TESIS DOCTORAL

Programa de Doctorado en Estadística Matemática y Aplicada

Departamento de Estadística e Investigación Operativa

UNIVERSIDAD DE GRANADA

Aportaciones en Comparación Estocástica y Modelos de Regresión Logística en Muestras Pareadas de Elementos de Sub-espacios de L_2 Finitamente Generados.

Cristhian Leonardo Urbano Leon

2024

Editor: Universidad de Granada. Tesis Doctorales
Autor: Cristhian Leonardo Urbano León
ISBN: 978-84-1195-423-5
URI: https://hdl.handle.net/10481/94736

Aportaciones en Comparación Estocástica y Modelos de Regresión Logística en Muestras Pareadas de Elementos de Sub-espacios de L_2 Finitamente Generados.

TESIS DOCTORAL

Presentada para optar al título de Doctor por la Universidad de Granada

Autor: Cristhian Leonardo Urbano Leon

Director:

Manuel Escabias Machuca

UNIVERSIDAD DE GRANADA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

GRANADA, ESPAÑA

2024

A mis padres.

Agradecimientos

Agradecimientos institucionales

Quiero empezar este apartado expresando mi agradecimiento al Ministerio de Ciencia Tecnología e Innovación de la república de Colombia, quien a través de la convocatoria 885 Doctorados en el Exterior, me brindó el respaldo financiero sin el cual, no habría sido posible al realización de mis estudios doctorales. Expreso también mis agradecimientos a la Universidad de Granada, a su Departamento de Estadística e Investigación Operativa y al Instituto Matemáticas IMAG, por brindarme los espacios, físicos y virtuales, necesarios para el desarrollo de mis actividades de investigación.

Agradecimientos Personales

Llegado a esta instancia de mi formación, es inevitable que confluyan lo académico y lo personal. Dicho esto, quiero agradecer de manera muy especial a mi director de tesis, Manuel Escabias (Manolo), cuya guía firme y respaldo a nivel académico y personal han hecho que su labor como mentor y su papel como amigo sean inmejorables. Quiero agradecer también muy "espacialmente" a mi colega y amiga, Paola Ovalle, por tantas y tantas horas de tertulia, difuminadas entre operadores espectrales y regresiones funcionales. Su apoyo y amistad son invaluables. A mi pareja, Isabel, le agradezco todo su amor, comprensión y sacrificio, los cuales se convirtieron en la fuerza impulsora detrás de

cada paso en este camino. Le agradezco a mi madre, Marian, por todo su amor y constante motivación. Su inquebrantable fe en mí ha sido mi fortaleza a lo largo de los años. A mi padre, Gonzalo, que siempre creyó en mí y me alentó a seguir mis sueños, le agradezco profundamente. Aunque su partida dejó un vacío imposible de llenar, su sabiduría, su fuerza inquebrantable y su amor incondicional son mi faro de inspiración.

Resumen

Los datos funcionales de segunda generación, se caracterizan por la imposibilidad de asumir la independencia entre las observaciones funcionales. Esto debido a que el fenómeno subyacente se encuentra condicionado a diseños como series de tiempo funcionales, estudios longitudinales funcionales, medidas repetidas funcionales, entre otros. En ese sentido, el trabajo aquí presentado pretende aportar metodologías en el marco de dicho tipo de datos, tratando diferentes tópicos en el proceso. En cada caso, se asume que los objetos funcionales, dentro de un mismo contexto, son elementos de un mismo subespacio de dimensión finita del espacio de funciones cuadrado integrables L_2 para un mismo dominio compacto, lo que permite la utilización de la metodología conocida como expansión básica.

En primera instancia, en este trabajo se aborda el tema de la variabilidad y la asociación entre conjuntos de datos funcionales, a partir de una varianza, covarianza y correlación escalares. Se discute brevemente la utilidad de las formas funcionales de algunos estadísticos muestrales propuestos en la literatura, al ser usados como herramientas comparativas. Se conduce también un estudio de simulación que aporta evidencia de la consistencia de las medidas de resumen escalares para datos funcionales estudiadas y se muestran algunos ejemplos de aplicación en datos reales.

En segundo lugar, en el campo de la comparación estocástica, motivados por los datos de la positividad de COVID-19 en Colombia, se utiliza la metodología de datos funcionales para examinar si existen diferencias significativas entre dos olas de contagio ocurridas entre

el 7 de julio de 2020 y el 20 de julio de 2021. Para este problema, inicialmente se utiliza una prueba t funcional puntual, posteriormente, se utiliza una prueba estadística alternativa para muestras funcionales pareadas. Esta prueba estadística alternativa, que genera un valor-p escalar, proporciona una idea global sobre la diferencia de medias funcionales en un contexto pareado, complementando las pruebas puntuales pareadas para datos funcionales existentes.

Por otra parte, ya en el problema de regresión, se presenta una propuesta para extender el modelo de regresión logística funcional – que modela una variable de respuesta escalar binaria a partir de un predictor funcional – al caso donde las observaciones provienen de un diseño de medidas repetidas funcionales. La extensión se aborda incluyendo un efecto aleatorio en el modelo. En este caso, el enfoque de expansión básica suele inducir un problema de multicolinealidad en el modelo multivariado emergente, que se resuelve con el uso de los componentes principales funcionales del predictor funcional, dando como resultado un nuevo modelo logístico funcional de medidas repetidas en componentes principales. La propuesta se contextualiza a través de un estudio de simulación en donde se evalúan los ajustes de cuatro modelos, en tres escenarios distintos para cuatro parámetros funcionales diferentes.

Productos de investigación asociados e indicadores de calidad

Como resultado de la investigación que conduce a esta tesis doctoral, se han obtenido en total los siguientes productos asociados:

Tipo de producto	**Cantidad**
Artículos de investigación	3
Ponencias en eventos científicos	7
Conferencias invitadas	1

La información detallada y los indicadores de calidad de cada una las publicaciones científicas se muestra a continuación:

- **Artículo de investigación 1:**

  Título: *"Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA".*

  Autores: *Cristhian Leonardo Urbano-Leon, Manuel Escabias.*

  Revista: *Stats.*

  Factores de impacto de la revista:

Año	Factor de Impacto	Rango	Cuartil	Área
2022	1.3	104/164	$Q3$	Estadística y Probabilidad

Estado: Publicado, 28 de octubre de 2022.

doi: 10.3390/stats5040059.

Posición de autor: Autor principal.

Este artículo se encuentra anexo en el Apéndice A.

- **Artículo de investigación 2:**

  Título: *"Scalar Variance and Scalar Correlation for Functional Data".*

  Autores: *Cristhian Leonardo Urbano-Leon, Manuel Escabias, Diana Paola Ovalle, Javier Olaya-Ochoa.*

  Revista: *Mathematics.*

  Factores de impacto de la revista:

Año	Factor de Impacto	Rango	Cuartil	Área
2022	2.4	23/330	$Q1$	Matemáticas

  Estado: Publicado, 9 de marzo de 2023.

  doi: 10.3390/math11061317.

  Posición de autor: Autor principal.

  Este artículo se encuentra anexo en el Apéndice B.

- **Artículo de investigación 3:**

  Título: *"Repeated Measures in Functional Logistic Regression".*

  Autores: *Cristhian Leonardo Urbano-Leon, Ana María Aguilera, Manuel Escabias.*

  Revista: *Mathematics and Computers in Simulation.*

  Factores de impacto de la revista:

Año	Factor de Impacto	Rango	Cuartil	Área
2022	4.6	4/267	$Q1$	Matemáticas, aplicada

Estado: Publicado, disponible en línea 10 de mayo de 2024.

doi: 10.1016/j.matcom.2024.05.002.

Posición de autor: Autor principal.

Este artículo se encuentra anexo en el Apéndice C.

Con respecto a las comunicaciones impartidas, la información relevante es la siguiente:

1. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"Comparison of Positivity in Two Epidemic Peaks of Covid-19 in Colombia with FDA".*

   Presentada en: *The 19th Conference of the Applied Stochastic Models and Data Analysis International Society ASMDA2021*

   Lugar del evento: Atenas, Grecia

   Fecha de realización del evento: 1 al 4 de junio de 2021

   Fecha de presentación: 3 de junio de 2021

   Sitio web del evento: http://www.asmda.es/asmda2021.html

2. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"Métodos de comparación en dos muestras pareadas de datos funcionales".*

   Presentada en: *XXXIX Congreso Nacional de Estadística e Investigación Operativa (SEIO2022)*

   Fecha de realización del evento: 7 al 10 de junio de 2022.

   Lugar del evento: Universidad de Granada, Granada, España

   Fecha de presentación: 8 de junio de 2022.

   Sitio web del evento: https://seio2022.confereasy.com/es/

3. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"A functional logistic regression model with non-independent*

functional variables".

Nombre del evento: *15th International Conference of the ERCIM WG on Computational and Methodological Statistics 16th International Conference on Computational and Financial Econometrics*

Lugar del evento: King's College London, Reino Unido

Fecha de realización del evento: 17 al 19 de diciembre de 2022.

Fecha de presentación: 18 de diciembre de 2022.

Sitio web del evento: https://www.cmstatistics.org/CMStatistics2022/

4. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"Comparison of paired curves from the gait cycle in different conditions".*

   Nombre del evento: *3rd International Workshop on Stochastic Processes and their Applications*

   Lugar del evento: Evento en modalidad virtual

   Fecha de realización del evento: 12, 17, 19, 24 y 26 de enero de 2023.

   Fecha de presentación: 24 de enero de 2023.

   Sitio web del evento: https://sites.google.com/view/iwspa2023-gt-seio/home

5. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"Repeated Measures in Functional Logistic Regression".*

   Nombre del evento: *Second International Conference on Mathematical And Computational Modelling, Approximation and Simulation New trends, recent developments and applications in environment and natural resources*

   Lugar del evento: Torino, Italia.

   Fecha de realización del evento: 29 de mayo al 1 de junio de 2023.

   Fecha de presentación: 30 de mayo de 2023.

   Sitio web del evento: https://www.macmas2023.unito.it/

6. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"Random effect inclusion in a functional logistic regression model".*

Nombre del evento: *I Joint Workshop on Functional Data Analysis and Nonparametric Statistics*

Lugar del evento: Miraflores de la Sierra, España.

Fecha de realización del evento: 7 al 10 de junio de 2023.

Fecha de presentación: 7 de junio de 2023.

Sitio web del evento: https://seio2023.com/es/

7. **Tipo de comunicación:** Ponencia en evento científico.

   Título de la ponencia: *"Modelo de Regresión Logística Funcional en Medidas Repetidas".*

   Nombre del evento: *XL Congreso Nacional de Estadística e Investigación Operativa y las XIV Jornadas de Estadística Pública*

   Lugar del evento: Elche, España.

   Fecha de realización del evento: 7 al 10 de noviembre de 2023.

   Fecha de presentación: 8 de noviembre de 2023.

   Sitio web del evento: https://seio2023.com/es/

Adicional a las anteriores charlas, se impartió la siguiente conferencia

- **Tipo de comunicación:** Conferencia invitada.

  Título de la conferencia: *"Avances en el Estudio del Modelo Funcional de Respuesta Binaria con Medidas Repetidas".*

  En el marco de: *Ciclo de Conferencias Estadística y Ciencia de Datos Patricia Román.*

  Lugar del evento: Granada, España.

  Fecha de presentación: 16 de junio de 2023.

  Sitio web del evento: https://ciclo-de-conferencias-patricia-roman

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y aportaciones

Dentro de la investigación científica, existe interés en el análisis de datos provenientes de observaciones de fenómenos dentro de un dominio continuo, que pueden ser representados como objetos funcionales. Estos datos representan información valiosa proveniente de espectros, señales, imágenes y demás características de fenómenos del mundo real. Así, el análisis de datos funcionales (FDA por sus siglas en inglés) provee un marco conceptual sobre el cual los datos, que ahora se supone son objetos funcionales, pueden ser tratados. Este hecho, con ayuda del exponencial desarrollo computacional de las últimas dos décadas, ha generado una explosiva popularización de los métodos para datos funcionales, ya que según Wang et al. (2016), el término FDA fue acuñado por Ramsay (1982), Ramsay & Dalzell (1991), pero los datos funcionales, como tal, ya habían sido tratados desde mediados del siglo XX en los trabajos de Grenander (1950) y Rao (1958).

En el FDA, la mayoría de las técnicas desarrolladas están basadas en el paradigma de extensión de conceptos de la estadística escalar a los objetos funcionales; razón por la cual, es posible encontrar en la literatura extensiones para datos funcionales de casi todas las ramas de la estadística, como los modelos de regresión, contrastes de hipótesis, análisis de la varianza, entre otros, y con enfoques tan diversos como paramétrico, no paramétrico o Bayesiano. Conforme a esto, los primeros desarrollos para los datos funcionales, que Wang et al. (2016) denomina datos funcionales de primera generación, corresponden a

todas esas técnicas estadísticas extendidas al FDA bajo el supuesto de independencia en las observaciones funcionales. No obstante, es común que en la investigación aplicada se presenten situaciones en donde este supuesto no puede ser asumido per se, en cuyo caso, se tiene lo que Koner & Staicu (2023) han denominado datos funcionales de segunda generación, que comprenden datos funcionales provenientes de estudios longitudinales, series de tiempo y estudios de datos espaciales correlacionados.

Como casos concretos de datos funcionales de segunda generación, se tienen las medidas repetidas funcionales, un diseño en donde las observaciones funcionales provienen de la medición reiterada de una misma característica de una misma unidad experimental, bajo las mismas o diferentes condiciones. También, pueden ser considerados de segunda generación aquellos diseños que proveen muestras pareadas, es decir, observaciones de un fenómeno cuya naturaleza genera un emparejamiento de los datos funcionales. Estas muestras pareadas pueden, bajo algunas circunstancias, ser vistas como un caso particular de las medidas repetidas cuando el número de repeticiones es dos.

Este trabajo, que se presenta como agrupación de publicaciones, se encuentra enmarcado dentro de los datos funcionales de segunda generación y tiene como objetivo aportar metodologías para el tratamiento de este tipo de datos funcionales. Así, en el artículo *"Scalar Variance and Scalar Correlation for Functional Data"* (ver Urbano-Leon et al. (2023)) se trata el problema de obtención de medidas de variabilidad y asociación en datos funcionales. En el artículo *"Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA"* (ver Urbano-Leon & Escabias (2022)) se aborda el problema de comparación estocástica por medio del contraste de medias funcionales en muestras pareadas. Mientras que en el artículo *"Repeated Measures in Functional Logistic Regression"* (ver Urbano-Leon et al. (2024)) se trata el problema de regresión logística funcional de medidas repetidas. En todos los casos, se asume que los datos funcionales son elementos de un subespacio del espacio de funciones cuadrado integrables L_2. Estos subespacios son generados a partir de una base finita de funciones, es decir, son finitamente generados.

Las aportaciones de cada uno de los artículos presentados se describen a continuación:

I *Scalar Variance and Scalar Correlation for Functional Data:*

Aportación 1 Con el ánimo de proveer medidas de resumen que sirvan para el estudio de la correlación de dos conjuntos de datos funcionales, se presenta una medida escalar de covarianza y de correlación para datos funcionales.

Aportación 2 Con el propósito de obtener una medida de variabilidad para los datos funcionales que preserve su uso como herramienta comparativa, en el mismo sentido que la varianza clásica, se presenta una medida de varianza escalar para datos funcionales.

Aportación 3 Se presentan desarrollos teóricos sobre algunas propiedades de las medidas estudiadas.

Aportación 4 Se presentan evidencia, bajo simulación, de la consistencia de la varianza propuesta.

Aportación 5 Se presentan varios ejemplos de aplicación de las medidas a modo de contextualización.

II *Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA:*

Aportación 6 Se presenta un contraste de hipótesis para la igualdad de medias funcionales en muestras funcionales pareadas, que proporciona un criterio de decisión global sobre si existen o no, diferencias significativas entre las dos medias funcionales muestrales observadas.

Aportación 7 El contraste propuesto se aplica a datos funcionales de la positividad de COVID-19 en dos olas de contagio en Colombia.

III *Repeated Measures in Functional Logistic Regression:*

Aportación 8 Se presenta una extensión del modelo logístico funcional, que pretende modelar una variable respuesta binaria a partir de un predictor funcional, bajo la consideración de que las observaciones provienen de un diseño de medidas repetidas, es decir, un modelo logístico funcional de medidas repetidas.

Aportación 9 Se muestra un estudio de simulación para comparar la estimación de cuatro diferentes funciones parámetro, partiendo de cuatro modelos en tres escenarios de simulación diferentes.

En este trabajo, la mayor parte del contenido de las publicaciones mencionadas se encuentra integrada como capítulos, mientras que los anexos, contienen su forma final publicada como artículos. Así, este documento está dividido de la siguiente forma. En el primer capítulo, se muestra la introducción con la finalidad de relacionar las ideas de cada una de las publicaciones, junto con la lista completa de aportaciones. En el segundo capítulo, se presentan algunos conceptos básicos sobre el FDA, que son recurrentes en los capítulos posteriores. En el tercer capítulo, se presenta la introducción, objetivos, metodología, resultados y conclusiones del artículo *Scalar Variance and Scalar Correlation for Functional Data*. En el cuarto capítulo, se presenta la introducción, objetivos, metodología, resultados y conclusiones del artículo *Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA*. En el quinto capítulo, se presenta la introducción, objetivos, metodología, resultados y conclusiones del artículo *Repeated Measures in Functional Logistic Regression*. En el sexto capítulo, se detallan las lineas abiertas producto de la investigación realizada. Finalmente, en los apéndices A, B y C, se encuentran los artículos 1, 2 y 3 respectivamente.

Capítulo 2

Análisis de datos funcionales

2.1. Introducción

El análisis de datos funcionales, es una rama de la estadística cuyo objetivo está marcado por la construcción, tratamiento y análisis de objetos funcionales, tanto de forma individual como en conjunto. Es decir, el FDA se ocupa de aquellas observaciones individuales continuas de un fenómeno, bajo el supuesto de que estas pueden ser expresadas mediante una función continua definida en un espacio compacto $\mathfrak{T}$ y recorrido $\mathfrak{T}'$. Así, los datos funcionales obtienen una caracterización dependiendo del espacio $\mathfrak{T}$ y $\mathfrak{T}'$, por ejemplo, si $\mathfrak{T} \subset \mathbb{R}^2 \wedge \mathfrak{T}' \subset \mathbb{R}$ se dice que cada dato funcional es una superficie contenida en $\mathbb{R}^3$ (ver Figura 2.1 panel izquierdo), pero si $\mathfrak{T} \subset \mathbb{R} \wedge \mathfrak{T}' \subset \mathbb{R}$, entonces cada dato funcional es una curva contenida en $\mathbb{R}^2$ (ver Figura 2.1 panel derecho).

Los datos funcionales son ideados originalmente como objetos abstractos continuos, lo que plantea serias complicaciones para su tratamiento debido a la imposibilidad de observar el continuo. En su lugar, lo que se tiene en la práctica, cuando el análisis de datos es el objetivo, es un conjunto de mediciones escalares provenientes de la observación en un conjunto finito de instantes de los datos funcionales. Esto provoca que, al contrario de los datos escalares, los datos funcionales necesiten de un pre-procesamiento en donde se asume un conjunto de supuestos que permite su análisis. Uno de los principales supuestos, es que los datos funcionales pertenecen a un espacio de funciones con ciertas características

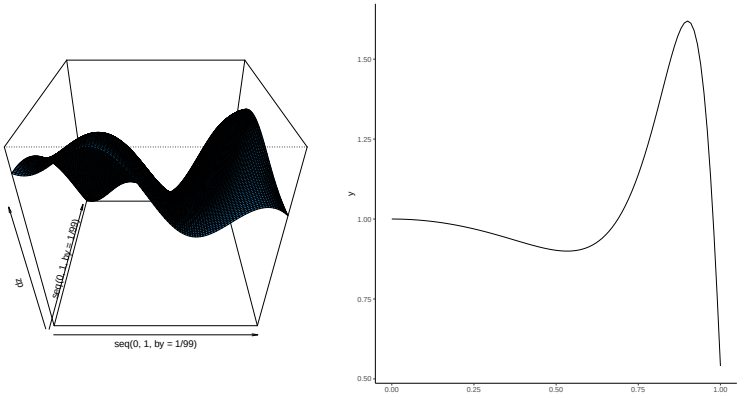


Figura 2.1: Ejemplos de datos funcionales. Superficie (izquierda), curva (derecha)

analíticas y topológicas deseables, equivalentes a los espacios euclidianos en los que la estadística clásica se ha desarrollado y que permiten tener nociones como tamaño, distancia y magnitud entre otras. En este trabajo, asumimos que los datos funcionales son elementos de espacios funcionales generados a partir de una base finita (espacios funcionales finitamente generados). Estos espacios, son concebidos como subespacios del espacio de funciones cuadrado integrables L_2, cuya dimensión es infinita.

2.2. Estructuras algebraicas para datos funcionales

En el FDA, uno de los supuestos principales es que los objetos funcionales son elementos de un espacio de funciones. Estos espacios son estructuras algebraicas pensadas con el fin de estudiar las interacciones entre elementos del conjunto que las conforman, a partir de ciertas operaciones definidas, siendo estas operaciones las que brindan la estructura del espacio y condicionan los desarrollos que en él se hacen. En el caso de los datos funcionales, la estructura algebraica más recurrente es la de espacio de Hilbert, que se define a continuación.

Definición 2.2.1. *Dado un espacio vectorial $\mathbb{H}$ sobre un campo $\mathbb{K}$, este se denomina un espacio pre-Hilbert si existe una función $\langle \cdot, \cdot \rangle : \mathbb{H} \times \mathbb{H} \longrightarrow \mathbb{K}$ bien definida, tal que para todo $v, u, w \in \mathbb{H}$ se tiene que*

- $\langle v + u, w \rangle = \langle v, w \rangle + \langle u, w \rangle$.

- $\langle \alpha u, w \rangle = \alpha \langle u, w \rangle$.
- $\langle v, v \rangle \geq 0$.
- $\langle v, v \rangle = 0 \Leftrightarrow v = 0_{\mathbb{H}}$,

siendo $0_{\mathbb{H}}$ *el elemento neutro del espacio* $\mathbb{H}$.

La inclusión de la función $\langle \cdot, \cdot \rangle$, denominada producto interno, permite al espacio vectorial $\mathbb{H}$ contar con una noción generalizada de ángulo entre dos elementos.

Definición 2.2.2. *Un espacio de Hilbert es un espacio vectorial completo y con producto interno.*

El producto interno $\langle \cdot, \cdot \rangle$ definido en un espacio vectorial, permite dotar a este de la noción de tamaño a partir de la norma inducida como $\|v\| = \langle v, v \rangle$ y una distancia por medio de una métrica inducida por $m(v, u) = \sqrt{\|v - u\|}$. Por esta razón, todo espacio de Hilbert es también un espacio normado y a su vez, un espacio métrico (ver Kreyszig (1989), Hsing & Eubank (2015)). Así, la propiedad topológica de completitud se refiere a que, en el espacio métrico subyacente, cada sucesión de elementos del espacio $\mathbb{H}$ contiene una sub-sucesión convergente a un elemento del espacio $\mathbb{H}$.

Por otro lado, dentro de los espacio vectoriales, de los cuales un espacio de Hilbert es un caso particular, el concepto de espacio vectorial generado por un conjunto obtiene un papel fundamental en la definición de la dimensión de un espacio. Este, se define como sigue.

Definición 2.2.3. *Dado un conjunto no vacío* Q *de elementos de un espacio vectorial* $\mathscr{V}$*, se define el espacio vectorial* $Gen(Q)$ *como la intersección de todos los espacios vectoriales que contienen a* Q*. Si* $Gen(Q) = \mathscr{V}$*, entonces se dice que* $\mathscr{V}$ *es generado por* Q*.*

Cuando un espacio vectorial $\mathscr{V} = Gen(Q)$ y $|Q| < \infty$ (el cardinal de Q es finito), se dice que el espacio $\mathscr{V}$ es finitamente generado y por lo tanto, tiene dimensión finita. Si para un espacio $\mathscr{V}$ no existe Q finito tal que $Gen(Q) = \mathscr{V}$, se dice que $\mathscr{V}$ tiene dimensión infinita. Además, si Q es finito y es un conjunto linealmente independiente (ver Carrell (2010)), entonces se dice que Q es una base para $\mathscr{V}$ y $|Q|$ es la dimensión de del espacio.

El concepto de independencia lineal, en el contexto de los espacio vectoriales con producto interior, como los espacios de Hilbert, suele ser tratado por medio del concepto de ortonormalidad (ver Definición 2.2.4), pues puede probarse que si un conjunto de elementos de un espacio vectorial son ortogonales, entonces estos son linealmente independientes. El reciproco de esta afirmación no es necesariamente cierta.

Definición 2.2.4. *Dado un espacio de Hilbert* $\mathbb{H}$ *con producto interno* $\langle , \rangle$*, dos elementos* v, u $\mathbb{H}$ *se dicen ortogonales si se verifica que* $\langle v, u \rangle = 0$*. Si además,* $\|v\| = \|u\| = 1$*, los elementos* v *y* u *se dicen ortonormales. Así, dada una sucesión* $\{v_j\}_{j=1}^{\infty} \subset \mathbb{H}$*, esta se dice que es una sucesión ortonormal si* $\|v_j\| = 1 \ \wedge \ \langle v_j, v_k \rangle = 0, \ \forall j, k \in \mathbb{N}, j \neq k.$

El siguiente teorema, cuya demostración puede ser consultada en Hsing & Eubank (2015), permitirá definir una base para un espacio de Hilbert a partir de una sucesión ortogonal de elementos y, posteriormente, para los datos funcionales como caso particular.

Teorema 2.2.1. *Dada una sucesión ortonormal* $\{v_i\}_{i=1}^{\infty} \subset \mathbb{H}$*, siendo* $\mathbb{H}$ *un espacio de Hilbert, con producto interno* $\langle , \rangle$*, se cumple que*

$$\sum_{j=1}^{\infty} \langle e, v_j \rangle^2 \leq \|e\|^2 \quad \forall e \in \mathbb{H}.$$

Además,

$$\sum_{j=1}^{\infty} \langle e, v_j \rangle \, v_j,$$

converge en $\mathbb{H}$*.*

Definición 2.2.5. *Dada una sucesión ortonormal* $\{v_j\}_{j=1}^{\infty}$ *de elementos de espacio de Hilbert* $\mathbb{H}$*, esta es llamada un sistema ortonormal completo (CONS por sus siglas en inglés) si* $Gen(\{v_j\}_{j=1}^{\infty}) = \mathbb{H}$*.*

La Definición 2.2.5 exige entonces que

$$\forall j, \ \langle w, v_j \rangle = 0 \Rightarrow w = 0,$$

lo que implica que en la CONS se encuentran todos los elementos ortonormales de su tipo. Así, si $\{v_j\}_{j=1}^{\infty}$ es una CONS en $\mathbb{H}$, cada elemento $x \in \mathbb{H}$ puede ser representado en

términos de una expansión

$$x = \sum_{j=1}^{\infty} \langle x, v_j \rangle \, v_j,$$

donde los $\langle x, v_j \rangle$, para $j = 1, 2, ...$ son conocidos como coeficientes generalizados de Fourier (ver Eubank (1998)), y para quienes se cumple la relación de Parseval dada por

$$\|x\|^2 = \sum_{j=1}^{\infty} \langle x, v_j \rangle^2 .$$

2.2.1. El espacio L_2

Como se dijo con anterioridad, en el FDA se asume que los objetos funcionales pertenecen a espacios de funciones. Uno de estos espacios, de hecho el más popular por sus propiedades teóricas, es precisamente el espacio de funciones cuadrado integrables en un dominio compacto $\mathfrak{T}$, denotado como $L_2(\mathfrak{T})$. Este, es un espacio vectorial sobre el campo $\mathbb{R}$, cuya estructura de espacio de Hilbert está dada bajo el producto interno definido como

$$\begin{aligned} \langle \cdot, \cdot \rangle : L_2(\mathfrak{T}) \times L_2(\mathfrak{T}) &\longrightarrow \mathbb{R} \\ (f, g) &\longrightarrow \langle f, g \rangle = \int_{\mathfrak{T}} f(t) g(t) dt. \end{aligned} \tag{2.1}$$

Además, $L_2(\mathfrak{T})$ es un espacio normado gracias a la norma inducida por

$$\|f\| = \langle f, f \rangle^{1/2} = \left(\int_{\mathfrak{T}} f^2(t) dt \right)^{1/2} \qquad \forall f \in L_2(\mathfrak{T}), \tag{2.2}$$

y es un espacio métrico con la métrica definida por

$$\mathscr{M}(f, g) = \|f - g\| = \left(\int_{\mathfrak{T}} \left(f(t) - g(t) \right)^2 dt \right)^{1/2} \qquad \forall f, g \in L_2(\mathfrak{T}). \tag{2.3}$$

Dado que el espacio $L_2(\mathfrak{T})$ es un espacio de Hilbert, existe una sucesión ortonormal completa $\Phi = \{\phi_j\}_{j=1}^{\infty}$ tal que

$$f = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \phi_j \quad \forall f \in L_2(\mathfrak{T}).$$

La sucesión de los coeficientes $\{\langle f, \phi_j \rangle\}_j^{\infty}$ es un elemento del espacio de sucesiones ℓ_2. Más aún, es fácil ver que la aplicación

$$\mathcal{N} : L_2(\mathfrak{T}) \longrightarrow \ell_2 \tag{2.4}$$

$$f = \sum_{j=1}^{\infty} \langle f, \phi_j \rangle \, \phi_j \longrightarrow \{\langle f, \phi_j \rangle\}_j^{\infty}, \tag{2.5}$$

es una función biyectiva, por lo que los espacios $L_2(\mathfrak{T})$ y ℓ_2 son espacios isomorfos.

La sucesión de coeficientes $\{\langle f, \phi_j \rangle\}_j^\infty$, por lo visto con anterioridad, cumple la relación de Parseval, por lo que estos en algún momento decaerán a cero. Esto permite entonces, caracterizar elementos dentro de subespacios $\mathcal{H} \subset L_2(\mathfrak{T})$ de dimensión $d < \infty$ por medio del truncamiento de la base Φ. Es decir, es posible definir cualquier subespacio de dimensión finita $\mathcal{H} \subset L_2(\mathfrak{T})$, como $\mathcal{H} = Gen\left(\{\phi_j\}_{j=1}^d\right)$. Además, la aplicación

$$\mathcal{N}' : \mathcal{H} \longrightarrow \mathbb{R}^d \tag{2.6}$$

$$f = \sum_{j=1}^{d} \langle f, \phi_j \rangle \, \phi_j \longrightarrow \{\langle f, \phi_j \rangle\}_j^d, \tag{2.7}$$

es también una biyección, por lo que en cualquier caso, $\mathcal{H}$ es isomorfo a $\mathbb{R}^d$.

Por otro lado, como puede verse en Hsing & Eubank (2015), las sucesiones

$$\Phi_1 = \left\{\phi_{1,0}(t) = 1, \phi_{1,k}(t) = \sqrt{2}\cos(k\pi t), k = 1, 2, \ldots\right\}, \tag{2.8}$$

$$\Phi_2 = \left\{\phi_{2,k}(t) = \sqrt{2}\sin(k\pi t), k = 1, 2 \ldots\right\}, \tag{2.9}$$

$$\Phi_3 = \left\{\phi_{3,0}(t) = 1, \phi_{3,2k-1}(t) = \sqrt{2}\sin(2k\pi t), \phi_{2k} = \sqrt{2}\cos(k\pi t), k = 1, 2, \ldots\right\} \tag{2.10}$$

proveen bases CONS para el espacio de funciones L_2.

2.3. Datos funcionales

Definición 2.3.1. *Dado un espacio muestral Ω, una sigma álgebra $\mathfrak{S}_\Omega$ de Ω, $\mathbb{H}$ un espacio de Hilbert, $\mathfrak{S}_\mathbb{H}$ una sigma álgebra de $\mathbb{H}$ y una función de medida $\mathbb{P} : \mathfrak{S}_\mathcal{H} \longrightarrow [0, 1]$, una variable aleatoria sobre un espacio de Hilbert $\mathbb{H}$, es una función medible $\mathscr{X} : \Omega \longrightarrow \mathbb{H}$ que preserva la estructura entre las dos sigma álgebras definidas, i.e.*

$$\forall F \in \mathfrak{S}_\mathbb{H} \exists \omega \in \mathfrak{S}_\Omega : X^{-1}(F) = \omega.$$

Además, el conjunto de variables aleatorias $\mathscr{X}$ definidas sobre este (denotado como L_2^Ω), y para las cuales

$$\mathbf{E}(\|\mathscr{X}\|^2) = \int_\Omega \|\mathscr{X}(\omega)\|^2 \, d\mathbb{P}(\omega) < \infty,$$

es en sí, un espacio funcional con estructura Hilbertiana cuyo producto interno puede definirse como

$$\langle \mathscr{X}, \mathscr{Y} \rangle_{L_2^\Omega} = \int_\Omega \mathscr{X}(\omega)\mathcal{Y}(\omega) d\mathbb{P}(\omega).$$

Note que la Definición 2.3.1 está dada en términos de un espacio de Hilbert arbitrario $\mathbb{H}$. Cuando dicho espacio es un espacio de funciones, se dice que función medible $\mathscr{X}$ de la Definición 2.3.1 es una variable aleatoria funcional (V.A.F) y que un dato funcional $\mathcal{X}_i$ es una observación de dicha variable (ver Ferraty & Vieu (2006)). En este punto, resulta importante señalar que este trabajo se enmarca en observaciones funcionales del espacio $L_2(\mathfrak{T})$ cuando $\mathfrak{T}$ es un subconjunto compacto de $\mathbb{R}$, el cual es referido como $L_2([a,b])$ o simplemente L_2 cuando no haya lugar a ambigüedad, y sus subespacios de dimensión finita arbitraria d, los cuales heredan la estructura Hilbertiana, referidos en este trabajo como $\mathcal{H}$. Además, siempre que no haya lugar a ambigüedad, en este trabajo de asume que una V.A.F está definida sobre el espacio de probabilidad arbitrario $(\Omega, \mathfrak{S}_\Omega, \mathbb{P})$.

Con respecto a las medidas para los datos funcionales, dada una variable aleatoria funcional $\mathscr{X}$, la media de $\mathscr{X}$ denotada como $\mathbf{E}(\mathscr{X})$, $\mu_{\mathscr{X}}$ o simplemente μ si no hay lugar a ambigüedad, es la función

$$\mu(t) = \mathbf{E}(\mathscr{X})(t) = \int_\Omega \mathscr{X}(\omega)(t) d\mathbb{P}(\omega). \tag{2.11}$$

Esta definición, según asegura Hsing & Eubank (2015), provee una extensión natural del concepto de media de una variable aleatoria, al caso funcional. Además, la función

$$C_{\mathscr{X}}(t,s) = \mathbf{E}\left[(\mathscr{X}(t) - \mu(t))\,(\mathscr{X}(s) - \mu(s))\right] = \int_\Omega (\mathscr{X}(t,\omega) - \mu(t))\,(\mathscr{X}(s,\omega) - \mu(s))\, d\mathbb{P}(\omega), \tag{2.12}$$

es llamada función de covarianza, que según Todorovic (1992), es continua siempre que

$$\lim_{h\to 0} \mathbf{E}\left[(\mathscr{X}(t+h) - \mathscr{X}(t))^2\right] = 0, \quad \forall t \in \mathfrak{T}. \tag{2.13}$$

Esta propiedad, conocida como continuidad en media cuadrática en el contexto de los procesos estocásticos, permite la definición del operador covarianza asociado a la variable aleatoria funcional $\mathscr{X}$ como

$$\begin{aligned} \mathcal{C}_{\mathscr{X}} = L_2(\mathfrak{T}) &\longrightarrow L_2(\mathfrak{T}) \\ f &\longrightarrow \mathcal{C}_{\mathscr{X}}(f) = \int_{\mathfrak{T}} C(t,s) f(s) ds. \end{aligned} \tag{2.14}$$

Si bien los datos funcionales son, en palabras de Wang et al. (2016), intrínsecamente de dimensión infinita, la definición de variable aleatoria funcional dada en la Definición 2.3.1 no se restringe solo a espacios funcionales de dimensión infinita. En ese sentido, dada la sucesión ortonormal $\Phi = \{\phi_j\}_{j=1}^{\infty}$ que genera al espacio $L_2(\mathfrak{T})$, el conjunto Φ_d de los primeros d términos de Φ, es tal que $Gen(\Phi_d) = \mathcal{H} \subset L_2(\mathfrak{T})$. Este espacio hereda el producto interno de $L_2(\mathfrak{T})$ y las demás medidas inducidas a partir de este, restringidas a $\mathcal{H}$. A la sucesión truncada Φ_d se le suele denominar como base truncada de $\mathcal{H}$ o simplemente base finita. De esta manera, los datos funcionales pueden ser representados por un único vector de coeficientes $\mathcal{H}$ bajo Φ_d, por medio del isomorfismo

$$\begin{aligned} \mathcal{N}' : \mathcal{H} &\longrightarrow \mathbb{R}^d \\ \mathcal{X}_i = \sum_{j=1}^{d} x_{ij}\phi_j &\longrightarrow (x_j)_{j=1}^{d}, \end{aligned} \tag{2.15}$$

para $i \in \mathcal{I}$ siendo $\mathcal{I}$ un conjunto de indices arbitrario. Una vez fijada una base Φ_d, es posible caracterizar a la variable aleatoria funcional $\mathscr{X}$ en términos de un vector aleatorio cuyas observaciones son los coeficientes básicos de las observaciones de $\mathscr{X}$, y que está definido sobre la sigma-álgebra de Borel $\mathcal{B}^d$ de $\mathbb{R}^d$. Con esto, el uso del vector de coeficientes básicos $(x_{ij})_{j=1}^{d} \in \mathbb{R}^d$ en lugar de la función $\mathcal{X}_i, \forall i \in \mathcal{I}$ resulta conveniente en términos operativos. A esta forma de tratamiento de datos funcionales se le conoce como expansión en base o expansión básica (EB), y ha sido ampliamente trabajada como puede verse en Acal et al. (2022) para detectar cambios en la contaminación del aire durante la pandemia de COVID-19 mediante el uso de ANOVA funcional; en Aguilera et al. (2021), y también Aguilera et al. (2010) donde utilizan los coeficientes de la base para estimar la regresión PLS funcional, en Aguilera-Morillo & Aguilera (2020) para el problema de la clasificación multiclase de datos biomecánicos, o incluso, en el uso de componentes principales de la regresión logística funcional en el trabajo de Escabias et al. (2004), donde los autores tratan algunos aspectos del tema. De esta manera, dada una variable aleatoria funcional $\mathscr{X}$ definida en sobre una espacio $\mathcal{H}$ de dimensión finita d, y Φ una base de $\mathcal{H}$. Se tiene que la esperanza $\mathscr{X}$, denotada por $E_{\Phi}[\mathscr{X}]$ está definida por

$$E_{\Phi}[\mathscr{X}] = \sum_{j=1}^{d} E[X_j]\,\phi_j \tag{2.16}$$

donde $(X_j)_{j=1}^d$ es un vector aleatorio en $\mathbb{R}^d$ cuyas observaciones son los vectores de coeficientes básicos de las observaciones de $\mathcal{X}$.

2.3.1. Medidas muestrales para datos funcionales

En el FDA es bien sabido (ver por ejemplo Ramsay & Silverman (2005)) que desde una perspectiva muestral, es decir, dado un conjunto de observaciones funcionales asociadas a una V.A.F, es posible, al igual que en la estadística escalar clásica, definir ciertas medidas a partir de este conjunto de observaciones, que son tratadas como muestrales. En ese sentido, se tiene la siguiente definición.

Definición 2.3.2. *Dados los conjuntos de observaciones funcionales* $\{\mathcal{X}_i\}_{i=1}^n$ *y* $\{\mathcal{Y}_i\}_{i=1}^n$ *asociados a las V.A.Fs* $\mathscr{X}$ *e* $\mathscr{Y}$ *respectivamente, se tiene que las funciones*

$$\overline{\mathcal{X}} = n^{-1} \sum_{i=1}^{n} \mathcal{X}_i \tag{2.17}$$

y

$$\widehat{C}_{\mathcal{X}}(t,s) = (n-1)^{-1} \sum_{i=1}^{n} \left(\mathcal{X}_i(s) - \overline{\mathcal{X}}(s)\right)\left(\mathcal{X}_i(t) - \overline{\mathcal{X}}(t)\right), \tag{2.18}$$

llamadas media muestral y covarianza muestral, son estimadores de la media funcional y de la función covarianza de las ecuaciones 2.11 y 2.12 respectivamente. Además, el operador

$$\widehat{\mathcal{C}}_{\mathcal{X}}(t) = \int_{\mathfrak{T}} \widehat{C}_{\mathcal{X}}(t,s) f(s) ds, \tag{2.19}$$

denominado operador de covarianza muestral, es un estimador del operador covarianza de la Ecuación 2.14.

Adicional a las ya mencionadas, suele ser de interés las funciones

$$\mathrm{Var}(\mathcal{X})(t) = (n-1)^{-1} \sum_{i=1}^{n} (\mathcal{X}_i(t) - \overline{\mathcal{X}})^2, \tag{2.20}$$

llamada función de varianza muestral puntual, y

$$\mathrm{Cov}(\mathcal{X})(t) = (n-1)^{-1} \sum_{i=1}^{n} (\mathcal{X}_i(t) - \overline{\mathcal{X}}(t))(\mathcal{Y}_i(t) - \overline{\mathcal{Y}}(t)), \tag{2.21}$$

llamada función de covarianza muestral cruzada puntual, donde $\overline{\mathcal{Y}}$ corresponde a la media muestral de $\mathscr{Y}$.

Por otro lado, una herramienta necesaria para los desarrollos presentados en este trabajo es el análisis de componentes principales funcionales (FPCA por sus siglas en inglés), que es una extensión del análisis en componentes principales clásico (ver Cuadras (1996)) al caso de observaciones funcionales. El FPCA posee la misma motivación que el análisis de componentes principales para datos multivariantes y además es el fundamento de la representación de Karhumen-Loève, que proporciona una forma de representar las observaciones funcionales en términos de variables aleatorias no correladas y una sucesión de funciones. Así, dada una variable aleatoria funcional $\mathscr{X}$ (que se asume centrada sin perdida de generalidad) y dado un conjunto de observaciones funcionales asociadas $\{\mathcal{X}_i\}_{i=1}^n$. La $j-$ésima componente principal $\xi_j = (\xi_{1j}, \xi_{2j}, \ldots, \xi_{nj})'$ esta dada por

$$\xi_{ij} = \int_{\mathfrak{T}} \mathcal{X}_i(t)\mathscr{F}_j(t)dt,$$

donde la función $\mathscr{F}_j$ es la $j-$ésima autofunción, resultado de la solución del problema de optimización de la ecuación propia

$$\int_{\mathfrak{T}} \widehat{C}(t,s)\mathscr{F}_j(s)ds = \lambda_j \mathscr{F}_j, \tag{2.22}$$

siendo λ_j el $j-$ésimo valor propio asociado y $\widehat{C}$ la función de covarianza muestral de la Ecuación 2.18. Así, dados los conjuntos $\{\xi_1, \xi_2, \ldots, \xi_{n-1}\}$ de componentes principales, $\{\mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_{n-1}\}$ de autofunciones y $\{\lambda_1, \lambda_2, \ldots, \lambda_{n-1}\}$ de autovalores, se puede obtener la representación

$$\mathcal{X}_i = \sum_{j=1}^{n-1} \xi_{ij}\mathscr{F}_j; \quad i = 1, 2, \ldots, n \tag{2.23}$$

a partir de la expresión

$$\widehat{C}(t,s) = \sum_{j=1}^{n-1} \lambda_j \mathscr{F}_j(t)\mathscr{F}_j(s). \tag{2.24}$$

Capítulo 3

Varianza y correlación escalar para datos funcionales

3.1. Introducción y objetivos

Las medidas para datos funcionales presentados en el capítulo 2, son el resultado de la extensión natural de conceptos de la estadística escalar a las funciones, lo cual deriva, como es lógico, en objetos funcionales. Sin embargo, esto puede generar ciertos inconvenientes conceptuales en algunos casos. Para ilustrar esto, considere en primer lugar la esperanza de una variable aleatoria, como es muy conocido para datos escalares (ver por ejemplo Dodge (2008)) esta viene motivada como un indicador de tendencia, en este caso central, cuya interpretación natural se da en el mismo contexto de los datos con los cuales se obtiene su estimación. En el caso funcional esto no cambia, pues dada una variable aleatoria funcional $\mathscr{X}$ de imágenes en $\mathcal{H}$ con dominio en $\mathfrak{T}$, la esperanza de $\mathscr{X}$ dada por $\mu_{\mathscr{X}}$, es un elemento de $\mathcal{H}$ como se muestra en la Ecuación 2.11. Así, una estimación $\overline{\mathcal{X}}$ de $\mathscr{X}$ (ver Ecuación 2.17) es un elemento de $\mathcal{H}$ y posee una interpretación directa como una función de tendencia en este espacio. Además, su estimación cuenta con la propiedad de centralidad (ver Kenny & Keeping (1954) para el caso escalar), que puede describirse en su forma funcional como

$$\sum_{i=1}^{n}(\mathcal{X}_i - \overline{\mathcal{X}}) = 0_{\mathcal{H}}, \tag{3.1}$$

donde $0_{\mathcal{H}}$ hace referencia a la función nula en $\mathcal{H}$, siendo $\{\mathcal{X}_i\}_{i=1}^n$ un conjunto de observaciones de la variable $\mathscr{X}$.

Ahora bien, en cuanto al concepto de varianza dentro los datos escalares, este tiene la motivación inicial de servir como una medida de variabilidad (ver Cohn (2013)), que si bien no cuenta con una interpretación directa, sirve como herramienta para definir conceptos como la consistencia y eficiencia de un estimador en el sentido que lo expone Fisher (1925, 1922), pues esta es un valor real. Lamentablemente en el caso funcional, la utilización de la varianza como herramienta comparativa se pierde, pues aunque se cuenta con una definición del operador covarianza, como se muestra en la Ecuación 2.14, o una definición de una curva de varianza, como en la Ecuación 2.20; al ser ambos objetos funcionales, el concepto de una función "mínima" carece de un sentido directo ya que las funciones no forman un conjunto bien ordenado (Kaplansky 1977, Rudin 1976, Apostol 1967). Además, parece existir un problema conceptual asociado con el hecho de que la varianza misma proporcione una medida de la distancia de los valores respecto a la media.

Por otro lado, en cuanto al concepto de covarianza, se espera que sea un indicador de la variación conjunta de dos variables aleatorias (ver por ejemplo Mathai & Rathie (1977), Galton (1889)). Dicho concepto mantiene una estrecha relación con el de varianza, pues es esta a menudo presentada como un caso muy particular de la covarianza (ver Dodge (2008) para el caso escalar). Sin embargo, aunque se puede establecer una curva de covarianza, como en la Ecuación 2.21, esta solo provee una interpretación punto a punto, dando una idea de las regiones en donde existe una variación conjunta en mayor o menor medida, pero no es un indicador de la variación conjunta de dos variables aleatorias funcionales.

Dicho esto, en este capitulo se presenta una propuesta de varianza, covarianza y correlación escalar para datos funcionales, que pretende adicionar a las ya existentes en la literatura, unas estadísticas para datos funcionales con coherencia conceptual y ciertas ventajas interpretativas, definiendo la varianza, covarianza y correlación para datos funcionales como valores reales, haciendo hincapie en que el enfoque aquí mostrado se basa en la representación de una función continua dentro de un subespacio arbitrario $\mathcal{H}$ de dimensión finita arbitraria d de $L_2(\mathfrak{T})$ y, por lo tanto, los resultados obtenidos pueden aplicarse sin pérdida de generalidad a cualquier tipo de base ortonormal y de cualquier

dimensión finita.

