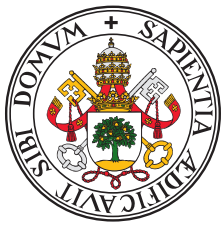

Universidad de Valladolid

Facultad de Ciencias

Departamento de Estadística e Investigación Operativa

Tesis Doctoral

Aplicaciones de los recortes imparciales en la comparación de distribuciones

Presentada por Pedro César Álvarez Esteban para optar al grado de doctor por la Universidad de Valladolid.

Dirigida por D. Eustasio del Barrio Tellado y D. Juan Antonio Cuesta Albertos.

Febrero 2009

Índice general

Agradecimientos

Esta es una de las páginas más agradables de escribir, primero porque significa que he concluido (casi) el trabajo, pero sobretodo porque tengo la oportunidad de dejar constancia de mi gratitud. En primer lugar, quiero expresar mis más profundo agradecimiento a Eustasio del Barrio y a Juan A. Cuesta por su continua dirección, ayuda, paciencia y comprensión durante la elaboración de este trabajo. Desde luego son excelentes investigadores, gracias a su esfuerzo y dedicación he aprendido un poco de Estadística. También quiero agradecer a Carlos Matrán que me ofreció este tema, al que también ha dedicado mucho esfuerzo y atención, por su generosidad. Es también un excelente investigador, y su intuición y capacidad de trabajo, admirables. Con los tres tengo además una deuda más personal por su apoyo y por estar cerca en algunos momentos duros.

No quiero olvidarme de Jesús Sáez, que me ha ayudado con la parte de programación matemática de esta tesis, y de mis compañeros en el departamento de Estadística e Investigación Operativa, por el buen ambiente de trabajo al que contribuyen cada día.

Por compartir muchas cosas con todos ellos me siento muy afortunado.

A mis padres.

A los que me quieren.

Notación.

$\mathcal{X}$	espacio general en el que se definen las medidas de probabilidad
$\mathbb{R}$	recta real
β	σ-álgebra de Borel en $\mathbb{R}$ o de forma general en $\mathcal{X}$
β^k	σ-álgebra de Borel en $\mathbb{R}^k$
ℓ	medida de Lebesgue en $\mathbb{R}$
ℓ^k	medida de Lebesgue en $\mathbb{R}^k$
$\mathcal{P}, \mathcal{P}(\mathcal{X})$	conjunto de medidas de probabilidad definidas en $(\mathcal{X}, \beta)$
$\mathcal{P}(\mathcal{X}, \beta)$	conjunto de medidas de probabilidad definidas en $(\mathcal{X}, \beta)$
$\mathcal{P}_p, \mathcal{P}_p(\mathcal{X})$	conjunto de medidas de probabilidad de $\mathcal{P}$ con momento de orden p finito
$\mathcal{P}_p', \mathcal{P}_p'(\mathcal{X})$	conjunto de medidas de probabilidad de $\mathcal{P}$ con momento de orden $p+\delta$ finito, para algún $\delta > 0$
$\mathcal{R}_\alpha(P)$	conjunto de recortes de nivel a lo sumo α de P
$\mathcal{C}_\alpha$	conjunto de funciones h definidas en $[0,1]$, absolutamente continuas, tales que $h(0)=0$, $h(1)=1$ y $0 \leq h' \leq \frac{1}{1-\alpha}$
$\tau_k(Q_n)$	conjunto de recortes enteros de la medida empírica Q_n
$\mathcal{W}_p$	distancia L_p de Wasserstein
$\mathcal{W}_\infty$	norma del supremo
$P \ll Q$	P es absolutamente continua respecto de Q
$\frac{dQ}{dP}$	derivada de Radon-Nikodym de Q respecto de P
φ	función de densidad de la normal estándar
Φ	función de distribución de la normal estándar
$\delta_{\cdot}$	delta de Dirac

$\lceil\ \rceil$	función entero superior
$\mathcal{L}(X)$	ley de probabilidad de la variable o vector aleatorio X
ν	medida genérica asociada al espacio (Ω, σ) en el que consideraremos definidas variables o vectores aleatorios cuando no se especifique otra cosa
$\rightarrow_{\mathcal{L}}$	convergencia en ley de variables o vectores aleatorios definidos en (Ω, σ)
$\rightarrow_p$	convergencia en probabilidad
$\rightarrow_w$	convergencia débil en $\mathcal{P}(\mathcal{X})$
$\mathrm{Sop}(P)$	soporte de la medida de probabilidad P
P_h	recorte de P siguiendo el patrón dado por $h \in \mathcal{C}_\alpha$
P_Z	medida de probabilidad asociada a la variable o vector aleatorio Z
$\|\cdot\|^p$	norma L_p en $\mathbb{R}^k$
$\|\cdot\|_\infty$	norma del supremo
$\mathcal{N}$	familia de distribuciones normales
$\tau_\alpha(P, Q)$	distancia α-recortada entre P y Q
$\tau_\alpha(P, \mathcal{N})$	distancia α-recortada entre P y la familia de distribuciones normales
$\tilde{\tau}_\alpha(P, \mathcal{N})$	distancia α-recortada estandarizada entre P y la familia de distribuciones normales

Capítulo 1

Introducción.

El objetivo de esta memoria es desarrollar una metodología de recortes imparciales en el ámbito de la comparación de distribuciones y los test de ajuste, y obtener algoritmos y resultados que permitan su aplicación en el análisis de datos e inferencia estadística.

La introducción del concepto de recorte imparcial parece deberse a Grubbs (1950). Sin embargo, no es hasta décadas más tarde cuando esta idea es redescubierta y generalizada en múltiples contextos por diversos autores: Rousseeuw (1985) en regresión y en localización y escala, Gordaliza (1991) en localización y Cuesta Albertos, Gordaliza y Matrán (1997a) en conglomerados. Este último fue el inicio de una serie de trabajos en los que se utilizan exitosamente y estudian distintos aspectos de los procedimientos de recortes imparciales en el análisis de datos multivariantes: Cuesta Albertos et al. (1998, 2002, 2008), García Escudero y Gordaliza (1999, 2005) y García Escudero et al. (1999a,b, 2003). Recientemente Cascos y López-Díaz (2008) han usado la noción de recorte en el estudio de regiones centrales.

La forma tradicional de robustificar una media se basa en eliminar la misma proporción de datos en las dos colas de la distribución. Esta forma de trabajar presupone que la posible contaminación es simétrica y siempre se produce en esa parte de la distribución. El problema se pone aún más de manifiesto cuando manejamos datos multivariantes donde no hay una dirección "a priori" en la que se haya de producir la contaminación. Más aún, en este caso es más clara la posibilidad de encontrarse con los denominados "inliers", puntos de contaminación en el "centro" de la distribución. La idea perseguida con los recortes imparciales es que no haya zonas predeterminadas en las que se recorte sino que sean los propios datos los que indiquen en qué zonas se debe recortar.

La introducción formal de los recortes en los procedimientos de comparación de distribuciones y en el ajuste de una muestra a una distribución prefijada se debe a Munk y Czado (1998) quienes realizaron este estudio en el contexto de ensayos de bioequivalencia con el objetivo de dar una cierta robustez al análisis de similitud allí planteado. Sin embargo, estos autores manejan el procedimiento tradicional de recorte arriba descrito, basado en eliminar sólo datos en las colas de la distribución. Es éste el punto de inicio de esta tesis doctoral, en el que desarrollamos una metodología de recortes dirigidos por los datos (y no por el investigador, de ahí el nombre de imparcial) cuyo objetivo es maximizar la "similitud" entre distribuciones.

Imaginemos que queremos comparar dos muestras de datos univariantes. Observamos que los correspondientes histogramas parecen diferentes, pero nos damos cuenta de que podemos eliminar una cierta fracción de los datos, digamos un 5 % de los datos de una muestra y otro 5 % de los datos de la segunda muestra, de tal forma que los datos que quedan en las muestras producen histogramas muy similares. En ese caso, podríamos decir que el (95 %) "núcleo" de las distribuciones subyacentes es similar. Este podría ser el caso, por ejemplo, cuando intentamos estudiar la similitud de dos poblaciones humanas con respecto a una característica dada. Ambas poblaciones pueden ser inicialmente iguales u homogéneas, pero la presencia de diferentes patrones de inmigración puede causar diferencias en la distribución de esta característica, mientras que por otra parte, los "núcleos" de ambas poblaciones permanecen iguales. Otra situación en la que podríamos estar interesados en comparar los "núcleos" de dos distribuciones sería cuando queremos verificar la igualdad de las distribuciones que generan dos muestras de una magnitud física pero consideramos que los dispositivos que la miden no son perfectos y producen algunas distorsiones cuando los valores verdaderos caen en cierto rango, no afectando al resto. Las distorsiones de los dos dispositivos de medida pueden ser de diferente tipo, pero si no afectan más que a una pequeña proporción de las observaciones, entonces el "núcleo" de la distribución sería el mismo.

Formalicemos un poco esta idea de recorte que da lugar al "núcleo" de una distribución. Cuando recortamos una proporción (de tamaño a lo sumo α con $0 < \alpha < 1$) de los datos de una muestra para tener un mayor parecido a la otra muestra, lo que hacemos es reemplazar la medida empírica $\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ con una nueva medida de probabilidad que da peso 0 a las observaciones que eliminamos y peso $\frac{1}{n-k}$ a cada observación que permanece en la muestra,

donde k es el número de observaciones recortadas y entonces $k \leq n\alpha$ y $\frac{1}{n-k} \leq \frac{1}{n}\frac{1}{1-\alpha}$. En vez de simplemente eliminar/mantener los datos podríamos incrementar el peso de los datos en las zonas "buenas" (por un factor acotado por $\frac{1}{1-\alpha}$) y disminuir la importancia de los datos en las zonas "malas", sin necesidad de eliminarlos completamente. La nueva medida empírica recortada puede escribirse entonces

$$\frac{1}{n}\sum_{i=1}^{n} b_i \delta_{x_i}, \text{ donde } 0 \leq b_i \leq \frac{1}{(1-\alpha)}, \text{ y } \frac{1}{n}\sum_{i=1}^{n} b_i = 1.$$

Si el generador aleatorio de la muestra fuese P, el equivalente poblacional al procedimiento de recorte correspondería a reemplazar la medida de probabilidad $P(B) = \int_B 1 \, dP$ por la nueva medida

$$\tilde{P}(B) = \int_B g \, dP \text{ donde } 0 \leq g \leq \frac{1}{(1-\alpha)} \text{ y } \int g \, dP = 1. \tag{1.1}$$

La ecuación (1.1) permite definir, por continuidad, los recortes totales (los correspondientes a tomar $\alpha = 1$) como aquellas probabilidades que son absolutamente continuas respecto de la que consideremos.

Una consecuencia intrínseca de la aleatoriedad es la variabilidad. Las muestras obtenidas de un experimento aleatorio diferirán y, obviamente, no podemos esperar que dos muestras ideales que provengan del mismo generador aleatorio sean iguales. Uno de los principales retos del estadístico es ser capaz de detectar desviaciones de esta igualdad ideal que no puedan de una forma razonable ser atribuidas a la aleatoriedad.

A menudo el investigador no está realmente interesado o preocupado por la coincidencia exacta, pero sí le gustaría poder garantizar que los dos generadores aleatorios no difieran mucho. El enfoque habitual en la literatura estadística a este "no diferir mucho" consiste en fijar un parámetro relacionado con la distribución de los generadores aleatorios (puede ser que la propia distribución en sí misma) y verificar cuándo determinada distancia entre los parámetros de las dos muestras está por debajo de un límite prefijado. En este caso, diremos que las distribuciones son similares.

Para la comparación de dos muestras o el ajuste de una muestra a una distribución prefijada se hace necesaria la utilización de un criterio de similitud. El criterio de similitud que empleamos en esta memoria, teniendo en cuenta que una muestra puede identificarse con la probabilidad empírica asociada, es el uso de alguna de las métricas probabilísticas con

mejores propiedades. Entre ellas cabe destacar las métricas L_p de Wasserstein que se caracterizan por metrizar la convergencia débil y de momentos, y la métrica L_∞ de Kolmogorov. En particular nos centraremos en la métrica L_2 de Wasserstein (que denotaremos $\mathcal{W}_2$). Es importante resaltar que, como se verá, el concepto de recorte y sus principales propiedades son independientes de la métrica que posteriormente se seleccione para las comparaciones.

La estructura de esta memoria es como sigue. En el Capítulo 2 situamos una serie de definiciones y resultados que no constituyen una aportación original de esta tesis pero de las que se hace uso a lo largo del desarrollo de esta memoria.

En la primera parte del Capítulo 3 se lleva a cabo la formalización del concepto de recorte de una distribución. Asimismo se estudian las principales propiedades de los mismos destacando el resultado que permite parametrizar la clase de recortes de cualquier medida de probabilidad en la recta real en términos de los recortes de la U(0,1). La conexión de nuestro problema con el Problema del Transporte Óptimo (PTO) permite la generalización del anterior resultado a $\mathbb{R}^k$ y a una probabilidad de referencia fija cualquiera, con tal de que sea absolutamente continua respecto de la medida de Lebesgue. Esta caracterización abre las puertas, entre otras cosas, a la obtención de algunos de los resultados asintóticos que se presentan en el Capítulo 5. A continuación se plantean los diversos tipos de problemas que pueden tener interés en la comparación de distribuciones, ya sean de una muestra o de dos. En el estudio del problema cuando se recorta en las dos distribuciones aparecen dos posibilidades, que se recorte libremente en las dos o siguiendo el mismo patrón.

En la siguiente sección se estudian las principales propiedades de los recortes óptimos. Con las propiedades estudiadas para el conjunto de recortes es fácil comprobar que el problema está bien definido. No es, en cambio, tan fácil probar la unicidad de los mismos. En esta sección se encuentran algunos de los resultados más destacables de esta memoria. La unicidad de la solución de un problema de minimización suele ser un requerimiento a la hora de estudiar el comportamiento asintótico, y con frecuencia ésta es difícil de verificar, por lo que se asume como hipótesis. Así ocurre por ejemplo en el estudio de las clásicas $k-$medias -uno de los problemas en los que se han introducido los recortes imparciales- (ver, por ejemplo, Pollard, 1981, 1982; Hartigan, 1978; Stute y Zhu, 1995), donde sólo algunos autores como Fleischer (1964) ó Li y Flury (1995) consideran este problema. En nuestro caso, haciendo uso nuevamente de la conexión con el PTO, probamos la unicidad en el caso de una y dos

muestras, bajo ciertas condiciones generales, y para la métrica L_2 de Wasserstein.

Desde el punto de vista computacional, los diferentes supuestos de recorte introducidos hasta aquí constituyen problemas de optimización de carácter diverso. En la Sección 3.4 se desarrollan algoritmos específicos que aprovechan las características particulares de cada caso y permiten encontrar la solución en un tiempo razonable. De esta manera los procedimientos de ajuste y comparación que se diseñan pueden ser implementados y utilizados en la práctica.

En el Capítulo 4 se muestra el funcionamiento de los recortes imparciales, en sus diferentes variantes, mediante unos cuantos ejemplos en los que se manejan varios modelos poblacionales. Estos ejemplos sirven asimismo para ilustrar la aplicación de esta metodología con fines exclusivamente descriptivos.

El comportamiento asintótico de los recortes y estadísticos introducidos se estudia en el Capítulo 5. En primer lugar se prueba la consistencia en métrica L_2 de Wasserstein de los recortes óptimos. Nuevamente el uso de resultados relacionados con el PTO permite la generalización a $\mathbb{R}^k$. A continuación se estudia la distribución límite de los estadísticos que miden la distancia L_2 de Wasserstein cuando se recorta con el mismo patrón (ya sea una o dos muestras) y se obtiene, haciendo uso de la aproximación fuerte, la normalidad asintótica de los mismos. Utilizando el mismo tipo de técnicas, en la siguiente sección se desarrolla un test de casi-normalidad univariante, generalizable fácilmente a cualquier familia de localización y escala. La utilización de los resultados anteriores para hacer inferencia queda ilustrada con varios ejemplos con datos reales y simulados en la Sección 5.4. Esta sección finaliza con sendas simulaciones que muestran el buen funcionamiento de la distribución asintótica incluso para tamaños muestrales moderados.

La obtención de la distribución límite en el caso de recortar sin restricciones es un problema en principio más difícil y por el momento abierto. Una forma de resolverlo sería conocer la tasa exacta de convergencia del que hemos llamado proceso (cuantil) empírico recortado. En la Sección 5.5 se incluyen algunos resultados en los que se obtiene la tasa exacta de convergencia para el caso uniforme cuando $\alpha = 1$, otro en el que se dan condiciones suficientes para la tasa exacta de la media cuando $\alpha = 1$ y finalmente, otro más general, que nos proporciona una tasa de convergencia en probabilidad y permite acotar la tasa exacta en el caso en el que exista un recorte de nivel inferior o igual a α que haga nula la distancia L_2 de Wasserstein. Este resultado da pie al desarrollo de una metodología bootstrap cuyos

fundamentos teóricos se reflejan en la última sección del Capítulo 5.

Finalmente, la memoria incluye un capítulo en el que se utiliza la metodología bootstrap desarrollada en el capítulo anterior en la comparación de distribuciones y en la búsqueda del núcleo común a n distribuciones. En el primer caso se incluyen tres simulaciones, mientras que en el segundo caso se analiza un ejemplo con datos reales.

En el Apéndice A se incluye el código en distintos lenguajes (R y AMPL, ver R, 2008; Fourer et al., 2003) de los programas utilizados para implementar los algoritmos.

Algunos de los resultados de esta tesis han sido ya presentados en congresos, publicados, o se encuentran sometidos en alguna revista. Las ideas que motivan este trabajo, la formalización del concepto de recorte, algunas propiedades de los mismos que aparecen en la Sección 3.1 y algunos de los algoritmos de la Sección 3.4 fueron presentados en la ponencia titulada "Trimming and Goodness-of-Fit" en Punta del Este (Uruguay) (ver Alvarez-Esteban et al., 2004). El primer trabajo publicado en una revista (ver Alvarez-Esteban et al., 2008a) recoge principalmente, además de la definición y propiedades básicas, los resultados asintóticos obtenidos para el recorte con el mismo patrón de la Sección 5.2, junto con un ejemplo para ilustrar su uso y una simulación (situados en la Sección 5.4). En Alvarez-Esteban et al. (2008b) se han incluido principalmente los resultados de unicidad y consistencia. Y en Alvarez-Esteban et al. (2008c) aparece el test de casi normalidad de la Sección 5.3 junto con un par de ejemplos y una simulación (en la Sección 5.4). Estos últimos resultados fueron presentados en el congreso ERCIM'08, en Neuchâtel (Suiza), (ver Alvarez-Esteban et al., 2008d).

Capítulo 2

Preliminares.

En este Capítulo incluimos las definiciones y resultados que no constituyendo una aportación original son necesarios durante el desarrollo de la memoria. Puesto que nuestro objetivo es facilitar la comprensión de los resultados que se encuentran en la parte de aportaciones, así como su uso en las correspondientes demostraciones, no se hará una exposición exhaustiva de los mismos. No pretendemos, pues, hacer una introducción a los diferentes tópicos que se mencionan.

En primer lugar se da la definición de las métricas que utilizaremos en esta memoria (métricas de Wasserstein y del supremo), así como algunas propiedades de las mismas de uso frecuente en los siguientes capítulos. En la Sección 2.2 se hace una introducción al Problema de Transporte Óptimo y se recogen algunos resultados relacionados de los que se hace uso, principalmente, en el Capítulo 3. A continuación se incluyen algunas definiciones y resultados de programación lineal y entera necesarios para un resultado que aparece en la Sección 3.4. Finalmente, se presentan las curvas de p-valores asintóticos, que usamos en los ejemplos analizados en la Sección 5.4.

2.1. Métricas.

Sea $(\mathcal{X}, \beta)$ un espacio de Banach completo y separable dotado de su σ-álgebra de Borel, donde $\|\cdot\|$ representa su norma. Llamamos $\mathcal{P}(\mathcal{X}, \beta)$ al conjunto de medidas de probabilidad en $\mathcal{X}$. Sea ahora $p \geq 1$ y $\mathcal{P}_p(\mathcal{X}) \subset \mathcal{P}(\mathcal{X}, \beta)$ el conjunto de medidas de probabilidad con momento de orden p finito sobre $\mathcal{X}$. Dada una variable aleatoria Z, denotaremos por P_Z a

la medida de probabilidad asociada. La métrica L_p de Wasserstein (también conocida como métrica de Mallows), que denotamos por $\mathcal{W}_p$, se define como

Definición 2.1. *Sean $P, Q \in \mathcal{P}_p(\mathcal{X})$,*

$$\mathcal{W}_p^p(P,Q) := \inf_{\pi \in \mathcal{M}(P,Q)} \left\{ \int \|x-y\|^p d\pi(x,y) \right\}, \tag{2.1}$$

donde $\mathcal{M}(P,Q)$ es el conjunto de medidas de probabilidad de Borel sobre $\mathcal{X} \times \mathcal{X}$ con marginales P y Q.

Una prueba de que $\mathcal{W}_p$ define una distancia sobre $\mathcal{P}_p(\mathcal{X})$ se puede encontrar en el Lema 8.1 de Bickel y Freedman (1981).

Aunque esta métrica se define asociada a un espacio de Banach separable general, $\mathcal{X}$, y algunas de las propiedades que veremos son válidas en dichos espacios, en la práctica nosotros la usaremos en el espacio euclídeo k-dimensional, $\mathbb{R}^k$, con la norma habitual.

A continuación se da una propiedad de representación que hace que esta distancia entre medidas de probabilidad sea "fácil" de obtener cuando estamos en la recta real. El siguiente resultado recoge el Lema 8.2 de Bickel y Freedman (1981).

Lema 2.2. *Sean $P, Q \in \mathcal{P}_p(\mathbb{R})$, $1 \leq p < \infty$, entonces,*

$$\mathcal{W}_p(P,Q) = \left[\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt \right]^{1/p}, \tag{2.2}$$

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

Inspirados por el anterior resultado, cuando manejamos medidas de probabilidad reales, es frecuente utilizar la notación

$$\mathcal{W}_p(F,G) := \mathcal{W}_p(P,Q),$$

donde F y G son las funciones de distribución de P y Q respectivamente. También, y esta vez en espacios generales, utilizaremos en algún caso (veáse por ejemplo, los Lemas 2.4 y 2.5 más adelante) la notación,

$$\mathcal{W}_p(X,Y) := \mathcal{W}_p(P,Q),$$

donde X e Y son dos variables/vectores aleatorios con leyes de probabilidad P y Q respectivamente.

El siguiente resultado, recogido en el Lema 8.3 de Bickel y Freedman (1981), prueba que la métrica $\mathcal{W}_p$ metriza la convergencia débil más la convergencia de momentos de orden p,

Lema 2.3. *Sean $P_n, P \in \mathcal{P}_p(\mathcal{X})$, $n \geq 1$, entonces son equivalentes:*

(a) $\mathcal{W}_p(P_n, P) \to 0$.

(b) $P_n \to_w P$ *y* $\int \|x\|^p P_n(dx) \to \int \|x\|^p P(dx)$.

(c) $P_n \to_w P$ *y* $\|x\|^p$ *es uniformemente* P_n*-integrable.*

En el Capítulo 5 utilizamos dos propiedades de las métricas de Wasserstein que recogemos en los dos siguientes lemas.

Lema 2.4. *(expresión (8.2) en Bickel y Freedman, 1981). Sean X, Y vectores aleatorios con valores en $\mathcal{X}$ y con momento de orden p finito. Entonces,*

$$\mathcal{W}_p(aX, aY) = |a| \cdot \mathcal{W}_p(X, Y), \qquad \text{para cualquier escalar } a.$$

En los dos resultados que siguen consideraremos la esperanza de vectores aleatorios con valores en $\mathcal{X}$ definida en el sentido de Bochner (ver, p.e., pag. 100 en Araujo y Giné, 1980).

Lema 2.5. *(Lema 8.7 de Bickel y Freedman, 1981). Sea $\mathcal{X}$ un espacio de Hilbert con producto interno $\langle \cdot, \cdot \rangle$ y $p = 2$. Sean X_i, Y_i; $i = 1, \ldots, n$ vectores aleatorios con valores en $\mathcal{X}$ y con momento de orden 2 finito. Supongamos que $\{X_i\}_i$ son independientes, $\{Y_i\}_i$ son independientes y además $E(X_i) = E(Y_i)$ para cada i. Entonces,*

$$\mathcal{W}_2^2\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\right) \leq \sum_{i=1}^n \mathcal{W}_2^2(X_i, Y_i).$$

Otra importante propiedad, también de la distancia $\mathcal{W}_2$, que utilizaremos a la hora de explicar el comportamiento de los recortes en la recta real a través de los ejemplos ofrecidos en el Capítulo 4, es,

Lema 2.6. *(Lema 8.8 de Bickel y Freedman, 1981). Sea $\mathcal{X}$ un espacio de Hilbert con producto interno $\langle \cdot, \cdot \rangle$. Supongamos que X e Y son vectores aleatorios con valores en $\mathcal{X}$ tal que $\|X\|$ y $\|Y\|$ tienen momento de orden 2 finito. Entonces,*

$$\mathcal{W}_2^2(X, Y) = \mathcal{W}_2^2\left(X - E(X), Y - E(Y)\right) + \|E(X) - E(Y)\|^2.$$

Para un estudio más exhaustivo de las métricas de Wasserstein, aparte del mencionado artículo de Bickel y Freedman (1981), se pueden consultar las monografías de Villani (2003, 2009) ó Rachev y Rüschendorf (1998), y también los artículos de Cuesta Albertos y Matrán

(1989) ó Cuesta Albertos et al. (1996). En del Barrio et al. (1999, 2000, 2005, 2007) encontramos aplicaciones de la distancia de Wasserstein, fundamentalmente en el ámbito de los tests de ajuste.

Aunque, como se ha mencionado en la introducción, nos centraremos en las métricas $\mathcal{W}_p$ (y en particular en $\mathcal{W}_2$), en la Sección 3.4 se dan algunos resultados para la métrica del supremo en la recta real, que denotaremos por $\mathcal{W}_\infty$, y definimos como sigue,

Definición 2.7. *Sean $P, Q \in \mathcal{P}(\mathbb{R})$, y F, G sus respectivas funciones de distribución,*

$$\mathcal{W}_\infty(P,Q) := \sup_{x\in\mathbb{R}} |F(x) - G(x)|. \tag{2.3}$$

2.2. El problema del transporte óptimo.

El Problema del Transporte Óptimo (PTO) fue formulado inicialmente por Monge en 1781 como un problema de transporte óptimo de masas. El problema consistía en trasportar arena con el menor costo posible entre dos lugares prefijados. En el planteamiento de Monge, se suponía que la arena a transportar estaba constituida por un conjunto de granos que, a su vez, formaban un montón sobre cierta zona del plano. La cuestión era trasladar estos granos a otra zona del plano, respetando la restricción de que puntos sobre la misma posición inicial deberían ir a la misma posición final. Es decir, si dos granos ocupaban una posición x en el lugar de origen y uno de ellos era transportado a la posición y, el otro también debía ser transportado a esa misma posición. Este problema fue reformulado a mediados del siglo XX por Kantorovich admitiendo la posibilidad de que hubiera división en el transporte de las masas. Es por ello que el PTO es también conocido como Problema de Monge-Kantorovich. En términos probabilísticos dicho problema puede formularse como sigue. Sean P y Q dos medidas de probabilidad en un espacio métrico separable $\mathcal{X}$. Dichas medidas representan respectivamente la distribución de masas original y final en cada posición de $\mathcal{X}$. Sea ahora c una función medible de coste, $c : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+$, tal que $c(x,y)$ representa el coste de transportar una unidad de masa de la posición x a la posición y. El problema consiste en encontrar

$$I(P,Q) = \inf_{\pi\in\mathcal{M}(P,Q)} \left\{ \int c(x,y) d\pi(x,y) \right\}, \tag{2.4}$$

donde, como en la Definición 2.1, $\mathcal{M}(P,Q)$ es el conjunto de medidas de probabilidad de Borel sobre $\mathcal{X} \times \mathcal{X}$ con marginales P y Q.

Comparando esta formulación con la Definición 2.1 de las métricas de Wasserstein, es obvio, que aquellas son las métricas naturales inducidas por el PTO cuando el coste es la norma L_p.

La expresión (2.4) puede ser reescrita en términos de vectores aleatorios, como

$$\inf_{K} \mathrm{E}\left[c(X,Y)\right], \tag{2.5}$$

donde K es el conjunto de pares de vectores aleatorios $\{(X,Y) : \mathcal{L}(X) = P, \mathcal{L}(Y) = Q\}$, y E es la esperanza.

Esto nos permite dar la siguiente definición,

Definición 2.8. *Si (X_0, Y_0) es un par de vectores aleatorios en los que se alcanza el inferior en la expresión (2.5), entonces diremos que (X_0, Y_0) es un emparejamiento óptimo para P y Q.*

Una cuestión interesante y que da lugar a un problema importante es saber bajo qué condiciones un emparejamiento óptimo como el anterior puede ser escrito de la forma $(X_0, T(X_0))$ para alguna función T. Es decir, volviendo al ejemplo de los granos de arena, bajo qué condiciones la solución al PTO formulado como en (2.4) equivale a la formulación inicial dada por Monge en que los granos que compartían la misma posición inicial debían compartir la misma posición final.

Un problema íntimamente relacionado con el anterior aparece cuando nos planteamos si dadas P y Q dos medidas de probabilidad en $\mathcal{X}$, existe una función T que lleve una en otra, es decir, tal que $Q = P \circ T^{-1}$. En ese contexto, McCann (1995) obtiene un resultado que asegura la existencia de funciones que llevan cualquier medida de probabilidad con ciertas condiciones de regularidad en cualquier otra probabilidad si $\mathcal{X} = \mathbb{R}^k$. Este resultado será clave a la hora de obtener la generalización a $\mathbb{R}^k$ del Teorema 3.15.

A continuación se ofrece una versión de dicho resultado. En ese teorema se emplea el siguiente tipo de funciones, cuya definición podemos encontrar, por ejemplo, en la monografía de Rachev y Rüschendorf (1998) (ver pag. 126 del primer volumen),

Definición 2.9. *Una función $T : D \subset \mathbb{R}^k \mapsto \mathbb{R}^k$ se dice que es cíclicamente monótona en D si y sólo si verifica, que dados $(x_i, T(x_i))$, $i = 1, \ldots, n$ del grafo de la función, si $x_{n+1} := x_1$, entonces*

$$\sum_{i=1}^{n} \left(\langle x_i, T(x_i) \rangle - \langle x_{i+1}, T(x_i) \rangle \right) \geq 0,$$

donde $\langle \cdot, \cdot \rangle$ *denota el producto escalar en* $\mathbb{R}^k$.

Teorema 2.10. *(McCann, 1995) Si* P *es cualquier medida de probabilidad en* $\mathbb{R}^k$ *absolutamente continua respecto de la medida de Lebesgue, y* Q *es cualquier medida de probabilidad en* $\mathbb{R}^k$*, entonces existe una (esencialmente) única función cíclicamente monótona* $T : \mathbb{R}^k \mapsto \mathbb{R}^k$ *tal que* $Q = P \circ T^{-1}$.

Las anteriores funciones T proporcionan el transporte óptimo de una probabilidad a otra cuando la función de coste es cuadrática (ver Villani, 2003; Rachev y Rüschendorf, 1998). Por ello, llamaremos función de transporte óptimo entre P y Q a cualquier función T de las dadas en el Teorema 2.10 que lleve P a Q. De hecho, cualquier función cíclicamente monótona proporciona transportes óptimos entre cualquier medida de probabilidad P' y $P' \circ T^{-1}$.

En la siguiente proposición se recogen algunos resultados relacionados con los emparejamientos óptimos y las funciones de transporte óptimo, cuya prueba puede encontrarse en Cuesta Albertos et al. (1997b,c). Estos resultados serán utilizados más adelante para obtener la unicidad de los recortes óptimos cuando se utiliza la métrica $\mathcal{W}_2$.

Sea ℓ^k la medida de Lebesgue en $\mathbb{R}^k$.

Proposición 2.11. *Supongamos que* $P, Q \in \mathcal{P}_2(\mathbb{R}^k)$*, y que* $P \ll \ell^k$*, y sea* (X, Y) *un emparejamiento óptimo para* (P, Q)*. Entonces se verifica que,*

(a) *El cardinal del soporte de una distribución condicional regular de* Y *dado que* $X = x$ *es uno,* P*-c.s.*

(b) *Existe un conjunto* D *con* $P(D) = 1$ *y una función medible y ciclícamente monótona* $T : D \to \mathbb{R}^k$ *tal que* $Y = T(X)$, $\nu-c.s.$

(c) *Si* (X, Y_1) *y* (X, Y_2) *son emparejamientos óptimos para* (P, Q)*, entonces* $Y_1 = Y_2$ $\nu-c.s.$

(d) *Si* T *es una función de transporte óptimo para* (P, Q)*, entonces* T *es continua en casi todos los puntos del soporte de* P.

Utilizando emparejamientos óptimos no es muy díficil ver que la métrica $\mathcal{W}_2$ verifica la siguiente propiedad de convexidad en $\mathbb{R}^k$,

Lema 2.12. *Sean $P, Q \in \mathcal{P}_2(\mathbb{R}^k)$. Entonces,*

$$\mathcal{W}_2^2(\gamma P + (1-\gamma)Q, R) \leq \gamma \mathcal{W}_2^2(P, R) + (1-\gamma)\mathcal{W}_2^2(Q, R),$$

para cada $\gamma \in (0, 1)$.

Finalmente, se dan dos resultados que recogen propiedades bien conocidas de los emparejamientos óptimos y la esperanza condicional de una variable aleatoria. El primer resultado es en realidad una consecuencia inmediata de la Proposición 2.11 y del hecho de que en $\mathbb{R}$ las funciones ciclícamente monótonas son las funciones crecientes. El segundo es una propiedad elemental de la esperanza condicionada. Dichos resultados serán usados para la prueba del Teorema 3.35 en la Sección 3.4.

Lema 2.13. *Sea Z una variable aleatoria real con momento de segundo orden finito. Si $f : \mathbb{R} \mapsto \mathbb{R}$ es una función creciente tal que $f(Z)$ tiene momento de segundo orden finito, entonces*

$$\mathcal{W}_2(P_Z, P_{f(Z)}) = \|Z - f(Z)\|_2.$$

Además, si la distribución de Z es continua y Z^* es otra variable aleatoria real tal que

$$\mathcal{W}_2(P_Z, P_{Z^*}) = \|Z - Z^*\|_2,$$

entonces, existe una función creciente $f : \mathbb{R} \mapsto \mathbb{R}$ tal que $Z^* = f(Z)$ casi seguro.

Lema 2.14. *Sea Z una v.a. real con momento de segundo orden finito. Si A es un conjunto de Borel, entonces*

$$E\left[(Z-a)^2 I_{\{Z \in A\}}\right] \geq E\left[(Z - E[Z/Z \in A])^2 I_{\{Z \in A\}}\right], \quad \text{para cada } a \in \mathbb{R}.$$

2.3. Programación entera.

En la Sección 3.4 se hará uso de algunas definiciones y resultados de programación lineal y entera que podemos encontrar, por ejemplo, en Papadimitriou y Steiglitz (1998) (ver Sección 13.2), y que mencionamos a continuación.

El problema estándar de programación lineal consiste en encontrar valores x_j, $j = 1, \ldots, n$ que maximicen (minimicen) la función objetivo

$$Z = \sum_{j=1}^{n} c_j x_j,$$

sujetos a las restricciones,

$$\sum_{j=1}^{n} a_{ij}x_j \leq b_i, \quad \forall i = 1, \ldots, m$$
$$x_j \geq 0, \quad j = 1, \ldots, n.$$

Dichas restricciones pueden escribirse en forma matricial como $Ax \leq b$, donde A es la matriz de dimensión $m \times n$ de coeficientes a_{ij}, $x = (x_1, x_2, \ldots, x_n)^t$ y $b = (b_1, b_2, \ldots, b_m)^t$.

Si además pedimos que los valores x_j sean enteros, entonces estamos ante un problema de programación entera.

Definición 2.15. *Una matriz cuadrada formada por valores enteros B (en adelante matriz entera), se dice unimodular si $det(B) = \pm 1$. Una matriz rectangular entera A se llama totalmente unimodular si cada submatriz cuadrada no singular de A es unimodular.*

Si llamamos $R_1(A) = \{x : Ax = b, \quad x \geq 0\}$, al politopo formado por el conjunto de soluciones factibles del correspondiente problema de programación lineal, tenemos el siguiente teorema,

Teorema 2.16. *Si A es totalmente unimodular, todos los vértices de $R_1(A)$ son enteros para cualquier vector de enteros b.*

Este resultado es inmediato si tenemos en cuenta que cada uno de estos vértices se obtiene a partir de una expresión como la que sigue,

$$x = \frac{B^{adj}b}{det(B)},$$

donde B^{adj} es la matriz adjunta de B, que es una submatriz cuadrada de A, formada por m columnas independientes. El resultado es entonces claro pues cada elemento es entero y se divide por $+1$ ó -1.

Si ahora llamamos $R_2(A) = \{x : Ax \leq b, \quad x \geq 0\}$, al politopo formado por el conjunto de soluciones factibles del correspondiente problema de programación lineal. Tenemos entonces el siguiente resultado,

Teorema 2.17. *Si A es totalmente unimodular, todos los vértices de $R_2(A)$ son enteros para cualquier vector de valores enteros b.*

La prueba es inmediata a partir del primer teorema sin más que añadir variables auxiliares para formular el problema con una igualdad.

Finalmente mencionamos un teorema que da una condición suficiente para que la matriz A de un politopo sea totalmente unimodular, y que utilizaremos en nuestro problema.

Teorema 2.18. *Una matriz entera A formada por elementos $a_{ij} = 0, \pm 1$ es totalmente unimodular si no aparecen más de dos elementos no nulos en cada columna, y las filas de A pueden ser divididas en dos subconjuntos I_1 e I_2 tal que:*

(1). *Si una columna tiene los dos elementos no nulos del mismo signo, sus filas están en diferentes conjuntos.*

(2). *Si una columna tiene los dos elementos no nulos de diferente signo, sus filas están en el mismo conjunto.*

2.4. Curvas de p-valores.

Presentamos aquí una "herramienta" que utilizamos en la Sección 5.4 como una ayuda en la evaluación de la similitud o disimilitud de las distribuciones subyacentes. Su introducción en este contexto se debe, hasta donde llega nuestro conocimiento, a Munk y Czado (1998).

Supongamos que tenemos dos distribuciones F y G, y una medida d de la similitud entre ellas. Ahora, queremos contrastar la hipótesis nula $H_0 : d(F, G) > \Delta_0$ (ó equivalentemente, $H_0 : d(F, G) < \Delta_0$), donde Δ_0 es un valor umbral fijado por el investigador, de tal forma que dos distribuciones para las que se verifica que $d(F, G) \leq \Delta_0$ se dice que son similares o semejantes. Y disimilares en caso contrario. Supongamos que disponemos de Z_n, un estadístico de contraste para H_0 del que además conocemos la distribución asintótica bajo H_0, y que la región de rechazo es del tipo $\{Z_n \leq z\}$. Entonces, la curva de p-valores asintóticos, $P(\Delta_0)$, se define como,

$$P(\Delta_0) := \sup_{(F,G)\in H_0} \lim_{n\to\infty} P_{F,G}\left(Z_n \leq z_0\right), \tag{2.6}$$

donde z_0 es el valor observado de Z_n', una copia del estadístico Z_n, basada en los n datos de que disponemos.

Esta curva de p-valores asintóticos puede ser usada de dos formas. Por una parte, dado un valor fijo de Δ_0, que controla el nivel de disimilitud, la curva nos permite encontrar el p-valor

asociado a la correspondiente hipótesis nula para decidir si las dos distribuciones son similares o no. Por otra parte, dado un nivel fijo para el contraste (p-valor), podemos encontrar el valor de Δ_0 tal que para cada $\Delta \geq \Delta_0$ deberíamos rechazar la hipótesis $H_0 : d(F, G) > \Delta$. De esta forma, podemos tener una idea más completa del grado de disimilitud entre dos distribuciones.

A la hora de manejar los valores Δ_0 el experimentador debería tener en cuenta como se interpreta el valor de la distancia d. Así, si $d = \mathcal{W}_2$, debemos tener en cuenta que en el caso de que F y G pertenezcan a la misma familia de localización, su distancia de Wasserstein es el valor absoluto de la diferencia de sus medias (ver Lema 2.6).

2.5. Miscelánea. Desigualdad de Bernstein.

Finalmente recogemos una versión de la conocida desigualdad de Bernstein que utilizamos en el Capítulo 5, y cuya prueba podemos encontrar, por ejemplo, en Massart (2007).

Lema 2.19 (Desigualdad de Bernstein). *Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias independientes con media cero y soporte acotado, es decir, $|Y_i| \leq M$ para todo i. Supongamos que $\sum_{i=1}^{n} \operatorname{Var}(Y_i) \leq V$. Entonces, para todo $\eta > 0$,*

$$P\left(|Y_1 + Y_2 + \cdots + Y_n| > \eta\right) \leq 2 \exp\left(\frac{-\frac{1}{2}\eta^2}{V + \frac{1}{3}M\eta}\right).$$

Capítulo 3

Distribuciones Recortadas y Recortes Imparciales.

En lo que sigue formalizaremos el concepto de recorte de una distribución de probabilidades, los recortes imparciales y veremos las diferentes formas de recortar. Estudiaremos las principales propiedades de los recortes óptimos y concluiremos con los algoritmos que permiten calcular en diferentes situaciones los recortes óptimos muestrales.

3.1. Definiciones y propiedades.

Cuando el investigador se aproxima por primera vez al problema de los recortes, es decir, a la posibilidad de que en los datos que estamos analizando haya algunos que sean fruto de la contaminación, sea ésta como sea, de lugar a "outliers" o "inliers", en lo que realmente está interesado es en detectar cuáles son y eliminarlos. Así por ejemplo es como se trabaja cuando se calculan medias recortadas en la recta real. Suponemos *a priori* que los datos erróneos aparecerán en las colas de la distribución y eliminamos una cierta proporción de ellos en cada cola.

Si a partir de estas consideraciones pretendemos generalizar el concepto de distribución recortada al caso poblacional (distribuciones continuas incluidas), podríamos pensar en utilizar indicadores de conjuntos para designar las zonas que eliminamos (completamente) y aquellas con las que nos quedamos. En este punto deberíamos además decidir sobre qué clase de conjuntos tomamos indicadores. Centrémonos en la recta real. Por ejemplo, si P es una

medida de probabilidad en $\mathbb{R}$, podríamos decir que Q es un recorte de P de nivel α si $Q \ll P$ y $\frac{dQ}{dP} = \frac{1}{1-\alpha} I_{\bigcup_{i=1}^{n} A_i}$, donde n es un número prefijado, A_i son intervalos de la recta real y $P(\bigcup_{i=1}^{n} A_i) = 1 - \alpha$. Es decir, eliminamos la masa de probabilidad en un número finito de intervalos (posiblemente no acotados). Trabajar de esta forma equivale a aceptar que van a existir algunos intervalos absolutamente libres de contaminación, lo que en cierta forma está reñido con la idea que pretendemos desarrollar en esta memoria de que sean los datos los que digan dónde se debe recortar, la idea de recorte imparcial.

De esta forma, podríamos pensar en otra alternativa, más general pero manejando todavía indicadores, y así definir los recortes como indicadores en conjuntos de Borel cualesquiera. Entonces, diríamos que Q es un recorte de P de nivel α si $Q \ll P$ y $\frac{dQ}{dP} = \frac{1}{1-\alpha} I_A$, donde $A \in \beta$, $P(A) = 1 - \alpha$ y β es la $\sigma-$álgebra de Borel. El problema que plantea definir los recortes de esta forma es que la clase

$$\mathcal{S}_\alpha(P) = \left\{ Q \in \mathcal{P}(\mathbb{R}, \beta) : Q \ll P,\ \frac{dQ}{dP} = \frac{1}{1-\alpha} I_A,\ P(A) = 1 - \alpha,\ A \in \beta \right\},$$

no es una clase cerrada para la topología de la convergencia débil. No es difícil pensar en una situación que lo demuestre, veámoslo en el siguiente ejemplo.

Ejemplo 3.1. Sea $P = U[0,1]$, y sean

$$\begin{aligned}
A_1 &= \left[0, \frac{1}{4}\right] \cup \left[\frac{1}{2}, 1\right], \\
A_2 &= \left[0, \frac{1}{8}\right] \cup \left[\frac{2}{8}, \frac{3}{8}\right] \cup \left[\frac{1}{2}, 1\right], \\
A_3 &= \left[0, \frac{1}{16}\right] \cup \left[\frac{2}{16}, \frac{3}{16}\right] \cup \left[\frac{4}{16}, \frac{5}{16}\right] \cup \left[\frac{6}{16}, \frac{7}{16}\right] \cup \left[\frac{1}{2}, 1\right], \\
\cdots &\quad \cdots \\
A_n &= \bigcup_{j=1}^{2^{n-1}} \left[\frac{2(j-1)}{2^{n+1}}, \frac{2j-1}{2^{n+1}}\right] \cup \left[\frac{1}{2}, 1\right], \\
\cdots &\quad \cdots
\end{aligned}$$

Es claro que $P(A_n) = \frac{3}{4}$, y que por tanto si Q_n son las medidas de probabilidad tales que

$$\frac{dQ_n}{dP} = \frac{1}{1 - \frac{1}{4}} I_{A_n},$$

entonces $\{Q_n\}_n \subset \mathcal{S}_{1/4}(P)$. Por otra parte, es inmediato que $Q_n \to_w Q = \frac{2}{3} I_{[0,1/2]} + \frac{4}{3} I_{[1/2,1]}$, por lo que $Q \notin \mathcal{S}_{1/4}(P)$. ■

Por ello, y porque la idea de eliminar parcialmente la masa de probabilidad de algunas regiones es muy útil a la hora de comparar distribuciones de probabilidad en términos de similitud, se van a definir las medidas de probabilidad recortadas de forma más general, tal y como aparece en la Definición 3.2.

En el contexto del Análisis de Datos, sigue siendo interesante poder detectar y eliminar los puntos/datos que pueden provenir de una contaminación, que hacen que una muestra no se ajuste a una distribución de referencia, o que no se ajuste a una segunda muestra que también puede estar contaminada. Cuando el recorte consista en eliminar completamente algunos datos hablaremos de recortes enteros.

Si tenemos una muestra aleatoria simple $X_1, X_2, \ldots, X_n$, y una realización de la misma $\{x_1, x_2, \ldots, x_n\}$; la medida empírica asociada es $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. Un recorte entero de dicha medida en el que se eliminan exactamente k puntos (de tamaño $\alpha = \frac{k}{n}$) consistirá en la medida de probabilidad

$$\frac{1}{n-k}\sum_{i=1}^{n} b_i \delta_{x_i}, \text{ con } b_i = 0 \text{ ó } 1 \text{ y } \sum_{i=1}^{n} b_i = n-k.$$

Y el conjunto de todos los recortes enteros de la medida Q_n en los que quitamos k puntos de los n iniciales ($k \leq n$) lo denotaremos por

$$\mathcal{T}_k(Q_n) = \left\{ \frac{1}{n-k}\sum_{i=1}^{n} b_i \delta_{x_i}, \text{ con } b_i = 0 \text{ ó } 1 \text{ y } \sum_{i=1}^{n} b_i = n-k \right\}.$$

Más adelante, en la Sección 3.4 se proporcionan algunos algoritmos para calcular los recortes imparciales enteros para algunas métricas.

Definición 3.2. *Sea $(\mathcal{X}, \beta)$ un espacio medible, P una medida de probabilidad en $(\mathcal{X}, \beta)$ y α tal que $0 \leq \alpha \leq 1$. Decimos que una medida de probabilidad Q en $(\mathcal{X}, \beta)$ es un recorte de nivel a lo sumo α de P si $Q \ll P$ y $\frac{dQ}{dP} \leq \frac{1}{1-\alpha}$ P-c.s.*

Denotaremos por $\mathcal{R}_\alpha(P)$ al conjunto de recortes de nivel a lo sumo α de P y por $\mathcal{P}(\mathcal{X}, \beta)$ el conjunto de medidas de probabilidad en $(\mathcal{X}, \beta)$. En los casos de valores extremos de α tendremos que $\mathcal{R}_1(P)$ será el conjunto de las medidas de probabilidad absolutamente continuas con respecto a P, mientras que $\mathcal{R}_0(P)$ estará integrado únicamente por P.

Esta definición incluye como caso particular a los recortes cuya densidad respecto a la probabilidad que se recorta es el indicador en un conjunto, es decir, $\mathcal{S}_\alpha(P) \subset \mathcal{R}_\alpha(P)$. Además, en el caso de distribuciones empíricas, también se tiene que $\mathcal{T}_k(Q_n) \subset \mathcal{R}_{k/n}(Q_n)$.

Existen otras definiciones equivalentes que manejaremos indistintamente a la hora de hablar de recortes y que mencionamos aquí.

Proposición 3.3. *Dadas P y Q en $\mathcal{P}(\mathcal{X},\beta)$, y $0 \leq \alpha < 1$, las siguientes afirmaciones son equivalentes:*

(a) *Q es un recorte de nivel a lo sumo α de P.*

(b) $\forall A \in \beta:\ (1-\alpha)Q(A) \leq P(A)$.

(c) *Existe $\gamma : \mathcal{X} \mapsto \mathbb{R}$ tal que $0 \leq \gamma \leq 1$ P-c.s. y $Q(A) = \frac{1}{1-\alpha}\int_A \gamma(x)P(dx)$.*

Demostración. Que (a) implica (b) es inmediato si tenemos en cuenta que dado $A \in \beta$:

$$Q(A) = \int_A \frac{dQ}{dP}dP \leq \int_A \frac{1}{1-\alpha}dP = \frac{1}{1-\alpha}P(A).$$

De (b) se deduce que $(1-\alpha)Q \ll P$. Existe pues $\gamma = \frac{d(1-\alpha)Q}{dP} \geq 0$ P-c.s., y tal que (despejando):

$$Q(A) = \frac{1}{1-\alpha}\int_A \frac{d(1-\alpha)Q}{dP}dP = \frac{1}{1-\alpha}\int_A \gamma dP \qquad \forall A \in \beta.$$

Como además

$$0 \leq \int_A \gamma dP = (1-\alpha)Q(A) \leq P(A) = \int_A 1dP \qquad \forall A \in \beta.$$

entonces $0 \leq \gamma \leq 1$ P-c.s., y tenemos que (b) implica (c). La demostración de que (c) implica (a) es inmediata por lo visto hasta aquí. ■

A continuación veamos un resultado que completa las caracterizaciones de los recortes dadas en la Proposición 3.3,

Proposición 3.4. *Sean P y Q dos medidas de probabilidad en un espacio métrico separable y completo $(\mathcal{X},\beta,d)$, entonces,*

$$Q \in \mathcal{R}_\alpha(P) \Leftrightarrow \forall h \geq 0 \text{ continua y acotada } \int hdQ \leq \frac{1}{1-\alpha}\int hdP.$$

Demostración. Supongamos que $Q \in \mathcal{R}_\alpha(P)$, y sea $h \geq 0$ una función continua y acotada general. Por la definición de recorte de una probabilidad y sus caracterizaciones dadas en la Proposición 3.3 se tiene trivialmente que

$$\int hdQ = \int h\frac{dQ}{dP}dP \leq \frac{1}{1-\alpha}\int hdP.$$

Para comprobar el recíproco, veamos en primer lugar que $Q(A) \leq \frac{1}{1-\alpha}P(A)$ si A es cerrado. En ese caso podemos aproximar $I_A(x)$ por la sucesión de funciones

$$h_n(x) = 1 - ((nd(x, A)) \wedge 1).$$

Estas funciones son obviamente positivas y acotadas. También son continuas por ser suma de una constante y el mínimo de funciones continuas. Y además se tiene que $h_n(x) \to I_A(x)$ cuando $n \to \infty$, por lo que,

$$Q(A) = \lim_{n\to\infty} \int h_n dQ \leq \lim_{n\to\infty} \frac{1}{1-\alpha} \int h_n dP = \frac{1}{1-\alpha} P(A).$$

De la propiedad de regularidad de las medidas de probabilidad en un espacio métrico separable y completo (ver Teorema 1.1.2 en Araujo y Giné, 1980) se sigue que $Q(A) \leq \frac{1}{1-\alpha}P(A), \quad \forall A \in \beta$ y por tanto $Q \in \mathcal{R}_\alpha(P)$. ■

Llamaremos *funciones de recorte de nivel* α a las obtenidas en el punto (c) de la Proposición 3.3. Es interesante analizar su significado. Son unas funciones que nos dan para cada punto del espacio $\mathcal{X}$ la fracción de masa de probabilidad que no recortamos, con la que nos quedamos. Si la función tomase únicamente los valores 0 ó 1, entonces sería el indicador de un conjunto A con $P(A) \geq 1 - \alpha$, y recortar equivaldría a eliminar toda la masa de probabilidad de A^c, o lo que es lo mismo, el recorte daría lugar a la probabilidad condicionada $P(\cdot|A)$.

