No. 11

ABRIL de 2019

quantil

Documentos de Trabajo

ISSN 1234-5790 Edición electrónica

Regresión con Proyecciones Aleatorias para Datos Funcionales

Paula Rodríguez Díaz

quantil | matemáticas aplicadas

Serie Documentos de Trabajo Quantil, 2019-10
Edición electrónica.

ABRIL de 2019

Comité editorial:
Álvaro J. Riascos, CoDirector General y Director Modelos Económicos e I&D
Diego Jara Pinzón, CoDirector General y Director Matemáticas Financieras
Natalia Iregui Parra, Directora Administrativa
Andrés F. Galeano Zurbaran, Director Asociado Matemáticas Financieras
Mateo Dulce Rubio, Director Asociado Minería de Datos
David Delgado, Líder de Tecnologías de la Información
Hamadys L. Benavides Gutiérrez, Investigadora

© 2019, Quantil S.A.S., Minería de Datos,
Carrera 7 # 77 - 07. Oficina 901, Bogotá, D. C., Colombia
Teléfonos: 3718132 – (310)6791459 – (320)8461236
E-mail: info@quantil.com.co
http://www.quantil.com.co

Impreso en Colombia – Printed in Colombia

La serie de Documentos de Trabajo Quantil se circula con propósitos de discusión y divulgación. Los artículos no han sido evaluados por pares ni sujetos a ningún tipo de evaluación formal por parte del equipo de trabajo de Quantil.

Publicado bajo licencia:

Atribución – Compartir igual

Creative Commons: https://co.creativecommons.org

Regresión con Proyecciones Aleatorias para Datos Funcionales*

Paula Rodríguez Díaz†

Resumen

En este proyecto se trabaja el problema de regresión en el contexto de datos funcionales. Teniendo en cuenta el método de regresión funcional por componentes principales, se desean encontrar las mejores funciones de proyección para mejorar el desempeño del modelo de regresión. Es decir que se espera encontrar funciones de proyección que tengan un mejor desempeño en regresión que las componentes principales funcionales. El método propuesto para encontrar estas funciones se lleva a cabo a partir de proyecciones sobre funciones aleatorias. Además, se hacen expermientos del método propuesto sobre dos bases de datos funcionales con respuesta escalar.

Palabras claves: Análisis de Datos Funcionales, Regresión Funcional, Componentes Principales Funcionales, Proyecciones Funcionales. Códigos JEL: C02, C49, C61.

*Agradezco de forma especial a Adolfo J. Quiroz por asesorar este trabajo y a María Elsa Correal por sus valiosos comentarios.

†Investigadora Junior, Quantil S.A.S. Correo electrónico: paula.rodriguez@quantil.com.co

1. Introducción

El Análisis de Datos Funcionales (FDA) propone tener un acercamiento funcional para el análisis y tratamiento de datos de alta dimensión. Dado que la naturaleza de estos datos en dimensión finita suele tener un comportamiento continuo, se pretende recuperar la función que describe cada observación. Estas funciones que identifican a cada uno de los datos en alta dimensión son considerados *datos funcionales*. Este tipo de datos ha sido ampliamente estudiado en los últimos años, siendo Ramsay and Dalzell (1991), Ramsay and Silverman (2005) y Ramsay et al. (2009) alguna de la literatura más popular en el tema.

Uno de los aspectos más estudiados y que tienen una gran relevancia en el ánalisis de datos funcionales, y en general en el análisis de datos, es obtener las direcciones que explican la mayor parte de la varianza de las observaciones para así reducir dimensionalidad y comprender más a fondo el comportamiento de los datos; a esto se le conoce como *Componentes Principales*. En Ramsay and Dalzell (1991) se extiende la noción de Componentes Principales a *Componentes Principales Funcionales*, dónde se busca una descomposición ortogonal del dato funcional para así explicar su comportamiento y varianza a través de un número de funciones más reducido y que además estén dadas por menos parámetros que la representación inicial del dato funcional.

Si bien las *Componentes Principales Funcionales* (FPC) explican el comportamiento y la variación de los datos funcionales, en este proyecto se presenta un método que pretende identificar funciones de proyección según la capacidad predictiva y explicativa de un modelo de regresión multivariado derivado de los datos funcionales en cuestión. El método consiste en generar un conjunto de funciones aleatorias las cuáles se relacionarán con las observaciones funcionales a partir de sus productos internos. Es decir, dados los datos funcionales $\mathcal{X}_1, \ldots, \mathcal{X}_n$ con respuesta escalar $\mathbf{Y} = (Y_1, ..., Y_n)'$ y las funciones aleatorias $\rho_1, \ldots, \rho_m$ se consideran los datos multivariados

$$\mathbf{Z}_j = \Big(\langle \mathcal{X}_1, \rho_j \rangle, \ldots, \langle \mathcal{X}_n, \rho_j \rangle\Big)' \in \mathbb{R}^n$$

para $j = 1, \ldots, m$. Por lo tanto, cada $\mathbf{Z}_j$ contiene la proyección de los datos $\mathcal{X}_1, \ldots, \mathcal{X}_n$ en dirección de la función ρ_j. De esta manera se considera el modelo de regresión multivariado

$$\mathbf{Y} = b_o + \sum_{j=1}^{m} b_j \mathbf{Z}_j + \epsilon. \tag{1.1}$$

A partir de esto se consideran regresiones multivariadas con selección de subconjuntos de predictores para después, dependiendo de la capacidad explicativa y predictiva del modelo de regresión, determinar cuáles son las funciones ρ que mejor explican las observaciones $\mathcal{X}_1, \ldots, \mathcal{X}_n$.

El método presentado es probado e implementado en dos bases de datos. La primera consiste en las funciones de altura instantánea de olas para predecir la energía total de la ola y la segunda consiste en predecir el promedio de precipitación anual en Canadá a partir de funciones de temperatura anual.

2. Marco Teórico

2.1. Introducción a Datos Funcionales

El *Análisis de Datos Funcionales* (FDA) estudia datos que son funciones pertenecientes a un espacio vectorial $\mathcal{S} = \{\mathcal{X}(t) \mid \mathcal{X} : T \to \mathbb{R}\}$ donde T es un dominio de tiempo o espacio. Usualmente se considera $\mathcal{S} = \mathcal{L}_2(T)$ el conjuto de funciones cuadrado integrables, el cual es un espacio vectorial con las operaciones usuales de suma y multiplicación escalar para funciones. Además, al considerar el espacio de funciones $\mathcal{L}_2(T)$, se puede definir el producto interno de dos funciones $\mathcal{X}$ y $\mathcal{Y}$ como

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \int_T \mathcal{X}(t)\mathcal{Y}(t)dt. \tag{2.1}$$

Este producto interno es análogo al producto interno de vectores en $\mathbb{R}^m$ dónde $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^m x_i y_i$. La idea del FDA es entonces extender nociones y metodologías del análisis de datos multivariado a datos funcionales. Teniendo en cuenta la estructura de $\mathcal{L}_2$ como espacio con producto interno, es posible extender métodos tales cómo la regresión e identificación de componentes principales al caso de datos funcionales de manera casi natural. Igualmente, se extienden nociones de medidas estadísticas tales como la media, varianza y correlación entre variables aleatorias funcionales.

Dada una muestra de datos funcionales $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ se define puntualmente la media, la varianza, la covarianza bivariada y la correlación sobre la muestra Ramsay and Dalzell (1991), Kokoszka and Reimherr (2017). Por lo tanto, se tiene que la media muestral es

$$\bar{\mathcal{X}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i(t), \tag{2.2}$$

y la varianza es

$$v(t) = \frac{1}{n-1}\sum_{i=1}^{n} [\mathcal{X}_i(t) - \bar{\mathcal{X}}(t)]^2. \tag{2.3}$$

La varianza brinda información sobre la variación de las curvas en cierto punto t, sin embargo no brinda información sobre cómo los valores de las curvas en t están

relacionados con los valores en otro punto s. Para esto se define la función de covarianza bivariada

$$C(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}[\mathcal{X}_i(s) - \bar{\mathcal{X}}(s)][\mathcal{X}_i(t) - \bar{\mathcal{X}}(t)]. \tag{2.4}$$

Con esto se tiene entonces que la función de correlación está dada por

$$r(s,t) = C(s,t)/[v(s)v(t)]^{1/2}. \tag{2.5}$$

Algunos ejemplos de datos funcionales pueden ser la concentración de azúcar en la sangre de un paciente n en un tiempo t, o el precio de cierta acción del mercado en un tiempo t. Sin embargo, aunque existan distintos eventos que presenten datos de naturaleza continua, en la práctica estos datos deben ser discretizados, obteniendo así datos de alta dimensión; es decir, cada observación está dada por un vector real de dimensión considerablemente alta. Por ejemplo, se tiene el precio de una acción del mercado cada hora durante un año ya que no es posible almacenar la información continua del precio.

Por lo tanto, en vez de tener observaciones continuas $\mathcal{X}_1, ..., \mathcal{X}_n$, se tienen observaciones discretas $\mathcal{X}_1(t_i), ..., \mathcal{X}_n(t_i)$ para $t_i \in T$, $i = 1, ..., m$. El FDA propone, a partir de estos datos finito dimensionales, recuperar las funciones $\mathcal{X}_1, ..., \mathcal{X}_n$ que describen el comportamiento de los datos discretos. Estas funciones recuperadas son consideradas *datos funcionales*. La manera de recuperar estas funciones se expone en la Sección 2.2. Para el análisis de datos funcionales, el FDA extiende los modelos clásicos de análisis de datos que usualmente se llevan a cabo para datos reales o categóricos al caso funcional.

2.2. Datos de Alta Dimensión Como Funciones

El proceso por el cuál se identifica cada observación de alta dimensión con una función es llamado *suavizamiento* de la función. Así como se propone en Ramsay and Dalzell (1991), el objetivo de este proceso consiste en pasar de un X finito dimensional a una representación $\hat{\mathcal{X}}$ que de cierto modo sea similar a la función $\mathcal{X}$ que realmente genera la observación X. La manera de representar estas funciones se hace a través de *bases funcionales*.

Definición 2.1 (Base funcional)**.** *Un conjunto de funciones $\{\phi_1, \phi_2, ...\}$ es una base funcional en $\mathcal{L}_2(T)$ si toda función $\mathcal{X} \in \mathcal{L}_2(T)$ tiene una única descomposición*

$$\mathcal{X}(t) = \sum_{j=1}^{\infty} c_j \phi_j(t) \tag{2.6}$$

con $c_j \in \mathbb{R}$ $\forall j$.

Sin embrago, en la práctica no es posible considerar una cantidad arbitraria de funciones

base, por lo tanto, se debe restringir la representación a una cantidad de funciones K lo suficientemente grande para que los errores sean tan pequeños como se desee. De esta manera se tiene la representación

$$\mathcal{X}(t) = \sum_{k=1}^{K} c_k \phi_k(t) + \epsilon(t). \tag{2.7}$$

Sin embargo, para datos observados no contamos con el continuo de puntos de $\mathcal{X}$. Por lo tanto, dados los puntos observados $X_i = \mathcal{X}(t_i)$ para $i = 1, \ldots, m$, se tiene la representación

$$X_i = \mathcal{X}(t_i) = \sum_{k=1}^{K} c_k \phi_k(t_i) + \epsilon(t_i) \ ; \ t_i \in T. \tag{2.8}$$

Teniendo en cuenta la representación 2.8, es claro que el suavizamiento de cada observación se puede ver como m problemas de regresión lineal dónde $\mathcal{X}(t_i)$ corresponde a la variable independiente y $\phi_1(t_i), \ldots, \phi_K(t_i)$ corresponden a las variables explicativas Ramsay et al. (2009). Según esto, los coeficientes c_k se pueden estimar por mínimos cuadrados de manera que

$$\hat{\boldsymbol{c}} = \arg\min \left(\sum_{i=1}^{p} \left(\mathcal{X}(t_i) - \sum_{k=1}^{K} c_k \phi_k(t) \right)^2 \right), \tag{2.9}$$

obteniendo así dato funcional suavizado

$$\hat{\mathcal{X}}(t) = \sum_{k=1}^{K} \hat{c}_k \phi_k(t_i) = \hat{\boldsymbol{c}}^T \boldsymbol{\phi}. \tag{2.10}$$

Es importante notar que dependiendo de las características y el comportamiento que presenten los datos que se desean suavizar, se escoge tanto la base funcional como el parámetro K y luego se hace la selección de los coeficientes para la representación como en 2.8 y 2.9.

En los principales estudios de Análisis de Datos Funcionales Ramsay and Dalzell (1991), Ramsay and Silverman (2005), Ramsay et al. (2009), se consideran principalmente **bases de Fourier** para la suavización de datos periódicos y **bases B-Splines** para la suavización de datos no periódicos. Además de éstas, también es útil tener en cuenta otras bases funcionales como las bases *Wavelets*, *Exponenciales* y *Polinomiales*, entre otras.

Bases de Fourier

La base de Fourier es el conjunto de funciones $\{1,\ sin(k\omega t),\ cos(k\omega t) \mid k \in \mathbb{N}^\star\}$. Dado que ésta es una base de funciones periódicas donde ω determina el perioro $2\pi/\omega$ de cada función, es adecuado usarla para suavizar datos de naturaleza periódica. Además, el uso de esta base es computacionalmente beneficioso cuando los puntos $t_1, ...t_m$ son equidistantes y m es una potencia de 2 ya que la Transformada Rápida de Fourier permite que el cálculo de los coeficientes c_k en 2.8 se resuleva en $\mathcal{O}(m \log m)$ operaciones. Ramsay and Silverman (2005).

