UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Estadística para datos en espacios no euclídeos

Algunas contribuciones

Leonardo Fabian Moreno Romero

Programa de Posgrado en Matemática
Facultad de Ciencias, Pedeciba
Universidad de la República

Montevideo – Uruguay
Abril de 2019

UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Estadística para datos en espacios no euclídeos

Algunas contribuciones

Leonardo Fabian Moreno Romero

Tesis de Doctorado Presentada al Programa de Postgrado en Matemática, Facultad de Ciencias de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Doctor en Matemática.

Director de tesis:
Ph.D. Prof. Ricardo Fraiman

Codirector:
Ph.D. Prof. Fabrice Gamboa

Montevideo – Uruguay
Abril de 2019

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Ph.D. Prof. Fabrice Gamboa

Ph.D. Prof. Ernesto Mordecki

Ph.D. Prof. Agnès Lagnoux

Ph.D. Prof. José Rafael León

Ph.D. Prof. Pablo Lessa

Ph.D. Prof. Ricardo Fraiman (Suplente)

Montevideo – Uruguay
Abril de 2019

Para mi padre...

Agradecimentos

En primera instancia quiero agradecer a mi esposa y mi familia quienes, sin todavía comprender mucho a que me dedico, me siguen apoyando día a día.

Por otro lado a amigos y compañeros de trabajo que me han alentado en todo momento para finalizar esta etapa, en particular a Marco Scavino y a Alejandro Cholaquidis (quién se tomó el trabajo de leer parte de la tesis).

Pero fundamentalmente a mis tutores, donde me es difícil llevar a palabras el agradecimiento enorme que tengo con ambos. Por abrirme las puertas, sentir que siempre estaban presentes y tener el privilegio de poder trabajar junto a ellos. Por contagiarme con su optimismo y alegría que me dieron constantemente una enseñanza no sólo en el aspecto profesional. A Ricardo que sin su apoyo esta tesis no hubiera sido posible, por darme la oportunidad de trabajar junto a él, de valorar mis ideas (en general defectuosas) y enseñarme que significa investigar. A Fabrice por su disponibilidad en todo momento (a pesar de la distancia) y apoyo incondicional, intentando siempre darme la mayor cantidad de opciones para permitirme crecer en mi formación como investigador.

Por último, a la CSIC por su apoyo mediante una beca CAP de doctorado, a la Facultad de Ciencias y al PEDECIBA por permitirme realizar este doctorado.

Prólogo

Como forma de titular esta tesis, podemos decir que intenta aportar sobre diversos aspectos de la estadística, en particular cuando los datos toman valores sobre determinados espacios no euclideanos o euclideanos pero de dimensión elevada donde la estadística clásica no esta diseñada para brindar respuestas eficientes. Es decir, el objetivo perseguido es transferir algunos métodos relevantes en la estadística clásica a ciertos espacios probabilísticos en donde los mecanismos tradicionales no pueden aplicarse de manera directa. En tal sentido es de relevante interés el estudio en dos posibles paradigmas:

- El espacio de probabilidad tiene estructura euclídea de dimensión alta o infinita (espacios funcionales).
- El espacio de probabilidad no tiene estructura vectorial. En este último punto nos restringiremos a las variedades Riemannianas.

En ambos casos el propósito base de este trabajo es poder poner en acción métodos ya desarrollados en $\mathbb{R}^d$ con d pequeño al caso en cuestión. En referencia al primer punto la idea clave es poder reducir la dimensión del espacio a través de proyecciones unidireccionales al azar. Con respecto al segundo punto, mediante una distancia conveniente (la distancia geodésica) y un transformación local adecuada (el mapa exponencial), la meta es poder generalizar algunas técnicas de estadística clásica al campo de las variedades Riemannianas.

A continuación esbozamos una breve panorámica general de la tesis desde el punto de vista de los principales conceptos que son los pilares en el desarrollo del trabajo.

La dimensionalidad del espacio

Una preocupación actual en la comunidad es la extensión de métodos clásicos del análisis estadístico cuando la dimensión del espacio es elevada, poder

sortear las limitaciones subyacentes es una tarea desafiante (ver Donoho et al. (2000)).

En la estadística tradicional, al analizar cierto fenómeno, es habitual que cada observación sea un vector conformado por las mediciones de un conjunto (en general reducido) de variables específicamente elegidas para el análisis. Actualmente, la recopilación intensiva, la automatización, la sistematización y la capacidad de almacenamiento provocan una "explosión" en las bases de datos.

La tendencia es un aumento en el número de observaciones pero aún más en el número de variables, es decir, existe una "hiperinformación" por instancia observada. Los datos son por ejemplo curvas, imágenes, películas o códigos genéticos, en donde cada observación está conformada por cientos, miles o incluso millones de atributos.

Los estadística clásica simplemente no está diseñada para afrontar esta tipología de datos. En particular, los métodos no paramétricos se ven afectados por la llamada "maldición de la dimensionalidad" (Bellman et al. (1961)), lo que provoca que en espacios de dimensión elevada, haya escasez de datos entorno de punto. Esto impacta de forma directa en la velocidad de la convergencia de los estimadores clásicos, causando que sean necesarias muestras de un tamaño irreal para obtener errores admisibles.

Una estrategia posible para afrontar este problema es poder encontrar una transformación adecuada del espacio de los datos a uno de menor dimensión, que preserve (en el mayor grado posible) la información relevante según la finalidad del problema. Diversos enfoques pueden se considerados para dicho fin. Algunas técnicas de reducción lineal como el *análisis de componentes principales* (PCA, Pearson (1901)) o el *análisis factorial* (AF, Spearman (1904)) ya llevan más un siglo de evolución y de diversas extensiones.

Otra técnica lineal de desarrollo más reciente son las *proyection pursuit* (ver Friedman and Tukey (1974)) basada en la proyección de la medida de probabilidad sobre una "grilla" de direcciones. También en esta línea encontramos un interesante y probabílistico método como lo son las *proyecciones al azar* (ver Dasgupta (1999) y Dasgupta and Gupta (2003)) en donde se proyectan los datos sobre un espacio k-dimensional elegido al azar, con k significativamente menor a la dimensión del espacio original.

Recientemente en Cuesta-Albertos et al. (2007) y Cuevas and Fraiman (2009) se desarrolla una metodología novedosa basada en proyecciones uni-

direccionales al azar. Ellos demuestran que es suficiente para caracterizar la distribución (bajo ciertos supuestos) sólo una proyección unidireccional, si esta es elegida al azar.

El segundo capítulo de la tesis se edifica sobre esta última línea metodologíca mencionada. Se construyen, mediante proyecciones al azar unidireccionales, un test de simetría y otro de independencia de sencilla aplicación, de baja complejidad computacional, con el debido sustento teórico y con una buena performance tanto en espacios de dimensión finita como infinita.

En referencia a la reducción de la dimensionalidad, otros procedimientos no lineales han sido diseñados en los úlimos años con el fin de detectar patrones más complejos (ver por ejemplo Tenenbaum et al. (2000) y Lee and Verleysen (2007)). En particular algunos métodos parten del supuesto que los datos, si bien se encuentran en un espacio de dimensión elevada, tienen una cierta estructura. Es decir, se encuentran en un entorno por ejemplo de una variedad Riemanniana inmersa en el espacio original pero de dimensión intrínseca mucho menor (ver Lin and Zha (2008)). Luego de estimada dicha variedad y proyectados los datos sobre ella (lo cual no es un objetivo perseguido en la tesis) es necesario poder hacer inferencia sobre esta variedad, lo cual podría enmarcarse como un segundo objetivo trazado en este trabajo.

La estructura del espacio

En muchos casos los datos se encuentran sobre un espacio de dimensión menor que el espacio original, como por ejemplo una variedad Riemanniana, pero con la desventaja que podríamos ya no contar con una estructura de espacio vectorial.

Dicha variedad puede ser estimada (como mencionamos anteriormente) o ya es intrínseca a los datos del problema. Es decir, en muchas aplicaciones los datos, por su génesis, toman valores en una variedad Riemanniana. Por ejemplo podemos pensar las mediciones de vientos de una estación meteorológica conformadas por su intensidad y dirección o las direcciones de los vientos de 2 estaciones distintas que se encuentran sobre un cilindro y sobre un toro bidimensional respectivamente. Otro ejemplo puede ser el cono determinado por la familia de las matrices definidas positivas (por ejemplo matrices de varianzas y covarianzas) o la variedad Grassmaniana determinada por los subespacios de dimensión k (por ejemplo obtenidos por las bases de un PCA al variar un

conjunto de covariables).

La tesis tiene entonces como segundo objetivo, en este escenario, extender dos conceptos importantes en estadística,

- El concepto de profundidad estadística, ahora cuando los datos se encuentran sobre una variedad Riemanniana (capítulo 3).
- El análisis de sensibilidad sobre un código con entradas estocásticas, pero ahora cuando el output esta en una variedad Riemanniana (capítulo 4).

Parte de los resultados de esta tesis se encuentran en los siguientes artículos:

Publicaciones basadas en el contenido de esta tesis:

- Ricardo Fraiman, Leonardo Moreno, and Sebastian Vallejo. "Some hypothesis tests based on random projection." Computational Statistics 32.3 (2017): 1165-1189.

  (Corresponde básicamente al contenido del capítulo 2 de la tesis.)

- Ricardo Fraiman, Fabrice Gamboa, and Leonardo Moreno. "Connecting pairwise geodesic spheres by depth: DCOPS." Journal of Multivariate Analysis 169 (2019): 81-94.

  (Corresponde básicamente al contenido del capítulo 3 de la tesis.)

A consideración editorial

- Ricardo Fraiman, Fabrice Gamboa, and Leonardo Moreno. "Sensitivity indices for output on a Riemannian manifold." arXiv preprint arXiv:1810.11591 (2018).

  (Corresponde básicamente al contenido del capítulo 4 de la tesis.)

Tabla de contenidos

Capítulo 1

Introducción

La introducción esta estructurada en dos secciones donde presentaremos las principales definiciones y proposiciones elementales para la lectura de los capítulos siguientes. La primer sección introduce algunas nociones básicas sobre geometría Riemanniana, el concepto de densidad sobre estos espacios y una discusión sobre la unicidad de la geodésica minimizante por dos puntos, conceptos de vital importancia para la comprensión de los capítulos 3 y 4. La segunda sección refiere a aquellos resultados relevantes sobre proyecciones al azar que son necesarios para el desarrollo del capítulo 2.

1.1. Probabilidad en variedades Riemannianas

Comenzaremos por desarrollar, tomando como referencia do Carmo (1992), aquellas definiciones básicas sobre variedades Riemannianas, necesarias para introducir el concepto de geodésica y del mapa exponencial (dichos conceptos serán utilizados en los capítulos 3 y 4 de la tesis). Posteriormente, como en Bhattacharya and Bhattacharya (2012) y Patrangenaru and Ellingson (2015), introduciremos el concepto de densidad sobre una variedad Riemanniana. Se discute además aquellas posibles condiciones suficientes, sobre la variedad o sobre la medida de probabilidad, para la unicidad de la geodésica determinada por dos puntos de la muestra. Por último se demuestra que la familia de bolas geodésicas de diámetro $\overline{pq}$ son Glivenko–Cantelli y también una clase determinante.

1.1.1. Algunas definiciones de Geometría Riemanniana

Definición 1.1.1 (Variedad diferenciable) *Diremos que $\mathcal{M}$ es una variedad diferenciable de dimensión m, si existe una familia de funciones inyectivas $x_\alpha : U_\alpha \subset \mathbb{R}^m \to \mathcal{M}$ de conjuntos abiertos U_α de $\mathbb{R}^m$ en $\mathcal{M}$ tal que,*

1. $\bigcup_\alpha x_\alpha(U_\alpha) = \mathcal{M}$,
2. *para cada par de índices α, β con $x_\alpha(U_\alpha) \cap x_\beta(U_\beta) = W \neq \emptyset$, los conjuntos $x_\alpha^{-1}(W)$ y $x_\beta^{-1}(W)$ son abiertos de $\mathbb{R}^m$ y las funciones $x_\beta^{-1} \circ x_\alpha$ son diferenciables.*
3. *El conjunto $\mathcal{A} := \{(U_\alpha, x_\alpha)\}$ es maximal relativo a las condiciones anteriores, es decir, si el par $(U_{\alpha_0}, x_{\alpha_0})$ verifica la condición (2) entonces $(U_{\alpha_0}, x_{\alpha_0}) \in \mathcal{A}$*

Podemos definir ahora funciones diferenciables cuando dominio y codominio de la función son variedades diferenciables.

Definición 1.1.2 (Función diferenciable entre variedades) *Sean $\mathcal{M}$ y $\mathcal{N}$ dos variedades de dimensión m y n respectivamente, diremos que la función $\phi : \mathcal{M} \to \mathcal{N}$ es diferenciable en $p \in \mathcal{M}$ si, dada una parametrización $y : V \subset \mathbb{R}^n \to \mathcal{N}$ en $\phi(p)$, existe otra $x : U \subset \mathbb{R}^m \to \mathcal{M}$ en p tal que $\phi(x(U)) \subset y(V)$ y además,*

$$y^{-1} \circ \phi \circ x : U \subset \mathbb{R}^m \to \mathbb{R}^n$$

es diferenciable en $x^{-1}(p)$.

Diremos que una función es diferenciable en un abierto si es diferenciable en todos los puntos del abierto.
Encontramos diversas medidas sobre las cuales poder integrar sobre una variedad, por ejemplo la medida de Hausdorff. En el trabajo integraremos respecto a la medida de volumen, para esto necesitamos que la variedad diferenciable sea orientable.

Definición 1.1.3 (Variedad orientable) *Sea $\mathcal{M}$ una variedad diferenciable. Decimos que $\mathcal{M}$ es orientable si admite una estructura diferenciable $\{(U_{\alpha,x_\alpha})\}$ tal que para cada par α, β con $x_\alpha(U_\alpha) \cap x_\beta(U_\beta) = W \neq \emptyset$, la diferencial del cambio de coordenadas $x_\beta^{-1} \circ x_\alpha$ tiene determinante positivo.*

Si $\mathcal{M}$ es orientable, una elección de una estructura diferenciable que verifique la definición es llamada orientación. Si además $\mathcal{M}$ es conexa existen exactamente dos orientaciones distintas sobre $\mathcal{M}$. Si ϕ es un difeomorfismo entre dos variedades $\mathcal{M}$ y $\mathcal{N}$ con $\mathcal{M}$ orientable, entonces ϕ induce una orientación en $\mathcal{N}$.

A continuación se exponen algunas definiciones necesarias para la definición del concepto de geodésica y del mapa exponencial.

Dada un curva diferenciable $\alpha : (-\epsilon, \epsilon) \to \mathcal{M}$, sea $\alpha(0) = p \in \mathcal{M}$ y consideremos $\mathcal{D} = \{f : \mathcal{M} \to \mathbb{R}, f \text{ es diferenciable en } p\}$, el vector tangente a la curva α en $t = 0$ es una función $\alpha'(0) : \mathcal{D} \to \mathbb{R}$ dada por

$$\alpha'(0)f = \frac{d(f \circ \alpha)}{dt}\Big|_{t=0}, \quad f \in \mathcal{D}.$$

El conjunto de todos los vectores tangentes a $\mathcal{M}$ en p los indicaremos como $T_p\mathcal{M}$ y lo llamaremos *espacio tangente* de $\mathcal{M}$ en p. Se puede observar que este espacio es un espacio vectorial de dimensión m y no depende de las parametrizaciones elegidas.

Definimos el haz tangente de $\mathcal{M}$, que denotaremos como $T\mathcal{M}$, al conjunto $T\mathcal{M} = \{(p, v) : p \in \mathcal{M}, v \in T_p\mathcal{M}\}$, que tiene estructura diferenciable de dimensión $2m$.

Definición 1.1.4 (Campos vectoriales) *Un campo vectorial X sobre una variedad diferenciable $\mathcal{M}$ es una función $X : M \to T\mathcal{M}$ que asocia a cada punto de M un vector $X(p) \in T_p\mathcal{M}$. Diremos que el campo es diferenciable si X es diferenciable.*

Otra forma de pensar un campo diferencial es como un mapeo $X : \mathcal{D} \to \mathcal{D}$ de la siguiente manera,

$$(Xf)(p) = \sum_i a_i(p) \frac{\partial f}{\partial x_i}(p),$$

donde f por abuso de notación representa la expresión de f en la parametrización x. Se puede verificar que Xf no depende de la elección de la parametrización. Esta manera de concebir los campos permiten que expresiones del tipo $X(Yf)$ tomen sentido.

Definición 1.1.5 (Operador corchete) *Dado dos campos vectoriales diferenciables sobre una variedad $\mathcal{M}$ definimos el operador corchete como el campo diferenciable,*

$$[X, Y] = XY - YX.$$

Se puede demostrar que dicho campo existe y es único.

Definición 1.1.6 (Métrica Riemanniana) *Una métrica Riemanniana sobre una variedad diferenciable $\mathcal{M}$ es una correspondencia la cual asocia a cada punto p de $\mathcal{M}$ un producto interno $g_p(\cdot, \cdot) = \langle \cdot, \cdot \rangle_p$ sobre el espacio tangente $T_p\mathcal{M}$, la cual varía diferencialmente en el siguiente sentido: si $x : U \subset \mathbb{R}^m \to \mathcal{M}$ es un sistema de coordenadas alrededor de p, con $x(x_1, \ldots, x_m) = q \in x(U)$ y $\frac{\partial}{\partial x_i}(q) = dx_q(0, \ldots, 1, \ldots, 0)$ (donde la i-ésima coordenada vale 1), entonces $\left\langle \frac{\partial}{\partial x_i}(q), \frac{\partial}{\partial x_j}(q) \right\rangle_p = g_{ij}(x_1, \ldots, x_m)$ es una función diferenciable sobre U.*

La función $g_{ij} = g_{ji}$ es llamada la representación local de la métrica Riemanniana en el sistema de coordenadas x. Una variedad diferenciable con una métrica Riemanniana g será llamada una variedad Riemanniana que denotaremos $(\mathcal{M}, g)$.

Se puede observar que una métrica Riemanniana nos permite definir una noción de volumen sobre una variedad diferenciable orientable $\mathcal{M}$ dada.

Anotemos ahora por $\mathcal{X}(\mathcal{M})$ al conjunto de todos los campos vectoriales C^∞ sobre $\mathcal{M}$ y por $\mathcal{D}(\mathcal{M})$ el anillo de las funciones con valores reales de clase C^∞ definidas sobre $\mathcal{M}$.

Definición 1.1.7 (Conexión afín) *Una conexión afín ∇ sobre una variedad diferenciable $\mathcal{M}$ es una función,*

$$\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M}).$$

La imagen de (X, Y) por el operador es denotada por $(X, Y) \overset{\nabla}{\to} \nabla_X Y$ y que satisface las siguientes propiedades,

1. $\nabla_{fX+gY} Z = f\nabla_X Z + g\nabla_Y Z.$
2. $\nabla_X (Y + Z) = \nabla_X Y + \nabla_X Z.$
3. $\nabla_X (fY) = f\nabla_X Y + X(f)Y,$

con $X, Y, Z \in \mathcal{X}(M)$ y $f, g \in \mathcal{D}(\mathcal{M})$

Es posible mostrar que dada una conexión afín ∇, existe una correspondencia única, llamada *derivada covariante*, que asocia un campo vectorial V a lo largo de una curva diferenciable $c : I \to \mathcal{M}$ otro campo vectorial $\frac{DV}{dt}$ a lo largo de c, que cumple

1. $\frac{D}{dt}(V + W) = \frac{DV}{dt} + \frac{DW}{dt}$.
2. $\frac{D}{dt}(fV) = \frac{df}{dt}V + f\frac{DV}{dt}$.
3. Si V esta inducido por un campo vectorial $Y \in \mathcal{M}$, es decir, $V(t) = Y(c(t))$, entonces $\frac{DV}{dt} = \nabla_{\frac{dc}{dt}} Y$.

Se puede observar que la conexión sólo depende de $X(p)$. Un campo vectorial a lo largo de una curva $c : I \to \mathcal{M}$ es llamado paralelo cuando $\frac{DV}{dt} = 0$ para todo $t \in I$. Se puede probar entonces que si $V_0 \in T_{c(t_0)}M$, existe un único campo paralelo V a lo largo de c (llamado transporte paralelo de $V(t_0)$ a lo largo de c), tal que $V(t_0) = V_0$

Definición 1.1.8 *Dada una variedad diferenciable $\mathcal{M}$ con una conexión afín ∇ y una métrica Riemanniana $\langle \cdot, \cdot \rangle$, diremos que la conexión es compatible con la métrica si $\langle P, P' \rangle = \alpha$, con α constante para todo P, P' campos paralelos a lo largo de c.*

Las siguientes proporciones son equivalentes,

- La métrica es compatible con la conexión.
- $\frac{d}{dt}\langle V, W \rangle = \langle \frac{DV}{dt}, W \rangle + \langle V, \frac{DV}{dt} \rangle$.
- $X\langle Y, Z \rangle = \langle \nabla_X Y, Z \rangle + X\langle Y, \nabla_X Z \rangle$.

Diremos que la conexión es simétrica si $\nabla_X Y - \nabla_Y X = [X, Y]$. En este contexto podemos establecer un Teorema de suma importancia en geometría Riemanniana.

Teorema 1.1.1 (Levi-Civita) *Dada una variedad Riemanniana $\mathcal{M}$, existe un única conexión afín ∇ (que llamaremos conexión Riemanniana) sobre $\mathcal{M}$ que satisface las condiciones,*

- ∇ *es simétrica*

- ∇ *es compatible con la métrica Riemanniana.*

A partir de las definiciones precedentes podemos llegar a un concepto relevante de la tesis como lo es el de geodésica.

Definición 1.1.9 *Una curva parametrizada $\gamma : I \to \mathcal{M}$ es una geodésica en el punto $t_0 \in I$ si $\frac{D}{dt}\left(\frac{d\gamma}{dt}\right) = 0$ en el punto t_0.*

Si γ es una geodésica para todo $t \in I$ diremos que es una geodésica. Abusando del lenguaje también nos referiremos como geodésica a la imagen $\gamma(I)$.

Sea $L(\alpha)$ la longitud de la curva α. Diremos que un segmento de geodésica $\gamma : [a, b] \to \mathcal{M}$ es minimizante si $L(\gamma) \leq L(\alpha)$ para toda curva α diferenciable a trozos. Dado $p \in \mathcal{M}$, existe un entorno W de p, tal que si dos puntos q_1 y q_2 pertenecen a W, se muestra que las geodésicas minimizan localmente la longitud de arco que determinan q_1 y q_2.

Existe un único campo vectorial G sobre $T\mathcal{M}$ cuyas trayectorias son de la forma $t \to (\gamma(t), \gamma'(t))$, donde γ es una geodésica sobre $\mathcal{M}$. G es llamado *campo geodésico* y su flujo el denominado el *flujo geodésico* sobre $T\mathcal{M}$.

Teorema 1.1.2 (Mapa exponencial) *Dado $p \in \mathcal{M}$, existe un entorno V de p en M, un número $\epsilon > 0$ y una función C^∞, $\gamma : (-2, 2) \times \mathcal{U} \to \mathcal{M}$, con $\mathcal{U} = \{(q, w) \in T\mathcal{M} : q \in M, w \in T_qM, \|w\| < \epsilon\}$, tal que $t \to \gamma(t, q, w)$ con $t \in (-2, 2)$, es la única geodésica de $\mathcal{M}$ que $t = 0$ pasa a través de q con velocidad w ($\gamma'(0, q, w) = w$), para cada $q \in V$ y $w \in T_qM$ con $\|w\| < \epsilon$.*

Por tanto podemos definir una función (mapa exponencial) $\exp p : \mathcal{U} \to M$ tal que $\exp p(q, v) = \gamma(1, q, v)$. Si nos restringimos a un abierto del tangente y denotamos $\exp p_q(v) = \exp p(q, v)$ se cumple que $\exp p_q$ es una función que va desde la bola abierta de centro $0 \in T_q\mathcal{M}$ y radio ϵ, $B_\epsilon(0)$ a $\mathcal{M}$. Es posible elegir ϵ de forma que el mapa $\exp p_q$ sea un difeomorfismo sobre un abierto de $\mathcal{M}$.

Es sencillo observar que globalmente las geodésicas minimizantes no tienen que ser necesariamente únicas. Por ejemplo en una esfera dos puntos opuestos son conectados por infinitas geodésicas minimizantes, (ver Figura 1.1).

Definimos el *cut locus* de p en $T_p\mathcal{M}$ como el conjunto de vectores $v \in T_p\mathcal{M}$ tal que el mapa $\exp p_p(tv)$ es geodesicamente minimizante si $0 \leq t \leq 1$, pero no lo es para $0 \leq t \leq 1 + \epsilon$ con $\epsilon > 0$.

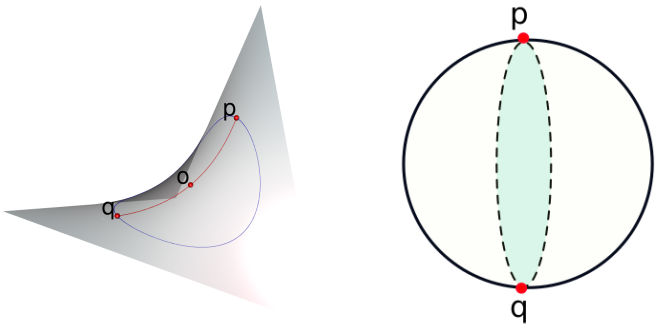

Figura 1.1: *Panel izquierdo*: Representación de una bola geodésica de diámetro $\overline{pq}$ y centro o en un paraboloide hiperbólico. *Panel derecho*: Gráfico de dos geodésicas minimzantes en una esfera que une dos puntos opuestos.

El *cut locus de p en $\mathcal{M}$*, que denotaremos por $C_{\mathcal{M}}(p)$, se define como la imagen mediante el mapa exponencial del cut locus de p en el espacio tangente. Además llamamos *radio de inyectividad de p* al máximo radio de una bola de centro p de forma que el mapa exponencial es un difeomorfismo. El *radio de inyectividad de la variedad*, que denotaremos r_{iny}, es el ínfimo de todos los radios de inyectividad para todos los puntos $p \in \mathcal{M}$.

Si suponemos que la variedad Riemanniana es simplemente conexa y orientada, y asumimos que el espacio $(\mathcal{M}, d_g)$, con d_g la distancia geodésica inducida por la métrica, es completo y separable. Por el Teorema de Hopf-Rinow (ver do Carmo (1992), pág. 146) podemos concluir que para cualquier par de puntos $p, q \in \mathcal{M}$ existe al menos una geodésica que conectan en $\mathcal{M}$ a p y q.

Si asumimos además las hipótesis del Teorema de Cartan-Hadamard (ver Petersen (2006), pág. 162), esto es, $\mathcal{M}$ es simplemente conexa, completa y con curvatura no positiva (para la definición de curvatura ver do Carmo (1992), pág. 88), podemos afirmar la unicidad de la geodésica.

1.1.2. Elementos aleatorios en una variedad Riemanniana

Sea que X un elemento aleatorio de $\mathcal{M}$, con distribución P, es decir, dado un espacio de probabilidad $(\Omega, \mathbb{P}, \mathcal{A})$ y $X : (\Omega, \mathbb{P}, \mathcal{A}) \to \mathcal{M}$, medible respecto a los borelianos B de $\mathcal{M}$ respecto a d_g y definimos $P(B) = \mathbb{P}\left(X^{-1}(B)\right)$. Si $(\mathcal{M}, d_g)$ es completo podemos afirmar la existencia de la geodésica que conecta dos puntos p y q en $\mathcal{M}$, pero no su unicidad (por ejemplo si no se

cumplen los supuestos mencionados en el Teorema de Cartan-Hadamard). Esta última condición se puede obtener mediante restricciones sobre la medida de probabilidad P. En términos generales, asumimos ciertos supuestos sobre la medida de probabilidad de forma que la geodésica determinada por dos puntos sea única casi seguramente. Si M es orientable, como mencionamos existe una medida de volumen $d\nu(y)$ en $\mathcal{M}$. Por tanto, si P es absolutamente continua respecto de ν (hipótesis que haremos), podemos definir una densidad f_Y tal que,

$$P(B) = \int_B f_Y(y) d\nu(y).$$

con ν la distribución del elemento aleatorio Y, (Ver Folland (2013), pág. 361).

Dado $q \in \mathcal{M}$ definimos el siguiente conjunto,

$$A_q := \left\{ y \in \mathcal{M} \Big/ \begin{array}{c} \text{hay mas de una geodésica} \\ \text{minimizante que conectan } y \text{ y } q \end{array} \right\}. \tag{1.1}$$

Asumimos entonces que podemos siempre encontrar un Boreliano $B_q \subset \mathcal{M}$ con $A_q \subset B_q$ tal que para cualquier $q \in \mathcal{M}$

$$\int_{B_q} f_Y(y) d\nu(y) = 0. \tag{1.2}$$

Si $P\left(Y \in C_{\mathcal{M}}(p)\right) = 0$, para todo $p \in \mathcal{M}$, la condición (1.2) se verifica automáticamente (ver Pennec (2006)).

Por ejemplo, en la esfera $S_d := \{u \in \mathbb{R}^d / \|u\| = 1\}$ el cut locus de un punto es su punto opuesto $-p$ y por tanto, si la medida sobre la variedad tiene densidad, se verifica (1.2).

En la tesis asumiremos que la variedad Riemanniana $(\mathcal{M}, g)$ con la distancia inducida d_g es simplemente conexa y orientada, y que el espacio métrico $(\mathcal{M}, d_g)$ es separable y completo. Y además asumiremos que dados dos puntos $p, q \in \mathcal{M}$ existe una única geodésica que los une respecto de la medida tensorial $d\xi(p, q) := f_Y(p) f_Y(q) d\nu(p) d\nu(q)$, donde Y es un elemento aleatorio con distribución ν.

1.1.3. Las bolas geodésicas de diámetro $\overline{pq}$

Dados dos puntos $p, q \in \mathcal{M}$ que determinan una única geodésica minimizante $\overline{pq}$, definimos la bola de diámetro $\overline{pq}$ (denotaremos B_{pq}) como la bola

cerrada de centro en el punto medio del segmento de geodésica que une p con q y radio $d_g(p,q)/2$ (ver Figura 1.1).

Se muestra en Christensen (1970) que la familia de bolas geodésicas de centros $p \in \mathcal{M}$ y radio positivo, $\mathcal{B}_p := \{B_p(r), p \in \mathcal{M} \text{ y } r > 0\}$, es una *clase determinante*. Es decir, si tenemos dos medidas de probabilidad sobre $\mathcal{M}$ que coinciden sobre todos los elementos de esta familia, entonces $\eta = \nu$. Si asumimos que la variedad es compacta, probaremos en la Propiedad 1 que un subconjunto de la familia anterior,

$$\mathcal{B}_{pq} := \{B_{pq}, p, q \in \mathcal{M} \text{ y existe una única geodésica entre } p \text{ y } q\},$$

es también una clase determinante si el radio de inyectividad r_{iny} de $\mathcal{M}$ es positivo. Para esto le pedimos una condición débil a la variedad que llamaremos B-continuidad,

HB) B-continuidad Dada una medida de probabilidad ν definida sobre la variedad $\mathcal{M}$ diremos que satisface la propiedad **HB** si $\nu(\partial A) = 0$ para todo conjunto de Borel cerrado A en $\mathcal{M}$, en donde ∂A es el borde topológico del conjunto A (ver Billingsley and Topsøe (1967)).

Diremos que $\mathcal{M}$ es una variedad diferenciable con borde si todo punto de $\mathcal{M}$ tiene un entorno homeomorfo con $\mathbb{R}^n$ o con $\mathbb{H}^n := \{x \in \mathbb{R}^n : x_n \geq 0\}$. La frontera de $\mathcal{M}$, que denotaremos $\partial\mathcal{M}$, esta determinada por aquellos puntos de $\mathcal{M}$ que tienen un entorno homeomorfo a $\mathbb{H}^n$ y no a $\mathbb{R}^n$.

Propiedad 1 *Sea $(\mathcal{M}, g)$ una variedad Riemanniana separable y compacta, con ν cualquier medida de probabilidad sobre $\mathcal{M}$ que verifica la propiedad **HB**, tal que $\nu(\partial\mathcal{M}) = 0$. Si el radio de inyectividad es positivo entonces $\mathcal{B}_{pq}$ es una clase determinante para las medidas ν.*

Demostración.