Se presenta además, un estudio de simulación con el fin de mostrar evidencia de la consistencia e insesgadez de la varianza propuesta. Adicional a ello, se describen dos ejemplos de aplicación en la parte final de este capítulo. En el primer ejemplo, se implementa la propuesta de correlación para determinar si existe correlación entre las variables aleatorias funcionales del ejemplo clásico de Ramsay & Silverman (2005), sobre la temperatura media anual promedio de Canadá en cuatro de sus regiones, al construir el análisis de varianza funcional (FANOVA). Esto se hace porque el análisis no especifica la existencia de independencia entre las variables aleatorias funcionales. El segundo ejemplo describe el uso de nuestra propuesta de varianza y correlación funcional como parte de un análisis de datos funcionales y descriptivos sobre material particulado de dos estaciones de monitoreo de calidad del aire en Cali, Colombia. Todos los resultados presentados en este capítulo se encuentran publicados en Urbano-Leon et al. (2023).

3.2. Metodología

Definición 3.2.1. *Dadas $\mathscr{X} : \Omega_1 \to \mathcal{H}$ e $\mathscr{Y} : \Omega_2 \to \mathcal{H}$ dos variables aleatorias funcionales con observaciones de dominio $\mathfrak{T}$, definidas en el espacio de probabilidad $(\Omega_1, \mathfrak{S}_1, P_1)$ y $(\Omega_2, \mathfrak{S}_2, P_2)$ respectivamente, se definen las variables aleatorias reales $\mathfrak{M}_{\mathscr{X}}$ y $\mathfrak{M}_{\mathscr{X}\mathscr{Y}}$ en $(\Omega_1, \mathfrak{S}_1, P_1')$ y $(\Omega_1 \times \Omega_2, \mathfrak{S}, P)$ respectivamente, como*

$$\begin{aligned} \mathfrak{M}_{\mathscr{X}} : \Omega_1 &\longrightarrow \mathbb{R} \\ \omega_1 &\longrightarrow \mathscr{M}\left(\mathscr{X}\left(\omega_1\right) - \mathbf{E}_{\mathcal{H}}\left[\mathscr{X}\right]\right) \end{aligned} \tag{3.2}$$

$$\begin{aligned} \mathfrak{M}_{\mathscr{X}\mathscr{Y}} : \Omega_1 \times \Omega_2 &\longrightarrow \mathbb{R} \\ (\omega_1, \omega_2) &\longrightarrow \left\langle \mathscr{X}\left(\omega_1\right) - \mathbf{E}_{\mathcal{H}}\left[\mathscr{X}\right], \mathscr{Y}\left(\omega_2\right) - \mathbf{E}_{\mathcal{H}}\left[\mathscr{Y}\right] \right\rangle \end{aligned} \tag{3.3}$$

Note que las variables aleatorias definidas en las ecuaciones 3.2 y 3.3 son en realidad transformaciones de variables aleatorias funcionales (siempre que la definición tenga sentido)

Dado que las variables aleatorias $\mathfrak{M}_{\mathscr{X}}$ y $\mathfrak{M}_{\mathscr{X}\mathscr{Y}}$ son de valor real, es posible definir algunas medidas para datos funcionales a partir de estas como sigue.

Definición 3.2.2. *Sea $\mathscr{X} : \Omega \to \mathcal{H}$ una variable aleatoria funcional cuyas observaciones tienen soporte $\mathfrak{T}$. Una medida de dispersión para datos funcionales está dada por*

$$\mathscr{V}_{\mathscr{X}} = \mathbf{E}_{\mathbb{R}}\left[(\mathfrak{M}_{\mathscr{X}})^2\right], \tag{3.4}$$

Definición 3.2.3. *Dadas $\mathscr{X} : \Omega_1 \to \mathcal{H}$ y $\mathscr{Y} : \Omega_2 \to \mathcal{H}$ dos variables aleatorias funcionales con observaciones de dominio $\mathfrak{T}$, se define una medida de asociación entre $\mathscr{X}$ y $\mathscr{Y}$ como*

$$\mathscr{V}_{\mathscr{X}\mathscr{Y}} = \mathbf{E}_{\mathbb{R}}\left[(\mathfrak{M}_{\mathscr{X}\mathscr{Y}})\right], \tag{3.5}$$

Las expresiones de las ecuaciones 3.4 e 3.5, proporcionan medidas escalares a partir de variables aleatorias de valor real asociadas a variables aleatorias funcionales. Estas medidas pueden ser estimadas a partir de conjuntos de observaciones funcionales a manera de muestra, por lo que se les denomina "muestrales" y se definen a continuación.

3.2.1. Varianza y covarianza escalares muestrales

Dados dos conjuntos de n observaciones $\{\mathcal{X}_i\}_i^n$ e $\{\mathcal{Y}_i\}_i^n$ contenidos en $\mathcal{H}$ de soporte en $\mathfrak{T}$, de las variables aleatorias funcionales $\mathscr{X}$ y $\mathscr{Y}$ se tiene que

$$SVFD_{\mathscr{X}} = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathfrak{T}}\left(\mathcal{X}_i - \overline{\mathcal{X}}\right)^2(t)dt\right) \tag{3.6}$$

y

$$SVFD_{\mathscr{X}\mathscr{Y}} = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathfrak{T}}\left(\mathcal{X}_i(t) - \overline{\mathcal{X}}(t)\right)\left(\mathcal{Y}_i(t) - \overline{\mathcal{Y}}(t)\right)dt\right), \tag{3.7}$$

donde $\overline{\mathcal{X}}$ e $\overline{\mathcal{Y}}$ son las medias muestrales de $\mathscr{X}$ e $\mathscr{Y}$ respectivamente conforme a la Ecuación 2.17; proporcionan, respectivamente medidas muestrales de variabilidad y asociación conjunta para los datos funcionales. Note además que $SVFD_{\mathscr{X}\mathscr{X}} = SVFD_{\mathscr{X}}$.

Ahora bien, suponga sin perdida de generalidad, que el espacio $\mathcal{H}$ es de dimensión d, generado por una base $\Phi = \{\phi_j\}_{j=1}^d$, entonces para cada $i = 1, 2, \ldots, n$ se tiene que

$$\mathcal{X}_i = \sum_{j=1}^{d} x_{ij}\phi_j; \;\; \wedge \;\; \mathcal{Y}_i = \sum_{j=1}^{d} y_{ij}\phi_j, \tag{3.8}$$

de donde

$$
\begin{aligned}
SVFD_{\mathscr{X}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathfrak{T}}\left(\sum_{j=1}^{d}x_{ij}\phi_j(t)-\sum_{j=1}^{d}\overline{X}_j\phi_j(t)\right)^2 dt\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathfrak{T}}\left(\sum_{j=1}^{d}\left(x_{ij}\phi_j(t)-\overline{X}_j\phi_j(t)\right)\right)^2 dt\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\mathfrak{T}}\left(\sum_{j=1}^{d}\left(x_{ij}-\overline{X}_j\right)^2\phi_j^2(t)+\sum_{k,j=1,k\neq j}^{d}\left(x_{ij}-\overline{X}_j\right)\left(x_{ik}-\overline{X}_k\right)\phi_j(t)\phi_k(t)\right)dt\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}\left(x_{ij}-\overline{X}_j\right)^2\int_{\mathfrak{T}}\phi_j^2(t)+\sum_{k,j=1,k\neq j}^{d}\left(x_{ij}-\overline{X}_j\right)\left(x_{ik}-\overline{X}_k\right)\int_{\mathfrak{T}}\phi_j(t)\phi_k(t)dt\right)\\
&= \sum_{j=1}^{d}\sum_{i=1}^{n}n^{-1}\left(x_{ij}-\overline{X}_j\right)^2\|\phi_j\|^2+\sum_{k,j=1,\neq k}^{d}\sum_{i=1}^{n}n^{-1}\left(x_{ij}-\overline{X}_j\right)\left(x_{ik}-\overline{X}_k\right)\langle\phi_j,\phi_k\rangle.
\end{aligned}
$$

Note que $\sum_{i=1}^{n}n^{-1}\left(x_{ij}-\overline{X}_j\right)^2$ es una estimación $S_{X_j}^2$ de la varianza del coeficiente $j-$ésimo del vector de coeficientes básicos, y que además, $\sum_{i=1}^{n}n^{-1}\left(x_{ij}-\overline{X}_j\right)\left(x_{ik}-\overline{X}_k\right)$ es una estimación $S_{X_jX_k}$ de la covarianza entre las componentes $j-$ésima y $k-$ésima, siempre que $j\neq k$, por lo que se tiene que

$$
SVFD_{\mathscr{X}}=\sum_{j=1}^{d}S_{X_j}^2\|\phi_j\|^2+\sum_{j=1,\neq k}^{d}S_{X_jX_k}\langle\phi_j,\phi_k\rangle. \tag{3.9}
$$

De la Ecuación 3.9, se desprende que si, la base Φ es una base ortonormal, entonces

$$
SVFD_{\mathscr{X}}=\sum_{j=1}^{d}S_{X_j}^2. \tag{3.10}
$$

Es decir, la medida de variabilidad $\mathscr{V}_{\mathscr{X}}$ de la variable aleatoria funcional $\mathscr{X}$, puede ser aproximada por la suma de las varianzas de cada uno de los componentes del vector aleatorio cuyas observaciones son los coeficientes básicos de los datos funcionales, puesto que $\|\phi_j\|^2=1$ y $\langle\phi_j,\phi_k\rangle=0\ \forall j,k=1,2,\ldots,d;j\neq k$. De igual manera, se tiene que

$$
SVFD_{\mathscr{X}\mathscr{Y}}=\sum_{j=1}^{d}\sum_{i=1}^{n}n^{-1}\left(x_{ij}-\overline{X}_j\right)\left(y_{ij}-\overline{Y}_j\right)\|\phi_j\|^2+\sum_{j=1,\neq k}^{d}\sum_{i=1}^{n}n^{-1}\left(x_{ij}-\overline{X}_j\right)\left(y_{ik}-\overline{Y}_k\right)\langle\phi_j,\phi_k\rangle
$$

siendo

$$
SVFD_{\mathscr{X}\mathscr{Y}}=\sum_{j=1}^{d}S_{X_jY_j}. \tag{3.11}
$$

cuando la base Φ es ortonormal.

3.2.2. Propiedades

Se muestran ahora algunas propiedades de la varianza escalar para datos funcionales definida en la Ecuación 3.6, las cuales son heredadas de la varianza para datos escalares clásica.

Proposición 3.2.1. *Sea $\mathcal{H}$ un subespacio de dimensión finita d de $L_2(\mathfrak{T})$, generado por la base ortonormal $\Phi = \{\phi_j\}_{j=1}^d$ y sea $\{\mathcal{X}_i\}_{i=1}^n \subseteq \mathcal{H}$ un conjunto de n observaciones de la variable aleatoria funcional $\mathscr{X}$, cuyos vectores de coeficientes básicos $(x_{i1}, x_{i2}, \ldots, x_{id})'$, son observaciones de cierto vector aleatorio $X = (X_j)_{j=1}^d$, entonces se siguen las propiedades de la varianza escalar para datos funcionales*

1. *$SVFD_{\mathscr{X}} \geq 0$.*
2. *$SVFD_{\mathscr{X}}$ es de valor mínimo bajo Φ.*
3. *Si $\{\mathcal{X}_i\}_{i=1}^n$ tienen el mismo vector de representación, entonces $SVFD(\mathcal{X}) = 0$.*

Demostración 3.2.1.

1. *Dado que $SVFD_{\mathscr{X}} = \sum_{j=1}^d S_{X_j}^2$ y que $S_{X_j}^2 \geq 0$ para cada $j = 1, 2, \ldots, d$, $SVFD_{\mathscr{X}} \geq 0$.*
2. *Dado que $SVFD_{\mathscr{X}} = \sum_{j=1}^d S_{X_j}^2$ y que cada S_{X_j} es de valor mínimo para cada $j = 1, 2, \ldots, d$, entonces $SVFD_{\mathscr{X}}$ es de valor mínimo.*
3. *Dado un conjunto $\{\mathcal{X}_i\}_{i=1}^n$ de datos funcionales, tales que sus vectores de representación son iguales, entonces*

$$\begin{aligned} x_{11} = x_{21} = x_{31}, \ldots, x_{n1} &= v_1 \\ x_{12} = x_{22} = x_{32}, \ldots, x_{n2} &= v_2 \\ &\vdots \quad \vdots \\ x_{1d} = x_{2d} = x_{3d}, \ldots, x_{nd} &= v_d. \end{aligned}$$

Por lo tanto, para todo $1 \leq j \leq d$, $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} = \frac{1}{n}\sum_{i=1}^{n} v_j = v_j$ *y*

$$S^2_{X_j} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i,j} - \overline{X}_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(v_j - v_j\right)^2 = 0.$$

Por lo tanto, $SVFD_{\mathscr{X}} = 0$.

El cumplimiento de esta última propiedad muestra que, en realidad, $SVFD_{\mathscr{X}}$ mide la dispersión de los datos funcionales.

Siguiendo esta misma linea de razonamiento, puede mostrarse que $\mathfrak{S}_{\mathscr{X}\mathscr{Y}}$ hereda propiedades clásicas de la covarianza de variables aleatorias reales. Más aún, es posible definir un coeficiente de correlación para datos funcionales como se muestra a continuación.

Definición 3.2.4. *Sean* $\{\mathcal{X}_i\}_{i=1}^n, \{\mathcal{Y}_i\}_{i=1}^n \subseteq \mathcal{H}$ *dos conjuntos de observaciones con soporte en* $\mathfrak{T}$ *de las variables aleatorias funcionales* $\mathscr{X}$ *e* $\mathscr{Y}$ *respectivamente, entonces el coeficiente de correlación escalar para los datos funcionales está definido por la expresión.*

$$\rho_{\mathscr{X}\mathscr{Y}} = \frac{SVFD_{\mathscr{X}\mathscr{Y}}}{\sqrt{SVFD_{\mathscr{X}}\, SVFD_{\mathscr{Y}}}}. \tag{3.12}$$

Note que la expresión de la Ecuación 3.12, produce un valor real en el intervalo $[-1, 1]$. Además, el cálculo de la correlación se puede realizar con la expresión

$$\rho_{\mathscr{X},\mathscr{Y}} = \frac{\sum_{j=1}^{d} S_{X_jY_j}}{\sqrt{\left(\sum_{j=1}^{d} S^2_{X_j}\right)\left(\sum_{j=1}^{d} S^2_{Y_j}\right)}}.$$

3.3. Resultados bajo simulación

Con el fin de mostrar que las medidas escalares aquí propuestas son en realidad una medida de la variabilidad de los conjuntos de curvas, se lleva a cabo un estudio de simulación a partir de una variedad de conjuntos de datos funcionales construidos dentro de un subespacio $\mathcal{H}$ de dimensión $d = 10$, generado por la base ortonormal Φ_1 definida en la Ecuación 2.8. En cada caso, se utilizan las expresiones de las ecuaciones 3.6, 3.7 y 3.12 para obtener las medidas estimadas de varianza, covarianza y correlación para los datos funcionales. Los casos de simulación provienen de dos funciones de diferentes tipos. El primer tipo, al que se le denomina Tipo A, es una función constante, mientras que el

segundo tipo, llamado Tipo B, es una función no constante, como se muestra en la Figura 3.1.

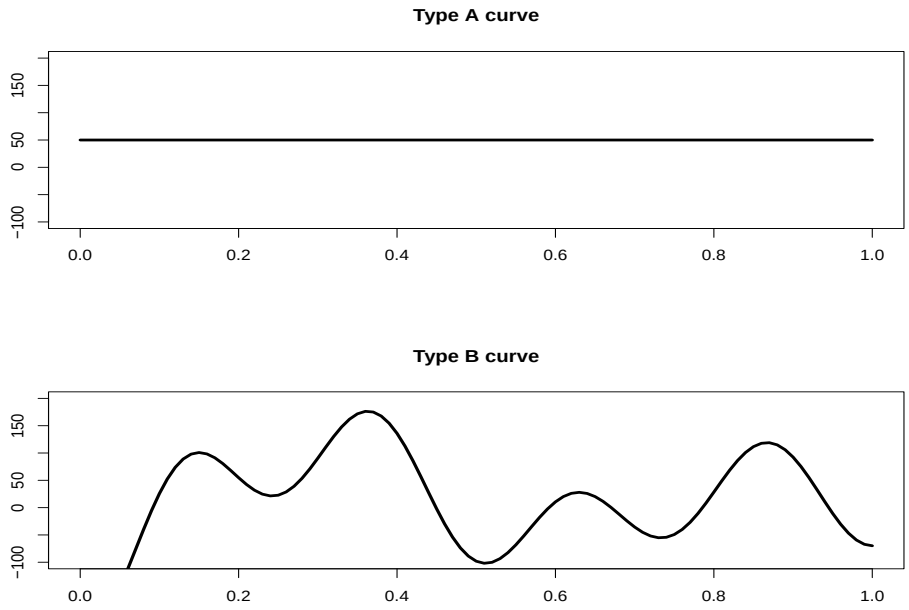


Figura 3.1: Curvas simuladas Tipo A (**superior**) y Tipo B (**inferior**).

En el primer caso de simulación, se consideran tres escenarios diferentes de dispersión constante a lo largo del soporte de las observaciones funcionales: Caso 1.1, considerado como de alta dispersión, Caso 1.2 de dispersión moderada y Caso 1.3 de baja dispersión. La Figura 3.2 muestra claramente la dispersión simulada en las curvas y cómo la medida de variabilidad propuesta captura las disminuciones y aumentos en la variabilidad de las curvas en cada escenario.

Como Caso 2, se construyen ahora curvas con una dispersión no uniforme en el dominio, y consideramos, como en el primer caso, tres escenarios diferentes: Caso 2.1 de alta dispersión, Caso 2.2 de dispersión moderada y Caso 2.3 de baja dispersión. Las curvas resultantes en este Caso 2 son más erráticas que las del Caso 1 en todos los escenarios. Sin embargo, la Figura 3.3 muestra como, nuevamente, la medida de variabilidad propuesta logra capturar esta dispersión entre curvas, disminuyendo a medida que disminuye la dispersión.

Se procede ahora a ilustrar que esta medida de variabilidad es consistente. Para ello, se construye un conjunto de datos funcionales con 500 funciones tipo B, a manera de población, y se obtienen muestras de diferentes tamaños a partir de esta, luego, se obtiene la diferencia absoluta entre las varianzas de la muestra y la población. Este proceso es replicado 500 veces para cada tamaño de muestra. En la Tabla 3.1, se reporta la media de los resultados obtenidos en cada tamaño de muestra asumido. Mientras que en la Figura 3.4, se muestra que la varianza propuesta converge en valor medio a la varianza poblacional a medida que aumenta el tamaño de la muestra.

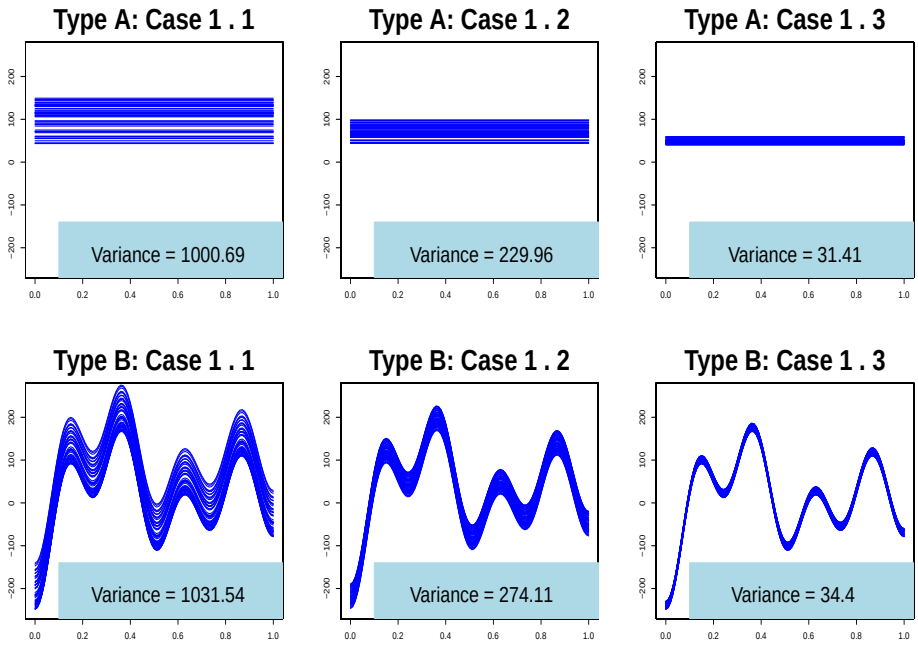


Figura 3.2: Variabilidad de los escenarios considerados para las curvas simuladas en el Caso 1 y varianzas escalares obtenidas. Curvas tipo A (Superior) y tipo B (Inferior).

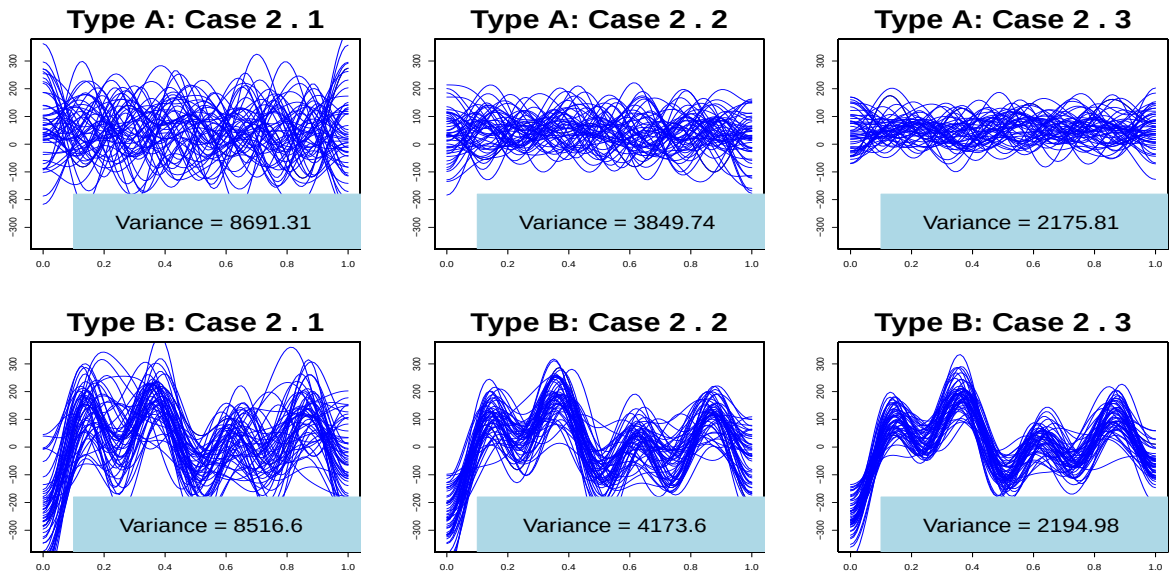


Figura 3.3: Escenarios de variabilidad considerados para curvas simuladas en el Caso 2 y varianzas escalares obtenidas. Curvas Tipo A (**superior**) y tipo B (**inferior**).

Tamaño de Muestra	**Media de las diferencias absolutas**
5	1828.16
10	926.43
20	403.40
50	178.40
100	81.47
150	34.31
200	23.74
250	15.29
300	13.6
400	3.74
450	3.83
490	0.45

Tabla 3.1: Consistencia

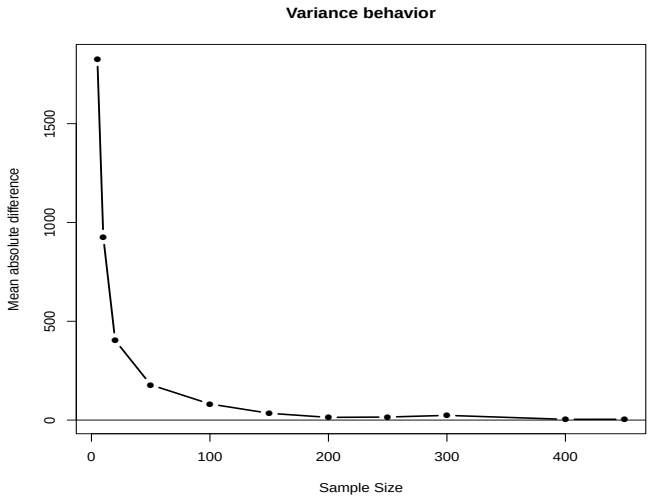


Figura 3.4: Comportamiento de la media de las diferencias absolutas cuando aumenta el tamaño de la muestra.

3.4. Ejemplos con datos reales

Con el fin de ilustrar el uso y las ventajas interpretativas de las medidas propuestas, se presenta a continuación dos ejemplos de aplicación con datos reales.

3.4.1. Variabilidad en curvas de temperatura

En este ejemplo, se presenta una forma de utilizar las medidas propuestas en este capítulo, en el contexto de un análisis funcional de varianza. Para ello, se toma el ejemplo clásico en datos funcionales de la información recopilada por Ramsay & Silverman (2005), correspondientes a la temperatura media mensual de 35 estaciones meteorológicas en Canadá. Estas estaciones están clasificadas en cuatro regiones según su ubicación: Atlántico, Continental, Pacífico y Ártico. En Ramsay & Silverman (2005), los autores implementan un análisis de la varianza funcional para evaluar la existencia de diferencias estadísticamente significativas entre la temperatura media anual en las cuatro áreas geográficas. Sin embargo, debido a la dinámica del fenómeno, podría existir una correlación en la temperatura de las cuatro áreas; por lo que las conclusiones de FANOVA podrían

verse afectadas. No obstante, esto no es considerado en el ejemplo clásico, puesto que la correlación cruzada funcional no permite concluir si existe correlación entre los datos funcionales de las cuatro áreas de manera general, pues esta constituye una superficie como se indica en la Figura 3.5. En contraste, la medida de correlación propuesta si permite determinarlo, como se muestra en la Tabla 3.2, en donde se evidencia una fuerte correlación entre las variables funcionales de temperatura ártica y continental. Además, la variable funcional temperatura Pacífico y Continental presenta correlación negativa, lo que indica que cuando hay aumento de temperatura en la zona del Pacífico, hay una disminución de temperatura en la zona Continental. En resumen, la matriz de correlaciones propuesta muestra que la temperatura en las cuatro zonas está correlacionada.

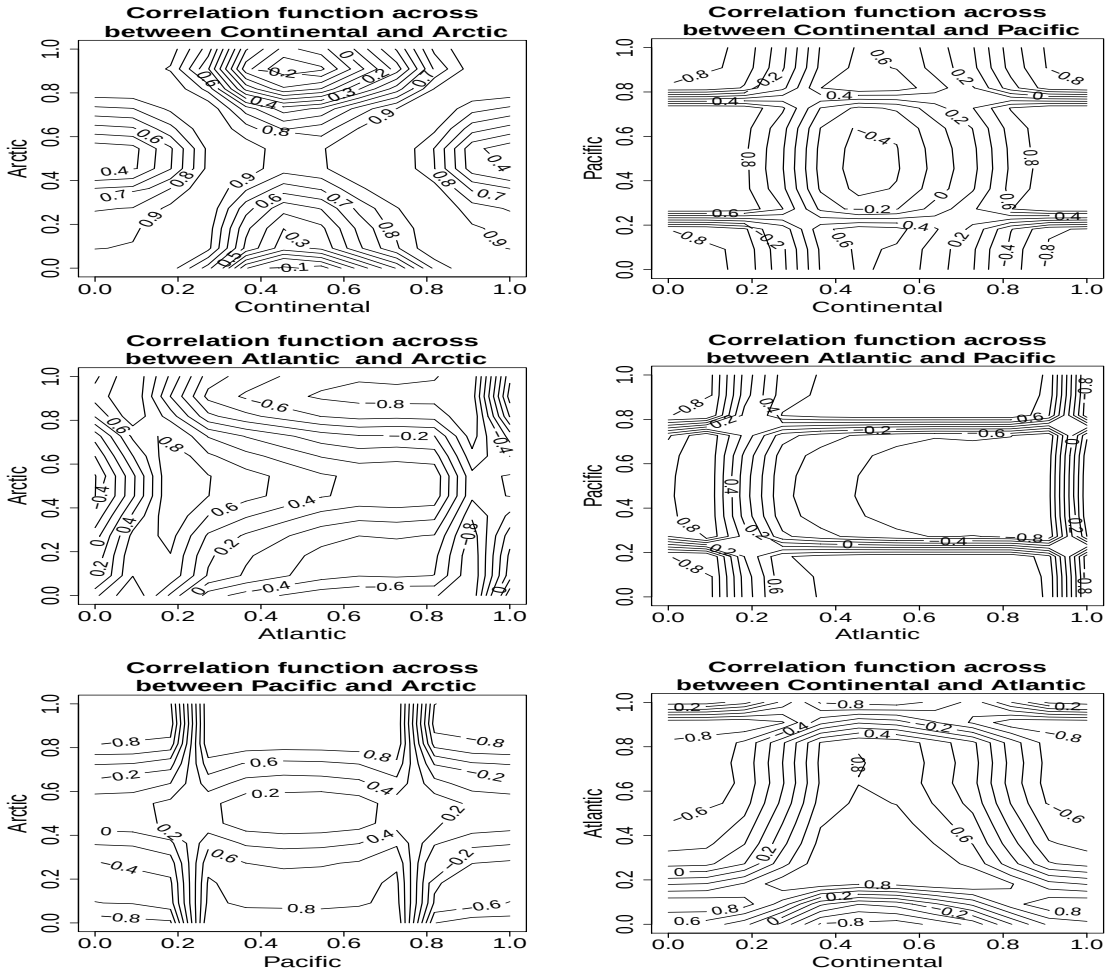


Figura 3.5: Correlaciones cruzadas de la variable funcional temperatura entre las áreas geográficas.

Zona	Altlantico	Continental	Pacifico	Arctico
Atlántico	1.00	0.45	−0.50	0.26
Continental	0.45	1.00	−0.62	0.82
Pacifico	−0.50	−0.62	1.00	−0.22
Ártico	0.26	0.82	−0.22	1.00

Tabla 3.2: Medidas de correlación con nueva propuesta

3.4.2. Variabilidad en curvas de contaminación

Se proporciona ahora un ejemplo de implementación sobre datos reales de contaminación en el aire, con el fin de mostrar las ventajas interpretativas de las estadísticas resumidas sugeridas.

Según World Health Organization (2005), las partículas suspendidas en el aire cuyo diámetro aerodinámico es inferior a 2.5μ ($PM_{2.5}$), son considerados contaminantes criterio, debido a que la exposición prolongada es perjudicial para la salud humana. En consecuencia, en muchas ciudades se han implementado estaciones de medición de $PM_{2.5}$ con el fin de monitorear la actividad diaria del contaminante. Como resultado, se ha logrado identificar una relación entre las actividades antropogénicas y la producción de $PM_{2.5}$ dentro de un área urbana. Por ejemplo, en la ciudad de Cali, Colombia, la entidad encargada del monitoreo de este contaminate es el Departamento Administrativo de Gestión Ambiental (DAGMA). Este utiliza dos estaciones de monitoreo de la calidad del aire que forman parte del sistema de vigilancia de la calidad del aire de la ciudad. Las estaciones reciben el nombre de *Base Aérea* (BA) y *Compartir* (CO), por el nombre de los barrios en donde se encuentran localizadas. Estas estaciones están ubicadas aproximadamente a 3,5 km de distancia entre sí y recopilan periódicamente registros de concentración de $PM_{2,5}$ cada 10 segundos pero solo el promedio por hora es reportado, por lo que solo es posible recoger como máximo veinticuatro mediciones diarias en cada estación. Para este ejemplo, se toman 29 días del año 2015 con sus veinticuatro mediciones en ambas estaciones de monitoreo. Para fines ilustrativos se construyen 58 curvas pareadas,

29 por estación, en el subespacio $\mathcal{H}$ de dimensión 8 generado por la base Φ_1 de la Ecuación 2.8. En la Figura 3.6, se muestran las curvas de los datos funcionales de las estaciones BA y CO.

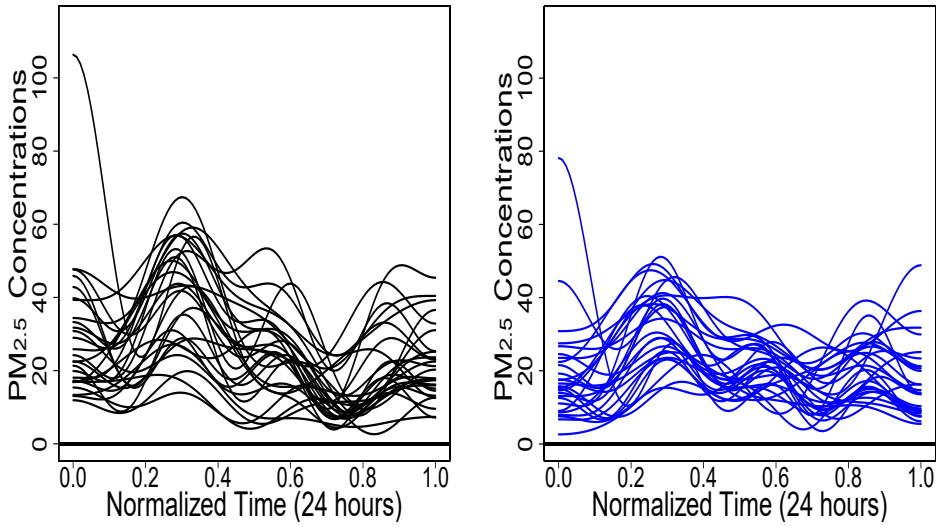


Figura 3.6: Curvas de $PM_{2.5}$ de la estación BA en negro (izquierda) y de la estación CO en Azul (derecha).

A su vez, la Figura 3.7 muestra las medias funcionales y las curvas de varianzas sugeridas por Ramsay & Silverman (2005). Aquí, es posible observar que una curva de varianza es insuficiente para decidir cuál de las dos estaciones experimenta una mayor variabilidad.

La Figura 3.8 muestra una curva de covarianzas punto a punto según la Ecuación 2.21, del mismo modo una curva de correlaciones puntuales, que resultan insuficientes para decidir si las variables funcionales están correlacionadas y en qué medida. Mientras que, las medidas escalares para datos funcionales propuestas, mostrada en la Tabla 3.3, permiten observar que la estación BA presenta más variabilidad. Esto se puede explicar por la elevada actividad industrial, el tráfico aéreo y terrestre en el sector. Además, la alta correlación entre las estaciones BA y CO, puede explicarse por su proximidad y sus datos pareados.

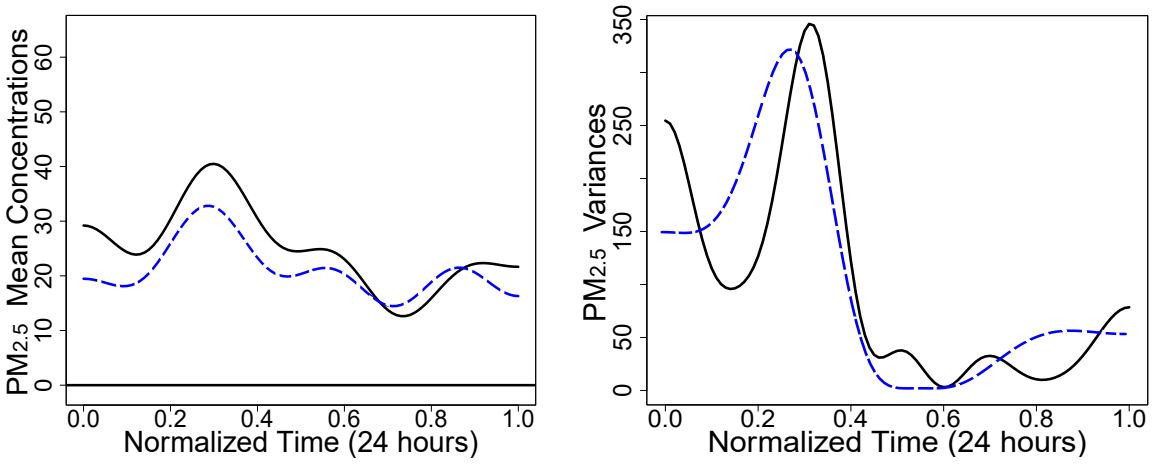


Figura 3.7: Para ambas figuras, en azul estación CO, en negro estación BA. Izquierda: Medias funcionales de la concentración de $PM_{2.5}$. Derecha: Función de varianza de la concentración de $PM_{2.5}$.

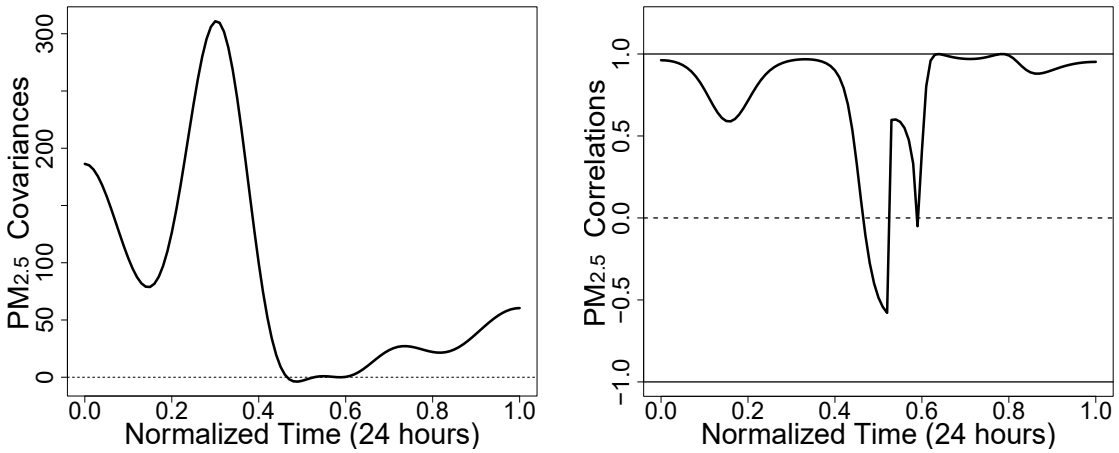


Figura 3.8: Curva de covarianzas (izquierda). Curva de Correlaciones (derecha).

Estación	**Variabilidad**	**Covarianza**	**Correlación**
Base Aérea	134.91	72.31	0.95
Compartir	76.24		

Tabla 3.3: Medidas propuestas para datos funcionales.

3.5. Conclusiones

Hasta el momento, al respecto de las medidas escalares de varianza, covarianza y correlación para datos funcionales, se puede concluir que:

- Mantienen coherencia con los conceptos de dispersión y asociación, lo que resulta en una ventaja interpretativa.
- La varianza propuesta permite comparar la variabilidad de dos grupos de datos funcionales, de tal manera que sea fácil de calcular e interpretar en términos de la cantidad.
- Permiten conocer qué tan fuerte es la variabilidad conjunta de dos conjuntos de datos funcionales, con facilidad interpretativa y operativa.
- La varianza aquí definida, aplicada a un conjunto de estimaciones, permite evaluar la consistencia y la eficiencia de un estimador funcional con respecto a otros.

Para utilizar las medidas propuestas, es necesario que los datos funcionales puedan representarse mediante una base funcional, lo que limita su uso únicamente a datos funcionales que tengan esta característica. Además, todos los datos funcionales deben representarse con la misma base y el mismo número de funciones.

Capítulo 4

Comparación en datos funcionales pareados

4.1. Introducción y objetivos

Dadas dos funciones de valor real y continuas f y g del mismo espacio funcional sobre el mismo dominio $\mathfrak{T}$, la comparación directa entre estas dos funciones es un problema matemático ya resuelto; pues debido a que las funciones no constituyen un conjunto ordenado (bajo el orden usual), solo es posible determinar si el par de funciones en cuestión son iguales o no por medio de sus imágenes, es decir, $f = g \Leftrightarrow f(x) = g(x) \forall x \in \mathfrak{T}$. No obstante, la comparación estocástica de datos funcionales hace referencia no a una comparación matemática directa entre las curvas o sus imágenes, sino que se ocupa de la comparación de características de uno o varios conjuntos de curvas, teniendo en cuenta la variabilidad inherente de los procesos que las generan, tanto a nivel individual como en conjunto. En el caso de datos escalares, uno de los métodos más utilizados para la comparación estocástica es el contraste de hipótesis, que permite, a partir de la información obtenida de un determinado fenómeno, la confrontación de una colección de dos o más suposiciones establecidas sobre una o más características de un conjunto de parámetros o sobre varios conjuntos de datos. En el contexto de los datos funcionales, el contraste de hipótesis posee la misma motivación que en el caso escalar. Debido a ello, una hipótesis inicial H_0 es contrastada con otra hipótesis, generalmente complementaria, sobre el mismo parámetro denominada hipótesis alternativa y denotada como H_1 Kenny & Keeping

(1951). Los contrastes de hipótesis son el fundamento de la comparación estocástica, por esta razón, han sido ampliamente estudiados en la literatura dentro de diferentes metodologías como por ejemplo en el análisis funcional de la varianza (FANOVA) en los trabajos de Ramsay & Silverman (2005), Cuevas & Fraiman (2004), Cuesta-Albertos & Febrero-Bande (2010), Acal & Aguilera (2023), o en bondad de ajuste como en el trabajo de Cuesta-Albertos et al. (2007, 2017), también para test de significación en Fan & Lin (1998) o en Shen & Faraway (2004) para una prueba F.

Dentro de los contrastes de hipótesis, el contraste de igualdad de medias funcionales $\mu_{\mathscr{X}}$ y $\mu_{\mathscr{Y}}$, que puede ser planteado como

$$\begin{aligned} H_0 : \quad & \mu_{\mathscr{X}} = \mu_{\mathscr{Y}} \\ H_1 : \quad & \mu_{\mathscr{X}} \neq \mu_{\mathscr{Y}}, \end{aligned} \tag{4.1}$$

ha sido ampliamente estudiado en la literatura por lo que se puede encontrar diferentes enfoques y metodologías. En Degras (2014), se utilizan bandas de confianza, en Cox & Follen (2015) los autores centran su estudio en el caso en donde la dimensión del espacio al que pertenecen los datos funcionales es mucho mayor que el tamaño de la muestra. En Horvath et al. (2013) abordan el problema desde los componentes principales. En Zhang & Chen (2007) los autores muestran algunos procedimientos para probar que, bajo normalidad, la función media es igual en dos grupos de curvas utilizando una prueba$-t$ punto a punto y posteriormente realizan una comparación basada en la norma de L_2. En Jiang et al. (2019) se aborda el problema de comparación de dos muestras de observaciones funcionales desde un punto de vista no paramétrico y se trabaja bajo el supuesto de independencia entre los grupos de curvas a comparar pero permitiendo un error funcional, utilizando funciones características empíricas. En Pomann et al. (2016), siguiendo la misma linea de Horvath & Rice (2015), los autores se centran en el problema de probar que dos grupos de datos funcionales poseen la misma distribución asumiendo la independencia entre muestras y considerando cada muestra proveniente de una marginal de un proceso de mixtura entre ambos procesos; entre otros. En todos los casos mencionados hasta el momento, se asume de una u otra manera la independencia entre las observaciones. No obstante, dentro de la investigación aplicada es posible encontrar diseños en donde una

misma característica es medida de manera repetida en dos ocasiones, dando lugar a lo que se conoce como un diseño pareado, caracterizado principalmente porque las observaciones se encuentran emparejadas. Así, tanto en el caso escalar como en el funcional, esta paridad entre las mediciones puede generar cierto grado de dependencia entre las observaciones funcionales. Esto produce que el supuesto de independencia no pueda ser asumido a priori, afectando significativamente los resultados en un contraste de hipótesis. En ese sentido, ya que en el trabajo de Cuevas & Fraiman (2004) sobre proyecciones aleatorias, los autores demuestran que el problema de probar diferencia de medias funcionales puede ser equivalente a probar diferencias de medias funcionales sobre proyecciones aleatoria en un espacio de dimensión menor, en el trabajo de Melendez et al. (2021) se valen de este hecho para extender la prueba de comparación de Wilcoxon al contexto funcional. Los autores reducen los datos funcionales, considerados como observaciones de variables aleatorias funcionales, a valores proyectados sobre un espacio de dimensión uno, por medio de

$$p = \int_T X(t)W(t)dt,$$

donde $W(t)$ un movimiento browniano. Dado que estos valores proyectados son, valores escalares, el problema se reduce a un caso de comparación univariante. Así, ordenando el valor absoluto de estos nuevos datos escalares, se conduce un test de wilcoxon escalar clásico.

En este capítulo, se aporta una propuesta para la comparación de medias de datos funcionales pareados. El contraste provee un criterio de decisión general para saber si existen o no diferencias significativas entre las medias de las mediciones funcionales pareadas. La metodología aquí expuesta, así como el ejemplo de aplicación, se encuentran publicados en Urbano-Leon & Escabias (2022).

4.2. Metodología

Dadas $\mathscr{X}$ e $\mathscr{Y}$ dos variables aleatorias funcionales con valores sobre el mismo subespacio $\mathcal{H}$. El interés se centra en presentar una propuesta de contraste de hipótesis

de igualdad de medias pareadas, partiendo de las hipótesis descritas como

$$\begin{aligned} H_0: \quad \mu_{\mathscr{X}} &= \mu_{\mathscr{Y}} \\ H_1: \quad \mu_{\mathscr{X}} &= \mu_{\mathscr{Y}}, \end{aligned} \tag{4.2}$$

en un contexto en donde a partir de $\mathscr{X}$ e $\mathscr{Y}$ se obtiene el conjunto de observaciones por parejas $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^n$, y que está constituido por las observaciones $\{\mathcal{X}_i\}_{i=1}^n$ y $\{\mathcal{Y}_i\}_{i=1}^n$ de $\mathscr{X}$ y $\mathscr{Y}$ respectivamente.

En este caso, se plantea que la $i-$ésima pareja de observaciones $(\mathcal{X}_i, \mathcal{Y}_i)$ podría poseer cierta similaridad en los recorridos de las observaciones de sus componentes, causada por el fenómeno de paridad subyacente. Esta similaridad, podría entorpecer el proceso de juzgar, por medio de los conjuntos de observaciones funcionales, si existe diferencias significativas entre las medias de ambas variables. Con eso en mente, la hipótesis H_0 puede ser reformulada en términos de la variable aleatoria $\mathscr{D} = \mathscr{X} - \mathscr{Y}$ como

$$\begin{aligned} H_0 : \mu_{\mathscr{X}} - \mu_{\mathscr{Y}} = 0_{\mathcal{H}} \quad &\longrightarrow \quad H_0 : \mu_{\mathscr{D}} = 0_{\mathcal{H}} \\ H_1 : \mu_{\mathscr{X}} - \mu_{\mathscr{Y}} \neq 0_{\mathcal{H}} \quad &\longrightarrow \quad H_0 : \mu_{\mathscr{D}} \neq 0_{\mathcal{H}}, \end{aligned} \tag{4.3}$$

donde $0_{\mathcal{H}}$ es la función nula en el subespacio $\mathcal{H}$, siendo el objetivo ahora, contrastar la similitud de la función $\mu_{\mathscr{D}}$ con la función nula en $\mathcal{H}$. Para ello, dado que $\mu_{\mathscr{D}} \in \mathcal{H}$, si la hipótesis nula se asume verdadera, es posible plantear un contraste alternativo a partir de la integral como

$$\begin{aligned} H_0': \quad \int_{\mathfrak{T}} \mu_{\mathscr{D}}(t)dt \quad &= 0 \\ H_1': \quad \int_{\mathfrak{T}} \mu_{\mathscr{D}}(t)dt \quad &\neq 0. \end{aligned} \tag{4.4}$$

Ahora bien, para el contraste de hipótesis de la Ecuación (4.4), se presenta el estadístico de prueba basado en la media de la integral de las diferencias (MID) de la Ecuación (4.5), definido a partir de dos conjuntos de observaciones funcionales de naturaleza pareada $\{\mathcal{X}_i\}_{i=1}^n$ y $\{\mathcal{Y}_i\}_{i=1}^n$, de las respectivas variables aleatorias funcionales $\mathscr{X}$ y $\mathscr{Y}$.

$$MID = n^{-1} \sum_{i=1}^{n} \int_{\mathfrak{T}} \mathcal{D}_i(t)dt, \tag{4.5}$$

donde $\mathcal{D}_i = \mathcal{X}_i - \mathcal{Y}_i$ es la diferencia de las $i-$ésimas observaciones de $\mathscr{X}$ e $\mathscr{Y}$, para cada $i = 1, 2, \ldots, n$.

