Regresión con proyecciones aleatorias para datos funcionales

Paula Rodríguez Díaz

Asesor: Adolfo J. Quiroz

Departamento de Matemáticas
Universidad de Los Andes

15 de noviembre de 2018

Overview

Motivación

- La mayoría de métodos en FDA se relacionan con métodos de estadística multivariada equivalentes
- La regresión funcional por componentes principales utiliza las proyecciones sobre los componentes principales funcionales para llevar a cabo una regresión multivariada
- Se quiere encontrar las funciones de proyección que lleven a un mejor desempeño de la regresión multivariada

Datos de alta dimension como funciones

Dados los datos finito dimensionales $X_1, ..., X_n \in \mathbb{R}^d$ con cada $X_i = (x_{i1}, ..., x_{id})$, se considera una función $\mathcal{X}_i$ para cada X_i tal que

$$\mathcal{X}_i(t_j) = x_{it_j} + \epsilon_{ij} \ , \ j = 1, ..., d$$

Datos de alta dimension como funciones

Dados los datos finito dimensionales $X_1, ..., X_n \in \mathbb{R}^d$ con cada $X_i = (x_{i1}, ..., x_{id})$, se considera una función $\mathcal{X}_i$ para cada X_i tal que

$$\mathcal{X}_i(t_j) = x_{it_j} + \epsilon_{ij} \ , \ j = 1, ..., d$$

Las funciones $\mathcal{X}_i(t)$ son los **datos funcionales** y se representan por medio de **bases funcionales**

Datos de alta dimension como funciones

Dados los datos finito dimensionales $X_1, \ldots, X_n \in \mathbb{R}^d$ con cada $X_i = (x_{i1}, \ldots, x_{id})$, se considera una función $\mathcal{X}_i$ para cada X_i tal que

$$\mathcal{X}_i(t_j) = x_{it_j} + \epsilon_{ij} \ , \ j = 1, \ldots, d$$

Las funciones $\mathcal{X}_i(t)$ son los **datos funcionales** y se representan por medio de **bases funcionales**

Definición (Base Funcional)

Un conjunto de funciones $\{\phi_1, \phi_2, \ldots\}$ es una base funcional en $\mathcal{L}_2(T)$ si toda función $\mathcal{X} \in \mathcal{L}_2(T)$ tiene una única descomposición

$$\mathcal{X}(t) = \sum_{j=1}^{\infty} c_j \phi_j(t) \ , \ c_j \in \mathbb{R} \ \forall j. \tag{1}$$

Datos de alta dimension como funciones

Dado un sistema de funciones base $\{\phi_1, ..., \phi_K\}$ se quiere tener los datos funcionales

$$\mathcal{X}_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \tag{2}$$

tales que

$$x_{ij} = \sum_{k=1}^{K} c_{ik}\phi_k(t_j) + \epsilon_{ij} \tag{3}$$

para cada $i = 1, ..., n$ y $j = 1, .., d$.

El suavizamiento de cada dato X_i corresponde a resolver d regresiones lineales.

Regresion lineal funcional

Dados los datos funcionales $\mathcal{X}_1, ..., \mathcal{X}_n$ y las respuestas escalares $Y_1, ..., Y_n$ se considera el modelo de regresion

$$Y_i = \alpha_0 + \int_{\mathcal{T}} \beta(t)\mathcal{X}_i(t)dt + \epsilon(t) \tag{4}$$

donde $\alpha_0 \in \mathbb{R}$ y $\beta \in \mathcal{L}_2(T)$.

Regresión funcional por bases funcionales y penalización

La regresión anterior se resuelve por mínimos cuadrados; penalizando irregularidades en la función de regresión estimada

$$\sum_{i=1}^{n}\left[Y_i-\alpha-\int \mathcal{X}_i(t)\beta(t)dt\right]^2+\lambda\int\left[\frac{d^2}{dt^2}\beta(t)\right]^2 dt. \tag{5}$$

Regresión funcional por bases funcionales y penalización

La regresión anterior se resuelve por mínimos cuadrados; penalizando irregularidades en la función de regresión estimada

$$\sum_{i=1}^{n}\left[Y_i-\alpha-\int \mathcal{X}_i(t)\beta(t)dt\right]^2+\lambda\int\left[\frac{d^2}{dt^2}\beta(t)\right]^2 dt. \tag{5}$$

Se fija una base $\psi_1, ..., \psi_B$ para representar la función $\beta(t)$ y se estiman los coeficientes $b_1, ..., b_B \in \mathbb{R}$ tales que

$$\beta(t)=\sum_{k=1}^{B} b_k\psi_k(t) \tag{6}$$

Regresión funcional por componentes principales

Cada dato funcional $\mathcal{X}_i$ se puede expresar como

$$\mathcal{X}_i(t) = \bar{\mathcal{X}}(t) + \sum_{j=1}^{p} f_{ij}\xi_j(t) \tag{7}$$

donde $\xi_j : T \to \mathbb{R}$ es la j-ésima componente principal funcional y f_{ij} es el *puntaje* de la componente j con respecto a la observación i.

