



Universidad de Valladolid

Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Matemáticas

Baricentros en el espacio de Wasserstein: aplicación a modelos estadísticos de deformación

Autor: Paula Gordaliza Pastor

Tutor: Eustasio del Barrio Tellado

Índice general

ABSTRACT.	7
RESUMEN.	9
AGRADECIMIENTOS.	11
INTRODUCCIÓN.	13
1. PRELIMINARES.	17
1.1. Teoría de transporte	17
1.1.1. El problema del transporte óptimo	19
1.1.2. La distancia de Wasserstein	20
1.2. Procesos empíricos y cuantiles	24
1.2.1. El proceso empírico	25
1.2.2. El proceso cuantil	27
1.3. El movimiento Browniano	28
1.4. La técnica de remuestreo <i>bootstrap</i>	36
1.4.1. El principio <i>bootstrap</i>	37
1.4.2. Aproximación Monte-Carlo de $\hat{\theta}^*$	37
2. BARICENTROS EN EL ESPACIO DE WASSERSTEIN Y EL MODELO	

DE DEFORMACIÓN PARA DISTRIBUCIONES.	39
2.1. Baricentros en el espacio de Wasserstein	40
2.2. Modelo de deformación para distribuciones.	43
2.2.1. Estimación del modelo de deformación paramétrico	46
2.2.2. Ejemplos de familias paramétricas de deformación	54
2.3. Bootstrap con la distancia de Wasserstein	57
2.3.1. Dependencia continua de la distancia de Wasserstein	58
2.3.2. Dependencia continua de la variación de Wasserstein	62
2.3.3. Bondad de ajuste del test para el modelo de deformación paramétrico	64
3. ASPECTOS COMPUTACIONALES.	69
3.1. Implementación del modelo de localización y escala	70
3.1.1. Caso I=2 muestras	71
3.1.2. Caso general	73
3.2. Bootstrap bajo el modelo de deformación	81
3.2.1. Bootstrap de residuos	82
4. SIMULACIONES Y RESULTADOS.	85
4.1. Simulaciones para estimar la frecuencia de rechazo	85
4.2. Simulaciones para comprobar la potencia del test	95
4.2.1. $\gamma \stackrel{d}{=} \varepsilon(1)$	96
4.2.2. $\gamma \stackrel{d}{=} Laplace(0, 1)$	97
4.2.3. $\gamma \stackrel{d}{=} t_3$	98
4.2.4. $\gamma \stackrel{d}{=} t_4$	99
CONCLUSIONES.	101

BIBLIOGRAFÍA.	103
ANEXOS	106
A. CÓDIGOS PARA IMPLEMENTAR LAS SIMULACIONES.	109
A.1. calculocriterio2	109
A.2. rechazo.BR2	109
A.3. calculocriteriog	110
A.4. rechazo.BR	111
A.5. rechazo.BR.potencia	111

ABSTRACT.

Wasserstein barycenters and variance-like criterion using Wasserstein distance are used in many problems to analyze the homogeneity of collections of distributions and structural relationships between the observations. We study the estimation of the quantiles of the empirical process of the Wasserstein's variation using a bootstrap procedure. Then we use these results for statistical inference on a distribution registration model for deformations in a parametric family of functions. Particularly, we calculate the value of the variation and the optimal parameters when the observations satisfy a scale-location model. We introduce a central limit theorem for the Wasserstein's variation, which is useful to build a test, based on this statistic, that checks the homogeneity of the distributions. Finally, we present results of simulations that have been carried out to verify the goodness of the test in different scenarios.

RESUMEN.

En el análisis de la homogeneidad de una colección de distribuciones y de relaciones estructurales entre las observaciones, son muy útiles los baricentros y la variación en distancia de Wasserstein. Estudiamos la estimación de los cuantiles del proceso empírico de la variación de Wasserstein mediante un procedimiento bootstrap. Después, usamos los resultados obtenidos para hacer inferencia estadística bajo un modelo de deformación paramétrico. En particular, calculamos el valor de la variación y de los parámetros óptimos cuando las observaciones satisfacen un modelo de localización y escala. Enunciamos un teorema central del límite para la variación de Wasserstein, útil para construir un test basado en dicho estadístico, que comprueba la homogeneidad de distribuciones. Finalmente, presentamos resultados de simulaciones llevadas a cabo, en diferentes escenarios, para verificar la bondad del test.

AGRADECIMIENTOS.

A Eustasio del Barrio, tutor de este Trabajo Fin de Grado, que se ha desarrollado en el marco de un proyecto de Beca de Colaboración del Ministerio de Educación, al que agradezco, tanto su dedicación en este trabajo, como su enseñanza en las diversas asignaturas que me ha impartido a lo largo del Grado.

Al Departamento de Estadística e Investigación Operativa, por cederme un lugar de trabajo y prestar sus instalaciones para llevar a cabo las simulaciones.

A todos mis otros profesores, que han contribuido también a mi formación en estos cuatro años, gracias a la cual he podido desarrollar este proyecto con solvencia.

A mis seres queridos, por su apoyo incondicional.

INTRODUCCIÓN.

Los seres humanos, con nuestras acciones cotidianas, estamos creando y almacenando información incesantemente y cada vez a un ritmo más elevado. A cada segundo que pasa, cantidades ingentes de datos de todo tipo (mediciones, mensajes, documentos, imágenes, sonidos, etc.) son generados y almacenados en el mundo. Nuestro cuerpo, en sí mismo, constituye un enorme contenedor de datos como los que, por ejemplo, determinan nuestra expresión génica. A esta acumulación masiva de datos también contribuyen las empresas, generando datos transaccionales y reuniendo información acerca de sus clientes, proveedores, operaciones, etc. De la misma manera sucede en el sector público, dada la necesidad de cada país de explotar las bases de datos del censo de población o de registros médicos; o en distintos sectores como el industrial, transporte, servicios o comercial, en los que sensores digitales son instalados en diferentes partes de un proceso que quiera ser monitorizado, con el objetivo de entender su funcionamiento y aprender de sus fallos, impulsando así su mejora. Si a todo esto le añadimos todas aquellas actividades que la mayoría de nosotros realizamos varias veces al día, tales como transacciones financieras, uso de redes sociales y correo electrónico, consulta de ubicaciones geográficas mediante coordenadas GPS, etc., expertos aseguran que se generan alrededor de 2.5 trillones de bytes de información diariamente en el mundo. Es más, el IBM estimó en 2014 que el 90 % de los datos existentes se habían generado en los dos últimos años.

La necesidad de desarrollar técnicas potentes y eficaces para la modelización y el análisis de grandes volúmenes de datos de diversa naturaleza es, por tanto, uno de los problemas que se han situado recientemente en la cabecera de la investigación en el campo de las matemáticas, particularmente de la estadística. La dificultad de este problema se encuentra tanto en la variedad como en el gran tamaño de los conjuntos de datos a analizar.

Entre los principales y más tempranos objetivos de la estadística, se encuentra

el dar sentido a la noción de comportamiento promedio. Esta tarea se ha vuelto particularmente interesante pues la diversidad en la tipología de la información ha suscitado el reto de tratar datos que poseen una geometría interna muy diferente a la Euclídea y, en consecuencia, la versión habitual de media en espacios euclídeos no es la más apropiada. De aquí surge la necesidad de definir una nueva noción de comportamiento promedio que dependerá del espacio en el que vivan los datos. Por otro lado, cuando nos enfrentamos a grandes muestras de datos generados por diferentes mecanismos o en distintos contextos, es habitual que estén afectados por diferentes fuentes de variabilidad que incluyen pequeñas perturbaciones como, por ejemplo, transformaciones de localización y escala o deformaciones más generales. En estos casos, se desaconseja el uso de los métodos estadísticos habituales y se requiere de nuevas técnicas que permitan un correcto análisis de los datos perturbados. Este problema surge de manera natural en un amplio rango de problemas estadísticos, tales como el análisis de datos funcionales, el cual es tratado en [16], [12], [23] o [5]; el procesamiento de imágenes en [26] o [4] y de formas en [20]; con aplicaciones en biología [7] o en reconocimiento de patrones [24], por ejemplo.

En este trabajo se considera el caso en el que los objetos de interés son distribuciones de probabilidad, estimadas a partir de diferentes fuentes de datos, que inducen deformaciones en la señal observada. En cuanto a la noción de comportamiento promedio, en [1] se introduce un concepto particular de media de una colección de probabilidades, que se corresponde con la solución del problema de minimización de las distancias cuadráticas de Wasserstein W_2 , y a la que se le ha dado el nombre de baricentro de Wasserstein. Respecto al problema de las deformaciones que acompañan a las muestras observadas, se trata de determinar si los datos, antes de ser deformados, procedían de una distribución común que es, en general, desconocida.

Como ejemplo de este problema estadístico de deformaciones, cabe destacar el que se presenta frecuentemente en la biología al estudiar datos de expresión génica obtenidos de técnicas de análisis de microarrays. Existen dos caracterizaciones de la variación que se espera observar al comparar múltiples arrays: *variación interesante* y *variación oscurecedora*. *Variación interesante* sería, por ejemplo, una gran diferencia en niveles de expresión de determinados genes entre un tejido enfermo y uno normal. Sin embargo, en los niveles de expresión observados también se incluyen variaciones que surgen de las diferencias en la preparación de las muestras (diferencias de etiquetado, por ejemplo) o en la producción y procesado de las muestras (diferencias de escaneado, por ejemplo); éstas serían *variaciones oscurecedoras*. Para hacer frente a éstas últimas, en [7] desarrollan la conocida técnica de la normalización cuantil, trabajo que

es continuado por [15].

Recientemente, metodologías de transporte óptimo han sido utilizadas para estudiar similitudes entre distribuciones de probabilidad observadas con deformaciones. En este sentido, en [11] se construye un test para comparar una colección de distribuciones ν_1, \dots, ν_I , y para ello se define una noción de variación entre distribuciones basada en la distancia de Wasserstein W_r entre probabilidades en \mathbb{R}^d , $d \geq 1$, con momento de orden $r \geq 1$ finito, que conforman el espacio que denotaremos por $\mathcal{W}_r(\mathbb{R}^d)$. La variación de Wasserstein mide la dispersión entre las distribuciones generalizando, por tanto, el concepto de varianza para puntos en \mathbb{R}^d , y se define como

$$V_r(\nu_1, \dots, \nu_I) = \inf_{\eta \in \mathcal{W}_r(\mathbb{R}^d)} \left(\frac{1}{I} \sum_{i=1}^I W_r^r(\nu_i, \eta) \right)^{\frac{1}{r}},$$

El interés de la variación de Wasserstein radica en su utilidad a la hora de valorar la veracidad de un modelo de deformación entre distribuciones. Esta idea se desarrolla en [2] y [21] cuando las deformaciones pertenecen a una familia paramétrica, y se extiende al caso general en [11]. En este último, se observan $i = 1, \dots, I$ muestras independientes $(X_{i,j})_{1 \leq j \leq n}$, de variables aleatorias en \mathbb{R}^d con distribución ν_i , que satisfacen el modelo de deformación

$$X_{i,j} = (\varphi_i^*)^{-1}(\varepsilon_{i,j}),$$

donde $\varepsilon_{i,j}$, $1 \leq j \leq n$, $1 \leq i \leq I$, son variables aleatorias independientes e igualmente distribuidas con distribución desconocida ν , y, para cada $i = 1, \dots, I$, la función φ_i^* representa la deformación mediante la cual se ha transformado ν para obtener la muestra i -ésima, de manera que $\varphi = (\varphi_1, \dots, \varphi_I)$ pertenece a una determinada familia de funciones \mathcal{G} . El criterio teórico que se propone para evaluar el ajuste de este modelo se basa en la 2-variación de Wasserstein

$$V_2^2(\nu_1(\varphi_1), \dots, \nu_I(\varphi_I)) = \inf_{\eta \in \mathcal{W}_2(\mathbb{R}^d)} \left(\frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i(\varphi_i), \eta) \right),$$

donde $\nu_i(\varphi_i)$, $i = 1, \dots, I$ representa la distribución de las observaciones tras deshacer la deformación.

Con las bases teóricas establecidas en [11] que marcan las pautas para aplicar *bootstrap* a la distancia de Wasserstein, en este trabajo se diseña un procedimiento general que permite dilucidar la validez del modelo de deformación en el caso paramétrico mediante métodos de simulación Monte-Carlo. El Capí-

tulo 1 pretende ser una recopilación de conceptos y resultados preliminares, necesarios para el desarrollo de los capítulos posteriores. En el Capítulo 2 se explican los aspectos teóricos del baricentro y la variación de Wasserstein, y del modelo de deformación, y se justifica el uso del bootstrap en este contexto, presentándose recientes resultados de [11]. El Capítulo 3 se dedica a la implementación práctica del criterio, especificándose su cálculo efectivo en el modelo de localización y escala. También, en este capítulo, se diseña un procedimiento general para valorar el ajuste de un modelo de deformación paramétrico, el cual se aplica al ejemplo particular del modelo de localización y escala. Por último, en el Capítulo 4 se muestran los resultados obtenidos al probar dicho procedimiento en diferentes escenarios, con los que se ha conseguido comprobar que el test construido alcanza la probabilidad de cubrimiento deseada y que además, cuenta con una alta potencia.

1 | PRELIMINARES.

Este capítulo pretende recopilar algunos conceptos que resultan necesarios para el desarrollo de la teoría expuesta en este trabajo, y que aparecerán a menudo a lo largo de los capítulos posteriores. La conexión entre las distintas secciones viene de la aplicación a modelos de deformación, y se verá más clara al leer los capítulos posteriores.

Se omitirán las demostraciones de la mayoría de los resultados teóricos que se enuncian en este capítulo. No obstante, se indicarán convenientemente algunas referencias bibliográficas donde éstos se presentan de manera detallada, siendo [28] la referencia principal para la sección 1.1; [10] para la sección 1.2; [6] para la sección 1.3; y [13], [17] y [25] para la última sección 1.4.

1.1 Teoría de transporte

Un problema que ha cobrado gran relevancia desde las primeras décadas del siglo XX es el del transporte óptimo. Este problema fue formalizado en 1781 por el matemático francés Gaspard Monge, y experimentó grandes avances gracias al matemático ruso Leonid Kantorovich durante la Segunda Guerra Mundial. Desde entonces, ha sido muy estudiado en la literatura. Nos limitaremos aquí a describirlo brevemente y establecer la relación con la herramienta que juega un papel fundamental en el tema que nos ocupa, y que es la distancia de Wasserstein.

Si \mathcal{X} es un espacio métrico completo y separable, denotaremos por $\mathcal{P}(\mathcal{X})$ el conjunto de medidas de probabilidad sobre \mathcal{X} .

| Definición 1.1. Sean \mathcal{X} e \mathcal{Y} dos espacios métricos completos y separables, $T : \mathcal{X} \rightarrow \mathcal{Y}$ una aplicación de Borel y $\mu \in \mathcal{P}(\mathcal{X})$. Se define la medida imagen de μ por T o impulso de μ a través de T (en inglés, ‘the push forward of μ through T ’) a la medida $T_{\#}\mu \in \mathcal{P}(\mathcal{Y})$ que verifica:

$$T_{\# \mu}(E) = \mu(T^{-1}(E)), \forall E \subset \mathcal{Y} \text{ conjunto de Borel.}$$

Esta medida se caracteriza por el siguiente hecho:

$$\int f dT_{\# \mu} = \int f \circ T d\mu, \text{ para toda aplicación de Borel } f : \mathcal{Y} \rightarrow \mathbb{R} \cup \{\pm\infty\},$$

donde tal identidad se entiende en el siguiente sentido: una de las integrales existe (posiblemente tomando el valor $\pm\infty$) si, y sólo si, la otra integral existe; y en este caso, los valores de ambas son el mismo.

Denotemos por $\Pi(\mu, \nu)$ el conjunto de probabilidades en el espacio producto $\mathcal{X} \times \mathcal{Y}$ con marginales μ y ν . Dada una medida $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, los siguientes enunciados son equivalentes a la condición $\pi \in \Pi(\mu, \nu)$:

1. Para todos conjuntos medibles $A \subset \mathcal{X}, B \subset \mathcal{Y}$, se tiene que $\pi(A \times \mathcal{Y}) = \mu(A)$ y $\pi(\mathcal{X} \times B) = \nu(B)$.
2. Para todas funciones integrables φ, ψ en \mathcal{X}, \mathcal{Y} , respectivamente, se verifica:

$$\int_{\mathcal{X} \times \mathcal{Y}} (\varphi(x) + \psi(y)) d\pi(x, y) = \int_{\mathcal{X}} \varphi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y).$$

Una observación interesante es que, a partir de dos medidas μ y ν , siempre se puede encontrar una distribución π tal que $\pi \in \Pi(\mu, \nu)$. Tal distribución es la que corresponde a considerar la ley conjunta de dos variables $X \sim \mu$ e $Y \sim \nu$ independientes, que resulta ser el producto tensorial $\mu \otimes \nu$. En el otro extremo, está el caso en que toda la información sobre la variable Y está contenida en la variable X . Más formalmente, existe una función medible $T : \mathcal{X} \rightarrow \mathcal{Y}$ tal que $Y = T(X)$. Este hecho se puede enunciar de las siguientes maneras equivalentes:

1. La ley π de (X, Y) está concentrada en el grafo de una función medible $T : \mathcal{X} \rightarrow \mathcal{Y}$.
2. T es un cambio de variable de μ a ν , es decir, para toda φ función ν -integrable,

$$\int_{\mathcal{Y}} \varphi(y) d\nu(y) = \int_{\mathcal{X}} \varphi(T(x)) d\mu(x).$$

A esta aplicación T se le conoce con el nombre de *transporte*. Informalmente, se podría decir que T transporta la masa representada por la medida μ a la masa representada por la medida ν . Veamos un ejemplo fácil para entender mejor estos conceptos.

Ejemplo 1.1. Sean μ y ν dos medidas de probabilidad en \mathbb{R} . Definimos sus funciones de distribución

$$F(x) = \int_{-\infty}^x d\mu, \quad G(y) = \int_{-\infty}^y d\nu,$$

y sus funciones cuantiles

$$\begin{aligned} F^{-1}(t) &= \inf\{x \in \mathbb{R} : F(x) > t\} \\ G^{-1}(t) &= \inf\{y \in \mathbb{R} : G(y) > t\}. \end{aligned}$$

Consideremos $T = G^{-1} \circ F$. Entonces, si μ no da probabilidad a puntos, se tiene que $T_{\#}\mu = \nu$.

1.1.1 El problema del transporte óptimo

Sean \mathcal{X} e \mathcal{Y} dos espacios métricos completos y separables, y consideremos una función de Borel $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{+\infty\}$ que denominaremos *función de coste*, pues el valor $c(x, y)$ se interpreta como el trabajo necesario para mover una unidad de masa desde la posición x hasta la posición y . El problema de minimización de Kantorovich corresponde a calcular

$$\inf E(c(X, Y)) \tag{1.1}$$

donde el par (X, Y) recorre el conjunto de vectores aleatorios con primera marginal μ y segunda marginal ν , o, equivalentemente,

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y),$$

donde π recorre el conjunto $\Pi(\mu, \nu)$. Tales medidas conjuntas se llaman planes de transferencia o transporte, y aquellas en las que se alcanza el ínfimo se conocen como *planes de transferencia óptimos*. En [28] puede verse que existe al menos una medida cumpliendo esta propiedad.

Observación 1.1. La formulación de Kantorovich es una relajación del problema original de Monge, que supone que la masa no se puede dividir, o, en otras palabras, que a cada localización x se le asocia un único destino y . Matemáticamente, esto consiste en exigir que Y sea una función de X en (1.1).

La solución al problema de Kantorovich depende claramente de la función de coste c . En particular, cuando c se escribe en términos de distancias (y $\mathcal{X} = \mathcal{Y}$), este problema da lugar a una métrica con muy buenas propiedades, que es la llamada métrica de Wasserstein.

1.1.2 La distancia de Wasserstein

La métrica de Wasserstein surge a partir del estudio del problema de transporte óptimo, y ha resultado ser una forma adecuada para comparar las distribuciones de probabilidad de una variable X y de otra Y , que se ha obtenido a partir de una transformación no uniforme en X , aleatoria o determinista. Fue propuesta por primera vez con este objetivo de medir distancias entre distribuciones de probabilidad por Kantorovich y Rubinstein en [19].

A pesar de que la distancia de Wasserstein ha resultado ser muy potente como herramienta teórica, su cálculo explícito no es inmediato en la mayoría de los casos. Un conocido resultado debido a Vallander en [27], del cual veremos una versión más adelante, facilita este cálculo cuando consideramos distribuciones de probabilidad en la recta real. Por otro lado, en el trabajo [9] se da el valor explícito de la distancia L^2 de Wasserstein entre dos distribuciones normales n -dimensionales cualesquiera, y se muestra que, en este caso, el problema puede ser reducido a calcular raíces cuadradas de matrices semidefinidas positivas.

Definición 1.2. Sea (E, d) un espacio métrico, se define el espacio de Wasserstein de orden r como

$$\mathcal{W}_r(E) = \left\{ \mu \in \mathcal{P}(E) : \int_E d(x_0, x)^r d\mu(x) < +\infty \right\},$$

donde el punto $x_0 \in E$ es arbitrario.

Como consecuencia directa de la desigualdad triangular y de las propiedades de linealidad y monotonía de la integral, es fácil ver que el espacio \mathcal{W}_r está bien definido, es decir, que no depende de la elección del punto $x_0 \in E$. En efecto, dados $x_0, y_0 \in E$, tales que $x_0 \neq y_0$,

$$\begin{aligned} \int_E d(y_0, x)^r d\mu(x) &\leq \int_E d(y_0, x_0)^r d\mu(x) + \int_E d(x_0, x)^r d\mu(x) \\ &= d(y_0, x_0)^r + \int_E d(x_0, x)^r d\mu(x), \end{aligned}$$

y por lo tanto,

$$\int_E d(y_0, x)^r d\mu(x) < +\infty \iff \int_E d(x_0, x)^r d\mu(x) < +\infty.$$

El espacio de Wasserstein $\mathcal{W}_r(E)$ es, en otras palabras, el conjunto de las medidas de probabilidad sobre E con momento de orden r finito.

| Definición 1.3. Dadas $\mu, \nu \in \mathcal{W}_r(E)$, se define la distancia de Wasserstein con coste L_r entre μ y ν como:

$$W_r(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int d(x, y)^r d\pi(x, y) \right)^{\frac{1}{r}},$$

donde $\Pi(\mu, \nu)$ el conjunto de probabilidades sobre $E \times E$ con primera marginal μ y segunda marginal ν .

Observemos que la distancia de Wasserstein de orden r es la raíz r -ésima del coste mínimo de transporte con función de coste $c(x, y) = d(x, y)^r$. Notemos además que el hecho de definirla en el espacio $\mathcal{W}_r(E)$ garantiza que sea finita, pues

$$d(x, y)^r \leq 2^{r-1} (d(x_0, x)^r + d(x_0, y)^r)$$

y, en consecuencia,

$$W_r(\mu, \nu)^r \leq 2^{r-1} \left(\int d(x_0, x)^r d\mu(x) + \int d(x_0, y)^r d\nu(y) \right).$$

Observación 1.2. A W_r también se le conoce en la literatura como la distancia de Monge-Kantorovich de orden r , o con exponente r . Cuando $r = 2$, a W_2 se le llama distancia cuadrática de Wasserstein. La distancia de Monge-Kantorovich con exponente 1, $W_1 = \mathcal{T}_1$, recibe el nombre de distancia de Kantorovich-Rubinstein.

Todos los resultados que aparecen a continuación son ciertos en espacios métricos más generales. Sin embargo, a partir de ahora, nos centraremos en el caso particular en el cual $E = \mathbb{R}^d$, $d \geq 1$ y $\|\cdot\|$ es la distancia Euclídea habitual, pues éste será el marco en el que se desarrollará el resto del trabajo. Observemos que, entonces, la distancia de Wasserstein entre dos probabilidades $\mu, \nu \in \mathcal{W}_r(\mathbb{R}^d)$ se puede escribir de la siguiente manera:

$$W_r^r(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int \|x - y\|^r d\pi(x, y). \quad (1.2)$$

Lema 1.1 (Existencia del plan de transferencia óptimo). Dadas $\mu, \nu \in \mathcal{W}_r(\mathbb{R}^d)$, $1 \leq r < \infty$, el inferior en (1.2) se alcanza para alguna medida $\pi \in \Pi(\mu, \nu)$.

Este resultado lo hemos anticipado en el contexto general del problema del transporte óptimo y su demostración puede encontrarse en [28]. Como ya hemos comentado, a esta medida π se le conoce como *plan de transferencia óptimo*.

Observación 1.3. Notemos que (1.2) se puede escribir de manera equivalente en términos de vectores aleatorios:

$$W_r^r(\mu, \nu) = \inf_{(X,Y): X \sim \mu, Y \sim \nu} E(\|X - Y\|^r). \quad (1.3)$$

De esta manera, el vector aleatorio (X, Y) en el cual se alcanza el inferior (1.3) se llamará *par óptimo* para las medidas μ y ν . (en inglés, se dice que “ (X, Y) is an optimal coupling of μ and ν ”).

El siguiente teorema, cuya demostración puede encontrarse en [28] con mayor generalidad, garantiza que W_r dota al espacio $\mathcal{W}_r(\mathbb{R}^d)$ de estructura de espacio métrico.

Teorema 1.1. Para $r \geq 1$, la distancia de Wasserstein $W_r(\mu, \nu)$ define una métrica en el espacio $\mathcal{W}_r(\mathbb{R}^d)$, $d \geq 1$.

Observación 1.4. Una propiedad fundamental de las distancias de Wasserstein es que están ordenadas. En efecto, como consecuencia inmediata de la desigualdad de Hölder, se tiene que

$$1 \leq p \leq q \implies W_p \leq W_q.$$

En particular, $W_1(\mu, \nu) \leq W_p(\mu, \nu)$, para todo $p \geq 1$. Sin embargo, no es posible establecer una desigualdad en el otro sentido, a no ser que consideremos medidas en un subconjunto $B \subset \mathbb{R}^d$ acotado, pues en este caso se tendría que:

$$\left(\int \|x - y\|^p d\pi \right)^{\frac{1}{p}} \leq \text{diam}(B)^{\frac{p}{p-1}} \left(\int \|x - y\| d\pi \right)^{\frac{1}{p}},$$

lo cual implica que $W_p(\mu, \nu) \leq C W_1(\mu, \nu)^{\frac{1}{p}}$, para $C = \text{diam}(B)^{\frac{p}{p-1}}$ y donde $\text{diam}(B) = \sup\{d(x, y) : x, y \in B\}$ es el diámetro de B . En conclusión, si $B \subset \mathbb{R}^d$ es acotado, todas las distancias de Wasserstein W_r , $r \geq 1$, definen la misma topología sobre $\mathcal{P}(B)$.

Convergencia en la distancia de Wasserstein

Veamos ahora una caracterización de la convergencia en el espacio de Wasserstein. El siguiente resultado se puede encontrar con mayor generalidad en [28]. La notación $\mu_k \rightarrow \mu$ significa que $(\mu_k)_{k \in \mathbb{N}}$ converge débilmente hacia μ , es decir, que $\int \varphi d\mu_k \rightarrow \int \varphi d\mu$, para cualquier función φ continua y acotada.

| Teorema 1.2. Sea $r \in (0, \infty)$ y consideremos $(\mu_k)_{k \in \mathbb{N}}$ una sucesión de medidas de probabilidad en $\mathcal{W}_r(\mathbb{R}^d)$ y μ otro elemento de $\mathcal{W}_r(\mathbb{R}^d)$, $d \geq 1$. Entonces, los siguientes enunciados son equivalentes:

(i) $W_r(\mu_k, \mu) \xrightarrow[k \rightarrow \infty]{} 0$.

(ii) $\mu_k \rightarrow \mu$ en el sentido débil, y $(\mu_k)_{k \in \mathbb{N}}$ satisface la siguiente condición de equicontinuidad en el infinito (en inglés, "tightness"): para algún (y por tanto, para todo) $x_0 \in \mathbb{R}^d$,

$$\lim_{R \rightarrow \infty} \left(\limsup_{k \rightarrow \infty} \int_{\|x_0 - x\| \geq R} \|x_0 - x\|^r d\mu_k(x) \right) = 0$$

(iii) $\mu_k \rightarrow \mu$ en el sentido débil, y hay convergencia del momento de orden r : para algún $x_0 \in \mathbb{R}^d$ (y, por tanto, para todo $x_0 \in \mathbb{R}^d$),

$$\int \|x_0 - x\|^r d\mu_k(x) \xrightarrow[k \rightarrow \infty]{} \int \|x_0 - x\|^r d\mu(x).$$

(iv) Para toda función φ continua en \mathbb{R}^d que satisface la condición de crecimiento: $|\varphi(x)| \leq C[1 + \|x_0 - x\|^r]$, para algún $x_0 \in \mathbb{R}^d$ y cierta constante $C \in \mathbb{R}$,

$$\int \varphi d\mu_k \xrightarrow[k \rightarrow \infty]{} \int \varphi d\mu.$$

La distancia de Wasserstein en la recta real ($E = \mathbb{R}$)

La distancia de Wasserstein entre probabilidades en \mathbb{R} adquiere una expresión más simple, que la hace útil en muchas situaciones y le confiere propiedades particulares. El siguiente teorema es fundamental en la teoría del transporte óptimo y de él se pueden encontrar numerosas demostraciones, siendo presentado por primera vez en [27]. Aquí enunciamos la versión de Villani [28].

| Teorema 1.3 (Teorema del transporte óptimo con coste cuadrático en \mathbb{R}). Sean μ y ν dos medidas de probabilidad en \mathbb{R} , con funciones de distribución F y G , respectivamente. Sea π la medida de probabilidad sobre \mathbb{R}^2 con función de distribución conjunta

$$H(x, y) = \min(F(x), G(y))$$

Entonces, $\pi \in \Pi(\mu, \nu)$ y π es óptima en el problema de transporte de Kantorovich entre μ y ν para la función de coste cuadrático $c(x, y) = |x - y|^2$.

Además, el valor del coste de transporte óptimo es

$$\mathcal{T}_2(\mu, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt,$$

donde F^{-1} y G^{-1} son las funciones cuantiles de F y G , respectivamente:

$$\begin{aligned} F^{-1}(t) &= \inf\{x \in \mathbb{R} : F(x) > t\} \\ G^{-1}(t) &= \inf\{x \in \mathbb{R} : G(x) > t\} \end{aligned}$$

En particular, esto implica que la distancia de Wasserstein entre dos probabilidades $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$ se puede escribir como:

$$W_2(\mu, \nu) = \left(\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt \right)^{\frac{1}{2}}.$$

Esta expresión, como veremos, juega un papel trascendental en todo este trabajo.

1.2 Procesos empíricos y cuantiles

En estadística, la función de distribución empírica es la función de distribución asociada a la medida empírica de la muestra, y es constante a trozos con saltos de magnitud $\frac{1}{n}$ en cada uno de los n puntos de la muestra.

Definición 1.4. Sean X_1, \dots, X_n variables aleatorias reales independientes e igualmente distribuidas (v.a.i.i.d.) con función de distribución $F(t)$. La función de distribución empírica se define como

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq t), \quad t \in \mathbb{R}.$$

Esta función estima a la función de distribución poblacional que subyace en la muestra. En efecto, para cada $t \in \mathbb{R}$ fijo, el indicador $I(X_i \leq t)$ es una variable aleatoria de Bernoulli de parámetro $p = F(t)$ y, por lo tanto, $nF_n(t)$ es una variable aleatoria binomial con media $nF(t)$ y varianza $n(F(t))(1 - F(t))$. Esto implica que $F_n(t)$ es un estimador insesgado de $F(t)$.