Note que el estadístico propuesto *MID*, sigue una distribución normal de media cero y varianza σ^2, lo cual está garantizado por el teorema central del límite. Por lo tanto, el contraste puede ser realizado a partir de una estandarización como

$$S.MID = \frac{MID - 0}{\sigma}, \tag{4.6}$$

donde, σ puede ser obtenido desde la expresión

$$S_d = \frac{\sigma}{\sqrt{n}}, \tag{4.7}$$

siendo S_d^2 la varianza muestral obtenida desde las observaciones $\left\{\int_{\mathfrak{T}} \mathcal{D}_i(t)dt\right\}_{i=1}^{n}$. Así, un valor-p escalar puede ser obtenido como $2P(Z \geq |S.MID|)$, donde Z es una variable aleatoria real, tal que $Z \sim N(0, 1)$.

4.3. Resultados

Durante la emergencia mundial por COVID-19, el gobierno de Colombia tomó la tasa de positividad como una variable clave para la toma de decisiones tempranas relacionadas con el manejo de la enfermedad. En este ejemplo, se toma la tasa de positividad diaria, definida como el porcentaje diario de personas diagnosticadas positivas para COVID-19 en torno al número total de personas examinadas. Según Dallal et al. (2021) y Fu et al. (2021), la tasa de positividad ayuda a determinar la presencia de olas de contagio, por lo que se asume que los casos confirmados de coronavirus en Colombia en un año sirven para identificar los momentos más críticos de un cambio de tendencia en ese período. Dado que la información discriminada por departamentos en Colombia solo se informó a partir de julio de 2020, tenemos información utilizable a partir del 19 de julio de 2020; podemos observar varias olas de contagio, pero centramos nuestra atención en dos de ellas, representadas en la Figura 4.1.

En el año 2021, Colombia fue testigo de fuertes protestas. La gente salió a las calles debido a múltiples factores sociales, lo que impidió el comportamiento de aislamiento y

otras medidas de precaución contra el COVID-19. Este hecho puede haber prolongado la duración de la ola de contagios, así como su magnitud, ya que las protestas comenzaron el 28 de abril de 2021 y duraron más de dos meses Restrepo (2022).

Debido a esto, consideramos dos estudios de caso. En el primer caso, asumimos que las dos olas de contagio tienen la misma duración. Sin embargo, en el segundo caso, consideramos que ambas olas tienen una duración diferente. El propósito de este estudio es determinar si existen diferencias significativas entre la primera y la segunda ola en cada estudio de caso por separado. Dado que los datos son continuos y además están emparejados, utilizamos la metodología de FDA con una prueba t funcional puntual, seguida de una propuesta de prueba de hipótesis basada en la integral de la diferencia en las curvas de tasa de positividad, para probar la igualdad de medias funcionales.

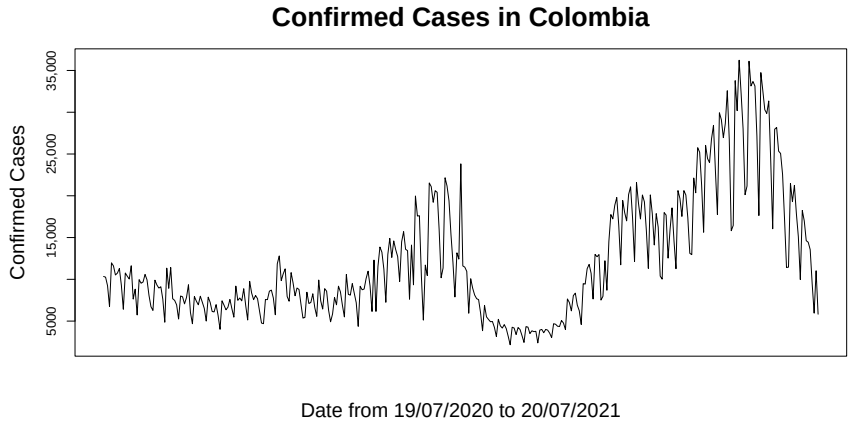


Figura 4.1: Dos olas de contagio por COVID-19 entre las fechas 19 de julio de 2021 y 20 de julio de 2022.

4.3.1. Acerca de los datos de COVID-19 en Colombia

En Colombia, los datos oficiales sobre COVID-19 son reportados por el instituto nacional de salud del ministerio de salud. No obstante, Colombia es un país dividido en 32 regiones llamadas departamentos y un distrito especial, que corresponde a la capital del país (Bogotá D.C.). Cada una de estas regiones reporta los casos positivos y demás variables de interés por separado al ministerio de salud, sin contar con un sistema adecuado para ello. Este inconveniente generó grandes problemas en el cruce

de información, provocando muchos casos de información incompleta. Por esta razón, en este estudio solo se considera los departamentes con información completa desde el 20 julio de 2020, al 20 julio de 2021, resultando en 23 departamentos y Bogotá D.C., para un total de 24 regiones. Estas son: Antioquia, Atlántico, Bolivar, Boyaca, Cáldas, Casanare, Cauca, Cesar, Córdoba, Cundinamarca, Guajira, Huila, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Tolima, Valle del Cauca, y Bogotá D.C.

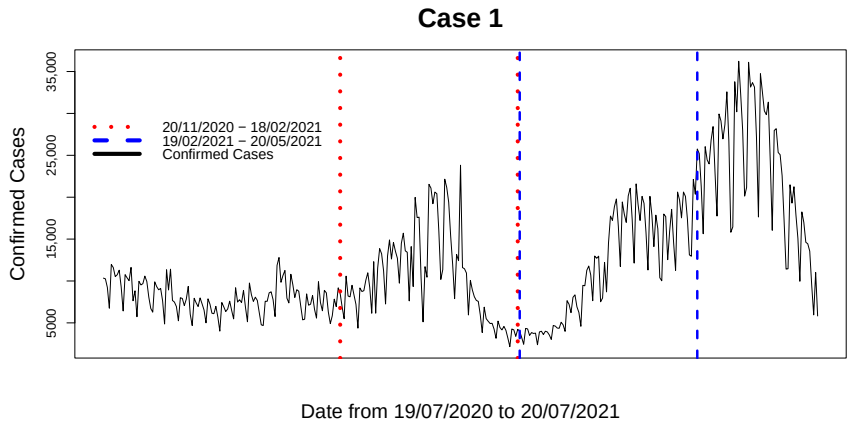


Figura 4.2: Caso 1: Dos olas de contagio por COVID-19 en Colombia sin paro nacional. Las líneas rojas punteadas marcan el inicio y el final de la primera ola de contagio, ocurrida entre el 20 y noviembre de 2020 y el 18 de febrero de 2021. Las líneas azules discontinuas marcan el inicio y el final de la segunda ola de contagio, ocurrida entre el 19 de febrero de 2021 y 20 de mayo de 2021.

Se sospecha que el paro nacional ocurrido en Colombia el 28 de abril de 2021, y que duró al menos dos meses Restrepo (2022), provocó que las personas no tomaran medidas de protección personal contra el COVID-19 y que esto alargó la duración de la ola de contagio por COVID-19 en Colombia. Por esta razón, se consideran dos escenarios de estudio. El primero, denominado Caso 1, asume las mediciones para la primera ola de contagio desde el 20 de noviembre de 2020, hasta el 18 de febrero de 2021; y una segunda ola de contagios del 19 de febrero de 2021 al 20 de mayo de 2021. Es decir, ambas olas de contagio duran tres meses cada una, desconociendo el efecto del paro nacional. El segundo escenario, denominado Caso 2, supone la primera ola de contagio del 20 de noviembre de

2020 al 18 de febrero de 2021, y la segunda ola de contagio del 19 de febrero de 2021 al 20 julio de 2021. En otras palabras, en el segundo escenario , la primera ola dura tres meses y la segunda ola dura 5 meses. Ver Figuras 4.2 y 4.3.

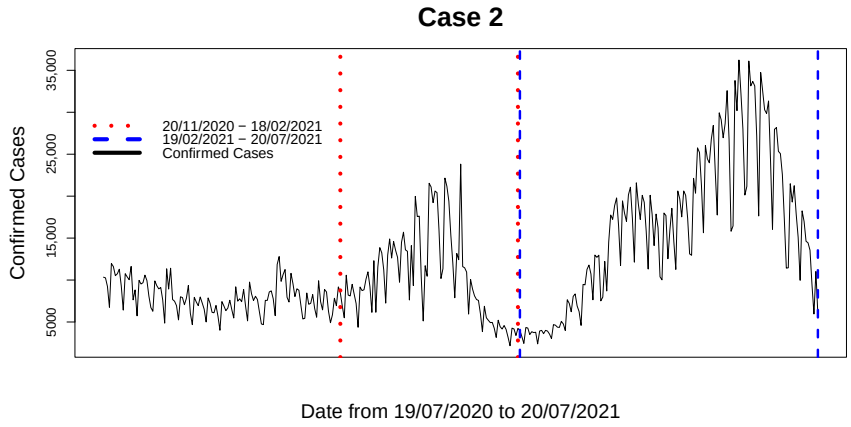


Figura 4.3: Caso 2: Dos olas de contagio por COVID-19 en Colombia con paro nacional. Las líneas rojas punteadas marcan el inicio y el final de la primera ola de contagio, que ocurrió entre el 20 de noviembre de 2020 y el 18 de febrero de 2021. Las líneas azules discontinuas marcan el inicio y el final de la segunda ola de contagio, que ocurrió entre 19 de febrero de 2021 y 20 de julio de 2021.

Es importante notar que, para cada caso, los datos de la primera y segunda ola de contagio poseen un comportamiento pareado. Para cada departamento, la primera ola es seguida por la segunda, es decir, un contexto de mediones funcionales antes y despues.

4.3.2. Datos funcionales construidos

La Figura 4.4 muestra los datos funcionales de la tasa de positividad COVID-19 en colombia en el Caso 1. Estos datos fueron construidos utilizando la base truncada CONS Φ_1 descrita en la Ecuación 2.8. Aquí, en las esquinas superior izquierda y derecha, se muestran los datos funcionales de las tasas de positividad en la primera y segunda ola respectivamente. En la esquina inferior izquierda, se muestran las medias funcionales, que son el objeto de comparación. En la esquina inferior derecha, se muestran las curvas diferencias.

De manera similar, los datos funcionales de la tasa de positividad de COVID-19 en Colombia en el caso 2 se muestran en la Figura 4.5. En los paneles superiores izquierdo y derecho se muestran los datos funcionales de la tasa de positividad para la primera y segunda oleada. Es posible apreciar una cierta diferencia en la tendencia de positividad entre las dos olas. Esto también se puede ver en las medias funcionales de ambas olas de contagio por COVID-19 que se muestran en el panel inferior izquierdo de ambas olas de contagio por COVID-19, que son el objetivo de comparación en el caso 2. Además, una tendencia diferente se observa entre las curvas de las diferencias en la tasa de positividad con respecto a las curvas diferencia del caso 1.

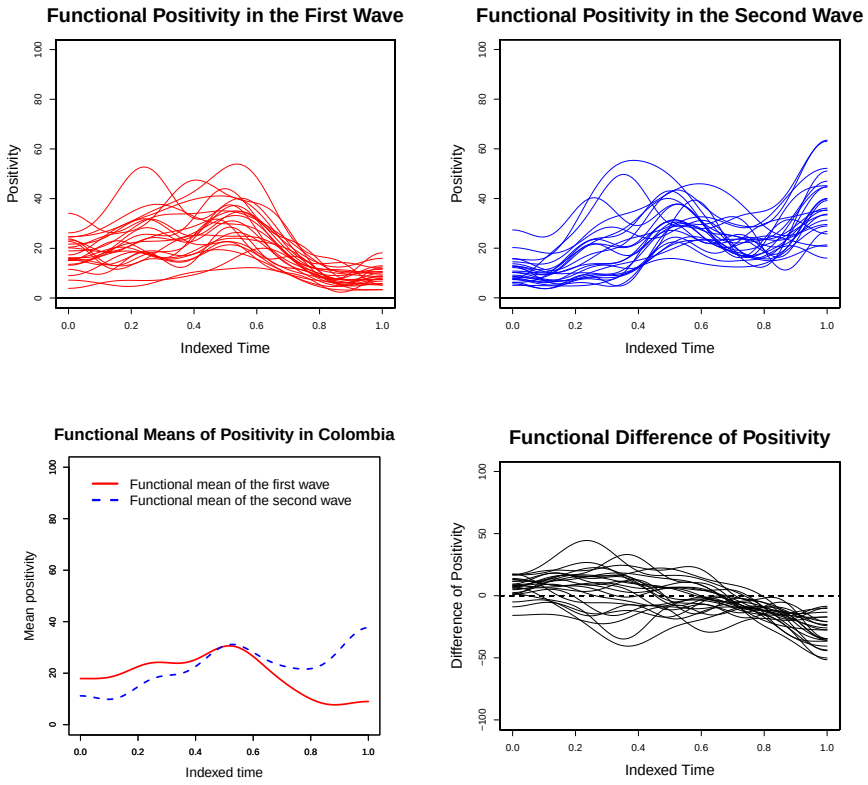


Figura 4.4: Datos de positividad funcional para COVID-19 en Colombia para el Caso 1: Positividad funcional en la primera ola (**arriba izquierda**); positividad funcional en la segunda ola (**arriba a la derecha**); medias funcionales de positividad en la primera ola en línea continua y la segunda ola en línea discontinua (**abajo izquierda**); curvas de diferencia de positividad (**abajo a la derecha**).

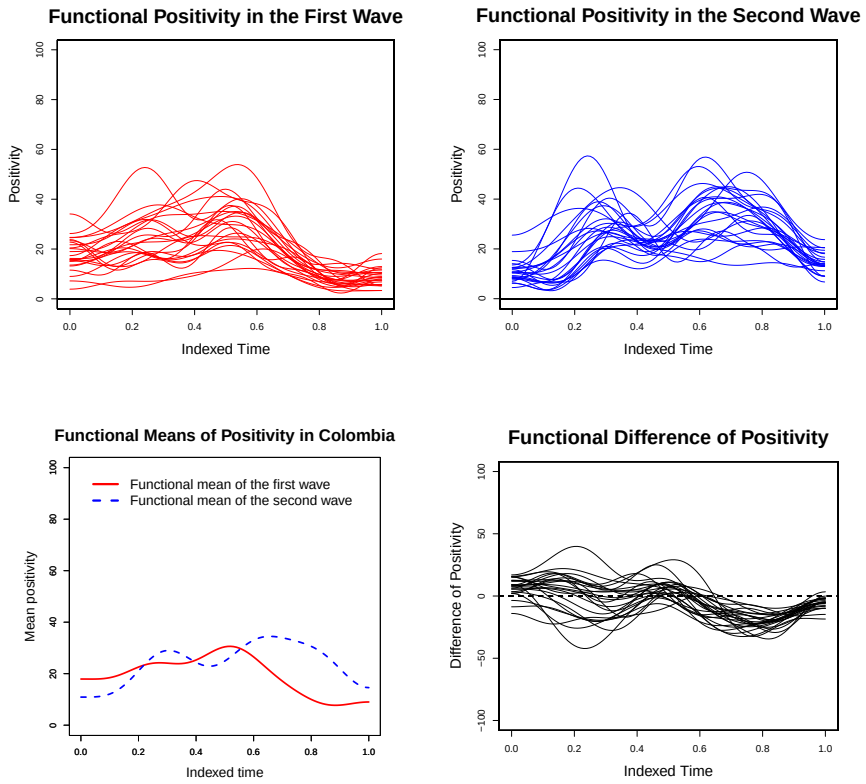


Figura 4.5: Datos de positividad funcional para COVID-19 en Colombia para el Caso 2: Positividad funcional en la primera ola (**arriba izquierda**); positividad funcional en la segunda ola (**arriba a la derecha**); medias funcionales de positividad en la primera ola en línea continua y la segunda ola en línea discontinua (**abajo izquierda**); curvas de diferencia de positividad (**abajo a la derecha**).

4.3.3. Contraste punto a punto

Como se indicó anteriormente, la prueba$-t$ puntual supone que, para cada valor del argumento de los datos funcionales en el dominio, se puede realizar una prueba$-t$ escalar con las imágenes de las funciones evaluadas en dicho punto. Es decir, una prueba$-t$ para los dos grupos de valores escalares $\{\mathcal{X}_i(t)\}_{i=1}^n$ y $\{\mathcal{Y}_i(t)\}_{i=1}^n$, resulta de evaluar las funciones $\mathcal{X}_i$ e $\mathcal{Y}_i$ en el mismo punto fijo t, para $i = 1, 2, \ldots, n$. En este caso, el contraste se define como

$$
\begin{aligned}
H_0: \quad & \mu_{\mathscr{X}}(t) - \mu_{\mathscr{Y}}(t) \quad = 0 \\
H_1: \quad & \mu_{\mathscr{X}}(t) - \mu_{\mathscr{Y}}(t) \quad \neq 0, \ \forall t \in \mathfrak{T},
\end{aligned}
\tag{4.8}
$$

Siendo $\mu_{\mathscr{X}}(t)$ y $\mu_{\mathscr{Y}}(t)$ parámetros escalares dependientes de cada valor t fijo. Este contraste se realiza mediante el estadístico

$$
\frac{\overline{\mathcal{X}}(t) - \overline{\mathcal{Y}}(t)}{sd/\sqrt{n}}, \tag{4.9}
$$

donde sd es la desviación estándar de los valores escalares $\{\mathcal{X}_i(t) - \mathcal{Y}_i(t)\}_{i=1}^{n}$. De esta manera, se toman 1000 valores dentro del intervalo $[0, 1]$ y se realiza el contraste en cada uno de ellos, obteniendo entonces 1000 valores$-p$, los cuales se muestran en la Figura 4.6 (panel izquierdo para el Caso 1 y panel derecho para el Caso 2).

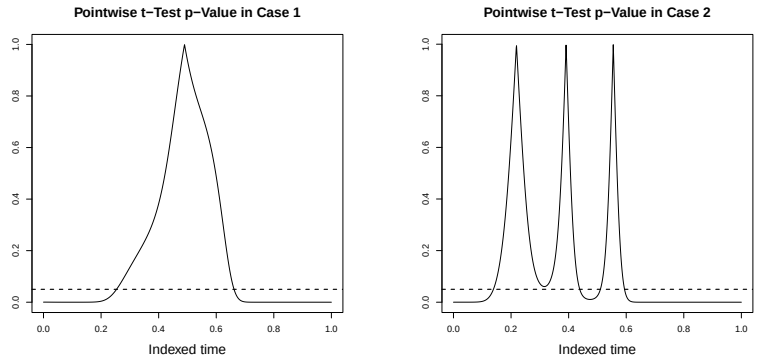


Figura 4.6: Resultados para el caso 1 (izquierda) y el caso 2 (derecha). En ambos casos, en linea continua los valores$-p$ obtenidos para t en $[0, 1]$ y en linea discontinua la referencia a 0.05.

Note que, hasta el momento, no es posible determinar si existen diferencias significativas entre las dos olas de contagio de COVID-19 a través de las curvas de positividad, de manera global.

4.3.4. Aplicación de la propuesta de contraste para muestras pareadas

Bajo la metodología expuesta anteriormente, se simulan dos grupos de curvas de tamaño 24 por pares bajo la hipótesis nula de que las medias funcionales son iguales, y se calcula el estadístico de prueba *MID* en la muestra. Este proceso se repitió 4000

veces por separado para cada caso de estudio, con lo que se obtuvieron 4000 valores del estadístico *MID*. Luego de realizar el proceso de simulación para obtener la distribución nula empírica, se calcula el valor del estadístico de prueba en los datos funcionales reales de la tasa de positividad para COVID-19 en Colombia en ambos casos de estudio. Utilizando la distribución empírica hallada bajo simulación, se establecen los valores críticos correspondientes a una significancia de 0.05 por medio de la frecuencia. Los histogramas de los valores encontrados, junto con el valor del estadístico y los respectivos valores críticos, se muestran en la Figura 4.7: en el panel izquierdo para el caso 1 y en el panel derecho para el caso 2.

En la Tabla 4.1, se muestran los valores$-p$ obtenidos en la prueba de Shapiro–Wilk realizada sobre los 24 datos de las integrales de la diferencia de los datos funcionales pareados en los dos casos de estudio. Además, se muestra el valor$-p$ del estadístico de prueba usando la distribución teórica del estadístico. También, se muestran los valores del estadístico de prueba en cada caso y sus respectivos valores$-p$ encontrados bajo simulación y los valores críticos de la distribución nula encontrados bajo simulación.

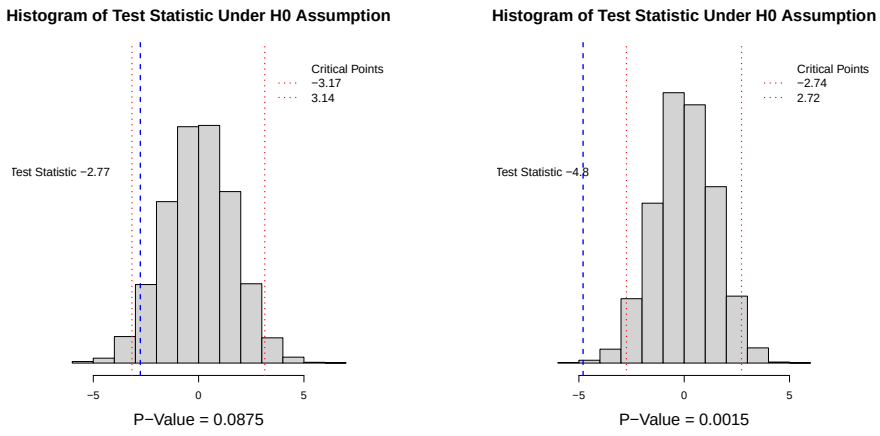


Figura 4.7: Histograma de los 4000 valores *MID* obtenidos bajo simulación y el estadístico de contraste *MID* obtenido a partir de los datos funcionales de positividad real en línea discontinua y los valores críticos en líneas punteadas para el Caso 1 (panel izquierdo) y el caso 2 (panel derecho).

Caso	**MID**	**Valor$-p$ bajo simulación**	**Valor$-p$ teórico**	**Valor-$-p$ Shapiro Wilk**
Case 1	-2.77	0.0875	0.08906	0.6029165
Case 2	-4.8	0.0015	0.00001	0.7006806

Tabla 4.1: Resumen de los resultados en ambos casos.

Note que ahora, con el uso de los valores$-p$ escalares encontrados con nuestra propuesta, es posible decidir sobre la existencia de diferencias significativas entre ambas olas de contagio de COVID-19 de manera global. Así, para el Caso 2, se puede decir que existen diferencias significativas entre las dos olas de contagio, ya que el valor$-p$ en este caso es 0,00001, bajo la distribución nula teórica, y 0,0015 bajo la distribución nula simulada. Por tanto, se rechaza la hipótesis de que las medias funcionales de la tasa de positividad son las mismas en ambas olas de contagio por COVID-19.

A su vez, para el primer caso, dado que los valores$-p$ son 0,08906 bajo la distribución nula teórica y 0,0875 bajo la distribución nula simulada, la hipótesis de que las medias funcionales de la tasa de positividad son las mismas en ambas ondas de contagio por COVID-19 no se rechaza, aunque por un margen muy pequeño respecto al valor de referencia de 0,05 de significancia.

4.4. Conclusiones

Es importante señalar que, como muestra la Figura 4.6, el contraste punto a punto para datos funcionales permite evaluar las secciones del dominio de las funciones donde existen diferencias significativas. En cuanto a los casos de estudio, se podrían identificar las fechas entre las que existe una mayor diferencia entre las dos olas de contagio. Sin embargo, dicha prueba es insuficiente para determinar si las dos oleadas de contagio son significativamente diferentes en cada estudio de caso.

Por otro lado, con el uso de la propuesta presentada para datos pareados, es posible tomar una decisión global basada en el valor$-p$ escalar. La aplicación de este contraste permite visualizar (como se puede ver en la Figura 4.7 (panel izquierdo)) que, para el

Caso 1, cuando se toma una significancia de 0.05, el estadístico de contraste no rechaza la hipótesis de igualdad de medias; es decir, con una significancia de 0,05, no hay evidencia de que existan diferencias significativas entre las dos ondas de contagio, aunque el valor$-p$ de 0,082 es relativamente cercano. Así, si la significancia se toma en 0,1, la decisión sería rechazar la hipótesis nula, aunque nuevamente, con un margen muy estrecho. Con respecto al Caso 2, como se muestra en la Figura 4.7(panel derecho), el valor$-p$ encontrado es 0.0015, por lo que es posible decir que hay evidencia suficiente de que las dos olas de contagio son significativamente diferente.

Por lo anterior, aunque el valor$-p$ en el caso 1 deja algunas dudas, es importante resaltar la diferencia entre los valores p en ambos casos desde un punto de vista más amplio, lo que parece apoyar la idea de que los dos estudios de caso son notablemente diferentes, y que el paro nacional en Colombia no debe ser ignorado al analizar el comportamiento epidemiológico, ya que los estudios de caso sugieren un posible cambio en la inclusión de datos positivos por incumplimiento de las medidas de cuidado durante el paro nacional.

Capítulo 5

Modelo de regresión logística funcional de medidas repetidas

5.1. Introducción y objetivos

Dentro de la investigación científica, modelar la relación de dependencia entre diferentes características asociadas a la observación de un cierto fenómeno resulta crucial. Este problema es usualmente tratado bajo el concepto de análisis de regresión, una técnica estadística nacida del problema de regresión que puede ser expresado mediante el modelo

$$Y = \mu(X),$$

donde Y representa una variable denominada variable dependiente, variable respuesta o simplemente variable; X representa un conjunto de variables llamadas variables explicativas, covariables o variables independientes, cuyo efecto sobre las observaciones de Y es denotado por $\mu()$. En este modelo de regresión, las variables relacionadas pueden ser de naturaleza continua o discreta. Con el surgimiento del FDA, resulta necesaria la inclusión de variables aleatorias funcionales al problema de regresión, dando origen a variantes exclusivas del FDA, por ejemplo, el modelo funcional-funcional que resulta cuando la variable respuesta y las variables explicativas son de naturaleza funcional; mientras que si se tiene es un modelo con variable respuesta escalar y covariables funcionales, se dice que se tiene un modelo de regresión escalar-funcional. Todas estas

formas del modelo de regresión funcional han sido tratadas desde diferentes enfoques, uno de ellos, conocido como enfoque paramétrico, parte de suponer que la relación $\mu()$ tiene una forma establecida que puede ser caracterizada por medio de un conjunto de parámetros desconocidos. Estos parámetros, que pueden ser funcionales o no según el modelo, proveen información del efecto de X sobre Y, por lo que se convierten entonces en el objetivo principal de la estimación.

En el campo de los modelos de regresión escalar-funcional, el modelo de regresión logística funcional es un caso muy particular y resulta esencial en fenómenos en donde es necesario modelar una respuesta categorizada en dos niveles, referidos usualmente como éxito/fracaso, a partir de observaciones de una covariable funcional. Este tipo de problema puede abordarse desde diferentes enfoques; sin embargo, este trabajo se enmarca dentro del enfoque paramétrico por lo que se asume la existencia de una función parámetro (o parámetro funcional) cuya interpretación proporciona una explicación de cómo los cambios a lo largo del dominio de las variables explicativas funcionales, generan cambios en el cociente de ventajas del éxito frente al fracaso (ver por ejemplo Escabias et al. (2013) para un caso de picos de los niveles de polen de olivo). Por esta razón, el objetivo principal en el enfoque paramétrico del modelo logístico funcional es obtener una estimación lo más precisa posible de la función parámetro. Dicha estimación puede ser obtenida aprovechando la estructura algebraica del espacio funcional del cual las observaciones se asume que son elementos, permitiendo así el uso de la expansión básica de los objetos funcionales para reducir el problema de estimación a un caso de estimación multivariante. El método más habitual para el problema de regresión logística funcional es el método de máxima verosimilitud (ML) (ver Escabias et al. (2004)). El uso de la expansión básica de los elementos funcionales ha sido tratado en diversos trabajos como Aguilera et al. (2010), Escabias et al. (2004), Acal & Aguilera (2023), Urbano-Leon et al. (2023) entre otros. No obstante, su uso en el contexto de regresión puede generar un problema conocido como multicolinealidad, es decir, correlación entre las variables explicativas del modelo, lo que lleva a grandes errores estándar y otros problemas en la estimación de los parámetros Fomby et al. (1984). Este inconveniente ha sido abordado por Escabias et al. (2004) y Escabias et al. (2022) con la introducción de la regresión

logística en componentes principales funcionales (FPCLR), aprovechando la estructura incorrelada de estas componentes garantizada teóricamente. El uso de las componentes principales funcionales, ha mostrado hasta ahora ser eficiente para tratar el problema de la multicolinealidad que subyace de la expansión básica, ofreciendo así buenas estimaciones de la función parámetro haciendo uso de ML.

Dentro de la investigación científica, puede resultar el caso en que una misma unidad experimental, sujeto o individuo, es medido en diferentes ocasiones dando lugar a un diseño de medidas repetidas. Esta estructura de repetición no permite asumir independencia en las observaciones, pues como aseguran Crowder & Hand (1990) y Davis (2002) para el caso escalar, el fenómeno de repetición puede generar una estructura de correlación en la matriz de diseño del modelo. Debido a esto, la estimación por ML, incluso en el modelo en componentes principales, puede no resultar adecuado.

Las medidas repetidas en el contexto funcional no son nuevas, han sido estudiadas como "análisis de medidas repetidas funcionales" que incluye diversas metodologías concebidas para el tratamiento de este tipo de diseños. Por ejemplo, en los trabajos de Martínez-Camblor & Corral (2011) y Smaga (2020) se discute el tema en el contexto de comparación de medias funcionales, siguiendo un enfoque principalmente no paramétrico, mientras que en Acal & Aguilera (2023) se trata el caso del análisis de varianza de medidas repetidas desde la metodología de expansión básica.

En el contexto de regresión, en el caso escalar, los diseños de medidas repetidas han sido usualmente abordados con la inclusión de efectos aleatorios en los modelos, como puede verse en Lindstrom & Bates (1990), Davis (2002), Hedeker & Gibbons (2006), lo que deriva en un tipo de modelos más generales que han sido denominados como modelos de efectos mixtos. Estos consideran que parte de la variabilidad del modelo justamente de la estructura de correlación causada por mediciones repetidas, planteando el uso de métodos alternativos de estimación como el método de máxima verosimilitud restringida (REML). En el caso funcional, los modelos de regresión de efectos mixtos han sido explorados para el caso de regresión funcional-funcional en trabajos como los de Guo (2002), Liu & Guo (2012) quienes los plantean de manera general; en el trabajo de Antoniadisa & Sapatinas (2007) quienes tratan el tema de la inferencia en estos mismos modelos explorando el uso

de wavelets y técnicas no paramétricas; también en Chen & Wang (2011) donde exploran el caso sobre los splines penalizados o en Scheipl et al. (2015), donde se enfocan de manera general en los modelos aditivos funcionales de efectos mixtos. En cambio, en el trabajo de Ma et al. (2019) se aborda el caso de los modelos mixtos en la regresión escalar-funcional, motivados por el estudio sobre datos funcionales provenientes de resonancia magnética. En todos los casos, se considera que el efecto aleatorio es también un objeto funcional. Hasta donde se sabe, el caso concreto de las medidas repetidas para el modelo de regresión logística funcional no ha sido considerado en la literatura, y mucho menos el efecto de la multicolinealidad en la estimación del modelo. Debido a esto, en este capítulo se propone una metodología para tratar el caso de medidas repetidas en el modelo logístico funcional, considerando la inclusión de un efecto aleatorio y el uso de las componentes principales funcionales como herramientas combinadas para tratar los inconvenientes en la estimación de la función parámetro. Los principales resultados mostrados en este capítulo se encuentran publicados en Urbano-Leon et al. (2024)

5.2. Generalidades sobre el modelo logístico escalar y funcional

5.2.1. El modelo logístico escalar

La regresión logística es una metodología que pretende modelar una variable respuesta binaria Y, en términos de una o más variables explicativas no aleatorias $X_1, X_2, \ldots, X_p$. La variable respuesta binaria Y suele estar asociada con el éxito o fracaso de un experimento aleatorio

$$Y(\omega) = \begin{cases} 1 & \text{si } \omega \text{ es un éxito} \\ 0 & \text{si } \omega \text{ es un fracaso.} \end{cases} \tag{5.1}$$

por lo que la distribución subyacente para modelizar esta variable es la distribución Bernoulli $Be(\pi)$. Este modelo resulta valioso en investigación debido a sus muchas aplicaciones en diferentes contextos.

Para la formulación del modelo de regresión logística, considérese una muestra de

n individuos y sea $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ el vector de observaciones de las variables explicativas para un cierto individuo i, y sea $y_1, y_2, \ldots, y_n$ las observaciones de la variable respuesta binaria, entonces las observaciones de la respuesta se expresan en términos de las observaciones de las variables explicativas como

$$y_i = \pi_i + \epsilon_i,\ i = 1, 2, \ldots, n, \tag{5.2}$$

donde la probabilidad de éxito de la respuesta $\pi_i(x_i) = P(Y = 1|x_i)$ se expresa como

$$\pi_i(x_i) = \frac{\exp\{\beta_0 + x_i'\beta\}}{1 + \exp\{\beta_0 + x_i'\beta\}},\ i = 1, 2, \ldots, n, \tag{5.3}$$

β_0 y $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ son los parámetros a estimar, y $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ son los términos de error del modelo. De manera equivalente, la probabilidad de éxito puede ser expresada como

$$\pi_i(x_i) = \frac{1}{1 + \exp\{-\beta_0 - x_i'\beta\}} = f(\beta_0 + x_i'\beta),\ i = 1, 2, \ldots, n, \tag{5.4}$$

donde f es la función logistica definida como

$$f(t) = \frac{1}{1 + \exp\{-t\}}. \tag{5.5}$$

Esta función está definida en $\mathbb{R}$, pero se encuentra acotada por 0 y 1 por lo que es adecuada para modelizar probabilidades. La gráfica 5.1 muestra la forma de la función logística.

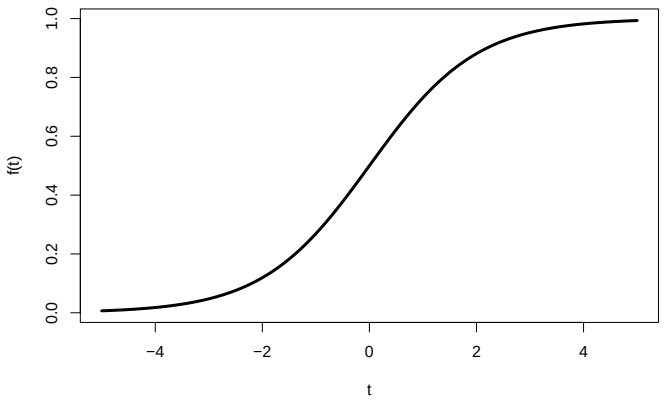


Figura 5.1: Función logística $f(t)$

Note que, debido a la naturaleza distribucional de la variable respuesta, cada $y_i|x_i \sim Be(\pi_i(x_i))$, que tiene por esperanza y varianza

$$E\left[y_i|x_i\right] = \pi_i(x_i),\ Var\left[y_i|x_i\right] = \pi_i(x_i)\left(1-\pi_i(x_i)\right), \tag{5.6}$$

indica que es imposible la utilización del modelo lineal clásico para variables de este tipo.

En este punto, cobra relevancia el concepto de "ventaja" (*Odds*), la cual es una medida que evalúa la ventaja que posee la probabilidad de ocurrencia de un evento con respecto a su no ocurrencia. En términos de la variable Y, es fácil ver que para la observación $i-$ésima el odds está dado por

$$O(x_i) = \frac{P(y_i=1|x_i)}{P(y_i=0|x_i)} = \frac{P(y_i=1|x_i)}{1-P(y_i=1|x_i)} = \frac{\pi_i(x_i)}{1-\pi_i(x_i)} = \exp\{\beta_0 + x_i'\beta\} \tag{5.7}$$

de donde se puede obtener la linealización

$$ln\left(\frac{\pi_i(x_i)}{1-\pi_i(x_i)}\right) = \beta_0 + x_i'\beta. \tag{5.8}$$

El término izquierdo en la Ecuación 5.8 es llamada función logit, una función de enlace muy conocida en la formulación del modelo logístico desde el punto de vista de los modelos lineales generalizados (GLM) (ver por ejemplo Dobson & Barnett (2008), Agresti (2015), Lindsey (2014)), definida por

$$logit(s) = ln\left(\frac{s}{1-s}\right), \quad \forall s \in [0,1]. \tag{5.9}$$

Así, denotanto por l_i a la función logit evaluada en π_i se tiene el modelo logístico

$$logit(\pi_i) = l_i = x_i\beta. \tag{5.10}$$

Esta expresión corresponde a la formulación clásica dentro de los modelos lineales generalizados, y permite una interpretación lineal de los parámetros contenidos en β, a través de los cambios en una unidad de cada una de las covariables contenidas en x_i. Con esto, si la $k-$ésima covariable varía en una unidad, dejando todas las demás fijas, es posible definir dos ventajas, una para x_{ik} y otra para $x_{ik}+1$ como

$$O(x_{ik}) = \exp\{z\}\exp\{x_{ik}\beta_k\} \qquad \wedge \qquad O(x_{ik}+1) = \exp\{z\}\exp\{(x_{ik}+1)\beta_k\}, \tag{5.11}$$

donde

$$z = \sum_{j=0, j\neq k}^{p} x_{ij}\beta_j.$$

Así, se tiene que el cociente de las ventajas (conocido como odds ratio) está dato por

$$\frac{O(x_{ik}+1)}{O(x_{ik})} = \frac{\exp\{z\}\exp\{(x_{ik}+1)\beta_k\}}{\exp\{z\}\exp\{x_{ik}\beta_k\}} = \exp\{\beta_k\}. \tag{5.12}$$

Esto permite obtener una interpretación de la exponencial del parámetro $k-$ésimo, en términos del cociente de ventajas a partir de la variación en una unidad de su correspondiente covariable.

5.2.2. Estimación en el modelo logístico

Dada la naturaleza del modelo logístico, el vector de parámetros β no puede ser estimado como en un modelo lineal, pues como puede verse en Hosmer & Lemeshow (2000) o en Agresti (2015), no se cumplen los supuestos para la correcta aplicación de mínimos cuadrados por lo que la estimación suele ser realizada por medio del método de máxima verosimilitud (ML por sus siglas en inglés). Este método sugiere tomar como estimador a los valores que maximizan la función de los parámetros basada en la verosimilitud, conocida como función de verosimilitud. En el caso particular del modelo de regresión logística, dada una muestra aleatoria $\{y_i\}_{i=1}^{n}$ donde cada y_i sigue una distribución de Bernoulli de parámetro $\pi_i(x_i)$ i.e. $y_i|x_i \sim Be\left(\pi_i(x_i)\right)$; se tiene que la función de verosimilitud está dada por

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} \left(1-\pi_i\right)^{1-y_i}. \tag{5.13}$$

En la expresión 5.13, y en lo que sigue, se denota $\pi_i = \pi_i(x_i)$ por simplicidad en la notación, siempre que no haya lugar a ambigüedad. Por otra parte, debido a que la función logaritmo es un una función monótona creciente, el problema de maximizar la función en términos de los parámetros de dicha probabilidad es equivalente a maximizar el logaritmo de la

función de verosimilitud, conocida como log-verosimilitud y dada por

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i ln\left(\frac{\pi_i}{1+\pi_i}\right) + ln\left(1-\pi_i\right) \tag{5.14}$$

$$= \sum_{i=1}^{n} \left[\left(y_i\left(\beta_0 + x_i'\beta\right)\right) - ln\left(1+\exp\left\{\beta_0 + x_i'\beta\right\}\right)\right] \tag{5.15}$$

De aquí, maximizar la función $\mathcal{L}(\beta)$ corresponde al ejercicio clásico de obtener las raíces de las ecuaciones dadas por las derivadas parciales

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \frac{\exp\left\{\beta_0 + x_i'\beta\right\}}{1+\exp\left\{\beta_0 + x_i'\beta\right\}} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\pi_i. \tag{5.16}$$

El problema deriva en resolver las ecuaciones de tipo

$$\sum_{i=1}^{n} x_{ij}\left(y_i - \pi_i\right) = 0,\ j = 0, 1, \ldots, p, \tag{5.17}$$

donde $x_{i0} = 1$. Estas ecuaciones son no lineales, por lo que carecen de una solución analítica y se recurre a métodos iterativos para su aproximación, siendo el método de Newton Raphson uno de los mas populares Agresti (2015).

5.2.3. Modelo de regresión logística funcional

Dentro del FDA, el modelo de regresión logística funcional pertenece a la clase de modelos de regresión escalar-funcional, que se caracterizan por tener respuesta escalar y predictor funcional, que se encuentran ligados por medio del uso de alguna operación que extrae una característica de las observaciones funcionales. En el caso particular de una variable respuesta binaria Y como en la Ecuación 5.1 la intensión de modelado se mantiene, aunque ahora desde un predictor funcional. Es decir, que para una muestra aleatoria $\{Y_i\}_{i=1}^n$ se tiene un conjunto de observaciones funcionales no aleatorias $\{\mathcal{X}_i\}_{i=1}^n \subset \mathcal{H}$ asociadas a una variable aleatoria $\mathscr{X}$, un parámetro escalar β_0 y un parámetro funcional $\beta \in \mathcal{H}$, relacionadas mediante la respuesta esperada para la observación $i-$ésima por

$$\pi_i = f\left(\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt\right), \tag{5.18}$$

donde f es la función logística de la Ecuación 5.5, y que queda especificado de manera equivalente como

$$\pi_i = \frac{\exp\left\{\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt\right\}}{1+\exp\left\{\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt\right\}}. \tag{5.19}$$

Note que la respuesta esperada es realmente una esperanza condicional dada una observación de la covariable funcional. Es decir que ahora, y en lo que sigue, $\pi_i = \mathbf{E}[Y|\mathscr{X} = \mathcal{X}_i]$. Además, dado que $\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt \in \mathbb{R}$, $\pi_i \in [0,1]$ por lo que puede plantearse la ventaja como

$$O(\mathcal{X}_i) = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi_i}{1-\pi_i} = \exp\left\{\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt\right\}, \tag{5.20}$$

lo cual, al igual que el modelo escalar, permite la linealización

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt. \tag{5.21}$$

Es decir

$$Logit(\pi_i) = \beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt, \tag{5.22}$$

lo que permite ver al modelo de regresión logística funcional como un caso particular de un modelo lineal generalizado funcional (por ejemplo ver James (2002)).

En este punto, vale la pena recordar que uno de los objetivos principales dentro del enfoque paramétrico es la interpretación del parámetro; el cual, en el caso del modelo logístico funcional es la función β. Dicha interpretación, como puede verse en Escabias et al. (2004), puede obtenerse mediante un cociente de ventajas bajo la consideración de que el predictor es funcional dentro de un dominio cerrado, por lo que es posible obtener un cociente de ventajas relacionado con la variación continua dentro de la observación funcional. Así, de la Ecuación 5.20 es fácil ver que dada una variación continua y acotada Δ definida en $\mathfrak{T}$, se tiene que

$$Odds\left(\mathcal{X}_i + \Delta\right) = Odds(\mathcal{X}_i)\exp\left\{\int_{\mathfrak{T}} \Delta(t)\beta(t)dt\right\}. \tag{5.23}$$

Por lo que el cociente de ventajas respecto a la variación Δ está dada por

$$\frac{Odds\left(\mathcal{X}_i + \Delta\right)}{Odds\left(\mathcal{X}_i\right)} = \exp\left\{\int_{\mathfrak{T}} \Delta(t)\beta(t)dt\right\}. \tag{5.24}$$

Es decir que es posible obtener una interpretación de la exponencial de la función parámetro en términos del cociente de ventajas respecto a una variación continua, la cual permite además, una interpretación en sub-intervalos de $\mathfrak{T}$ restringiendo la integral a estos. Debido a esto, resulta evidente que una correcta estimación de la función parámetro es necesaria para llevar a cabo dichas interpretaciones.

5.2.4. Estimación en el modelo logístico funcional

Dada la naturaleza funcional del predictor $\mathcal{X}\beta$, la Ecuación de verosimilitud 5.17 conlleva inevitablemente a la solución de

$$\sum_{i=1}^{n} \mathcal{X}_i(t)\,(y_i - \pi_i) = 0,\ \forall t \in \mathfrak{T}, \tag{5.25}$$

es decir, la solución de 5.25 en cada uno de los infinitos puntos del dominio $\mathfrak{T}$, es por esto, que los métodos escalares puntuales no resultan apropiados. Sin embargo, dado que las observaciones de la variable funcional $\mathscr{X}$ y la función parámetro β están definidas en el mismo espacio $\mathcal{H} \subset L_2[\mathfrak{T}]$, donde $\mathcal{H} = Gen(\Phi)$, siendo $\Phi = \{\phi_j\}_{j=1}^{d}$ una base para $\mathcal{H}$, entonces existen los vectores de $\mathbb{R}^d$, $\mathcal{B} = (\beta_1, \beta_2, \ldots, \beta_d)'$ y $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})'$ para $i = 1, 2, \ldots, n$ tales que

$$\beta = \sum_{j=1}^{d} \beta_j \phi_j = \mathcal{B}'\phi \quad \wedge \quad \mathcal{X}_i = \sum_{j=1}^{d} x_{ij}\phi_j = x_i'\phi;\ i = 1, 2, \ldots, n, \tag{5.26}$$

donde hemos denotado por $\phi = (\phi_1, \phi_2, \ldots, \phi_d)' \in \mathcal{H}^d$. De este modo, el modelo de regresión logística de la Ecuación 5.22 se puede reescribir como

$$\begin{aligned}
l_i = logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \int_{\mathfrak{T}} \mathcal{X}_i(t)\beta(t)dt && (5.27)\\
&= \beta_0 + \langle \mathcal{X}_i, \beta \rangle && (5.28)\\
&= \beta_0 + \left\langle \sum_{j=1}^{d} x_{ij}\phi_j, \sum_{j=1}^{d} \beta_j\phi_j \right\rangle && (5.29)\\
&= \beta_0 + \sum_{j=1}^{d} x_{ij} \left\langle \phi_j, \sum_{j=1}^{d} \beta_j\phi_j \right\rangle && (5.30)\\
&= \beta_0 + \sum_{j=1}^{d} x_{ij}\beta_j \left\|\phi_j\right\|^2 + \sum_{j=1, k\neq j}^{d} x_{ij}\beta_j \left\langle \phi_j, \phi_k \right\rangle . && (5.31)
\end{aligned}$$

Es decir que para cada observación $i-$ésima se tiene una combinación lineal de valores dados por los coeficientes básicos de $\mathcal{X}_i$ y β, junto con los productos internos de las funciones de Φ, por lo que para todas las n observaciones, el modelo puede escribirse en forma matricial como

$$L = \mathbf{1}\beta_{\mathbf{0}} + \mathbf{A}\mathbf{\Psi}\mathcal{B}, \tag{5.32}$$

donde $L = (l_1, l_2, \ldots, l_n)'$ es el vector de transformaciones logit, $\mathbf{1}$ es un vector de unos en $\mathbb{R}^n$, A es la matriz $n \times d$ de coeficientes básicos de las observaciones funcionales, B es el vector de coeficientes básicos de la función β y Ψ es la matriz cuadrada $d \times d$ que contiene los cuadrados de las normas de las funciones de Φ en la diagonal y los productos internos de estas en el resto de entradas. Es decir, que se puede obtener una estimación

$$\hat{\beta} = \sum_{j=1}^{d} \hat{\beta}_j \phi_j,$$

es decir, a partir de la estimación de sus coeficientes básicos $\left(\hat{\beta}_j\right)_{j=1}^{d}$ desde el modelo de regresión logística multivariante planteado en la Ecuación 5.32.