Se muestra en Christensen (1970) que los conjuntos

$$\mathcal{B} := \{B_p(r)/p \in \mathcal{M} \text{ y } r > 0\}, \text{ y } \mathcal{B}_\delta := \{B_p(r)/p \in \mathcal{M} \text{ y } 0 < r < \delta\}, \quad (1.3)$$

son clases determinantes.

Nuestro objetivo es probar que la familia de bolas

$$\mathcal{B}_{pq} := \{B_{pq}, p, q \in \mathcal{M} \text{ y existe una única geodésica entre } p \text{ y } q\},$$

es también una clase determinante si el radio de inyectividad r_{iny} de $\mathcal{M}$ es positivo. Es conocido que en un espacio métrico separable la familia de los cerrados $\mathcal{C}$ es una clase determinante (ver Billingsley (1999), pág. 7). Por tanto es suficiente probar que si dos medidas de probabilidad ν y η coinciden en $\mathcal{B}_{pq}$, entonces $\nu(A) = \eta(A)$ para todo $A \in \mathcal{C}$. Sea ∂A el borde topológico de A y $\partial A^{\epsilon} := \bigcup_{x \in \partial A} B(x, \epsilon)$ para $\epsilon > 0$. Ahora descomponemos la medida ν del conjunto A,

$$\nu(A) = \nu\left(A \setminus (\partial A^{\epsilon} \cup \partial \mathcal{M}^{\epsilon})\right) + \nu\left(A \cap (\partial A^{\epsilon} \cup \partial \mathcal{M}^{\epsilon})\right).$$

Tomemos $\epsilon^* := \text{mín}(\epsilon/2, r_{iny})$. Por la propiedad (1.3) sabemos que $\mathcal{B}_{\epsilon^*}$ es una clase determinante, a través del mapa exponencial podemos inferir que la familia de bolas $\mathcal{B}_{\epsilon^*}$, que determinan la ν–probabilidad de $A \setminus (\partial A^{\epsilon^*} \cup \partial \mathcal{M}^{\epsilon^*})$, también pertenecen a la familia $\mathcal{B}_{pq}$. Por lo tanto,

$$\nu\left(A \setminus (\partial A^{\epsilon^*} \cup \partial \mathcal{M}^{\epsilon^*})\right) = \eta\left(A \setminus (\partial A^{\epsilon^*} \cup \partial \mathcal{M}^{\epsilon^*})\right).$$

Finalmente, mediante el Teorema de convergencia dominada, podemos concluir que,

$$\nu(A) = \eta(A) - \eta(\partial A \cup \partial \mathcal{M}) + \nu(\partial A \cup \partial \mathcal{M}) = \eta(A).$$

□

Por otro lado, si $K \in \mathcal{M}$ es un conjunto compacto de $\mathcal{M}$ la familia de bolas $\mathcal{B}_{pq}$ es una clase de Glivenko–Cantelli en K; es decir,

$$\sup_{p,q \in K} |P\left(B_{pq}\right) - P_n\left(B_{pq}\right)| \to 0 \quad \text{c.s.} \quad \text{cuando } n \to +\infty,$$

donde P_n es la medida empírica correspondiente a una sucesión de variables aleatorias iid con distribución P. Esto es consecuencia inmediata de que la familia de bolas de centro p, que es una clase de Glivenko–Cantelli (ver Szabados (1989)), contiene a la familia de bolas de diámetro $\overline{pq}$.

1.2. Proyecciones al azar

La técnica de proyecciones unidimensionales *projection pursuit* (ver Friedman and Tukey (1974) y Friedman and Stuetzle (1981)) es un mecanismo que

permite la reducción de la dimensión del espacio donde se encuentran los datos. Para una referencia general sobre este tópico ver Jones and Sibson (1987). En particular en Blough (1989) podemos encontrar una aplicación de estas técnicas en un test de simetría. Sin embargo este mecanismo de "barrido" de direcciones univariadas son en general muy sensibles a la dimensionalidad del espacio y sus bondades justificadas sólo por resultados empíricos.

Una posible solución a estos problemas es brindada en Cuesta-Albertos et al. (2007). La principal idea en la que se basa este trabajo es que, bajo ciertas hipótesis débiles, es posible una generalización del Teorema de Cramér–Wold a través de proyecciones univariadas al azar. Los mismos autores proponen una aplicación a pruebas de bondad de ajuste (ver Cuesta-Albertos et al. (2006)). Posteriormente, en Cuevas and Fraiman (2009) extienden los resultados a espacios de Banach.

A diferencia de las *proyection pursuit*, el enfoque de proyecciones al azar esta basado en el hecho que para determinar una medida de probabilidad en un espacio de dimensión finita o infinita, bajo algunos débiles supuestos, es sólo necesario conocer la distribución de probabilidad en una única proyección, si esta es elegida aleatoriamente.

Sea $P \in \mathbb{P}\left(\mathbb{R}^d\right)$ una probabilidad sobre $\mathbb{R}^d$. Si denotamos π_h a la función proyección ortogonal de $\mathbb{R}^d$ sobre el subespacio generado por el vector $h \in \mathbb{R}^d$ ($\|h\| = 1$), y sea B un conjunto Boreliano de ese subespacio, entonces la media de probabilidad inducida sobre B (que denotaremos $P_{\langle h \rangle}$) es dada por

$$P_{\langle h \rangle}(B) = P\left[\pi_h^{-1}(B)\right].$$

Definamos ahora un conjunto que cumple un papel relevante en esta sección. Dadas $P, Q \in \mathbb{P}\left(\mathbb{R}^d\right)$, se define el siguiente subconjunto de $\mathbb{R}^d$

$$\mathcal{E}(P, Q) = \{h \in \mathbb{R}^d / P_{\langle h \rangle} = Q_{\langle h \rangle}\}. \tag{1.4}$$

En primera instancia se puede observar que el Teorema de Cramér–Wold (ver Cramér and Wold (1936)) puede ser expresado en términos del conjunto $\mathcal{E}(P, Q)$,

$$\mathcal{E}(P, Q) = \mathbb{R}^d \Leftrightarrow P = Q.$$

En un espacio de Banach E infinito dimensional, el conjunto $\mathcal{E}(P, Q)$ dado

en la ecuación (1.4) puede ser definido considerando, en lugar de un producto interno, funciones del espacio dual E^*. Es decir, si $\mathbf{X}$ e $\mathbf{Y}$ son dos elementos aleatorios de un espacio de Banach separable E con distribuciones de probabilidad P y Q respectivamente, definimos en este caso

$$\mathcal{E}(P,Q) := \mathcal{E}(\mathbf{X},\mathbf{Y}) = \{f \in E^* / \ f(\mathbf{X}) \text{ y } f(\mathbf{Y}) \text{ tienen la misma distribución}\}.$$

En este contexto en Padgett and Taylor (1973) se enuncia una versión funcional del Teorema de Cramér–Wold,

$$\mathcal{E}(\mathbf{X},\mathbf{Y}) = E^* \Leftrightarrow \ \mathbf{X} \text{ y } \mathbf{Y} \text{ tienen la misma distribución} \Leftrightarrow P = Q. \tag{1.5}$$

Definamos ahora ciertas condiciones necesarias para el enunciado de los Teoremas.

Condición de Carleman. La siguiente condición (denominada condición de Carleman) permite caracterizar una distribución a partir de sus momentos y será uno de los supuestos claves asumidos para poder deducir resultados mediante proyecciones al azar.

Sea $P \in \mathbb{P}\left(\mathbb{R}^d\right)$ tal que sus momentos absolutos $m_n = \int \|x\|^n P(dx)$ son finitos. Diremos que P verifica la *condición de Carleman* si $\sum_{n\geq 1} m_n^{-1/n} = \infty$. En Shohat and Tamarkin (1943) se muestra que si P verifica la condición de Carleman entonces ella es determinada por sus momentos.

Llamaremos a un polinomio $p(x)$ homogéneo de grado r si se cumple que $p(\lambda x) = \lambda^r p(x)$ para todo $\lambda \in \mathbb{R}$.

Definición 1.2.1 *Diremos que S es una hipersuperficie proyectiva de $\mathbb{R}^d$ sí y sólo sí existe un polinomio homogéneo (no nulo) $p(x)$ en $\mathbb{R}^d$ tal que*

$$S = \{x \in \mathbb{R}^d / p(x) = 0\}. \tag{1.6}$$

El Teorema 1.2.1 expuesto en Cuesta-Albertos et al. (2007) es uno de los pilares en las pruebas de hipótesis que posteriormente desarrollaremos. Este Teorema permite caracterizar una distribucón mediante proyecciones al azar.

Teorema 1.2.1 (Cuesta, Fraiman y Ransford, 2007) *Sean $P, Q \in \mathbb{P}\left(\mathbb{R}^d\right)$, con $d \geq 2$. Si se cumple que,*

- *P esta determinada por sus momentos.*
- *$\mathcal{E}(P,Q)$ no esta contenido en una hipersuperficie proyectiva de $\mathbb{R}^d$,*

entonces $P = Q$.

En particular, si el conjunto $\mathcal{E}(P,Q)$ tiene H-medida positiva en $\mathbb{R}^d$, en donde H es una medida absolutamente continua respecto a la medida de Lebesgue, entonces podemos afirmar que no esta contenido en ninguna hipersuperficie proyectiva.

En Cuevas and Fraiman (2009) el Teorema 1.2.1 se generaliza para cuando los elementos aleatorios están definidos sobre un espacio de Banach separable E. Sea P una medida de probabilidad sobre E y sea $f \in E^*$. Definimos la distribución univariada P_f como,

$$P_f(A) = P\left(f^{-1}(A)\right),$$

para todo conjunto Boreliano A de $\mathbb{R}$.

Una medida μ de probabilidad es de Radon si para todo boreliano B se cumple que $\mu(B) = \sup\{\mu(K) : K \subset B, K \text{ compacto}\}$ y diremos que μ, definida sobre un espacio de Banach E, es Gaussiana (y no degenerada) si la medida $\mu \circ f^{-1}$ es Gaussiana (y no degenerada) en $\mathbb{R}$ para toda $f \in E^*$.

Teorema 1.2.2 (Cuevas y Fraiman, 2009) *Dada μ una medida Gaussiana de Radon no degenerada en E^*. Sean Q y M dos medidas de probabilidad en E tal que,*

- *Los momentos absolutos $m_n = \int \|x\|^n dM(x)$ son finitos y verifican la condición de Carleman.*
- *El conjunto $\mathcal{E}(M,Q) = \{h \in E^*/Q_h = M_h\}$ tiene μ-medida positiva.*

entonces $Q = M$.

En el capítulo siguiente aplicaremos los conceptos desarrollados en esta subsección para la construcción de dos test de hipótesis no paramétricos.

Capítulo 2

Dos pruebas de hipótesis mediante proyecciones al azar

En este capítulo introducimos dos tipos de pruebas estadísticas de suma relevancia,

- *Las pruebas de simetría.*
- *Las pruebas de independencia.*

Mediante proyecciones al azar dos nuevos test son propuestos. En ambos casos su implementación se realizará tanto en dimensión finita como infinita. Se observará también que en dimensión finita ambos test mantienen su buena performance cuando la dimensión del espacio es elevada.

Es claro que los conceptos de simetría e independencia tienen una significativa importancia en la estadística multivariada. La simetría multivariada es una noción crucial, en particular en estadística no paramétrica. A modo de ejemplo, los problemas multivariados de posición requieren de dicho concepto.

En Brandwein and Strawderman (1991) se plantea una prueba para testear la simetría esférica de una distribución. En Hallin and Paindaveine (2002) se realiza un test sobre un concepto más débil de simetría como lo es la simetría elíptica. En otros trabajos también han sido diseñados diversas pruebas para contrastar otros tipos de simetría multivariada, por ejemplo para testear si la distribución presenta simetría central (ver Sen and Puri (1967), Ley (2010), Aki (1993), Neuhaus and Zhu (1998), Einmahl and Gan (2016) y Heathcote et al. (1995)).

Otro supuesto muy postulado en estadística es la independencia de los datos muestrales, por ejemplo en estudios clínicos donde se supone independencia de la muestra extraída (ver Albert et al. (2001)). Sin embargo dicho supuesto necesita ser corroborado.

En tal sentido muchos procedimientos para testear independencia han sido propuestos. Sin embargo gran parte de ellos toman como supuesto la normalidad (ver por ejemplo Wilks (1935) y Puri and Sen (1971)) o no son aplicables cuando la dimensión del espacio es superior al tamaño de la muestra.

Otros mecanismos, que se basan en extensiones multivariadas de las pruebas de rango, son en general ineficientes para detectar asociaciones no monótonas entre las variables (ver Székely et al. (2007)).

El capítulo esta organizado de la siguiente manera, se comienza por definir independencia y presentar los diferentes formas en que se puede conceptualizar la simetría de una variable en un contexto multivariado. Además se explicitan condiciones necesarias y suficientes de simetría y de independencia mediante proyecciones univariadas y una caracterización de estos conceptos mediante proyecciones al azar. A partir de estas equivalencias se propone un test de simetría y uno de independencia aplicables en espacios de dimensión finita o infinita. Se demuestran las bondades de las pruebas (como por ejemplo la distribución libre del estadístico bajo la hipótesis nula y la consistencia del test) y mediante un estudio de simulación se determinan sus niveles de potencia bajo diferentes alternativas. Por último se muestra una aplicación del test de independencia sobre un conjunto de datos reales.

2.1. Conceptos básicos

En esta sección se definen aquellos conceptos básicos sobre simetría e independencia. La definiciones son dadas sobre un espacio de Banach. Comencemos por definir la independencia de elementos aleatorios.

Definición 2.1.1 (Independencia de elementos aleatorios)
Sean dos elementos aleatorios $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *definidos sobre un espacio de probabilidad* $(\Omega, \mathcal{A}, \mathbb{P})$ *que toman valores sobre un espacio de Banach. Diremos que* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *son variables aleatorias independientes sí y sólo sí los sucesos* $\boldsymbol{X}^{-1}(A)$ *e* $\boldsymbol{Y}^{-1}(B)$ *son independientes para todo par de conjuntos abiertos* $A, B \in E$.

En referencia al concepto de simetría, existen diversas maneras de extender el concepto de simetría univariada al campo multivariado. En Serfling (2006) se encuentra un compendio de ellas y sus principales propiedades. En esta sección haremos una breve introducción de estos conceptos desde el más restrictivo al más general.

Definición 2.1.2 (Simetría Esférica) *Diremos que un vector aleatorio* $\boldsymbol{X} \in \mathbb{R}^d$ *tiene simetría esférica respecto al origen sí y sólo sí* $\boldsymbol{X}$ *y* $A\boldsymbol{X}$ *presentan la misma distribución para cualquier matriz* $A_{d\times d}$ *ortogonal.*

Definición 2.1.3 (Simetría Elíptica) *Diremos que un vector aleatorio* $\boldsymbol{X} \in \mathbb{R}^d$ *tiene simetría elíptica respecto al origen sí y sólo sí* $\boldsymbol{X}$ *y* $A^{'}\boldsymbol{Y}$ *presentan la misma distribución para alguna matriz* $A_{k\times d}$ *tal que* $A^{'}A = \Sigma$ *siendo* $rank(\Sigma) = k \leq d$ *y con* $\boldsymbol{Y} \in \mathbb{R}^k$ *un vector aleatorio con simetría esférica respecto al origen. Denotamos* $A^{'}$ *a la matriz traspuesta de* A*.*

La definición de simetría que utilizaremos en las secciones posteriores es la simetría central, que enunciamos a continuación en un espacio de Banach.

Definición 2.1.4 (Simetría Central) *Sea* E *un espacio de Banach. Diremos que un elemento aleatorio* $\boldsymbol{X}$ *en* E *tiene simetría central respecto al origen sí y sólo sí* $\boldsymbol{X}$ *y* $-\boldsymbol{X}$ *presentan la misma distribución.*

Podemos observar que la simetría central es el más débil de los tres conceptos y además tiene la ventaja de poder ser caracterizado por proyecciones unidimensionales.

2.2. Caracterización a través de proyecciones unidimensionales

Para el desarrollo de nuestras pruebas de independencia y simetría basadas en proyecciones al azar es necesario en primera instancia poder caracterizar dichos conceptos en términos de sus proyecciones unidireccionales. Es decir, podemos caracterizar completamente, a partir de lo desarrollado en la Sección 1.2 de la introducción, los conceptos de simetría (central) e independencia en el espacio original a través de las distribuciones obtenidas de las proyecciones

unidireccionales. Empleando como insumo la linealidad de f, la versión funcional del Teorema de Cramer–Wold (ver ecuación (1.5)) y el Lema 2.3.5 en Padgett and Taylor (1973) se derivan los dos siguientes enunciados.

Lema 1 (Caracterización de la simetría central) *Sea* $\boldsymbol{X}$ *un elemento aleatorio en un espacio de Banach separable* E*. Entonces* $\boldsymbol{X}$ *es un elemento aleatorio centralmente simétrico respecto al origen sí y sólo sí las variables aleatorias* $f(\boldsymbol{X})$ *y* $-f(\boldsymbol{X})$ *tienen la misma distribución para toda* $f \in E^*$*.*

Por tanto, a partir de este lema podemos decir que si todas las proyecciones univariadas de $\mathbf{X}$ son variables aleatorias simétricas centralmente entonces podemos deducir que $\mathbf{X}$ tiene simetría central en el espacio original. Respecto a la independencia se obtienen conclusiones análogas.

Lema 2 (Caracterización de la independencia) *Sean* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *dos elementos aleatorios en un espacio de Banach separable* E*. Entonces* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *son independientes sí, y sólo sí, para toda* $f, g \in E^*$ *se cumple que* $f(\boldsymbol{X})$ *y* $g(\boldsymbol{Y})$ *son variables aleatorias independientes.*

Si el espacio es de Hilbert $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ (incluyendo el caso de dimensión finita), a partir del Teorema de representación de Riesz, podemos caracterizar los elementos del espacio dual a través del producto interno, es decir, para toda $f \in \mathcal{H}^*$ existe un vector $h \in \mathcal{H}$ tal que $f(X) = \langle h, X \rangle$. Por tanto los lemas precedentes, usando proyecciones ortogonales, son válidos en $\mathcal{H}$. La pregunta que surge en este punto es si son necesarias todas las proyecciones o este conjunto puede ser reducido. En este sentido apuntan los resultados de la sección siguiente.

2.2.1. Caracterización mediante proyecciones al azar

A partir de los resultados previamente mencionados podemos abordar el problema de caracterizar la simetría y la independencia mediante proyecciones al azar. Sea $(\Omega, \mathcal{A}, \mathbb{P})$ un espacio de probabilidad y $\mathbf{X}$ un elemento aleatorio que toma valores en un espacio de Banach separable E. Diremos que los momentos absolutos de $\mathbf{X}$ son finitos si los momentos absolutos de la medida de probabilidad inducida por $\mathbf{X}$ (que denotamos $M = \mathbb{P} \circ \mathbf{X}^{-1}$) lo son.

Teorema 2.2.1 *Dada* μ *una medida de Radon Gaussiana no degenerada definida en* E^**. Sea* $\boldsymbol{X}$ *un elemento aleatorio en* E *con las siguientes propiedades*

- *Los momentos absolutos de* $\boldsymbol{X}$ *son finitos y verifican la condición de Carleman.*
- *El conjunto* $\mathcal{E}(\boldsymbol{X}) = \{f \in E^* / \ f(\boldsymbol{X})$ *es una variable aleatoria simétrica*$\}$ *tiene* μ*-medida positiva.*

entonces $\boldsymbol{X}$ *es un elemento aleatorio centralmente simétrico.*

Demostración.

Definamos por M y Q las medidas de probabilidad inducidas por los elementos aleatorios $\mathbf{X}$ y $-\mathbf{X}$ respectivamente. Entonces

$$\begin{aligned}\mathcal{E}(M,Q) &= \{h \in E^* / Q_h = M_h\} \\ &= \{h \in E^* / \ f(\mathbf{X}) \text{ y } -f(\mathbf{X}) \text{ tienen la misma distribución}\} \\ &= \mathcal{E}(\mathbf{X}).\end{aligned}$$

A partir del Teorema 1.2.2, puesto que $\mathcal{E}(M,Q)$ tiene μ-medida positiva, podemos concluir entonces que $Q = M$. Por tanto las distribuciones de $\mathbf{X}$ y $-\mathbf{X}$ coinciden, lo que significa que $\mathbf{X}$ es un elemento aleatorio centralmente simétrico. □

Teorema 2.2.2 *Dado* $(\Omega, \mathcal{B}, \mathbb{P})$ *un espacio de probabilidad,* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *dos elementos aleatorios definidos en un espacio de Banach separable* E *y* μ *una medida de Radon Gaussiana no degenerada definida en* $(E \times E)^*$*. Asumiendo que los momentos absolutos de* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *son finitos y la siguiente serie satisface*

$$\sum_{n \geq 1} \text{mín} \left\{ m_{\boldsymbol{X}}^{-1/n}(n), m_{\boldsymbol{Y}}^{-1/n}(n) \right\} = \infty. \tag{2.1}$$

Definimos como $f(\boldsymbol{X}) = h(\boldsymbol{X}, 0)$ *y* $g(\boldsymbol{Y}) = h(0, \boldsymbol{Y})$ *entonces tenemos que* $h(\boldsymbol{X}, \boldsymbol{Y}) = f(\boldsymbol{X}) + g(\boldsymbol{Y})$*. Si se cumple que el conjunto*

$$\begin{aligned}&\mathcal{E}(\boldsymbol{X}, \boldsymbol{Y}) = \\ &\{h \in (E \times E)^* / \ f(\boldsymbol{X}) \text{ y } g(\boldsymbol{Y}) \text{ son variables aleatorias independientes}\}\end{aligned}$$

tiene μ*-medida positiva, entonces* $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *son independientes.*

Demostración.

Se definen dos medias de probabilidad M y Q en $E \times E$ que quedan determinadas por los valores que toman en el conjunto de los semiespacios $A = \{(x, y) \in E \times E / h(x, y) \leq t, h \in (E \times E)^*\}$, de la siguiente manera

$$\begin{aligned} M(A) &= \int \mathbf{1}_{\{h(x,y)\leq t\}} dP_{\mathbf{X}}(x) dP_{\mathbf{Y}}(y), \quad \forall h \in (E \times E)^*, \quad \forall t \in \mathbb{R}, \\ Q(A) &= \int \mathbf{1}_{\{h(x,y)\leq t\}} dP_{(\mathbf{X},\mathbf{Y})}(x, y), \quad \forall h \in (E \times E)^*, \quad \forall t \in \mathbb{R}. \end{aligned} \tag{2.2}$$

donde $P_{\mathbf{X}}$ y $P_{\mathbf{Y}}$ son las distribuciones de $\mathbf{X}$ e $\mathbf{Y}$ respectivamente, mientras que $P_{(\mathbf{X},\mathbf{Y})}$ es la distribución de $(\mathbf{X}, \mathbf{Y})$. En el espacio producto $E \times E$ se considera la norma del máximo, es decir

$$\|(x, y)\| = \text{máx}\{\|x\|_E, \|y\|_E\}, \tag{2.3}$$

Por lo tanto,

$$\begin{aligned} m_M(n) &= \int \|(x, y)\|^n dP_{\mathbf{X}}(x) dP_{\mathbf{Y}}(y) \\ &\leq \left[\underbrace{\int \|x\|^n dP_{\mathbf{X}}(x)}_{m_{\mathbf{X}}(n)} + \underbrace{\int \|y\|^n dP_{\mathbf{Y}}(y)}_{m_{\mathbf{Y}}(n)} \right] \\ &\leq 2 \,\text{máx}\{m_{\mathbf{X}}(n), m_{\mathbf{Y}}(n)\}. \end{aligned}$$

Además,

$$\begin{aligned} m_M^{-1/n}(n) &\geq 2^{-1/n} \left[\text{máx}\{m_{\mathbf{X}}(n), m_{\mathbf{Y}}(n)\}\right]^{-1/n} \\ &\geq 2^{-1/n} \left[\text{mín}\{m_{\mathbf{X}}^{-1/n}(n), m_{\mathbf{Y}}^{-1/n}(n)\}\right]. \end{aligned}$$

Podemos deducir entonces que la medida M tiene sus momentos finitos y también verifica la condición de Carleman. Por otro lado, sea el conjunto $\mathcal{E}(M, Q) = \{h \in (E \times E)^* / Q_h = M_h\}$ y sea $h \in \mathcal{E}(\mathbf{X}, \mathbf{Y})$.

Entonces,

$$
\begin{aligned}
Q_h\left((-\infty,t]\right) &= P\left(f(\mathbf{X})+g(\mathbf{Y})\leq t\right)\\
&= \int_{E\times E}\mathbf{1}_{\{f(x)+g(y)\leq t\}}dP_{(\mathbf{X},\mathbf{Y})}(x,y)\\
&= \int_{\mathbb{R}\times\mathbb{R}}\mathbf{1}_{\{u+v\leq t\}}dP_{(f(\mathbf{X}),g(\mathbf{Y}))}(u,v)\\
&= \int_{\mathbb{R}\times\mathbb{R}}\mathbf{1}_{\{u+v\leq t\}}dP_{f(\mathbf{X})}(u)dP_{g(\mathbf{Y})}(v)\\
&= M_h\left((-\infty,t]\right).
\end{aligned}
$$

Por lo tanto $h \in \mathcal{E}(M,Q)$, lo que implica que $\mathcal{E}(M,Q)$ tiene μ-medida positiva. A través del Teorema 1.2.2 se deduce que $M = Q$, y se concluye entonces que $\mathbf{X}$ e $\mathbf{Y}$ son elementos aleatorios independientes. □

Ejemplo 1 *Consideremos por ejemplo una medida de probabilidad sobre* $\mathbb{R}^3$*, que se distribuye de manera uniforme sobre la bola de centro* $(0,0,1)$ *y radio* 1*. Es claro que dicha distribución no es simétrica respecto al origen, pero las distribuciones obtenidas de las proyecciones sobre las infinitas direcciones contenidas en el plano xy son simétricas. No obstante este ejemplo no conlleva ninguna contradicción puesto que el plano xy tiene medida de Lebesgue nula en* $\mathbb{R}^3$*.*

Ejemplo 2 *Consideremos ahora un caso de perfecta dependencia en* $\mathbb{R}^2$*, donde ambos vectores coinciden*

$$
\boldsymbol{X} = \boldsymbol{Y} = (X_1, X_2) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right]. \tag{2.4}
$$

Si proyectamos $\boldsymbol{X}$ *e* $\boldsymbol{Y}$ *en las direcciones determinadas por los* $u = (u_1, u_2)$ *y* $v = (u_3, u_4)$ *respectivamente, las variables proyectadas son independientes sí y sólo sí* u *y* v *son ortogonales. Es decir*

$$
\begin{aligned}
\mathcal{E}(\boldsymbol{X},\boldsymbol{Y}) &= \{(u,v)\in\mathbb{R}^4/\ \langle \boldsymbol{X},u\rangle \text{ y } \langle \boldsymbol{Y},v\rangle \text{ son indep.}\}\\
&= \{(u_1,u_2,u_3,u_4)\in\mathbb{R}^4/u_1u_3+u_2u_4=0\}.
\end{aligned}
$$

Por tanto existe un conjunto infinito de pares de direcciones donde las proyecciones son independientes, pero este conjunto esta contenido en una hipersuperficie proyectiva de grado 2 *en* $\mathbb{R}^4$*, que también tiene medida de Lebesgue* 0*.*

2.3. Un test de simetría central

Sea ahora $\{\mathbf{X}; \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ una muestra de elementos aleatorios independientes en un espacio de Banach separable E donde la distribución es determinada por sus momentos. Desarrollaremos un test de simetría central en E. Consideramos la siguiente prueba de hipótesis

$$\begin{aligned} &H_0) \ \mathbf{X} \text{ y } -\mathbf{X} \text{ tienen la misma distribución,} \\ &H_1) \ \mathbf{X} \text{ y } -\mathbf{X} \text{ no tienen la misma distribución,} \end{aligned} \tag{2.5}$$

El resultado de la prueba queda determinado a partir de los siguientes pasos

- Se elige $h \in E^*$ al azar a partir de una μ-medida Gaussiana como la definida en el Teorema 2.2.1, . En el caso finito dimensional es suficiente elegir una dirección h al azar a partir de una medida de probabilidad H (absolutamente continua respecto a la medida de Lebesgue) on $\mathbb{R}^d$ (en ambos casos se normaliza h, es decir se considera $h/\|h\|$).

- Fijado h, se construye una muestra iid de variables aleatorias (dependientes de h),

$$\{h(\mathbf{X}_1), h(\mathbf{X}_2), \ldots, h(\mathbf{X}_n)\}. \tag{2.6}$$

  En el caso finito dimensional, la muestra iid $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ se proyecta sobre un espacio unidireccional generado por el vector h elegida al azar, escribimos $h(\mathbf{X}) = \langle \mathbf{X}, h \rangle$. Se obtiene así una muestra de variables aleatorias en $\mathbb{R}$.

- Dado un nivel de significación α, se realiza un test unidimensional del tipo de Kolmogorov–Smirnov desarrollado en Sen and Chatterjee (1973) a partir de la muestra de datos proyectados obtenida en (2.6). Si denotamos por F^h a la distribución acumulada de la variable aleatoria $h(\mathbf{X}_1)$ y sea $F^h(x^-) := \lim_{t \to x^-} F^h(t)$. El test unidireccional de Kolmogorov–Smirnov para la muestra proyectada es,

$$\begin{aligned} &H_0) \ \ F^h(x) + F^h(-x) - 1 = 0 \ \forall x \in \mathbb{R}, \\ &H_1) \ \ |F^h(x) + F^h(-x) - 1| > 0 \text{ para algún } x \in \mathbb{R}. \end{aligned} \tag{2.7}$$

El estadístico del test $D^h(n)$ esta dado por,

$$D^h(n) = \sup_{x \geq 0} |F_n^h(x) + F_n^h(-x^-) - 1| = \text{máx}\left[D_-^h(n), D_+^h(n)\right],$$

con

$$D_-^h(n) = \sup_{x \geq 0} \left(1 - F_n^h(x) - F_n^h(-x^-)\right) \text{ y}$$
$$D_+^h(n) = \sup_{x \geq 0} \left(F_n^h(x) + F_n^h(-x^-) - 1\right).$$

Aquí F_n^h es la distribución empírica de la muestra univariada

$$\{h(\mathbf{X}_1), h(\mathbf{X}_2), \ldots, h(\mathbf{X}_n)\}.$$

Es decir $F_n^h := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{h(X_i) \leq x\}}, \quad (x \in \mathbb{R})$.

Para valores "grandes" del estadístico $D^h(n)$ es rechazada la hipótesis nula H_0 en (2.5). Este test lo denotaremos por RPK_1.

2.3.1. Distribución exacta y asintótica de $D^h(n)$ bajo H_0

Sea $\mathcal{F}_0$ el conjunto de todas las distribuciones simétricas en E. Determinaremos ahora, utilizando los resultados expuestos en Sen and Chatterjee (1973), la distribución exacta del estadístico bajo H_0, verificando además que es de distribución libre, es decir, la distribución del estadístico (bajo H_0) no depende de la medida H utilizada para la elección de la dirección h ni tampoco de la distribución de los datos $F \in \mathcal{F}_0$.

Teorema 2.3.1 (Distribución del estadístico bajo H_0) *Para toda $F \in \mathcal{F}_0$ y para cualquier "dirección" $h \in E^*$, se cumple que*

$$P\left(nD_-^h(n) \geq k\right) = P\left(nD_+^h(n) \geq k\right) = 2\sum_{i=0}^{s} \binom{n}{i} 2^{-n} - \delta_k \binom{n}{s} 2^{-n}, \quad (2.8)$$

donde $s = [n/2k] - 1$, $k = 1, 2, \ldots, n$ y δ_k es 0 si $n - k$ es par, mientras que vale 1 si $n - k$ es impar.

Por tanto,

$$P\left(nD^h(n) \geq k\right) = \begin{cases} 1 & si\ k = 0, 1 \\ 2\sum_{j=0}^{u}(-1)^j P\left[nD_-^h(n^-) \geq (2j+1)k\right] & si\ k > 1, \end{cases}$$

donde $u = [n/2k] - 1$.