Regresión funcional por componentes principales

Cada dato funcional $\mathcal{X}_i$ se puede expresar como

$$\mathcal{X}_i(t) = \bar{\mathcal{X}}(t) + \sum_{j=1}^{p} f_{ij}\xi_j(t) \tag{7}$$

donde $\xi_j : T \to \mathbb{R}$ es la j-ésima componente principal funcional y f_{ij} es el *puntaje* de la componente j con respecto a la observación i.

Se considera el modelo de regresión

$$Y_i = b_0 + \sum_{j=1}^{p} f_{ij}b_j + \epsilon_i \tag{8}$$

para $i = 1, ..., n$. Un modelo de regresión múltiple estándar.

Regresión funcional por componentes principales

Como $\{\xi_1, ..., \xi_p\}$ es un sistema ortogonal de funciones entonces $f_{ij} = \langle \xi_j, \mathcal{X}_i - \bar{\mathcal{X}} \rangle$.

Con esto se puede recuperar la función de regresión estimada

$$\widehat{\beta}(t) = \sum_{j=1}^{p} \widehat{b}_j \xi_j(t)$$

considerando el modelo de regresión funcional

$$Y_i = \alpha_0 + \int_{\mathcal{T}} \beta(t) \mathcal{X}_i(t) dt + \epsilon(t)$$

Método Propuesto

Dadas las observaciones funcionales $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ y las variables de respuesta $Y_1, ..., Y_n$ se quiere encontrar un conjunto de funciones $\{\rho_1^\star(t), ..., \rho_m^\star(t)\}$ tal que el modelo de regresión

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = b_0 + \sum_{j=1}^{m} b_j \begin{bmatrix} \langle \mathcal{X}_1, \rho_j^\star \rangle \\ \vdots \\ \langle \mathcal{X}_n, \rho_j^\star \rangle \end{bmatrix} + \epsilon \tag{9}$$

tenga un R^2*ajustado* a lo más cercano posible a 1.

Método Propuesto

Dadas las observaciones funcionales $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ y las variables de respuesta $Y_1, ..., Y_n$ se quiere encontrar un conjunto de funciones $\{\rho_1^\star(t), ..., \rho_m^\star(t)\}$ tal que el modelo de regresión

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = b_0 + \sum_{j=1}^{m} b_j \begin{bmatrix} \langle \mathcal{X}_1, \rho_j^\star \rangle \\ \vdots \\ \langle \mathcal{X}_n, \rho_j^\star \rangle \end{bmatrix} + \epsilon \tag{9}$$

tenga un R^2*ajustado* a lo más cercano posible a 1.

Se transforma el modelo de regresión funcional en un modelo de regresión multivariada equivalente determinado por las funciones de proyección $\rho_1^\star(t), ..., \rho_m^\star(t)$.

Método Propuesto

Generar *funciones aleatorias* en un espacio de funciones determinado y seleccionar adecuadamente aquellas funciones que conllevan a un mejor ajuste del modelo de regresión.

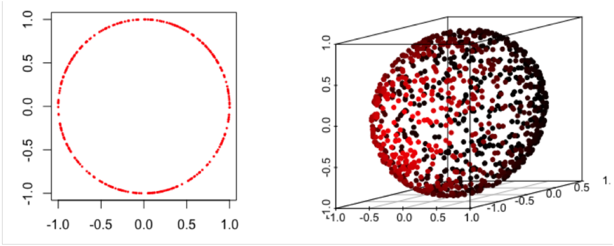
Generación de funciones aleatorias

Se generan funciones aleatorias en el espacio generado por las funciones base $\{\phi_1, \ldots, \phi_d\}$.

Una función en este espacio es de la forma

$$g = \sum_{i=1}^{d} c_i \phi_i = \mathbf{c}' \boldsymbol{\phi}.$$

Por lo tanto, basta con generar coeficientes aleatorios $\mathbf{c} \in S^{d-1}$.