Base B-Splines

Definición 2.2 (Función Spline)**.** *Una función $\phi : T \to \mathbb{R}$ es una Función Spline si existe una partición $\{\tau_l\}_{l=1,\ldots,L}$ de T y funciones polinomiales $\{\beta_k\}_{k=1,\ldots,L}$ con cada $\beta_k : \tau_k \to \mathbb{R}$ de grado m tales que*

(i) $\phi \mid_{\tau_k} = \beta_k$ *y*
(ii) Para todo par de polinomios adyacentes β_j, β_i sus derivadas $\frac{\mathrm{d}^q}{\mathrm{d}t^q}\beta_j$ y $\frac{\mathrm{d}^q}{\mathrm{d}t^q}\beta_i$ coinciden en el punto $t \in T$ tal que $\phi(t) = \beta_j(t) = \beta_i(t)$ para $q = 1, \ldots, m-2$

Una *Base Spline* de grado m para la partición $\tau = \{\tau_l\}_{l=1,\ldots,L}$ de T es un conjunto de funciones Spline tales que cualquier función Spline definida por m y τ puede expresarse como combinación lineal de las funciones en la base.

Sin embargo, existen distintas maneras en que se puede construir una base Spline. La construcción más popular en FDA es el sistema desarrollado por de Boor (2001) con el cuál se crea la base B-Spline. Lo interesante de usar este tipo de base para suavizar datos no periódicos es que pueden ser evaluados de manera igualmente eficiente que los polinomios pero además logran determinar variaciones locales de manera más precisa, lo cuál propone un beneficio extra frente a usar una base polinomial.

2.3. Componentes Principales Para Datos Funcionales

Luego de los pasos preliminares de tratamiento de datos funcionales es común que se desee observar los atributos principales que caracterizan a las observaciones funcionales en cuestión. Al igual que en el caso de datos multivariados se puede llevar a cabo un *análisis de componentes principales* para reducir la dimensionalidad de los datos y así entender el comportamiento de éstos utlizando una menor cantidad de parámetros. En el caso multivariado las componentes principales corresponden a las direcciones ortogonales que explican mayor parte de la varianza de las observaciones. Naturalmente, en el caso funcional, las *componentes principales funcionales* corresponden a una sistema de funciones ortogonales que explican la mayor parte del comportamiento de las observaciones

funcionales. Por lo tanto, teniendo en cuenta la descomposición en bases funcionales, la idea del análisis de componentes principales es encontrar una descomposición ortogonal para la función de covarianza bivariada de los datos.

Según esto, y como se muestra en Ramsay and Dalzell (1991), el objetivo principal del análisis de componentes principales funcionales es determinar los componentes dominantes de la función de covarianza presentada en (2.4) por medio de una descompisición ortogonal

$$C(s,t) = \sum_{j=1}^{p} \lambda_j \xi_j(t) \xi_j(s). \tag{2.11}$$

Las funciones $\xi_1, ..., \xi_p$ corresponden a las p componentes principales funcionales y los valores propios λ_j indican la cantidad de varianza que se le atribuye a cada componente principal. Dado que se está considerando $\{\xi_1, ..., \xi_p\}$ un sistema ortogonal entonces se tiene que $\lambda_j \xi_j(s) = \langle C(s,t), \xi_j(t) \rangle$ para cada j. Por lo tanto, el problema de encontrar los componentes principales funcionales es equivalente a solucionar las p ecuación de funciones propias y valores propios

$$\langle C(s,), \xi_j \rangle = \lambda_j \xi_j(s) \tag{2.12}$$

sujeto a

$$\|\xi_j\| = 1 \quad \text{y} \quad \langle \xi_j, \xi_i \rangle = 0 \quad , \quad i = 1, ..., j-1. \tag{2.13}$$

Teniendo en cuenta la representación en bases funcionales de cada dato funcional $\mathcal{X}_i(t) = \boldsymbol{c}_i \boldsymbol{\phi}^T$, asi como en (2.10), se considera una representación en la misma base $\{\phi_k\}_{k=1}^K$ para las componentes principales. Por lo tanto, para el problema de encontrar las componentes principales, se desea ahora determinar los coeficientes $\boldsymbol{d}_j \in \mathbb{R}^K$ donde $\xi_j = \boldsymbol{d}_j \boldsymbol{\phi}^T$ para cada j. Por lo tanto, asumiendo que los datos están centrados y teniendo en cuenta la definición de $C(s,t)$ en (2.4), el problema de encontrar las componentes principales se convierte en solucionar:

$$
\begin{aligned}
& \int C(s,)\xi_j dt = \lambda_j \xi_j(s) && (2.14)\\
\Leftrightarrow \quad & \int \frac{1}{n-1}\sum_{i=1}^{n} \mathcal{X}_i(s)\mathcal{X}_i(t)\xi_j dt = \lambda_j \xi_j(s) && (2.15)\\
\Leftrightarrow \quad & \int \frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{c}_i^T \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(t)\boldsymbol{c}_i \boldsymbol{d}_j^T \boldsymbol{\phi}(t) dt = \lambda_j \boldsymbol{d}_j^T \boldsymbol{\phi}(s) && (2.16)\\
\Leftrightarrow \quad & \left(\frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{c}_i \boldsymbol{c}_i^T\right)\left(\int \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(t)dt\right)\boldsymbol{d}_j^T \boldsymbol{\phi}(s) = \lambda_j \boldsymbol{d}_j^T \boldsymbol{\phi}(s) && (2.17)\\
\Leftrightarrow \quad & \left(\frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{c}_i \boldsymbol{c}_i^T\right)\left(\int \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(t)dt\right)\boldsymbol{d}_j^T = \lambda_j \boldsymbol{d}_j^T && (2.18)\\
\Leftrightarrow \quad & \boldsymbol{C}\Phi \boldsymbol{d}_j^T = \lambda_j \boldsymbol{d}_j^T && (2.19)
\end{aligned}
$$

donde $\boldsymbol{C} = \frac{1}{n-1}\sum_{i=1}^{n} \boldsymbol{c}_i \boldsymbol{c}_i^T$ y $\Phi = \int \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(t)dt$. Finalmente el problema de solucionar una ecuación de funciones propias y valores propios para determinar componentes principales funcionales se tranforma en un problema de encontrar vectores y valores propios en un contexto vectorial sujeto a condiciones de ortongonalidad. Para más detalles de cómo resolver este sistema ver Ramsay and Silverman (2005).

Luego de haber obtenido los componentes principales funcionales ξ_j, es de gran utilidad considerar los puntajes de cada componente $f_{ij} = \langle \mathcal{X}_i - \bar{\mathcal{X}}, \xi_j \rangle$. Estos puntajes indican la cantidad relativa que hay de la componente j en la observación i. Además, son útiles tanto para entender el comportamiento de las observaciones como para llevar a cabo modelos de regresión multivariada que reflejen la naturaleza funcional de los datos.

2.4. Regresión Funcional Para Respuestas Escalares

En esta sección se considerarán extensiones del modelo de regresión al caso en que se tienen los datos funcionales $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ como variables independientes y las respuestas escalares correspondientes $Y_1, ..., Y_n \in \mathbb{R}$. Es decir, se considerará el modelo

$$
Y_i = \alpha + \int_T \mathcal{X}_i(t)\beta(t)dt + \epsilon \tag{2.20}
$$

donde $\alpha \in \mathbb{R}$ corresponde al intercepto, $\beta : T \to \mathbb{R}$ y $\epsilon(t)$ representa una función de errores aleatorios funcionales iid de media 0.

La manera en que se propone resolver este tipo de modelos es transformándolo en un modelo de regresión multivariado equivalente. Sin embargo, la asociación de un modelo de regresión funcional a un modelo de regresión multivariada no es único, y por lo tanto en la literatura se han propuesto distintos métodos para llevar a cabo regresiones funcionales. Para este caso consideraremos los métodos propuestos en Ramsay and Silverman (2005) dónde el primer método, y el más común, consiste en considerar la expansión en bases funcionales y llevar a cabo una penalización sobre $\beta(t)$, mientras que el segundo método consiste en utilizar los puntajes de las componentes principales funcionales f_{ij}.

2.4.1. Regresión por bases funcionales y penalización

El enfoque de regresión funcional por medio de bases funcionales consiste en considerar una base funcional lo suficientemente grande y rica para expresar la función estimada de regresión $\hat{\beta}(t)$. Es decir, si se considera la base funcional $\{\psi_k\}_{k=1}^{B}$, se desea ahora determinar los coeficientes $b_k \in \mathbb{R}$ tales que

$$\hat{\beta}(t) = \sum_{k=1}^{B} b_k \psi_k = \mathbf{b}^T \boldsymbol{\psi}. \tag{2.21}$$

De esta manera, y teniendo en cuenta la expresión por medio de bases funcionales de las observaciones, el modelo de regresión funcional (2.20) se convierte en

$$\begin{aligned} Y_i &= \alpha + \int \mathbf{c}_i^T \boldsymbol{\phi}(t) \boldsymbol{\psi}(t)^T \mathbf{b} \; dt + \epsilon & (2.22) \\ &= \alpha + \mathbf{c}_i^T \Big(\int \boldsymbol{\phi}(t) \boldsymbol{\psi}(t)^T \; dt \Big) \mathbf{b} + \epsilon & (2.23) \\ &= \alpha + \mathbf{c}_i^T \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\psi}} \mathbf{b} + \epsilon & (2.24) \end{aligned}$$

donde $\mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\psi}}$ es la matriz $K \times B$ de productos internos de las funciones en $\boldsymbol{\phi}$ y $\boldsymbol{\psi}$. Considerando $\mathbf{Z}_i = [1 \quad \mathbf{c}_i^T \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\psi}}]$ y $\zeta = (\boldsymbol{\alpha}, b_1, .., b_B)$ se tiene entonces el modelo de regresión

$$\widehat{\mathbf{Y}} = \mathbf{Z} \; \widehat{\boldsymbol{\zeta}}. \tag{2.25}$$

Éste es ahora un modelo de regresión multivariada estándar que permite estimar α y los coeficientes $\mathbf{b}$ de $\beta(t)$ por mínimos cuadrados obteniendo así la estimación

$$\widehat{\boldsymbol{\zeta}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y}. \tag{2.26}$$

Luego, con la estimación de los coeficientes, se puede construir por bases funcionales la función de regresión estimada $\hat{\beta}(t)$. Sin embargo, la escogencia del tamaño de la base funcional para $\hat{\beta}(t)$ hace que el modelo no sea lo suficientemente flexible pues al aumentar B la complejidad del modelo aumenta considerablemente y se puede incurrir en

sobreparametrización, mientras que al disminuirlo se puede estar perdiendo información importante que debería contemplar el comportamiento de la función de regresión. La estrategia para evitar esta situación, mencionada en Ramsay and Silverman (2005), es considerar una base lo suficientemente rica pero haciendo uso de una penalización sobre el comportamiento de la función de regresión para evitar fluctuaciones locales excesivas. Por lo tanto, considerando el modelo (2.20), se desea encontrar (α, β) tales que minimicen

$$\sum_{i=1}^{n} \left[Y_i - \alpha - \int \mathcal{X}_i(t)\beta(t)dt \right]^2 + \lambda \int \left[\frac{d^2}{dt^2}\beta(t) \right]^2 dt \tag{2.27}$$

donde $\lambda \in \mathbb{R}$ es un parámetro que debe ser calibrado, usualmente por validación cruzada. Para el modelo con penalización, la expresión para la estimación de los coeficientes de $\beta(t)$ sería

$$\widehat{\boldsymbol{\zeta}} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{R})^{-1}\mathbf{Z}^T\mathbf{Y}. \tag{2.28}$$

donde $\mathbf{R}$ es la matriz $\int \left[\frac{d^2}{dt^2}\beta(t) \right]^2 dt$ aumentada con una fila y columna de ceros.