Demostración.

Bajo H_0, la variable $\mathbf{X}_1$ es simétrica para toda $h \in E^*$. Por el Teorema 1 podemos deducir que $h(\mathbf{X}_1) \in \mathbb{R}$ es también simétrica y por tanto verifica la condición dada en (2.7) para la hipótesis nula en el caso unidireccional.

Sea $h(\mathbf{Y}_1) \geq h(\mathbf{Y}_2) \geq \ldots \geq h(\mathbf{Y}_n)$ los estadísticos de orden (ordenados de mayor a menor) de los valores absolutos $|h(\mathbf{X}_1)|, |h(\mathbf{X}_2)|, \ldots, |h(\mathbf{X}_n)|$. Sea $t_{n,i}^h = F_n(-h(\mathbf{Y}_i))$, entonces $0 \leq t_{n,1}^h \leq t_{n,2}^h \leq \ldots \leq t_{n,n}^h \leq F^h(0) = 1/2$. Puesto que F es continua, entonces no ocurren empates c.s. y por lo tanto

$$0 < t_{n,1}^h < t_{n,2}^h < \ldots < t_{n,n}^h < F^h(0) < 1/2 \quad \text{c.s.}$$

A partir de la transformación canónica podemos definir $V_n^h(t) = n^{1/2}[G_n^h(t) - t]$ con $0 < t < 1$, donde $G_n^h(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,t)}\left(F^h(X_i)\right)$, y además llamar $\tilde{V}_n^h(t)$ a

$$\tilde{V}_n^h(t) = V_n^h(t^-) + V_n^h(1-t), \quad 0 \leq t \leq 1/2.$$

De esta manera $\tilde{V}_n^h(t)$ es un proceso estocástico definido en $(0, 1/2)$ con n saltos de 1 o -1 en los puntos $t_{n,1}^h, t_{n,2}^h, \ldots, t_{n,n}^h$. Se ahora $p_{i,j} = P\left(h(\mathbf{Y}_{n-i+1}) = |h(\mathbf{X}_j|\right)$. Entonces se cumple que

$$\begin{aligned} &P\left(h(\mathbf{Y}_{n-i+1}) \quad \text{coincida con valor positivo de} \quad h(\mathbf{X}_j)\right) = \\ &= \sum_{j=1}^{n} p_{ij} P\left(h(\mathbf{X}_j) > 0 \Big/ Y_{n-i+1}^h = |h(\mathbf{X}_j)|\right) = 1/2 \sum_{j=1}^{n} p_{i,j} = 1/2. \end{aligned} \tag{2.9}$$

Denotamos por $sg(.)$ y $|\cdot|$ a las funciones signo y valor absoluto respectivamente. Es sabido que los vectores

$$\left(sg\left(h(\mathbf{X}_1)\right), \ldots, sg\left(h(\mathbf{X}_n)\right)\right) \text{ y } \left(|h(\mathbf{X}_1)|, \ldots, |h(\mathbf{X}_n)|\right)$$

son independientes bajo la hipótesis nula.

Por tanto, los saltos de $n^{1/2}\tilde{V}_n^h(t)$ y $t_{n,1}^h, t_{n,2}^h, \ldots, t_{n,n}^h$ también son independientes. Para finalizar, podemos deducir entonces que, bajo H_0, las distribuciones de los estadísticos $D_-^h(n)$ y $D_+^h(n)$ tiene la misma distribución que el máximo de una caminata al azar simétrica de n pasos desde el origen, y por tanto $nD^h(n)$ tiene la misma distribución que el máximo del valor absoluto de una caminata al azar (ver Takács (1967)) y se concluye la tesis del Teorema.

□

También es posible obtener la distribución asintótica del estadístico bajo la hipótesis nula H_0 (ver der Vaart (1998), sección 19.3).

2.3.2. Consistencia del test

En esta sección se deduce la consistencia universal del test bajo cualquier alternativa no simétrica, es decir, el test es capaz de detectar cualquier alternativa no simétrica para una muestra suficientemente grande. Enunciamos la consistencia en $\mathbb{R}^d$, pero es análogo en un espacio de Banach E.

Teorema 2.3.2 (Consistencia del test) *Sea $\{\boldsymbol{X}_n\}_{n\geq 1}$ una sucesión de vectores aleatorios iid con distribución determinada por sus momentos y tal que $\boldsymbol{X}_1$ no es centralmente simétrica. Entonces, dada una medida H absolutamente continua respecto a la medida de Lebesgue en $\mathbb{R}^d$, se cumple que*

$$H\left\{h \in \mathbb{R}^d : P\left(\liminf_{n\to+\infty} D^h(n) > 0\right) = 1\right\} = 1.$$

Demostración

Anotemos por P y Q a las distribuciones de las variables $\mathbf{X}$ y $-\mathbf{X}$ respectivamente. Entonces el conjunto $\mathcal{E}(P,Q)$ tiene H-medida nula en $\mathbb{R}^d$, puesto que si el conjunto tuviera H-medida positiva, por el Teorema 2.2.1, $\mathbf{X}$ y $-\mathbf{X}$ tendrían la misma distribución, lo que contradice las hipótesis del Teorema.

Para cada $h \in \mathcal{E}^c(P,Q)$, por el Lema 1, la variable $h(\mathbf{X})$ no es simétrica, entonces si definimos $\delta(F^h)$ como

$$\delta(F^h) = \sup_{x\geq 0} |F^h(x) + F^h(-x) - 1|,$$

tenemos que

$$\delta(F^h) = \begin{cases} 0 & \text{si } F^h \in \mathcal{F}_0^h \\ \\ > 0 & \text{si } F^h \in \mathcal{F}_1^h. \end{cases} \tag{2.10}$$

Por tanto, bajo H_1, podemos deducir que existe $t_h \in \mathbb{R}$ tal que

$$P\left(x \in \mathbb{R}^d / \langle x, h\rangle \leq t_h\right) \neq Q\left(x \in \mathbb{R}^d / \langle x, h\rangle \leq t_h\right). \tag{2.11}$$

Es decir,

$$F^h(t_h) + F^h(-t_h) - 1 \neq 0. \tag{2.12}$$

Por el Teorema de Glivenko–Cantelli (ver der Vaart (1998), sección 19.1), $\sup_x |F_n^h(x) - F^h(x)| \xrightarrow{c.s} 0$, lo que implica que

$$\begin{aligned} D^h(n) &\geq |F_n^h(t_h) + F_n^h(-t_h) - 1| \\ &\geq |F^h(t_h) + F^h(-t_h) - 1| - |F_n^h(t_h) - F^h(t_h)| - |F_n^h(-t_h) - F^h(-t_h)| \\ &\geq \frac{\delta(F^h)}{2}, \end{aligned}$$

casi seguramente cuando $n \to +\infty$. □

2.3.3. Potencia del test

Ya hemos demostrado dos propiedades deseables en todo test estadístico, como los son su distribución libre y su consistencia bajo cualquier alternativa. Pero las muestras son de tamaño finito y en estos casos las pruebas podrían presentar baja potencia. Este problema ya es analizado en Cuesta-Albertos et al. (2006). Encontramos dos motivos que van en detrimento de la potencia del test. En primera instancia, bajo H_1, el conjunto de direcciones donde la proyecciones tienen distribución simétrica es de H-medida nula. Sin embargo al considerar las estimaciones "plug-in", es decir, al reemplazar la distribución teórica por su versión empírica podemos encontrar un entorno de "direcciones vecinas" a una dirección dada, con H-medida positiva, donde H_0 no es rechazada cuando debería serlo.

Por otro lado, las pruebas no paramétricas univariadas del tipo Kolmogorov–Smirnov, si bien son universalmente consistentes, no son óptimas en términos de la potencia para cualquier alternativa. Una solución para

este segundo problema es modificar el estadístico de manera de hacerlo óptimo en referencia a determinadas alternativas (ver por ejemplo Mason and Schuenemeyer (1983)). En referencia al primer problema, propondremos dos posibles soluciones (la performance de estas son evaluadas mediante un pequeño estudio de simulación).

Una primer alternativa, presentada en Cuesta-Albertos et al. (2006), es considerar (en lugar de sólo una) un conjunto finito de proyecciones al azar, calcular el estadístico para lo datos proyectados sobre cada una de ellas, y computar como estadístico del test el máximo de todos ellos. Esta prueba basada en j proyecciones lo denotaremos RPK_j.

Una segunda alternativa es elegir de manera "adecuada" la distribución absolutamente continua H. Entendemos por conveniente a aquellas distribuciones que asignan más masa de probabilidad a aquellas direcciones donde la variable en estudio podría ser más asimétrica. Para ello, a partir de un subconjunto de la muestra, calculamos como índice de simetría en cada dirección el valor absoluto de la mediana y de manera proporcional a este índice podemos estimar la densidad de H. Considerando esta distribución para sortear la dirección, el test sigue siendo libre bajo H_0 (se mantiene el nivel de significación), y logramos (como se verá en las simulaciones) aumentar la potencia de la prueba. A este test lo denotaremos por RPKW.

En la sección referente a simulaciones mostraremos como ambos enfoques mejoran la potencia del test. Es claro que en la primer alternativa el test pierde la propiedad de distribución libre. Por tanto para controlar el nivel de significación del test original se realiza la corrección de Bonferroni en RPKj.

2.4. Un test de Independencia

Sea $\{(\mathbf{X},\mathbf{Y});(\mathbf{X}_1,\mathbf{Y}_1),(\mathbf{X}_2,\mathbf{Y}_2),\ldots,(\mathbf{X}_n,\mathbf{Y}_n)\}$ un conjunto de elementos aleatorios iid en $E\times E$, en donde E es un espacio de Banach separable y además suponemos que la distribución de $(\mathbf{X},\mathbf{Y})$ es determinada por sus momentos. Propondremos un test para contrastar

$$\begin{aligned} H_0)&\ \ \mathbf{X} \text{ y } \mathbf{Y} \text{ son independientes,} \\ H_1)&\ \ \mathbf{X} \text{ y } \mathbf{Y} \text{ no son independientes.} \end{aligned} \tag{2.13}$$

El objetivo, al igual que en el test de simetría, es poder extrapolar métodos

que ya han sido desarrollados en dimensón finita a espacios funcionales. En particular utilizaremos a nivel finito dimensional un test de cópulas, cuya teoría es intrínsecamente concebida en dimensión finita. Se explicitan los resultados en el espacio $E \times E$, sin embargo estos pueden ser extendidos con facilidad al caso de un producto de más de dos espacios.

Sean $(\mathbf{X}, \mathbf{Y}); (\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ una muestra de variables aleatorias iid de elementos aleatorios en $E \times E$, queremos determinar si $\mathbf{X}$ e $\mathbf{Y}$ son independientes. Bajo algunas condiciones sobre los momentos de la medida de probabilidad y a partir de una medida Gaussiana no degenerada μ en el espacio dual de $(E \times E)$, $(E \times E)^*$, vamos a mostrar que si el conjunto

$$\mathcal{E}(\mathbf{X}, \mathbf{Y}) =$$
$$\{h \in (E \times E)^* / f(X) = h(\mathbf{X}, 0) \text{ y } g(X) = h(0, \mathbf{Y}) \text{ son elementos aleatorios independientes}\},$$

tiene μ-medida positiva, con μ una medida Gauussiana en $(E \times E)^*$, entonces $\mathbf{X}$ e $\mathbf{Y}$ son independientes.

La metodología propuesta es la siguiente,

- Elegimos al azar $h \in (E \times E)^*$ a partir de la medida μ. Definimos $f(x) = h(x, 0)$ y $g(y) = h(0, y)$. Observar que $f, g \in E^*$.

- Fijadas f y g, se realiza un test de independencia para las variables aleatorias $f(\mathbf{X})$ y $g(\mathbf{Y})$ basados en la cópula empírica (ver Genest et al. (2007)). El proceso empírico asociado a la cópula de independencia es de la forma

$$\sqrt{n}[C_n^{f(\mathbf{X}),g(\mathbf{Y})}(u, v) - u.v] \quad \text{tal que} \quad (u, v) \in [0, 1]^2.$$

La distribución asintótica del test es calculada bajo la hipótesis nula. A partir de los resultados desarrollados en Fermanian et al. (2004) y Fermanian (2005) es mostrada su consistencia universal bajo cualquier alternativa.

2.4.1. Distribución asintótica bajo H_0 y consistencia del test

Sea $f(x) = h(x, 0)$ y $g(y) = h(0, y)$. Dada la dirección h sorteada la muestra iid obtenida es

$$\{(f(\mathbf{X}_1), g(\mathbf{Y}_1)), \ldots, (f(\mathbf{X}_n), g(\mathbf{Y}_n))\}.$$

Usamos en este caso un estadístico del tipo Cramér–von Mises,

$$W_n = \int_{[0,1]^2} \mathbb{C}_n^2(u, v) du dv, \tag{2.14}$$

con

$$\mathbb{C}_n(u, v) = \sqrt{n}[C_n^{f(X),g(Y)}(u, v) - u.v] \quad \text{tal que} \quad (u, v) \in [0, 1]^2,$$

en donde $C_n^{f(X),g(Y)}(u, v)$ es la cópula empírica

$$C_n(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{R_{i1} \leq nu_1\}} \mathbf{1}_{\{R_{i2} \leq nu_2\}},$$

con $R_{i1} = \sum_{k=1}^{n} \mathbf{1}_{\{f(X_k) \leq f(X_i)\}}$ y $R_{i2} = \sum_{k=1}^{n} \mathbf{1}_{\{g(X_k) \leq g(X_i)\}}$ para $1 \leq i \leq n$.

En Fermanian et al. (2004) se presentan algunos resultados para pruebas de bondad de ajuste mediante cópulas, donde el test de independencia es un caso particular. En Genest and Rémillard (2008) se muestra, bajo ciertas hipótesis de regularidad, que el test es universalmente consistente.

2.5. Simulaciones

2.5.1. Simulaciones para el test de simetría

En esta subsección comparamos el test de simetría con otros posibles test de la literatura. La performance de dichos test será evaluada en $\mathbb{R}^d$ con $d = 2, 3$ y 50 para muestras de tamaño $n = 100$. Consideramos las siguientes pruebas,

1. (Test RPK_{10}.) Se eligen uniformemente 10 direcciones al azar h en la esfera unidad de $\mathbb{R}^d$. Se determina la muestra proyectada $h(X_i) :=< \mathbf{X}_i, h >$ en cada una de las 10 direcciones. Se determina el p-valor de cada una de las 10 pruebas univariadas. Cada test univariado se rechaza

si el p-valor es inferior a 0.005. Si en algunas de las direcciones el test univariado es rechazado, entonces también se rechaza H_0 de la prueba original. Para determinar la potencia se reitera el procedimiento 10000 veces y se determina la proporción de rechazos.

2. (Test RPK_1.) Se realiza el mismo test que el anterior pero tomando sólo una dirección al azar y $\alpha = 0.05$.

3. (Test RPKW.) Como primer paso se eligen 100 direcciones al azar en la esfera unitaria. Se proyectan el 20 % los datos en cada una de estas direcciones y se calcula la mediana en valor absoluto como una medida de simetría. A partir de esta función, que a cada dirección le asocia un real positivo, mediante splines cúbicos se estima la función de regresión. A partir de la normalización de esta función se determina una densidad que es utilizada para modelar la distribución H. Se elige una direción H para realizar el test con un nivel de significación $\alpha = 0.05$. A modo de ejemplo, a partir de una muestra de tamaño 30 de una normal bivariada $N(\boldsymbol{\mu} = (1/2, 1/2), \Sigma = \mathbf{I}_2)$, en la Figura 2.1 se determina, bajo el procedimiento descrito, la estimación de la densidad de H. Se puede apreciar que dicha densidad acumula más masa en la direcciones en donde la muestra es mas asimétrica respecto al origen (en este ejemplo en la dirección de 45 grados).

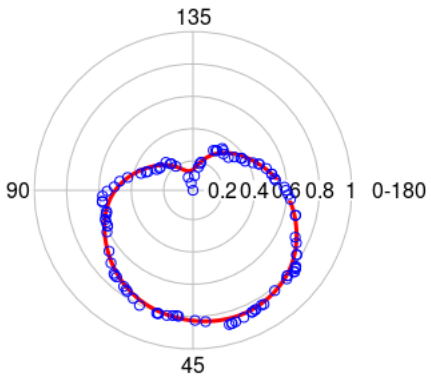

Figura 2.1: Estimación de la regresión circular del valor absoluto de la mediana a partir de 100 direcciones al azar.

4. (Test de Marden) Esquematizamos ahora el test desarrollado en Marden (1999). Dada la muestra iid $\{X_1, \ldots, X_n\}$, se comienza por reor-

denar la muestra según la distancia euclídea al origen, anotemos $\{X_{A_1}, X_{A_2}, \ldots, X_{A_n}\}$ a la muestra reordenada. El estadístico del test lo denotamos $M^{(n)}$, donde

$$M^{(n)} = \sqrt{\frac{2}{n}} \sum_{i=2}^{n} U_{A_i}' U_{A_{i-1}}, \tag{2.15}$$

con $U_{A_i} = X_{A_i}/\|X_{A_i}\|$. Se rechaza la hipótesis nula a valores grandes del estadístico. La distribución asintótica del test es normal estándar bajo la hipótesis nula de simetría esférica.

5. (Ley's Test.) Test propuesto en Ley (2010). Esta basado en una medida de profundidad D en $\mathbb{R}^d$ (ver Serfling and Zuo (2000)). La muestra $\{X_{A_1}, X_{A_2}, \ldots, X_{A_n}\}$ es ordenada en función de la profundidad D de las observaciones en la muestra simetrizada $\{\pm X_1, \pm X_2, \pm \ldots, \pm X_n\}$. El estadístico que proponen es

$$R_D^{(n)} = 1 + \sum_{i=k+1}^{n} \mathbf{1}_{\{\mathbf{0} \in S(U_{A_i}, U_{A_{i-1}}, \ldots, U_{A_{i-k}})\}}, \tag{2.16}$$

donde $S(x_1, x_2, \ldots, x_n)$ es le simplex convexo definido por $(x_1, x_2, \ldots, x_n)$.

A partir de la hipótesis nula de simetría central respecto al origen, con algún otro supuesto débil sobre la distribución (ver Dyckerhoff et al. (2015)), se cumple que $n^{-1/2}(4R_D^{(n)} - n - 2)$ tiene distribución asintótica normal centrada con varianza $\sigma^2 = 11/3$.

Para el caso finito dimensional realizaremos simulaciones en dos posibles escenarios considerando como dimensiones del espacio $d = 2, 3$ y 50. Para el caso infinito dimensional es planteado un tercer escenario. En los escenarios 1 y 2 testeamos simetría central respecto al origen y en el escenario 3 simetría central respecto a la función nula.

- (Escenario 1.) Se toma una muestra iid $\{X_1, \ldots, X_n\}$ a partir de una distribución de Student multivariada con 4 grados de libertad, $t_4\Big(\boldsymbol{\mu} = \mu \mathbf{1}_d, \Sigma = \mathbf{I}_d\Big)$, en donde μ es un número real en el intervalo $[0; 0.5]$ y d es la dimensión espacial, (ver Kotz and Nadarajah (2004)).

- (Escenario 2.) En este caso se considera una muestra iid $\{X_1, \ldots, X_n\}$ de una distribución normal "skew" multivariada (ver Azzalini and Valle (1996)) $SN_d(\Sigma, \boldsymbol{\beta})$ con densidad

$$f(x) = 2\phi(x; \Sigma)\Phi(\boldsymbol{\beta}^t x) \quad x \in \mathbb{R}^d,$$

donde $\phi(x; \Sigma)$ es la densidad de una normal multivariada de dimensión d con media cero y matriz de varianzas y covarianzas Σ, $\Phi(\cdot)$ es la distribución acumulada de una $N(0, 1)$ y $\beta \in \mathbb{R}^d$ es el parámetro de asimetría (ver Figura 2.2). Tomamos $\Sigma = \mathbf{I}_d$ y $\boldsymbol{\beta} = \mu(1, 2, 1, \ldots) \in \mathbb{R}^d$, con μ variando entre 0 y 0.5.

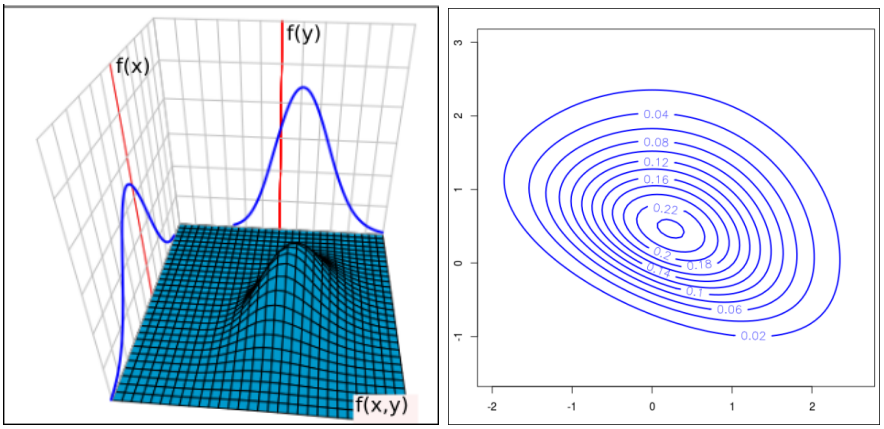

Figura 2.2: Representación de la distribución normal "skew" bivariante con $\beta = (1, 2)$ y $\Sigma = \mathbf{I}_2$. Panel Izquierdo: gráfico de la densidad conjunta $f(x, y)$ y de las densidades marginales $f(x)$ y $f(y)$. Panel Derecho: Curvas de nivel de $f(x, y)$.

- (Escenario 3.) Para el caso en dimensión infinita sorteamos una muestra de tamaño $n = 100$ a partir de

$$X(t) = W(t) + \mu t, \tag{2.17}$$

donde $W(t)$ es un movimiento Browninano en $[0, 1]$ y $\mu \in [0, 1/2]$.

Las funciones de potencia para las diferentes pruebas son dadas en la Tabla 2.1 y en la Tabla 2.2. Se puede observar en todos los casos la mejor performance del test por proyecciones al azar sobre sus competidores.

Los métodos basados en proyecciones al azar RPK_{10} y RPKW son los que presentan potencia más elevadas frente a las distintas hipótesis alternativas. Si

bien ambos test mencionados presentan potencias similares el último (RPKW) presenta la ventaja de retener la propiedad de distribución libre.

No son realizables en dimensión 50 las pruebas de Marden y Ley por los elevados tiempos computacionales que son requeridos. En la Tabla 2.3 se muestran las funciones de potencia de RPK_1 y RPK_{10} en dimensión 50 y en el espacio $C(0,1)$.

Tabla 2.1: Potencia de los test de simetría en el Escenario 1 en dimensión 2 y 3.

μ	Student en dim 2					Student en dim 3				
	RPK_1	Marden	Ley	RPKW	RPK_{10}	RPK_1	Marden	Ley	RPKW	RPK_{10}
0	0.05	0.05	0.06	0.05	0.03	0.04	0.05	0.05	0.05	0.04
0.05	0.07	0.08	0.08	0.07	0.04	0.09	0.05	0.09	0.06	0.06
0.10	0.10	0.09	0.10	0.12	0.09	0.14	0.06	0.11	0.16	0.15
0.15	0.11	0.10	0.11	0.31	0.22	0.21	0.10	0.12	0.31	0.34
0.20	0.14	0.14	0.13	0.54	0.41	0.35	0.14	0.17	0.63	0.58
0.25	0.21	0.28	0.22	0.83	0.59	0.44	0.21	0.18	0.67	0.78
0.30	0.30	0.37	0.31	0.92	0.80	0.51	0.31	0.25	0.70	0.91
0.35	0.38	0.41	0.38	0.94	0.90	0.58	0.48	0.32	0.75	0.97
0.40	0.52	0.54	0.52	0.95	0.96	0.62	0.60	0.45	0.82	0.99
0.45	0.62	0.66	0.68	0.98	0.99	0.69	0.74	0.66	0.90	1.00
0.50	0.76	0.77	0.80	0.99	1.00	0.74	0.79	0.70	0.97	1.00

Tabla 2.2: Potencia de los test de simetría en el Escenario 2 en dimensión 2 y 3.

μ	Skew normal en dim 2					Skew normal en dim 3				
	RPK_1	Marden	Ley	*RPKW*	RPK_{10}	RPK_1	Marden	Ley	*RPKW*	RPK_{10}
0.00	0.05	0.04	0.05	0.05	0.03	0.05	0.05	0.06	0.05	0.03
0.05	0.12	0.05	0.07	0.05	0.07	0.07	0.06	0.07	0.06	0.04
0.10	0.23	0.07	0.08	0.20	0.20	0.16	0.10	0.08	0.12	0.14
0.15	0.38	0.10	0.16	0.43	0.40	0.29	0.15	0.12	0.40	0.38
0.20	0.53	0.20	0.20	0.55	0.67	0.39	0.18	0.19	0.53	0.61
0.25	0.62	0.30	0.31	0.63	0.84	0.48	0.29	0.23	0.59	0.83
0.30	0.70	0.37	0.40	0.73	0.95	0.55	0.38	0.33	0.71	0.91
0.35	0.73	0.49	0.55	0.85	0.98	0.57	0.50	0.44	0.80	0.97
0.40	0.74	0.63	0.66	0.99	0.99	0.62	0.58	0.51	0.88	0.99
0.45	0.77	0.72	0.76	0.99	1.00	0.63	0.67	0.58	0.93	0.99
0.50	0.78	0.79	0.83	1.00	1.00	0.66	0.70	0.69	0.97	1.00

Desempeño de la prueba de simetría en el caso bivariado

La principal bondad del test de simetría mediante proyecciones al azar es su sencilla aplicación e implementación en dimensiones elevadas e incluso infinita. Sin embargo también se observa su buena performance en dimensiones

Tabla 2.3: Potencia de los test de simetría en los Escenarios 1 y 2 en dimensión 50. Potencia de los test de simetría en el Escenario 3.

μ	Student en dim 50.		Skew normal en dim 50.		$X(t)$ en $C[0,1]$.	
	RPK_1	RPK_{10}	RPK_1	RPK_{10}	RPK_1	RPK_{10}
0.00	0.04	0.05	0.05	0.05	0.05	0.05
0.05	0.07	0.10	0.06	0.06	0.06	0.06
0.10	0.12	0.27	0.08	0.09	0.11	0.10
0.15	0.20	0.53	0.28	0.27	0.19	0.19
0.20	0.30	0.78	0.33	0.38	0.30	0.30
0.25	0.38	0.91	0.45	0.50	0.43	0.46
0.30	0.45	0.96	0.52	0.59	0.57	0.61
0.35	0.50	0.98	0.53	0.70	0.69	0.74
0.40	0.56	0.99	0.59	0.78	0.79	0.86
0.45	0.61	1.00	0.60	0.85	0.87	0.93
0.50	0.64	1.00	0.60	0.92	0.91	0.97

bajas. A modo de ejemplo, como en Einmahl and Gan (2016), consideremos las siguientes hipótesis alternativas bivariadas,

- H_1) (A): $X_1, X_2 \sim exp(1) - 1$; X_1,X_2 v.a. independientes.
- H_1) (B): $X_1, X_2 \sim pareto(1) - 1$; X_1,X_2 v.a. independientes.
- H_1) (C): $\tilde{\Theta} \sim U(0, 2\pi)$, $R|\tilde{\Theta} \sim exp(1)$ si $\tilde{\Theta} \leq 5\pi/4$, $R|\tilde{\Theta} \sim exp(10)$ si $\tilde{\Theta} > 5\pi/4$, $(X_1, X_2) = (R\cos(\tilde{\Theta}), R\sin(\tilde{\Theta}))$.
- H_1) (D): $X_1, X_2 \sim \chi^2_1 - 1$; X_1,X_2 v.a. independientes.
- H_1) (E): $X_1 \sim N(0, 1)$, $X_2 \sim exp(1) - 1$; X_1,X_2 v.a. independientes,

en donde exp, $pareto$, U, χ^2 son las distribuciones exponencial, Pareto, Uniforme y Chi-cuadrado respectivamente.

A partir de las alternativas anteriores se analiza la potencia de los test. Se compara con la propuesta realizada en Einmahl and Gan (2016) donde se rechaza al hipótesis nula para valores elevados del estadístico T_n, que se define de la siguiente manera: Sea $(\tilde{\Theta}, R)$ las coordenadas polares de el vector (X_1, X_2), se considera

$$\Theta = \begin{cases} \tilde{\Theta} & \text{si } \tilde{\Theta} \in [0, \pi) \\ \tilde{\Theta} - \pi & \text{si } \tilde{\Theta} \in [\pi, 2\pi), \end{cases}$$

$$\delta = \begin{cases} 1 & \text{si } \tilde{\Theta} \in [0, \pi) \\ -1 & \text{si } \tilde{\Theta} \in [\pi, 2\pi), \end{cases}$$

$$U_i = F_{R|\Theta}(R_i|\Theta_i), \qquad \hat{U}_i = \hat{F}_{R|\Theta}(R_i|\Theta_i),$$

donde $F_{R|\Theta}$ es la distribución acumulada de R dado Θ y $\hat{F}_{R|\Theta}$ es un estimador de $F_{R|\Theta}$. Entonces el estadístico T_n es dado por

$$T_n = \int_0^1 \int_0^\pi \int_{\theta_1}^\pi \left(\hat{W}_n(\theta_2, u) - \hat{W}_n(\theta_1, u)\right)^2 d\hat{G}(\theta_2) d\hat{G}(\theta_1) du + \\ + \int_0^1 \int_0^\pi \int_0^{\theta_1} \left(\hat{W}_n(\pi, u) - \hat{W}_n(\theta_1, u) - \hat{W}_n(\theta_2, u)\right)^2 d\hat{G}(\theta_2) d\hat{G}(\theta_1) du,$$

donde $\hat{W}_n(\theta, u) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \mathbb{1}_{[0,\theta]\times[0,u]} \left(\Theta_i, \hat{U}_i\right)$, y $\hat{G}$ representa la distribución empírica de $\Theta_1, \ldots, \Theta_n$. Para un nivel de significación de $\alpha = 0.05$, se computan 2000 replicaciones para tamaños de muestra $n = 100$ y $n = 200$. Los resultados se encuentran expresados en la Tabla 2.4, donde se puede apreciar que nuestro test es también un competidor a considerar en dimensión 2.

Tabla 2.4: Proporción de rechazos para las diferentes alternativas

Distribucón	$n = 100$.			$n = 200$.		
H_1)	T_n	Ley	RPK_{10}	T_n	Ley	RPK_{10}
(A)	0.99	0.97	0.90	1.00	0.99	0.99
(B)	0.85	1.00	0.87	1.00	1.00	0.95
(C)	0.81	0.77	0.95	1.00	0.97	1.00
(D)	1.00	0.99	0.99	1.00	1.00	1.00
(E)	0.86	0.31	0.77	0.99	0.59	0.90

2.5.2. Simulaciones para el test de independencia

Al igual que en simetría el desempeño del test es evaluado mediante simulaciones. El test contra el cual realizamos las comparaciones es introducido en Székely et al. (2007), donde se encuentra un relevante estudio sobre independencia de vectores aleatorios. En Székely and Rizzo (2013) el test es ajustado

para mejorar su performance en dimensiones elevadas. A este le llamaremos el test de distancia covarianza y lo denotaremos por $DIST.COV$

Dada la función compleja γ en $\mathbb{R}^p \times \mathbb{R}^q$, la norma $\|\cdot\|_w$ se define por,

$$\|\gamma(s,t)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(s,t)|^2 w(t,s)dtds,$$

donde $w(t,s)$ es una función de pesos arbitraria. Dado un par de vectores aleatorios $(X,Y) \in \mathbb{R}^p \times \mathbb{R}^q$, en Székely et al. (2007) se introduce una medida de independencia entre ellos llamada *distancia covarianza*,

$$\begin{aligned} \mathcal{V}(X,Y;w) &= \|\psi_{X,Y}(t,s) - \psi_X(t)\psi_Y(s)\|_w^2, \\ \mathcal{V}(X;w) &= \mathcal{V}(X,X;w), \end{aligned} \tag{2.18}$$

donde $\psi_X(t), \psi_Y(t)$ representan las funciones características de los vectores aleatorios X e Y respectivamente, mientras que $\psi_{X,Y}$ es la función de distribución conjunta. Además ellos incluyen en su trabajo una versión estandarizada de dependencia que denominan *distancia correlación* dada por,

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}(X,Y;w)}{\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}}, \tag{2.19}$$

donde

$$\|\cdot\|_w^2 = \int_{\mathbb{R}^{p+q}} |\cdot|^2 w(t,s)dtds.$$

La función de pesos $w(t,s)$ es elegida apropiadamente,

$$w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}, \quad \text{con } c_d = \frac{\pi^{(1+d)/2}}{\Gamma\left((1+d)/2\right)}. \tag{2.20}$$

En Székely and Rizzo (2013) modifican este estadístico y determinan su distribución asintótica bajo el supuesto de independencia entre X y Y. Este test se encuentra implementado en el paquete *energy* de *R*.