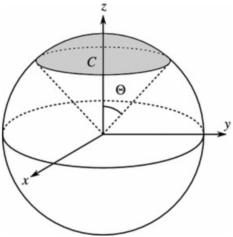


Generación de funciones aleatorias

Si f^* es la 'mejor' función de proyección en el espacio de funciones generado por $\{\phi_1, ..., \phi_d\}$, con

$$f^* = \sum_{i=1}^{d} a_i \phi_i = \mathbf{a}'\phi,$$

se quiere generar N coeficientes aleatorios en S^{d-1} tal que al menos de ellos esté a distancia angular menor o igual θ de $\mathbf{a}$.



Generación de funciones aleatorias

Sea $\Delta(S^{d-1}, \theta)$ la hipercapa centrada en $\mathbf{a} \in S^{d-1}$ con ángulo colatitudinal $\theta \in (0, 2\pi)$.

Generación de funciones aleatorias

Sea $\Delta(S^{d-1}, \theta)$ la hipercapa centrada en $\mathbf{a} \in S^{d-1}$ con ángulo colatitudinal $\theta \in (0, 2\pi)$.

Si $\delta = A(S^{d-1}, \theta)/A(S^{d-1})$ es la razón de área de la hipercapa $\Delta(S^{d-1}, \theta)$ y S^{d-1}, entonces la probabilidad de que un punto aleatorio uniforme en S^{d-1} no esté en $\Delta(S^{d-1}, \theta)$ es $(1 - \delta)$. Por lo tanto,

$$P\Big(\mathbf{c}_1, ..., \mathbf{c}_N \notin \Delta(S^{d-1}, \theta)\Big) = (1 - \delta)^N. \tag{10}$$

Generación de funciones aleatorias

Sea $\Delta(S^{d-1}, \theta)$ la hipercapa centrada en $\mathbf{a} \in S^{d-1}$ con ángulo colatitudinal $\theta \in (0, 2\pi)$.

Si $\delta = A(S^{d-1}, \theta)/A(S^{d-1})$ es la razón de área de la hipercapa $\Delta(S^{d-1}, \theta)$ y S^{d-1}, entonces la probabilidad de que un punto aleatorio uniforme en S^{d-1} no esté en $\Delta(S^{d-1}, \theta)$ es $(1-\delta)$. Por lo tanto,

$$P\Big(\mathbf{c}_1, ..., \mathbf{c}_N \notin \Delta(S^{d-1}, \theta)\Big) = (1-\delta)^N. \tag{10}$$

Tomando $N = \frac{1}{\delta}\log(\frac{1}{\delta^2})$ se tiene que $(1-\delta)^N = (1-\delta)^{\frac{1}{\delta}\log(\frac{1}{\delta^2})} \approx \delta^2$ cuando $\delta \ll 1$.

Generación de funciones aleatorias

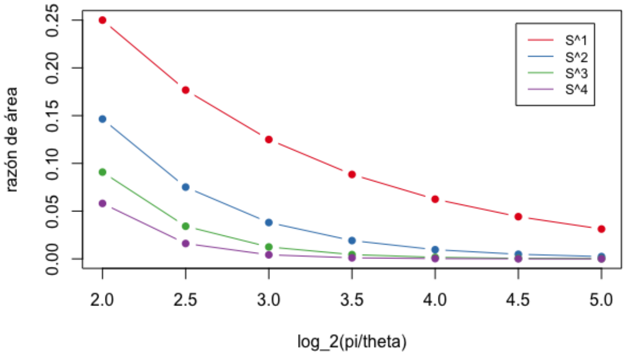
La razón de área de la hipercapa $\Delta(S^{d-1}, \theta)$ y S^{d-1} se puede calcular como

$$\delta_{d-1,\theta} = \frac{1}{2} I_{\sin^2\theta}\left(\frac{d-1}{2}, \frac{1}{2}\right). \tag{11}$$

donde $I_x(a, b)$ es la función beta incompleta regularizada.