Existen numerosos resultados que cuantifican la velocidad de convergencia de la función de distribución empírica hacia la función de distribución, de los cuales mencionamos los dos más importantes. Por la Ley Fuerte de los Grandes Números, se tiene que el estimador $F_n(t)$ converge casi seguro hacia $F(t)$, para cada valor de $t \in \mathbb{R}$:

$$F_n(t) \xrightarrow{c.s.} F(t), \text{ cuando } n \longrightarrow +\infty,$$

lo cual implica que $F_n(t)$ es un estimador consistente de $F(t)$. El Teorema de Glivenko-Cantelli prueba que esta convergencia ocurre uniformemente en t :

$$\|F_n - F\|_\infty = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{c.s.} 0, \text{ cuando } n \longrightarrow +\infty.$$

1.2.1 El proceso empírico

Sean U_1, \dots, U_n v.a.i.i.d. $\mathcal{U}(0, 1)$, con función de distribución común $G(t) = t$ y función de distribución empírica $G_n(t)$.

| Definición 1.5. *En esta situación, definimos el proceso empírico uniforme $\alpha_n(t) := \sqrt{n}(G_n(t) - G(t)) = \sqrt{n}(G_n - t), 0 \leq t \leq 1$.*

Veamos que $\alpha_n(t) \xrightarrow{d} \mathcal{N}(0, t(1-t))$, para cada $0 \leq t \leq 1$. En primer lugar, es claro que

$$\begin{aligned} \alpha_n(t) &= \sqrt{n}(G_n(t) - t) = \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n (I(U_i \leq t) - P(U_i \leq t)) \right] \\ &= \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n (I(U_i \leq t) - P(U_i \leq t)) \right]. \end{aligned}$$

Consideremos, para cada $i \in 1, \dots, n$, la variable aleatoria $Z_i = I(U_i \leq t) - P(U_i \leq t) = I(U_i \leq t) - t$, por ser $U_i \sim \mathcal{U}(0, 1)$. Sabemos que la variable aleatoria $I(U_i \leq t)$ tiene una distribución de Bernoulli de parámetro $P(U_i \leq t) = t$. Por lo tanto, Z_1, \dots, Z_n resulta ser una colección de v.a.i.i.d. centradas. Además, en virtud del Teorema Central del Límite,

$$\frac{1}{\sqrt{n}} \left[\sum_{i=1}^n (I(U_i \leq t) - P(U_i \leq t)) \right] \xrightarrow{d} \mathcal{N}(0, \text{Var}(Z_i)),$$

donde $\text{Var}(Z_i) = \text{Var}(I(U_i \leq t)) = t(1-t)$.

En general, dados dos instantes de tiempo $0 \leq s < t \leq 1$, se tiene que:

$$\begin{aligned}
& Cov(\alpha_n(s), \alpha_n(t)) \\
&= Cov\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (I(U_i \leq s) - P(U_i \leq s)), \frac{1}{\sqrt{n}} \sum_{i=1}^n (I(U_i \leq t) - P(U_i \leq t))\right) \\
&= Cov\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (I(U_i \leq s) - ns), \frac{1}{\sqrt{n}} \sum_{i=1}^n (I(U_i \leq t) - nt)\right) \\
&= Cov\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (I(U_i \leq s) - \sqrt{ns}), \frac{1}{\sqrt{n}} \sum_{i=1}^n (I(U_i \leq t) - \sqrt{nt})\right) \\
&= Cov\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n I(U_i \leq s), \frac{1}{\sqrt{n}} \sum_{i=1}^n I(U_i \leq t)\right) = \frac{1}{n} Cov\left(\sum_{i=1}^n I(U_i \leq s), \sum_{i=1}^n I(U_i \leq t)\right).
\end{aligned}$$

Por otro lado, como U_1, \dots, U_n son independientes, se verifica que $Cov(U_i, U_j) = 0$, si $i \neq j$. Luego,

$$Cov(\alpha_n(s), \alpha_n(t)) = \frac{1}{n} \sum_{i=1}^n Cov(I(U_i \leq s), I(U_i \leq t)) = Cov(I(U_1 \leq s), I(U_1 \leq t)),$$

donde la última igualdad es consecuencia de que U_1, \dots, U_n son igualmente distribuidas. Ahora, por definición de covarianza,

$$\begin{aligned}
Cov(I(U_1 \leq s), I(U_1 \leq t)) &= E(I(U_1 \leq t)I(U_1 \leq s)) - E(I(U_1 \leq t))E(I(U_1 \leq s)) \\
&= E(I((U_1 \leq t) \cup (U_1 \leq s))) - ts = E(I(U_1 \leq s \wedge t)) - ts = s \wedge t - st.
\end{aligned}$$

Consideremos ahora X_1, \dots, X_n variables aleatorias independientes e igualmente distribuidas con función de distribución F . Sabemos que

$$(X_1, \dots, X_n) \stackrel{d}{=} (F^{-1}(U_1), \dots, F^{-1}(U_n)),$$

donde U_1, \dots, U_n son v.a.i.i.d $\mathcal{U}(0, 1)$. Análogamente al proceso empírico uniforme, definimos

$$\alpha_n^F(t) := \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n (I(X_i \leq t) - F(t)) \right), \text{ para } t \in \mathbb{R}.$$

La propiedad de la función cuantil:

$$F^{-1}(u) \leq t \Leftrightarrow u \leq F(t) \text{ para cada } u \in (0, 1),$$

equivale a escribir:

$$I(X_1 \leq t) = I(\mathcal{U}_1 \leq F(t)) \text{ para cada } t \in \mathbb{R}.$$

Por lo tanto, el proceso empírico asociado a X_1, \dots, X_n en un instante de tiempo $t \in \mathbb{R}$, resulta ser el proceso empírico uniforme en el instante de tiempo $F(t)$:

$$\begin{aligned} \alpha_n^F(t) &= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n (I(X_i \leq t) - F(t)) \right) \stackrel{d}{=} \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n (I(\mathcal{U}_1 \leq F(t)) - F(t)) \right) \\ &= \alpha_n(F(t)), t \in \mathbb{R}. \end{aligned} \quad (1.4)$$

El estadístico de Kolmogorov

Un ejemplo de aplicación de lo que acabamos de ver aparece en el cálculo de la distribución asintótica del estadístico de Kolmogorov-Smirnov. El test de Kolmogorov-Smirnov (Test K-S) es un test no paramétrico de igualdad de distribuciones unidimensionales y con función de distribución continua. Este test tiene dos versiones, el test K-S de una muestra, que permite comparar una colección de observaciones con una distribución de referencia, y el test K-S de dos muestras, que se usa para comparar dos muestras entre sí.

Sean X_1, \dots, X_n v.a.i.i.d con función de distribución desconocida F y sea F_0 una función continua en \mathbb{R} . El contraste de hipótesis que se plantea en el test K-S de una muestra es el siguiente:

$$\begin{aligned} \mathcal{H}_0 &: F = F_0, \\ \text{contra la alternativa:} \\ \mathcal{H}_a &: F \neq F_0. \end{aligned}$$

El estadístico de K-S para una función de distribución F_0 dada es

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

Entonces, si \mathcal{H}_0 es cierta, se tiene que $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$, y por el Teorema de Glivenko-Cantelli, $D_n \rightarrow 0$ casi seguro. En la sección 1.3 comentamos un refinamiento de este resultado, para lo cual es útil observar que $D_n = \|\alpha_n(F_0)\|_\infty$.

1.2.2 El proceso cuantil

Sean U_1, \dots, U_n v.a.i.i.d. $\mathcal{U}(0, 1)$, con función de distribución común $G(t) = t$. Sean $G_n(t)$ y $G_n^{-1}(t)$ la función de distribución empírica y la función cuantil empírica, respectivamente, asociadas a la muestra U_1, \dots, U_n .

| Definición 1.6. En esta situación, se define el proceso cuantil uniforme por $\rho_n(t) := \sqrt{n} \left(t - G_n^{-1}(t) \right)$, $0 < t < 1$

Sean X_1, \dots, X_n v.a.i.i.d. con función de distribución común F , función de distribución empírica F_n y función cuantil empírica F_n^{-1} . El proceso cuantil asociado a X_1, \dots, X_n viene dado por:

$$\rho_n^F(t) := \sqrt{n} \left(F_n^{-1}(t) - F^{-1}(t) \right).$$

En este contexto, denotemos por ν a la medida de probabilidad subyacente a la muestra X_1, \dots, X_n y por ν_n a la correspondiente medida empírica. Existe la siguiente relación entre la distancia de Wasserstein entre una medida de probabilidad y su correspondiente versión empírica asociada a una muestra dada, y el proceso cuantil relativo a dicha muestra:

$$\begin{aligned} W_2^2(\nu_n, \nu) &= \int_0^1 (F_n^{-1} - F^{-1})^2 dt = \frac{1}{n} \int_0^1 \left[\sqrt{n} (F_n^{-1} - F^{-1}) \right]^2 dt \\ &\implies nW_2^2(\mu_n, \mu) = \int_0^1 (\rho_n^F)^2 dt. \end{aligned}$$

1.3 El movimiento Browniano

En numerosos resultados asintóticos en estadística, las distribuciones límite se expresan bien en términos de procesos Gaussianos, entre los cuales destaca el Puente Browniano. Consideraremos el espacio probabilístico $(\Omega, \mathcal{F}, \mathbb{P})$.

| Definición 1.7. El proceso estocástico $\{B(t)\}_{t \geq 0}$ es un Movimiento Browniano (MB) o Proceso de Wiener con varianza σ^2 en el espacio $(\Omega, \mathcal{F}, \mathbb{P})$, si cumple las siguientes condiciones:

1. $B(0) = 0$
2. Tiene incrementos independientes, es decir, para cualquier conjunto de instantes de tiempo $0 \leq t_1 < t_2 < \dots < t_n$, se tiene que $B(t_1), B(t_2) - B(t_1), \dots, B(t_n) - B(t_{n-1})$ son variables aleatorias independientes.
3. Si $s < t$, entonces la variable aleatoria $B(t) - B(s)$ tiene una distribución $\mathcal{N}(0, \sigma^2(t - s))$
4. Las trayectorias del proceso son funciones continuas, $t \mapsto B(t)$.

Si $\sigma = 1$ decimos que el movimiento es estándar. Notemos que si $B(t)$ es estándar entonces $\beta B(t)$ es un MB con varianza β^2 . Luego, a partir de ahora, asumiremos sin pérdida de generalidad que $B(t)$ es estándar.

Observación 1.5 (Algunas propiedades del MB).

1. Una propiedad útil del MB es la de reescalamiento: dado $\alpha > 0$ se tiene que:

$$\{B(\alpha t) : t \geq 0\} = \{\sqrt{\alpha}B(t) : t \geq 0\}.$$

2. El MB es un caso particular de los procesos gaussianos: las distribuciones finito dimensionales de un MB son normales multivariantes, es decir, cualesquiera que sean $0 \leq t_1 < t_2 < \dots < t_k \leq 1$,

$$(B(t_1), \dots, B(t_k)) \sim \mathcal{N}_k(\mu, \Sigma).$$

En efecto, observemos que

$$\begin{pmatrix} B(t_1) \\ B(t_2) \\ \vdots \\ B(t_k) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} B(t_1) \\ B(t_2) - B(t_1) \\ \vdots \\ B(t_k) - B(t_{k-1}) \end{pmatrix},$$

y, por lo tanto, la ley del vector aleatorio $(B(t_1), \dots, B(t_k))$ es normal por ser una transformación lineal del vector

$$(B(t_1), B(t_2) - B(t_1), \dots, B(t_k) - B(t_{k-1})),$$

que tiene ley normal ya que sus componentes son normales independientes.

3. La media y la autocovarianza de un MB se calculan fácilmente:

$$E(B(t)) = 0, \text{ y}$$

$$\begin{aligned} \text{Cov}(B(s), B(t)) &= E(B(s)B(t)) - E(B(s))E(B(t)) = E(B(s)B(t)) \\ &= E(B(s)(B(t) - B(s) + B(s))) = E(B(s)(B(t) - B(s))) + E(B(s)^2) \\ &= E(B(s))E(B(t) - B(s)) + \text{Var}(B(s)) = 0 + s = \min(s, t), \text{ si } s \leq t. \end{aligned}$$

$$\begin{aligned} \text{De manera análoga, } \text{Cov}(B(s), B(t)) &= E(B(t)(B(s) - B(t) + B(t))) \\ &= E(B(t)(B(s) - B(t))) + E(B(t)^2) = E(B(t))E(B(s) - B(t)) + \text{Var}(B(t)) \\ &= t = \min(s, t), \text{ si } s \geq t. \end{aligned}$$

Por lo tanto, $\text{Cov}(B(s), B(t)) = \min(s, t)$, $\forall 0 \leq s, t \leq 1$. En conclusión, si $0 \leq t_1 < \dots < t_k \leq 1$,

$$(B(t_1), \dots, B(t_k)) \sim \mathcal{N}_k \left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} t_1 & t_1 & \dots & t_1 \\ t_1 & t_2 & \dots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & \dots & t_k \end{bmatrix} \right).$$

Definición 1.8. Se dice que el proceso estocástico $\{X(t)\}_{0 \leq t \leq 1}$ es un *Puente Browniano (PB)* si tiene trayectorias continuas y sus distribuciones finito dimensionales son normales, centradas y con $Cov(X(s), X(t)) = s \wedge t - st$.

El PB es un proceso que se genera a partir del MB. En efecto, veamos que si $\{W(t)\}_{t \geq 0}$ es un MB, entonces el proceso estocástico $\{B(t)\}_{t \geq 0}$ dado por $B(t) = W(t) - tW(1), t \geq 0$ es un PB. Para ello, dados $0 \leq t_1 < \dots < t_k \leq 1$, tenemos que comprobar que se satisfacen las siguientes condiciones:

1. $(B(t_1), \dots, B(t_k))$ es normal
2. $(B(t_1), \dots, B(t_k))$ es centrada
3. $Cov(B(s), B(t)) = s \wedge t - st$.

Las condiciones 1 y 2 son consecuencia de que

$$\begin{pmatrix} B(t_1) \\ B(t_2) \\ \vdots \\ B(t_k) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -t \\ 0 & 1 & 0 & \cdots & 0 & -t \\ 0 & 0 & 1 & \cdots & 0 & -t \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -t \end{pmatrix} \begin{pmatrix} W(t_1) \\ W(t_2) \\ \vdots \\ W(t_k) \\ W(1) \end{pmatrix},$$

y del hecho de que $(W(t_1), \dots, W(t_k), W(1))$ es normal y centrado, por ser W un MB. Por último, dados $0 \leq s, t \leq 1$,

$$\begin{aligned} Cov(B(s), B(t)) &= Cov(W(s) - sW(1), W(t) - tW(1)) \\ &= Cov(W(s), W(t)) - sCov(W(1), W(t)) - tCov(W(s), W(1)) + stCov(W(1), W(1)) \\ &= s \wedge t - s(1 \wedge t) - t(s \wedge 1) + st = s \wedge t - st - st + st = s \wedge t - st. \end{aligned}$$

Observación 1.6 (Convergencia del proceso empírico, del estadístico de Kolmogorov y del proceso cuantil hacia PB). Por lo que acabamos de ver, deducimos que las distribuciones finito dimensionales del proceso empírico uniforme α_n convergen en distribución hacia las distribuciones finito dimensionales de un PB. Más generalmente, teniendo en cuenta (1.4), las distribuciones finito dimensionales del proceso empírico $\{\alpha_n^F\}_{n \geq 0}$ convergen en distribución hacia las distribuciones finito dimensionales del proceso estocástico $B \circ F$.

El proceso α_n se puede interpretar como un elemento aleatorio con valores en el espacio $D[0, 1]$ formado por las funciones de $[0, 1]$ en \mathbb{R} que son continuas por la derecha con límites por la izquierda. En $D[0, 1]$ se puede considerar la métrica de Skorohod

$$d(x, y) = \inf \left\{ \varepsilon > 0 : \inf_{\lambda \in \Lambda} \left[\sup_{t \in [0,1]} |\lambda(t) - t| \leq \varepsilon \right] \right\},$$

donde $\Lambda = \{\lambda : [0, 1] \rightarrow [0, 1] \text{ tal que } \lambda(0) = 0, \lambda(1) = 1, \lambda \text{ continua y estrictamente creciente}\}$.

Un resultado clásico en estadística es que α_n converge en distribución al puente Browniano B , en el sentido de que si $H : D[0, 1] \rightarrow \mathbb{R}$ es continua en $\mathcal{C}[0, 1]$, entonces $H(\alpha_n) \xrightarrow{d} H(B)$ en el sentido usual (ver Capítulo 3 en [6]). Como la aplicación $x \mapsto \|x\|_\infty$ es continua, se deduce que

$$D_n = \|\alpha_n(F_0)\|_\infty \xrightarrow{d} \|B \circ F_0\|_\infty = \sup_{t \in \mathbb{R}} |B(F_0(t))|.$$

Como además F_0 es continua

$$D_n \xrightarrow{d} \|B\|_\infty = \sup_{t \in [0,1]} |B(t)|,$$

que no depende de F_0 . Este último resultado se conoce como Teorema de Kolmogorov.

La convergencia débil de funcionales del proceso empírico y del proceso cuantil hacia los funcionales correspondientes del puente Browniano se da también para funcionales continuos respecto a otras topologías. Algunos resultados de este tipo serán enunciados durante la memoria (véase, por ejemplo, (2.4) en la sección 2.3).

Concluimos esta sección con dos proposiciones de gran utilidad a la hora de establecer los resultados asintóticos que veremos a lo largo del trabajo, concretamente en la sección 2.2.

Proposición 1.1. Sean $\{B(t)\}_{t \in I}$ un PB y una función f continua en $I = [0, 1]$. Se tiene que

$$\int_0^1 B(t)f(t) \rightsquigarrow \mathcal{N} \left(0, \int_{I \times I} (s \wedge t - st) f(s)f(t) ds dt \right).$$

Demostración. Si consideramos la integral del enunciado como una integral de Riemann, podemos escribir

$$\int_0^1 B(t)f(t) dt = \lim_{n \rightarrow \infty} \left[\frac{1}{n} \sum_{i=1}^n B\left(\frac{i}{n}\right) f\left(\frac{i}{n}\right) \right].$$

Ahora, observemos que por ser B un PB, el vector aleatorio

$$\left(B\left(\frac{1}{n}\right), B\left(\frac{2}{n}\right), \dots, B\left(\frac{n-1}{n}\right), B(1) \right)$$

tiene distribución normal, y como $\forall i = 1, \dots, n$ y $\forall n \in \mathbb{N}$, $f\left(\frac{i}{n}\right)$ es un valor fijo, se tiene que $\frac{1}{n} \sum_{i=1}^n B\left(\frac{i}{n}\right) f\left(\frac{i}{n}\right)$ es una variable aleatoria normal.

Concluimos que $\int_0^1 B(t)f(t)dt$ es normal por ser límite de una sucesión de v.a. normales. Además,

$$\begin{aligned} E\left(\int_0^1 |B(t)||f(t)| dt\right) &= \int_0^1 E(|B(t)||f(t)|) dt = \int_0^1 |f(t)| E|B(t)| dt \\ &\leq \int_0^1 |f(t)| \sqrt{E(B(t)^2)} dt = \int_0^1 \sqrt{\frac{2}{\pi}} \sqrt{t(1-t)} |f(t)| dt \\ &\leq \sqrt{\frac{2}{\pi}} \left(\int_0^1 t(1-t) dt\right)^{\frac{1}{2}} \left(\int_0^1 |f(t)|^2 dt\right)^{\frac{1}{2}} < +\infty, \end{aligned}$$

donde la primera desigualdad es consecuencia de la desigualdad de Jensen, y el penúltimo paso es consecuencia de la desigualdad de Cauchy-Schwarz. Luego, por el Teorema de Fubini, está justificado el siguiente intercambio en el orden de integración en los siguientes cálculos

$$\begin{aligned} E\left(\int_0^1 B(t)f(t)dt\right) &= \int_0^1 E(B(t))f(t)dt = 0, \\ \text{Var}\left(\int_0^1 B(t)f(t)dt\right) &= E\left[\left(\int_0^1 B(t)f(t)dt\right)^2\right] - \left[E\left(\int_0^1 B(t)f(t)dt\right)\right]^2 \\ &= E\left[\left(\int_0^1 B(t)f(t)dt\right)\left(\int_0^1 B(t)f(t)dt\right)\right] = E\left(\int_0^1 \int_0^1 B(t)B(s)f(t)f(s)dt ds\right) \\ &= \int_0^1 \int_0^1 f(t)f(s)E(B(t)B(s)) dt ds. \end{aligned}$$

Como $E(B(t)) = E(B(s)) = 0$, $E(B(t)B(s)) = \text{Cov}(B(t)B(s)) = s \wedge t - st$ y, en consecuencia,

$$\text{Var}\left(\int_0^1 B(t)f(t)dt\right) = \int_0^1 \int_0^1 (s \wedge t - st)f(t)f(s)dt ds.$$

■

Proposición 1.2. Si $\{B(t)\}_{t \in I}$ un PB, existen constantes reales, positivas y no nulas $\{\lambda_n\}_{n=0}^\infty$ tales que

$$\int_0^1 B(t)^2 dt \stackrel{d}{=} \sum_{n=0}^{\infty} \lambda_n Z_n^2, \text{ donde } Z_n \sim \mathcal{N}(0, \lambda_n), \text{ para cada } n \in \mathbb{N}.$$

Demostración. Sea $\{X(t)\}_{t \in I}$, donde $I = [0, 1]$ y $X(\cdot) \in L^2(I)$ es un proceso Gaussiano centrado. La función de covarianza del proceso viene dada por

$$\begin{aligned} k : I \times I &\longrightarrow \mathbb{R} \\ (s, t) &\longmapsto k(s, t) = E(X(s)X(t)). \end{aligned}$$

Si consideremos $\{\varphi_n\}_{n \geq 0}$ un sistema ortonormal y completo en $L^2(I)$, entonces, el proceso X se puede escribir como

$$X = \sum_{n=0}^{\infty} \langle X, \varphi_n \rangle \varphi_n.$$

Además, por la Identidad de Parseval,

$$\|X\|^2 = \int_I X^2 dt = \sum_{n=0}^{\infty} \langle X, \varphi_n \rangle^2.$$

Definamos ahora el operador

$$\begin{aligned} \mathcal{K} : L^2(I) &\longrightarrow L^2(I) \\ f &\longmapsto (\mathcal{K}f)(t) = \int_I k(s, t)f(s)ds. \end{aligned}$$

Supongamos que $\{\psi_n\}_{n \geq 0}$ es un sistema ortonormal completo formado por autofunciones del operador \mathcal{K} , es decir, $\mathcal{K}\psi_n = \lambda_n\psi_n$, para todo $n \geq 0$. Veamos que si $X_n = \langle X, \psi_n \rangle$, entonces $X_n \sim \mathcal{N}(0, \lambda_n)$, para cada $n \geq 0$. En primer lugar, X_n es centrada:

$$\begin{aligned} X_n = \langle X, \psi_n \rangle &= \int_I X(t)\psi_n(t)dt \\ \implies E(X_n) &= E\left(\int_I X(t)\psi_n(t)dt\right) = \int_I E(X(t))\psi_n(t)dt = 0. \end{aligned}$$

De aquí se sigue que:

$$\begin{aligned} \text{Var}(X_n) &= E(X_n^2) - (E(X_n))^2 = E(X_n^2) = E\left(\int_{I \times I} X(s)X(t)\psi_n(s)\psi_n(t) ds dt\right) \\ &= \int_I \left(\int_I k(s,t)\psi_n(s) ds\right) \psi_n(t) dt = \int_I \lambda_n \psi_n(t)\psi_n(t) dt = \lambda_n \int_I \psi_n^2(t) dt \\ &= \lambda_n \|\psi_n\|^2 = \lambda_n \end{aligned}$$

Por último, X_n es normal por ser límite de sumas de variables aleatorias normales y con esto queda probado que $X_n \sim \mathcal{N}(0, \lambda_n)$.

Además, observemos que si $n \neq m$,

$$\text{Cov}(X_n, X_m) = \int_I \left(\int_I k(s,t)\psi_n(s) ds\right) \psi_m(t) dt = \lambda_n \int_I \psi_n(s)\psi_m(t) dt = 0.$$

En consecuencia, $\{X_n\}_{n=0}^\infty$ son variables aleatorias normales independientes.

Por otro lado, como $X_n \sim \mathcal{N}(0, \lambda_n)$, se tiene que $\frac{X_n^2}{\lambda_n} \sim \chi_1^2$, y en conclusión

$$\|X\|^2 = \sum_{n=0}^\infty X_n^2 \stackrel{d}{=} \sum_{n=0}^\infty \lambda_n Z_n^2, \text{ donde } \{Z_n\}_{n=0}^\infty \text{ son v.a.i.i.d. } \mathcal{N}(0, 1).$$

■

Observación 1.7. Se puede decir más acerca de la distribución en (ii) si el proceso gaussiano es un PB. Si $\{X(t)\}_{t \in I}$ es un PB, ya hemos visto que la función de covarianzas viene dada por $k(s,t) = s \wedge t - st$, $0 \leq s, t \leq 1$. Para cada $f \in L^2(I)$,

$$\begin{aligned} (\mathcal{K}f)(t) &= \int_0^1 (s \wedge t - st)f(s) ds = \int_0^t (s \wedge t - st)f(s) ds + \int_t^1 (s \wedge t - st)f(s) ds \\ &= \int_0^t (1-t)sf(s) ds + \int_t^1 t(1-s)f(s) ds = (1-t) \int_0^t sf(s) ds + t \int_t^1 (1-s)f(s) ds. \end{aligned}$$

Si además f es una autofunción asociada al autovalor λ , $\lambda \neq 0$, del operador \mathcal{K} , entonces

$$(\mathcal{K}f)(t) = \lambda f(t) \implies \lambda f(t) = (1-t) \int_0^t sf(s) ds + t \int_t^1 (1-s)f(s) ds. \quad (1.5)$$

De esta expresión y del Teorema Fundamental del Cálculo Integral, se deduce

que f es derivable y

$$\begin{aligned}\lambda f'(t) &= (1-t)tf(t) - \int_0^t sf(s)ds - t(1-t)f(t) + \int_t^1 (1-s)f(s)ds \\ &= - \int_0^t sf(s)ds + \int_t^1 f(s)ds - \int_t^1 sf(s)ds = \int_t^1 f(s)ds - \int_0^1 sf(s)ds.\end{aligned}$$

De nuevo, por el Teorema Fundamental del Cálculo Integral, f' es derivable y

$$\lambda f''(t) = f(1) - f(t).$$

Por otro lado, de (1.5) se deduce que los valores de f en la frontera de I son

$$\begin{aligned}\lambda f(0) = 0 &\Rightarrow f(0) = 0, \\ \lambda f(1) = 0 &\Rightarrow f(1) = 0.\end{aligned}$$

Por lo tanto, hemos obtenido que una autofunción $f \in L^2(I)$ del operador \mathcal{K} con autovalor no nulo λ asociado, debe satisfacer la ecuación diferencial

$$\begin{cases} f''(t) = -\frac{1}{\lambda}f(t) \\ f(0) = f(1) = 0 \end{cases} \quad (1.6)$$

Sabemos, de las asignaturas de Ecuaciones Diferenciales impartidas en el Grado (ver [14]), que para cada $n > 0$, la autofunción dada por:

$$f_n(t) = \sin(n\pi t),$$

es solución de (1.6) para el autovalor $\lambda_n = \frac{1}{n^2\pi^2}$. De aquí obtenemos el sistema de autofunciones ortonormal y completo $\{\psi_n\}_{n>0}$, con $\psi_n(t) = \sqrt{2}\sin(n\pi t)$ (observemos que $\|\sin(n\pi t)\|_{L^2(I)} = \int_0^1 \sin(n\pi t)^2 dt = \int_0^1 \frac{1 - \cos(2n\pi t)}{2} dt = \frac{1}{2}$).

En definitiva,

$$\|B\|^2 = \sum_{n=0}^{\infty} X_n \stackrel{d}{=} \sum_{n=0}^{\infty} \frac{1}{n^2\pi^2} Z_n^2, \text{ donde } \{Z_n\}_{n=0}^{\infty} \text{ son v.a.i.i.d. } \mathcal{N}(0, 1).$$

1.4 La técnica de remuestreo *bootstrap*

El *bootstrap* es un procedimiento de inferencia estadística que consiste en construir la distribución de un estadístico sin más que recurrir a los valores observados. La idea básica del *bootstrap* es que, en ausencia de información sobre la distribución, la muestra observada contiene toda la información disponible sobre la distribución subyacente, y por tanto, remuestrear la muestra es lo más parecido a remuestrear directamente de la distribución. El término que da nombre a este procedimiento fue introducido por Efron en 1979 en alusión a la expresión anglosajona: “*pulling oneself up by one’s bootstrap*”, que podría traducirse como: “*salir adelante por esfuerzos propios*”.

Supongamos que a partir de una muestra $\mathbf{X} = (X_1, \dots, X_n)$ se quiere estimar un parámetro θ de la distribución, y sea $\hat{\theta} = s(\mathbf{X})$ el estadístico que estima a θ . Conocer la distribución del estadístico $\hat{\theta}$ es de gran ayuda para poder realizar inferencias sobre el parámetro; sin embargo, en muchas ocasiones, el problema de hallar dicha distribución o de establecer resultados asintóticos de convergencia del estadístico es de gran complejidad.

Si la verdadera distribución P de la cual ha sido extraída la muestra \mathbf{X} fuese conocida, se podrían obtener muestras $\mathbf{X}^1, \dots, \mathbf{X}^B$ de P y, usando métodos de simulación de tipo Monte-Carlo, estimar la distribución del estadístico $\hat{\theta}$. Pero como P es desconocida y no podemos generar muestras a partir de ella, la alternativa que sugiere el bootstrap consiste en remuestrear la muestra original \mathbf{X} , es decir, generar muestras a partir de la distribución empírica P_n asociada a \mathbf{X} . A una remuestra obtenida de esta manera se le conoce con el nombre de *muestra o réplica bootstrap* de \mathbf{X} . Se puede decir, entonces, que el principio del bootstrap es:

La población es a la muestra como la muestra es a las muestras bootstrap.

Los resultados teóricos básicos de probabilidad que avalan esta afirmación son la Ley de los Grandes Números y el Teorema de Glivenko-Cantelli que, como vimos al principio de la sección 1.2, garantizan que P_n es un estimador consistente de P . A primera vista, el planteamiento que propone el *bootstrap* resulta sorprendente, e incluso artificioso, pero se ha demostrado a lo largo de la literatura que tiene completo sentido. Nos referimos a [13] para los aspectos más generales de esta técnica, y a [17] y [25] para aquellos más específicos acerca de la consistencia de las aproximaciones bootstrap.

1.4.1 El principio *bootstrap*

Supongamos que queremos sortear una muestra de v.a.i.i.d. $\mathbf{X}^* = (X_1^*, \dots, X_n^*)$ a partir de P_n . Como sabemos, P_n da probabilidad $\frac{1}{n}$ a cada observación X_i (para valores que aparecen más de una vez en la muestra, la probabilidad es un múltiplo de $\frac{1}{n}$). Por lo tanto, cuando observamos P_n , la i -ésima observación X_i de la muestra original es elegida con probabilidad $\frac{1}{n}$. Esto sugiere el siguiente procedimiento para generar la réplica bootstrap \mathbf{X}^* de tamaño m , con $0 < m \leq n$:

- Sortear índices i_1, \dots, i_m independientes de la distribución uniforme en el conjunto $\{1, \dots, m\}$.
- Tomar $X_j^* = X_{i_j}$ y $\mathbf{X}^* = (X_1^*, \dots, X_m^*)$.

En otras palabras, para construir la muestra bootstrap \mathbf{X} se muestrea con reemplazamiento la muestra original \mathbf{X} .