Como habíamos mencionado en la introducción el concepto de recorte es aplicable en espacios generales, sin embargo, en gran parte de esta memoria acabaremos centrándonos en los recortes en la recta real. En ese sentido es interesante el siguiente resultado puesto que simplificará la comprobación de que una medida de probabilidad es un recorte de otra, al reducir la comprobación de la condición exigida en el apartado (b) de la Proposición 3.3 a la semiálgebra de los intervalos cuando trabajamos en la recta real.

Proposición 3.5. *Dadas P y Q en $\mathcal{P}(\mathcal{X}, \beta)$, es condición necesaria y suficiente para que Q sea un recorte de P de nivel a lo sumo α, que se verifique $(1-\alpha)Q(A) \leq P(A)$ $\forall A \in \mathcal{S}$, siendo $\mathcal{S}$ una semi-álgebra de $\mathcal{X}$ que genera β.*

Demostración. Es claro que sólo es preciso probar la condición suficiente pues $\mathcal{S} \subset \beta$. Si se verifica la desigualdad del enunciado para todo conjunto de una semi-álgebra $\mathcal{S}$, es

claro, por la aditividad de las medidas de probabilidad P y Q, que se verificará para todos los conjuntos del álgebra generada por la misma:

$$\mathcal{A} = \mathcal{A}(\mathcal{S}) = \left\{ \bigcup_{i=1}^{n} A_i : A_i \cap A_j = \varnothing \quad \forall i \neq j \right\}.$$

Por otra parte, si se considera la clase $\mathcal{M} = \{A \in \beta : \ (1-\alpha)Q(A) \leq P(A)\}$, es fácil ver que es una clase monótona. Dada una sucesión de conjuntos $\{A_n\}_n \subset \mathcal{M}$, si $A_n \uparrow A \Rightarrow A \in \mathcal{M}$. Esto es inmediato ya que $(1-\alpha)Q(A_n) \leq P(A_n)\ \forall A_n$, y además, $Q(A_n) \uparrow Q(A)$ y $P(A_n) \uparrow P(A)$, luego $(1-\alpha)Q(A) \leq P(A)$. De igual forma se comprueba la monotonía hacia abajo.

Así pues tenemos $\mathcal{A} \subset \mathcal{M}$ y $\mathcal{M}$ una clase monótona. Aplicando el Teorema de la Clase Monótona o Teorema de Halmos (ver, por ejemplo, pag. 43 de Billingsley, 1995), concluimos que $\sigma(\mathcal{A}) = \beta = \mathcal{M}$.

■

Veamos ahora una proposición que recoge algunas propiedades del conjunto de recortes $\mathcal{R}_\alpha(P)$ de una probabilidad P.

Proposición 3.6. *Para cualquier medida de probabilidad, P,*

(a) $\mathcal{R}_{\alpha_1}(P) \subset \mathcal{R}_{\alpha_2}(P)$ *si* $\alpha_1 \leq \alpha_2$.

(b) $\mathcal{R}_\alpha(P)$ *es un conjunto convexo.*

(c) *Si* $\alpha < 1$ *y* $(\mathcal{X}, \beta)$ *es un espacio métrico separable, donde* β *es la* σ*-álgebra de Borel, entonces* $\mathcal{R}_\alpha(P)$ *es cerrado para la topología de la convergencia débil en* $\mathcal{P}(\mathcal{X}, \beta)$.

(d) *Si* $\mathcal{X}$ *es además completo, entonces* $\mathcal{R}_\alpha(P)$ *es compacto.*

Demostración. (a) es trivial a partir de la definición de recorte o cualquiera de las caracterizaciones dadas en la Proposición 3.3.

(b) también es inmediato pues si $\lambda \in [0,1]$ y $Q_1,\ Q_2 \in \mathcal{R}_\alpha(P)$,

$$(1-\alpha)\left[(1-\lambda)Q_1(A) + \lambda Q_2(A)\right] \leq (1-\lambda)P(A) + \lambda P(A) = P(A), \quad A \in \beta.$$

(c) Para probar este apartado es suficiente ver que si $\{Q_n\}_n$ es una sucesión tal que $Q_n \in \mathcal{R}_\alpha(P)$ y $Q_n \to_w Q$ entonces $Q \in \mathcal{R}_\alpha(P)$. Para ello supongamos que $\{Q_n\}_n$ es tal sucesión. Entonces, por el Teorema Portmanteau, se tiene que $Q(A) \leq \liminf_{n\to\infty} Q_n(A) \leq \frac{1}{1-\alpha}P(A)$

para cada conjunto abierto A de β. El resultado se sigue del apartado (b) de la Proposición 3.3 y la regularidad de las medidas de probabilidad en espacios métricos.

(d) Si además $\mathcal{X}$ es completo entonces P es *tight*. De la definición de $\mathcal{R}_\alpha(P)$ es inmediato que es uniformemente tight y, por el Teorema de Prokhorov (ver, por ejemplo, el Teorema 1.2.10 en Araujo y Giné, 1980), es relativamente compacto y por tanto compacto. ■

El siguiente resultado establece una propiedad de los recortes imparciales que podríamos denominar propiedad de continuidad, y es que el límite débil de una sucesión de recortes de una sucesión de medidas de probabilidad convergentes es a su vez un recorte del límite de esta última sucesión. Es decir,

Teorema 3.7. *Sean P, y $\{P_n\}_n$ medidas de probabilidad en un espacio métrico separable y completo. Si $\alpha < 1$, $\{P_n\}_n$ es una sucesión tight y $P_{n,\alpha} \in \mathcal{R}_\alpha(P_n)$ para cada n, entonces $\{P_{n,\alpha}\}_n$ es tight. Además, si $P_n \to_w P$ y $P_{n,\alpha} \to_w P_0$, entonces $P_0 \in \mathcal{R}_\alpha(P)$.*

Demostración. La primera parte es trivial a partir del apartado (b) de la Proposición 3.3. Mientras que la segunda parte se obtiene también fácilmente a partir del Teorema Portmanteau para la convergencia débil y la Proposición 3.4. Como $P_{n,\alpha} \in \mathcal{R}_\alpha(P_n)$ tenemos que $\forall h \geq 0$ continua y acotada,

$$\int h dP_{n,\alpha} \leq \frac{1}{1-\alpha} \int h dP_n \text{ para cada } n. \tag{3.1}$$

Por otra parte, como $P_n \to_w P$ y $P_{n,\alpha} \to_w P_0$, tendremos que para toda función h continua y acotada, $\lim_{n\to\infty} \int h dP_n = \int h dP$ y $\lim_{n\to\infty} \int h dP_{n,\alpha} = \int h dP_0$.

Sea ahora $h \geq 0$, continua y acotada, tomando límites en ambos miembros de la desigualdad (3.1) tendremos que,

$$\int h dP_0 = \lim_{n\to\infty} \int h dP_{n,\alpha} \leq \lim_{n\to\infty} \frac{1}{1-\alpha} \int h dP_n = \frac{1}{1-\alpha} \int h dP.$$

Y por tanto, $P_0 \in \mathcal{R}_\alpha(P)$. ■

El siguiente resultado establece una propiedad de continuidad respecto del tamaño de recorte.

Proposición 3.8. *Sea P una medida de probabilidad en un espacio métrico separable y completo. Si $\{\alpha_n\}_n \subset [0,1)$ tal que $\alpha_n \downarrow \alpha_0$, y $\{P_n\}_n$ es una sucesión de medidas de probabilidad tal que $P_n \in \mathcal{R}_{\alpha_n}(P)$ para cada n y $P_n \to_w P_0$. Entonces $P_0 \in \mathcal{R}_{\alpha_0}(P)$.*

Demostración. Combinando que $P_n \in \mathcal{R}_{\alpha_n}(P)$ y el Teorema Portmanteau, tenemos que para cada abierto A de β,

$$P_0(A) \leq \liminf_n P_n(A) \leq \liminf_n \frac{1}{1-\alpha_n} P(A) = \frac{1}{1-\alpha_0} P(A).$$

El resultado se sigue entonces del apartado (b) de la Proposición 3.3 y la regularidad de las medidas de probabilidad en espacios métricos. ■

Por lo visto en la Proposición 3.6 es claro que si Q es un recorte de nivel a lo sumo α_1 de P, entonces también lo es de nivel a lo sumo α_2, para todo α_2 tal que $\alpha_1 \leq \alpha_2$. Tiene entonces cierto interés definir el nivel exacto de un recorte asociado a P, así como estudiar alguna propiedad de las funciones de recorte arriba vistas de cara a entender un poco mejor el funcionamiento de los recortes.

Definición 3.9. *Se dice que un recorte Q de P es exactamente de nivel α si*

$$\alpha = \min\{\beta \in (0,1) : \ (1-\beta)Q(A) \leq P(A), \quad \forall A \in \beta\}.$$

Sabemos que si Q es un recorte de P de nivel exactamente α_1, también lo es de nivel a lo sumo α_2, donde $\alpha_1 < \alpha_2 \leq 1$. Por tanto, existirán $\gamma_1,\ \gamma_2 : \mathcal{X} \mapsto [0,1]$ con $\int \gamma_1(x)P(dx) = 1-\alpha_1$ y $\int \gamma_2(x)P(dx) = 1-\alpha_2$, y tal que:

$$Q(A) = \frac{1}{1-\alpha_1}\int_A \gamma_1(x)P(dx) = \frac{1}{1-\alpha_2}\int_A \gamma_2(x)P(dx).$$

La cuestión es qué relación hay entre γ_1 y γ_2. A partir de la anterior expresión es claro que

$$\gamma_2 = \frac{1-\alpha_2}{1-\alpha_1}\gamma_1 = k\gamma_1 \quad P\text{-c.s.},$$

con $0 < k < 1$.

Nos interesa por tanto tener una cierta idea de cómo son las funciones de recorte γ que dan lugar a recortes de un nivel exacto.

Definición 3.10. *Sea P una medida de probabilidad y $\alpha \in [0,1]$. Una función $\gamma : \mathcal{X} \mapsto \mathbb{R}$ con $0 \leq \gamma \leq 1$ P-c.s. y tal que $\int \gamma(x)P(dx) = 1-\alpha$ se dice maximal respecto de P si define un recorte de nivel exactamente α.*

Proposición 3.11. *Una función de recorte γ es maximal si y sólo si su supremo esencial calculado con respecto de P es uno.*

Demostración. Si γ es una función de recorte, se tiene entonces que $\gamma \leq 1$ P-c.s. Ahora, las relaciones siguientes son equivalentes:

(i) $ess\ sup_x\ \gamma(x) < 1$ P-c.s.,

(ii) existe $\varepsilon_0 > 0$ tal que $P(\gamma > 1 - \varepsilon_0) = 0$,

(iii) existe $\varepsilon_0 > 0$ tal que $\gamma \leq 1 - \varepsilon_0 < 1$ P-c.s., y

(iv) existe $k > 1$ y podemos definir γ' tal que $\gamma' = k\gamma \leq 1$ P-c.s., con lo que γ' define un recorte de P de nivel a lo sumo $\alpha' < \alpha$,

por lo tanto, γ no sería maximal.

■

Hemos visto en la Proposición 3.6 que el conjunto $\mathcal{R}_\alpha(P)$ de los recortes de nivel a lo sumo α de P es cerrado para la topología de la convergencia débil en $\mathcal{P}(\mathcal{X}, \beta)$. Siguiendo los mismos pasos de la demostración podríamos ver que si tenemos una sucesión $\{Q_n\}_n$ de recortes de nivel exactamente α de P y $Q_n \to_w Q$ entonces Q es un recorte de nivel a lo sumo α, pero no podemos asegurar que en general sea de nivel exactamente α, y por lo tanto la clase de recortes de nivel exactamente α de una probabilidad dada P no sería cerrada para la convergencia débil. El siguiente ejemplo con probabilidades en la recta real ilustra la anterior afirmación.

Ejemplo 3.12. Sean $P = U[0,1]$ y Q_n, $n > 1$, la medida de probabilidad definida en $[0,1]$ con función de densidad $f_n(x) = 2I_{[0,1/n]}(x) + \frac{n-2}{n-1} I_{[1/n,1]}(x)$, $0 \leq x \leq 1$. Es claro que $Q_n \ll P$ y que $\frac{dQ_n}{dP}(x) = f_n(x) \leq \frac{1}{1-0.5}$. Además la última desigualdad se verifica con igualdad si $x \in [0, 1/n]$, por tanto Q_n es un recorte de nivel exactamente 0.5 de P. Por otra parte, también es inmediato que $Q_n \to_w P$, y P es un recorte de nivel exactamente 0 de sí mismo. Con lo que podemos decir que el conjunto de recortes de nivel exactamente 0.5 de P no es una clase cerrada.

■

Podrían buscarse condiciones sobre las medidas de probabilidad P que dan lugar a que la clase de recortes de nivel exactamente α sea cerrada, pero no parece suficientemente interesante, más aún cuando tenemos ejemplos de otras propiedades que no se verifican para esta clase. En cambio, el manejo del conjunto $\mathcal{R}_\alpha(P)$ funciona bien y tiene buenas propiedades. A continuación mostramos otro ejemplo en el que se ve que la clase de recortes de nivel exactamente α tampoco verifica la propiedad de continuidad vista en el Teorema 3.7 para el conjunto $\mathcal{R}_\alpha(P)$.

Ejemplo 3.13. Sea $P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ y $P_0 = \frac{1}{4}\delta_0 + \frac{3}{4}\delta_1$. Tenemos entonces,

$$\begin{aligned} P_0(\{0\}) &= \frac{1}{4} \leq \frac{1}{1-\frac{1}{3}}\frac{1}{2} = \frac{1}{1-\frac{1}{3}}P(\{0\}), \\ P_0(\{1\}) &= \frac{3}{4} = \frac{1}{1-\frac{1}{3}}\frac{1}{2} = \frac{1}{1-\frac{1}{3}}P(\{1\}). \end{aligned} \tag{3.2}$$

Por tanto P_0 es un recorte de nivel exactamente $\alpha = \frac{1}{3}$ de P, al verificarse con igualdad para dicho α la expresión (3.2).

Si tomamos ahora $P_n = \frac{1}{n}\delta_{\frac{-1}{n}} + \left(\frac{1}{2} - \frac{2}{n}\right)\delta_0 + \left(\frac{1}{2} + \frac{1}{n}\right)\delta_1$ y $P_{n,\alpha} = \frac{2}{n}\delta_{\frac{-1}{n}} + \left(\frac{1}{4} - \frac{4}{n}\right)\delta_0 + \left(\frac{3}{4} + \frac{2}{n}\right)\delta_1$ tendremos que $P_n \to_w P$ y $P_{n,\alpha} \to_w P_0$ (y además uniformemente). Y también que,

$$P_{n,\alpha}\left(\left\{\frac{-1}{n}\right\}\right) = \frac{2}{n} = \frac{1}{1-\frac{1}{2}}\frac{1}{n} = \frac{1}{1-\frac{1}{2}}P_n\left(\left\{\frac{-1}{n}\right\}\right), \tag{3.3}$$

$$\begin{aligned} P_{n,\alpha}(\{0\}) &= \frac{1}{4} - \frac{4}{n} \leq \frac{1}{1-\frac{1}{2}}\left(\frac{1}{2} - \frac{2}{n}\right) = \frac{1}{1-\frac{1}{2}}P_n(\{0\}), \\ P_{n,\alpha}(\{1\}) &= \frac{3}{4} + \frac{2}{n} \leq \frac{1}{1-\frac{1}{2}}\left(\frac{1}{2} + \frac{1}{n}\right) = \frac{1}{1-\frac{1}{2}}P_n(\{1\}). \end{aligned}$$

Y por tanto $P_{n,\alpha}$ es un recorte de tamaño exactamente $\alpha = \frac{1}{2}$ de P_n para cada n, al verificarse con igualdad la expresión (3.3) para ese valor de α. ■

Cuando trabajamos en la recta real, resulta muy útil el uso del conjunto $\mathcal{C}_\alpha$, la clase de las funciones absolutamente continuas $h : [0,1] \mapsto [0,1]$ tal que, $h(0) = 0$, $h(1) = 1$, con derivada h' tal que $0 \leq h' \leq \frac{1}{1-\alpha}$. La compacidad de este conjunto en la topología $\|\cdot\|_\infty$ se prueba en la siguiente proposición y será clave en algunas de las demostraciones que veremos más adelante.

Proposición 3.14. *Sea $\alpha \in [0,1)$. El conjunto $\mathcal{C}_\alpha$ de todas las funciones absolutamente continuas $h : [0,1] \mapsto [0,1]$ tal que, $h(0) = 0$, $h(1) = 1$, con derivada h' tal que $0 \leq h' \leq \frac{1}{1-\alpha}$ es compacto para la topología $\|\cdot\|_\infty$.*

Demostración. El conjunto $\mathcal{C}_\alpha$ está uniformemente acotado en 0 ($h(0) = 0$ para cada $h \in \mathcal{C}_\alpha$) y es uniformemente equicontinuo ($|h(y) - h(x)| \leq \frac{1}{1-\alpha}|y - x|$ para cada $h \in \mathcal{C}_\alpha$). Entonces, por el Teorema de Arzelá-Ascoli, $\mathcal{C}_\alpha$ es relativamente compacto para $\|\cdot\|_\infty$ y es suficiente con ver que $\mathcal{C}_\alpha$ es cerrado. Supongamos entonces que $\{h_n\}_n$ son tal que $h_n \in \mathcal{C}_\alpha$ y $\|h_n - h\|_\infty \to 0$. Entonces

$$0 \leq h(y) - h(x) = \lim_{n\to\infty} (h_n(y) - h_n(x)) \leq \frac{1}{1-\alpha}(y - x), \quad \text{si } 0 \leq x \leq y \leq 1.$$

Esto implica que h es absolutamente continua y $0 \leq h' \leq \frac{1}{1-\alpha}$ casi seguro. De lo que $h \in \mathcal{C}_\alpha$, lo que completa la demostración. ∎

Teorema 3.15. *Para cualquier medida de probabilidad en la recta real, P,*

(a) $\mathcal{R}_\alpha(P) = \{P^* \in \mathcal{P} : P^*(-\infty, t] = h\left(P(-\infty, t]\right), \quad h \in \mathcal{C}_\alpha\}$.

(b) $\mathcal{R}_\alpha(U[0,1]) = \{P^* \in \mathcal{P} : P^*(-\infty, t] = h(t), 0 \leq t \leq 1, \quad h \in \mathcal{C}_\alpha\}$.

Demostración. Sea $\mathcal{A} = \{P^* \in \mathcal{P} : P^*(-\infty, t] = h\left(P(-\infty, t]\right), \quad h \in \mathcal{C}_\alpha\}$. Dado $P^* \in \mathcal{A}$, la continuidad absoluta de h supone que

$$P^*(s, t] = h\left(P(-\infty, t]\right) - h\left(P(-\infty, s]\right) = \int_{P(-\infty,s]}^{P(-\infty,t]} h'(x)\, dx \leq \frac{1}{1-\alpha} P(s, t].$$

Entonces, $P^* \ll P$ y $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$. De esta manera, $P^* \in \mathcal{R}_\alpha(P)$.

De forma opuesta, dado $P^* \in \mathcal{R}_\alpha(P)$, si F es la función de distribución de P y definimos $h(t) = \int_0^t \frac{dP^*}{dP}(F^{-1}(s))ds$, es inmediato que $h \in \mathcal{C}_\alpha$ y,

$$P^*(-\infty, t] = \int_{-\infty}^{t} \frac{dP^*}{dP}(s)dF(s) = \int_0^{F(t)} \frac{dP^*}{dP}(F^{-1}(s))ds = h\left(P(-\infty, t]\right).$$

Por tanto, $P^* \in \mathcal{A}$, y la primera parte está probada.

La parte (b) es inmediata a partir de (a). ∎

La parte (b) del Teorema 3.15 dice que la clase $\mathcal{C}_\alpha$ es la clase de todas las funciones de distribución de los recortes de nivel a lo sumo α de la distribución $U[0,1]$. Entonces, (a) nos da una caracterización de los recortes de nivel a lo sumo α de cualquier distribución de la recta real en términos de los recortes de nivel a lo sumo α de la distribución $U[0,1]$.

En adelante cuando escribamos P_h estaremos designando a la medida de probabilidad con función de distribución $h(P(-\infty, t])$, es decir, un recorte de P. El conjunto de recortes de nivel a lo sumo α de P puede entonces ser escrito también como $\mathcal{R}_\alpha(P) = \{P_h : h \in \mathcal{C}_\alpha\}$.

El siguiente teorema establece las bases para la generalización del anterior resultado a un espacio medible $\mathcal{X}$ general, y en particular a $\mathbb{R}^k$. Dicho resultado se da en el Corolario 3.17, en el que una medida de probabilidad cualquiera, Q, absolutamente continua respecto de la medida de Lebesgue, juega el papel que en la recta real tiene la distribución $U[0,1]$.

Teorema 3.16. *Sea Q una medida de probabilidad cualquiera de $\mathcal{P}(\mathcal{X}, \beta)$, y sea $T : \mathcal{X} \mapsto \mathcal{X}$ una función medible. Si T lleva Q en otra medida de probabilidad P, es decir $P = Q \circ T^{-1}$, entonces*

$$\mathcal{R}_\alpha(P) = \left\{P^* \in \mathcal{P}(\mathcal{X}, \beta) : P^* = Q^* \circ T^{-1},\, Q^* \in \mathcal{R}_\alpha(Q)\right\}.$$

Demostración. Si $\alpha = 1$ y Q^* es una medida de probabilidad cualquiera, absolutamente continua con respecto a Q, entonces si definimos $P^* := Q^* \circ T^{-1}$ tendremos que $P^* \ll P$, puesto que si $P(B) = 0$, entonces $Q(T^{-1}(B)) = 0$ y en consecuencia $P^*(B) = Q^*(T^{-1}(B)) = 0$. Viceversa, si $\alpha = 1$ y $P^* \ll P$, podemos definir $w(y) = \frac{dP^*}{dP}(T(y))$ y $Q^*(B) = \int_B w(y)Q(dy)$. Aplicando ahora la fórmula del cambio de variable para $B \in \beta$ tendremos,

$$Q^* \circ T^{-1}(B) = Q^*\left(T^{-1}(B)\right) = \int_{T^{-1}(B)} \frac{dP^*}{dP}(T(y))Q(dy) = \int_B \frac{dP^*}{dP}(x)P(dx) = P^*(B).$$

Si $\alpha < 1$ y $Q^* \in \mathcal{R}_\alpha(Q)$, entonces para cada $B \in \beta$, aplicando (b) en la Proposición 3.3,

$$Q^* \circ T^{-1}(B) = Q^*\left(T^{-1}(B)\right) \leq \frac{1}{1-\alpha} Q\left(T^{-1}(B)\right) = \frac{1}{1-\alpha} P(B), \tag{3.4}$$

y por tanto $Q^* \circ T^{-1} \in \mathcal{R}_\alpha(P)$.

Y viceversa, si $\alpha < 1$ y $P^* \in \mathcal{R}_\alpha(P)$, entonces $P^*(B) = \int_B \frac{dP^*}{dP}(x)P(dx), \quad \forall B \in \beta$. Si definimos $Q^*(B) := \int_B \frac{dP^*}{dP}(T(y))Q(dy)$, es claro que $Q^* \ll Q$, y como $P = Q \circ T^{-1}$, también que $\frac{dP^*}{dP}(T(y)) \leq \frac{1}{1-\alpha}$ Q-c.s. Luego $Q^* \in \mathcal{R}_\alpha(Q)$. Además mediante un cambio de variable es inmediato que si $B \in \beta$,

$$Q^* \circ T^{-1}(B) = Q^*\left(T^{-1}(B)\right) = \int_{T^{-1}(B)} \frac{dP^*}{dP}(T(y))Q(dy) = \int_B \frac{dP^*}{dP}(x)P(dx) = P^*(B).$$

Luego (3.4) sigue siendo válido ahora y tenemos el resultado. ∎

Aunque el resultado anterior puede establecerse (con igual demostración) para probabilidades definidas en distintos espacios ($Q \in \mathcal{P}(\mathcal{X}, \beta), T : \mathcal{X} \mapsto \mathcal{X}', P \in \mathcal{P}(\mathcal{X}', \beta')$), su uso a lo largo de esta memoria se reducirá al caso aquí planteado.

Obviamente, existen medidas de probabilidad en $\mathcal{X}$ que no pueden ser llevadas en cualquier otra probabilidad de $\mathcal{X}$. Pensemos por ejemplo, en una medida de probabilidad con al menos un punto discreto, es claro que cualquier medida de probabilidad transportada a partir de ésta tendrá al menos un punto discreto. Por otra parte, y como consecuencia del Teorema 2.10 mencionado en el Capítulo 2, si $\mathcal{X} = \mathbb{R}^k$ solamente algunas funciones medibles son apropiadas para manejar representaciones del conjunto de medidas de probabilidad de $\mathcal{X}$ en términos de una medida de probabilidad dada Q. Estas funciones son precisamente las funciones ciclícamente monótonas (ver Definición 2.9).

Así pues, a partir de los Teoremas 3.16 y 2.10 se obtiene el siguiente corolario,

Corolario 3.17. *Sean $P, Q \in \mathcal{P}(\mathbb{R}^k, \beta)$ cualesquiera, tal que Q es absolutamente continua respecto a la medida de Lebesgue. Entonces,*

$$\mathcal{R}_\alpha(P) = \left\{ P^* \in \mathcal{P}(\mathbb{R}^k, \beta) : P^* = Q^* \circ T^{-1},\, Q^* \in \mathcal{R}_\alpha(Q) \right\},$$

donde T es la (esencialmente) única función de transporte óptimo entre Q y P.

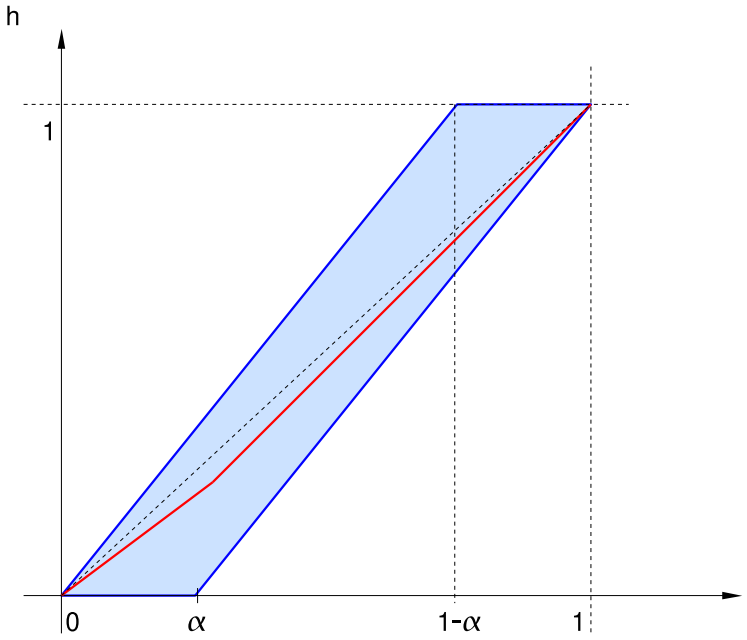

Figura 3.1: Banda en la que se mueven los recortes de nivel a lo sumo α de la $U[0, 1]$.

Como aplicación del Teorema 3.15 (extendido por el Corolario 3.17), en la Figura 3.1 se muestra en azul la banda en la que se mueven las funciones $h \in \mathcal{C}_\alpha$. Parten del punto de coordenadas $(0,0)$, llegan al $(1,1)$, son absolutamente continuas y su máxima pendiente en el interior de la banda no puede exceder de la pendiente de las líneas que delimitan la banda. En rojo aparece una de esas posibles funciones h. En la Figura 3.2 aparece en rojo una función h correspondiente a recortar enteramente una fracción de masa de probabilidad en las colas y otra en un subintervalo del interior de la $U[0,1]$. Las zonas planas de la función h indican los intervalos en el eje horizontal en el que se produce este recorte.

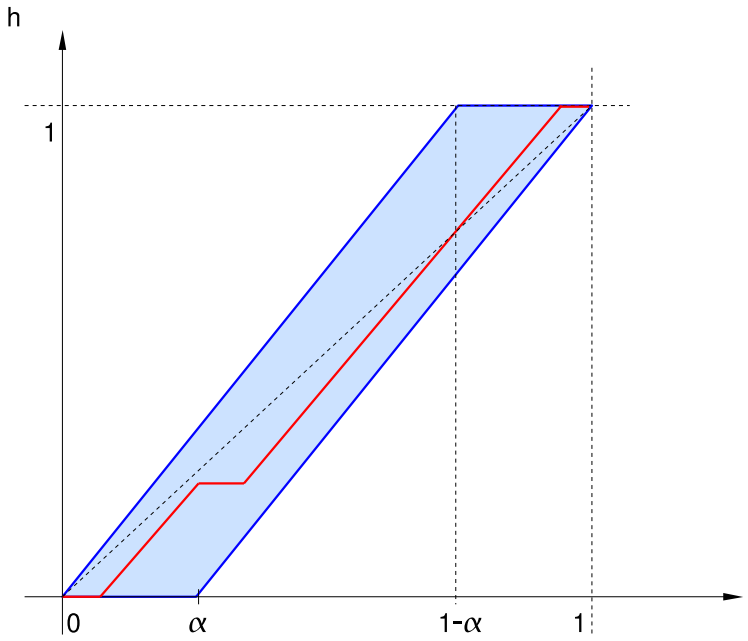

Figura 3.2: Ejemplo de función h.

Con la caracterización de recorte dada en la Proposición 3.3 es muy fácil ver que en el caso de que la contaminación de una muestra (resp. población) se produzca a partir de una mezcla, los recortes son capaces de recuperar la distribución original sin contaminar. Si tenemos una distribución P contaminada con una fracción ε de otra medida de probabilidad Q,

$$(1-\varepsilon)\,P + \varepsilon\,Q,$$

con un recorte de nivel a lo sumo ε de esta mixtura conseguimos eliminar Q y recuperar P. Es tan fácil como ver que si $\alpha \geq \varepsilon\,(0 < \alpha < 1)$ entonces:

$$(1-\alpha)\,P(A) \leq (1-\varepsilon)\,P(A) + \varepsilon\,Q(A), \qquad \forall A \in \beta.$$

La cuestión ahora sería saber si en algún caso con una fracción de recorte inferior a la contaminación conseguimos eliminarla. Por ejemplo, si los soportes de ambas fuesen completamente disjuntos es claro que no podría ser.

Proposición 3.18. *Sean P y Q dos medidas de probabilidad y $0 < \varepsilon,\ \alpha < 1$,*

(a) *Si $\alpha \geq \varepsilon$, entonces $P \in \mathcal{R}_\alpha\left((1-\varepsilon)P + \varepsilon Q\right)$. Además, cuando $\alpha < \varepsilon$, son equivalentes,*

(i) $P \in \mathcal{R}_\alpha\left((1-\varepsilon)P + \varepsilon Q\right)$, *y*

(ii) $P \in \mathcal{R}_{\frac{\alpha}{\varepsilon}}(Q)$.

(b) *Si $\beta = \frac{\alpha}{\varepsilon + \alpha(1-\varepsilon)}$, son equivalentes,*

(i) $(1-\varepsilon)P + \varepsilon Q \in \mathcal{R}_\alpha(P)$, *y*

(ii) $Q \in \mathcal{R}_\beta(P)$.

Demostración. La primera parte de (a) se ha visto antes del enunciado. Para la segunda parte, sea $0 < \alpha < \varepsilon$, $A \in \beta$, entonces son equivalentes,

(i) $(1-\alpha)\,P(A) \leq (1-\varepsilon)\,P(A) + \varepsilon\,Q(A)$,

(ii) $(\varepsilon - \alpha)\,P(A) \leq \varepsilon\,Q(A)$, y

(iii) $(1 - \frac{\alpha}{\varepsilon})\,P(A) \leq Q(A)$.

Para la parte (b), tenemos las siguientes equivalencias,

(i) $(1-\varepsilon)\,P(A) + \varepsilon\,Q(A) \leq \frac{1}{1-\alpha}P(A)$,

(ii) $Q(A) \leq \frac{(1-\alpha)\varepsilon + \alpha}{(1-\alpha)\varepsilon}\,P(A)$, y

(iii) $\left(1 - \frac{\alpha}{\varepsilon + \alpha(1-\varepsilon)}\right)\,Q(A) \leq P(A)$.

■

3.2. Recortes imparciales y comparación de distribuciones. Tipos de problemas.

Como ya se mencionó en la introducción, el uso de los recortes o distribuciones recortadas tiene una doble vertiente. Por su inspiración en ideas ya aplicadas en otros campos de la Estadística Robusta va a permitir el diseño de procedimientos de contraste o ajuste que son robustos frente a contaminaciones o perturbaciones en los datos. Y por otro lado, va a permitir el diseño de procedimientos que permitan evaluar de una forma innovadora la similitud o disimilitud entre dos muestras, o de una muestra y una distribución de referencia.

Este segundo enfoque de la metodología que se introduce en esta sección puede en particular aplicarse a lo que se ha dado en llamar dentro de la Estadística, análisis de bioequivalencia o biodisponibilidad (ver, por ejemplo, Chow y Liu, 1992).

Los estudios de bioequivalencia juegan un papel muy importante en el proceso de desarrollo de nuevos medicamentos. Son usados tanto por las grandes empresas farmacéuticas en la búsqueda de formulaciones más apropiadas para ser comercializadas cuando ya se dispone de una formulación de partida para la que se ha probado su seguridad y eficacia, como por las empresas farmacéuticas dedicadas al desarrollo de medicamentos genéricos. En todos estos casos, en vez de repetir todos los ensayos clínicos necesarios para validar la nueva formulación, se estudian las llamadas características farmaco-cinéticas, que se resumen en dos parámetros para cada individuo: la concentración máxima (C_{max}) del principio activo en plasma y el área bajo la curva tiempo-concentración en plasma del principio activo (AUC). Son estas dos variables las que se utilizan para comparar si las dos formulaciones, la de referencia y la nueva, tienen el mismo efecto terapéutico sobre los pacientes, y, si es así, se dirá que son bioequivalentes.

Así pues lo relevante en estos estudios no es que los parámetros-resumen (medias, varianzas, la propia distribución,...) de las distribuciones correspondientes a las características arriba mencionadas, medidas en la muestra de pacientes que toma el medicamento de referencia y en la que toma el nuevo medicamento, sean iguales, sino que den lugar a un mismo efecto terapéutico. Y para que ocurra esto no es necesario que dichos parámetros sean iguales, basta con que sean similares. El grado de similitud vendrá establecido por los investigadores o, como suele ser habitual, por la autoridades regulatorias (FDA en el caso de Estados Unidos y la Agencia Europea de Medicamentos en el caso europeo).

En este contexto los recortes imparciales pueden ser contemplados no sólo como una forma de robustificar el procedimiento de una forma más flexible (respecto al recorte que consiste en eliminar los datos de las colas), sino también como una forma de ampliar el concepto de bioequivalencia. Es posible que, en ciertos casos, y debido a la posible existencia de subgrupos dentro de la población de estudio u otras razones, sea suficiente con que las dos formulaciones tengan el mismo efecto terapéutico para, por ejemplo, el 95 % de la población.

El adjetivo de imparciales se añade para recalcar el hecho de que no es el investigador el que decide qué puntos de la muestra se eliminan o su importancia es disminuida, sino que son los propios datos y la estructura del tipo de problema que se plantea los que deciden cuales serán estos puntos.

De forma general podemos distinguir dos tipos de problemas, los de comparación de una muestra contra una población de referencia y los de comparación de dos muestras.

3.2.1. Problemas de una muestra.

Los tipos de problemas de una muestra que podemos abordar con la metodología de los recortes imparciales son variados, dependiendo del objetivo y alcance del estudio que pretendamos llevar a cabo. Pero principalmente podemos citar dos: Por una parte la robustificación de los procedimientos tradicionales cuando se tiene una distribución de referencia y por otra, la valoración del grado de similitud de una muestra con una distribución de referencia. Estas dos ideas se pueden tratar, a su vez, de diferentes formas, que van desde la valoración meramente descriptiva que se puede hacer en un contexto de Análisis de Datos (ver Capítulo 4), a la elaboración de diferentes contrastes de hipótesis si se pretende hacer inferencias.

En este último caso, existen también varias posibilidades dependiendo de la formulación que hagamos en términos de la hipótesis nula a contrastar. En un primer momento podemos pensar en distintas posibilidades que van desde la más sencilla que sería contrastar la hipótesis nula simple $\mathcal{F} = \{P\}$, a una hipótesis nula compuesta del tipo $\mathcal{F} = \mathcal{R}_\alpha(P)$. Pero antes de seguir con otras posibilidades, formalicemos un poco el problema.

Sean P y Q dos medidas de probabilidad en $\mathcal{P}(\mathcal{X}, \beta)$. En adelante P jugará el papel de distribución de referencia y Q el de la distribución que queremos comparar con P. Si d es una métrica (en un cierto subconjunto de $\mathcal{P}(\mathcal{X}, \beta)$) podemos definir diferentes "distancias

recortadas" entre P y Q, es decir, medidas de cuanto de lejos está P de Q una vez que eliminamos o disminuimos la influencia de ciertas regiones de $\mathcal{X}$ en P, en Q o en ambas a la vez. Concretamente definimos,

$$T_1(P,Q,d,\alpha) = \inf_{R\in\mathcal{R}_\alpha(P)} d(R,Q), \tag{3.5}$$

$$T_2(P,Q,d,\alpha) = \inf_{R\in\mathcal{R}_\alpha(Q)} d(P,R), \tag{3.6}$$

$$T_3(P,Q,d,\alpha) = \inf_{R_1\in\mathcal{R}_\alpha(P),R_2\in\mathcal{R}_\alpha(Q)} d(R_1,R_2). \tag{3.7}$$

Aunque T_1, T_2, T_3 no sean en general distancias nos referiremos en adelante a ellas como *distancias recortadas*. Es obvio que tenemos las siguientes desigualdades:

$$T_1(P,Q,d,\alpha) \geq T_3(P,Q,d,\alpha), \text{ y } T_2(P,Q,d,\alpha) \geq T_3(P,Q,d,\alpha).$$

Volviendo al análisis de similitud ya mencionado, otro tipo de hipótesis nula que puede tener sentido querer contrastar sería $H_0 : T_i(P,Q,d,\alpha) \geq \Delta_0$ (o también $H_0 : T_i(P,Q,d,\alpha) \leq \Delta_0$), $i = 1,2,3$, donde Δ_0 es un valor umbral preestablecido que indicaría que dos distribuciones son similares si después de recortadas distan menos de Δ_0 .

En la práctica trataremos con una muestra de variables/vectores aleatorios independientes e igualmente distribuidos $X_1,\ldots,X_n$, y será la distribución empírica asociada la que juegue el papel de Q en T_1, T_2 ó T_3 para los fines arriba mencionados. Veamos brevemente algunos ejemplos de posibles usos de estos estadísticos. Si P_n denota la medida empírica basada en $X_1,\ldots,X_n$ entonces $T_1(P,P_n,d,\alpha)$ podría ser usado para contrastar la hipótesis nula $H_0 : \mathcal{L}(X_i) \in \mathcal{R}_\alpha(P)$. También podríamos definir un test robusto para la hipótesis nula $H_0 : \mathcal{L}(X_i) = P$ rechazando H_0 para valores grandes de $T_2(P,P_n,d,\alpha)$.

En principio parece que las versiones empíricas de T_1 y T_2 son más adecuadas para los problemas de una muestra, mientras que T_3 sería más apropiado para los problemas de dos muestras. Pero tampoco sería descabellado utilizar T_3 en un problema de una muestra. Imaginemos, por ejemplo, una situación en la que queremos saber si el modelo generador de los datos es esencialmente normal, pues sospechamos que en una zona pequeña de valores se producen perturbaciones (desconocidas y sin interés) que lo alejan de la normalidad. En este caso nos gustaría poder recortar en la muestra para eliminar posibles contaminaciones de cualquier tipo que hagan el procedimiento más robusto. Pero también podríamos querer recortar en el modelo de referencia para eliminar el efecto de dichas perturbaciones, y poder hablar así de casi normalidad.

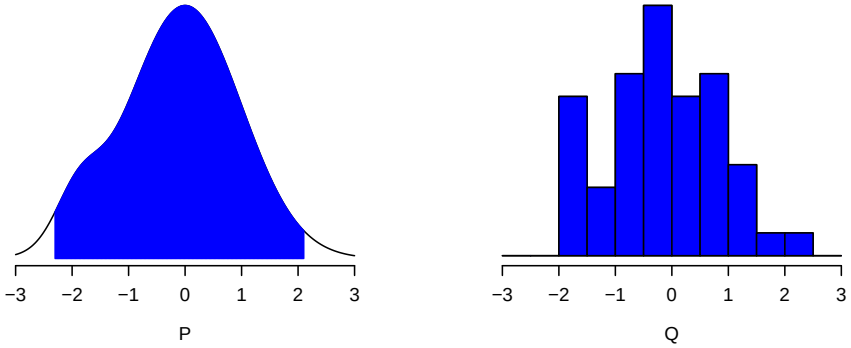

Figura 3.3: Ejemplo en el que se recorta un 10 % una distribución $N(0,1)$ (P), según el esquema definido por T_1, para hacerla más parecida a una muestra de 50 valores al azar de una $N(0,1)$ (distribución Q), que no se recorta. En azul, las distribuciones después de recortar.

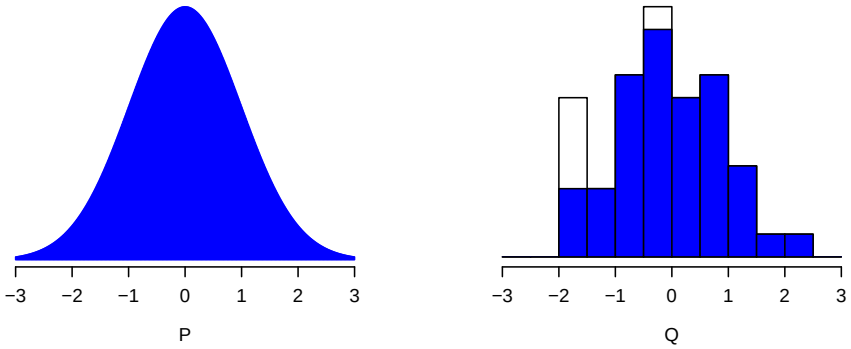

Figura 3.4: Ejemplo en el que se recorta un 10 % la muestra de 50 valores al azar de una $N(0,1)$ (distribución Q) para hacerla más parecida a la $N(0,1)$ (P), que no se recorta (esquema de recorte definido por T_2). En azul, las distribuciones después de recortar.

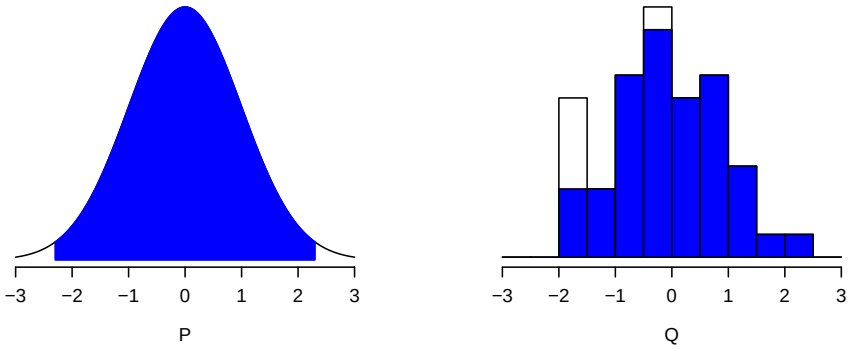

Figura 3.5: Ejemplo en el que se recortan un 10 % ambas distribuciones, según el esquema definido por T_3. En azul, las distribuciones después de recortar.

En las Figuras 3.3, 3.4 y 3.5 se muestran ejemplos de cada una de estas tres formas de recortar. En dichas figuras, la medida P corresponde a una $N(0,1)$ y la medida Q a la distribución empírica de 50 valores generados al azar de una $N(0,1)$. Obsérvese que lo que más aleja de la normalidad a los datos de la muestra es el exceso de datos en la cola izquierda, es por ello que el uso del procedimiento T_2 conduce, sobre todo, a un recorte en esta zona. Consideramos especialmente interesante la Figura 3.3. En ella ha aparecido una segunda moda en el lado izquierdo de la densidad de P después de recortar. Ello ha sido debido a que, para acomodarse al exceso de datos en la cola izquierda de la muestra, la función de recorte está por debajo de uno excepto en esta zona. Al dividir por $(1-\alpha)$ aparece una moda, aproximadamente, en -2. Hablando un poco imprecisamente, podríamos decir que la interpretación más razonable de los datos de la muestra es que contienen una muestra tomada de P y otra proveniente de una contaminación situada alrededor de -2. Si recortamos la muestra para que se parezca a P, automáticamente se elimina la contaminación. Pero si decidimos dejar la muestra intacta y recortar P, el procedimiento hace que P se transforme para que aparezca una segunda subpoblación alrededor de -2.

Como hemos visto el concepto de distribución recortada es independiente del tipo de métrica utilizada. Hasta aquí nos hemos referido a una métrica genérica. Es claro que dependiendo de cómo se quiera valorar las distancias entre las medidas de probabilidad se adoptará una u otra. Así por ejemplo si nos preocupa la máxima diferencia entre funciones de distribución utilizaremos la métrica $\mathcal{W}_\infty$ (ver Definición 2.7). También pueden ser útiles métricas de tipo L_p como las métricas de Wasserstein introducidas en la Definición 2.1. Como ya se mencionó en la introducción, en esta memoria nos centraremos en la distancia $\mathcal{W}_2$ de Wasserstein con el objeto de no "oscurecer" el papel de los recortes y su utilidad con la diversificación de la casuística que tendríamos si tratásemos todas las métricas en pie de igualdad. No obstante, en la Sección 3.4 se ofrecen algunos procedimientos algorítmicos especialmente diseñados para otras métricas.

Cuando manejemos la distancia L_2 de Wasserstein, $d = \mathcal{W}_2$, nos circunscribiremos necesariamente al conjunto de medidas de probabilidad con momento de orden dos finito que denotaremos por $\mathcal{P}_2(\mathcal{X}, \beta)$. Y en el caso de estar en la recta real, podremos manejar la representación cuantil dada por el Lema 2.2. Ésta fue la métrica utilizada en el cálculo de los recortes que se muestran en las Figuras 3.3, 3.4 y 3.5.

En el caso de estar en la recta real, la caracterización de los recortes en términos de recortes de la $U[0,1]$ dada en el Teorema 3.15, nos permite reescribir las distancias recortadas (3.5)-(3.7) de la siguiente forma,

$$T_1(P,Q,d,\alpha) = \inf_{h\in\mathcal{C}_\alpha} d(P_h,Q), \tag{3.8}$$

$$T_2(P,Q,d,\alpha) = \inf_{h\in\mathcal{C}_\alpha} d(P,Q_h), \tag{3.9}$$

$$T_3(P,Q,d,\alpha) = \inf_{h_1\in\mathcal{C}_\alpha, h_2\in\mathcal{C}_\alpha} d(P_{h_1},Q_{h_2}). \tag{3.10}$$

donde P_h designa, como ser recordará, el recorte de nivel a lo sumo α de P que tiene función de distribución $h(P(-\infty,t])$, $h\in\mathcal{C}_\alpha$. Y de forma similar para Q_h.

3.2.2. Problemas de dos muestras.

Los tipos de problemas de dos muestras que se pueden abordar con los recortes imparciales siguen las mismas líneas que los de una muestra, donde la distribución de referencia es sustituida por una segunda muestra. Los enfoques, de igual forma, van desde el más puro análisis descriptivo a la elaboración de procedimientos de inferencia para contrastar ciertas hipótesis.

Algunas de las hipótesis nulas que tiene sentido plantearse en este apartado son

$$H_0 : P = Q, \tag{3.11}$$

$$H_0 : \exists\, P_\alpha \in \mathcal{R}_\alpha(P) \text{ y } \exists\, Q_\alpha \in \mathcal{R}_\alpha(Q) \text{ con } P_\alpha = Q_\alpha, \tag{3.12}$$

$$H_0 : T_3(P,Q,d,\alpha) \geq \Delta_0 \quad (\text{resp. } \leq)\ . \tag{3.13}$$

Para las tres hipótesis nulas previas podemos utilizar la versión empírica de T_3. Si estamos en el caso (3.11), estaríamos diseñando un procedimiento robusto que nos proteja frente a posibles contaminaciones. En los casos (3.12) y (3.13) incidiríamos más en la valoración de la similitud de las distribuciones, por supuesto, pero también protegiéndonos frente a posibles contaminaciones.

En el caso de dos muestras pueden existir situaciones que aconsejen recortar las dos muestras siguiendo un mismo "patrón" de recorte en ambas. Pensemos, por ejemplo, en una situación en la que queremos comparar dos muestras de datos de una misma magnitud física pero sabemos que los dispositivos de medida no son perfectos e introducen distorsiones (diferentes) cuando los valores están en cierto rango común a ambos dispositivos, no afectando

al resto de valores. En este caso sería razonable un recorte en la misma zona de cuantiles de ambas muestras. Esto nos lleva a plantear una cuarta distancia recortada, que en el caso de manejar distribuciones en la recta real es fácil de formalizar siguiendo la caracterización dada en el Teorema 3.15 como sigue,

$$T_4(P,Q,d,\alpha) = \inf_{h\in\mathcal{C}_\alpha} d(P_h,Q_h). \tag{3.14}$$

Y que a partir del Corolario 3.17 podemos generalizar a $\mathbb{R}^k$ como se muestra a continuación. Sean $P_0, P, Q \in \mathcal{P}(\mathbb{R}^k)$, $P_0 \ll \ell^k$ y T_P, T_Q las respectivas funciones de transporte óptimo entre P_0 y P, y entre P_0 y Q, es decir, $P = P_0 \circ T_P^{-1}$ y $Q = P_0 \circ T_Q^{-1}$. Podemos definir la distancia recortada entre P y Q de acuerdo al patrón dado por P_0 como,

$$T_4(P,Q,d,\alpha) = \inf_{P_0^*\in\mathcal{R}_\alpha(P_0)} d(P_0^*\circ T_P^{-1}, P_0^*\circ T_Q^{-1}). \tag{3.15}$$

Es obvio que esta forma de recortar es más restrictiva que la manejada en T_3. Por lo tanto,

$$T_3(P,Q,d,\alpha) \le T_4(P,Q,d,\alpha).$$

Si $d = \mathcal{W}_2$, y P, $Q \in \mathcal{P}_2(\mathbb{R})$ entonces (3.14) se puede escribir como,

$$\begin{aligned} T_4(P,Q,d,\alpha) &= \inf_{h\in\mathcal{C}_\alpha}\left(\int_0^1 (F^{-1}(h^{-1}(x)) - G^{-1}(h^{-1}(x)))^2 dx\right)^{1/2} \\ &= \inf_{h\in\mathcal{C}_\alpha}\left(\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt\right)^{1/2}, \end{aligned} \tag{3.16}$$

donde F^{-1} y G^{-1} son las funciones cuantiles de P y Q respectivamente.