2.4.2. Regresión por componentes principales

Otro de los métodos más comunes en la literatura para asociar un modelo de regresión multivariado a un modelo de regresión funcional es utilizar los puntajes de las componentes principales. Dadas las observaciones $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ en cuestión y su función de covarianza $v(s,t)$ considere las p primeras componentes principales funcionales $\xi_1(t), ..., \xi_p(t)$ así como se definieron en la sección de 2.3 Componentes Principales Para Datos Funcionales. Se considera entonces la descomposición

$$\mathcal{X}_i(t) = \bar{\mathcal{X}}(t) + \sum_{j=1}^{p} f_{ij}\xi_j \tag{2.29}$$

y con esto, el modelo de regresión multivariado asosiado al modelo de regresión funcional 2.20 es

$$Y_i = b_0 + \sum_{j=1}^{p} f_{ij}b_j + \epsilon_i. \tag{2.30}$$

Este modelo es un modelo de regresión múltiple multivariada estándar el cuál se resuelve por mínimo cuadrados. De esta manera se tiene que la estimación de $\mathbf{b} = (b_0, b_1, ..., b_p)$ está dada por

$$\widehat{\mathbf{b}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{Y} \tag{2.31}$$

donde $\mathbf{F} = (f_{ij})_{n\times p}$. Recuerde que las componentes principales funcionales forman un sistema ortogonal, por lo tanto en este caso $f_{ij} = \langle \xi_j, \mathcal{X}_i - \bar{\mathcal{X}} \rangle$ y con esto se tiene el

modelo

$$Y_i = b_0 + \int \sum_{j=1}^{p} b_j \xi_j (\mathcal{X}_i - \bar{\mathcal{X}}) dt + \epsilon_i. \tag{2.32}$$

Note que la expresión anterior tiene la misma forma que el modelo (2.20), lo cuál implica que $\beta(t) = \sum_{j=1}^{p} b_j \xi_j(t)$. Por lo tanto, luego de llevar a cabo la regresión por componentes principales, se puede recuperar la función de regresión funcional estimada de manera que

$$\widehat{\beta}(t) = \sum_{j=1}^{p} \widehat{b}_j \xi_j(t). \tag{2.33}$$

3. Método Propuesto

Como se mostró anteriormente, la mayoria de métodos desarrollados con datos funcionales son llevados a un problema de estadística multivariada equivalente. En particular, el método de regresión funcional por componentes principales presentado en la sección 2.4.2 consiste en resolver una regresión multivarida sobre los puntajes de los componentes principales. Teniendo en cuenta este método, es natural preguntarse si las funciones de componentes principales son precisamente las indicadas para proyectar los datos funcionales y obtener un buen ajuste en el modelo de regresión. La motivación para el desarrollo de este método es entonces encontrar las funciones ideales para proyectar los datos funcionales y obtener el mejor ajuste posible en el modelo de regresión multivariado dado por los productos internos (puntajes) con estas funciones.

Por lo tanto, 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

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = b_0 + \sum_{j=1}^{m} b_j \begin{bmatrix} \langle \mathcal{X}_1, \rho_j^\star \rangle \\ \vdots \\ \langle \mathcal{X}_n, \rho_j^\star \rangle \end{bmatrix} + \epsilon \tag{3.1}$$

tenga el mejor ajuste posible; i.e. que su $R^2 ajustado$ correspondiente sea lo más cercano posible a 1.

3.1. Funciones ideales de proyección y funciones aleatorias

Dados los datos funcionales $\mathcal{X}_i(t), ..., \mathcal{X}_n(t) \in \mathcal{L}_2(T)$ y las respuestas escalares $Y_1, ..., Y_n \in \mathbb{R}$, suponemos que existen funciones ideales $\rho_1^\star(t), ..., \rho_m^\star(t) \in \mathcal{L}_2(T)$ y puntajes correspondientes $f_{ij} = \langle \mathcal{X}_i, \rho_j^\star \rangle$ para $i = 1, ...n$ y $j = 1, ..., m$ tales que maximizan el ajuste del modelo de regresión 3.1.

Sin embargo, en la práctica no es posible considerar todas las posibles funciones para determinar cuáles son aquellas funciones ideales de proyección. Por lo tanto, teniendo en cuenta la expansión por medio de bases funcionales, restringimos el espacio de funciones considerado a funciones que tengan una descomposición en d funciones de una base determinada $\{\phi_k\}_k$ que sea lo suficientemente rica. Por lo tanto, se pasaría de determinar las funciones $\rho_1^\star(t), ..., \rho_m^\star(t)$ a determinar los coeficientes $c_{jl}^\star \in \mathbb{R}$ tales que

$$\rho_j^\star(t) = \sum_{l=1}^{d} c_{jl}^\star \phi_{k_l}. \tag{3.2}$$

El método que se propone para encontrar estas funciones ideales de proyección es **generar *funciones aleatorias* en el espacio de funciones considerado y a partir de éstas seleccionar adecuadamente aquellas funciones que representan un mejor ajuste del modelo de regresión**, i.e. aquellas funciones que más se acercan al comportamiento de $\rho_1^\star(t), ..., \rho_m^\star(t)$. Según lo anterior, el espacio de funciones considerado hasta el momento es el generado por $\{\phi_{k_1}, ..., \phi_{k_d}\} \subset \{\phi_k\}_k$; sea $S_{k_1,...,k_d}$ este conjunto de funciones. Note que generar funciones aletorias con la forma de (3.2), en $S_{k_1,...,k_d}$, es equivalente a generar los coeficientes aleatorios $\mathbf{c} = (c_1, ..., c_d) \in \mathbb{R}^d$ y luego considerar la combinación lineal con $\phi_{k_1}, ..., \phi_{k_d}$. Para el método propuesto, teniendo en cuenta que se van a considerar los pesos dados por productos internos, basta con generar coeficientes $\mathbf{c} \in \mathbb{R}^d$ tales que $\|\mathbf{c}\| = 1$, es decir $\mathbf{c} \in S^{d-1}$.

Teniendo las funciones $\phi_{k_1}, ..., \phi_{k_d}$ provenientes de la base funcional $\{\phi_k\}$, se quiere generar funciones aleatorias en $S_{k_1,...,k_d}$ para encontrar la función $f^* \in S_{k_1,...,k_d}$ que conlleva a una mejor proyección de los datos funcionales según su ajuste en el modelo de regresión. Note que esto se hace para un $d \in \mathbb{N}$ fijo e indices $k_1, ..., k_d$ fijos.

Dado que las funciones a generar aleatoriamente son de la forma

$$\rho(t) = \sum_{l=1}^{d} c_l \phi_{k_l}(t) = \mathbf{c}^T \Phi(t), \tag{3.3}$$

basta con generar coeficientes $\mathbf{c}_1, ..., \mathbf{c}_N \in S^{d-1}$ aleatorios para así generar las funciones aleatorias $\rho_1(t), ..., \rho_N(t)$.

3.1.1. Datos uniformes en la esfera unitaria

Recuerde que la esfera unitaria en $\mathbb{R}^d$ se define como $S^{d-1} = \{\mathbf{x} = (x_1, ..., x_d) \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$. Ahora, para generar puntos aleatorios en S^{d-1} se generan datos $\mathbf{c}_1, ..., \mathbf{c}_N \in \mathbb{R}^d$ i.i.d con distribución normal multivariada $N^{d+1}(0, I)$ y luego cada dato es normalizado de forma tal que esté en S^{d-1}. Es decir, consideramos los puntos $\mathbf{c}_1 \leftarrow \frac{\mathbf{c}_1}{\|\mathbf{c}_1\|}, ..., \mathbf{c}_N \leftarrow \frac{\mathbf{c}_N}{\|\mathbf{c}_N\|}$. En las Figuras 1 y 2 se pueden observar muestras uniformes generadas sobre S^1 y S^2 respectivamente.

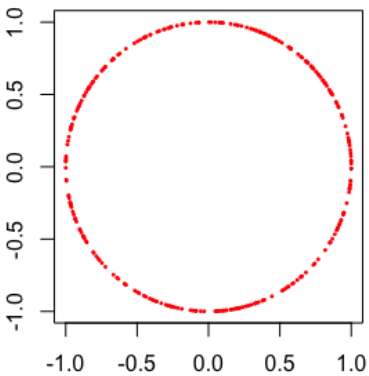

Figura 1: *Muestra uniforme en* S^1.

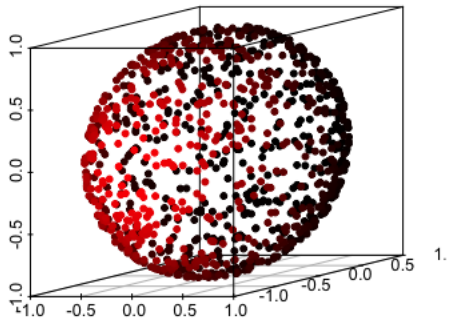

Figura 2: *Muestra uniforme en* S^2.

Según lo anterior, si f^* tiene coeficientes $\mathbf{c}^* = (c_1^*, ..., c_d^*) \in S^{d-1}$, es decir que $f^* = \sum_{l=1}^{d} c_l^* \phi_{k_l} = \mathbf{c}^* \Phi$, quisieramos generar N puntos uniformes en S^{d-1} de manera tal que nos acerquemos lo suficiente a $\mathbf{c}^*$.

3.1.2. Distancia entre coeficientes aleatorios y coeficientes óptimos

Consideraremos la distancia angular en S^{d-1}, teniendo asi que si $\Delta(S^{d-1}, \theta)$ es la hipercapa centrada en $\mathbf{c}^*$ con ángulo colatitudinal $\theta \in [0, 2\pi]$ entonces queremos que

$$P\Big(\mathbf{c}_1, ..., \mathbf{c}_N \notin \Delta(S^{d-1}, \theta)\Big) \approx 0 \tag{3.4}$$

Es decir, que la probabilidad de que todos los puntos generados aleatoriamente estén a distancia mayor que θ de $\mathbf{c}^*$ tienda a 0. Si $A(S^{d-1})$ es el área superficial de S^{d-1}, $A(S^{d-1}, \theta)$ es el área de la hipercapa $\Delta(S^{d-1}, \theta)$ y $\delta = A(S^{d-1}, \theta)/A(S^{d-1})$ 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,

$$P\Big(\mathbf{c}_1, ..., \mathbf{c}_N \notin \Delta(S^{d-1}, \theta)\Big) = (1-\delta)^N \tag{3.5}$$

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

Según esto, calculamos el valor de N que se debe considerar para cada dimensión. Tenemos que el área superficial de la esfera unitaria d-dimensional está dada por

$$A(S^{d-1}) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}. \tag{3.6}$$

Además, Li (2011) desarrolló fórmulas exactas para el área de hipercapas en esferas d-dimensionales obteniendo que

$$A(S^{d-1}, \theta) = \frac{1}{2}A(S^{d-1})I_{\sin^2\theta}\Big(\frac{d-1}{2}, \frac{1}{2}\Big) \tag{3.7}$$

donde $I_x(a,b)$ es la función beta incompleta regularizada. Por lo tanto, se tiene que la razón $\delta_{d-1,\theta} = A(S^{d-1}, \theta)/A(S^{d-1})$ está dada por

$$\delta_{d-1,\theta} = \frac{1}{2}I_{\sin^2\theta}\Big(\frac{d-1}{2}, \frac{1}{2}\Big). \tag{3.8}$$

Teniendo en cuenta esta expresión, en el Cuadro 1 se presenta el cálculo de la razón δ_{d-1} en función del ángulo θ para esferas unitarias d-dimensionales. Además, en la Figura 3 se observa el valor de $\delta_{d-1,\theta}$ para $\theta = \pi/2^k$ y $3 \leq k \leq 8$. Note que la razón del área de una capa esférica dada por un mismo ángulo θ disminuye cuando ésta se considera en dimensiones mayores. Además, recordando que la probabilidad de que todos los puntos de la muestra uniforme $\mathbf{c}_1, ..., \mathbf{c}_N$ con $N = \frac{1}{\delta}\log(\frac{1}{\delta^2})$ estén a distancia angular mayor que θ de $\mathbf{c}^*$ es δ^2, tenemos que para $\theta \leq \frac{\pi}{32}$ y $d \geq 2$

$$P\Big(\mathbf{c}_1, ..., \mathbf{c}_N \notin \Delta(S^{d-1}, \theta)\Big) < 1{,}6 \times 10^{-3}.$$

Tabla 1: *Razón entre área de la capa esférica con ángulo θ en S^{d-1} y el area total de S^{d-1}.*

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	$\frac{1}{4}(2 - 3\cos\theta + cos^3\theta)$

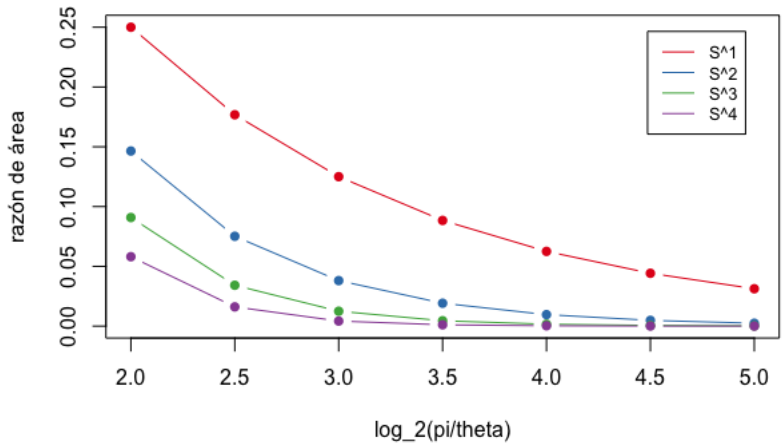

Figura 3: *Razón de área δ de la capa esférica con ángulo $\theta = \pi/2^k$ para esferas unitarias de dimensión 1, 2, 3 y 4.*

Según lo anterior, también podemos determinar el tamaño necesario de la muestra aleatoria para que al menos un punto de la muestra este a distancia menor que θ de $\mathbf{c}^*$ con probabilidad cercana a 1. En la Figura 4 se observa el valor de $N = \frac{1}{\delta}\log(\frac{1}{\delta^2})$ para distintos valores de θ y para los casos de esferas de dimensión $1, 2, 3$ y 4. Dado que la razón de área de una capa dada por un mismo ángulo θ disminuye considerablemente al aumentar la dimensión, se tiene que la muestra aleatoria también debe ser mayor para dimensiones más altas. Si bien los resultados en cuanto al tamaño de la muestra no son muy alentadores para dimensiones altas, en la práctica el comportamiento del modelo, generando muestras más pequeñas que las indicadas, presenta un buen comportamiento para centenares de funciones aleatorias. Esto se puede observar en los ejemplos estudiados del Capítulo 4.