Dado que se asume que las observaciones son independientes, la estimación $\hat{\beta}$ puede ser obtenida mediante máxima verosimilitud como en el caso escalar. Sin embargo, ya que la matriz $A\Psi$ es obtenida a partir de los coeficientes básicos, no se puede garantizar que las columnas de estos sean incorreladas por lo que se presenta un problema conocido en el contexto de regresión multivariante como multicolinealidad (ver por ejemplo Fomby et al. (1984)). Este problema genera múltiples inconvenientes que derivan en grandes errores estándar en las estimaciones de los parámetros (coeficientes básicos de la función parámetro) del modelo. Estos errores estándar provocan que la estimación obtenida de los parámetros podría ser imprecisa, lo que perjudica la interpretación de la función parámetro. Ahora bien, este inconveniente dentro del modelo logístico funcional ha sido tratado en trabajos como el de Escabias et al. (2004), en donde se propone que para lidiar con el problema de multicolinealidad debida a la reformulación del modelo logístico en términos de los coeficientes básicos, el modelo puede replantearse a un modelo en términos de las componentes principales funcionales muestrales. Así, dadas las componentes principales funcionales de las observaciones funcionales $\{\mathcal{X}_i\}_{i=1}^{n}$ descritas como.

$$\xi_{i,w} = \int_a^b \left(\mathcal{X}_i(t) - \overline{\mathcal{X}}(t)\right) \mathscr{F}_w(t) dt,$$

donde las funciones $\mathscr{F}_w \in \mathcal{H}$, $(w = 1, 2, \ldots, d)$ son las funciones propias provenientes de la Ecuación 2.22. Dado que los vectores ξ_w y ξ_j son incorrelados para todo $w \neq j$, y cada

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

$$\mathcal{X}_i \approx \overline{\mathcal{X}} + \sum_{w=1}^{p<d} \xi_{i,w} \mathscr{F}_w,$$

entonces, el modelo logístico funcional en componentes principales está dado por

$$l_i = \beta_0 + \int_{\mathfrak{T}} \left(\overline{\mathcal{X}}(t) + \sum_{w=1}^{p} \xi_{iw} \mathscr{F}_w(t) \right) \beta(t) dt \; i = 1, 2, \ldots, n, \tag{5.33}$$

que en su forma matricial es

$$L = \mathbf{1}\gamma_0 + \Gamma\gamma, \tag{5.34}$$

donde Γ es la matriz de componentes principales $(\xi_{i,w})$, $\gamma_0 = \alpha + \int_a^b \overline{\mathcal{X}}(t)\beta(t)dt$ y γ el vector de parámetros con elementos $\gamma_w = \int_a^b f_w(t)\beta(t)dt$. Esta formulación evita el problema de multicolinealidad a través de los componentes principales no correlacionados. Por lo que es posible el uso de métodos clásicos, como la máxima verosimilitud, para obtener una estimación $\hat{\gamma}$ del vector de parámetros γ y por medio de esta, una estimación del vector de parámetros original $\mathcal{B}$, que corresponde a los coeficientes del parámetro funcional β, por medio de $\hat{\mathcal{B}} = V\hat{\gamma}$, con V la matriz de coeficientes básicos de las funciones propias $\mathscr{F}_w$ in $\mathcal{H}$. Esta forma de evitar la multicolinealidad en el modelo de regresión logística por medio de las componentes principales funcionales, ya ha sido tratado en Escabias et al. (2004), en donde los autores proponen dos enfoques diferentes de análisis de componentes principales funcionales para este fin.

Una vez obtenida una estimación de la función parámetro β, resulta necesario comprobar el desempeño del modelo. Para ello, en la regresión logística funcional se puede hacer uso de la tasa de clasificaciones correctas (CCR por sus siglas en inglés), la cual indica el porcentaje de valores cuya predicción coincide perfectamente con la observación original. La CCR puede ser usada tanto en el modelo logístico escalar como en el funcional, ya que depende de la respuesta que, como se sabe, es escalar en ambos casos. La CCR está definida como

$$CCR = \frac{1}{n} \sum_{i=1}^{n} I(y_i = \hat{y}_i), \tag{5.35}$$

donde y_i e $\hat{y}_i$ son la $i-$ésima observación binaria y su correspondiente predicción respectiva. Por otro lado, es posible evaluar la precisión de las estimaciones del parámetro funcionales

mediante el error cuadrático integrado (*ISE* por sus siglas en inglés), definido como

$$ISE = \int_a^b (\beta(t) - \hat{\beta}(t))^2 dt. \tag{5.36}$$

5.3. Metodología: Modelo de regresión logística funcional en medidas repetidas

Dentro de la investigación científica, un diseño de medidas repetidas surge cuando una misma unidad experimental, es medida repetidamente bajo iguales o diferentes condiciones, lo que produce un conjunto de observaciones cuya independencia no puede ser asumida a priori (ver Davis (2002)). Existen diferentes diseños de medidas repetidas, por ejemplo, cuando una única unidad experimental es medida repetidamente a lo largo del tiempo, que usualmente ha sido tratado en la literatura como una serie de tiempo o un dato longitudinal. Por otro lado, existen otros diseños en donde un conjunto de sujetos son medidos en un número determinado de instantes, hecho que suele ser tratado en el contexto de regresión con la inclusión de efectos aleatorios y conocido como modelos de efectos mixtos. También, un caso particular de un diseño de medidas repetidas es aquel en donde un mismo conjunto de sujetos son medidos antes y después de un tratamiento, en cuyo caso, el número de repeticiones es dos y suele ser denominado un diseño pareado o un diseño de muestras pareadas. En el contexto de los datos funcionales, los diseños de medidas repetidas guardan concordancia con sus homólogos en la estadística clásica, solo que ahora, cada medición corresponde a un dato funcional. Así, es posible tener diseños de medidas repetidas en datos funcionales como series de tiempo funcionales, muestras funcionales pareadas o conjuntos de medidas repetidas funcionales. Este tipo de diseños se encuentran enmarcados, según Koner & Staicu (2023) y Wang et al. (2016), dentro de los denominados datos funcionales de segunda generación que se diferencian de los datos funcionales típicos (de primera generación) en que no es posible asumir la independencia de las observaciones y debe tenerse en cuenta la posible dependencia que el diseño de medidas repetidas pueda generar.

En los términos de las secciones anteriores de este capítulo, suponga que un conjunto

de N sujetos es medido sobre un mismo dominio continuo (mediciones funcionales) repetidamente n_i veces en cada sujeto, es decir, para $i = 1, 2, \ldots, N$. En este caso, se tiene un diseño de medidas repetidas funcionales en donde cada curva $\mathcal{X}_{is}$ representa la $s-$ésima repetición funcional para el $i-$ésimo sujeto, obteniendo así, el conjunto de observaciones funcionales $\bigcup_{i=1}^{N} \{\mathcal{X}_{is}\}_{s=1}^{n_i} \subset \mathcal{H}$, donde el cardinal del conjunto $\bigcup_{i=1}^{N} \{\mathcal{X}_{is}\}_{s=1}^{n_i}$ puede ser hallado por la suma $\sum_{i=1}^{N} n_i = n$, siendo n_i el número condiciones experimentales bajo las cuales el individuo fue medido, es decir, el número de repeticiones para cada individuo, y n el total de observaciones funcionales. Note que en principio, el número de repeticiones no es necesariamente igual para cada individuo como en un diseño pareado.

Suponga ahora que se tiene un conjunto de respuestas dicotómica $\{y_{is}\}_{i=1,s=1}^{N,n_i}$, las cuales se encuentran relacionadas con el conjunto de observaciones funcionales repetidas. Cada una de estas observaciones binarias, puede ser categorizada como

$$y_{is} = \begin{cases} 1 & \text{si} \quad \text{éxito} \\ 0 & \text{si} \quad \text{fracaso.} \end{cases} \tag{5.37}$$

En términos prácticos, la Tabla 5.1 muestra el formato de los datos en un diseño de medidas repetidas funcionales de respuesta binaria.

Al igual que en el modelo logístico multivariante y el modelo logístico funcional, para el caso de medidas repetidas cada y_{is} tiene distribución Bernoulli $y_{is}|\mathcal{X}_{is} \sim Be(\pi_{is})$, con parámetro $\pi_{is} = P(Y_{is} = 1|\mathcal{X}_{is})$ para cada $i = 1, 2, \ldots, N$ y $s = 1, 2, \ldots, n_i$. Como se dijo con anterioridad, el problema de medidas repetidas en el caso escalar puede ser tratado por medio de la inclusión de un efecto aleatorio dentro del modelo. Según indican Larsen et al. (2000), el efecto aleatorio puede ser visto como una covariable no medida, como una forma de modelar la heterogeneidad o como una forma de modelar datos correlacionados. En el caso de las medidas repetidas, dicho efecto aleatorio es un valor no observado de una variable aleatoria Normal centrada y de varianza constante $U \sim N(0, \sigma_U)$ que pretende capturar la variabilidad que se genera dentro de las observaciones repetidas de una misma unidad experimental. Así, el modelo logístico funcional para medidas repetidas en términos de la $s-$ésima observación del $i-$ésimo individuo puede plantearse como

$$y_{is} = \pi_{is} + \epsilon_{is}, \ i = 1, 2, \ldots, N; s = 1, 2, \ldots, n_i, \tag{5.38}$$

donde

$$\pi_{is} = \frac{\exp\left\{\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_{is}(t)\beta(t)dt + z_{is}u_i\right\}}{1 + \exp\left\{\beta_0 + \int_{\mathfrak{T}} \mathcal{X}_{is}(t)\beta(t)dt + z_{is}u_i\right\}}. \tag{5.39}$$

Al igual que en el modelo logístico clásico, el modelo se puede expresar de manera lineal en términos de la transformación logit como

$$l_{is} = ln\left[\frac{\pi_{is}}{1-\pi_{is}}\right] = \beta_0 + \int_{\mathfrak{T}} \mathcal{X}_{is}(t)\beta(t)dt + z_{is}u_i,\ i = 1, 2, \ldots, N; s = 1, 2, \ldots, n_i, \tag{5.40}$$

siendo $\beta_0 + \int_a^b \mathcal{X}_{is}(t)\beta(t)dt$ un efecto fijo (no aleatorio), u_i un efecto aleatorio asociado con el individuo $i-$ésimo (efecto aleatorio del sujeto i) y z_{is} una variable indicadora de la repetición.

Sujeto	**Repetición**	**Respuesta binaria**	**Observación funcional**
1	1	y_{11}	$\mathcal{X}_{11}$
1	2	y_{12}	$\mathcal{X}_{11}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	n_1	y_{2n_1}	$\mathcal{X}_{1n_1}$
2	1	y_{21}	$\mathcal{X}_{21}$
2	2	y_{22}	$\mathcal{X}_{21}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	n_2	y_{2n_2}	$\mathcal{X}_{2n_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	y_{N1}	$\mathcal{X}_{N1}$
N	2	y_{N2}	$\mathcal{X}_{N1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	n_N	y_{Nn_N}	$\mathcal{X}_{Nn_N}$

Tabla 5.1: Esquema de datos en un diseño de medidas repetidas funcionales de respuesta binaria

5.3.1. Estimación en el modelo logístico de medidas repetidas

Note que el modelo de la Ecuación 5.40 corresponde a la formulación clásica del modelo logístico funcional de la Ecuación 5.22 con la inclusión de un efecto aleatorio, que en este caso es un valor escalar proveniente de una distribución normal de media 0 y desviación σ_U, guardando concordancia con la naturaleza escalar de la respuesta en lugar de ser una función de efecto aleatorio. No obstante, las componentes del efecto fijo β y $\mathcal{X}_{is}$ siguen siendo objetos funcionales, por lo que el objetivo es obtener una estimación $\hat{\beta}$ de β. Así, al suponer que los objetos funcionales pertenecen al subespacio $\mathcal{H}$, existen los vectores $\mathcal{B} = (\beta_1, \beta_2, \ldots, \beta_d)'$ y $x_{is} = (x_{is1}, x_{is2}, \ldots, x_{isd})'$ de $\mathbb{R}^d$, tales que

$$\mathcal{X}_{is} = \sum_{j=1}^{d} x_{isj}\phi_j = x'_{is}\phi \ \wedge \ \beta = \sum_{j=1}^{d} \beta_j\phi_j = \mathcal{B}'\phi, \ i = 1, 2, \ldots, N, s = 1, 2, \ldots, n_i, \quad (5.41)$$

entonces, de la Ecuación 5.40 se sigue que

$$\begin{aligned} \int_{\mathfrak{T}} \mathcal{X}_{is}(t)\beta(t)dt &= \int_{\mathfrak{T}} \left[\sum_{j=1}^{d} x_{isj}\phi_j(t)\right]\left[\sum_{j=1}^{d} \beta_j\phi_j(t)\right] dt \\ &= \left[\sum_{j=1}^{d} x_{isj}\beta_j \left\|\phi_j\right\|^2\right] + \left[\sum_{j=1,k\neq j}^{d} x_{isj}\beta_k \left\langle\phi_j, \phi_k\right\rangle\right]. \end{aligned} \quad (5.42)$$

Por lo que el modelo logístico funcional para medidas repetidas puede ser expresado en forma matricial como

$$L = \mathbf{1}\beta_0 + A\Psi\mathcal{B} + ZU, \quad (5.43)$$

donde $\mathbf{1}$ es un vector de unos, L es el vector de las n transformaciones logit l_{is}, A es la matriz de coeficientes básicos de las curvas, Ψ es la matriz de productos internos de los elementos de la base Φ, $\mathcal{B}$ el vector de coeficientes básicos del parámetro funcional β, U es el vector de efectos aleatorios y Z es la matriz de diseño asociada a la repetición de los valores de U. Así, obteniendo una estimación $\hat{\mathcal{B}} = \left(\hat{\beta}_j\right)_{j=1}^{d}$ del vector $\mathcal{B}$ se puede proveer una estimación de la función parámetro *beta* como

$$\hat{\beta} = \sum_{j=1}^{d} \hat{\beta}_j\phi_j.$$

Es decir, el modelo de regresión logística funcional para medidas repetidas permite su tratamiento por medio de los coeficientes básicos. No obstante, la estimación a partir de

aquí se ve afectada por el diseño de medidas repetidas, pues no es posible asumir que las observaciones dentro de un mismo individuo sean independientes, por lo que tampoco resulta apropiado asumir que la verosimilitud es el producto de las probabilidades de las observaciones individuales como en 5.13, pues al hacerlo, la estimación por máxima verosimilitud podría ser imprecisa o sesgada. Pese a esto, se puede suponer que toda la correlación inherente a la repetición de las mediciones dentro de un mismo individuo se encuentra dentro del efecto aleatorio del mismo sujeto, es decir, las observaciones de un mismo sujeto son condicionalmente independientes dado un valor de su efecto aleatorio, siendo entonces dicho efecto una condición del individuo como se muestra en la Tabla 5.2.

Sujeto	**Repetición**	**Respuesta binaria**	**Observación funcional**	**Efecto Aleatorio**
1	1	y_{11}	$\mathcal{X}_{11}$	u_1
1	2	y_{12}	$\mathcal{X}_{11}$	u_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	n_1	y_{2n_1}	$\mathcal{X}_{1n_1}$	u_1
2	1	y_{21}	$\mathcal{X}_{21}$	u_2
2	2	y_{22}	$\mathcal{X}_{21}$	u_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	n_2	y_{2n_2}	$\mathcal{X}_{2n_2}$	u_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	1	y_{N1}	$\mathcal{X}_{N1}$	u_N
N	2	y_{N2}	$\mathcal{X}_{N1}$	u_N
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	n_N	y_{Nn_N}	$\mathcal{X}_{Nn_N}$	u_N

Tabla 5.2: Esquema de datos en un diseño de medidas repetidas funcionales de respuesta binaria con inclusión de un efecto aleatorio.

Este supuesto, conocido como hipótesis de la independencia condicional (ver Hedeker & Gibbons (2006) para el caso escalar), permite suponer que la probabilidad de que la variable aleatoria Y tome el valor 1 para la repetición $s-$ésima del $i-$ésimo individuo

está condicionada al efecto del individuo mismo, por lo que

$$p(\,Y_{is}=1|U=u_i)=\pi_{is}=\frac{\exp\left\{\beta_0+\int_{\mathfrak{T}}\mathcal{X}_{is}(t)\beta(t)dt+z_{is}u_i\right\}}{1+\exp\left\{\beta_0+\int_{\mathfrak{T}}\mathcal{X}_{is}(t)\beta(t)dt+z_{is}u_i\right\}}. \tag{5.44}$$

Siguiendo la linea de razonamiento de Hedeker & Gibbons (2006) para el caso escalar, la probabilidad conjunta condicional de las mediciones repetidas de un mismo individuo puede obtenerse por medio del producto

$$\prod_{s=1}^{n_i}\left(\pi_{is}\right)^{y_{is}}\left(1-\pi_{is}\right)^{1-Y_{is}}. \tag{5.45}$$

A partir de aquí, la probabilidad conjunta de un individuo (incondicional), puede ser vista como

$$\int_u\left[\prod_{s=1}^{n_i}\left(\pi_{is}\right)^{y_{is}}\left(1-\pi_{is}\right)^{1-Y_{is}}\right]\left(f(u;\sigma_U)\right)du, \tag{5.46}$$

donde $f(u,\sigma_U)$ representa la densidad de U. Esto se corresponde con la probabilidad marginal de las observaciones repetidas del sujeto $i-$ésimo. Así, la verosimilitud para todos los N sujetos puede verse como

$$\prod_{i=1}^{N}\int_u\left[\prod_{s=1}^{n_i}\left(\pi_{is}\right)^{y_{is}}\left(1-\pi_{is}\right)^{1-Y_{is}}\right]\left(f(u;\sigma_U)\right)du. \tag{5.47}$$

En este punto, conviene recordar que la función parámetro β y las observaciones funcionales poseen la representación dada en 5.26 por ser elementos de $\mathcal{H}$, por lo que el predictor lineal puede expresarse como en la Ecuación 5.42, quedando las probabilidades π_{is} en términos de $\mathcal{B}$ y σ_U de las Ecuaciones 5.46 y 5.47 como

$$\pi_{is}(\mathcal{B};\sigma_U)=\frac{\exp\left\{\beta_0+\sum_{j=1}^{d}x_{is,j}\beta_j\left\|\phi_j\right\|+\sum_{k,j=1,k\neq j}^{d}x_{is,j}\beta_k\left\langle\phi_j,\phi_k\right\rangle+u_i\right\}}{1+\exp\left\{\beta_0+\sum_{j=1}^{d}x_{is,j}\beta_j\left\|\phi_j\right\|+\sum_{k,j=1,k\neq j}^{d}x_{is,j}\beta_k\left\langle\phi_j,\phi_k\right\rangle+u_i\right\}}. \tag{5.48}$$

Así, la verosimilitud para el modelo de medidas repetidas está dada por

$$L(\mathcal{B};\sigma_U)=\prod_{i=1}^{N}\int_u\left[\prod_{s=1}^{n_i}\left(\pi_{is}(\mathcal{B};\sigma_U)\right)^{y_{is}}\left(1-\pi_{is}(\mathcal{B};\sigma_U)\right)^{1-y_{is}}\right]f(u;\sigma_U)du, \tag{5.49}$$

cuya maximización puede ser obtenida mediante la logverosimilitud

$$\mathcal{L}(\mathcal{B};\sigma_U)=ln(L(\mathcal{B};\sigma_U)), \tag{5.50}$$

haciendo uso de métodos numéricos iterativos para la solución del problema de maximización y métodos de cuadratura para las integrales involucradas. Este método, es conocido en el caso escalar como método de verosimilitud restringida (REML por sus siglas en inglés).

La estimación obtenida de la función parámetro por medio de la maximización de la función en la Ecuación 5.50 soluciona el problema de las medidas repetidas, no obstante, dado que se realiza por medio de los coeficientes básicos de los objetos funcionales, el problema de la multicolinealidad subyacente puede estar presente. Debido a esto, una alternativa es plantear el modelo de regresión logística funcional para medidas repetidas en términos de las componentes principales funcionales. Así, dadas las componentes principales funcionales de las observaciones $\{\mathcal{X}_{is}\}_{i=1,s=1}^{N,n_i}$ descritas como

$$\xi_{isv} = \int_{\mathfrak{T}} \left(\mathcal{X}_{is}(t) - \overline{\mathcal{X}}(t)\right) \mathscr{F}_v(t)dt,$$

donde las funciones propias $\mathscr{F}_v \in \mathcal{H}$, $(v = 1, 2, \ldots, d)$ y los vectores ξ_v y ξ_ω son independientes para todo $v \neq \omega$, se obtiene la representación

$$\mathcal{X}_{is} \approx \overline{\mathcal{X}} + \sum_{v=1}^{p} \xi_{isv}\mathscr{F}_v.$$

Entonces, de la Ecuación 5.40 se obtiene el modelo logístico funcional para medidas repetidas en componentes principales como

$$l_{is} = \beta_0 + \int_{\mathfrak{T}} \left(\overline{\mathcal{X}}(t) + \sum_{v=1}^{p} \xi_{isv}\mathscr{F}_v(t)\right) \beta(t)dt + z_{is}u_i,\ i = 1, 2, \ldots, N; s = 1, 2, \ldots, n_i, \tag{5.51}$$

el cual puede ser expresado en forma matricial como

$$L = \mathbf{1}\gamma_0 + \Gamma\gamma + ZU, \tag{5.52}$$

donde Γ es la matriz de componentes principales (ξ_{isv}), $\gamma_0 = \beta_0 + \int_{\mathfrak{T}} \overline{\mathcal{X}}(t)\beta(t)dt$ y γ el vector de parámetros con elementos $\gamma_v = \int_{\mathfrak{T}} f_v(t)\beta(t)dt$. De esta manera, se puede obtener una estimación $\hat{\gamma}$ por medio de la maximización de la verosimilitud condicionada a los efectos aleatorios estandarizados individuales, y con esta, obtener una estimación del vector de parámetros original $\mathcal{B}$ de los coeficientes de la función parámetro β por medio de $\hat{\mathcal{B}} = V\widehat{\gamma}$, con V la matriz de coeficientes básicos de las funciones propias $\mathscr{F}_v$ in $\mathcal{H}$.

5.4. Resultados de simulación

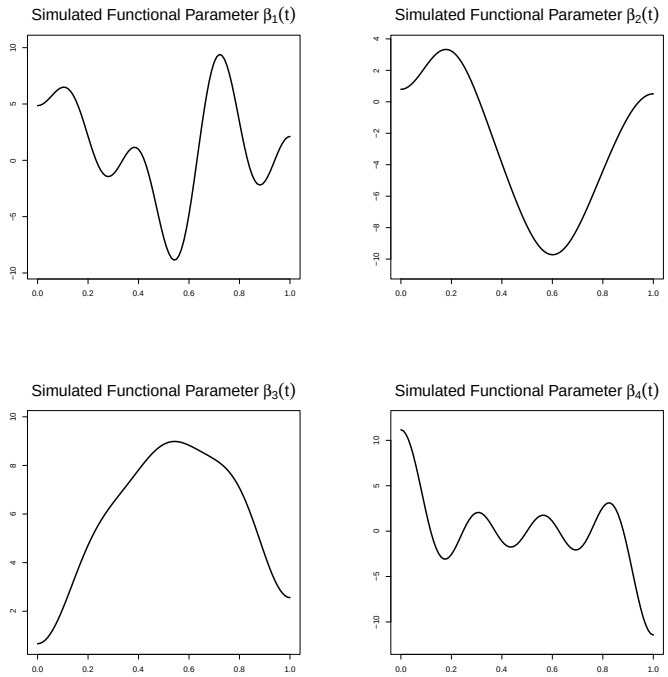


Figura 5.2: Superior izquierda: Función parámetro $\beta_1(t)$ generada a partir de los valores $s = 10$, $w_1 = 15$ y $w_2 = 5$. Superior derecha: función parámetro $\beta_2(t)$ generada a partir de los valores $s = 10$, $w_1 = 3$ y $w_2 = 5$. Inferior izquierda: función parámetro $\beta_3(t)$ generada a partir de los valores $s = 80$, $w_1 = 0.3$, $w_2 = 1.5$. Inferior derecha: función parámetro $\beta_4(t)$ generada a partir de los valores $s = 70$, $w_1 = 25$, $w_2 = 25$

Con el fin de evaluar el comportamiento del método propuesto, se lleva a cabo un estudio de simulación considerando, inicialmente, tres escenarios diferentes:

- Escenario 1: Modelo logístico funcional sin medidas repetidas
- Escenario 2: Modelo logístico funcional con medidas repetidas
- Escenario 3: Modelo logístico funcional con medidas repetidas y multicolinealidad.

Para todos los escenarios, se consideran cuatro diferentes funcionaes parámetro β en el subespacio $\mathcal{H}$ generado por la base Φ_1 descrita en la Ecuación 2.8 truncada en $d = 8$.

Cada función beta es generada a partir de la expresión $s\left(sin(w_1 \cdot t)\right)\left(cos(w_2 \cdot t)\right)$, donde s, w_1, y w_2 son valores que al ser modificados, producen cambios en la escala, oscilación y rugosidad del β simulado. Los cuatro tipos de parámetros funcionales simulados en $\mathcal{H}$ pueden ser vistos en la Figura 5.2

5.4.1. Escenario 1

Para el escenario 1, ya que no se tiene en cuenta la multicolinealidad ni las medidas repetidas, se genera un conjunto de $n = 750$ curvas $\{\mathcal{X}_\iota\}_{\iota=1}^{n}$ a partir de sus coeficientes básicos por medio de una distribución uniforme, i.e. $(a_{\iota,j})_{j=1}^{d} = A_i \sim Unif[0.5, 3]$. Una muestra de 100 curvas simuladas puede ser vista en la Figura 5.3.

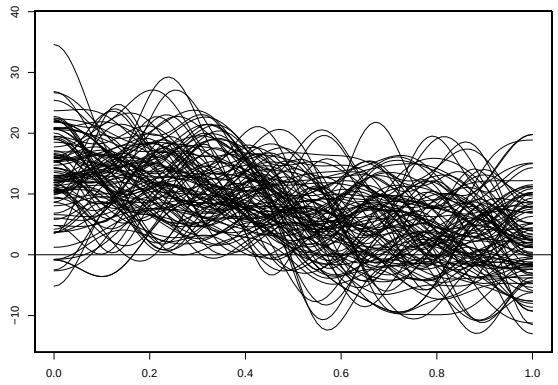


Figura 5.3: Una muestra de 100 datos funcionales en $\mathcal{H}$ simulada a partir de sus coeficientes básicos.

Después de fijar la función parámetro correspondiente β_1, β_2, β_3 o β_4, y simular los datos funcionales del predictor, se calcula el predictor lineal l_ι (para cada observación $\iota-$ésima) dado por la Ecuación 5.42. Para simular la respuesta desde el predictor lineal, se utiliza una distribución Bernoulli cuyas probabilidades están dadas por $\exp(l_\iota)/(1 + \exp(l_\iota))$. Luego, cuatro diferentes modelos son ajustados para estimar el parámetro funcional β.

- Modelo 1: $L = \mathbf{1}\alpha + A\Psi\mathcal{B}$, i.e. el modelo propuesto en la Ecuación 5.40 sin la adición de un efecto aleatorio –llamado modelo clásico (CL_Model)–. Las estimaciones son

obtenidas mediante ML.

- Modelo 2: $L = \mathbf{1}\alpha + A\Psi\mathcal{B} + ZU$, i.e. el modelo propuesto en la Ecuación 5.40 con la adición de un efecto aleatorio –llamado modelo de medidas repetidas (RM_Model) –. Las estimaciones son obtenidas mediantes $REML$.

- Modelo 3: $L = \mathbf{1}\gamma_0 + \Gamma\gamma$ i.e. el modelo propuesto en la Ecuación 5.33 donde no se adiciona un efecto aleatorio –llamado modelo clásico de componentes principales (PC_Model) –. Las estimaciones son realizadas a partir de ML.

- Modelo 4: $L = \mathbf{1}\gamma_0 + \Gamma\gamma + ZU$ i.e. el modelo propuesto en la Ecuación 5.51 con efectos aleatorios –llamado modelo de medidas repetidas en componentes principales (RMPC_Model) –. Se obtienen las estimaciones por medio de $REML$.

En PC_Model y $PCRM\_Model$, el número de componentes principales es fijado como el número de componentes que acumulan un 99 % de la variabilidad explicada. La Figura 5.4 muestra un ejemplo de estimación $\hat{\beta}_1$ del parámetro funcional simulado β_1 bajo los cuatro modelos del Escenario 1.

Para cada uno de los modelos logit, el comportamiento y la precisión de las estimaciones son evaluados mediante el CCR y el ISE descritos en la Ecuación 5.35 y 5.36 respectivamente. No obstante, debido a que para un mismo parámetro funcional, el proceso bajo un mismo modelo se lleva a cabo 100 veces, se obtienen 100 estimaciones funcionales del mismo parámetro, por lo tanto, se opta por utilizar el promedio de CCR e ISE, denominados como $MCCR$ y $MISE$ respectivamente, junto con la desviación estándar del CCR ($SDCCR$). Además, dado que las estimaciones para un mismo parámetro conforman un conjunto de curvas, es posible utilizar la varianza escalar para datos funcionales ($SVFD$) descrita en la Ecuación 3.6 del Capítulo 2 para proporcionar un valor escalar de la variabilidad del conjunto estimaciones funcionales. Esta medida resulta útil en este caso para evaluar el comportamiento de los conjuntos de estimaciones en cada uno de los cuatro modelos asumidos.

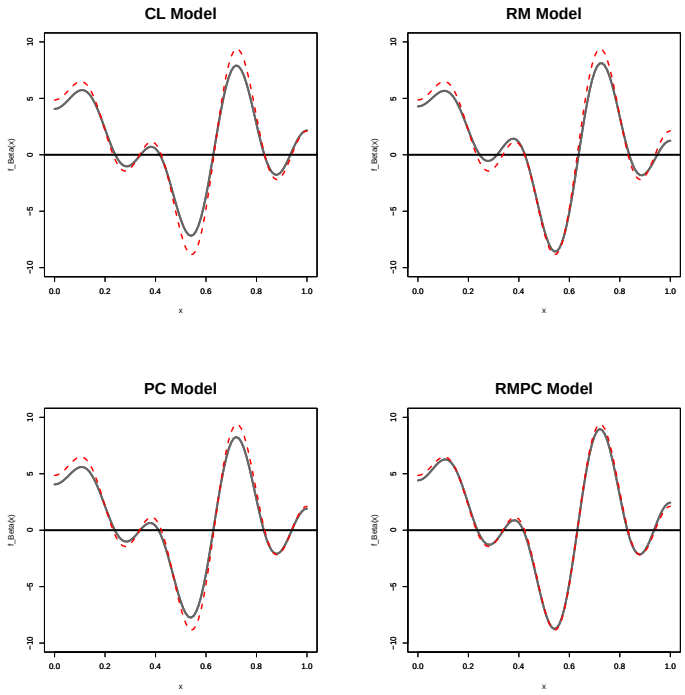


Figura 5.4: Para todas las figuras: En línea punteada roja, el parámetro funcional objetivo β_1; en líneas sólidas negras, las estimaciones funcionales $\hat{\beta}_1$. En la esquina superior izquierda para el primer modelo CL_Model. En la esquina superior derecha para el segundo modelo RM_Model. En la esquina inferior izquierda, para el tercer modelo PC_Model. En la esquina inferior derecha, para el cuarto modelo $RMPC\_Model$.

La Figura 5.5 muestra los resultados de las 100 estimaciones de las funciones parámetros asumidas para los cuatro modelos en el Escenario 1. Aquí, no hay multicolinealidad y no hay estructura de correlación debido a mediciones repetidas del mismo individuo. Como se esperaba en este caso, los cuatro modelos producen estimaciones muy similares, lo que se puede verificar a través de medidas de precisión en la Tabla 5.3 en la esquina superior izquierda. Los resultados gráficos con coherentes con los resultados esperados.

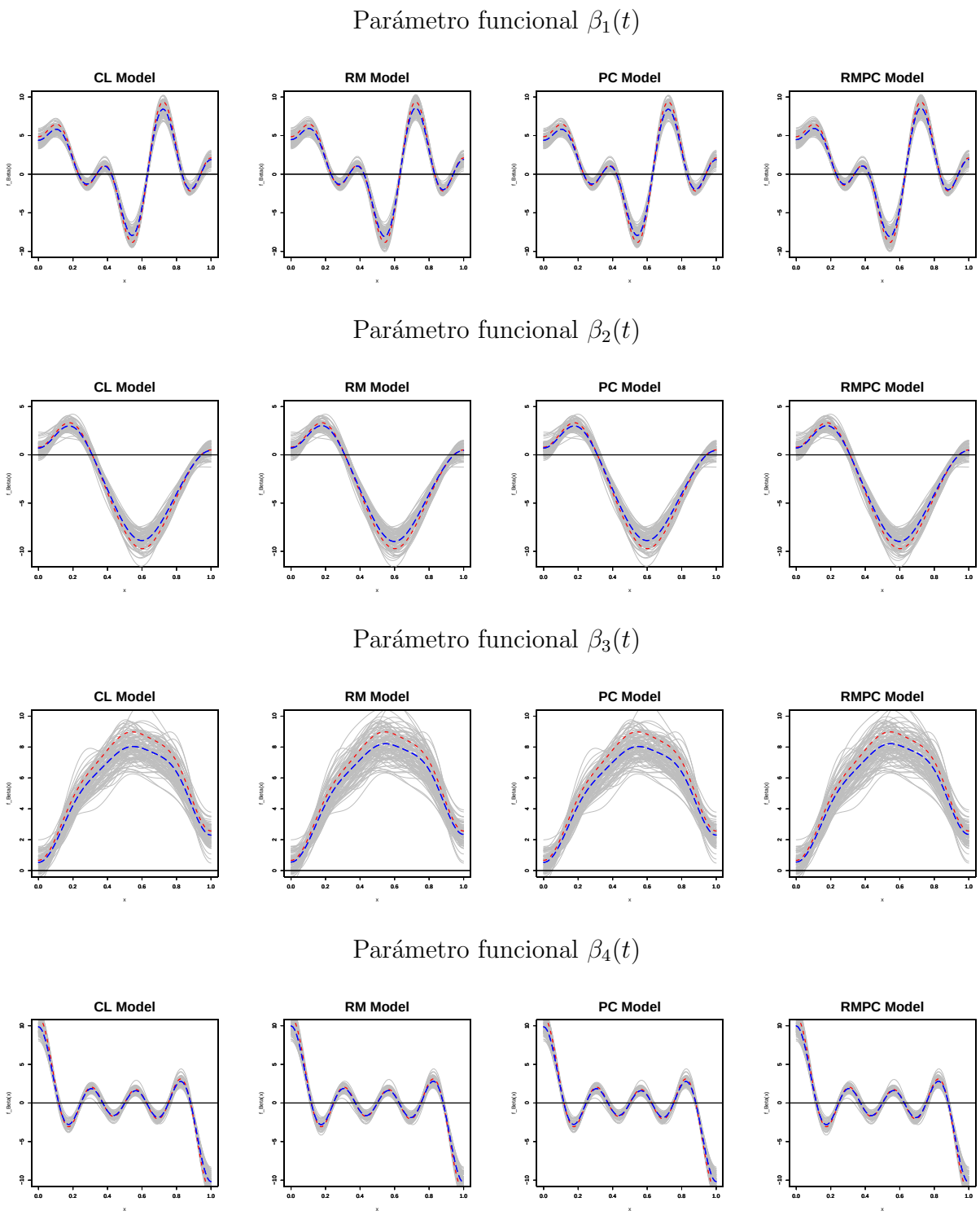


Figura 5.5: Escenario 1: La línea punteada roja muestra el parámetro funcional objetivo, las líneas grises son las 100 estimaciones funcionales, la línea discontinua azul la media funcional de las 100 estimaciones. Cada franja muestra los resultados de los modelos ajustados CL_Model, RM_Model, PC_Model y $RMPC\_Model$.

Con el fin de comparar los ajustes de los cuatro modelos en el escenario 1, la Tabla 5.3 muestra las medidas de precisión para las cuatro funciones parámetro consideradas. Es posible observar que los resultados muestran cierta regularidad incluso con cambios en la forma del parámetro funcional. En todos los casos, las *MCCR* son altas, rondando el 90 %, y hay diferencias casi insignificantes en *SVFD*, *ISB* y *MISE* entre los modelos en los cuatro parámetros funcionales.

Medida de Precisión	Parámetro funcional β_1				parámetro funcional β_2			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
SVFD	0.24	0.27	0.24	0.27	0.29	0.31	0.29	0.31
ISB	0.23	0.16	0.23	0.16	0.21	0.17	0.21	0.17
MISE	0.47	0.43	0.47	0.43	0.50	0.48	0.50	0.48
MCCR	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91
SDCCR	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.02
	Parámetro funcional β_3				Parámetro funcional β_4			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
SVFD	0.47	0.58	0.47	0.58	0.19	0.20	0.19	0.20
ISB	0.43	0.28	0.43	0.28	0.20	0.17	0.20	0.17
MISE	0.90	0.85	0.90	0.85	0.39	0.37	0.39	0.37
MCCR	0.92	0.93	0.92	0.93	0.89	0.89	0.89	0.89
SDCCR	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Tabla 5.3: Medidas de precisión en el Escenario 1 para las cuatro funciones β objetivo.

5.4.2. Escenario 2

En este escenario de simulación, las curvas predictoras son simuladas dentro del mismo subespacio $\mathcal{H}$ generado por la misma base finita Φ que en el Escenario 1. Se asume que se cuenta con $N = 50$ individuos y, para cada uno de ellos, se asume que fueron tomadas $n_i = 15$ medidas repetidas funcionales, en otras palabras, la misma cantidad de repeticiones en cada individuo. Se cuenta entonces con un total de $n = 750$ curvas.

El efecto aleatorio de las medidas repetidas se simula con el uso de una distribución normal, en concreto, $U \sim N(0, 3.5)$. La matriz de covarianza para los coeficientes básicos se genera sin multicolinealidad, pero las respuestas contienen la estructura de repetición como resultado de la inclusión del efecto aleatorio. Dicha respuesta, es simulada en los mismos términos que en el Escenario 1 y, al igual que en el Escenario 1, el proceso se es replicado 100 veces. Los ajustes y la precisión se prueban utilizando las mismas medidas y técnicas que en ese escenario.

La Figura 5.6, muestra los resultados de las 100 estimaciones de los parámetros funcionales para los cuatro modelos en el Escenario 2. Aquí, es posible observar que los modelos CL_Model y PC_Model (que utilizan estimación de ML) muestran un sesgo en la media funcional de las estimaciones, mientras que en los modelos RM_Model y $RMPC\_Model$ (que utilizan estimación de REML) las medias funcionales de las estimaciones están más cercanas. Esto debido a que, en este escenario, no se considera multicolinealidad entre las columnas de la matriz de coeficientes básicos, pero sí una estructura de correlación en la respuesta, causada por la repetición. También, se puede observar cómo la inclusión de las componentes principales funcionales en los modelos, en este escenario, no tiene efecto en la precisión y el sesgo de las estimaciones de los parámetros funcionales en comparación con no incluirlos. Por otro lado, al comparar los resultados obtenidos para los diferentes parámetros funcionales, se puede suponer que la forma de estos podría estar influyendo en la precisión y el sesgo de las estimaciones. Por ejemplo, se puede observar que en parámetros funcionales como β_2 y β_3, las discrepancias al no usar efectos aleatorios en los modelos son mayores que en β_1 y β_4. En cualquier caso, la inclusión de un efecto aleatorio en el modelo parece mejorar las estimaciones.

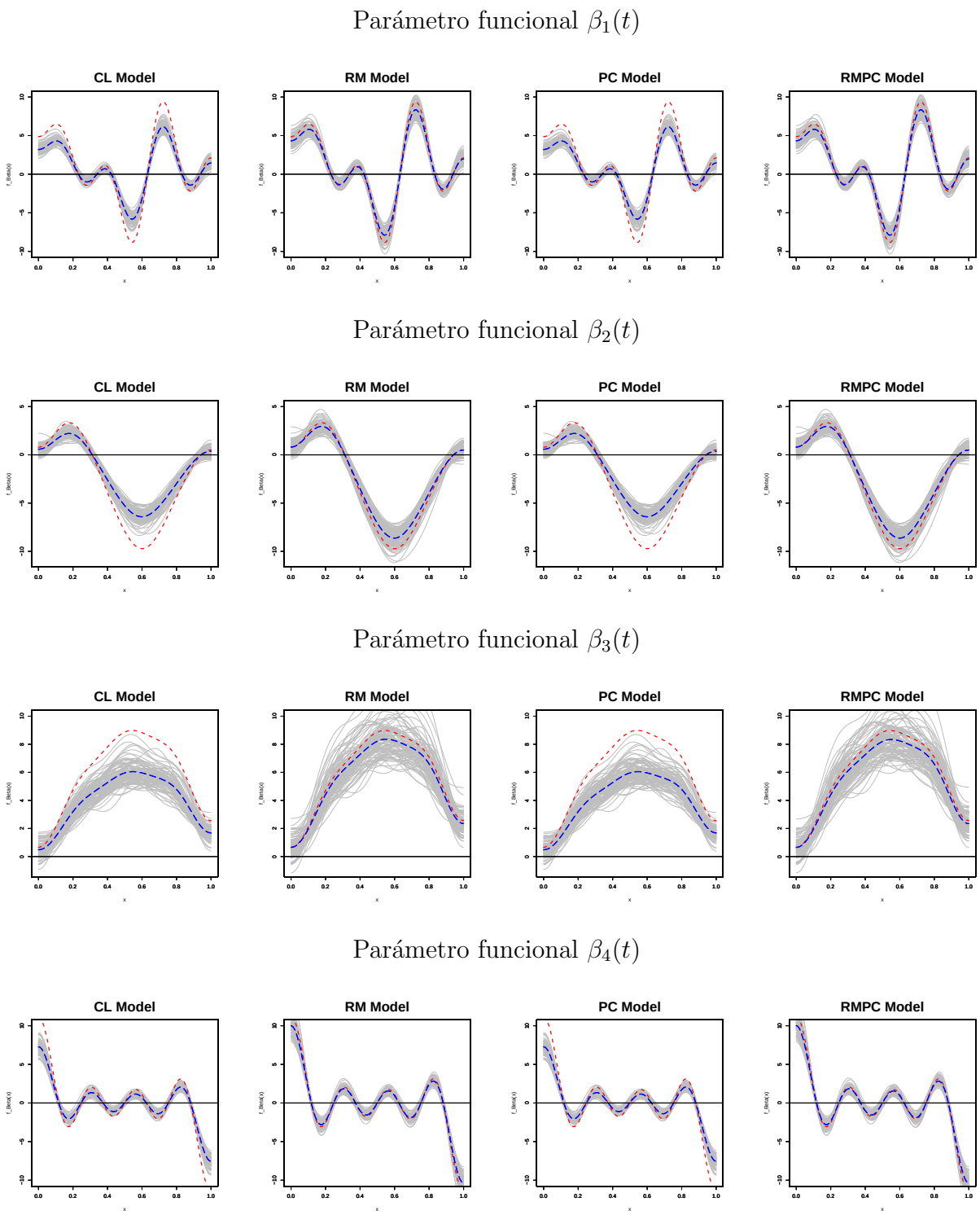


Figura 5.6: Escenario 2: Las figuras, la línea punteada roja muestra el parámetro funcional objetivo, las líneas sólidas grises las 100 estimaciones funcionales, la línea discontinua azul la media funcional de las 100 estimaciones. Cada franja muestra para cada parámetro funcional los resultados de los modelos ajustados CL_Model, RM_Model, PC_Model y $RMPC\_Model$ respectivamente.

Medida de precisión	Parámetro funcional β_1				Parámetro funcional β_2			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
SVFD	0.16	0.29	0.16	0.29	0.19	0.34	0.19	0.34
ISB	**2.65**	**0.26**	2.65	0.27	3.09	0.32	3.09	0.32
MISE	**2.81**	**0.55**	2.81	0.55	3.28	0.66	3.28	0.66
MCCR	0.87	0.92	0.87	0.92	0.88	0.93	0.88	0.93
SDCCR	0.02	0.01	0.02	0.01	0.02	0.01	0.02	0.01
	Parámetro funcional β_3				Parámetro funcional β_4			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
SVFD	0.35	0.92	0.35	0.85	0.12	0.21	0.12	0.21
ISB	4.50	0.18	4.50	0.20	1.89	0.15	1.89	0.15
MISE	4.84	1.09	4.84	1.04	2.01	0.36	2.01	0.36
MCCR	0.90	0.94	0.90	0.94	0.85	0.90	0.85	0.90
SDCCR	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01

Tabla 5.4: Medidas de precisión en el Escenario 2 para los cuatro parámetros funcionales β_i objetivo.

En la Tabla 5.4 (superior izquierda), tal y como se esperaba, la capacidad de predicción de los cuatro modelos es muy similar, la *MCCR* es alta en todos los casos. No obstante, se observa un incremento en el sesgo y el error en las estimaciones de los modelos CL_Model y PC_Model con respecto a estos mismos modelos en el Escenario 1. Este incremento se presenta como una consecuencia de las medidas repetidas. Pese a esto, la reducción de 2.65 a 0.26 en *ISB* y de 2.81 a 0.55 en *MISE*, muestra la importancia de considerar la inclusión de un efecto aleatorio en el modelo logístico funcional para medidas repetidas, con el fin de mejorar las estimaciones del parámetro funcional y obtener una interpretación precisa de este en términos del cociente de ventaja. Los resultados de este escenario, al igual que en el Escenario 1, muestran regularidad en la disminución del sesgo *ISB* y el error promedio *MISE* en los modelos con efecto aleatorio, incluso ante cambios en la forma de la función parámetro β_2, β_3 y β_4. Además, en cuanto a la *SVFD*, al igual que en el Escenario 1,

se observa un aumento de esta en los modelos con efecto aleatorio debido al método de estimación utilizado, pues el REML puede incrementar la varianza al mismo tiempo que produce una reducción del sesgo de las estimaciones. Este efecto es reiterado en las cuatro funciones parámetro ajustadas. Mientras que el excesivo aumento de la varianza para las estimaciones de β_3, en los modelos RM_Model y $RMPC\_Model$, podría estar indicando una influencia de la forma del parámetro funcional.