La solución del problema de optimización (3.16) es bastante directa. Consideremos la función $t \longmapsto |F^{-1}(t) - G^{-1}(t)|$ como una variable aleatoria definida en el intervalo $(0,1)$ provisto de la medida de Lebesgue, ℓ. Denotemos ahora por

$$L_{F,G}(x) := \ell\{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| \le x\},\ x \ge 0, \tag{3.17}$$

a su función de distribución, y por $L_{F,G}^{-1}$ a su función cuantil. Si $L_{F,G}$ es continua en $L_{F,G}^{-1}(1-\alpha)$ entonces $L_{F,G}(L_{F,G}^{-1}(1-\alpha)) = 1-\alpha$, y

$$\inf_{h\in\mathcal{C}_\alpha}\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h_0'(t)dt, \tag{3.18}$$

donde

$$h_0'(t) = \frac{1}{1-\alpha} I_{[0,L_{F,G}^{-1}(1-\alpha)]}(|F^{-1}(t) - G^{-1}(t)|). \tag{3.19}$$

En ese caso h_0 es de hecho el *único* minimizador del criterio funcional.

Cuando $L_{F,G}$ no sea continua en $L_{F,G}^{-1}(1-\alpha)$, entonces, de la definición de la función cuantil, tendremos,

$$\begin{aligned}\ell\{t\in(0,1):\,|F^{-1}(t)-G^{-1}(t)|<L_{F,G}^{-1}(1-\alpha)\}&\\ \leq\quad 1-\alpha\leq\ell\{t\in(0,1):\,|F^{-1}(t)-G^{-1}(t)|\leq L_{F,G}^{-1}(1-\alpha)\},&\end{aligned}$$

donde, al menos una de las desigualdades es estricta. Por lo tanto podemos también asegurar la existencia de un conjunto A_0 (no necesariamente único) tal que $\ell(A_0)=1-\alpha$ y

$$\begin{aligned}\{t\in(0,1):\,|F^{-1}(t)-G^{-1}(t)|<L_{F,G}^{-1}(1-\alpha)\}&\\ \subset\quad A_0\subset\{t\in(0,1):\,|F^{-1}(t)-G^{-1}(t)|\leq L_{F,G}^{-1}(1-\alpha)\}.&\end{aligned}$$

Obviamente, si para tales conjuntos A_0 consideramos la función $h_0\in\mathcal{C}_\alpha$, tal que $h_0'=\frac{1}{1-\alpha}I_{A_0}$, entonces el inferior en (3.16) se alcanza para h_0. Así pues, el problema (3.16) es equivalente a resolver:

$$\inf_A\left(\frac{1}{1-\alpha}\int_A\left(F^{-1}(t)-G^{-1}(t)\right)^2dt\right)^{1/2},\tag{3.20}$$

donde ahora el conjunto A varía en los conjuntos de la σ-álgebra de Borel en $(0,1)$ con medida de Lebesgue igual a $1-\alpha$.

El minimizador, h_0, nos proporciona, teniendo en cuenta el Teorema 3.15, las funciones de distribución de las medidas recortadas para las que se alcanza el inferior en (3.16), $h_0(F(x))$ y $h_0(G(x))$ respectivamente.

Podemos comparar con otras distancias recortadas como, por ejemplo, la introducida por Munk y Czado (1998). Como señalamos anteriormente, el criterio propuesto por estos autores consiste en recortar $\alpha/2$ en cada una de las colas de las dos distribuciones. Por lo tanto, este recorte es más restrictivo que el propuesto por T_4 (y también obviamente, al dado por T_3), y se tiene,

$$T_4(P,Q,\mathcal{W}_2,\alpha)^2\leq\frac{1}{1-\alpha}\int_{\alpha/2}^{1-\alpha/2}(F^{-1}(t)-G^{-1}(t))^2dt.\tag{3.21}$$

En realidad es fácil comprobar que el lado derecho de la anterior expresión es igual a $\mathcal{W}_2^2(P_\alpha,Q_\alpha)$, cuando P_α es la medida de probabilidad con función de distribución

$$F_\alpha(t)=\begin{cases}0 & \text{si}\quad -\infty<t<F^{-1}\left(\frac{\alpha}{2}\right),\\ \frac{1}{1-\alpha}\left(F(t)-\frac{\alpha}{2}\right) & \text{si}\quad F^{-1}\left(\frac{\alpha}{2}\right)\leq t<F^{-1}\left(1-\frac{\alpha}{2}\right),\\ 1 & \text{si}\quad F^{-1}\left(1-\frac{\alpha}{2}\right)\leq t<\infty,\end{cases}$$

y de forma similar para Q_α. Claramente, $P_\alpha \ll P$ y $\frac{dP_\alpha}{dP}$ es igual a $\frac{1}{1-\alpha}$ si $F^{-1}\left(\frac{\alpha}{2}\right) \leq t < F^{-1}\left(1-\frac{\alpha}{2}\right)$ y 0 en el resto. Y también $Q_\alpha \ll Q$ y $\frac{dQ_\alpha}{dQ}$ es igual a $\frac{1}{1-\alpha}$ si $G^{-1}\left(\frac{\alpha}{2}\right) \leq t < G^{-1}\left(1-\frac{\alpha}{2}\right)$ y 0 en el resto. De lo que se deduce que P_α y Q_α son recortes de P y Q respectivamente. Además, si $h'(t) = \frac{1}{1-\alpha} I_{[\alpha/2,1-\alpha/2]}(t)$, se tiene que $F_\alpha(t) = h\left(F(t)\right)$ y $G_\alpha(t) = h\left(G(t)\right)$, y por tanto (3.21).

3.3. Propiedades generales de los recortes óptimos.

En esta sección describiremos algunas de las propiedades de los recortes cuando se plantean como mejores aproximantes en el sentido descrito por las expresiones (3.5)-(3.7) y (3.14).

3.3.1. Existencia.

Veamos en primer lugar que los problemas planteados en (3.5)-(3.7) y (3.14) están bien definidos y admiten solución.

Teorema 3.19 (**Problema de una muestra**). *Sea $(\mathcal{X},\beta)$ un espacio métrico separable y completo dotado de su σ-álgebra de Borel, y sean $P, Q \in \mathcal{P}(\mathcal{X},\beta)$ y d una métrica en $\mathcal{P}(\mathcal{X},\beta)$ que metriza la convergencia débil. Entonces existe al menos una medida de probabilidad $P_0 \in \mathcal{R}_\alpha(P)$ tal que*

$$d(P_0, Q) = \min_{R \in \mathcal{R}_\alpha(P)} d(R, Q).$$

Demostración. Es inmediato si tenemos en cuenta que se verifican las condiciones de la Proposición 3.6, con lo que el conjunto $\mathcal{R}_\alpha(P)$ es compacto. ∎

Teorema 3.20 (**Problema de dos muestras**). *Sea $(\mathcal{X},\beta)$ un espacio métrico separable y completo dotado de su σ-álgebra de Borel, y sean $P, Q \in \mathcal{P}(\mathcal{X},\beta)$ y d una métrica en $\mathcal{P}(\mathcal{X},\beta)$ que metriza la convergencia débil. Entonces existen al menos dos medidas de probabilidad $P_0 \in \mathcal{R}_\alpha(P)$ y $Q_0 \in \mathcal{R}_\alpha(Q)$ tales que*

$$d(P_0, Q_0) = \min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} d(R_1, R_2).$$

Demostración. Tenemos por una parte que al verificarse las condiciones de la Proposición 3.6, tanto $\mathcal{R}_\alpha(P)$ como $\mathcal{R}_\alpha(Q)$ son compactos. Por otra, si consideramos en $\mathcal{P}(\mathcal{X},\beta)\times\mathcal{P}(\mathcal{X},\beta)$ la topología producto derivada de la topología de la convergencia débil en cada espacio por separado, el conjunto $\mathcal{R}_\alpha(P)\times\mathcal{R}_\alpha(Q)$ es compacto, al serlo por componentes (por el Teorema de Tychonoff, ver por ejemplo, Ash, 1972).

La demostración concluye teniendo en cuenta que d es continua para la convergencia en distribución, y estamos hallando el inferior de dicha función en el compacto $\mathcal{R}_\alpha(P)\times\mathcal{R}_\alpha(Q)$, y por tanto se alcanza en el mismo. ■

Los recortes óptimos que aparecen en los Teoremas 3.19 y 3.20 no tienen porque ser en principio únicos, pero la convexidad del conjunto de recortes garantiza que el conjunto de mejores aproximantes es convexo si la métrica d es convexa (en el sentido de que, $d(\gamma P+(1-\gamma)Q,R)\leq\gamma d(P,R)+(1-\gamma)d(Q,R)$ para todo $\gamma\in[0,1]$). Esto ocurre por ejemplo para la métrica "bounded Lipschitz", o como vimos en el Lema 2.12, para la métrica $\mathcal{W}_2$ de Wasserstein.

Centrémonos en la métrica de Wasserstein $\mathcal{W}_p$, $p\geq 1$, dada en la Definición 2.1. Es posible ver que $\mathcal{W}_p$ define una distancia en el conjunto $\mathcal{P}_p\subset\mathcal{P}(\mathcal{X},\beta)$ de medidas de probabilidad con momento de orden p finito siempre que $\mathcal{X}$ sea un espacio métrico separable (ver Lema 8.1 en Bickel y Freedman, 1981). Ahora, si P tiene momento de orden p finito y $Q\in\mathcal{R}_\alpha(P)$ entonces

$$\int\|x\|^p dQ(x)\leq\frac{1}{1-\alpha}\int\|x\|^p dP(x).$$

Esto demuestra que $\mathcal{R}_\alpha(P)\subset\mathcal{P}_p$ si $P\in\mathcal{P}_p$.

El siguiente resultado nos da una versión de la Proposición 3.6 para la métrica $\mathcal{W}_p$,

Proposición 3.21. *Sea $(\mathcal{X},\beta)$ un espacio métrico separable y sea $P\in\mathcal{P}_p$. Entonces $\mathcal{R}_\alpha(P)$ es compacto en la topología generada por $\mathcal{W}_p$.*

Demostración. Por una parte tenemos que el conjunto $\mathcal{R}_\alpha(P)$ es uniformemente *tight* pues como vimos en la Sección 3.1,

$$Q(A)=\int_A\frac{dQ}{dP}dP\leq\frac{1}{1-\alpha}\int_A dP=\frac{1}{1-\alpha}P(A),\quad A\in\beta,$$

por lo que dado un conjunto infinito $\mathcal{R}\subset\mathcal{R}_\alpha(P)$ podemos extraer una sucesión $Q_n\in\mathcal{R}$ que converge débilmente. Sea Q dicho límite. Entonces $\mathcal{W}_p(Q_n,Q)\to 0$ si y sólo si $\|x\|^p$ es

uniformemente Q_n-integrable (ver Lema 2.3). Fijando $t > 0$, tendremos,

$$\int_{\{\|x\|>t\}} \|x\|^p dQ_n(x) = \int_{\{\|x\|>t\}} \|x\|^p \frac{dQ_n}{dP}(x) dP(x) \leq \frac{1}{1-\alpha} \int_{\{\|x\|>t\}} \|x\|^p dP(x).$$

Esto completa la demostración. ■

3.3.2. Unicidad.

En esta subsección nos centraremos en el espacio euclídeo k-dimensional con la norma usual y denotaremos por ℓ^k a la medida de Lebesgue en este espacio. Además, manejaremos la métrica $\mathcal{W}_2$ de Wasserstein para el cálculo de los recortes.

La unicidad del recorte imparcial, solución de los problemas de mínimo planteados en la anterior sección, es de capital importancia a la hora de tratar de obtener resultados asintóticos como la consistencia y otros. En la literatura de clasificación robusta se puede constatar que no es fácil, en general, probar la unicidad, por lo que lo habitual es suponerla basándose en argumentos diversos. Sólo en algún caso muy particular como por ejemplo en el caso de la 1-media recortada imparcial para distribuciones elípticas unimodales existe un resultado general de este tipo (ver García Escudero et al., 1999a). En nuestro caso no siempre habrá unicidad como muestra el siguiente ejemplo.

Ejemplo 3.22. Si $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ y $Q = \delta_0$. Supongamos que queremos encontrar el recorte de P de nivel a lo sumo α que mejor se aproxima a Q en métrica $\mathcal{W}_p$ de Wasserstein. Entonces, si $R \in \mathcal{R}_\alpha(P)$, es inmediato que $\mathcal{W}_p^p(R, Q) = \int_0^1 |F_R^{-1}(t)|^p dt = \int |x|^p dR(x) = 1$. Por lo que cada $R \in \mathcal{R}_\alpha(P)$ nos da una mejor aproximación a Q y el conjunto de mejores aproximantes es el de todos los posibles recortes de P, $\mathcal{R}_\alpha(P)$. ■

No obstante, bajo ciertas condiciones de regularidad bastante generales, podremos asegurar la unicidad de la mejor aproximación. A continuación se presenta un ejemplo significativo de unicidad en la recta real.

Ejemplo 3.23. Sea $Q = U(0, 1)$, la distribución uniforme en el intervalo unitario de la recta real, y $P \in \mathcal{P}_2(\mathbb{R})$. Sea F la función de distribución asociada a P. Hemos visto en el Teorema 3.15 que el conjunto $\mathcal{R}_\alpha(P)$ puede ser escrito como $\mathcal{R}_\alpha(P) = \{P_h : h \in \mathcal{C}_\alpha\}$, donde P_h es

la medida de probabilidad con función de distribución $F_h(x) = h(F(x))$ y $\mathcal{C}_\alpha$ es la clase de funciones dada en dicho teorema.

Sean ahora $h_1, h_2 \in \mathcal{C}_\alpha$ tales que $P_{h_1} \neq P_{h_2}$ y definamos $v(r) = \mathcal{W}_2^2(rP_{h_1} + (1-r)P_{h_2}, Q)$, $r \in (0,1)$, donde $h_i \in \mathcal{C}_\alpha$, $i = 1, 2$. Entonces,

$$\begin{aligned} v(r) &= \int_0^1 (y - F^{-1}((rh_1 + (1-r)h_2)^{-1}(y)))^2 dy \\ &= \frac{1}{3} + \int_0^1 F^{-1}((rh_1 + (1-r)h_2)^{-1}(y))^2 dy - 2\int_0^1 y\, F^{-1}((rh_1 + (1-r)h_2)^{-1}(y))\, dy \\ &= \frac{1}{3} + \int_0^1 F^{-1}(t)^2 (rh_1'(t) + (1-r)h_2'(t)) dt \\ &\quad -2\int_0^1 F^{-1}(t)(rh_1(t) + (1-r)h_2(t))(rh_1'(t) + (1-r)h_2'(t)) dt \\ &= Ar^2 + Br + C, \end{aligned}$$

donde

$$A = -2\int_0^1 F^{-1}(t)(h_1(t) - h_2(t))(h_1'(t) - h_2'(t)) dt.$$

Si integramos por partes podemos ver que

$$A = \int_0^1 (h_1(t) - h_2(t))^2 dF^{-1}(t) = \int (h_1(F(x)) - h_2(F(x)))^2 dx.$$

Por lo tanto $A > 0$ (en otro caso $h_1 \circ F = h_2 \circ F$, en contra de $P_{h_1} \neq P_{h_2}$) y v es estrictamente convexa. Y por ello existe un único $P_\alpha \in \mathcal{R}_\alpha(P)$ tal que

$$\mathcal{W}_2(P_\alpha, Q) = \min_{R \in \mathcal{R}_\alpha(P)} \mathcal{W}_2(R, Q).$$

■

En la Sección 2.2 se recoge el Lema 2.12 que muestra la convexidad de la métrica $\mathcal{W}_2$ en $\mathbb{R}^k$, en el siguiente sentido,

$$\mathcal{W}_2^2(\gamma P + (1-\gamma)Q, R) \leq \gamma \mathcal{W}_2^2(P, R) + (1-\gamma)\mathcal{W}_2^2(Q, R), \text{ para cada } \gamma \in (0,1),$$

y para cualesquiera $P, Q, R \in \mathcal{P}_2(\mathbb{R}^k)$.

Aquí probamos, a partir de la propiedad (a) de la Proposición 2.11 y el siguiente resultado, que si además se cumple que $P \ll \ell^k$, entonces se verifica la convexidad estricta.

Teorema 3.24. *Sean $P_i, Q_i, i = 1, 2$, medidas de probabilidad de $\mathcal{P}_2(\mathbb{R}^k)$ tal que $P_i \ll \ell^k, i = 1, 2$. Si $Q_1 \neq Q_2$ y no existe una función de transporte óptimo común T tal que $Q_1 = P_1 \circ T^{-1}$ y $Q_2 = P_2 \circ T^{-1}$, entonces, para cada γ en $(0, 1)$,*

$$\mathcal{W}_2^2(\gamma P_1 + (1-\gamma)P_2, \gamma Q_1 + (1-\gamma)Q_2) < \gamma \mathcal{W}_2^2(P_1, Q_1) + (1-\gamma)\mathcal{W}_2^2(P_2, Q_2).$$

Demostración. Supongamos que f_i es la función de densidad de P_i, y $(X_i, T_i(X_i))$, $i = 1, 2$ son los emparejamientos óptimos para $(P_i, Q_i), i = 1, 2$. Si definimos $P_\gamma := \gamma P_1 + (1-\gamma)P_2$ y $Q_\gamma := \gamma Q_1 + (1-\gamma)Q_2$, entonces $f_\gamma := \gamma f_1 + (1-\gamma)f_2$ será la función de densidad de la medida de probabilidad P_γ. Definimos ahora en el soporte de P_γ la siguiente función aleatoria:

$$T(x) = \begin{cases} T_1(x) & \text{con probabilidad } \gamma f_1(x)/(\gamma f_1(x) + (1-\gamma)f_2(x)), \\ T_2(x) & \text{con probabilidad } (1-\gamma)f_2(x)/(\gamma f_1(x) + (1-\gamma)f_2(x)). \end{cases}$$

Si X_γ es cualquier vector aleatorio con ley de probabilidad P_γ, tenemos que:

$$\begin{aligned} \nu[T(X_\gamma) \in A] &= \int \nu[T(X_\gamma) \in A | X_\gamma = x] P_\gamma(dx) \\ &= \int I_A[T_1(x)] \frac{\gamma f_1(x)}{\gamma f_1(x) + (1-\gamma)f_2(x)} P_\gamma(dx) \\ &\quad + \int I_A[T_2(x)] \frac{(1-\gamma)f_2(x)}{\gamma f_1(x) + (1-\gamma)f_2(x)} P_\gamma(dx) \\ &= \gamma \int I_A[T_1(x)] f_1(x) dx + (1-\gamma) \int I_A[T_2(x)] f_2(x) dx \\ &= \gamma \nu[T_1(X_1) \in A] + (1-\gamma)\nu[T_2(X_2) \in A] \\ &= \gamma Q_1(A) + (1-\gamma)Q_2(A) = Q_\gamma(A). \end{aligned}$$

Por tanto, T transporta P_γ en Q_γ. De aquí, utilizando el mismo argumento, tenemos:

$$\begin{aligned} \mathcal{W}_2^2(P_\gamma, Q_\gamma) &\leq \int \|X_\gamma - T(X_\gamma)\|^2 \, d\nu \\ &= \gamma \int \|X_1 - T_1(X_1)\|^2 \, d\nu + (1-\gamma) \int \|X_2 - T_2(X_2)\|^2 \, d\nu \\ &= \gamma \mathcal{W}_2^2(P_1, Q_1) + (1-\gamma)\mathcal{W}_2^2(P_2, Q_2). \end{aligned}$$

Esto prueba que $\mathcal{W}_2^2(P_\gamma, Q_\gamma) < \gamma \mathcal{W}_2^2(P_1, Q_1) + (1-\gamma)\mathcal{W}_2^2(P_2, Q_2)$ a menos que T sea una función de transporte óptimo para (P_γ, Q_γ). Pero del apartado (a) de la Proposición 2.11 se deduce que una función aleatoria no puede ser una función de transporte óptimo, y

por tanto, T debe ser no aleatoria, lo que nos lleva a que

$$T(x) = \begin{cases} T_1(x) & \text{si } x \in \text{Sop}(P_1) - \text{Sop}(P_2), \\ T_1(x) & (= T_2(x)) \text{ si } x \in \text{Sop}(P_1) \cap \text{Sop}(P_2), \\ T_2(x) & \text{si } x \in \text{Sop}(P_2) - \text{Sop}(P_1). \end{cases}$$

Finalmente, esto contradiría nuestra hipótesis de partida ya que significaría que T es una función de transporte óptimo común para (P_1, Q_1) y (P_2, Q_2). Y el resultado queda probado.

■

Tomando $P_1 = P_2$ en el anterior teorema obtenemos el siguiente corolario que muestra la convexidad estricta de la función $\mathcal{W}_2^2(P, \cdot)$ siempre que P sea absolutamente continua.

Corolario 3.25. *Sean P, Q_1, Q_2, medidas de probabilidad de $\mathcal{P}_2(\mathbb{R}^k)$ tal que $P \ll \ell^k$. Si $Q_1 \neq Q_2$, entonces, para cada γ en $(0, 1)$,*

$$\mathcal{W}_2^2(P, \gamma Q_1 + (1-\gamma)Q_2) < \gamma \mathcal{W}_2^2(P, Q_1) + (1-\gamma)\mathcal{W}_2^2(P, Q_2).$$

Este corolario nos proporciona de forma inmediata un resultado que garantiza la unicidad del recorte óptimo en los problemas de una muestra.

Teorema 3.26. *Sean P y Q dos medidas de probabilidad en $\mathcal{P}_2(\mathbb{R}^k)$. Supongamos que $P \ll \ell^k$. Entonces, existe una única medida de probabilidad Q_0 tal que*

$$Q_0 = \text{ arg mín } \left\{ \mathcal{W}_2^2(P, R) : R \in \mathcal{R}_\alpha(Q) \right\}.$$

Demostración. La existencia de un minimizador está garantizada por el Teorema 3.19, y la unicidad por ser la minimización de una función estrictamente convexa (Corolario 3.25) en un compacto.

■

Para probar la unicidad de los recortes mejores aproximantes cuando recortamos en las dos distribuciones necesitamos alguna notación adicional así como algún resultado previo. Dado $v_0 \in \mathbb{R}^k$ tal que $\|v_0\| = 1$, llamaremos H_0 al hiperplano ortogonal a v_0. La proyección ortogonal sobre H_0 será denotada como π_0 y para cada $y \in \mathbb{R}^k$, definimos $r_y := \langle y - \pi_0(y), v_0 \rangle$. Dado un conjunto medible $B \subset \mathbb{R}^k$, y $z \in H_0$, definimos también

$$B_z := \{y \in B : \pi_0(y) = z\}, \text{ y } z_{v_0} := \{r_y : y \in B_z\}.$$

Dada una distribución de probabilidades P en $\mathbb{R}^k$, denotaremos por P° a la distribución marginal de P en H_0 y con P_z a la distribución condicional regular de P dado z, donde $z \in H_0$. Esta probabilidad condicional induce de forma obvia una probabilidad en la recta real a través de la isometría $\mathcal{I}_z$ entre $(\mathbb{R}^k)_z$ y $\mathbb{R}$, dada por $y \to r_y$. A esta medida de probabilidad la llamaremos λ_z y su función de distribución (resp. función cuantil) será $F(x|z)$ (resp. $q_z(t)$). En el siguiente lema reflejamos la medibilidad conjunta de estas dos funciones.

Lema 3.27. *Las funciones $(x,z) \mapsto F(x|z)$ y $(t,z) \mapsto q_z(t)$ son conjuntamente medibles en sus argumentos.*

Demostración. Debemos tener cuenta que si $F(x,y)$ es la función de distribución conjunta en $\mathbb{R}\times\mathbb{R}^{k-1}$ y $G(z)$ es la marginal en $\mathbb{R}^{k-1}$, estas funciones son medibles (para medidas de probabilidad con soporte finito es obvio y la generalización se lleva a cabo con argumentos estándar). Por otra parte, consideremos las medidas η_x y μ asociadas respectivamente a las funciones crecientes $F(x,\cdot)$ y $G(\cdot)$. Como consecuencia del Teorema de Diferenciación para Medidas de Radon (ver por ejemplo las Secciones 1.6.2 y 1.7.1 en Evans y Gariepy, 1992), si consideramos para cada $z = (z_1,...,z_{k-1}) \in \mathbb{R}^{k-1}$, la sucesión de rectángulos $A_n(z) := \{(y_1,...,y_{k-1}) : z_i - \frac{1}{n} < y_i \leq z_i + \frac{1}{n},\ i = 1,...,k-1\}$, tenemos la siguiente convergencia casi seguro,

$$F(x|z) = \lim_{n\to\infty} \frac{\eta_x(A_n(z))}{\mu(A_n(z))},$$

que proporciona la medibilidad.

Finalmente, la medibilidad de $q_z(t)$ se obtiene de la siguiente propiedad de la función cuantil: $x \leq q_z(t)$ si y sólo si $F(x|z) \leq t$.

■

El siguiente resultado proporciona una propiedad muy interesante y útil cuando se recorta en ambas distribuciones. De acuerdo con este resultado, las funciones de recorte asociadas a los mejores aproximantes son básicamente indicadores de conjuntos con, quizás, la excepción de puntos que permanecen fijos en el transporte de la masa de probabilidad. En particular, es imposible que existan puntos parcialmente recortados en $\mathrm{Sop}(P)-\mathrm{Sop}(Q)$.

Teorema 3.28. *Sea $\alpha > 0$, y sean $P, Q \in \mathcal{P}(\mathbb{R}^k)$. Supongamos que $P \ll \ell^k$. Si $P_1 \in \mathcal{R}_\alpha(P)$ y $Q_1 \in \mathcal{R}_\alpha(Q)$ son tales que*

$$\mathcal{W}_2^2(P_1, Q_1) = \min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2^2(R_1, R_2) > 0,$$

y T es una función de transporte óptimo para (P_1, Q_1), entonces $T(x) = x$ P-c.s. en el conjunto $\mathcal{A} := \{x \in \mathbb{R}^k : a_1(x) \in (0,1)\}$, donde $a_1 := (1-\alpha)f_1$ y f_1 es la función de densidad de P_1 con respecto a P.

Demostración. Razonemos por reducción al absurdo y supongamos que $P(\mathcal{A} \cap \{x \in \mathbb{R}^k : \|T(x) - x\| > 0\}) > 0$. Veamos que entonces podemos construir una medida $P^* \in \mathcal{R}_\alpha(P)$ de tal forma que (P^*, Q_1) mejora a (P_1, Q_1).

Sea $\hat{P}$ la distribución condicional de P dado el anterior conjunto de probabilidad positiva.

Del apartado (d) de la Proposición 2.11 tenemos que T es continua en casi todo punto del soporte. Si x_0 es un punto del soporte de $\hat{P}$ en el cual T es continua, entonces, para cada $\epsilon > 0$ existe $\delta > 0$ tal que $T(B(x_0, \delta)) \subset B(T(x_0), \epsilon)$. Llamemos $A = B(x_0, \delta) \cap \mathcal{A}$.

Sea $v_0 = (T(x_0) - x_0)/\|T(x_0) - x_0\|$ y H_0 el hiperplano ortogonal a v_0 que contiene a x_0. Con la notación introducida previamente, tomando ϵ suficientemente pequeño, podemos suponer que, $m := \text{ínf}_{y \in B(T(x_0), \epsilon)} r_y$ es mayor que $M := \sup_{y \in B(x_0, \delta)} r_y$. Por tanto,

$$\|T(y) - \pi_0[T(y)]\| > r_y, \text{ para cada } y \in A. \tag{3.22}$$

Por otra parte, tenemos que

$$P[A] = \int_{H_0} P_z(A_z) P^\circ(dz) = \int_{H_0} \lambda_z(z_{v_0}) P^\circ(dz). \tag{3.23}$$

Puesto que x_0 pertenece al soporte de $\hat{P}$, tendremos que $P[A] > 0$, y así

$$P^\circ\{z \in H_0 : \lambda_z(z_{v_0}) > 0\} > 0. \tag{3.24}$$

Sea $z \in H_0$ tal que $\lambda_z(z_{v_0}) > 0$. Si $y_1, y_2 \in A_z$ satisfacen que $r_{y_1} < r_{y_2}$, la ortogonalidad entre $(\pi_0(y) - x_0)$ y $(y - \pi_0(y))$ para cada $y \in \mathbb{R}^k$ y (3.22) nos llevan a que

$$\begin{aligned} \|y_1 - T(y_1)\|^2 &= \|T(y_1) - \pi_0[T(y_1)] + \pi_0(y_1) - y_1 + \pi_0(T(y_1)) - \pi_0(y_1)\|^2 \\ &= \left(r_{T(y_1)} - r_{y_1}\right)^2 + \|\pi_0[T(y_1)] - z\|^2 \\ &> \left(r_{T(y_1)} - r_{y_2}\right)^2 + \|\pi_0[T(y_1)] - \pi_0(y_2)\|^2 \\ &= \|y_2 - T(y_1)\|^2. \end{aligned} \tag{3.25}$$

Ahora, consideramos la partición del conjunto $A = A^- \cup A^+$ dada por

$$\begin{aligned} A^- &:= \{y \in A : F(r_y|\pi_0(y)) \leq 1/2\}, \text{ y} \\ A^+ &:= \{y \in A : F(r_y|\pi_0(y)) > 1/2\}. \end{aligned}$$

Del Lema 3.27 tenemos que estos conjuntos son medibles. Para casi todo $z \in H_0$ que satisfaga $\lambda_z(z_{v_0}) > 0$, dichos conjuntos definen un valor R_z, tal que los conjuntos

$$\begin{aligned} A_z^- &:= \{y \in A_z : r_y < R_z\}, & A_z^+ &:= \{y \in A_z : r_y > R_z\}, \\ z_{v_0}^- &:= \{r_y : y \in A_z^-\}, & z_{v_0}^+ &:= \{r_y : y \in A_z^+\}, \end{aligned}$$

verifican $\lambda_z[z_{v_0}^-] = \lambda_z[z_{v_0}^+] > 0$. Sean λ_z^- y λ_z^+ las probabilidades obtenidas a partir de λ_z condicionando a los conjuntos $z_{v_0}^-$ y $z_{v_0}^+$ respectivamente, y sean $F^-(x|z)$ y $F^+(x|z)$ (resp. $q_z^-(t)$ y $q_z^+(t)$) sus correspondientes funciones de distribución (resp. funciones cuantiles). Entonces, utilizando la isometría $\mathcal{I}_z$ y la forma de obtener funciones de transporte óptimo en la recta real, la función $\Gamma : A^- \mapsto A^+$ definida como

$$\Gamma(y) = \mathcal{I}_{\pi_0(y)}^{-1}\left[q_{\pi_0(y)}^+\left[F^-\left(r_y\ |\pi_0(y)\right)\right]\right],$$

es una función de transporte óptimo entre P_z^- y P_z^+ para casi todo punto $z \in H_0$ que satisfaga $P_z(z_{v_0}) > 0$. Para finalizar la construcción, consideremos la función $a^* : \mathbb{R}^k \mapsto \mathbb{R}$ definida como sigue:

$$a^*(y) = \begin{cases} a_1(y) & \text{si } y \notin A, \\ a_1(y) - \text{mín}\{1 - a_1[\Gamma(y)], a_1(y)\} & \text{si } y \in A^-, \\ a_1(y) + \text{mín}\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\} & \text{si } y \in A^+. \end{cases}$$

Desde este punto, la demostración se desarrolla en tres pasos:

Paso 1. *$f^* := a^*/(1-\alpha)$ es una función de densidad con respecto a P que define una medida de probabilidad $P^* \in \mathcal{R}_\alpha(P)$.*

Obviamente $a^*(\mathbb{R}^k) \subset [0,1]$. Por otra parte,

$$\begin{aligned} \int_{\mathbb{R}^k} a^*(y)P(dy) &= \int_{\mathbb{R}^k} a_1(y)P(dy) \\ &\quad - \int_{A^-} \text{mín}\left\{1 - a_1[\Gamma(y)], a_1(y)\right\} P(dy) \\ &\quad + \int_{A^+} \text{mín}\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P(dy). \end{aligned} \tag{3.26}$$

Para casi todo $z \in H_0$ que satisfaga $P_z(A_z) > 0$, por construcción, se tiene que la ley de probabilidades de a_1 bajo P_z^+, $P_z^+ \circ a_1^{-1}$, coincide con la ley $P_z^- \circ (a_1(\Gamma))^{-1}$, mientras que $P_z^+ \circ (a_1(\Gamma^{-1}))^{-1} = P_z^- \circ a_1^{-1}$. Por tanto el último término verifica

$$\begin{aligned}
&\int_{A^+} \text{mín}\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P(dy) \\
&\qquad = \int_{H_0} \left(\int_{A_z^+} \text{mín}\left\{1 - a_1(y), a_1[\Gamma^{-1}(y)]\right\} P_z(dy) \right) P^\circ(dz) \\
&\qquad = \int_{H_0} \left(\int_{A_z^-} \text{mín}\left\{1 - a_1(\Gamma(y)), a_1(y)\right\} P_z(dy) \right) P^\circ(dz) \\
&\qquad = \int_{A^-} \text{mín}\left\{1 - a_1[\Gamma(y)], a_1(y)\right\} P(dy), \qquad (3.27)
\end{aligned}$$

que, junto con (3.26) lleva a que $\int_{\mathbb{R}^k} a^*(y)P(dy) = \int_{\mathbb{R}^k} a_1(y)P(dy) = 1 - \alpha$, lo cual prueba este paso.

Paso 2. *Existe una función aleatoria, T^*, que transporta P^* en Q_1.*

Consideremos la función aleatoria T^* definida como $T^*(y) = T(y)$ en el complementario de A^+ y, tal que para cada $y \in A^+$, toma los valores $T(y)$ ó $T[\Gamma(y)]$ con probabilidades $f_1(y)/f^*(y)$ $(= a_1(y)/a^*(y))$ y $[f^*(y) - f_1(y)]/f^*(y)$ $(= [a^*(y) - a_1(y)]/a^*(y))$ respectivamente. Estos valores son positivos ya que, por construcción, $a^*(y) > a_1(y)$ en A^+.

El argumento para probar que la función T^* transporta P^* en Q_1 es análogo al desarrollado en el Teorema 3.24, teniendo en cuenta que $P_z^+ \circ a_1^{-1} = P_z^- \circ (a_1(\Gamma))^{-1}$.

Paso 3. $\mathcal{W}_2^2(P_1, Q_1) > \mathcal{W}_2^2(P^*, Q_1)$.

De la construcción de T^* y la desigualdad (3.25), tenemos que

$$\begin{aligned}
\mathcal{W}_2^2(P^*, Q_1) &\leq \int_{\mathbb{R}^k} \|y - T^*(y)\|^2 P^*(dy) \\
&= \int_{(A^+)^c} \|y - T(y)\|^2 P^*(dy) \\
&\quad + \int_{A^+} \left(\|y - T(y)\|^2 \frac{f_1(y)}{f^*(y)} + \|y - T[\Gamma^{-1}(y)]\|^2 \frac{f^*(y) - f_1(y)}{f^*(y)} \right) f^*(y) P(dy) \\
&< \int_{(A^- \cup A^+)^c} \|y - T(y)\|^2 f_1(y) P(dy) + \int_{A^-} \|y - T(y)\|^2 f^*(y) P(dy) \\
&\quad + \int_{A^+} \left(\|y - T(y)\|^2 f_1(y) + \|\Gamma^{-1}(y) - T[\Gamma^{-1}(y)]\|^2 (f^*(y) - f_1(y)) \right) P(dy).
\end{aligned}$$

Además, de la definición de la función Γ, y utilizando la relación $P_z^+ \circ (a_1(\Gamma^{-1}))^{-1} =$

$P_z^- \circ (a_1)^{-1}$, obtenemos que

$$\begin{aligned}\int_{A^+} \|\Gamma^{-1}(y) - T[\Gamma^{-1}(y)]\|^2 (f^*(y) - f_1(y)) P(dy) \\ = \quad -\int_{A^-} \|y - T(y)\|^2 (f^*(y) - f_1(y)) P(dy),\end{aligned}$$

que, por la construcción de f^*, nos da

$$\mathcal{W}_2^2(P^*, Q_1) < \mathcal{W}_2^2(P_1, Q_1),$$

en contradicción con la optimalidad del par (P_1, Q_1).

■

Teorema 3.29. *Sea $\alpha > 0$ y $P, Q \in \mathcal{P}_2(\mathbb{R}^k)$, tal que $P \ll \ell^k$. Si $P_1 \in \mathcal{R}_\alpha(P)$ y $Q_1 \in \mathcal{R}_\alpha(Q)$ son tales que,*

$$(P_1, Q_1) = \text{ arg mín } \left\{ \mathcal{W}_2^2(R_1, R_2) : R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q) \right\}, \tag{3.28}$$

y $\mathcal{W}_2^2(P_1, Q_1) > 0$, entonces, el par (P_1, Q_1) es único.

Demostración. Supongamos que (P_1, Q_1) y (P_2, Q_2) son dos pares diferentes de distribuciones que verifican (3.28), y sea $a_i := (1-\alpha) f_i$, $i = 1, 2$, donde f_i es la función de densidad de P_i con respecto a P. Usando las combinaciones convexas $P_{\delta_i} = \delta_i P_1 + (1-\delta_i) P_2$ y $Q_{\delta_i} = \delta_i Q_1 + (1-\delta_i) Q_2$, $i = 1, 2$, con $\delta_1 \neq \delta_2$, a partir del Teorema 3.24, podemos suponer que P_1 y P_2 tienen soporte común, y que T es una función de transporte óptimo común a ambas soluciones. Es decir que $Q_i = P_i \circ T^{-1}$, para $i = 1, 2$. Además, en el conjunto $\{a_1 \neq a_2\}$ se verifica que $0 < a_1(x) < 1$, por lo que del Teorema 3.28 se deduce que $T(x) = x$ en este conjunto. Pero entonces es fácil probar que existen conjuntos $A \subset \{a_1 = a_2\}$ y $B \subset \{a_1 < a_2\}$ tales que, definiendo

$$a^*(x) = \begin{cases} 0 & \text{si } x \in A, \\ a_2(x) & \text{si } x \in B, \\ a_1(x) & \text{si } x \notin A \cup B, \end{cases}$$

se tiene que, $f^* := a^*/(1-\alpha)$ es una función de densidad de una medida de probabilidad,

P^*, de $\mathcal{R}_\alpha(P)$, $Q^* := P^* \circ T^{-1}$ pertenece a $\mathcal{R}_\alpha(Q)$ y:

$$\begin{aligned}\mathcal{W}_2^2(P^*, Q^*) &= \int_{\mathbb{R}^k} \|x - T(x)\|^2 f^*(x) P(dx) \\ &= \int_{\{a_1=a_2\}-A} \|x - T(x)\|^2 f_1(x) P(dx) \\ &< \int_{\{a_1=a_2\}} \|x - T(x)\|^2 f_1(x) P(dx) = \mathcal{W}_2^2(P_1, Q_1).\end{aligned}$$

■

El siguiente resultado es muy interesante porque caracteriza el recorte de menor nivel que hace cero la distancia entre los recortes de P y Q, y además garantiza su unicidad.

Teorema 3.30. *Sean $P, Q \in \mathcal{P}_2(\mathbb{R}^k)$. Supongamos que $\mathcal{W}_2(P_\alpha, Q_\alpha) = 0$ para algún $0 < \alpha < 1$, y $P_\alpha \in \mathcal{R}_\alpha(P)$ y $Q_\alpha \in \mathcal{R}_\alpha(Q)$. Si*

$$A = \left\{\alpha' \in [0, \alpha] : \mathcal{W}_2(R_1, R_2) = 0, R_1 \in \mathcal{R}_{\alpha'}(P), R_2 \in \mathcal{R}_{\alpha'}(Q)\right\}, \tag{3.29}$$

entonces existe $\alpha_0 \in [0, \alpha]$ tal que $\alpha_0 = \min A$.

Si $P, Q \ll \ell^k$, y $f(x)$ y $g(x)$ son sus respectivas funciones de densidad, entonces existe una única medida de probabilidad $R_0 \in \mathcal{R}_{\alpha_0}(P) \cap \mathcal{R}_{\alpha_0}(Q)$, cuya función de densidad es

$$m(x) = \frac{1}{1 - \alpha_0} \min\left(f(x), g(x)\right), \tag{3.30}$$

y $\alpha_0 = 1 - \int_{\mathbb{R}^k} \min\left(f(x), g(x)\right) dx$.

Demostración. El inferior de A existe porque A es no vacío y está contenido en $[0, \alpha]$ que es un conjunto compacto. Sea α_0 dicho inferior. Veamos que efectivamente es un mínimo. Para ello tomamos una sucesión $\{\alpha_n\}_n \subset A$ tal que $\{\alpha_n\}_n \downarrow \alpha_0$. Existen por tanto $P_n \in \mathcal{R}_{\alpha_n}(P)$ y $Q_n \in \mathcal{R}_{\alpha_n}(Q)$ tales que $\mathcal{W}_2(P_n, Q_n) = 0$ para cada n. De (a) en la Proposición 3.6 tenemos que $\mathcal{R}_{\alpha_n}(P) \subset \mathcal{R}_\alpha(P)$ y $\mathcal{R}_{\alpha_n}(Q) \subset \mathcal{R}_\alpha(Q)$ para cada n, ya que $\alpha_n \leq \alpha$ para cada n. Ahora, por la compacidad de $\mathcal{R}_\alpha(P)$ y $\mathcal{R}_\alpha(Q)$, tenemos que de cada subsucesión de $\{(P_n, Q_n)\}_n$ podemos extraer una subsucesión convergente (unificando índices): $\{(P_{n_k}, Q_{n_k})\}_{n_k}$ tal que $P_{n_k} \to_w P^*$ y $Q_{n_k} \to_w Q^*$, donde obviamente $P^* \in \mathcal{R}_\alpha(P)$ y $Q^* \in \mathcal{R}_\alpha(Q)$. Pero también, en virtud de la Proposición 3.8, $P^* \in \mathcal{R}_{\alpha_0}(P)$ y $Q^* \in \mathcal{R}_{\alpha_0}(Q)$. De la desigualdad triangular para $\mathcal{W}_2$ tenemos que $\mathcal{W}_2(P_{n_k}, Q_{n_k}) \to \mathcal{W}_2(P^*, Q^*)$. Pero como $\mathcal{W}_2(P_{n_k}, Q_{n_k}) = 0$ para cada n_k, sólo puede ser $\mathcal{W}_2(P^*, Q^*) = 0$ y se tiene la primera parte.

Para la segunda parte, tenemos que si R es una medida de probabilidad con función de densidad r, entonces, $R \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$ para algún α, si y sólo si (de la definición de recorte),

$$r(x) = \frac{dR}{d\ell^k}(x) = \frac{dR}{dP}(x)\frac{dP}{d\ell^k}(x) \leq \frac{1}{1-\alpha} f(x).$$

Y de igual forma con g. Por lo que, R será un recorte común, si y sólo si,

$$r(x) \leq \frac{1}{1-\alpha} \min\left(f(x), g(x)\right). \tag{3.31}$$

Además, integrando en la anterior expresión, α debe verificar

$$\alpha \geq 1 - \int_{\mathbb{R}^k} \min\left(f(x), g(x)\right) dx.$$

El resultado se obtiene entonces de forma obvia, tomando $m(x) = \frac{1}{1-\alpha_0} \min\left(f(x), g(x)\right)$. La unicidad también es inmediata, pues si existiese $R_1 \in \mathcal{R}_{\alpha_0}(P) \cap \mathcal{R}_{\alpha_0}(Q)$ con $R_1 \neq R_0$, y r_1 fuese su densidad. Entonces $r_1(x)$ debería verificar (3.31) para $\alpha = \alpha_0$, y para que $R_1 \neq R_0$ debería ser que $r_1(x) < m(x)$, en un conjunto de medida de Lebesgue positiva. Pero entonces r_1 no sería una densidad porque $\int_{\mathbb{R}^k} r(x)dx < \int_{\mathbb{R}^k} m(x)dx = 1$. Lo que concluye la demostración. ■

A partir de este resultado es claro que si $R \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$, entonces, fuera del conjunto $\mathrm{Sop}(P) \cap \mathrm{Sop}(Q)$, se recorta toda la masa de probabilidad. Además, conviene resaltar que si la medida de Lebesgue de $\mathrm{Sop}(P) \cap \mathrm{Sop}(Q)$ es estrictamente positiva, y $P \neq Q$, entonces existe un valor de $\alpha \in (0,1)$ tal que $\mathcal{W}_2(P_\alpha, Q_\alpha) = 0$.

3.4. Algoritmos.

En esta sección se abordan los diferentes problemas que se plantean dependiendo de si tenemos una o dos muestras, de la métrica utilizada, y del tipo de recorte (entero o general). En algunos casos se llega a un problema de optimización lineal que puede resolverse utilizando el método simplex (o alternativamente, un método de punto interior), en otros casos se llega a un problema de programación cuadrática o convexa en general. Finalmente, para algún caso particular, relacionado con la métrica $\mathcal{W}_\infty$, se da un algoritmo específicamente diseñado para resolver el problema.

3.4.1. Problemas de una muestra.

Recorte general utilizando la métrica $\mathcal{W}_2$.

En este apartado nos ceñiremos a la recta real en la que disponemos de la representación cuantil de la distancia $\mathcal{W}_2$ de Wasserstein (ver Lema 2.2). Sean pues P y Q dos medidas de probabilidad de $\mathcal{P}_2(\mathbb{R},\beta)$, y sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple tal que $\mathcal{L}(X_i) = Q$. Si $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ es la medida empírica asociada a la muestra, el problema que se pretende resolver es el de encontrar

$$Q_{n,\alpha} = \text{ arg mín } \left\{\mathcal{W}_2^2(P,R) : R \in \mathcal{R}_\alpha(Q_n)\right\}, \tag{3.32}$$

donde, de acuerdo con la Definición 3.2, si $X_{(1)}, \ldots, X_{(n)}$ representa la muestra ordenada,

$$\mathcal{R}_\alpha(Q_n) = \left\{ \sum_{i=1}^n b_i \delta_{X_{(i)}} : \sum_{i=1}^n b_i = 1,\ 0 \le b_i \le \frac{1}{n(1-\alpha)} \right\}.$$

Tomemos ahora $R = \sum_{i=1}^n b_i \delta_{X_{(i)}}$ un elemento genérico de $\mathcal{R}_\alpha(Q_n)$. Si denotamos por G^{-1} su función cuantil entonces $G^{-1}(y) = X_{(i)}$ para $c_{i-1} < y \le c_i$, donde $c_i = b_1 + \cdots + b_i$, $i = 1, \ldots, n-1$, $c_0 = 0$ y $c_n = 1$.

Si F^{-1} es la función cuantil asociada a la medida de probabilidad P y operamos con $\mathcal{W}_2^2(P,R)$ tendremos

$$\begin{aligned}
\mathcal{W}_2^2(P,R) &= \int_0^1 (G^{-1}(y) - F^{-1}(y))^2 dy \\
&= \int_0^1 F^{-1}(y)^2 dy + \int_0^1 (G^{-1}(y)^2 - 2G^{-1}(y)F^{-1}(y)) dy \\
&= \sigma^2(F) + \mu^2(F) + \sum_{i=1}^n \left(X_{(i)}^2 (c_i - c_{i-1}) - 2X_{(i)} \left(h(c_i) - h(c_{i-1})\right)\right),
\end{aligned}$$

donde $h(c) = \int_0^c F^{-1}(y)dy$, y $\mu(F)$ y $\sigma^2(F)$ designan la media y la varianza de P respectivamente. Reorganizando los términos y teniendo en cuenta que $h(1) = \mu(F)$ obtenemos

$$\mathcal{W}_2^2(P,R) = \sigma^2(F) + (X_{(n)} - \mu(F))^2 - \sum_{i=1}^{n-1} g_i(c_i), \tag{3.33}$$

donde $g_i(c) = 2(X_{(i+1)} - X_{(i)}) \left(c\frac{(X_{(i)} + X_{(i+1)})}{2} - h(c)\right)$.

En el caso $\alpha = 1$ no existe ninguna restricción sobre los coeficientes c_i. Entonces el problema de minimizar la expresión (3.33) es un problema de variables separadas y basta con maximizar cada una de las funciones g_i por separado. La derivada de estas funciones es

$$g_i'(c) = 2(X_{(i+1)} - X_{(i)}) \left(\frac{(X_{(i)} + X_{(i+1)})}{2} - F^{-1}(c) \right), \tag{3.34}$$

por lo que cada una de estas funciones es cóncava y el óptimo se alcanza en el punto que anula la derivada

$$\tilde{c}_i = F\left(\frac{X_{(i)} + X_{(i+1)}}{2}\right). \tag{3.35}$$

Esto nos da un primer resultado para el caso extremo en el que $\alpha = 1$, es decir, con recorte total.

Proposición 3.31. *Sea $P \in \mathcal{P}_2(\mathbb{R})$ y sean F y f sus funciones de distribución y densidad respectivamente. Sea $X_1, \ldots, X_n$ una m.a.s. cuya medida empírica asociada es $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_{(i)}}$. Si $\alpha = 1$, la solución del problema de minimización*

$$Q_{n,1} = \arg\min\left\{\mathcal{W}_2^2(P,R) : R \in \mathcal{R}_1(Q_n)\right\},$$

viene dada por

$$\tilde{Q}_{n,1} = \sum_{i=1}^n \tilde{b}_i \delta_{X_{(i)}},$$

donde $\tilde{b}_i = \tilde{c}_i - \tilde{c}_{i-1}$ y $\tilde{c}_i = F\left(\frac{X_{(i)}+X_{(i+1)}}{2}\right)$. Además la expresión de la distancia en ese caso viene dada por,

$$\begin{aligned}\mathcal{W}_2^2(P,\tilde{Q}_{n,1}) &= \int_{-\infty}^{\frac{X_{(2)}+X_{(1)}}{2}} (t - X_{(1)})^2 f(t)dt + \int_{\frac{X_{(n)}+X_{(n-1)}}{2}}^{\infty} (t - X_{(n)})^2 f(t)dt \\ &\quad + \sum_{i=2}^{n-1} \int_{\frac{X_{(i)}+X_{(i-1)}}{2}}^{\frac{X_{(i+1)}+X_{(i)}}{2}} (t - X_{(i)})^2 f(t)dt. \end{aligned} \tag{3.36}$$

Demostración. Que el minimizador es el dado es inmediato por lo visto hasta la expresión (3.35).