3.1.3. Distancia entre funciones aleatorias y función de proyección ideal

En la sección anterior se encontró el número de coeficientes aleatorios N que se deben generar para que la distancia angular entre los coeficientes óptimos y alguno de los coeficientes aleatorios sea a lo más θ. Recuerde que se generan coeficientes aleatorios en S^{d-1} para representar una función como combinación lineal de un sistema ortonormal de funciones $\{\phi_1, ..., \phi_d\}$.

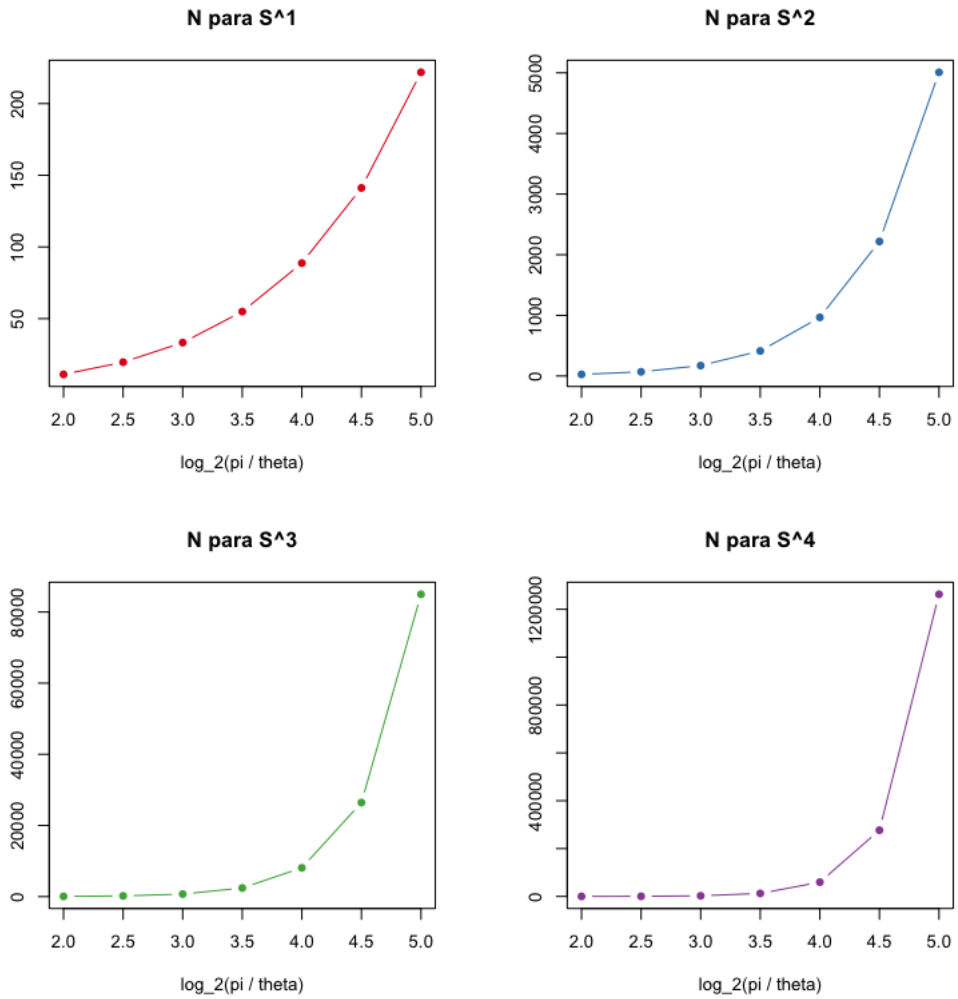

Figura 4: *Tamaño de la muestra aleatoria a generar* (N) *según el ángulo colatitudinal de la esfera.*

Suponga que la función ideal de proyección en $S_{\phi_1,\ldots,\phi_d}$ es

$$f^* = \sum_{k=1}^{d} a_k \phi_k = \mathbf{a}'\boldsymbol{\phi}.$$

Si se genera un conjunto de coeficientes aleatorios $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_N\} \subset S^{d-1}$ con N escogido como en la sección anterior según un ángulo θ, entonces se tiene que existe un $\mathbf{c} \in \mathcal{C}$ tal que la distancia angular entre $\mathbf{c}$ y $\mathbf{a}$ es menor o igual a θ. Considere entonces la función

$$g = \sum_{k=1}^{d} c_k \phi_k = \mathbf{c}'\boldsymbol{\phi}.$$

Teniendo en cuenta que la distancia angular entre $\mathbf{a}$ y $\mathbf{c}$ es menor o igual a θ, se quiere determinar la distancia en $\mathcal{L}_2$ entre f^* y g. Recuerde que ésta última está dada por

$$\|f^* - g\|_2^2 = \int \mid f^* - g \mid^2 dt \tag{3.9}$$

$$= \int \Big[\sum_{k=1}^{d} (a_k - c_k)\phi_k \Big]^2 dt. \tag{3.10}$$

Como $\{\phi_1, ..., \phi_d\}$ es un sistema ortonormal de funciones entonces $\int \phi_i \phi_j dt = 0$ para $i \neq j$ y $\int \phi_i^2 dt = 1$, teniendo así que

$$\|f^* - g\|_2^2 = \sum_{k=1}^{d} \Big[\int (a_k - c_k)^2 \phi_k^2 dt \Big] \tag{3.11}$$

$$= \sum_{k=1}^{d} (a_k - c_k)^2 \tag{3.12}$$

$$= \|\mathbf{a} - \mathbf{c}\|^2, \tag{3.13}$$

donde la última norma corresponde a la norma Euclidiana. Teniendo en cuenta que $\mathbf{a} \cdot \mathbf{c} = \|\mathbf{a}\| \, \|\mathbf{c}\| \cos(\mathbf{a}, \mathbf{c})$ y que $\mathbf{a}, \mathbf{c} \in S^{d-1}$, se tiene que

$$\begin{aligned} \|\mathbf{a} - \mathbf{c}\|^2 &= 2(1 - \mathbf{a} \cdot \mathbf{c}) \\ &= 2(1 - \cos(\mathbf{a}, \mathbf{c})). \end{aligned}$$

Además, como el ángulo entre $\mathbf{a}$ y $\mathbf{c}$ es menor o igual que $\theta \leq \pi/2$, se llega a que

$$\|f^* - g\|_2^2 \leq 2(1 - \cos(\theta)). \tag{3.14}$$

Lo anterior indica que al tomar una menor distancia angular $\theta \leq \pi/2$ para los coeficientes, se llega a una cota cada vez menor para la distancia en $\mathcal{L}_2$ de las funciones dadas como combinación lineal de una misma base con los coeficiente dados. En la Figura 5 se observa la distancia entre dos funciones que son combinación lineal de una misma base al variar el ángulo de sus coeficientes.

3.2. Selección de mejores funciones de proyección

En la sección anterior se expuso la manera de generar N funciones aleatorias de manera tal que al menos una de ellas esté lo suficientemente cerca de la función ideal de proyección. Sin embargo, esto se hizo fijando una base funcional inicial $\{\phi_k\}_k$, un número de funciones $d \in \mathbb{N}$ y unos índices $k_1, ..., k_d$. De esta manera se generaban las funciones aleatorias $\rho_1(t), ..., \rho_N(t)$ en el espacio generado por $\{\phi_{k_1}, ..., \phi_{k_d}\}$. Esto mismo se puede

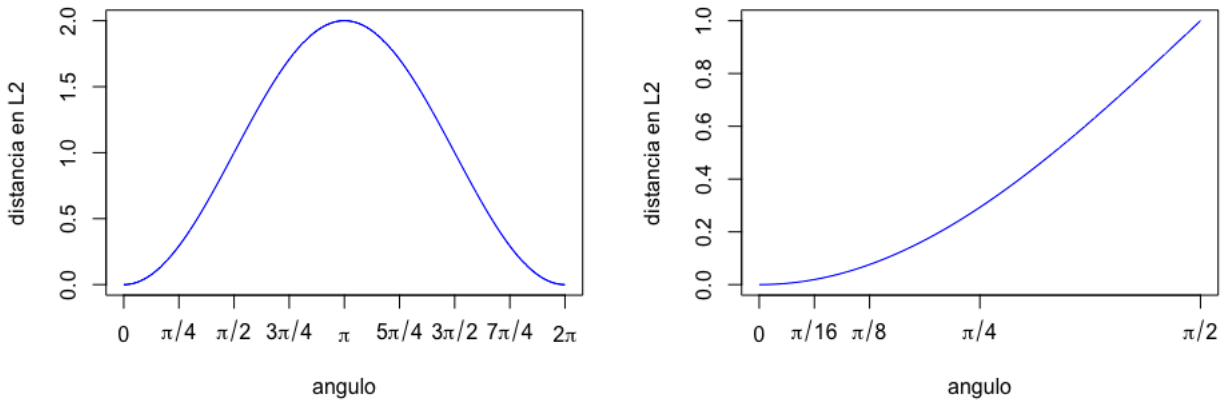

Figura 5: *Distancia en $\mathcal{L}_2$ entre dos funciones según distancia angular de sus coeficientes*

hacer para varios $d \in \mathbb{N}$ y distintas combinaciones de indices $I = \{k_1, ..., k_d\}$ dónde, para cada pareja (d, I), se generan N_d funciones aleatorias según el ángulo colatitudinal escogido. Considere entonces $\mathcal{P} = \{\rho_1(t), ..., \rho_N(t)\}$ el conjunto de funciones aleatorias generadas variando el par (d, I). Teniendo en cuenta el desarrollo de la sección anterior se tendría que para cada (d, I) considerados en la construcción de $\mathcal{P}$, existe una función $\rho_{(d,I)} \in \mathcal{P}$ que se acerca lo suficientemente a la función ideal de proyección del espacio generado por $\{\phi_j : j \in I\}$. Ahora, lo que se quiere es determinar cuáles y cuántas son aquellas funciones en $\mathcal{P}$ que conllevan a proyecciones con mejor ajuste en el modelo de regresión.

Para encontrar las funciones de proyección de los datos $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ que generen el mejor ajuste posible en el modelo de regresión que explica las variables escalares $Y_1, ..., Y_n$, se tendrán en cuenta aspectos de significancia y ajuste según el indicador $R^2 ajustado$. Teniendo las funciones aleatorias $\mathcal{P}$ se lleva a cabo el Algoritmo 1.

Selección de predictores

Dado que con el método se pretende identifiar las mejores funciones de proyección brindándole cierto puntaje a cada una, es necesario llevar a cabo el punto 2. (b) del Algoritmo 1 donde se seleccionan los mejores predictores. Note que para cada subconjunto P_i esto se hace para evitar que una función que no conlleva a una proyección conveniente obtenga un buen puntaje por estar en un modelo con una función que si presenta las características deseadas. La manera clásica de identificar la importancia de predictores por medio de regresión es fijandose en su significancia según el p-valor obtenido. Sin embargo, para seleccionar predictores que se creen estar más relacionados con la respuesta, es popular usar *selección paso a paso hacia atrás/delante* James et al. (2014). El procedimiento a seguir para la selección paso a paso hacia atrás se presenta en

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 $k = 1, ..., K : Z_{i,k} \leftarrow \big(\langle \mathcal{X}_1, \rho_{i_k}\rangle, ..., \langle \mathcal{X}_n, \rho_{i_k}\rangle\big)$
   (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
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.

el Algoritmo 2 y el procedimiento que se propone para seleccionar el mejor subconjunto de predictores según su significancia se presenta en el Algoritmo 3.

Algoritmo 2: 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 $R^2 ajustado$.

Algoritmo 3: 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 $\mathcal{M}_p$

Medida de desempeño según R^2 ajustado Recuerde que el objetivo es encontrar

las funciones de proyección de $\mathcal{X}_1(t), ..., \mathcal{X}_n(t)$ que mejor expliquen las variables de respuesta $Y_1, ..., Y_n$. En regresión multivariada el indicador de desempeño más común para identifcar ajuste de los datos es el coeficiente de detrminación R^2. Recuerde que éste está dado por

$$R^2 = 1 - \frac{RSS}{TSS} \tag{3.15}$$

donde $RSS = \sum(Y_i - \widehat{Y}_i)^2$ y $TSS = \sum(Y_i - \bar{Y})^2$. Sin embargo, como el RSS siempre disminuye cuando más predictores son añadidos al modelo, entonces el R^2 siempre aumenta cuando más predictores son añadidos. Por lo tanto, el R^2 no puede usarse para comparar modelos con diferente cantidad de predictores. Dado que en el punto 4 del Algoritmo 1 se desea comparar modelos que posiblemente tienen un número distinto de predictores, el indicador de desempeño que se utiliza es $R^2 ajustado$, el cual está dado por

$$R^2 ajustado = \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{3.16}$$

donde d es la cantidad de predictores. Como se menciona en James et al. (2014), la intuición detrás del $R^2 ajustado$ es que una vez todos los predictores correctos han sido incluidos en el modelo, añadir predictores adicionales disminuirá el valor del $R^2 ajustado$. Otras medidas de desempeño que podrían ser utilizada son C_p, *Criterio de Información de Akaike (AIC)* y *Criterio de información de Bayes (BIC)*.