La comparación se realiza en tres escenarios,

- (Escenario 1.)

  Se considera una muestra de 20 vectores aleatorios iid $(X^{(k)}, Y^{(k)})$ de los vectores X e Y. Las distribuciones marginales son uniformes. La relación de dependencia está dada a través de la cópula de Gumbel con $\delta = 2$, es

decir

$$C(x, y; \delta) = \exp\left(-\left[(-\log x)^{\delta} + (-\log y)^{\delta}\right]^{1/\delta}\right) \quad \delta \geq 1,$$

ver Figura 2.3.

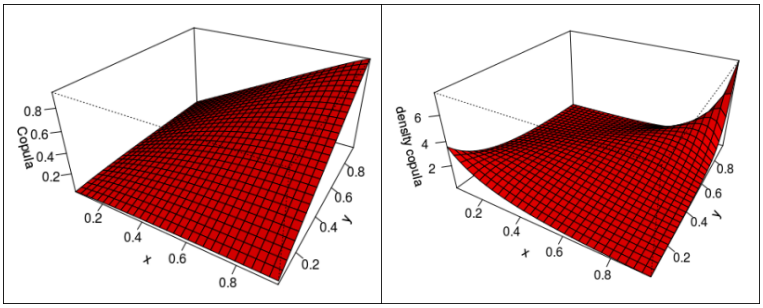

Figura 2.3: Para $\delta = 2$. Panel Izquierdo: Cópula de Gumbel. Panel Derecho: Densidad de la cópula de Gumbel.

La construcción de alternativas se realiza reemplazando las últimas j coordenadas del vector Y por las primeras j coordenadas del vector X para diferentes valores de $j \in \{0, 5, 9, 13, 17, 21, 25, 29\}$.

- (Escenario 2.)

  Al igual que en el caso anterior se simula una muestra iid de tamaño 20, $(X^{(k)}, Y^{(k)})$ de vectores aleatorios de dimensión 50, con marginales uniformes y una cópula de Gumbel con $\delta = 2$. Para construir una muestra de alternativas dependientes se reemplaza la coordenada del vector Y por la coordenada del vector X, si las coordenada de la primera supera un determinado umbral u que toma valores entre 0 y 1.

- (Escenario 3.)

  Se simula una muestra de tamaño 20, de vectores aleatorios U, V y Z de dimensión 50. Son vectores independientes, cada uno con marginales uniformes y asociadas a través de la cópula de Gumbel con $\delta = 2$. Sea $X = (1 - \epsilon)U + \epsilon Z$ y $Y = (1 - \epsilon)V + \epsilon Z$ con $\epsilon \in [0, 1]$. Cuando $\epsilon = 0$ se corresponde con el caso de independencia.

Son comparados los desempeños de los siguientes test en los tres escenarios descriptos.

- (RPK.INDEP$_{50}$ Test.) Para cada muestra, se consideran 50 direcciones independientes al azar sobre la esfera unidad y de manera idéntica otras 50 direcciones que denominaremos h y f respectivamente. Las muestras X e Y son proyectadas sobre las direcciones h_j y f_j, $j = 1, \ldots, 50$ respectivamente. A partir de los datos proyectados se realiza el test desarrollado en Genest et al. (2007), en donde se calcula el p-valor del test en $\mathbb{R}^2$ en cada par de direcciones (h_j, f_j). Si en al menos uno de los pares de direcciones el test es rechazado, entonces se rechaza H_0 del test original. Este procedimiento se realiza 10.000 veces y se contabiliza la proporción de p-valores por debajo de $\alpha = 0.001$ (corrección por Bonferroni) en cada test.

- (DIST.COV. Test) Test desarrollado en Székely and Rizzo (2013), ($\alpha = 0.05$).

Tabla 2.5: Potencia del test de independencia en los escenarios 1, 2 y 3.

Escenario 1			Escenario 2			Escenario 3		
k	RPK.INDEP$_{50}$	DIST.COV	u	RPK.INDEP$_{50}$	DIST.COV	ϵ	RPK.INDEP$_{50}$	DIST.COV
0	0.01	0.05	1	0.04	0.04	0.00	0.04	0.08
5	0.10	0.12	0.86	0.19	0.21	0.14	0.05	0.11
9	0.55	0.13	0.71	0.60	0.51	0.29	0.10	0.12
13	0.82	0.24	0.57	0.85	0.87	0.43	0.21	0.31
17	0.98	0.26	0.43	0.96	0.97	0.57	0.72	0.92
21	1.00	0.63	0.26	0.98	0.99	0.71	1.00	1.00
25	1.00	0.78	0.14	1.00	1.00	0.86	1.00	1.00
29	1.00	0.93	0	1.00	1.00	1.00	1.00	1.00

Como se observa en la Tabla 2.5, en los Escenarios 2 y 3 el test por proyecciones al azar RPK.INDEP$_{50}$ y el test DIST.COV tienen un comportamiento similar. En el Escenario 1 el test RPK.INDEP$_{50}$ tiene un mejor desempeño.

2.6. Datos Reales: Actividad neuronal en individuos alcohólicos

La base de datos de estudios consiste en ondas registradas en un electroencefalograma (EEG) para 88 pacientes. Pensamos estas señales como datos funcionales. En la base de estudio encontramos dos grupos, 44 pacientes alcohólicos y 44 pacientes de control. El objetivo del problema es encontrar evidencia que el consumo de alcohol en forma desmedida provoca cierta pérdida

de asociación o sincronización entre los dos hemisferios del cerebro humano, y en particular cuáles son aquellas áreas más afectadas. El conjunto de datos utilizado es abierto y se encuentra disponible en el repositorio UCI [1] Una descripción detallada de la base de datos es brindada en Zhang et al. (1995). La base esta conformada por las medidas obtenidas de 64 electrodos ubicados sobre el cuero cabelludo midiendo la actividad neuronal a 256 Hz por cada 1 segundo. En el estudio se consideran 27 tripletas de nodos como se muestra en la Figura 2.4. Cada sujeto es expuesto a un estímulo (S1) o a dos estímulos consecutivos (S1 y S2), en que cada uno de ellos consiste en mostrar una pintura. Cada observación se encuentra discretizada en 256 medidas tomadas cada un segundo. Los valores emitidos en EEG están expresados en micro volts. Se remueve el ruido a través de la descomposición espectral usando la transformada rápida de Fourier (fast Fourier transform). La asociación para cada una de las tripletas es estudiada como se muestra en la Figura 2.4. Fijada la tripleta de electrodos los pasos a seguir son los siguientes

Selección de los datos:

- Se seleccionan los datos correspondientes a un sujeto para la tripleta elegida.
- Cada dato esta conformado por tres observaciones, una por cada electrodo. Y para cada segundo contamos con 256 medidas.
- Por tanto, contamos con 60 vectores de 256 medidas para cada tripleta de electrodos

Proyección de la tripleta:

- Para cada nodo, asumimos que tenemos 60 trayectorias, cada una de ellas discretizada en 256 valores por segundo.
- Se elige una dirección aleatoria $h \in E^*$ generada a partir de un Proceso Browniano discretizado, es decir, partiendo de 0 se consideran incrementos independientes a partir de una variable $N(0, 1/256)$.
- Se obtiene así para cada nodo una muestra de 60 valores reales. Si consideramos las tripletas tenemos muestras de tamaño 60 en $\mathbb{R}^3$ (ver Figura 2.5), donde $\mathbf{X}_i = (X_{1,i}, X_{2,i}, X_{3,i})$ con $i = 1, 2, \ldots, 60$.

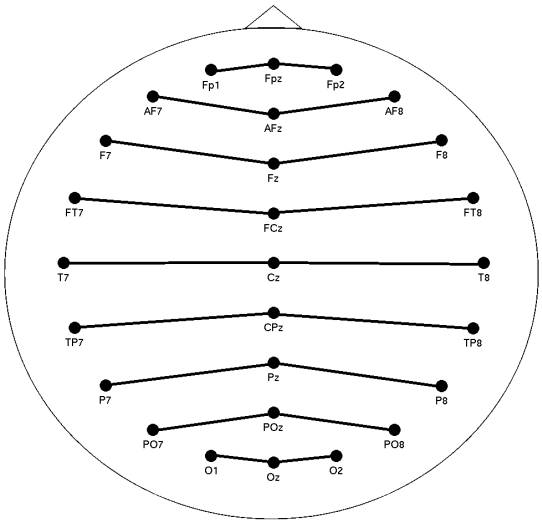

Figura 2.4: Asociación de los electrodos en el test de independencia

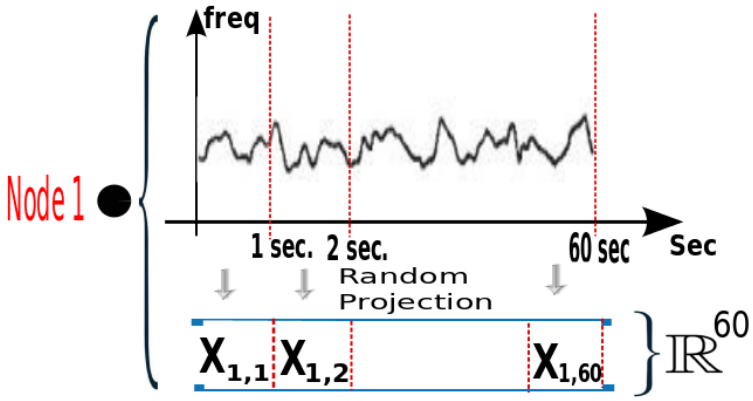

Figura 2.5: Esquema de la construcción de la muestra a través de proyecciones al azar

Cálculo del p-valor para cada tripleta:

- Usando nuestro test, obtenemos 4 p-valores, uno para cada par de nodos y otro para el total de la tripleta.

Este proceso se reitera para los 44 sujetos que integran el grupo de control y los 44 individuos alcohólicos. Por tanto se obtienen 44×4 p-valores para los sujetos de control y la misma cantidad para los alcohólicos.

Se realiza la comparación de los resultados:

[1] https://archive.ics.uci.edu/ml/datasets/EEG+Database

- Se realizan los *boxplot* de los p-valores obtenidos para cada una de las 4 combinaciones posibles por tripleta.
- Se representan en color gris los *boxplot* de los sujetos de control y en blanco los alcohólicos.
- Cada *boxplot* es generado a partir de los 44 pacientes correspondientes a cada grupo.

Este procedimiento se reitera con cada una de las 9 tripletas y los resultados obtenidos se visualizan en la Figura 2.6.

En enfermedades neurodegenerativas como la enfermedad de Alzheimer (EA), se han observado tres efectos principales en el EEG: ralentización del EEG, disminución de la complejidad de las señales del EEG y perturbaciones en la sincronía del EEG (ver Dauwels et al. (2010)). Dawels hace una revisión de algunas posibles medidas de la sincronía EEG consideradas en el contexto del diagnóstico de EA. En nuestro trabajo podríamos considerar el valor de p de la prueba de independencia desarrollada como una medida de asociación o sincronía. Un valor mayor de p indicaría una mayor pérdida de asociación.

Como se puede observar e la Figura 2.6, los *boxplot* del grupo de control presentan una menor mediana mayor y una dispersión mas baja, lo que podría estar dando un indicio que esta población presenta una mayor sincronización. Por otro lado, se puede apreciar que el área frontal del cerebro exhibe menos pérdida de asociación que el resto.

2.7. Conclusiones del capítulo

- En este capítulo se expusieron dos test no paramétricos, uno de simetría y otro de independencia, sustentados en las proyecciones al azar. Mediante un estudio de simulación se muestra el buen desempeño de ambas pruebas en referencia a otros competidores.
- Ambos test pueden ser implementados de manera simple tanto en dimensión finita como infinita, presentando un alto desempeño en lo referente a tiempos computacionales.
- A través de la prueba de independencia se detecta un pérdida de sincronización en las ondas de un EEG en pacientes alcohólicos en referencia al grupo de control.

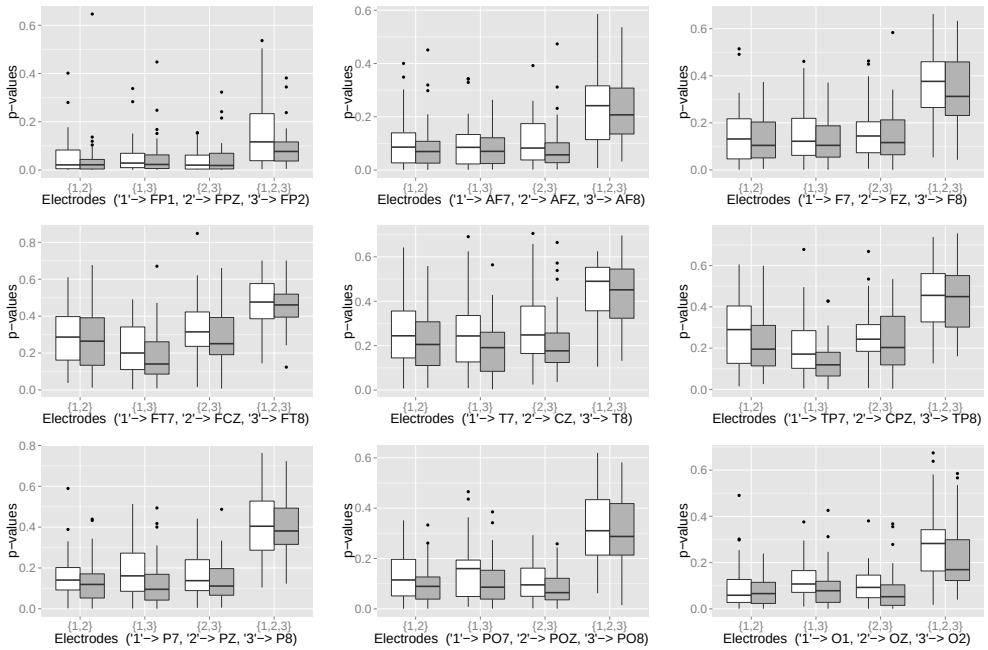

Figura 2.6: *Boxplot* de los p-valor obtenidos para cada uno de los grupos. En cada uno de los 9 paneles (uno por cada tripleta) se muestran los *boxplot* de los 4 test posibles por tripleta. Se representa en color gris los del grupo de control y en blanco los alcohólicos.

Capítulo 3

Profundidad estadística sobre variedades Riemannianas

El concepto de profundidad estadística es una herramienta de gran importancia en la estadística moderna. En la últimas cuatro décadas ha surgido la noción de profundidad de los datos, como un mecanismo de ordenamiento respecto a un centro en un marco multivariado, y más recientemente para datos en dimensión infinita. La idea de profundidad permite generalizar nociones robustas de posición como los son por ejemplo la mediana o la media recortada. La posibilidad de poder asignar una medida de profundidad a los datos ha tenido suma utilidad en diversas aplicaciones. Por ejemplo, en investigaciones recientes, es frecuente su uso en problemas de clasificación supervisada, de clasificación no supervisada, de test de hipótesis y de detección de oultiers (ver Liu et al. (1999)).

En primera instancia el concepto de profundidad fue introducido en Tukey (1975) para una muestra de datos bivariados llamada profundidad por semiespacios y posteriormente se extendió a dimensiones mas elevadas en Donoho and Gasko (1992).

Fueron desarrollados en años posteriores diversas nociones y formas de entender la profundidad estadística. Por ejemplo en Barnett (1976) se desarrolla el concepto de profundidad mediante envolturas convexas. Otras medidas de profundidad son por ejemplo la de Oja (Oja (1983)), la profundidad simplicial (Liu (1990)), la profundidad espacial (Vardi and Zhang (2000)) y la profundidad esférica (Elmore et al. (2006)). Recientemente también se han introducido diversos medidas de profundidad para datos funcionales (ver por ejemplo Frai-

man and Muniz (2001) y Cuevas and Fraiman (2009)).

Ya han sido estudiadas de manera exhaustiva las propiedades de las medidas clásicas de profundidad como lo son la profundidad simplicial, la profundidad por semi-espacios y la esférica. En tal sentido son conocidas y listadas (ver Liu (1990) y Serfling and Zuo (2000)) aquellas buenas propiedades generales que una medida de profundidad debería tener. Dichas propiedades son las siguientes,

P1) Invariancia con respecto a un grupo de transformaciones. En general se considera el grupo de transformaciones afines o un grupo más reducido como lo es el grupo de las transformaciones ortogonales:

P11: Invariancia Afin: La profundidad no debería depender del sistema de coordenadas, en particular, de la escala de medida utilizada en cada uno de los ejes de coordenadas.

P12: Invariancia Ortogonal: La profundidad no debería depender de la posición del sistema de coordenadas ni de la escala global utilizada.

P2) Maximalidad respecto a un centro: Diremos que la distribución es simétrica respecto a un centro si es invariante frente a un grupo de transformaciones. Según quienes sean estas transformaciones encontramos diferentes conceptos de simetría, por ejemplo, la simetría angular, central, esférica, elíptica son algunos de ellos. En todos los casos si estamos en el caso univariado estos conceptos coinciden con la noción usual de simetría para datos unidimensionales.

Para aquellas distribuciones que tengan definidas un único *centro*, presentando la distribución alguna tipo de simetría respecto a este, la profundidad se debería maximizar en el centro de simetría.

P3) Monotonía respecto al punto más profundo: Si nos alejamos del punto más profundo a través de un rayo la profundidad debería disminuir.

P4) Desvanecimiento en el infinito: Si la distancia de un punto al punto más profundo se va al infinito, entonces la profundidad del punto debería tender a cero.

Por otro lado en Serfling (2006) se mencionan dos propiedades que según Serfling todo procedimiento no paramétrico de análisis multivariado en la ac-

tualidad debería tener, en particular una medida no paramétrica de profundidad. Por un lado que tenga en cuenta la *dimensionalidad intrínseca* de los datos. Es decir, si los datos se encuentran en $\mathbb{R}^d$ pero tienen una estructura con un dimensión nominal menor que d. El procedimiento estadístico debería incorporar este hecho. Por otro lado, el estimador debe ser sencillo de calcular en tiempos computacionales razonables, en función del tamaño de la muestra y la dimensionalidad de los datos.

Referente a estos últimos aspectos, con excepción de algunas pocas medidas de profundidad (como lo son la profundidad esférica y la profundidad espacial), el cálculo efectivo de las clásicas medidas de profundidad es un problema de elevada complejidad computacional, lo que imposibilita su cálculo ya en dimensiones moderadas ($d \geq 4$). Este problema surge por ejemplo en el tratamiento de imágenes (ver Pizer and Marron (2017)).

Por tanto, es necesario observar que medidas de profundidad son aplicables cuando los datos están en un espacio euclídeo de alta dimensión, dimensión infinita, o donde no encontramos una estructura de espacio vectorial, problemas de actual interés en la comunidad estadística. Por ejemplo en Pennec (2006) se pueden encontrar aplicaciones recientes a diagnósticos médicos, en donde la morfología y visualización de objetos en 3D son de particular importancia y los datos se encuentran sobre una variedad Riemanniana. Es necesario entonces, para un análisis estadístico apropiado, poder extender conceptos desarrollados en la estadística clásica a espacios más generales. Por tanto, nuestro principal objetivo en este capítulo es definir una medida de profundidad (extenderemos la profundidad esférica) para datos en un marco más general. Dicha extensión la denominaremos DCOPS (connecting pairwise geodesic spheres by depth).

En dicho contexto, cuando los datos se encuentran sobre una variedad Riemanniana, es importante poder encontrar un medida de centralidad (ver Arnaudon et al. (2013) y Fletcher et al. (2009)).

Cabe señalar que si bien el concepto de profundidad para datos funcionales (FDA) ya ha sido desarrollado por varios autores, como se menciona en Cuevas (2014), es una línea de investigación que esta lejos de estar cerrada. Se muestra que, bajo un conjunto de supuestos débiles, el estimador de la profundidad esférica propuesto en Elmore et al. (2006) es consistente en el paradigma de FDA.

En este capítulo se enuncian y demuestran las propiedades deseables para una profundidad (aquellas que conserven el sentido sobre una variedad) que

detenta DCOPS. Se estudia la consistencia y TCL para el estimador. Para el caso funcional a partir de resultados fundamentales desarrollados en Billingsley and Topsøe (1967) se prueba la convergencia uniforme del estimador propuesto. En la sección final se estudia en detalle un caso de particular relevancia, los datos se encuentran en la variedad Riemanniana de las matrices definidas positivas.

3.1. Profundidad esférica

El concepto de profundidad esférica es introducida en Elmore et al. (2006). La idea es similar a la profundidad simplicial pero en lugar de contabilizar el número de simplices con vértices en los puntos de la muestra que contienen a el punto, se considera el número de bolas con diámetro determinado por pares de datos. Este mecanismo presenta la ventaja que el número de bolas posible depende del tamaño de la muestra pero no de la dimensión del espacio, lo que si sucede en la profundidad simplicial. De manera más precisa, en un espacio euclídeo $\mathbb{R}^d$ la profundidad esférica se define de la siguiente manera:

Sean dos puntos $x, y \in \mathbb{R}^d$, si denotamos por B_{xy} a la bola cerrada de diámetro el segmento $\overline{xy}$, es decir, la bola cerrada con centro en el punto medio entre x e y (lo denotaremos $xy/2$) y radio $d(x,y)/2$, donde $d(\cdot,\cdot)$ es la distancia euclidiana en $\mathbb{R}^d$. Dada un vector aleatorio X en $\mathbb{R}^d$ se define la profundidad esférica de $x \in R^d$ como,

$$\mathrm{BD}(x) := \Pr\left(x \in B_{X_1 X_2}\right),$$

en donde X_1 y X_2 son copias de X independientes.

Ahora, dada una muestra aleatoria $X_1, \ldots, X_n$ de variables independientes con la misma distribución que X, la versión empírica de BD es dada por un U–estadístico de orden 2, es decir, dado $x \in \mathbb{R}^d$,

$$\widehat{\mathrm{BD}}_n(x) := U_2^n\left(D_x\right) = \frac{1}{\binom{n}{2}} \sum_{1 \le i_1 < i_2 \le n} \mathbf{1}_{B_{X_{i_1} X_{i_2}}}(x),$$

donde

$$D_x := \{(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d : x \in B_{x_1 x_2}\},$$

y

$$U_2^n(A) := \frac{1}{\binom{n}{2}} \sum_{i_1<i_2} \mathbf{1}_A(x_{i_1}, x_{i_2}), \text{ con } A \in \mathcal{B}\left(\mathbb{R}^d \times \mathbb{R}^d\right).$$

Anotamos $\mathcal{D}$ a la familia de todos los conjuntos D_x, es decir, $\mathcal{D} := \{D_x : x \in \mathbb{R}^d\}$. Cuando $d = 1$ es sencillo ver que BD coincide con la profundidad simplicial y el punto más profundo con la mediana.

Esta definición es también bien definida en un espacio de Hilbert $\mathcal{H}$ separable, donde la distancia es la inducida por el producto interno. Para este caso se puede obtener una interpretación geométrica expresando la definición en términos del producto interno como,

$$\mathrm{BD}(x) = \Pr\left(\langle X_1 - x, X_2 - x\rangle \leq 0\right),$$

en donde X_1 y X_2 son dos elementos aleatorios independientes en $\mathcal{H}$ copias de X, y $\langle\cdot,\cdot\rangle$ es el producto interno en $\mathcal{H}$.

Se observa que la complejidad computacional del estimador de esta medida es sólo del orden $O(d \times n^2)$. Es decir, es lineal en la dimensión del espacio d y cuadrática en función del tamaño de la muestra n. Una de las posibles debilidades de la profundidad esférica es que es invariante frente a transformaciones ortogonales pero no afines. Sin embargo, en nuestro contexto (las variedades Riemannianas) sólo tiene sentido realizar transformaciones ortogonales de la variedad dentro del espacio en la cuál se encuentra inmersa.

El capítulo se organiza de la siguiente manera. En la Sección 3.2 se detallan algunos ejemplos y las propiedades básicas de DCOPS cuando los datos se encuentran sobre una variedad Riemanniana. En la Subsección 3.2.2 se deduce la convergencia uniforme y la distribución asintótica del estimador. En la Subsección 3.2.3 se estudia el DCOPS para el caso de análisis de datos funcionales (FDA). En la Sección 3.3 se realizan diversas simulaciones sobre datos en variedades Riemannianas con especial foco cuando los datos son matrices definidas positivas.

3.2. DCOPS en variedades Riemannianas

Sea $(\mathcal{M}, g)$ una variedad Riemanniana y d_g la distancia inducida por la métrica g. En el resto de capítulo asumiremos que la variedad tiene una única componente conexa y es orientada. Consideramos que el espacio métrico

$(\mathcal{M}, d_g)$ es separable y completo. Sea $X : \Omega \to \mathcal{M}$ un elemento aleatorio con densidad f_X con respecto a una medida ν, asumimos que la medida tensorial de probabilidad sobre la variedad $d\xi(p,q) := f_X(p)f_X(q)d\nu(p)d\nu(q)$, si elegimos dos puntos $(p,q) \in \mathcal{M} \times \mathcal{M}$ sobre la variedad con dicha medida de probabilidad, entonces existe y es única la geodésica minimizante determinada por p y q con probabilidad 1.

Para cada par de puntos $p, q \in \mathcal{M}$ que determinan una única geodésica minimizante $\overline{pq}$, definimos la bola de diámetro $\overline{pq}$ (la denotaremos por B_{pq}) como la bola cerrada con centro en el punto medio de la geodésica que une p y q y radio $d_g(p,q)/2$.

Se define DCOPS en un punto $p \in \mathcal{M}$, que denotaremos por $\mathrm{BD}(p)$, como

$$\mathrm{BD}(p) := \Pr\left(p \in B_{X_1X_2}\right) = \int_{\mathcal{M}\times\mathcal{M}} \mathbf{1}_{B_{x_1x_2}}(p) f_X(x_1) f_X(x_2) d\nu(x_1) \times d\nu(x_2),$$

donde X_1 y X_2 son dos elementos aleatorios independientes copias de X.

Dada ahora la muestra de elementos independientes $X_1, \ldots, X_n$ copias de X, la versión empírica de $\mathrm{BD}(p)$ es dada, al igual que en $\mathbb{R}^d$, por el U–estadístico asociado de orden 2,

$$\widehat{\mathrm{BD}}_n(p) = \frac{1}{\binom{n}{2}} \sum_{1 \le i_1 < i_2 \le n} \mathbf{1}_{B_{X_{i_1}X_{i_2}}}(p). \tag{3.1}$$

3.2.1. Un par de ejemplos

En esta sección exponemos dos simples ejemplos donde los datos se encuentran sobre un toro y sobre la representación de un rostro humano.

En el toro $\mathbb{T}^p = [0, 2\pi)^p$ consideramos la distribución multivariada de von Mises, que denotaremos por $\mathcal{MVM}(\mu, \kappa, \Delta)$ (ver por ejemplo Mardia and Voss (2014) y Mardia et al. (2008)). La densidad evaluada en $\theta \in \mathbb{T}^p$ esta dada por

$$f(\theta; \mu, \kappa, \Delta) = \frac{1}{Z(\kappa, \Delta)} \exp\{\kappa^\top c(\theta) + s(\theta)\Delta s(\theta)/2\},$$

donde $\mu \in \mathbb{T}^p$ (parámetro llamado media), $\kappa \in \mathbb{R}^d$ de componentes positivas (parámetro denominado de concentración), $\Delta = (\lambda_{i,j})$ es una matriz simétrica en $\mathbb{R}^{d\times d}$ con entradas en la diagonal nulas ($\lambda_{i,i} = 0$ para todo $i \in \{1, \ldots, d\}$) y $Z(\kappa, \Delta)$ es la constante de normalización. Las funciones c_i y s_i se encuen-

tran definidas como $c_i(\theta) = \cos(\theta_i - \mu_i)$ y $s_i(\theta) = \sin(\theta_i - \mu_i)$ para todo $i \in \{1, \ldots, d\}$. En el panel izquierdo de la Figura 3.1 se representan 100 datos simulados a partir de la distribución $\mathcal{MVM}_1(\mu^*, \kappa^*, \Delta^*)$ con

$$\mu^* = (\pi/2, 0), \quad \kappa^* = (20, 20), \quad \Delta^* = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

En el panel derecho se simulan 60 observaciones a partir del modelo determinado por una mezcla de distribuciones,

$$0.9\mathcal{MVM}_1(\mu^*, \kappa^*, \Delta^*) + 0.1\mathcal{MVM}_2(\mu^{**}, \kappa^{**}, \Delta^{**}),$$

con $\mu^{**} = (7/4\pi, 0)$, $\kappa^{**} = (100, 100)$ y $\Delta^{**} = \Delta^*$. Se determina la profundidad empírica DCOPS de cada uno de los puntos representados en la Figura 3.1 mediante una escala de colores. Se observa como la profundidad decrece cuando nos alejamos por rayos geodésicos del centro μ_1. Por otro lado, los datos generados por la segunda componente de la mezcla (outliers) presentan un bajo valor de DCOPS, propiedad también deseable en toda medida de profundidad.

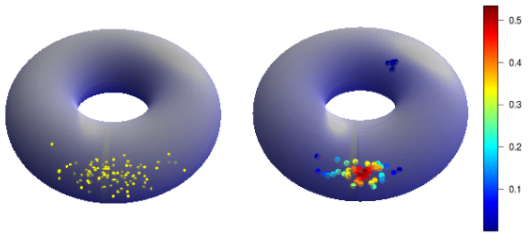

Figura 3.1: Panel Izquierdo: 100 puntos muestrales generados a partir de la distribución $\mathcal{MVM}_1(\mu^*, \kappa^*, \Delta^*)$. Panel Derecho: $\widehat{DB}$ de una muestra de tamaño 60 generados a partir de la mezcla $0.9\mathcal{MVM}_1(\mu^*, \kappa^*, \Delta^*) + 0.1\mathcal{MVM}_2(\mu^{**}, \kappa^{**}, \Delta^{**})$.

Consideremos ahora como variedad una superficie en $\mathbb{R}^3$, por ejemplo la representación de un rostro humano, ver Figura 3.2. Sobre el rostro que llamaremos $\mathcal{M}$ se fija un punto p (que representamos con color violeta en la Figura 3.2), si d_g es la distancia geodésica, definimos la densidad f sobre $\mathcal{M}$ como,

$$f(x) := \frac{\psi(x)}{\int_{\mathcal{M}} \psi(t) d\nu(t)}, \tag{3.2}$$

con $\psi(x) = 1/\{1 + d_g(p, x)\}$. Se determina una muestra de 50 puntos alea-

torios sobre el rostro, para ello se considera una grilla de valores sobre la superficie y se discretiza la densidad planteada f. En el panel izquierdo de la Figura 3.2 se observa los valores simulados con la profundidad DCOPS asociada, mientras que en el panel derecho se representan todos los puntos del rostro con su profundidad DCOPS asociada. Como se puede observar la profundidad decrece a medida que nos alejamos del punto fijo p y la medida de profundidad, al utilizar las distancias geodésicas, incorpora en sus valores la geometría del rostro.

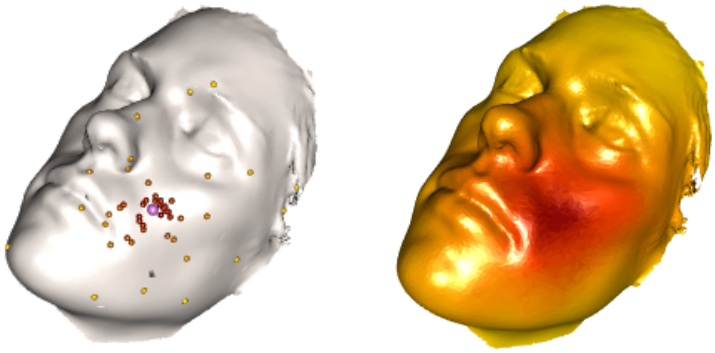

Figura 3.2: Panel Izquierdo: Valores de $\widehat{DB}$ para 50 puntos simulados. En violeta representamos el punto más profundo. Panel Derecho: Intensidad de $\widehat{DB}$ en todos los puntos del rostro. Cuanto más es la intensidad del color rojo es mayor el valor de $\widehat{DB}$.