Dimensión $d-1$	**Razón de área** $\delta_{d-1,\theta}$
1	θ/π
2	$\frac{1}{2}(1-\cos\theta)$
3	$\frac{1}{2\pi}(2\theta - \sin 2\theta)$
4	$\frac{1}{4}(2 - 3\cos\theta + cos^3\theta)$

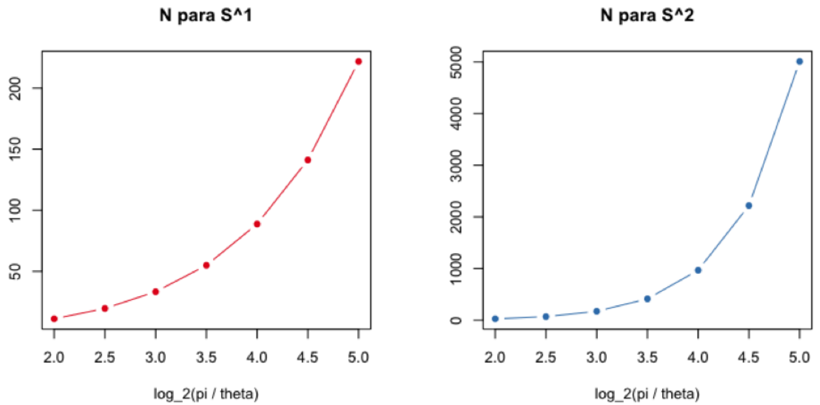
Generación de funciones aleatorias



Tomando $N = \frac{1}{\delta}\log(\frac{1}{\delta^2})$ se tiene que para $\theta \leq \frac{\pi}{8}$ y $d \geq 2$,

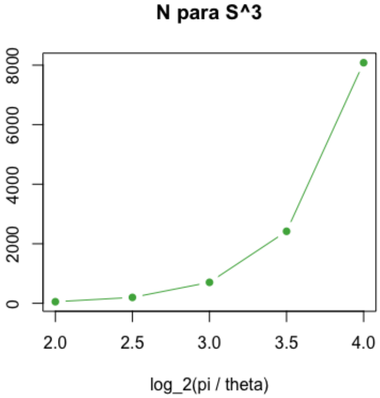
$$P\Big(\mathbf{c}_1, \ldots, \mathbf{c}_N \notin \Delta(S^{d-1}, \theta)\Big) < 2{,}5 \times 10^{-3}$$

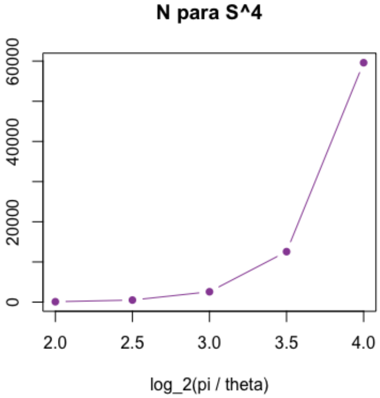
Generación de funciones aleatorias



Ejemplo 1 Ejemplo 2

Generación de funciones aleatorias





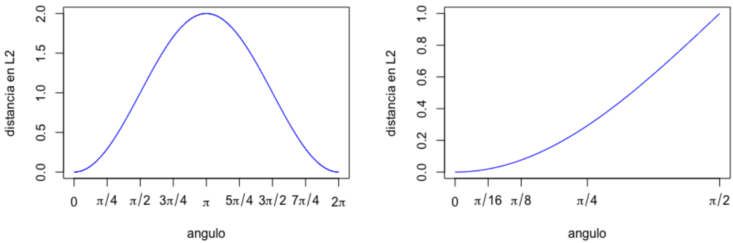
Distancia entre funciones aleatorias y mejor función de proyección

Proposición

Sea $\phi = (\phi_1, ..., \phi_d)$ *un sistema ortogonal de funciones. Si* $\mathbf{a}, \mathbf{c} \in S^{d-1}$ *son tales que* $\cos^{-1}(\mathbf{a}, \mathbf{c}) \leq \theta$ *con* $\theta \in [0, \pi/2]$ *entonces*

$$\|f^* - g\|_2^2 \leq 2(1 - \cos(\theta))$$

con $f^* = \mathbf{a}'\phi$ *y* $g = \mathbf{c}'\phi$. *Donde* $\|\cdot\|_2$ *es la norma en* $\mathcal{L}_2$