3.3. Reconstrucción de la función de regresión estimada

Siguiendo el procedimiento del método de regresión funcional por componentes principales, se desea pasar del modelo de regresión multivariado al modelo de regresión funcional equivalente y reconstruir la función de regresión estimada $\beta(t)$. Ya habiendo encontrado las funciones ideales de proyección $\{\rho_1(t), ..., \rho_m(t)\}$ por medio del método propuesto, se soluciona por mínimos cuadrados el modelo de regresión

$$\begin{aligned} Y_i &= b_0 + \sum_{j=1}^{m} b_j \langle \mathcal{X}_i, \rho_j \rangle \\ &= b_0 + \sum_{j=1}^{m} b_j \int \mathcal{X}_i(t)\rho_j(t)dt \\ &= b_0 + \int \mathcal{X}_i(t) \sum_{j=1}^{m} b_j \rho_j(t)dt. \end{aligned}$$

Note que la última igualdad es equivalente a la expresión de modelo de regresión funcional $Y_i = \alpha + \int \mathcal{X}_i(t)\beta(t)dt$. Por lo tanto, para el modelo de regresión funcional tenemos que

la función de regresión estimada está dada por

$$\widehat{\beta}(t) = \sum_{j=1}^{m} b_j \rho_j(t). \tag{3.17}$$

4. Implementación del Método

En este capítulo se implementará el método mencionado en el capítulo anterior para determinar funciones ideales de proyección según regresión sobre dos bases de datos funcionales. La primera consiste en funciones de altura instantánea de olas para predecir la energía total de la ola. La segunda base de datos a trabajar contiene las curvas de temperatura anual de distintas regiones de Canadá para luego predecir el logaritmo de la precipitación promedio anual de cada región.

4.1. Ejemplo 1: Altura instantánea de olas para predecir su energía

Para este caso se cuenta con una base de datos que contiene la altura durante 101 momentos de 865 olas del mar. Dada la forma de cada ola y su evidente similitud en forma con funciones trigonométricas, los datos de altura instantánea se suavizaron utilizando una base funcional de Fourier de tamaño 21. En la Figura 6 se pueden observar 50 de los 865 datos de posición de las olas después de haber sido suavizados.

Se implementará el método propuesto de tres maneras distintas sobre esta base de datos. Además para cada caso se utilizara ambos procedimiento expuetsos para la selección de predictores: por significancia y por selección paso a paso hacía atrás con $R^2 ajustado$. Finalmente se llevará a cabo una regresió funcional por componentes principales para determinar la eficacia del método propuesto.

4.1.1. Método propuesto

Dado que el modelo propuesto depende de la base funcional escogida inicialmente, de la dimensión escogida para generar las funciones aleatorias y del método de selección de predictores, se lleva a cabo el procedimiento para 3 casos diferentes. Cada caso utiliza una base funcional diferente pero en los tres se consideran combinaciones lineales de dos funciones para generar las funciones aleatorias; es decir, se generan coeficientes aleatorios en S^1. Dependiendo del tamaño de la base funcional inicial se decide considerar distintas combinaciones lineales de pares de funciones diferentes. Para cada una de estas combinaciones se generan N funciones aleatorias generando N coeficientes aleatorios.

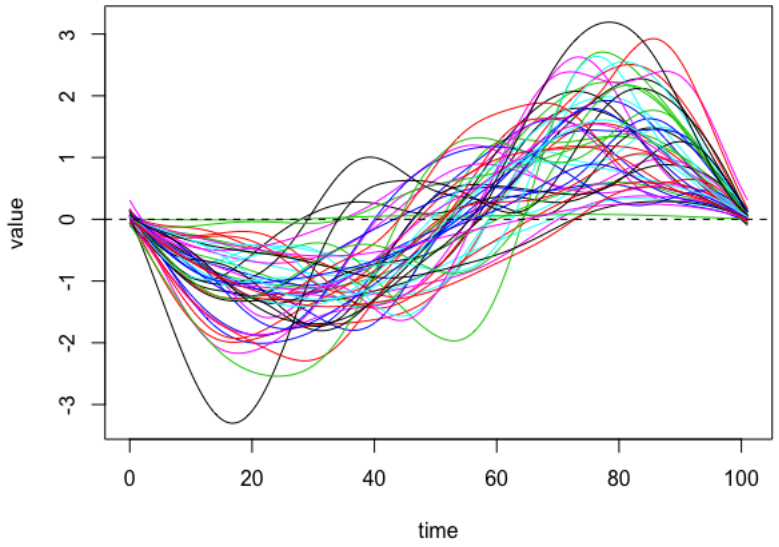

Figura 6: *Datos funcionales de posición de olas suavizados.*

Además, para cada caso se llevan a cabo ambos procedimientos de selección de predictores expuestos en el capítulo anterior. Las características de cada caso considerado se presentan resumidas en la Tabla 2.

Tabla 2: *Parámetros del modelo para cada caso considerado*

	Base Inicial	**# Pares de Funciones**	**Funciones por Indices (N)**
Caso 1	Fourier 11	100	300
Caso 2	Indicadora 10 [1]	45	500
Caso 3	Indicadora 25	6	1.000

Resultados Caso 1. Para este caso se generaron funciones aleatorias que fuesen combinación lineal de dos funciones de la base $\{1, \sin(x), \cos(x), ..., \sin(5x), \cos(5x)\}$. Específicamente se generaron 300 coeficientes aleatorios en S^1 para cada una de las 100 combinaciones de funcionaes consideradas. Con esto, se obtuvieron 30.000 funciones aleatorias. A partir de estas funciones se crearon 2.000 subconjuntos de 20 funciones cada uno y se llevó a cabo el método haciendo selección de variables en cada subconjunto por significancia y por selección paso a paso hacia atrás con $R^2 ajustado$.

[1] La base Indicadora M se refiere a funciones indicadoras en los intervalor $[1, M], [M+1, 2M+1], ...$

Como se puede observar en la Figura 7, la mayoría de submodelos, que inicialmente tenían 20 predictores, quedaron con 4 predictores significativos después de hacer la selección de mejores predictores según significancia en regresión. Esto indica que, basándose en significancia en regresión, los datos son de dimensión 4 aproximadamente. Luego de escoger las mejores 20 funciones según su desempeño en cada submodelo, se llevó a cabo una selección de predictores por significancia y se redujo a 5 funciones, reafirmando nuevamente que la dimensionalidad del problema está entre 4 y 5. En la Figura 8 se observan las mejores funciones de proyección obtenidas. La regresión con proyecciones llevada a cabo con estas 5 funciones tiene un $R^2 ajustado = 0{,}8912$.

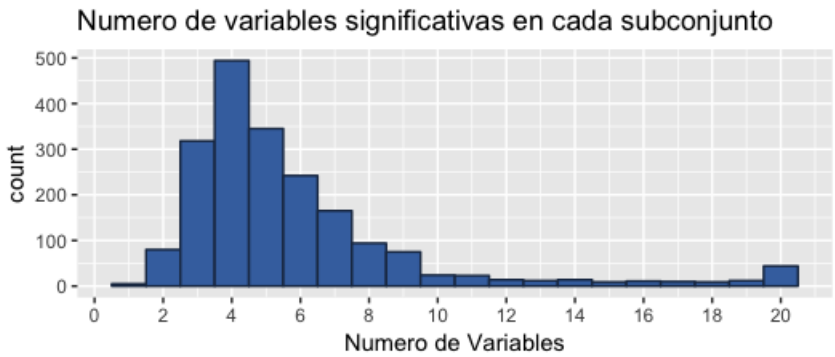

Figura 7: *Histograma de cantidad de predictores en cada submodelo después de selección por significancia (Caso 1)*

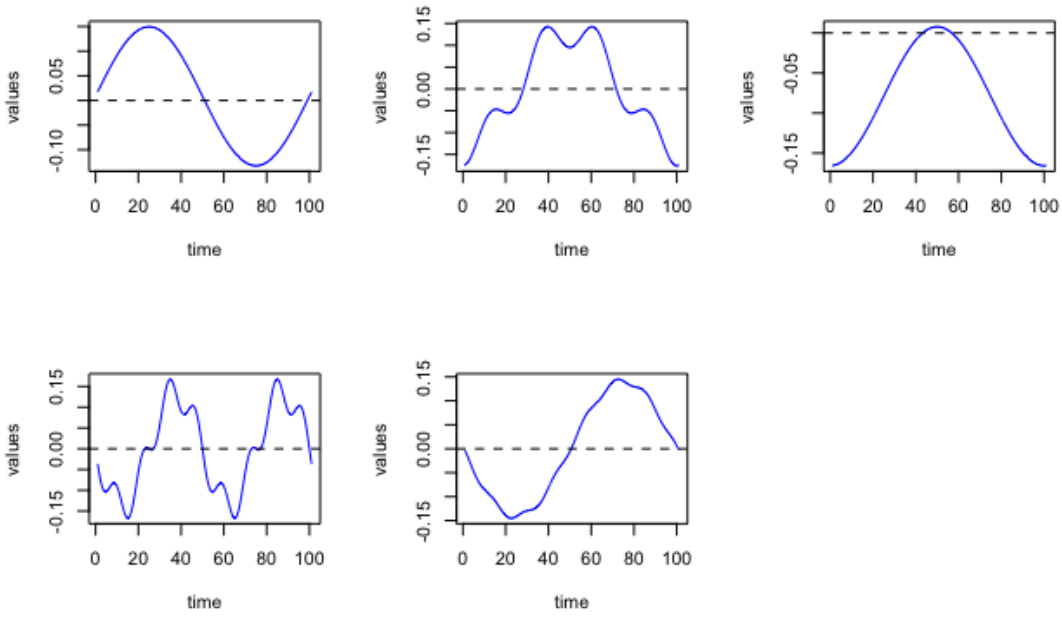

Figura 8: *Mejores funciones de proyección para selección por significancia (Caso 1)*

Para la escogencia de predictores utilizando selección paso a paso hacia atrás y escogiendo por $R^2 ajustado$ se obtuvo que la mayoría de submodelos quedaron con 7 predictores, esto se puede observar en la Figura 9. Además, al obtener las mejores funciones según este modelo nuevamente se llegan a 7 funciones. Esto indica que según aspectos de $R^2 ajustado$ el problema tiene dimensión 7. Para este caso es evidente que el método varía significativamente en cuánto a las funciones otenidas cuando se utilizan distintos procedimientos para la selecicón de predictores. Sin embargo, el $R^2 ajustado$ obtenido con las mejores funciones de proyección para este caso es de 0,8919, el cuál es similar al obtenido por significancia. Por lo tanto, aunque ambos métodos determinen funciones distintas como las mejores funciones de proyección, se podría decir que ambos procedimientos son consistentes pues obtienen funciones que conllevan a modelos de regresión con $R^2 ajustado$ similar. Dado que las funciones de proyección obtenidas por el método propuesto utilizando significancia son menos y además son más suaves que las obtenidas por selección paso a paso hacia atrás, se prefieren las primeras sobre las segundas.

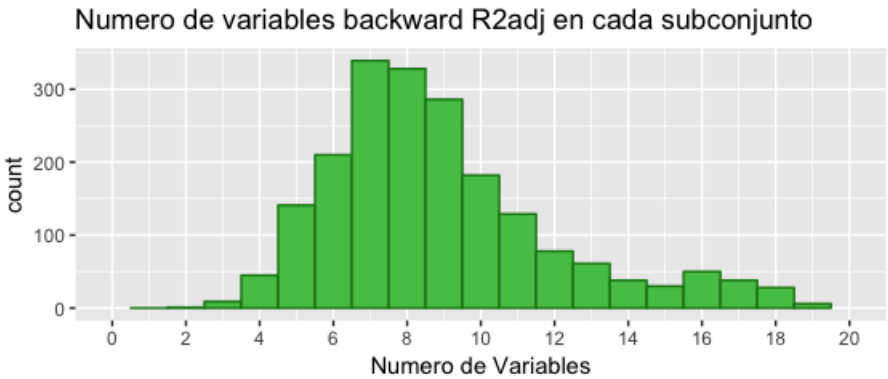

Figura 9: *Histograma de cantidad de predictores en cada submodelo después de selección por backward $R^2 ajustado$ (Caso 1)*

Resultados Caso 2. Para este caso se generaron funciones aleatorias que fuesen combinación lineal de dos funciones indicadoras de intervalos de la forma $[10t+1, 10(t+1)+1]$ para $t = 0, ..., 9$. Para cada par de funciones se generaron 500 coeficientes en S^1 de manera tal que se obtuvieron 22.500 funciones aleatorias. A partir de estas funciones se crearon 2.000 subconjuntos de 20 funciones cada uno y se llevó a cabo el método con ambos procedimientos expuestos de selección de predictores.

