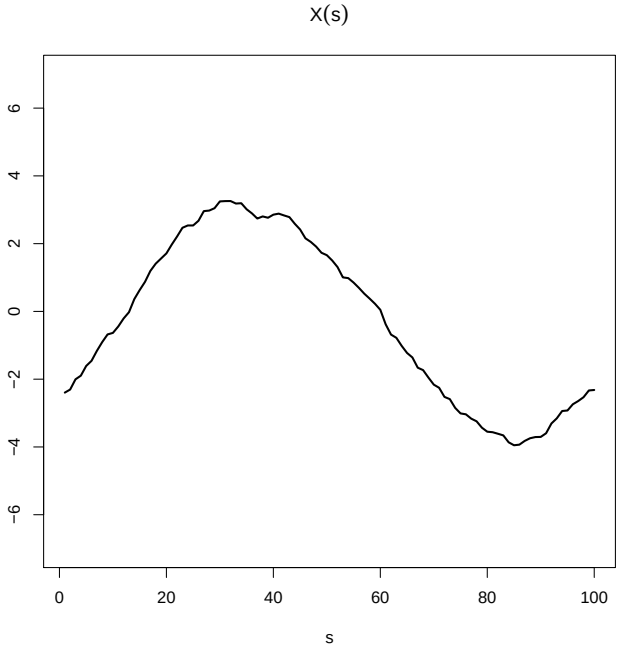


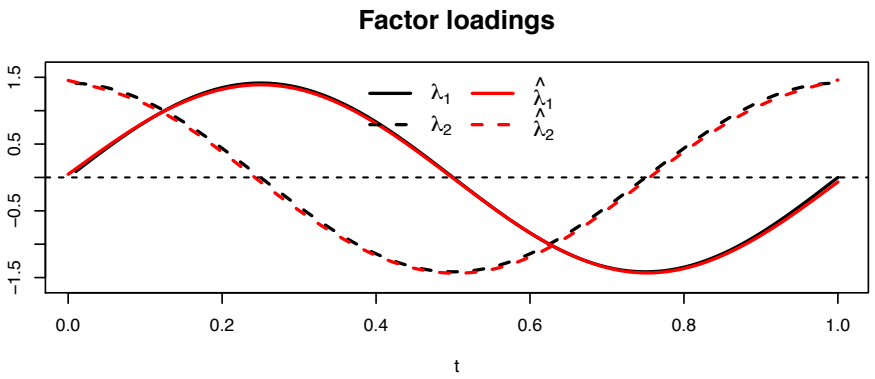


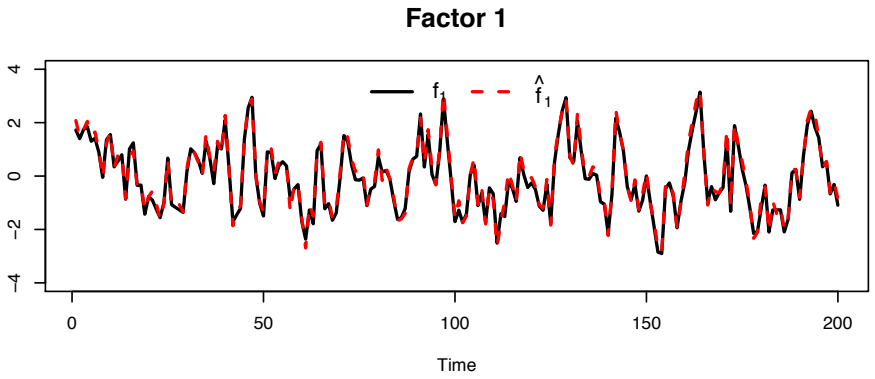
Ejemplo vía Simulación

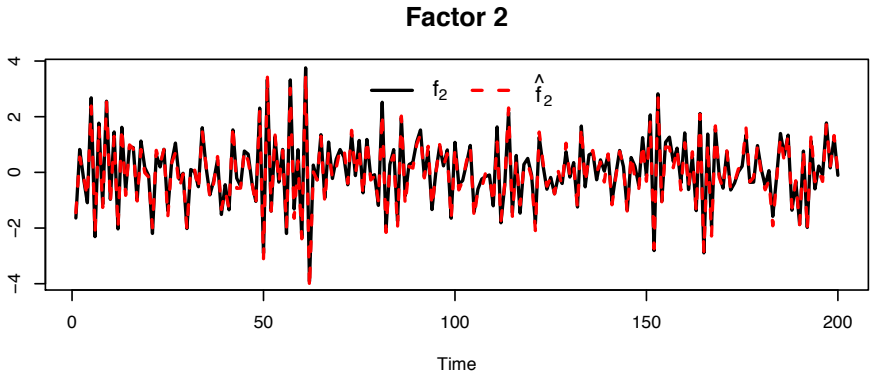


Ejemplo vía Simulación



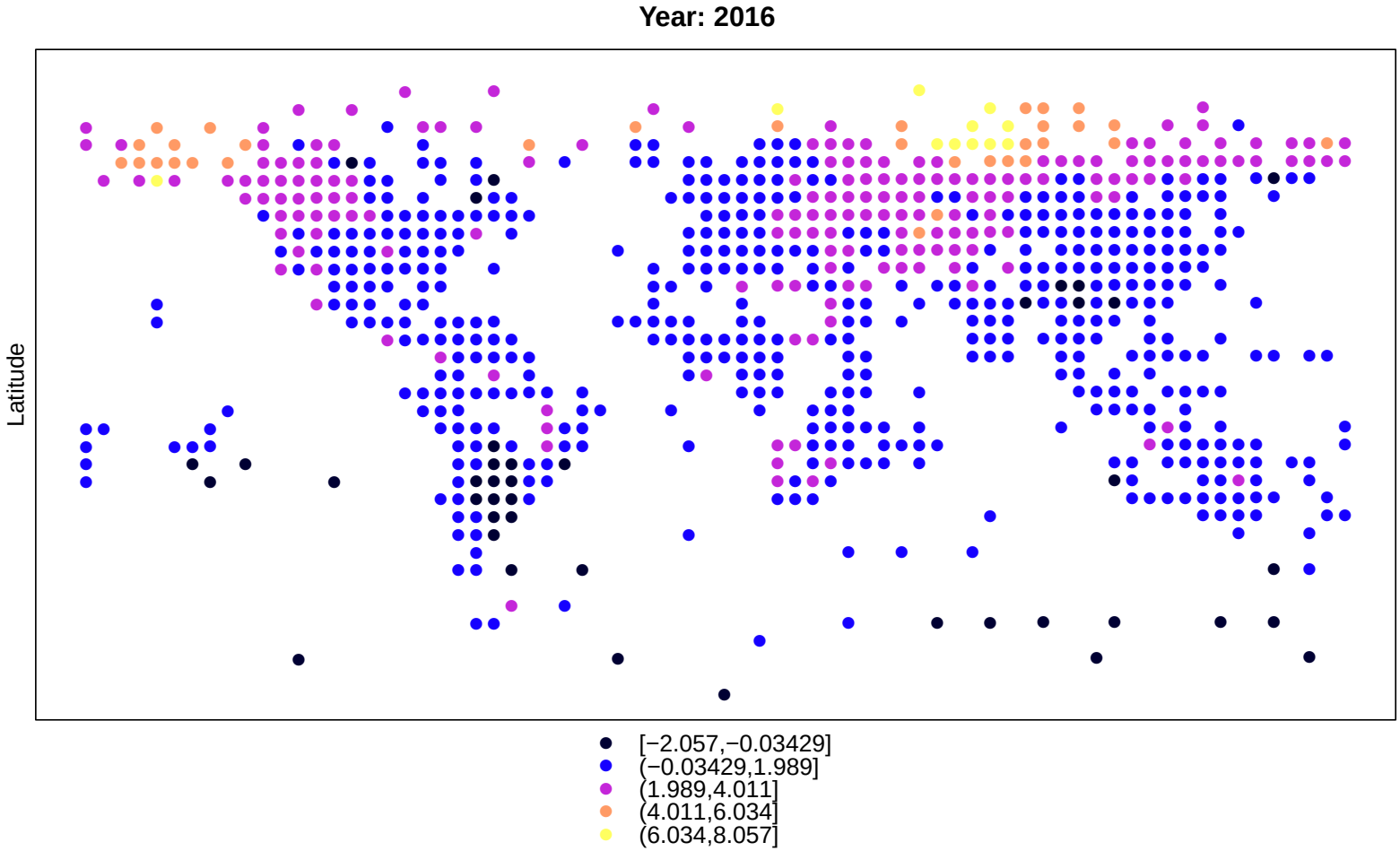






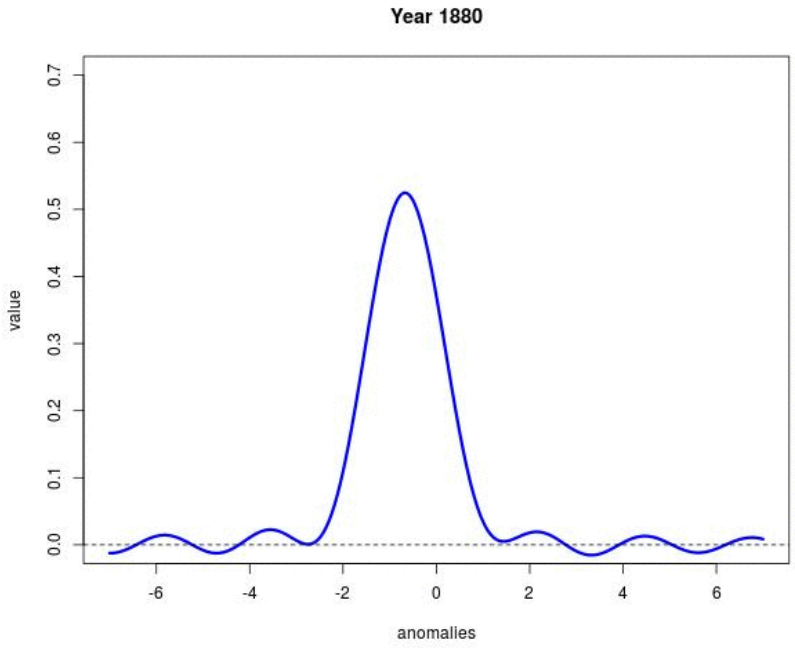
Ejemplo con Datos Reales

Datos de HadCRUT4 de 1880-2016



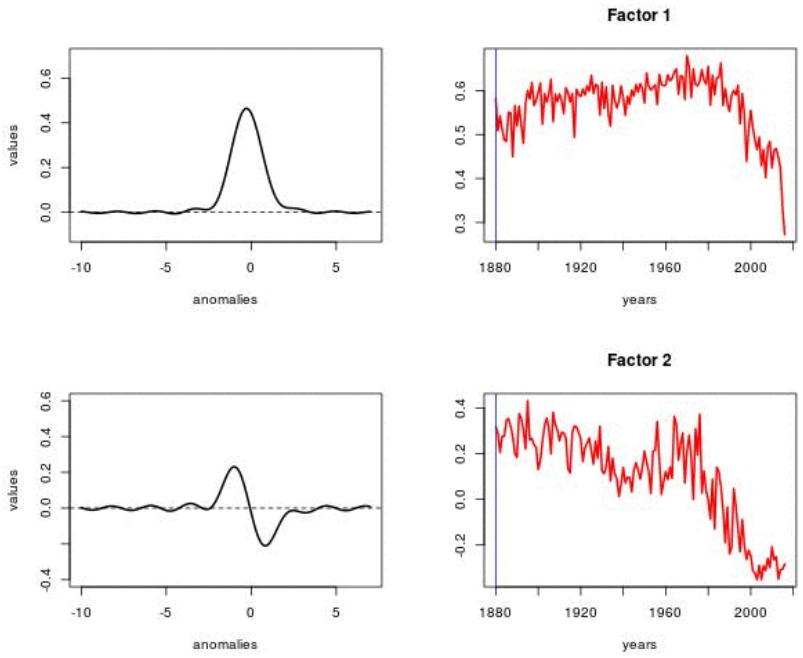
Ejemplo con Datos Reales

Datos de HadCRUT4



Ejemplo con Datos Reales

Estimadores



bibliography

Martinez-Hernandez I., Genton, M.G., and Gonzalez-Farias, G. 2018. Robust depth-based estimation of the functional autoregressive model. Comput. Statist. Data Anal. https://doi.org/10.1016/j.csda.2018.06.003

Bai, J. and Ng, S. 2002. Determining the Number of Factors in Approximate Factor Models. Econometrica 70:191-221.

Fama, E. F.; French, K. R. 1993. Common risk factors in the returns on stocks and bonds. J. Financial Economics

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. 2005. The generalized dynamic factor model: One-sided estimation and forecasting. Journal of the American Statistical Association 100:830-840.

Hays, S., Shen, H., and Huang, J. Z. 2012. Functional dynamic factor models with application to yield curve forecasting. Ann. Appl. Stat. 6:870-894.

Kokoszka, P., Miao, H., and Zhang, X. 2015. Functional dynamic factor model for intraday price curves. Journal of Financial Econometrics 13:456-477.

Liebl, D. 2013. Modeling and forecasting electricity spot prices: a functional data perspective. Ann. Appl. Stat. 7:1562-1592

Ramsay, J. O. and Silverman, B. W. 2005. Functional data analysis. Springer Series in Statistics. Springer, New York, second edition.

Gracias!!

Proyecto de Investigación parcialmente financiado por CONACYT CB-2015-01-252996 y Statistics Program, King Abdullah University of Science and Technology.