5.4.3. Escenario 3

En este escenario, se considera la presencia de multicolinealidad en la matriz A de los coeficientes básicos debida a la expansión básica. Así, se utiliza una distribución Normal para simular los coeficientes de las curvas del predictor. En otras palabras, $(a_{is,j})_{j=1}^{d} = A_j \sim N(0, \Sigma), i = 1, 2, \ldots, N$, donde la matriz de covarianza Σ es una matriz definida positiva no diagonal. Como en el Escenario 2, se asume un total de $n = 750$, correspondientes a $N = 50$ y $n_i = 15$ repeticiones funcionales para cada uno de ellos. La respuesta simulada, replicas, ajustes y evaluación de los modelos, son llevados a cabo del mismo modo en como se realiza en los dos escenarios previos a este.

La Figura 5.7 muestra los resultados obtenidos de las estimaciones funcionales en el Escenario 3, en donde se considera multicolinealidad y una estructura de correlación debida a la repetición por medio de la inclusión de un efecto aleatorio. Note que, si bien en este escenario los cuatro modelos tienen dificultad para obtener una correcta estimación del parámetro funcional objetivo, el impacto de la multicolinealidad es mucho mayor en los modelos CL_Model y RM_Model que producen malas estimaciones. Mientras que, los modelos que obtienen la estimación a través de las componentes principales funcionales, como lo son PC_Model y $RMPC\_Model$, producen estimaciones más estables y con una disminución del sesgo, el error y la varianza, en relación a los primeros modelos mencionados. Además, como en el Escenario 2, en este escenario se sospecha que la forma de la función parámetro, puede estar influyendo en la precisión de los ajustes, pues resulta evidente que para las funciones parámetro β_2 y β_3, ningún modelo provee estimaciones satisfactorias. No obstante, es posible concluir que para todos los parámetros

funcionales asumidos, la consideración del efecto aleatorio por si sola es insuficiente si es que existe multicolinealidad entre los coeficientes básicos de las observaciones funcionales del predictor. Por otro lado, se deja entrever que el uso de componentes principales por si solo también resulta insuficiente si el problema subyacente contiene mediciones repetidas. Es entonces la combinación de ambas metodología, la que produce una ganancia significativa en términos de la estimación de la función parámetro. Esto puede corroborarse con los resultados descritos en la Tabla 5.5.

Medida de precisión	Parámetro funcional β_1				Parámetro funcional β_2			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
SVFD	334.90	676.02	2.13	3.83	41.20	82.05	2.21	4.56
ISB	6.30	7.19	7.08	4.00	6.04	2.35	10.42	5.58
MISE	337.82	676.38	9.19	7.80	46.82	83.58	12.61	10.10
MCCR	0.90	0.94	0.89	0.93	0.84	0.91	0.83	0.90
SDCCR	0.03	0.02	0.03	0.02	0.04	0.02	0.04	0.03
	Parámetro funcional β_3				Parámetro funcional β_4			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
SVFD	147.34	165.11	5.75	7.50	1853.57	2692.48	1.24	19.52
ISB	4.18	2.48	11.51	8.84	20.05	22.81	3.56	0.78
IMSE	150.02	165.93	17.20	16.26	1854.89	2688.09	4.79	20.10
MCCR	0.91	0.93	0.90	0.92	0.95	0.97	0.94	0.97
SDCCR	0.03	0.02	0.04	0.03	0.01	0.01	0.01	0.01

Tabla 5.5: Medidas de precisión en el Escenario 3 para los cuatro parámetros funcionales β objetivo.

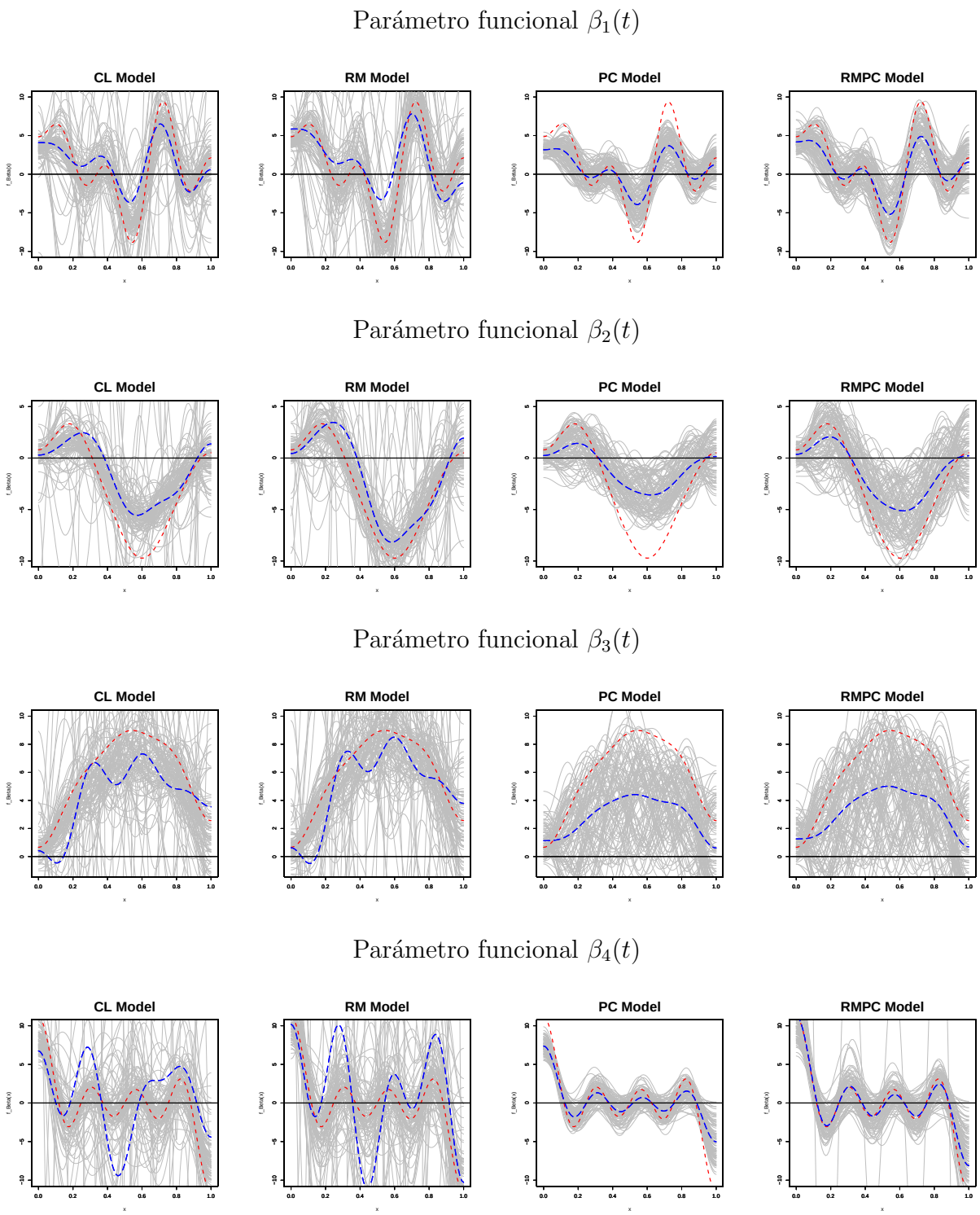


Figura 5.7: Escenario 3: La línea punteada roja muestra el parámetro funcional objetivo, las líneas sólidas grises las 100 estimaciones funcionales, la línea discontinua azul la media funcional de las 100 estimaciones. Cada franja muestra para cada parámetro funcional los resultados de los modelos ajustados CL_Model, RM_Model, PC_Model y $RMPC\_Model$ respectivamente.

5.5. Conclusiones

Es importante recalcar que la contribución fundamental buscada en este capítulo es la estimación adecuada del parámetro funcional, conforme a esto, es posible lo siguiente.

- La inclusión de un efecto aleatorio en el modelo logístico funcional parece efectiva para mejorar la estimación del parámetro funcional en el caso de medidas funcionales repetidas. Como se puede ver en todos los escenarios, la inclusión del efecto aleatorio mejora significativamente la predicción de la respuesta, así como la estimación del parámetro funcional y, por lo tanto, mejora la interpretación.

- El uso de componentes principales funcionales permite mejorar la estimación del parámetro funcional, incluso en el modelo de efectos aleatorios en presencia de multicolinealidad.

- Aunque los resultados bajo simulación muestran cierta consistencia en el rendimiento de los modelos cuando se cambia el parámetro funcional, en presencia de multicolinealidad y medidas repetidas, la forma real del parámetro funcional podría estar influyendo en el modelo a utilizar.

Capítulo 6

Líneas abiertas

Debido a la naturaleza de los datos funcionales de segunda generación, es necesario el desarrollo de metodologías que permitan considerar la posible interdependencia inherente en los diseños que los originan. La investigación que produce los resultados expuestos en este documento avanza en esa dirección. No obstante, esta presenta desafíos en cada uno de los tópicos abordados, lo que lleva a la identificación de posibles líneas de investigación futuras, con el fin de ampliar el conocimiento en este tema.

- Si bien las medidas de resumen escalares presentadas en el Capítulo 3 parecen proveer una forma de medir la variabilidad y la asociación entre conjuntos de observaciones funcionales, resulta necesario ampliar la discusión sobre el significado de estos conceptos en un entorno funcional. Esto se debe a que, a diferencia de lo que ocurre con los datos escalares, los datos funcionales podrían tener diferentes tipos de variabilidad.

- Es necesario llevar a cabo nuevos estudios de simulación para ampliar el entendimiento de los conceptos de varianza y correlación escalar en datos funcionales. Además, aún no sabemos cómo evaluar la confianza de las estimaciones, lo que requiere nuevos experimentos de simulación en esa dirección.

- Aún no se han estudiado las implicaciones de la medida de correlación escalar para datos funcionales en un contexto de regresión funcional ni en los contrastes

de hipótesis para datos funcionales.

- En cuanto al contraste de hipótesis presentado en el Capítulo 4, resulta necesario plantear metodologías dirigidas al estudio de la potencia y su sensibilidad ante diferentes cambios en la distancia y forma de las observaciones funcionales pareadas.

- Para este mismo contraste, es necesario estudiar el efecto de la no reciprocidad del test, así como explorar metodologías que permitan determinar intervalos de confianza del contraste o formas equivalentes.

- Con respecto al modelo logístico funcional de medidas repetidas presentado en el Capítulo 5, los estudios futuros estarán enfocados en examinar la sensibilidad del modelo a los cambios en las estructuras de variabilidad interna de los parámetros funcionales, así como a la forma de la función parámetro.

- En el modelo logístico funcional de medidas repetidas, es necesario evaluar el impacto sobre la estimación de la función parámetro que se produciría al utilizar diferentes tipos de componentes principales funcionales.

- Se puede explorar la inclusión de efectos aleatorios funcionales en el contexto específico de medidas repetidas, para el modelo logístico funcional.

- Resulta necesario explorar el caso ordinal y de respuesta múltiple derivados a partir del modelo logístico de medidas repetidas planteado.

Bibliografía

Acal, C. & Aguilera, A. M. (2023), 'Basis expansion approaches for functional analysis of variance with repeated measures', *Advances in Data Analysis and Classification* **186**, 291–321.

Acal, C., Aguilera, A. M., Sarra, A., Evangelista, A., Di-Battista, T. & Palermi, S. (2022), 'Functional anova approaches for detecting changes in air pollution during the covid-19 pandemic.', *Stochastic Environmental Research and Risk Assessment* pp. 1083–1101.

Agresti, A. (2015), *Foundations of Linear and Generalized Linear Models*, 1 edn, John Wiley & Sons, New Jersey.

Aguilera, A. M., Acal, C. J., Aguilera-Morillo, M. d. C., Jiménez-Molinos, F. & Roldán, J. B. (2021), 'Homogeneity problem for basis expansion of functional data with applications to resistive memories', *Mathematics and Computers in Simulation* **186**, 41–51.

Aguilera, A. M., Escabias, M., Preda, C. & Saporta, G. (2010), 'Using basis expansions for estimating functional pls regression: Applications with chemometric data', *Chemometrics and Intelligent Laboratory Systems* **104**(2), 289–305.

Aguilera-Morillo, M. d. C. & Aguilera, A. M. (2020), 'Multi-class classification of biomechanical data: A functional lda approach based on multi-class penalized functional pls', *Statistical Modelling* **20**, 592–616.

Antoniadisa, A. & Sapatinas, T. (2007), 'Estimation and inference in functional mixed-effects models', *Computational Statistics & Data Analysis* **51**, 4793 – 4813.

Apostol, T. M. (1967), *Calculus. One-Variable Calculus, with an Introduction to Linear Algebra*, Vol. 1, 2nd edn, John Wiley and Sons, Inc., New York.

Carrell, j. B. (2010), *Groups, Matrices and Vector Spaces: A Group of Theoretic Approac to Linear Algebra*, Springer.

Chen, H. & Wang, Y. (2011), 'A penalized spline approach to functional mixed effects model analysis', *Biometrics* **67**, 861 – 870.

Cohn, D. (2013), *Measure theory* , 2nd edn, Birkhauser, Boston.

Cox, D. D. Lee, J. S. & Follen, M. (2015), ' A two sample test for functional data. ', *Communications for Statistical Applications and Methods* **22**, 121–135.

Crowder, M. & Hand, D. (1990), *Analysis of Repeated Measures (1st ed.)*, Chapman and Hall.

Cuadras, C. M. (1996), *Nuevos métodos de análisis multivariante*, CMC Edicions, Barcelona, Spaña.

Cuesta-Albertos, J. & Febrero-Bande, M. (2010), 'A simple multiway anova for functional data', *Test* **19**, 537–557.

Cuesta-Albertos, J., Fraiman, R. & Ransford, T. (2007), 'A sharp form of the cramér-wold theorem', *Journal of Theoretical Probability* **20**, 201–209.

Cuesta-Albertos, J., García-Portugués, E., Febrero-Bande, M. & Gonzalez-Manteiga, W. (2017), 'Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes', *Annals of Statistics* **47**.

Cuevas, A. Febrero, M. & Fraiman, R. (2004), ' A anova test for functional data ', *Computational Statistics and Data Analysis* **47**, 111–122.

Dallal, A. A., AlDallal, U. & Dallal, J. A. (2021), 'Positivity rate: an indicator for the spread of covid-19', *Current Medical Research and Opinion* **37**(12), 2067–2076.

Davis, C. S. (2002), *Statistical Methods for the Analysis of Repeated Measurements*, Springer-Verlag New York, Inc.

Degras, D. (2014), 'Simultaneous Confidence Bands for the Mean of Functional Data. ', *R package version* **1**.

Dobson, A. J. & Barnett, A. G. (2008), *An Introduction to Generalized Linear Models*, third edn, John Wiley & Sons, New York.

Dodge, Y. (2008), *The Concise Encyclopedia of Statistics*, Springer, New York, NY, pp. 15–18.

Escabias, M., Aguilera, A. M. & Acal, C. (2022), 'logitfd: an r package for functional principal component logit regression', *R journal* **14**(3), 231–248.

Escabias, M., Aguilera, A. M. & Valderrama, M. J. (2004), 'Principal component estimation of functional logistic regression: Discussion of two different approaches', *Journal of Nonparametric Statistics* **16**, 365–384.

Escabias, M., Valderrama, M. J., Aguilera, A. M., Santofimia, M. H. & Aguilera-Morillo, M. C. (2013), 'Stepwise selection of functional covariates in forecasting peak levels of olive pollen.', *Stochastic Environmental Research and Risk Assessment* **27**, 367–376.

Eubank, R. L. (1998), *Nonparametric Regression and Spline Smoothing; Second Edition* , Marcel Dekker Inc, New York.

Fan, J. & Lin, S. (1998), ' Test of significance when data are curves ', *American Statistics. Assoc* **93**, 111–122.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis Theory and Practice* , Springer.

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions Of The Royal Society A. Mathematical, Physical and Engineering Sciences* **222**, 309–368.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Oliver and Boyd, London.

Fomby, T. B., Johnson, S. R. & Hill, R. C. (1984), *Advanced Econometric Methods* , Springer New York, New York, NY.

Fu, Y., Li, Y., Guo, E., He, L., Liu, J., Yang, B., Li, F., Wang, Z., Li, Y., Xiao, R., Liu, C., Huang, Y., Wu, X., Lu, F., You, L., Qin, T., Wang, C., Li, K., Wu, P., Ma, D., Sun, C. & Chen, G. (2021), 'Dynamics and Correlation Among Viral Positivity, Seroconversion, and Disease Severity in COVID-19', *Annals of Internal Medicine* **174**(4), 453–461.

Galton, F. (1889), 'Co-relations and their measurement, chiefly from anthropometric data', *Proceedings of the Royal Society of London* **45**, 273–279.

Grenander, U. (1950), 'Stochastic processes and statistical inference', *Arkiv för Matematik* **1**(3), 195–277.

Guo, W. (2002), 'Functional mixed effects models', *Biometrics* **58**(1), 121–128.
URL: *http://www.jstor.org/stable/3068297*

Hedeker, D. & Gibbons, R. D. (2006), *Longitudinal Data Analysis*, 1 edn, John Wiley & Sons, New Jersey.

Horvath, L., Huskova, M. & Rice, G. (2013), 'Test of independence for functional data', *Journal of Multivariate Analysis* **117**, 100–119.

Horvath, L. & Rice, G. (2015), 'An introduction to functional data analysis and a principal component approach for testing the equality of mean curves', *Mat Complut* **28**, 505–548.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied logistic regression*, 2 edn, John Wiley & Sons, New York.

Hsing, T. & Eubank, R. (2015), *Theoretical foundations of functional data analysis, with an introduction to linear operators*, first edn, Wiley.

James, G. M. (2002), 'Generalized linear models with functional predictors', *Journal of the Royal Statistical Society B* **64**(3), 411 – 432.

Jiang, Q., Huskova, M., Simos, G. & Zhu, L. (2019), 'Asymptotics, finite-sample comparisons and applications for two-sample tests with functional data', *Journal of Multivariate Analysis* **170**, 202–220.

Kaplansky, I. (1977), *Set Theory and Metric Spaces*, 2nd edn, Ams Chelsea Publishing.

Kenny, J. F. & Keeping, E. S. (1951), *Mathematics of Statistics. Part Two*, 2nd edn, D. VAN NOSTRAND COMPANY, INC.

Kenny, J. F. & Keeping, E. S. (1954), *Mathematics of Statistics. Part One*, 3rd edn, D. Van Nostrand Company, Inc.

Koner, S. & Staicu, A.-M. (2023), 'Second-generation functional data', *Annual Review of Statistics and its Application* **10**, 547–572.

Kreyszig, E. (1989), *Introductory Functional Analysis with applications*, 2006 edn, Wiley.

Larsen, K., Petersen, J. H., Budtz-Jørgensen, E. & Endahl, L. (2000), 'Interpreting parameters in the logistic regression model with random effects', *Biometrics* **53**(6), 909–914.

Lindsey, J. (2014), *Applying Generalized Linear Models*, Springer, New Jersey.

Lindstrom, M. J. & Bates, D. M. (1990), 'Nonlinear mixed effects models for repeated measures data', *Biometrics* **46**(3), 673–687.

Liu, Z. & Guo, W. (2012), 'Functional mixed effects models', *WIREs Computational Statistics* **4**(6), 527–534.

Ma, W., Xiao, L., Liu, B. & Lindquist, M. A. (2019), 'A functional mixed model for scalar on function regression with application to a functional MRI study', *Biostatistics* **22**(3), 439–454.

Martínez-Camblor, P. & Corral, N. (2011), 'Repeated measures analysis for functional data', *Computational Statistics & Data Analysis* **55**(12), 3244–3256.

Mathai, A. M. & Rathie, P. N. (1977), *Probability and Statistics*, The Macmillan Press Ltd.

Melendez, R., Giraldo, R. & Leiva, V. (2021), 'Sign, wilcoxon and mann-whitney tests for functional data: An approach based on random projections', *Matematics* **44**(9), 1–11.

Pomann, G.-M., Staicu, A.-M. & Ghosh, S. (2016), 'A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis ', *Journal of the Royal Statistical Society* **65**(3), 395–414.

Ramsay, J. (1982), 'When the data are functions', *Psychometrika* **47**(4), 379–396.

Ramsay, J. O. & Dalzell, C. J. (1991), 'Some tools for functional data analysisl', *Journal of the Royal Statistical Society* **3**, 593–572.

Ramsay, J. O. & Silverman, B. W. (2005), *Funcional Data Analysis*, 2da. edn, Springer.

Rao, C. R. (1958), 'Some statistical methods for comparison of growth curves', *Biometrics* **14**, 1–17.

Restrepo, J. (2022), 'Colombia 2021: Between Crises and Hope', *Revista de Ciencia PolÃtica* **42**(2), 255–280.

Rudin, W. (1976), *Principles of Mathematical Analysis*, 3rd edn, McGraw-Hill Inc.

Scheipl, F., Staicu, A. M. & Greven, S. (2015), 'Functional additive mixed models', *Journal of Computational and Graphical Statistics* **24**(2), 477–501.

Shen, Q. & Faraway, J. (2004), ' An f test for linear models whit functional responses ', *Statistica Sinica* **14**, 1239–1257.

Smaga, L. (2020), 'A note on repeated measures analysis for functional data', *AStA Adv Stat Anal* **104**(2), 117–139.

Todorovic, P. (1992), *An Introduction to Stochastic Processes and Their Applications*, Springer.

Urbano-Leon, C. L., Aguilera, A. M. & Escabias, M. (2024), 'Repeated measures in functional logistic regression', *Mathematics and Computers in Simulation* **225**, 66–77. **URL:** *https://www.sciencedirect.com/science/article/pii/S0378475424001757*

Urbano-Leon, C. L. & Escabias, M. (2022), 'Comparison of positivity in two epidemic waves of covid-19 in colombia with fda', *Stats* **5**(4), 993–1003.

Urbano-Leon, C. L., Escabias, M. & Ovalle-Muñoz, Diana Paola Olaya-Ochoa, J. (2023), 'Scalar variance and scalar correlation for functional data', *Mathematics* **11**(1317), 1–20.

Wang, J. L., Chiou, J. M. & Müller, H. G. (2016), 'Review of functional data analysis', *Annual Review of Statistics and its Application* pp. 1–41.

World Health Organization (2005), *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide.*, World Health Organization, United States.

Zhang, J. & Chen, J. (2007), 'Statical Inferences For Functional Data ', *The Annals of Statistics* **35**(3).

Apéndice A

Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA

Titulo: *"Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA".*

Autores: *Cristhian Leonardo Urbano-Leon, Manuel Escabias.*

Revista: *Stats.*

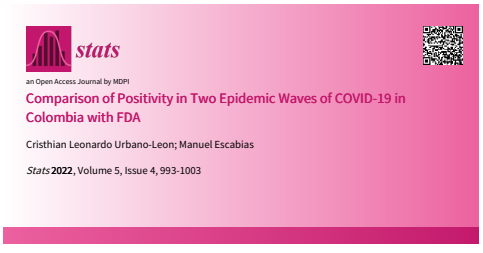
Factores de impacto de la revista:

Estado: Publicado, 28 de octubre de 2022.

doi: 10.3390/stats5040059

Posición de autor: Autor principal.

Año	Factor de Impacto	Rango	Cuartil	Área
2022	1.3	104/164	$Q3$	Estadística y Probabilidad



stats

MDPI

Article

Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA

Cristhian Leonardo Urbano-Leon † and Manuel Escabias *,†

Department of Statistics and Operations Research, University of Granada, 18071 Granada, Spain
* Correspondence: escabias@ugr.es
† These authors contributed equally to this work.

Abstract: We use the functional data methodology to examine whether there are significant differences between two waves of contagion by COVID-19 in Colombia between 7 July 2020 and 20 July 2021. A pointwise functional *t*-test is initially used, then an alternative statistical test proposal for paired samples is presented, which has a theoretical distribution and performs well in small samples. Our statistical test generates a scalar *p*-value, which provides a global idea about the significance of the positivity curves, complementing the existing punctual tests, as an advantage.

Keywords: COVID-19; FDA; hypothesis testing for paired functional data; Colombian national strike

Citation: Urbano-Leon, C.L.; Escabias, M. Comparison of Positivity in Two Epidemic Waves of COVID-19 in Colombia with FDA. *Stats* **2022**, *5*, 993–1003. https://doi.org/10.3390/stats5040059

Academic Editor: Wei Zhu

Received: 2 October 2022
Accepted: 24 October 2022
Published: 28 October 2022

1. Introduction

Functional data analysis (FDA) is a branch of statistics that has had great development in recent years due to its multiple applications in different fields of science [1–7]. One of the reasons for its popularity is that all consecutive observations of a continuous phenomenon can be viewed as a single curve, since the objective of this field is to analyze sets of curves. FDA has a wide range of descriptive and inferential techniques to accomplish this [8,9].

During the global emergency caused by COVID-19, the Colombian government chose the positivity rate as a key variable for early decisions regarding the management of disease, which can be calculated over multiple periods (daily, weekly, monthly, etc.). In this study, the rate of positivity refers to the daily percentage of positive COVID-19 tests for the total number of tests processed. Its trend helps determine the presence of a wave epidemic outbreak [10,11].

Because the positivity rate is related to waves of contagion [10,12], we assume that confirmed coronavirus cases in Colombia in one year serve to identify the most critical moments of a trend change in that time. As the information discriminated by departments in Colombia was only reported from July 2020, we have usable information from 19 July 2020 onward; we can see several contagion waves, but we focus our attention on two of them, depicted in Figure 1.

In the year 2021, Colombia witnessed strong protests. People took to the streets due to multiple social factors, which prevented isolation behavior and other care measures against COVID-19. This fact may have prolonged the duration of the contagion wave as well as its magnitude, since the protests began in 28 April 2021 and lasted for more than two months [13].

Because of this, we consider two case studies. In the first case, we assume that the two waves of contagion have the same duration. However, in the second case, we consider that both waves have a different duration. The purpose of this study is to determine whether there are significant differences between the first and second waves in each case study separately. Because the data are continuous and additionally paired, we used the FDA methodology with a pointwise functional *t*-test, followed by a hypothesis testing proposal based on the integral of the difference in positivity rate curves, to test the equality of functional means.

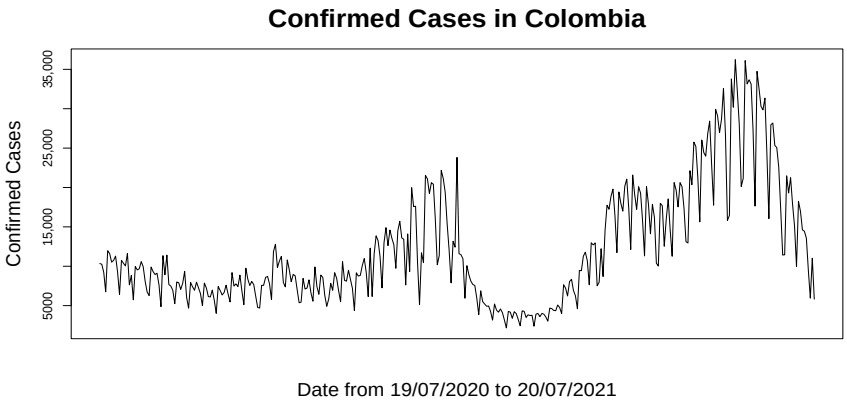


Figure 1. Two waves of contagion by COVID-19 between the dates 19 July 2021 and 20 July 2022.

The rest of the paper is composed as follows: Section 2 offers a contextualization of the data used in the study, contextualization of the functional data, the use of the punctual *t*-test, and the theoretical component of our proposed test, as well as the basis of our simulation; Then, Section 3 shows the functional data built and the results of the tests carried out. Subsequently, Section 4 presents some comments on the results and, finally, we offer our acknowledgments.

2. Materials and Methods

In this section, we present a contextualization on topics of the article, and we propose a hypothesis testing for functional data.

2.1. About the COVID-19 Dataset in Colombia and the Two Case Studies

In Colombia, data on COVID-19 are officially reported by the National Health Institute (NHI) of the Colombian Ministry of Health. However, Colombia is a country divided in 32 regional sections called departments and one special capital district called Bogotá D.C. Each of these reports information on the number and type of tests performed and the number of positive cases to the NHI. This information is not properly reported in some cases, thus leaving a problem of incomplete information. In this study, we only consider departments whose information—from 20 July 2020, to 20 July 2021—is complete. Therefore, we have 23 departments whose information we use and, including Bogotá D.C., 24 regions in total. They are: Antioquia, Atlántico, Bolivar, Boyaca, Cáldas, Casanare, Cauca, Cesar, Córdoba, Cundinamarca, Guajira, Huila, Magdalena, Meta, Nariño, Norte de Santander, Putumayo, Quindío, Risaralda, Santander, Sucre, Tolima, Valle del Cauca, and Bogotá D.C.

It is suspected that the national strike that occurred in Colombia on 28 April 2021, and lasted for at least two months [13], caused people not to take personal protection measures against COVID-19 and that this lengthened the duration of the wave of contagion by SARS-CoV-2 in Colombia. Two study scenarios were considered. The first, called Case 1, assumes the measurements for the first wave of contagion from 20 November 2020, to 18 February 2021; and a second contagion wave from 19 February 2021 to 20 May 2021. That is, both waves of contagion last three months each, ignoring the national strike. The second scenario, called Case 2, assumes the first wave of contagion from 20 November 2020 to 18 February 2021, and the second contagion wave from 19 February 2021 to 20 July 2021. In other words, in the second scenario, the first wave lasts three months, and the second wave lasts 5 months. See Figures 2 and 3.

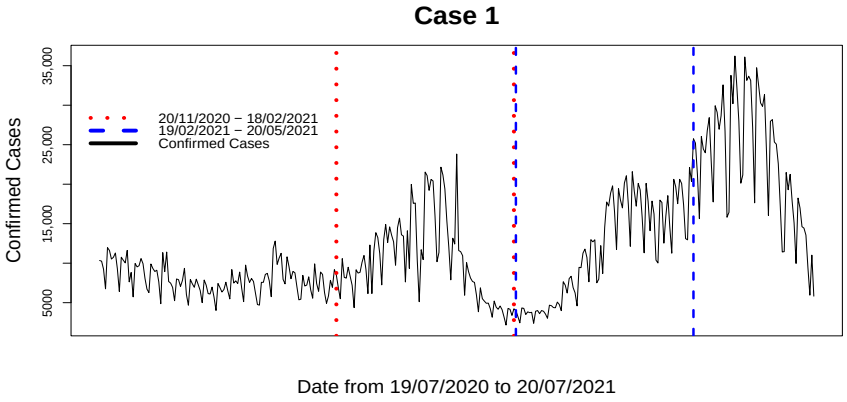


Figure 2. Case 1: Two waves of contagion by COVID-19 in Colombia without a national strike. The dotted red lines mark the start and end of the first wave of contagion, which occurred between 20 November 2020 and 18 February 2021. The dashed blue lines mark the start and end of the second wave of contagion, which occurred between 19 February 2021 and 20 May 2021.

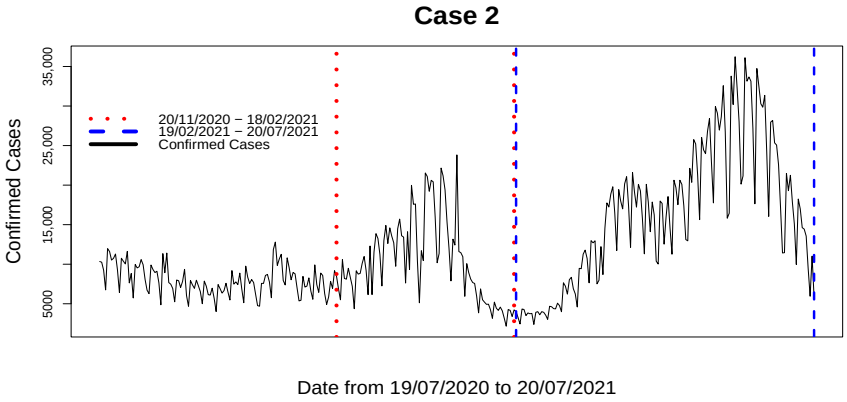


Figure 3. Case 2: Two waves of contagion by COVID-19 in Colombia with a national strike. The dotted red lines mark the start and end of the first wave of contagion, which occurred between 20 November 2020 and 18 February 2021. The dashed blue lines mark the start and end of the second wave of contagion, which occurred between 19 February 2021 and 20 July 2021.

It is important to note that, for each case, the data from the first and second waves have a paired behavior. For each department, the first wave is followed by the second, which constitutes before and after observations.

2.2. About Functional Data Analysis

Functional data analysis is a statistical methodology in which the objects of study are not scalar values, but continuous functions [7,14], considered as observations of a stochastic process $\{X(t) : t \in T\}$. Thus, the set of functional observations $x_1(t), x_2(t), \ldots, x_n(t)$ constitute a simple random sample of it, and each observation is called a functional datum.

The FDA allows the appropriation of a certain mathematical theory about the functions, collected in the functional analysis, since the functional observations considered are smooth curves and square-integrable. That is, if $\{x_i(t)\}_{i=1}^n$ is a functional random sample, associated with the stochastic process $\{X(t) : t \in T\}$, so that Equation (1) holds for $i = 1, 2, \ldots n$.

$$\int_T x_i^2(t)dt < \infty. \tag{1}$$

Thus, the functional data are elements of a Hilbert vector space over the field $\mathbb{R}$ of the real numbers. This space is denoted $\mathcal{L}_2([a,b])$, where the interval $[a,b]$, is the domain of the elements of $\mathcal{L}_2$, which, without loss of generality, can be moved to the interval [0, 1], [15,16].

In FDA, the underlying stochastic process $\{X(t) : t \in T\}$ is defined as a second-order stochastic process [17], so its expected value exists in its functional form, defined as in Equation (2):

$$\begin{aligned} \mu: \quad T &\longrightarrow \quad \mathbb{R} \\ t &\longrightarrow \quad \mu(t) = \int_{\Omega} X(t,\omega) dP(\omega), \end{aligned} \tag{2}$$

on the probability space $(\Omega, \mathcal{A}, P)$.

Although the FDA considers that observations of the stochastic process are continuous functions, the curves of these functions must be obtained through punctual observations of the phenomenon. For this, there are different methodologies; however, we use the vector space structure of $\mathcal{L}_2([0,1])$ to assume that the observations are elements of a finite-dimensional subspace $\mathcal{H}$ of the space $\mathcal{L}_2([0,1])$. This allows us to assume the existence of a finite basis $B = \{\phi_j(t)\}_{j=1}^{k}$ of size k for the subspace $\mathcal{H}$, where each element of B is a basis function, and k is the dimension of $\mathcal{H}$. Thus, each functional datum $x_i(t)$ is uniquely expressed as a linear combination of elements of B, and elements of $\mathbb{R}$ called coefficients, as in Equation (3):

$$x_i(t) = \sum_{j=1}^{k} c_{i,j}\phi_j(t), \tag{3}$$

where $x_i(t)$ is the i-th functional datum, $\phi_j(t)$ is the j-th basis function, $c_{i,j}$ is the j-th coefficent for the i-th functional datum, for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, and $t \in [0, 1]$.

Obtaining the values of $c_{i,j} \in \mathbb{R}$ are carried out in this work by least squares [1]. This allows for an estimate of the mean function $\mu(t)$, through the Expression (4):

$$\bar{\mathcal{X}}(t) = n^{-1} \sum_{i=1}^{n} x_i(t), \tag{4}$$

where $\{x_i(t)\}_{i=1}^{n}$ is a set of functional data [8].

There are different types of bases to generate functional data. In this study, we consider a basis of functions of the form defined as in Equation (5):

$$\phi_j(t) = \sqrt{2}Cos(j-1)\pi t, \tag{5}$$

for $j = 2, 3, \ldots, k$, since, together with the constant function $\phi_1(t) = 1$, the set of functions $\{\phi_j\}_{j=1}^{k}$ constitutes a finite orthonormal basis for a vector subspace of dimension k of the Hilbert space $\mathcal{L}_2([0,1])$ [18,19].

2.3. Hypothesis Testing in Functional Data

A hypothesis test for functional data stems from the same theoretical foundation as a hypothesis test for scalar data. Accordingly, an initial hypothesis is generated about a population parameter, known as the null hypothesis and denoted as H_0, which is contrasted with a hypothesis, generally complementary about the same parameter, called the alternative hypothesis and denoted as H_1 [20].

Since the objective of our study is to determine the existence or not of significant differences between the two waves of contagion of COVID-19, and as the positivity rate has a continuous behavior, the functional data methodology is used. We use the functional mean as a parameter to define a hypothesis test, defining as μ_X and μ_Y the functional means of the stochastic processes $\{X(t) : t \in T\}$ and $\{Y(t) : t \in T\}$ associated with the positivity

rate of COVID-19 in the first and second waves of contagion, respectively. With this, the contrast of hypotheses is raised in Equation (6):

$$\begin{aligned} H_0: &\quad \mu_X &= \mu_Y \\ H_1: &\quad \mu_X &\neq \mu_Y. \end{aligned} \tag{6}$$

Based on the data samples, a statistical test is generated and calculated. The value of the statistic is located within the null distribution, which is the probability distribution that would apply to the statistic if the null hypothesis was correct; next, using the null distribution, a p-value is calculated, which indicates the probability that the test statistic is at least as extreme as the observed statistic [20].

On the tests of hypotheses in functional data, one can find very diverse literature from different approaches such as the [21–28]. However, as can be seen from early work on functional data, hypothesis testing for functional data can be performed using a pointwise t-test [1], based on the idea of fixing a value $t \in [0,1]$. Thus, the hypothesis test of the Equation (6) is performed for each of the infinite points $t \in [0,1]$, and since the values of the images of the functions $x_i(t)$ and $y_i(t)$ are scalars for each fixed t, application of a t-test for scalar data is allowed.

We make the statistical comparison in a first instance with a pointwise t-test, which is a natural extension of a t-test but now in the functional context. This methodology has the limitation that, when performing the scalar tests on the domain values $[0,1]$, the p-value is a continuous function, so it is difficult to generate a global conclusion on the contrast. For this reason, we now propose a different approach to hypothesis testing for functional data, which produces a global p-value that helps decide on the entire domain and not about sections of it.

2.4. Another Hypothesis Test Approach for Functional Data

In our case study, the data of interest, in addition to being of a continuous nature, exhibit paired behavior. That is, for each department, there is a curve of the first wave and a curve of the second wave. Therefore, we have 24 pairs of curves. Thus, we proceed as in the scalar case and restate the contrast as in Equation (7):

$$\begin{aligned} H_0: &\quad \mu_X - \mu_Y &= 0_{\mathcal{L}_2([0,1])} \\ H_1: &\quad \mu_X - \mu_Y &\neq 0_{\mathcal{L}_2([0,1])}, \end{aligned} \tag{7}$$

where in this case $0_{\mathcal{L}_2([0,1])}$ refers to the zero function in $\mathcal{L}_2([0,1])$.

The similarity of the difference curve of any two continuous functions with the zero function is an indicator that both functions are similar. Therefore, if the null hypothesis is true, the integral of the difference curve must be zero, and we can obtain the contrast of Equation (8):

$$\begin{aligned} H_0': &\quad \textstyle\int_T (\mu_X(t) - \mu_Y(t))dt &= \displaystyle\int_T 0_{\mathcal{L}_2(T)} dt \\ H_1': &\quad \textstyle\int_T (\mu_X(t) - \mu_Y(t))dt &\neq \displaystyle\int_T 0_{\mathcal{L}_2(T)dt}. \end{aligned} \tag{8}$$

For the hypothesis test of Equation (8), we present a test statistic based on the average of the integral of the functional differences, denoted by the acronym for Mean Integral of Differences (MID), which is arrived at by using a bit of algebra on the sample estimates of the parameters. Thus, given two sets of functional data of a paired nature $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, the MID contrast statistic is defined by Equation (9):

$$MID = n^{-1} \sum_{i=1}^{n} \int_0^1 d_i(t)dt, \tag{9}$$

where $d_i(t)$ is defined as in Equation (10):

$$d_i(t) = x_i(t) - y_i(t), \tag{10}$$

for each $i = 1, 2, \ldots, n$.

The form of our proposed statistical test, $MID \sim N(0, \sigma)$, follows a normal distribution with mean zero and variance σ^2 , guaranteed by the central limit theorem. The contrast can be done by standardizing MID using Expression (11):

$$S.MID = \frac{MID - 0}{\sigma}. \tag{11}$$

Thus, σ can be obtained from Equation (12):

$$S_d = \frac{\sigma}{\sqrt{n}}, \tag{12}$$

where S_d^2 is the sample variance obtained from set $\left\{ \int_0^1 d_i(t)dt \right\}_{i=1}^{n}$. In this way, a scalar-value can be computed as $2P(Z \geq |S.MID|)$, where Z is a real random variable, such that $Z \sim N(0, 1)$.

In addition to the theoretical approach, to apply our contrast statistic to the specific study cases, we decided to also run a simulation process to find a null distribution and perform the test—considering that the null hypothesis is that the two paired functional samples come from populations with the same mean. Thus, we simulate paired scalar points from a common functional mean for the two groups and use them to construct the curves that constitute the functional data samples of size 24; then, we apply the test statistic. The process is repeated 4000 times for that sample size.

Moreover, a quick Shapiro–Wilk test is performed on the values of the integral of the differences of the functions to assess whether there is evidence that the resulting data do not come from a normal distribution. The p-value for this test is reported later in the results section, which supports the use of the proposed methodology.

3. Results

In this section, we present the results of our study applying the FDA methodology to the positivity data for COVID-19 in Colombia in both cases considered.

3.1. Constructed Functional Data

In Figure 4, we show the functional data of the COVID-19 positivity rate in Colombia in case 1, using the orthonormal basis of the Equation (5). Here, in the top left and right panels, functional data of positivity rate are shown for the first and second waves, respectively; meanwhile, in the bottom left panel, the functional means of positivity rate of both waves of contagion by COVID-19 are shown, which are the goal of the comparison, to be conducted through the curves of difference of the positivity ratio in both waves of contagion by COVID-19, shown in the lower right panel.

Similarly, the functional data of COVID-19 positivity rate in Colombia in case 2 are depicted in Figure 5. We respectively show in the upper left and right panels the functional data of positivity rate for the first and second waves. It is possible to appreciate a certain difference in the trend of positivity between the two waves. This can also be seen in the functional means of both waves of contagion by COVID-19 shown in the bottom left panel of both waves of contagion by COVID-19, which are the goal of comparison in case 2. In addition, a different trend is observed between the curves of the differences in the rate of positivity shown in the figure, with respect to the curves of difference in case 1.

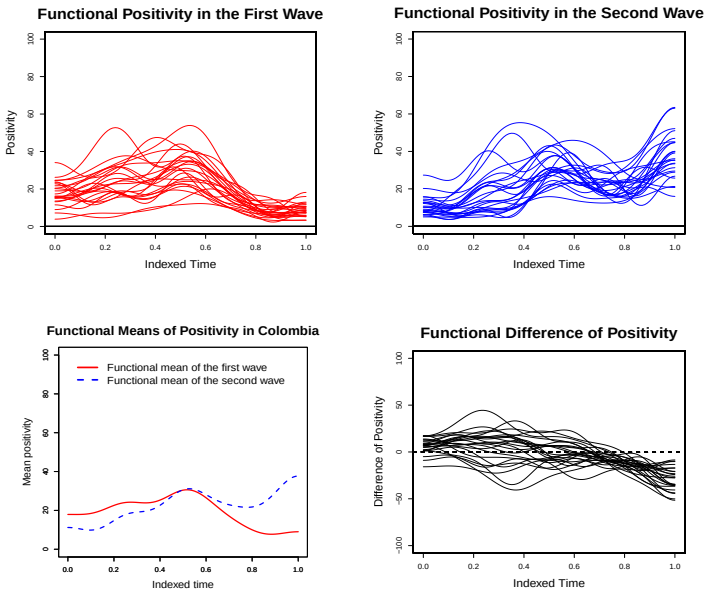


Figure 4. Functional positivity data for COVID-19 in Colombia for Case 1: Functional positivity in the first wave (**top left**); functional positivity in the second wave (**top right**); functional means of positivity in the first wave in solid line and the second wave in dashed line (**bottom left**); positivity difference curves (**bottom right**).

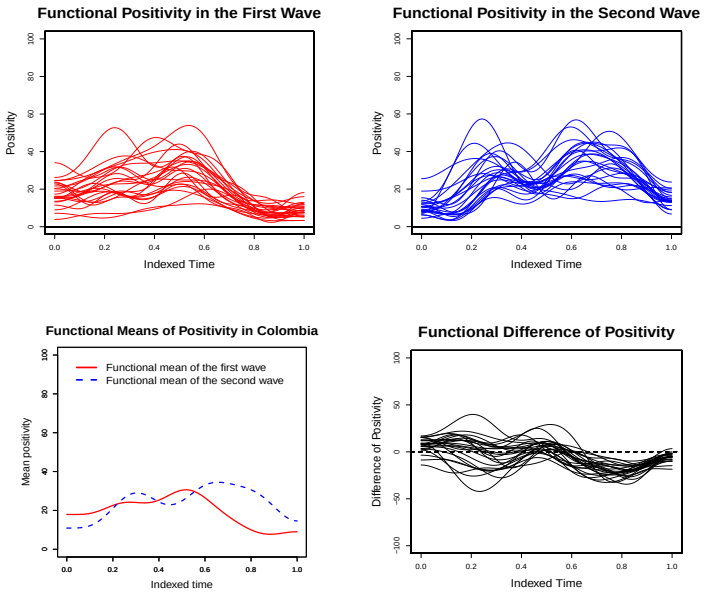


Figure 5. Functional positivity data for COVID-19 in Colombia for Case 2: Functional positivity in the first wave (**top left**); functional positivity in the second wave (**top right**); functional means of positivity in the first wave in solid line and the second wave in dashed line (**bottom left**); positivity difference curves (**bottom right**).

3.2. Pointwise Hypothesis Contrast for Curves

As stated above, the pointwise t-test assumes that, for each $t \in [0,1]$, a scalar t-test can be performed with the images of the functions evaluated at point t. That is, a t-test for the two groups of scalar values $\{x_i(t)\}_{i=1}^n$ and $\{y_i(t)\}_{i=1}^n$, results from evaluating the functions x_i and y_i at the same fixed point t, for $i = 1, 2, \ldots, n$. In this case, the contrast is defined as in Equation (13):

$$\begin{aligned} H_0: &\quad \mu_{X(t)} - \mu_{Y(t)} = 0 \\ H_1: &\quad \mu_{X(t)} - \mu_{Y(t)} \neq 0, \end{aligned} \tag{13}$$

where $\mu_{X(t)}$ and $\mu_{Y(t)}$ are scalar parameters, since t is a fixed value. This test is performed using the test statistic of Equation (14):

$$\frac{\bar{X}(t) - \bar{Y}(t)}{sd/\sqrt{n}}, \tag{14}$$

where sd is the standard deviation of the scalar values $\{x_i(t) - y_i(t)\}_{i=1}^{n}$. In this way, we take 1000 values of t within the interval $[0, 1]$ and perform the test for each of these. We then obtain 1000 p-values, which are shown in Figure 6: in the left panel for case 1 and in the right panel for case 2.