Selección de mejores funciones de proyección

Algoritmo 1: Selección de mejores funciones de proyección

1. Teniendo el conjunto de funciones aleatorias $\mathcal{P}$, crear $P_1, ..., P_M$ subconjuntos de K funciones escogidas uniformemente de $\mathcal{P}$. $P_i = \{\rho_{i_1}, ..., \rho_{i_K}\}$
2. Para $i = 1, 2, ..., M$
   (a) Para $k = 1, ..., K$: $Z_{i,k} \leftarrow \left(\langle \mathcal{X}_1, \rho_{i_k}\rangle, ..., \langle \mathcal{X}_n, \rho_{i_k}\rangle\right)$
   (b) Hacer selección de predictores sobre $Z_{i,1}, ..., Z_{i,K}$ y llamar $\mathcal{M}^i$ el conjunto de predictores seleccionados
3. Para $i = 1, 2, ..., M$: Llevar a cabo regresión multivariada con variables independientes $\mathcal{M}^i$ y variables dependientes $Y_1, ..., Y_n$. Almacenar $R^2 ajustado$ en R_i

Selección de mejores funciones de proyección

Algoritmo 2: Selección de mejores funciones de proyección

4. **Puntaje por subconjunto:** Asignar un puntaje en función de R_i a cada subconjunto de variables $\mathcal{M}^i$.

5. **Puntaje por predictor (función):** Dar puntaje a cada función de $\mathcal{P}$ según puntaje de los $\mathcal{M}^i$ en los que están sus predictores corespondientes.

6. Seleccionar los K predictores con mejor puntaje y hacer la misma selección de predictores hecha en el punto 2 (b). Las funciones correspondientes a los predictores obtenidos son consideradas las mejores funciones de proyeccón.

Selección de predictores (funciones)

Algoritmo 3: Selección paso a paso hacia atrás por R^2*ajustado*

1. Sea $\mathcal{M}_p$ el modelo completo que contiene los p predictores
2. Para $k = p, p-1, ..., 1$:
   (a) Considerar los k modelos que contienen todos menos uno de los predictores en $\mathcal{M}_k$, para un total de $k-1$ predictores
   (b) Escoger el mejor de los k modelos según R^2 y llamarlo $\mathcal{M}_{k-1}$
3. Escoger el mejor modelo entre $\mathcal{M}_0, ..., \mathcal{M}_p$ según R^2*ajustado*.

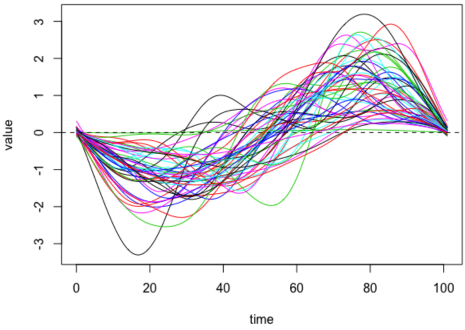
Selección de predictores (funciones)

Algoritmo 4: Selección de predictores por significancia en regresión

1. Sea $\mathcal{M}_p$ el modelo completo que contiene los p predictores y pv_{max} el máximo valor de los p-valores.
2. Mientras que $pv_{max} > 0{,}05$:
   (a) Sea $\mathcal{M}_{p-1}$ el modelo con los predictores de $\mathcal{M}_p$ menos el predictor que tiene p-valor pv_{max}
   (b) Llevar a cabo una regresión en el modelo $\mathcal{M}_{p-1}$ y llamar pv_{max} al máximo p-valor obtenido entre los $p-1$ predictores.
   (c) $p \leftarrow p-1$
3. Cuando $pv_{max} \leq 0{,}5$ se escoge el modelo $\mathcal{M}_p$

Ejemplo 1: Altura instantánea de olas

- Altura instantánea de 865 olas del mar durante 101 momentos
- Predecir la energía total de cada ola
- Suavizamiento utilizando una base de Fourier de tamaño 21



Ejemplo 1: Método Propuesto

- Se generan funciones aleatorias como combinación lineal de distintos sistemas de funciones
- Se consideraron 3 casos con sistemas de funciones distintos
- Para cada sistema se generan funciones aleatorias como combinación lineal de 2 funciones
- En cada caso se hizo selección de predictores paso a paso hacia atrás por R^2*ajustado* y por significancia

tamaño muestra (N)

	Base Inicial	**# combinaciones**	***N* por combinación**
Caso 1	Fourier 11	100	300
Caso 2	Indicadora 10	45	500
Caso 3	Indicadora 25	6	1.000