Sea ahora $\theta = t(P)$ un parámetro de la distribución y $\hat{\theta} = s(\mathbf{X})$ un estimador de θ , del cual queremos conocer la distribución para poder evaluar sus propiedades estadísticas, tales como el sesgo o el error estándar, y tener así información sobre el verdadero parámetro θ , por ejemplo, intervalos de confianza. Para ello, a partir de una muestra bootstrap \mathbf{X}^* , se forman las cantidades equivalentes:

$$\begin{aligned}\theta^* &= t(P_n), \text{ el parámetro en el mundo bootstrap;} \\ \hat{\theta}^* &= s(\mathbf{X}^*), \text{ la réplica bootstrap de } \hat{\theta}.\end{aligned}$$

La distribución muestral del estimador θ^* se estima mediante su equivalente bootstrap, es decir, $P_n(\hat{\theta} \in A) = P^*(\hat{\theta}^* \in A)$, para todo $A \subset \mathbb{R}$. En la figura 1.1 se resume esta técnica de manera esquemática.

1.4.2 Aproximación Monte-Carlo de $\hat{\theta}^*$.

En muchas ocasiones, estamos interesados en una característica concreta de la distribución muestral del estimador $\hat{\theta}$ que puede obtenerse directamente de las réplicas bootstrap, como son, por ejemplo, la media, el error estándar o el sesgo. En muchas otras circunstancias, requerimos de la distribución bootstrap del estimador.

A pesar de que la distribución P_n de las muestras bootstrap $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$ es conocida, el estudio de la distribución muestral bootstrap de $\hat{\theta}^*$ sigue siendo

intratable en la mayoría de los casos. Por lo tanto, la alternativa habitual para estimar la distribución bootstrap del estimador θ^* es aproximarla mediante métodos de simulación Monte-Carlo:

1. Generar B muestras bootstrap independientes $\mathbf{X}^{*1}, \dots, \mathbf{X}^{*B}$:

$$\mathbf{X}^{*b} = (X_1^{*b}, \dots, X_m^{*b}) \sim P_n \text{ i.i.d., con } b = 1, \dots, B.$$

2. Calcular $\hat{\theta}^{*b} = s(\mathbf{X}^{*b}), b = 1, \dots, B$.
3. Estimar la distribución muestral de $\hat{\theta}$ mediante la distribución empírica de las réplicas bootstrap $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$:

$$P_n(\hat{\theta} \in A) = \frac{1}{B} \sum_{b=1}^B I(\hat{\theta}^{*b} \in A), \text{ para } A \subset \mathbb{R}^d, d \geq 1.$$

En este procedimiento nos basaremos para poder estimar la distribución bootstrap del estadístico de interés en este trabajo.

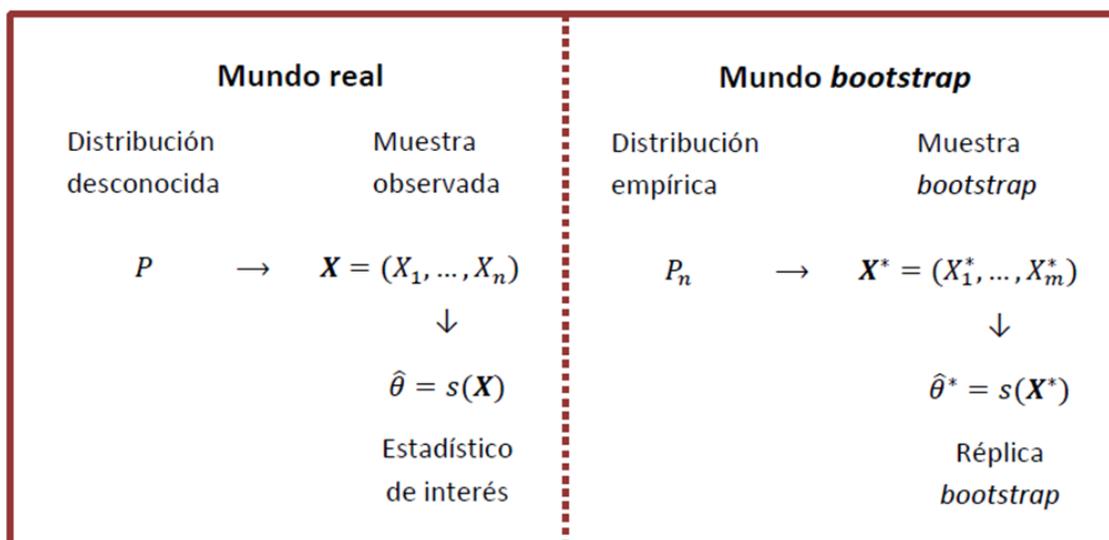


Figura 1.1: Principio bootstrap

2 | BARICENTROS EN EL ESPACIO DE WASSERSTEIN Y EL MODELO DE DEFORMACIÓN PARA DISTRIBUCIONES.

Este capítulo da una breve descripción del baricentro y de la variación en el espacio de Wasserstein que, como se dijo en el Capítulo 1, cuenta con la estructura de espacio métrico cuando en él se considera la distancia de Wasserstein. El baricentro es una generalización a espacios métricos, no necesariamente lineales, del concepto de media de una colección de puntos o de una distribución. El caso particular del espacio de Wasserstein se describe en la sección 2.1. Como veremos, estos conceptos tienen interés en este trabajo en cuanto a su aplicabilidad en la valoración del ajuste de un modelo de deformación para distribuciones de probabilidad. En muchas situaciones, se quieren analizar datos que provienen de fuentes que introducen, por diversas razones, diferentes deformaciones en las señales observadas. Para tratar la información contenida en ellos y poder estudiar las relaciones estructurales existentes entre las diferentes muestras, es necesario alinear dichas distribuciones. En este sentido, se introduce en la sección 2.2 el modelo de deformación, herramienta que da buenos resultados en la resolución de este problema. Las demostraciones que se omitan sobre la primera parte pueden encontrarse en [1], y los resultados de la segunda parte están probados en [11]. Finalmente, los resultados más importantes de este trabajo y sus demostraciones se reúnen en la sección 2.3, que garantizan que las distribuciones bootstrap de la variación de Wasserstein son estimadores consistentes.

2.1 Baricentros en el espacio de Wasserstein

Sea (E, d, Ω) un espacio métrico medible y consideremos $\mathcal{P}(E)$ el conjunto de medidas de probabilidad sobre E . Dada una colección de distribuciones de probabilidad ν_1, \dots, ν_I en $\mathcal{P}(E)$, existen numerosas maneras de definir un promedio ponderado $\nu_B \in \mathcal{P}(E)$ de estas medidas con respecto a unos pesos

$\lambda_1, \dots, \lambda_I$, tales que $\lambda_i \geq 0$, para todo $1 \leq i \leq I$ y $\sum_{i=1}^I \lambda_i = 1$. La forma más intuitiva de hacerlo resulta al tomar la combinación lineal convexa:

$$\nu_B = \sum_{i=1}^I \lambda_i \nu_i.$$

Como consecuencia de que el espacio de medidas de probabilidad $\mathcal{P}(E)$ es un subconjunto convexo del espacio lineal de medidas finitas, se tiene que ν_B está bien definido y $\nu_B \in \mathcal{P}(E)$. Sin embargo, al dotar al espacio $\mathcal{P}(E)$ con una métrica cualquiera, la definición de comportamiento promedio dada por ν_B podría no ser la más apropiada. En este sentido, se ha desarrollado una alternativa a esta noción de media ponderada cuando los objetos a promediar son distribuciones de probabilidad, y que propone trabajar en el espacio de Wasserstein $\mathcal{W}_2(E)$.

De ahora en adelante, consideraremos el subconjunto de $\mathcal{P}(E)$ formado por las probabilidades en \mathbb{R}^d con momento de orden 2 finito, es decir, $\mathcal{W}_2(\mathbb{R}^d)$, y $d \geq 1$. Como sabemos, en \mathbb{R}^d con la distancia Euclídea, el baricentro de los puntos $x_1, \dots, x_n \in \mathbb{R}^d$ con pesos respectivos $\lambda_1, \dots, \lambda_n$, es el punto $x_B \in \mathbb{R}^d$ que minimiza el funcional

$$E(x) = \sum_{i=1}^n \lambda_i \|x - x_i\|^2.$$

Este punto x_B es único y es la media ponderada

$$x_B = \sum_{i=1}^n \lambda_i x_i.$$

El mismo procedimiento se propone en [1] al considerar distribuciones de probabilidad en \mathbb{R}^d , en lugar de puntos, y la distancia cuadrática de Wasserstein en lugar de la Euclídea.

Definición 2.1. Se dice que la medida $\nu \in \mathcal{W}_2(\mathbb{R}^d)$ es un baricentro de Wasserstein de las medidas ν_1, \dots, ν_I con pesos $\lambda_1, \dots, \lambda_I$, tales que $\lambda_i \geq 0$, para todo $i \in \{1, \dots, I\}$, y $\sum_{i=1}^I \lambda_i = 1$, si ν minimiza el funcional:

$$E(\nu) = \sum_{i=1}^I \lambda_i W_2^2(\nu, \nu_i).$$

Denotaremos esta medida por $\nu_B^{(\lambda)} = \text{bar}((\nu_i, \lambda_i)_{1 \leq i \leq I})$, salvo cuando $\lambda_i = \frac{1}{I}$, $\forall i = 1, \dots, I$, en cuyo caso la representaremos simplemente por ν_B . El baricentro así definido existe, y una condición suficiente para su unicidad es, por ejemplo, que una de las medidas ν_i admita densidad con respecto a la medida de Lebesgue.

Cuando sólo hay dos medidas de probabilidad ($I = 2$), este problema se conoce como interpolación de McCann [22] y propone como baricentro $\text{bar}((\nu_1, \lambda), (\nu_2, 1 - \lambda))$, con $\lambda \in (0, 1)$, la probabilidad

$$\nu_B = ((1 - \lambda)Id + \lambda T)_{\# \nu_1},$$

donde T es la función de transporte óptimo que lleva ν_1 sobre ν_2 .

Si en el caso unidimensional ($d = 1$) consideramos más de dos probabilidades, la expresión del baricentro adquiere también una forma bastante simple: ν_B es la probabilidad cuya función cuantil es

$$F_B^{-1} = \sum_{i=1}^I \lambda_i F_i^{-1}.$$

Otra situación sencilla en la que el baricentro se ha caracterizado de manera semiexplícita (las ecuaciones del sistema matricial (2.1) no son lineales) es la del caso Gaussiano: $\nu_i = \mathcal{N}(0, S_i)$, $i = 1, \dots, I$. Supongamos que la matriz de covarianzas S_i es definida positiva $\forall i$. En [1] se muestra que el baricentro viene dado por $\nu_B = \mathcal{N}(0, \bar{S})$, donde \bar{S} es la única matriz definida positiva que es solución de la ecuación matricial

$$\sum_{i=1}^I \lambda_i \left(S_i^{\frac{1}{2}} S_i S_i^{\frac{1}{2}} \right)^{\frac{1}{2}} = \bar{S}. \quad (2.1)$$

Una vez dada esta noción de comportamiento promedio de una colección de medidas de probabilidad, podemos estar interesados en conocer cómo de dispersas se encuentran tales medidas con respecto a su baricentro. Volviendo al

contexto euclídeo, dados puntos $x_1, \dots, x_n \in \mathbb{R}^d$, sabemos que una medida de su dispersión viene dada por su varianza

$$S^2 = \sum_{i=1}^n \|x_i - x_B\|^2.$$

Por analogía a este caso, y con el objetivo de definir una medida global de la separación de una colección de probabilidades $(\nu_i)_{1 \leq i \leq I}$, en [11] se introduce la siguiente definición de 2-variación de Wasserstein.

Definición 2.2. Sean ν_1, \dots, ν_I probabilidades en $\mathcal{W}_2(\mathbb{R}^d)$. Se denomina 2-variación (o variación) de Wasserstein de ν_1, \dots, ν_I a la cantidad:

$$V_2(\nu_1, \dots, \nu_I) := \inf_{\eta \in \mathcal{W}_2(\mathbb{R}^d)} \left(\frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \eta) \right)^{\frac{1}{2}}.$$

Por lo visto anteriormente, la medida que minimiza el funcional:

$$\eta \mapsto \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \eta)$$

es el baricentro ν_B de ν_1, \dots, ν_I , con pesos $\lambda_i = \frac{1}{I}, \forall i = 1, \dots, I$. Luego, podemos escribir que:

$$V_2(\nu_1, \dots, \nu_I) = \left(\frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \nu_B) \right)^{\frac{1}{2}}.$$

En el caso unidimensional ($d = 1$), por el Teorema 1.3, tenemos la expresión

$$V_2(\nu_1, \dots, \nu_I) = \left(\frac{1}{I} \sum_{i=1}^I \int_0^1 (F_i^{-1}(t) - F_B^{-1}(t))^2 dt \right)^{\frac{1}{2}}.$$

La variación generaliza la noción de varianza empírica y, tal y como ocurre para ésta,

$$V_2(\nu_1, \dots, \nu_I) = 0 \Leftrightarrow \nu_1 = \dots = \nu_I.$$

Esto nos sugiere que la variación de Wasserstein es un buen candidato para evaluar la igualdad entre distribuciones de probabilidad. En particular, veremos que resulta útil para valorar la existencia de un modelo de deformación en una colección de distribuciones.

2.2 Modelo de deformación para distribuciones.

En estadística, es de gran importancia encontrar la manera de medir la relación estructural entre una colección de observaciones. Esta cuestión se vuelve particularmente interesante al estimar probabilidades observadas de fuentes que han inducido deformaciones en los datos, pues estas distorsiones, tales como traslaciones, modelos de localización y escala o deformaciones más generales, previenen el uso de los métodos estadísticos habituales. Para estudiar este problema, se quieren diseñar métodos de alineamiento para las distribuciones que les eliminen las perturbaciones. Sin embargo, la literatura es escasa en cuanto al análisis de esta situación. Mencionamos el método de la normalización cuantil desarrollado en [7] y, posteriormente en [15], dirigido a cuestiones computacionales en el ámbito de la biología. En esta sección, se desarrolla el modelo de deformación para probabilidades en $\mathcal{W}_2(\mathbb{R})$, herramienta que ha resultado útil para llevar a cabo esta tarea. Presentamos, de manera general, los aspectos básicos de este modelo y el procedimiento de alineamiento de probabilidades y, posteriormente, nos centramos en el caso paramétrico para exponer una serie de resultados asintóticos.

Sean $j = 1, \dots, n$ observaciones de $i = 1, \dots, I$ variables aleatorias independientes $X_{i,j}$ con distribución ν_i :

$$\begin{array}{ccccccc} X_{1,1}, & X_{1,2}, & \dots, & X_{1,n} & i.i.d. & \sim & \nu_1 \\ X_{2,1}, & X_{2,2}, & \dots, & X_{2,n} & i.i.d. & \sim & \nu_2 \\ \vdots & \vdots & \ddots & \vdots & & & \vdots \\ X_{I,1}, & X_{I,2}, & \dots, & X_{I,n} & i.i.d. & \sim & \nu_I \end{array}$$

El modelo de deformación consiste en suponer que cada muestra $(X_{i,j})_{1 \leq j \leq n}$, $i = 1, \dots, I$, se ha extraído de una distribución común desconocida ν , a la que se le han aplicado diferentes variaciones. Más concretamente, supongamos que existen funciones φ_i^* , $i = 1, \dots, I$, a las que llamaremos *deformaciones* o *funciones deformadoras*, tales que:

$$\nu_i = \nu \circ (\varphi_i^*)^{-1}, i = 1, \dots, I.$$

Estas funciones φ_i^* pueden estar indexadas o no por un parámetro, dando lugar a un modelo de deformación paramétrico o no paramétrico, respectivamente. En general, consideremos una familia $\mathcal{G} = \mathcal{G}_1 \times \dots \times \mathcal{G}_I$ de deformaciones tales que, para cada $i = 1, \dots, I$ y para cada función $h \in \mathcal{G}_i$, se verifica que:

(A1) $h : (c_i, d_i) \longrightarrow (a, b)$ es invertible y creciente con $-\infty \leq a < b \leq +\infty$,

$$-\infty \leq c \leq c_i < d_i < d \leq +\infty.$$

En esta situación, el modelo de deformación para distribuciones de probabilidad se enuncia de la manera siguiente:

existen $(\varphi_1^*, \dots, \varphi_I^*) \in \mathcal{G}$ y variables aleatorias $(\varepsilon_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$ i.i.d. con distribución ν tales que:

$$X_{i,j} = (\varphi_i^*)^{-1}(\varepsilon_{i,j}), \forall 1 \leq i \leq I, \forall 1 \leq j \leq n \quad (2.2)$$

Conviene fijar los siguientes aspectos de la notación que seguiremos en el resto del trabajo. De ahora en adelante, G denotará la función de distribución de la ley ν , que tiene soporte el intervalo (a, b) y, para cada $i = 1, \dots, I$, $G_{n,i}$ denotará la función de distribución empírica asociada a la muestra $(\varepsilon_{i,j})_{1 \leq j \leq n}$. Por otra parte, para cada $i = 1, \dots, I$, sea

$$F_i : (c_i, d_i) \longrightarrow (0, 1)$$

la función de distribución asociada a la medida ν_i , con densidad f_i con respecto a la medida de Lebesgue. Por último, sean $\nu_{n,i}$ y $F_{n,i}$ la medida empírica y la función de distribución empírica, respectivamente, asociadas a la muestra $(X_{i,j})_{1 \leq j \leq n}$:

En [2] se presenta un procedimiento para alinear una colección de leyes de probabilidad que responden a un modelo de deformación paramétrico (2.3), que se extiende al modelo de deformación general (2.2) en el trabajo posterior [21]. Como la distribución ν es desconocida, debemos invertir el operador φ_i^{-1} de deformación de cada muestra y calcular, para cada observación $X_{i,j}$, su imagen a través de la deformación candidata φ_i :

$$Z_{i,j}(\varphi_i) = \varphi_i(X_{i,j}), 1 \leq i \leq I; 1 \leq j \leq n.$$

La notación $Z_{i,j}(\varphi_i)$ quiere resaltar la dependencia de φ_i . Observemos que, bajo la hipótesis $(\mathcal{A}1)$, se tiene que:

$$Z_{i,j}(\varphi_i) \sim \nu_i(\varphi_i), \text{ con función de distribución } F_i \circ \varphi_i^{-1} := F_{\varphi_i}.$$

En efecto,

$$\begin{aligned} F_{\varphi_i}(x) &= P[Z_{i,j}(\varphi_i) \leq x] = P[\varphi_i(X_{i,j}) \leq x] = P[X_{i,j} \leq \varphi_i^{-1}(x)] \\ &= F_i(\varphi_i^{-1}(x)) = F_i \circ \varphi_i^{-1}(x), \forall x \in \mathbb{R}. \end{aligned}$$

Notemos que bajo el modelo (2.2), todas las medidas ν_i han sido deformadas a partir de una medida común ν y, por lo tanto, al invertir las deformaciones se tiene que

$$\nu_B(\varphi^*) = \nu = \nu_i(\varphi_i^*), 1 \leq i \leq I,$$

y, en consecuencia,

$$V_2(\nu_1(\varphi_1^*), \dots, \nu_I(\varphi_I^*)) = 0.$$

Por lo tanto, para construir un test que permita valorar si el modelo de deformación es correcto, una idea que surge de manera intuitiva es tomar como criterio de decisión la 2-variación de Wasserstein de las distribuciones $(\nu_i(\varphi_i))_{1 \leq i \leq I}$. Definamos entonces, para facilitar la notación, $U(\varphi) := V_2^2(\nu_1(\varphi_1), \dots, \nu_I(\varphi_I))$ y consideremos el siguiente escenario para contrastar una relación de deformación entre las distribuciones

$$\inf_{\varphi \in \mathcal{G}} U(\varphi) = 0 \quad (\mathcal{H}_0)$$

$$\inf_{\varphi \in \mathcal{G}} U(\varphi) > 0 \quad (\mathcal{H}_1)$$

Con estas hipótesis, si el test rechaza la hipótesis nula hay evidencia estadística de que el modelo de deformación no se verifica de manera exacta. Observemos que este contraste de hipótesis lleva implícita la resolución del problema de optimización consistente en

minimizar el criterio $U(\varphi)$, con φ variando en \mathcal{G} .

Para ello, debemos buscar el elemento $\varphi^* = (\varphi_1^*, \dots, \varphi_I^*) \in \mathcal{G}$ que minimiza el coste de alineamiento de las distribuciones de las variables $(Z_{i,j}(\varphi_i))_{1 \leq i \leq I}$ con la distribución de su baricentro $\nu_B(\varphi)$. Es importante tener en cuenta que, para cada $\varphi = (\varphi_1, \dots, \varphi_I) \in \mathcal{G}$, el baricentro de las distribuciones $(\nu_i(\varphi_i))_{1 \leq i \leq I}$ es la probabilidad $\nu_B(\varphi)$ con función cuantil:

$$F_B^{-1}(\varphi)(t) = \frac{1}{I} \sum_{k=1}^I \varphi_k \circ F_k^{-1}(t), \text{ con } 0 \leq t \leq 1.$$

Sin embargo, las medidas teóricas ν_i son desconocidas, pues lo único que conocemos son los valores observados $(X_{i,j})_{1 \leq j \leq n}$ de la distribución ν_i y, por tanto, de las leyes $(\nu_i(\varphi_i))_{1 \leq i \leq I}$ sólo conocemos las cantidades $(Z_{i,j}(\varphi_i))_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$,

y con ello es con lo que debemos trabajar. Escribamos entonces $\nu_{n,i}(\varphi_i)$ y $F_{n,\varphi_i} = F_{n,i} \circ \varphi_i^{-1}$ para denotar la medida y la distribución empíricas asociadas a la muestra $(Z_{i,j}(\varphi_i))_{1 \leq j \leq n}$.

La medida empírica del baricentro de $(\nu_{n,i}(\varphi_i))_{1 \leq i \leq I}$ es:

$$\nu_{n,B}(\varphi)(t) = \frac{1}{I} \sum_{i=1}^I \nu_{n,i}(\varphi_i)(t),$$

con función cuantil empírica asociada

$$F_{n,B}^{-1}(\varphi)(t) = \frac{1}{I} \sum_{k=1}^I \varphi_k \circ F_{n,k}^{-1}(t), \text{ con } 0 \leq t \leq 1.$$

Entonces, debemos estimar el elemento $\varphi^* = (\varphi_1^*, \dots, \varphi_I^*) \in \mathcal{G}$ que minimiza el coste de alineamiento de las distribuciones $(\nu_{n,i}(\varphi_i))_{1 \leq i \leq I}$ de las variables $(Z_{i,j}(\varphi_i))_{1 \leq i \leq I}$ con la distribución de su baricentro $\nu_{n,B}(\varphi)$, por el elemento $\hat{\varphi}_n = (\hat{\varphi}_{n,1}, \dots, \hat{\varphi}_{n,I})$ que minimiza la versión empírica de la variación

$$\begin{aligned} U_n(\varphi) &= V_{2,n}^2(\nu_{n,1}(\varphi_1), \dots, \nu_{n,I}(\varphi_I)) = \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_{n,1}(\varphi_1), \dots, \nu_{n,I}(\varphi_I)) \\ &= \frac{1}{I} \sum_{i=1}^I \left[\int_0^1 (F_{n,\varphi_i}^{-1}(t) - F_{n,B}^{-1}(\varphi)(t))^2 dt \right] \\ &= \frac{1}{I} \sum_{i=1}^I \left[\int_0^1 \left(\varphi_i \circ F_{n,i}^{-1}(t) - \frac{1}{I} \sum_{k=1}^I \varphi_k \circ F_{n,k}^{-1}(t) \right)^2 dt \right]. \end{aligned}$$

En consecuencia, nos basaremos en el estadístico $\inf_{\varphi \in \mathcal{G}} U_n(\varphi)$ para hacer inferencias sobre el modelo de deformación (2.2).

2.2.1 Estimación del modelo de deformación paramétrico

En muchas ocasiones, se puede suponer que las funciones que perturban los datos antes de ser observados adquieren formas más específicas. Consideremos entonces el caso en el que la distribución $\nu \in \mathcal{W}_2(\mathbb{R})$ sigue siendo desconocida, pero las funciones de deformación pertenecen a una familia paramétrica. Supongamos que la familia de deformaciones \mathcal{G} está indexada por un parámetro $\lambda \in \Lambda \subset \mathbb{R}^p$, de manera que si la función $\varphi \in \mathcal{G}$, entonces

$$\begin{aligned}\varphi : \Lambda \times (c, d) &\longrightarrow (a, b) \\ (\lambda, x) &\longmapsto \varphi_\lambda(x).\end{aligned}$$

La formulación del modelo de deformación (2.2) en el caso paramétrico sería:

existen parámetros $\theta^* = (\theta_1^*, \dots, \theta_I^*) \in \Theta = \Lambda^I$ y variables aleatorias $(\varepsilon_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$ i.i.d. con distribución ν , tales que:

$$X_{i,j} = (\varphi_{\theta_i^*})^{-1}(\varepsilon_{i,j}), \forall 1 \leq i \leq I, \forall 1 \leq j \leq n, \quad (2.3)$$

y la versión del estadístico viene dada por

$$\inf_{\theta \in \Theta} U_n(\theta) = \inf_{\theta \in \Theta} \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_{n,i}(\theta_i), \dots, \nu_{n,I}(\theta_I)).$$

En [11] (ver Teorema 4.1) se prueba, bajo el modelo (2.3) y ciertas condiciones, la convergencia hacia 0 del estadístico multiplicado por un factor de escala:

$$\inf_{\theta \in \Theta} \sqrt{n} U_n(\theta) \xrightarrow{d} 0.$$

Para poder hacer inferencias es necesario conocer de forma más precisa el comportamiento asintótico de $\inf_{\theta \in \Theta} U_n(\theta)$ asumiendo que el modelo (2.3) es válido. Enunciamos a continuación condiciones bajo las cuales

$$\hat{\theta}_n := \operatorname{argmin}_{\theta \in \Theta} U_n(\theta)$$

es un estimador consistente del verdadero parámetro de deformación. Posteriormente, veremos que, bajo condiciones adicionales, $n \inf_{\theta \in \Theta} U_n(\theta)$ converge en distribución a un límite no degenerado.

(A1) Para todo $\lambda \in \Lambda$, la función

$$\begin{aligned}\varphi_\lambda : (c, d) &\rightarrow (a, b) \\ x &\longmapsto \varphi_{\lambda(x)}\end{aligned}$$

es invertible, creciente y tal que

$$\begin{aligned}-\infty &\leq a < b \leq +\infty, \\ -\infty &\leq c \leq c_i < d_i \leq d \leq +\infty.\end{aligned}$$

(A2) G es de clase \mathcal{C}^2 en (a, b) y tal que $G'(x) = g(x) > 0, \forall x \in (a, b)$ y

$$\sup_{a < x < b} \frac{G(x)(1 - G(x))g'(x)}{g(x)^2} < \infty.$$

- (A3) φ es continua con respecto a x y a λ ,
 $\forall \lambda \in \Lambda, \varphi_\lambda$ es de clase C^1 con respecto a x ,
 Λ es compacto.
- (A4) $d\varphi$ es acotada en $\Lambda \times [c_i, d_i]$, continuo con respecto a λ y tal que

$$\sup_{\lambda \in \Lambda} \left| d\varphi_\lambda(x_n^\lambda) - d\varphi_\lambda(x) \right| \xrightarrow{\sup_{\lambda \in \Lambda} |x_n^\lambda - x| \rightarrow 0} 0.$$

- (A5) $\forall 1 \leq j \leq J, E[|X_j|^r] < \infty$ para algún exponente $r > 4$.
- (A6) $\forall 1 \leq j \leq J, E \left[\sup_{\lambda \in \Lambda} |\varphi_\lambda(X_j)|^r \right] < \infty$, para algún $r > 4$.
- (A7) (**Condición de identificabilidad**) El punto θ^* donde U alcanza el mínimo es único, y está en el interior de Θ .

Como estamos suponiendo que se satisface el modelo de deformación (2.3), la hipótesis (A6) implica que ε tiene algún momento de orden mayor que 4 finito, y la hipótesis (A7) equivale a que θ^* sea el único cero de U .

De la siguiente proposición se deduce que $\hat{\theta}_n$ es un buen candidato para estimar el parámetro del modelo θ^* .

Proposición 2.1. Si se cumplen las hipótesis (A1)-(A7), entonces

$$\hat{\theta}_n \rightarrow \theta^* \text{ en probabilidad.}$$

Bajo condiciones adicionales de regularidad se pueden establecer resultados de interés acerca de la convergencia del estadístico $\inf_{\theta \in \Theta} U_n(\theta)$. Estas condiciones son

- (A8) $\forall 1 \leq i \leq I, \varphi_{\theta_i^*}^{-1}$ es de clase C^1 con respecto a x
 $d\varphi_{\theta_i^*}^{-1}$ es acotada en $[a, b]$
 φ es de clase C^2 con respecto a x y a λ .
- (A9) $\forall 1 \leq i \leq I, E \left[\sup_{\lambda \in \Lambda} \left| \partial^2 \varphi_\lambda(x) \left(\varphi_{\theta_i^*}^{-1}(\varepsilon) \right) \right|^2 \right] < \infty$
- (A10) $\int \frac{t(1-t)}{g^2(G^{-1}(t))} dt < \infty$

El operador d indica derivación respecto de x , mientras que ∂ es el operador de derivación respecto del parámetro λ .

(A8) y (A9) imponen condiciones sobre las funciones de deformación $\varphi_i, i = 1, \dots, I$, mientras que (A10) restringe las posibilidades para la distribución

G de ε . Por otro lado, la condición **(A10)** la verifican aquellas distribuciones con soporte compacto, estrictamente positivas y con densidad absolutamente continua. Pero también existen variables aleatorias con soporte no acotado que la satisfacen, de las cuales se pueden encontrar ejemplos en [10].

Observación 2.1. Los resultados de esta sección, se basan en aproximaciones fuertes del proceso cuantil

$$\rho_n(t) = \sqrt{ng} \left(G^{-1}(t) \right) \left(G_n^{-1}(t) - G^{-1}(t) \right), 0 < t < 1, i = 1, \dots, I.$$

El resultado clave es el siguiente (ver [8]):

Teorema 2.1. *Bajo la hipótesis **(A2)**, existe, en un espacio de probabilidad suficientemente rico, versiones independientes de los procesos $\rho_{n,i}$ y familias independientes de Puentes Brownianos $\{B_{n,i}\}_{n=1}^{\infty}, i = 1, \dots, I$ satisfaciendo*

$$n^{\frac{1}{2}-\nu} \sup_{\frac{1}{n} \leq t \leq 1 - \frac{1}{n}} \frac{|\rho_{n,i}(t) - B_{n,i}(t)|}{(t(1-t))^\nu} = \begin{cases} O_p(\log(n)), & \nu = 0 \\ O_p(1), & 0 < \nu \leq \frac{1}{2} \end{cases}$$

La notación O_p significa lo siguiente: dadas una sucesión de números reales $\{a_n\}_{n \geq 0}$ y una sucesión de variables aleatorias $\{X_n\}_{n \geq 0}$, se dice que $X_n = O_p(a_n)$ si para cada $\varepsilon > 0$, existe $M > 0$ tal que $P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \varepsilon, \forall n \geq 1$. Por otro lado, se conoce como una *aproximación fuerte* a un resultado que garantiza la existencia de versiones del proceso cuantil y de procesos Gaussianos próximas con probabilidad 1, lo que permite reducir el estudio de la convergencia débil de ciertos funcionales del proceso cuantil a estudiar la convergencia puntual de funcionales de procesos Gaussianos. Los detalles se encuentran en [11].