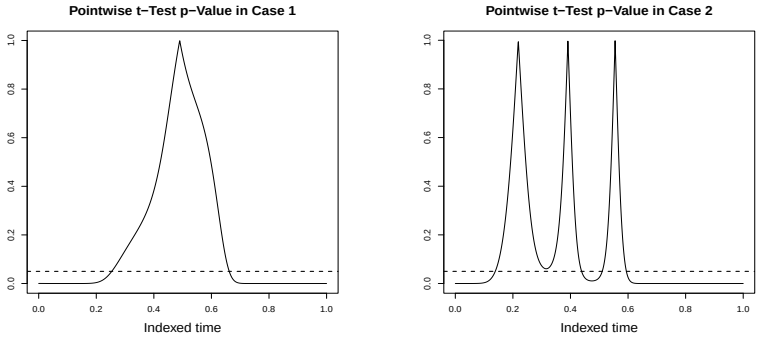


Figure 6. *P*-values obtained for the values of t in interval [0, 1] in the solid line and the reference line of 0.05 of significance in the dashed line for case 1 (**left**) and case 2 (**right**).

Note that, so far, it is not possible to determine whether there are significant differences between the two waves of COVID-19 contagion through the positivity curves, in a global way.

3.3. Another Hypothesis Test Approach for Functional Data

Under the previously exposed methodology, two groups of curves of size 24 were simulated in pairs, under the null hypothesis that the functional means are equal, and the MID test statistic was calculated in the sample. This process was repeated 4000 times separately for each case of study, with which 4000 values of the MID statistic were obtained. After performing the simulation process to obtain the null distribution in the form of a histogram, the value of the test statistic was calculated in the real functional data of positivity rate for COVID-19 in Colombia in both cases of study. Using the histogram found under simulation, the critical values corresponding to a significance of 0.05 were found by frequency. The histograms of the values found, together with the value of the statistic and the respective critical values, are shown in Figure 7: in the left panel for case 1 and in the right panel for case 2.

In addition to the above, in Table 1, we show the p-values obtained in the Shapiro–Wilk test performed on the 24 pieces of data from the integrals of the difference of the paired functional data in the two study cases. In addition, we show the p-value of the test statistic using the theoretical distribution of the test statistic, and we also show the values of the test statistic in each case and their respective p-values found under simulation and the critical values of the distribution null found under simulation.

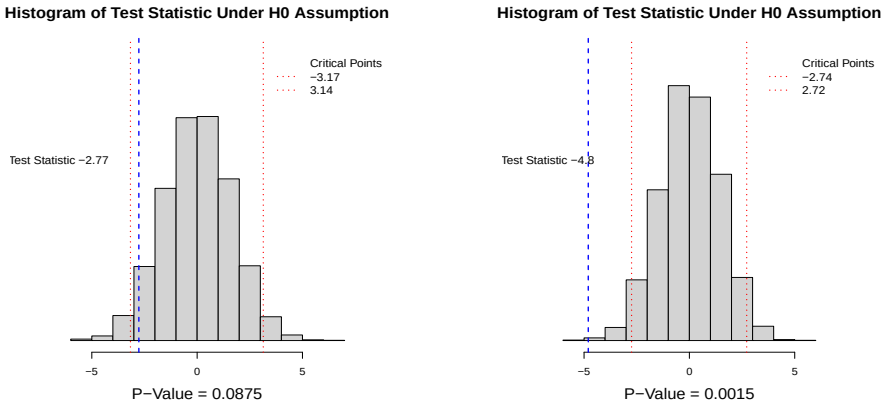


Figure 7. Histogram of the 4000 MID values obtained under simulation, and the MID contrast statistic calculated on the real positivity functional data in dashed line and the critical values in dotted lines for case 1 (**left panel**) and case 2 (**right panel**).

Table 1. Summary of our test results in both cases.

Case	Test Statistic	*p*-Value under Simulation	*p*-Value under Theoretical Distribution	*p*-Value of Shapiro–Wilk Test
Case 1	−2.77	0.0875	0.08906	0.6029165
Case 2	−4.8	0.0015	0.00001	0.7006806

Note that now, with the use of the scalar *p*-values found with our proposal, it is possible to decide on the existence of significant differences between both waves of COVID-19 contagion in a global way. Thus, for case 2, we can say that there are significant differences between the two waves of contagion, since the *p*-value in this case is 0.00001, under the theoretical null distribution, and 0.0015 under the simulated null distribution. Therefore, the hypothesis that the functional means of the positivity rate are the same in both waves of contagion by COVID-19 is rejected.

In turn, for the first case, since the *p*-values are 0.08906 under the theoretical null distribution, and 0.0875 under the simulated null distribution, the hypothesis that the functional means of the positivity rate are the same in both waves of contagion by COVID-19 is not rejected. It is not rejected by a very small margin with respect to the reference value of 0.05 significance.

4. Discussion

It is important to point out that, as Figure 6 shows, the point *t*-test for functional data allows us to evaluate the sections of the domain of the functions where there are significant differences. In terms of the cases of study, the dates between which there is a greater difference between the two contagion waves could be identified. Nevertheless, the pointwise *t*-test is insufficient to determine whether or not the two contagion waves are significantly different in each case study.

Our proposed test for hypotheses testing for paired data allow a global decision to be made based on the scalar *p*-value, however. The application of our contrast statistic allows us to visualize—as can be seen in Figure 7 (left panel)—that, for Case 1, when a significance of 0.05 is taken, the contrast statistic does not reject the hypothesis of equal means; i.e., at a significance of 0.05, there is no evidence that there are significant differences between the two contagion waves, even though the *p*-value of 0.082 is relatively close. Thus, if the significance is taken at 0.1, the decision would be to reject the null hypothesis, although again, with a very close margin. With regard to Case 2, as shown in Figure 7(rigth panel), the *p*-value found with our statistic is 0.0015, so we can say that there is sufficient evidence that the two contagion waves are significantly different.

Because of the above, although the p-value in case 1 leaves some doubt, it is important to highlight the difference between the p-values in both cases from a broader point of view, which seems to support the idea that the two case studies are remarkably different, and that the national strike in Colombia should not be ignored when analyzing epidemiological behavior, since the case studies suggest a possible change in the inclusion of positive data due to noncompliance with care measures during the national strike.

Author Contributions: Both authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study correspond to official data of the confirmed positive cases of COVID-19 in Colombia, data PCR tests for COVID-19 in Colombia, and data on antigen tests for COVID-19 in Colombia, respectively available at: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data (accessed on 14 September 2022), https://www.datos.gov.co/Salud-y-Protecci-n-Social/Pruebas-PCR-procesadas-de-COVID-19-en-Colombia-Dep/8835-5baf (accessed on 14 September 2022), and https://www.datos.gov.co/Salud-y-Protecci-n-Social/Ant-geno-procesadas-de-COVID-19-en-Colombia-Depart/ci85-cyhe/data (accessed on 14 September 2022).

Acknowledgments: This paper is partially supported by the research group FQM-307 of the Government of Andalusia (Spain) and by the project PID2020-113961GB-I00 of the Spanish Ministry of Science and Innovation (also supported by the FEDER programme). The authors also acknowledge the financial support of the Consejería de Conocimiento, Investigación y Universidad, Junta de Andalucía (Spain), and the FEDER programme for project A-FQM-66-UGR20. Additionally, the authors acknowledge financial support by the IMAG–María de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ramsay, J.; Silverman, B. *Functional Data Analysis*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2005.
2. Stewart, K.J.; Darcy, D.P.; Daniel, S.L. Opportunities and Challenges Applying Functional Data Analysis to the Study of Open Source Software Evolution. *Stat. Sci.* **2006**, *21*, 167–178. [CrossRef]
3. Jank, W.; Shmueli, G. Functional Data Analysis in Electronic Commerce Research. *Stat. Sci.* **2006**, *21*, 155–166. [CrossRef]
4. Ferraty, F. *Recent Advances in Functional Data Analysis and Related Topics*; Springer: Berlin/Heidelberg, Germany, 2011.
5. Horváth, L.; Kokoszka, P. *Inference for Functional Data with Applications*; Springer Series in Statistics; Springer: New York, NY, USA, 2012.
6. Sørensen, H.; Goldsmith, J.; Sangalli, L.M. An introduction with medical applications to functional data analysis. *Stat. Med.* **2013**, *32*, 5222–5240. [CrossRef] [PubMed]
7. Srivastava, A.; Klassen, E.P. *Functional and Shape Data Analysis*; Springer Series in Statistics; Springer: New York, NY, USA, 2016.
8. Ramsay, J.O.; Silverman, B.W. *Applied Functional Data Analysis: Methods and Case Studies*; Springer Series in Statistics; Springer: New York, NY, USA, 2002.
9. Hsing, T.; Eubank, R. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, 1st ed.; John Wiley & Sons Ltd.: West Sussex, UK, 2015.
10. Dallal, A.A.; AlDallal, U.; Dallal, J.A. Positivity rate: An indicator for the spread of COVID-19. *Curr. Med. Res. Opin.* **2021**, *37*, 2067–2076. [CrossRef] [PubMed]
11. Fu, Y.; Li, Y.; Guo, E.; He, L.; Liu, J.; Yang, B.; Li, F.; Wang, Z.; Li, Y.; Xiao, R.; et al. Dynamics and Correlation Among Viral Positivity, Seroconversion, and Disease Severity in COVID-19. *Ann. Intern. Med.* **2021**, *174*, 453–461. [CrossRef]
12. Furuse, Y.; Ko, Y.K.; Ninomiya, K.; Suzuki, M.; Oshitani, H. Relationship of Test Positivity Rates with COVID-19 Epidemic Dynamics. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4655. [CrossRef] [PubMed]
13. Juliana, R. Colombia 2021: Between Crises and Hope. *Rev. Cienc. PolíTica* **2022**, *42*, 255–280. [CrossRef]
14. Clarkson, D.B.; Fraley, C.; Gu, C.C.; Ramsay, J.O. *S+ Functional Data Analysis*; Springer: New York, NY, USA, 2005.
15. Rudin, W. *Functional Analysis*, 2nd ed.; McGraw Hill: Singapore, 1991.
16. Royden, H.L.; Fitzpatrick, P.M. *Real Analysis*, 4th ed.; Prentice Hall: Hoboken, NJ, USA, 2010.
17. Escabias, M. Reducción de Dimensión en Regresión Logistica Funcional. Ph.D. Thesis, Universidad de Granada, Granada, Spain, 2002.
18. Eubank, R.L. *Nonparametric Regression and Spline Smoothing*, 2nd ed.; Marcel Dekker Inc.: New York, NY, USA, 1998.

19. Olaya, J. *Metodos de Regresión no Paramétrica*; Universidad del Valle: Cali, Colombia, 2012.
20. Kenny, J.F.; Keeping, E.S. *Mathematics of Statistics. Part Two*, 2nd ed.; D. Van Nostrand Company, Inc.: New York, NY, USA, 1951.
21. Cox, D.D.; Lee, J.S.; Follen, M. A two sample test for functional data. *Commun. Stat. Appl. Methods* **2015**, *22*, 121–135.
22. Fan, J.; Lin, S. Test of significance when data are curves. *Am. Stat. Assoc.* **1998**, *93*, 111–122. [CrossRef]
23. Zhang, J.; Chen, J. Statical Inferences For Functional Data. *Ann. Stat.* **2007**, *35*. [CrossRef]
24. Cuevas, A.; Febrero, M.; Fraiman, R. A anova test for functional data. *Comput. Stat. Data Anal.* **2004**, *47*, 111–122. [CrossRef]
25. Degras, D. Simultaneous Confidence Bands for the Mean of Functional Data. *Wiley Interdiscip. Rev. Comput. Stat.* **2017**, *9*, e1397. [CrossRef]
26. Cuesta-Albertos, J.; Febrero-Bande, M. A simple multiway anova for funcional data. *Bus. Econ.* **2007**, *19*, 537–557.
27. Qiu, Z.; Chen, J.; Zhang, J.T. Two-sample tests for multivariate functional data with applications. *Comput. Stat. Data Anal.* **2021**, *157*, 107–160. [CrossRef]
28. Melendez, R.; Giraldo, R.; Leiva, V. Sign, Wilcoxon and Mann-Whitney Tests for Functional Data: An Approach Based on Random Projections. *Matematics* **2021**, *9*, 44. [CrossRef]

Apéndice B

Scalar Variance and Scalar Correlation for Functional Data

Titulo: *"Scalar Variance and Scalar Correlation for Functional Data".*

Autores: *Cristhian Leonardo Urbano-Leon, Manuel Escabias, Diana Paola Ovalle, Javier Olaya-Ochoa.*

Revista: *Mathematics.*

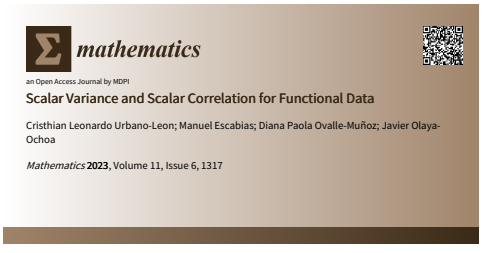
Factores de impacto de la revista:

Estado: Publicado, 9 de marzo de 2023.

doi: 10.3390/math11061317

Posición de autor: Autor principal.

Año	Factor de Impacto	Rango	Cuartil	Área
2022	2.4	23/330	$Q1$	Matemáticas



mathematics

MDPI

Article

Scalar Variance and Scalar Correlation for Functional Data

Cristhian Leonardo Urbano-Leon [1,*,†], Manuel Escabias [1,†], Diana Paola Ovalle-Muñoz [1,†] and Javier Olaya-Ochoa [2,†]

[1] Department of Statistics and Operations Research, University of Granada, 18071 Granada, Spain
[2] School of Statistics, University of Valle, Cali 760042, Colombia
* Correspondence: e.leonardourbano@go.ugr.es
† These authors contributed equally to this work.

Abstract: In Functional Data Analysis (FDA), the existing summary statistics so far are elements in the Hilbert space L_2 of square-integrable functions. These elements do not constitute an ordered set; therefore, they are not sufficient to solve problems related to comparability such as obtaining a correlation measurement or comparing the variability between two sets of curves, determining the efficiency and consistency of a functional estimator, among other things. Consequently, we present an approach of coherent redefinition of some common summary statistics such as sample variance, sample covariance and correlation in Functional Data Analysis (FDA). Regarding variance, covariance and correlation between functional data, our summary statistics lead to numbers instead of functions which is helpful for solving the aforementioned problems. Furthermore, we briefly discuss the functional forms coherence of some statistics already present in the FDA. We formally enumerate and demonstrate some properties of our functional summary statistics. Then, a simulation study is presented briefly, with evidence of the consistency of the proposed variance. Finally, we present the implementation of our statistics through two application examples.

Keywords: correlation for functional data; covariance for functional data; FDA; summary statistics in functional data; variance for functional data

MSC: 62R10

Citation: Urbano-Leon, C.L.; Escabias, M.; Ovalle-Muñoz, D.P.; Olaya-Ochoa J. Scalar Variance and Scalar Correlation for Functional Data. *Mathematics* **2023**, *11*, 1317. https://doi.org/10.3390/math11061317

Academic Editor: Alicia Nieto-Reyes

Received: 3 February 2023
Revised: 2 March 2023
Accepted: 4 March 2023
Published: 9 March 2023

1. Introduction

Functional Data Analysis (FDA) is a branch of statistics that has played a growing role since the book by [1] due to its multiple applications [2–8]. In fact, according to [9], the term Functional Data Analysis is due to [10,11], although some previous work on the subject is credited to [12,13]. FDA takes, as a starting point, discrete measurements of a continuous phenomenon to construct smooth curves using modified numerical analysis techniques. With these, the set of scalar data is converted into a new object called a functional datum, which is a continuous function [8,14]. This allows us to bring into statistical analysis some theoretical aspects from functional analysis, where some sets of functions with certain characteristics can form algebraic structures [15,16]. These structures can provide optimal properties for the analysis and measurement of continuous function curves. The Hilbert space, which is formed by square-integrable functions in a closed interval $[a, b]$; $a, b \in \mathbb{R}$, is a structure that plays a key role in this context. It is usually denoted as $L_2[a, b]$ and the main reason for playing such an important role is because Hilbert spaces are usually seen as extensions of the Euclidean space [15,17] (pages 249 and 19, respectively) because they have distance and size measures [18], which are desirable properties.

Until now, most existing FDA theory has been constructed based on the extension of scalar statistics concepts to functions, giving functional objects as a result. For instance, a widely accepted definition of summary statistics for functional data is given in [1], who define the sample functional mean, variance, covariance and correlation as continuous functions in $L_2[a, b]$, as shown in Definition 1.

Definition 1. ***Measures in Functional Data:*** *Given the sets of functional data* $\{\mathcal{X}_i\}_{i=1}^n$ *and* $\{\mathcal{Y}_i\}_{i=1}^n$ *defined in* $t \in [a, b]$*, the summary functions are defined as:*

- ***Mean:***

$$\overline{\mathcal{X}}(t) = \frac{\sum_{i=1}^n \mathcal{X}_i(t)}{n}$$

- ***Variance:***

$$Var(\mathcal{X})(t) = \frac{\sum_{i=1}^n (\mathcal{X}_i(t) - \overline{\mathcal{X}}(t))^2}{(n-1)}$$

- ***Standard Deviation:***

$$Sd(\mathcal{X})(t) = \sqrt{Var(\mathcal{X}(t))}$$

- ***Covariance:***

$$Cov(\mathcal{X}, \mathcal{Y})(t) = \frac{\sum_{i=1}^n (\mathcal{X}_i(t) - \overline{\mathcal{X}}(t))(\mathcal{Y}_i(t) - \overline{\mathcal{Y}}(t))}{(n-1)}$$

- ***Correlation:***

$$Corr(\mathcal{X}_i, \mathcal{Y}_i)(t) = \frac{Cov(\mathcal{X}, \mathcal{Y})(t)}{Sd(\mathcal{X})(t) Sd(\mathcal{Y})(t)}$$

In Definition 1, we show an expression for the covariance between $\mathcal{X}$ and $\mathcal{Y}$, which is usually defined in the literature as "cross-covariance". This name is suggested by Ramsay and Silverman in their 2005 book because they define the covariance as a summary of the dependence of records across different argument values.

It should be noted, however, that functions and scalars are different mathematical objects with different properties, which generates some conceptual discussions. Let us consider in the first place the functional mean. As is very well known for scalar data [19] (pp. 15–18), the Arithmetic Mean is a central tendency indicator that must be interpretable in the same context as the data. In this sense, the fact that the functional mean is a function results in a conceptually coherent concept of "tendency" because the functional mean describes the expected behavior of a set of functions related to a functional random variable. Moreover, it also has the property of "centrality" [20] (p. 76), which can be described in its functional form with the fulfilment of Equation (1), where $\mathcal{X}_0$ is the null function in the definition interval of the functional dataset $\{\mathcal{X}_i\}_{(i=1)}^n$.

$$\sum_{i=1}^n (\mathcal{X}_i - \overline{\mathcal{X}}) = \mathcal{X}_0 \tag{1}$$

However, if we follow the approach of [1], proving this fact requires an exhaustive walkthrough of the infinite points of the function's definition interval, which adds complexity to a proof that would become simpler upon performing the new approach we propose in this paper.

Let us consider now the functional variance. It should be considered that the usual motivation for the concept of variance is as a measure of dispersion [21] (p. 309), whose initial intention is to bring a comparative way for the use of consistency and efficiency concepts [22,23], whose definition is based on the basic premise that the sum of the squares of the deviations from the mean is a minimum [19] (pp. 555–557). Namely, given a set of scalar data $\{x_i\}_{i=1}^n$, $n \in \mathbb{N}$, associated with the random variable X and whose arithmetic mean is $\overline{X}$, the sum of Expression (2) is minimum value when $a = \overline{X}$ [20] (p. 84), so that the variance in Expression (3) is also minimum value because it reflects the expected behavior of the deviations from the mean.

$$\sum_{i=1}^n (x_i - a)^2 \tag{2}$$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 \tag{3}$$

Similarly, given a functional dataset $\{\mathcal{X}_i\}_{i=1}^n$ and a functional datum $\mathcal{Y}$, the sum in Expression (4) must be minimum value when $\mathcal{Y} = \overline{\mathcal{X}}$, where $\overline{\mathcal{X}}$ is the functional mean of $\{\mathcal{X}_i\}_{i=1}^n$.

$$\sum_{i=1}^{n}(\mathcal{X}_i - \mathcal{Y})^2 \tag{4}$$

However, Ramsay and Silverman's functional variance given in Definition 1 does not comply with this functional version of such property. This is true because the concept of a "minimum" function lacks validity due to the functional character of the objects and that the functions do not form a well ordered set [24–26].

Furthermore, a conceptual problem seems to be associated with the fact that the variance itself gives a measure of the distance from the values to the mean [21]. According to [17], a measure always has to be a nonnegative real number; therefore, the functional variance given in Definition 1 is not truly a measure of dispersion of the functional data but rather a curve that offers point-to-point variances within the functional data definition interval.

Another important flaw of the variance curve is that, due to the lack of order inside a functional space, if there are two sets of curves, it is hard to decide which of the two sets has a larger dispersion. This is possible, however, with a point-to-point comparison, as is done if we look at the functional variance as a curve of point-to-point variances.

On the other hand, on the functional covariance and correlation, while the concept of covariance is not thought of as a measure, it is expected to be an indicator of the joint variation of two random variables [27,28]. It maintains a close relation to the concept of variance because the variance is a very particular case of the covariance [19] (pp. 126–128). Therefore, if the functional variance is a real number, the covariance must also be because there should be coherence between the concepts. However, covariance in Definition 1 is just a point-to-point covariance function whose interpretation gives an idea of the regions where there is a larger or a smaller joint variation of the curves within the definition interval of the functional data, but it is not an indicator of the joint variation of two functional random variables.

In turn, correlation is defined as a coefficient that is linked to the concept of covariance [19] (pp. 115–119) and for that reason, it must be defined as a real number, even in functional data. However, it is important to point out that the concept of "linear" correlation within functional data does not become clearly visible and its "linear" interpretation needs further study.

In this sense, we present a new approach for the treatment of functional data that attempts to provide summary statistics for functional data with conceptual coherence and operational advantages, defining the variance, covariance and correlation for functional data as real numbers. For this, we use one of the most notable particularities of the Hilbert spaces like vector space because, as claimed by [29–32], the elements of a vector space can be uniquely completely described by a linear combination of a set of orthogonal elements called a basis. However, for $L_2[a, b]$, this set is infinite. For that reason, all FDA theory is not performed over the entire $L_2[a, b]$ but over a subspace of finite dimension [33] because the number of functions used as a basis for the construction of the functional data is finite [1,34].

Thus far, this fact has been little explored in FDA theoretical development, although there are recent works, such as [35], that estimate a test statistic using the basis coefficients, the use of coefficients in the homogeneity problem by [36], and also [37] who use the basis coefficients for estimating functional PLS regression, or even previously, with the use of principal components of functional logistic regression the authors of [38] deal with some aspects of the subject. However, our approach addresses the functional data from a vector

perspective because each functional datum corresponds to a single coefficient vector that characterizes it, which allows transferring some operations between functions to operations between coefficient vectors, component-by-component, providing operational advantages in formal proofs.

It is important to point out that our approach is based on the representation of a continuous function within an arbitrary subspace of finite dimension of $L_2[a,b]$ and therefore the results obtained can be applied without loss of generality to any type of orthonormal basis and of any finite dimension. Thus, in addition to the theory proposed, a simulation study with curves represented in a subspace spanned by an orthonormal basis of cosines in the [0, 1] interval is presented briefly and two implementation examples are described in the final section. In the first example, we implement our correlation proposition to determine whether there is independence between the functional random variables, used by [1] when constructing a functional analysis of variance (FANOVA) of the Canadian average annual temperature in four of its regions. This is done because the analysis does not specify the existence of independence between the functional random variables. The second example describes the use of our variance and functional correlation proposition as part of a functional and descriptive data analysis on particulate matter from two air quality monitoring stations in Cali, Colombia.

At this point, we recall some very well known background information, which we will need to present the new summary statistics definitions.

Definition 2. ***System of Generators, Basis and Dimension:***

- *Given a vector space $\mathbb{V}$ over a field $\mathbb{K}$ and given $S \subset \mathbb{V}$, it is decided that $\mathbb{V}$ is spanned by S if every element of $\mathbb{V}$ can be written as a linear combination of the elements of S [32].*
- *A vector space is said to be of finite dimension if it can be generated by a finite set of elements [39].*
- *A set $B \subset \mathbb{V}$ is a basis for $\mathbb{V}$ if B spans $\mathbb{V}$ and is also linearly independent. [32]*

Definition 3. ***Hilbert Space:*** *In a very general manner, a Hilbert space is a vector space over the field of real numbers in which a norm and an inner product have been defined [34,40].*

Definition 4. ***Functional Space:*** *A functional space is a vector space whose elements are functions.*

Definition 5. ***The Functional Hilbert Space*** $L_2[a,b]$***:*** *The $L_2[a,b]$ space is a vector space over the field of real numbers whose elements are square-integrable functions in the closed interval $[a,b]$; $a,b \in \mathbb{R}$ and where given $f,g \in L_2[a,b]$, an inner product, a norm and a distance are defined as:*

- *Inner product: $\langle f,g\rangle = \int_a^b f(x)g(x)dx$.*
- *Norm: $\|f\| = \langle f,f\rangle^{1/2} = (\int_a^b f^2(x)dx)^{1/2}$*
- *Distance: $d(f,g) = \sqrt{\int_a^b (f(x)-g(x))^2 dx}$*

With this, the $L_2[a,b]$ space is a functional Hilbert space [29,41].

Definition 6. ***Orthonormal and Orthogonal Bases:*** *Two elements of a vector space are orthogonal if and only if the inner product between them is zero. In the same way, an element of a vector space is said to be normal if and only if its norm value is equal to 1. Thus, a basis is said to be orthogonal if and only if it is composed of elements that are orthogonal two-by-two. If, in addition, such elements satisfy the condition of normality, the basis is said to be orthonormal [30,42].*

Definition 7. ***Functional Random Variable:*** *A functional random variable $\mathcal{X}$ is a random variable that takes values in a functional space, where an observation of $\mathcal{X}$ is called a functional datum [43].*

2. Summary Statistics for Functional Data

With the motivation of providing FDA theory with summary statistics that have interpretive and operational advantages, in this section, we propose a new approach for the treatment of functional data that considers them as elements of the same functional subspace $\mathcal{H}$ of finite dimension of the Hilbert space $L_2[a,b]$, which allows transforming operations between functions to operations between elements of vectors of coefficients component-by-component because the $L_2[a,b]$ space is in particular a vector space and therefore if $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^{p}$ is a finite basis of $\mathcal{H}$ and $\mathcal{X} \in \mathcal{H}$, defined on Equation (5), is a functional datum,

$$\mathcal{X} = \sum_{j=1}^{p} a_j \mathcal{F}_j \text{ For } a_j \in \mathbb{R} \text{ with } j = 1, 2, 3, \ldots, p. \tag{5}$$

where $\mathcal{X}$ is a linear combination of the elements of $\mathcal{B}$ and therefore $\mathcal{X}$ is uniquely completely determined by the vector $(a_1, a_2, a_3, \ldots, a_p)_{\mathcal{B}}$, which is called the $\mathcal{B}-$basis representation vector.

As mentioned, Ref. [35] estimate a test statistic using the basis coefficients. Previously, on a density estimation problem, Ref. [33] used a finite-dimensional approximation of the functional data, just like the one we propose here. However, none of them moved toward the representation of summary statistics for functional data using the vectors of coefficients used for the functional data representation shown in Equation (5). Next, the new definitions of summary statistics are proposed, as well as their properties.

2.1. Sum of Functional Data

As observed in works by [44,45] and in most of the textbooks on linear algebra and functional analysis, the sum of the elements of a space of finite dimension can be defined by the sum of their representation coefficients in the same basis, as is presented in Proposition 1 and whose proof is immediate, using the representation of each functional datum and the association and commutation properties of $L_2[a,b]$ under the sum.

Proposition 1. *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a,b]$ with basis $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^{p}$ and $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ a set of n functional data with representation vectors $(a_{i,1}, a_{i,2}, \ldots, a_{i,p})_{\mathcal{B}}$, for some $a_{i,j} \in \mathbb{R}$ with $i = 1, 2, \ldots n$ and $j = 1, 2, 3, \ldots, p$. Then, the sum $\sum_{i=1}^{n} \mathcal{X}_i$ is an element of $\mathcal{H}$ with representation vector:*

$$\left(\sum_{i=1}^{n} (a_{i,1}), \sum_{i=1}^{n} (a_{i,2}), \ldots, \sum_{i=1}^{n} (a_{i,p}) \right)_{\mathcal{B}}$$

Proof. Given that $(a_{i,1}, a_{i,2}, \ldots, a_{i,p})_{\mathcal{B}}$ with $a_{i,j} \in \mathbb{R}$, $i = 1, 2, \ldots n$ and $j = 1, 2, 3, \ldots, p$ are representation vectors,

$$\begin{aligned} \sum_{i=1}^{n} \mathcal{X}_i &= \sum_{i=1}^{n} \left(\sum_{j=1}^{p} a_{i,j} \mathcal{F}_j \right) \\ &= \sum_{j=1}^{p} \left(\sum_{i=1}^{n} a_{i,j} \right) \mathcal{F}_j \end{aligned}$$

Therefore, given $\mathcal{F}_j \in \mathcal{H}$ for every $1 \leq j \leq p$, then $\sum_{i=1}^{n} \mathcal{X}_i \in \mathcal{H}$ and its representation vector is

$$\left(\sum_{i=1}^{n} (a_{i,1}), \sum_{i=1}^{n} (a_{i,2}), \ldots, \sum_{i=1}^{n} (a_{i,p}) \right)_{\mathcal{B}}$$

□

Proposition 1 indicates that the sum of the n functional data $\{\mathcal{X}_i\}_{i=1}^n$ is completely determined by the sum of the representation vectors component-by-component.

2.2. Mean of Functional Data

It is possible to obtain a definition of the mean for functional data from the representation coefficients, as we show in Proposition 2 and whose proof applies Proposition 1 as well as the representation coefficients.

Proposition 2. *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a,b]$ with basis $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^p$ and $\{\mathcal{X}_i\}_{i=1}^n \subseteq \mathcal{H}$ a set of n functional data with representation vectors $(a_{i,1}, a_{i,2}, \ldots, a_{i,p})_{\mathcal{B}}$ for $a_{i,j} \in \mathbb{R}$ with $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,p$. Then, the functional mean for the functional dataset $\{\mathcal{X}_i\}_{i=1}^n$ is given by:*

$$\overline{\mathcal{X}} = \sum_{j=1}^{p} (\overline{A}_j)\mathcal{F}_j \tag{6}$$

where $\overline{A}_j = \frac{1}{n}\left(\sum_{i=1}^n a_{i,j}\right)$ for $j = 1,2,\ldots,p$

Proof. Given $\overline{A}_j = n^{-1}\left(\sum_{i=1}^n a_{i,j}\right)$ is the mean of the j^{th} coefficients with $j = 1,2,\ldots,p$, then:

$$\begin{aligned}
\overline{\mathcal{X}} &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{X}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p} a_{i,j}\mathcal{F}_j\right) \\
&= \sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n} a_{i,j}\right)\mathcal{F}_j = \sum_{j=1}^{p}(\overline{A}_j)\mathcal{F}_j
\end{aligned}$$

□

Proposition 2 indicates that the representation vector of the functional mean is completely determined by the representation vector of component-by-component mean coefficients. Namely, for a set of n functional data whose representation vectors are $(a_{i,1}, a_{i,2}, \ldots, a_{i,p})_{\mathcal{B}}$ with $1 \leq i \leq n$, the functional mean is another functional datum, whose representation vector is $(\overline{A}_1, \overline{A}_2, \ldots, \overline{A}_p)_{\mathcal{B}}$.

In addition, it should be noted that the functional mean of Proposition 2 and the one in Definition 1 are the same functions because, as mentioned above, it is coherent with the concept of tendency. However, under this new approach, the functional mean has operational advantages, some of which are immediately observed in the proof of the functional version of the property of centrality (p. 76, [20]), illustrated by Proposition 3.

Proposition 3. *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a,b]$, $\{\mathcal{X}_i\}_{i=1}^n \subseteq \mathcal{H}$ a set of n functional data and $\overline{\mathcal{X}}$ its functional mean, then:*

$$\sum_{i=1}^{n}(\mathcal{X}_i - \overline{\mathcal{X}}) = \mathcal{X}_0$$

where $\mathcal{X}_0$ is the null function in $[a,b]$.

Proof. Let $\mathcal{B} = \{\mathcal{F}_i\}_{i=1}^p$ be a basis of $\mathcal{H}$. Therefore, there are $a_{i,j} \in \mathbb{R}$ such that $\mathcal{X}_i = \sum_{j=1}^p a_{i,j}\mathcal{F}_j$ for every $i = 1,2,\ldots n$. In addition, by Proposition 2, $\overline{\mathcal{X}} = \sum_{j=1}^p \overline{\mathcal{A}_j}\mathcal{F}_j$, where $\overline{\mathcal{A}_j} = \frac{1}{n}\sum_{i=1}^n a_{i,j}$; then, by Proposition 1:

$$\begin{aligned}\sum_{i=1}^{n}(\mathcal{X}_i - \overline{\mathcal{X}}) &= \sum_{i=1}^{n}\left(\sum_{j=1}^{p}(a_{i,j} - \overline{A_j})\mathcal{F}_j\right)\\ &= \sum_{j=1}^{p}\left(\sum_{i=1}^{n}(a_{i,j} - \overline{A_j})\right)\mathcal{F}_j\\ &= \sum_{j=1}^{p}(0)\mathcal{F}_j = \mathcal{X}_0\end{aligned}$$

because $\sum_{i=1}^{n}(a_{i,j} - \overline{A_j}) = 0$ for $j = 1, 2, \ldots, p$, "$a_{i,j}$" and "$\overline{A_j}$" are scalars and therefore satisfy the property of centrality. □

2.3. Variance for Functional Data

We define the variance for functional data as the average of the squared distances of each function to the functional mean. The distance used is the distance between the functions of $L_2[a, b]$, shown above in Definition 5, which gives a scalar number as a result and therefore the variance of Definition 8 is a scalar and maintains the concept of the variance as a measure of dispersion because it is the expected behavior of the distances of the functions to the functional mean, whose interpretation is performed in a general manner over the entire set of functions.

Definition 8. ***Variance for Functional Data:*** *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a, b]$ and let $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ be a set of n functional data associated with the functional random variable $\mathcal{X}$; then, the scalar variance for this functional data set is defined as:*

$$\mathbb{V}ar(\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_a^b (\mathcal{X}_i - \overline{\mathcal{X}})^2(t)dt\right) \tag{7}$$

This definition allows having a scalar measure of dispersion, around the functional mean, of a set of functions. The most notable operational advantage of Definition 8 is given by Theorem 1.

Theorem 1. *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a, b]$, $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^{p}$ an orthonormal basis of $\mathcal{H}$ and $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ a set of n functional data associated with the functional random variable $\mathcal{X}$, with representation vectors $(a_{i,1}, a_{i,2}, \ldots, a_{i,p})_{\mathcal{B}} : 1 \leq i \leq n$, then:*

$$\mathbb{V}ar(\mathcal{X}) = \sum_{j=1}^{p} V_j \tag{8}$$

where $V_j = \frac{1}{n}\sum_{i=1}^{n}(a_{i,j} - \overline{\mathcal{A}_j})^2$.

Theorem 1 indicates that if the representation basis is orthonormal, the variance for the functional data can be simply calculated as the sum of the variances of the coefficients, component-by-component.

To prove Theorem 1, we need first to prove a couple of very important results presented in Lemmas 1 and 2.

Lemma 1. *Let $\mathcal{H}$ be a subspace of finite dimension p of $L_2[a, b]$, $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^{p}$ a basis of $\mathcal{H}$, $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ a set of functional data with representation vectors $(a_{i,1}, a_{i,2}, \ldots, a_{i,p})_{\mathcal{B}} : 1 \leq i \leq n$ and $\mathcal{Y} \in \mathcal{H}$ a functional datum with representation vector $(b_1, b_2, \ldots, b_p)_{\mathcal{B}}$, then:*

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i-\mathcal{Y})^2=\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-b_j)^2\right)\mathcal{F}_j^2+2$$
$$\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}\left(\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-b_j)(a_{i,(j+1)}-b_{(j+1)})\right)\mathcal{F}_j\mathcal{F}_{(j+1)}$$

The proof of Lemma 1 follows from the representation of each of the elements of $\mathcal{H}$ on the selected basis and from the properties of summation.

Proof. Given that $\mathcal{Y}\wedge\mathcal{X}_i\in\mathcal{H}$ for every $i=1,2,\ldots,n$, then:

$$\frac{1}{n}\sum_{i=1}^{n}[\mathcal{X}_i-\mathcal{Y}]^2=$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p}a_{i,j}\mathcal{F}_j-\sum_{j=1}^{p}b_j\mathcal{F}_j\right]^2$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p}(a_{i,j}\mathcal{F}_j-b_j\mathcal{F}_j)\right]^2$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{p}(a_{i,j}-b_j)\mathcal{F}_j\right]^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}(a_{i,j}-b_j)^2\mathcal{F}_j^2\right)$$
$$+\frac{1}{n}\sum_{i=1}^{n}2\left(\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}\Big((a_{i,j}-b_j)(a_{i,(j+1)}-b_{(j+1)})\Big)\mathcal{F}_j\mathcal{F}_{(j+1)}\right)$$
$$=\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-b_j)^2\right)\mathcal{F}_j^2$$
$$+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}\left(\frac{1}{n}\sum_{i=1}^{n}\Big((a_{i,j}-b_j)(a_{i,(j+1)}-b_{(j+1)})\Big)\right)\mathcal{F}_j\mathcal{F}_{(j+1)}$$

□

If in Lemma 1 we replace $\mathcal{Y}$ by the functional mean of $\{\mathcal{X}_i\}_{i=1}^n$, the Lemma 2 is obtained.

Lemma 2. *Let $\mathcal{H}$ be a subspace of finite dimension p of $L_2[a,b]$, $\mathcal{B}=\{\mathcal{F}_j\}_{j=1}^p$ a basis of $\mathcal{H}$, $\{\mathcal{X}_i\}_{i=1}^n\subseteq\mathcal{H}$ a set of functional data with representation vectors $\left(a_{i,1},a_{i,2},\ldots,a_{i,p}\right)_{\mathcal{B}}:1\leq i\leq n$ and $\overline{\mathcal{X}}$ the functional mean of $\{\mathcal{X}_i\}_{i=1}^n$; then, if $\overline{\mathcal{A}_j}=\frac{1}{n}\sum_{i=1}^n a_{i.j}$, for $1\leq j\leq p$, the mean difference of squares can be decomposed as:*

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i-\overline{\mathcal{X}})^2=\sum_{j=1}^{p}V_j\mathcal{F}_j+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\mathcal{F}_j\mathcal{F}_{j+1}$$

where $V_j=\frac{1}{n}\sum_{i=1}^n(a_{i,j}-\overline{\mathcal{A}_j})^2$ and $S_{a_j,a_{j+1}}=\frac{1}{n}\sum_{i=1}^n(a_{i,j}-\overline{\mathcal{A}_j})(a_{i,j+1}-\overline{\mathcal{A}}_{j+1})$.

Proof. By Lemma 1:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i-\overline{\mathcal{X}})^2=\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{\mathcal{A}_j})\right)^2\mathcal{F}_j^2$$
$$+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}\left(\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{\mathcal{A}_j})(a_{i,(j+1)}-\overline{\mathcal{A}}_{(j+1)})\right)\mathcal{F}_j\mathcal{F}_{(j+1)}$$

but $V_j=\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{\mathcal{A}_j})^2$ and $S_{a_j,a_{j+1}}=\frac{1}{n}\sum_{i=1}^{n}(a_{i,j}-\overline{\mathcal{A}_j})(a_{i,j+1}-\overline{\mathcal{A}}_{j+1})$, with which:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i-\overline{\mathcal{X}})^2=\sum_{j=1}^{p}V_j\mathcal{F}_j^2+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\mathcal{F}_j\mathcal{F}_{j+1}$$

□

We thus provide a proof of Theorem 1.

Proof. By Definition 8:

$$\mathbb{V}ar(\mathcal{X})=\frac{1}{n}\sum_{i=1}^{n}\left(\int_a^b(\mathcal{X}_i-\overline{\mathcal{X}})^2(t)dt\right)$$

and by the integral's properties, we know that:

$$\frac{1}{n}\sum_{i=1}^{n}\int_a^b(\mathcal{X}_i-\overline{\mathcal{X}})^2(t)dt=\int_a^b\frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i-\overline{\mathcal{X}})^2(t)dt$$

but from Lemma 2, we have that:

$$\int_a^b\frac{1}{n}\sum_{i=1}^{n}(\mathcal{X}_i-\overline{\mathcal{X}})^2(t)dt=$$

$$=\int_a^b\left(\sum_{j=1}^{p}V_j\mathcal{F}_j^2(t)+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\mathcal{F}_j(t)\mathcal{F}_{j+1}(t)\right)dt$$
$$=\sum_{j=1}^{p}V_j\int_a^b(\mathcal{F}_j^2)(t)dt+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\int_a^b(\mathcal{F}_j(t)\mathcal{F}_{j+1}(t))dt$$
$$=\sum_{j=1}^{p}V_j\left\|\mathcal{F}_j\right\|+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\left\langle\mathcal{F}_j\mathcal{F}_{j+1}\right\rangle$$

By hypothesis, $\mathcal{B}$ is an orthonormal basis; that is, $\left\|\mathcal{F}_j\right\|=1$ and $\left\langle\mathcal{F}_j\mathcal{F}_m\right\rangle=0$; for each $j,m=1,2,\ldots,p\wedge j\neq m$, therefore, we have that:

$$\sum_{j=1}^{p}V_j\left\|\mathcal{F}_j\right\|^2+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\left\langle\mathcal{F}_j\mathcal{F}_{j+1}\right\rangle=$$

$$=\sum_{j=1}^{p}V_j+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}0$$
$$=\sum_{j=1}^{p}V_j$$

□

We highlight that as part of the proof of Theorem 1, the Equation (9) is obtained:

$$\mathbb{V}ar(\mathcal{X})=\sum_{j=1}^{p}V_j\left\|\mathcal{F}_j\right\|^2+2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}S_{a_j,a_{j+1}}\left\langle\mathcal{F}_j\mathcal{F}_{j+1}\right\rangle,\tag{9}$$

which means that we can apply the result to a nonnormal basis and even to generator sets that are not necessarily orthogonal.

Let us display now some properties of the functional variance under this new approach, which satisfies the same properties of a variance for scalar data because it inherits them from the variances of coefficients, as shown below.

Proposition 4. *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a,b]$, $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^{p}$ an orthonormal basis of $\mathcal{H}$ and $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ a set of n functional data associated with the functional random variable $\mathcal{X}$, with representation vectors $\left(a_{i,1}, a_{i,2}, \ldots, a_{i,p}\right)_{\mathcal{B}} : 1 \leq i \leq n$. Then, the following properties for the functional variance are followed:*

1. $\mathbb{V}ar(\mathcal{X}) \geq 0$.
2. $\mathbb{V}ar(\mathcal{X})$ *is of minimum value.*
3. *If $\{\mathcal{X}_i\}_{i=1}^{n}$ have the same representation vectors, then $\mathbb{V}ar(\mathcal{X}) = 0$.*

Proof. Properties

1. Given that $\mathbb{V}ar(\mathcal{X}) = \sum_{j=1}^{p} V_j$ and that $V_j \geq 0$ for every $j = 1, 2, \ldots, p$, $\mathbb{V}ar(\mathcal{X}) \geq 0$.
2. Given that $\mathbb{V}ar(\mathcal{X}) = \sum_{j=1}^{p} V_j$ and that each V_j is of minimum value for every $j = 1, 2, \ldots, p$, then $\mathbb{V}ar(\mathcal{X})$ is of minimum value.
3. Given $\{\mathcal{X}_i\}_{i=1}^{n}$ functional data, such that their representation vectors are equal, then:

$$\begin{aligned} a_{1,1} = a_{2,1} = a_{3,1}, \ldots, a_{n,1} &= w_1 \\ a_{1,2} = a_{2,2} = a_{3,2}, \ldots, a_{n,2} &= w_2 \\ \vdots \quad \vdots & \\ a_{1,p} = a_{2,p} = a_{3,p}, \ldots, a_{n,p} &= w_p \end{aligned}$$

Then, for every $1 \leq j \leq p$, $\overline{A}_j = \frac{1}{n}\sum_{i=1}^{n} a_{i,j} = \frac{1}{n}\sum_{i=1}^{n} w_j = w_j$ and

$$V_j = \frac{1}{n}\sum_{i=1}^{n}\left(a_{i,j} - \overline{A}_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(w_j - w_j\right)^2 = 0$$

Therefore, $V_j = 0$ for each $j = 1, 2, \ldots, p$ and consequently $\mathbb{V}ar(\mathcal{X}) = 0$.

□

The fulfilment of this last property shows that, in fact, $\mathbb{V}ar(\mathcal{X})$ measures the dispersion of the functional data.

2.4. Covariance and Correlation in Functional Data

Following the same line of reasoning as for the variance for functional data associated with Definition 8, the covariance for functional data is shown in Definition 9.

Definition 9. ***Covariance in Functional Data:*** *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[0,1]$ and $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ and $\{\mathcal{Y}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ two sets of n functional data associated with the functional random variables $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then the scalar covariance for these two sets of functional data is defined as:*

$$\mathbb{C}ov(\mathcal{X}, \mathcal{Y}) = \frac{1}{n}\sum_{i=1}^{n}\int_a^b \left(\mathcal{X}_i - \overline{\mathcal{X}}\right)\left(\mathcal{Y}_i - \overline{\mathcal{Y}}\right)(t)dt \tag{10}$$

Like the variance case, this definition allows having a scalar value of the joint variability of two functional random variables in relation to the functional mean in each case and presents the operational advantage given by Theorem 2.