La expresión para la distancia se obtiene de forma inmediata después del cambio de variable $t = F^{-1}(y)$ en cada una de las integrales, y teniendo en cuenta que $\tilde{c}_0 = 1 - \tilde{c}_n = 0$,

$$\begin{aligned}\mathcal{W}_2^2(P,\tilde{Q}_{n,1}) &= \sum_{i=1}^n \int_{\tilde{c}_{i-1}}^{\tilde{c}_i} (X_{(i)} - F^{-1}(y))^2 dy \\ &= \int_{-\infty}^{\frac{X_{(2)}+X_{(1)}}{2}} (t - X_{(1)})^2 f(t)dt + \int_{\frac{X_{(n)}+X_{(n-1)}}{2}}^{\infty} (t - X_{(n)})^2 f(t)dt \\ &\quad + \sum_{i=2}^{n-1} \int_{\frac{X_{(i)}+X_{(i-1)}}{2}}^{\frac{X_{(i+1)}+X_{(i)}}{2}} (t - X_{(i)})^2 f(t)dt. \end{aligned}$$

■

Para que el problema (3.32) sea un problema de programación lineal deberíamos tener funciones $g_i(c)$ lineales. Por lo tanto, h también debe ser lineal y como $h(0) = 0$, ha de ser $h(c) = kc$, $\forall c \in (0,1)$, $k \in \mathbb{R}$. Derivando, es inmediato ver que esto ocurre solamente si la medida asociada a F es $P = \delta_k$ (las medidas concentradas en un punto).

Otro caso que tiene especial interés es cuando la distribución de referencia es la uniforme en el intervalo unitario.

Proposición 3.32. *Si $P = U[0,1]$, y $X_1, \ldots, X_n$ es una m.a.s. cuya medida empírica asociada es $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_{(i)}}$. El problema dado por la expresión (3.32) es un problema de programación cuadrática con variables separadas y restricciones lineales.*

Demostración. Es inmediato pues si $P = U[0,1]$, entonces $h(c) = \frac{c^2}{2}$ y las funciones $g_i(c) = (X_{(i+1)} - X_{(i)})(c(X_{(i)} + X_{(i+1)}) - c^2)$ son cuadráticas. Entonces, de acuerdo con lo visto en la expresión (3.33), encontrar el minimizador de (3.32) equivale a encontrar

$$(c_1, \ldots, c_{n-1}) = \text{ arg máx } \left\{ \sum_{i=1}^{n} g_i(c_i) : 0 \leq c_i - c_{i-1} \leq \frac{1}{n(1-\alpha)}; i = 1, \ldots, n \right\}.$$

Lo que finaliza la prueba. ■

En el caso de estar ante un problema de programación cuadrática con restricciones lineales como el planteado en la proposición anterior, la condiciones de Kuhn-Tucker son lineales. Por ello la solución al problema de optimización se puede encontrar escribiendo el problema como un problema lineal complementario que se resuelve con algoritmos de pivoteo específicos como el descrito en la Sección 11.1 de Bazaraa et al. (1993).

Estos problemas se pueden resolver en la mayor parte de *solvers* comerciales como CPLEX ó XPRESS, y también en no comerciales como MINOS. El paquete estadístico R también incorpora una librería llamada *quadprog* para llevar a cabo este tipo de optimizaciones. En el Apéndice A aparece una función que hemos programado en R para resolver este problema. Aparecen también los ficheros necesarios para resolver el mismo problema utilizando el lenguaje de modelización AMPL junto con el *solver* MINOS.

En el caso más general el problema (3.32) es un problema de programación convexa (pues las funciones g_i son cóncavas), de variables separadas, con restricciones lineales y en el que si P es absolutamente continua sabemos que hay unicidad (ver Teorema 3.26). En esta

situación el hecho de tener además una derivada explícita si tenemos F^{-1} (ver expresión (3.34)), hace que el problema tenga las mejores condiciones para ser resuelto con *solvers* que incluyen programación no lineal como MINOS.

Recorte general utilizando el algoritmo para dos muestras.

Gracias a los resultados de consistencia que se obtienen en el Capítulo 5 una alternativa para resolver este problema consiste en reemplazar la distribución de referencia poblacional por una realización de una muestra $Y_1, \ldots, Y_m$ de variables (o vectores) aleatorias i.i.d. de tamaño m suficientemente elevado y resolver el problema de forma aproximada utilizando el algoritmo para dos muestras que se da en la siguiente sección. Este procedimiento es válido en general en $\mathbb{R}^k$ y para la métrica $\mathcal{W}_2$ (condiciones bajo las que se verá la consistencia del recorte unilateral). También es viable para las métricas $\mathcal{W}_p$ ($p \neq 2$), pero su validez dependerá de que haya consistencia, propiedad que cabe esperar aunque por lo que nosotros sabemos, aún no ha sido probada.

Recorte entero utilizando la métrica $\mathcal{W}_2$.

Sean P y Q dos medidas de probabilidad de $\mathcal{P}_2(\mathbb{R}, \beta)$, y sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple tal que $\mathcal{L}(X_i) = Q$. Si $Q_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_{(i)}}$ es la medida empírica asociada a una realización $\{x_1, x_2, \ldots, x_n\}$ de la muestra, el problema que vamos a resolver en este apartado es el de encontrar el recorte entero, tal y como se definió en la Sección 3.1, de Q_n que mejor aproxima P, es decir, si $k \leq n$, buscamos

$$Q_{n,k} = \text{ arg mín } \left\{\mathcal{W}_2^2(P, R) : R \in \mathcal{T}_k(Q_n)\right\}, \tag{3.37}$$

donde

$$\mathcal{T}_k(Q_n) = \left\{ \frac{1}{n-k}\sum_{i=1}^n b_i \delta_{x_{(i)}}, \text{ con } b_i = 0 \text{ ó } 1 \text{ y } \sum_{i=1}^n b_i = n-k \right\}.$$

Veamos en primer lugar el caso particular en el que $P = U[0,1]$. Vamos a comenzar por el siguiente problema: Sea $\mathcal{D}_m$ el conjunto de medidas de probabilidad discretas y uniformes sobre m puntos del intervalo $[0,1]$. Cada una de estas medidas tiene asociado un vector $(y_1, y_2, \ldots, y_m)$ con $0 \leq y_1 \leq y_2 \leq \cdots \leq y_m \leq 1$, de modo que

$$\mathcal{D}_m = \left\{ Q : Q = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}, \ 0 \leq y_1 \leq y_2 \leq \cdots \leq y_m \leq 1 \right\}.$$

Sea Q un elemento genérico de $\mathcal{D}_m$ entonces,

$$\mathcal{W}_2^2(Q, U[0,1]) = \int_0^1 (t - F_Q^{-1}(t))^2 dt = \sum_{i=1}^{m} \int_{\frac{i-1}{m}}^{\frac{i}{m}} (t - y_i)^2 dt. \tag{3.38}$$

Si $\tilde{P}_m = \text{ arg mín } \left\{\mathcal{W}_2^2(Q, U[0,1]) : Q \in \mathcal{D}_m\right\}$, es claro a partir de la expresión (3.38) y de las propiedades de la media de una variable aleatoria que el mínimo se alcanza para los valores de y_i que son los puntos medios de los intervalos $\left[\frac{i-1}{m},\ \frac{i}{m}\right]$ y por tanto,

$$\tilde{P}_m = \frac{1}{m} \sum_{i=1}^{m} \delta_{\frac{i-0.5}{m}}. \tag{3.39}$$

En estas condiciones tenemos el siguiente resultado,

Proposición 3.33. *Sea* $\tilde{P}_m = \frac{1}{m}\sum_{i=1}^{m} \delta_{\frac{i-0.5}{m}}$ *y* $\mathcal{A} \subset \mathcal{D}_m$ *un subconjunto finito. Entonces,*

$$\text{arg mín } \left\{\mathcal{W}_2^2(Q, U[0,1]) : Q \in \mathcal{A} \subset \mathcal{D}_m\right\} = \text{ arg mín } \left\{\mathcal{W}_2^2(Q, \tilde{P}_m) : Q \in \mathcal{A} \subset \mathcal{D}_m\right\}$$

y además $\mathcal{W}_2^2(Q, U[0,1]) = \mathcal{W}_2^2(Q, \tilde{P}_m) + \frac{1}{12m^2} \quad \forall Q \in \mathcal{D}_m$.

Demostración. Si $Q \in \mathcal{D}_m$,

$$\begin{aligned}
\mathcal{W}_2^2(Q, U[0,1]) &= \sum_{i=1}^{m} \int_{\frac{i-1}{m}}^{\frac{i}{m}} (t - y_i)^2 dt = \sum_{i=1}^{m} \frac{1}{3}\left[\left(\frac{i}{m} - y_i\right)^3 - \left(\frac{i-1}{m} - y_i\right)^3\right] \\
&= \sum_{i=1}^{m} \left[\frac{1}{3m^3} + \frac{1}{m}\left(\frac{i}{m} - y_i\right)\left(\frac{i-1}{m} - y_i\right)\right] \\
&= \sum_{i=1}^{m} \left[\frac{1}{3m^3} + \frac{1}{m}\left(\frac{i-0.5}{m} - y_i + \frac{0.5}{m}\right)\left(\frac{i-0.5}{m} - y_i - \frac{0.5}{m}\right)\right] \\
&= \sum_{i=1}^{m} \left[\frac{1}{3m^3} + \frac{1}{m}\left(\frac{i-0.5}{m} - y_i\right)^2 - \frac{1}{m}\left(\frac{0.5}{m}\right)^2\right] \\
&= \frac{1}{12m^2} + \frac{1}{m}\sum_{i=1}^{m}\left(\frac{i-0.5}{m} - y_i\right)^2 = \frac{1}{12m^2} + \mathcal{W}_2^2(Q, \tilde{P}_m).
\end{aligned}$$

Lo que completa la prueba si tenemos en cuenta que el mínimo se alcanza, pues lo buscamos en un subconjunto finito, $\mathcal{A}$, de medidas de probabilidad. ■

Este resultado nos permite garantizar que buscar el recorte entero de una muestra que mejor aproxima la distribución uniforme en el intervalo unidad es equivalente a buscar el recorte entero mejor aproximante de dicha muestra a una distribución discreta y uniforme, lo que se recoge en la siguiente proposición,

Proposición 3.34. *Sean $P = U[0,1]$ y $Q_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_{(i)}}$ la medida empírica asociada a una realización de una muestra aleatoria simple $X_1, X_2, \ldots, X_n$. Entonces, si $k \leq n$,*

$$\tilde{Q}_{n,k} = \text{ arg mín } \left\{\mathcal{W}_2^2(P,R) : R \in \mathcal{T}_k(Q_n)\right\} = \text{ arg mín } \left\{\mathcal{W}_2^2(\tilde{P}_{n-k},R) : R \in \mathcal{T}_k(Q_n)\right\}$$

donde $\tilde{P}_{n-k}$ es como se ha obtenido en (3.39), y además

$$\mathcal{W}_2^2(U[0,1],\tilde{Q}_{n,k}) = \frac{1}{12(n-k)^2} + \mathcal{W}_2^2(\tilde{P}_{n-k},\tilde{Q}_{n,k}).$$

Demostración. Es inmediato aplicando la Proposición 3.33 cuando $m = n-k$ y $\mathcal{A} = \mathcal{T}_k(Q_n)$. Es obvio que $\mathcal{T}_k(Q_n) \subset \mathcal{D}_{n-k}$ y que $\mathcal{T}_k(Q_n)$ es finito ya que su cardinal es $\binom{n}{k}$ (formas de eliminar k puntos de n). ■

Hemos pasado de un problema de comparación de una distribución continua contra una discreta (una medida empírica), a un problema de comparación de dos discretas recortando en una de ellas. En la siguiente sección veremos que este es un problema de programación lineal entera y por tanto abordable con *solvers* que incorporen programación entera como CPLEX ó XPRESS. En el paquete estadístico R disponemos de una librería llamada *lpSolve* que permite resolver problemas de programación lineal en general, y además incorpora una función para programación lineal entera en el caso de que el problema se pueda escribir como un problema de transporte entero. Este es nuestro caso como se verá más adelante.

El siguiente paso es tratar de ver si el resultado obtenido en la Proposición 3.34 es generalizable a otras medidas de probabilidad P y en qué condiciones. El siguiente resultado nos da la clave de dicha generalización.

Teorema 3.35. *Sea P una medida de probabilidad continua de $\mathcal{P}_2(\mathbb{R},\beta)$ y, dado $n \in \mathbb{N}$, sean $-\infty = q_0, q_1, ..., q_n = \infty$ los cuantiles $0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$ de P. Sea P_n la distribución uniforme en los puntos $a_1, ..., a_n$ donde, si X es una v.a. tal que $\mathcal{L}(X) = P$, entonces*

$$a_i = E[X/X \in (q_{i-1}, q_i]],\ i = 1, \ldots, n.$$

Sea $\mathcal{A}$ el conjunto de medidas de probabilidad con soporte finito tal que, si $Q \in \mathcal{A}$ y denotamos por S_Q el soporte de Q, se tiene,

$$Q(s) = \frac{k_s}{n},\ k_s \in \{1, ..., n\},\ s \in S_Q.$$

Entonces, se cumple que

$$\text{arg mín}\left\{\mathcal{W}_2^2(Q,P) : Q \in \mathcal{A}\right\} = \text{ arg mín}\left\{\mathcal{W}_2^2(Q,P_n) : Q \in \mathcal{A}\right\},$$

en el sentido de que, si uno de ellos existe, el otro también, y entonces ambos coinciden.

Demostración. Consideremos el espacio probabilístico $(\mathbb{R}, \beta, P)$. Sea $\mathcal{X}^0$ el conjunto de todas las v.a., X, que satisfacen que existen $x_1, ..., x_n \in \mathbb{R}$ tales que

$$X = \sum_{i=1}^{n} x_i I_{(q_{i-1}, q_i]} \text{ y } x_1 \leq ... \leq x_n.$$

Denotemos por $X_P = \sum_{i=1}^{n} a_i I_{(q_{i-1}, q_i]}$. Entonces la ley de la v.a. X_P es P_n, $X_P \in \mathcal{X}^0$ y por el Lema 2.14 se cumple que

$$\|I_d - X_P\|_2 = \inf_{X \in \mathcal{X}^0} \|I_d - X\|_2. \tag{3.40}$$

Por otro lado, de la primera parte del Lema 2.13, si $X_1, X_2 \in \mathcal{X}^0$, entonces

$$\mathcal{W}_2(P_{X_1}, P_{X_2}) = \|X_1 - X_2\|_2 \text{ y } \mathcal{W}_2(P_{X_1}, P) = \|X_1 - I_d\|_2. \tag{3.41}$$

Sea ahora $\mathcal{X}$ el conjunto de todas las v.a. X tal que existe $X^0 \in \mathcal{X}^0$ que satisface que $P_X = P_{X^0}$. Es claro que $\mathcal{X}$ es un subespacio lineal tal que $\mathcal{X}^0 \subset \mathcal{X}$ y, de la segunda parte del Lema 2.13 y la expresión (3.40), X_P es la proyección de la identidad en $\mathcal{X}$. Así pues, de (3.41) y el Teorema de Pitágoras, tenemos que, si $X \in \mathcal{X}^0$, entonces

$$\mathcal{W}_2^2(P_X, P) = \|X - I_d\|_2^2 = \|X - X_n\|_2^2 + \|X_n - I_d\|_2^2 = \mathcal{W}_2^2(P_X, P_n) + \mathcal{W}_2^2(P_n, P).$$

De aquí, y puesto que $\mathcal{W}_2^2(P_n, P)$ es una constante y $\mathcal{A} \subset \left\{P_X : X \in \mathcal{X}^0\right\}$, tenemos que la distribución de $\mathcal{A}$, si existe alguna, que minimiza el lado izquierdo coincide con la que minimiza el primer sumando del lado derecho. ■

Teorema 3.36. *Sea P una medida de probabilidad continua de $\mathcal{P}_2(\mathbb{R}, \beta)$ y $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{(i)}}$ la medida empírica asociada a una realización de una muestra aleatoria simple $X_1, X_2, \ldots, X_n$. Sea $k \leq n$ y, $-\infty = q_0, q_1, ..., q_{n-k} = \infty$, los cuantiles $0, \frac{1}{n-k}, \frac{2}{n-k}, \ldots, 1$ de P. Entonces,*

$$\tilde{Q}_{n,k} = \text{ arg mín } \left\{\mathcal{W}_2^2(P,R) : R \in \mathcal{T}_k(Q_n)\right\} = \text{ arg mín } \left\{\mathcal{W}_2^2(\tilde{P}_{n-k}, R) : R \in \mathcal{T}_k(Q_n)\right\},$$

donde

$$\tilde{P}_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k}\delta_{a_i}, \quad a_i = E[X/X \in (q_{i-1}, q_i]], \text{y } X \text{ una v.a. tal que } \mathcal{L}(X) = P.$$

Y además,

$$\mathcal{W}_2^2(P, \tilde{Q}_{n,k}) = \mathcal{W}_2^2(P, \tilde{P}_{n-k}) + \mathcal{W}_2^2(\tilde{P}_{n-k}, \tilde{Q}_{n,k}).$$

Demostración. Es inmediato aplicando el Teorema 3.35 cuando $\mathcal{A} = \mathcal{T}_k(Q_n)$. Además el mínimo existe pues, como vimos en el caso uniforme, el conjunto $\mathcal{T}_k(Q_n)$ es finito. ■

Recorte entero utilizando la métrica del supremo.

Sean P y Q dos medidas de probabilidad en $\mathbb{R}$, P continua, y sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria simple tal que $\mathcal{L}(X_i) = Q$. Sea $Q_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_{(i)}}$ la medida empírica asociada a una realización $\{x_1, x_2, \ldots, x_n\}$ de la muestra.

En este apartado vamos a resolver el problema de encontrar el recorte entero de la medida empírica Q_n que mejor se aproxima a la distribución de referencia P, pero utilizando la norma del supremo que denotamos por $\mathcal{W}_\infty$. Es decir, pretendemos encontrar

$$Q_{n,k} = \text{ arg mín } \{\mathcal{W}_\infty(P, R) : R \in \mathcal{T}_k(Q_n)\}. \tag{3.42}$$

Si manejamos probabilidades en $\mathbb{R}$, de igual forma que podemos caracterizar los recortes generales a partir de las funciones de distribución de recortes de la distribución uniforme, mediante la clase $\mathcal{C}_\alpha$ (ver Teorema 3.15), podemos caracterizar los recortes enteros de Q_n, mediante una subclase de $\mathcal{C}_\alpha$, ya que $\mathcal{T}_k(Q_n) \subset \mathcal{R}_{k/n}(Q_n)$.

Es sencillo comprobar que los recortes de tamaño $\alpha = \frac{k}{n}$ de la $U[0,1]$ que generan recortes enteros de una empírica Q_n, son de la forma $P_{n,k} = \frac{1}{1-\alpha}\sum_{i=1}^{n} b_i \delta_{\left[\frac{i-1}{n}, \frac{i}{n}\right]}$ donde $b_i = 0$ ó 1, y $\sum_{i=1}^{n} b_i = n - k$. Denotaremos por $\mathcal{C}_{n,k}$ a la clase de sus funciones de distribución, es decir, la clase de las funciones $h : [0,1] \mapsto [0,1]$ absolutamente continuas, con $h(0) = 1 - h(1) = 0$, y $h'(t) = 0$ ó $\frac{1}{n(1-\alpha)}$, $t \in \left[\frac{i-1}{n}, \frac{i}{n}\right]$, pero $h'(t) = 0$ exactamente en k subintervalos del tipo $\left[\frac{i-1}{n}, \frac{i}{n}\right]$. Es obvio entonces que $\mathcal{C}_{n,k} \subset \mathcal{C}_{k/n}$, y que además $\#\mathcal{C}_{n,k} = \binom{n}{k}$.

Sea F_n la función de distribución de Q_n. Si R es un elemento genérico de $\mathcal{T}_k(Q_n)$, su función de distribución, F_R, será una función escalonada con saltos en los $n - k$ puntos no

recortados, y tendremos al igual que en el cálculo del estadístico de Kolmogorov,

$$\begin{aligned}\mathcal{W}_\infty(P,R) &= \sup_{x\in\mathbb{R}}|F_R(x)-F_P(x)| = \sup_{x\in\mathbb{R}}|h(F_n(x))-F_P(x)| \\ &= \max_{i=1,\ldots,n}\left(\left|h\left(\frac{i}{n}\right)-F_P(x_{(i)})\right|,\left|h\left(\frac{i-1}{n}\right)-F_P(x_{(i)})\right|\right),\end{aligned}$$

donde $h \in \mathcal{C}_{n,k}$ de acuerdo con la caracterización de los recortes enteros vista previamente.

Si consideramos ahora el estadístico de distancia asociado al problema (3.42) tendremos,

$$\begin{aligned}D_{n,k} &= \min_{R\in\mathcal{T}_k(Q_n)}\mathcal{W}_\infty(P,R) \\ &= \min_{h\in\mathcal{C}_{n,k}}\max_{i=1,\ldots,n}\left(\left|h\left(\frac{i}{n}\right)-F_P(x_{(i)})\right|,\left|h\left(\frac{i-1}{n}\right)-F_P(x_{(i)})\right|\right). \qquad (3.43)\end{aligned}$$

Teorema 3.37. *Si P es una medida de probabilidad en $\mathbb{R}$ continua y $X_1, X_2, \ldots, X_n$ una m.a.s, tal que $\mathcal{L}(X_i) = P$, entonces el estadístico $D_{n,k}$ es de distribución libre.*

Demostración. Es inmediato siguiendo los mismos pasos que en el resultado correspondiente para el estadístico de Kolmogorov. Como P es continua, las v.a. $Z_i = F_P(X_i)$ forman una m.a.s. de la $U[0,1]$, y esto llevado a la expresión (3.43) produce,

$$D_{n,k} \overset{d}{=} \min_{h\in\mathcal{C}_{n,k}}\max_{i=1,\ldots,n}\left(\left|h\left(\frac{i}{n}\right)-Z_{(i)}\right|,\left|h\left(\frac{i-1}{n}\right)-Z_{(i)}\right|\right),$$

donde el lado derecho no depende de la distribución P (la clase de funciones $\mathcal{C}_{n,k}$ no depende de P), y tenemos el resultado. ∎

Un resultado similar se tendría si en vez de manejar recortes enteros, $h \in \mathcal{C}_{n,k}$, manejásemos la norma del supremo con recortes generales, $h \in \mathcal{C}_\alpha$.

El cálculo del estadístico $D_{n,k}$ es un problema de optimización combinatoria. Tenemos $\binom{n}{k}$ posibles recortes enteros y tenemos que ver cuál de ellos hace mínima la distancia. Para tamaños pequeños de n (digamos $n < 30$) y valores de k bajos, es factible plantearse el cálculo exhaustivo. Por ejemplo, en el paquete R, existen librerías y comandos que generan todas las posibles combinaciones y en unos pocos segundos pueden computar el valor del estadístico. Sin embargo, para muestras un poco más grandes esto ya no es posible.

El algoritmo para el cálculo de $D_{n,k}$ que se presenta está inspirado en los algoritmos de camino de mínima distancia. La Figura 3.6 ilustra un ejemplo en el que disponemos de 6

datos y queremos eliminar 4. Representa un "grafo" con todas las posibilidades, que son los posibles caminos que llevan de los puntos de inicio a los puntos finales. Los datos se ordenan de menor a mayor y se va recorriendo la muestra en este sentido. Sobre cada punto se decide si se recorta o no (estado rojo y azul resp.). Si se recorta se pasa a la siguiente fila y se avanza de columna, y si no se recorta se avanza de columna, pero en la misma fila.

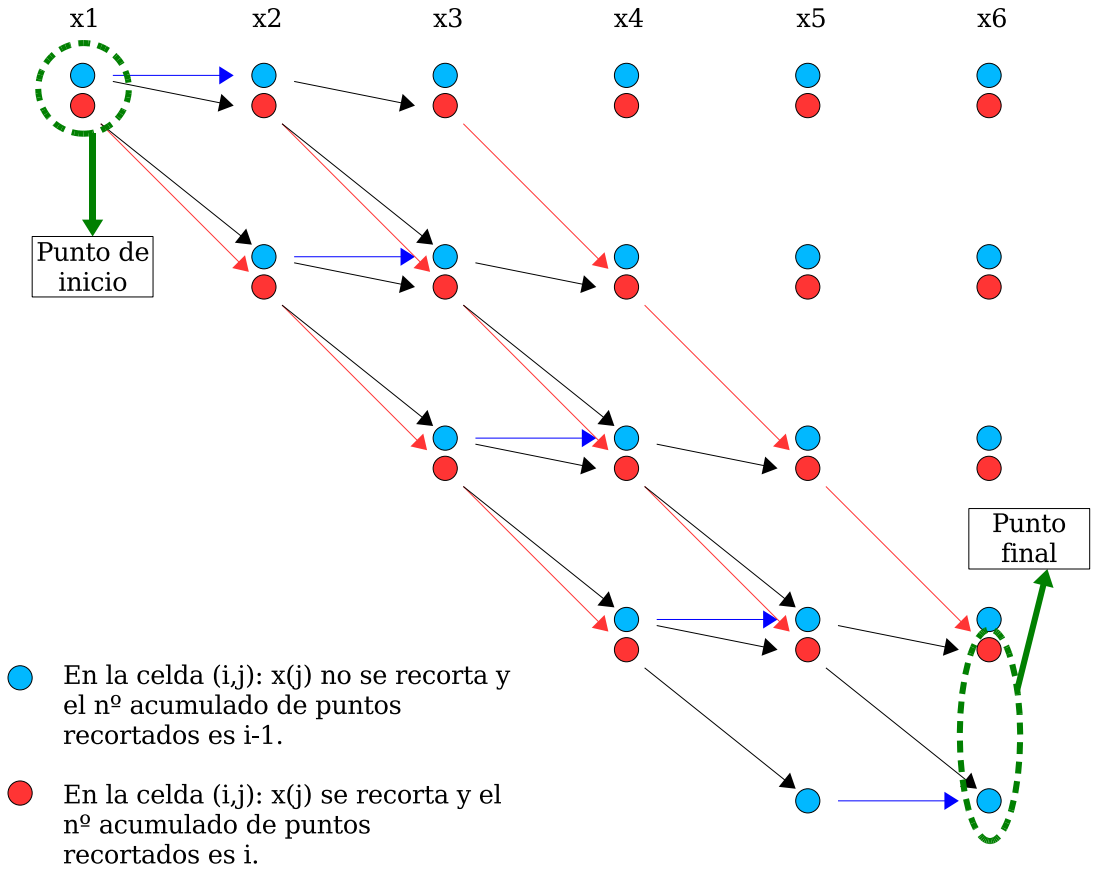

Figura 3.6: Gráfico asociado al problema cuando $n = 6$ y $k = 4$.

Así pues, si asociamos una matriz al gráfico, en la celda (i, j) el índice de filas i refleja que ya tenemos recortados $i - 1$ puntos al llegar a este nodo, y el de columnas indica que estamos examinando el dato $x_{(j)}$. Por otro lado, como resultado del análisis del elemento $x_{(j)}$, resultará que tenemos recortados $i - 1$ ó i elementos.

Asociado a cada nodo o celda (i, j) se va a calcular el "coste" que supone para el valor del estadístico $D_{n,k}$ que el camino final elegido pase por ese punto. Como en cada nodo tenemos dos posibles estados necesitamos calcular dos coeficientes. Llamaremos $c^*_{i,j}$ a la aportación al valor del estadístico si se recorta el dato $x_{(j)}$ habiendo recortado previamente $i - 1$ datos,

por lo que

$$c^*_{i,j} = \left| \frac{j-i}{n-k} - F(x_{(j)}) \right|,$$

ya que si se recorta el punto, no se produce un salto en la función de distribución empírica asociada. Y llamaremos $c_{i,j}$ al valor correspondiente cuando el punto $x_{(j)}$ no se recorta y el número acumulado de puntos recortados hasta $x_{(j)}$ es $i-1$. Entonces,

$$c_{i,j} = \text{máx}\left(\left| \frac{j-i+1}{n-k} - F(x_{(j)}) \right|, \left| \frac{j-i}{n-k} - F(x_{(j)}) \right| \right).$$

A continuación, y también asociado a cada nodo (i,j), se calculan otros dos coeficientes, de forma recursiva, que representan el "coste" de llegar hasta uno u otro estado del nodo. Si $d_{i,j}$ representa el coste de llegar al nodo (i,j) cuando no se recorta el punto $x_{(j)}$, teniendo en cuenta las dos posibilidades de llegar hasta ese nodo y estado, tal y como aparece en la Figura 3.7, entonces

$$d_{i,j} = \text{mín}\{\text{máx}\{c_{i,j}, d^*_{i-1,j-1}\}, \text{máx}\{c_{i,j}, d_{i,j-1}\}\}.$$

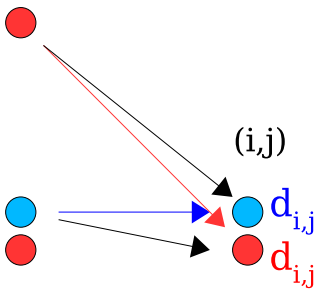

Figura 3.7: Posibles caminos para llegar al nodo (i,j).

Y si $d^*_{i,j}$ representa el correspondiente coste cuando el punto $x_{(j)}$ no se recorta, entonces, de forma similar,

$$d^*_{i,j} = \text{mín}\{\text{máx}\{c^*_{i,j}, d^*_{i-1,j-1}\}, \text{máx}\{c^*_{i,j}, d_{i,j-1}\}\}.$$

En el caso de estar en alguno de los nodos del borde, el cálculo de los coeficientes $d_{i,j}$ y $d^*_{i,j}$, se simplifica siguiendo el mismo esquema, de acuerdo con la Figura 3.6.

El valor final del estadístico $D_{n,k}$ se obtiene al final del "camino" y será el mínimo de los valores $d_{i,j}$ en los puntos finales: $D_{n,k} = \text{mín}\left(d_{k+1,n}, d^*_{k,n}\right)$.

Comentarios sobre el algoritmo. Según lo expuesto, para la aplicación de este algoritmo se deben obtener las cuatro matrices de coeficientes, $\{c^*_{i,j}\}_{ij}$, $\{d^*_{i,j}\}_{ij}$, $\{c_{i,j}\}_{ij}$, $\{d_{i,j}\}_{ij}$.

A la vista del gráfico de la Figura 3.6 es obvio que las 4 matrices necesarias para la obtención de $D_{n,k}$ pueden ser relativamente dispersas. De la dimensión inicial $(k+1)\times n$ y $k\times n$ sólo están ocupadas una banda diagonal superior de dimensión $(k+1)\times(n-k)$. Dependiendo de los tamaños de n y k, en la implementación computacional de este algoritmo es posible ahorrar espacio, recodificando las matrices. En el Apéndice A se incluye una función en R que calcula el valor del estadístico $D_{n,k}$.

Nota 3.38. Este algoritmo se puede generalizar al caso de dos muestras con relativamente poco esfuerzo. Al añadir la segunda muestra necesitamos manejar un "grafo" en tres dimensiones. En una de ellas representamos los puntos de las dos muestras conjuntamente ordenados (por ejemplo, $x_1 \le x_2 \le y_1 \le x_3 \le y_2 \le \ldots \le y_m$). Las otras dos dimensiones se utilizan para representar el número de puntos recortados que se acumulan en cada muestra (de forma similar a como se utilizan las filas en el algoritmo de una muestra). De esta forma se manejan matrices de tres dimensiones, además de un vector "bandera" que indica a que muestra pertenece cada punto de la ordenación.

3.4.2. Problemas de dos muestras.

Recorte general utilizando la métrica $\mathcal{W}_p$ en $\mathbb{R}^k$.

En este caso, si manejamos distribuciones discretas generales, y en particular medidas empíricas, no necesitamos usar la representación cuantil de la distancia de Wasserstein -que, además, es válida sólo en la recta real- puesto que es factible usar la definición directamente. Esto se traduce en que vamos a poder dar un resultado válido en $\mathbb{R}^k$ y para distancias $\mathcal{W}_p$, $p \ge 1$.

Sean $P = \sum_{i=1}^n p_i\delta_{x_i}$ y $Q = \sum_{j=1}^m q_j\delta_{y_j}$ dos medidas de probabilidad discretas en $\mathbb{R}^k$, i.e., $x_i \in \mathbb{R}^k$, $p_i \ge 0$, $i = 1,\ldots,n$ con $\sum_{i=1}^n p_i = 1$ e $y_j \in \mathbb{R}^k$, $q_j \ge 0$, $j = 1,\ldots,m$ con $\sum_{j=1}^m q_j = 1$. Si $0 < \alpha_1,\ \alpha_2 < 1$ (admitimos un tamaño de recorte diferente en cada distribución), el problema para el que vamos a dar un algoritmo es el de encontrar un par de medidas discretas $(P_{\alpha_1}, Q_{\alpha_2})$ tales que

$$\mathcal{W}_p^p(P_{\alpha_1}, Q_{\alpha_2}) = \text{mín}\left\{\mathcal{W}_p^p(R_1,R_2) : R_1 \in \mathcal{R}_{\alpha_1}(P),\ R_2 \in \mathcal{R}_{\alpha_2}(Q)\right\}. \tag{3.44}$$

Teniendo en cuenta la Definición 2.1 de la métrica $\mathcal{W}_p$ de Wasserstein, si R y S son dos elementos genéricos de $\mathcal{R}_{\alpha_1}(P)$ y $\mathcal{R}_{\alpha_2}(Q)$ respectivamente

$$\mathcal{W}_p^p(R,S) = \int \|x-y\|^p d\pi = \sum_{i=1}^{n}\sum_{j=1}^{m} \|x_i - y_j\|^p \pi_{ij},$$

donde

$$0 \leq \sum_{j=1}^{m} \pi_{ij} \leq \frac{p_i}{1-\alpha_1},\ i = 1,\ldots,n; \tag{3.45}$$

$$0 \leq \sum_{i=1}^{n} \pi_{ij} \leq \frac{q_j}{1-\alpha_2},\ j = 1,\ldots,m; \quad \text{y} \tag{3.46}$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} = 1. \tag{3.47}$$

Entonces el problema (3.44) se puede escribir como un problema de optimización lineal con restricciones lineales,

$$\mathcal{W}_p^p(P_{\alpha_1}, Q_{\alpha_2}) = \text{mín} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\pi_{ij},$$

sujeto a las restricciones de no negatividad ($\pi_{ij} \geq 0$) y las dadas por las expresiones (3.45), (3.46) y (3.47); y donde $c_{ij} = \|x_i - y_j\|^p$ son datos.

Esto prueba el siguiente resultado.

Proposición 3.39. *Sean $P = \sum_{i=1}^{n} p_i \delta_{x_i}$ y $Q = \sum_{j=1}^{m} q_j \delta_{y_j}$ dos medidas de probabilidad discretas en $\mathbb{R}^k$. Entonces el problema dado por la expresión (3.44) es un problema de programación lineal que puede resolverse utilizando el algoritmo simplex.*

El problema (3.44) es en realidad un problema de transporte, en el que hay n posibles puntos de oferta y m posibles puntos de demanda. π_{ij} representa la cantidad, en este caso de masa de probabilidad, que se transporta del origen i al punto de demanda j. Del punto i de oferta no puede salir más cantidad de $\frac{p_i}{1-\alpha_1}$ y el punto j de demanda no puede recibir más de $\frac{q_j}{1-\alpha_2}$. Finalmente, debe haber igualdad entre lo que se envía y recibe.

El anterior resultado es válido en particular cuando tenemos dos muestras de vectores aleatorios en $\mathbb{R}^k$ y sus distribuciones empíricas asociadas, haciendo $p_i = \frac{1}{n},\ i = 1,\ldots,n$ y $q_j = \frac{1}{m},\ j = 1,\ldots,m$.

También es válido en el caso en el que decidamos recortar sólo en una de las dos medidas. Así, si decidimos no recortar en la segunda, sólo tenemos que cambiar el grupo de restricciones dado por (3.46) por las restricciones de igualdad: $\sum_{i=1}^{n} \pi_{ij} = q_j,\ j = 1,\ldots,m$.

En el Apéndice A se incluye una función en R para el cálculo del recorte óptimo en el caso de dos muestras. Esta función es relativamente rápida para tamaños de muestra de hasta $n = 200$ (tamaño del problema $n \times m \simeq 4 \times 10^4$, si $n = m$). Para tamaños superiores hay que recurrir a un *solver* comercial como por ejemplo CPLEX. Se incluye en el Apéndice el programa en AMPL+CPLEX que se ha utilizado para los ejemplos del Capítulo 4. Este programa se ha utilizado para tamaños de problema de 10^6 ($n = m = 1000$) y ofrece soluciones en unos pocos minutos.

Recorte general utilizando la métrica $\mathcal{W}_2$ y el mismo patrón en ambas muestras.

El problema (3.16) alcanza el inferior para una función $h_0 \in \mathcal{C}_\alpha$ (ver (3.19)) tal que:

$$h_0'(t) = \frac{1}{1-\alpha} I_{[0, L_{F,G}^{-1}(1-\alpha)]}(|F^{-1}(t) - G^{-1}(t)|).$$

El procedimiento para hallar tanto h_0 como el valor del estadístico T_4 dado por (3.18) es numérico. En primer lugar se evalúa $|F^{-1}(t) - G^{-1}(t)|$ en una malla lo suficientemente densa en el intervalo $[0, 1]$, usando el $(1-\alpha)$-cuantil de dichos valores para determinar el valor de corte $L_{F,G}^{-1}(1-\alpha)$. De esta forma tenemos una aproximación numérica del conjunto $\left\{t \in (0,1) : |F^{-1}(t) - G^{-1}(t)| \leq L_{F,G}^{-1}(1-\alpha))\right\}$. A continuación se calcula también de forma numérica la integral dada en (3.18).

Para aplicar este procedimiento a dos muestras basta con reemplazar en todo lo anterior F y G por F_n y G_m, las funciones de distribución empíricas asociadas a cada una de las muestras.

En el Capítulo 5 se incluye un ejemplo que ha sido tratado con esta metodología en el que se calculan las funciones h_0' para distintos tamaños de α, así como los correspondientes valores del estadístico y algunas otras salidas de interés. Se hace referencia también al procedimiento de estimación de la varianza asintótica asociada a los resultados que se ven en el mismo capítulo. En el Apéndice A se incluyen las funciones y programas en R que sirven para realizar estos cálculos.

Recorte entero utilizando la métrica $\mathcal{W}_p$ en $\mathbb{R}^k$.

Por las mismas razones esgrimidas en el caso de recortes generales visto anteriormente, en este caso podremos en principio resolver problemas en $\mathbb{R}^k$ y para distancias $\mathcal{W}_p$, $p \geq 1$. Nos

centraremos en las distribuciones empíricas (y no discretas en general) porque representan el caso más interesante.

Sean $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ y $Q_= \frac{1}{m}\sum_{j=1}^{m}\delta_{y_j}$ dos medidas de probabilidad empíricas en $\mathbb{R}^k$, i.e., $x_i \in \mathbb{R}^k,\ i = 1,\dots,n$ y $y_j \in \mathbb{R}^k,\ j = 1,\dots,m$. Si $k_1 \leq n$ y $k_2 \leq m$ (admitimos eliminar un número distinto de puntos en cada muestra), el problema que se plantea es el de encontrar un par de medidas discretas (P_{n,k_1}, Q_{m,k_2}) tales que

$$\mathcal{W}_p^p(P_{n,k_1}, Q_{m,k_2}) = \text{mín}\left\{\mathcal{W}_p^p(R_1,R_2) : R_1 \in \mathcal{T}_{k_1}(P_n),\ R_2 \in \mathcal{T}_{k_2}(Q_m)\right\}. \tag{3.48}$$

Siguiendo los mismos pasos que en el caso general llegamos a que el problema de optimización que tenemos que resolver consiste en

$$\text{mín}\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\pi_{ij}, \tag{3.49}$$

sujeto a las habituales restricciones de no negatividad ($\pi_{ij} \geq 0$), y,

$$0 \leq \sum_{j=1}^{m}\pi_{ij} = \frac{a_i}{n-k_1},\ a_i = 0 \text{ ó } 1,\ i = 1,\dots,n; \tag{3.50}$$

$$0 \leq \sum_{i=1}^{n}\pi_{ij} = \frac{b_j}{m-k_2},\ b_j = 0 \text{ ó } 1,\ j = 1,\dots,m; \tag{3.51}$$

$$\sum_{i=1}^{n} a_i = n - k_1, \tag{3.52}$$

$$\sum_{j=1}^{m} b_j = m - k_2, \tag{3.53}$$

y donde $c_{ij} = \|x_i - y_j\|^p$ son datos. Esto prueba el siguiente resultado.

Proposición 3.40. *Sean $P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ y $Q_= \frac{1}{m}\sum_{j=1}^{m}\delta_{y_j}$ dos medidas de probabilidad empíricas en $\mathbb{R}^k$. Entonces el problema dado por la expresión (3.48) es un problema de programación lineal entera.*

Para resolver el problema (3.48) podemos utilizar *solvers* comerciales como el XPRESS o el CPLEX, o también utilizar la librería *lpSolve* del paquete estadístico R.

Si sólo admitimos recortar una de las dos muestras, digamos la primera, entonces basta con reemplazar las restricciones (3.51) y (3.53) por $\sum_{i=1}^{n}\pi_{ij} = \frac{1}{m}$, $j = 1,\dots,m$.

Esto último ocurre en particular en el caso del problema planteado en el Teorema 3.36, en el que reducimos el problema de comparar una muestra que permitimos recortar contra

una distribución de referencia, a comparar esa misma muestra contra una uniforme discreta especial (que no recortamos).

Existe un caso particular del problema de recortes enteros en dos muestras que merece especial atención, y es cuando $r = n - k_1 = m - k_2$, lo que en particular ocurre si $n = m$ y $k_1 = k_2$. En esa situación vamos a ver que la solución del problema de programación entera dado por (3.48) se obtiene resolviendo la relajación lineal, i.e., mediante el algoritmo simplex.

Supongamos entonces que $r = n - k_1 = m - k_2$. Realizando el cambio de variable $z_{ij} = r\pi_{ij}$, el problema (3.48) se puede reformular como sigue,

$$\text{mín} \frac{1}{r} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} z_{ij},$$

sujeto a las habituales restricciones de no negatividad ($z_{ij} \geq 0$), y,

$$\begin{aligned} 0 \leq \sum_{j=1}^{m} z_{ij} = a_i,\ a_i = 0 \text{ ó } 1,\ i = 1, \ldots, n; \\ 0 \leq \sum_{i=1}^{n} z_{ij} = b_j,\ b_j = 0 \text{ ó } 1,\ j = 1, \ldots, m; \\ \sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j = r, \end{aligned}$$

y donde $c_{ij} = \|x_i - y_j\|^p$ son datos. Y la relajación lineal del mismo,

$$\text{mín} \frac{1}{r} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} z_{ij}, \tag{3.54}$$

sujeto a

$$0 \leq z_{ij} \leq 1;\ 0 \leq a_i \leq 1;\ 0 \leq b_j \leq 1;\ i = 1, \ldots, n;\ j = 1, \ldots, m; \tag{3.55}$$

$$0 \leq \sum_{j=1}^{m} z_{ij} = a_i,\ i = 1, \ldots, n; \tag{3.56}$$

$$0 \leq \sum_{i=1}^{n} z_{ij} = b_j,\ j = 1, \ldots, m; \tag{3.57}$$

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j = r. \tag{3.58}$$

Denotemos por $\mathbf{I_n}$ a la matriz identidad de dimensión $n \times n$, por $\underline{\mathbf{1}}_\mathbf{n}$ y $\underline{\mathbf{0}}_\mathbf{n}$ a los vectores columna de unos y ceros, respectivamente, de longitud n.

Sea $\underline{\mathbf{x}} = (z_{11}, z_{12}, \ldots, z_{1m}, z_{21}, \ldots, z_{2m}, \ldots, z_{n1}, \ldots, z_{nm}, |a_1, \ldots, a_n, |b_1, \ldots, b_m)$ el vector de variables asociadas al problema de optimización. Si $h = nm + n + m$, sea $\mathbf{D}$ la matriz de dimensiones $(n + m + 2) \times h$ siguiente

$$\left[\begin{array}{cccccccccccccc|cccc|cccc}
1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 & \ldots & \ldots & 0 & 0 & \ldots & 0 & 0 & -1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \ldots & \ldots & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & -1 & 0 & 0 & \ldots & 0 \\
\hline
1 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 & \ldots & \ldots & 1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & -1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 & 0 & 1 & \ldots & 0 & \ldots & \ldots & 0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 & 0 & -1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 1 & \ldots & \ldots & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & -1 \\
\hline
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1
\end{array}\right]$$

y sean $\underline{\mathbf{b}}_{\mathbf{1}} = \underline{\mathbf{1}}_{\mathbf{h}}$ y $\underline{\mathbf{b}}_{\mathbf{2}} = (\underline{\mathbf{0}}^{\mathbf{t}}_{\mathbf{n+m}}, r, r)^t$. Entonces las restricciones (3.55) pueden ser reescritas como $\mathbf{I_h}\underline{\mathbf{x}} \leq \underline{\mathbf{b}}_{\mathbf{1}}$ y las restricciones (3.56)-(3.58) como $\mathbf{D}\underline{\mathbf{x}} = \underline{\mathbf{b}}_{\mathbf{2}}$.

Si ahora completamos el vector de variables $\underline{\mathbf{x}}$ con las variables auxiliares (de holgura) para las restricciones "$\leq$" y llamamos $\underline{\mathbf{y}}$ al nuevo vector de variables, y llamamos $\mathbf{A}$ y $\underline{\mathbf{b}}$, a la matriz por cajas y vector, respectivamente, siguientes

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I_h} & \mathbf{I_h} \\ \hline \mathbf{D} & \underline{\mathbf{0}}_{\mathbf{(n+m+2)\times h}} \end{array}\right] \quad \text{y} \quad \underline{\mathbf{b}} = \left[\begin{array}{c} \underline{\mathbf{b}}_{\mathbf{1}} \\ \hline \underline{\mathbf{b}}_{\mathbf{2}} \end{array}\right],$$

entonces las restricciones (3.55)-(3.58) pueden ser escritas de forma compacta como

$$\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}},$$

y podemos formular el siguiente resultado,

Teorema 3.41. *Si $r = n - k_1 = m - k_2$, la solución del problema dado por la expresión (3.48) es la misma que la del problema de optimización dado por (3.54) sujeto a las restricciones (3.55)-(3.58).*

Demostración. Es inmediato por lo visto hasta aquí. Por una parte hemos visto que el politopo dado por las restricciones (3.55)-(3.58) se puede escribir como

$$R(\mathbf{A}) = \{\underline{\mathbf{x}} : \mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{b}}, \quad \underline{\mathbf{x}} \geq \underline{\mathbf{0}}\}.$$

Por otra, en virtud del Teorema 2.18 la matriz $\mathbf{D}$ es unimodular, pues sus filas pueden ser divididas en dos conjuntos, I_1 (las marcadas en rojo) e I_2 (las marcadas en azul), que cumplen las condiciones del teorema. Entonces, la matriz $\mathbf{A}$, construida añadiendo matrices identidad, también lo es. Como el vector $\underline{\mathbf{b}}$ está formado por valores enteros, aplicando el Teorema 2.16 todos los vértices de $R(\mathbf{A})$ son enteros, y por tanto, la solución de la relajación lineal es entera, y se tiene el resultado pues estamos minimizado la misma función objetivo.

■

Una consecuencia inmediata e interesante de este resultado es que cuando estamos comparando y recortando sin restricciones dos distribuciones empíricas, si $n = m$ y $n(1-\alpha)$ es entero, entonces siempre puede tomarse un recorte entero, es decir, las distribuciones recortadas consisten en "empíricas" sobre un subconjunto de los puntos originales. Por otro lado, esta es la única posibilidad en los casos en los que el valor de la función objetivo es estrictamente positivo.

Capítulo 4

Aspectos descriptivos de la metodología.

En este capítulo se pretende, por una parte, dar una idea general del funcionamiento de los recortes imparciales en sus diferentes modalidades, y por otra, ilustrar su aplicación como metodología descriptiva en la comparación de distribuciones. Para ello se han elegido 6 modelos poblacionales diferentes: N(0,1), N(-3,1), N(0,3), $U(-\sqrt{3},\sqrt{3})$ (uniforme centrada en 0 y con varianza unidad), χ^2_4 - 4 y 0.8N(0,1)+0.2N(4,1), de los que se han generado muestras de tamaño 1000. A continuación se ha procedido a realizar algunas comparaciones 2 a 2 de estas muestras variando el tamaño de recorte ($\alpha = 0.05$, 0.1 y 0.2). Los resultados se muestran en las Figuras 4.1 a 4.8. En las mismas la primera fila de gráficos corresponde al caso en el que se recorta en ambas distribuciones, la segunda y tercera, al caso en el que sólo se recorta en una distribución, y la cuarta fila al caso en el que se recortan ambas distribuciones siguiendo el mismo patrón. La parte de los histogramas en blanco corresponde al primer 5 % de observaciones recortadas. En amarillo aparecen las que se añaden cuando $\alpha = 0.1$, es decir, las observaciones recortadas en el primer 10 % aparecen en blanco y amarillo. En naranja aparecen las observaciones que se recortan al pasar del 10 % al 20 %, y finalmente, en rojo aparecen las observaciones no recortadas cuando $\alpha = 0.2$.

En la Figura 4.1 se comparan las muestras de la N(0,1) y la N(-3,1) y sirve para ilustrar el efecto que tiene en los recortes una diferente localización de las distribuciones. Cuando el recorte se produce sin restricciones (tres primeras filas), se eliminan observaciones de la cola derecha de la N(0,1) y de la cola izquierda de la N(-3,1). Puestos en cada distribución,

las que están más separadas del rango de valores de la otra distribución. Así pues, cuando la localización de las distribuciones es lo suficientemente distinta, lo primero que intentan los recortes es acercar la localización de las distribuciones resultantes. Esto en realidad es el reflejo de una conocida propiedad de la distancia $\mathcal{W}_2$, recogida en el Lema 2.6. Si en este caso el recorte se realiza con el mismo patrón (cuarta fila), las observaciones que se eliminan aparecen en cambio de forma mayoritaria en la parte central (también hay alguna en la cola derecha).

En la Figura 4.2, comparamos dos distribuciones con la misma localización - N(0,1) vs N(0,3) -, pero con diferente dispersión. En este caso la parte de la N(0,3) (la más dispersa), que es más diferente de la N(0,1), se halla precisamente en la parte de la distribución que contribuye a esa mayor dispersión y que está fuera del rango de valores de la N(0,1), y esta es la parte en la que se recorta -ambas colas-. En cambio, en la muestra de la N(0,1), la parte más diferente de la N(0,3) se encuentra en la parte central donde se recortan observaciones rebajando la densidad.

Comparamos ahora la N(0,1) con la $U(-\sqrt{3}, \sqrt{3})$ (Figura 4.3), dos distribuciones con la misma localización y dispersión pero, obviamente, diferente forma. Cuando recortamos en ambas distribuciones los recortes de cada distribución intentan acomodarse a la forma de la otra distribución. El recorte en la normal se produce en dos partes. Por un lado en las colas que se salen del rango de la uniforme, y por otro, en la parte central haciendo más plana la densidad resultante. Por su parte, el recorte en la uniforme hace aparecer la forma de campana de la normal (gráfico de la derecha en la primera fila). Hasta ahora (Figuras 4.1 y 4.2) cuando el recorte se producía en una sola distribución, el resultado era el mismo (aprox.) que cuando el recorte se llevaba a cabo en las dos. Este es un claro ejemplo de que esto no es una pauta general. Cuando recortamos sólo en la normal, la proporción de observaciones recortadas en el centro es mayor que en el caso de recortes en ambas. Y también es ligeramente diferente si recortamos sólo en la uniforme. En este caso, también el recorte intenta hacer aparecer la forma de campana de la normal, pero con una diferencia importante con respecto al caso de recorte en ambas. En los extremos del rango de la uniforme apenas se recorta para "emparejar" esta zona con las colas de la normal (que no se recorta). Si recortamos con el mismo patrón se produce un recorte raro sin una clara relación con la forma de la distribución.