A lo largo de esta sección consideraremos $\Lambda \subset \mathbb{R}$, sin embargo, los resultados siguen siendo ciertos si $\Lambda \subset \mathbb{R}^d$, $d \geq 1$, y las pruebas serían las mismas con las únicas modificaciones relativas a la dimensión. Añadiendo la siguiente hipótesis, se consigue un refinamiento de la *Proposición 2.1*.

(TCL) $\forall 1 \leq i \leq I$, la función $R_i := \partial \varphi_{\theta_i^*} \circ \varphi_{\theta_i^*}^{-1}$ es continua y acotada en $[a, b]$.

Definimos ahora la matriz $\Phi = [\Phi_{i,j}]_{1 \leq i, j \leq I}$, cuyos elementos vienen dados por

$$\Phi_{i,j} = -\frac{2}{I^2} \langle R_i, R_j \rangle_\nu, \text{ si } i \neq j;$$

$$\Phi_{i,i} = \frac{2(I-1)}{I^2} \|R_i\|_\nu,$$

donde $\|\cdot\|_v$ y $\langle \cdot, \cdot \rangle_v$ denotan la norma y el producto interno, respectivamente, en $L^2(v)$. La matriz Φ , en forma extendida, se escribe como sigue:

$$\Phi = \frac{2}{I^2} \begin{pmatrix} (I-1)\|R_1\|_v & -\langle R_1, R_2 \rangle_v & \dots & -\langle R_1, R_I \rangle_v \\ -\langle R_1, R_2 \rangle_v & (I-1)\|R_2\|_v & \dots & -\langle R_2, R_I \rangle_v \\ \vdots & \vdots & \ddots & \vdots \\ -\langle R_1, R_I \rangle_v & -\langle R_2, R_I \rangle_v & \dots & (I-1)\|R_I\|_v \end{pmatrix}$$

Dado $\mathbf{x} \in \mathbb{R}^I$, se tiene que

$$\begin{aligned} \|\Phi \mathbf{x}\| &= \mathbf{x}^T \Phi \mathbf{x} = \frac{2}{I^2} \begin{bmatrix} x_1 & x_2 & \dots & x_I \end{bmatrix} \begin{bmatrix} (I-1)\|R_1\|_v x_1 - \sum_{j \neq 1} \langle R_1, R_j \rangle_v x_j \\ (I-1)\|R_2\|_v x_1 - \sum_{j \neq 2} \langle R_1, R_j \rangle_v x_j \\ \vdots \\ (I-1)\|R_I\|_v x_1 - \sum_{j \neq I} \langle R_1, R_j \rangle_v x_j \end{bmatrix} \\ &= \frac{2}{I^2} \left\{ (I-1)\|R_1\|_v x_1^2 - \sum_{j \neq 1} \langle R_1, R_j \rangle_v x_j x_1 + (I-1)\|R_1\|_v x_2^2 - \right. \\ &\quad \left. - \sum_{j \neq 1} \langle R_1, R_j \rangle_v x_j x_2 + \dots + (I-1)\|R_I\|_v x_I^2 - \sum_{j \neq I} \langle R_I, R_j \rangle_v x_j x_I \right\} \\ &= \frac{2}{I^2} \left\{ \sum_{i=1}^I (I-1)\|R_i\|_v x_i^2 - 2 \sum_{\substack{i,j=1 \\ i < j}}^I \langle R_i, R_j \rangle_v x_i x_j \right\} \\ &= \frac{2}{I^2} \int \left(\sum_{i=1}^I (I-1)x_i^2 R_i^2 - 2 \sum_{\substack{i,j=1 \\ i < j}}^I x_i x_j R_i R_j \right) dv \\ &= \frac{2}{I^2} \int \left\{ (x_1^2 R_1^2 + x_2^2 R_2^2 - 2x_1 x_2 R_1 R_2) + [(x_1^2 R_1^2 + x_3^2 R_3^2 - 2x_1 x_3 R_1 R_3) + \right. \\ &\quad \left. + (x_1^2 R_1^2 + x_3^2 R_3^2 - 2x_1 x_3 R_1 R_3)] + \dots + [(x_1^2 R_1^2 + x_I^2 R_I^2 - 2x_1 x_I R_1 R_I) + \dots \right. \\ &\quad \left. \dots + (x_{I-1}^2 R_{I-1}^2 + x_I^2 R_I^2 - 2x_{I-1} x_I R_{I-1} R_I) \right\} dv = \frac{2}{I^2} \int \sum_{\substack{i,j=1 \\ i < j}}^I (x_i R_i - x_j R_j)^2 dv. \end{aligned}$$

Por lo tanto, se deduce que $\mathbf{x}^T \Phi \mathbf{x} \geq 0$ y Φ es semidefinida positiva. Más aún, observemos que

$$\begin{aligned}
\frac{2}{I^2} \int \sum_{\substack{i,j=1 \\ i < j}}^I (x_i R_i - x_j R_j)^2 d\nu &= 0 \text{ si, y sólo si,} \\
\sum_{\substack{i,j=1 \\ i < j}}^I (x_i R_i - x_j R_j)^2 &= 0 \text{ } \nu\text{-c.s., si, y sólo si,} \\
(x_i R_i - x_j R_j) &= 0 \text{ } \nu\text{-c.s., } \forall i < j, \text{ si, y sólo si,} \\
x_i R_i &= x_j R_j \text{ } \nu\text{-c.s., } \forall i < j.
\end{aligned}$$

La última condición equivale a que R_i sean proporcionales ν -c.s., $\forall i$. Entonces, Φ es definida positiva y, en consecuencia, invertible, salvo en el caso de que todos los R_i sean proporcionales ν -c.s., $i = 1, \dots, I$. Éste es el último requisito que se deberá imponer. Estamos en las condiciones de enunciar la siguiente proposición, que es del tipo Teorema Central del Límite.

Proposición 2.2. Bajo las hipótesis **(A1)** hasta **(A9)** y **(TCL)**, y si, además, Φ es invertible, entonces

$$\sqrt{n} (\hat{\theta}_n - \theta^*) \xrightarrow{d} \Phi^{-1} Y,$$

donde $Y \stackrel{d}{=} (Y_1, \dots, Y_I)$, con

$$Y_i = \frac{2}{I} \int_0^1 R_i \circ G^{-1} \frac{\tilde{B}_i}{g \circ G^{-1}},$$

y $\tilde{B}_i = B_i - \frac{1}{I} \sum_{k=1}^I B_k$ y $(B_i)_{1 \leq i \leq I}$ son puentes Brownianos independientes.

Observación 2.2. Hay que tener en cuenta que, para poder establecer ambos resultados, se han tenido que imponer una serie de condiciones, de las cuales una puede no resultar demasiado realista a la hora de estudiar modelos con ciertas funciones de deformación. Esta hipótesis es la condición de identificabilidad **(A7)**. Resulta que, para determinados modelos de deformación puede suceder que dados η y $\theta \in \Lambda$,

$$\varphi_\eta \circ \varphi_\theta = \varphi_\zeta, \text{ para cierto } \zeta \in \Lambda.$$

Por lo tanto, bajo el modelo (2.3), existe $\theta^* = (\theta_1^*, \dots, \theta_I^*) \in \Theta$ y v.a.i.i.d. $(\varepsilon_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$ tales que

$$X_{i,j} = \varphi_{\theta_i^*}^{-1}(\varepsilon_{i,j}),$$

y para cualquier $\eta \in \Lambda$,

$$\begin{aligned} X_{i,j} &= \varphi_{\theta_i^*}^{-1}(\varepsilon_{i,j}) = \varphi_{\theta_i^*}^{-1}(\varphi_\eta^{-1}(\varphi_\eta(\varepsilon_{i,j}))) = \varphi_{\theta_i^*}^{-1} \circ \varphi_\eta^{-1}(\varphi_\eta(\varepsilon_{i,j})) \\ &= (\varphi_\eta \circ \varphi_{\theta_i^*})^{-1}(\varphi_\eta(\varepsilon_{i,j})) = \varphi_{\zeta_i}^{-1}(\tilde{\varepsilon}_{i,j}), \end{aligned}$$

para ciertos $\zeta_i \in \Lambda$, $i = 1, \dots, I$, donde $\tilde{\varepsilon}_{i,j} = \varphi_\eta(\varepsilon_{i,j})$ son también v.a.i.i.d. Esto implica que $\zeta = (\zeta_1, \dots, \zeta_I) \in \Theta$ es también un cero de U .

Esta situación se presenta, por ejemplo, cuando el modelo de deformación es de localización y escala (notemos que en este modelo $\Lambda \subset \mathbb{R}^2$):

$$X_{i,j} = \varphi_{\theta_i^*}^{-1}(\varepsilon_{i,j}) = \mu_i^* + \sigma_i^* \varepsilon_{i,j},$$

donde $\theta_i^* = (\mu_i^*, \sigma_i^*)$ y $\varphi_{\theta_i^*}(x) = \frac{x - \mu_i^*}{\sigma_i^*}$. Efectivamente, dado $\eta = (\tilde{\mu}, \tilde{\sigma}) \in \Lambda$,

$$\varphi_\eta \circ \varphi_{\theta_i^*}(x) = \frac{\frac{x - \tilde{\mu}}{\tilde{\sigma}} - \mu_i}{\sigma_i} = \frac{x - (\tilde{\mu} + \tilde{\sigma} \mu_i)}{\sigma_i},$$

y tomando $\zeta_i = (\tilde{\mu} + \tilde{\sigma} \mu_i, \sigma_i)$, tenemos que $\varphi_\eta \circ \varphi_{\theta_i^*} = \varphi_{\zeta_i}$.

Afortunadamente, esto tiene una fácil solución que consiste en seleccionar como referencia uno de los parámetros θ_i^* , digamos el primero, y asumir que θ_1^* es conocido. Si $\tilde{\theta} = (\theta_2, \dots, \theta_I)$, la función a minimizar pasa a ser

$$\tilde{U}(\tilde{\theta}) = U(\theta_1^*, \tilde{\theta}),$$

y el estadístico en estudio se convierte en $\inf_{\tilde{\theta} \in \Lambda^{I-1}} \tilde{U}_n(\tilde{\theta}) = \inf_{\tilde{\theta} \in \Lambda^{I-1}} U_n(\theta_1^*, \tilde{\theta})$.

Con estos cambios, **(A7)** se reescribe de la siguiente manera:

($\tilde{\mathbf{A7}}$) $\tilde{\theta}^* = (\theta_2^*, \dots, \theta_I^*)$ es el único cero de \tilde{U} .

Sea $\hat{\theta}_n \in \Lambda^{I-1}$ el punto donde \tilde{U}_n alcanza el mínimo: $\min_{\tilde{\theta} \in \Lambda^{I-1}} \tilde{U}_n(\tilde{\theta}) = \tilde{U}_n(\hat{\theta}_n)$. La

Proposición 2.1 se traslada de manera literal a este contexto, cambiando $\hat{\theta}_n$ por $\hat{\theta}_n$ y θ^* por $\tilde{\theta}^*$. De la misma manera, la *Proposición 2.2* sigue siendo cierta:

$$\sqrt{n}(\hat{\theta}_n - \tilde{\theta}^*) \xrightarrow{d} \tilde{\Phi}^{-1} \tilde{Y},$$

con $\tilde{Y} \stackrel{d}{=} (Y_2, \dots, Y_I)$ y $\tilde{\Phi} = [\Phi_{i,j}]_{2 \leq i, j \leq I}$, si la matriz $\tilde{\Phi}$ es invertible. Dado

$\mathbf{x} \in \mathbb{R}^{I-1}$, y con un razonamiento similar al de la matriz Φ , tenemos que:

$$\mathbf{x}^T \tilde{\Phi} \mathbf{x} = \frac{2}{I^2} \int \left(\sum_{\substack{i,j=2 \\ i < j}}^I (x_i R_i - x_j R_j)^2 + \sum_{i=2}^I x_i^2 R_i^2 \right) d\nu \geq 0.$$

Más aún,

$$\frac{2}{I^2} \int \left(\sum_{\substack{i,j=2 \\ i < j}}^I (x_i R_i - x_j R_j)^2 + \sum_{i=2}^I x_i^2 R_i^2 \right) d\nu = 0 \text{ si, y sólo si,}$$

$$\sum_{\substack{i,j=2 \\ i < j}}^I (x_i R_i - x_j R_j)^2 + \sum_{i=2}^I x_i^2 R_i^2 = 0 \text{ } \nu\text{-c.s. si, y sólo si,}$$

$$\forall 2 \leq i < j \leq I, (x_i R_i - x_j R_j) = 0 \text{ } \nu\text{-c.s. y } \forall 2 \leq i \leq I, x_i R_i = 0 \text{ } \nu\text{-c.s.}$$

Esta última condición equivale a que $R_i = 0$ ν -c.s., $\forall 2 \leq i \leq I$, y en conclusión, $\tilde{\Phi}$ es definida positiva salvo si $R_i = 0$ ν -c.s. para $i = 2, \dots, I$.

Recordemos que el objetivo de esta sección es conocer el comportamiento asintótico de la variación de Wasserstein bajo el modelo de deformación (2.3). El siguiente teorema constituye el resultado principal de [11], en lo que al modelo de deformación paramétrico se refiere, y especifica la velocidad de convergencia hacia 0 del estadístico $\inf_{\theta \in \Theta} U_n(\theta)$, cuando el modelo es cierto. Además, proporciona la distribución asintótica de dicho estadístico.

Teorema 2.2. *Bajo las hipótesis (A1)-(A10), (TCL) y si Φ es invertible,*

$$n \inf_{\theta \in \Theta} U_n(\theta) \xrightarrow{d} \frac{1}{I} \sum_{i=1}^I \int_0^1 \left(\frac{\tilde{B}_i}{g \circ G^{-1}} \right)^2 - \frac{1}{2} Y^T \Phi^{-1} Y,$$

con $Y = (Y_1, \dots, Y_I)$, $Y_i = \frac{2}{J} \int_0^1 R_i \circ G^{-1} \frac{\tilde{B}_i}{g \circ G^{-1}}$, $\tilde{B}_i = B_i - \frac{1}{I} \sum_{k=1}^I B_k$, donde $(B_i)_{1 \leq i \leq I}$ son puentes Brownianos independientes.

Observemos que la distribución límite depende de la distribución G de la que han sido obtenidas las muestras observadas que, en general, es desconocida. Por tanto, deberemos recurrir a técnicas estadísticas de remuestreo bootstrap para tener información sobre el comportamiento asintótico del estadístico. Para ello, en la siguiente sección 2.3 se introducen los teoremas necesarios para

garantizar la consistencia de las aproximaciones bootstrap de la distribución del estadístico.

2.2.2 Ejemplos de familias paramétricas de deformación

Dado $\lambda \in \Lambda$, diremos que una función

$$\begin{aligned} f_\lambda : (c, d) &\rightarrow (a, b) \\ x &\mapsto f_{\lambda(x)} \end{aligned}$$

es una deformación admisible si satisface las hipótesis de la sección 2.2.1. A continuación, examinamos algunos ejemplos de familias paramétricas de deformaciones admisibles.

Ejemplo 2.1 (Modelo de localización y escala). Consideremos la familia de deformaciones indexada por un parámetro $\theta \in \Lambda \subset \mathbb{R}^2$:

$$\begin{aligned} \varphi : \Lambda \times [c, d] &\longrightarrow [a, b] \\ (\theta, x) &\longmapsto \varphi_\theta(x) = \frac{x - \mu}{\sigma} = \frac{x}{\sigma} - \frac{\mu}{\sigma}, \sigma > 0. \end{aligned}$$

Los parámetros μ y σ se conocen habitualmente como parámetros de localización y escala, respectivamente, del modelo.

En primer lugar, notemos que, para todo $\theta = (\mu, \sigma) \in \Lambda$, como $\sigma > 0$ se tiene que φ_θ es invertible en \mathbb{R} , con función inversa $\varphi_\theta^{-1}(x) = \mu + \sigma x$, y además, φ_θ es estrictamente creciente en \mathbb{R} por ser lineal en x con pendiente positiva. Por tanto, la hipótesis **(A1)** se satisface.

Es trivial comprobar que φ es de clase C^∞ con respecto a sus dos argumentos, luego para que se cumpla **(A3)** basta escoger el conjunto de parámetros Λ un subconjunto compacto de $\mathbb{R} \times \mathbb{R}^+$.

La función $d\varphi(\theta, x) = \frac{1}{\sigma}$ es acotada en $\Lambda \times [c, d]$ y continua con respecto a θ .

Por otro lado, para todo $\theta \in \Lambda$, la función $d\varphi_\theta$ no depende de x y $|d\varphi_\theta(x_n^\lambda) - d\varphi_\theta(x)| = 0$, para toda sucesión $\{x_n\}_{n \geq 0}$ de números reales definida en $[c, d]$, con lo que se verifica **(A4)**.

Siguiendo la *Observación 2.2*, en el caso del modelo de localización y escala, debemos comprobar la condición **(A7)**. En este caso, el criterio $\tilde{U}(\tilde{\theta})$ es una forma cuadrática definida positiva y, por lo tanto, el punto $\tilde{\theta}^* = (\theta_2^*, \dots, \theta_l^*)$ en el que alcanza el mínimo es único. Para ver esto, notemos que la función

cuantil de $\nu_i(\theta_i)$ es

$$F_{\theta_i}^{-1} = (F_i \circ \varphi_{\theta_i})^{-1} = \varphi_{\theta_i} \circ F_i = \frac{G^{-1} - \mu_i}{\sigma_i} = \alpha_i G^{-1} - \beta_i, \text{ con } \alpha_i = \frac{1}{\sigma_i} \text{ y } \beta_i = \frac{\mu_i}{\sigma_i}.$$

Denotemos por $\check{\theta}_i = (\alpha_i, \beta_i)$ y $\check{\theta} = (\check{\theta}_1, \dots, \check{\theta}_I)$.

Como hemos visto, la función cuantil del baricentro es

$$F_B^{-1} = \frac{1}{I} \sum_{i=1}^I F_{\theta_i}^{-1} = \bar{\alpha} G^{-1} - \bar{\beta}, \text{ con } \bar{\alpha} = \frac{1}{I} \sum_{i=1}^I \alpha_i \text{ y } \bar{\beta} = \frac{1}{I} \sum_{i=1}^I \beta_i,$$

de donde deducimos que

$$\begin{aligned} W_2^2(\nu_i(\check{\theta}_i), \nu_B(\check{\theta})) &= \int_0^1 \left((\alpha_i - \bar{\alpha}) G^{-1}(t) - (\beta_i - \bar{\beta}) \right)^2 dt \\ &= (\alpha_i - \bar{\alpha})^2 \int (G^{-1}(t))^2 dt + (\beta_i - \bar{\beta})^2 - 2(\alpha_i - \bar{\alpha})(\beta_i - \bar{\beta}) \int G^{-1}(t) dt \\ &= (\alpha_i - \bar{\alpha})^2 \int (G^{-1}(t))^2 dt + (\beta_i - \bar{\beta})^2. \end{aligned}$$

Por tanto, $W_2^2(\nu_i(\check{\theta}_i), \nu_B(\check{\theta}))$, $i = 1, \dots, I$ es una forma cuadrática definida positiva, y, por consiguiente, lo es $\tilde{U}(\check{\theta}) \equiv \tilde{U}(\check{\theta})$ por ser suma de formas cuadráticas definidas positivas.

Las funciones de localización y escala también satisfacen **(A8)**, pues para $1 \leq i \leq I$, la función $\varphi_{\theta_i^*}(x) = \mu_i^* + \sigma_i^* x$ es de clase $C^\infty([a, b])$ y $d\varphi_{\theta_i^*}^{-1}(x) = \sigma_i^*$ es acotada en $[a, b]$. Finalmente, $\partial^2(\varphi_\theta) = \text{Hess}(\varphi_\theta) = 0$, de donde se sigue **(A9)**.

Ahora pasamos a comprobar la condición **(TCL)**: para todo $i = 1, \dots, I$, la función $R_i(x) = \nabla_\theta(\varphi) = \frac{1}{\sigma_i^*}(-1, x)$ es continua y acotada en $[a, b]$.

Los elementos que forman la matriz $\tilde{\Phi} = [\Phi_{i,j}]_{\substack{2 \leq i \leq I \\ 2 \leq j \leq I}}$ son

$$\begin{aligned} \Phi_{i,j} &= -\frac{2}{I^2} \langle R_i, R_j \rangle_\nu = -\frac{2}{I^2} \int \frac{1}{\sigma_i^* \sigma_j^*} (-1, x) (-1, x)^T d\nu \\ &= -\frac{2}{I^2} \frac{1}{\sigma_i^* \sigma_j^*} \int (1 + x^2) d\nu = -\frac{2}{I^2} \frac{1}{\sigma_i^* \sigma_j^*} (1 + E(\varepsilon^2)) \end{aligned}$$

$$\begin{aligned}
\Phi_{i,i} &= -\frac{2(I-1)}{I^2} \|R_i\|_v = \frac{2(I-1)}{I^2} \int \frac{1}{(\sigma_i^*)^2} (-1, x)(-1, x)^T dv \\
&= \frac{2(I-1)}{I^2} \frac{1}{(\sigma_i^*)^2} \int (1+x^2) dv = \frac{2(I-1)}{I^2} \frac{1}{(\sigma_i^*)^2} (1 + E(\varepsilon^2)).
\end{aligned}$$

Por consiguiente, la matriz $\tilde{\Phi}$ completa es:

$$\tilde{\Phi} = \frac{2}{I^2} (1 + E(\varepsilon^2)) \begin{pmatrix} \frac{I-1}{(\sigma_2^*)^2} & \frac{-1}{\sigma_2^* \sigma_3^*} & \cdots & \frac{-1}{\sigma_2^* \sigma_I^*} \\ \frac{-1}{\sigma_2^* \sigma_3^*} & \frac{I-1}{(\sigma_3^*)^2} & \cdots & \frac{-1}{\sigma_3^* \sigma_I^*} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-1}{\sigma_2^* \sigma_I^*} & \frac{-1}{\sigma_3^* \sigma_I^*} & \cdots & \frac{I-1}{(\sigma_I^*)^2} \end{pmatrix}.$$

Dado $x \in \mathbb{R}^{I-1}$,

$$\begin{aligned}
\|\tilde{\Phi}x\| &= \frac{2}{I^2} \left[\sum_{i=2}^I (I-1) \|R_i\|_v x_i^2 - 2 \sum_{\substack{i,j=2 \\ i < j}}^I \langle R_i, R_j \rangle_v x_i x_j \right] \\
&= \frac{2}{I^2} \left[\sum_{i=2}^I (I-1) \frac{1}{(\sigma_i^*)^2} (1 + E(\varepsilon^2)) x_i^2 - 2 \sum_{\substack{i,j=2 \\ i < j}}^I \frac{1}{\sigma_i^* \sigma_j^*} (1 + E(\varepsilon^2)) x_i x_j \right] \\
&= \frac{2}{I^2} (1 + E(\varepsilon^2)) \left[\sum_{i=2}^I (I-1) \frac{x_i^2}{(\sigma_i^*)^2} - 2 \sum_{\substack{i,j=2 \\ i < j}}^I \frac{x_i x_j}{\sigma_i^* \sigma_j^*} \right] \\
&= \frac{2}{I^2} (1 + E(\varepsilon^2)) \left[\sum_{\substack{i,j=2 \\ i < j}}^I \left(\frac{x_i}{(\sigma_i^*)} - \frac{x_j}{(\sigma_j^*)} \right)^2 + \sum_{i=2}^I \frac{x_i^2}{(\sigma_i^*)^2} \right] \geq 0.
\end{aligned}$$

Luego $\tilde{\Phi}$ es definida positiva y por tanto invertible, pues es simétrica.

Este modelo se ha explicado con detalle pues constituye el ejemplo con el que testaremos el método que se presenta en el Capítulo 3 para evaluar el modelo de deformación paramétrico (2.3). Dos casos particulares de esta familia de deformación son:

- el modelo de localización ($\sigma = 1$ y $\lambda = -\mu$): $\varphi_\lambda(x) = x + \lambda$
- el modelo de escala ($\mu = 0$): $\varphi_\theta(x) = \frac{1}{\sigma}x$

Ejemplo 2.2 (Transformación logarítmica). Es fácil comprobar que si consideramos el parámetro θ variando en un conjunto compacto $\Theta \subset (0, +\infty)$, la familia de deformaciones

$$\begin{aligned} \varphi_\theta : (0, +\infty) &\longrightarrow \mathbb{R} \\ x &\longmapsto \varphi_\theta(x) = \theta \log(x) \end{aligned}$$

es admisible. Notemos que φ_θ es invertible en $(0, +\infty)$, para todo $\theta \neq 0$, con inversa $\varphi_\theta^{-1}(x) = \exp(\frac{x}{\theta})$, y φ_θ es estrictamente creciente si $\theta > 0$.

Ejemplo 2.3 (Composición $\varphi_\theta(x) = f \circ \tilde{\varphi}_\theta(x)$). Consideremos una función admisible $\tilde{\varphi}_\theta : [a, b] \rightarrow [c, d]$. Entonces, si $f : [c, d] \rightarrow [e, f]$ es una función creciente e invertible, la función $\varphi_\theta(x) = f \circ \tilde{\varphi}_\theta(x)$ es también admisible, cambiando el intervalo $[a, b]$ por $[c, d]$. La importancia de este ejemplo radica en la gran amplitud de deformaciones admisibles que incluye. Por ejemplo, el modelo *logit* que viene dado por

$$\varphi_\theta(x) = \frac{1}{1 + \exp(-\theta x)}$$

y que se ha demostrado muy útil para explicar muchas situaciones reales, puede obtenerse mediante la composición del modelo de escala con la función $f(x) = \frac{1}{1 + \exp(-x)}$.

2.3 Bootstrap con la distancia de Wasserstein

Como hemos visto, la distribución asintótica del estadístico $\inf_{\theta \in \Theta} U_n(\theta)$ en el [Teorema 2.2](#) depende de elementos desconocidos y, por lo tanto, debemos construir mediante bootstrap una estimación de su distribución límite. Para ello, necesitamos conocer algunos resultados sobre la distancia de Wasserstein presentados en [11], que posteriormente aplicaremos a las leyes bootstrap de tal estadístico. Las versiones de los resultados que aquí se presentan son para el caso de distancias cuadráticas de Wasserstein; sin embargo, las demostraciones para el caso general de distancias de Wasserstein con coste $L_r, r \geq 1$, son las mismas cambiando 2 por r convenientemente.

En los dos primeros apartados de esta sección, consideramos distribuciones en \mathbb{R}^d con momento de orden 2 finito, es decir, distribuciones en el espacio $\mathcal{W}_2(\mathbb{R}^d)$. En el último apartado, nos centraremos en el caso unidimensional

($d = 1$), para obtener un resultado que nos asegurará que la distribución bootstrap del estadístico de interés es un estimador consistente.

Recordemos que la distancia de Wasserstein con coste L_2 entre dos probabilidades $\mu, \nu \in \mathcal{W}_2(\mathbb{R}^d)$ se define como

$$W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int \|y - z\|^2 d\pi(y, z),$$

donde $\|\cdot\|$ es la norma euclídea en \mathbb{R}^d .

2.3.1 Dependencia continua de la distancia de Wasserstein

El siguiente teorema pone de manifiesto que las leyes de los costes de transporte entre las distribuciones empíricas dependen continuamente del coste de transporte entre las distribuciones poblacionales.

| Teorema 2.3. Sean ν, ν', η medidas de probabilidad en $\mathcal{W}_2(\mathbb{R}^d)$. Sean Y_1, \dots, Y_n vectores aleatorios i.i.d. con ley común ν , y Y'_1, \dots, Y'_n vectores aleatorios i.i.d. con ley común ν' . Si ν_n, ν'_n son las correspondientes distribuciones empíricas, entonces:

$$W_2(\mathcal{L}(W_2(\nu_n, \eta)), \mathcal{L}(W_2(\nu'_n, \eta))) \leq W_2(\nu, \nu').$$

Demostración. Denotemos por $\|\cdot\|_2$ la norma en el espacio $\mathcal{L}^2(\{1, \dots, n\} \times \mathbb{R}^d, \pi)$.

Definimos $T_n = W_2(\nu_n, \eta)$ y $T'_n = W_2(\nu'_n, \eta)$, y denotamos por $\Pi_n(\eta)$ el conjunto de probabilidades en $\{1, \dots, n\} \times \mathbb{R}^d$ con primera marginal igual a la distribución uniforme sobre $\{1, \dots, n\}$ y segunda marginal η . Consideremos la cantidad

$$a(\pi) = \left(\int_{\{1, \dots, n\} \times \mathbb{R}^d} \|Y_i - z\|^2 d\pi(i, z) \right)^{\frac{1}{2}}.$$

Notemos que entonces:

$$\begin{aligned} T_n = W_2(\nu_n, \eta) &= \inf_{\pi \in \Pi(\nu_n, \eta)} \left(\int \|y - z\|^2 d\pi(y, z) \right)^{\frac{1}{2}} \\ &= \inf_{\pi \in \Pi_n(\eta)} \left(\int_{\{1, \dots, n\} \times \mathbb{R}^d} \|Y_i - z\|^2 d\pi(y, z) \right)^{\frac{1}{2}} = \inf_{\pi \in \Pi_n(\eta)} a(\pi), \end{aligned}$$

donde la penúltima igualdad es consecuencia de que ν_n es la medida empírica asociada a (Y_1, \dots, Y_n) .

Análogamente, para la muestra (Y'_1, \dots, Y'_n) , definimos

$$a'(\pi) = \left(\int_{\{1, \dots, n\} \times \mathbb{R}^d} \|Y'_i - z\|^2 d\pi(i, z) \right)^{\frac{1}{2}},$$

y se tiene que $T'_n = \inf_{\pi \in \Pi_n(\eta)} a'(\pi)$.

Por otro lado, $|a(\pi) - a'(\pi)| = \left| \|Y_i - z\|_2 - \|Y'_i - z\|_2 \right|$, y aplicando la segunda desigualdad triangular,

$$|a(\pi) - a'(\pi)| \leq \|Y_i - z - Y'_i + z\|_2 = \|Y_i - Y'_i\|_2.$$

Ahora, observemos que en virtud del Teorema de Fubini,

$$\begin{aligned} \|Y_i - Y'_i\|_2 &= \left(\int_{\{1, \dots, n\} \times \mathbb{R}^d} \|Y_i - Y'_i\|^2 d\pi(i, z) \right)^{\frac{1}{2}} \\ &= \left(\int_{\{1, \dots, n\}} \|Y_i - Y'_i\|^2 d\nu_n(i) \right)^{\frac{1}{2}} = \left(\frac{1}{n} \sum_{i=1}^n \|Y_i - Y'_i\|^2 \right)^{\frac{1}{2}} \end{aligned}$$

En definitiva, $|a(\pi) - a'(\pi)| \leq \left(\frac{1}{n} \sum_{i=1}^n \|Y_i - Y'_i\|^2 \right)^{\frac{1}{2}}$ y, teniendo en cuenta las expresiones obtenidas para T_n y T'_n , llegamos a que:

$$|T_n - T'_n|^2 \leq \frac{1}{n} \sum_{i=1}^n \|Y_i - Y'_i\|^2.$$

Sea ahora (Y, Y') un par óptimo para ν y ν' , de manera que $E[\|Y - Y'\|^2] = W_2^2(\nu, \nu')$, y sean $(Y_1, Y'_1), \dots, (Y_n, Y'_n)$ realizaciones i.i.d. de (Y, Y') . Entonces, se tiene que:

$$\begin{aligned} E(|T_n - T'_n|^2) &\leq E\left(\frac{1}{n} \sum_{i=1}^n \|Y_i - Y'_i\|^2\right) = \frac{1}{n} \sum_{i=1}^n E\left(\|Y_i - Y'_i\|^2\right) \\ &= \frac{1}{n} \sum_{i=1}^n W_2^2(\nu, \nu') = W_2^2(\nu, \nu'). \end{aligned}$$

Como además $E(|T_n - T'_n|^2) = W_2^2(\mathcal{L}(T_n), \mathcal{L}(T'_n))$, se verifica la desigualdad:

$$W_2(\mathcal{L}(W_2(\nu_n, \eta)), \mathcal{L}(W_2(\nu'_n, \eta))) = W_2^2(\mathcal{L}(T_n), \mathcal{L}(T'_n)) \leq W_2(\nu, \nu').$$

■

Este teorema es de gran importancia práctica pues, combinado junto con técnicas bootstrap de remuestreo, nos permite conocer información acerca de la distribución de la distancia de Wasserstein entre una medida de probabilidad y su versión empírica. La siguiente propiedad de la distancia de Wasserstein, resulta de gran utilidad a la hora de formalizar esta idea.