Cuadro: *Parámetros del modelo para cada caso considerado*

Ejemplo 1: Predictores por significancia

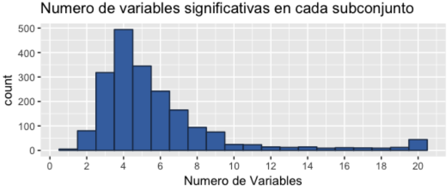


Figura: *Predictores Significativos Caso 1.*

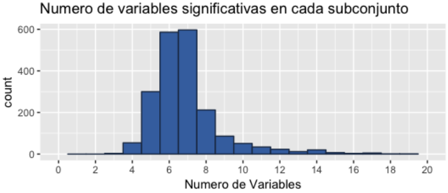


Figura: *Predictores Significativos Caso 2.*

Ejemplo 1: Mejores funciones de proyección por significancia

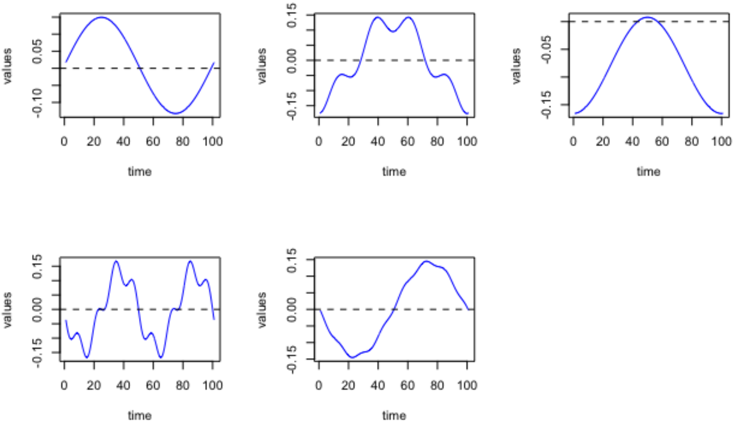


Figura: Mejores funciones de proyección Caso 1

Ejemplo 1: Mejores funciones de proyección por significancia

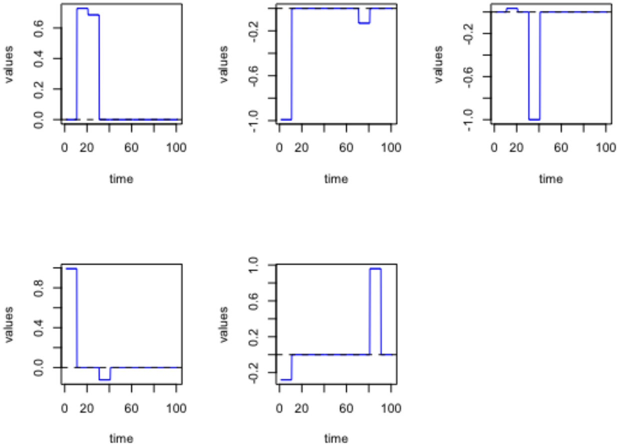


Figura: Mejores funciones de proyección Caso 2

Ejemplo 1: Predictores por paso a paso hacia atrás

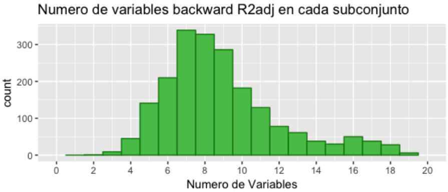


Figura: *Predictores Backward Caso 1.*

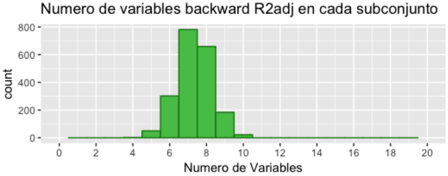


Figura: *Predictores Backward Caso 2.*

Ejemplo 1: Mejores funciones de proyección por significancia

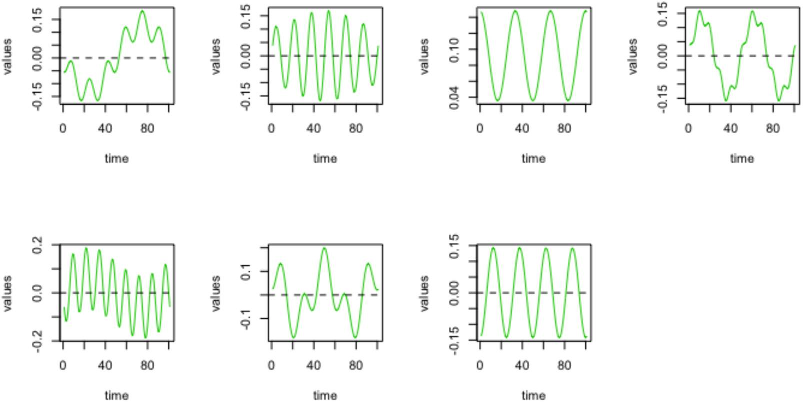


Figura: Mejores funciones de proyección Caso 1

Ejemplo 1: Mejores funciones de proyección por significancia