En la Figura 4.4 se compara la normal estándar con una χ^2_4 centrada en 0. El resultado del recorte en la normal hace que se eliminen observaciones hacia la zona central promoviendo que la forma de la distribución resultante sea más parecida a la de la χ^2_4, es decir, asimétrica a la derecha. En cambio en la χ^2_4, las observaciones que se eliminan se sitúan al principio en la cola de la derecha, que es la que más contribuye a la asimetría, y cuando se llega a $\alpha = 0.1$, se recorta también en la zona de la izquierda. El recorte con el mismo patrón (cuarta fila) produce resultados un poco diferentes. Por un lado, la distribución resultante de recortar la χ^2_4 es algo más simétrica que cuando recortamos sin restricciones (comparar con la primera o tercera fila). Por otro lado el recorte en la N(0,1) es diferente al caso sin restricciones, pues se recorta en la cola derecha, y observaciones que están alrededor del percentil 20 %.

En la Figura 4.5 se compara una N(0,1) con una N(0,1) contaminada un 20 % con una N(4,1). El recorte sin restricciones en ambas distribuciones muestra la capacidad del mismo para recuperar la N(0,1) en la mezcla, recortando aquellas observaciones que provienen de la contaminación. En cambio en la muestra de la N(0,1) recorta aproximadamente en todo el rango de valores rebajando la densidad de forma proporcional a ésta, es decir, respetando el modelo original. Cuando se recorta en la N(0,1) y no en la mezcla, el recorte se produce esencialmente en la parte izquierda de la distribución. Es también diferente el recorte utilizando el mismo patrón. Recorta la cola derecha en ambas distribuciones.

Cuando comparamos la N(0,3) con la $\chi^2_4 - 4$ (Figura 4.6), dos distribuciones con la misma localización, casi igual varianza, y un rango muestral de valores relativamente similar, el recorte en las dos distribuciones afecta más claramente a la forma. En la normal se recortan observaciones de tal forma que la distribución resultante es asimétrica a la derecha, aproximándose a la forma de la χ^2_4. En la $\chi^2_4 - 4$ se eliminan inicialmente algunos puntos de la cola derecha, pero enseguida se eliminan observaciones de la parte central buscando la simetría que la acerque a la normal. Si el recorte se produce en una sola distribución se obtienen resultados similares a los anteriores. Y finalmente, si el recorte es con el mismo patrón, los resultados son algo diferentes. Se recortan observaciones en ambas colas y si aumentamos α hasta 0.2 aparecen algunas observaciones recortadas en la zona central.

En la Figura 4.7 se compara la $\chi^2_4 - 4$ con la mezcla $0.8N(0,1) + 0.2N(4,1)$. En este caso, cuando se recorta sin restricciones, también se observa un recorte que tiende a acomodar la forma de la mezcla con la asimetría de la $\chi^2_4 - 4$ (ver primera fila, gráfico de la derecha).

Mientras que en la $\chi^2_4 - 4$ se recortan las observaciones más extremas en la cola derecha para pasar a perfilar la distribución en la cola izquierda.

Finalmente, parece interesante comparar dos muestras provenientes de la misma distribución. En la Figura 4.8 se comparan dos muestras de una N(0,1). El recorte se produce en ambas distribuciones eliminando observaciones que rebajan la densidad de forma proporcional a ésta (al igual que se había observado en la Figura 4.5). Además, el recorte tiende a suavizar el histograma de la distribución resultante, recortando más en aquellas clases en las que por efecto de la aleatoriedad se ha producido una mayor o menor frecuencia que en la correspondiente clase de la otra distribución. Obsérvese para ello el recorte que se produce en la clase (-0.5,0] del histograma del lado derecho de la primera fila.

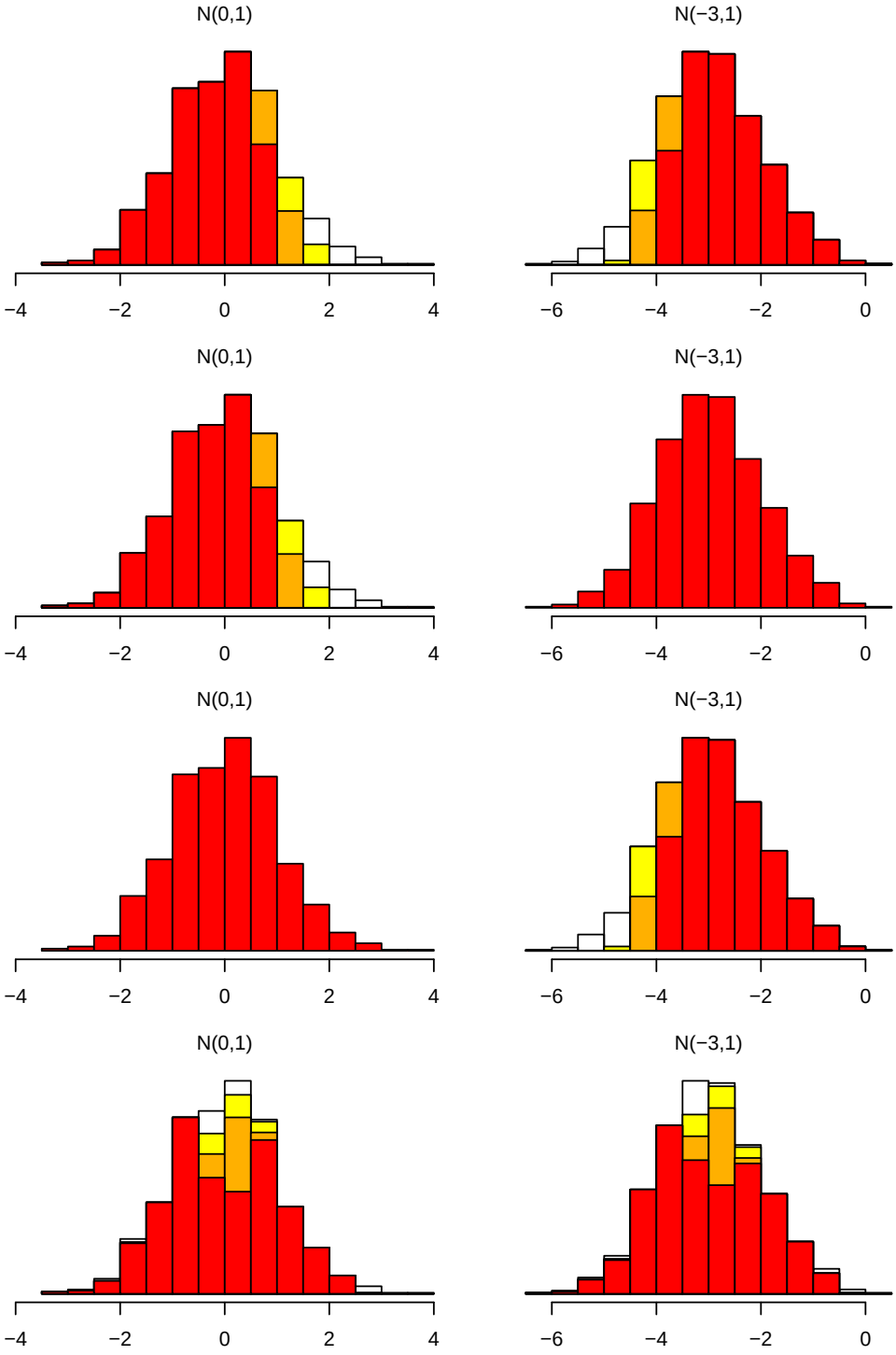

Figura 4.1: Comparación de una N(0,1) vs N(-3,1). Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

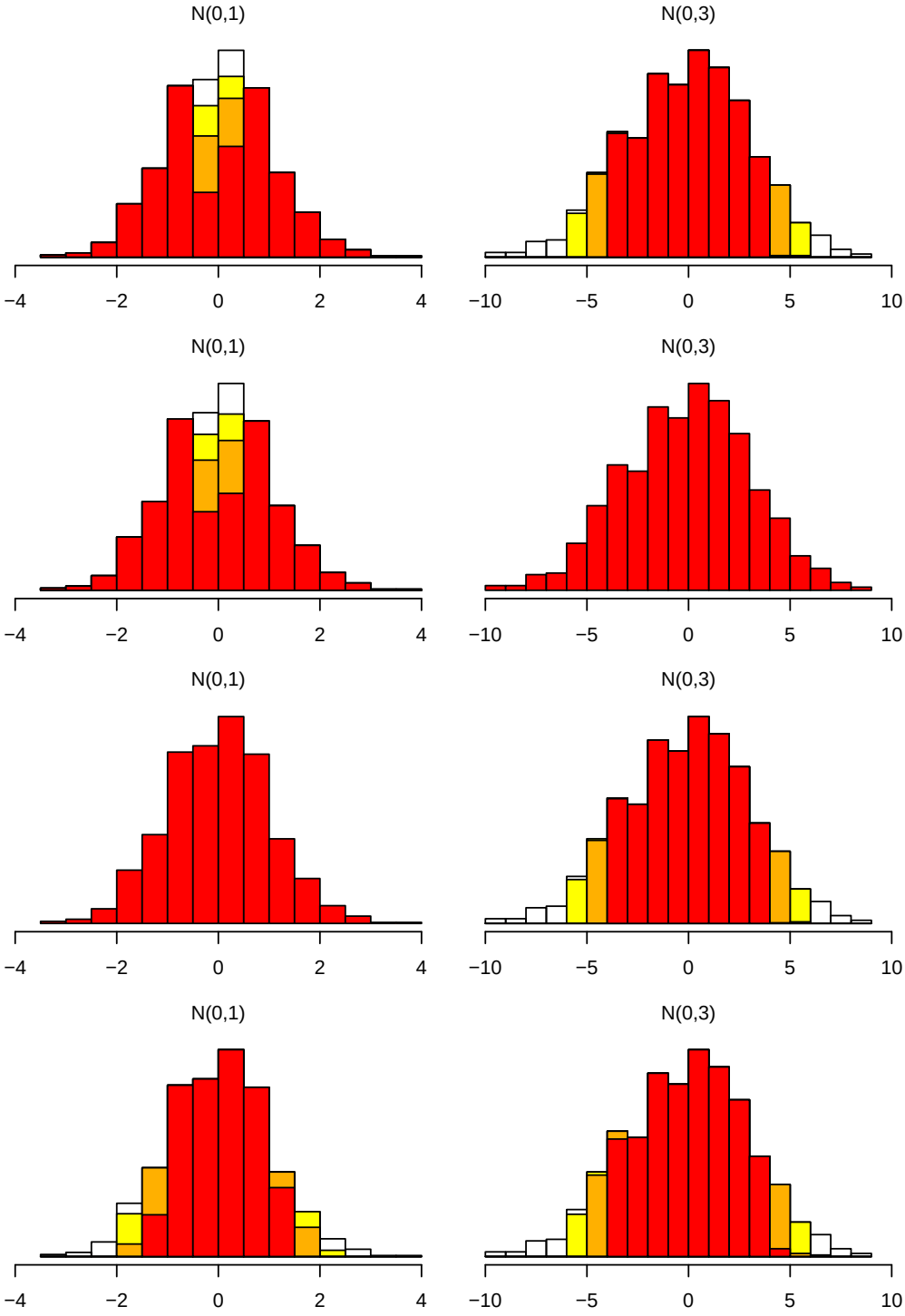

Figura 4.2: Comparación de una N(0,1) vs N(0,3). Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

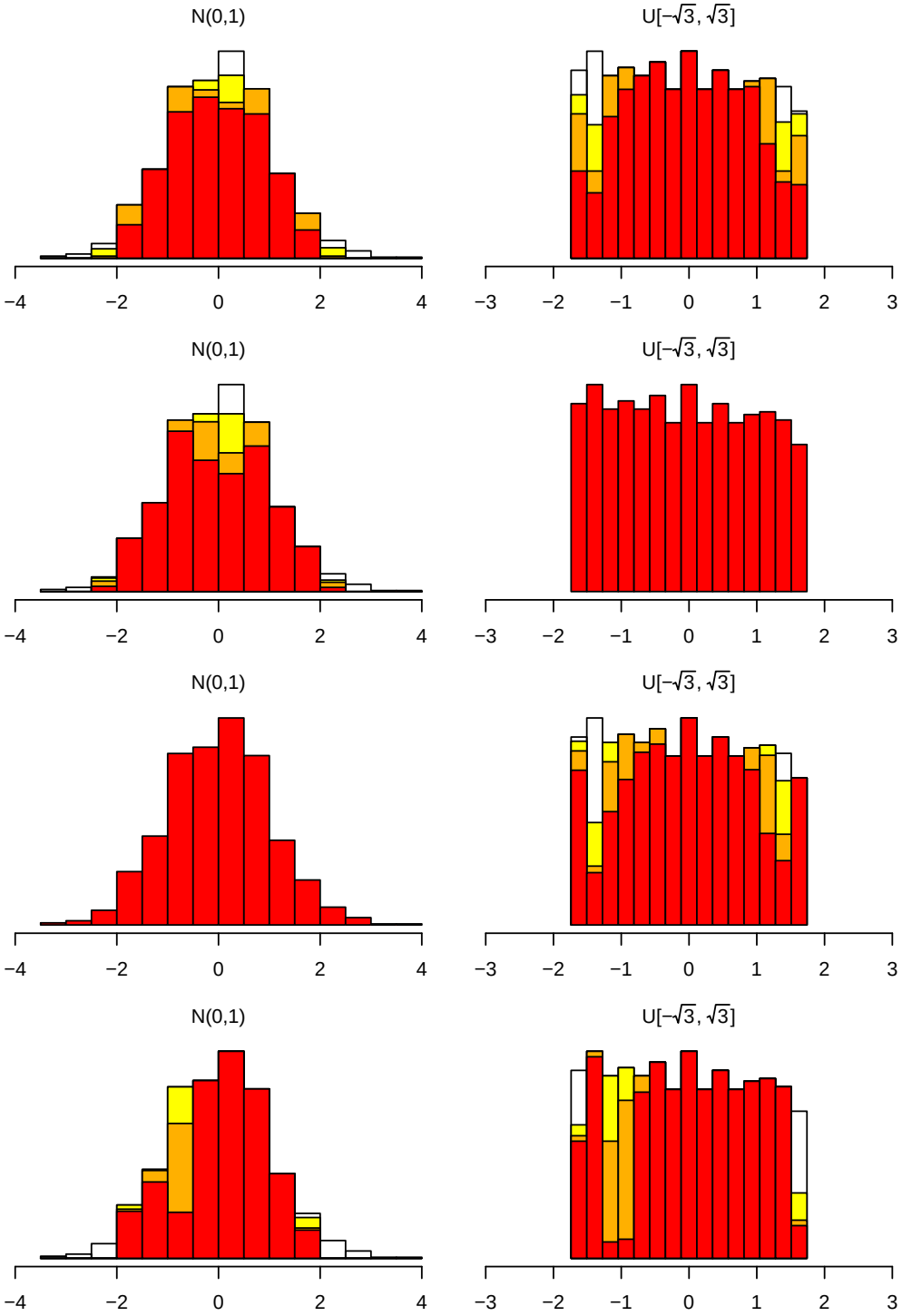

Figura 4.3: Comparación de una N(0,1) vs $U(-\sqrt{3}, \sqrt{3})$. Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

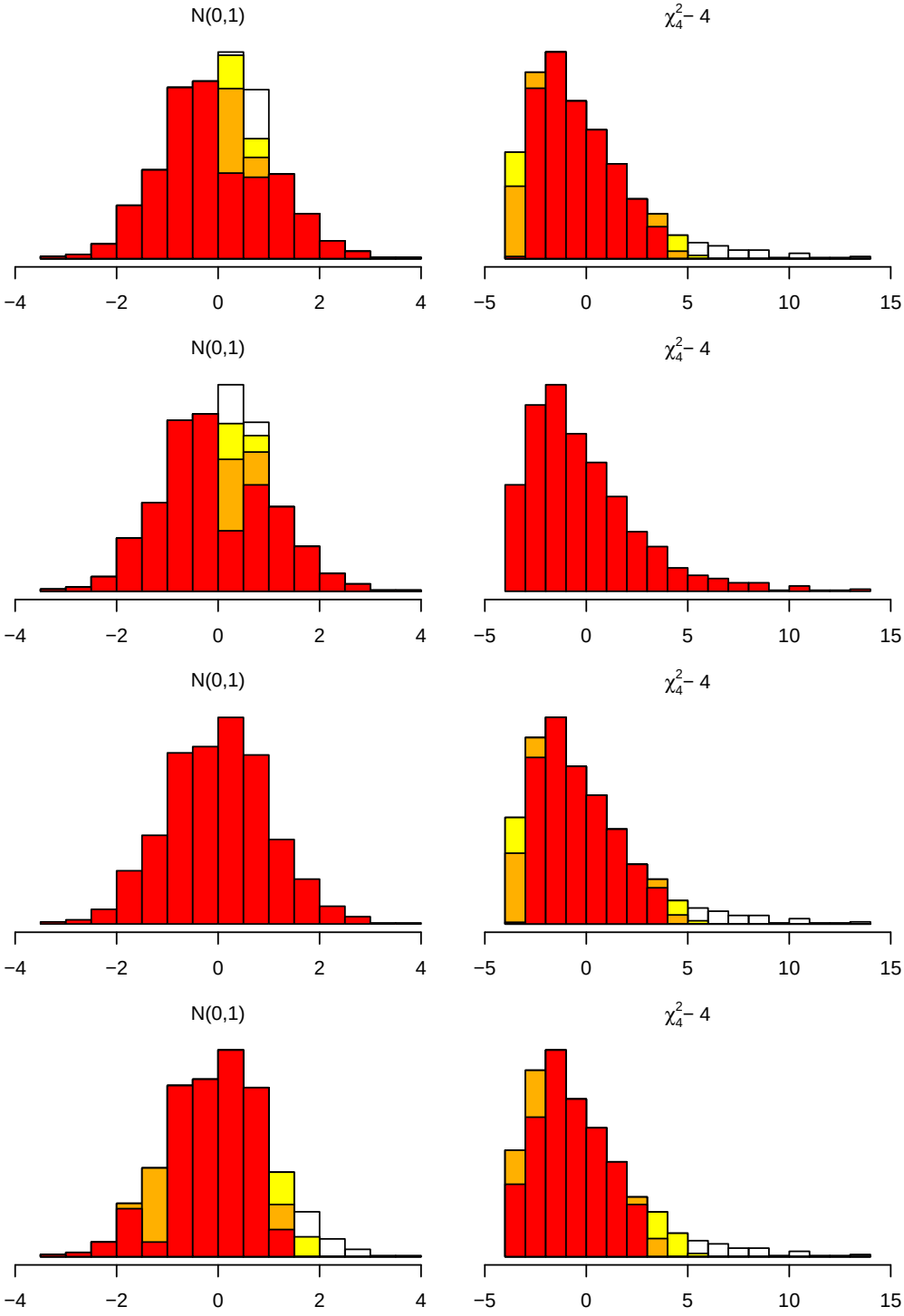

Figura 4.4: Comparación de una N(0,1) vs $\chi^2_4 - 4$. Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

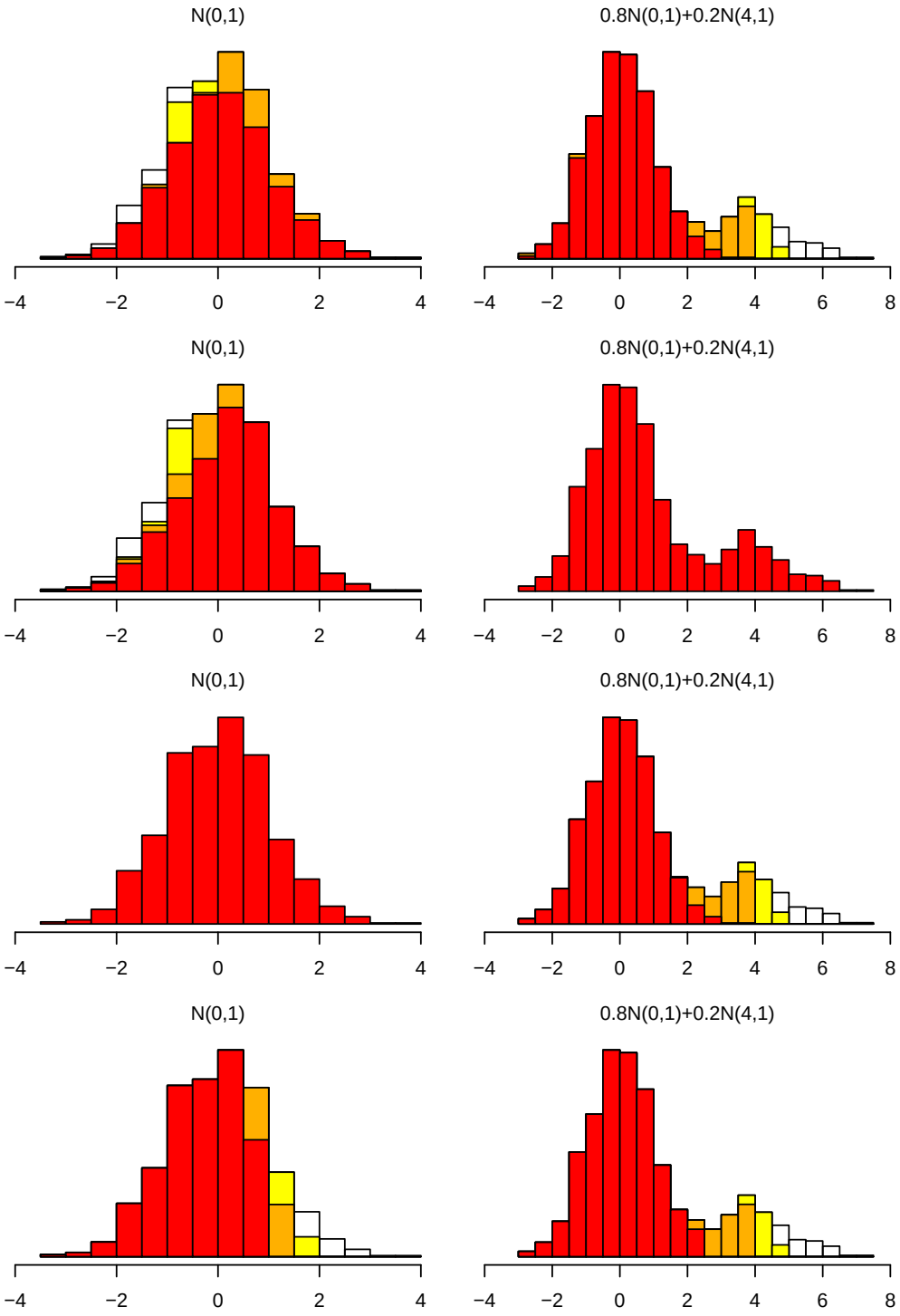

Figura 4.5: Comparación de una N(0,1) vs 0.8N(0,1)+0.2N(4,1). Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

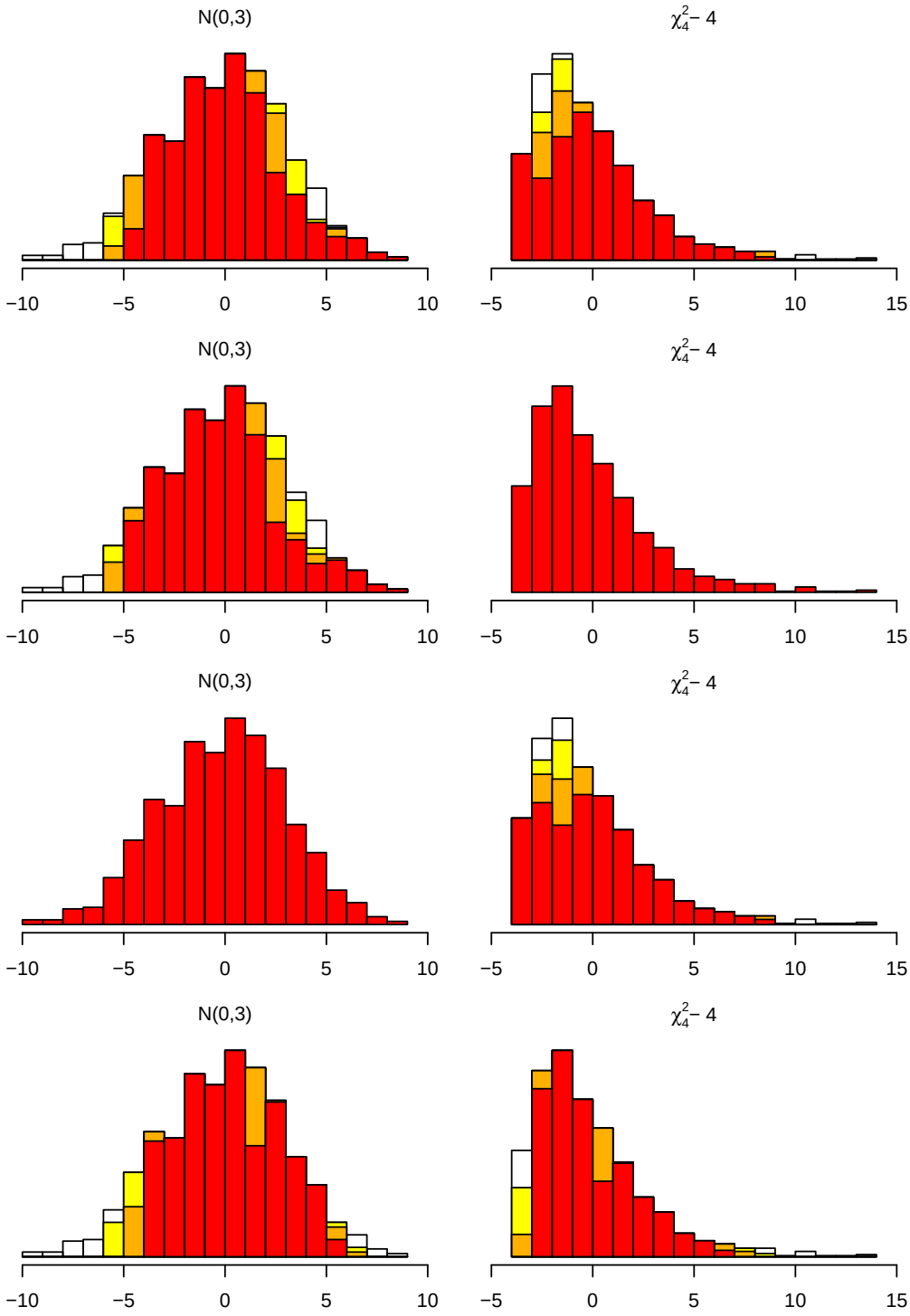

Figura 4.6: Comparación de una N(0,3) $\chi^2_4 - 4$. Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

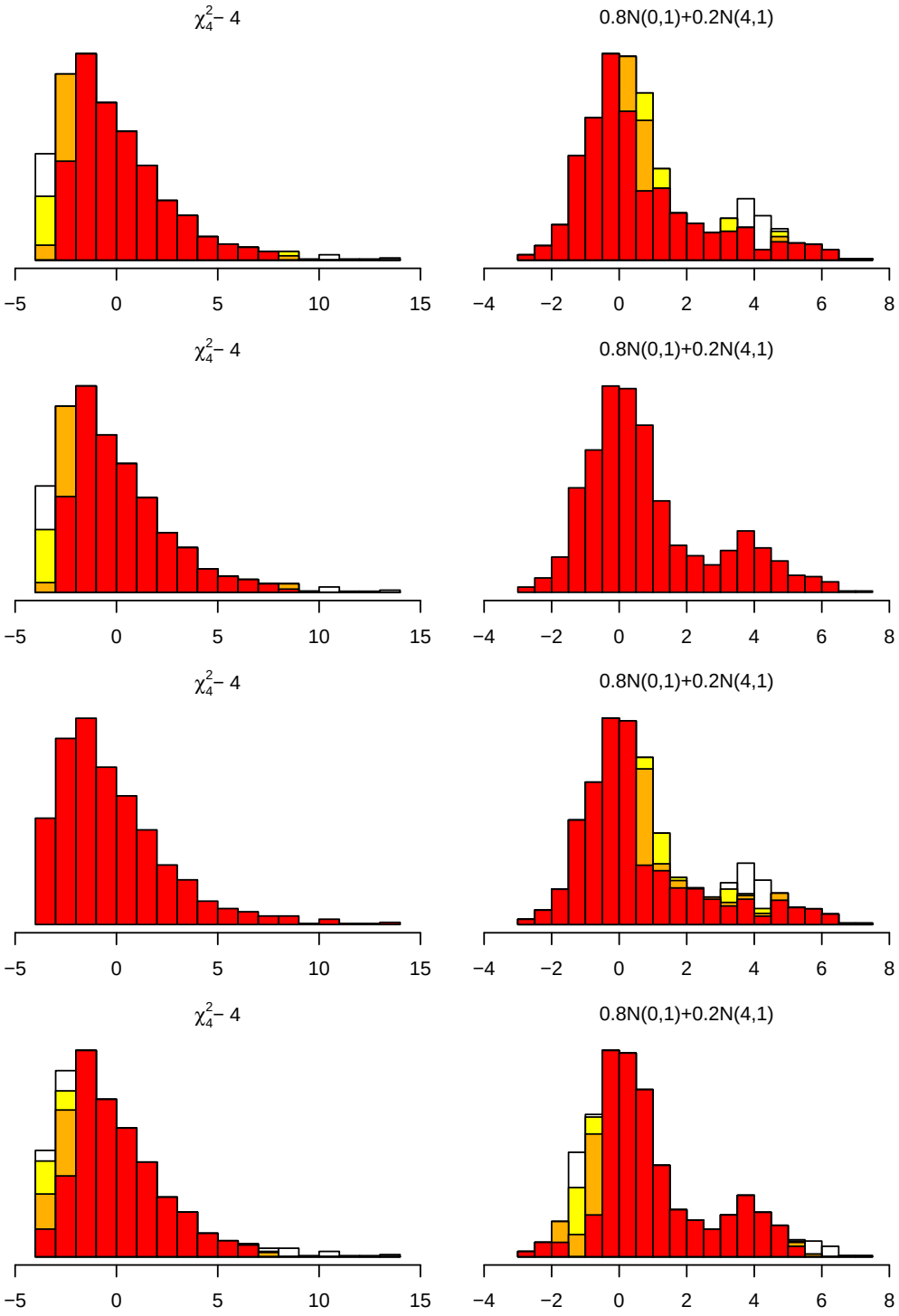

Figura 4.7: Comparación de una $\chi^2_4 - 4$ vs 0.8N(0,1)+0.2N(4,1). Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

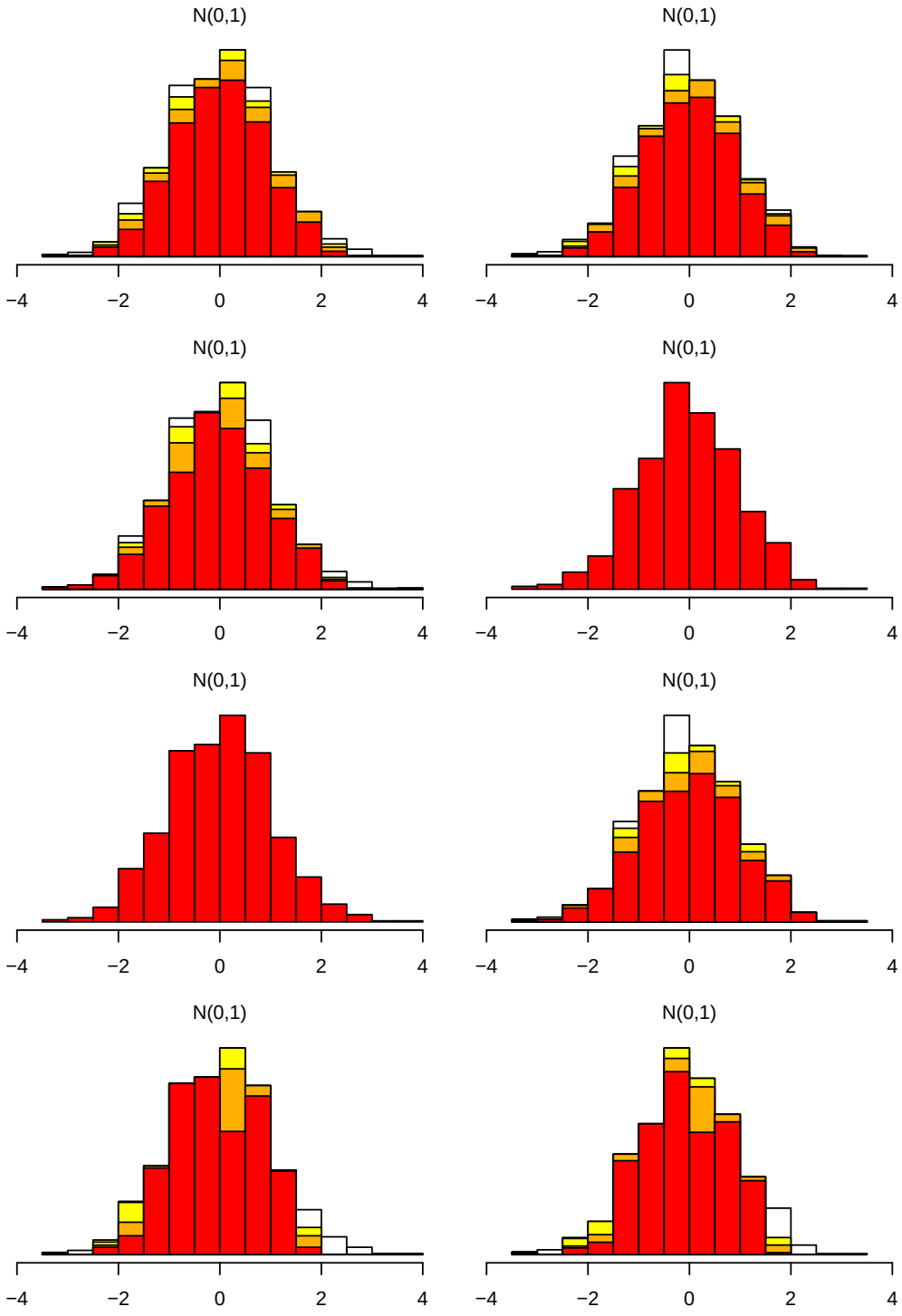

Figura 4.8: Comparación de dos muestras indep. de una N(0,1). Se generan 1000 datos de cada población, y se recortan un 5 % (observaciones recortadas en blanco), un 10 % (blanco + amarillo) y un 20 % (blanco + amarillo + naranja). La primera fila de gráficos corresponde al recorte en ambas muestras, la segunda y tercera, al recorte en una muestra, y la cuarta, al recorte con el mismo patrón.

Capítulo 5

Comportamiento asintótico.

En este Capítulo se incluyen los resultados obtenidos en relación con el comportamiento asintótico de los recortes imparciales. En la primera sección se aborda la consistencia en el caso de recortes en una y dos muestras sin restricciones utilizando la distancia $\mathcal{W}_2$. A continuación, en la Sección 5.2, se obtiene la distribución asintótica asociada a la distancia $\mathcal{W}_2$ cuando se recorta con el mismo patrón. Las ideas manejadas en esta sección se usan para extender el resultado cuando recortamos con el mismo patrón y manejamos la distancia a una familia de distribuciones de localización y escala. Esto nos permite obtener la distribución asintótica del correspondiente estadístico en el caso normal, a lo que dedicamos la Sección 5.3. En la sección siguiente ilustramos la aplicación práctica de los anteriores resultados mediante ejemplos y simulaciones. En la Sección 5.5 se dan algunos resultados parciales relacionados con tasas de convergencia en el caso del recorte unilateral sin restricciones. Esta sección concluye con un resultado que explica el sobreajuste que se observa en el proceso (cuantil) empírico recortado. Dicho resultado sirve como base para el desarrollo de una metodología bootstrap que se presenta en la última sección.

5.1. Consistencia para recortes con diferentes patrones.

Habitualmente los resultados de consistencia que se obtienen tanto en la literatura de recortes imparciales como en la literatura relacionada con estadística robusta suponen unicidad. En nuestro caso, y gracias a los resultados obtenidos en la Subsección 3.3.2 no necesitamos de dicha hipótesis.

En primer lugar se dan unos resultados previos que nos permitirán demostrar la consistencia de los recortes en $\mathbb{R}^k$ para la distancia $\mathcal{W}_2$ de Wasserstein.

Lema 5.1. *Sean $P, \{P_n\}_n \in \mathcal{P}_p(\mathbb{R}^k)$, $p \geq 1$, tal que $\mathcal{W}_p(P_n, P) \to 0$ y sea $\{P_{n,\alpha}\}_n$ una sucesión de medidas de probabilidad tal que $P_{n,\alpha} \in \mathcal{R}_\alpha(P_n)$ y $P_{n,\alpha} \to_w P_0$. Entonces, $\mathcal{W}_p(P_{n,\alpha}, P_0) \to 0$.*

Demostración. En la Subsección 3.3.1, en los comentarios que preceden a la Proposición 3.21, vimos que $\mathcal{R}_\alpha(P) \subset \mathcal{P}_p$ si $P \in \mathcal{P}_p$. En estas condiciones, tanto $P_{n,\alpha}$ como P_0 ($P_0 \in \mathcal{R}_\alpha(P)$ en virtud del Teorema 3.7) tienen momentos de orden p finitos. De esta forma, y en virtud del Lema 2.3, lo único que tenemos que ver es que $\|x\|^p$ es uniformemente $P_{n,\alpha}$-integrable. Pero esto es inmediato pues al darse $\mathcal{W}_p(P_n, P) \to 0$ tenemos que $\|x\|^p$ es uniformemente P_n-integrable, y fijado $t > 0$,

$$\int_{\|x\|>t} \|x\|^p P_{n,\alpha}(dx) = \int_{\|x\|>t} \|x\|^p \frac{dP_{n,\alpha}}{dP_n}(x) P_n(dx) \leq \frac{1}{1-\alpha} \int_{\|x\|>t} \|x\|^p P_n(dx).$$

Luego $\|x\|^p$ es uniformemente $P_{n,\alpha}$-integrable. ∎

El siguiente resultado puede probarse en su versión para la recta real utilizando la caracterización dada por el Teorema 3.15, mientras que para su versión más general en $\mathbb{R}^k$ se precisa del Teorema 3.16. Se presentan aquí ambos resultados (y no sólo el más general) con fines ilustrativos.

Proposición 5.2. *Sean P una medida de probabilidad en $\mathbb{R}$, y $\{P_n\}_n$ una sucesión de medidas de probabilidad reales tales que $P_n \to_w P$. Entonces, para cada $Q \in \mathcal{R}_\alpha(P)$ existe una sucesión $\{Q_n\}_n$ tal que $Q_n \to_w Q$ y $Q_n \in \mathcal{R}_\alpha(P_n)$ para cada $n \in \mathbb{N}$.*

Demostración. Aplicando el Teorema 3.15 tendremos que si F y G son las funciones de distribución de P y Q respectivamente, entonces existe una función absolutamente continua $h_0 \in \mathcal{C}_\alpha$, tal que $G(x) = h_0(F(x))$ para casi todo $x \in \mathbb{R}$. Definiendo $G_n(x) = h_0(F_n(x))$, donde F_n es la función de distribución de P_n para cada n, tenemos que si Q_n es la medida de probabilidad con función de distribución G_n, entonces es inmediato que $Q_n \in \mathcal{R}_\alpha(P_n)$. Además, como h_0 es continua, $h_0(F_n(x)) \to h_0(F(x))$ para cada x de continuidad de F y por tanto $Q_n \to_w Q$. ∎

Proposición 5.3. *Sean P una medida de probabilidad en $\mathbb{R}^k$ absolutamente continua, y $\{P_n\}_n$ una sucesión de medidas de probabilidad tales que $P_n \to_w P$. Entonces, para cada $Q \in \mathcal{R}_\alpha(P)$ existe una sucesión $\{Q_n\}_n$ tal que $Q_n \to_w Q$ y $Q_n \in \mathcal{R}_\alpha(P_n)$ para cada $n \in \mathbb{N}$.*

Demostración. Para cada n, sea T_n la función de transporte óptimo entre P y P_n. Como $P \ll \ell^k$, si $Q \in \mathcal{R}_\alpha(P)$, utilizando el Teorema 3.16 tenemos que definiendo $Q_n = Q \circ T_n^{-1}$, entonces $Q_n \in \mathcal{R}_\alpha(P_n)$. Además, como $P_n \to_w P$, aplicando el Teorema 3.4 en Cuesta Albertos et al. (1997c), tenemos la convergencia de las funciones de transporte óptimo $T_n \to Id$ P-c.s., donde Id es la identidad (función de transporte óptimo entre P y P). Luego también $T_n \to Id$ Q-c.s. , y si X es una variable aleatoria tal que $\mathcal{L}(X) = Q$, entonces $\mathcal{L}(T_n(X)) \to \mathcal{L}(X)$, y por tanto $Q_n \to_w Q$. ∎

Con estos resultados y la unicidad estamos en condiciones de demostrar la consistencia tanto si recortamos en un lado únicamente como si lo hacemos en ambos (Teorema 5.6). Este resultado es una consecuencia inmediata de los teoremas 5.4 y 5.5 que demostramos a continuación y que podrían tener interés independiente.

Teorema 5.4 (Recorte unilateral)**.** *Sean P, Q y $\{P_n\}_n$ medidas de probabilidad en $\mathcal{P}_2(\mathbb{R}^k)$ tal que $\mathcal{W}_2(P_n, P) \to 0$. Se tiene que,*

(a) *Si $Q \ll \ell^k$ y $P_{n,\alpha} := \underset{R_n \in \mathcal{R}_\alpha(P_n)}{\arg\min}\ \mathcal{W}_2(R_n, Q)$, entonces*

$$\mathcal{W}_2(P_{n,\alpha}, P_\alpha) \to 0, \ \textit{donde}\ P_\alpha := \underset{R \in \mathcal{R}_\alpha(P)}{\arg\min}\ \mathcal{W}_2(R, Q).$$

(b) *Si $P \ll \ell^k$ y $Q_{n,\alpha} \in \mathcal{R}_\alpha(Q)$ verifica que $\mathcal{W}_2(P_n, Q_{n,\alpha}) = \min_{R \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(P_n, R)$, entonces*

$$\mathcal{W}_2(Q_{n,\alpha}, Q_\alpha) \to 0, \ \textit{donde}\ Q_\alpha := \underset{R \in \mathcal{R}_\alpha(Q)}{\arg\min}\ \mathcal{W}_2(P, R).$$

Demostración. Los dos apartados tienen una demostración similar por lo que consideraremos sólo el apartado (a). Como $P_n \to_w P$ ($\mathcal{W}_2$ metriza la convergencia débil), entonces $\{P_n\}_n$ es *tight* y, en consecuencia (Teorema 3.7), también lo es $\{P_{n,\alpha}\}_n$. Así pues, para cada subsucesión de $\{P_{n,\alpha}\}_n$ podemos encontrar una subsucesión convergente. Veamos que todas ellas convergen al mismo elemento y este es P_α, único en virtud del Teorema 3.26 (puesto que $Q \ll \ell^k$). Sea una de estas subsucesiones $\{P_{n_k,\alpha}\}_{n_k}$ de forma que $P_{n_k,\alpha} \to_w P_1$,

donde $P_1 \in \mathcal{R}_\alpha(P)$ en virtud del Teorema 3.7. Si $P_1 \neq P_\alpha$, entonces, por la unicidad, $\mathcal{W}_2(P_\alpha, Q) < \mathcal{W}_2(P_1, Q)$. Aplicando la Proposición 5.3 y el Lema 5.1, podemos encontrar una sucesión $\{R_{n,\alpha}\}_n$ de forma que $R_{n,\alpha} \in \mathcal{R}_\alpha(P_n)$ para cada n y $\mathcal{W}_2(R_{n,\alpha}, P_\alpha) \to 0$. Finalmente, aplicando la desigualdad triangular de la distancia de Wasserstein tendremos que,

$$\mathcal{W}_2(P_\alpha, Q) < \mathcal{W}_2(P_1, Q) \leq \mathcal{W}_2(P_1, P_{n_k,\alpha}) + \mathcal{W}_2(P_{n_k,\alpha}, Q) \leq \mathcal{W}_2(P_1, P_{n_k,\alpha}) + \mathcal{W}_2(R_{n_k,\alpha}, Q),$$

y como $\mathcal{W}_2(P_1, P_{n_k,\alpha}) \to 0$ (en virtud del Lema 5.1) y $\mathcal{W}_2(R_{n_k,\alpha}, Q) \to \mathcal{W}_2(P_\alpha, Q)$, llegamos a un absurdo y tenemos que $P_{n,\alpha} \to_w P_\alpha$. Y por tanto, $\mathcal{W}_2(P_{n,\alpha}, P_\alpha) \to 0$ (aplicando nuevamente el Lema 5.1). ■

En el anterior resultado la condición de que Q (ó P en el segundo apartado) sea absolutamente continua puede relajarse si estamos en la recta real (ver la Proposición 5.2).

De forma similar se prueba el correspondiente resultado cuando recortamos en los dos lados.

Teorema 5.5 (Recorte bilateral). *Sean P, Q, $\{P_n\}_n$ y $\{Q_n\}_n$ medidas de probabilidad en $\mathcal{P}_2(\mathbb{R}^k)$, tales que $P \ll \ell^k$, $\mathcal{W}_2(P_n, P) \to 0$ y $\mathcal{W}_2(Q_n, Q) \to 0$.*

Si $P_{n,\alpha} \in \mathcal{R}_\alpha(P_n)$ y $Q_{n,\alpha} \in \mathcal{R}_\alpha(Q_n)$ son tales que

$$\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha}) = \min_{R_{n,1} \in \mathcal{R}_\alpha(P_n), R_{n,2} \in \mathcal{R}_\alpha(Q_n)} \mathcal{W}_2(R_{n,1}, R_{n,2}),$$

y $P_\alpha \in \mathcal{R}_\alpha(P)$, $Q_\alpha \in \mathcal{R}_\alpha(Q)$ son tales que

$$\mathcal{W}_2(P_\alpha, Q_\alpha) = \min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2) > 0,$$

entonces $\mathcal{W}_2(P_{n,\alpha}, P_\alpha) \to 0$ y $\mathcal{W}_2(Q_{n,\alpha}, Q_\alpha) \to 0$.

Demostración. Tanto $\{P_n\}_n$ como $\{Q_n\}_n$ son *tight* y, por tanto, aplicando el Teorema 3.7, también lo son $\{P_{n,\alpha}\}_n$ y $\{Q_{n,\alpha}\}_n$. Veamos que convergen respectivamente a P_α y Q_α, y para ello razonamos por reducción al absurdo como en el Teorema 5.4. Si el resultado no fuese cierto, existirían dos subsucesiones con el mismo conjunto de índices tales que, $P_{n_k,\alpha} \to_w P_1$, $Q_{n_k,\alpha} \to_w Q_1$, y $(P_\alpha, Q_\alpha) \neq (P_1, Q_1)$. Por la unicidad del límite (ver Teorema 3.29), y dado que $P_1 \in \mathcal{R}_\alpha(P)$ y $Q_1 \in \mathcal{R}_\alpha(Q)$ (en virtud del Teorema 3.7), tendremos que $\mathcal{W}_2(P_\alpha, Q_\alpha) < \mathcal{W}_2(P_1, Q_1)$. Aplicando la Proposición 5.3 y el Lema 5.1, podemos encontrar sucesiones

$\{R_{n,\alpha}\}_n$ y $\{S_{n,\alpha}\}_n$ de forma que $R_{n,\alpha} \in \mathcal{R}_\alpha(P_n)$ y $S_{n,\alpha} \in \mathcal{R}_\alpha(Q_n) \quad \forall n$ y $\mathcal{W}_2(R_{n,\alpha}, P_\alpha) \to 0$ y $\mathcal{W}_2(S_{n,\alpha}, Q_\alpha) \to 0$. Utilizando ahora la desigualdad triangular de la distancia de Wasserstein tendremos que,

$$\begin{aligned} \mathcal{W}_2(P_\alpha, Q_\alpha) < \mathcal{W}_2(P_1, Q_1) &\leq \mathcal{W}_2(P_1, P_{n_k,\alpha}) + \mathcal{W}_2(P_{n_k,\alpha}, Q_{n_k,\alpha}) + \mathcal{W}_2(Q_{n_k,\alpha}, Q_1) \\ &\leq \mathcal{W}_2(P_1, P_{n_k,\alpha}) + \mathcal{W}_2(R_{n_k,\alpha}, S_{n_k,\alpha}) + \mathcal{W}_2(Q_{n_k,\alpha}, Q_1), \end{aligned}$$

y como $\mathcal{W}_2(P_1, P_{n_k,\alpha}) \to 0, \mathcal{W}_2(Q_{n_k,\alpha}, Q_1) \to 0$ y $\mathcal{W}_2(R_{n_k,\alpha}, S_{n_k,\alpha}) \to \mathcal{W}_2(P_\alpha, Q_\alpha)$, llegamos a un absurdo y tenemos el resultado. ■

Si en el resultado anterior, tomamos $Q_n = Q$ para cada n, entonces tenemos un resultado equivalente al apartado (a) del Teorema 5.4 pero para dos muestras.

La Ley Fuerte de los Grandes Números, el Teorema de Glivenko-Cantelli y el Lema 2.3 garantizan, a través de un argumento de integrabilidad uniforme, (ver, p.e., Cuesta Albertos y Matrán, 1986), que cuando $\{P_n^\omega\}_n$ es la sucesión de distribuciones de probabilidad empíricas basadas en la sucesión $\{X_n\}_n$ de vectores aleatorios independientes e igualmente distribuidos, con ley de probabilidad $P \in \mathcal{P}_2(\mathbb{R}^k)$, entonces $\mathcal{W}_2(P_n^\omega, P) \to 0$ para casi todo ω. Por ello el siguiente teorema sobre la consistencia de los recortes mejores aproximantes es inmediato a partir de los teoremas 5.4 y 5.5. Este resultado permite el uso de simulaciones de tipo Monte-Carlo para aproximar las medidas de disimilitud entre probabilidades dadas en el Capítulo 3 por las expresiones (3.5)-(3.7).