Proposición 2.3. Sean X e Y dos variables aleatorias cualesquiera. Entonces, para todo $a > 0$,

$$W_2(\mathcal{L}(aX), \mathcal{L}(aY)) = aW_2(\mathcal{L}(X), \mathcal{L}(Y))$$

Demostración. Sea $a > 0$. Por definición de la distancia de Wasserstein,

$$\begin{aligned} W_2(\mathcal{L}(aX), \mathcal{L}(aY)) &= \left(\inf_{\pi \in \Pi(\mathcal{L}(aX), \mathcal{L}(aY))} \int \|x - y\|^2 d\pi(x, y) \right)^{\frac{1}{2}} \\ &= a \left(\inf_{\pi \in \Pi(\mathcal{L}(aX), \mathcal{L}(aY))} \frac{1}{a^2} \int \|x - y\|^2 d\pi(x, y) \right)^{\frac{1}{2}} \\ &= a \left(\inf_{\pi \in \Pi(\mathcal{L}(aX), \mathcal{L}(aY))} \int \left\| \frac{x}{a} - \frac{y}{a} \right\|^2 d\pi(x, y) \right)^{\frac{1}{2}}. \end{aligned}$$

Ahora, teniendo en cuenta que $\frac{1}{a}\mathcal{L}(aX) \stackrel{d}{=} \mathcal{L}(X)$, concluimos

$$\begin{aligned} W_2(\mathcal{L}(aX), \mathcal{L}(aY)) &= a \left(\inf_{\pi \in \Pi(\mathcal{L}(X), \mathcal{L}(Y))} \int \|x - y\|^2 d\pi(x, y) \right)^{\frac{1}{2}} \\ &= aW_2(\mathcal{L}(X), \mathcal{L}(Y)). \end{aligned}$$

■

Consideremos Y_1, \dots, Y_n v.a.i.i.d. $\nu \in \mathcal{W}_2(\mathbb{R}^d)$ con distribución empírica asociada ν_n , e $Y_1^*, \dots, Y_{m_n}^*$ i.i.d. (condicionalmente dadas Y_1, \dots, Y_n) con ley ν_n , es decir, una réplica bootstrap de tamaño m_n de Y_1, \dots, Y_n , con distribución empírica $\nu_{m_n}^*$. Si escribimos $\mathcal{L}^*(Z)$ para designar la ley de una variable aleatoria Z condicionada a Y_1, \dots, Y_n , entonces, tomando en el *Teorema 2.3* $\nu' = \nu_n$, tenemos que:

$$W_2(\mathcal{L}^*(W_2(\nu_{m_n}^*, \nu)), \mathcal{L}(W_2(\nu_{m_n}, \nu))) \leq W_2(\nu_n, \nu).$$

Sea $\{r_n\}_{n=0}^\infty$ una sucesión tal que:

- (i) $r_n > 0$, para todo $n \geq 0$
- (ii) $\lim_{n \rightarrow \infty} \frac{r_{m_n}}{r_n} = 0$

$$(iii) \quad W_2(\nu_n, \nu) = O_p\left(\frac{1}{r_n}\right).$$

Por la *Proposición 2.3*,

$$\begin{aligned} & W_2\left(\mathcal{L}^*(r_{m_n} W_2(\nu_{m_n}^*, \nu)), \mathcal{L}(r_{m_n} W_2(\nu_{m_n}, \nu))\right) \\ &= r_{m_n} W_2\left(\mathcal{L}^*(W_2(\nu_{m_n}^*, \nu)), \mathcal{L}(W_2(\nu_{m_n}, \nu))\right) \leq r_{m_n} W_2(\nu_n, \nu) = \frac{r_{m_n}}{r_n} r_n W_2(\nu_n, \nu). \end{aligned}$$

Teniendo en cuenta las condiciones (i)-(iii), se concluye que, cuando n tiende hacia infinito,

$$\frac{r_{m_n}}{r_n} r_n W_2(\nu_n, \nu) \longrightarrow 0, \text{ en probabilidad.}$$

Esto significa que la distribución de $r_{m_n} W_2(\nu_{m_n}^*, \nu)$ es próxima, en distancia de Wasserstein, a la distribución de $r_{m_n} W_2(\nu_{m_n}, \nu)$, condicionalmente dados Y_1, \dots, Y_n . Si además de la condición (iii), se tiene la convergencia

$$r_n W_2(\nu_n, \nu) \xrightarrow[n \rightarrow \infty]{} \gamma(\nu),$$

para una distribución $\gamma(\nu)$ no degenerada, entonces, en particular, se sigue verificando

$$r_{m_n} W_2(\nu_{m_n}, \nu) \xrightarrow[n \rightarrow \infty]{} \gamma(\nu).$$

La razón por la cual es necesario que $\gamma(\nu)$ sea no degenerada se detalla en [18] (ver *Lemma 1*).

Por lo tanto, si $\hat{q}_{m_n}(\alpha)$ denota el cuantil α de la distribución condicionada $\mathcal{L}^*(r_{m_n} W_2(\nu_{m_n}^*, \nu))$, entonces, bajo ciertas condiciones de regularidad sobre la distribución $\gamma(\nu)$, se tiene que

$$\mathbb{P}(r_n W_2(\nu_n, \nu) \leq \hat{q}_{m_n}(\alpha)) \xrightarrow[n \rightarrow \infty]{} \alpha.$$

Esto muestra que los cuantiles bootstrap, es decir, los cuantiles de la ley condicional de $r_{m_n} W_2(\nu_{m_n}^*, \nu)$ son estimadores consistentes de los cuantiles de la distribución de $r_n W_2(\nu_n, \nu)$. Dichos cuantiles bootstrap pueden ser aproximados por simulación de Monte-Carlo.

Resulta que si consideramos distribuciones en \mathbb{R} , la distribución $\gamma(\nu)$ adquiere una forma conocida si la distribución ν cumple determinadas condiciones de integrabilidad y derivabilidad. Concretamente, en [10] se prueba que bajo

ciertas condiciones, tomando $r_n = \sqrt{n}$,

$$\sqrt{n}W_2(\nu_n, \nu) \xrightarrow{d} \left(\int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))} dt \right)^{\frac{1}{2}}. \quad (2.4)$$

Por consiguiente, en este caso, será posible aproximar los cuantiles de la distribución por $\hat{q}_{m_n}(\alpha)$.

2.3.2 Dependencia continua de la variación de Wasserstein

Un resultado similar al anterior prueba que las leyes de las 2-variaciones de Wasserstein de las medidas empíricas dependen de manera continua de las medidas poblacionales subyacentes. En este contexto, es útil observar que se verifica la siguiente igualdad:

$$V_2^2(\nu_1, \dots, \nu_I) = \inf_{\pi \in \Pi(\nu_1, \dots, \nu_I)} \int T(y_1, \dots, y_I) d\pi(y_1, \dots, y_I), \quad (2.5)$$

donde $\Pi(\nu_1, \dots, \nu_I)$ es el conjunto de probabilidades en \mathbb{R}^d con marginales ν_1, \dots, ν_I , y $T(y_1, \dots, y_I) = \min_{z \in \mathbb{R}^d} \frac{1}{I} \sum_{i=1}^I \|y_i - z\|^2$. En efecto,

$$\begin{aligned} V_2^2(\nu_1, \dots, \nu_I) &= \inf_{\eta \in \mathcal{W}_2(\mathbb{R}^d)} \left(\frac{1}{I} \sum_{i=1}^I W_2^2(\eta, \nu_i) \right) \\ &= \inf_{\eta \in \mathcal{W}_2(\mathbb{R}^d)} \left(\frac{1}{I} \sum_{i=1}^I \inf_{\pi \in \Pi(\eta, \nu_i)} \int \|x - y\|^2 d\pi(x, y) \right) \\ &= \inf_{\eta \in \mathcal{W}_2(\mathbb{R}^d)} \left[\inf_{\pi \in \Pi(\eta, \nu_1, \dots, \nu_I)} \left(\frac{1}{I} \sum_{i=1}^I \int \|x - y_i\|^2 d\pi(x, y_1, \dots, y_I) \right) \right] \\ &= \inf_{\pi \in \Pi(\nu_1, \dots, \nu_I)} \left[\int \left(\min_{z \in \mathbb{R}^d} \frac{1}{I} \sum_{i=1}^I \|y_i - z\|^2 \right) d\pi(y_1, \dots, y_I) \right]. \end{aligned}$$

Teorema 2.4. Sean $Y_{i,1}, \dots, Y_{i,n_i}$ e $Y'_{i,1}, \dots, Y'_{i,n'_i}$ dos colecciones de v.a.i.i.d. con distribución ν_i y ν'_i , respectivamente, con $1 \leq i \leq I$. Consideremos, para cada $i \in 1, \dots, I$, las medidas de distribución empíricas asociadas a dichas colecciones $\nu_{n_i,i}, \nu'_{n'_i,i}$. Se verifica la desigualdad

$$W_2^2(\mathcal{L}(V_2(\nu_{n_1,1}, \dots, \nu_{n_I,I})), \mathcal{L}(V_2(\nu'_{n_1,1}, \dots, \nu'_{n_I,I}))) \leq \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \nu'_i)$$

Demostración. A lo largo de la prueba, escribiremos $V_{2,n} = V_2(\nu_{n_1,1}, \dots, \nu_{n_I,I})$ y $V'_{2,n} = V_2(\nu'_{n_1,1}, \dots, \nu'_{n_I,I})$. De (2.5) y por ser ν_{n_i} la medida empírica asociada a la muestra $(Y_{i,1}, \dots, Y_{i,n_i})$

$$\begin{aligned} V_{2,n}^2 &= \inf_{\pi \in \Pi(\nu_{n_1}, \dots, \nu_{n_I})} \int T(y_1, \dots, y_I) d\pi(y_1, \dots, y_I) \\ &= \inf_{\pi \in \Pi(U_1, \dots, U_I)} \int T(j_1, \dots, j_I) d\pi(j_1, \dots, j_I), \end{aligned}$$

donde U_i es la distribución uniforme discreta en el conjunto $\{1, \dots, n_i\}$ y $T(j_1, \dots, j_I) = \min_{z \in \mathbb{R}^d} \frac{1}{I} \sum_{i=1}^I \|Y_{i,j_i} - z\|^2$. Sea $T'(j_1, \dots, j_I)$ la versión equivalente de esta función calculada para la muestra (Y'_1, \dots, Y'_n) . Tenemos que:

$$|T(j_1, \dots, j_I)^{\frac{1}{2}} - T'(j_1, \dots, j_I)|^2 \leq \frac{1}{I} \sum_{i=1}^I \|Y_{i,j_i} - Y'_{i,j_i}\|^2,$$

lo cual implica

$$\begin{aligned} &\left| \left(\int T(j_i, \dots, j_I) d\pi(j_1, \dots, j_I) \right)^{\frac{1}{2}} - \left(\int T'(j_i, \dots, j_I) d\pi(j_1, \dots, j_I) \right)^{\frac{1}{2}} \right|^2 \\ &\leq \int \frac{1}{I} \sum_{i=1}^I \|Y_{i,j_i} - Y'_{i,j_i}\|^2 d\pi(j_1, \dots, j_I) = \frac{1}{I} \sum_{i=1}^I \int \|Y_{i,j_i} - Y'_{i,j_i}\|^2 d\pi(j_1, \dots, j_I) \\ &= \frac{1}{I} \sum_{i=1}^I \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \|Y_{i,j} - Y'_{i,j}\|^2 \right). \end{aligned}$$

En definitiva, hemos llegado a que

$$|V_{2,n} - V'_{2,n}|^2 \leq \frac{1}{I} \sum_{i=1}^I \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \|Y_{i,j} - Y'_{i,j}\|^2 \right).$$

Si ahora tomamos (Y_i, Y'_i) un par óptimo para las medidas ν_i y ν'_i , y consideramos copias $(Y_{i,1}, Y'_{i,1}), \dots, (Y_{i,n_i}, Y'_{i,n_i})$ i.i.d., entonces, tenemos que

$$\begin{aligned} E(|V_{2,n} - V'_{2,n}|^2) &\leq \frac{1}{I} \sum_{i=1}^I \left(\frac{1}{n_i} \sum_{j=1}^{n_i} E(\|Y_{i,j} - Y'_{i,j}\|^2) \right) \\ &= \frac{1}{I} \sum_{i=1}^I \frac{1}{n_i} n_i W_2^2(\nu_i, \nu'_i) = \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \nu'_i). \end{aligned}$$

Finalmente, como $E(|V_{2,n} - V'_{2,n}|^2) = W_2^2(\mathcal{L}(V_{2,n}), \mathcal{L}(V'_{2,n}))$, se concluye la prueba. ■

2.3.3 Bondad de ajuste del test para el modelo de deformación paramétrico

Recordemos que las inferencias sobre el modelo de deformación paramétrico (2.3) se basan en el estadístico

$$u_n^2 := \inf_{\theta \in \Theta} U_n(\theta) = \inf_{\theta \in \Theta} \frac{1}{I} \sum_{i=1}^I W_2^2(v_{n,i}(\theta), v_{n,B}(\theta)),$$

donde $v_{n,i}(\theta)$ es la medida empírica asociada a la muestra $(Z_{i,j}(\theta))_{1 \leq j \leq n} = (\varphi_{\theta_i}(X_{i,j}))_{1 \leq j \leq n}$, y $(X_{i,j})_{1 \leq j \leq n}$ son i.i.d. v_i .

Supongamos que observamos otras I muestras independientes $(X'_{i,j})_{1 \leq j \leq n}$, $1 \leq i \leq I$, i.i.d. con distribución v'_i . Sean $v'_{n,i}$, u'_n y U'_n las versiones equivalentes para estas muestras, y $\mathcal{L}(u_n)$ y $\mathcal{L}(u'_n)$ las leyes de las variables aleatorias u_n y u'_n , respectivamente. Bajo ciertas condiciones de regularidad sobre las funciones de deformación, se puede asegurar que la distancia de Wasserstein entre las distribuciones de los estadísticos u_n y u'_n , asociados a las dos colecciones distintas de muestras $(X_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$ y $(X'_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$, es controlada por la distancia de Wasserstein entre las distribuciones poblacionales subyacentes en dichas muestras. Esto es lo que afirma el siguiente resultado.

Teorema 2.5. *En la situación anterior y bajo las hipótesis (A1), (A3) y (A4)*

$$W_2^2(\mathcal{L}(u_n), \mathcal{L}(u'_n)) \leq \sup_{\substack{x \in (c,d) \\ \lambda \in \Lambda}} |d\varphi_\lambda(x)|^2 \frac{1}{I} \sum_{i=1}^I W_2^2(v_i, v'_i).$$

Demostración. Como en la demostración del Teorema 2.4,

$$\begin{aligned} U_n(\theta) &= \inf_{\pi \in \Pi(v_{n,1}(\theta_1), \dots, v_{n,I}(\theta_I))} \int T(y_1, \dots, y_I) d\pi(y_1, \dots, y_I) \\ &= \inf_{\pi \in \Pi(U_1, \dots, U_I)} \int T(j_1, \dots, j_I) d\pi(j_1, \dots, j_I), \end{aligned}$$

donde para cada $i = 1, \dots, I$, la variable aleatoria U_i es uniforme discreta en el conjunto $\{1, \dots, n\}$, y $T(j_1, \dots, j_I) = \min_{y \in \mathbb{R}} \frac{1}{I} \sum_{i=1}^I |Z_{i,j}(\theta_i) - y|^2$. Sea $T'(j_1, \dots, j_I)$

la misma función calculada para la muestra $(Z'_{i,j}(\theta_i))_{1 \leq j \leq n}$. Por el teorema del valor medio de Lagrange, se tiene que

$$|Z_{i,j}(\theta_i) - Z'_{i,j}(\theta_i)|^2 = |\varphi_{\theta_i}(X_{i,j}) - \varphi_{\theta_i}(X'_{i,j})|^2 = |\varphi_{\theta_i}(\xi)|^2 |X_{i,j} - X'_{i,j}|^2,$$

para un punto intermedio $\xi \in (\min\{X_{i,j}, X'_{i,j}\}, \max\{X_{i,j}, X'_{i,j}\})$. Por las hipótesis impuestas, $\|\varphi'_{\theta_i}\|_{\infty} := \sup_{\substack{x \in (c,d) \\ \lambda \in \Lambda}} |d\varphi_{\lambda}(x)| < \infty$, y por lo tanto, se tiene la siguiente cota superior:

$$|Z_{i,j}(\theta_i) - Z'_{i,j}(\theta_i)|^2 \leq \|\varphi'_{\theta_i}\|_{\infty}^2 |X_{i,j} - X'_{i,j}|^2.$$

De esta acotación, deducimos que

$$\left| T(j_1, \dots, j_I)^{\frac{1}{2}} - T'(j_1, \dots, j_I)^{\frac{1}{2}} \right|^2 \leq \|\varphi'_{\theta_i}\|_{\infty}^2 \frac{1}{I} \sum_{i=1}^I |X_{i,j} - X'_{i,j}|^2.$$

Ahora, por la monotonía de la integral, se verifica que

$$\begin{aligned} & \int \left| T(j_1, \dots, j_I)^{\frac{1}{2}} - T'(j_1, \dots, j_I)^{\frac{1}{2}} \right|^2 d\pi(j_1, \dots, j_I) \\ & \leq \|\varphi'_{\theta_i}\|_{\infty}^2 \frac{1}{I} \sum_{i=1}^I \int |X_{i,j} - X'_{i,j}|^2 d\pi(j_1, \dots, j_I) = \|\varphi'_{\theta_i}\|_{\infty}^2 \frac{1}{I} \sum_{i=1}^I \frac{1}{n} \sum_{j=1}^n |X_{i,j} - X'_{i,j}|^2 \end{aligned}$$

Por otro lado, por la desigualdad triangular

$$\begin{aligned} & \int \left| T(j_1, \dots, j_I)^{\frac{1}{2}} - T'(j_1, \dots, j_I)^{\frac{1}{2}} \right|^2 d\pi(j_1, \dots, j_I) \\ & \geq \left| \left(\int T(j_1, \dots, j_I)^{\frac{1}{2}} d\pi(j_1, \dots, j_I) \right)^{\frac{1}{2}} - \left(\int T'(j_1, \dots, j_I)^{\frac{1}{2}} d\pi(j_1, \dots, j_I) \right)^{\frac{1}{2}} \right|^2 \\ & \geq |V_{2,n} - V'_{2,n}|^2 \end{aligned}$$

y, en consecuencia,

$$(u_n - u'_n)^2 \leq \|\varphi'_{\theta_i}\|_{\infty}^2 \frac{1}{I} \sum_{i=1}^I \left(\frac{1}{n} \sum_{j=1}^n |X_{i,j} - X'_{i,j}|^2 \right).$$

Sea (X_i, X'_i) un par óptimo para ν y ν' y consideremos $(X_{i,j}, X'_{i,j}), j = 1, \dots, n$,

copias i.i.d. de (X_i, X'_i) . Tomando esperanzas en la desigualdad anterior,

$$\begin{aligned} W_2^2(\mathcal{L}(u_n), \mathcal{L}(u'_n)) &= E \left[(u_n - u'_n)^2 \right] \leq E \left[\|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I \left(\frac{1}{n} \sum_{j=1}^n |X_{i,j} - X'_{i,j}|^2 \right) \right] \\ &= \|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I \left[\frac{1}{n} \sum_{j=1}^n E \left(|X_{i,j} - X'_{i,j}|^2 \right) \right] = \|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \nu'_i). \end{aligned}$$

■

Una consecuencia muy útil de ese teorema surge al considerar, para cada $i = 1, \dots, I$, muestras bootstrap $X_{i,1}^*, \dots, X_{i,m_n}^*$ i.i.d. de tamaño m_n con distribución común $\nu_{n,i}$. Sean $\nu_{m_n,i}^*$, $i = 1, \dots, I$, la medida empírica asociada a la muestra $X_{i,1}^*, \dots, X_{i,m_n}^*$ (condicionalmente dadas $X_{i,1}, \dots, X_{i,n}$) y $u_{m_n}^* = \left(\inf_{\theta \in \Theta} U_{m_n}^* \right)^{\frac{1}{2}}$ la versión equivalente del estadístico. Aplicando el *Teorema 2.5* con $\nu'_i = \nu_{n,i}$, tenemos que

$$W_2^2(\mathcal{L}(u_{m_n}), \mathcal{L}(u_{m_n}^*)) \leq \|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \nu_{n,i}).$$

Utilizando la *Proposición 2.3*, vemos que

$$\begin{aligned} W_2^2(\mathcal{L}(\sqrt{m_n}u_{m_n}), \mathcal{L}^*(\sqrt{m_n}u_{m_n}^*)) &= m_n W_2^2(\mathcal{L}(u_{m_n}), \mathcal{L}^*(u_{m_n}^*)) \leq \\ &\leq m_n \|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_i, \nu_{n,i}) = \frac{m_n}{n} \|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I n W_2^2(\nu_i, \nu_{n,i}). \end{aligned}$$

En [11] (ver demostración del Teorema 4.1) se prueba que $W_2(\nu_i, \nu_{n,i}) = O_p\left(\frac{1}{\sqrt{n}}\right)$ (notemos que si $d=1$, esto se deduce también de la expresión (2.4)). Por lo tanto, si m_n y n son tales que $\frac{m_n}{n} \xrightarrow{n \rightarrow \infty} 0$, se tiene la convergencia en distribución

$$\frac{m_n}{n} \|\varphi'_{\theta_i}\|_\infty^2 \frac{1}{I} \sum_{i=1}^I n W_2^2(\nu_i, \nu_{n,i}) \xrightarrow{d} 0.$$

Esto nos dice que las distribuciones del estadístico $\sqrt{m_n}u_{m_n}$ y de su versión bootstrap $\sqrt{m_n}u_{m_n}^*$ (condicionalmente dadas $X_{i,1}, \dots, X_{i,n}$) están próximas en distancia de Wasserstein, bajo ciertas condiciones (en el caso $d=1$, bajo las condiciones que garantizan (2.4)).

Si denotamos por $\gamma(G, \theta^*)$ a la distribución límite en el *Teorema 2.2*, se tiene que

$$\sqrt{nu_n} \xrightarrow{d} (\gamma(G, \theta^*))^{\frac{1}{2}} \implies \sqrt{m_n} u_{m_n} \xrightarrow{d} (\gamma(G, \theta^*))^{\frac{1}{2}},$$

y, por lo que acabamos de ver, esto también implica que

$$\sqrt{m_n} u_{m_n}^* \xrightarrow{d} (\gamma(G, \theta^*))^{\frac{1}{2}}.$$

Finalmente, por ser $x \mapsto x^2$ continua, se deduce que

$$m_n \inf_{\theta \in \Theta} U_{m_n}^* \xrightarrow{d} \gamma(G, \theta^*).$$

Este razonamiento junto con el teorema de Glivenko-Cantelli y resultados de convergencia de los cuantiles empíricos, prueban lo siguiente:

Corolario 2.1. Si $m_n \rightarrow \infty$ y $\frac{m_n}{n} \rightarrow 0$, entonces bajo las hipótesis **(A1)**-**(A10)** y **(TCL)**, se tiene que

$$\mathcal{L}^* \left(m_n \inf_{\theta \in \Theta} U_{m_n}^*(\theta) \right) \xrightarrow{d} \gamma(G, \theta^*), \text{ en probabilidad.}$$

En particular, si $\hat{q}_{m_n}(\alpha)$ denota el cuantil condicional α (dadas $X_{i,1}, \dots, X_{i,n}, i = 1, \dots, I$) de $m_n \inf_{\theta \in \Theta} U_{m_n}^*(\theta)$ y la función cuantil de $\gamma(G, \theta^*)$ es continua con respecto a α ,

$$\mathbb{P} \left(n \inf_{\theta \in \Theta} U_n(\theta) \leq \hat{q}_{m_n}(\alpha) \right) \xrightarrow{n \rightarrow \infty} \alpha.$$

El *Corolario 2.1* proporciona un procedimiento para evaluar la bondad de ajuste del modelo de deformación paramétrico pues, bajo las hipótesis **(A1)**-**(A10)** y **(TCL)**, permite construir un test para contrastar:

$$\inf_{\theta \in \Theta} U(\theta) = 0 \quad (\mathcal{H}_0)$$

$$\inf_{\theta \in \Theta} U(\theta) > 0 \quad (\mathcal{H}_1)$$

En base al estadístico $n \inf_{\theta \in \Theta} U_n(\theta)$, es posible conocer el nivel de una región asintótica de la forma $\left\{ n \inf_{\theta \in \Theta} U_n(\theta) > \lambda_n \right\}$, en virtud del *Corolario 2.1*.

Si $\hat{q}_{m_n}(1-\alpha)$ denota el cuantil condicional $(1-\alpha)$ (dadas las muestras $X_{i,1}, \dots, X_{i,n}$, $i = 1, \dots, I$) de la distribución bootstrap $m_n \inf_{\theta \in \Theta} U_{m_n}^*(\theta)$, entonces

$$\mathbb{P} \left(n \inf_{\theta \in \Theta} U_n(\theta) > \hat{q}_{m_n}(1-\alpha) \right) \xrightarrow{n \rightarrow \infty} \alpha.$$

En conclusión, $\left\{ n \inf_{\theta \in \Theta} U_n(\theta) > \hat{q}_{m_n}(1-\alpha) \right\}$ es una región asintótica de nivel α y el cuantil $\hat{q}_{m_n}(1-\alpha)$ puede aproximarse por un método de simulación Monte-Carlo. En el siguiente capítulo, concretamente en la sección 3.2, se propone un procedimiento general para diseñar estas simulaciones.

3 | ASPECTOS COMPUTACIONALES.

En este capítulo se dan ciertas ideas útiles a la hora de realizar el cálculo efectivo de la función criterio, y se resuelve este problema en el caso concreto del modelo de deformación de localización y escala. En una segunda parte, se especifica un procedimiento general para obtener la distribución bootstrap del estadístico $\inf_{\theta \in \Theta} U_n(\theta)$, presentando con detalle el ejemplo de localización y escala. Como vimos en la sección 1.4 dedicada a conocimientos básicos sobre bootstrap, esta técnica se basa en remuestrear la muestra y este remuestreo se puede hacer de diferentes maneras. Una de ellas, es las que presentamos en este capítulo bajo el nombre de *bootstrap de residuos*, para la cual se ha diseñado un procedimiento específico que, de nuevo, se ha particularizado al modelo de localización y escala.

Recordemos que la expresión formal de la función criterio considerada para valorar la bondad de ajuste del modelo de deformación (2.3) en el que se registran datos de I señales diferentes es:

$$\inf_{\theta \in \Theta} U_n(\theta) = \inf_{\theta \in \Theta} \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_{n,i}(\theta_i), \nu_{n,B}(\theta)).$$

donde $\nu_{n,B}(\theta)$ es el baricentro de Wasserstein de la colección de probabilidades $(\nu_{n,i}(\theta_i))_{i=1}^I$ asociadas a las muestras $(Z_{i,j}(\theta_i))_{j=1}^n$.

Como la medida de referencia ν está definida en un subconjunto de \mathbb{R} , el valor de U_n se calcula fácilmente teniendo en cuenta la expresión para la distancia de Wasserstein dada por el *Teorema 1.3*:

$$U_n(\theta) = \frac{1}{I} \sum_{i=1}^I W_2^2(\nu_{n,i}(\theta_i), \nu_{n,B}(\theta)) = \frac{1}{I} \sum_{i=1}^I \int_0^1 (F_{n,i}(\theta_i)^{-1}(t) - F_{n,B}(\theta)^{-1}(t))^2 dt,$$

donde $F_{n,i}(\theta_i)$ es la función de distribución empírica y asociada a la muestra

$(Z_{i,j}(\theta_i))_{j=1}^n = (\varphi_{\theta_i}(X_{i,j}))_{j=1}^n$, y $F_{n,B}(\theta)$ es la correspondiente versión asociada a la muestra baricentro $(\frac{1}{I} \sum_{i=1}^I Z_{i(j)}(\theta_i))_{j=1}^n$.

En general, observemos que si F_n denota la función de distribución empírica asociada a la muestra (Y_1, \dots, Y_n) , entonces

$$F_n^{-1}(t) = Y_{(j)}, \text{ para } \frac{j-1}{n} < t \leq \frac{j}{n},$$

donde $Y_{(1)} \leq \dots \leq Y_{(i)} \leq \dots \leq Y_{(n)}$ son los estadísticos de orden asociados a la muestra (Y_1, \dots, Y_n) . Entonces, si $Z_{i(j)}(\theta_i)$ es el estadístico de orden asociado a la muestra $(Z_{ij}(\theta_i))_{j=1}^n$,

$$F_{n,i}^{-1}(\theta_i)(t) = Z_{i(j)}(\theta_i), \text{ para } \frac{j-1}{n} < t \leq \frac{j}{n},$$

y

$$F_{n,B}^{-1}(\theta)(t) = \frac{1}{I} \sum_{i=1}^I Z_{i(j)}(\theta_i), \text{ para } \frac{j-1}{n} < t \leq \frac{j}{n}.$$

Por lo tanto, la expresión para $U_n(\theta)$ viene dada por

$$\begin{aligned} U_n(\theta) &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[F_{n,i}^{-1}(\theta_i) \left(\frac{j}{n} \right) - F_{n,B}^{-1}(\theta) \left(\frac{j}{n} \right) \right]^2 \\ &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[Z_{i(j)}(\theta_i) - \frac{1}{I} \sum_{i=1}^I Z_{i,j}(\theta_i) \right]^2. \end{aligned}$$

3.1 Implementación del modelo de localización y escala

En esta sección, mostramos el ejemplo del modelo de localización y escala, para el que hemos obtenido una expresión explícita, tanto para los parámetros que minimizan U_n , como para dicho valor mínimo. En primer lugar, trataremos el caso más simple en el que se registran 2 muestras diferentes, en el que el problema resulta más manejable. Posteriormente, se desarrolla el caso general.