3.2.2. Algunas propiedades de DCOPS en las variedades Riemannianas

En esta sección describimos algunas de las propiedades deseables, que tienen sentido en el marco de variedades Riemannianas, de DCOPS y de su distribución empírica.

2.1.1 Invariancia Ortogonal:

DCOPS es invariante bajo transformaciones ortogonales en el espacio euclídeo $\mathbb{R}^d$ donde la variedad Riemanniana se encuentra inmersa.

2.1.2 Desvanecimiento en el infinito:

Enunciamos un Teorema análogo al desvanecimiento en el infinito para DCOPS y su versión empírica.

Teorema 3.2.1 *Sea q un punto fijo de la variedad Riemanniana $\mathcal{M}$ y $X : \Omega \to \mathcal{M}$ es un elemento aleatorio con densidad f_X*

a) $\sup_{d_g(p,q)>M} \mathrm{BD}(p) \to 0$ *cuando* $M \to \infty$*;*

b) $\lim_{M\to\infty} \sup_{d_g(p,q)>M} \widehat{\mathrm{BD}}_n(p) = 0$ *casi seguramente.*

Demostración.

a) Comencemos por mostrar que si $d_g(p, q) > M$ entonces

$$\mathrm{BD}(p) = P(p \in B_{X_1X_2}) \leq 2P\{d_g(q, X) > M/4\},$$

lo cuál demostraremos por contradicción.

Dados $x_1, x_2 \in B(q, M/4)$, si denotamos $x_1x_2/2$ al punto medio sobre la geodésica determinada por x_1 y x_2, entonces

$$d_g\left(q, x_1x_2/2\right) \leq d_g(q, x_1) + d_g\left(x_1, x_1x_2/2\right) = d_g(q, x_1) + d_g(x_1, x_2)/2 < M/2.$$

Observemos que $p \notin B_{x_1x_2}$ puesto que si $d_g\left(p, x_1x_2/2\right) < d_g(x_1, x_2)/2$ se cumple que

$$d_g(q, p) \leq d_g\left(q, x_1x_2/2\right) + d_g\left(x_1x_2/2, p\right) < M2 + M/4 < M,$$

lo cuál sería una contradicción. Partiendo de la hipótesis que $(\mathcal{M}, d_g)$ es separable, se cumple que $d_g(X, q)$ es una variable aleatoria. Por lo tanto

$$\mathrm{BD}(x) \leq P\left[\{X_1 \notin B(q, M/4)\} \cup \{X_2 \notin B(q, M/4)\}\right] \leq 2P\{d_g(X, q) > M/4\}.$$

Finalmente por el Teorema de Prokhorov, como X es tensa, se deduce que $\lim_{M\to\infty} P\{d_g(X, q) > M/4\} = 0$. □

b) Supongamos ahora que dado cualquier $M > 0$, existe un conjunto de probabilidad positiva $\Omega_1 \subset \Omega$ y sea $\delta = P(\Omega_1)$, para los cuales podemos encontrar $\gamma > 0$ y $x_0 \in \mathcal{M}$ con $d_g(q, x_0) > M$ tal que $\widehat{\mathrm{BD}}_n(x_0, w) > \gamma > 0$ para todo $w \in \Omega_1$. Elegimos M_0 tal que $P\{d_g(q, X) > M_0/4\} < \delta/(2n)$. Puesto que $\Omega_1 \subset \bigcup_{i=1}^{n} \{d_g(q, X_i) > M_0/4\}$ obtenemos $\delta \leq nP\{d_g(q, X) > M_0/4\} < \delta/2$, lo que determina una contradicción □

2.1.3 Continuidad:

En el siguiente Teorema demostraremos que si dos puntos son lo suficientemente cercanos respecto a la distancia geodésica entonces sus profundidades también son cercanas. Asumiremos para la demostración la condición siguiente.

Condición ***VC***: Sea una variedad Riemanniana compacta $\mathcal{M}$ y denotamos $\bar{B}(p,r)$ a la bola geodésica de centro $p \in \mathcal{M}$ y radio $r > 0$. Diremos que $\mathcal{M}$ satisface la condición **VC** si la familia de conjuntos $\mathcal{A} = \{A_p : p \in \mathcal{M}\}$ con

$$A_p = \{(x,y) \in \mathcal{M} \times \mathcal{M} : xy/2 \in \bar{B}\{p, d_g(x,y)/2\}\},$$

tiene dimensión de Vapnik–Chervonenkis finita ($VC < \infty$) (para la definición de VC ver Steele (1975) y Dudley (1978)).

Teorema 3.2.2 *Dados $p, q \in \mathcal{M}$ y $X : \Omega \to \mathcal{M}$ un elemento aleatorio con densidad f_X, se tiene que*

a) $|\mathrm{BD}(p) - \mathrm{BD}(q)| \to 0$ *cuando* $d_g(p,q) \to 0$.

b) Si $\mathcal{M}$ una variedad Riemanniana compacta que cumple la condición ***VC****. Si la densidad f_X es una función acotada. Entonces, para $\epsilon > 0$ tenemos que*

$$\sup_{d_g(p,q)<\epsilon} |\widehat{\mathrm{BD}}_n(p) - \widehat{\mathrm{BD}}_n(q)| < \gamma(\epsilon) + R_n,$$

en donde $\gamma(\epsilon)$ es una función determinística que tiende a 0 cuando $\epsilon \to 0$ y R_n una v.a que converge c.s. a 0 cuando $n \to \infty$.

Demostración.

a) Dados dos puntos $p, q \in \mathcal{M}$ anotemos por S_{pq} a la esfera geodésica de diámetro $\overline{pq}$, es decir

$$S_{pq} := \{x \in \mathcal{M} : d_g(pq/2, x) = d_g(p,q)/2\}.$$

Bajo las hipótesis de Teorema tenemos que para todo $p \in \mathcal{M}$, $P(p \in S_{X_1X_2}) =$

0. Consideremos ahora la siguiente diferencia,

$$|\mathrm{BD}(p) - \mathrm{BD}(q)| = |P\left(p \in B_{X_1X_2}\right) - P\left(q \in B_{X_1X_2}\right)| = \\ |P\left(p \in B_{X_1X_2}, q \notin B_{X_1X_2}\right) - P\left(q \in B_{X_1X_2}, p \notin B_{X_1X_2}\right)| \leq \\ \text{máx}\left\{P\left(p \in B_{X_1X_2}, q \notin B_{X_1X_2}\right), P\left(q \in B_{X_1X_2}, p \notin B_{X_1X_2}\right)\right\}.$$

El suceso de que la bola $B_{X_1X_2}$ contenga únicamente a uno de los puntos p y q esta contenido en el evento que la geodésica $\overline{pq}$ determinada por p y q corte a la esfera $S_{X_1X_2}$. Ahora, a partir del Teorema de Convergencia Dominada se deduce

$$|\mathrm{BD}(p) - \mathrm{BD}(q)| \leq P\left(S_{X_1X_2} \cap \overline{pq} \neq \emptyset\right) \to 0$$

cuando $d_g(p,q) \to 0$. Este argumento concluye la demostración del Teorema. □

b)Sea $R^* = \mathrm{diam}(\mathcal{M})$ y denotamos por $A \triangle B$ a la diferencia simétrica entre los conjunto A y B. Entonces

$$|\widehat{\mathrm{BD}}_n(p) - \widehat{\mathrm{BD}}_n(q)| \leq \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} \left|\mathbf{1}_{B_{X_{i_1}X_{i_2}}}(p) - \mathbf{1}_{B_{X_{i_1}X_{i_2}}}(q)\right| \leq \\ \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} \left|\mathbf{1}_{B\left(p,d_g(X_{i_1},X_{i_2})/2\right) \triangle B\left(q,d_g(X_{i_1},X_{i_2})/2\right)}\left(X_{i_1}X_{i_2}/2\right)\right|.$$

Observemos que $h_{p,q}(x,y) = \mathbf{1}_{B(p,d_g(x,y)/2) \triangle B(q,d_g(x,y)/2)}\left(xy/2\right)$ es el núcleo de un U-estadístico de orden 2. Por tanto se cumple que

$$|\widehat{\mathrm{BD}}_n(p) - \widehat{\mathrm{BD}}_n(q)| \leq \\ \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} |h_{p,q}(X_{i_1}, X_{i_2}) - \mathrm{E}\{h_{p,q}(X_{i_1}, X_{i_2})\}| + \mathrm{E}\{h_{p,q}(X_1, X_2)\}.$$

Ahora, puesto que la familia de conjuntos $\mathcal{A}_2 = \{A_{pq} = A_p \triangle A_q\}_{(p,q) \in E \times E}$ tiene dimensión VC finita, a partir del Corolario 3.3 en Arcones and Gine (1993) se deduce que

$$R_n := \sup_{(p,q) \in E \times E} \frac{1}{\binom{n}{2}} \sum_{1 \leq i_1 < i_2 \leq n} |h_{p,q}(X_{i_1}, X_{i_2}) - \mathrm{E}\{h_{p,q}(X_{i_1}, X_{i_2})\}| \xrightarrow{c.s} 0$$

cuando $n \to \infty$, y por el Teorema de Convergencia Dominada, cuando $\epsilon \to 0$,

se cumple que

$$\mathrm{E}\{h_{p,q}(X_1, X_2)\} \leq 2C\nu\{B(p, R^*) \triangle B(q, R^*)\} := \gamma(\epsilon) \to 0.$$

Esto concluye la demostración. □

Observación 1 Si estamos en $\mathbb{R}^d$ la condición **VC** se satisface automáticamente, puesto que la dimensión VC del conjunto $\nu_{\mathcal{A}}$ se encuentra acotada por $3d + 1$. Esta cota se deduce del hecho de que el conjunto A_p lo podemos expresar como

$$\begin{aligned} A_p &= \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \langle p, x + y\rangle - 2\langle x, y\rangle + 2\|p\|^2 \geq 0\} \\ &= \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : g_p(x, y) \geq 0\} \end{aligned}$$

con $g_p(x, y) = \langle p, x+y\rangle - 2\langle x, y\rangle + 2\|p\|^2$. La familia de funciones $\mathcal{G} = \{g_p : p \in \mathbb{R}^d\}$ es un espacio vectorial de dimensión finita generada por $g_0(x, y) = 1$; $g_{1,i}(x, y) = x_i$; $g_{2,i}(x, y) = y_i$; $g_{3,i}(x, y) = x_i y_i$, con $i \in \{1, \ldots, d\}$, $x = (x_1, \ldots, x_d)$ y $y = (y_1, \ldots, y_d)$. Entonces a partir del Teorema 13.9, pág. 221 expuesto en Devroye et al. (2013) podemos deducir que $\nu_{\mathcal{A}} \leq 3d + 1$.

El caso de FDA es analizado como un caso particular en la Sección 3.2.3.

Observación 2 Encontrar aquellas familias de variedades Riemannianas la condición VC es cierta es un tema de investigación abierto, un primer acercamiento a estos tópicos se pueden encontrar en Ferri and Frosini (2008). Sin embargo, al igual como lo hicimos en $\mathbb{R}^d$, para algunas variedades simples como la esfera, el toro o el cilindro la condición VC puede ser verificada directamente. Consideremos el caso de la esfera S_3 unitaria (con centro en el origen y radio 1) inmersa en $\mathbb{R}^3$. Dados dos puntos $x = (x_1, x_2, x_3)$ e $y = (y_1, y_2, y_3)$, entonces el par $(x, y) \in A_p$ si se cumple la condición

$$g_p(x, y) := \|(x - y)/2\|^2 - \sum_{i=1}^{3} \{p_i - (x_i + y_i)/2\}^2 \geq 0,$$

con $p = (p_1, p_2, p_3)$. Por tanto el conjunto $\mathcal{G}$ es generado por las funciones $g_0(x, y) = 1$, $g_1(x, y) = \|(x - y)/2\|^2$, $g_i^{(k)}(x, y) = \{(x_i + y_i)/2\}^d$ for $i \in \{1, 2, 3\}$ y $k \in \{1, 2\}$.

En el caso del toro, si d_{S_2} es la distancia geodésica en S_2 y $x = (x^{(1)}, x^{(2)}), y = (y^{(1)}, y^{(2)}) \in S_2 \times S_2$. Entonces se cumple que $(x, y) \in A_p$ si se cumple la siguiente desigualdad,

$$g_p(x, y) := d^2(x, y)/4 - d^2_{S_2}\{p^{(1)}, (xy/2)^{(1)}\} - d^2_{S_2}\{p^{(2)}, (xy/2)^{(2)}\} \geq 0.$$

Dada la relación existente en la esfera entre la medida de una cuerda y el arco geodésico minimizante

$$d_{S_2}(x^{(k)}, y^{(k)}) = Kd(x^{(k)}, y^{(k)}),$$

donde K es una constante conocida. Es claro, con argumentos similares a los usados en la esfera, que el conjunto $\mathcal{G}$ es de dimensión VC finita en el toro. El razonamiento en el cilindro es completamente análogo.

Comencemos probando la convergencia uniforme de la versión empírica a la versión poblacional. Fijado $p \in \mathcal{M}$, el estimador propuesto $\widehat{\mathrm{BD}}_n(p)$ en (3.1) es insesgado y la consistencia fuerte se deduce del hecho que el núcleo del U–estadístico $h(x, y) = \mathbf{1}_{B(x,y)}(p)$ es acotado (ver Serfling (1980), pág. 201). La convergencia uniforme es enunciada en el siguiente Teorema.

Teorema 3.2.3 *Sea $\mathcal{M}$ una variedad Riemanniana que verifica la condición **VC**. Entonces, cuando $n \to \infty$, se cumple que*

$$\sup_{p \in \mathcal{M}} |\widehat{\mathrm{BD}}_n(p) - \mathrm{BD}(p)| \to 0 \quad c.s.$$

Demostración.

Dado $\epsilon > 0$, a partir del Teorema 3.2.1, para M y n lo suficientemente grandes se cumple que

$$\sup_{d_g(x,q) \geq M} |\widehat{\mathrm{BD}}_n(x) - \mathrm{BD}(x)| \leq \sup_{d_g(x,q) \geq M} |\widehat{\mathrm{BD}}_n(x)| + \sup_{d_g(x,q) \geq M} |\mathrm{BD}(x)| < \epsilon/3 \quad \text{c.s.}$$

Consideramos ahora el conjunto compacto $\bar{B}(q, M)$, podemos elegir $\delta > 0$ lo suficientemente pequeño y utilizar el Teorema 3.2.2. Puesto que $\mathcal{M}$ es separable, podemos encontrar un conjunto finito $D := \{x_1, \ldots, x_m\} \subset \bar{B}(q, M)$

tal que $\cup_{j=1}^{m} B(x_j, \delta) \supset \bar{B}(q, M)$. Para $x \in B(x_j, \delta)$ se cumple que,

$$
\begin{aligned}
&|\widehat{\mathrm{BD}}_n(x) - \mathrm{BD}(x)| \leq \\
&\quad |\widehat{\mathrm{BD}}_n(x) - \widehat{\mathrm{BD}}_n(x_i)| + |\widehat{\mathrm{BD}}_n(x_i) - \mathrm{BD}(x_i)| + |\mathrm{BD}(x_i) - \mathrm{BD}(x)| \leq \\
&\quad 2\epsilon + R_n + \sup_{y \in D} |\widehat{\mathrm{BD}}_n(y) - \mathrm{BD}(y)|.
\end{aligned}
$$

De este hecho podemos deducir entonces que,

$$
\sup_{x \in \bar{B}(q,M)} |\widehat{\mathrm{BD}}_n(x) - \mathrm{BD}(x)| \leq 2\epsilon + R_n + \sup_{y \in D} |\widehat{\mathrm{BD}}_n(y) - \mathrm{BD}(y)|.
$$

Ahora, puesto que el núcleo del U–estadístico de orden 2 toma valores entre 0 y 1, a partir de la desigualdad de Hoeffding, podemos afirmar que para $s > 0$ y $n > 2$,

$$
P\{|\widehat{\mathrm{BD}}_n(y) - \mathrm{BD}(y)| > s\} \leq 2\exp(-ns^2/2).
$$

Entonces a partir de esta desigualdad y la subaditividad de la probabilidad concluímos que

$$
\begin{aligned}
&P\left\{\sup_{y \in D} |\widehat{\mathrm{BD}}_n(y) - \mathrm{BD}(y)| > s\right\} \leq \\
&\quad P\{(|\widehat{\mathrm{BD}}_n(y) - \mathrm{BD}(y)| > s\} \leq 2m\exp(-ns^2/2).
\end{aligned}
$$

Finalmente, el Lema de Borel–Cantelli implica que $\sup_{y \in D} |\widehat{\mathrm{BD}}_n(y) - \mathrm{BD}(y)|$ converge c.s. a 0. Por tanto, este resultado unido a la desigualdad deducida en el comienzo de la demostración, dan por demostrado el Teorema. □

La distribución asintótica de nuestro estimador $\widehat{\mathrm{BD}}_n$ puede ser derivada a través de los resultados para U–estadísticos obtenidos en Arcones and Gine (1993). Anotemos por

$D_x = \{(x_1, x_2) \in \mathcal{M} \times \mathcal{M} : x \in B_{x_1 x_2}\}$, $\mathcal{D} = \{D_x : x \in \mathcal{M}\}$ y $\ell^\infty(\mathcal{D})$ el conjunto de todas las funciones $f : \mathcal{D} \to \mathbb{R}$ acotadas. Consideremos en lo que sigue la profundidad empírica como un U–proceso indexado en los conjuntos $\mathcal{D}$, $\widehat{\mathrm{BD}}_n := \{U_2^n(D_x) : D_x \in \mathcal{D}\}$, es decir, para cada x, $U_2^n(D_x)$ es un U–estadístico.

Teorema 3.2.4 (Distribución asintótica) *Dada $\mathcal{M}$ una variedad Riemanniana compacta que verifica la condición **VC**. Sea $\{X_n : n \geq 1\}$ una su-*

cesión de variables aleatorias independientes copias de X y P a la distribución de X. Entonces, el proceso estocástico $n^{1/2}(\widehat{\mathrm{BD}}_n - \mathrm{BD})$ converge en ley a $2G_P$ en $\ell^\infty(\mathcal{D})$, donde G_P es el puente Browniano asociado a la medida P e indexado en la familia $\mathcal{D}$. Es decir, G_P es un proceso Gaussiano centrado con covarianzas

$\mathrm{E}\{G_P(D_x)G_P(D_y)\} = P_2(D_x \cap D_y) - P_2(D_x)P_2(D_y)$ *para todo* $D_x, D_y \in \mathcal{D}$*), en donde*

$$P_2(D) := \int_{\mathcal{M}\times\mathcal{M}} \mathbf{1}_D(x_1, x_2)\, dP(x_1) \times dP(x_2),$$

con $D \in \mathcal{D}$*.*

Demostración.

Por hipótesis sabemos que el conjunto $\mathcal{D}$ tiene dimensión VC finita. Por último, esta condición implica que se cumplen los supuestos de la Proposición 10 en Giné (1996) para el caso particular de un U–estadístico de orden 2. La distribución asintótica se derivada directamente entonces del Teorema 4.10 en Arcones and Gine (1993). □

Cabe observar que el concepto de convergencia en ley utilizado es en el sentido desarrollado en Hoffmann-Jørgensen (1991).

En particular, si fijamos un punto x diferente al punto más profundo, la distribución marginal del proceso límite $2G_P$ es dada por $2G_p(D_x) \sim \mathcal{N}[0, \sigma_d^2(x)]$ con $\sigma_d^2(x) = 4P_2(D_x)\{1 - P_2(D_x)\}$.

En el caso de que la variedad es $\mathbb{R}^d$, si $X \sim \mathcal{N}(0, I_d)$ y denotamos por

$$S(x, y) = \{t \in \mathbb{R}^d : \langle y - x, x - t\rangle \geq 0\},$$

entonces,

$$P_2(D_x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} g_x(y) e^{-\|y\|^2/2} dy,$$

donde

$$g_x(y) = P\{S(x, y)\} = \frac{1}{\sqrt{2\pi}} \int_{\langle x-y, x\rangle/\|x-y\|}^{+\infty} e^{-t^2/2} dt.$$

En la Figura 3.3 se representan las varianzas marginales $\sigma_d^2(\ell u)$ en función de $\ell \in \mathbb{R}$. En este caso u es cualquier vector unitario de $\mathbb{R}^d$. Se observa en la figura como la variabilidad aumenta cuando la dimensión del espacio crece.

Si la distribución es esférica respecto al punto más profundo c, entonces la distribución límite marginal en c se corresponde con el caso degenerado para U-estadísticos, es decir, $n\{\widehat{\mathrm{BD}}_n(c) - \mathrm{BD}(c)\}$ converge en ley a $\sum_{j=1}^{+\infty} \lambda_j(\chi_j^2 - 1)$

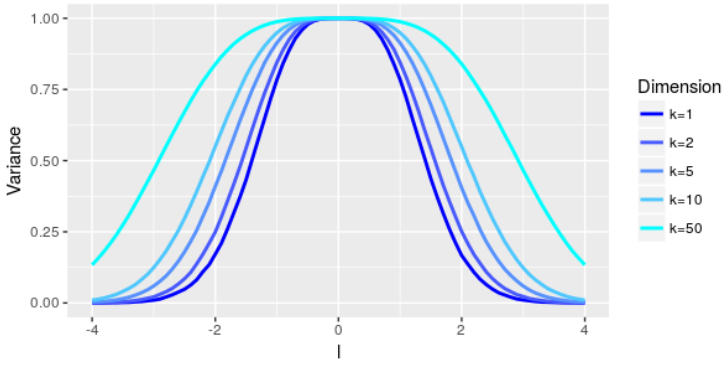

Figura 3.3: Gráficos de las funciones $\sigma_d^2(\ell u) = 4P_2(D_{\ell u})\{1 - P_2(D_{\ell u})\}$ cuando el parámetro ℓ varía en el intervalo $[-4, 4]$ y $d \in \{1, 2, 5, 10, 50\}$. El vector$u \in \mathbb{R}^d$ cumple que la $\|u\| = 1$ y $D_{\ell u} := \{(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d : B_{x_1 x_2} \ni \ell u\}$.

donde $\chi_1^2, \chi_2^2, \ldots$ son variables independientes con distribución Chi–cuadrado con un grado de libertad (ver pág. 195 en Serfling and Zuo (2000)).

3.2.3. Un resultado de consistencia de DCOPS en FDA

Consideremos en esta subsección que los elementos aleatorios se encuentran en un espacio de Hilbert separable $\mathcal{H}$. Como se muestra en el Teorema 3.2.5, en este caso la condición **VC** no es necesaria para demostrar la convergencia uniforme de nuestro estimador y la podemos reemplazar por una condición más débil que denominaremos condición HB.

> *Condición HB (B-continuidad)*: Dada una medida de probabilidad P en un espacio de Hilbert separable diremos que satisface la condición HB si para todo $r > 0$ e $y \in \mathcal{H}$ se cumple que $P\{\partial B(y, r)\} = 0$. En este caso denotamos por ∂A a la frontera en el sentido topológico del conjunto A.

Teorema 3.2.5 *Dada una sucesión de elementos aleatorios $\{X_n : n \geq 1\}$ que toma valores en espacio de Hilbert separable $(\mathcal{H}, \|\cdot\|)$. Sea P una medida de probabilidad que verifica la condición HB y P_n la distribución empírica de $\{X_1, \ldots, X_n\}$. Si se cumple que $P_n \to P$ débilmente cuando $n \to \infty$, entonces podemos afirmar que*

$$\sup_{p \in \mathcal{H}} |\widehat{\mathrm{BD}}_n(p) - \mathrm{BD}(p)| \to 0 \quad c.s.$$

La demostración del Teorema anterior se basa en un interesante resultado expuesto en Billingsley and Topsøe (1967), que además es cierto en espacios más generales como los espacios de Banach (ver Teorema 1 y el Ejemplo 3 en Billingsley and Topsøe (1967)). Enunciamos dicho resultado en el Teorema expuesto a continuación.

Teorema 3.2.6 ***(Billingsley y Topsøe).*** *Sea S un espacio métrico separable y $\mathcal{B}(S)$ la familia de todas la funciones reales, acotadas y medibles definidas sobre S. Sea además $\mathcal{F} \subset \mathcal{B}(S)$ un subclase de funciones de $\mathcal{B}(S)$. Entonces, cuando $n \to \infty$, podemos afirmar que*

$$\sup_{f\in\mathcal{F}} \left| \int f dP_n - \int f dP \right| \to 0,$$

para toda sucesión de medidas probabilidad P_n que converjan débilmente a P, sí y sólo sí $\sup\{|f(z) - f(t)| : f \in \mathcal{F}, z, t \in S\} < \infty$, y para todo $\epsilon > 0$, se cumple que

$$\lim_{\delta\to 0} \sup_{f\in\mathcal{F}} P\left[\{x : \omega_f\{B(x,\delta)\} > \epsilon\}\right] = 0,$$

donde $\omega_f(A) = \sup\{|f(x) - f(y)| : x, y \in A \subset S\}$ y $B(x,\delta)$ es la bola abierta de centro x y radio $\delta > 0$.

Demostración del Teorema 3.2.5.

Comencemos por considerar el espacio producto de espacios de Hilbert $\mathcal{H}^2 := \mathcal{H} \times \mathcal{H}$ y el subconjunto de funciones,

$$\mathcal{F} = \{f_x : x = (x_1, x_2) \in \mathcal{H}^2\} \text{ donde } f_x(z) = \mathbf{1}_{\{\langle x_1 - z, x_2 - z\rangle \leq 0\}}.$$

Es suficiente probar el Teorema para el siguiente V-estadístico asociado,

$$\widetilde{\mathrm{BD}}_n(z) := \frac{1}{n^2} \sum_{1\leq i_1, i_2 \leq n} \mathbf{1}_{B_{X_{i_1} X_{i_2}}}(z) = (1 - 1/n)\widehat{\mathrm{BD}}_n(z).$$

Es evidente que

$$\sup\{|f_x(z) - f_x(t)| : x \in \mathcal{H}^2 \text{ y } z, t \in \mathcal{H}\} \leq 1 < \infty.$$

Por otro lado, dado $\epsilon > 0$, se puede observar

$$\left|\mathbf{1}_{\{\langle t_1' - y, t_2' - y\rangle \leq 0\}} - \mathbf{1}_{\{\langle t_1 - y, t_2 - y\rangle \leq 0\}}\right| > \epsilon \Leftrightarrow \langle t_1' - y, t_2' - y\rangle \langle t_1 - y, t_2 - y\rangle \leq 0.$$

Dado $x := (x_1, x_2) \in \mathcal{H}^2$, $\|(t_1, t_2)\|_{\mathcal{H}^2} := \max(\|t_1\|_{\mathcal{H}}, \|t_2\|_{\mathcal{H}})$ y (t_1, t_2), $(t_1^{'}, t_2^{'}) \in B\{(x_1, x_2), \delta\}$ la bola en $\mathcal{H}^2$ con centro en (x_1, x_2) y radio δ con respecto a la distancia inducida por la norma $\|\cdot\|_{\mathcal{H}^2}$. Consideremos el caso donde $\langle t_1^{'} - y, t_2^{'} - y\rangle \geq 0$ y $\langle t_1 - y, t_2 - y\rangle \leq 0$, puesto que el otro caso $\langle t_1^{'} - y, t_2^{'} - y\rangle \leq 0$ y $\langle t_1 - y, t_2 - y\rangle \geq 0$ se deduce de manera análoga. Observemos que los puntos $(t_1 + t_2)/2$ y $(t_1^{'} + t_2^{'})/2$ se encuentran en la bola $B\{(x_1 + x_2)/2, \delta\}$. Ahora, si $\mathcal{H}$ es un espacio de Hilbert, entonces para todo $y, t_1, t_2 \in \mathcal{H}$ se cumple que $y \in B_{t_1 t_2} \Leftrightarrow \langle t_1 - y, t_2 - y\rangle \leq 0$. Por lo tanto,

$$y \in B\{(t_1 + t_2)/2, \|t_1 - t_2\|/2\} \cap B^c\{(t_1^{'} + t_2^{'})/2, \|t_1^{'} - t_2^{'}\|/2\}.$$

Además, de la desigualdad

$$\|t_1 - t_2\|/2 \leq \|t_1 - x_1\|/2 + \|x_1 - x_2\|/2 + \|x_2 - t_2\|/2 \leq \|x_1 - x_2\|/2 + \delta,$$

se desprende que

$$\|y - (x_1 + x_2)/2\| \leq \|y - (t_1 + t_2)/2\| + \|(t_1 + t_2)/2 - (x_1 + x_2)/2\| \leq \|(t_1 - t_2)/2\| + \delta \leq \|x_1 - x_2\|/2 + 2\delta.$$

En consecuencia, $y \in B\{(x_1 + x_2)/2, \|x_1 - x_2\|/2 + 2\delta\}$. De manera análoga, tenemos que $\|t_1^{'} - t_2^{'}\|/2 \geq \|x_1 - x_2\|/2 - \delta$ y por lo tanto $y \notin B\{(x_1 + x_2)/2, \|x_1 - x_2\|/2 - 2\delta\}$, para δ lo suficientemente pequeño.

Consideremos un conjunto compacto $K_\gamma \in \mathcal{H}$ de manera que si definimos $X = (X_1, X_2)$, se cumpla que $P\{X \in (K_\gamma \times K_\gamma)^c\} < \gamma$.

Ahora sea $M = (X_1 + X_2)/2$ y $R = \|X_1 - X_2\|/2$, ahora tenemos que

$$\lim_{\delta \to 0} \sup_y P\left(\sup_{t, t' \in B(x, \delta)} \left|\mathbf{1}_{\{\langle t_1^{'} - y, t_2^{'} - y\rangle \leq 0\}} - \mathbf{1}_{\{\langle t_1 - y, t_2 - y\rangle \leq 0\}}\right| > \epsilon\right) \leq \lim_{\delta \to 0} \sup_y P\{y \in B(M, R + 2\delta) \setminus B(M, R - 2\delta)\},$$

y el término derecho puede ser acotado por

$$\lim_{\delta \to 0} \sup_y P\{y \in B(M, R + 2\delta) \setminus B(M, R - 2\delta), X \in K_\gamma \times K_\gamma\} + \gamma.$$

Puesto que K_γ es un conjunto compacto, entonces existe $L > 0$ tal que $K_\gamma \subset B(0, L)$. Además sea $K_{\gamma,L}$ un conjunto compacto tal que $K_\gamma \subset K_{\gamma,L} \ominus B(0, 2L+1)$ en donde $K_{\gamma,L} \ominus B(0, 2L+1) := \{z \in K_{\gamma,L} : d(z, K_{\gamma,L}^c) > 2L+1\}$.

Llamamos ψ_δ a la función $\psi_\delta(y) = P\{y \in B(M, R+2\delta) \setminus B(M, R-2\delta), X \in K_\gamma \times K_\gamma\}$. Si $y \in K_{\gamma,L}^c$, entonces se verifica

$$\begin{aligned} d\{(x_1+x_2)/2, y\} &\geq d(y, x_1) - d\{x_1, (x_1+x_2)/2\} \geq L+1 \geq \\ &\geq L + 2\delta \geq \|x_1 - x_2\|/2 + 2\delta, \end{aligned}$$

para todo $x_1, x_2 \in K_\gamma$, lo cuál implica que $\psi_\delta(y) = 0$ para todo $y \in K_{\gamma,L}^c$. Por lo tanto se cumple que $\sup_{y \in \mathcal{H}} \psi_\delta(y) = \sup_{y \in K_{\gamma,L}} \psi_\delta(y) = \psi_\delta(y_\delta^*)$ para todo $\gamma > 0$, con y_δ^* el punto donde se alcanza el máximo en el conjunto compacto $K_{\gamma,L}$.