Theorem 2. *Let $\mathcal{H}$ be a subspace of finite dimension p of the Hilbert space $L_2[a,b]$, $\mathcal{B} = \{\mathcal{F}_j\}_{j=1}^{p}$ an orthonormal basis of $\mathcal{H}$ and $\{\mathcal{X}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ and $\{\mathcal{Y}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ two sets of n functional data associated with the functional random variables $\mathcal{X}$ and $\mathcal{Y}$, with representation vectors $\left(a_{i,1}, a_{i,2}, \ldots, a_{i,p}\right)_{\mathcal{B}}$ and $\left(b_{i,1}, b_{i,2}, \ldots, b_{i,p}\right)_{\mathcal{B}} : 1 \leq i \leq n$, respectively. Then:*

$$\mathbb{C}ov(\mathcal{X}, \mathcal{Y}) = \sum_{j=1}^{p} C_j, \tag{11}$$

where $C_j = \frac{1}{n}\sum_{i=1}^{n}\left(a_{i,j} - \overline{\mathcal{A}_j}\right)\left(b_{i,j} - \overline{\mathcal{B}_j}\right)$, being $\overline{\mathcal{B}_j} = \frac{1}{n}\sum_{i=1}^{n} b_{i,j}$ and $\overline{\mathcal{A}_j} = \frac{1}{n}\sum_{i=1}^{n} a_{i,j}$ for each $j = 1, 2, \ldots, p$

Proof. By definition, we have that:

$$\mathbb{C}ov(\mathcal{X}, \mathcal{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_a^b \left(\mathcal{X}_i - \overline{\mathcal{X}}\right)\left(\mathcal{Y}_i - \overline{\mathcal{Y}}\right)(t)dt\right)$$

and by the integral's properties, we know that:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\int_a^b \left(\mathcal{X}_i - \overline{\mathcal{X}}\right)\left(\mathcal{Y}_i - \overline{\mathcal{Y}}\right)(t)dt\\ &= \int_a^b \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{X}_i - \overline{\mathcal{X}}\right)\left(\mathcal{Y}_i - \overline{\mathcal{Y}}\right)(t)dt.\end{aligned}$$

Because of Lemma 1, we have that:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{X}_i - \overline{\mathcal{X}}\right)\left(\mathcal{Y}_i - \overline{\mathcal{Y}}\right) =$$

$$\begin{aligned}&\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}(a_{i,j} - \overline{A}_j)(b_{i,j} - \overline{B}_j)\right)\mathcal{F}_j^2\\ &+ 2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left((a_{i,j} - \overline{A}_j)(b_{i,(j+1)} - \overline{B}_{(j+1)})\right)\right)\mathcal{F}_j\mathcal{F}_{(j+1)}\end{aligned}$$

By applying the integral, we obtain:

$$\sum_{j=1}^{p} C_j\|\mathcal{F}_j\| + 2\sum_{k=1}^{p-1}\sum_{j=k}^{p-1} S_{a_k,a_{k+1}}\langle\mathcal{F}_k\mathcal{F}_{k+1}\rangle$$

where $C_j = \frac{1}{n}\sum_{i=1}^{n}(a_{i,j} - \overline{\mathcal{A}_j})(b_{i,j} - \overline{\mathcal{B}_j})$ and $S_{a_p,b_{p+1}} = \frac{1}{n}\sum_{i=1}^{n}(a_{i,p} - \overline{\mathcal{A}_p})(b_{i,p+1} - \overline{\mathcal{B}}_{p+1})$. However, because the basis is orthonormal, we have that:

$$\mathbb{C}ov(\mathcal{X}, \mathcal{Y}) = \sum_{j=1}^{p} C_j \cdot 1 + 0 = \sum_{j=1}^{p} C_j \tag{12}$$

□

We notice that under this approach, it makes sense that $\mathbb{C}ov(\mathcal{X}, \mathcal{X}) = \mathbb{V}ar(\mathcal{X})$ and that $\mathbb{C}ov(\mathcal{X}, \mathcal{Y})$, by virtue of Theorem 2, as well as $\mathbb{V}ar(\mathcal{X})$, by virtue of Theorem 1, inherit the properties of the classical covariance in scalar data through their coefficients, but even more so, it is possible to define a correlation coefficient, as shown in Definition 10.

Definition 10. ***Correlation in Functional Data:*** *Let $\mathcal{H}$ be a subspace of dimension p of $L_2[a,b]$ and $\{\mathcal{X}_i\}_{i=1}^{n}, \{\mathcal{Y}_i\}_{i=1}^{n} \subseteq \mathcal{H}$ two sets of functional data associated with the functional random*

variables $\mathcal{X}$ and $\mathcal{Y}$, whose scalar variances are $\mathbb{V}ar(\mathcal{X})$ and $\mathbb{V}ar(\mathcal{Y})$, respectively, and with scalar covariance $\mathbb{C}ov(\mathcal{X},\mathcal{Y})$; then the scalar correlation coefficient is defined by the expression:

$$\mathbb{C}or(\mathcal{X},\mathcal{Y}) = \frac{\mathbb{C}ov(\mathcal{X},\mathcal{Y})}{\sqrt{\mathbb{V}ar(\mathcal{X})}\sqrt{\mathbb{V}ar(\mathcal{Y})}} \tag{13}$$

As a result of the proposed approach, it can be seen that $\mathbb{C}or(\mathcal{X},\mathcal{Y}) \in \mathbb{R}$ and that $-1 \leq \mathbb{C}or(\mathcal{X},\mathcal{Y}) \leq 1$. In addition, under the compliance of the hypothesis of Theorems 1 and 2, the calculation of the correlation can be performed with the expression:

$$\mathbb{C}or(\mathcal{X},\mathcal{Y}) = \frac{\sum_{j=1}^{p} C_j}{\sqrt{\left(\sum_{j=1}^{p} V_{a_j}\right)\left(\sum_{j=1}^{p} V_{b_j}\right)}}$$

3. Simulation

In order to show that our variance is actually a measure of the variability of the curves, without loss of generality, we conduct a brief simulation study, which builds a variety of functional data sets and we represent it in the ten-dimensional subspace spanned by the orthonormal basis $\left\{1, \sqrt{2}Cos((j-1)\pi x\right\}_{j=2}^{9}$. The Supplementary Materials contains the code for the simulation, which was carried out in the R language. In each case, we use our Theorem 1 to obtain the variance measure. Simulation cases come from two functions of different types. The first type, which we call Type A, is a constant function, whereas the second type, called Type B, is a non-constant function, as shown in Figure 1.

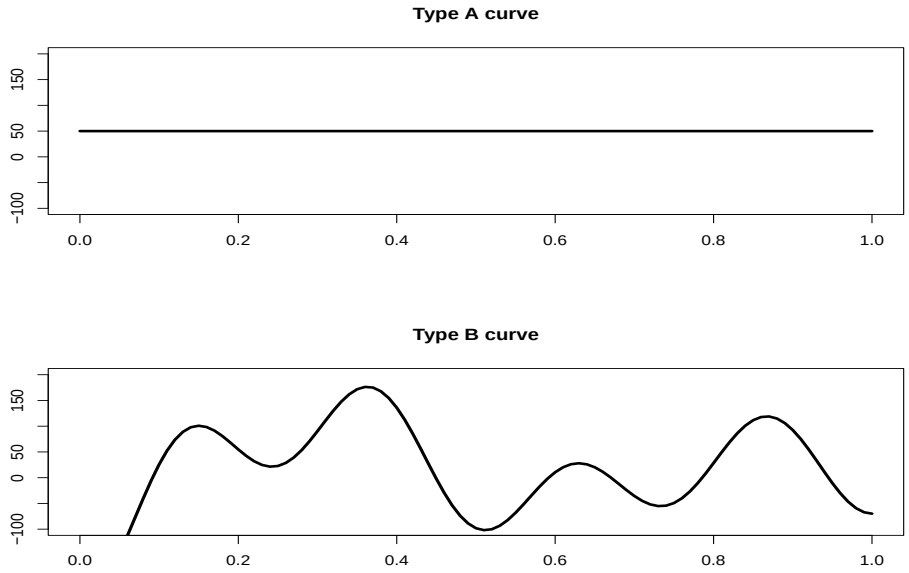


Figure 1. Type A (**top**) and Type B (**bottom**) curves for simulation.

In the first simulation case, we consider three different scenarios of constant dispersion across the domain of the curves: case 1.1 high dispersion, case 1.2 moderate dispersion and case 1.3 low dispersion. Figure 2 clearly shows the simulated dispersion in the curves and how our proposal captures the decreases in variability in the proposed scenarios.

We construct functions with non-uniform dispersion across the domain of functions in the second simulation case and as in the first case, we consider three different scenarios: case 2.1 high dispersion, case 2.2 moderate dispersion and case 2.3 low dispersion. The resulting curves are more erratic than those in the first case in each type of curve. Nevertheless,

Figure 3 shows our variance measure can capture this variability between curves since the variance decreases as the dispersion decreases.

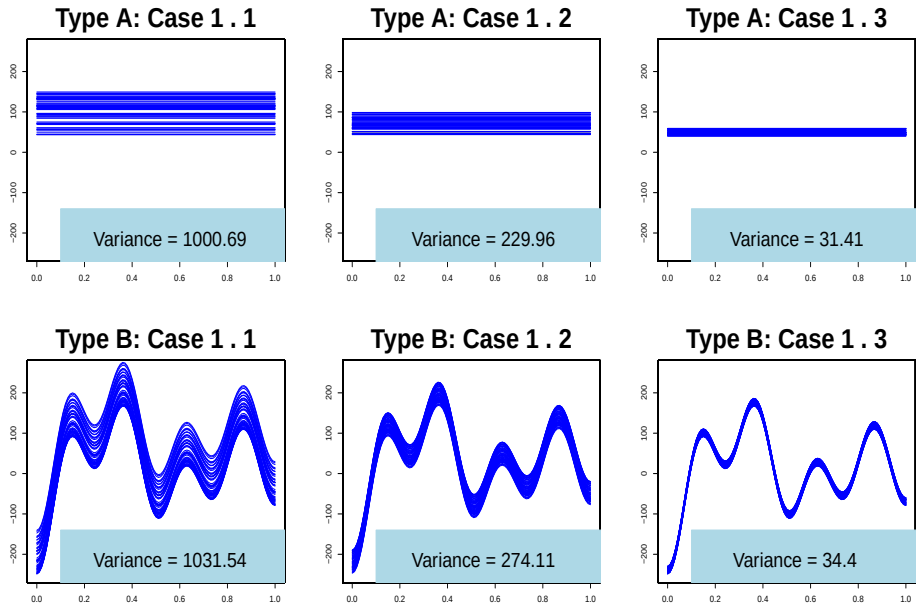


Figure 2. Scenarios variability considered for simulated curves in case 1 and scalar variances obtained. Type A (**top**) and Type B (**bottom**) curves.

Now, we use the simulation to illustrate that our variance approach is consistent. For this, we construct a functional data set with 500 type B functions. In this way, we obtain samples of different sizes, then calculate the absolute difference between the variances of the sample and the population. This step is repeated 500 times for each sample size. In Table 1, we report the mean of results for each sample size; in Figure 4, we show that our proposed variance converges in mean value to the population variance as the sample size increases.

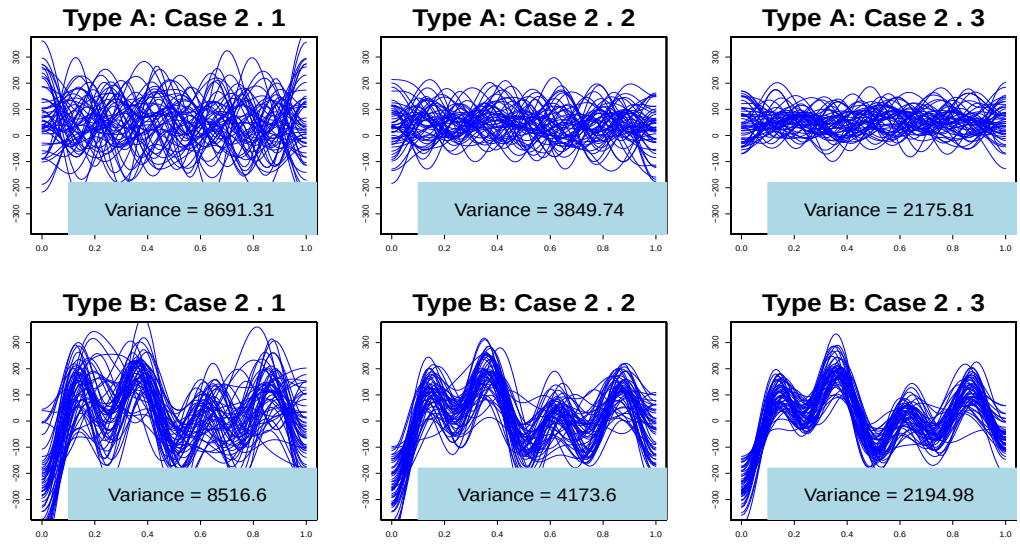


Figure 3. Scenarios of variability considered for simulated curves in case 2 and scalar variances obtained. Type A (**top**) and Type B (**bottom**) curves.

Table 1. Mean of the absolute difference between population variance and sample variance for each sample size.

Sample Size	Mean Absolute Difference	Sample Size	Mean Absolute Difference
5	1828.16	200	23.74
10	926.43	250	15.29
20	403.40	300	13.6
50	178.40	400	3.74
100	81.47	450	3.83
150	34.31	490	0.45

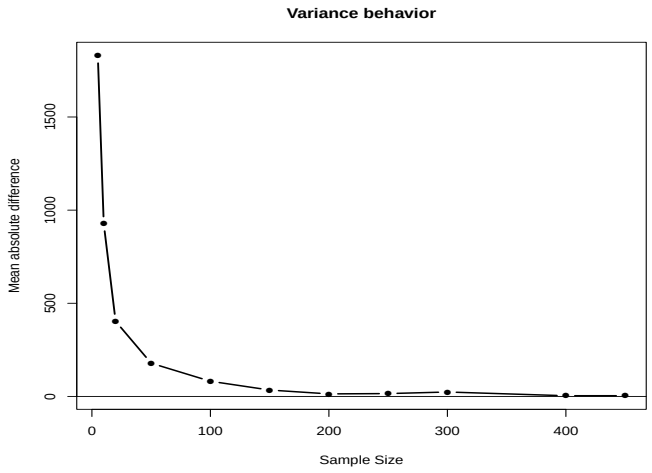


Figure 4. Behavior of the mean of the absolute differences when the sample size increases.

4. Application Examples

To show the interpretive advantages, we present two application example cases on actual data.

4.1. Example 1

In this example, we present a way of using our summary statistics in a functional analysis of variance (FANOVA). Therefore, we analyse the data collected by [1] corresponding to monthly temperature from 35 weather stations in Canada. Such stations are sorted, in four regions, according to their location: Atlantic, Continental, Pacific and Arctic.

Ref. [1] implement a FANOVA in order to evaluate the existence of statistically significant differences between the average annual temperature in the four geographic areas. However, because of the phenomenon dynamics, there might be a correlation in the four areas' temperature; hence, the conclusions from the FANOVA might be affected. Nonetheless, this is not considered by [1] since their functional cross-correlation does not permit to conclude whether there is a correlation between the functional data from the four areas, as indicated in Figure 5. In contrast, our correlation approach permits to determine if it is present between the four areas.

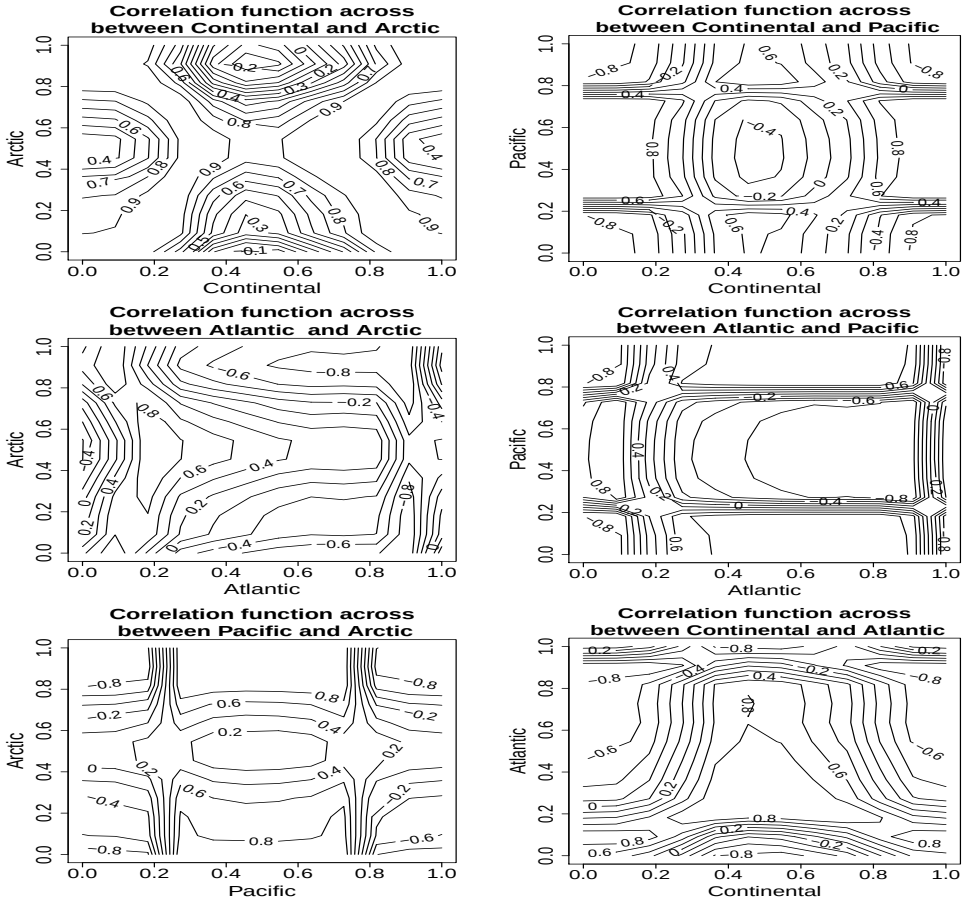


Figure 5. Cross-correlations of the functional variable temperature between the geographic areas.

In addition, Table 2 shows evidence of a strong correlation between the functional variables Arctic and Continental temperature. Furthermore, the functional variable Pacific and Continental temperature presents negative correlation, which indicates that when there is temperature rise in the Pacific area, there is a temperature decrease in the Continental area. In summary, our correlation matrix demonstrates that the temperature in the four areas is not independent since they present correlations different to zero.

Table 2. New correlation proposition of the functional variable temperature between the geographic areas.

Zone	Altlantic	Continental	Pacific	Arctic
Altlantic	1.00	0.45	−0.50	0.26
Continental	0.45	1.00	−0.62	0.82
Pacific	−0.50	−0.62	1.00	−0.22
Arctic	0.26	0.82	−0.22	1.00

4.2. Example 2

We provide an example of implementation on actual air pollution data in order to show the interpretive advantages of the summary statistics we suggest.

According to [46], particles in the air whose aerodynamic diameter is less than 2.5 μm ($PM_{2.5}$) may be considered as criteria air pollutants; therefore, prolonged exposure is harmful to human health. Consequently, $PM_{2.5}$ monitoring stations have been implemented in different cities around the world to monitor the pollutant's daily behaviour. As a result,

a relationship between anthropogenic activities and $PM_{2.5}$ production within an urban area has been identified. In consequence, the existing variability is of interest.

For instance, the Administrative Department of Environmental Management (DAGMA) monitors the amount of $PM_{2.5}$ in Cali, Colombia. To this end, they use two air quality monitoring stations which are part of the air quality surveillance system in the city. These stations are called *Base Aérea* (BA) and *Compartir* (CO). These stations are located approximately 3.5 km away from each other and regularly collect records of $PM_{2.5}$ concentration every 10 s. However, they only report the hourly average. Thus, we may collect at most twenty-four measurements per day from each station. In addition, we aim to find out which station presents more variability and if there is any correlation between the daily behaviour measurements.

Because of this problem, it is necessary to consider all the daily behaviour that is recorded by both stations. This and the variable's continuous nature make necessary to carry out a functional and descriptive analysis.

For this analysis, we include 29 days of year 2015 with all their twenty-four measurements at both monitoring stations. For illustration purposes, we construct 58 curves, 29 per station, in the subspace $\mathcal{H}$ of $L_2[0,1]$ of dimension 8 using a Fourier orthonormal basis. In total, we have 29 paired curves. The dimensions of both $\mathcal{H}$ and the basis obey construction techniques of functional data. Although such techniques are not the objective of this work, it is worth mentioning that the suggested approach is independent of them; therefore, it can be applied without loss of generality to any type of basis.

In Figure 6, we show the curves of the functional data of BA and CO stations. The X-axis has been set to $[0,1]$.

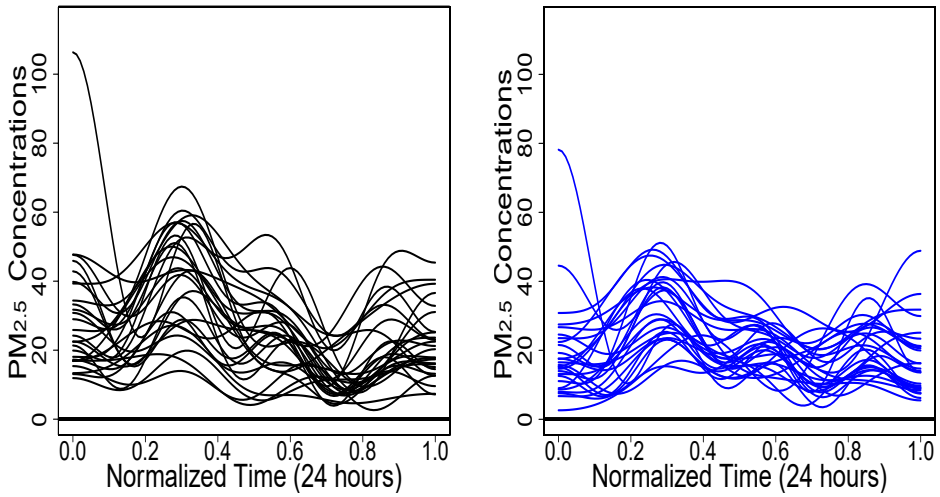
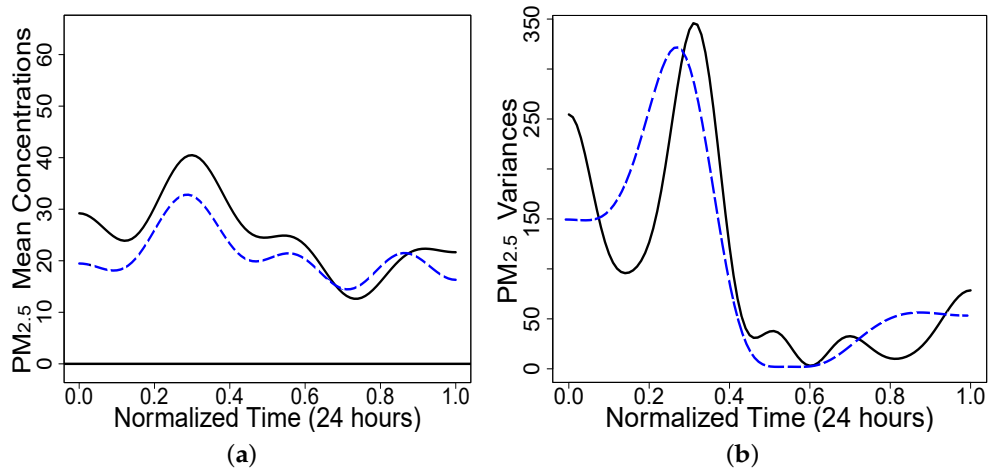


Figure 6. $PM_{2.5}$ functional data from station BA on the left and from station CO on the right.

In turn, Figure 7 shows the functional means and the curves of variances suggested by [1]. Moreover, Figure 7b demonstrates that it is not possible to decide which of the two stations experiences a larger variability.

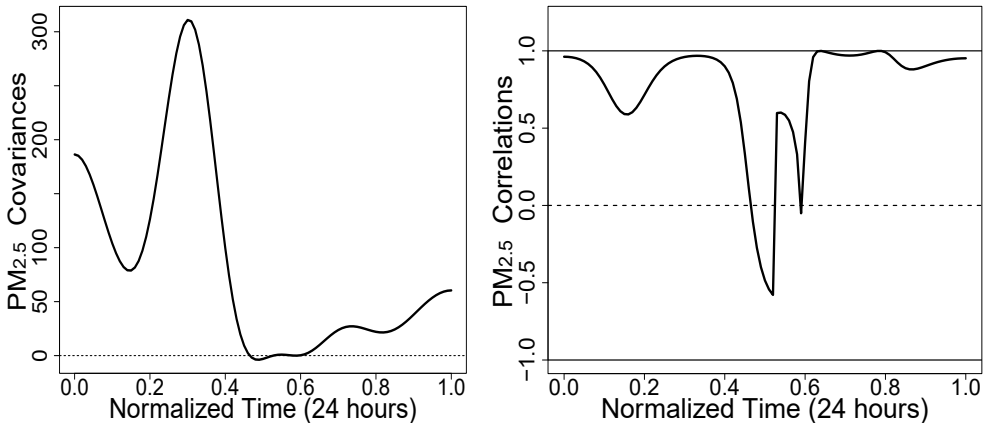


(**a**) (**b**)

Figure 7. Functional Means and Curves of Variances at both stations (**a**) Base Aérea (black solid line) and Compartir (blue dashed line) Functional Means. (**b**) Variance Curves in Base Aérea (black solid line) and Compartir (blue dashed line).

In addition, Figure 8a,b indicate a curve of covariances and a curve of correlations, respectively. However, these descriptive statistics suggested by [1] are not sufficient to decide whether the functional variables are correlated and to what extent.

We observe that the functional descriptive analysis suggested by [1] is not sufficient to provide an answer to the specific case in Cali, Colombia. However, our summary statistics for functional data, shown in Table 3, solve the problem because it allows us to observe that station BA presents more variability. This can be explained by BA's high industrial activity and its air and land traffic. Moreover, the high correlation between stations BA and CO can be also explained by their proximity and their paired data. This suggests that the pollutant might be airborne.



(**a**) (**b**)

Figure 8. Curves of covariances and curves of correlations from Definition 1; (**a**) Base Aérea and Compartir Covariance. (**b**) Base Aérea and Compartir Correlation.

Table 3. Proposed measures for functional data.

Station	Variance	Covariance	Correlation
Base Aérea	134.91	72.31	0.95
Compartir	76.24		

5. Conclusions

Because of the functional character of objects in FDA, it is reasonable to believe that a first approach would be the extension of concepts in their functional form. However, scalars and functions do not have the same properties or interpretations. Therefore, it is necessary to create treatment methods that are coherent according to the nature of the objects. In this sense, the proposed treatment method has important advantages because it puts aside the point-to-point treatment of the curve to treat the functional datum as a complete unit by means of the representation coefficients, given that each coefficient modifies a characteristic of the function and not just a point of it.

Some of the most important advantages provided by this approach are:

- It maintains the concepts of tendency, centrality, dispersion and association coherently, which results in an interpretive advantage.
- Given that each coefficient oversees a specific characteristic of the curve, when taking the arithmetic mean of the functional data through the coefficients per group of components, what is being taken is an "expected behavior" of each of the characteristics, which is conceptually coherent in terms of tendency.
- It offers operational advantages, given that the set of representation coefficients is finite and their elements are scalars, which facilitates the treatment of functional data because many of the operations with them can be transferred to the coefficients.
- The proposed variance allows the comparison of the variability of two groups of functional data in such a way that it is easy to calculate and interpret in terms of the "amount".
- It allows knowing how strong the joint variability of two datasets of functional data is, with interpretive and operational ease.
- It allows characterizing the functional data in terms of a probability distribution function for the coefficients, component-by-component, which facilitates the controlled simulation of the functional data and their analysis.
- Our new proposition of variance for functional data is also useful to determine which functional estimators are consistent within a set, i.e., which one decreases its variance when the number of functional observations increases. Furthermore, our proposition helps to determine which estimator shows less variance for the same number of functional observations; i.e., which one is more efficient.

To use our summary statistics, it is necessary that the functional data can be represented by a function basis, which limits its use to only functional data that have this characteristic. Moreover, all functional data must be represented with the same basis and function number.

Another limitation is that, because it is a new proposal, we still do not know how to assess the confidence of the estimates, which requires new simulation experiments.

To conclude, it is important to point out the need to create a theory based on methods for the analysis of functional data and to take advantage of the mathematical richness of continuous functions. Therefore, we present a formal theory to validate the proposed approach that can be taken as a starting point for future works, hoping that FDA can be transformed into a new form of statistics, say functional statistics.

Supplementary Materials: The R code used in the simulation is available at https://www.mdpi.com/article/10.3390/math11061317/s1 , pages S1 to S5.

Author Contributions: Conceptualization, C.L.U.-L.; Methodology, C.L.U.-L., M.E., D.P.O.-M., J.O.-O.; Formal analysis, C.L.U.-L.; writing—original draft preparation, C.L.U.-L.; Writing—review and editing, C.L.U.-L., M.E., D.P.O.-M., J.O.-O.; supervision, M.E., J.O.-O. All authors have read and agreed to the published version of the manuscript.

Funding: This paper and the APC was funded by the research group FQM-307 of the Government of Andalucia (Spain) and by the project A-FQM-66-UGR20 of the Consejería de Conocimiento, Investigación y Universidad, Junta de Andalucia (Spain) and the FEDER programme.

Data Availability Statement: The data set used in Section 4.1 is available in the package "fda" in the software R project (see in [47]) and the data set analysed in Section 4.2 are available in the web sites http://datos.cali.gov.co/dataset/datos-de-calidad-del-aire-en-la-estacion-compartir-2014-a-2021-en-santiago-de-cali/resource/34e0f68a-d488-4c83-8782-42a3d0126540 and http://datos.cali.gov.co/dataset/datos-de-calidad-del-aire-en-la-estacion-base-aerea-2013-a-2021-en-\santiago-de-cali/resource/6ab2160a-2bee-478b-aef2-d56887610513 for stations Compartir and Base Aérea respectively, accessed on 21 October 2022.

Acknowledgments: The authors acknowledge the support by the research group FQM-307 of the Government of Andalucia (Spain), and the project A-FQM-66-UGR20 of the Consejería de Conocimiento, Investigación y Universidad, Junta de Andalucia (Spain), and the FEDER programme. The authors acknowledge to "Administrative Department of Environmental Management" (DAGMA for its acronym in Spanish) for providing the data set used in Section 4.2.

Conflicts of Interest: Not applicable.

References

1. Ramsay, J.; Silverman, B. *Funcional Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2005.
2. Ramsay, J.; Silverman, B. *Applied Functional Data Analysis: Methods and Case Studies*; Springer: New York, NY, USA, 2002.
3. Stewart, K.J.; Darcy, D.P.; Daniel, S.L. Opportunities and Challenges Applying Functional Data Analysis to the Study of Open Source Software Evolution. *Stat. Sci.* **2006**, *21*, 167–178. [CrossRef]
4. Jank, W.; Shmueli, G. Functional Data Analysis in Electronic Commerce Research. *Stat. Sci.* **2006**, *21*, 155–166. [CrossRef]
5. Ferraty, F. *Recent Advances in Functional Data Analysis and Related Topics*; Springer: New York, NY, USA, 2011.
6. Horváth, L.; Kokoszka, P. *Inference for Functional Data with Applications*; Springer: New York, NY, USA, 2012.
7. Sørensen, H.; Goldsmith, J.; Sangalli, L.M. An introduction with medical applications to functional data analysis. *Stat. Med.* **2013**, *32*, 5222–5240. [CrossRef] [PubMed]
8. Srivastava, A.; Klassen, E.P. *Functional and Shape Data Analysis*; Springer: New York, NY, USA, 2016.
9. Wang, J.L.; Chiou, J.M.; Muller, H.G. Review of Functional Data Analysis. *Annu. Rev. Stat. Appl.* **2015**, *3*, 1–41.
10. Ramsay, J. When the data are functions. *Psychometrika* **1982**, *47*, 379–396. [CrossRef]
11. Ramsay, J.; Dalzell, C.J. Some tools for Functional Data Analysisl. *J. R. Stat. Soc.* **1991**, *3*, 572–593.
12. Grenander, U. Stochastic processes and statistical inference. *Ark. Mat.* **1950**, *1*, 195–277. [CrossRef]
13. Rao, C.R. Some statistical methods for comparison of growth curves. *Biometrics* **1958**, *14*, 1–17. [CrossRef]
14. Clarkson, D.B.; Fraley, C.; Gu, C.C.; Ramsay, J.O. *S+ Functional Data Analysis*; Springer: New York, NY, USA, 2005.
15. Rudin, W. *Functional Analysis*, 2nd ed.; Mc Graw Hill: Singapore, 1991.
16. Royden, H.L.; Fitzpatrick, P.M. *Real Analysis*, 4th ed.; Pearson Education Asia Limited: Beijing, China, 2010.
17. Billingsley, P. *Probability and Measure*, 3rd ed.; John Wiley and Sons: New York, NY, USA, 1995.
18. Conway, J.B. *A Course in Functional Analysis*, 2nd ed.; Springer: New York, NY, USA, 1990.
19. Dodge, Y. Arithmetic Mean. In *The Concise Encyclopedia of Statistics*; Springer: New York, NY, USA, 2008; pp. 15–18. [CrossRef]
20. Kenny, J.F.; Keeping, E.S. *Mathematics of Statistics. Part One*, 3rd ed.; D. Van Nostrand Company, Inc.: New York, NY, USA, 1954.
21. Cohn, D. *Measure Theory*, 2nd ed.; Birkhauser: Boston, MA, USA, 2013.
22. Fisher, R.A. *Statistical Methods for Research Workers*; Oliver and Boyd: London, UK, 1925.
23. Fisher, R.A. On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **1922**, *222*, 309–368.
24. Kaplansky, I. *Set Theory and Metric Spaces*, 2nd ed.; Allyn and Bacon Inc.: Boston, MA, USA, 1977.
25. Rudin, W. *Principles of Mathematical Analysis*, 3rd ed.; McGraw-Hill Inc.: New York, NY, USA, 1976.
26. Apostol, T.M. *Calculus. One-Variable Calculus, with an Introduction to Linear Algebra*, 2nd ed.; John Wiley and Sons, Inc.: New York, NY, USA, 1967; Volume 1.
27. Mathai, A.M.; Rathie, P.N. *Probability and Statistics*; The Macmillan Press Ltd.: London, UK, 1977.
28. Galton, F. Co-relations and their measurement, chiefly from anthropometric data. *Proc. R. Soc. Lond.* **1889**, *45*, 273–279.
29. MacCluer, B. *Elementary Functional Analysis*; Springer: New York, NY, USA, 2009.
30. Lax, P.D. *Functional Analysis*; Wiley Interscience: New York, NY, USA, 2002.
31. Rynne, B.P.; Youngson, M.A. *Linear Functional Analysis*, 2nd ed.; Springer Undergraduate Mathematics Series: London, UK, 2008.
32. Judson, T.W. *Abstract Algebra. Theory and Applications*; Orthogonal Publishing L3C: Dallas, TX, USA, 2018.
33. Gasser, T.; Hall, P.; Presnell, B. Nonparametric estimation of the mode of a distribution of random curves. *J. R. Stat. Soc. Ser.* **1998**, *60*, 681–691. [CrossRef]
34. Hsing, T.; Eubank, R. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*; Wiley: Chichester, UK, 2015.
35. Staicua, A.; Lahiri, S.; Carrol, R. Significance tests for functional data with complex dependence structure. *J. Stat. Plan. Inference* **2015**, *156*, 1–13. [CrossRef] [PubMed]

36. Aguilera, A.M.; Acal, C.; Aguilera-Morillo, M.C.; Jiménez-Molinos, F.; Roldán, J.B. Homogeneity problem for basis expansion of functional data with applications to resistive memories. *Math. Comput. Simul.* **2021**, *186*, 41–51. [CrossRef]
37. Aguilera, A.M.; Escabias, M.; Preda, C.; Saporta, G. Using basis expansions for estimating functional PLS regression: Applications with chemometric data. *Chemom. Intell. Lab. Syst.* **2010**, *104*, 289–305. [CrossRef]
38. Escabias, M.; Aguilera, A.M.; Valderrama, M.J. Principal component estimation of functional logistic regression: Discussion of two different approaches. *J. Nonparametric Stat.* **2004**, *16*, 365–384. [CrossRef]
39. Muscat, J. *Functional Analysis. An Introduction to Metric Spaces, Hilbert Spaces and Banach Algebras*; Springer: Cham, Switzerland, 2014.
40. Kreyszig, E. *Introductory Functional Analysis with Applications*; Wiley.: New York, NY, USA, 2006.
41. Hansen, V.L. *Functional Analysis Entering Hilbert Space*, 2nd ed.; World Scientific Publishing Co. Pte. Ltd.: Danvers, MA, USA, 2016.
42. Roman, S. *Advanced Linear Algebra*, 3rd ed.; Springer: New York, NY, USA, 2008.
43. Ferraty, F.; Vieu, P. *Nonparametric Functional Data Analysis Theory and Practice*; Springer: New York, NY, USA, 2006.
44. Herstein, I.N. *Abstract Algebra*, 3rd ed.; Prentice-Hall: Hoboken, NJ, USA, 1996.
45. Fraleigh, J.; Beauregard, R.A. *Linear Algebra*, 3rd ed.; Addison Wesley Longman: Boston, MA, USA, 1995.
46. World Health Organization. *WHO Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide*; World Health Organization: Geneva, Switzerland, 2005.
47. Ramsay, J.; Graves, S.; Hooker, G. Package 'fda', 2022. Available online: https://cran.r-project.org/web/packages/fda/fda.pdf (accessed on 10 December 2022).

Apéndice C

Repeated Measures in Functional Logistic Regression

Titulo: *"Repeated Measures in Functional Logistic Regression".*

Autores: *Cristhian Leonardo Urbano-Leon, Ana María Aguilera, Manuel Escabias.*

Revista: *Mathematics and Computers in Simulation.*

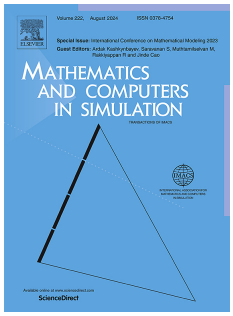
Factores de impacto de la revista:

Estado: Publicado, disponible en línea 10 de mayo de 2024.

doi: 10.1016/j.matcom.2024.05.002

Posición de autor: Autor principal.

Año	Factor de Impacto	Rango	Cuartil	Área
2022	4.6	4/267	$Q1$	Matemáticas, aplicada



Mathematics and Computers in Simulation 225 (2024) 66–77

Contents lists available at ScienceDirect

Mathematics and Computers in Simulation

journal homepage: www.elsevier.com/locate/matcom

Original articles

Repeated measures in functional logistic regression

Cristhian Leonardo Urbano-Leon *, Ana María Aguilera, Manuel Escabias

Department of Statistics and Operations Research, and Institute of Mathematics, University of Granada, Granada, Spain

ARTICLE INFO

Keywords:
Functional data
Functional logistic regression
Random effects
Repeated measures

ABSTRACT

We present a proposal to extend the functional logistic regression model – which models a binary scalar response variable from a functional predictor – to the case where the functional observations are not independent because the same functional variable is measured in the same individuals in different experimental conditions (repeated measures). The extension is addressed by including a random effect in the model. The functional approach of this model assumes that all functional objects are elements of the same finite-dimensional subspace of the space of square-integrable functions L_2 in the same compact domain allowing the functions to be treated through the basis coefficients on the basis that spans the subspace to which functional objects belong (basis expansion). This methodology usually induces a multicollinearity problem in the multivariate model that emerges, which is solved with the use of the functional principal components of the functional predictor, resulting in a new functional principal component random effects model. The proposal is contextualized through a simulation study that contains three simulation scenarios for four different functional parameters and considering the lack of independence.

1. Introduction

Functional data analysis (FDA) is a branch of statistics where the main studied objects are continuous functions, and not only scalar values as in classical statistics. FDA has its beginnings in the works of [28], but its popularity has increased since the works of [14,26,27] as a result of its multiple applications in a number of scientific disciplines and technological advances, allowing for increasingly precise measurements of continuous phenomena.

The developed theory for FDA has been possible thanks to the extension of concepts and methods from scalar statistics to functions, as can be seen in the works of [13,16,17,27]. Accordingly, one of the crucial techniques for scientific research is functional regression analysis, which attempts to find the relationship between a variable called dependent from one or more variables called covariates or explanatory variables. In functional regression it is possible to find different combinations between the types of variables and covariates. This is the case of functional logistic regression, which aims to model a binary random variable from a set of functional observations. This type of model is important in problems where the response can be categorized into two levels, commonly referred to as success and failure (see for example [11] in the case of peak levels of olive pollen). In this context [23] evaluates three approaches for functional logistic regression: dimension reduction using functional principal component analysis, penalized functional regression, and wavelet expansions in combination with Least Absolute Shrinking and Selection Operator penalization. Authors conclude that none of the three methods convince in their ability to reconstruct the parameter function, showing the difficulty of an accurate estimation of the functional parameter in this type of models.

* Corresponding author.
E-mail addresses: leonardourbano@correo.ugr.es (C.L. Urbano-Leon), aaguiler@ugr.es (A.M. Aguilera), escabias@ugr.es (M. Escabias).

https://doi.org/10.1016/j.matcom.2024.05.002
Received 15 November 2023; Received in revised form 30 April 2024; Accepted 3 May 2024
Available online 10 May 2024

There are different approaches to modelling a binary response, but in this work we consider a parametric approach that assumes the existence of a functional parameter. The interpretation of this parameter explains not only the relationship between response and covariates, but also provides an explanation of how changes in parameter generate changes in the odds ratio. Thus, the main objective of this approach is to obtain a correct estimation of the functional parameter that can be carried out by different methods, where the most appropriate for the functional logistic regression problem is the maximum likelihood (ML) method. ML requires observations to be independent, however, when the observations come from the same experimental unit or subject measured repeatedly in different conditions or in different periods of time, the framework is repeated measures. In this context, to assume independence is unrealistic [7] because the phenomenon of repetition can generate a correlation structure in the design matrix of the model [8]. This problem can be addressed by adding a random effect as is done in the scalar case (see [19]), where it is considered that part of the collected variability comes from the correlation structure caused by repeated measurements. Repeated measures in FDA context have a limited development in literature, its study focuses mainly on the curves comparison problem for instance. In [22] authors study the k-sample problem when the data are from the same subjects, proposing a statistic that takes into account the variability between groups. Subsequently, in [31,32] authors consider the variability in each group for a similar problem. Additionally, in [1] a basis expansion approach is used for functional analysis of variance with repeated measures. Most of these methods treat these problems of repeated measures in FDA by adding a random effect into the model, i.e. in functional mixed models (FMM) context. FMM have been developed in literature by some authors, e.g., in [20] a linear mixed effects model is formulated from a non-parametric context. In a same way, in [30] authors address a functional additive mixed models. In [25] Bayesian perspective is used by authors to introduce a functional logistic mixed-effects model for estimating learning curves in longitudinal experiments. In the last three cases, the random effect considered is functional. Alternatively, scalar random effects are included in [21] for a functional linear mixed model in the context of scalar on function regression – functional predictor and scalar response–.

On the other hand, as a consequence of the algebraic structure of some functional spaces, it is possible to consider the functional data as elements of a finite-dimensional subspace of square integrable functions space $L_2[a,b]$. This consideration allows the use of all vector space properties, as the representation of any element in terms of a basis of fixed functions. This representation produces, for each functional datum, a unique vector of scalars that are the coefficients of the linear combination of the elements of the basis that span the subspace to which the functional data belong. This treatment of functional data allows models to be reduced to a multivariate scalar problem, as already done in works such as those seen in [1,5,10,33], among other examples. However, the treatment of functional data through their basis coefficients within the context of regression can generate a problem known as multicollinearity i.e. correlation among the explanatory variables of a model, leading to high standard errors and another problems in the model parameters estimation [15]. This concern in the context of the functional logistic regression was worked on by [9,10], where functional principal components logistic regression (FPCLR) was introduced, and an extended FPCLR model and R-package were developed. FPCLR model provides the advantage that the new vectors of coefficients no longer present the problem of multicollinearity, since no correlation is theoretically guaranteed. Additionally, FPCLR model allows the reduction of the dimensionality of the problem, through the choice of a reduced number of functional principal components.

Functional logistic models for repeated measures on basis coefficients have problems of correlation attributable to repetition, and multicollinearity caused by the same basis representation for predictor and functional parameter (see [10]). The approach proposed here consists of the combination of two methodologies to address these issues: the random effect inclusion in the model to capture some of the variability attributable to repetition of functional observations, and the use of the functional principal components to deal with the possible multicollinearity problem that may exist in the model. As far as we know, the case of repeated measures for functional logistic regression model has not been considered in literature, much less the effect of multicollinearity in the model estimation.

This paper is divided into 4 sections. The introductory section shows some background in the context of functional data analysis, repeated measures and functional mixed models. Section 2 presents the theoretical framework on functional data and functional logistic regression model for repeated measures. Section 3 develops a simulation study with three different scenarios on four different functional parameters. Finally, Section 4 contains a summary and discussion of the main results and conclusion obtained.

2. Methodology

2.1. Functional data

There are different approaches to extend concepts from scalar statistics to continuous functions. One of these approaches considers that a functional datum $\mathcal{X}$ is an observation of a second order stochastic process $\{\mathcal{X}(t) : t \in [a,b]\}$, i.e. it satisfies the property of Eq. (1)

$$\int_a^b \mathcal{X}^2(t)dt < \infty. \tag{1}$$

We assume that $\mathcal{X} \in \mathcal{H} : \mathcal{H} \subset L_2[a,b] \wedge dim(\mathcal{H}) = d \in \mathbb{N}$, where $\mathcal{H}$ and $L_2[a,b]$ are vector spaces over the field $\mathbb{R}$, whose elements are square integrable functions on the same domain $[a,b]$, and which has a Hilbert space structure (see [18,29]) with inner product defined as in Eq. (2):

$$\langle f,g\rangle = \int_a^b f(t)g(t)dt,\ \forall f,g \in L_2[a,b]. \tag{2}$$

This inner product induces the usual norm and distance m defined by $\|f\| = \sqrt{\langle f, f\rangle}$ and $m(f,g) = \|f-g\|$ respectively. Here, d represents the dimension of the subspace $\mathcal{H}$. These assumptions allow the use of functional finite basis concept of $\mathcal{H}$. Consequently, given $\{\mathcal{X}_i\}_{i=1}^{n}$ a set of n functional data, then $\forall \mathcal{X}_i \in \mathcal{H}$, $\exists \left(a_{ij}\right)_{j=1}^{d} \in \mathbb{R}^d : \mathcal{X}_i(t) = \sum_{j=1}^{d} a_{ij}\phi_j(t)$, where the set $\boldsymbol{\Phi} = \{\phi_j\}_{j=1}^{d} \subset \mathcal{H}$ is a basis for $\mathcal{H}$, and the $a_{ij} \in \mathbb{R}$ are called basis coefficients or coefficients of representation of the $i-th$ functional datum in basis $\boldsymbol{\Phi}$.

The vector of basis coefficients for each function is unique, then it is possible to establish an isomorphism between the spaces $\mathcal{H}$ and $\mathbb{R}^d$. As a consequence, the use of a vector of basis coefficients $(a_{ij})_{j=1}^{d} \in \mathbb{R}^d$ instead of function $\mathcal{X}_i$ is conceptually coherent and simplifies the data treatment and the simulation process, reducing some functional problems to the multivariate scope, as can be seen for example in [2] for detecting changes in air pollution during the COVID-19 pandemic by using functional ANOVA.

It is important to note that different proposals exist for functional basis in order to obtain the representation in basis coefficients such as Fourier, B spline (see [27]), CONS basis (see [12]) or wavelets (see [4]).

2.2. Functional logistic model for repeated measurements

Let $L_2[a,b]$ be the vector space over $\mathbb{R}$, with Hilbert space structure, of square integrable functions defined on the interval $[a,b]$, and $\mathcal{H} \subset L_2[a,b]$ a subspace such that $dim(\mathcal{H}) = d \in \mathbb{N}$. Suppose N individuals measured in different experimental conditions for the same continuous domain, where the curve $\mathcal{X}_{is}$ is given by the $s-th$ functional repetition for the $i-th$ individual. The set of functional observations $\bigcup_{i=1}^{N}\{\mathcal{X}_{is}\}_{s=1}^{n_i} \subset \mathcal{H}$, being $card(\bigcup_{i=1}^{N}\{\mathcal{X}_{is}\}_{s=1}^{n_i}) = \sum_{i=1}^{N} n_i = n$ the total number of functional observations. Furthermore, let us suppose that $\{y_{is}\}_{i=1,s=1}^{N,n_i}$ is a set of binary responses that represent the success or failure of a phenomenon related to the functional observations, and which are defined as in the Eq. (3)

$$y_{is} = \begin{cases} 1 & if \quad success \\ 0 & if \quad failure. \end{cases} \tag{3}$$

Each y_{is}, $i = 1,2,\dots,N$; $s = 1,2,\dots\dots,n_i$ is an observation of a random variable Y, such that $Y|\mathcal{X}_{is} \sim Be(\pi_{is})$, where $Be(\pi_{is})$ represents a Bernoulli probability distribution with parameter $\pi_{is} = P(Y=1|\mathcal{X}_{is})$. The issue of repeated measures has been addressed in literature mainly through mixed models. These models add different random effects to the models themselves (see [8]). With this in mind, we propose the mixed functional logistic model for repeated measurements treatment and that is represented by Eq. (4)

$$l_{is} = ln\left[\frac{\pi_{is}}{1-\pi_{is}}\right] = \alpha + \int_a^b \mathcal{X}_{is}(t)\beta(t)dt + z_{is}u_i,\ i = 1,2,\dots,N; s = 1,2,\dots,n_i, \tag{4}$$

where l_{is} is the logarithm of the odds of success over failure, $E\left[Y|\mathcal{X}_{is}\right] = \pi_{is}$, $\int_a^b \mathcal{X}_{is}(t)\beta(t)dt$ is a fixed effect, u_i is the vector of random effects and z_{is} is a repetition indicator vector. This is the classical formulation of the mixed logit model, seen as a Generalized Linear Model (GLM) with logit transformation as link function (see [3,8] for scalar case).