Con la forma específica para las funciones de deformación dada en el *Ejemplo 2.1*, el modelo (2.3) se enuncia de la siguiente manera:

existen $\theta^* = (\theta_1^*, \dots, \theta_I^*) \in \Theta = \Lambda^I$ y $\Lambda \subset (\mathbb{R} \times \mathbb{R}^+)$, con $\theta_i^* = (\mu_i^*, \sigma_i^*) \in \Lambda$, y

$(\varepsilon_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$ v.a.i.i.d. con distribución ν , tales que

$$X_{i,j} = \varphi_{\theta_i^*}^{-1}(\varepsilon_{i,j}) = \mu_i^* + \sigma_i^* \varepsilon_{i,j}, 1 \leq i \leq I, 1 \leq j \leq n. \quad (3.1)$$

3.1.1 Caso $l=2$ muestras

Consideremos primero el caso en el que se observan dos muestras independientes:

$$\begin{aligned} X_{1,1}, \dots, X_{1,n} &\text{ i.i.d. } \nu_1 \\ X_{2,1}, \dots, X_{2,n} &\text{ i.i.d. } \nu_2. \end{aligned}$$

Sea $\theta = ((\mu_1, \sigma_1), (\mu_2, \sigma_2)) \in \Theta$, donde $\Theta = \Lambda^2$ y $\Lambda \subset \mathbb{R} \times \mathbb{R}^+$. Según lo anterior, tenemos la siguiente expresión para la versión empírica del criterio de interés:

$$\begin{aligned} U_n(\theta) &= \frac{1}{2n} \sum_{i=1}^2 \sum_{j=1}^n \left[\frac{X_{i(j)} - \mu_i}{\sigma_i} - \frac{1}{2} \sum_{k=1}^2 \frac{X_{k(j)} - \mu_k}{\sigma_k} \right]^2 \\ &= \frac{1}{2n} \left\{ \sum_{j=1}^n \left[\frac{X_{1(j)} - \mu_1}{\sigma_1} - \frac{1}{2} \left(\frac{X_{1(j)} - \mu_1}{\sigma_1} + \frac{X_{2(j)} - \mu_2}{\sigma_2} \right) \right]^2 + \right. \\ &\quad \left. + \sum_{j=1}^n \left[\frac{X_{2(j)} - \mu_2}{\sigma_2} - \frac{1}{2} \left(\frac{X_{1(j)} - \mu_1}{\sigma_1} + \frac{X_{2(j)} - \mu_2}{\sigma_2} \right) \right]^2 \right\} \\ &= \frac{1}{2n} \left\{ \frac{1}{4} \sum_{j=1}^n \left(\frac{X_{1(j)} - \mu_1}{\sigma_1} - \frac{X_{2(j)} - \mu_2}{\sigma_2} \right)^2 + \frac{1}{4} \sum_{j=1}^n \left(\frac{X_{2(j)} - \mu_2}{\sigma_2} - \frac{X_{1(j)} - \mu_1}{\sigma_1} \right)^2 \right\} \\ &= \frac{1}{4n} \sum_{j=1}^n \left(\frac{X_{1(j)} - \mu_1}{\sigma_1} - \frac{X_{2(j)} - \mu_2}{\sigma_2} \right)^2. \end{aligned}$$

Por la *Observación 2.2*, asumimos que $\theta_1^* = (\mu_1^*, \sigma_1^*)$ es conocido y, entonces, $U_n(\theta_1, \theta_2) = U_n(\theta_1^*, \theta_2) \equiv U_n(\theta_2)$. Ahora, con el objetivo de simplificar los cálculos, consideremos el parámetro

$$(\alpha_2, \beta_2) \in \mathbb{R}^+ \times \mathbb{R} \text{ con } \alpha_2 = \frac{1}{\sigma_2} \text{ y } \beta_2 = \frac{\mu_2}{\sigma_2},$$

y las constantes

$$\alpha_1^* = \frac{1}{\sigma_1^*} \text{ y } \beta_1^* = \frac{\mu_1^*}{\sigma_1^*},$$

y reparametrizamos la función U_n :

$$\begin{aligned} U_n(\alpha_2, \beta_2) &= \frac{1}{4n} \sum_{j=1}^n \left[\alpha_1^* X_{1(j)} - \beta_1^* - (\alpha_2 X_{2(j)} - \beta_2) \right]^2 \\ &= \frac{1}{4n} \sum_{j=1}^n \left[\alpha_1^* X_{1(j)} - \alpha_2 X_{2(j)} - (\beta_1^* - \beta_2) \right]^2. \end{aligned} \quad (3.2)$$

Fijado $\alpha_2 \in \mathbb{R}^+$, hallaremos primero el punto $\hat{\beta}_2$ que minimiza la expresión (3.2) en $\beta_2 \in \mathbb{R}$, es decir,

$$\hat{\beta}_2 = \operatorname{argmin}_{\beta_2 \in \mathbb{R}} U_n(\alpha_2, \beta_2),$$

y, posteriormente, obtendremos la expresión de α_2^* que hace mínimo $U_n(\alpha_2, \hat{\beta}_2)$. Esto es equivalente a decir que $(\hat{\alpha}_2, \hat{\beta}_2) = \operatorname{argmin}_{(\alpha_2, \beta_2) \in \mathbb{R}^+ \times \mathbb{R}} U_n(\alpha_2, \beta_2)$.

De (3.2) y de la definición de esperanza de una cantidad finita de valores, deducimos que, para cada $\alpha_2 \in \mathbb{R}^+$ fijo,

$$\hat{\beta}_2 = \operatorname{argmin}_{\beta_2 \in \mathbb{R}} U_n(\beta_2) \iff \hat{\beta}_2 = \beta_1^* - \alpha_1^* \bar{X}_1 + \alpha_2 \bar{X}_2.$$

Por lo tanto,

$$\begin{aligned} U_n(\alpha_2, \hat{\beta}_2) &\equiv U_n(\alpha_2) = \frac{1}{4n} \sum_{j=1}^n \left[\alpha_1^* X_{1(j)} - \alpha_2 X_{2(j)} - (\alpha_1^* \bar{X}_1 - \alpha_2 \bar{X}_2) \right]^2 \\ &= \frac{1}{4n} \sum_{j=1}^n \left[\alpha_1^* (X_{1(j)} - \bar{X}_1) - \alpha_2 (X_{2(j)} - \bar{X}_2) \right]^2 \\ &= \frac{1}{4} \left[(\alpha_1^*)^2 \hat{V}ar(X_1) + \alpha_2^2 \hat{V}ar(X_2) - 2\alpha_1^* \alpha_2 \hat{C}ov(X_1, X_2) \right], \end{aligned}$$

donde $\hat{V}ar$ y $\hat{C}ov$ denotan varianzas y covarianzas muestrales.

Observemos que ésta es una función cuadrática en α_2 y además, el coeficiente que acompaña a α_2^2 es positivo. En consecuencia, el punto $\hat{\alpha}_2$ donde $U_n(\alpha_2)$ alcanza su mínimo absoluto es:

$$\frac{\partial U_n(\alpha_2)}{\partial \alpha_2} \Big|_{\alpha_2 = \hat{\alpha}_2} = 0 \iff 2\hat{\alpha}_2 \hat{V}ar(X_2) - 2\hat{C}ov(X_1, X_2) = 0 \iff \hat{\alpha}_2 = \frac{\hat{C}ov(X_1, X_2)}{\hat{V}ar(X_2)}.$$

En definitiva, hemos probado que el punto en el que la función U_n alcanza el mínimo es

$$\hat{\theta}_2 = (\hat{\alpha}_2, \hat{\beta}_2) = \left(\frac{\hat{C}ov(X_1, X_2)}{\hat{V}ar(X_2)}, \beta_1^* - \alpha_1^* \bar{X}_1 + \hat{\alpha}_2 \bar{X}_2 \right)$$

$$= \left(\frac{\widehat{Cov}(X_1, X_2)}{\widehat{Var}(X_2)}, \beta_1^* - \alpha_1^* \bar{X}_1 + \frac{\widehat{Cov}(X_1, X_2)}{\widehat{Var}(X_2)} \bar{X}_2 \right),$$

y, en este punto, el valor mínimo del criterio resulta ser

$$U_n(\hat{\theta}_2) = \frac{1}{4} \left((\alpha_1^*)^2 - \frac{\widehat{Cov}(X_1, X_2)^2}{\widehat{Var}(X_2)} \right).$$

3.1.2 Caso general

Consideremos ahora n realizaciones de I variables independientes

$$\begin{array}{cccc} X_{1,1}, & X_{1,2}, & \dots, & X_{1,n} & i.i.d. \sim \nu_1 \\ X_{2,1}, & X_{2,2}, & \dots, & X_{2,n} & i.i.d. \sim \nu_2 \\ \vdots & \vdots & \ddots & \vdots & \\ X_{I,1}, & X_{I,1}, & \dots, & X_{I,1} & i.i.d. \sim \nu_I \end{array}$$

y repitamos el mismo razonamiento que en el caso anterior. La expresión general del criterio para evaluar la validez del modelo de deformación es

$$U_n(\theta) = \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\frac{X_{i(j)} - \mu_i}{\sigma_i} - \frac{1}{I} \sum_{k=1}^I \frac{X_{k(j)} - \mu_k}{\sigma_k} \right]^2.$$

De manera análoga, tomamos $\check{\theta} = (\alpha_i, \beta_i)_{i=1}^I$, donde $\alpha_i = \frac{1}{\sigma_i}, \beta_i = \frac{\mu_i}{\sigma_i}, i = 1, \dots, I$, y reparametrizamos

$$\begin{aligned} U_n(\check{\theta}) &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I (\alpha_k X_{k(j)} - \beta_k) \right]^2 \\ &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right]^2. \end{aligned}$$

Para cada $i = 1, \dots, I$, definimos

$$u_i := \left[\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right]^2.$$

Sean $l \in \{2, \dots, I\}$, $\alpha = (\alpha_i)_{i=1}^I$ y $\beta = (\beta_i)_{i=1}^I$, tenemos que:

- Si $l \neq i$, $\frac{\partial u_i(\alpha, \beta)}{\partial \alpha_l} = 2 \left(\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) \left(-\frac{1}{I} X_{l(j)} \right)$
- Si $l = i$, $\frac{\partial u_i(\alpha, \beta)}{\partial \alpha_l} = 2 \left(\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) \left(X_{l(j)} - \frac{1}{I} X_{l(j)} \right)$.

De aquí deducimos que

$$\begin{aligned}
\frac{\partial U_n(\alpha, \beta)}{\partial \alpha_l} &= \frac{1}{nI} \sum_{j=1}^n \sum_{i=1}^I \frac{\partial u_i(\alpha, \beta)}{\partial \alpha_l} \\
&= \frac{2}{nI} \sum_{j=1}^n \left[\sum_{i=1}^I -\frac{1}{I} X_{l(j)} \left(\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) + \right. \\
&\quad \left. + \left(\alpha_l X_{l(j)} - \beta_l - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) X_{l(j)} \right] \\
&= \frac{2}{nI} \sum_{j=1}^n \left\{ \sum_{i=1}^I -\frac{1}{I} \left(\alpha_i X_{i(j)} X_{l(j)} - \beta_i X_{l(j)} \right) + \alpha_l X_{l(j)}^2 - \beta_l X_{l(j)} \right\} \\
&= \frac{2}{I} \left\{ \sum_{i=1}^I \left[\frac{-\alpha_i}{I} \frac{1}{n} \sum_{j=1}^n X_{i(j)} X_{l(j)} + \frac{\beta_i}{I} \frac{1}{n} \sum_{j=1}^n X_{l(j)} \right] + \alpha_l \frac{1}{n} \sum_{j=1}^n X_{l(j)}^2 - \beta_l \frac{1}{n} \sum_{j=1}^n X_{l(j)} \right\}.
\end{aligned}$$

Entonces, $\frac{\partial U_n(\alpha, \beta)}{\partial \alpha_l} = 0$, si y sólo si,

$$\sum_{i=1}^I \left\{ \frac{-\alpha_i}{I} \frac{1}{n} \sum_{j=1}^n X_{i(j)} X_{l(j)} + \frac{\beta_i}{I} \frac{1}{n} \sum_{j=1}^n X_{l(j)} \right\} + \alpha_l \frac{1}{n} \sum_{j=1}^n X_{l(j)}^2 - \beta_l \frac{1}{n} \sum_{j=1}^n X_{l(j)} = 0$$

si, y sólo si,

$$\begin{aligned}
\frac{-1}{I} \sum_{i=1}^I \alpha_i \frac{1}{n} \sum_{j=1}^n X_{i(j)} X_{l(j)} + \frac{1}{I} \left(\sum_{i=1}^I \beta_i \right) \frac{1}{n} \sum_{j=1}^n X_{l(j)} + \alpha_l \frac{1}{n} \sum_{j=1}^n X_{l(j)}^2 - \beta_l \frac{1}{n} \sum_{j=1}^n X_{l(j)} = 0, \\
l = 1, \dots, I. \quad (3.3)
\end{aligned}$$

Con estos cálculos hemos obtenido I ecuaciones lineales en los I parámetros $(\alpha_l, \beta_l)_{l=1}^I$. Por otro lado,

- Si $l \neq i$, $\frac{\partial u_i(\alpha, \beta)}{\partial \beta_l} = 2 \left(\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) \frac{1}{I}$

- Si $l = i$, $\frac{\partial u_i(\alpha, \beta)}{\partial \beta_l} = 2 \left(\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) \left(-1 + \frac{1}{I} \right)$,

y, por lo tanto, se tiene que

$$\begin{aligned} \frac{\partial U_n(\alpha, \beta)}{\partial \beta_l} &= \frac{1}{nI} \sum_{j=1}^n \sum_{i=1}^I \frac{\partial u_i(\alpha, \beta)}{\partial \beta_l} \\ &= \frac{2}{nI} \sum_{j=1}^n \left[\sum_{i=1}^I \frac{1}{I} \left(\alpha_i X_{i(j)} - \beta_i - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) - \right. \\ &\quad \left. - \left(\alpha_l X_{l(j)} - \beta_l - \frac{1}{I} \sum_{k=1}^I \alpha_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \beta_k \right) \right] \\ &= \frac{2}{I} \left[\sum_{i=1}^I \left(\frac{1}{I} \alpha_i \frac{1}{n} \sum_{j=1}^n X_{i(j)} - \frac{1}{I} \beta_i \right) - \alpha_l \frac{1}{n} \sum_{j=1}^n X_{l(j)} + \beta_l \right]. \end{aligned}$$

Luego, $\frac{\partial U_n(\alpha, \beta)}{\partial \beta_l} = 0$, si y sólo si,

$$\frac{1}{I} \sum_{i=1}^I \alpha_i \frac{1}{n} \sum_{j=1}^n X_{i(j)} - \frac{1}{I} \sum_{j=1}^n \beta_j - \alpha_l \frac{1}{n} \sum_{j=1}^n X_{l(j)} + \beta_l = 0, l = 1, \dots, I. \quad (3.4)$$

Así, hemos obtenido otras I ecuaciones lineales en los parámetros $(\alpha_l, \beta_l)_{l=1}^I$ que, con las (3.3), hacen un total de $2I$ ecuaciones lineales en $2I$ parámetros.

Para facilitar la notación, definimos las constantes:

$$A_{k,l} = \frac{1}{n} \sum_{j=1}^n X_{k(j)} X_{l(j)}, k, l = 1, \dots, I.$$

Es claro que $A_{k,l} = A_{l,k}$, para todo par de índices $j, k = 1, \dots, I$. Ahora las ecuaciones (3.3) y (3.4) se escriben de la siguiente manera:

$$-\frac{1}{I} \sum_{i=1}^I A_{i,l} \alpha_i + \frac{1}{I} \bar{X}_l \sum_{i=1}^I \beta_i + A_{l,l} \alpha_l - \bar{X}_l \beta_l = 0, l = 1, \dots, I \quad (3.5)$$

$$\frac{1}{I} \sum_{i=1}^I \bar{X}_i \alpha_i - \frac{1}{I} \sum_{i=1}^I \beta_i - \bar{X}_l \alpha_l + \beta_l = 0, l = 1, \dots, I. \quad (3.6)$$

Notemos que para que se verifique la condición de identificabilidad ($\tilde{\mathbf{A7}}$), debemos suponer que uno de los parámetros θ_i^* del modelo es conocido, para algún $i \in 1, \dots, I$. Asumiremos, como en el caso anterior, que conocemos el valor del primero $\theta_1^* = (\mu_1^*, \sigma_1^*)$ y, consecuentemente, también el valor de $\check{\theta}^* = (\alpha_1^*, \beta_1^*)$. Esto nos lleva a reescribir las ecuaciones (3.5) y (3.6):

$$-\sum_{i=2}^I A_{i,l} \alpha_i + \bar{X}_l \sum_{i=2}^I \beta_i + A_{l,l} I \alpha_l - \bar{X}_l I \beta_l = A_{1,l} \alpha_1^* - \bar{X}_l \beta_1^*, l = 2, \dots, I \quad (3.7)$$

$$\sum_{i=2}^I \bar{X}_i \alpha_i - \sum_{i=2}^I \beta_i - \bar{X}_l I \alpha_l + I \beta_l = -\bar{X}_l \alpha_1^* + \beta_1^*, l = 2, \dots, I. \quad (3.8)$$

Luego, tenemos en total $2(I-1)$ ecuaciones lineales en los $2(I-1)$ parámetros $(\alpha_l, \beta_l)_{l=2}^I$. Escrito en forma matricial, el sistema de ecuaciones lineales es el siguiente:

$$\left[\begin{array}{c|c} M_1 & M_2 \\ \hline M_2^T & M_3 \end{array} \right] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = D, \quad (3.9)$$

donde:

$$D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}, \text{ con } D_1 = \begin{bmatrix} \alpha_1^* A_{1,2} - \beta_1^* \bar{X}_2 \\ \alpha_1^* A_{1,3} - \beta_1^* \bar{X}_3 \\ \vdots \\ \alpha_1^* A_{1,I} - \beta_1^* \bar{X}_I \end{bmatrix} \text{ y } D_2 = \begin{bmatrix} -\alpha_1^* \bar{X}_1 + \beta_1^* \\ -\alpha_1^* \bar{X}_1 + \beta_1^* \\ \vdots \\ -\alpha_1^* \bar{X}_1 + \beta_1^* \end{bmatrix},$$

$$M_1 = \begin{bmatrix} (I-1)A_{2,2} & -A_{2,3} & \dots & -A_{2,J} \\ -A_{2,3} & (I-1)A_{3,3} & \dots & -A_{3,J} \\ \vdots & \vdots & \ddots & \vdots \\ -A_{2,J} & -A_{3,J} & \dots & (I-1)A_{I,I} \end{bmatrix},$$

$$M_2 = \begin{bmatrix} (1-I)\bar{X}_2 & \bar{X}_2 & \dots & \bar{X}_2 \\ \bar{X}_3 & (1-I)\bar{X}_3 & \dots & \bar{X}_3 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{X}_I & \bar{X}_I & \dots & (1-I)\bar{X}_I \end{bmatrix},$$

$$M_3 = \begin{bmatrix} I-1 & -1 & \dots & -1 \\ -1 & I-1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & I-1 \end{bmatrix}.$$

Para cada $i = 1, \dots, 2(I-1)$, denotemos por F_i a la fila i -ésima de la matriz asociada al sistema. Este sistema adquiere una forma más sencilla tras aplicar la siguiente secuencia de operaciones por filas a su matriz ampliada asociada:

- (1) para cada $i = 1, \dots, I-1$, sustituimos la fila F_i por $F_i + \bar{X}_{i+1}F_{i+I-1}$
- (2) intercambiamos la fila F_I por $F_I^{(1)} = F_I + \sum_{k=I+1}^{2(I-1)} F_k$
- (3) para cada $i = I+1, \dots, 2(I-1)$, reemplazamos la fila F_i por $-\frac{1}{I}F_i + F_I^{(1)}$
- (4) intercambiamos la fila $F_I^{(1)}$ por $(-1)\left(F_I^{(1)} + \sum_{k=I+1}^{2(I-1)} F_k\right)$.

Estas transformaciones nos llevan al siguiente sistema de ecuaciones equivalentes:

$$\left[\begin{array}{cccccccc|c} (I-1)A_{2,2} & -A_{2,3} & \dots & -A_{2,I} & (1-I)\bar{X}_2 & \bar{X}_2 & \dots & \bar{X}_2 & A_{1,2}\alpha_1^* - \bar{X}_2\beta_1^* \\ -A_{2,3} & (I-1)A_{3,3} & \dots & -A_{3,I} & \bar{X}_3 & (1-I)\bar{X}_3 & \dots & \bar{X}_3 & A_{1,3}\alpha_1^* - \bar{X}_3\beta_1^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -A_{2,J} & -A_{3,J} & \dots & (I-1)A_{I,I} & \bar{X}_I & \bar{X}_I & \dots & (1-I)\bar{X}_I & A_{1,I}\alpha_1^* - \bar{X}_I\beta_1^* \\ (1-I)\bar{X}_2 & \bar{X}_3 & \dots & \bar{X}_I & I-1 & -1 & \dots & -1 & -\bar{X}_1\alpha_1^* + \beta_1^* \\ \bar{X}_2 & (1-I)\bar{X}_3 & \dots & \bar{X}_I & -1 & I-1 & \dots & -1 & -\bar{X}_1\alpha_1^* + \beta_1^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \bar{X}_2 & \bar{X}_3 & \dots & (1-I)\bar{X}_I & -1 & -1 & \dots & I-1 & -\bar{X}_1\alpha_1^* + \beta_1^* \end{array} \right]$$

(1)

$$\left[\begin{array}{cccccccc|c} (I-1)(A_{2,2} - \bar{X}_2^2) & -A_{2,3} + \bar{X}_2\bar{X}_3 & \dots & -A_{2,J} + \bar{X}_2\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,2} - \bar{X}_1\bar{X}_2)\alpha_1^* \\ -A_{2,3} + \bar{X}_2\bar{X}_3 & (I-1)(A_{3,3} - \bar{X}_3^2) & \dots & -A_{3,I} + \bar{X}_3\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,3} - \bar{X}_1\bar{X}_3)\alpha_1^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -A_{2,I} + \bar{X}_2\bar{X}_I & -A_{3,I} + \bar{X}_3\bar{X}_I & \dots & (I-1)(A_{I,I} - \bar{X}_I^2) & 0 & 0 & \dots & 0 & (A_{1,I} - \bar{X}_1\bar{X}_I)\alpha_1^* \\ (1-I)\bar{X}_2 & \bar{X}_3 & \dots & \bar{X}_I & I-1 & -1 & \dots & -1 & -\bar{X}_1\alpha_1^* + \beta_1^* \\ \bar{X}_2 & (1-I)\bar{X}_3 & \dots & \bar{X}_I & -1 & I-1 & \dots & -1 & -\bar{X}_1\alpha_1^* + \beta_1^* \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \bar{X}_2 & \bar{X}_3 & \dots & (1-I)\bar{X}_I & -1 & -1 & \dots & I-1 & -\bar{X}_1\alpha_1^* + \beta_1^* \end{array} \right]$$

(2)

$$\left[\begin{array}{cccccccc|cccc} (I-1)(A_{2,2} - \bar{X}_2^2) & -A_{2,3} + \bar{X}_2\bar{X}_3 & \dots & -A_{2,I} + \bar{X}_2\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,2} - \bar{X}_1\bar{X}_2)\alpha_1^* & \\ -A_{2,3} + \bar{X}_2\bar{X}_3 & (I-1)(A_{3,3} - \bar{X}_3^2) & \dots & -A_{3,I} + \bar{X}_3\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,3} - \bar{X}_1\bar{X}_3)\alpha_1^* & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ -A_{2,I} + \bar{X}_2\bar{X}_I & -A_{3,I} + \bar{X}_3\bar{X}_I & \dots & (I-1)(A_{I,I} - \bar{X}_I^2) & 0 & 0 & \dots & 0 & (A_{1,I} - \bar{X}_1\bar{X}_I)\alpha_1^* & \\ (1-I)\bar{X}_2 & \bar{X}_3 & \dots & \bar{X}_I & 1 & 1 & \dots & 1 & (1-I)\bar{X}_1\alpha_1^* + (I-1)\beta_1^* & \\ \bar{X}_2 & (1-I)\bar{X}_3 & \dots & \bar{X}_I & -1 & I-1 & \dots & -1 & -\bar{X}_1\alpha_1^* + \beta_1^* & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ \bar{X}_2 & \bar{X}_3 & \dots & (1-I)\bar{X}_I & -1 & -1 & \dots & I-1 & -\bar{X}_1\alpha_1^* + \beta_1^* & \end{array} \right]$$

↓ (3)

$$\left[\begin{array}{cccccccc|cccc} (I-1)(A_{2,2} - \bar{X}_2^2) & -A_{2,3} + \bar{X}_2\bar{X}_3 & \dots & -A_{2,I} + \bar{X}_2\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,2} - \bar{X}_1\bar{X}_2)\alpha_1^* & \\ -A_{2,3} + \bar{X}_2\bar{X}_3 & (I-1)(A_{3,3} - \bar{X}_3^2) & \dots & -A_{3,I} + \bar{X}_3\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,3} - \bar{X}_1\bar{X}_3)\alpha_1^* & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ -A_{2,I} + \bar{X}_2\bar{X}_I & -A_{3,I} + \bar{X}_3\bar{X}_I & \dots & (I-1)(A_{I,I} - \bar{X}_I^2) & 0 & 0 & \dots & 0 & (A_{1,I} - \bar{X}_1\bar{X}_I)\alpha_1^* & \\ -\bar{X}_2 & -\bar{X}_3 & \dots & -\bar{X}_I & 1 & 1 & \dots & 1 & (1-I)\bar{X}_1\alpha_1^* + (I-1)\beta_1^* & \\ 0 & \bar{X}_3 & \dots & 0 & 0 & -1 & \dots & 0 & \bar{X}_1\alpha_1^* - \beta_1^* & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & \dots & \bar{X}_I & 0 & 0 & \dots & -1 & \bar{X}_1\alpha_1^* - \beta_1^* & \end{array} \right]$$

↓ (4)

$$\left[\begin{array}{cccccccc|cccc} (I-1)(A_{2,2} - \bar{X}_2^2) & -A_{2,3} + \bar{X}_2\bar{X}_3 & \dots & -A_{2,I} + \bar{X}_2\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,2} - \bar{X}_1\bar{X}_2)\alpha_1^* & \\ -A_{2,3} + \bar{X}_2\bar{X}_3 & (I-1)(A_{3,3} - \bar{X}_3^2) & \dots & -A_{3,I} + \bar{X}_3\bar{X}_I & 0 & 0 & \dots & 0 & (A_{1,3} - \bar{X}_1\bar{X}_3)\alpha_1^* & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ -A_{2,I} + \bar{X}_2\bar{X}_I & -A_{3,I} + \bar{X}_3\bar{X}_I & \dots & (I-1)(A_{I,I} - \bar{X}_I^2) & 0 & 0 & \dots & 0 & (A_{1,I} - \bar{X}_1\bar{X}_I)\alpha_1^* & \\ \bar{X}_2 & 0 & \dots & 0 & -1 & 0 & \dots & 0 & \bar{X}_1\alpha_1^* - \beta_1^* & \\ 0 & \bar{X}_3 & \dots & 0 & 0 & -1 & \dots & 0 & \bar{X}_1\alpha_1^* - \beta_1^* & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & \dots & \bar{X}_I & 0 & 0 & \dots & -1 & \bar{X}_1\alpha_1^* - \beta_1^* & \end{array} \right]$$

De esta manera, de las ecuaciones lineales correspondientes a las filas $I + 1, \dots, 2(I-1)$, obtenemos el siguiente conjunto de relaciones entre los parámetros $(\alpha_i)_{i=2}^I$ y $(\beta_i)_{i=2}^I$ del modelo:

$$\beta_i = \bar{X}_i\alpha_i - \bar{X}_1\alpha_1^* + \beta_1^*, i = 2, \dots, I. \quad (3.10)$$

Por otro lado, tenemos que resolver el siguiente sistema en los parámetros $(\alpha_i)_{i=2}^I$.

$$\left[\begin{array}{cccc|c} (I-1)(A_{2,2} - \bar{X}_2^2) & -A_{2,3} + \bar{X}_2\bar{X}_3 & \dots & -A_{2,I} + \bar{X}_2\bar{X}_I & (A_{1,2} - \bar{X}_1\bar{X}_2)\alpha_1^* \\ -A_{2,3} + \bar{X}_2\bar{X}_3 & (I-1)(A_{3,3} - \bar{X}_3^2) & \dots & -A_{3,I} + \bar{X}_3\bar{X}_I & (A_{1,3} - \bar{X}_1\bar{X}_3)\alpha_1^* \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -A_{2,I} + \bar{X}_2\bar{X}_I & -A_{3,I} + \bar{X}_3\bar{X}_I & \dots & (I-1)(A_{I,I} - \bar{X}_I^2) & (A_{1,I} - \bar{X}_1\bar{X}_I)\alpha_1^* \end{array} \right]$$

Si tenemos en cuenta la definición de las constantes $A_{k,l}$, $k, l = 1, \dots, I$, resulta que

$$A_{k,k} - \bar{X}_k^2 = \sum_{j=1}^n X_{k(j)}^2 - \bar{X}_k^2 = \hat{V}ar(X_k), \quad k = 1, \dots, I$$

$$A_{k,l} - \bar{X}_k\bar{X}_l = \sum_{j=1}^n X_{k(j)}X_{l(j)} - \bar{X}_k\bar{X}_l = \hat{C}ov(X_k, X_l), \quad k, l = 1, \dots, I,$$

y, entonces, el sistema de ecuaciones anterior viene dado por:

$$\left[\begin{array}{cccc|c} (I-1)\hat{V}ar(X_2) & -\hat{C}ov(X_2, X_3) & \dots & -\hat{C}ov(X_2, X_I) & \hat{C}ov(X_1, X_2)\alpha_1^* \\ -\hat{C}ov(X_2, X_3) & (I-1)\hat{V}ar(X_3) & \dots & -\hat{C}ov(X_2, X_I) & \hat{C}ov(X_1, X_3)\alpha_1^* \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\hat{C}ov(X_2, X_I) & -\hat{C}ov(X_3, X_I) & \dots & (I-1)\hat{V}ar(X_I) & \hat{C}ov(X_1, X_I)\alpha_1^* \end{array} \right]. \quad (3.11)$$

Observemos que dado un vector no nulo $\mathbf{z} = (z_2, \dots, z_I)^T \in \mathbb{R}^{I-1}$, si llamamos M a la matriz de coeficientes del sistema (3.11),

$$\begin{aligned} \mathbf{z}^T M \mathbf{z} &= \sum_{i=2}^I (I-1)\hat{V}ar(X_i)z_i^2 - 2 \sum_{\substack{i,j=2 \\ i < j}}^I z_i z_j \hat{C}ov(X_i, X_j) \\ &= \sum_{k=1}^n \left[\sum_{i=2}^I (I-1)(X_{i,k} - \bar{X}_i)^2 z_i^2 - 2 \sum_{\substack{i,j=2 \\ i < j}}^I z_i z_j (X_{i,k} - \bar{X}_i)(X_{j,k} - \bar{X}_j) \right] \\ &= \sum_{k=1}^n \left[\sum_{\substack{i,j=2 \\ i < j}}^I (z_i(X_{i,k} - \bar{X}_i) - z_j(X_{j,k} - \bar{X}_j))^2 + \sum_{i=2}^I z_i^2 (X_{i,k} - \bar{X}_i)^2 \right] \geq 0, \end{aligned}$$

y M es semidefinida positiva. Además, la igualdad se alcanza si, y sólo si,

$$\begin{aligned} \left(z_i(X_{i,k} - \bar{X}_i) - z_j(X_{j,k} - \bar{X}_j) \right)^2 &= 0, \forall k = 1, \dots, n, \forall i, j = 2, \dots, I \text{ y} \\ z_i^2(X_{i,k} - \bar{X}_i)^2 &= 0, \forall k = 1, \dots, n, \forall i = 2, \dots, I \end{aligned}$$

o, equivalentemente,

$$z_i(X_{i,k} - \bar{X}_i) = 0, \forall k = 1, \dots, n, \forall i = 2, \dots, I.$$

Como $\mathbf{z} \neq \mathbf{0}$, existe un índice $i_0 \in \{2, \dots, I\}$ tal que $z_{i_0} \neq 0$, y necesariamente

$$(X_{i_0,k} - \bar{X}_{i_0}) = 0, \forall k = 1, \dots, n,$$

lo cual implica que $\sigma_{i_0} = 0$ y contradice la hipótesis de que las varianzas en el modelo de localización y escala son estrictamente positivas.