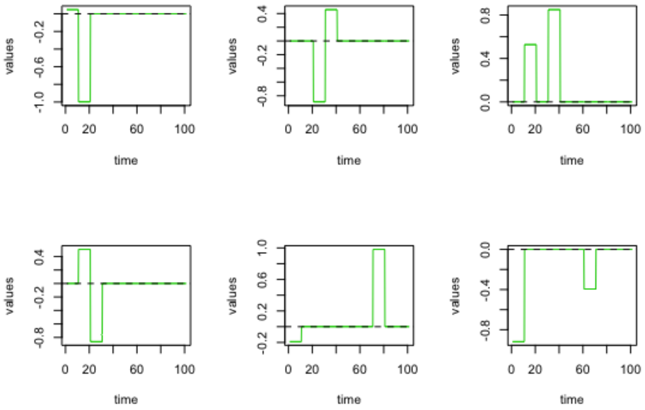


Figura: Mejores funciones de proyección Caso 2

Ejemplo 1: Resultados método propuesto

Método Propuesto	# de funciones	R^2*ajustado*
Caso 1.1	5	0,8912
Caso 1.2	7	0,8919
Caso 2.1	5	0,8896
Caso 2.2	6	0,8876
Caso 3.1	4	0,8864
Caso 3.2	4	0,8864

El ajuste de los modelos de regresión disminuye al utilizar funciones más sencillas. Sin embargo, la diferencia es considerablemente pequeña.

Ejemplo 1: Comparación con regresión por componentes principales

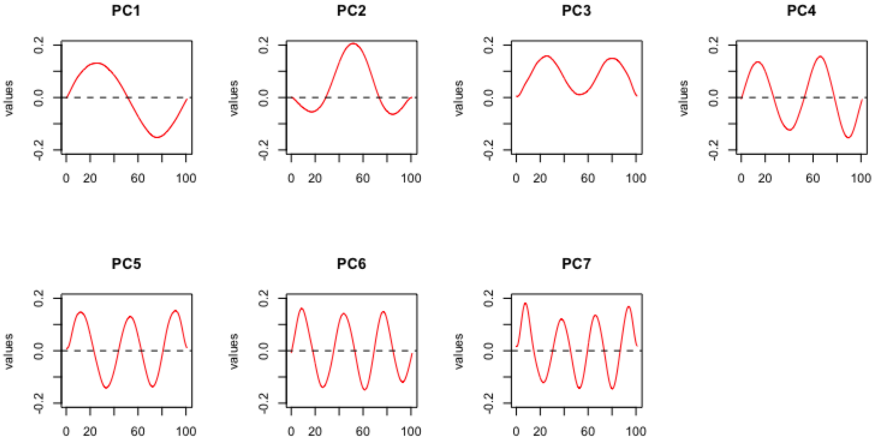


Figura: Componentes principales funcionales

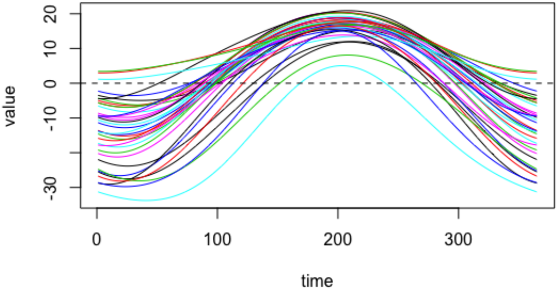
Ejemplo 1: Comparación con regresión por componentes principales

# de componentes	$R^2 ajustado$
1	0,889
2	0,889
3	0,889
4	0,890
5	0,891
6	0,891
7	0,891

- Las funciones indicadoras encontradas en el Caso 2 y 3 tienen un desempeño similar al de las componentes principales funcionales
- La cantidad de funciones encontradas por el método propuesto es similar a la cantidad de CP que explican el 95 % de la varianza.