En la Figura 11 se observa que la mayoría de submodelos quedaron con 7 predictores significativos después de hacer la selección de mejores predictores según significancia en regresión. En el caso anterior se obtuvo una dimensión de 4 aproximadamente puesto que se tenía una base más rica que logra extraer más información de las observaciones a través de los productos internos. En este caso se tienen funciones indicadoras y por lo tanto tiene sentido que se necesiten más funciones para tener un buen desempeño. En la Figura 12 se observan las mejores funciones de proyección obtenidas; en este caso fueron

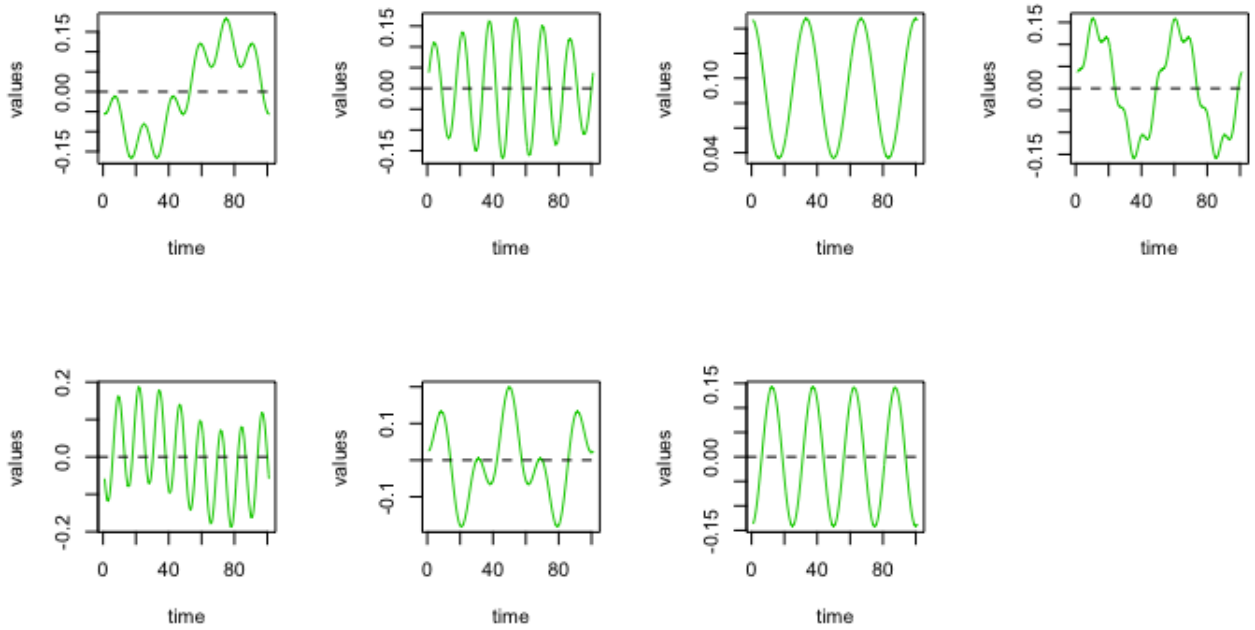

Figura 10: *Mejores funciones de proyección para selección por backward $R^2 ajustado$ (Caso 1)*

5 funciones. La regresión con proyecciones llevada a cabo con estas 5 funciones tiene un $R^2 ajustado = 0{,}8896$. Por lo tanto, aunque la mayoría de submodelos quedaran con 7 predictores significativos, el hecho de que las mejores funciones sean 5 indica nuevamente que la dimensión del problema es 5 aproximadamente según el criterio de significancia.

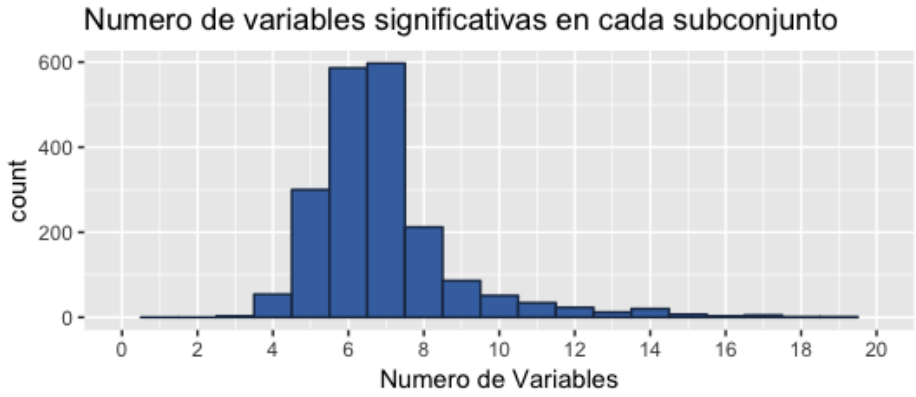

Figura 11: *Histograma de cantidad de predictores en cada submodelo después de selección por significancia (Caso 2)*

Al hacer la selección de mejores predictores utilizando el procedimiento de selección paso a paso hacia atrás se obtuvo que la mayoría de submodelos se explican bien con

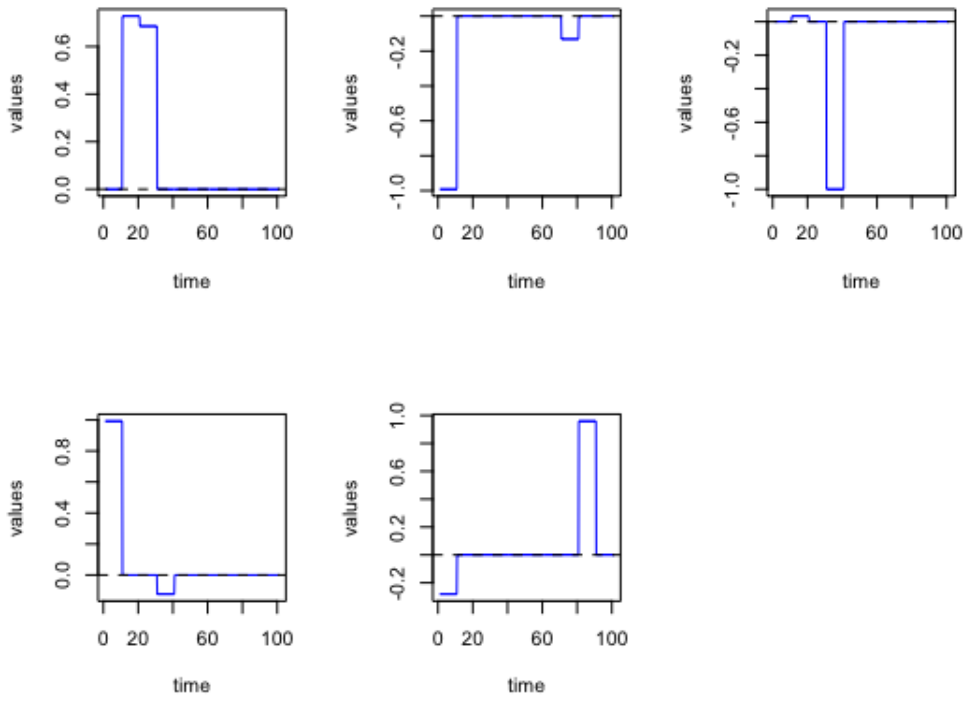

Figura 12: *Mejores funciones de proyección para selección por significancia (Caso 2)*

7 predictores. El resultado fue similar a la selección por significancia, lo cual indica nuevamente que el uso de funciones más simples implica utilizar un mayor número de éstas para tener un buen ajuste en el modelo de regresión. Se encontró que las mejores funciones de proyección son 6 funciones; éstas se presentan en la Figura 14. El $R^2 ajustado$ correspondiente a estas 6 funciones fue de 0,8876. Note además que en ninguno de los dos casos considerados se obtuvieron funciones que fuesen combinación lineal de la función indicadora en el intervalo $[50, 60]$. Esto tiene sentido pues es precisamente en este intervalo donde la altura de las olas es 0.

Resultados Caso 3. Para este caso se generaron funciones aleatorias que fuesen combinación lineal de dos funciones indicadoras de intervalos de la forma $[25t+1, 25(t+1)+1]$ para $t = 0, ..., 3$. Para cada par de funciones se generaron 1,000 coeficientes en S^1 de manera tal que se obtuvieron 6.000 funciones aleatorias. A partir de estas funciones se crearon 1.000 subconjuntos de 20 funciones cada uno y se llevó a cabo el método con ambos procedimientos expuestos de selección de predictores.

Como se puede observar en la Figura 15 la mayoría de submodelos quedaron con los mismos 4 predictores al hacer la selección por significancia. Sin embargo, en este caso la reducción de predictores se debe principalmente a la linealidad entre ellos pues

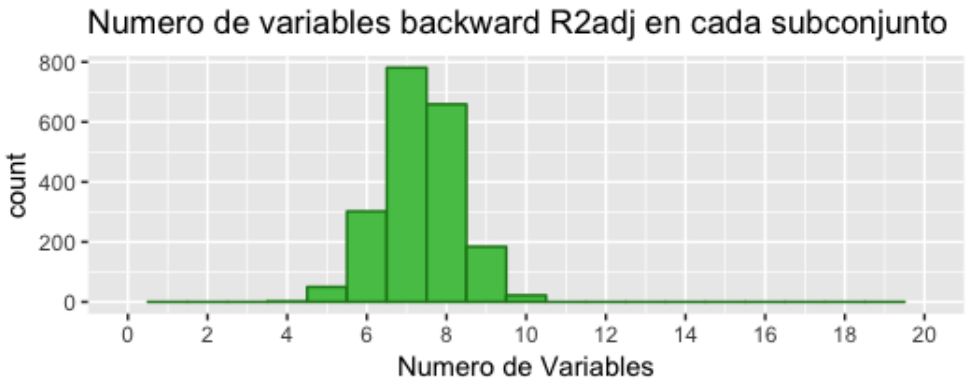

Figura 13: *Histograma de cantidad de predictores en cada submodelo después de selección por backward $R^2 ajustado$ (Caso 2)*

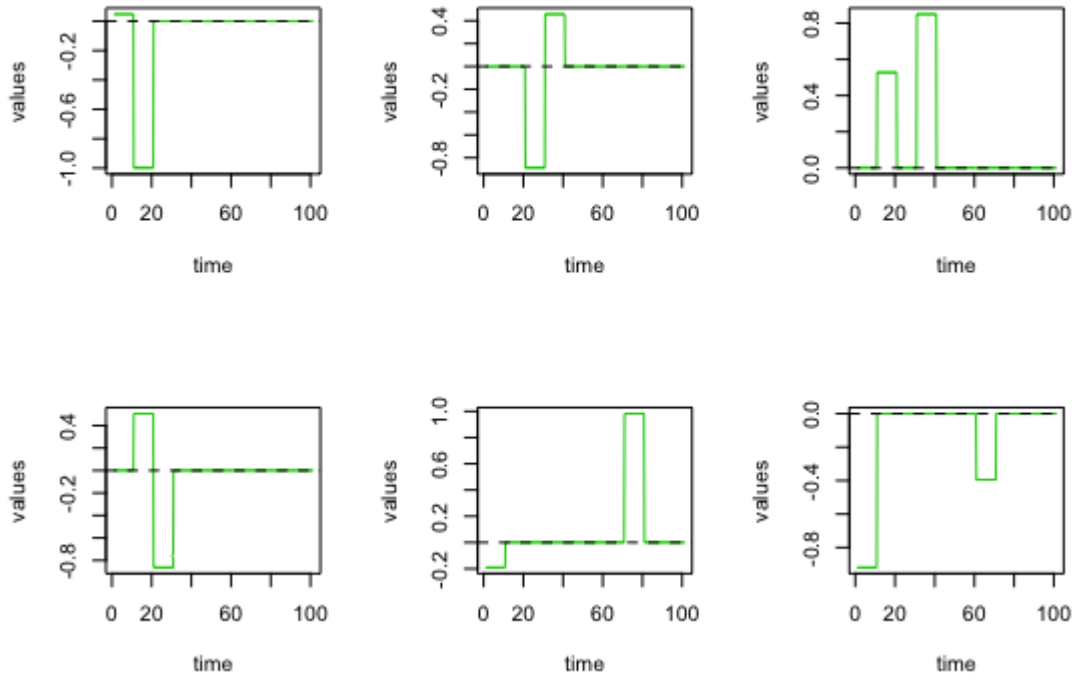

Figura 14: *Mejores funciones de proyección para selección por backward $R^2 ajustado$ (Caso 2)*

la base considerada inicialmente tiene solamente 4 funciones. Para este caso las mejores funciones de proyección tienen un $R^2 ajustado$ de 0,8864; estas funciones se pueden ver en la Figura 16. Note que con las cuatro funciones de proyección se logra cubrir todo el intervalo $[1, 101]$, por lo tanto entre las cuatro se está intentando capturar toda la información posible de las observaciones.