En conclusión, M es definida positiva, luego invertible y el sistema (3.11) tiene solución única, denotémosla por $\hat{\alpha} = (\hat{\alpha}_2, \dots, \hat{\alpha}_I)$. Entonces, de (3.10), se sigue que

$$\hat{\beta} = (\hat{\beta}_i)_{i=2}^I, \text{ con } \hat{\beta}_i = \bar{X}_i \hat{\alpha}_i - \bar{X}_1 \alpha_1^* + \beta_1^*, i = 2, \dots, I,$$

y $(\hat{\alpha}, \hat{\beta})$ es la solución del problema de optimización

$$\min_{\substack{\alpha_i, \beta_i \\ 2 \leq i \leq I}} U_n(\alpha_1^*, \beta_1^*, (\alpha_i, \beta_i)_{2 \leq i \leq I}).$$

Finalmente, si escribimos $\hat{\theta} = (\hat{\alpha}_i, \hat{\beta}_i)_{i=1}^I$, con $\hat{\alpha}_1 = \alpha_1^*$ y $\hat{\beta}_1 = \beta_1^*$, el valor mínimo del criterio se computa como sigue

$$\begin{aligned} U_n(\hat{\theta}) &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\hat{\alpha}_i X_{i(j)} - \hat{\beta}_i - \frac{1}{I} \sum_{k=1}^I \hat{\alpha}_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I \hat{\beta}_k \right]^2 \\ &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\hat{\alpha}_i X_{i(j)} - \bar{X}_i \hat{\alpha}_i + \bar{X}_1 \hat{\alpha}_1 - \hat{\beta}_1 - \frac{1}{I} \sum_{k=1}^I \hat{\alpha}_k X_{k(j)} + \frac{1}{I} \sum_{k=1}^I (\bar{X}_k \hat{\alpha}_k - \bar{X}_1 \hat{\alpha}_1 + \hat{\beta}_1) \right]^2 \\ &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\hat{\alpha}_i (X_{i(j)} - \bar{X}_i) - \frac{1}{I} \sum_{k=1}^I \hat{\alpha}_k (X_{k(j)} - \bar{X}_k) \right]^2. \end{aligned}$$

Como resumen de los cálculos anteriores, hemos probado el siguiente resultado:

Proposición 3.1. En el modelo de localización y escala (3.1), conocido el parámetro $\theta_1^* = (\alpha_1^*, \beta_1^*)$ de la primera muestra, se tiene que

$$\begin{aligned} & \min_{\substack{\alpha_i, \beta_i \\ 2 \leq i \leq I}} U_n(\alpha_1^*, \beta_1^*, (\alpha_i, \beta_i)_{2 \leq i \leq I}) \\ &= \frac{1}{nI} \sum_{i=1}^I \sum_{j=1}^n \left[\hat{\alpha}_i (X_{i(j)} - \bar{X}_i) - \frac{1}{I} \sum_{k=1}^I \hat{\alpha}_k (X_{k(j)} - \bar{X}_k) \right]^2, \end{aligned}$$

donde $\hat{\alpha}_1 = \alpha_1^*$ y $\hat{\alpha} = M^{-1}b$, con $\hat{\alpha} = (\hat{\alpha}_2, \dots, \hat{\alpha}_I)$, y M y b son la matriz de coeficientes y el término independiente, respectivamente, del sistema de ecuaciones (3.11).

3.2 Bootstrap bajo el modelo de deformación

Tras introducir la idea general en la sección 1.4.2 del capítulo de preliminares para aproximar la distribución de un estadístico mediante bootstrap, veremos una particularización de ésta para obtener la distribución bootstrap de nuestro estadístico de interés $u_n^2 = \inf_{\theta \in \Theta} U_n(\theta)$. Esto nos permitirá decidir sobre la validez del modelo de deformación, que, como hemos visto, equivale a estudiar cómo de grande es el valor observado de dicho estadístico. Finalmente, explicaremos con detalle el procedimiento bootstrap que se ha seguido para conseguirlo.

Supongamos que observamos

$$\begin{array}{ccccccc} X_{1,1} & X_{1,2} & \dots & X_{1,n} & i.i.d. & \sim & \nu_1 \\ X_{2,1} & X_{2,2} & \dots & X_{2,n} & i.i.d. & \sim & \nu_2 \\ \vdots & \vdots & \ddots & \vdots & & & \vdots \\ X_{I,1} & X_{I,2} & \dots & X_{I,n} & i.i.d. & \sim & \nu_I \end{array}$$

Fijado un número B de remuestras bootstrap, repetiremos el siguiente procedimiento, al que nos referiremos en adelante con \star .

1. Para cada $b = 1, \dots, B$,
 - 1.1. Para $i = 1, \dots, I$, generar una réplica bootstrap $X_{i,1}^{*b}, \dots, X_{i,m}^{*b}$, de tamaño fijado $0 < m \leq n$, de la muestra original $X_{i,1}, \dots, X_{i,n}$
 - 1.2. Calcular el valor $(u_m^{*b})^2 = \inf_{\theta \in \Theta} U_m^{*b}(\theta)$. En el modelo de localización y escala, este valor viene dado por la *Proposición 3.1*.
2. Ordenar los valores $(u_m^{*b})^2$, $b = 1, \dots, B$, dando lugar a:

$$(u_m^{*(1)})^2 \leq \dots \leq (u_m^{*(B)})^2,$$

y tomar $\hat{q}_m(1-\alpha) = u_m^{*(B(1-\alpha))}$, el cuantil $1-\alpha$ de la distribución bootstrap del estadístico $\inf_{\theta \in \Theta} U_n(\theta)$, con $\alpha = 0.05$ habitualmente.

3. Por último, comprobar si $nu_n^2 > m \left(u_m^{*(B(1-\alpha))} \right)^2$.

Este proceso se repite un número grande K de veces, y se calcula la frecuencia estimada de rechazo:

$$\hat{p}_n = \frac{\sum_{k=1}^K I \left(nu_{n,k}^2 > m \left(u_{m,k}^{*(B(1-\alpha))} \right)^2 \right)}{K}.$$

Dependiendo del problema en concreto, el paso 1. puede llevarse a cabo de maneras alternativas a la descrita en la sección 1.4.1, que es la más intuitiva de todas. A ella nos referiremos, a partir de ahora, por *bootstrap de muestras*, y es el equivalente a lo que en modelos de regresión se conoce como *bootstrap de pares*. La otra forma que presentamos en este trabajo es lo que se llama en la literatura *bootstrap de residuos*, y se basa en suponer que el modelo de deformación paramétrico (2.3) es cierto.

3.2.1 Bootstrap de residuos

En este caso, el proceso de remuestreo para generar las réplicas bootstrap del paso 1. de ★, consiste en remuestrear la muestra original bajo la hipótesis de que el modelo de deformación es correcto. Con lo cual, suponemos que las señales observadas

$$\begin{array}{ccccccc} X_{1,1} & X_{1,2} & \dots & X_{1,n} & i.i.d. & \sim & \nu_1 \\ X_{2,1} & X_{2,2} & \dots & X_{2,n} & i.i.d. & \sim & \nu_2 \\ \vdots & \vdots & \ddots & \vdots & & & \vdots \\ X_{I,1} & X_{I,2} & \dots & X_{I,n} & i.i.d. & \sim & \nu_I \end{array}$$

satisfacen el modelo (2.3). Este procedimiento, al que llamaremos **Boot.R.** consiste en:

1. Estimar $\theta^* = (\theta_1^*, \dots, \theta_I^*)$ por $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_I)$, tales que minimizan el coste de alineamiento

$$\sum_{i=1}^I W_2^2(\nu_{n,i}(\theta_i), \nu_{n,B}(\theta)).$$

2. Calcular los residuos estimados

$$\hat{\varepsilon}_{i,j} = \varphi_{\hat{\theta}_i}(X_{i,j}), \quad i = 1, \dots, I; j = 1, \dots, n$$

3. Formar la distribución empírica \hat{G}_{nI} asociada a los nI valores $\{\hat{\varepsilon}_{i,j}\}_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$.
4. Fijado el tamaño deseado para las remuestras, digamos m , construir mI réplicas bootstrap de los residuos $\{\varepsilon_{i,j}^*\}_{\substack{1 \leq i \leq I \\ 1 \leq j \leq m}}$, esto es, mI valores simulados a partir de la función de distribución empírica \hat{G}_{nI} .
5. Generar, para cada $i = 1, \dots, I$, la réplica bootstrap $(X_{i,j}^*)_{1 \leq j \leq m}$ de la siguiente manera

$$\begin{array}{ccccccc}
\varepsilon_{1,1}^* & \varepsilon_{1,2}^* & \cdots & \varepsilon_{1,m}^* & \longrightarrow & X_{1,j}^* = \varphi_{\hat{\theta}_1}^{-1}(\varepsilon_{1,j}^*), & j = 1, \dots, m \\
\varepsilon_{2,1}^* & \varepsilon_{2,2}^* & \cdots & \varepsilon_{2,m}^* & \longrightarrow & X_{2,j}^* = \varphi_{\hat{\theta}_2}^{-1}(\varepsilon_{2,j}^*), & j = 1, \dots, m \\
\vdots & \vdots & \ddots & \vdots & & & \\
\varepsilon_{I,1}^* & \varepsilon_{I,2}^* & \cdots & \varepsilon_{I,I}^* & \longrightarrow & X_{I,j}^* = \varphi_{\hat{\theta}_I}^{-1}(\varepsilon_{I,j}^*), & j = 1, \dots, m
\end{array}$$

Ejemplo 3.1 (Bootstrap bajo el modelo de localización y escala).

El procedimiento **Boot.R.** para el modelo (3.1), teniendo en cuenta la reparametrización dada en la sección 3.1, sería:

1. Estimar α_i^* , $i = 2, \dots, I$, por las soluciones del sistema (3.11), sean éstas $\hat{\alpha}_i$, y β_i^* por $\hat{\beta}_i = \bar{X}_i \hat{\alpha}_i - \bar{X}_1 \alpha_1^* + \beta_1^*$, $i = 2, \dots, I$. (ver (3.10)).
2. Hallar $\hat{\sigma}_i = \frac{1}{\hat{\alpha}_i}$ y $\hat{\mu}_i = \frac{\hat{\beta}_i}{\hat{\alpha}_i}$, $i = 1, \dots, I$.
3. Calcular los residuos estimados

$$\hat{\varepsilon}_{i,j} = \frac{X_{i,j} - \hat{\mu}_i}{\hat{\sigma}_i}, \quad i = 1, \dots, I; j = 1, \dots, n,$$

y formar la distribución empírica \hat{G}_{nI} asociada a los nI valores $\{\hat{\varepsilon}_{i,j}\}_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$.

4. Dado m , $0 < m \leq n$, construir mI réplicas bootstrap de los residuos $\{\varepsilon_{i,j}^*\}_{\substack{1 \leq i \leq I \\ 1 \leq j \leq m}}$.
5. Para cada $i = 1, \dots, I$, la réplica bootstrap $(X_{i,j}^*)_{1 \leq j \leq m}$ viene dada por

$$\begin{array}{ccccccc}
\varepsilon_{1,1}^* & \varepsilon_{1,2}^* & \cdots & \varepsilon_{1,m}^* & \longrightarrow & X_{1,j}^* = \hat{\mu}_1 + \hat{\sigma}_1 \varepsilon_{1,j}^*, & j = 1, \dots, m \\
\varepsilon_{2,1}^* & \varepsilon_{2,2}^* & \cdots & \varepsilon_{2,m}^* & \longrightarrow & X_{2,j}^* = \hat{\mu}_2 + \hat{\sigma}_2 \varepsilon_{2,j}^*, & j = 1, \dots, m \\
\vdots & \vdots & \ddots & \vdots & & & \\
\varepsilon_{I,1}^* & \varepsilon_{I,2}^* & \cdots & \varepsilon_{I,I}^* & \longrightarrow & X_{I,j}^* = \hat{\mu}_I + \hat{\sigma}_I \varepsilon_{I,j}^*, & j = 1, \dots, m
\end{array}$$

4 | SIMULACIONES Y RESULTADOS.

En este capítulo, probaremos el método propuesto que utiliza el bootstrap de residuos con diferentes simulaciones que se sitúan en dos escenarios distintos. Para ello, se ha trabajado con el software R, y se han realizado dos tipos de experimentos:

- En primer lugar, se han realizado simulaciones con muestras que verifican el modelo de deformación de localización y escala (3.1), para comprobar la frecuencia de rechazo estimada.
- En segundo lugar, para comprobar la potencia del test, se ha probado el método con muestras que no verifican el modelo de deformación (3.1).

Los códigos implementados en ambos casos se encuentran en el apartado de Anexos.

Las mismas simulaciones se han llevado a cabo cuando el procedimiento de remuestreo consiste en el bootstrap de muestras. Sin embargo, los resultados obtenidos no han sido tan satisfactorios y, por esta razón, no se ha incluido en la memoria. Intuitivamente, parece claro que las aproximaciones mediante el bootstrap de residuos deberán ser mucho más precisas, pues la distribución empírica \hat{G}_{nI} , a partir de la cual generamos las réplicas bootstrap, cuenta con I veces más puntos que cada una de las distribuciones empíricas a partir de las cuales se generan las réplicas con el bootstrap de muestras.

4.1 Simulaciones para estimar la frecuencia de rechazo

Hemos realizado las simulaciones en 4 escenarios diferentes, según el número de señales observadas I ; en concreto, hemos tomado $I = 2, 3, 5, 10$. Para ello, hemos elegido los siguientes pares de parámetros $(\mu_i^*, \sigma_i^*)_{1 \leq i \leq I}$:

- (0, 1), (5, 2), para $I = 2$
- (0, 1), (5, 2), (3, 1), para $I = 3$
- (0, 1), (5, 2), (3, 1), (1,5, 3), (7, 4), para $I = 5$

- $(0, 1), (5, 2), (3, 1), (1.5, 3), (7, 4), (2.5, 0.5), (1, 1.5), (4, 3), (6, 5), (2, 1)$, para $I = 10$.

Tras generar nI observaciones $(\varepsilon_{i,j})_{\substack{1 \leq i \leq I \\ 1 \leq j \leq n}}$ de una distribución $\mathcal{N}(0, 1)$, hemos construido I muestras de variables aleatorias normales

$$X_1 \sim \mathcal{N}(\mu_1^*, \sigma_1^{*2}), \dots, X_I \sim \mathcal{N}(\mu_I^*, \sigma_I^{*2}), \text{ donde}$$

$$X_{i,j} = \mu_i^* + \sigma_i^* \varepsilon_{i,j}, \quad i = 1, \dots, I, \quad j = 1, \dots, n.$$

Suponiendo conocido el primer par $(\mu_1^*, \sigma_1^*) = (0, 1)$, hemos estimado los parámetros $(\hat{\alpha}_i, \hat{\beta}_i)_{i=2}^I$ que minimizan el coste de alineamiento y hemos calculado dicho coste mínimo, tal y como se ha mostrado en la sección 3.1.1 si $I = 2$, y en la sección 3.1.2 si $I = 3, 5, 10$. Seguidamente, hemos aplicado el procedimiento Boot.R del *Ejemplo 3.1* para generar réplicas bootstrap de las muestras construidas, a las que finalmente hemos aplicado \star . Todo ello se ha repetido $K = 1000$ veces y se han obtenido las frecuencias de rechazo \hat{p} que se muestran en la Tabla 4.1.

Lo que queremos constatar es que el nivel de cubrimiento se aproxima al valor nominal $\alpha = 0.05$, como afirma el *Colorario 2.1*. Las simulaciones se han llevado a cabo con tamaño de la muestra original $n = 50, 100, 200, 500, 1000, 2000, 5000$, y tamaño de la muestra bootstrap $m_n = n^e$, con $e = 0.6, 0.7, 0.8, 0.9, 0.95, 1$.

Como se observa en la Tabla 4.1, cuanto mayor es el número de muestras observadas, peor se alcanza el nivel de cubrimiento deseado, siendo $e = 0.6, 0.7$ los valores más desfavorables en todos los casos. Concretamente, los resultados más exitosos se han obtenido en los siguientes casos:

- Para $I = 2$: el nivel es alcanzado con $e = 0.8, 0.9, 0.95$, y 1, independientemente de n . Ver Figura 4.7.
- Para $I = 3$: tomando $e = 0.9, 0.95, 1$, independientemente de n . Ver Figura 4.8.
- Para $I = 5$: tomando $e = 0.9, 0.95, 1$, y $n \geq 500$. Ver Figura 4.9.
- Para $I = 10$: tomando $e = 0.95$ si $n \geq 2000$ y $e = 1$ si $n \geq 500$. Ver Figura 4.10.

Esto es fácilmente apreciable también en las Figuras 4.1, 4.2, 4.3, 4.4, 4.5 y 4.6.

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,144	0,079	0,038	0,046	0,041	0,03
	100	0,148	0,067	0,07	0,05	0,04	0,033
	200	0,129	0,085	0,068	0,043	0,037	0,044
	500	0,138	0,089	0,05	0,048	0,035	0,036
	1000	0,127	0,086	0,063	0,055	0,039	0,032
	2000	0,129	0,104	0,071	0,048	0,043	0,038
	5000	0,039	0,042	0,041	0,049	0,043	0,055
3	50	0,295	0,194	0,115	0,078	0,054	0,034
	100	0,273	0,163	0,089	0,053	0,034	0,039
	200	0,238	0,15	0,077	0,054	0,047	0,031
	500	0,226	0,122	0,07	0,057	0,042	0,029
	1000	0,217	0,107	0,092	0,069	0,042	0,035
	2000	0,221	0,128	0,077	0,053	0,043	0,035
	5000	0,205	0,145	0,082	0,06	0,025	0,047
5	50	0,659	0,428	0,281	0,129	0,111	0,081
	100	0,583	0,337	0,192	0,104	0,083	0,053
	200	0,538	0,281	0,159	0,081	0,078	0,029
	500	0,449	0,267	0,138	0,063	0,056	0,04
	1000	0,415	0,238	0,129	0,064	0,051	0,037
	2000	0,354	0,212	0,115	0,06	0,053	0,032
	5000	0,322	0,203	0,108	0,057	0,061	0,039
10	50	0,996	0,971	0,873	0,702	0,553	0,456
	100	0,994	0,902	0,708	0,433	0,33	0,226
	200	0,958	0,802	0,521	0,247	0,184	0,119
	500	0,914	0,663	0,388	0,149	0,093	0,063
	1000	0,864	0,532	0,286	0,119	0,084	0,046
	2000	0,813	0,473	0,239	0,103	0,063	0,051
	5000	0,756	0,449	0,217	0,088	0,061	0,041

Tabla 4.1: Frecuencia de rechazo simulada bajo el modelo de deformación

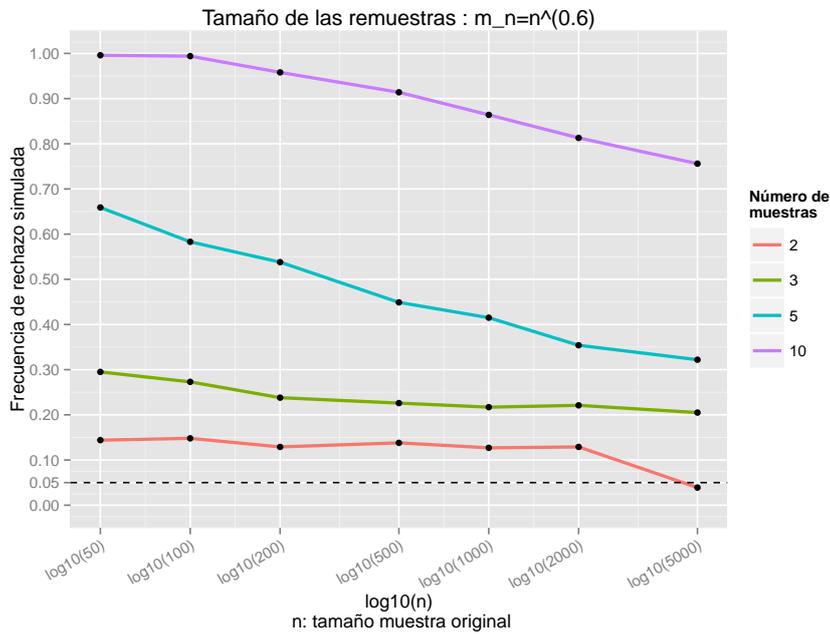


Figura 4.1: Frecuencia de rechazo simulada frente a tamaño de la muestra original y número de muestras, para un tamaño de remuestra fijado $m_n = n^{0,6}$

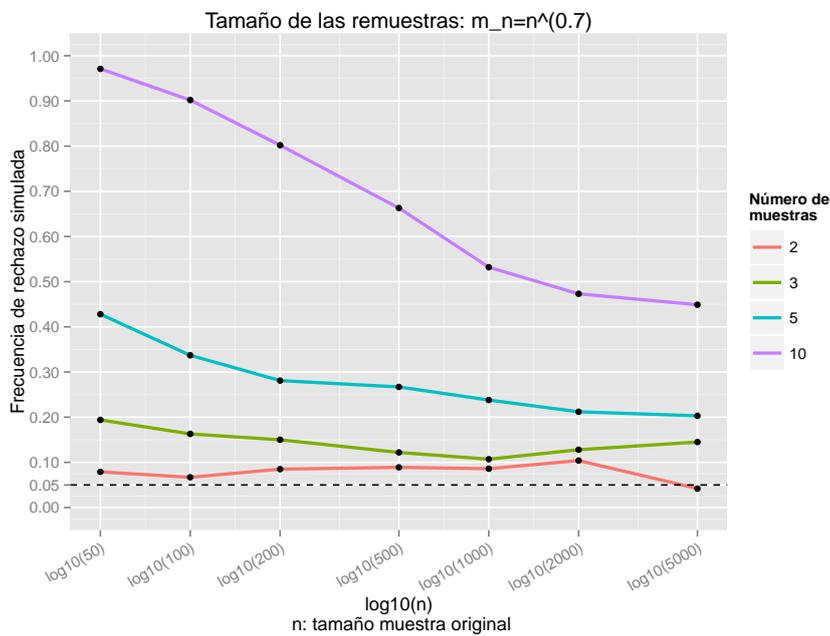


Figura 4.2: Frecuencia de rechazo simulada frente a tamaño de la muestra original y número de muestras, para un tamaño de remuestra fijado $m_n = n^{0,7}$

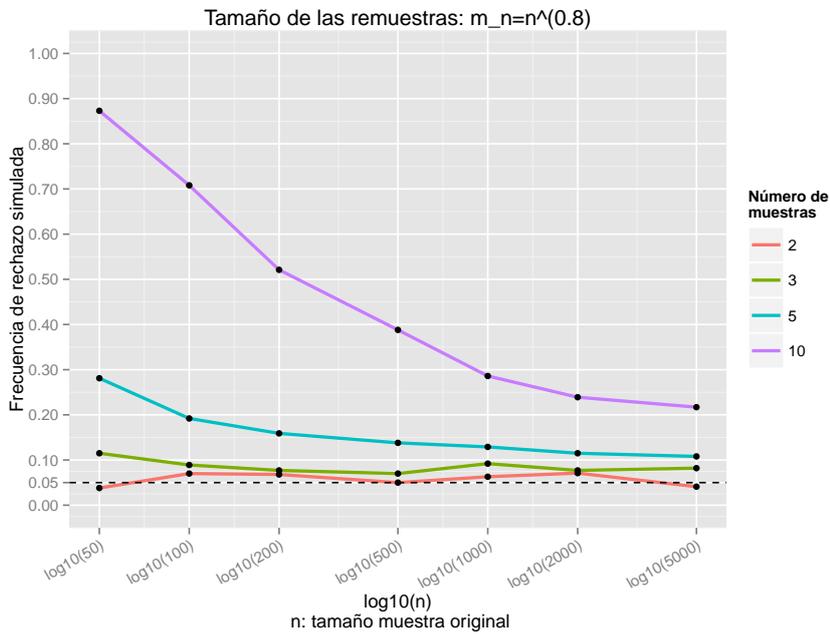


Figura 4.3: Frecuencia de rechazo simulada frente a tamaño de la muestra original y número de muestras, para un tamaño de remuestra fijado $m_n = n^{0.8}$

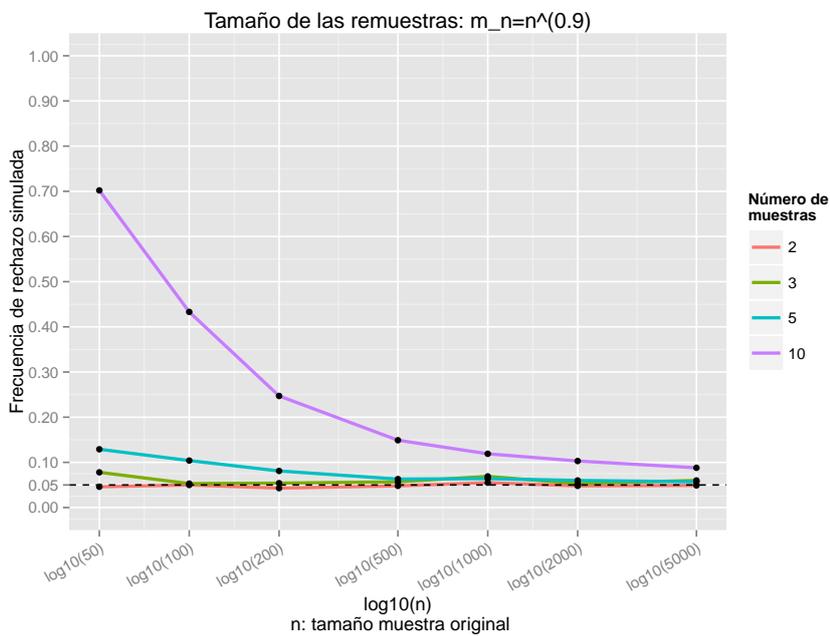


Figura 4.4: Frecuencia de rechazo simulada frente a tamaño de la muestra original y número de muestras, para un tamaño de remuestra fijado $m_n = n^{0.9}$

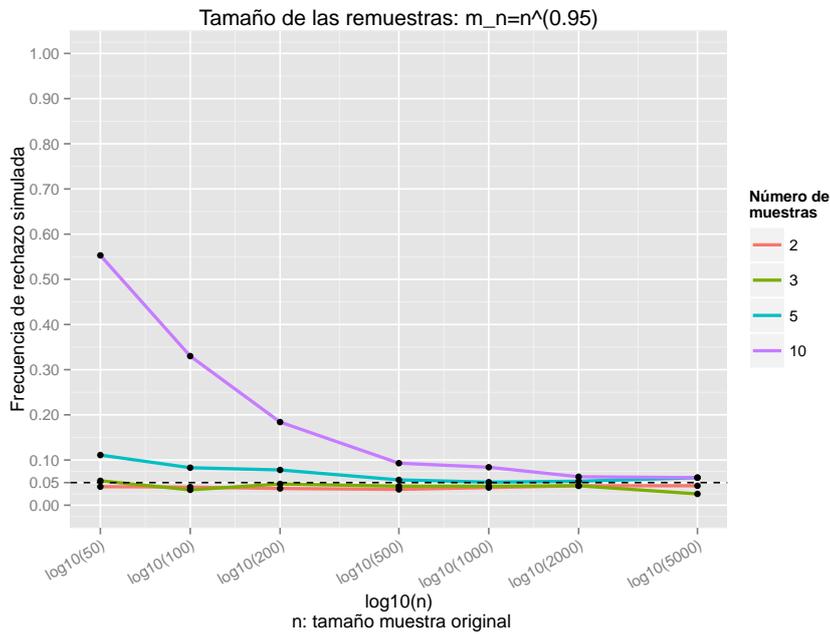


Figura 4.5: Frecuencia de rechazo simulada frente a tamaño de la muestra original y número de muestras, para un tamaño de remuestra fijado $m_n = n^{0.95}$

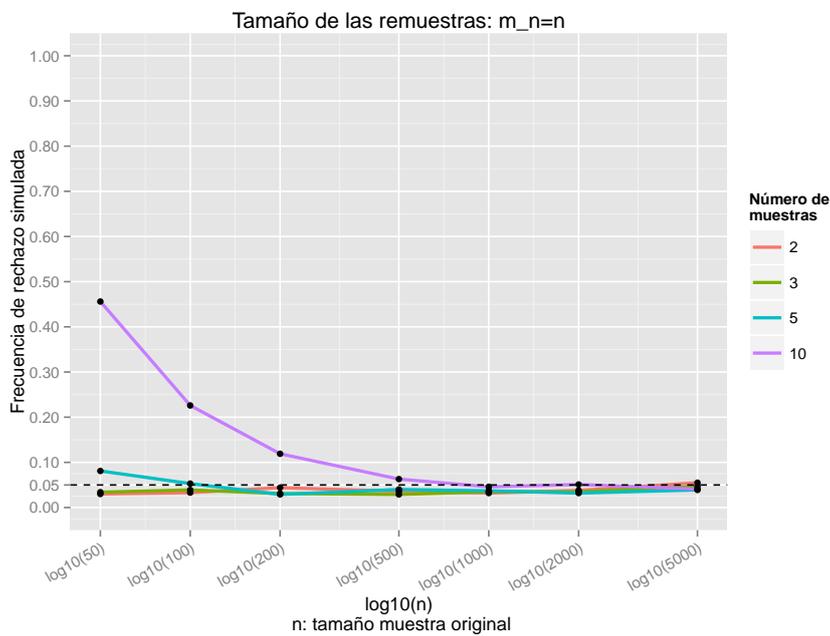


Figura 4.6: Frecuencia de rechazo simulada frente a tamaño de la muestra original y número de muestras, para un tamaño de remuestra fijado $m_n = n$

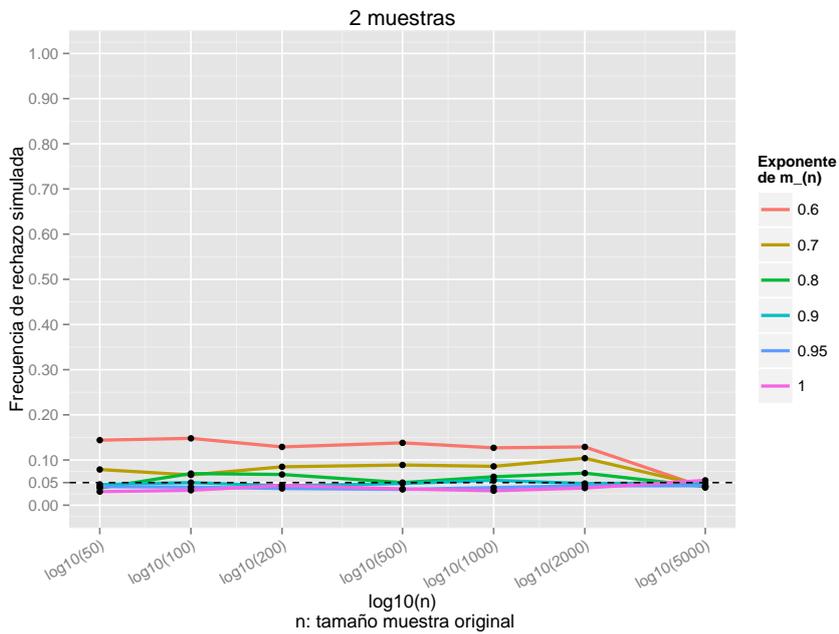


Figura 4.7: Frecuencia de rechazo simulada frente al tamaño de la muestra original y de la remuestra, para el caso de 2 muestras