Para finalizar la demostración alcanzaría mostrar que para todo γ fijo se cumple que $\lim_{\delta \to 0} \psi_\delta(y_\delta^*) = 0$.

Supongamos que esta premisa fuera falsa, es decir, que existe $\eta > 0$, $y_n \in K_{\gamma,L}$ y $\delta_n \to 0$ tal que

$$\psi_{\delta_n}(y_n) = P\{M \in B(y_n, R+2\delta_n) \setminus B(y_n, R-2\delta_n), X \in K_\gamma \times K_\gamma\} > \eta.$$

Puesto que el conjunto $[0,1] \times K_{\gamma,L}$ es compacto existe una subsucesión convergente, que denotaremos (δ_n, y_n), es decir, $(\delta_n, y_n) \to (0, y)$ para algún $y \in K_{\gamma,L}$. Además se cumple que

$$\psi_{\delta_n}(y_n) = \mathbb{E}[\mathbb{E}\{\mathbf{1}_{\{B(y_n, R+2\delta_n) \setminus B(y_n, R-2\delta_n)\}}(M) \mathbf{1}_{\{K_\gamma \times K_\gamma\}}(X) | R = r\}],$$

entonces, a partir de la condición HB, el Teorema de Convergencia Dominada y la continuidad de la función ψ_δ se deduce que

$$\begin{aligned} P\{M \in B(y_n, R+2\delta_n) \setminus B(y_n, R-2\delta_n), X \in K_\gamma \times K_\gamma\} \to \\ \to \mathbb{E}\{P(M \in B(y, R), X \in K_\gamma \times K_\gamma \big| R = r)\} = 0, \end{aligned}$$

cuando $n \to \infty$, lo cúal lleva a una contradicción y el Teorema es demostrado. □

De este Teorema anterior podemos deducir de manera inmediata los si-

guientes resultados en espacios de Hilbert.

Corolario 3.2.1 *(Consistencia del punto más profundo) Bajo las hipótesis del Teorema 3.2.5, si* $\mathrm{BD}(p)$ *tiene un único mínimo bajo* P*, entonces*

$$\hat{\theta}_n = arg\ \underset{p}{máx}\widehat{\mathrm{BD}}_n(p)\ converge\ a\ \ \theta = arg\ \underset{p}{máx}\mathrm{BD}(p)\ \ c.s.$$

Corolario 3.2.2 *(Robustez del punto más profundo) Bajo las hipótesis del Corolario 3.2.1,* $\hat{\theta}_n$ *es un estimador cualitativamente robusto en el sentido desarrollado en Hampel (1971).*

3.3. Algunos ejemplos simulados

Al principio de esta sección presentaremos simulaciones en dos posibles escenarios cuando los datos pertenecen a una variedad Riemanniana. Analizaremos datos sobre una esfera y sobre el cono de matrices definidas positivas. Por último analizaremos la performance de la profundidad esférica en un espacio euclídeo cuando la dimensión del espacio es elevada.

3.3.1. Datos simulados en la esfera

Consideremos una muestra iid de tamaño 100 en la esfera $S_3 = \{x \in \mathbb{R}^3 : \|x\| = 1\}$ con la distribución de von Mises–Fisher determinada por la densidad,

$$f(x, \mu, \kappa) = C(x)e^{\kappa\mu^\top x}\mathbf{1}_{S_3}(x),$$

donde $\kappa \geq 0$ y $\mu \in S_3$ son los denominados parámetros de concentración y media direccional respectivamente, donde $C(x)$ es la constante de normalización (ver Mardia and Jupp (2000)). En la Figura 3.4 representamos la profundidad empírica normalizada de 100 puntos generados con $\kappa = 15$ y $\mu = (1, 0, 0)$, a un incremento en la intensidad del color rojo mayor es la profundidad. Observamos que la mayores profundidades se encuentran entorno a la media direccional y decrece hacia el punto antipodal en la esfera.

Con la misma muestra se representa en la misma figura la profundidad empírica angular de Tukey (abreviaremos ATD) introducida en Liu and Singh (1992). Se observa que las profundidades empíricas calculadas sobre los valores de la muestra y sobre los puntos de la esfera presentan los mismos patrones. Sin

embargo, los tiempos computacionales de cálculo en la estimación de DCOPS son considerablemente menores.

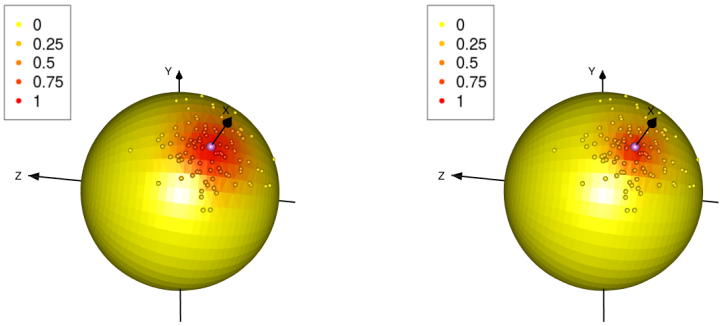

Figura 3.4: Profundidad empírica para una muestra de tamaño 100 de datos simulados a partir de la distrbución von Mises–Fisher con parámetros $\kappa = 15$ y $\mu = (1, 0, 0)$. Intensidades altas de colores rojo indican profundidades empíricas elevadas. Panel–Izquierdo: DCOPS empírica normalizada. Panel–Derecho: ATD empírica normalizada.

Estudiamos ahora el mismo ejemplo pero con presencia de outliers. Para ello simulamos 120 puntos de un mezcla de distribuciones von Mises–Fisher dada por

$$f(x) = 0.9f(x, (1, 0, 0), 10) + 0.1f(x, (0, 0, 1), 50).$$

En la Figura 3.5, al igual que en la Figura 3.4 se representan los datos simulados con las respectivas profundidades empíricas determinadas mediante DCOPS Y ATD.

Ambas profundidades muestran un comportamiento similar, asignando profundidades empíricas bajas a los outliers. Con esto queremos mostrar que en aquellos escenarios donde ya hay medidas de profundidad definidas DCOPS puede ser un posible competidor.

3.3.2. Ejemplo de datos simulados sobre el cono de las matrices definidas positivas

Varios problemas de estadística tienen como datos un conjunto de matrices definidas positivas, por ejemplo, métodos robustos para la estimación de

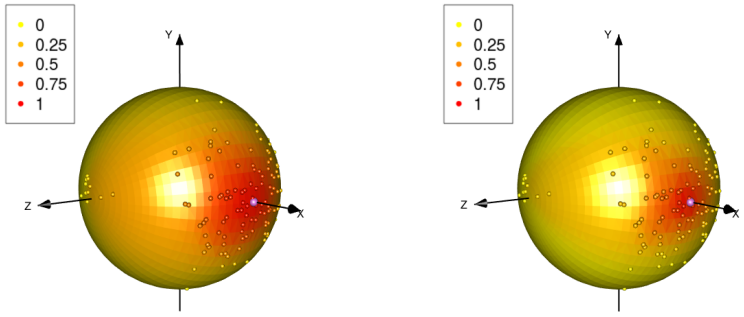

Figura 3.5: Profundidades empíricas en la esfera a partir de una muestra de tamaño 120 de una mezcla de distribuciones von Mises–Fisher. Intensidades altas de color rojo rojo indica una mayor profundidad empírica. Panel–Izquierdo: DCOPS empírica normalizada. Panel–Derecho: ATD empírica normalizada.

las matrices de varianzas y covarianzas, componentes principales, métodos de aprendizaje automático (ver por ejemplo Chen et al. (2018) y Croux et al. (2017)). En este capítulo anotemos por $(\mathbb{P}_d, g)$ la variedad Riemanniana de las matrices definidas positivas. Dadas dos matrices $A, B \in \mathbb{P}_d$ existe una única geodésica determinada por A y B (ver Moakher (2005)),

$$\gamma(s) := A^{1/2}(A^{-1/2}BA^{-1/2})^s A^{1/2}.$$

Se puede deducir el punto medio y la distancia geodésica entre A y B que denotaremos por $A\#B$ y $d_g(A, B)$ respectivamente,

$$A\#B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2} \quad \text{y} \quad d_g(A, B) = \|\ln(A^{-1/2}BA^{-1/2})\|,$$

donde $\|\cdot\|$ es la norma de Hilbert–Schmidt.

Consideramos la distribución de Wishart $\mathcal{W}_3(\Sigma, m)$ en $\mathbb{P}_3$ con parámetros $m = 20$ y $\Sigma = I_3$. Una matriz aleatoria S con distribución de Wishart con parámetros Σ y m puede ser generada a través de una muestra multivariada de un vector Gaussiano $\mathcal{N}(0, \Sigma)$ a partir de $S = X_1X_1' + \cdots + X_mX_m'$. El valor esperado de la distribución de Wishart esta dada por $\mathbb{E}(S) = m \times \Sigma$.

A partir de dicha distribución se genera una muestra de 100 matrices. Para cada valor de la muestra se determina su profundidad empírica $\widehat{\text{BD}}_n$ y

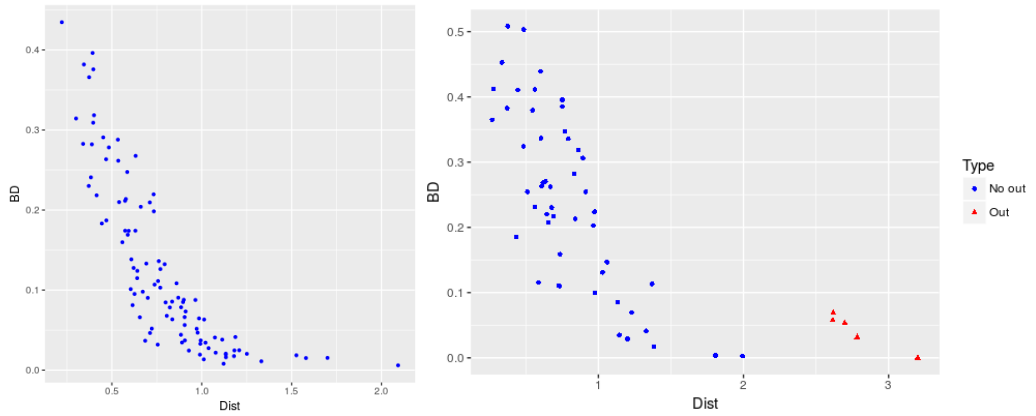

Figura 3.6: Panel Izquierdo: Gráfico de $\widehat{\mathrm{BD}}_n$ en función de la distancia geodésica al valor esperado en un muestra de tamaño 100 a partir de una distribución $\mathcal{W}_3(I_3, 20)$. Panel Derecho: Gráfico de $\widehat{\mathrm{BD}}_n$ en función de la distancia geodésica al valor esperado (de la distribución no contaminada) en un muestra de tamaño 120 a partir de la distribución $0.9\mathcal{W}_3(I_3, 20) + 0.1\mathcal{W}_3(I_3/10, 50)$.

la distancia geodésica al valor esperado de la distribución. Los resultados son mostrados en la Figura 3.6. Se puede observar como la profundidad empírica decrece a medida que no alejamos del valor esperado.

Ahora tomamos un muestra contaminada de tamaño 120 generada a partir de una mezcla de distribuciones Wishart,

$$0.9\mathcal{W}_3(I_3, 20) + 0.1\mathcal{W}_3(I_3/10, 50).$$

En la Figura 3.6 se muestra como $\widehat{\mathrm{BD}}_n$ asigna a los outliers profundidades bajas.

3.3.3. Datos simulados en $\mathbb{R}^d$ con $d = 5$ y $d = 20$

Como mencionamos anteriormente para valores elevados de d, las estimaciones de la profundidad simplicial y la profundidad por semi–espacios son de una alto costo computacional e imposibles de calcular en tiempos razonables. Por tanto, en esta sección se necesitan otras medidas de profundidad para poder comparar con DCOPS. En tal sentido la comparación será realizada con dos medidas de profundidad introducidas en Liu (1992) y en Cuevas and Fraiman (2009) respectivamente. Ambas nociones de profundidad se basan en el concepto de proyecciones unidimensionales.

Sea $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^d$ una muestra iid con la misma distribución que $\mathbf{X}$.

En Liu (1992), dada una dirección $\mathbf{u} \in \mathbb{R}^d$ ($\|\mathbf{u}\| = 1$), la muestra se proyecta ortogonalmente en dicha dirección obteniendo así una muestra unidimensional, $\aleph_{n,u} = \{X_{i,u} = \langle \mathbf{X}_i, \mathbf{u} \rangle \text{ para todo } i \in \{1, \ldots, n\}\}$.

Un medida de "outlieriedad" para $\mathbf{x} \in \mathbb{R}^d$ es dada por

$$\mathrm{OU}_n(\mathbf{x}) := \sup_{\|\mathbf{u}\|=1} \frac{|\langle \mathbf{x}, \mathbf{u} \rangle - \mu_{\aleph_{n,u}}|}{\tau_{\aleph_{n,u}}}, \tag{3.3}$$

en donde $\mu_{\aleph_{n,u}}$ y $\tau_{\aleph_{n,u}}$ son la mediana y la desviación mediana de la muestra proyectada $\aleph_{n,u}$ respectivamente (ver pág. 5 en Maronna et al. (2006)).

Obtenemos la versión poblacional remplazando en la ecuación (3.3) $\mu_{\aleph_{n,u}}$ y $\tau_{\aleph_{n,u}}$ por μ_u y τ_u, siendo μ_u y τ_u la mediana y la desviación mediana de la variable $\langle X, \mathbf{u} \rangle$.

En Liu (1992) se sugiere utilizar como una medida de profundidad,

$$\mathrm{PD}_{1,n}(x) := \{1 + \mathrm{OU}_n(x)\}^{-1}.$$

En la práctica podemos obtener una aproximación de $\mathrm{PD}_{1,n}(x)$ tomando el máximo sobre un número elevado de direcciones elegidas al azar sobre la esfera unidad.

La segunda noción de profundidad es introducida para datos funcionales en Cuevas and Fraiman (2009), pero es plausible su cálculo en datos multivariados, y en particular si la dimensión del espacio es elevada. Si denotamos por $g_{\mathbf{u}}(x) = \langle x, \mathbf{u} \rangle$, entonces la profundidad es definida como

$$\mathrm{PD}_2(x) := \int_{S_d} F_{g_{\mathbf{u}}(X)}\{g_{\mathbf{u}}(x)\}[1 - F_{g_{\mathbf{u}}(X)}\{g_{\mathbf{u}}(x)\}]d\mathbf{u},$$

donde $F_{g_{\mathbf{u}}(X)}$ es la distribución acumulada de la variable aleatoria $g_{\mathbf{u}}(X)$. Esta profundidad puede ser estimada a través de un promedio sobre K direcciones elegidas al azar en la esfera unidad,

$$\mathrm{PD}_{2,n}(x) := \frac{1}{K} \sum_{i=1}^{K} \hat{F}_{g_{\mathbf{u}_i}}\{g_{\mathbf{u}_i}(x)\}[1 - \hat{F}_{g_{\mathbf{u}_i}}\{g_{\mathbf{u}_i}(x)\}],$$

donde $\hat{F}_{g_{\mathbf{u}_i}}$ es la distribución empírica de la muestra $g_{\mathbf{u}_i}(X_1), \ldots, g_{\mathbf{u}_i}(X_n)$.

En primera instancia, tomamos una muestra de tamaño 1000 en $\mathbb{R}^d$ para $d \in \{5, 20\}$ generada de una distribución normal $\mathcal{N}(0, I_d)$. En la Figura 3.7 se representa la profundidad de los puntos que se encuentran en el rayo $y =$

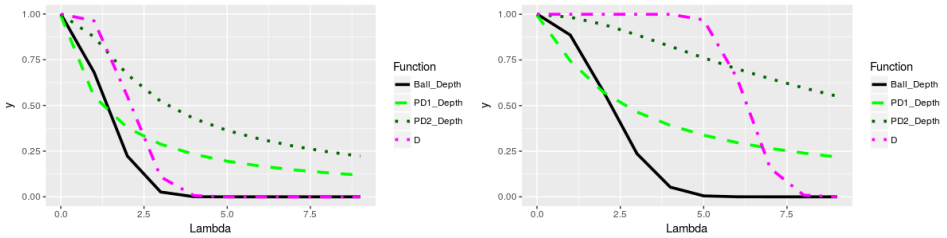

Figura 3.7: Gráfica de las profundidades empíricas normalizadas $\mathrm{PD}_{1,n}, 4\mathrm{PD}_{2,n}, 2\mathrm{BD}_n$ y de la probabilidad D en $\mathcal{N}(0, I_d)$. Las profundidades son calculadas sobre el rayo $y = \lambda(1, 0, 0, \ldots, 0) \in \mathbb{R}^d$, $\lambda > 0$ y representadas en función de la distancia del punto al origen. Panel Izquierdo: $d = 5$ y $n = 1000$. Panel Derecho : $d = 20$ y $n = 1000$

$\lambda(1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ con $\lambda > 0$ de las estimaciones de las profundidades (normalizadas) $\mathrm{PD}_{1,n}$, $4\mathrm{PD}_{2,n}$ y $2\mathrm{BD}_n$ en función de la distancia euclídeana al origen. En el mismo gráfico se representa la función $D(\lambda) = P\left(\|X\| > \lambda\right)$ con $X \sim \mathcal{N}(0, I_d)$. En ambos casos se puede observar el buen comportamiento de la versión empírica de DCOPS, puesto que a medida que el punto se aleja del origen a través del rayo la profundidad decrece rápidamente a cero.

Datos contaminados

Al igual que en simulaciones anteriores mediante una mezcla contaminamos la muestra,

$$(1 - \alpha)\mathcal{N}_2\left(\mu_1, \Sigma_1\right) + \alpha\mathcal{N}_2\left(\mu_2, \Sigma_2\right), \tag{3.4}$$

Se genera una muestra de tamaño 5000 a partir de la distribución (3.4) con $\mu_1 = (0, \ldots, 0)$, $\mu_2 = 2(1, \ldots, 1) \in \mathbb{R}^{10}$, $\Sigma_1 = \Sigma_2 = I_d$ y $\alpha = 0.1$. En la Figura 3.8 se representan las profundidades empíricas normalizadas $\mathrm{PD}_{1,n}$, $4\mathrm{PD}_{2,n}$ y $2\mathrm{BD}_n$ de 100 puntos generados con la misma distribución en función de la distancia λ al origen. Los outliers son marcados en el gráfico con un círculo rojo. Se observa en dicha figura como la versión empírica BD_n asigna a los outliers profundidades moderadas y en tal sentido tiene un mejor desempeño que sus competidores.

3.4. Conclusiones del capítulo

En el capítulo se extiende una noción de profundidad a variedades Riemannianas con las siguientes cualidades

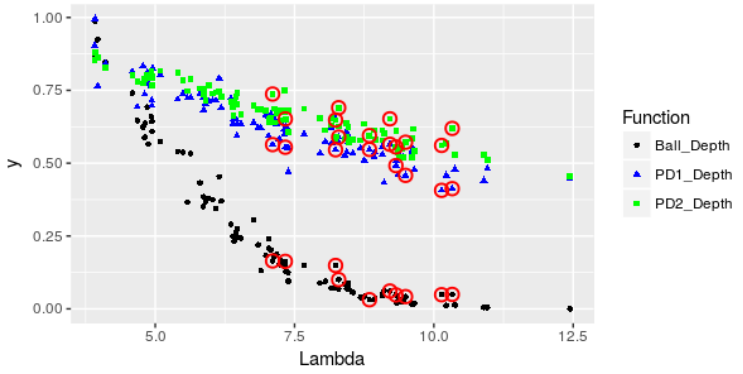

Figura 3.8: Gráfica de las profundidades empíricas normalizadas $\mathrm{PD}_{1,n}$, $4\mathrm{PD}_{2,n}$, $2\mathrm{BD}_n$ a partir de una muestra de tamaño 5000 de la distribución 3.4 evaluada en 100 puntos simulados con la misma distribución. Los outliers son marcados con un círculo rojo.

- En espacios euclídeos de dimensión finita verifica aquellas propiedades deseables que una medida de profundidad debe tener.

- Estas propiedades deseables se conservan también en espacios de Hilbert. En este marco la consistencia fuerte del estimador es demostrada.

- En el caso que el espacio es una variedad Riemanniana son probadas aquellas propiedades que tiene sentido en este paradigma. También se demuestra una ley fuerte y un Teorema Central del Límite para la versión empírica de la profundidad estudiada.

- Para analizar las bondades de la profundidad son realizadas simulaciones en diversas variedades Riemannianas (como por ejemplo, la esfera, el toro, y el cono de las matrices definidas positivas).

- Por último, a través de simulaciones, se realiza el cálculo del estimador en espacios de dimensión elevada en tiempos admisibles (la complejidad computacional es del orden de $d \times n^2$, donde d es la dimensión del espacio y n el tamaño de la muestra). Si embargo lo que consume tiempo computacional es el cálculo de la geodésica, pero consideramos que dicho enfoque es el apropiado en este paradigma.

- Es importante destacar que existen otras medidas de profundidad propuestas en espacios métricos separables (ver por ejemplo Carrizosa (1996) y Liu and Modarres (2011)). Sin embargo, en general las propiedades

deseables de toda profundidad no son formalmente probadas. Se pospone para un trabajo a futuro estudiar la consistencia, un TCL para estos estimadores y un estudio de simulación donde se comparen estas profundidades con DCOPS. En esta dirección, en Shahsavarifar and Bremner (2018) realiza un estudio de simulación de estos estimadores pero en $\mathbb{R}^d$.

Capítulo 4

Sensibilidad en variedades Riemannianas

Este capítulo de la tesis se enfoca sobre el estudio de la influencia o efecto en términos globales de una cierta entrada sobre la salida de un modelo matemático o de un código computacional. El trabajo se centra en medir esta influencia cuando la salida pertenece a una variedad Riemanniana dada. El estudio de esta influencia es denominado *Análisis global de la sensibilidad* (AS en lo que sigue). Reseñas recientes en esta área son dadas en Kucherenko (2005) y Iooss and Lemaître (2015).

Son diversas las aplicaciones del análisis de sensibilidad a diversos modelos relacionados por ejemplo a la física, la industria e incluso a modelos ambientales (ver Saltelli et al. (2000), Rocquigny et al. (2008) y Pianosi et al. (2016) respectivamente).

El concepto de sensibilidad surge en primera instancia mediante un enfoque local. Bajo perturbaciones locales de una entrada alrededor de un valor predeterminado x^* de la entrada, en general la media, se evalúa el impacto en el modelo mediante la estimación de la derivada parcial. Es decir, dado el modelo

$$Z = f(X), \tag{4.1}$$

en donde $f : \mathbb{R}^d \to \mathbb{R}$ es una función desconocida y $X = (X_1, \ldots, X_d)$ un vector aleatorio de $\mathbb{R}^d$. El análisis de sensibilidad local busca una estimación de $\left(\frac{\partial f}{\partial x}\right)_{X=x^*}$. Una aplicación de esta metodología a modelos climáticos es dada por ejemplo en Cacuci (1981). Los modelos locales de sensibilidad en general

se basan en supuestos de linealidad y de normalidad, además de sólo medir la influencia de la entrada en términos locales. A principios de los años 80, con el fin de prescindir de estas restricciones, surgen los índices de sensibilidad denominados globales. Estos brindan un indicador del efecto o influencia de uno o algunos de los factores de entrada sobre la salida, incluso al variar los factores restantes. Este análisis tiene por objetivo, como se señala en Saltelli et al. (2000) y Iooss and Lemaître (2015), por un lado identificar y priorizar aquellas entradas mas influyentes y por otro obtener modelos más simples, fijando las entradas de menor influencia en un valor fijo nominal. En la sección 4.1 se explicitan dos tipos de índices globales que se denominan según su construcción en,

- Índices basados en los momentos (en la varianza),
- Índices independiente de los momentos (basados en distancias sobre las distribuciones).

En la sección 4.2 se propone, basados en un índice del tipo Cramér–von Mises, un nuevo índice cuya salida se encuentra sobre una variedad Riemanniana.

4.1. Algunos índices globales de sensibilidad

En esta sección se describen los índices globales de sensibilidad más clásicos en la literatura.

4.1.1. Índice de Sobol

El índice de Sobol (ver Sobol (1993)) es unos de los índices mas utilizados en la literatura. La idea es cuantificar el AS mediante los momentos de segundo orden del sistema conformado por la variables de entrada y de salida. Este se basa en la descomposición de Hoeffding (ver Hoeffding (1948) y Owen (1994)) por la cual la varianza de la salida puede ser descompuesta de manera única en términos de dimensión creciente bajo condiciones de independencia en las variables de entrada. El primer aporte en dicha descomposición ANOVA es realizada por Hoeffding en 1948 para U–estadísticos y su extensión al caso

funcional es desarrollada en Antoniadis (1984) y Sobol (1993). Dado un vector aleatorio X en $\mathbb{R}^d$ y P_X su medida asociada de probabilidad denotaremos

$$\mathbb{L}^2\left(\mathbb{R}^d, P_X\right) := \{f : \mathbb{R}^d \to \mathbb{R} / \int f^2 dP_X < \infty\}$$

Teorema 4.1.1 (Descomposición de Sobol-Hoeffding) *Sea $Y = f(X)$ con $X = (X_1, \ldots, X_d)$ un vector aleatorio y $f \in \mathbb{L}^2\left(\mathbb{R}^d, P_X\right)$. Se supone además que las variables de $X_1, \ldots, X_p$ son independientes, es decir, la medida de probabilidad de P_X es la medida producto ($P_X = P_{X_1} \times \ldots, \times P_{X_d}$). Entonces la función f puede ser descompuesta de manera única en una suma de funciones de dimensión creciente,*

$$\begin{aligned} f(X_1, \ldots, X_d) &= f_0 + \sum_{i=1}^{d} f_i(X_i) + \sum_{1 \le i < j \le d} f_{i,j}(X_i, X_j) + f_{1,\ldots,d}(X_1, \ldots, X_d) \\ &= \sum_{\nu \subset \{1,\ldots,d\}} f_\nu(X^\nu) \end{aligned} \tag{4.2}$$

en donde $X^\nu := \{X_i / i \in \nu\}$ y las funciones f_ν verifican las condiciones,

$$1)\mathbb{E}\left[f_\nu(X^\nu)\right] = \int f_\nu(x_\nu) dP_{X^\nu} = 0 \quad \forall i \in \nu, \forall \nu \subset \{1, \ldots, d\},$$

$$2)\mathbb{E}\left[f_\nu(X^\nu) f_\upsilon(X^\upsilon)\right] = \int f_\nu(x_\nu) f_\upsilon(x_\upsilon) dP_X = 0 \quad \forall \nu, \upsilon \subset \{1, \ldots, d\}, \nu \neq \upsilon.$$

Por ejemplo, si $d = 2$ se puede descomponer $f(X_1, X_2) = f_0 + f_1(X_1) + f_2(X_2) + f_{1,2}(X_1, X_2)$, con

$$\begin{aligned} &f_0 = \mathbb{E}\left[f(X_1, X_2)\right], \quad f_i(X_i) = \mathbb{E}\left[f(X_1, X_2)|X_i\right] - f_0, \ i = 1, 2. \\ &f_{1,2}(X_1, X_2) = f(X_1, X_2) - f_0 - f_1(X_1) - f_2(X_2) = \\ &= f(X_1, X_2) - \mathbb{E}\left[f(X_1, X_2)|X_1\right] - \mathbb{E}\left[f(X_1, X_2)|X_2\right] + \mathbb{E}\left[f(X_1, X_2)\right] \end{aligned}$$

Por la independencia de X_1 y X_2, a través de las propiedades de la esperanza condicional, son evidentes las condiciones de ortogonalidad,

$$\begin{aligned} &\mathbb{E}\left[f(X_1, X_2)\right] = \mathbb{E}\left[f(X_1, X_2) f_1(X_1)\right] = \mathbb{E}\left[f(X_1, X_2) f_2(X_2)\right] = \\ &= \mathbb{E}\left[f_1(X_2) f_2(X_2)\right] = \mathbb{E}\left[f_1(X_1)\right] = \mathbb{E}\left[f_2(X_2)\right] = 0. \end{aligned}$$

En general se tienen las definiciones explícitas para $f_0 = \mathbb{E}[f(X)]$ y $f_\nu = \mathbb{E}[f(X)|X_\nu] - \sum_{\upsilon \subsetneq \nu} f_\upsilon$, $|\nu| \geq 1$.

A partir de la ecuación (4.2) podemos descomponer la varianza de Y, y ver el aporte relativo de cada término en ella,

$$1 = \frac{\sum_{\nu \subset \{1,\dots,d\}} \mathrm{Var}[f_\nu(X^\nu)]}{\mathrm{Var}(Y)}.$$

Para cuantificar la contribución del conjunto de variables ν en las fluctuaciones de la salida Y se definen los índices de Sobol como

$$S_\nu := \frac{\mathrm{Var}[f_\nu(X^\nu)]}{\mathrm{Var}(Y)}.$$

Cuando $|\nu| = 1$ se denominan índices de Sobol de primer orden y en ese caso denotamos

$$S_i := \frac{\mathrm{Var}[\mathbb{E}(f(X)|X_i)]}{\mathrm{Var}[f(X)]},$$

y los índices de Sobol globales de cada variable como

$$S_i^{\mathrm{tot}} := \sum_{i \in \nu \subset \{1,\dots,d\}} \frac{\mathrm{Var}[f_\nu(X^\nu)]}{\mathrm{Var}[f(X)]}.$$

Este índice es generalizable cuando las entradas son variables dependientes (ver por ejemplo Chastaing et al. (2015)) pero en general los nuevos índices obtenidos son de compleja interpretación.

Desde un enfoque estadístico es importante brindar un estimador eficiente en términos computacionales de estos índices, así como también poder cuantificar la precisión de los mismos.

Varias metodologías se han utilizado para estimar los índices Sobol (para un enfoque "plug-in" ver por ejemplo Da Veiga et al. (2009) y Da Veiga and Gamboa (2013)).

En esta sección desarrollaremos el método denominado *Pick–Freeze* (ver Sobol (1993) y Sobol (2001)), que será de utilidad en el resto del capítulo.

El método consiste en escribir la varianza de la media condicional como una covarianza. Si llamamos X^ν al vector aleatorio que coincide con X en las ν componentes, es decir, $X_\nu^\nu = X_\nu$ y además $X_i^\nu = X_i'$ si $i \notin \nu$, donde X_i' es una copia de X_i independiente. Se anota

$$Y^\nu := f(X^\nu).$$

Si se cumple que $E(Y) = 0$, entonces,

$$\text{Var}\left(E(Y/X_\nu)\right) = E\left(E^2(Y/X_\nu)\right) = \text{cov}(Y, Y^\nu). \tag{4.3}$$

La igualdad anterior es debido a que Y y Y^ν son condicionalmente independientes respecto a X_ν,

$$\begin{aligned}\text{cov}(Y, Y^\nu) &= E(YY^\nu) - E(Y)E(Y^\nu) = E(YY^\nu) - E^2(Y) = \\ &= E\left[E(YY^\nu/X_\nu)\right] = E\left[E(Y/X_\nu)E(Y^\nu/X_\nu)\right] = E\left[E^2(Y/X_\nu)\right].\end{aligned}$$

Si estimamos la covarianza por Montecarlo al igual que en Janon et al. (2014), obtenemos el estimador Pick–Freeze de (4.3),

$$T^\nu = \frac{1}{N}\sum_{j=1}^{N} Y_j Y_j^\nu - \left(\frac{1}{2N}\sum_{j=1}^{N}(Y_j + Y_j^\nu)\right)^2, \tag{4.4}$$

donde Y_j y Y_j^ν, son N copias independientes de Y y Y^ν respectivamente.
Por tanto obtenemos un estimador de S^ν (que denotaremos $\hat{S}^\nu$),

$$\hat{S}^\nu = \frac{1/N\sum_{j=1}^{N} Y_j Y_j^\nu - \left(1/(2N)\sum_{j=1}^{N}(Y_j + Y_j^\nu)\right)^2}{1/(2N)\sum_{j=1}^{N}\left[Y_j^2 + (Y^\nu)_j^2\right] - \left(1/(2N)\sum_{j=1}^{N}(Y_j + Y_j^\nu)\right)^2}.$$

En Janon et al. (2014) se demuestra la consistencia fuerte del estimador y, suponiendo finitud del momento cuarto de la variable Y, un Teorema Central de Límite para el estimador.