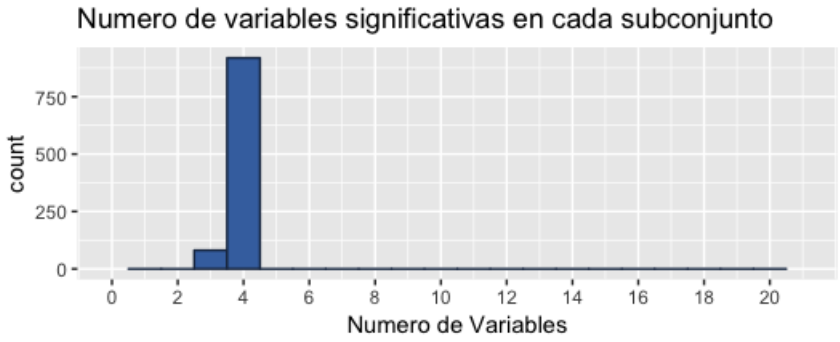

Figura 15: *Histograma de cantidad de predictores en cada submodelo después de selección por significancia (Caso 3)*

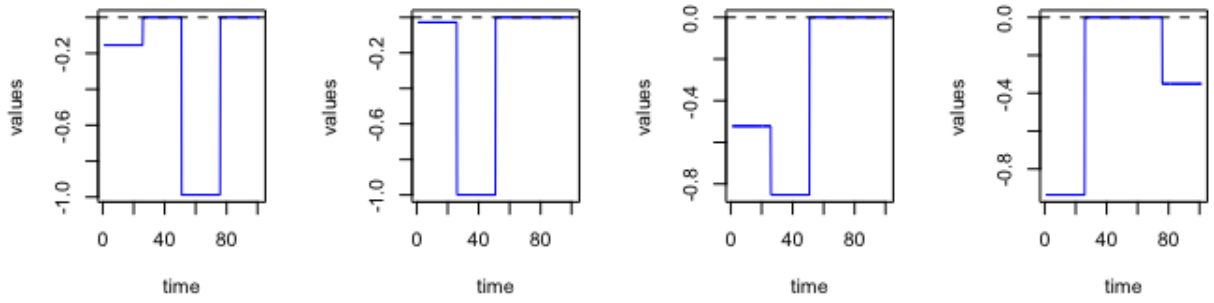

Figura 16: *Mejores funciones de proyección para selección por significancia (Caso 3)*

Al hacer la selección de mejores predictores utilizando el procedimiento de selección paso a paso hacia atrás se obtuvo que la mayoría de submodelos se explican bien con 4 predictores. Nuevamente se obtiene este resultado ya que el conjunto de funciones iniciales es de tan solo 4 funciones. Se encontró que las mejores funciones de proyección son 4 funciones; éstas se presentan en la Figura 18. El $R^2 ajustado$ correspondiente fue de 0,8864. Además, al igual que en el caso por significancia, se tiene que las 4 funciones encontradas buscan cubrir todo el intervalo $[1, 101]$.

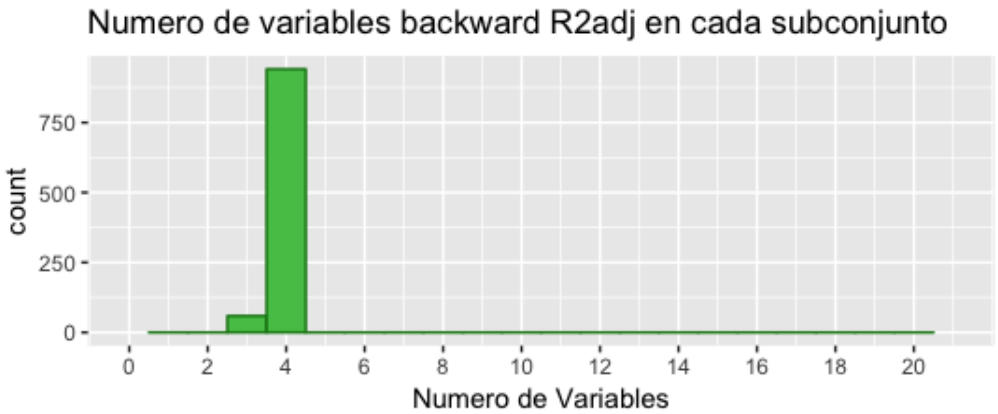

Figura 17: *Histograma de cantidad de predictores en cada submodelo después de selección por significancia (Caso 3)*

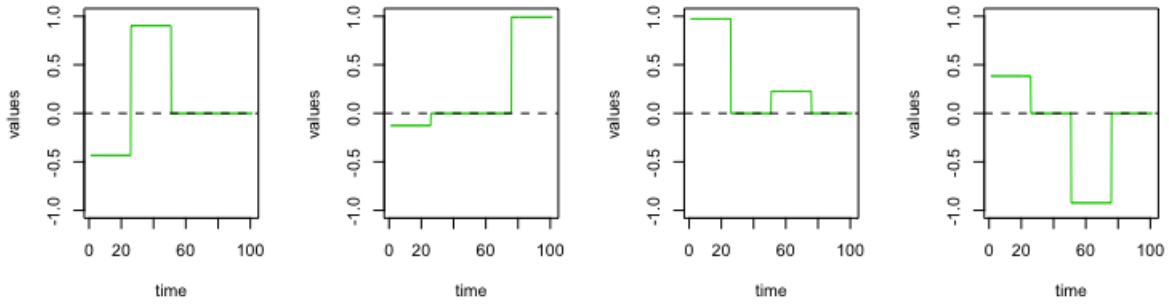

Figura 18: *Mejores funciones de proyección para selección por significancia (Caso 3)*

4.1.2. Comparación con regresión funcional por componentes principales

En el caso de regresión funcional por componentes principales se lleva a cabo el mismo modelo de regresión que se lleva a cabo en el método propuesta pero utilizando la funciones de componentes principales. Por lo tanto, se realiza una regresión por componentes para comparar las funciones de proyección encontradas y los componentes principales funcionales.

En la Tabla 3 se muestra el porcentaje de variación que explica cada componente principal. Se tiene que las primeras cuatro componentes principales explican el $94,6\,\%$ de la varianza y con las primeras siete componentes se explica el $97,1\,\%$ de la varianza. Dado que en los casos implimentados del método propuesto se encontraron entre 4 y 7 funciones ideales de proyección, encontramos que el método identifica una cantidad de funciones que tiene sentido en cuánto a la dimensión de los datos originales.

Tabla 3: *Varianza explicada por los componentes principales funcionales - Olas*

Componente	Prop. de Var.	Prop. Acumulada
1	0,605	0,605
2	0,174	0,780
3	0,113	0,893
4	0,053	0,946
5	0,025	0,971
6	0,012	0,984
7	0,006	0,990

En la Figura 19 se presentan las primeras 7 componentes principales de los datos de altura instantánea de olas. Note que la primera componente principal PC1 explica el comportamiento inverso de la ola, la segunda componente PC2 explica el comportamiento en el segundo tercio del intervalo y la tercera PC3 explica la variación en cada mitad del intervalo. Se llevó a cabo una regresión funcional por componentes principales teniendo en cuenta desde la primera hasta la séptima componentes principal. Los $R^2 ajustados$ obtenidos en cada caso se presentan en la Tabla 4.

Tabla 4: $R^2 ajustado$ *para regresión funcional por componentes principales - Olas*

# 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

Si bien el desempeño de la regresión funcional utilizando tan solo la primera componente principal es considerablemente bueno, cabe notar que las funciones indicadoras encontradas en el Caso 2 y Caso 3 del método propuesto tienen un desempeño similar a éste siendo funciones mucho más sencillas. Además, aunque el $R^2 ajustado$ es muy similar en todos los casos estudiados, encontramos que a partir de funciones aleatorias se pueden

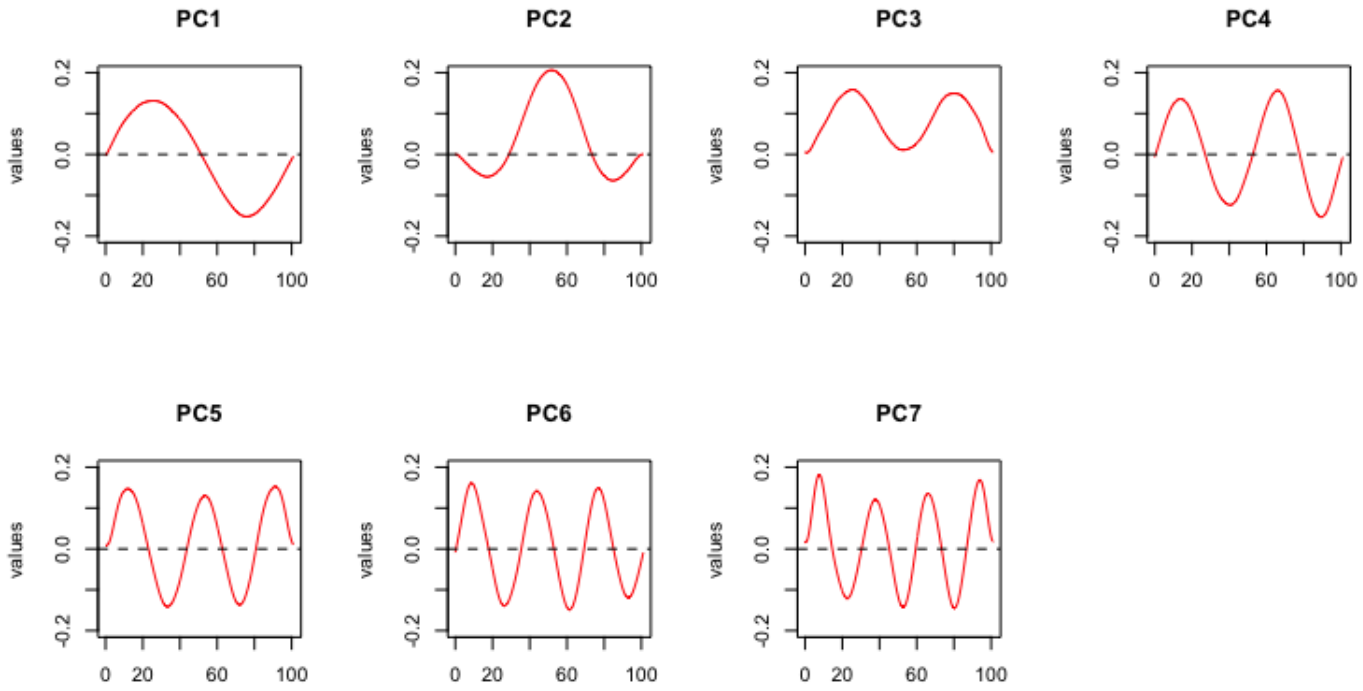

Figura 19: *Primeras siete componentes principales funcionales - Olas*

Tabla 5: *Número de mejores funciones de proyección y $R^2 ajustado$ correspondiente para cada caso considerado*

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

identificar funciones que expliquen el modelo de la misma manera que lo hacen los componentes principales funcionales. En la Tabla 5 se presentan los resultados obtenidos en cuánto a las funciones de proyección encontradas y su $R^2 ajustado$ correspondiente para cada uno de los casos considerados.

4.2. Ejemplo 2: Temperatura anual en Canadá para predecir precipitación promedio

La base de datos trabajada para este ejemplo consiste en la temperatura diaria por año, promediada entre los años 1960 y 1994, de 35 estaciones de clima en Canadá. Es decir, inicialmente se cuanta con 35 vectores de dimensión 365. La idea es utilizar estos datos para predecir el logaritmo de la precipitación anual en cada una de las 35 estaciones. Esta base de datos está disponible en el paquete **fda** de **R** Ramsay et al. (2009).

La suavización de los datos se hizo utilizando una base de Fourier de 5 funciones. En los experimentos realizados también se probó para una base de Fourier de 11, 21 y 35 y los resultados en cuanto al $R^2 ajustado$ fueron similares en todos los casos. Por lo tanto, se presentan los resultados para la base de tamaño 5 pues permite tener una mejor interpretación de los componentes principales. En la figura 20 se presentan los 35 datos funcionales luego de haber sido suavizados.

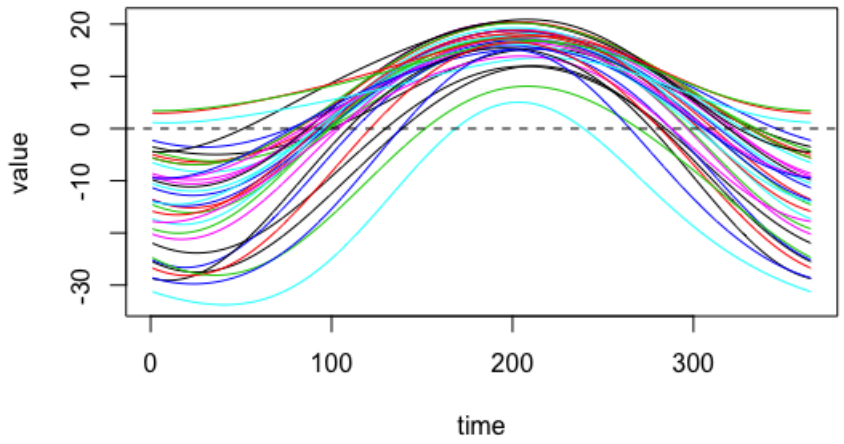

Figura 20: *Observaciones funcionales de temperatura anual en Canadá*

Dado que las componentes principales funcionales se expresan como combinación lineal de la misma base en la cual se suavizaron lo datos, utilizaremos esa misma base inicial para el método propuesto de manera tal que se pueda hacer una comparación directa con los componentes principales funcionales.