Teorema 5.6 (Consistencia). *Sean $\{X_n\}_n$, $\{Y_n\}_n$ dos sucesiones de vectores aleatorios i.i.d., definidos en (Ω, σ, ν), tales que $\mathcal{L}(X_n) = P$, $\mathcal{L}(Y_n) = Q$ y $P, Q \in \mathcal{P}_2(\mathbb{R}^k)$. Sean P_n^ω, Q_n^ω las distribuciones empíricas basadas en las muestras $\{X_1(\omega), ...X_n(\omega)\}$ y $\{Y_1(\omega), ...Y_n(\omega)\}$.*

(a) *Si $Q \ll \ell^k$ y $P_{n,\alpha}^\omega := \underset{R \in \mathcal{R}_\alpha(P_n^\omega)}{\arg\min}\ \mathcal{W}_2(R, Q)$, entonces*

$$\mathcal{W}_2(P_{n,\alpha}^\omega, P_\alpha) \to 0 \quad \nu\text{-c.s., donde } P_\alpha := \underset{R \in \mathcal{R}_\alpha(P)}{\arg\min}\ \mathcal{W}_2(R, Q).$$

(b) *Si $P \ll \ell^k$ y $Q_{n,\alpha}^\omega \in \mathcal{R}_\alpha(Q)$ verifica que $\mathcal{W}_2(P_n^\omega, Q_{n,\alpha}^\omega) = \min_{R \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(P_n^\omega, R)$, entonces*

$$\mathcal{W}_2(Q_{n,\alpha}^\omega, Q_\alpha) \to 0 \quad \nu\text{-c.s., donde } Q_\alpha := \underset{R \in \mathcal{R}_\alpha(Q)}{\arg\min}\ \mathcal{W}_2(P, R).$$

(c) *Si* P *ó* $Q \ll \ell^k$, *y* $P_{n,\alpha}^{\omega} \in \mathcal{R}_\alpha(P_n^{\omega})$ *y* $Q_{n,\alpha}^{\omega} \in \mathcal{R}_\alpha(Q)$ *verifican*

$$\mathcal{W}_2(P_{n,\alpha}^{\omega}, Q_{n,\alpha}^{\omega}) = \min_{R_1 \in \mathcal{R}_\alpha(P_n^{\omega}), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2),$$

y P_α *y* Q_α *son tal que*

$$\mathcal{W}_2(P_\alpha, Q_\alpha) = \min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2) > 0.$$

entonces $\mathcal{W}_2(P_{n,\alpha}^{\omega}, P_\alpha) \to 0$ *y* $\mathcal{W}_2(Q_{n,\alpha}^{\omega}, Q_\alpha) \to 0$ ν*-c.s.*

(d) *Si* P *ó* $Q \ll \ell^k$, *y* $P_{n,\alpha}^{\omega} \in \mathcal{R}_\alpha(P_n^{\omega})$ *y* $Q_{n,\alpha}^{\omega} \in \mathcal{R}_\alpha(Q_n^{\omega})$ *verifican*

$$\mathcal{W}_2(P_{n,\alpha}^{\omega}, Q_{n,\alpha}^{\omega}) = \min_{R_1 \in \mathcal{R}_\alpha(P_n^{\omega}), R_2 \in \mathcal{R}_\alpha(Q_n^{\omega})} \mathcal{W}_2(R_1, R_2),$$

y P_α *y* Q_α *son tal que*

$$\mathcal{W}_2(P_\alpha, Q_\alpha) = \min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R_1, R_2) > 0.$$

entonces $\mathcal{W}_2(P_{n,\alpha}^{\omega}, P_\alpha) \to 0$ *y* $\mathcal{W}_2(Q_{n,\alpha}^{\omega}, Q_\alpha) \to 0$ ν*-c.s.*

5.2. Recorte con el mismo patrón. Distribución asintótica.

En la Subsección 3.2.2 se introdujo el problema de comparación de dos distribuciones de probabilidad en $\mathbb{R}$ mediante la "distancia recortada" T_4 (ver (3.14)). Vimos también que en el caso de manejar la distancia $\mathcal{W}_2$ de Wasserstein, T_4 se puede escribir (ver (3.16)) como

$$T_4(P, Q, \mathcal{W}_2, \alpha) = \inf_{h \in \mathcal{C}_\alpha} \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t) dt.$$

Como se probó allí, la solución de este problema de optimización es única en el caso de que la función de distribución $L_{F,G}$ definida en (3.17) sea continua en $L_{F,G}^{-1}(1-\alpha)$ y el minimizador, $h_0 \in \mathcal{C}_\alpha$, cumple que (ver (3.19)),

$$h_0'(t) = \frac{1}{1-\alpha} I_{[0, L_{F,G}^{-1}(1-\alpha)]}(|F^{-1}(t) - G^{-1}(t)|).$$

En esta sección estudiaremos el comportamiento asintótico de las distintas versiones empíricas de T_4. Como hasta ahora, $\mathcal{P}_s$ denotará el conjunto de medidas de probabilidad en $\mathbb{R}$ con momento de orden s finito y $\mathcal{P}_r' = \cup_{s>r} \mathcal{P}_s$. En lo que sigue también denotaremos por $\tau_\alpha(P, Q) = \tau_\alpha(F, G) := T_4(P, Q, \mathcal{W}_2, \alpha)$, donde $P, Q \in \mathcal{P}_2$, y F y G son sus respectivas funciones de distribución.

Sea $X_1, \ldots, X_n$ (resp. $Y_1, \ldots, Y_m$) una muestra aleatoria simple con función de distribución común F (resp. G) y con función de distribución empírica asociada F_n (resp. G_m). Definimos $T_{n,\alpha} := \tau_\alpha(F_n, G)$, para los problemas de una muestra, y, $T_{n,m,\alpha} := \tau_\alpha(F_n, G_m)$ para los problemas de dos muestras. Los resultados que aparecen en esta sección muestran que bajo condiciones débiles sobre F y G, $T_{n,\alpha}$ y $T_{n,m,\alpha}$ son asintóticamente normales. Estos resultados se usarán más adelante (ver el Ejemplo 1 en este mismo Capítulo) para aproximar los valores críticos del test $H_0 : \tau_\alpha(F, G) \geq \Delta_0^2$ contra $H_a : \tau_\alpha(F, G) < \Delta_0^2$, donde Δ_0 es un valor umbral preestablecido.

Comenzamos con un par de lemas técnicos previos.

Lema 5.7. *Si $F, G \in \mathcal{P}_4'$ y g es la función de densidad de G, entonces*

(i) $\sqrt{n}\int_0^{1/n}(F^{-1}(t))^2dt \to 0$ *y* $\sqrt{n}\int_{1-1/n}^{1}(F^{-1}(t))^2dt \to 0.$

(ii) $\sqrt{n}\int_0^{1/n}(F_n^{-1}(t))^2dt \to 0$ *y* $\sqrt{n}\int_{1-1/n}^{1}(F_n^{-1}(t))^2dt \to 0$ *en probabilidad.*

(iii) $\int_0^1 \frac{\sqrt{t(1-t)}}{g(G^{-1}(t))}|F^{-1}(t) - G^{-1}(t)|dt < \infty.$

Además, si G satisface la condición

$$\sup_{x\in\mathbb{R}}\left|\frac{G(x)(1-G(x))g'(x)}{g^2(x)}\right| < \infty, \tag{5.1}$$

entonces

(iv) $\frac{1}{\sqrt{n}}\int_{1/n}^{1-1/n}\frac{t(1-t)}{g^2(G^{-1}(t))}dt \to 0.$

Demostración.

(i) Para la primera integral la desigualdad de Schwarz da

$$\begin{aligned} 0 &\leq \sqrt{n}\int_0^{\frac{1}{n}}(F^{-1}(t))^2dt \leq \sqrt{n}\left(\int_0^{\frac{1}{n}}(F^{-1}(t))^4dt\right)^{1/2}\left(\int_0^{\frac{1}{n}}1dt\right)^{1/2} \\ &= \left(\int_0^{\frac{1}{n}}(F^{-1}(t))^4dt\right)^{1/2} = \left(\int_0^1(F^{-1}(t))^4 I_{[0,1/n]}(t)dt\right)^{1/2}. \end{aligned} \tag{5.2}$$

Ahora, como F tiene momento de orden 4 finito entonces $\int_0^1(F^{-1}(t))^4dt < \infty$. El resultado se obtiene entonces aplicando el Teorema de la Convergencia Dominada al último término de (5.2).

La segunda convergencia es completamente similar.

(ii) Consideremos ahora la segunda expresión. Podemos asumir sin pérdida de generalidad que F está concentrada en la parte positiva de la recta real, porque el otro caso es trivial. Como $\sqrt{n}\int_{1-1/n}^{1}(F_n^{-1}(t))^2dt = \left(n^{-1/4}\,\text{máx}_{1\le i\le n}\,X_i\right)^2$, basta con que probemos que

$$n^{-1/4}\max_{1\le i\le n} X_i \to 0$$

en probabilidad, o, equivalentemente, que $F(\varepsilon n^{1/4})^n \to 1$ para todo $\varepsilon > 0$. Tomando logaritmos y utilizando un infinitésimo para el logaritmo, tal y como se hace habitualmente en la teoría de extremos (ver, p.e., Resnick, 1987), tenemos que

$$n\log(F(\varepsilon n^{1/4})) \simeq -n(1-F(\varepsilon n^{1/4})).$$

Luego basta con que veamos que $n(1-F(\varepsilon n^{1/4})) \to 0$ cuando $n\to\infty$.

Ahora, si X es una v.a. con función de distribución F, como $\mathrm{E}X^r < \infty$ para $r = 4$ y $0 \le X^r I_{(x,\infty)}(X) \le X^r$, aplicando el Teorema de la Convergencia Dominada se tiene que

$$\lim_{x\to\infty} x^r P(X>x) \le \lim_{x\to\infty}\int_0^\infty X^r I_{(x,\infty)}(X)dP = \int_0^\infty \left(\lim_{x\to\infty} X^r I_{(x,\infty)}(X)\right)dP = 0. \quad (5.3)$$

Finalmente, el resultado se obtiene a través del cambio de variable $x = \varepsilon n^{1/4}$ en

$$\lim_{n\to\infty} n(1-F(\varepsilon n^{1/4})) = \lim_{n\to\infty} nP(X>\varepsilon n^{1/4}) = \lim_{x\to\infty}\varepsilon^{-1}x^4P(X>x) = 0.$$

(iii) Suponemos esta vez que G está concentrada en la parte positiva de la recta real. Si realizamos el cambio de variable $y = G^{-1}(t)$ tenemos que

$$\begin{aligned}\int_0^1 \frac{\sqrt{t(1-t)}}{g(G^{-1}(t))}|F^{-1}(t)-G^{-1}(t)|dt &= \int_0^\infty \sqrt{G(y)(1-G(y))}|F^{-1}(G(y))-y|dy \\ &\le \int_0^\infty \sqrt{(1-G(y))}|F^{-1}(G(y))|dy + \int_0^\infty y\sqrt{(1-G(y))}dy.\end{aligned}$$

Fijamos $r > 4$ tal que F, G tienen momento de orden r finito. Ahora, aplicando la desigualdad de Markov tenemos que $1-G(y) \le \frac{\mu_r}{y^r}$ donde μ_r es el momento de orden r de G. Por lo que $y\sqrt{1-G(y)} \le y\frac{C}{y^{r/2}} = Cy^{(r-2)/2}$ para una constante $C > 0$. Como $(r-2)/2 > 1$, entonces $y^{(r-2)/2}$ es integrable y por tanto $\int_0^\infty y\sqrt{(1-G(y))}dy < \infty$. Así pues, sólo nos queda probar que

$$\int_0^\infty \sqrt{1-G(y)}|F^{-1}(G(y))|dy < \infty.$$

Para ello, usando el mismo argumento que en (5.3), esta vez para $r > 4$, tenemos que

$$\lim_{t\to 1}(1-t)|F^{-1}(t)|^r = \lim_{y\to\infty} y^r(1-F(y)) = \lim_{y\to\infty} y^r P(X>y) = 0.$$

Por tanto, para valores grandes de y tenemos por una parte que $|F^{-1}(G(y))| \leq (1-G(y))^{-1/r}$ lo que combinado con la desigualdad de Markov nos da que

$$\sqrt{1-G(y)}|F^{-1}(G(y))| \leq (1-G(y))^{(r-2)/2r} \leq \mu_r^{(r-2)/2r} y^{-(r-2)/2}.$$

Finalmente el resultado se sigue del hecho de que $(r-2)/2 > 1$.

(iv) Suponemos para simplificar que G tiene como soporte $(0,\infty)$. Con el cambio de variable $y = G^{-1}(t)$, si hacemos $x = G^{-1}(1-1/n)$, tenemos

$$\begin{aligned}\frac{1}{\sqrt{n}}\int_{1/n}^{1-1/n} \frac{t(1-t)}{g^2(G^{-1}(t))}dt &= \frac{1}{\sqrt{n}}\int_{G^{-1}(1/n)}^{G^{-1}(1-1/n)} \frac{G(y)(1-G(y))}{g(y)}dy \\ &\leq \frac{1}{\sqrt{n}}\int_0^{G^{-1}(1-1/n)} \frac{(1-G(y))}{g(y)} = \sqrt{1-G(x)}\int_0^x \frac{(1-G(y))}{g(y)}dy.\end{aligned}$$

Por tanto podemos reducir la demostración a probar que

$$\sqrt{1-G(x)}\int_0^x \frac{(1-G(y))}{g(y)}dy \to 0 \text{ cuando } x \to \infty.$$

Observando ahora que $h(y) = \frac{(1-G(y))}{g(y)}$ tiene derivada $h'(y) = -1 - \frac{(1-G(y))g'(y)}{g^2(y)}$ que, por (5.1), está uniformemente acotada, tenemos que $h(y)$ es sublineal. Entonces,

$$\limsup_{x\to\infty} \sqrt{1-G(x)}\int_0^x \frac{(1-G(y))}{g(y)}dy \leq K \limsup_{x\to\infty} \sqrt{1-G(x)}x^2 = 0,$$

puesto que G tiene momento de orden 4 finito (usando nuevamente (5.3)). Lo que concluye la demostración. ■

Definimos el siguiente conjunto

$$\mathcal{C}_\alpha(F,G) = \left\{h \in \mathcal{C}_\alpha : \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt = 0\right\}. \tag{5.4}$$

Observamos que $\mathcal{C}_\alpha(F,F) = \mathcal{C}_\alpha$, pero para $F \neq G$ tenemos que $\mathcal{C}_\alpha(F,G)$ es un subconjunto de $\mathcal{C}_\alpha$. También se cumple que $\mathcal{C}_\alpha(F,G) \neq \emptyset$ si y sólo si $\tau_\alpha(F,G) = 0$. De hecho, el tamaño de $\mathcal{C}_\alpha(F,G)$ depende de la medida de Lebesgue del conjunto $\{t \in (0,1) : F^{-1}(t) \neq G^{-1}(t)\}$. Y $\tau_\alpha(F,G) = 0$ si y sólo si la medida de Lebesgue de este último conjunto es menor o igual que α; si es igual a α entonces $\mathcal{C}_\alpha(F,G)$ está formado por una única función, h, tal que $h'(t) = \frac{1}{1-\alpha} I_{(F^{-1}=G^{-1})}(t)$. El Lema 5.8 a continuación prueba que $\mathcal{C}_\alpha(F,G)$ es compacto para la topología $\|\cdot\|_\infty$.

Lema 5.8. *Si $F, G \in \mathcal{P}_2$ entonces el conjunto $\mathcal{C}_\alpha(F,G)$ definido en (5.4) es compacto para la topología $\|\cdot\|_\infty$.*

Demostración. Basta con probar que $\mathcal{C}_\alpha(F,G)$ es cerrado, ya que $\mathcal{C}_\alpha(F,G) \subset \mathcal{C}_\alpha$ y $\mathcal{C}_\alpha$ es compacto por el Lema 3.14. Esto podemos reducirlo a probar que

$$\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h_n'(t)dt \to \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)dt,$$

siempre que $h_n \in \mathcal{C}_\alpha$ y $\|h_n - h\|_\infty \to 0$. O, de forma equivalente, a probar que

$$\int_0^1 (F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 dy \to \int_0^1 (F^{-1}(h^{-1}(y)) - G^{-1}(h^{-1}(y))^2 dy. \tag{5.5}$$

Por la continuidad de F^{-1} y G^{-1} (excepto, quizás, en un conjunto numerable de puntos) tenemos $(F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 \to (F^{-1}(h^{-1}(y)) - G^{-1}(h^{-1}(y))^2$ en casi todo $y \in (0,1)$. Para probar (5.5) sólo queda ver la integrabilidad uniforme de $(F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2$. Pero esto se deduce de la siguiente desigualdad

$$\begin{aligned}
&\sup_n \int_{\{(F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 > x\}} (F^{-1}(h_n^{-1}(y)) - G^{-1}(h_n^{-1}(y))^2 dy \\
&\quad = \sup_n \int_{\{(F^{-1}(t) - G^{-1}(t))^2 > x\}} (F^{-1}(t) - G^{-1}(t))^2 h_n'(t)dt \\
&\quad \leq \frac{1}{1-\alpha} \int_{\{(F^{-1}(t) - G^{-1}(t))^2 > x\}} (F^{-1}(t) - G^{-1}(t))^2 dt.
\end{aligned}$$

■

Teorema 5.9 (Una muestra). *Supongamos que F, $G \in \mathcal{P}_4'$, G^{-1} es continua, $L_{F,G}$ es continua en $L_{F,G}^{-1}(1-\alpha)$ y que F tiene una densidad continuamente diferenciable f tal que*

$$\sup_{x \in \mathbb{R}} \left| \frac{F(x)(1-F(x))f'(x)}{f^2(x)} \right| < \infty. \tag{5.6}$$

Entonces $\sqrt{n}(T_{n,\alpha} - \tau_\alpha(F,G))$ es asintóticamente normal con media cero y con varianza

$$\sigma_\alpha^2(F,G) = 4\left(\int_0^1 l^2(t)dt - \left(\int_0^1 l(t)dt \right)^2 \right), \tag{5.7}$$

donde

$$l(t) = \int_{F^{-1}(1/2)}^{F^{-1}(t)} (x - G^{-1}(F(x)))h_0'(F(x))dx,$$

y h_0 está determinada por (3.19).

Esta varianza asintótica puede ser estimada de forma consistente por

$$S^2_{n,\alpha}(G) = \frac{4}{(1-\alpha)^2}\frac{1}{n}\sum_{i,j=1}^{n-1}(i \wedge j - \tfrac{ij}{n})a_{n,i}a_{n,j},$$

donde

$$a_{n,i} = (X_{(i+1)} - X_{(i)})((X_{(i+1)} + X_{(i)})/2 - G^{-1}(\tfrac{i}{n}))I_{[0,L^{-1}_{F_n,G}(1-\alpha)]}(|X_{(i)} - G^{-1}(\tfrac{i}{n})|).$$

Demostración. Denotaremos por $\rho_n(t) = \sqrt{n}f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$ al proceso cuantil ponderado, donde f es la función de densidad de F.

Podemos trabajar en un espacio probabilístico lo suficientemente rico como para que existan versiones de $\{X_n\}_n$ y puentes brownianos B_n que satisfagan

$$n^{1/2-\nu}\sup_{\frac{1}{n}\le t\le 1-\frac{1}{n}}\frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} = \begin{cases} O_P(\log n), & \text{if } \nu = 0 \\ O_P(1), & \text{si } 0 < \nu \le 1/2 \end{cases}. \tag{5.8}$$

La existencia de tal espacio probabilístico es una consecuencia de (5.1) (ver, por ejemplo el Teorema 6.2.1 en Csörgő y Horváth, 1993).

Si llamamos $M_n(h) = \sqrt{n}\int_0^1 (F_n^{-1}(t) - G^{-1}(t))^2 h'(t)dt$ y

$$\begin{aligned} N_n(h) &= 2\int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{B_n(t)}{f(F^{-1}(t))}(G^{-1}(t) - F^{-1}(t))h'(t)dt \\ &\quad + \sqrt{n}\int_{\frac{1}{n}}^{1-\frac{1}{n}}(G^{-1}(t) - F^{-1}(t))^2 h'(t)dt. \end{aligned}$$

Podemos observar que

$$\begin{aligned} \sup_{h\in\mathcal{C}_\alpha}|M_n(h) - N_n(h)| &\le \sqrt{n}\int_0^{\frac{1}{n}}(F_n^{-1}(t) - G^{-1}(t))^2 dt \\ &\quad + \sqrt{n}\int_{1-\frac{1}{n}}^{1}(F_n^{-1}(t) - G^{-1}(t))^2 dt \\ &\quad + \frac{1}{\sqrt{n}}\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{|\rho_n(t) - B_n(t)|^2}{f^2(F^{-1}(t))}dt + \\ &\quad \frac{1}{\sqrt{n}}\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{B_n(t)^2}{f^2(F^{-1}(t))}dt \\ &\quad + 2\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{|\rho_n(t) - B_n(t)|}{f(F^{-1}(t))}|G^{-1}(t) - F^{-1}(t)|dt \\ &=: A_{n,1} + A_{n,2} + A_{n,3} + A_{n,4} + A_{n,5}. \end{aligned}$$

El hecho de que $F, G \in \mathcal{P}_4'$ y el Lema 5.7 implican que $A_{n,1} \to 0$ y $A_{n,2} \to 0$ en probabilidad. De (5.8) tenemos que

$$A_{n,3} \leq O_P(1)\frac{1}{\sqrt{n}}\int_{1/n}^{1-1/n}\frac{t(1-t)}{f^2(F^{-1}(t))}dt$$

y la última integral converge a 0 por el Lema 5.7. En consecuencia, $A_{n,3} \to 0$ en probabilidad. De forma similar, $A_{n,4} \to 0$ en probabilidad. Finalmente, (5.8) nos da

$$A_{n,5} \leq O_P(1)n^{\nu-1/2}\int_{\frac{1}{n}}^{1-\frac{1}{n}}\frac{(t(1-t))^{\nu}}{f(F^{-1}(t))}|G^{-1}(t)-F^{-1}(t)|dt$$

para algún $\nu \in (0, 1/2)$. El Lema 5.7 muestra que $\int_0^1 \frac{(t(1-t))^{1/2}}{f(F^{-1}(t))}|G^{-1}(t)-F^{-1}(t)|dt < \infty$. De esto y del Teorema de la Convergencia Dominada se deriva que el lado derecho de la última expresión tiende a 0 en probabilidad.

Juntando los anteriores estimadores vemos que $\sup_{h\in\mathcal{C}_\alpha}|M_n(h)-N_n(h)| \to 0$ en probabilidad y, en consecuencia, $\sqrt{n}(T_{n,\alpha}-R_{n,\alpha}) \to 0$ en probabilidad, donde $\sqrt{n}R_{n,\alpha} = \text{ínf}_{h\in\mathcal{C}_\alpha} N_n(h)$. De esta forma, la demostración estará completa si probamos que $\sqrt{n}(\tilde{R}_{n,\alpha} - \tau_\alpha(F,G))$ es asintóticamente $N(0, \sigma_\alpha^2(F,G))$, donde

$$\begin{aligned}\sqrt{n}\tilde{R}_{n,\alpha} &= \inf_{h\in\mathcal{C}_\alpha}\left[2\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}h'(t)dt\right.\\ &\quad\left.+\sqrt{n}\int_0^1 (G^{-1}(t)-F^{-1}(t))^2h'(t)dt\right]. \qquad (5.9)\end{aligned}$$

Si llamamos

$$\begin{aligned}h_n &= \arg\min_{h\in\mathcal{C}_\alpha}\left(\int_0^1 (F^{-1}(t)-G^{-1}(t))^2h'(t)dt\right.\\ &\quad\left.+\frac{2}{\sqrt{n}}\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}h'(t)dt\right).\end{aligned}$$

Claramente $h_n'(t) \to h_0'(t)$ para casi todo t. Más aún, la optimalidad de h_n prueba que $B_n \leq 0$, donde,

$$\begin{aligned}B_n &:= \sqrt{n}\tilde{R}_{n,\alpha} - \left(2\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}h_0'(t)dt\right.\\ &\quad\left.+\sqrt{n}\int_0^1 (G^{-1}(t)-F^{-1}(t))^2h_0'(t)dt\right),\end{aligned}$$

pero, por otra parte,

$$\begin{aligned}B_n &= \sqrt{n}\left(\int_0^1 (F^{-1}(t)-G^{-1}(t))^2h_n'(t)dt - \int_0^1 (F^{-1}(t)-G^{-1}(t))^2h_0'(t)dt\right)\\ &\quad+2\int_0^1 B(t)\frac{G^{-1}(t)-F^{-1}(t)}{f(F^{-1}(t))}(h_n'(t)-h_0'(t))dt =: B_{n,1}+B_{n,2}\end{aligned}$$

y $B_{n,1} \geq 0$ por la optimalidad de h_0, mientras que $B_{n,2} = o_P(1)$ por el Teorema de la Convergencia Dominada. Por tanto, $B_n \to 0$ en probabilidad, lo que demuestra que

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(F,G)) \to_w 2\int_0^1 B(t)\frac{G^{-1}(t) - F^{-1}(t)}{f(F^{-1}(t))}h_0'(t)dt. \tag{5.10}$$

Integrando por partes obtenemos

$$\int_0^1 B(t)\frac{G^{-1}(t) - F^{-1}(t)}{f(F^{-1}(t))}h_0'(t)dt = \int_0^1 l(t)dB(t)$$

y esto prueba la normalidad asintótica y la expresión (5.7) para la varianza.

Para la afirmación acerca del estimador de la varianza veamos que

$$S_{n,\alpha}^2 = 4\left(\int_0^1 l_n^2(t)dt - \left(\int_0^1 l_n(t)dt\right)^2\right),$$

donde

$$l_n(t) = \int_{F_n^{-1}(1/2)}^{F_n^{-1}(t)} (x - G^{-1}(F_n(x)))h_n'(F_n(x))dx$$

y

$$h_n(t) = \arg\min_{h\in\mathcal{C}_\alpha} \int (F_n^{-1} - G^{-1})^2 h'.$$

Reflejamos aquí sólo el cálculo de $\int_0^1 l_n(t)dt$ ya que de forma completamente similar se obtiene $\int_0^1 l_n^2(t)dt$, y operando, el resultado. Además, para simplificar la exposición cambiamos el extremo inferior de integración en $l_n(t)$ por $F_n^{-1}(1/n)$. Nótese que este cambio no afecta al resultado pues la diferencia entre ambas integrales es una constante que se cancela en el cálculo de $S_{n,\alpha}^2$.

$$\begin{aligned}
\int_0^1 l_n(t)dt &= \sum_{i=1}^{n-1}\int_{\frac{i}{n}}^{\frac{i+1}{n}} \left(\int_{F_n^{-1}(1/n)}^{F_n^{-1}(t)} \left(x - G^{-1}(F_n(x))\right) h_n'(F_n(x))dx\right) dt \\
&= \sum_{i=1}^{n-1}\int_{\frac{i}{n}}^{\frac{i+1}{n}} \left(\sum_{j=1}^{i}\int_{X_{(j)}}^{X_{(j+1)}} \left(x - G^{-1}(\tfrac{j}{n})\right)\frac{1}{1-\alpha} I_{[0,L_{F_n,G}^{-1}(1-\alpha)]}(|X_{(j)} - G^{-1}(\tfrac{j}{n})|)dx\right) dt \\
&= \frac{1}{n(1-\alpha)}\sum_{i=1}^{n-1}\sum_{j=1}^{i}(X_{(j+1)} - X_{(j)})\left(\tfrac{X_{(j+1)}+X_{(j)}}{2} - G^{-1}(\tfrac{j}{n})\right) I_{[0,L_{F_n,G}^{-1}(1-\alpha)]}(|X_{(j)} - G^{-1}(\tfrac{j}{n})|).
\end{aligned}$$

Para completar la demostración, basta con ver que, $l_n(t) \to l(t)$ para casi todo $t \in (0,1)$ con probabilidad 1, y la integrabilidad uniforme de $l_n^2(t)$.

De la Ley de los Grandes Números y las condiciones sobre F tenemos que, como variables aleatorias definidas en $(0,1)$, $|F_n^{-1}(t) - G^{-1}(t)| \to |F^{-1}(t) - G^{-1}(t)|$ para todo $t \in (0,1)$ con probabilidad 1. En consecuencia, $L_{F_n,G}(x) \to L_{F,G}(x)$ para todo x de continuidad de $L_{F,G}$,

y, por la continuidad de $L^{-1}_{F,G}$ (derivada de la continuidad de F^{-1} y G^{-1}), $L^{-1}_{F_n,G}(1-\alpha) \to L^{-1}_{F,G}(1-\alpha)$. Por tanto,

$$\begin{aligned} h_n'(F_n(x)) &= \frac{1}{1-\alpha} I_{\{|F_n^{-1}-G^{-1}|\le L^{-1}_{F_n,G}(1-\alpha)\}}(F_n(x)) = \frac{1}{1-\alpha} I_{\{|Id-G^{-1}\circ F_n|\le L^{-1}_{F_n,G}(1-\alpha)\}}(x) \\ &\to h_0'(F(x)) = \frac{1}{1-\alpha} I_{\{|Id-G^{-1}\circ F|\le L^{-1}_{F,G}(1-\alpha)\}}(x), \end{aligned}$$

para casi todo x con probabilidad 1. En consecuencia, para todo t y casi todo x, con probabilidad 1 se tiene,

$$\begin{aligned} g_n(x) := (x - G^{-1}(F_n(x)))h_n'(F_n(x)) I_{[F_n^{-1}(1/2),F_n^{-1}(t)]}(x) \\ \to (x - G^{-1}(F(x)))h_0'(F(x)) I_{[F^{-1}(1/2),F^{-1}(t)]}(x). \end{aligned}$$

Esta primera parte finaliza con la integrabilidad uniforme de $g_n(x)$. Puesto que $h_n' \le \frac{1}{1-\alpha}$, es suficiente con ver que $xI_{[F_n^{-1}(1/2),F_n^{-1}(t)]}(x)$ y $G^{-1}(F_n(x))I_{[F_n^{-1}(1/2),F_n^{-1}(t)]}(x)$ son uniformemente integrables, pero esto es inmediato puesto que ambas son funciones acotadas en el compacto $[F_n^{-1}(1/2), F_n^{-1}(t)]$.

Finalmente, la integrabilidad uniforme de $l_n^2(t)$ se sigue del hecho de que F y G tienen momento de orden 4 finito y la acotación,

$$\begin{aligned} |l_n(t)|^2 &\le \frac{1}{(1-\alpha)^2}\left[\int_{F_n^{-1}(1/2)}^{F_n^{-1}(t)} x dx - \int_{F_n^{-1}(1/2)}^{F_n^{-1}(t)} G^{-1}(F_n(x))dx\right]^2 \\ &\le \frac{2}{(1-\alpha)^2}\left[\frac{1}{2}\left(F_n^{-1}(t)^2 + F_n^{-1}(\tfrac{1}{2})^2\right)^2 \right. \\ &\qquad \left. + \left(\left(|G^{-1}(\tfrac{1}{2})| + |G^{-1}(t)|\right)\left(|F_n^{-1}(\tfrac{1}{2})| + |F_n^{-1}(t)|\right)\right)^2\right]. \end{aligned}$$

■

En el resultado anterior, (5.6) es una condición natural introducida por Csörgő y Révész (1978) para aproximar el proceso empírico cuantil general por procesos gaussianos. Dicha condición se verifica por las denominadas *densidades monótonas en las colas* entre las que están todas las de uso habitual (ver Parzen, 1979).

Teorema 5.10 (Dos muestras). *Bajo las condiciones del Teorema 5.9, si G satisface también (5.1) y $\frac{n}{n+m} \to \lambda \in (0,1)$ entonces $\sqrt{\frac{nm}{n+m}}(T_{n,m,\alpha} - \tau_\alpha(F,G))$ es asintóticamente una normal centrada con varianza $(1-\lambda)\sigma_\alpha^2(F,G) + \lambda\sigma_\alpha^2(G,F)$. Esta varianza asintótica puede ser estimada de forma consistente por $S^2_{n,m,\alpha} = \frac{m}{n+m}S^2_{n,\alpha}(G_m) + \frac{n}{n+m}S^2_{m,\alpha}(F_n)$.*

Demostración. La demostración de este resultado es muy similar a la del Teorema 5.9.

■

Si $\tau_\alpha(F,G) = 0$, entonces el Teorema 5.9 se reduce a $\sqrt{n}T_{n,\alpha} \to 0$ en probabilidad (obsérvese que $\tau_\alpha(F,G) = 0$ implica $(x - G^{-1}(F(x)))^2 h_0'(F(x)) = 0$ para casi todo x y, por tanto, $\sigma_\alpha^2(F,G) = 0$). Aunque en general esto sería suficiente para las aplicaciones, el Teorema 5.12 da la tasa exacta de convergencia así como la distribución límite. Previamente necesitamos de un lema técnico.

Lema 5.11. *Sea (X,d) un espacio métrico compacto, $A \subset X$ compacto y $\{f_n\}$, f funciones de valores reales y continuas en X tal que*

(i) $\sup_{x\in A} |f_n(x) - f(x)| \to 0$, *cuando* $n \to \infty$,

(ii) *para* $x \in X - A$ *existe* $\varepsilon_x > 0$ *tal que* $\inf_{d(y,x)<\varepsilon_x} f_n(y) \to \infty$, *cuando* $n \to \infty$,

(iii) *si* $x_n \to x \in A$ *existe una subsucesión,* $\{x_m\}$, *tal que* $f_m(x_m) \to f(x)$.

Entonces,

$$\min_{x\in X} f_n(x) \to \min_{x\in A} f(x).$$

Demostración. Elegimos x_n tal que $f_n(x_n) = \min_{x\in X} f_n(x)$. Entonces existe una subsucesión convergente $x_m \to x_0 \in X$. Si $x_0 \in X - A$ entonces fijamos $\varepsilon > 0$ tal que $\inf_{d(y,x_0)<\varepsilon} f_n(y) \to \infty$. Puesto que $x_m \to x_0$ y $\min_{x\in A} f_n(x) \to \min_{x\in A} f(x)$ tenemos que $f_m(x_m) > 2\min_{x\in A} f_m(x)$ para m suficientemente grande, lo que contradice la elección de x_m. Por tanto, $x_0 \in A$. Ahora, tomando una subsucesión de esta última (que siguemos denotando por x_m) tenemos que $f_m(x_m) \to f(x_0)$. Por tanto,

$$\min_{x\in A} f(x) \le f(x_0) = \lim_m f_m(x_m) = \lim_m \min_{x\in X} f_m(x) \le \lim_m \min_{x\in A} f_m(x) = \min_{x\in A} f(x),$$

y todas las desigualdades de arriba son, de hecho, igualdades. Lo que completa la demostración.

■

Teorema 5.12. *Si $\tau_\alpha(F,G) = 0$, F satisface la condición (5.1) y*

$$\int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} dt < \infty, \tag{5.11}$$

entonces

$$nT_{n,\alpha} \underset{w}{\to} \min_{h\in\mathcal{C}_\alpha(F,G)} \int_0^1 \frac{B(t)^2}{f^2(F^{-1}(t))} h'(t)\, dt,$$

donde $\{B(t)\}_{0<t<1}$ *es un puente browniano.*

Nota 5.13. Razonando de la misma forma que en la demostración del Lema 5.8 se puede ver que c.s.

$$h \mapsto \int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))} h'(t)dt$$

es $\|\cdot\|_\infty$-continua como función de h. Por tanto, alcanza su mínimo valor en el conjunto compacto $\mathcal{C}_\alpha(F,G)$. Esto justifica la expresión para la distribución límite en el Teorema 5.12.

Demostración.

Definimos $D_n(h) := n\int_0^1 (F_n^{-1}(t) - G^{-1}(t))^2 h'(t)\, dt$ y $D(h) := \int_0^1 \frac{B^2(t)}{f^2(F^{-1}(t))} h'(t)dt$ para $h \in \mathcal{C}_\alpha$. Tenemos que

$$\begin{aligned} D_n(h) &= \int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} h'(t)\, dt + n\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h'(t)\, dt \\ &\quad +2\sqrt{n}\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))}(F^{-1}(t) - G^{-1}(t))h'(t)dt. \end{aligned}$$

Observamos también que $nT_{n,\alpha} = D_n(h_n)$ para algún $h_n \in \mathcal{C}_\alpha$ Si $h \in \mathcal{C}_\alpha(F,G)$ entonces el segundo y tercer sumandos en el lado derecho se anulan y $D_n(h) = \int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))} h'(t)\, dt$. Por (5.1) y (5.11) tenemos la convergencia débil de $\rho_n(\cdot)/f(F^{-1}(\cdot))$ a $B(\cdot)/f(F^{-1}(\cdot))$ como elementos aleatorios en $L_2(0,1)$ (ver, por ejemplo, el Teorema 4.6 en del Barrio et al., 2005). Por el Teorema de Representación de Skorohod (ver, por ejemplo, el Teorema 11.7.1 en Dudley, 1989) existen versiones de $\rho_n(\cdot)/f(F^{-1}(\cdot))$ y $B(\cdot)/f(F^{-1}(\cdot))$ (para las cuales mantenemos la misma notación) tal que

$$\|\rho_n(\cdot)/f(F^{-1}(\cdot)) - B(\cdot)/f(F^{-1}(\cdot))\|_2 \to 0,$$

c.s. Ahora, para estas versiones tenemos

$$\sup_{h\in\mathcal{C}_\alpha(F,G)} |D_n(h) - D(h)| \le \frac{1}{1-\alpha}\int_0^1 \left|\frac{\rho_n^2(t)}{f^2(F^{-1}(t))} - \frac{B^2(t)}{f^2(F^{-1}(t))}\right| dt \to 0,$$

c.s., mientras que para $h_0 \in \mathcal{C}_\alpha - \mathcal{C}_\alpha(F,G)$ tenemos casi seguro que $D_n(h) \to \infty$ uniformemente en un entorno lo suficientemente pequeño de h_0. Además, si $h_n \to h \in \mathcal{C}_\alpha(F,G)$ entonces podemos extraer una subsucesión tal que $n\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 h_n'(t)dt \to 0$. El resultado se sigue entonces del Lema 5.11. ∎

5.3. Normal más próxima recortada con el mismo patrón.

En esta sección se presentan algunos resultados asintóticos que permiten la elaboración de un test de normalidad (o casi normalidad) basado en recortar (utilizando el mismo patrón) las dos distribuciones, la que da origen a la muestra en cuestión y la de la normal de referencia más próxima.

La generalización a otras familias de localización y escala en la recta real es inmediata a partir del desarrollo que se hará.

Sea $P \in \mathcal{P}_2(\mathbb{R})$, F su función de distribución y $\mathcal{N}$ la familia de distribuciones normales en la recta real. Siguiendo el esquema de recorte introducido en (3.14) podemos definir la distancia α-recortada a la familia de normales como,

$$\begin{aligned} \tau_\alpha(P,\mathcal{N}) \equiv \tau_\alpha(F,\mathcal{N}) &:= \inf_{h\in\mathcal{C}_\alpha, Q\in\mathcal{N}} \mathcal{W}_2^2(P_h, Q_h) = \inf_{h\in\mathcal{C}_\alpha,\, \mu\in\mathbb{R},\, \sigma\geq 0} v_{\mu,\sigma}(h) \\ &= \inf_{h\in\mathcal{C}_\alpha} v(h), \end{aligned} \tag{5.12}$$

donde

$$v_{\mu,\sigma}(h) := \mathcal{W}_2^2\left(P_h, (N(\mu,\sigma))_h\right) = \int_0^1 (F^{-1}(t) - \mu - \sigma\Phi^{-1}(t))^2 h'(t)dt, \tag{5.13}$$

$$v(h) := \inf_{\mu\in\mathbb{R},\, \sigma\geq 0} v_{\mu,\sigma}(h). \tag{5.14}$$

Diferenciando la expresión (5.13), es inmediato ver que para cada $h \in \mathcal{C}_\alpha$, el inferior en μ y σ se alcanza en

$$\mu(h) = \int_0^1 (F^{-1}(t) - \sigma(h)\Phi^{-1}(t))h'(t)dt, \tag{5.15}$$

$$\begin{aligned} \sigma(h) = \sigma_\Phi^{-2}(h) \Bigg(&\int_0^1 F^{-1}(t)\Phi^{-1}(t)h'(t)dt \\ &- \int_0^1 F^{-1}(t)h'(t)dt \int_0^1 \Phi^{-1}(t)h'(t)dt \Bigg), \end{aligned} \tag{5.16}$$

donde

$$\sigma_\Phi^2(h) = \int_0^1 (\Phi^{-1}(t))^2 h'(t)dt - \left(\int_0^1 \Phi^{-1}(t)h'(t)dt\right)^2. \tag{5.17}$$

Consideremos ahora las funciones $h(t)$ como variables aleatorias en $(0,1)$ provisto de la medida de Lebesgue. Es inmediato comprobar que, después del correspondiente cambio de variable, la expresión (5.16) se puede escribir como,

$$\sigma(h) = \frac{\mathrm{Cov}\left(F^{-1}\circ h^{-1}, \Phi^{-1}\circ h^{-1}\right)}{\mathrm{Var}\left(\Phi^{-1}\circ h^{-1}\right)} \tag{5.18}$$

Operando y haciendo ese mismo cambio de variable llegamos a que $v(h)$ se puede escribir entonces como,

$$v(h) = \mathrm{Var}\left(F^{-1} \circ h^{-1}\right) - \frac{\mathrm{Cov}\left(F^{-1} \circ h^{-1}, \Phi^{-1} \circ h^{-1}\right)^2}{\mathrm{Var}\left(\Phi^{-1} \circ h^{-1}\right)} \tag{5.19}$$

Así pues la minimización en (5.12) es equivalente a obtener el inferior en h de la expresión (5.19). En el siguiente resultado vemos algunas propiedades relacionadas con la función $v(h)$.

Lema 5.14. *Si $\mu(h)$, $\sigma(h)$ y $\sigma^2_\Phi(h)$ están definidos como en (5.15)-(5.17), $\alpha < \frac{1}{2}$, y F tiene momento de orden 2 finito, entonces*

(a) *$\sigma^2_\Phi(h)$ está acotada inferior y superiormente como función de $h \in \mathcal{C}_\alpha$. Más aún,*

$$\sigma^2_\Phi(h_1) = \mathrm{Var}\ \left(\Phi^{-1} \circ h_1^{-1}\right) \leq \sigma^2_\Phi(h) \leq \mathrm{Var}\ \left(\Phi^{-1} \circ h_2^{-1}\right) = \sigma^2_\Phi(h_2), \quad \forall h \in \mathcal{C}_\alpha,$$

donde $h_1'(t) = \frac{1}{1-\alpha} I_{[\alpha/2, 1-\alpha/2]}(t)$ y $h_2'(t) = \frac{1}{1-\alpha} I_{[0,(1-\alpha)/2] \cup [(1+\alpha)/2, 1]}(t)$. Además,

$$\sigma^2_\Phi(h_1) = 1 + \frac{2}{1-\alpha} \Phi^{-1}\left(\frac{\alpha}{2}\right) \varphi\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)\right) \quad y \quad \sigma^2_\Phi(h_2) = 1 + \frac{2}{1-\alpha} \Phi^{-1}\left(\frac{1+\alpha}{2}\right) \varphi\left(\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right).$$

(b) *$\mu(h)$, $\sigma(h)$ y $v(h)$ son funciones acotadas y continuas de h para la topología uniforme en $\mathcal{C}_\alpha$.*

(c) *$v(h)$ alcanza el mínimo en $\mathcal{C}_\alpha$.*

(d) *$v(h)$ es cóncava como función de h.*

Demostración. La cota superior del apartado (a) se obtiene fácilmente puesto que,

$$\sigma^2_\Phi(h) = \int_0^1 (\Phi^{-1}(t))^2 h'(t) dt - \left(\int_0^1 \Phi^{-1}(t) h'(t) dt\right)^2 \leq \int_0^1 (\Phi^{-1}(t))^2 h'(t) dt, \quad h \in \mathcal{C}_\alpha.$$

Ahora, basta con analizar las funciones de nivel de $\left(\Phi^{-1}(t)\right)^2$ para obtener que el máximo en el lado derecho de la anterior expresión se alcanza para $h_2 \in \mathcal{C}_\alpha$ tal que $h_2'(t) = \frac{1}{1-\alpha} I_{[0,(1-\alpha)/2] \cup [(1+\alpha)/2, 1]}(t)$.

Para obtener la cota inferior, utilizamos que

$$\sigma^2_\Phi(h) = \mathrm{Var}\ \left(\Phi^{-1} \circ h^{-1}\right) = \int_0^1 \left(\Phi^{-1}(t) - \mu_\Phi(h)\right)^2 h'(t) dt,$$

donde

$$\mu_\Phi(h) = \int_0^1 \Phi^{-1}(t) h'(t) dt, \tag{5.20}$$

es la media de la distribución normal estándar α-recortada siguiendo el patrón marcado por h. Observamos que $\mu_\Phi(h)$ está acotada,

$$\frac{-1}{1-\alpha}\varphi(\Phi^{-1}(\alpha)) = \frac{1}{1-\alpha}\int_0^{1-\alpha}\Phi^{-1}(t)dt \leq \mu_\Phi(h) \leq \frac{1}{1-\alpha}\int_\alpha^1 \Phi^{-1}(t)dt = \frac{1}{1-\alpha}\varphi(\Phi^{-1}(\alpha)).$$

Para cada $h \in \mathcal{C}_\alpha$ fijo,

$$\sigma^2_\Phi(h) \geq \frac{1}{1-\alpha}\int_0^1 \left(\Phi^{-1}(t) - \mu_\Phi(h)\right)^2 I_{(|\Phi^{-1}-\mu_\Phi(h)|\leq k(\mu_\Phi(h),\alpha))}(t)dt,$$

donde $k(\mu_\Phi(h),\alpha)$ es una constante tal que $\ell\left\{t \in (0,1) : |\Phi^{-1}(t) - \mu_\Phi(h)| \leq k(\mu_\Phi(h),\alpha)\right\} = 1-\alpha$. Si fijamos μ, α y k, tenemos que $\left\{t \in (0,1) : |\Phi^{-1}(t) - \mu| \leq k\right\} = [\Phi(\mu-k), \Phi(\mu+k)]$, y por tanto $k(\mu,\alpha)$ es el único valor de k tal que $\Phi(\mu-k) - \Phi(\mu+k) = 1-\alpha$. De esta forma,

$$\sigma^2_\Phi(h) \geq \frac{1}{1-\alpha}\int_{\Phi(\mu_\Phi(h)-k)}^{\Phi(\mu_\Phi(h)+k)} \Phi^{-1}(t)^2 dt - \left(\frac{1}{1-\alpha}\int_{\Phi(\mu_\Phi(h)-k)}^{\Phi(\mu_\Phi(h)+k)} \Phi^{-1}(t)dt\right)^2.$$

Ahora, si para cada $\mu_\Phi(h) \in \mathbb{R}$ definimos $a = \Phi(\mu_\Phi(h) - k)$, tenemos que $\min_{h\in\mathcal{C}_\alpha} \sigma^2_\Phi(h) \geq \min_{a\in[0,\alpha]} G(a)$, donde

$$G(a) = \frac{1}{1-\alpha}\int_a^{a+1-\alpha} \Phi^{-1}(t)^2 dt - \left(\frac{1}{1-\alpha}\int_a^{a+1-\alpha} \Phi^{-1}(t)dt\right)^2.$$

Derivando,

$$\begin{aligned} G'(a) &= \frac{1}{1-\alpha}\left(\left(\Phi^{-1}(a+1-\alpha)\right)^2 - \left(\Phi^{-1}(a)\right)^2\right) \\ &\quad -\frac{2}{(1-\alpha)^2}\left(\Phi^{-1}(a+1-\alpha) - \Phi^{-1}(a)\right)\int_a^{a+1-\alpha}\Phi^{-1}(t)dt \\ &= \frac{2}{1-\alpha}\left(\Phi^{-1}(a+1-\alpha) - \Phi^{-1}(a)\right)H(a), \end{aligned}$$

donde

$$H(a) = \left(\frac{\Phi^{-1}(a+1-\alpha) + \Phi^{-1}(a)}{2} - \frac{1}{1-\alpha}\int_a^{a+1-\alpha}\Phi^{-1}(t)dt\right).$$

Tenemos que $H(a)$ tiene el mismo signo que $G'(a)$ y ambas se anulan en los mismos puntos. Además, $H(a)$ es estrictamente creciente pues

$$H'(a) = \frac{1}{2}\left(\frac{1}{\varphi(\Phi^{-1}(a+1-\alpha))} + \frac{1}{\varphi(\Phi^{-1}(a))}\right) - \frac{1}{1-\alpha}\left(\Phi^{-1}(a+1-\alpha) - \Phi^{-1}(a)\right) > 0,$$

para cada $a \in (0,\alpha)$ (ya que $\frac{1-\alpha}{2}\frac{1}{\varphi(x)} > \frac{1}{4\varphi(x)} > x$, si $x > 0$ y $\alpha < \frac{1}{2}$). Entonces, como $H(\frac{\alpha}{2}) = 0$, éste será el único punto que anula $G'(a)$. Por otra parte, tenemos que

$$G''(\alpha/2) = \frac{2}{1-\alpha}\left[\frac{2\Phi^{-1}(1-\alpha/2)}{\varphi\left(\Phi^{-1}(1-\alpha/2)\right)}\right] - \frac{2}{(1-\alpha)^2}4\left(\Phi^{-1}(1-\alpha/2)\right)^2 > 0,$$

puesto que analizar el signo de la anterior expresión, equivale a analizar el de $\frac{2x}{\varphi(x)} - \frac{4x^2}{1-\alpha}$, si hacemos $x = \Phi^{-1}(1-\alpha/2)$, y éste es positivo pues $\frac{1}{\varphi(x)} - \frac{x}{3/2-\Phi(x)} > \frac{1}{\varphi(x)} - 2x > 0$ si $x > 0$ ($\frac{1}{2\varphi(x)} > x$).

En consecuencia se tiene que G alcanza su valor mínimo en $a = \frac{\alpha}{2}$, y

$$\min_{h\in\mathcal{C}_\alpha} \sigma^2_\Phi(h) = \frac{1}{1-\alpha}\int_{\alpha/2}^{1-\alpha/2} \Phi^{-1}(t)^2 dt = \sigma^2_\Phi(h_1).$$

Los valores de $\sigma^2_\Phi(h_1)$ y $\sigma^2_\Phi(h_2)$ se obtienen de forma inmediata integrando $\Phi^{-1}(t)^2$ y teniendo en cuenta su simetría respecto a 0.

Para probar el apartado (b), veremos primero que dadas dos funciones cuantiles de cuadrado integrable, F^{-1} y G^{-1}, el funcional $a(h) = \int_0^1 F^{-1}(h^{-1}(t))G^{-1}(h^{-1}(t))dt$ es continuo en $\mathcal{C}_\alpha$ para la topología uniforme. Para ello, tomamos $\{h_n\}_n$, $h_0 \in \mathcal{C}_\alpha$ tal que $\|h_n - h_0\| \to 0$. Es claro entonces que $h_n(F(x)) \to h_0(F(x))$ para cada x, y en consecuencia, $F^{-1}(h_n^{-1}(t)) \to F^{-1}(h_0^{-1}(t))$ para casi todo $t \in (0,1)$. Y de forma similar para G^{-1}. Para tener que $a(h_n) \to a(h_0)$ basta con ver que $(F^{-1}\circ h^{-1})(G^{-1}\circ h^{-1})$ es uniformemente integrable. Pero esto es inmediato, pues dado $k > 0$,

$$\begin{aligned}
&\sup_{h\in\mathcal{C}_\alpha}\int_0^1 |F^{-1}(h^{-1}(t))G^{-1}(h^{-1}(t))| I(|F^{-1}(h^{-1}(t))G^{-1}(h^{-1}(t))| > k)dt \\
&\qquad = \sup_{h\in\mathcal{C}_\alpha}\int_0^1 |F^{-1}(y)G^{-1}(y)| I(|F^{-1}(y)G^{-1}(y)| > k)h'(y)dy \\
&\qquad \le \frac{1}{1-\alpha}\int_0^1 |F^{-1}(y)G^{-1}(y)| I(|F^{-1}(y)G^{-1}(y)| > k)dy \to 0
\end{aligned}$$

cuando $k \to \infty$. De esto se deduce que $\mathrm{Cov}(\Phi^{-1}\circ h^{-1}, F^{-1}\circ h^{-1})$, $\mathrm{Var}(F^{-1}\circ h^{-1})$ y $\mathrm{Var}(\Phi^{-1}\circ h^{-1})$ son continuas en h. Del apartado anterior tenemos que $\min_{h\in\mathcal{C}_\alpha} \mathrm{Var}(\Phi^{-1}\circ h^{-1}) = \sigma^2_\Phi(h_1) > 0$. Como consecuencia también son continuas $\mu(h)$, $\sigma(h)$ y $v(h)$.