The $\beta \in \mathcal{H}$ is the functional parameter whose accurate estimation is the objective of the methods proposed here. As can be seen in [10] for the functional logistic regression model, this functional parameter allows an interpretation of the relationship between the binary response and the functional predictor in terms of odds ratio. Then, since $\mathcal{X}_{is}, \beta \in \mathcal{H}; i = 1,2,\dots,N, s = 1,2,\dots,n_i$, there are vectors $(b_j)_{j=1}^{d}$, and $(a_{is,j})_{j=1}^{d}$, such that

$$\mathcal{X}_{is} = \sum_{j=1}^{d} a_{is,j}\phi_j \qquad \wedge \qquad \beta = \sum_{j=1}^{d} b_j\phi_j. \tag{5}$$

Then in Eq. (4) it follows

$$\begin{aligned} \int_a^b \mathcal{X}_{is}(t)\beta(t)dt &= \int_a^b \left[\sum_{j=1}^{d} a_{isj}\phi_j(t)\right]\left[\sum_{j=1}^{d} b_j\phi_j(t)\right]dt \\ &= \left[\sum_{j=1}^{d} a_{isj}b_j\left\|\phi_j\right\|^2\right] + \left[\sum_{j=1,k\neq j}^{d} a_{isj}b_k\left\langle\phi_j,\phi_k\right\rangle\right]. \end{aligned} \tag{6}$$

Thus, the functional logit model for repeated measures can be written in matrix form as

$$L = \mathbf{1}\alpha + A\Psi\mathcal{B} + ZU, \tag{7}$$

with $\mathbf{1}$ being a vector of ones, L the vector of the n logit transformations l_{is}, A the matrix of basis coefficients of curves, Ψ the matrix of inner products of the elements of the basis $\boldsymbol{\Phi}$, and $\mathcal{B}$ the parameter vector to be estimate that coincides with the vector of basis coefficients of functional parameter β. U is the random effects vector and Z the design matrix associated to U that contains the repetition framework.

In this way the functional logistic regression model for repeated measures is transformed into a multivariate logistic model (for repeated measures). Assuming a spherical Gaussian distribution for the random effects, the estimation of the basis coefficients $b_j, j = 1,2,\dots,d$ of the functional parameter β can be obtained by classic methods for repeated measures as restricted maximum likelihood (REML), and penalized iteratively re-weighted least squares (PIRLS) (see [6]). However, this formulation poses drawbacks caused by possible multicollinearity in the new explanatory variables, since it is not possible to ensure that the basis coefficients of curves are independent by columns. To deal with the multicollinearity issue, it is possible to restate the problem in terms of

the functional principal components (FPCs), see [10]. Let us thus consider the functional principal components of sample curves $\{\mathcal{X}_{is}\}_{i=1,s=1}^{N,n_i}$ as

$$\xi_{is,w} = \int_a^b \left(\mathcal{X}_{is}(t) - \overline{\mathcal{X}}(t)\right) f_w(t)dt, \tag{8}$$

where the functions $f_w \in \mathcal{H}$, $(w = 1, 2, \ldots, d)$ are the solutions to the *eigenequation*

$$\int_a^b C(r,t) f_w(r) ds = \lambda f_w(t), \tag{9}$$

with C being the functional sample covariance defined by

$$C(r,t) = n^{-1} \sum_{i=1}^{n} \left(\mathcal{X}_i(r) - \overline{\mathcal{X}}(r)\right)\left(\mathcal{X}_i(t) - \overline{\mathcal{X}}(t)\right). \tag{10}$$

The vectors ξ_w and ξ_j are independent for all $w \neq j$, and it is well known that each functional datum $\mathcal{X}_{is}$ can be approximated in terms of a reduced set of p eigenfunctions f_w and $\xi_{is,w}$ by

$$\mathcal{X}_{is} \approx \overline{\mathcal{X}} + \sum_{w=1}^{p} \xi_{is,w} f_w. \tag{11}$$

Then, from Eq. (4) we can obtain the functional principal components logit model for repeated measures (in terms of logit transformation) by

$$l_{is} = \alpha + \int_a^b \left(\overline{\mathcal{X}}(t) + \sum_{w=1}^{p} \xi_{isw} f_w(t)\right)\beta(t)dt + z_{is}u_i, \; i = 1, 2, \ldots, N; s = 1, 2, \ldots, n_i. \tag{12}$$

The model in the Eq. (12) can be expressed in matrix form as

$$L = \mathbf{1}\gamma_0 + \Gamma\gamma + ZU, \tag{13}$$

where Γ is the matrix of the functional principal components $(\xi_{is,w})$, $\gamma_0 = \alpha + \int_a^b \overline{\mathcal{X}}(t)\beta(t)dt$ and γ the vector of parameters with elements $\gamma_w = \int_a^b f_w(t)\beta(t)dt$. The model in Eq. (13) avoids the problem of multicollinearity since the principal components are incorrelated, so it is possible to use all the usual methods to obtain an estimate $\hat{\gamma}$ of the parameter vector γ, and through this an estimate of the original parameter vector $\mathcal{B}$, through $\hat{\mathcal{B}} = V\hat{\gamma}$, with V being the matrix of the basis coefficients of eigenfunctions f_w in $\mathcal{H}$.

3. Simulation

In order to evaluate the performance of the proposed methods we have developed a simulation study, considering three different scenarios:

- Scenario 1: Functional logit model without repeated measures.
- Scenario 2: Functional logit model with repeated measures.
- Scenario 3: Functional logit model with repeated measures, and multicollinearity

For all scenarios, four functional parameters β in $\mathcal{H}$ (subspace spanned by finite basis in Eq. (14)) were generated from expression $s\left(sin(w_1 \cdot t)\right)\left(cos(w_2 \cdot t)\right)$, where s, w_1, and w_2 are scalar values that, when modified, generate changes in the scale, oscillation and roughness of the β function. The 4 types of functional parameters considered can be seen in Fig. 1.

3.1. Scenario 1

The functional curves considered in this scenario belong to subspace $\mathcal{H}$ spanned by finite basis Φ with $d = 8$. The elements of the basis come from a complete orthonormal sequence (CONS), which provides multiple operational advantages thanks to its orthonormality, see [12,24]. These elements are described in Eq. (14), and shown in the left panel of Fig. 2.

$$\phi_j(x) = \begin{cases} 1 & if \quad j = 1 \\ \sqrt{2}\cos((j-1)\pi x) & if \quad 2 \leq j \leq d. \end{cases} \tag{14}$$

So, $n = 750$ curves $\{\mathcal{X}_i\}_{i=1}^{n}$ were considered by simulating their basis coefficients with uniform values, i.e. $(a_{i,j})_{j=1}^{d} = A_i \sim Unif[0.5, 3]$. A sample of 100 simulated curves can be seen in Fig. 2 (Right).

After simulating the predictor curves, we calculate the linear predictor l_i given by Eq. (6) using the functional parameters β_1, β_2, β_3 and β_4. After adding an error term $E \sim N(0, Id_{n\times n})$ to the linear predictor the response was simulated by using a Bernoulli distribution with probabilities given by $\exp(l_i)/(1 + \exp(l_i))$. Four different models were then fitted, and the β functional parameter estimated:

- Model 1: $L = \mathbf{1}\alpha + A\Psi\mathcal{B}$, i.e. the proposed model in Eq. (4) without random effects – called Classic Model (CL_Model) –. The estimates were obtained by ML.

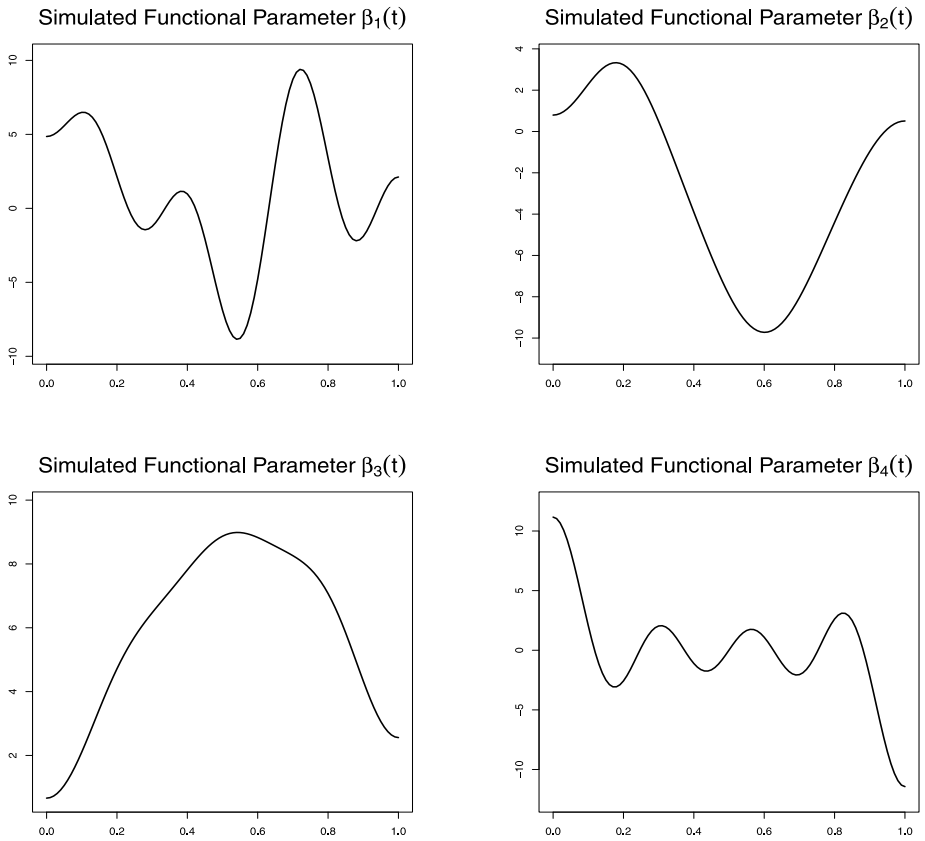


Fig. 1. Top left: Functional parameter $\beta_1(t)$ generated with the values $s = 10$, $w_1 = 15$ and $w_2 = 5$. Top right: Functional parameter $\beta_2(t)$ generated with the values $s = 10$, $w_1 = 3$ and $w_2 = 5$. Bottom left: Functional parameter $\beta_3(t)$ generated with the values $s = 80$, $w_1 = 0.3$, $w_2 = 1.5$. Bottom right: Functional parameter $\beta_4(t)$ generated with the values $s = 70$, $w_1 = 25$, $w_2 = 25$.

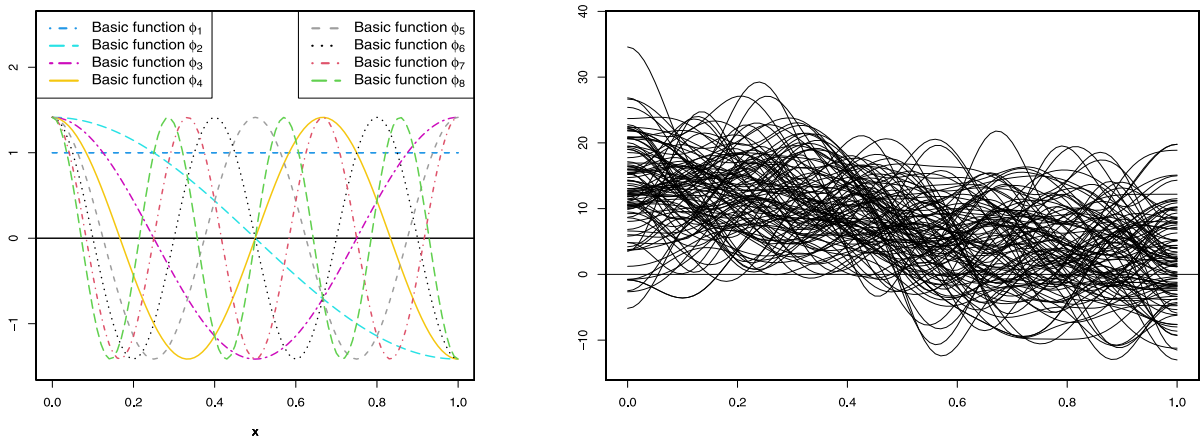


Fig. 2. Left: 8 functions of the basis Φ of subspace $\mathcal{H}$. Right: a sample of 100 functional data simulated as elements of subspace H.

- Model 2: $L = \mathbf{1}\alpha + A\Psi\mathcal{B} + ZU$, i.e. the proposed model in Eq. (4) with random effects – called Repeated Measures Model (RM_Model) –. The estimates were obtained by $REML$.
- Model 3: $L = \mathbf{1}\gamma_0 + \Gamma\gamma$ i.e. the proposed model in Eq. (12) model without random effects – called Classic Model on the Principal Components (PC_Model) –. The estimates were obtained by ML.
- Model 4: $L = \mathbf{1}\gamma_0 + \Gamma\gamma + ZU$ i.e. the proposed model in Eq. (12) with random effects – called Repeated Measurements Model on the Principal Components (RMPC_Model) –. The estimates were obtained by $REML$.

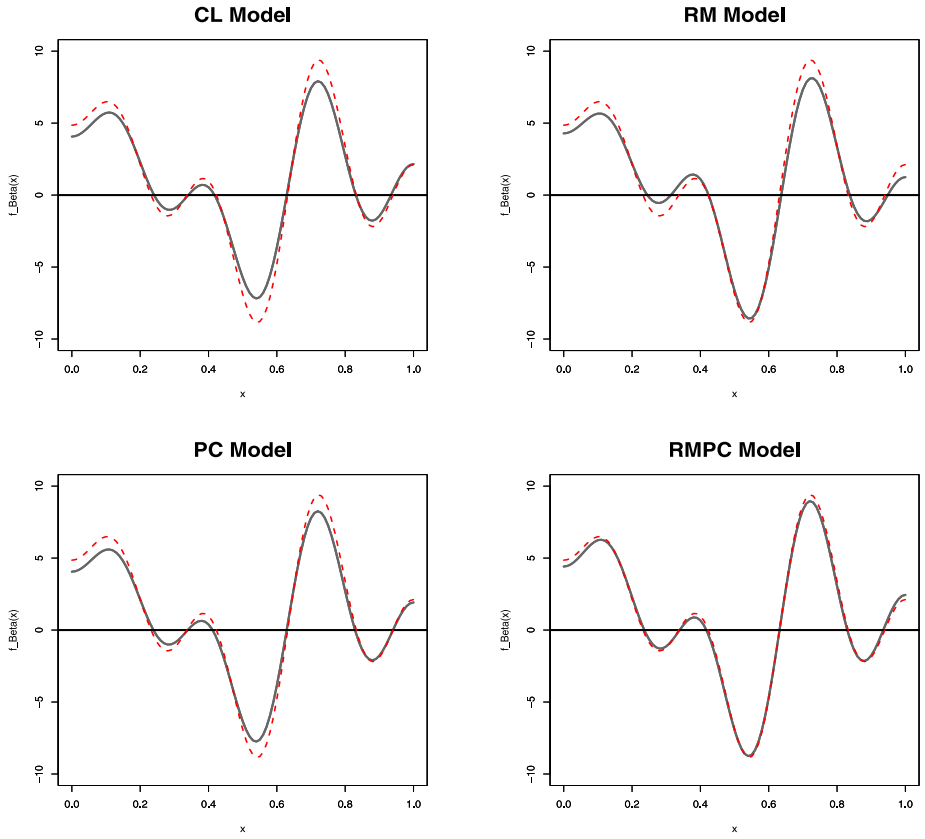


Fig. 3. For all figures: In red dashed line, the target functional parameter β_1, in black solid lines, the functional estimations $\hat{\beta}_1$. On the top left for the first model CL_Model. On top right for the second model RM_Model. Bottom left, for the third model PC_Model. Bottom right, for the fourth model $RMPC\_Model$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In PC_Model and $PCRM\_Model$, the number of functional principal components used was fixed as to explain 99% of the total variability. Fig. 3 shows an example of the simulated functional parameter β_1 and estimated $\hat{\beta}_1$ with four models for Scenario 1.

The performance of the logit models was carried out by the correct classification rate (CCR), an indicator of the percentage of items whose prediction perfectly matches the original observation. The CCR is a good indicator of model accuracy and can be calculated using the Eq. (15)

$$CCR = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i), \tag{15}$$

where y_i is the binary observation for the $i-th$ individual, and $\hat{y}_i$ is the prediction made for the same individual. Thus, the proportion of model successes is obtained with all the predictions of the model, moreover, the accuracy of the estimations of the functional parameters was tested by the integrated squared error ISE, defined by Eq. (16)

$$ISE = \int_a^b (\beta(t) - \hat{\beta}(t))^2 dt, \tag{16}$$

This process was carried out 100 times, wherein we obtained 100 functional parameter estimations $\{\hat{\beta}_i\}_{i=1}^{100}$, CCR and ISE for each model, The evaluation of the 100 simulations was tested by the averages of CCR and ISE – referred to as $MCCR$ and $MISE$ respectively – in addition to the standard deviation of the CCR ($SDCCR$). Furthermore, scalar variance for functional data ($SVFD$) developed by [33] was used to provide a scalar value of the variability of a set of curves within a single finite-dimensional subspace. This is useful for comparing the consistency of the estimates from the four models. The scalar variance for functional data is defined in Eq. (17)

$$SVFD = \frac{1}{n-1}\sum_{i=1}^{n}\int_a^b (\hat{\beta}_i(t) - \bar{\hat{\beta}}(t))^2 dt \tag{17}$$

where $\hat{\beta}_i(t)$ is the $i-th$ estimation of functional parameter $\beta_i(t)$, and $\bar{\hat{\beta}}(t)$ is the functional mean of the estimations. One of the advantages of $SVFD$ is that it can be calculated directly from the basis coefficients, offering operational advantages, even more so

Table 1
Accuracy measures in Scenario 1 for four target β functions.

Accuracy measure	Functional parameter β_1				Functional parameter β_2			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
$SVFD$	0.24	0.27	0.24	0.27	0.29	0.31	0.29	0.31
ISB	0.23	0.16	0.23	0.16	0.21	0.17	0.21	0.17
$MISE$	0.47	0.43	0.47	0.43	0.50	0.48	0.50	0.48
$MCCR$	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91
$SDCCR$	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.02
	Functional parameter β_3				Functional parameter β_4			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
$SVFD$	0.47	0.58	0.47	0.58	0.19	0.20	0.19	0.20
ISB	0.43	0.28	0.43	0.28	0.20	0.17	0.20	0.17
$MISE$	0.90	0.85	0.90	0.85	0.39	0.37	0.39	0.37
$MCCR$	0.92	0.93	0.92	0.93	0.89	0.89	0.89	0.89
$SDCCR$	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02

when the basis is orthonormal. Accordingly, the scalar variance for functional data in the subspace spanned by Φ orthonormal basis can be calculated as in Eq. (18)

$$SVFD = \sum_{j=1}^{d} V_j \tag{18}$$

where V_j is the variance of the $j-th$ vector of basis coefficients from estimations of $\beta(t)$. The bias in the functional estimations was calculated using the integrated squared bias ISB according to Eq. (19)

$$ISB = \int_a^b (\beta(t) - \bar{\hat{\beta}}(t))^2 dt, \tag{19}$$

where $\beta(t)$ is the simulated functional parameter, and $\bar{\hat{\beta}}$ is the mean of the functional estimations of $\beta(t)$. The ISB provides a general scalar measure of the bias in the estimates.

Fig. 4 shows the results of the 100 estimations of simulated functional parameters, for the four models in simulation Scenario 1. Here, there is no multicollinearity, and there is no correlation structure due to repeated measurements of the same individual. As expected in this case, the four models produce similar estimates, which can be verified through accuracy measures in the top left Table 1. The graphic results are consistent in all functional parameters.

In order to compare the accuracy of the fits for the four models in simulation Scenario 1, Table 1 shows the accuracy measures for the four simulated parameters considered. It is possible to note that the results in fact show stability even with changes in the form of the functional parameter. In all cases CCR are high, hanging around 90%, and there are almost negligible differences in $SVFD$, ISB and $MISE$ among models and functional parameters.

3.2. Scenario 2

In this scenario, the predictor curves were simulated by assuming the same subspace $\mathcal{H}$ spanned by the same finite basis Φ as Scenario 1. In this case, we assumed $N = 50$ individuals and $n_i = 15$ repetitions for $i = 1, 2 \ldots, N$, $n = 750$ curves in total. For the repeated measures the random effects were simulated by using a Gaussian distribution, i.e. $U \sim N(0, 3.5)$. The covariance matrix was generated without multicollinearity but the responses had random effect because of repeated measurements. The response was also simulated in the same terms as Scenario 1. As in Scenario 1, the process was replicated 100 times, the fits and the accuracy were tested by using the same measures and techniques as in that scenario.

Fig. 5 shows the results of the 100 estimates of the functional parameters for the four models in simulation Scenario 2. Here, no multicollinearity between the columns of the design matrix was considered, but there existed a correlation structure caused by repetition and that was added through the random effects simulation. Here, it is possible to observe that the models CL_Model and PC_Model – which use ML estimation – show a bias in the functional mean of the estimates, while in the models RM_Model and $RMPC\_Model$ – which use REML estimation – the functional means of the estimates are closer. You can also observe how including functional principal components in the models, in this scenario, has no effect on the accuracy and bias of the functional parameter estimates compared to not including them. On the other hand, when comparing the results obtained for the different functional parameters, one might suspect that the shape of them could be influencing the accuracy and bias of the estimates. For example, it can be observed that in functional parameter like β_2 and β_3, the discrepancies from not using random effects in the models are greater than in β_1 and β_4. In any case, the inclusion of a random effect in the model always improves the estimates.

In Table 2 (top left), as expected, the prediction ability of the four models is very accurate with similar and high CCR in all models. However, it is possible to observe an increase in bias and error in the estimates of the CL_Model and PC_Model models with respect to scenario 1, since the increase in the bias of the estimates is just a consequence of repeated measures. Despite this, the decreases from 2.65 to 0.26 in ISB and from 2.81 to 0.55 in $MISE$ show the importance of including the random effect in the

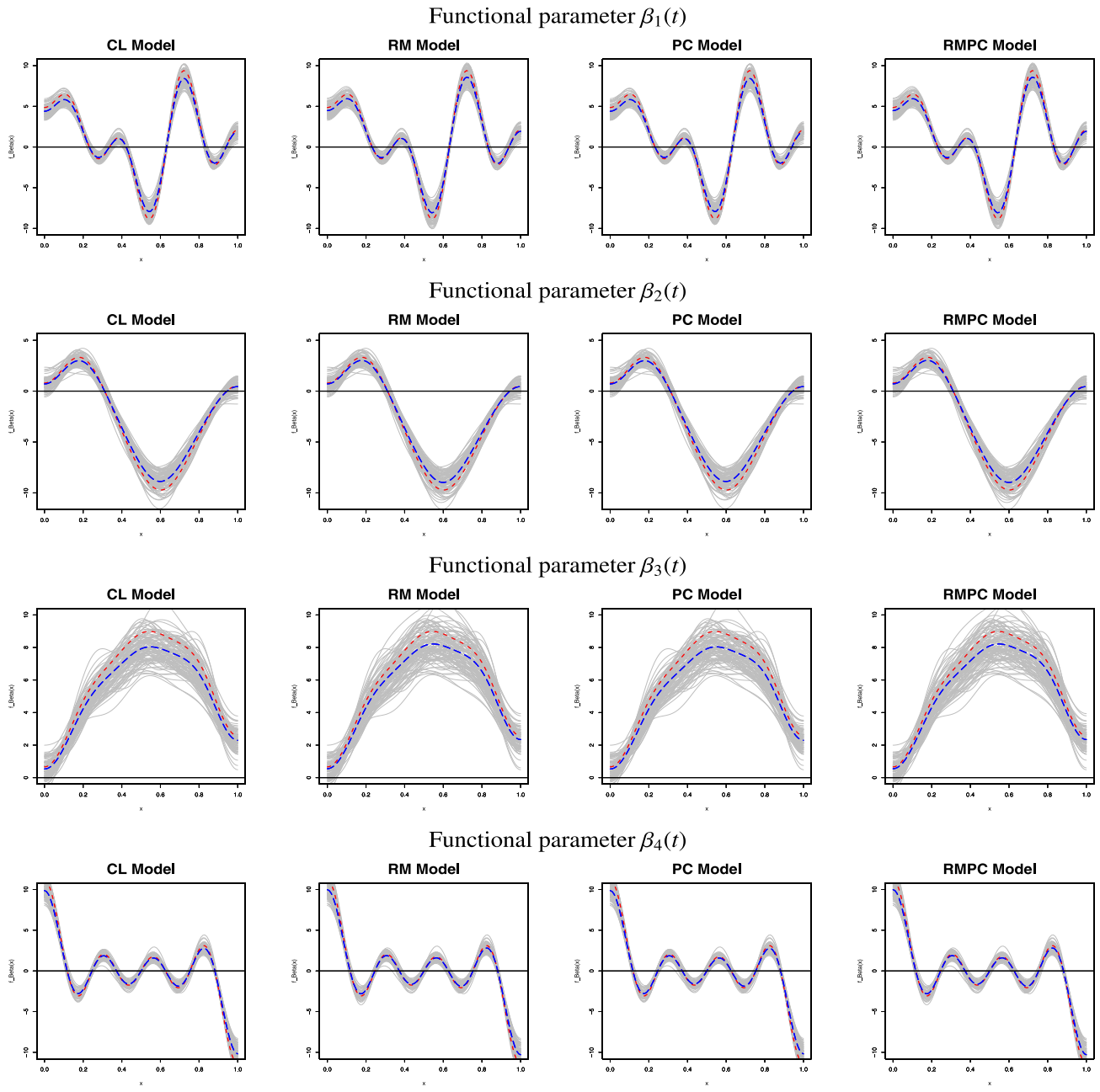


Fig. 4. Scenario 1: For all figures, in red dashed line the target functional parameter, in grey solid lines, the 100 functional estimations, in blue long dashed line, the functional mean of the 100 estimations. Each line shows for each functional parameter the results of fitted models CL_Model, RM_Model, PC_Model and $RMPC\_Model$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

logit model for repeated measures to improve the estimation of the functional parameter, and for obtaining a precise interpretation of the functional parameter in terms of odds ratios. As in Scenario 1, the results show stability decreasing ISB and $MISE$ in models with random effect, even with changes in the form of the functional parameter given by β_2, β_3 and β_4. In terms of the $SVFD$, as in scenario 1, an increase is observed in the models with random effect, which is because the REML method can increase the variance by reducing the bias in the estimates. This can be seen in the tested functional parameters β_1, β_2, β_3, and β_4, although the excessive increase in the $SVFD$ in β_3 for the RM_Model and $RMPC\_Model$ may be an indication that the it is influenced by the shape of the functional parameter.

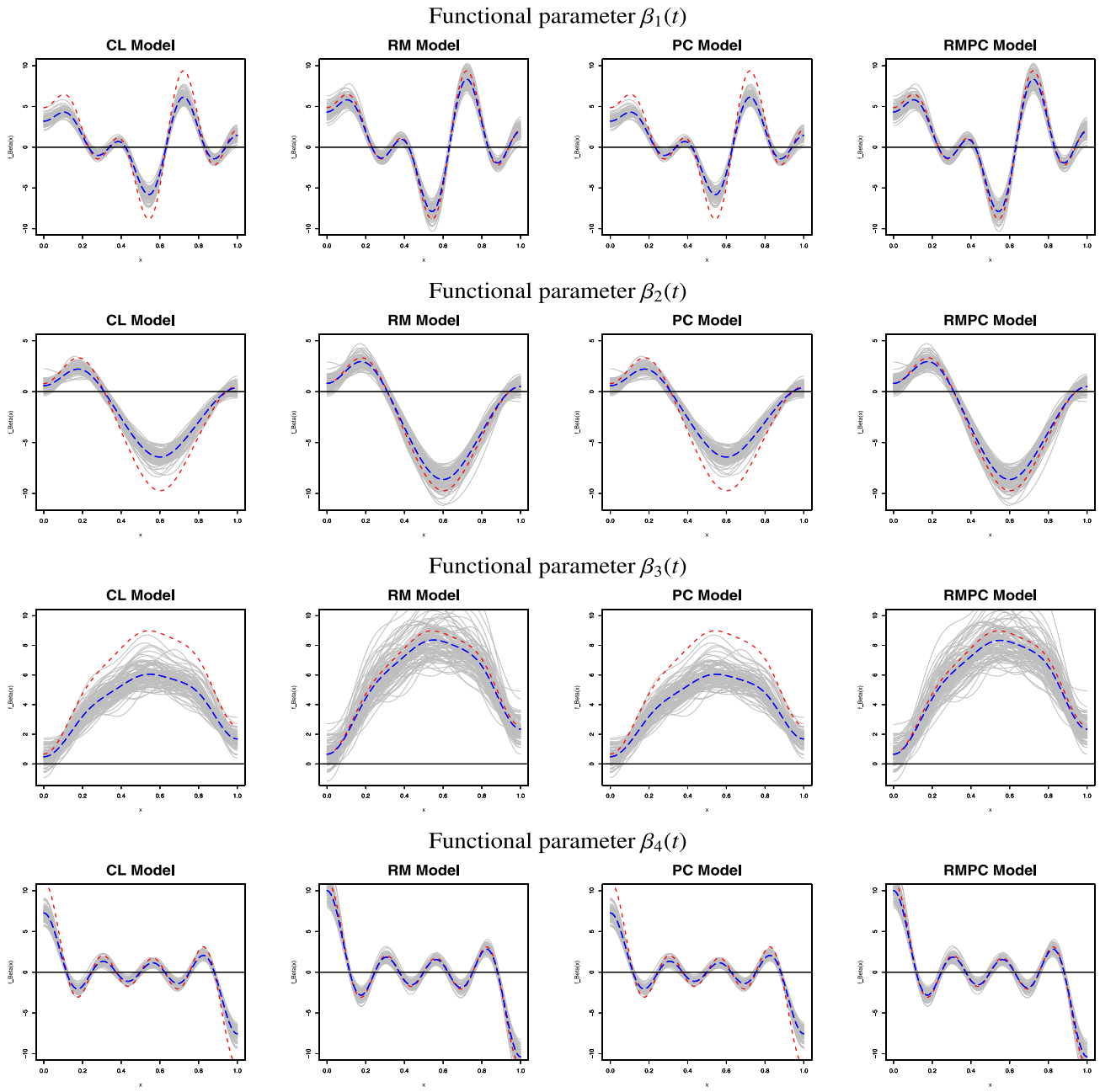


Fig. 5. Scenario 2: For all figures, in red dashed line the target functional parameter, in grey solid lines, the 100 functional estimations, in blue long dashed line, the functional mean of the 100 estimations. Each line shows for each functional parameter the results of fitted models CL_Model, RM_Model, PC_Model and $RMPC\_Model$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Scenario 3

In this scenario we deal with a more realistic case, where multicollinearity exists due to basis expansion representation of the functional objects of our models. Thus, $N = 50$ individuals and $n_i = 15$ repetitions for $i = 1, 2 \ldots, N$, $n = 750$ curves in total were now simulated with repeated measures and multicollinearity. In this scenario, a Normal distribution instead of Uniform was used for basis coefficients simulation of the functional predictors, i.e. $(a_{is,j})_{j=1}^{d} = A_j \sim N(0, \Sigma), i = 1, 2, \ldots, N$, where covariance matrix Σ was generated with multicollinearity. The response simulation, replication, fits, and accuracy evaluation were carried out as in the two previous scenarios.

Fig. 6 shows the results of the estimations of the functional parameters for simulation of this Scenario 3, which considers multicollinearity and correlation structure because of repetition. Here it can be seen that the four models have difficulty estimating

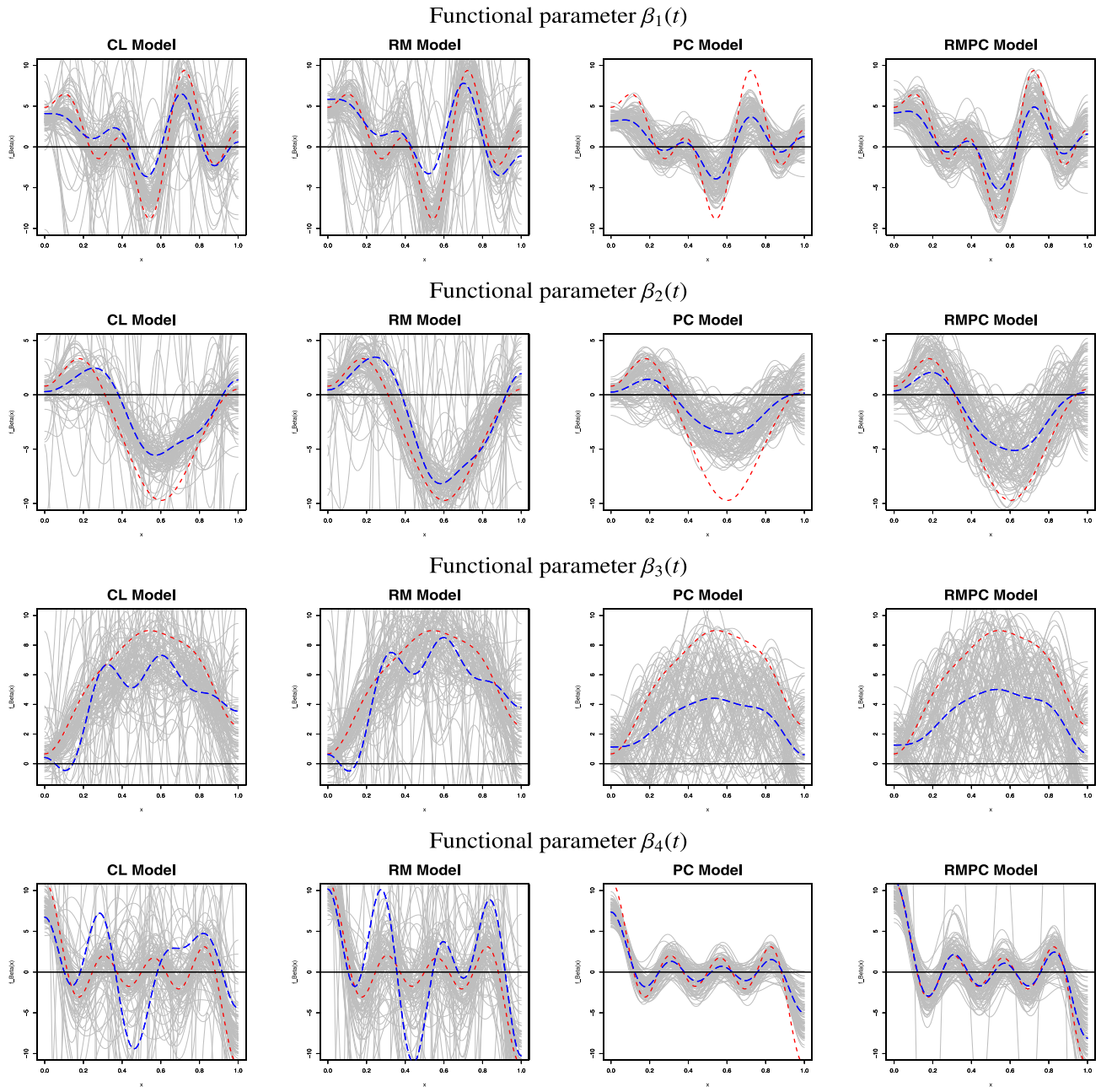


Fig. 6. Scenario 3: For all figures, in red dashed line the target functional parameter, in grey solid lines, the 100 functional estimations, in blue long dashed line, the functional mean of the 100 estimations. Each line shows for each functional parameter the results of fitted models CL_Model, RM_Model, PC_Model and $RMPC\_Model$ respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Accuracy measures in Scenario 2 for four target β functions.

Accuracy measure	Functional parameter β_1				Functional parameter β_2			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
$SVFD$	0.16	0.29	0.16	0.29	0.19	0.34	0.19	0.34
ISB	**2.65**	**0.26**	2.65	0.27	3.09	0.32	3.09	0.32
$MISE$	**2.81**	**0.55**	2.81	0.55	3.28	0.66	3.28	0.66
$MCCR$	0.87	0.92	0.87	0.92	0.88	0.93	0.88	0.93
$SDCCR$	0.02	0.01	0.02	0.01	0.02	0.01	0.02	0.01
	Functional parameter β_3				Functional parameter β_4			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
$SVFD$	0.35	0.92	0.35	0.85	0.12	0.21	0.12	0.21
ISB	4.50	0.18	4.50	0.20	1.89	0.15	1.89	0.15
$MISE$	4.84	1.09	4.84	1.04	2.01	0.36	2.01	0.36
$MCCR$	0.90	0.94	0.90	0.94	0.85	0.90	0.85	0.90
$SDCCR$	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01

Table 3
Accuracy measures in Scenario 3 for four target β functions.

Accuracy measure	Functional parameter β_1				Functional parameter β_2			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
$SVFD$	334.90	676.02	2.13	3.83	41.20	82.05	2.21	4.56
ISB	6.30	7.19	7.08	4.00	6.04	2.35	10.42	5.58
$MISE$	337.82	676.38	9.19	7.80	46.82	83.58	12.61	10.10
$MCCR$	0.90	0.94	0.89	0.93	0.84	0.91	0.83	0.90
$SDCCR$	0.03	0.02	0.03	0.02	0.04	0.02	0.04	0.03
	Functional parameter β_3				Functional parameter β_4			
	CL	RM	PC	RMPC	CL	RM	PC	RMPC
$SVFD$	147.34	165.11	5.75	7.50	1853.57	2692.48	1.24	19.52
ISB	4.18	2.48	11.51	8.84	20.05	22.81	3.56	0.78
$MISE$	150.02	165.93	17.20	16.26	1854.89	2688.09	4.79	20.10
$MCCR$	0.91	0.93	0.90	0.92	0.95	0.97	0.94	0.97
$SDCCR$	0.03	0.02	0.04	0.03	0.01	0.01	0.01	0.01

the target functional parameter, with the cases in the CL_Model and RM_Model, being notable since multicollinearity produces bad estimations with dramatic differences. Although the principal components models (PC_Model and $RMPC\_Model$) improve by decreasing the bias, error and variance of estimates, and produce stable results if compared to the other two models, this is not enough for β_2 and β_3, where no methods are able to provide suitable estimates. As in scenario 2, in this scenario, there is suspicion that the shape of the parameter function may influence the accuracy of the estimates. However, for all parameter functions considered, indicate that only the inclusion of a random effect does not improve the estimates in the presence of multicollinearity, so the use of functional principal components is necessary for a more precise estimation. On the other hand, the use of functional principal components alone does not improve the estimates in the case of repeated measures. It is the combination of both methodologies that produces a significant gain in the estimates. These conclusions can also be checked in Table 3.

4. Conclusions

In this work we propose functional principal components logistic regression for modelling a binary response variable from a functional covariate, when the observations are of repeated functional type. It is important to note here that the fundamental contribution sought is the appropriated estimation of the functional parameter because, as Section 1 indicates, the goal of the functional logistic model – from a parametric perspective – is the interpretation of the functional parameter, which will be realistic as long as the estimates recover the shape of the target functional parameter. From this point of view, our conclusions about the model from simulation results are the following:

- The inclusion of a random effect in the functional logistic model is effective for improving the estimation of the functional parameter in the case of functional repeated measures. As can be seen in all scenarios the inclusion of the random effect significantly improves the prediction of the response as well as the estimation of the functional parameter and, therefore, improves the interpretation.
- The use of functional principal components allows the estimation of the functional parameter to be improved, even in the random effects model in the presence of multicollinearity.
- Although the results in Section 3 show some regularity in model performance when the functional parameter is changed, the shape of the parameter could be influencing the estimates. This is more evident in Scenario 3, where it is shown that in the

presence of multicollinearity and repeated measures, no model produce suitable estimates for parameter functions with low variability along their trajectory, like β_2 and β_3. However, in the presence of multicollinearity and repeated measures, using the principal component model with random effects ($RMPC\_Model$) is appropriate for functions with higher variability, such as β_1 and β_4.

According to this last item, future model evaluation studies should examine the sensitivity of the model to changes in internal variability structures of functional parameters.

CRediT authorship contribution statement

Cristhian Leonardo Urbano-Leon: Conceptualization, Formal analysis, Investigation, Methodology, Simulation, Visualization, Writing – original draft, Writing – review & editing. **Ana María Aguilera:** Supervision, Writing – review & editing. **Manuel Escabias:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision.

Acknowledgements

The authors acknowledge the support by PID2020-113961GB-I00 project of the Spanish Ministry of Science and Innovation (also supported by the FEDER program), research group FQM-307 of the Autonomous Government of Andalusia (Spain) and the IMAG Maria de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033. Funding for open access charge: Universidad de Granada / CBUA.

References

[1] C. Acal, A.M. Aguilera, Basis expansion approaches for functional analysis of variance with repeated measures, Adv. Data Anal. Classif. 186 (2023) 291–321, http://dx.doi.org/10.1007/s11634-022-00500-y.
[2] C. Acal, A.M. Aguilera, A. Sarra, A. Evangelista, T. Di-Battista, S. Palermi, Functional ANOVA approaches for detecting changes in air pollution during the COVID-19 pandemic, Stoch. Environ. Res. Risk Assess. (2022) 1083–1101, http://dx.doi.org/10.1007/s00477-021-02071-4.
[3] A. Agresti, Foundations of Linear and Generalized Linear Models, first ed., John Wiley & Sons, New Jersey, 2015.
[4] A.M. Aguilera, M. Escabias, F.A. Ocaña, M.J. Valderrama, Functional wavelet–based modelling of dependence between lupus and stress, Methodol. Comput. Appl. Probab. 17 (4) (2015) 1015–1028, http://dx.doi.org/10.1007/s11009-014-9424-5.
[5] A.M. Aguilera, M. Escabias, C. Preda, G. Saporta, Using basis expansions for estimating functional PLS regression: Applications with chemometric data, Chemometr. Intell. Lab. Syst. 104 (2) (2010) 289–305, http://dx.doi.org/10.1016/j.chemolab.2010.09.007.
[6] D. Bates, R: Computational Methods for Mixed Models, R Foundation for Statistical Computing, Vienna, Austria, 2023, URL https://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf.
[7] M. Crowder, D. Hand, Analysis of Repeated Measures, 1st ed., Chapman and Hall, 1990.
[8] C.S. Davis, Statistical Methods for the Analysis of Repeated Measurements, Springer-Verlag New York, Inc., 2002.
[9] M. Escabias, A.M. Aguilera, C. Acal, LogitFD: An R package for functional principal component logit regression, R J. 14 (3) (2022) 231–248, http://dx.doi.org/10.32614/RJ-2022-053.
[10] M. Escabias, A.M. Aguilera, M.J. Valderrama, Principal component estimation of functional logistic regression: Discussion of two different approaches, J. Nonparametr. Stat. 16 (2004) 365–384.
[11] M. Escabias, M.J. Valderrama, A.M. Aguilera, M.H. Santofimia, M.C. Aguilera-Morillo, Stepwise selection of functional covariates in forecasting peak levels of olive pollen., Stoch. Environ. Res. Risk Assess. 27 (2013) 367–376, http://dx.doi.org/10.1007/s00477-012-0655-0.
[12] R.L. Eubank, Nonparametric Regression and Spline Smoothing, Second Edition, Marcel Dekker Inc, New York, 1998.
[13] F. Ferraty, Recent Advances in Functional Data Analysis and Related Topics, Springer, 2011.
[14] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis Theory and Practice, Springer, 2006.
[15] T.B. Fomby, S.R. Johnson, R.C. Hill, Advanced Econometric Methods, Springer, New York, NY, 1984.
[16] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Springer, 2012.
[17] T. Hsing, R. Eubank, Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators, First, Wiley, 2015.
[18] P.D. Lax, Functional Analysis, Wiley Interscience, 2002.
[19] M.J. Lindstrom, D.M. Bates, Nonlinear mixed effects models for repeated measures data, Biometrics 46 (3) (1990) 673–687.
[20] Z. Liu, W. Guo, Functional mixed effects models, WIREs Comput. Stat. 4 (6) (2012) 527–534, http://dx.doi.org/10.1002/wics.1226.
[21] W. Ma, L. Xiao, B. Liu, M.A. Lindquist, A functional mixed model for scalar on function regression with application to a functional MRI study, Biostatistics 22 (3) (2019) 439–454, http://dx.doi.org/10.1093/biostatistics/kxz046.
[22] P. Martínez-Camblor, N. Corral, Repeated measures analysis for functional data, Comput. Statist. Data Anal. 55 (12) (2011) 3244–3256.
[23] S.N. Mousavi, H. Sorensen, Functional logistic regression: A comparison of three methods, J. Stat. Comput. Simul. 88 (2) (2018) 250–268, http://dx.doi.org/10.1080/00949655.2017.1386664.
[24] J. Olaya, Metodos de Regresión no Paramétrica, Universidad del Valle, Colombia, 2012.
[25] G. Paulon, R. Reetzke, B. Chandrasekaran, A. Sarkar, Functional logistic mixed-effects models for learning curves from longitudinal binary data, J. Speech, Lang., Hear. Res. 62 (1) (2019) 543–553, http://dx.doi.org/10.1044/2018_JSLHR-S-ASTM-18-0283.
[26] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, 2002.
[27] J.O. Ramsay, B.W. Silverman, Funcional Data Analysis, 2da., Springer, 2005.
[28] C.R. Rao, Some statistical methods for comparison of growth curves, Biometrics 14 (1958) 1–17.
[29] W. Rudin, Functional Analysis, second ed., Mc Graw Hill, 1991.
[30] F. Scheipl, A.M. Staicu, S. Greven, Functional additive mixed models, J. Comput. Graph. Statist. 24 (2) (2015) 477–501, http://dx.doi.org/10.1080/10618600.2014.901914.
[31] L. Smaga, Repeated measures analysis for functional data using box-type approximation: With applications, REVSTAT-Stat. J. 17 (4) (2019) 523–549, http://dx.doi.org/10.57805/revstat.v17i4.279.
[32] L. Smaga, A note on repeated measures analysis for functional data, AStA Adv. Stat. Anal. 104 (2) (2020) 117–139, http://dx.doi.org/10.1007/s10182-018-00348-8.
[33] C.L. Urbano-Leon, M. Escabias, D.P. Ovalle-Munoz, J. Olaya-Ochoa, Scalar variance and scalar correlation for functional data, Mathematics 11 (1317) (2023) 1–20, http://dx.doi.org/10.3390/math11061317.