También este índice ha sido generalizado cuando la salida Y es un vector aleatorio o de dimensión infinita como son los espacios de Hilbert (ver Gamboa et al. (2014) y Marrel et al. (2011)).

Por ejemplo, en el caso multivariado, el estimador propuesto es de la forma

$$S_\nu(M, f) = \frac{\text{Tr}\,(MC_\nu)}{\text{Tr}\,(M\Sigma)},$$

en donde Tr es el operador traza, M una matriz de pesos, Σ y C_ν las matrices de covarianzas de $f(X)$ y $f_\nu(X_\nu)$ respectivamente.

Sin embargo esta metodología presentada resume el impacto en el sistema

de las variables de entrada en referencia al segundo momento centrado de la salida Y y no cuantifica la sensibilidad del modelo sobre toda la distribución de Y (ver Da Veiga (2015)).

En referencia a este aspecto, en los últimos años índices alternativos de sensibilidad han sido propuestos, denominados del tipo *momento-independientes*

4.1.2. Índices del tipo momento-independientes

En general este tipo de índices, para medir el impacto de una variable de entrada X_i, se basan en alguna medida de discrepancia entre la distribución de la salida Y y la distribución condicional Y dado X_i (ver Borgonovo (2017)).

Por ejemplo, en Pianosi and Wagener (2015) se plantea una medida de sensibilidad basada en el estadístico de Kolmogorov–Smirnov,

$$S_i = \text{stat}_{x_i} \max_y \left| F_Y(y) - F_{Y|X_i=x_i}(y) \right|,$$

donde F_Y y $F_{Y|X_i=x_i}$ son la distribución y la distribución condicional de la salida Y y la función *stat* es un estadístico particular, por ejemplo, la media o la mediana. Sin embargo en su artículo no es demostrada la consistencia ni la distribución asintótica del estimador propuesto.

En Gamboa et al. (2018), una idea similar es propuesta pero a través del estadístico de Cramér–von Mises. Para este caso ellos proponen un estimador del tipo Pick–Freeze (ver sección 4.2) para el cual prueban su consistencia y encuentran su normalidad asintótica.

Otras propuestas plantean índices basados en medidas de entropía (ver Liu et al. (2006) y Borgonovo (2007)).

En Borgonovo et al. (2016) se desarrolla un índice admisible cuando la salida del sistema toma valores en espacios más generales. En este caso los autores proponen medir la sensibilidad de la entrada X_i mediante

$$S_i := E_{X_i}\left[d\left(P_Z, P_{Z|X_i}\right)\right]. \tag{4.5}$$

En este caso, $d(\cdot,\cdot)$ es una medida de disimiliaridad entre dos medidas de probabilidad, $\mathbb{E}_{X_i}(\cdot)$ es el valor esperado respecto a la variable X_i y P_Z (resp. $P_{Z|X_i}$) es la probabilidad incondicional (respectivamente condicional). Sin embargo la estimación de este índice propuesto no es sencilla.

Por tanto, el objetivo del capítulo es formular un índice que sea admisible

cuando la salida se encuentra en un espacio métrico más general que $\mathbb{R}^p$ (como lo son las variedades Riemannianas) pero de sencilla estimación.

4.2. Un índice de sensibilidad con salida en una variedad

Para la construcción de este índice se extienden los conceptos desarrollados en Gamboa et al. (2018). En este trabajo, de manera similar a Borgonovo et al. (2016) pero *intercambiando* (para la estimación) en la ecuación (4.5) el valor esperado con la medida de similaridad, se obtiene un estimador del índice de sensibilidad más sencillo (del tipo Pick–Freeze).

En nuestro marco, se compara la distribución de P_Y con la distribución condicional $P_{Y|X_i}$, pero en lugar de comparar sobre semirrectas como en Gamboa et al. (2018) se considera una familia de bolas geodésicas sobre la variedad Riemanniana. De esta manera se construye un índice intrínseco a la variedad, es decir, sólo depende de la estructura y dimensión de la variedad pero no de la dimensión del espacio euclídeo donde se encuentra inmersa, a diferencia de lo desarrollado en Gamboa et al. (2014) y Gamboa et al. (2018).

4.2.1. Construcción del índice

Dado un vector aleatorio $\mathbb{X} = (X_1, \ldots, X_d) \in \mathbb{R}^d$ con su ley de probabilidad $\mathbb{P} := \mathbb{P}_1 \times \ldots \times \mathbb{P}_d$. Se define el elemento aleatorio Z como la salida del código

$$Z = f(X_1, \ldots, X_d),$$

en donde $f : \mathbb{R}^d \to \mathcal{M}$ es una función continua y $\mathcal{M}$ es una variedad Riemanniana de dimensión k. Se tiene por objetivo establecer una medida global de sensibilidad sobre la salida Z en referencia a perturbaciones en las variables de entrada $(X_1, \ldots, X_d)$.

Como muestra Nash (1954) es sabido que toda variedad Riemanniana puede ser inmersa en $\mathbb{R}^p$ para un p lo suficientemente grande. Por tanto este problema puede ser también enmarcado en un contexto de salida multivariada, y medir la sensibilidad mediante un índice extrínseco a ella. No obstante este último enfoque no toma en cuenta la geometría del problema y no captura la *menor dimensionalidad* del espacio donde pertenece la salida Z. Comencemos por

estudiar el problema en un caso muy particular considerando la recta real como un variedad para luego extender los resultados al caso general. En la sección siguiente dados dos vectores aleatorios con valor esperado finito X_1 y X_2 denotaremos $\mathbb{E}_{X_2}\left(f(X_1, X_2)\right) := \mathbb{E}\left(f(X_1, X_2)/X_1\right)$.

4.2.2. Un caso muy particular: La recta real

A modo de ejemplificar e introducir de manera mas clara los conceptos, se hace foco en primera instancia en un caso realmente sencillo que es considerar la recta real como una variedad Riemanniana, es decir, $\mathcal{M} = \mathbb{R}$.

Sea F a la distribución acumulada de la variable aleatoria Z, $F(t) = P(Z \leq t)$ $(t \in \mathbb{R})$. Además anotemos por F^ν a la distribución de la v.a. Z condicionada a un subconjunto de coordenadas del vector aleatorio $(X_1, \ldots, X_d)$. Es decir si $\nu = \{i_1, \ldots, i_k\} \subset \{1, \ldots, d\}$, definimos

$$\begin{aligned} X_\nu &:= (X_{i_1}, \ldots, X_{i_k}), \\ F^\nu(t) &:= \mathbb{E}\left(\mathbb{1}_{(\infty,t]}(Z)\middle|X_\nu\right). \end{aligned}$$

Siguiendo esta notación, en Gamboa et al. (2018) se considera un índice de sensibilidad normalizado del tipo Cramér–von Mises que denotaremos C_2^ν. Dicho índice es definido de la siguiente manera,

$$C_2^\nu = \frac{N_{2,}^\nu}{\int_{\mathbb{R}} F(x)\left(1 - F(x)\right) dF(x)} = 6N_2^\nu, \tag{4.6}$$

en donde

$$\begin{aligned} N_2^\nu &= \int \mathbb{E}\left(\left[F(t) - F^\nu(t)\right]^2\right) dF(t) \\ &= \int \left(\mathbb{E}\left[\int \mathbb{1}_{(\infty,t]}(z) d(F - F^\nu)(z)\right]^2\right) dF(t) \\ &= \mathbb{E}\left\{\mathbb{E}\left(\left[\mathbb{E}\left(\mathbb{1}_{(\infty,t]}(Z)\right) - \mathbb{E}\left(\mathbb{1}_{(\infty,t]}(Z)\middle|X_\nu\right)\right]^2\right)\right\}. \end{aligned}$$

Introducimos para la construcción de nuestro índice un modificación en la función indicatriz $\mathbb{1}_{(\infty,t]}$. Sustituímos dicha función por otra función indicatriz pero que vale 1 no sobre semirrectas sino sobre intervalos que denotaremos $h_{s,t}$,

$$h_{s,t}(x) := \mathbb{1}_{\{s\leq x\leq t\}} + \mathbb{1}_{\{t\leq x\leq s\}} = \mathbb{1}_{\{\text{mín}\{s,t\}\leq x\leq \text{máx}\{s,t\}\}},$$

con $s, t \in \mathbb{R}$.

Podemos pensar la función $h_{s,t}$ como la indicatriz de la bola de diámetro $\overline{st}$ en $\mathbb{R}$. Por tanto, si definimos la funciones

$$H(s,t) := \mathbb{E}\left[h_{s,t}(Z)\right] \quad \text{y} \quad H^{\nu}(s,t) = \mathbb{E}\left[h_{s,t}(Z)\middle| X_{\nu}\right], \tag{4.7}$$

es claro que $\mathbb{E}_{X_\nu}\left[H^{\nu}(s,t)\right] = H(s,t)$.

Podemos entonces definir un nuevo índice de sensibilidad en sistemas con salida real y de sencilla generalización cuando la salida se encuentra en una variedad Riemannianna (bajo ciertas hipótesis).

Definición 4.2.1 *Se define un nuevo índice de sensibilidad (lo llamaremos índice de sensibilidad por bolas y denotaremos B_2^{ν}) como*

$$B_2^{\nu} := \frac{S_2^{\nu}}{\int_{\mathbb{R}^2} H(x,y)\left(1-H(x,y)\right) dF(y)dF(x)} = 6S_2^{\nu}, \tag{4.8}$$

en donde

$$S_2^{\nu} := \mathbb{E}_{Z_1,Z_2}\left[\mathbb{E}_{X_\nu}\left\{\left[H(Z_1,Z_2) - H^{\nu}(Z_1,Z_2)\right]^2\right\}\right], \tag{4.9}$$

siendo Z_1 y Z_2 dos v.a. copias independientes de la variable Z. Es claro que la constante 6 es determinada por el hecho de que

$$\int_{t>s}(t-s)(1-(t-s))dtds = \int_{t\leq s}(t-s)(1-(t-s))dtds = 1/12.$$

Se puede reescribir el numerador S_2^{ν} en términos de valores esperados como

$$S_2^{\nu} = \int_{\mathcal{M}\times\mathcal{M}} \mathbb{E}_{X_\nu}\left\{\left(H(z_1,z_2) - H^{\nu}(z_1,z_2)\right)^2\right\} dF(z_1)\times dF(z_2).$$

4.2.3. Generalización del índice de sensibilidad por bolas a una variedad Riemanniana

Sea $\mathcal{M}$ es una variedad Riemanniana de dimensión k en el marco de las hipótesis enunciadas en la introducción, es decir, dados dos puntos $\{z_1, z_2\} \subset \mathcal{M}$, la geódesica minimizante definida por ellos existe y es única. Entonces se encuentra bien definida la función $h_{z_1,z_2} : \mathcal{M} \to \{0,1\}$ tal que

$$h_{z_1,z_2}(t) := \mathbb{1}_{B_{z_1 z_2}}(t),$$

con $B_{z_1 z_2}$ la bola geodésica de diámetro $\overline{z_1 z_2}$.

Si Z_1 y Z_2 son dos copias independientes del elemento aleatorio Z, el índice de sensibilidad por bolas es construído de manera análoga que en la Definición 4.2.1,

$$B_2^\nu := \frac{S_2^\nu}{D_2^\nu}, \tag{4.10}$$

en donde

$$S_2^\nu := \mathbb{E}_{Z_1,Z_2}\left[\mathrm{Var}_{X_\nu}\left\{\mathbb{E}_Z\left[h_{Z_1,Z_2}(Z)\middle|X_\nu\right]\right\}\right],$$

y

$$D_2^\nu := \mathbb{E}\left[H(Z_1,Z_2)\left(1-H(Z_1,Z_2)\right)\right].$$

Este índice coincide con el definido en la subsección anterior cuando $\mathcal{M}=\mathbb{R}$.

A partir de la Propiedad 1 que probamos en la Introducción podemos observar la universalidad del índice, es decir, si el índice es nulo la variable de entrada no tiene un impacto sobre la salida del sistema. Dicho enunciado se formaliza en la observación siguiente,

Observación 3 *Si* $B_2^\nu = 0$ *podemos concluir que* $\mathbb{E}_Z\left[h_{z_1,z_2}(Z)\middle|X_\nu\right] = \mathbb{E}_Z\left[h_{z_1,z_2}(Z)\right]$ *c.s. para todo* $(z_1,z_2)\in\Omega\subset\mathcal{M}\times\mathcal{M}$ *con* $P(\Omega)=1$. *Por tanto si el índice es nulo las medidas de probabilidad* $P_{Z|X_\nu}$ *y* P_Z *coinciden.*

4.2.4. Estimación

Se propone un estimador para el índice B_2^ν. Al igual que en Sobol (1993) y Sobol (2001) el estimador de la varianza de la esperanza condicional es construído mediante el método *Pick–Freeze*. Puesto que la función valor esperado $\mathbb{E}_{Z_1,Z_2}(\cdot)$ es simétrica respecto a las variables aleatorias Z_1 y Z_2, se formula el estimador a través de la teoría de U–estadísticos. En la subsección siguiente se detallan los pasos para la construcción del estimador.

Estimación por el método *Pick–Freeze*

Se desarrolla la estimación del numerador de B_2^ν y de manera análoga se estima el denominador. Mediante los siguientes pasos se construye el estimador propuesto,

Primer paso. Anotemos por $\mathcal{P}_{N,p}$ a la familia de todos los subconjuntos de p elementos del conjunto $\{1,\ldots,N\}$, es decir,

$$\mathcal{P}_{N,p} = \{(i_1,\ldots,i_p) \in \{1,\ldots,N\}^p / i_1 < \ldots < i_p\}.$$

Se anota por $\tau = (i_1,\ldots,i_p) \in \mathcal{P}_{N,p}$ y por $\mathbf{W}_\tau = \left(W_{i_1},\ldots,W_{i_p}\right)$ una muestra de p copias independientes de la variable Z. El estimador a partir de la ecuación (4.4) se obtiene de la siguiente manera,

1. Se elige al azar

   i) una muestra de copias independientes de (Z, Z^ν) de tamaño N, $\{(Z_j, Z_j^\nu),\ \ 1 \le j \le N\}$,

   ii) una muestra de copias independientes de Z de tamaño N, que denotaremos $\{W_k, 1 \le k \le N\}$ que además son también independientes de la muestra $\{(Z_j, Z_j^\nu),\ \ 1 \le j \le N\}$.

2. El estimador de S_2^ν es construído como un U–estadístico de orden 2, es decir

$$\hat{S}_2^\nu = \frac{1}{\binom{N}{2}} \sum_{\tau \in \mathcal{P}_{N,2}} \left\{ \frac{1}{N} \sum_{j=1}^{N} h_{\mathbf{W}_\tau}(Z_j)\, h_{\mathbf{W}_\tau}\left(Z_j^\nu\right) - \right.$$
$$\left. - \left(\frac{1}{2N} \sum_{i=1}^{N} \left[h_{\mathbf{W}_\tau}(Z_i) + h_{\mathbf{W}_\tau}(Z_i^\nu) \right] \right)^2 \right\}.$$

De manera análoga se estima el denominador del índice D_2^ν,

$$\hat{D}_2^\nu := \frac{1}{\binom{N}{2}} \sum_{\tau \in \mathcal{P}_{N,2}} \left\{ \frac{1}{2N} \sum_{j=1}^{N} \left(h_{\mathbf{W}_\tau}(Z_j) + h_{\mathbf{W}_\tau}\left(Z_j^\nu\right) \right) - \right.$$
$$\left. - \left(\frac{1}{2N} \sum_{i=1}^{N} \left[h_{\mathbf{W}_\tau}(Z_i) + h_{\mathbf{W}_\tau}(Z_i^\nu) \right] \right)^2 \right\}.$$

El cálculo computacional del índice es sencillo debido a que sólo es necesario el cálculo de funciones indicatrices. Sin embargo el uso de U-estadísticos provoca que el número de términos sea del orden de N^3 y por otro lado en cada término es necesario el cálculo de una geodésica minimizante. Por tanto los tiempos computacionales requeridos podrían ser elevados. Planteado el estimador es necesario demostrar la consistencia, la cuál será desarrollada en la siguiente subsección. Para ello se prueba una desigualdad exponencial y a partir de ella por el Lema de Borel–Cantelli se deduce inmediatamente la consistencia fuerte.

En referencia a la precisión de la estimación las bandas de confianza serán determinadas mediante un procedimiento de remuestreo, sin embargo es factible una Teorema Central Funcional para el estimador de manera análoga al desarrollado en Gamboa et al. (2018) a partir del Método Delta funcional.

4.2.5. Propiedades asintóticas de $\hat{B}_2^\nu$.

Se analiza de manera separada la convergencia fuerte del numerador $\hat{S}_2^\nu$ y denominador $\hat{D}_2^\nu$ del índice. Nosotros probaremos la consistencia del numerador a través de una desigualdad del tipo exponencial. Sin embargo la demostración de la consistencia también puede ser deducida utilizando la desigualdad de Rosenthal para U–estadístcos de orden 2 (ver Ibragimov and Sharakhmetov (1999)), debido a la finitud de los momentos de orden 4 de las variables indicatrices.

Teorema 4.2.1 (Desigualdad exponencial) *Dado $s > 0$, entonces podemos encontrar $N_0 \in \mathbb{N}$ de manera que si $N > N_0$ se cumple que,*

$$P\left(\left|\hat{S}_2^\nu - S_2^\nu\right| > s\right) \leq 16 \exp\left\{-\frac{N\left(\frac{s}{9}\right)^2}{8}\right\}. \tag{4.11}$$

Demostración.

Se quiere mostrar que dado $s > 0$, es posible encontrar N_0 tal que para todo $N > N_0$ se cumple que,

$$P\left(\left|\hat{S}_2^\nu - S_2^\nu\right| > 9s\right) \leq 16 \exp\left\{-\frac{Ns^2}{8}\right\}. \tag{4.12}$$

Para ello demostraremos que

$$P\left(\hat{S}_2^\nu - S_2^\nu > 9s\right) \leq 8\exp\left\{-\frac{Ns^2}{8}\right\},$$

y la desigualdad en la otra cola de la distribución es análoga.

Para $1 \leq j,k \leq N$ y $\tau \in \mathcal{P}_{N,2}$ definimos

- $\mathbf{W}_\tau = (W_{k_1}, W_{k_2})$
- $\mathbf{Z}_j = \left(Z_j, Z_j^\nu\right)$
- $G(\mathbf{Z}_j, \mathbf{W}_\tau) = h_{\mathbf{W}_\tau}(Z_j)h_{\mathbf{W}_\tau}(Z_j^\nu)$
- $J(\mathbf{Z}_j, \mathbf{W}_\tau) = \frac{1}{2}\left[h_{\mathbf{W}_\tau}(Z_j) + h_{\mathbf{W}_\tau}(Z_j^\nu)\right]$
- $H(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{W}_\tau) = J(\mathbf{Z}_i, \mathbf{W}_\tau)J(\mathbf{Z}_j, \mathbf{W}_\tau)$

La demostración se divide para una mejor comprensión en tres pasos.

- **Paso 1** De manera similar que en Gamboa et al. (2018) se reescribe la expresión para $\hat{S}_2^\nu - S_2^\nu$. En este caso en términos de U–estadísticos obtenemos,

$$\begin{aligned}
\hat{S}_2^\nu &= \frac{1}{N\binom{N}{2}} \sum_{\substack{j\in\{1,\ldots,N\}\\ \tau\in\mathcal{P}_{N,2}}} G(\mathbf{Z}_j, \mathbf{W}_\tau) - \frac{1}{N^2\binom{N}{2}} \sum_{\substack{\{i,j\}\in\{1,\ldots,N\}\\ \tau\in\mathcal{P}_{N,2}}} H(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{W}_\tau) \\
&= \frac{1}{N\binom{N}{2}} \sum_{\substack{j\in\{1,\ldots,N\}\\ \tau\in\mathcal{P}_{N,2}}} \{G(\mathbf{Z}_j, \mathbf{W}_\tau) - E\left[G(\mathbf{Z}_j, \mathbf{W}_\tau)\right]\} - \\
&\quad - \frac{1}{N^2\binom{N}{2}} \sum_{\substack{\{i,j\}\in\{1,\ldots,N\}\\ \tau\in\mathcal{P}_{N,2}}} \{H(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{W}_k) - E\left[H(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{W}_\tau)\right]\} + \\
&+ E\left[G(\mathbf{Z}_1, \mathbf{W}_{\tau_1})\right] - \left(1 - \frac{1}{N}\right) E\left[H(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{W}_{\tau_1})\right] - \frac{1}{N} E\left[H(\mathbf{Z}_1, \mathbf{Z}_1, \mathbf{W}_{\tau_1})\right],
\end{aligned}$$

y

$$\begin{aligned}
S_2^\nu &= E_{\mathbf{W}_1}\left\{V_{X_\nu}\left(H^\nu(\mathbf{W}_1)\right)\right\} \\
&= E_{\mathbf{W}_1}\left\{V_{X_\nu}\left(E_Z(h_{\mathbf{W}_1}(Z)/X_\nu)\right)\right\} = E_{\mathbf{W}_1}\left\{\mathrm{cov}\left(h_{\mathbf{W}_1}(Z_1), h_{\mathbf{W}_1}(Z_1^\nu)\right)\right\} \\
&= E_{\mathbf{W}_1}\left\{E_{\mathbf{Z}_1}\left(h_{\mathbf{W}_1}(Z_1)h_{\mathbf{W}_1}(Z_1^\nu)\right)\right\} - E_{\mathbf{W}_1}\left\{\left[E_{\mathbf{Z}_1}\left(h_{\mathbf{W}_1}(Z_1)\right)\right]^2\right\}.
\end{aligned}$$

Ahora se descompone el error en tres términos que denominaremos (A), (B) y (C) respectivamente,

$$\hat{S}_2^\nu - S_2^\nu = \underbrace{\frac{1}{N\binom{N}{2}} \sum_{\substack{j\in\{1,\dots,N\}\\ \tau\in\mathcal{P}_{N,2}}} \{G(\mathbf{Z}_j, \mathbf{W}_\tau) - E\left[G(\mathbf{Z}_j, \mathbf{W}_\tau)\right]\} -}_{(A)}$$

$$\underbrace{-\frac{1}{N^2\binom{N}{2}} \sum_{\substack{\{i,j\}\in\{1,\dots,N\}\\ \tau\in\mathcal{P}_{N,2}}} \{H(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{W}_k) - E\left[H(\mathbf{Z}_i, \mathbf{Z}_j, \mathbf{W}_\tau)\right]\} +}_{(B)}$$

$$\underbrace{+\frac{1}{N}\left\{E\left[H(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{W}_{\tau_1})\right] - E\left[H(\mathbf{Z}_1, \mathbf{Z}_1, \mathbf{W}_{\tau_1})\right]\right\}.}_{(C)}$$

En lo siguiente, para un N lo suficientemente grande, se acotan los términos (A) y (B).

En los pasos 2 y 3 las acotaciones son determinadas a través de la desigualdad de Hoeffding para variables independientes reales (ver Hoeffding (1963)). Es decir, si $W_1, \dots, W_N$ son variables aleatorias independientes y centradas, cada una de ellas con recorrido en el intervalo $[a_i, b_i]$ respectivamente para $i = 1, \dots, N$, entonces

$$P\left(\sum_{i=1}^{N} X_i > s\right) \leq \exp\left\{-\frac{s^2}{\sum_{i=1}^{N}(b_i - a_i)}\right\} \quad (s > 0). \tag{4.13}$$

También utilizaremos la extensión de la desigualdad de Hoeffding para U–estadísticos de orden 2 (ver por ejemplo Serfling (1980), Teorema A [5.6]). Es decir, si $s = s(X_1, X_2)$ es el núcleo de un U–estadístico U_n tal que $E(s(X_1, X_2)) = \theta$ y $a \leq s(x_1, x_2) \leq b$, para $s > 0$ y $N > 2$ se cumple que,

$$P\left(U_n - \theta > s\right) \leq \exp\left\{-\frac{Ns^2}{(b-a)^2}\right\}. \tag{4.14}$$

En nuestro caso los núcleos de los U–estadísticos serán variables indicadoras centradas, por tanto su recorrido estará acotado en el intervalo $[-1, 1]$.

- **Paso 2** [Acotación del término (A)]

Sea $\tilde{G}(Z_i, \mathbf{W}_\tau) = G_c(Z_i, \mathbf{W}_\tau) - E_{Z_i}(G_c(Z_i, \mathbf{W}_\tau))$,

$$A = \underbrace{\frac{1}{N\binom{N}{2}} \sum_{\substack{i\in\{1,\dots,N\} \\ \tau\in\mathcal{P}_{N,2}}} \tilde{G}(Z_i, \mathbf{W}_\tau)}_{A_1} + \underbrace{\frac{1}{N\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} E_Z\left[G_c(Z, \mathbf{W}_\tau)\right]}_{A_2}$$

$$\begin{aligned} P(A_1 > s) &= P\left(\frac{1}{N\binom{N}{2}} \sum_{\substack{i\in\{1,\dots,N\} \\ \tau\in\mathcal{P}_{N,2}}} \tilde{G}(Z_i, \mathbf{W}_\tau) > s\right) \\ &= P\left(\sum_{i=1}^{N} \frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} \tilde{G}(Z_i, \mathbf{W}_\tau) > Ns\right) \\ &\leq \exp\left\{-\frac{s^2 N}{4}\right\} \leq \exp\left\{-\frac{Ns^2}{8}\right\} \quad \text{mientras que,} \end{aligned}$$

$$\begin{aligned} P(A_2 > s) &= P\left(\frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} E_Z\left[G_c(Z, \mathbf{W}_\tau)\right] > s\right) \\ &\leq \exp\left\{\frac{-2Ns^2}{8}\right\} \leq \exp\left\{-\frac{Ns^2}{8}\right\}. \end{aligned}$$

A su vez se cumple que $\{A_1 + A_2 > 2s\} \subset \{A_1 > s\} \cup \{A_2 > s\}$, por tanto

$$P(A > 2s) \leq 2\exp\left\{-\frac{Ns^2}{8}\right\}. \tag{4.15}$$

- **Paso 3** [Acotación del término (B)]

Sea $\tilde{H}(Z_i, Z_j, \mathbf{W}_\tau) = H_c(Z_i, Z_j, \mathbf{W}_\tau) - E_{Z_j}(H_c(Z_i, Z_j, \mathbf{W}_\tau))$,

$$B = \underbrace{\frac{1}{N^2\binom{N}{2}} \sum_{\substack{\{i,j\}\in\{1,\dots,N\} \\ \tau\in\mathcal{P}_{N,2}}} \tilde{H}(Z_i, Z_j, \mathbf{W}_\tau) +}_{B_1}$$
$$\underbrace{+ \frac{1}{N\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} E_Z\left[H_c(Z, Z, \mathbf{W}_\tau)\right] +}_{B_2}$$
$$\underbrace{+ \frac{N-1}{N^2\binom{N}{2}} \sum_{\substack{j\in\{1,\dots,N\} \\ \tau\in\mathcal{P}_{N,2}}} E_Z\left[H_c(Z, Z_j, \mathbf{W}_\tau)\right].}_{B_3}$$

Tenemos entonces que

$$\left\{\sum_{i=1}^{3} B_i > 6s\right\} \subset \{B_1 > 3s\} \cup \{B_2 > s\} \cup \{B_3 > 2s\},$$

Por lo tanto podemos deducir que,

$$\begin{aligned} P(B_1 > 3s) \leq P &\left(\sum_j \frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} \tilde{H}(Z_j, Z_j, \mathbf{W}_\tau) > N^2 s\right) + \\ &+ P\left(\sum_{j>i} \frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} \tilde{H}(Z_i, Z_j, \mathbf{W}_\tau) > N^2 s\right) + \\ &+ P\left(\sum_{j<i} \frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} \tilde{H}(Z_i, Z_j, \mathbf{W}_\tau) > N^2 s\right) \leq \\ &\leq \exp\left\{-\frac{s^2 N^3}{4}\right\} + 2\exp\left\{\frac{-2Ns^2N^2}{8\binom{N}{2}^2}\right\} \leq 3\exp\left\{-\frac{Ns^2}{8}\right\}, \end{aligned}$$

además,

$$\begin{aligned} P(B_2 > s) &\leq P\left(\frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} E_Z\left[H_c(Z, Z, \mathbf{W}_\tau)\right] > Ns\right) \\ &\leq \exp\left\{-\frac{Ns^2N^2}{8}\right\} \leq \exp\left\{-\frac{Ns^2}{8}\right\}, \end{aligned}$$

y por último,

$$
\begin{aligned}
P(B_3 > 2s) &\leq P\left(\frac{N-1}{N^2\binom{N}{2}} \sum_{\substack{j\in\{1,\dots,N\}\\ \tau\in\mathcal{P}_{N,2}}} E_Z\left[H_c(Z, Z_j, \mathbf{W}_\tau)\right] > 2s\right) \\
&\leq P\left(\sum_{j\in\{1,\dots,N\}} \frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} E_Z\left[H_c(Z, Z_j, \mathbf{W}_\tau)\right] > 2s\frac{N^2}{N-1}\right) \\
&\leq P\Bigg(\sum_{j\in\{1,\dots,N\}} \frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} \Big\{E_Z\left[H_c(Z, Z_j, \mathbf{W}_\tau)\right] - \\
&\quad - E_{Z,Z_j}\left[H_c(Z, Z_j, \mathbf{W}_\tau)\right]\Big\} > \frac{sN^2}{N-1}\Bigg) + \\
&\quad + P\left(\frac{1}{\binom{N}{2}} \sum_{\tau\in\mathcal{P}_{N,2}} E_{Z_1,Z_2}\left[H_c(Z_1, Z_2, \mathbf{W}_\tau)\right] > s\frac{N}{N-1}\right) \\
&\leq \exp\left\{-\frac{s^2N^4}{4N(N-1)^2}\right\} + \exp\left\{-\frac{2Ns^2N^2}{8(N-1)^2}\right\} \leq 2\exp\left\{-\frac{Ns^2}{8}\right\},
\end{aligned}
$$

Acoplando los resultados obtenidos en los desigualdades anteriores podemos deducir entonces que,

$$
P(B > 6s) \leq 6\exp\left\{-\frac{Ns^2}{8}\right\}. \tag{4.16}
$$

Por tanto, si usamos las cotas obtenidas en los pasos 2 y 3, es cierta la desigualdad (4.12).

□

A partir del Lema Borel-Cantelli se deduce entonces el siguiente corolario.

Corolario 4.2.1 (Consistencia del estimador) *$\hat{S}_2^\nu$ es un estimador consistente de S_2^ν.*

De forma análoga se demuestra la consistencia del denominador y por consiguiente del estimador.

4.3. Simulaciones

Las simulaciones se desarrollan en base a tres ejemplos. El primero pretende mostrar la precisión de nuestro estimador cuando la salida es un número real, lo que significa que puede ser un competidor a tener en cuenta cuando la salida es un escalar. En el segundo ejemplo la salida se encuentra sobre una variedad muy simple como lo es el círculo unitario inmerso en $\mathbb{R}^2$. En ambos ejemplos se compara nuestro estimador con el expuesto en Gamboa et al. (2018). En el último ejemplo se considera una variedad Riemanniana inmersa en $\mathbb{R}^3$ en donde el estimador propuesto en Gamboa et al. (2018) no brinda información acerca de la sensibilidad.