Que $\mu(h)$, $\sigma(h)$ y $v(h)$ están acotadas también es inmediato. Veamos en primer lugar que lo está $\sigma(h)$. Utilizando (5.18) para $\sigma(h)$, el apartado (a) y acotaciones estándar tenemos que

$$\sup_{h\in\mathcal{C}_\alpha} |\sigma(h)| \le \sup_{h\in\mathcal{C}_\alpha} \frac{1}{\sigma^2_\Phi(h_1)} \left|\mathrm{Cov}(\Phi^{-1}\circ h^{-1}, F^{-1}\circ h^{-1})\right| \le \frac{1}{\sigma^2_\Phi(h_1)}\frac{1}{1-\alpha}\left|\mathrm{E}(\Phi^{-1}F^{-1})\right| < \infty.$$

A partir de las expresiones (5.15) y (5.19), de forma completamente similar, obtenemos también que $\mu(h)$ y $v(h)$ están acotadas.

El apartado (c) es una consecuencia inmediata de (b) y que $\mathcal{C}_\alpha$ es compacto para la topología uniforme (Proposición 3.14).

Finalmente, el apartado (d) es una consecuencia directa de que $v_{\mu,\sigma}(h)$ es lineal en h, y por tanto, si $\lambda \in (0,1)$ y $h_1, h_2 \in \mathcal{C}_\alpha$,

$$\begin{aligned} v\left(\lambda h_1 + (1-\lambda)h_2\right) &= \min_{\mu,\sigma}\left\{v_{\mu,\sigma}\left(\lambda h_1 + (1-\lambda)h_2\right)\right\} = \min_{\mu,\sigma}\left\{\lambda v_{\mu,\sigma}(h_1) + (1-\lambda)v_{\mu,\sigma}(h_2)\right\} \\ &\geq \min_{\mu,\sigma}\left\{\lambda v_{\mu,\sigma}(h_1)\right\} + \min_{\mu,\sigma}\left\{(1-\lambda)v_{\mu,\sigma}(h_2)\right\} = \lambda v(h_1) + (1-\lambda)v(h_2). \end{aligned}$$

■

Nota 5.15. Nótese que la hipótesis de que $\alpha < \frac{1}{2}$ en el anterior lema se usa exclusivamente para obtener el mínimo en el apartado (a). Aún sin dicha hipótesis, es claro que siempre $\min_{h\in\mathcal{C}_\alpha} \sigma_\Phi(h)^2 > 0$. Dicho mínimo no podría ser nunca cero pues recortando una distribución absolutamente continua (la normal estándar en este caso) no se puede obtener nunca una distribución concentrada en un punto. Por tanto los apartados (b)-(d) son válidos con independencia del valor de α.

Nota 5.16. Un desarrollo totalmente paralelo se puede realizar si reemplazamos en (5.12) la clase $\mathcal{C}_\alpha$ de funciones de recorte generales por la clase $\mathcal{D}_\alpha$ donde

$$\mathcal{D}_\alpha = \left\{h \in \mathcal{C}_\alpha : h'(t) = \frac{1}{1-\alpha} I_{[\delta,\delta+1-\alpha]}(t),\ 0 \leq \delta \leq \alpha\right\}$$

Esta clase de funciones de recorte corresponde al caso de recorte asimétrico en el cual una proporción α de masa de probabilidad se elimina de las colas de la distribución: δ en la cola de la izquierda y el resto hasta llegar a α en la de la derecha. Incluye como caso particular el caso de recorte simétrico en el que $\delta = \frac{\alpha}{2}$. Resultados similares a los que se obtienen en esta sección se pueden deducir en este caso particular a partir de la compacidad de la clase $\mathcal{D}_\alpha \subset \mathcal{C}_\alpha$.

Combinando los apartados (c) y (d) del Lema 5.14 tenemos que el inferior en (5.12) se alcanza en algún punto extremal de $\mathcal{C}_\alpha$, es decir, en alguna h, tal que $h'(t) = \frac{1}{1-\alpha} I_A$. Lo que, en principio, no podemos garantizar es que dicho minimizador sea único. Es posible pensar en situaciones, caracterizadas por la bimodalidad, en las que no hay unicidad. Sin embargo, es previsible que bajo situaciones de unimodalidad sí haya unicidad. En cualquier caso, no abordamos en esta tesis esta cuestión, y de aquí en adelante asumiremos que $h_0 := \arg\min_{h\in\mathcal{C}_\alpha} v(h)$ es único.

Finalizamos esta sección con el diseño de un test para contrastar la hipótesis

$$H_0 : \tau_\alpha(P,\mathcal{N}) \geq \Delta_0^2 \quad \text{vs.} \quad H_a : \tau_\alpha(P,\mathcal{N}) < \Delta_0^2. \tag{5.21}$$

Para ello, si $X_1,\ldots,X_n$ es una muestra aleatoria simple de observaciones con función de distribución F, consideremos el estadístico $T_{n,\alpha} := \tau_\alpha(F_n,\mathcal{N})$, donde F_n es la función de distribución empírica asociada a la muestra. De igual forma que en la versión poblacional tendremos entonces que $\tau_\alpha(F_n,\mathcal{N}) = \text{ínf}_{h\in\mathcal{C}_\alpha} v_n(h)$, donde

$$v_n(h) := \min_{Q\in\mathcal{N}} \mathcal{W}_2^2((P_n)_h, Q_h) = \text{Var}\left(F_n^{-1}\circ h^{-1}\right) - \frac{\text{Cov}\left(F_n^{-1}\circ h^{-1}, \Phi^{-1}\circ h^{-1}\right)^2}{\text{Var}\left(\Phi^{-1}\circ h^{-1}\right)}.$$

Podemos aplicar el Lema 5.14 a $v_n(h)$ con lo que tendremos que el inferior se alcanza. Denotaremos por $h_n := \arg\,\text{mín}_{h\in\mathcal{C}_\alpha} v_n(h)$ y

$$\sigma_n = \frac{\int_0^1 F_n^{-1}\Phi^{-1}h_n' - \int_0^1 F_n^{-1}h_n' \int_0^1 \Phi^{-1}h_n'}{\int_0^1(\Phi^{-1})^2 h_n' - (\int_0^1 \Phi^{-1})h_n')^2}, \quad \mu_n = \int_0^1 (F_n^{-1} - \sigma_n\Phi^{-1})h_n'. \tag{5.22}$$

Para la obtención de la distribución asintótica asociada a $T_{n,\alpha}$, observemos que,

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P,\mathcal{N})) = \sqrt{n}\left(\min_{h\in\mathcal{C}\alpha} v_n(h) - \min_{h\in\mathcal{C}\alpha} v(h)\right). \tag{5.23}$$

Por ello, para obtener el resultado del Teorema 5.19 estudiaremos los procesos

$$M_n(h) := \sqrt{n}(v_n(h) - v(h)), \quad h\in\mathcal{C}_\alpha, \tag{5.24}$$

y

$$M(h) := 2\int_0^1 \frac{B(t)}{f(F^{-1}(t))}(F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))h'(t)dt,$$

donde $\{B(t)\}_{t\in(0,1)}$ es un puente browniano. Observemos que $\{M(h)\}_{h\in\mathcal{C}_\alpha}$ es un proceso gaussiano centrado con función de covarianza

$$K(h_1,h_2) = 4\int_0^1 l_1(t)l_2(t)dt - 4\int_0^1 l_1(t)dt\int_0^1 l_2(t)dt,$$

donde

$$l_i(t) = \int_{F^{-1}(1/2)}^{F^{-1}(t)} (x - \mu(h_i) - \sigma(h_i)\Phi^{-1}(F(x)))h_i'(F(x))dx, \quad i = 1,2.$$

Este resultado se sigue de que, integrando por partes, $M(h_i) = -2\int_0^1 l_i(t)dB(t)$.

A continuación se da un resultado clave para la demostración del Teorema 5.19.

Proposición 5.17. *Supongamos que $P \in \mathcal{P}_4'$, y F es su función de distribución con función de densidad f verificando la condición*

$$\sup_{x\in\mathbb{R}} \left| \frac{F(x)(1-F(x))f'(x)}{f^2(x)} \right| < \infty. \tag{5.25}$$

Entonces, M es un funcional tight y M_n converge débilmente a M en $(\mathcal{C}_\alpha, \|\cdot\|_\infty)$.

Demostración. Al igual que en la demostración del Teorema 5.9, podemos asumir sin pérdida de generalidad que existe una sucesión de puentes brownianos B_n que verifican,

$$n^{1/2-\nu} \sup_{\frac{1}{n}\le t\le 1-\frac{1}{n}} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} = \begin{cases} O_P(\log n), & \text{si } \nu = 0 \\ O_P(1), & \text{si } 0 < \nu \le 1/2 \end{cases} \tag{5.26}$$

donde, como allí, $\rho_n(t) = \sqrt{n}(F_n^{-1}(t) - F^{-1}(t))f(F^{-1}(t))$.

Operando es fácil ver que,

$$\begin{aligned} M_n(h) &= 2\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))}(F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))h'(t)dt \\ &+ \frac{1}{\sqrt{n}}\left[\int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))}h'(t)dt - \left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))}h'(t)dt\right)^2\right. \\ &\left. + \frac{1}{\sigma_\Phi^2(h)}\left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))}\Phi^{-1}(t)h'(t)dt - \left(\int_0^1 \frac{\rho_n(t)}{f(F^{-1}(t))}h'(t)dt\right)\mu_\Phi(h)\right)^2\right], \end{aligned} \tag{5.27}$$

donde $\mu_\Phi(h)$ y $\sigma_\Phi^2(h)$ están definidas como en (5.20) y (5.17).

Definimos ahora,

$$N_n(h) := 2\int_0^1 \frac{B_n(t)}{f(F^{-1}(t))}(F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))h'(t)dt.$$

Se tiene entonces que $\|M_n - N_n\|_{\mathcal{C}_\alpha} := \sup_{h\in\mathcal{C}_\alpha} |M_n(h) - N_n(h)| \to 0$ en probabilidad. Para comprobarlo, escribimos,

$$\sup_{h\in\mathcal{C}_\alpha} \frac{1}{\sqrt{n}}\int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))}h'(t)dt \le \frac{1}{1-\alpha}\frac{1}{\sqrt{n}}\int_0^1 \frac{\rho_n^2(t)}{f^2(F^{-1}(t))}dt \to_p 0,$$

donde esta última convergencia se deriva de la hipótesis sobre los momentos de F y el Lema 5.7 siguiendo el mismo esquema que en la demostración del Teorema 5.9. Por tanto, podemos descartar (uniformemente) el segundo término de la expresión (5.27) para M_n. Una cota similar y el hecho de que $\min_{h\in\mathcal{C}_\alpha} \sigma_\Phi^2(h) > 0$ es suficiente para controlar el tercer término de dicha expresión. En consecuencia, basta con probar que

$$\sup_{h\in\mathcal{C}_\alpha} \int_0^1 \frac{|\rho_n(t) - B_n(t)|}{f(F^{-1}(t))}|F^{-1}(t) - \mu(h) - \sigma(h)\Phi^{-1}(t))|dt \to_p 0.$$

Pero esto se sigue de forma inmediata del hecho de que $\mu(h)$ y $\sigma(h)$ están acotados (Lema 5.14), razonando de la misma forma que en la demostración del Teorema 5.9.

Puesto que N_n y M están igualmente distribuidos, para completar la demostración, sólo tenemos que probar que M es tight o, equivalentemente, que es uniformemente equicontinuo en probabilidad para alguna métrica d para la que $\mathcal{C}_\alpha$ esté totalmente acotado (ver los Teoremas 1.5.7 y 1.10.2 en van der Vaart y Wellner, 1996). Sea d la métrica uniforme en $\mathcal{C}_\alpha$ (entonces $\mathcal{C}_\alpha$ es compacto). Así pues, tenemos que probar que dados $\varepsilon, \eta > 0$ existe $\delta > 0$ tal que

$$P\left(\sup_{\|h_1-h_2\|_\infty<\delta} |M(h_1) - M(h_2)| > \varepsilon\right) < \eta.$$

Utilizando la desigualdad de Markov y la compacidad de $\mathcal{C}_\alpha$ concluimos que es suficiente con probar que el funcional $h \mapsto E|M(h)|$ es continuo con respecto a la métrica $\|\cdot\|_\infty$. Pero esta continuidad se obtiene, de forma inmediata, utilizando los mismos argumentos que para el apartado (b) del Lema 5.14. ■

La Proposición 5.17 tiene la siguiente e importante consecuencia,

Corolario 5.18. *Bajo las hipótesis de la Proposición 5.17*

$$\sup_{h\in\mathcal{C}_\alpha} |v_n(h) - v(h)| \to 0, \quad \text{en probabilidad.}$$

Además, si $h_0 = \arg\min_{h\in\mathcal{C}_\alpha} v(h)$ *es único, tenemos que* $\|h_n - h_0\| \to 0$, en probabilidad.

Demostración. Como consecuencia de la Proposición 5.17 tenemos que $\sup_h |M_n(h)| = \sqrt{n}\sup_h |v_n(h) - v(h)|$ converge débilmente a $\sup_{h\in\mathcal{C}_\alpha} |M(h)|$, y por tanto tenemos que $\sup_{h\in\mathcal{C}_\alpha} |v_n(h) - v(h)| \to_p 0$. Para la segunda parte utilizamos la compacidad, por la que podemos extraer subsucesiones de h_n tales que $h_{n_k} \to h_1$ y también asegurar que $v_{n_k}(h) \to v(h)$ uniformemente. Puesto que $v_{n_k}(h_{n_k}) \leq v_{n_k}(h)$ vemos que, tomando límites, h_1 tiene que ser un minimizador de v y por tanto, por la unicidad, $h_1 = h_0$. Lo que concluye la demostración. ■

Teorema 5.19. *Si* P *satisface las condiciones de la Proposición 5.17 y* h_0 *es único, entonces*

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P, \mathcal{N})) \underset{w}{\to} N(0, \sigma_\alpha^2(P, \mathcal{N}))$$

donde

$$\begin{aligned}\sigma_\alpha^2(P,\mathcal{N}) &= 4\left(\int_0^1 l^2(t)dt - \left(\int_0^1 l(t)dt\right)^2\right),\\ l(t) &= \int_{F^{-1}(1/2)}^{F^{-1}(t)} (x-\mu(h_0)-\sigma(h_0)\Phi^{-1}(F(x)))h_0'(F(x))dx,\end{aligned}$$

y $\mu(h_0)$, $\sigma(h_0)$ *están definidos como en (5.15) y (5.16) respectivamente.*

Si $S_{n,\alpha}^2 := 4\left(\int_0^1 l_n^2(t)dt - \left(\int_0^1 l_n(t)dt\right)^2\right)$, *donde*

$$l_n(t) = \int_{F_n^{-1}(1/2)}^{F_n^{-1}(t)} (x-\mu_n-\sigma_n\Phi^{-1}(F_n(x)))h_n'(F_n(x))dx$$

y μ_n, σ_n *están dados por (5.22), entonces* $S_{n,\alpha}^2 \to \sigma_\alpha^2(P,\mathcal{N})$ *en probabilidad.*

Demostración. A partir de (5.23) tenemos que,

$$\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P,\mathcal{N})) = \sqrt{n}(v_n(h_n) - v(h_0)) = M_n(h_0) - \sqrt{n}(v_n(h_0) - v_n(h_n)).$$

La optimalidad de h_0 implica que $v(h_n) - v(h_0) \geq 0$, y la de h_n, que $v_n(h_0) - v_n(h_n) \geq 0$. Por otra parte, de la Proposición 5.17, el Corolario 5.18 y la equicontinuidad tenemos que

$$\sqrt{n}(v(h_n) - v(h_0)) + \sqrt{n}(v_n(h_0) - v_n(h_n)) = M_n(h_0) - M_n(h_n) \to 0, \quad \text{en probabilidad.}$$

Ahora, como los dos sumandos de la parte izquierda de la expresión anterior son positivos, sólo puede suceder que $\sqrt{n}(v_n(h_0) - v_n(h_n)) \to 0$, en probabilidad. En consecuencia, $\sqrt{n}(T_{n,\alpha} - \tau_\alpha(P,\mathcal{N}))$ converge en distribución a $M(h_0)$. Lo que concluye la primera parte del teorema. La segunda parte se obtiene con la ayuda del Corolario 5.18 y argumentos de continuidad como en el Lema 5.14, de forma completamente similar a como se hizo en el Teorema 5.9. ■

Es inmediato comprobar que $\tau_\alpha(P,\mathcal{N})$ es invariante por cambios de localización, pero no de escala. Si queremos eliminar esta dependencia de la escala, tal y como es habitual en el ajuste de modelos de localización y escala (ver, por ejemplo, del Barrio et al., 1999), podemos buscar una modificación de $\tau_\alpha(P,\mathcal{N})$ que sea invariante frente a este tipo de cambios. Así, podemos definir, suponiendo la unicidad de h_0,

$$\tilde{\tau}_\alpha(P,\mathcal{N}) := \frac{\tau_\alpha(P,\mathcal{N})}{r_\alpha(P)}, \tag{5.28}$$

donde $r_\alpha(P) = \mathrm{Var}(F^{-1} \circ h_0^{-1})$. Si ahora tenemos en cuenta la expresión (5.19), entonces vemos que $\tilde{\tau}_\alpha(P,\mathcal{N}) = 1 - \mathrm{Corr}^2(F^{-1} \circ h_0^{-1}, \Phi^{-1} \circ h_0^{-1})$. De esta expresión es claro que $\tilde{\tau}_\alpha(P,\mathcal{N})$ es invariante frente a cambios de localización y escala, y además que satisface $0 \leq \tilde{\tau}_\alpha(P,\mathcal{N}) \leq 1$. Los valores de $\tilde{\tau}_\alpha(P,\mathcal{N})$ admiten ahora una interpretación mucho más clara que los de $\tau_\alpha(P,\mathcal{N})$. $\tilde{\tau}_\alpha(P,\mathcal{N}) = 0$ significa un ajuste perfecto a la normalidad después de recortar, mientras que valores grandes (próximos a 1) indican un alto grado de no normalidad incluso después de recortar. De aquí en adelante nos referiremos a $\tilde{\tau}_\alpha(P,\mathcal{N})$ como la distancia recortada estandarizada a la normalidad.

Si admitimos que la evaluación de la normalidad de una distribución después de recortar no debiera depender de la escala de medida de los datos, entonces deberíamos reemplazar el contraste de la hipótesis de (5.21) por

$$H_0 : \tilde{\tau}_\alpha(P,\mathcal{N}) \geq \Delta_0^2 \quad \text{vs.} \quad H_a : \tilde{\tau}_\alpha(P,\mathcal{N}) < \Delta_0^2, \tag{5.29}$$

donde ahora Δ_0^2 lo elegiremos en $(0,1)$.

Como veremos en la Sección 5.4, utilizaremos las curvas de p-valores (ver Sección 2.4) para manejar los posibles valores de Δ_0^2. Estas curvas las construimos utilizando el p-valor asintótico del estadístico

$$Z_{n,\alpha} := \frac{T_{n,\alpha} - \Delta_0^2}{S_{n,\alpha}}, \tag{5.30}$$

cuya distribución asintótica si $\Delta_0^2 = \tau_\alpha(P,\mathcal{N})$ es N(0,1).

En este caso, de acuerdo con (2.6), para cada Δ_0 calcularemos

$$P(\Delta_0) = \sup_{F \in H_0} \lim_{n\to\infty} P_F\left(Z_{n,\alpha} \leq z_0\right) = \Phi\left(\sqrt{n}\tfrac{t_{n,\alpha} - \Delta_0^2}{s_{n,\alpha}}\right),$$

donde $z_0 = \sqrt{n}\frac{t_{n,\alpha} - \Delta_0^2}{s_{n,\alpha}}$ es el valor observado de $Z_{n,\alpha}$.

En la práctica estaremos interesados en contrastar (5.29) en vez de (5.21). Podemos reescribir (5.29) como

$$H_0 : \tau_\alpha(P,\mathcal{N}) \geq \Delta_0^2 r_\alpha(P) \quad \text{vs.} \quad H_a : \tau_\alpha(P,\mathcal{N}) < \Delta_0^2 r_\alpha(P),$$

una familia de problemas de contraste que pueden ser analizados usando la curva de p-valores $P(\Delta_0 r_\alpha^{1/2}(P))$. Puesto que $r_\alpha(P)$ es desconocido, podemos reemplazarlo por un estimador consistente $R_{n,\alpha} = \int_0^1 (F_n^{-1})^2 h_n' - \left(\int_0^1 F_n^{-1} h_n'\right)^2$ y obtener la curva estimada de p-valores para (5.29):

$$\tilde{P}(\Delta_0) := P(\Delta_0 R_{n,\alpha}^{1/2}), \quad 0 < \Delta_0 < 1.$$

5.4. Ejemplos y Simulaciones.

En esta sección se presentan varios ejemplos con datos reales y simulados donde se muestra una aplicación de los resultados obtenidos en las Secciones 5.2 y 5.3. Se incluyen también sendas simulaciones que ilustran el funcionamiento en lo que a la potencia se refiere de los tests propuestos en dichas secciones. En el caso de las comparaciones entre dos muestras que se realizan en el primer ejemplo, comparamos además los resultados que obtenemos con los que se obtendrían si recortásemos en las colas de la distribución siguiendo la metodología de Munk y Czado (1998).

5.4.1. Ejemplo 1, comparación de muestras.

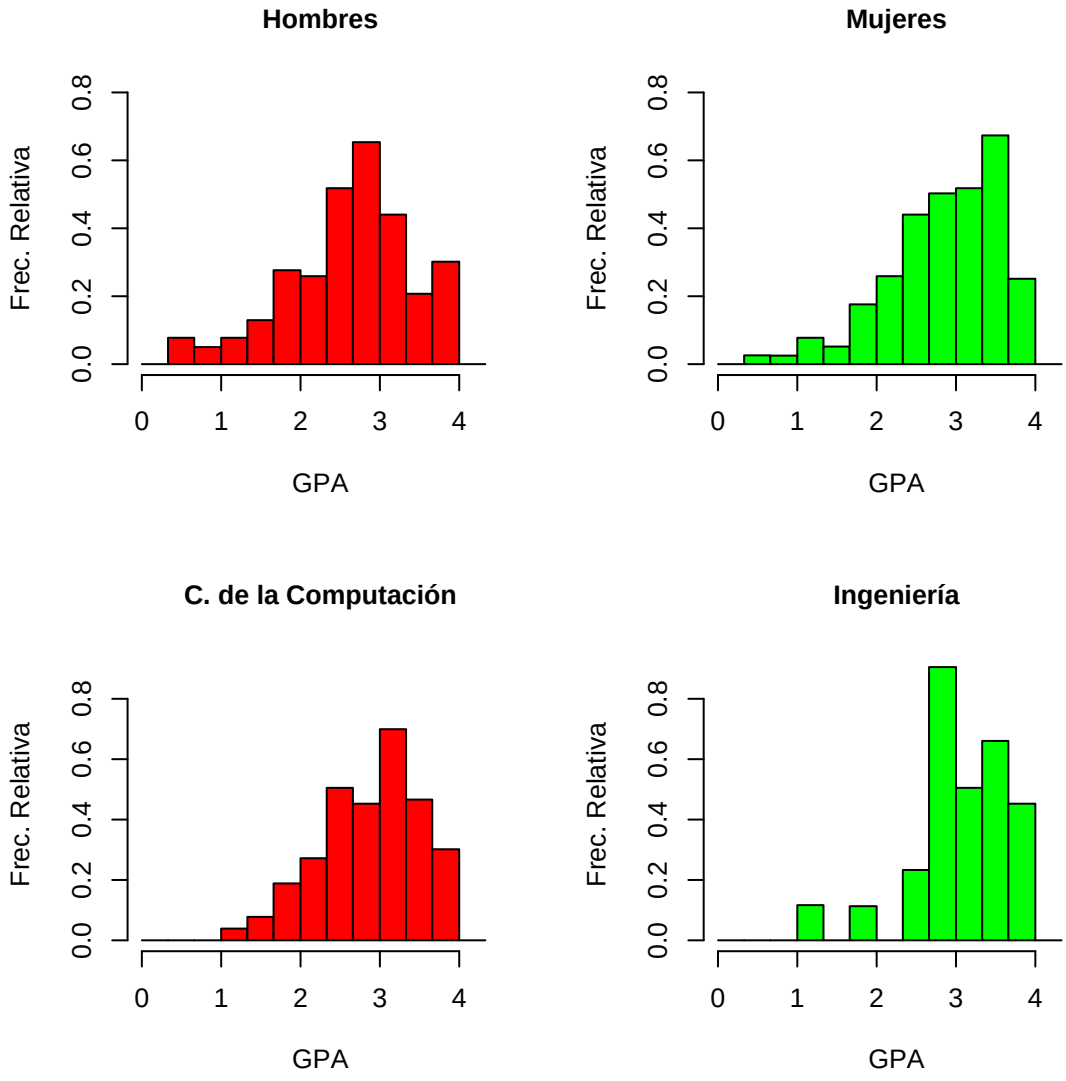

Figura 5.1: Histogramas para distintas submuestras de la variable GPA.

Los datos utilizados en este ejemplo provienen del fichero "majors.dat" disponible entre los conjuntos de datos para ejemplos que vienen en la distribución de muchos paquetes de estadística, y que en nuestro caso hemos tomado directamente de su fuente, el libro de Moore y McCabe (2003). El análisis está basado en la variable GPA (College Grade Point Average) observada en un conjunto de 234 estudiantes. Esta variable toma valores entre 0 y 4. Por otra parte, los estudiantes se pueden clasificar en grupos atendiendo al Sexo (Hombres y Mujeres) y a la Especialidad cursada (Ciencias de la Computación, Ingeniería y Otras Ciencias). Estamos interesados en estudiar la similitud distribucional de la variable GPA entre Hombres ($n = 117$) y Mujeres ($m = 117$), y entre estudiantes con la especialidad en Ciencias de la Computación ($n = 78$) y estudiantes de Ingeniería ($m = 78$). La Figura 5.1 muestra un histograma de cada muestra.

La comparación de estas muestras utilizando procedimientos clásicos produce los resultados que se muestran en las Tablas 5.1 y 5.2. Dado que el test de Shapiro-Wilks rechaza la normalidad en las cuatro muestras, usamos procedimientos no paramétricos como el test de Kolmogorov-Smirnov (KS) o el de Wilcoxon-Mann-Whitney (WMW) para analizar la hipótesis nula de que ambas muestras provienen de la misma distribución en las dos comparaciones que estamos realizando: Hombres versus Mujeres (H vs M), y estudiantes de Ciencias de la Computación versus estudiantes de Ingeniería (C vs I). En los dos casos, los p-valores de ambos tests nos llevan a rechazar claramente la hipótesis nula.

Test	***p*-valor**			
	H	M	C	I
Shapiro-Wilks	0.0176	0.0217	0.0360	0.0001

Tabla 5.1: P-valores para el test de normalidad de Shapiro-Wilks.

Test	***p*-valor**	
	GPA: H vs M	GPA: C vs I
Kolmogorov-Smirnov	0.0028	0.0040
Wilcoxon-Mann-Whitney	0.0004	0.0175

Tabla 5.2: P-valores para los tests clásicos de comparación de dos muestras.

El siguiente paso en nuestro análisis es utilizar los recortes imparciales con el mismo patrón de recorte introducidos en (3.14). La Figura 5.2 muestra las funciones de recorte

óptimas para las dos comparaciones en cuestión. En esta figura, y para cada comparación, representamos el valor de $|F_n^{-1}(t) - G_m^{-1}(t)|$ y los valores de corte $L^{-1}_{F_n,G_m}(1-\alpha)$ para $\alpha =$ 0.05, 0.1 y 0.2.

El primer gráfico (a la izquierda) muestra que el recorte óptimo se produce en la cola inferior de la distribución, pero no exactamente desde el mínimo valor de la muestra. Además cuando el tamaño de recorte aumenta ($\alpha = 0.1$ y 0.2) la zona recortada no es un intervalo e incluye puntos alrededor de los percentiles 20 %, 40 %, 60 % y 70 %. En la segunda comparación se ve que los puntos que hay que recortar para hacer más similares las dos muestras se encuentran entre los percentiles 10 % y 30 % aproximadamente. Este ejemplo, por tanto, nos sirve para ilustrar una situación de disimilitud no simétrica y por tanto las limitaciones del recorte simétrico. De hecho, en la primera comparación la zona menos similar se encuentra alrededor de la cola inferior, pero no en la superior, donde en realidad se encuentran los valores más similares.

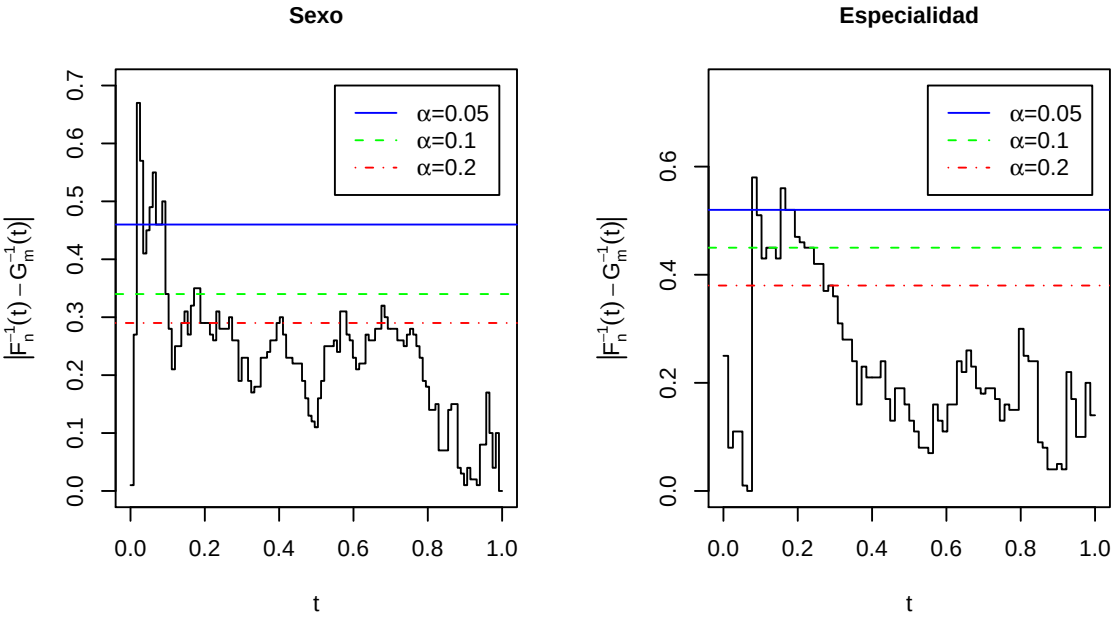

Figura 5.2: Funciones de recorte óptimas.

Para mejorar la visión de la evaluación de la similitud o disimilitud de las distribuciones subyacentes podemos usar la curva de p-valores asintóticos (ver Sección 2.4) asociada al contraste de hipótesis nula $H_0 : \tau_\alpha(F,G) \geq \Delta_0^2$ contra la alternativa $H_a : \tau_\alpha(F,G) < \Delta_0^2$.

Dado que estamos comparando dos muestras, utilizamos el estadístico

$$Z_{n,m,\alpha} = \sqrt{\frac{nm}{n+m}} \frac{(T_{n,m,\alpha} - \Delta_0^2)}{s_{n,m,\alpha}}, \tag{5.31}$$

cuya distribución asintótica (ver Teorema 5.10) si $\tau_\alpha(F,G) = \Delta_0^2$ es N(0,1) (por tanto, $Z_{n,m,\alpha} \to +\infty$ si $\Delta < \tau_\alpha(F,G)$ y $Z_{n,m,\alpha} \to -\infty$ si $\Delta > \tau_\alpha(F,G)$).

La curva de p-valores asintóticos, $P(\Delta_0)$, definida en (2.6), se calculará en este caso como

$$P(\Delta_0) = \sup_{(F,G)\in H_0} \lim_{n,m\to\infty} P_{F,G}\left(Z_{n,m,\alpha} \leq z_0\right) = \Phi\left(z_0\right),$$

donde $z_0 = \sqrt{\frac{nm}{n+m}} \frac{(t_{n,m,\alpha} - \Delta_0^2)}{s_{n,m,\alpha}}$ es el valor observado de $Z_{n,m,\alpha}$.

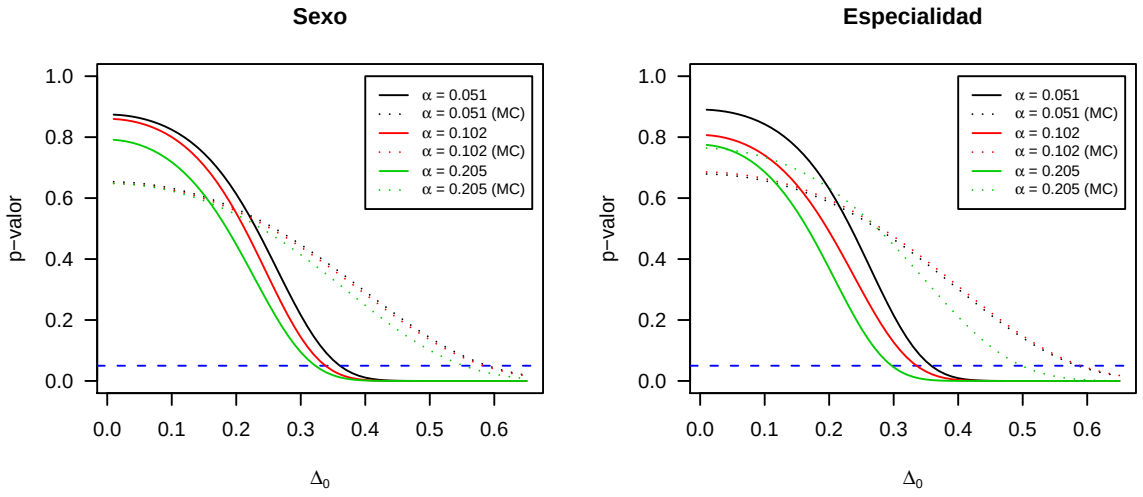

Figura 5.3: Curva de p-valores usando recortes imparciales con el mismo patrón y recortes simétricos (MC).

La Figura 5.3 muestra la mejora que se obtiene en la evaluación de la similitud de dos muestras cuando utilizamos los recortes imparciales con el mismo patrón en vez del recorte simétrico. En esta figura se muestran las curvas de p-valores asintóticos para el recorte imparcial y el recorte simétrico en ambas comparaciones para diferentes niveles de recorte ($\alpha = 0.05$, 0.1 y 0.2). En cada gráfico se han representado los valores de $P(\Delta_0)$ vs Δ_0 para los distintos casos y se ha dibujado una línea horizontal (en azul) para marcar un nivel de referencia para el contraste (0.05). Fijando el nivel del contraste en 0.05, cuando se usan recortes imparciales, la puntuación GPA obtenida por Hombres y Mujeres es similar para valores de Δ_0 inferiores o iguales al rango de valores que va desde 0.32 a 0.36, dependiendo del tamaño de recorte. Estos valores representan entre el 11.4 %($= 100 \times 0.32/2.815$) y el

12.8 % de la media de las medianas de las dos muestras. Sin embargo, cuando se usa el recorte simétrico, la línea horizontal corta a las curvas de p-valores para valores de Δ_0 que van desde 0.56 a 0.59. Esto representa entre el 20 % y el 21 % de la media de las medianas de ambas muestras. Un análisis similar en la comparación de la puntuación GPA obtenida dependiendo de la Especialidad nos lleva a valores de Δ_0 que van desde 0.29 a 0.36, cuando utilizamos recortes imparciales, es decir, entre el 9.6 % y el 11.9 % de la media de las medianas. En cambio, cuando utilizamos el recorte simétrico estos porcentajes están entre el 16.6 % y el 19.5 %.

Nota 5.20 (Cálculo de $Z_{n,m,\alpha}$)**.** Para obtener los valores de $T_{n,m,\alpha}$ calculamos en primer lugar $|F_n^{-1}(t) - G_m^{-1}(t)|^2$ en una malla suficientemente densa del intervalo $[0,1]$, usando el $(1-\alpha)$-cuantil de estos valores para determinar $L^{-1}_{F_n,G_m}(1-\alpha)$. A continuación calculamos la integral correspondiente de forma numérica. Y para el cálculo de la estimación de la varianza asintótica, $S^2_{n,m,\alpha}$, operamos de forma similar.

5.4.2. Ejemplo 2, casi normalidad con datos simulados.

Antes de analizar un ejemplo con datos reales y para ilustrar mejor el uso de las curvas de p-valores asintóticos en la evaluación de la casi normalidad se han generado 6 muestras aleatorias de tamaño 100 correspondientes a 6 modelos diferentes (dos normales, dos mixturas diferentes de normales que después de recortar están cercanas a la normalidad, un modelo chi-cuadrado y un modelo exponencial). En este caso usaremos las que en la Sección 5.3 hemos denominado curvas estimadas de p-valores y que denotamos por $\tilde{P}(\Delta_0)$.

La Figura 5.4 muestra las curvas de p-valores correspondientes a las 6 muestras para diferentes tamaños de recorte, incluido el caso de no recorte ($\alpha = 0$). Aunque la distribución asintótica no se ha calculado explícitamente para este caso, es claro que se puede obtener de manera inmediata utilizando los mismos argumentos que para el Teorema 5.19 y coincide además con el caso límite $\alpha = 0$ en este teorema.

El gráfico (a) de esta figura se corresponde con una N(0,1). El comportamiento es claro, la distancia estandarizada antes de recortar está muy próxima a cero. Además, se observa un pequeño pero apreciable descenso después del primer 5 % de recorte -probablemente debido a algún tipo de efecto de suavizado de la aleatoriedad en la muestra-. Finalmente, se observa una estabilización de la distancia recortada estandarizada cuando se incrementa el tamaño

de recorte ($\alpha = 0.05, 0.1, 0.2$). En otras palabras, no existe mejora en el grado de normalidad incrementando el tamaño de recorte. El gráfico (b), que muestra las curvas de p-valores para un modelo N(10,4), es bastante similar al (a), demostrando que la estandarización llevada a cabo en (5.28) funciona.

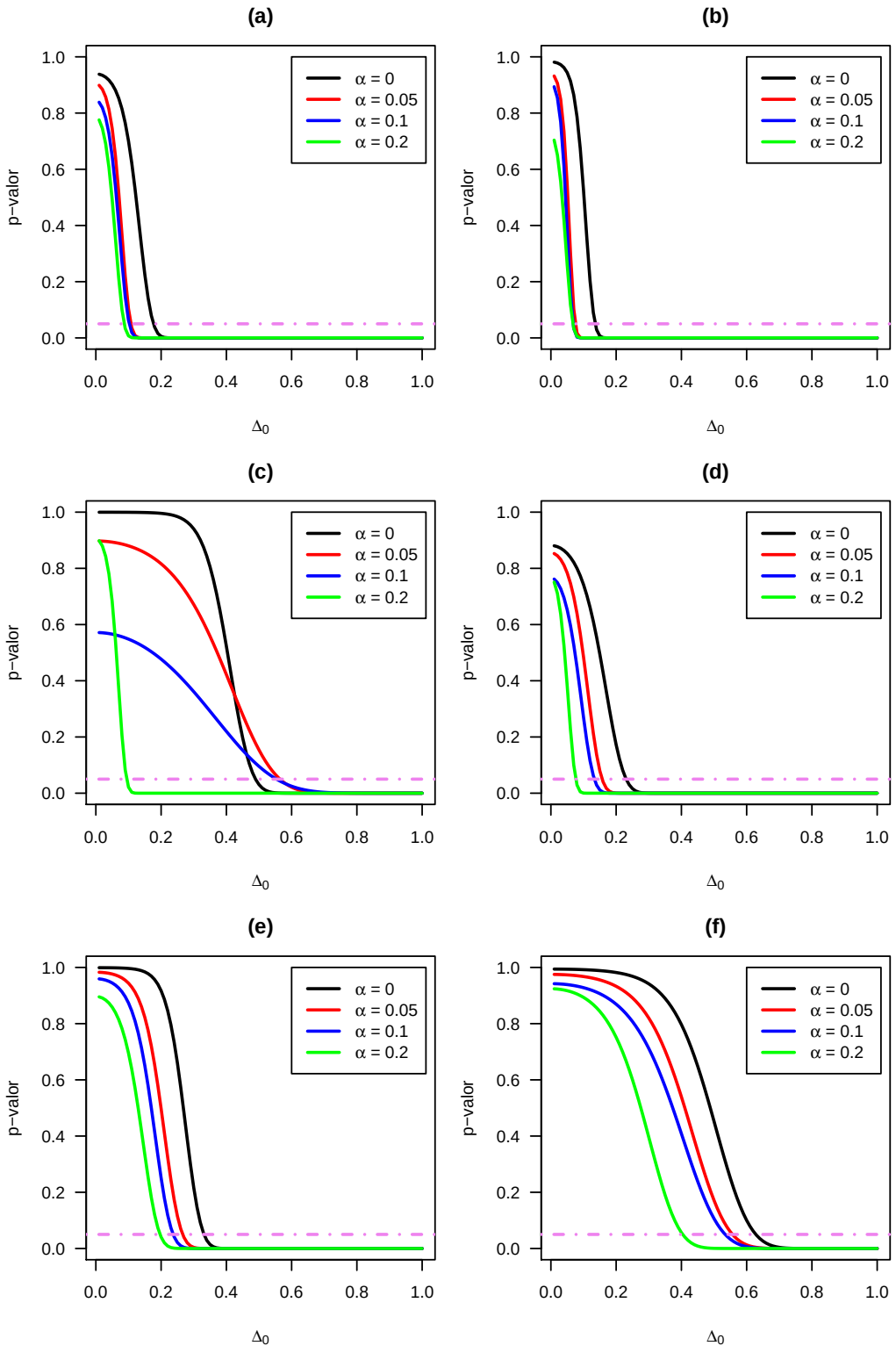

Figura 5.4: Curva de p-valores asintóticos para datos simulados: (a) N(0,1); (b) N(10,4); (c) 0.9 N(0,1) + 0.1 N(-5,1); (d) 0.9 N(0,1) + 0.1 N(-3,1); (e) χ^2_4; y (f) exp(1). La línea de puntos es una línea de referencia ($p = 0.05$).

Los gráficos (c) y (d) se corresponden con los modelos de mezclas $0.9N(0,1)+0.1N(-5,1)$ y $0.9N(0,1)+0.1N(-3,1)$, respectivamente. El comportamiento en ambos es claramente diferente. En el primer caso la contaminación es claramente detectada. Así, si fijamos el nivel de significación $p = 0.05$, la distancia estandarizada significativa es aproximadamente $\Delta_0 = 0.5$ antes de recortar, y este valor desciende hasta $\Delta_0 = 0.07$ cuando $\alpha = 0.2$, similar a los valores observados en el caso normal. Los cruces que se observan en la curva de p-valores cuando $\alpha = 0.05$ ó 0.1 están relacionados con la variabilidad en la estimación de la varianza asintótica $\sigma^2(P,\mathcal{N})$. Estos cruces se observan de vez en cuando cuando el tamaño muestral es bajo o moderado, pero desaparecen en cuanto aumentamos éste.

En el segundo caso, gráfico (d), la contaminación sólo es detectada tímidamente puesto que la distancia estandarizada significativa cuando $p = 0.05$ es sólo ligeramente mayor que la observada en los modelos normales ($\Delta_0 = 0.24$ en (d), miestras que $\Delta_0 = 0.17$ en (a) ó $\Delta_0 = 0.14$ en (b)). Esta distancia desciende hasta valores similares a los de los modelos normales cuando $\alpha = 0.05$ ó 0.1.

Los gráficos (e) y (f) en la Figura 5.4, se corresponden con situaciones en las que la normalidad no se alcanza para niveles razonables de recorte, un modelo χ^2_4 y un modelo exponencial, respectivamente. En ambos casos la distancia estandarizada significativa antes de recortar está claramente lejos de la del modelo normal ($\Delta_0 = 0.34$ y $\Delta_0 = 0.64$ vs $\Delta_0 = 0.17$ en (a)). También se observa una clara mejora en esta distancia cuando se aumenta el tamaño de recorte. Sin embargo, en las dos situaciones esta distancia no alcanza valores comparables a los del modelo normal, incluso cuando $\alpha = 0.2$. En el caso del modelo chi-cuadrado la diferencia respecto al modelo normal es menor que en el caso del modelo exponencial en el que la distancia recortada estandarizada está bastante lejos de la del modelo normal ($\Delta_0 = 0.41$ en (f) vs $\Delta_0 = 0.09$ en (a)). De esta forma, estas muestras no pueden ser consideradas normales a ningún nivel razonable de recorte.

Para la construcción de la curva $\tilde{P}(\Delta_0)$ es preciso el cálculo de los valores de $Z_{n,\alpha}$ definido como en (5.30), además de los de $R^{1/2}_{n,\alpha}$. Estos cálculos se describen en la siguiente nota.

Nota 5.21 (Cálculo de $Z_{n,\alpha}$)**.** En primer lugar, para calcular el valor de $T_{n,\alpha}$ observamos que

$$T_{n,\alpha} = \min_{h\in\mathcal{C}_\alpha,\mu\in\mathbb{R},\sigma\geq 0} \int_0^1 (F_n^{-1} - \mu - \sigma\Phi^{-1})^2 h' = \min_{\mu\in\mathbb{R},\sigma\geq 0} V_n(\mu,\sigma),$$

donde

$$V_n(\mu,\sigma) = \min_{h\in\mathcal{C}_\alpha} \int_0^1 (F_n^{-1} - \mu - \sigma\Phi^{-1})^2 h'.$$

Si $\sigma > 0$ entonces $V_n(\mu,\sigma) = \int_0^1 (F_n^{-1} - \mu - \sigma\Phi^{-1})^2 h'_{n,\mu,\sigma}$, donde

$$h'_{n,\mu,\sigma} = \frac{1}{1-\alpha} I_{|F_n^{-1}-\mu-\sigma\Phi^{-1}|\leq k_{n,\mu,\sigma}}$$

y $k_{n,\mu,\sigma}$ es el (único) k tal que el conjunto $\{t \in (0,1) : |F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t)| \leq k\}$ tiene medida de Lebesgue $1-\alpha$. Usamos esto para calcular numéricamente $V_n(\mu,\sigma)$ como sigue

1. Calculamos los valores de $|F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t)|$ en una malla (fina) de $[0,1]$.

2. Aproximamos $k_{n,\mu,\sigma}$ por el cuantil $(1-\alpha)$ de dichos valores.

3. Aproximamos $V_n(\mu,\sigma)$ por la media de $(F_n^{-1}(t) - \mu - \sigma\Phi^{-1}(t))^2 h'_{n,\mu,\sigma}(t)$ en la malla.

La minimización de $V_n(\mu,\sigma)$ da el valor de $T_{n,\alpha}$. Para llevar a cabo este paso buscamos en una malla de valores de (μ,σ). Este paso puede ser sustituido por un procedimiento basado en métodos de tipo gradiente de los que están disponibles en R (ver, por ejemplo, *nlm*, *optim*), usando como valores iniciales los valores muestrales.

Si μ_n y σ_n son los minimizadores de $V_n(\mu,\sigma)$ obtenidos con el anterior algoritmo, entonces, tomamos

$$h_n = h_{n,\mu_n,\sigma_n}.$$

Finalmente, aproximamos

$$S^2_{n,\alpha} = 4\left[\int_0^1 l_n(t)^2 dt - \left(\int_0^1 l_n(t)dt\right)^2\right]$$

calculando las integrales de forma numérica, donde

$$l_n(t) := \int_{F_n^{-1}(1/n)}^{F_n^{-1}(t)} \left(x - \mu_n - \sigma_n\Phi^{-1}(F_n(x))\right) h'_n(F_n(x))dx$$

se evalúa numéricamente promediando en una malla en $(0,1)$ como se ha mencionado antes. De forma similar calculamos el valor de $R_{n,\alpha}$.

Todos estos cálculos han sido implementados en R en un programa que se adjunta en el Apéndice. Estos cálculos han sido codificados en forma vectorizada y el resultado es que para tamaños n de muestra moderados (100-500) el tiempo requerido en un PC es de unos cuantos segundos, mientras que para tamaños grandes (5000) es de un par de minutos, dependiendo de la malla para (μ,σ). La malla en $[0,1]$ para el cálculo de $V_n(\mu,\sigma)$ se ha obtenido dividiendo $[0,1]$ en 10^5 intervalos de igual longitud.

5.4.3. Ejemplo 3, casi normalidad con datos reales.

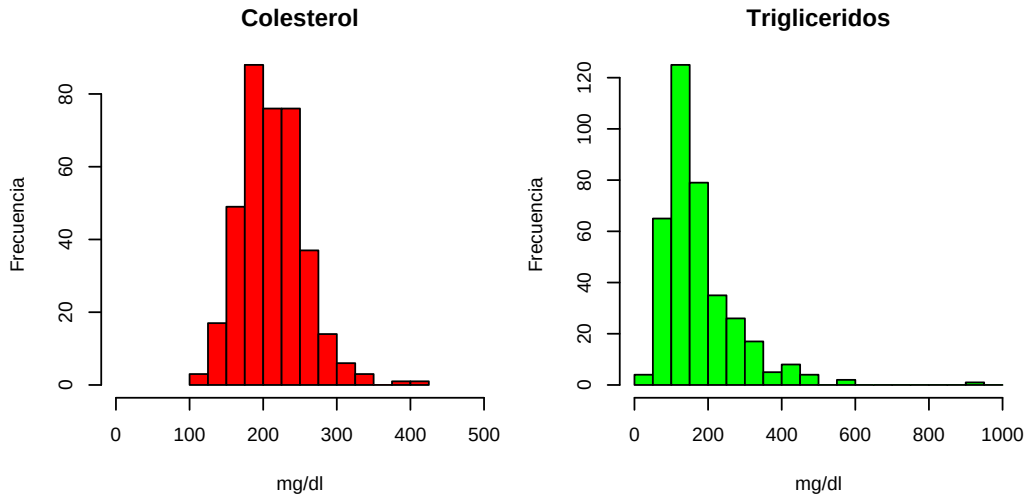

Figure 5.5: Histograma para las variables Colesterol y Trigliceridos.

Los datos de este ejemplo han sido tomados del libro de Hand et al. (1994). Se utilizan las variables concentración de colesterol y de trigliceridos en plasma sanguíneo (mg/dl) recogidas en una muestra de $n = 371$ pacientes, y el objetivo es investigar si estas muestras pueden provenir de un modelo normal para algún nivel razonable de recorte. En la Figura 5.5 se representa el histograma de cada muestra. En la muestra correspondiente al colesterol se aprecia una ligera asimetría a la derecha con dos posibles *outliers*. En cambio, en la muestra de trigliceridos la asimetría a la derecha es clara, así como la existencia de un *outlier* también en la misma cola. Usando procedimientos clásicos como el test de Shapiro-Wilks rechazaríamos la hipótesis de normalidad, incluso si eliminamos los *outliers* antes señalados ($p = 0.0306$ y $p < 0.0000$, respectivamente).

	Colesterol			Trigliceridos		
α	μ_n	σ_n	$\tilde{\tau}_\alpha(P_n, \mathcal{N})$	μ_n	σ_n	$\tilde{\tau}_\alpha(P_n, \mathcal{N})$
0	213.31	42.07	0.026	173.94	88.90	0.193
0.05	212.08	39.64	0.005	165.92	71.27	0.112
0.10	211.75	39.88	0.004	161.16	63.87	0.089
0.20	211.02	41.68	0.002	153.15	53.04	0.048

Tabla 5.3: Distancias estandarizadas, medias y desviaciones estandar de la distribución normal α-recortada más próxima.