Ejemplo 2: Temperatura anual en Canadá

- Temperatura diaria de 35 regiones en Canadá durante un año
- Predecir el logaritmo de la precipitación promedio anual en cada región
- Suavizamiento utilizando una base de Fourier de tamaño 5

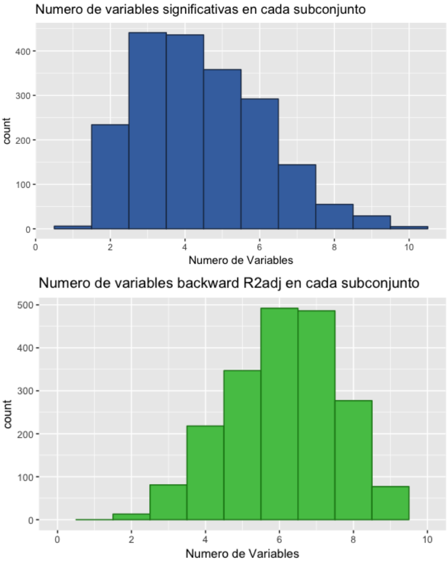


Ejemplo 2: Método Propuesto

- Se generaron funciones aleatorias como combinación lineal de 3 funciones del sistema $\{1, \sin(x), \cos(x), \sin(2x), \cos(2x)\}$.
- Para cada combinación posible se generaron $N = 1{,}000$ funciones aleatorias.
- Se llevó cabo el método propuesto seleccionando predictores con paso a paso hacia atrás por R^2*ajustado* y por significancia.

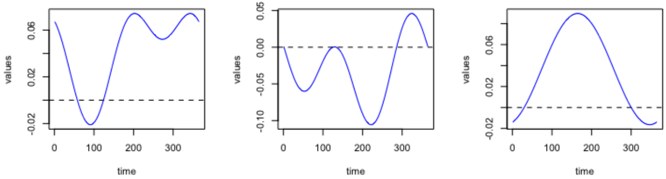
tamaño muestra (N)

Ejemplo 2: Selección de predictores

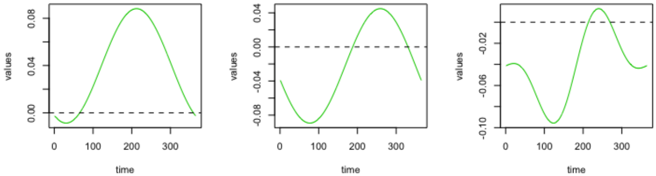


Ejemplo 2: Mejores funciones de proyección

Selección de predictores por significancia, $R^2 adj = 0{,}769$:



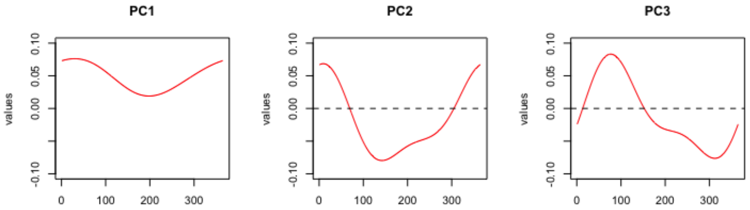
Selección de predictores paso a paso hacía atrás por $R^2 ajustado$, $R^2 adj = 0{,}769$:



Ejemplo 2: Comparación con regresión funcional por componentes principales

Se lleva a cabo una regresión funcional por componentes principales con la misma cantidad de funciones que se encontraron en el método propuesto.

Componentes principales funcionales ($R^2 adj = 0{,}691$)



Conclusiones

- El método propuesto brinda un acercamiento a la dimensión del problema de regresión
- La cantidad de mejores funciones de proyección encontradas es similar a la cantidad de CPF necesarios para explicar el 98 % de la varianza aproximadamente.
- En el Ejemplo 1 se encontraron funciones que se expresan en bases funcionales más sencillas que los CPF y sin embargo alcanzan un desempeño similar.
- En el Ejmplo 2 se encontraron funciones de de proyección que conllevan a un $R^2 ajustado$ mayor que el de regresión por CPF. La diferencia fue de 0,07.

References

J. O. Ramsay and C. J. Dalzell (1991)
Some Tools for Functional Data Analysis
Journal of the Royal Statistical Society. Series B (Methodological) 53(3), 39–572.

J. O. Ramsay and W. Silverman (2005)
Functional Data Analysis

Ramsay, J. and Hooker, G. and Graves, S. (2009)
Functional Data Analysis with R and MATLAB

James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert (2014)
An Introduction to Statistical Learning: With Applications in R

Kokoszka, P. and Reimherr, M. (2017)
Introduction to Functional Data Analysis