4.2.1. Método Propuesto

Para el método propuesto se generarán funciones aleatorias que sean combinación lineal de tres funciones de la base $\{1, \sin(x), \cos(x), \sin(2x), \cos(2x)\}$; es decir que generamos coeficientes aleatorios en S^2 para cada una de las 10 posibles combinaciones con tres funciones. Para cada combinación se generaron 1.000 coeficientes aleatorios y con esto se consideró un conjunto de 10.000 funciones aleatorias. A partir de estas funciones se crearon 2.000 subconjuntos de 10 funciones cada uno y se llevó a cabo el modelo haciendo selección de variables en cada subconjunto por significancia y por selección paso a paso hacia atrás con $R^2 ajustado$.

Haciendo selección de predictores por significancia se obtuvo que el modelo se explica lo suficientemente bien con 3 y 4 predictores, esto se puede observar en la Figura 21. Además, se obtuvo que las mejores funciones de proyección son 3, las cuáles se pueden observar en la Figura 22. El $R^2 ajustado$ obtenido al proyectar en estas tres funciones es de 0,769.

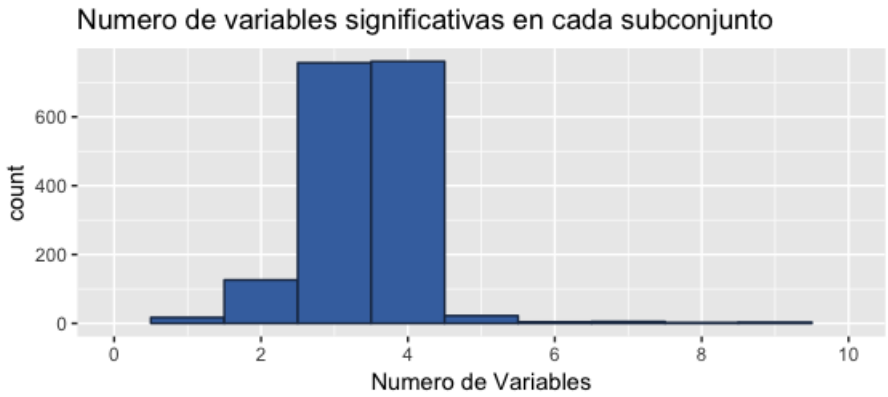

Figura 21: *Histograma de cantidad de predictores en cada submodelo después de selección por significancia datos Canada*

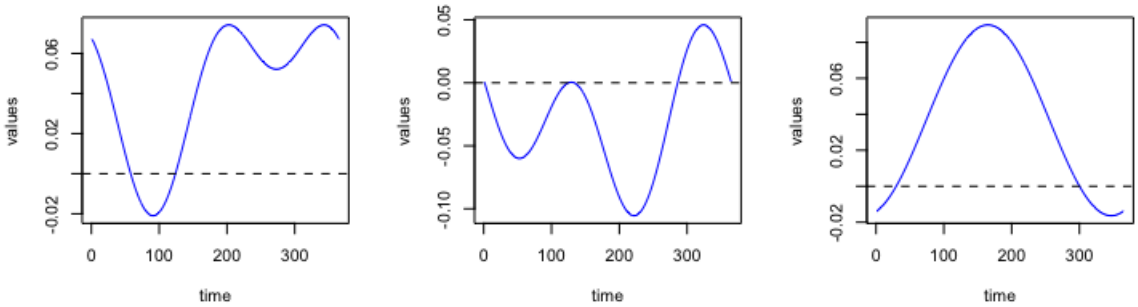

Figura 22: *Mejores funciones de proyección para selección por significancia datos Canada*

Haciendo selección de predictores pasa a paso hacia atrás por $R^2 ajustado$ se obtuvo que el modelo se explica lo suficientemente bien con 4 predictores, reafirmando nuevamente que la dimensión de los datos es aproximadamente 4; esto se puede observar en la Figura 23. En este caso nuevamente se obtuvo que hay 3 mejores funciones de proyección; éstas se pueden ver en la Figura 24. El $R^2 ajustado$ obtenido al proyectar sobre estas tres funciones fue de 0,769.

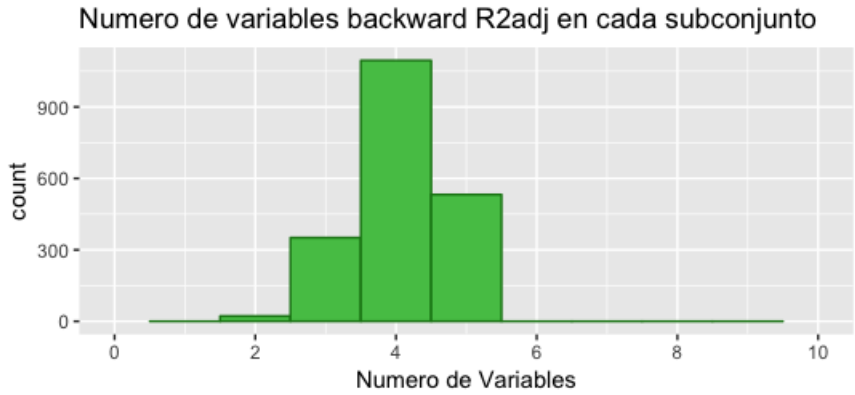

Figura 23: *Histograma de cantidad de predictores en cada submodelo después de selección por backward $R^2 ajustado$ datos Canada*

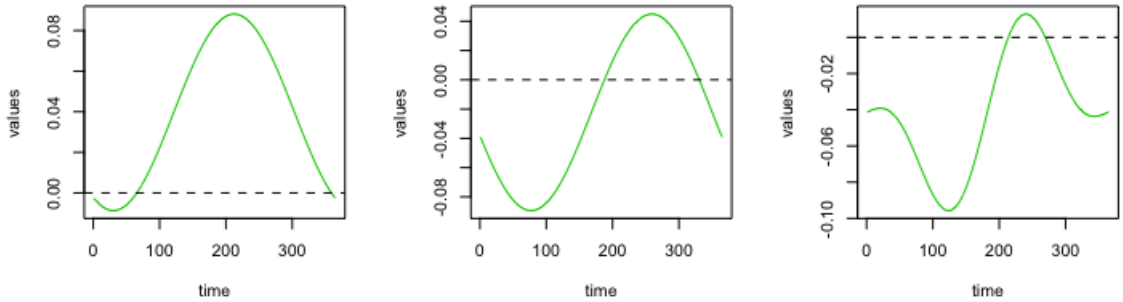

Figura 24: *Mejores funciones de proyección para selección por backward $R^2 ajustado$ datos Canada*

4.2.2. Comparación con regresión funcional por componentes principales

Si bien el desempeño del método propuesto no parece ser lo suficientemente bueno al obtener valores de 0,7686 y 0,7694 para el $R^2 ajustado$, se lleva a cabo una regresión funcional por componentes principales para poder comparar el desempeño de las funciones encontradas contra las funciones de componentes principales. En la Tabla 6 se muestra el porcentaje de variación que explica cada componente principal. Se tiene que las primeras tres componentes principales explican casi toda la varianza de las observaciones, esto

indica que el método propuesto identificó acertadamente la cantidad de funciones necesarias para explicar los datos.

Tabla 6: *Varianza explicada por los componentes principales funcionales - Canadá*

Componente	Prop. de Var.	Prop. Acumulada
1	0,892	0,892
2	0,085	0,977
3	0,018	0,995
4	0,004	0,999

En la Figura 25 se presentan los tres primeros componentes. Note que el primer componente principal PC1 presenta un comportamiento muy similar al de las observaciones y con esto explica el $89,2\,\%$ de la variación. Sin embargo, aunque las segunda y tercera componente explican menor parte de la variación, $8,5\,\%$ y $1,8\,\%$ respectivamente, se puede entender por su comportamiento que PC2 expresa la variación de temperatura entre invierno y verano mientras que PC3 expresa la variación entre el primer y segundo semestre del año. Resolviendo el modelo de regresión para explicar el logaritmo de la precipitación anual con respecto a las variables de pesos de cada una de las tres componente principal se tiene un $R^2 ajustado$ de 0,6909. Por lo tanto, el método propuesto, tanto con selección por significancia como selección paso a paso hacia atrás, encuentra funciones de proyección que conllevan a un mejor ajuste del modelo de regresión.

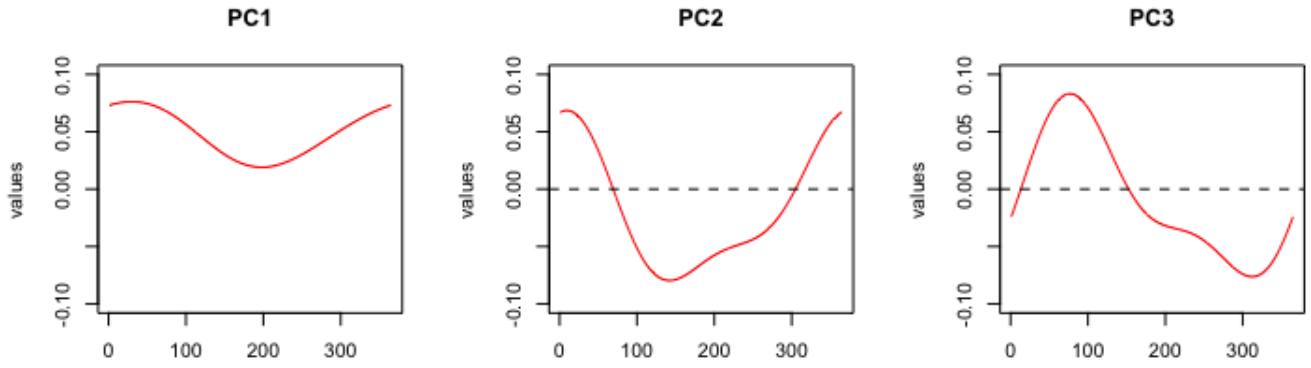

Figura 25: *Primeras tres componentes principales funcionales - Canadá*

5. Conclusiones

En este trabajo se propuso un método para encontrar, a partir de funciones aleatorias, las mejores funciones de proyección para llevar a cabo una regresión funcional. El procedimiento de regresión funcional llevado a cabo en el método propuesto es similar al de regresión funcional por componentes principales con la diferencia que, en vez de utilizar los componentes principales funcionales, se utilizan funciones aleatorias. Como se pudo observar en los ejemplos expuestos, el método propuesto no solo determina funciones de proyección sino que también, con pasos intermedios del procedimiento, se tiene un acercamiento a la dimensión del problema de regresión.

En ambos ejemplos en los cuales se llevó a cabo el método propuesto se obtuvo que la cantidad de mejores funciones de proyección encontradas es similar a la cantidad de componentes principales funcionales necesarios para explicar el 98 % de la varianza aproximadamente. Además, en el primer ejemplo se pudo observar que cuando se utilizan funciones simples como las indicadoras, las funciones que son linealmente independientes entre sí son también significativas para el modelo. En cuanto a las funciones de proyección encontradas se obtuvo en los ejemplos que el modelo de regresión dado por los productos internos de las observaciones con estas funciones tiene un desempeño según el $R^2 ajustado$ similar o mejor al que se obtiene en la regresión funcional por componentes principales. Sin embargo, se deja esta comparación general para estudios posteriores.

En el primer ejemplo se obtuvo que la regresión utilizando las funciones de proyección encontradas por el método propuesto cosiderando distintas bases funcionales y la regresión funcional por componentes principales, obtienen un desempeño similar al tener un $R^2 ajustado$ entre 0,88 y 0,89 para todos los casos. Si bien el método no logró encontrar funciones que superen el desempeño de los componentes principales funcionales, se encontraron funciones que se expresan en bases funcionales más sencillas que los componentes principales y sin embargo alcanzan un desempeño similar. El buen desempeño del modelo de regresión para las funciones de proyección encontradas en las tres bases consideradas se debe también a que la variable de respuesta es la energía de la curva, es decir que corresponde a la integral de su cuadrado. Por lo tanto, los produtos internos de las observaciones con funciones arbitrarias brindan información relevante sobre la variable de respuesta.

Para el segundo ejemplo se utilizó la misma base inicial con la cual se representan los componentes principales funcionales. Esto permitió comparar directamente los resultados del método propuesto con los resultados de la regresion por componentes principales funcionales. Mientras que el $R^2 ajustado$ de la regresión por componentes principales funcionales fue de $0,69$, el $R^2 ajustado$ de la regresión utilizando las funciones de proyección encontradas fue de $0,78$. Por lo tanto, a partir de funciones aleatorias se logró identificar funciones que mejoran el ajuste del modelo de regresión funcional a través de proyecciones.

Si bien el método solo se implementó para funciones aleatorias que sean combinación lineal de 2 o 3 funciones de una base inicial dada, se tuvo que esto fue suficiente para obtener buenos desempeños en los dos experimentos hechos. Sin embargo, si los datos funcionales en cuestión son más variables se espera que se tengan mejores resultados del método si se consideran funciones aleatorias más complejas.

Note que los componentes principales funcionales se construyen con información de los datos funcionales mientras que el método propuesto obtiene las mejores funciones de proyección a partir del desempeño de las proyecciones sobre funciones aleatorias. A pesar de esta diferencia en la manera de obtener funciones en las cuáles proyectar, se observa que los resultados en dimensión del problema y desempeño de la regresión son consistentes para ambos métodos.

Referencias

de Boor, C. (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences. Springer New York.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Li, S. (2011). Concise Formulas for the Area and Volume of a Hyperspherical Cap. *Asian Journal of Mathematics and Statistics*, 4.

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

Ramsay, J. O. and Dalzell, C. J. (1991). Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572.

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