4.3.1. Ejemplo 1: Salida en la recta real

Comencemos por desarrollar un ejemplo donde el índice de Sobol no aporta información sobre la sensibilidad, mientras que si lo hacen C_2^ν y B_2^ν. Consideremos el sistema determinado por la ecuación $Z = \alpha X_1 + X_2$ con $\alpha > 0$, y en donde las variables aleatorias X_1 y X_2 son independientes. Se asume que $X_1 \sim$ Bernoulli(p) y $X_2 \sim F$ donde F es la distribución acumulada de una variable aleatoria continua con segundo momento finito. Denotamos con $m = E(X_2)$ y $\sigma^2 = V(X_2) = \alpha^2 p(1-p)$. Determinemos B_2^1, si denotamos $P(z_1, z_2) = F(z_2) - F(z_1)$, entonces podemos expresar

$$\begin{aligned} H(s,t) &= \mathbb{1}_{t<s} + F_Z(t) - F_Z(s)) \\ &= \mathbb{1}_{t<s} + (1-p)P(s,t) + pP(s-\alpha, t-\alpha), \end{aligned}$$

y

$$H^1(s,t) = \begin{cases} \mathbb{1}_{t<s} + P(s,t) & \text{si } X_1 = 0, \\ \mathbb{1}_{t<s} + P(s-\alpha, t-\alpha) & \text{si } X_1 = 1. \end{cases}$$

Por lo tanto,

$$H(s,t) - H^1(s,t) = \begin{cases} p\left[P(s,t) - P(s-\alpha, t-\alpha)\right] & \text{si } X_1 = 0, \\ (1-p)\left[P(s-\alpha, t-\alpha) - P(s,t)\right] & \text{si } X_1 = 1, \end{cases}$$

$$E_1\left[\left(H(s,t) - H^1(s,t)\right)^2\right] = p(1-p)\left[P(s,t) - P(s-\alpha, t-\alpha)\right]^2,$$

y además

$$\begin{aligned} S_2^1 &= p(1-p)E\left[(P(Z_1-\alpha,Z_1)-P(Z_2-\alpha,Z_2))^2\right] \\ &= 2p(1-p)V\left[F(Z)-F(Z-\alpha)\right]. \end{aligned}$$

Si $X_2 \sim U(0,b)$, con $b=\sqrt{12\alpha^2 p(1-p)}$, podemos entonces obtener

$$E\left(F(Z)-F(Z-\alpha)\right) = \begin{cases} \lim_{N\to\infty} \frac{1}{N^2}\frac{\alpha}{b}\left(1-\frac{1}{2}\frac{\alpha}{b}\right) & \text{si } \alpha \leq b, \\ 1/2 & \text{si } \alpha > b. \end{cases}$$

$$E\left[(F(Z)-F(Z-\alpha))^2\right] = \begin{cases} \left(\frac{\alpha}{b}\right)^2\left(1-\frac{2}{3}\frac{\alpha}{b}\right) & \text{si } \alpha \leq b, \\ 1/3 & \text{si } \alpha > b. \end{cases}$$

$$V\left(F(Z)-F(Z-\alpha)\right) = \begin{cases} \left(\frac{\alpha}{b}\right)^3\left(\frac{1}{3}-\frac{1}{4}\frac{\alpha}{b}\right) & \text{si } \alpha \leq b, \\ 1/12 & \text{si } \alpha > b. \end{cases}$$

Por tanto,

$$B_2^1 = 2p(1-p)\begin{cases} \left(\frac{\alpha}{b}\right)^3\left(\frac{1}{3}-\frac{1}{4}\frac{\alpha}{b}\right) & \text{si } \alpha \leq b, \\ 1/12 & \text{si } \alpha > b. \end{cases}$$

y como en Gamboa et al. (2018) se tiene que,

$$C_2^1 = p(1-p)\begin{cases} \left(\frac{\alpha}{b}\right)^2\left(1-\frac{2}{3}\frac{\alpha}{b}\right) & \text{si } \alpha \leq b, \\ 1/3 & \text{si } \alpha > b, \end{cases}$$

En la Figura 4.1 se comparan ambos estimadores con el verdadero valor del índice en cada caso al variar el parámetro $p \in [0,1]$. Se consideran muestras de tamaño 100, 500 y 1000. Para cada p mediante remuestreo por Bootstrap para U–estadísticos (ver Arcones and Gine (1992)) se construyen las bandas de confianza al 95 % que también se representan en la Figura 4.1.

Para cuantificar la precisión de cada estimador, en cada uno de los casos, se calcula la raíz del error cuadrático medio (denotamos MSD) para cada estimador, ver Tabla 4.1, es decir

$$MSD_{\hat{Y}} = \sqrt{\int_0^1 \left(Y(p)-\hat{Y}(p)\right)^2 dp}.$$

Se puede observar en la Figura mencionada y también en dicha Tabla que el

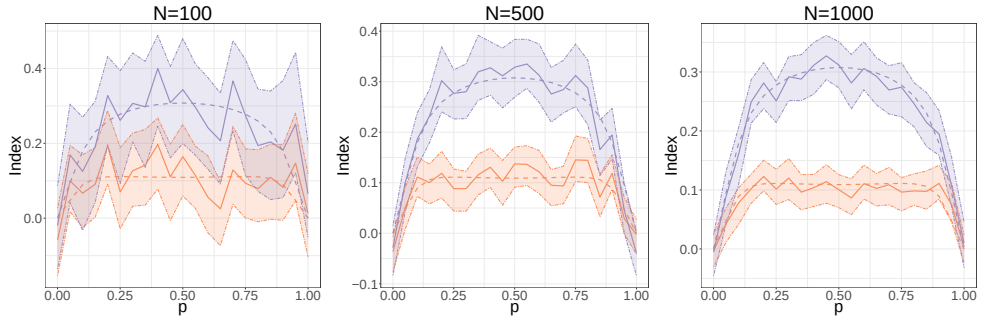

Figura 4.1: Cálculo de nuestro índice y del índice tipo Cramér–von Mises para el Ejemplo 4.3.1. En todos los casos simbolizamos con (– –) los índices teóricos y con (—) sus estimaciones. Con color violeta representamos el índice de CVM mientras que con color rojo representamos nuestro índice de sensibilidad por bolas. Las bandas de confianza al 95 % obtenidas por Bootstrap se representan sombreadas. Panel Izquierdo: $N = 100$. Panel–Central: $N = 500$. Panel–Derecho: $N = 1000$.

MSD de $\hat{B}_2^1$ es levemente inferior al obtenido en $\hat{C}_2^1$ en todos los casos. Se obtienen resultados similares cuando comparamos $\hat{B}_2^2$ y $\hat{C}_2^2$.

Size	$MSD_{\hat{B}_2^1}$	$MSD_{\hat{C}_2^1}$
$N = 100$	0.051	0.067
$N = 500$	0.022	0.028
$N = 1000$	0.013	0.018

Tabla 4.1: Cálculo del MSD para los estimadores $\hat{B}_{2,1}^1$ y $\hat{C}_2^1$ para tamaños de muestra $N = 100, 500$ y 1000.

Por tanto parecería que el nuevo índice propuesto tiene un comportamiento similar al desarrollado en Gamboa et al. (2018) si la salida se encuentra en la recta real.

4.3.2. Ejemplo 2: La salida se encuentra en una sencilla variedad Riemanniana inmersa en $\mathbb{R}^2$

Consideremos ahora el caso en que la variable de salida Z toma valores sobre el círculo unitario S_1 de $\mathbb{R}^2$. Para ello se asume que el vector de entrada tiene distribución

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right].$$

Para este caso so construye la salida como la variable X normalizada, es decir

$$Z := \frac{X}{\|X\|}.$$

La distribución de Z ha sido estudiada de manera exhaustiva por diversos autores (para una referencia ver por ejemplo Mardia (1972), Kendall (1974) y Watson (1983)). Se considera $\mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$. Se estima nuestro índice de sensibilidad y el propuesto por Gamboa et al. (2018) al variar el parámetro μ_1 en el intervalo $[-5, 0]$. En la Figura 4.2 se representan dichas estimaciones para una muestra de tamaño $N = 300$. Al representar las bandas de confianza observamos que la variabilidad de $\hat{B}_i^\nu$ es menor a la obtenida en $\hat{C}_i^\nu$ tanto para $i = 1$ como para $i = 2$.

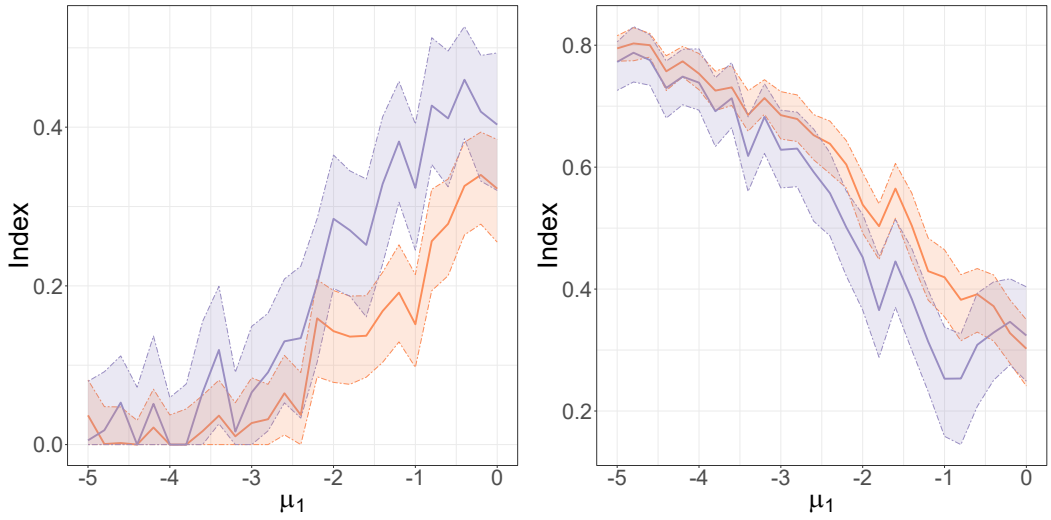

Figura 4.2: Cálculo de las estimaciones de los índices B_2^ν y C_2^ν con $\nu = 1, 2$ en la ecuación 4.3.2 con $\mu_1 \in [-5, 0]$ y $N = 300$. En todos los casos simbolizamos con (—) la estimación y en sombreado las bandas de confianza determinadas al 95 % por Bootstrap. Se representan B_2^ν y C_2^ν en colores violeta y rojo respectivamente. Panel–Izquierdo: $\nu = 1$. Panel–Derecho: $\nu = 2$.

4.3.3. Ejemplo 3: La variable de salida se encuentra inmersa $\mathbb{R}^3$

Se considera ahora una variedad Riemanniana inmersa en $\mathbb{R}^3$ definida de la siguiente manera,

$$\mathcal{M} = \left\{(x, y, z) \in \mathbb{R}^3 / xyz = 1,\ x, y, z > 0\right\}.$$

Se define un código cuya salida reside en $\mathcal{M}$,

$$Z = f(X, Y) := (X + Y, \frac{1}{X}, \frac{X}{X + Y})$$

donde X e Y son variables aleatorias independientes con distribución exponencial con media $\mu_1 > 0$. En este caso las funciones $\mathbb{1}_{\{z \leq w\}} = 0$ para todo $z, w \in M, z \neq w$. Por tanto la estimación $\hat{C}_2^{\nu}$ basada en los puntos de la muestra que se encuentran en el octante inferior es siempre nula y por consiguiente no brinda información acerca de la sensibilidad. Cabe destacar también que la extensión del índice de Sobol definido en Gamboa et al. (2014) no puede ser calculado, puesto que la segunda componente de Z, $\frac{1}{X}$, no tiene momento de segundo orden finito.

Al variar la media de las distribuciones μ_1 en el intervalo $[0, 5]$, se calcula el índice de sensibilidad por bolas congelando ambas variables, es decir, B_2^1 y B_2^2 (ver Figura 4.3). En el ejemplo se genera una muestra *Pick–Freeze* (Z_j, Z_j^{ν}), $j = 1, 2 \ldots, N = 1000$ de variables independientes a la distribución Z. Por otro lado se generan otras 1000 muestras de Z que corresponden a las variables W_k, $k = 1, \ldots, 1000$. independientes de la muestra *Pick–Freeze*. Se observa en la Figura 4.3 com el índice detecta el impacto sobre la salida para diferentes valores del parámetro μ_1 y la monotonía de este a medida que μ_1 aumenta. Las bandas de confianza se construyen mediante Bootstrap al 95 %.

4.4. Sensibilidad de la matriz de rigidez en materiales isotrópos

Si consideramos la matriz de rigidez Z para materiales isotrópicos (ver Landau and Lifshitz (1965), pág. 13) en función de las constantes de Lamé λ y μ sin el efecto de la temperatura, se tiene que la matriz Z esta dada por

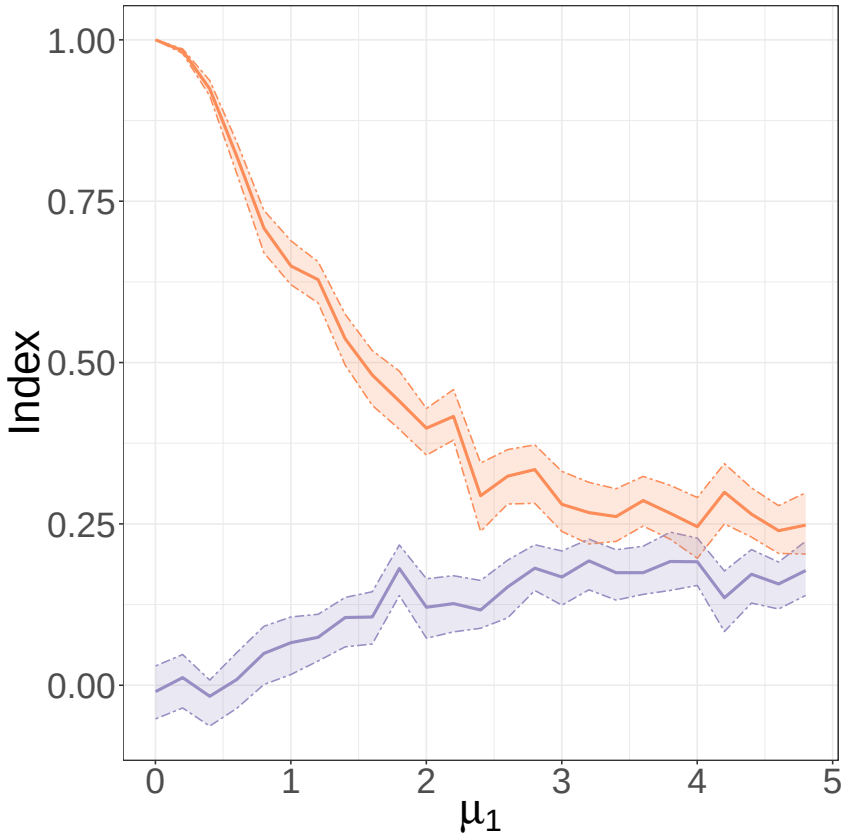

Figura 4.3: Cálculo del índice $\hat{B}_2^{\nu}$ para $\nu = 1, 2$. En color rojo y violeta se representan $\hat{B}_2^1$ y $\hat{B}_2^2$ respectivamente. Las bandas de confianza son construídas mediante Bootstrap al 95 % y se representan sombreadas.

$$Z = \begin{pmatrix} K + 4\mu/3 & K - 2\mu/3 & K - 2\mu/3 & 0 & 0 & 0 \\ K - 2\mu/3 & K + 4\mu/3 & K - 2\mu/3 & 0 & 0 & 0 \\ K - 2\mu/3 & K - 2\mu/3 & K + 4\mu/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu \end{pmatrix}$$

en donde $K = \lambda + 2\mu/3$ es el módulo volumétrico. El parámetro μ es llamado módulo de rigidez. Puesto que K y μ sólo toman valores positivos son modelados con una distribución con soporte en $\mathbb{R}^+$.

El conjunto de matrices de rigidez se considera como una subvariedad de la variedad Riemanniana de las matrices de semidefindas positivas con la métrica g, $(\mathbb{P}_d, g)$, ver Moakher (2005). Dadas dos matrices $A, B \in \mathbb{P}_d$ sabemos que existe una única geodésica que determinan A y B dada por,

$$\gamma(t) := A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^t A^{1/2}.$$

Por tanto podemos calcular el punto medio entre A y B, que denotaremos $A\#B$, y su distancia geodésica,

$$A\#B = A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2} A^{1/2} \tag{4.17}$$

$$d(A,B) = \| \log\left(A^{-1/2}BA^{-1/2}\right) \|, \tag{4.18}$$

donde $\|\cdot\|$ es la norma de Hilbert–Schmidt entre matrices. Observar que el punto medio es simplemente la media geométrica de las matrices.

En el ejemplo consideramos K y μ variables aleatorias independientes y nos enfocamos en dos posibles escenarios,

Caso 1 $K \sim \gamma(1/\lambda_K, \lambda_K)$ $\mu \sim \gamma(1/\lambda_\mu, \lambda_\mu)$

Caso 2 $K \sim \mathrm{U}(1-\lambda_K, 1+\lambda_K)$ $\mu \sim \mathrm{U}(1-\lambda_\mu, 1+\lambda_\mu)$

donde γ y U son las distribuciones Gamma y Uniforme respectivamente.

En las Tablas 4.2 y 4.3 se observan los valores de los índices $\hat{B}_2^1$ y $\hat{B}_2^2$ para el caso 1 y 2 respectivamente para diferentes valores de los parámetros λ_K y λ_μ.

Ambas Tablas muestran que el índice proporciona información relevante acerca de la sensibilidad de la salida Z con respecto a los parámetros de entrada K y μ. En particular se observa como el aumento de la variabilidad de la variable en cuestión impacta sobre el valor del índice, a medida que aumenta la variabilidad (λ_μ o λ_K) el índice de sensibilidad propuesto se incrementa. En las simulaciones tomamos 500 muestras *Pick–Freeze* y 500 muestras W independientes de las anteriores.

Tabla 4.2: Valores de los índices en una matriz isótropa con parámetros con distribución Gamma (caso 1).

Distribución	Caso 1: B^1				Caso 1: B^2			
$\lambda_\mu \backslash \lambda_K$	**0.001**	**0.01**	**0.1**	**1**	**0.001**	**0.01**	**0.1**	**1**
0.001	0.625	0.212	0.016	0.001	0.083	0.435	0.855	0.980
0.01	0.925	0.593	0.215	0.033	0.004	0.072	0.458	0.865
0.1	0.987	0.912	0.587	0.184	0.001	0.006	0.137	0.518
1	0.999	0.990	0.930	0.600	0.000	0.007	0.210	0.311

Tabla 4.3: Valores de los índices en una matriz isótropa con parámetros con distribución Uniforme (caso 2).

Distribución	Caso 2: B^1				Caso 2: B^2			
$\lambda_\mu \backslash \lambda_K$	**0.001**	**0.01**	**0.1**	**1**	**0.001**	**0.01**	**0.1**	**1**
0.001	0.620	0.008	0.001	0.000	0.109	0.848	0.997	1.000
0.01	0.989	0.623	0.003	0.001	0.018	0.092	0.849	0.997
0.1	1.000	0.989	0.623	0.624	0.016	0.016	0.102	0.846
1	1.000	1.000	0.990	0.987	0.015	0.016	0.100	0.211

4.5. Conclusiones del capítulo

En el capítulo se introduce una nueva medida de sensibilidad global que brinda información sobre el impacto de las variables de entrada cuando la salida del sistema se encuentra sobre una variedad Riemanniana. En resumen podemos mencionar las siguientes cualidades del índice,

- El índice propuesto es una posible extensión del propuesto en Gamboa et al. (2018) y permite su extensión cuando la salida se encuentra sobre una variedad Riemanniana.
- Este índice esta basado en toda la distribución del sistema y no sólo en sus momentos.
- La manera en que se construye el índice, mediante el uso de geodésicas minimizantes, permite incorporar la geometría del problema en las medidas obtenidas.
- El estimador *Pick–Freeze* propuesto, construído como un U–estadístico, es sencillo de calcular. Es demostrada su consistencia fuerte.
- Mediante diferentes escenarios de simulación se muestra la buena performance de nuestro índice, inclusive sobre la recta real. Usando remuestreo por el método de Bootstrap para U–estadísticos son construídos las bandas de confianza.
- En la sección final se analiza la aplicación de nuestro índice a un problema físico. Es analizada la sensibilidad de la matriz isótropa de rigidez en referencia a diferentes modelizaciones de las constantes de Lamé.

Capítulo 5

Conclusiones Finales

A modo de resumen enumeramos los principales aportes que consideramos la tesis contiene,

- Se construye, a través de proyecciones al azar, un test de simetría y otro de independencia aplicables tanto en dimensión finita como infinita. Para ambos casos se demuestra su consistencia, su distribución libre y un TCL. Se muestra una aplicación a datos reales del test de independencia.

- Se extiende una medida de profundidad (la profundidad esférica) para el caso infinito dimensional y para el caso en el que los datos se encuentran sobre una variedad Riemanniana. Son probadas aquellas propiedades deseables de una medida de profundidad, la consistencia de estimador y su distribución asintótica.

- Se define una nueva medida de sensibilidad aplicable cuando el output se encuentra en un espacio euclídeo (finito o infinito dimensional) o sobre una variedad Riemanniana. Se propone como estimador un $U-$ estadístico basado en el método Pick–Freeze y se demuestra su consistencia.

En todos los casos, mediante simulaciones, se muestra el buen desempeño de los estimadores planteados (en diversos escenarios) frente a otros competidores. En particular, cuando el espacio es de dimensión finita (independientemente de la dimensión del espacio), es de dimensión infinita o es una variedad Riemanniana, los métodos propuestos mantienen un comportamiento adecuado.

Referencias bibliográficas

Aki, S. (1993). On nonparametric tests for symmetry in R^m. *Annals of the Institute of Statistical Mathematics*, 45:787–800.

Albert, P., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interactions. *American Journal of Epidemiology*, 154:687–693.

Antoniadis, A. (1984). Analysis of variance on function spaces. *Statistics: A Journal of Theoretical and Applied Statistics*, 15(1):59–71.

Arcones, M. A. and Gine, E. (1992). On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674.

Arcones, M. A. and Gine, E. (1993). Limit theorems for U–processes. *The Annals of Probability*, 21(3):1494–1542.

Arnaudon, M., Barbaresco, F., and Yang, L. (2013). Riemannian medians and means with applications to radar signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 7(4):595–604.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83:715–726.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*, 139(3):318–355.

Bellman, R., Bellman, R., and Collection, K. M. R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton legacy library. Princeton University Press.

Bhattacharya, A. and Bhattacharya, R. (2012). *Nonparametric inference on manifolds: with applications to shape spaces*, volume 2. Cambridge University Press.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition. A Wiley-Interscience Publication.

Billingsley, P. and Topsøe, F. (1967). Uniformity in weak convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 7(1):1–16.

Blough, D. K. (1989). Multivariate symmetry via projection pursuit. *Annals of the Institute of Statistical Mathematics*, 41:461–475.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771–784.

Borgonovo, E. (2017). *Sensitivity Analysis: An Introduction for the Management Scientist*, volume 251. Springer.

Borgonovo, E., Hazen, G. B., and Plischke, E. (2016). A common rationale for global sensitivity measures and their estimation. *Risk Analysis*, 36(10):1871–1895.

Brandwein, A. and Strawderman, W. (1991). Generalizations of James-Stein estimators under spherical symmetry . *The Annals of Statistics*, 19:1639–1650.

Cacuci, D. G. (1981). Sensitivity theory for nonlinear systems. i. nonlinear functional analysis approach. *Journal of Mathematical Physics*, 22(12):2794–2802.

Carrizosa, E. (1996). A characterization of halfspace depth. *Journal of multivariate analysis*, 58(1):21–26.

Chastaing, G., Gamboa, F., and Prieur, C. (2015). Generalized sobol sensitivity indices for dependent variables: numerical methods. *Journal of Statistical Computation and Simulation*, 85(7):1306–1333.

Chen, M., Gao, C., Ren, Z., et al. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960.

Christensen, J. (1970). On some measures analogous to Haar measure. *Mathematica Scandinavica*, 26:103–106.

Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. *Journal London Mathematical Society*, 11:290–294.

Croux, C., Garcia-Escudero, L.A., Gordaliza, A., Ruwet, C., and Martin, R. (2017). Robust principal component analysis based on trimming around affine subspaces. *Statistica Sinica*, 27(3):1437–1459.

Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2006). Random projections and goodness of fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society*, 37:477–501.

Cuesta-Albertos, J. A., Fraiman, R., and Ransford, T. (2007). A sharp form of the Cramer–Wold theorem. *Journal of Theoretical Probability*, 20:201–209.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Cuevas, A. and Fraiman, R. (2009). On depth measures and dual statistics. A methodology for dealing with general data. *J. Multivariate Analysis*, 100(4):753–766.

Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305.

Da Veiga, S. and Gamboa, F. (2013). Efficient estimation of sensitivity indices. *Journal of Nonparametric Statistics*, 25(3):573–595.

Da Veiga, S., Wahl, F., and Gamboa, F. (2009). Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics*, 51(4):452–463.

Dasgupta, S. (1999). Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE.

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

Dauwels, J., Vialatte, F., and Cichocki, A. (2010). Diagnosis of alzheimers disease from EEG signals: Where are we standing. *Current Alzheimer Research*, 7:487–505.

der Vaart, A. W. V. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

do Carmo, M. (1992). *Riemannian Geometry*. Mathematics. Birkhäuser, Boston Basel Berlin.

Donoho, D. L. et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(32):375.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929.

Dyckerhoff, R., Ley, C., and Paindaveine, D. (2015). Depth-based runs test for bivariate central symmetry. *Annals of the Institute of Statistical Mathematics*, 67:917–941.

Einmahl, J. and Gan, Z. (2016). Testing for central symmetry. *Journal of Statistical Planning and Inference*, 169:27–33.

Elmore, R. T., Hettmansperger, T. P., and Xuan, F. (2006). Spherical data depth and a multivariate median. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:87.

Fermanian, J. (2005). Goodness of fit tests for copulas. *Journal of multivariate analysis*, 95(1):119–152.

Fermanian, J.-D., Radulovic, D., and Wegkamp, M. (2004). Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860.

Ferri, M. and Frosini, P. (2008). VC-dimension on manifolds: a first approach. *Mathematical methods in the applied sciences*, 31(5):589–605.

Fletcher, P. T., Venkatasubramanian, S., and Joshi, S. (2009). The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152.

Folland, G. B. (2013). *Real analysis: modern techniques and their applications*. John Wiley & Sons.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23:881–890.

Gamboa, F., Janon, A., Klein, T., Lagnoux, A., et al. (2014). Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8(1):575–603.

Gamboa, F., Klein, T., and Lagnoux, A. (2018). Sensitivity analysis based on Cramér–von Mises distance. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):522–548.

Genest, C., Quessy, J.-F., and Rémillard, B. (2007). Asymptotic Local Efficiency of Cramér-Von Mises Tests for Multivariate Independence. *The Annals of Statistics*, 35:166–191.

Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques*, 44:1096–1127.

Giné, E. (1996). Empirical processes and applications: an overview. *Bernoulli*, 2(1):1–28.

Hallin, M. and Paindaveine, D. (2002). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks . *The Annals of Statistics*, 30:1103–1133.

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896.

Heathcote, C., Rachev, S., and Cheng, B. (1995). Testing Multivariate Symmetry. *Journal of Multivariate Analysis*, 54(1):91–112.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat*, 19:293–325.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30.

Hoffmann-Jørgensen, J. (1991). *Stochastic processes on Polish spaces*. Number 39. Aarhus Universitet. Matematisk Institut.

Ibragimov, R. and Sharakhmetov, S. (1999). Analogues of Khintchine, Marcinkiewicz-Zygmund and Rosenthal Inequalities for Symmetric Statistics. *Scandinavian Journal of Statistics*, 26(4):621–633.

Iooss, B. and Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems*, pages 101–122. Springer.

Janon, A., Klein, T., Lagnoux, A., Nodet, M., and Prieur, C. (2014). Asymptotic normality and efficiency of two sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364.

Jones, M. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society Series A*, 150:1–36.

Kendall, D. G. (1974). Pole-seeking brownian motion and bird navigation. *Journal of the Royal Statistical Society*, 36:261–294.

Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Kucherenko, S. (2005). Global sensitivity indices for nonlinear mathematical models. review. *Wilmott Mag*, 1:5661.

Landau, L. D. and Lifshitz, E. M. (1965). *Theory of elasticity*. Nauka.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Ley, C. (2010). *Univariate and multivariate symmetry: statistical inference and distributional aspects*. PhD thesis, Université libre de Bruxelles.

Lin, T. and Zha, H. (2008). Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809.

Liu, H., Chen, W., and Sudjianto, A. (2006). Relative entropy based method for probabilistic sensitivity analysis in engineering design. *Journal of Mechanical Design*, 128(2):326–336.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414.

Liu, R. Y. (1992). Data depth and multivariate rank tests. *L1-statistical analysis and related methods (Y. Dodge, ed.)*, pages 279–294.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3):783–840.

Liu, R. Y. and Singh, K. (1992). Ordering directional data: Concepts of data depth on circles and spheres. *Ann. Statist.*, 20(3):1468–1484.

Liu, Z. and Modarres, R. (2011). Lens data depth and median. *Journal of Nonparametric Statistics*, 23(4):1063–1074.

Marden, J. (1999). *Multivariate Analysis, Design of Experiments, and Survey Sampling*, chapter 14. Multivariate Rank Test, pages 401–432. CRC Press.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, 36(1):99–109.

Mardia, K. V. and Jupp, P. E. (2000). *Directional statistics*. John Wiley & Sons.

Mardia, K. V. and Voss, J. (2014). Some fundamental properties of a multivariate von Mises distribution. *Communications in Statistics-Theory and Methods*, 43(6):1132–1144.

Maronna, R., Martin, R. D., and Yohai, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

Marrel, A., Iooss, B., Jullien, M., Laurent, B., and Volkova, E. (2011). Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics*, 22(3):383–397.

Mason, D. M. and Schuenemeyer, J. H. (1983). A Modified Kolmogorov-Smirnov Test Sensitive to Tail Alternatives. *Annals of Statistics*, 11:933–946.

Moakher, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747.

Nash, J. (1954). c^1–isometric imbedding. *Annals of Mathematics*, 60(3):383–396.

Neuhaus, G. and Zhu, L. (1998). Permutation Tests for Reflected Symmetry . *Journal of Multivariate Analysis*, 67(2):129–153.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332.

Owen, A. (1994). Lattice sampling revisited: Monte carlo variance of means over randomized orthogonal arrays. *The Annals of Statistics*, 22(2):930–945.

Padgett, W. J. and Taylor, R. L. (1973). *Laws of Large Number for Normed Linear Spaces and Certain Fréchet Spaces*. Springer-Verlag.

Patrangenaru, V. and Ellingson, L. (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Pennec, X. (2006). Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154.

Petersen, P. (2006). *Riemannian geometry*, volume 171. Springer.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., and Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, 79:214–232.

Pianosi, F. and Wagener, T. (2015). A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. *Environmental Modelling & Software*, 67:1–11.

Pizer, S. M. and Marron, J. (2017). Object statistics on curved manifolds. In *Statistical Shape and Deformation Analysis*, pages 137–164. Elsevier.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley Series in Probability and Statistics.

Rocquigny, E. D., Devictor, N., and Tarantola., S. (2008). *Uncertainty in industrial practice*. Wiley Online Library.

Saltelli, A., Chan, K., and Scott, E. (2000). *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

Sen, P. K. and Chatterjee, S. K. (1973). On Kolmogorov-Smirnov type test for symmetry. *Annals of the Institute of Statistical Mathematics*, 25:288–300.

Sen, P. K. and Puri, M. L. (1967). On the theory of rank order tests for location in the multivariate one sample problem . *The Annals of Mathematical Statistics*, 38:1216–1228.

Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, New York, NY [u.a.], [nachdr.] edition.

Serfling, R. (2006). Multivariate symmetry and asymmetry. In *Encyclopedia of Statistical Sciences, Second Edition*, volume 8, pages 5338–5345. J. Wiley & Sons.

Serfling, R. and Zuo, Y. (2000). General notions of statistical depth function. *The Annals of Statistics*, 28(2):461–482.

Shahsavarifar, R. and Bremner, D. (2018). Computing the planar *beta*-skeleton depth. *arXiv preprint arXiv:1803.05970*.

Shohat, J. A. and Tamarkin, J. D. (1943). *The problem of moments*. Mathematical Surveys and Monographs.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Modeling Comput. Experiment*, 1(4):407–414 (1995).

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1):271–280.

Spearman, C. (1904). "general intelligence,.ºbjectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Steele, J. M. (1975). *Combinatorial entropy and uniform limit laws*. Department of Mathematics, Stanford University.

Szabados, T. (1989). On the Glivenko-Cantelli theorem for balls in metric spaces. *Studia Scientiarum Mathematicarum Hungarica*, 24:473–481.

Székely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Székely, G. J., Rizzo, M. L., and Bakirov, N. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35:2769–2794.

Takács, L. (1967). *Combinatorial methods in the theory of stochastic processes*. John Wiley.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

Watson, G. S. (1983). . *Statistics on Spheres*. Wiley.

Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica*, 3:309–326.

Zhang, X., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538.