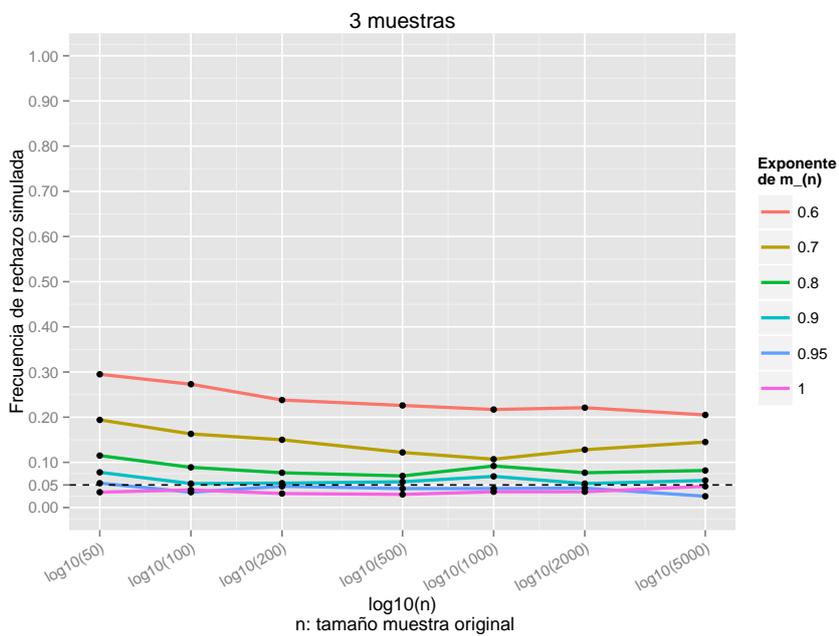


Figura 4.8: Frecuencia de rechazo simulada frente al tamaño de la muestra original y de la remuestra, para el caso de 3 muestras

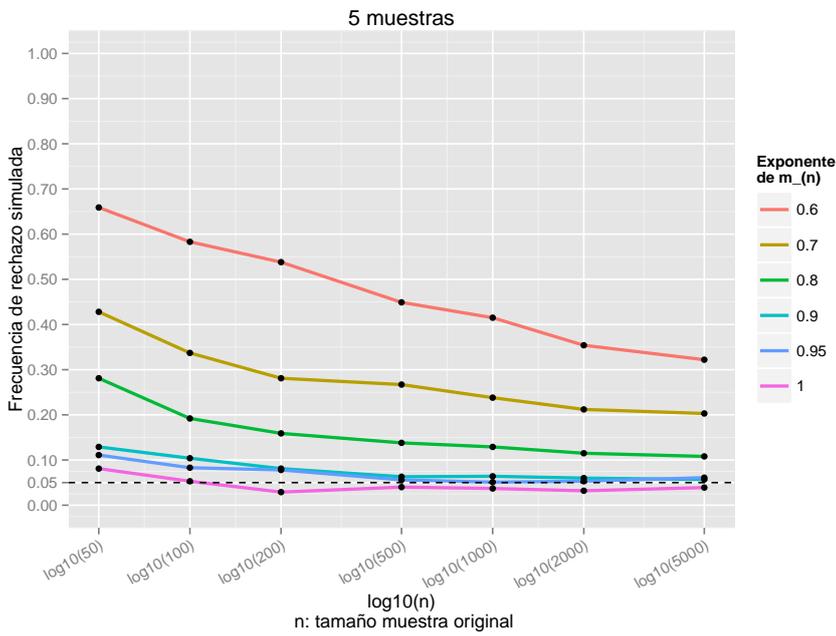


Figura 4.9: Frecuencia de rechazo simulada frente al tamaño de la muestra original y de la remuestra, para el caso de 5 muestras

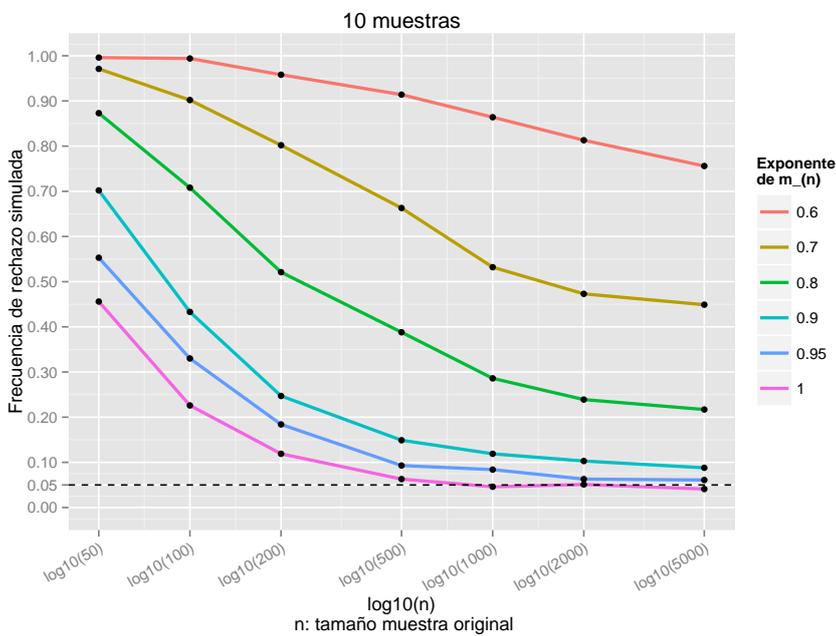


Figura 4.10: Frecuencia de rechazo simulada frente al tamaño de la muestra original y de la remuestra, para el caso de 10 muestras

Teniendo en cuenta la aproximación Binomial-Normal, una variable aleatoria $X \sim \text{Bin}(K, p)$ puede aproximarse por $X \rightsquigarrow \mathcal{N}(Kp, \sqrt{Kp(1-p)})$. Esta aproximación es válida si $Kp > 5$ y $K(1-p) > 5$ y, en ese caso, se tiene que

$$\hat{p} = \frac{X}{K} \rightsquigarrow \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{K}}\right) \text{ asintóticamente.}$$

En consecuencia, se pueden obtener los siguientes intervalos de confianza asintóticos de nivel $1 - \delta$ para las proporciones:

$$\left[\hat{p} \pm z_{\frac{\delta}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{K}} \right].$$

La Tabla 4.2 recoge los intervalos de confianza de nivel $\delta = 0.05$ para la probabilidad de cubrimiento α en el caso de que la muestra original sea de tamaño $n = 5000$, y su contenido se representa en la Figura 4.11.

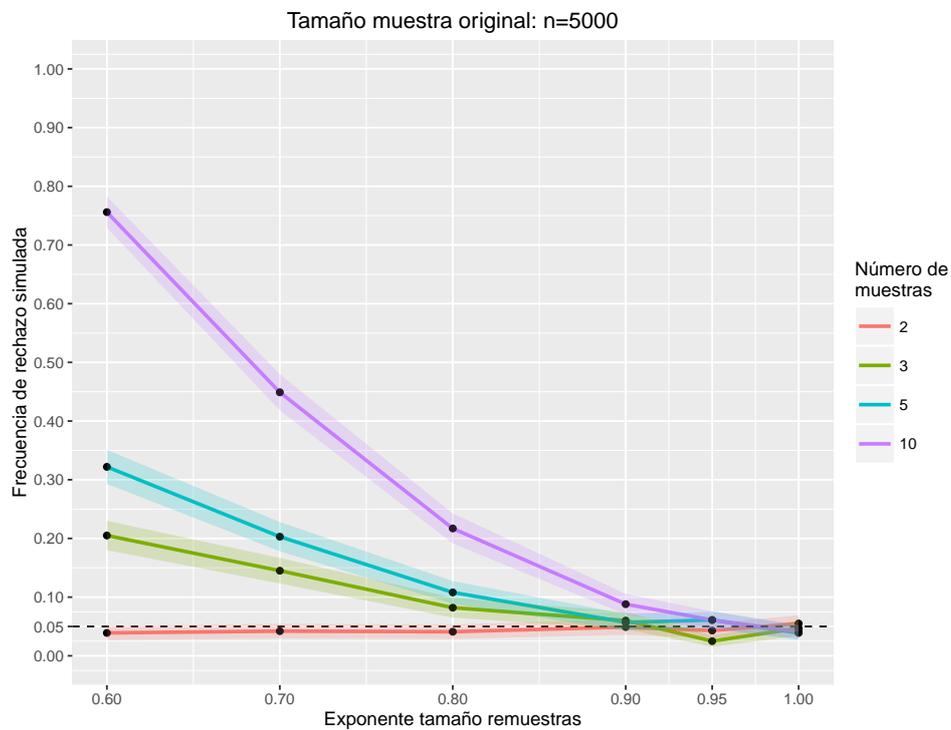


Figura 4.11: Intervalos de confianza para las frecuencias estimadas de rechazo, con $n=5000$

I	e	frecuencia	IC
2	0.6	0.039	0.027 - 0.051
	0.7	0.042	0.030 - 0.054
	0.8	0.041	0.029 - 0.053
	0.9	0.049	0.036 - 0.062
	0.95	0.043	0.030 - 0.056
	1	0.055	0.041 - 0.070
3	0.6	0.205	0.180 - 0.230
	0.7	0.145	0.123 - 0.167
	0.8	0.082	0.065 - 0.099
	0.9	0.06	0.045 - 0.075
	0.95	0.025	0.015 - 0.035
	1	0.047	0.034 - 0.060
5	0.6	0.322	0.293 - 0.351
	0.7	0.203	0.178 - 0.228
	0.8	0.108	0.089 - 0.127
	0.9	0.057	0.043 - 0.071
	0.95	0.061	0.046 - 0.076
	1	0.039	0.027 - 0.051
10	0.6	0.756	0.729 - 0.783
	0.7	0.449	0.418 - 0.480
	0.8	0.217	0.191 - 0.243
	0.9	0.088	0.070 - 0.106
	0.95	0.061	0.046 - 0.076
	1	0.041	0.029 - 0.053

Tabla 4.2: Intervalos de confianza para las frecuencias estimadas de rechazo con $n=5000$

4.2 Simulaciones para comprobar la potencia del test

La potencia de un contraste de hipótesis es la probabilidad de que la hipótesis nula sea rechazada cuando es falsa, es decir, la probabilidad de no cometer un error de tipo II. Para valorar esta propiedad del test, hemos realizado las simulaciones en los 4 escenarios anteriores según el número de señales observadas $I = 2, 3, 5, 10$, tomando todas ellas normales, construidas como se ha explicado en el caso anterior, menos una que ha sido generada a partir de una distribución diferente γ . En concreto, hemos calculado la potencia estimada cuando las señales observadas proceden de las siguientes distribuciones:

- $\mathcal{N}(0, 1)$ y γ , para $I = 2$
- $\mathcal{N}(0, 1)$, $\mathcal{N}(5, 2^2)$ y γ , para $I = 3$
- $\mathcal{N}(0, 1)$, $\mathcal{N}(5, 2^2)$, $\mathcal{N}(3, 1)$, $\mathcal{N}(1.5, 3^2)$ y γ , para $I = 5$
- $\mathcal{N}(0, 1)$, $\mathcal{N}(5, 2^2)$, $\mathcal{N}(3, 1)$, $\mathcal{N}(1.5, 3^2)$, $\mathcal{N}(7, 4^2)$, $\mathcal{N}(2.5, 0.5^2)$, $\mathcal{N}(1, 1.5^2)$, $\mathcal{N}(4, 3^2)$, $\mathcal{N}(6, 5^2)$ y γ , para $I = 10$.

Para la distribución diferente γ se ha elegido una distribución exponencial con parámetro 1, una distribución doble exponencial con parámetro 1 (también llamada distribución de Laplace) y una distribución T de Student con 3 y 4 grados de libertad. Igual que en la situación anterior, las simulaciones se han llevado a cabo con tamaño de la muestra original $n = 50, 100, 200, 500, 1000, 2000, 5000$, y tamaño de la muestra bootstrap $m_n = n^e$, con $e = 0.6, 0.7, 0.8, 0.9, 0.95, 1$. Los resultados obtenidos en cada caso se recogen en las Tablas 4.3, 4.4, 4.5 y 4.6, respectivamente.

En todos los casos, la potencia es alta o muy alta. Analicemos cada uno de ellos por separado. Como puede verse en la Tabla 4.3, en el caso de que $\gamma \stackrel{d}{=} \varepsilon(1)$ la frecuencia de rechazo es muy elevada, prácticamente igual a 1 en todas las simulaciones. Esto no es sorprendente, pues la distribución exponencial es la que más se diferencia de la normal, de entre las cuatro candidatas propuestas para γ , pues ni siquiera es simétrica. Las otras tres distribuciones sí lo son, y el rechazo es similar cuando $\gamma \stackrel{d}{=} \text{Laplace}(0, 1)$ y $\gamma \stackrel{d}{=} t_3$, tal y como se muestra en la Tabla 4.4 y en la Tabla 4.5. Por último, sabemos que a medida que aumentan los grados de libertad de la distribución T, más se parece a una $\mathcal{N}(0, 1)$. En consecuencia, los valores de rechazo contenidos en la Tabla 4.6, que corresponden a $\gamma \stackrel{d}{=} t_4$, son menos elevados.

Hay que tener en cuenta que estos valores obtenidos son especialmente importantes en casos en los que se ha comprobado que se alcanza el nivel. Por ejemplo, cuando tenemos 5 muestras, son informativos los valores de potencia elevados para configuraciones de los parámetros n y $m_n = n^e$ tales que $n \geq 500$ y $e \geq 0.9$.

4.2.1 $\gamma \stackrel{d}{=} \varepsilon(1)$

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,961	0,919	0,897	0,864	0,829	0, 838
	100	1	0,998	0,998	0,995	0,994	0,993
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
3	50	0,987	0,971	0,97	0,953	0,939	0,91
	100	1	1	0,999	1	0,999	0,999
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
5	50	1	0,996	0,988	0,976	0,971	0,955
	100	1	1	1	1	1	1
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
10	50	1	1	1	1	0,996	0,985
	100	1	1	1	1	1	1
	200	1	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1

Tabla 4.3: Potencia simulada cuando las muestras son normales menos una exponencial

4.2.2 $\gamma \stackrel{d}{=} \text{Laplace}(0, 1)$

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,426	0,33	0,3	0,241	0,223	0,163
	100	0,658	0,534	0,468	0,365	0,361	0,3
	200	0,855	0,824	0,751	0,665	0,613	0,602
	500	0,998	0,998	0,993	0,982	0,965	0,962
	1000	1	1	1	1	0,999	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
3	50	0,657	0,533	0,422	0,331	0,282	0,223
	100	0,831	0,708	0,586	0,514	0,461	0,377
	200	0,946	0,915	0,841	0,778	0,709	0,661
	500	1	0,998	0,997	0,994	0,989	0,977
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
5	50	0,895	0,741	0,633	0,471	0,394	0,333
	100	0,936	0,874	0,728	0,623	0,519	0,443
	200	0,994	0,947	0,903	0,847	0,786	0,696
	500	1	1	1	0,996	0,992	0,985
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
10	50	1	0,997	0,97	0,875	0,79	0,703
	100	0,997	0,985	0,949	0,854	0,765	0,643
	200	1	0,996	0,968	0,924	0,859	0,789
	500	1	1	1	0,996	0,996	0,975
	1000	1	1	1	1	1	0,999
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1

Tabla 4.4: Potencia simulada cuando las muestras son normales salvo una doble exponencial

4.2.3 $\gamma \stackrel{d}{=} t_3$

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,566	0,445	0,429	0,352	0,321	0,307
	100	0,775	0,704	0,647	0,576	0,503	0,454
	200	0,942	0,927	0,882	0,833	0,771	0,697
	500	1	0,997	0,995	0,991	0,989	0,957
	1000	1	1	1	1	1	0,986
	2000	1	1	1	1	1	0,999
	5000	1	1	1	1	1	0,997
3	50	0,745	0,653	0,546	0,46	0,402	0,349
	100	0,881	0,821	0,738	0,65	0,592	0,563
	200	0,98	0,958	0,928	0,891	0,873	0,794
	500	1	1	0,999	0,997	0,997	0,978
	1000	1	1	1	1	1	0,995
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
5	50	0,91	0,813	0,682	0,593	0,525	0,45
	100	0,972	0,909	0,822	0,751	0,686	0,621
	200	0,995	0,984	0,967	0,915	0,887	0,836
	500	1	1	1	0,999	0,999	0,995
	1000	1	1	1	1	1	1
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	1
10	50	1	0,997	0,953	0,894	0,827	0,758
	100	0,999	0,993	0,969	0,907	0,862	0,79
	200	1	0,998	0,995	0,961	0,941	0,903
	500	1	1	1	1	0,998	0,988
	1000	1	1	1	1	1	0,998
	2000	1	1	1	1	1	0,999
	5000	1	1	1	1	1	1

Tabla 4.5: Potencia simulada cuando las muestras son normales menos una t_3

4.2.4 $\gamma \stackrel{d}{=} t_4$

I	n	$m_n = n^{0,6}$	$m_n = n^{0,7}$	$m_n = n^{0,8}$	$m_n = n^{0,9}$	$m_n = n^{0,95}$	$m_n = n$
2	50	0,398	0,353	0,292	0,207	0,182	0,183
	100	0,623	0,52	0,429	0,341	0,29	0,228
	200	0,826	0,717	0,65	0,589	0,526	0,41
	500	0,989	0,978	0,954	0,928	0,878	0,787
	1000	1	1	0,999	1	0,984	0,955
	2000	1	1	1	1	1	0,985
	5000	1	1	1	1	1	0,993
3	50	0,634	0,495	0,4	0,295	0,263	0,222
	100	0,756	0,666	0,56	0,465	0,399	0,336
	200	0,914	0,859	0,778	0,663	0,602	0,521
	500	0,998	0,989	0,985	0,972	0,928	0,868
	1000	1	1	1	1	0,999	0,963
	2000	1	1	1	1	1	0,989
	5000	1	1	1	1	1	1
5	50	0,851	0,709	0,583	0,426	0,359	0,316
	100	0,919	0,825	0,668	0,546	0,493	0,316
	200	0,959	0,908	0,842	0,738	0,684	0,578
	500	1	0,997	0,994	0,973	0,934	0,888
	1000	1	1	1	1	0,999	0,968
	2000	1	1	1	1	1	1
	5000	1	1	1	1	1	0,999
10	50	1	0,986	0,941	0,813	0,774	0,653
	100	1	0,988	0,925	0,806	0,738	0,606
	200	1	0,991	0,948	0,854	0,813	0,679
	500	1	1	0,998	0,985	0,954	0,886
	1000	1	1	1	1	0,997	0,949
	2000	1	1	1	1	1	0,974
	5000	1	1	1	1	1	0,995

Tabla 4.6: Potencia simulada cuando las muestras son normales y t_4

CONCLUSIONES.

La gran diversidad en la tipología de la información que se quiere modelizar y analizar actualmente, ha hecho necesario reconsiderar la definición de la noción de comportamiento medio. Además, en muchas situaciones, esta información la constituyen grandes conjuntos de datos que han sido generados por diferentes mecanismos, cada uno de los cuales ha inducido una perturbación diferente en los mismos y, para estudiarlos correctamente se debe recurrir a técnicas estadísticas diferentes a las habituales.

En este trabajo, se ha tratado el caso en el cual los elementos a estudiar son probabilidades y el problema consiste en dar con métodos de alineamiento para deshacer las deformaciones y poder comparar la igualdad de las distribuciones adecuadamente. La técnica que se propone para modelar esta situación viene dada por los modelos de deformación, que suponen que las distribuciones observadas han sido deformadas de una distribución común desconocida. Se ha explicado que, en este caso, el baricentro de Wasserstein es una buena elección para representar el comportamiento promedio y que la variación de Wasserstein es el elemento clave para valorar la bondad en el ajuste del modelo.

Para poder decidir sobre la procedencia común de las diferentes muestras deformadas observadas, se ha analizado el test propuesto en [11] para valorar el ajuste del modelo, el cual lleva implícita la resolución de un problema de optimización. Este problema ha sido resuelto en este trabajo para el caso concreto del modelo de deformación paramétrico de localización y escala, obteniéndose expresiones cerradas para los parámetros de localización y escala óptimos que determinan el valor explícito del estadístico base en dicho test.

Por otro lado, se ha especificado un procedimiento general de simulación tipo Monte-Carlo para aproximar este estadístico. En tal procedimiento, se han tenido en cuenta dos tipos de técnicas de remuestreo: bootstrap de muestras y bootstrap de residuos. Se ha tratado de explicar por qué la primera de ellas no ha resultado satisfactoria, dejando abierta una futura investigación. Finalmente, las diferentes simulaciones implementadas con el bootstrap de residuos han mostrado que el test diseñado alcanza la probabilidad de cubrimiento deseada y que cuenta con una elevada potencia. En cuanto a la probabilidad de cubrimiento, los experimentos realizados han confirmado que, a medida que aumenta

el número de señales registradas, es más difícil que el test se aproxime al nivel nominal esperado, siendo muy satisfactorios los resultados obtenidos para 2,3 y 5 muestras. También se ha comprobado que el test funciona mejor a mayor tamaño de la muestra original, y tomando un tamaño de remuestreo próximo a éste. Finalmente, la potencia es elevada en todas las situaciones simuladas, siendo especialmente interesantes los casos en los que se ha comprobado que se alcanza el nivel.

Bibliografía

- [1] AGUEH, M. AND CARLIER, G. (2011). *Barycenters in the Wasserstein space. SIAM J. Math. Anal.* 43, 904-924.
- [2] AGULLÓ-ANTOLÍN, M., CUESTA-ALBERTOS, J. A., LESCORNEL, H. AND LOUBES, J.-M. (2015). *A parametric registration model for warped distributions with Wasserstein's distance. J. Multivariate Anal.* 135 117-130.
- [3] ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. AND MATRÁN, C. (2008). *Trimmed comparison of distributions. J. Amer. Statist. Assoc.* 103, 697-704.
- [4] AMIT, Y., GRENANDER, U., PICCIONI, M. (1991). *Structural Image Restoration through deformable template. Journal of the American Statistical Association* 86, 376-387.
- [5] BERCU, B. AND FRAYSSE, P. (2012). *A Robbins-Monro procedure for estimation in semiparametric regression models. Ann. Statist.* 40 (2), 666-693.
- [6] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley.
- [7] BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M., SPEED, T. P. (2003). *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. Bioinformatics* 19 (2), 185-193.
- [8] CSÖRGO, M. AND HORVÁTH, L. (1993). *Weighted approximations in probability and statistics.* Wiley.
- [9] CUESTA-ALBERTOS, J. A., MATRÁN-BEA, C., TUERO-DÍAZ, A. (1993). *On Lower Bounds for the L_2 Wasserstein metric in a Hilbert space. Journal of Theoretical Probability* 9(2), 263-283
- [10] DEL BARRIO, E., DEHEUVELS, P. AND VAN DE GEER, S. (2007). *Lectures on Empirical Processes: Theory and Statistical Applications.* European Mathematical Society.

- [11] DEL BARRIO, E., LESCORNEL, H. AND LOUBES, J.-M. (2016). *Central Limit Theorem and bootstrap procedure for Wasserstein's variations with application to structural relationships between ditributions.* <hal-01269785v2>
- [12] DUPUY, J.-F., LOUBES, J.-M., MAZA, E. (2011). *Non parametric estimation of the structural expectation of a stochastic increasing function.* Stat. Comput. 21 (1), 121-136.
- [13] EFRON, B. AND TIBSHIRANI, R. (1994). *An Introduction to the Bootstrap.* Chapman & Hall.
- [14] EVANS, L. C. *Partial Differential Equations.* Graduate Studies in Mathematics.
- [15] GALLON, S., LOUBES, J.-M., MAZA, E. (2013). *Statistical properties of the quantile normalization method for density curve alignment.* Mathematical Biosciences 242 (2), 129-142.
- [16] GAMBOA, F., LOUBES, J.-M., MAZA, E. (2007). *Semi-parametric estimation of shifts.* Electron. J. Stat. 1, 616-640.
- [17] HALL, P. (1992). *The bootstrap and Edgeworth Expansion.* Springer.
- [18] JANSSEN, A. AND PAULS, T. (2003). *How do bootstrap and permutation test work?* The Annals of Statistics. 31(3), 768-806.
- [19] KANTOROVICH, L. AND RUBINSTEIN, G. (1958). *On a space of completely additive functions.* 52-59.
- [20] KENDALL, D. G., BARDEN, D., CARNE, T. K., LE., H. (1999). *Shape and shape theory.* Wiley.
- [21] LESCORNEL, H. AND LOUBES, J.-M. (2014). *Estimation of deformations between distributions by minimal Wasserstein distance.* <hal - 00749519v2>
- [22] MCCANN, R.J. (1997). *A convexity principle for interacting gases.* Advances in Mathematics, 128, 153-179
- [23] RAMSAY, J. O. AND SILVERMAN, B. W. (2005). *Functional data analysis, 2nd Edition.* Springer. Series in Statistics. Springer, New York.
- [24] SAKOE, H. AND CHIBA, S. (1978). *Dynamic Programming Algorithm Optimization for Spoken Word Recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1), 43-49.
- [25] SHAO, J. AND TU, D. (1995). *The Jackknife and the Bootstrap.* Springer.
- [26] TROUVE, A. AND YOUNES, L. (2005). *Metamorphoses Through Lie Group Action.* Foundations of Computational Mathematics 5(2), 173-198.

- [27] VALLANDER, S.S. (1973). *Calculation of the Wasserstein distance between probability distributions on the line*. Theory Probab. Appl. 784-786.
- [28] VILLANI, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

ANEXOS

A | CÓDIGOS PARA IMPLEMENTAR LAS SIMULACIONES.

A.1 calculocriterio2

Esta función implementa los cálculos desarrollados en la sección 3.1.1 para obtener los parámetros óptimos y calcular el valor mínimo del criterio cuando sólo hay dos muestras ($I = 2$) a analizar.

```
calculocriterio2<-function(muestra1, muestra2){
  n<-length(muestra1)
  s1=1
  mu1=0
  alpha1<-1/s1
  beta1<-mu1/s1
  muestra1ord<-sort(muestra1)
  muestra2ord<-sort(muestra2)
  alpha2<-cov(muestra1ord,muestra2ord)/var(muestra2ord)
  betha2<-beta1-mean(muestra1ord)+alpha2*mean(muestra2ord)
  s2<-1/alpha2
  mu2<-s2*beta2
  media<-c(mu1, mu2)
  desv.tipica<-c(s1, s2)
  Un<-((var(muestra1ord)-cov(muestra1ord,muestra2ord)^2/var(muestra2ord))/4

  solucion<-list()
  solucion$Un<-Un
  solucion$media<-media
  solucion$desv.tipica<-desv.tipica

  return(solucion)
}
```

A.2 rechazo.BR2

Esta función implementa los algoritmos Boot.R y ★, para el caso de $I = 2$, y calcula la frecuencia de rechazo.

```

library(00misc)

rechazo.BR2<-function(n,e){

  mn=floor(n^e)
  muestral<-rnorm(n)
  muestra2<-2+5*rnorm(n)

  #muestra2<-rexp(n)
  #muestra2<-rlaplace(n,1)
  #muestra2<-rt(n,3)
  #muestra2<-rt(n,4)
  cc<-calculocriterio2(muestral, muestra2)
  U<-cc$Un
  media.opt=cc$media
  desv.tipica.opt=cc$desv.tipica
  z=c(muestral, (muestra2-media.opt[2])/desv.tipica.opt[2]) #residuos estimados
  bootstrap.U<-function(a){
    muestra1.bootstrap<-sample(z, mn, replace=TRUE)
    muestra2.bootstrap<-media.opt[2] + desv.tipica.opt[2]*sample(z, mn, replace=TRUE)
    U.bootstrap<-calculocriterio2(muestra1.bootstrap, muestra2.bootstrap)$Un
    U.bootstrap
  }
  replicas.B<-sapply(1:10000, bootstrap.U)
  resultado<-1*(n*U>mn*quantile(replicas.B, probs=0.95))
  resultado
}

```

Para simular la potencia del test en este caso, basta con cambiar `muestra2` por una de las cuatro opciones que aparecen inmediatamente debajo, que son las distribuciones que hemos considerado para hallar esta característica del test.

A.3 calculocriteriog

Esta función implementa los cálculos desarrollados en la sección 3.1.2 para obtener los parámetros óptimos y calcular el valor mínimo del criterio en el caso de un número arbitrario I de muestras.

```

calculocriteriog<-function(muestras){
  n<-ncol(muestras)
  I<-nrow(muestras)
  mu1=0
  s1=1
  alpha1<-1/s1
  beta1<-mu1/s1
  muestrasord<-t(apply(muestras, 1, sort))
  sigma<-cov(t(muestrasord))
  B<-rowMeans(muestras)
  A<-(-sigma[2:I, 2:I]+I*diag(diag(sigma[2:I, 2:I])))
  D<-alpha1*sigma[1, 2:I]
  CHOL<-chol(A)
  y<-forwardsolve(t(CHOL), D)
  alpha<-backsolve(CHOL, y)
  beta<- (B[2:I]*alpha - B[1]*alpha1 + beta1)
}

```

```

desv.tipica.opt<-1/alpha
media.opt<-beta/alpha
media<-c(mu1, media.opt)
desv.tipica<-c(s1, desv.tipica.opt)
alpha<-c(alpha1, alpha)
bar2<-colMeans(alpha*(muestrasord-B))
Un<-sum((t(alpha*(muestrasord-B))-bar2)^2)/(n*I)

solucion<-list()
solucion$Un<-Un
solucion$media<-media
solucion$desv.tipica<-desv.tipica
return(solucion)
}

```

A.4 rechazo.BR

Esta función implementa los algoritmos `Boot.R` y ★ para el caso general, y calcula la frecuencia de rechazo cuando las muestras son todas normales.

```

rechazo.BR<-function(n,e){
  mn=floor(n^e)
  media.ini<-c(0,5,3)#,1.5,7,2.5,1,4,6,2)
  desv.tipica.ini<-c(1,2,1)#,3,4,0.5,1.5,3,5,1)
  I=length(media.ini)
  muestras<-media.ini + desv.tipica.ini*matrix(rnorm(I*n), nrow=I, ncol=n)
  cc<-calulocriteriog(muestras)
  U<-cc$Un
  media.opt=cc$media
  desv.tipica.opt=cc$desv.tipica
  bootstrap.U<-function(a){
    z<-(muestras-media.opt)/desv.tipica.opt #residuos estimados
    muestras.bootstrap<-media.opt + desv.tipica.opt*matrix(sample(z,I*mn, replace=TRUE), nrow=I, ncol=mn)
    U.bootstrap<-calulocriteriog(muestras.bootstrap)$Un
    U.bootstrap
  }
  replicas.B<-sapply(1:10000,bootstrap.U)
  resultado<-1*(n*U>mn*quantile(replicas.B,probs=0.95))
  resultado
}

```

A.5 rechazo.BR.potencia

Esta función implementa los algoritmos `Boot.R` y ★, para el caso general, y calcula la frecuencia de rechazo cuando las muestras son todas normales salvo una. Éste es el código empleado para conocer la potencia del test.

```

library(OOmisc)

rechazo.BR.potencia<-function(n,e){
  mn=floor(n^e)

```

```

media.ini<-c(0,5,3,1.5,7)#,2.5,1,4,6)
desv.tipica.ini<-c(1,2,1,3,4)#,0.5,1.5,3,5)
I=length(media.ini)
muestrasnormales<-media.ini + desv.tipica.ini*matrix(rnorm(I*n), nrow=I, ncol=n)
muestradiferente<-rexp(n)
#muestradiferente<-rlaplace(n,1)
#muestradiferente<-rt(n,3)
#muestradiferente<-rt(n,4)

muestras<-rbind(muestrasnormales,muestradiferente)
cc<-calulocriteriog(muestras)
U<-cc$Un
media.opt=cc$media
desv.tipica.opt=cc$desv.tipica
bootstrap.U<-function(a){
  z<-(muestras-media.opt)/desv.tipica.opt
  muestras.bootstrap<-media.opt + desv.tipica.opt*matrix(sample(z,I*mn, replace=TRUE), nrow=I, ncol=mn)
  U.bootstrap<-calulocriteriog(muestras.bootstrap)$Un
  U.bootstrap
}
replicas.B<-sapply(1:10000,bootstrap.U)
resultado<-1*(n*U>mn*quantile(replicas.B,probs=0.95))
resultado
}

```