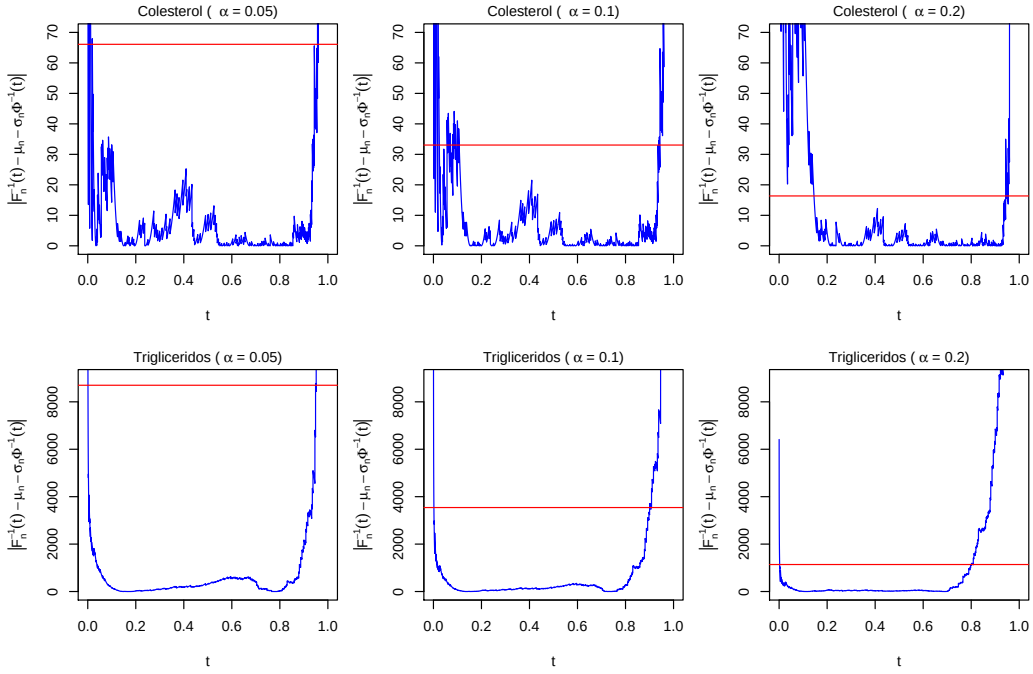

Figura 5.6: Funciones de recorte óptimas para la muestra del colesterol y los triglicéridos, para diferentes tamaños de recorte ($\alpha = 0.05$, 0.1 y 0.2).

En lugar de dichos procedimientos, utilizamos la metología de recortes imparciales con el mismo patrón para obtener la mejor aproximación normal α-recortada a cada muestra. La Tabla 5.3 muestra las distancias estandarizadas, medias y desviaciones estándar de los mejores aproximantes para ambas muestras y para diferentes tamaños de recorte. Una primera idea de las diferencias que vamos a encontrar entre ambas muestras se adivina a la vista de los valores de las distancias a la normalidad en cada una de ellas y su reducción con el tamaño de recorte. En la Figura 5.6 se representan las funciones de recorte óptimo para estas muestras y diferentes valores de α (0.05, 0.1 y 0.2). En cada gráfico se dibuja el valor de $|F_n^{-1}(t) - \mu_n - \sigma_n \Phi^{-1}(t)|$ y los valores de corte k_{n,μ_n,σ_n} donde μ_n y σ_n son la media y la desviación estándar de la distribución normal más cercana (ver Tabla 5.3) estimada usando el algoritmo que se describe en la Nota 5.21. Estos gráficos muestran que para hacer más normal ambas muestras el recorte debe realizarse en las colas de la distribución, pero no de forma simétrica, y no siempre eliminando todas las observaciones de las colas. Además, el gráfico correspondiente a la muestra de colesterol y $\alpha = 0.2$ sugiere que si incrementamos el tamaño de recorte entonces se recortarían algunas observaciones en el centro de la muestra.

La Figura 5.7 muestra las curvas de p-valores $\tilde{P}(\Delta_0)$ para las dos muestras y diferentes tamaños de recorte. Además hay un tercer gráfico que se corresponde con las curvas obtenidas para una muestra aleatoria ($n = 371$) de una normal estándar. Este último gráfico se ha incluido como una referencia para ayudar en la evaluación de la normalidad de las muestras del ejemplo.

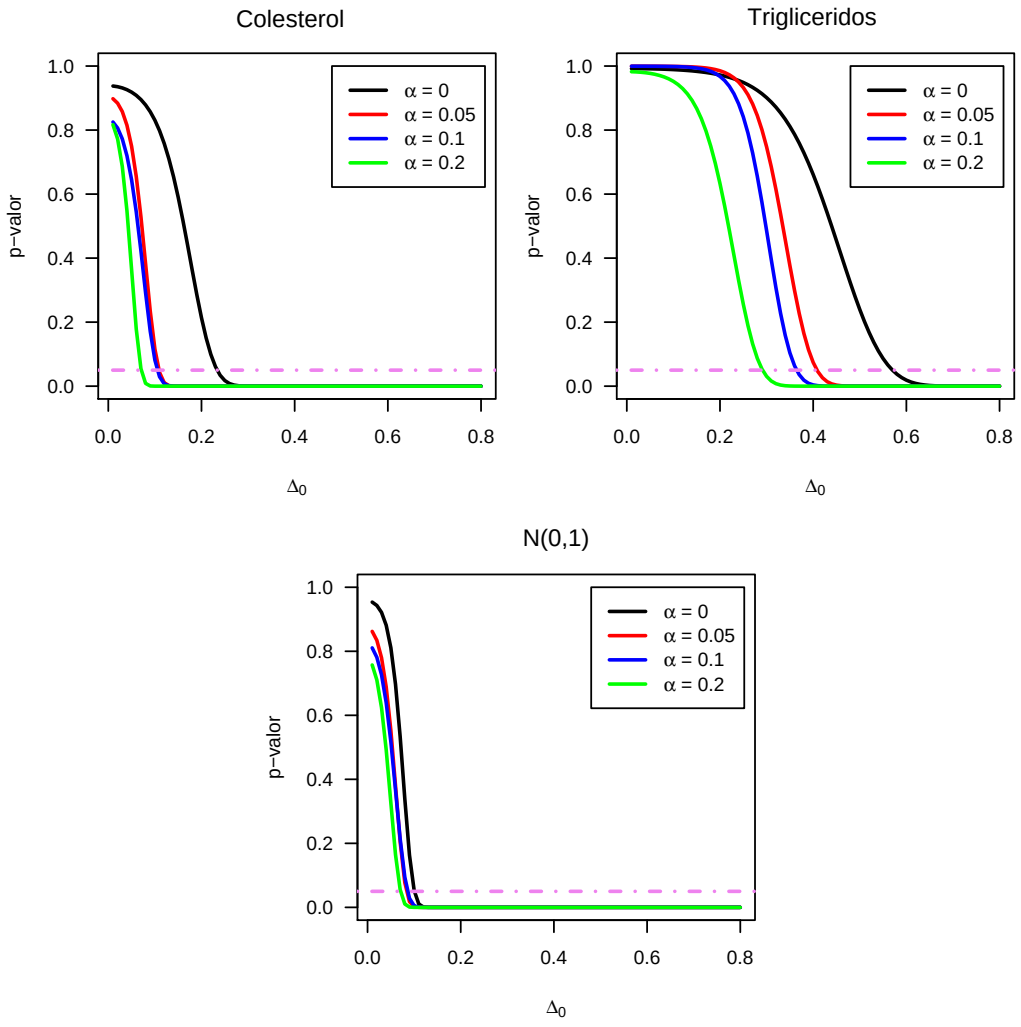

Figura 5.7: Curvas de p-valores asintóticos para el colesterol, los triglicéridos y una muestra de 371 observaciones generadas de una N(0,1). La línea punteada es una línea de referencia ($p = 0.05$).

Aunque en ambos casos se observa una importante mejora después de recortar el primer 5 % de los datos (fijando $p = 0.05$, se pasa de $\Delta_0 = 0.24$ cuando $\alpha = 0$ a $\Delta_0 = 0.11$ cuando $\alpha = 0.05$ para el colesterol; y de $\Delta_0 = 0.57$ a $\Delta_0 = 0.41$ para los triglicéridos), las dos muestras tienen un comportamiento diferente. Mientras que en el primer caso la distancia recortada estandarizada alcanza valores similares a los del modelo normal (tercer gráfico), en el segundo caso esto no ocurre. En esta última muestra la distancia recortada estandarizada no alcanza los mismos niveles aunque recortemos un 20 % ó un 30 % (no mostrada en el gráfico). Así pues, la muestra de colesterol puede ser considera normal a nivel $\alpha = 0.05$ (y por tanto, casi normal), sin embargo, la muestra de triglicéridos no puede ser considerada normal a niveles razonables de recorte.

5.4.4. Simulaciones.

En esta subsección se presentan dos pequeños estudios de simulación que ilustran el comportamiento para muestras finitas de los procedimientos desarrollados en las Secciones 5.2 y 5.3 para el contraste de las hipótesis nulas $H_0 : \tau_\alpha(F, G) \geq \Delta_0^2$ y $H_0 : \tilde{\tau}_\alpha(F, \mathcal{N}) \geq \Delta_0^2$, respectivamente.

El primer estudio que se presenta se refiere al caso de dos muestras. Se han considerado dos modelos normales contaminados de forma diferente, dos tamaños de recorte y varios valores del valor umbral Δ_0. En cada situación se han generado 10000 réplicas del valor del estadístico $Z_{n,m,\alpha}$ definido en (5.31) para varios valores de $n = m$. Dichos valores se han comparado con el 0.05-cuantil de la distribución normal estándar, rechazando H_0 cuando el valor observado era menor que esta cantidad. La Tabla 5.4 muestra las frecuencias de rechazo obtenidas. A partir de dicha tabla se deduce un buen comportamiento en el uso de la distribución asintótica incluso para tamaños de muestra moderados, con frecuencias de rechazo bajas para valores de Δ_0 más pequeños que la verdadera distancia y frecuencias de rechazo altas en caso contrario. Se puede observar también que cuando el valor de Δ_0 es igual a la verdadera distancia entre las distribuciones poblacionales la frecuencia de rechazo se aproxima al nivel nominal del contraste.

$P = 0.95N(0,1) + 0.05N(5,1)$ $Q = 0.95N(0,1) + 0.05N(-5,1)$				$P = 0.9N(0,1) + 0.1N(5,1)$ $Q = 0.9N(0,1) + 0.1N(-5,1)$			
	Δ_0^2	n	freq.		Δ_0^2	n	freq.
	0.1	100	0.0038		0.5	100	0.0062
		200	0.0030			200	0.0028
		500	0.0000			500	0.0000
		1000	0.0000			1000	0.0000
		5000	0.0000			5000	0.0000
	0.25	100	0.0282		1	100	0.0554
		200	0.0208			200	0.0556
		500	0.0084			500	0.0470
		1000	0.0028			1000	0.0270
		5000	0.0000			5000	0.0071
	0.404	100	0.0782		1.116	100	0.0886
$\alpha = 0.05$		200	0.0860	$\alpha = 0.1$		200	0.0852
		500	0.0934			500	0.0878
$(\tau_\alpha(P,Q)$		1000	0.0966	$(\tau_\alpha(P,Q)$		1000	0.0914
$\simeq 0.404)$		5000	0.1080	$\simeq 1.116)$		5000	0.0893
	0.5	100	0.1124		1.25	100	0.1128
		200	0.1440			200	0.1470
		500	0.2130			500	0.1932
		1000	0.2986			1000	0.2538
		5000	0.6893			5000	0.5503
	1	100	0.4220		1.75	100	0.2813
		200	0.6793			200	0.4621
		500	0.9468			500	0.7373
		1000	0.9989			1000	0.9304
		5000	1.0000			5000	1.0000

Tabla 5.4: Potencia simulada para el estadístico $Z_{n,m,\alpha}$.

Finalizamos esta sección con otro pequeño estudio, basado en simulaciones, de la potencia del contraste de casi normalidad que se deriva de la distribución asintótica obtenida en el Teorema 5.19. En este estudio se consideran también dos modelos poblacionales diferentes: $P_1 = 0.9N(0,1) + 0.1N(-3,1)$ y $P_2 = \chi^2_2$, el primero de ellos más cercano a la normalidad (casi normal) y el segundo más alejado de esta. Se contrasta la hipótesis nula

$H_0 : \tilde{\tau}_\alpha(P_i,\mathcal{N}) \geq \Delta_0^2$ contra la alternativa $H_a : \tilde{\tau}_\alpha(P_i,\mathcal{N}) < \Delta_0^2$ para diferentes valores de Δ_0 y dos tamaños de recorte ($\alpha = 0.05$ y $\alpha = 0.1$). Para ello, en cada caso, se obtienen 10000 réplicas del estadístico $Z_{n,\alpha}$ dado por (5.30) para varios valores de n, y seguimos el mismo criterio de rechazo de H_0 que en el anterior estudio. Las Tablas 5.5 y 5.6 contienen estas frecuencias de rechazo para P_1 y P_2 respectivamente. En ambos casos, y al igual que en el anterior estudio, la simulación demuestra que el comportamiento del test, incluso para tamaños de muestra moderados, es bastante bueno.

$P_1 = 0.9N(0,1) + 0.1N(-3,1)$				$P_1 = 0.9N(0,1) + 0.1N(-3,1)$			
	Δ_0^2	n	freq.		Δ_0^2	n	freq.
	0.001	100	0.0000		0.001	100	0.0000
		200	0.0000			200	0.0000
		500	0.0000			500	0.0000
		1000	0.0000			1000	0.0000
		5000	0.0000			5000	0.0000
		10000	0.0000			10000	0.0000
	0.01	100	0.0012		0.005	100	0.0003
		200	0.0012			200	0.0010
		500	0.0000			500	0.0016
		1000	0.0000			1000	0.0010
		5000	0.0000			5000	0.0000
		10000	0.0000			10000	0.0000
	0.0225	100	0.0478		0.0079	100	0.0058
		200	0.0484			200	0.0206
$\alpha = 0.05$		500	0.0355	$\alpha = 0.1$		500	0.0266
		1000	0.0392			1000	0.0304
$(\tilde{\tau}_\alpha(P_1, \mathcal{N})$		5000	0.0445	$(\tilde{\tau}_\alpha(P_1, \mathcal{N})$		5000	0.0326
$\simeq 0.0225)$		10000	0.0520	$\simeq 0.0079)$		10000	0.0460
	0.05	100	0.3410		0.025	100	0.2654
		200	0.5038			200	0.4404
		500	0.8114			500	0.8068
		1000	0.9780			1000	0.9748
		5000	1.0000			5000	1.0000
		10000	1.0000			10000	1.0000
	0.1	100	0.8298		0.05	100	0.7048
		200	0.9732			200	0.9182
		500	1.0000			500	0.9988
		1000	1.0000			1000	1.0000
		5000	1.0000			5000	1.0000
		10000	1.0000			10000	1.0000

Tabla 5.5: Frecuencias de rechazo para $P_1 = 0.9N(0,1) + 0.1N(-3,1)$.

$P_2 = \chi^2_2$				$P_2 = \chi^2_2$			
	Δ_0^2	n	freq.		Δ_0^2	n	freq.
	0.05	100	0.0004		0.01	100	0.0000
		200	0.0000			200	0.0000
		500	0.0000			500	0.0000
		1000	0.0000			1000	0.0000
		5000	0.0000			5000	0.0000
		10000	0.0000			10000	0.0000
	0.1	100	0.0274		0.05	100	0.0014
		200	0.0122			200	0.0008
		500	0.0028			500	0.0000
		1000	0.0000			1000	0.0000
		5000	0.0000			5000	0.0000
		10000	0.0000			10000	0.0000
	0.1272	100	0.0724		0.1022	100	0.0770
		200	0.0534			200	0.0898
$\alpha = 0.05$		500	0.0590	$\alpha = 0.1$		500	0.1042
		1000	0.0570			1000	0.0886
$(\tilde{\tau}_\alpha(P_2, \mathcal{N})$		5000	0.0.418	$(\tilde{\tau}_\alpha(P_2, \mathcal{N})$		5000	0.0578
$\simeq 0.1272)$		10000	0.0378	$\simeq 0.1022)$		10000	0.0486
	0.15	100	0.1296		0.15	100	0.2724
		200	0.1594			200	0.4378
		500	0.2528			500	0.6770
		1000	0.3718			1000	0.8828
		5000	0.9196			5000	1.0000
		10000	0.9953			10000	1.0000
	0.25	100	0.5231		0.25	100	0.7260
		200	0.8198			200	0.9352
		500	0.9954			500	0.9996
		1000	1.0000			1000	1.0000
		5000	1.0000			5000	1.0000
		10000	1.0000			10000	1.0000

Tabla 5.6: Frecuencias de rechazo para $P_2 = \chi^2_2$.

5.5. Recorte sin restricciones y sobreajuste.

En la Sección 5.2 vimos que la tasa de convergencia de la distancia $\mathcal{W}_2$ entre la distribución empírica α-recortada con el mismo patrón y la verdadera distribución (Teorema 5.12) era de orden n. En esta sección estudiamos qué ocurre con esta tasa cuando tratamos con un recorte general. A través de resultados parciales relacionados con el recorte cuando $\alpha = 1$ y de una simulación se mostrará que en este caso el problema es considerablemente más difícil.

Sean $X_1, \ldots, X_n$ variables aleatorias i.i.d. con distribución común $P \in \mathcal{P}_2(\mathbb{R})$, $P \ll \ell$. Sean F y f, sus funciones de distribución y de densidad, respectivamente. Como habitualmente, denotaremos por P_n la distribución empírica basada en $X_1, \ldots, X_n$ y por $P_{n,\alpha}$ $(0 \leq \alpha \leq 1)$ el único (Teorema 3.26) recorte de nivel a lo sumo α de P_n que verifica

$$\mathcal{W}_2(P_{n,\alpha}, P) = \min_{R \in \mathcal{R}_\alpha(P_n)} \mathcal{W}_2(R, P).$$

Entonces, de forma obvia tenemos $\mathcal{W}_2(P_{n,\alpha_1}, P) \geq \mathcal{W}_2(P_{n,\alpha_2}, P)$ si $\alpha_1 \leq \alpha_2$ y de esto se deduce que

$$n\mathcal{W}_2^2(P_{n,\alpha}, P) \leq n\mathcal{W}_2^2(P_n, P)$$

y por tanto que $n\mathcal{W}_2^2(P_{n,\alpha}, P)$ está estocásticamente acotado para cada $\alpha \in [0,1]$ si se tiene que $\int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} dt < \infty$ (ver del Barrio et al., 2005).

A continuación estudiaremos la tasa para $\mathcal{W}_2^2(P_{n,1}, P)$ y veremos como es en algunos casos n^2 en vez de n. Una de las ventajas de considerar $\mathcal{W}_2^2(P_{n,1}, P)$ es que podemos calcular explícitamente el minimizador, como se vió en la Proposición 3.31, donde además se obtuvo la expresión (3.36) para el cuadrado de la distancia. De esta forma, y para algunos casos particulares, la anterior expresión es suficiente para obtener la tasa exacta de convergencia de $\mathcal{W}_2^2(P_{n,1}, P)$, como podemos ver en el siguiente ejemplo

Ejemplo 5.22. Si tomamos $P = U(0,1)$ y $U_1, \ldots, U_n$ es una muestra de variables aleatorias i.i.d. de P, entonces $F(t) = F^{-1}(t) = t$ y

$$\int_{U_{(i)}}^{\frac{U_{(i)}+U_{(i+1)}}{2}} (t - U_{(i)})^2 dt = \int_{\frac{U_{(i)}+U_{(i+1)}}{2}}^{U_{(i+1)}} (t - U_{(i+1)})^2 dt = \frac{(U_{(i+1)} - U_{(i)})^3}{24},$$

$$\int_0^{U_{(1)}} (t - U_{(1)})^2 dt = \frac{U_{(1)}^3}{3}, \quad \int_{U_{(n)}}^1 (t - U_{(n)})^2 dt = \frac{(1 - U_{(n)})^3}{3}.$$

Por tanto,

$$\mathcal{W}_2^2(P_{n,1},P)=\frac{U_{(1)}^3}{3}+\frac{(1-U_{(n)})^3}{3}+\frac{1}{12}\sum_{i=1}^{n-1}(U_{(i+1)}-U_{(i)})^3.$$

A partir de aquí, si utilizamos la representación del estadístico de orden uniforme en términos de las sumas parciales de variables exponenciales independientes, $(U_{(1)},\ldots,U_{(n)})\overset{d}{=}(S_1/S_{n+1},\ldots,S_n/S_{n+1})$ (donde $S_n=\sum_{i=1}^n X_i$, $\mathcal{L}(X_i)=exp(1)$), tenemos que,

$$n^2\mathcal{W}_2^2(P_{n,1},P)\overset{d}{=}\frac{n^3}{S_{n+1}^3}\left(\frac{1}{3}\frac{X_1}{n}+\frac{1}{3}\frac{X_n}{n}+\frac{1}{12}\frac{\sum_{i=1}^n X_i^3}{n}\right).$$

Utilizando ahora la Ley de los Grandes Números, $\frac{S_{n+1}^3}{n^3}\underset{c.s}{\to}1$ y $\frac{\sum_{i=1}^n X_i^3}{n}\underset{c.s}{\to}EX_1^3=6$, y entonces es inmediato que

$$n^2\mathcal{W}_2^2(P_{n,1},P)\underset{c.s.}{\to}\frac{1}{2}.$$

5.5.1. Tasas medias de convergencia.

El argumento del Ejemplo 5.22 no es fácilmente generalizable y de hecho la tasa de $\mathcal{W}_2^2(P_{n,1},P)$ es menor que n^2 incluso para algunas distribuciones con soporte acotado. A continuación se da una expresión para el valor esperado de $\mathcal{W}_2^2(P_{n,1},P)$ de la que obtendremos una condición suficiente para que la tasa media de convergencia sea n^2.

A lo largo de esta subsección $B(\alpha,\beta)$ representará la función Beta en (α,β),

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

Comenzamos con un lema en el que se obtiene una descomposición bastante útil de $\mathrm{E}(\mathcal{W}_2^2(P_{n,1},P))$.

Lema 5.23. *Dado un número natural n, y la función de distribución F, definimos*

$$\begin{aligned}
A_n(F) &:= n\int_0^1(1-x)^{n-1}\left(\int_0^x(F^{-1}(t)-F^{-1}(x))^2dt\right)dx\\
B_n(F) &:= n\int_0^1 x^{n-1}\left(\int_x^1(F^{-1}(t)-F^{-1}(x))^2dt\right)dx\\
C_n(F) &:= n(n-1)\iint_{0<x<y<1}(1-(y-x))^{n-2}\left(\int_x^{F_{x,y}}(F^{-1}(t)-F^{-1}(x))^2dt\right)dxdy\\
D_n(F) &:= n(n-1)\iint_{0<x<y<1}(1-(y-x))^{n-2}\left(\int_{F_{x,y}}^{y}(F^{-1}(t)-F^{-1}(y))^2dt\right)dxdy,
\end{aligned}$$

donde $F_{x,y}=F((F^{-1}(x)+F^{-1}(y))/2)$. Se verifica que

$$E(\mathcal{W}_2^2(P_{n,1},P))=A_n(F)+B_n(F)+C_n(F)+D_n(F). \tag{5.32}$$

Demostración. Podemos suponer sin pérdida de generalidad que $X_i = F^{-1}(U_i)$ donde $\{U_i\}$ son v.a.i.i.d. uniformes en $(0,1)$. Sean $a_n(F)$, $b_n(F)$, $c_n(F)$, y $d_n(F)$ los cuatro términos de la expresión para $\mathcal{W}_2^2(P_{n,1}, P)$ en la Proposición 3.31. Podemos escribir entonces

$$a_n(F) = \int_0^{U_{(1)}} (F^{-1}(t) - F^{-1}(U_{(1)}))^2 dt.$$

Teniendo en cuenta que la función de densidad de $U_{(1)}$ es $n(1-x)^{n-1}$ para $0 < x < 1$ concluimos que $\mathrm{E}(a_n(F)) = A_n(F)$, primer término en (5.32). Un argumento similar se aplica a $b_n(F)$ para obtener $B_n(F)$. Respecto de $C_n(F)$ observamos que

$$\int_{F(X_{(i)})}^{F((X_{(i)}+X_{(i+1)})/2)} (F^{-1}(t) - X_{(i)})^2 dt = \int_{U_{(i)}}^{F(F^{-1}(U_{(i)})+F^{-1}(U_{(i+1)})/2)} (F^{-1}(t) - F^{-1}(U_{(i)})^2 dt.$$

La densidad conjunta de $(U_{(i)}, U_{(i+1)})$ es $g_i(x,y) = n(n-1)\binom{n-2}{i-1} x^{i-1}(1-y)^{n-i-1}$ para $0 < x < y < 1$. Luego, $\mathrm{E}(c_n(F))/(n(n-1))$ es igual a

$$\sum_{i=1}^{n-1} \iint_{0<x<y<1} \binom{n-2}{i-1} x^{i-1}(1-y)^{n-i-1} \left(\int_x^{F((F^{-1}(x)+F^{-1}(y))/2)} (F^{-1}(t) - F^{-1}(x))^2 dt \right) dxdy.$$

Aplicando ahora la fórmula binomial tenemos que $\mathrm{E}(c_n(F)) = C_n(F)$. Finalmente, $D_n(F)$ se obtiene de forma similar.

■

La expresión (5.33) en el siguiente corolario se obtiene como caso particular del Teorema 5.25 pero la presentamos aquí, de forma separada porque la utilizaremos para la demostración de dicho teorema, ya que parte de la demostración consiste en reducir el problema al caso uniforme.

Corolario 5.24. *Si U es la distribución uniforme en $(0,1)$, entonces $E(\mathcal{W}_2^2(P_{n,1}, U)) = \frac{n+9}{2(n+1)(n+2)(n+3)}$, y*

$$\lim_n n^2 E(\mathcal{W}_2^2(P_{n,1}, U)) = \frac{1}{2}. \tag{5.33}$$

Demostración. De acuerdo con las definiciones dadas en el Lema 5.23 tenemos que $B_n(U) = A_n(U)$ y también que

$$\begin{aligned} A_n(U) &= n \int_0^1 (1-x)^{n-1} \left(\int_0^x (t-x)^2 dt \right) dx = \frac{n}{3} \int_0^1 (1-x)^{n-1} x^3 dx = \frac{n}{3} B(n,4) \\ &= \frac{2}{(n+1)(n+2)(n+3)}. \end{aligned}$$

Por otra parte, también se tiene $D_n(U) = C_n(U)$ y

$$\begin{aligned} C_n(U) &= n(n-1)\iint_{0<x<y<1}(1-(y-x))^{n-2}\left(\int_x^{(x+y)/2}(t-x)^2dt\right)dxdy \\ &= \frac{n(n-1)}{24}\int_0^1\left(\int_0^y(1-(y-x))^{n-2}(y-x)^3dx\right)dy. \end{aligned} \tag{5.34}$$

Integrando por partes 3 veces obtenemos que

$$\begin{aligned} \int_0^y(1-(y-x))^{n-2}(y-x)^3dx &= -\frac{1}{n-1}(1-y)^{n-1}y^3 - \frac{3}{n(n-1)}(1-y)^n y^2 \\ &\quad -\frac{6}{n(n-1)(n+1)}(1-y)^{n+1}y \\ &\quad +\frac{6}{n(n-1)(n+1)(n+2)}\left(1-(1-y)^{n+2}\right), \end{aligned}$$

de donde, tenemos que

$$\begin{aligned} C_n(U) &= \frac{n(n-1)}{24}\left[-\frac{1}{n-1}B(n,4) - \frac{3}{n(n-1)}B(n+1,3) - \frac{6}{n(n-1)(n+1)}B(n+2,2)\right. \\ &\qquad \left. +\frac{6}{n(n-1)(n+1)(n+2)}\left(1-\frac{1}{n+3}\right)\right] \\ &= \frac{1}{4(n+2)(n+3)}. \end{aligned}$$

Teniendo en cuenta (5.32), sumamos esos valores y obtenemos el resultado.

■

Teorema 5.25. *Supongamos que P tiene función de densidad, f, continua y estrictamente positiva en $Sop(P) = [l,u]$. Entonces,*

$$\lim_n n^2 E(\mathcal{W}_2^2(P_{n,1},P)) = \frac{1}{2}\int_0^1 \frac{1}{f^2(F^{-1}(y))}dy.$$

Demostración. Usaremos la descomposición que hemos obtenido en el Lema 5.32. De las hipótesis tenemos que existen m, M tales que $m \le f(x) \le M$ para cada $x \in [l,u]$. Utilizando el Teorema del Valor Medio tenemos que si $0 < u \le v < 1$, entonces

$$|F^{-1}(v) - F^{-1}(u)| = (v-u)\frac{1}{f(\varepsilon_{u,v})} \le \frac{1}{m}(v-u), \text{ donde } \varepsilon_{u,v} \in (u,v). \tag{5.35}$$

En las siguientes desigualdades, H será una constante que puede cambiar de una expre-

sión a la siguiente. A partir de (5.35), obtenemos que

$$\begin{aligned} A_n(F) &\leq Hn\int_0^1 (1-x)^{n-1}\left(\int_0^x (x-t)^2 dt\right) dx \\ &= Hn\int_0^1 (1-x)^{n-1}x^3 dx \\ &= HnB(n,4) \qquad (5.36) \\ &= H\frac{1}{(n+3)(n+2)(n+1)} = O(n^{-3}). \qquad (5.37) \end{aligned}$$

De forma similar obtendríamos que $B_n(F) = O(n^{-3})$ y , en consecuencia tenemos que

$$\lim_n n^2 \mathrm{E}(\mathcal{W}_2^2(P_{n,1}, P)) = \lim_n n^2 (C_n(F) + D_n(F)). \qquad (5.38)$$

Si definimos

$$V_F(x,y) = \int_x^{F((F^{-1}(x)+F^{-1}(y))/2)} (F^{-1}(t) - F^{-1}(x))^2 dt,$$

tenemos que

$$C_n(F) = n(n-1)\int\int_{0<x<y<1} (1-(y-x))^{n-2} V_F(x,y) dxdy.$$

Si b es un número positivo fijo, entonces

$$C_n(F) \geq n(n-1)\int_{\frac{b}{n}}^1 \int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2} V_F(x,y) dxdy.$$

Analicemos ahora el término $V_F(x,y)$ si x, y satisfacen que $0 < y - x < bn^{-1}$. En primer lugar, puesto que $F((F^{-1}(x) + F^{-1}(y))/2) \leq y$, si $t \in (x, F((F^{-1}(x) + F^{-1}(y))/2))$, entonces $t \in (x,y)$ y tenemos que

$$F^{-1}(t) - F^{-1}(x) = (t-x)\frac{1}{f(F^{-1}(\varepsilon_t))}$$

La continuidad uniforme de f, el hecho de que sea estrictamente positiva y que $|t-y| < b/n$ implican que existe $k_n \leq 1$, con $\lim_n k_n = 1$ tal que

$$\frac{1}{f(F^{-1}(\varepsilon_t))} \geq \frac{\sqrt{k_n}}{f(F^{-1}(y))}.$$

Respecto al límite superior de integración en V_F tenemos que

$$F^{-1}(y) = F^{-1}(x) + (y-x)\frac{1}{f(F^{-1}(\varepsilon_y))},$$

donde $\varepsilon_y \in (x,y)$ y, entonces,

$$\begin{aligned} F((F^{-1}(x)+F^{-1}(y))/2) &= F\left(F^{-1}(x)+\frac{y-x}{2}\frac{1}{f(F^{-1}(\varepsilon_y))}\right) \\ &= F\left(F^{-1}(x)\right)+\frac{y-x}{2}\frac{f(\varphi_y)}{f(F^{-1}(\varepsilon_y))} \\ &= x+\frac{y-x}{2}\frac{f(\varphi_y)}{f(F^{-1}(\varepsilon_y))} \end{aligned} \tag{5.39}$$

donde $\varphi_y \in (x,(y-x)/2m)$, y, en consecuencia, la distancia entre φ_y e y también es de orden n^{-1}. Por tanto, la continuidad de f y que sea estrictamente positiva, aseguran que existe una nueva sucesión h_n tal que $h_n \leq 1$ con $\lim_n h_n = 1$ y

$$h_n \leq \frac{f(\varphi_y)}{f(F^{-1}(\varepsilon_y))}.$$

De todo ello junto, se deduce que

$$V_F(x,y) \geq \int_x^{x+\frac{h_n(y-x)}{2}} \frac{k_n}{f^2(F^{-1}(y))}(t-x)^2dt = \frac{k_n h_n^3}{24 f^2(F^{-1}(y))}\left(y-x\right)^3.$$

Y, entonces,

$$C_n(F) \geq n(n-1)\frac{k_n h_n^3}{24}\int_{\frac{b}{n}}^1 \frac{1}{f^2(F^{-1}(y))}\left(\int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2}(y-x)^3dx\right)dy.$$

La integral interior en esta expresión se puede calcular integrando por partes (tres veces) y obtenemos que

$$\begin{aligned} \int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2}(y-x)^3dx &= -\frac{1}{n-1}\left(1-\frac{b}{n}\right)^{n-1}\left(\frac{b}{n}\right)^3 - \frac{3}{n(n-1)}\left(1-\frac{b}{n}\right)^n\left(\frac{b}{n}\right)^2 \\ &\quad -\frac{6}{n(n-1)(n+1)}\left(1-\frac{b}{n}\right)^{n+1}\frac{b}{n} \\ &\quad +\frac{6}{n(n-1)(n+1)(n+2)}\left[1-\left(1-\frac{b}{n}\right)^{n+2}\right] \end{aligned}$$

Por tanto,

$$\begin{aligned} &\lim_n n^3(n-1)\int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2}(y-x)^3dx \\ &\quad = \lim_n\left[-\left(1-\frac{b}{n}\right)^{n-1}b^3 - 3\left(1-\frac{b}{n}\right)^n b^2 - \frac{6n}{(n+1)}\left(1-\frac{b}{n}\right)^{n+1}b\right. \\ &\qquad \left. +\frac{6n^2}{(n+1)(n+2)}\left(1-\left(1-\frac{b}{n}\right)^{n+2}\right)\right] \\ &\quad = 6-e^{-b}(b^3+3b^2+6b+6), \end{aligned}$$

y, en consecuencia,

$$\liminf_n n^2 C_n(F) \geq \left[6 - e^{-b}(b^3 + 3b^2 + 6b + 6)\right] \frac{1}{24} \int_0^1 \frac{1}{f^2(F^{-1}(y))} dy.$$

Pero, la desigualdad previa es válida para cada $b > 0$ y entonces, también es válida si tomamos límites, cuando $b \to \infty$, lo que da

$$\liminf_n n^2 C_n(F) \geq \frac{1}{4} \int_0^1 \frac{1}{f^2(F^{-1}(y))} dy. \tag{5.40}$$

De forma análoga, obtendríamos que

$$\liminf_n n^2 D_n(F) \geq \frac{1}{4} \int_0^1 \frac{1}{f^2(F^{-1}(y))} dy.$$

Teniendo en cuenta (5.38), tenemos que, para finalizar la demostración, sólo tenemos que probar que

$$\lim_{b\to\infty} \lim_{n\to\infty} \left(C_n(F) - n(n-1) \int_{\frac{b}{n}}^1 \int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2} V_F(x,y) dxdy \right) = 0. \tag{5.41}$$

Para ello, observemos que si n es un número natural, b es positivo y definimos

$$A_{n,b} = (0,1) \times (0,1) - \left(\frac{b}{n}, 1\right) \times \left(y - \frac{b}{n}, y\right),$$

entonces tenemos que

$$\begin{aligned} C_n(F) - n(n-1) \int_{\frac{b}{n}}^1 \int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2} V_F(x,y) dxdy & \\ = \quad n(n-1) \iint_{A_{n,b}} (1-(y-x))^{n-2} V_F(x,y) dxdy. & \end{aligned} \tag{5.42}$$

Ahora, sea $0 \leq x \leq y$. De la desigualdad (5.39) se obtiene que

$$F((F^{-1}(x) + F^{-1}(y))/2) \leq x + \frac{M}{m} \frac{y-x}{2},$$

que junto con (5.35) da

$$V_F(x,y) \leq \int_x^{x+\frac{M}{2m}(y-x)} \frac{1}{m^2}(t-x)^2 dt = \frac{M^3}{24m^5}(y-x)^3.$$

De esta desigualdad y (5.42) tenemos que

$$\begin{aligned} & n(n-1) \iint_{A_{n,b}} (1-(y-x))^{n-2} V_F(x,y) dxdy \\ & \quad \leq \quad \frac{M^3}{24m^5} n(n-1) \iint_{A_{n,b}} (1-(y-x))^{n-2} (y-x)^3 dxdy \\ & \quad = \quad \frac{M^3}{24m^5} n(n-1) \iint_{A_{n,b}} (1-(y-x))^{n-2} V_U(x,y) dxdy \\ & \quad = \quad \frac{M^3}{24m^5} \left(C_n(U) - n(n-1) \int_{\frac{b}{n}}^1 \int_{y-\frac{b}{n}}^y (1-(y-x))^{n-2} V_U(x,y) dxdy \right) \end{aligned}$$

de donde deducimos que para probar (5.41) para una distribución general F es suficiente con probarlo si F está uniformemente distribuida en $(0,1)$. Entonces, la demostración concluye puesto que esto se deduce de forma trivial de (5.40) aplicado a la distribución uniforme y el Corolario 5.24.

■

5.5.2. El proceso empírico recortado.

La clave para obtener la tasa de convergencia de $\mathcal{W}_2^2(P_{n,\alpha}, P)$ y las distribuciones asintóticas correspondientes en el problema de recortes con el mismo patrón es la aparición de un funcional que depende directamente del proceso empírico cuantil, $\rho_n(t) = \sqrt{n}(F_n^{-1}(t) - F^{-1}(t))$. Una aproximación similar en el caso que nos ocupa requeriría el conocimiento del comportamiento asintótico del que podríamos denominar proceso empírico cuantil recortado,

$$\varrho_n(t) := g(n,\alpha)(F_{n,\alpha}^{-1}(t) - F^{-1}(t)) = g(n,\alpha)(F_n^{-1}(h_{n,\alpha}^{-1}(t)) - F^{-1}(t)),$$

donde $g(n,\alpha)$ es una función que depende de n y posiblemente de α (dependencia sugerida por el estudio que se muestra a continuación), y $h_{n,\alpha} \in \mathcal{C}_\alpha$.

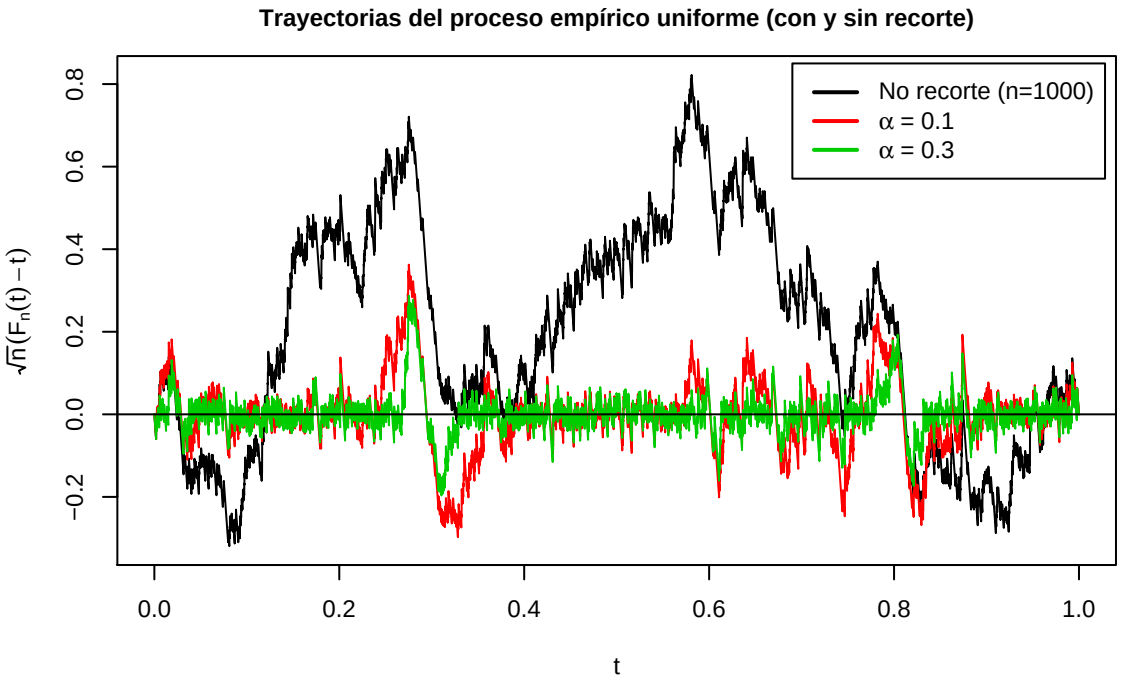

Figura 5.8: Una trayectoria del proceso empírico uniforme y el recortado.

Este es un problema abierto, pero la simulación que se muestra a continuación sugiere que el tamaño de recorte, α, puede tener algún tipo de intervención en la tasa de convergencia. En la Figura 5.8 se muestra una trayectoria del proceso empírico uniforme cuando $n = 1000$ (línea negra) y el proceso empírico recortado cuando $g(n, \alpha) = \sqrt{n}$ con los mismos datos y dos tamaños de recorte. Se observa que la aparición del recorte (paso de $\alpha = 0$ a $\alpha = 0.1$) produce una disminución bastante pronunciada de la "amplitud" en la trayectoria del proceso (línea roja) que se corresponde con una especie de "sobreajuste". Si aumentamos el tamaño de recorte hasta $\alpha = 0.3$ esta amplitud sigue disminuyendo (línea verde), pero la disminuación no es tan marcada como en la aparición del recorte.

A continuación se da un resultado (Teorema 5.26) que justifica el comportamiento observado en la Figura 5.8 para el proceso $\sqrt{n}(F_n^{-1}(h_{n,\alpha}^{-1}(t)) - t)$. En dicho resultado se ve que este proceso converge en L_2 en probabilidad a 0, a diferencia del proceso empírico cuantil que converge (débilmente) a un puente browniano.

Como hasta ahora $X_1, \ldots, X_n$ denotará una muestra de variables aleatorias i.i.d. con distribución común $P \in \mathcal{P}_2(\mathbb{R})$, y F y f serán sus funciones de distribución y densidad, respectivamente. De forma similar, G y g serán las funciones de distribución y densidad de otra distribución $Q \in \mathcal{P}_2(\mathbb{R})$. Si P_n es la medida empírica asociada a la muestra anterior, sabemos que tanto $P_{n,\alpha} = \arg\min_{R \in \mathcal{R}_\alpha(P_n)} \mathcal{W}_2(R, Q)$ como $P_\alpha = \arg\min_{R \in \mathcal{R}_\alpha(P)} \mathcal{W}_2(R, Q)$ están bien definidas, y dado que Q tiene densidad, son únicas (ver los resultados del Capítulo 2).

En el Teorema 5.26 se trata el caso en el que $\mathcal{W}_2(P_\alpha, Q) = 0$. En el caso particular en el que $P = Q$ y $\alpha = 0$ (y por tanto, $P_{n,0} = P_n$) se sabe que, bajo suficientes condiciones de integrabilidad, $n\mathcal{W}_2^2(P_n, Q) = O_P(1)$ (ver, por ejemplo, del Barrio et al., 2005). De aquí, a partir de la acotación $\mathcal{W}_2(P_{n,\alpha}, Q) \leq \mathcal{W}_2(P_n, Q)$, tenemos que $n\mathcal{W}_2^2(P_{n,\alpha}, Q) = O_P(1)$. El siguiente teorema prueba que, de hecho, $n\mathcal{W}_2^2(P_{n,\alpha}, Q) = o_P(1)$, incluso si $P \neq Q$. En primer lugar supondremos que P y Q satisfacen

$$\mathcal{W}_2(P_{\alpha'}, Q) = 0 \quad \text{para algún} \quad \alpha' \in (0, \alpha), \tag{5.43}$$

es decir, Q puede ser recuperada a partir de P recortando una fracción de su masa menor que α. La segunda condición, más técnica, será que

$$f \text{ y } g \text{ con soporte en } [a, b], \text{ son continuas y estrictamente positivas;}$$
$$\text{y } f \text{ tiene una derivada continua.} \tag{5.44}$$

Mientras que (5.43) parece ser necesaria para el resultado que sigue, (5.44) está lejos de ser óptima. Dado que no es nuestro objetivo aquí no perseguiremos afinar esta última condición.

Teorema 5.26. *Si $\mathcal{W}_2(P_{\alpha'}, Q) = 0$ para algún $\alpha' \in (0, \alpha)$ y se verifica (5.44) entonces*

$$n\mathcal{W}_2^2(P_{n,\alpha}, Q) \to 0 \quad \textit{en probabilidad} \tag{5.45}$$

Demostración. En primer lugar observemos que (5.43) implica que $h_\alpha(F(x)) = G(x)$ y de esto y la condición (5.44) obtenemos que

$$\delta \leq h'_\alpha(t) = \frac{g(F^{-1}(t))}{f(F^{-1}(t))} \leq \frac{1}{1-\alpha} - \delta,\ 0 \leq t \leq 1 \tag{5.46}$$

para algún $\delta > 0$ y también que h'_α es una función continua. Ahora, si escribimos

$$\mathcal{W}_2^2(P_{n,\alpha}, Q) = \min_{h \in \mathcal{C}_\alpha} \mathcal{W}_2^2((P_n)_h, Q) = \min_{h \in \mathcal{C}_\alpha} \int_0^1 (F_n^{-1}(h^{-1}(t)) - G^{-1}(t))^2 dt \tag{5.47}$$

tendremos que $n\mathcal{W}_2^2(P_{n,\alpha}, Q) = \min_{h \in \mathcal{C}_\alpha} M_n(h)$, donde

$$M_n(h) = \int_0^1 \left(\frac{\rho_n(t)}{f(F^{-1}(t))} - \sqrt{n}(G^{-1}(h(t)) - F^{-1}(t)) \right)^2 h'(t) dt$$

y $\rho_n(t) = \sqrt{n} f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$ es el proceso cuantil con pesos. Sin pérdida de generalidad podemos suponer que $\{X_n\}_n$ están definidas en un espacio de probabilidad lo suficientemente rico como para que existan puentes brownianos B_n que satisfacen

$$n^{1/2-\nu} \sup_{\frac{1}{n} \leq t \leq 1-\frac{1}{n}} \frac{|\rho_n(t) - B_n(t)|}{(t(1-t))^\nu} = \begin{cases} O_P(\log n), & \text{si } \nu = 0 \\ O_P(1), & \text{si } 0 < \nu \leq 1/2 \end{cases} \tag{5.48}$$

(esto es posible gracias a la condición (5.44) y al Teorema 6.2.1 en Csörgő y Horváth (1993)). Ahora, si definimos

$$\tilde{N}_n(h) = \int_0^1 \left(\frac{B_n(t)}{f(F^{-1}(t))} - \sqrt{n}(G^{-1}(h(t)) - F^{-1}(t)) \right)^2 h'(t) dt,$$

tenemos que

$$\sup_{h \in \mathcal{C}_\alpha} |M_n(h)^{1/2} - \tilde{N}_n(h)^{1/2}| \leq \left(\frac{1}{1-\alpha} \int_0^1 \left(\frac{\rho_n(t) - B_n(t)}{f(F^{-1}(t))} \right)^2 dt \right)^{1/2} = o_P(1). \tag{5.49}$$

Esta última igualdad se obtiene de (5.48), ya que, si tomamos $\nu = 0$ y usamos (5.44), tenemos

$$\int_{1/n}^{1-1/n} \left(\frac{\rho_n(t) - B_n(t)}{f(F^{-1}(t))} \right)^2 dt \leq \frac{\log n}{\sqrt{n}} \int_0^1 \frac{1}{f^2(F^{-1}(t))} dt O_P(1) = o_P(1).$$

A partir de la expresión (5.49) tenemos que (5.45) es cierta si probamos que $\text{mín}_{h\in\mathcal{C}_\alpha}\tilde{N}_n(h)\to 0$ es probabilidad o, de forma equivalente, si probamos que $\text{mín}_{h\in\mathcal{C}_\alpha}N_n(h)\to 0$ en probabilidad, donde

$$N_n(h)=\int_0^1\left(\frac{B(t)}{f(F^{-1}(t))}-\sqrt{n}(G^{-1}(h(t))-F^{-1}(t))\right)^2h'(t)dt$$

y B es un puente browniano fijo. Para ver que $\text{mín}_{h\in\mathcal{C}_\alpha}N_n(h)\to 0$ en probabilidad debemos tener en cuenta que $\text{mín}_{h\in\mathcal{C}_\alpha}N_n(h)\le\frac{1}{1-\alpha}\,\text{mín}_{k\in\mathcal{G}_{\alpha,n}}R_n(k)$, donde

$$R_n(k)=\int_0^1\left(\frac{B(t)}{f(F^{-1}(t))}-\sqrt{n}(G^{-1}(h_\alpha(t)+k(t)/\sqrt{n})-F^{-1}(t))\right)^2dt$$

y $\mathcal{G}_{\alpha,n}$ es el conjunto de las funciones reales absolutamente continuas en $[0,1]$ tales que $k(0)=k(1)=0$ y $-\sqrt{n}h'_\alpha(t)\le k'(t)\le\sqrt{n}\left(\frac{1}{1-\alpha}-h'_\alpha(t)\right)$ para casi todo t. Obsérvese que $\mathcal{G}_{\alpha,n}\subset\mathcal{G}_{\alpha,n+1}$ para cada n y también que, por (5.46), $\mathcal{G}:=\cup_{n\ge1}\mathcal{G}_{\alpha,n}$ es el conjunto de todas las funciones absolutamente continuas en $[0,1]$ tales que $k(0)=k(1)=0$ y k' está (esencialmente) acotada. Puesto que $F^{-1}(t)=G^{-1}(h_\alpha(t))$, se tiene de forma fácil a partir de (5.43) y (5.44) que, para $k\in\mathcal{G}$,

$$R_n(k)\to R(k):=\int_0^1\left(\frac{B(t)-k(t)/h'_\alpha(t)}{f(F^{-1}(t))}\right)^2dt$$

y por lo tanto $\text{mín}_{k\in\mathcal{G}_{\alpha,n}}R_n(h)\to 0$ (con lo que $n\mathcal{W}_2^2(P_{n,\alpha},Q)\to 0$) será cierto si probamos que $\text{ínf}_{k\in\mathcal{G}}R(k)=0$. Pero esto se puede comprobar fácilmente si tenemos en cuenta, por ejemplo, que eligiendo k_n como la función que interpola $B(t)h'_\alpha(t)$ en los nodos i/n, $i=0,\ldots,n$ y es lineal entre ellos, tendremos que $k_n\in\mathcal{G}$ y $R(k_n)\to 0$. Esto completa la demostración de (5.45).

■

5.6. Bootstrap.

En esta sección se incluyen los resultados teóricos que justifican el proceso de bootstrap que vamos a aplicar en el Capítulo 6.

Es conveniente resaltar que los resultados teóricos que aquí se obtienen son ciertamente restrictivos, y aunque tenemos evidencias empíricas que apuntan a que los mismos son válidos en situaciones más generales (por ejemplo, para tamaños muestrales diferentes), su extensión

requeriría un desarrollo que excede los objetivos de esta sección, que no son más que dar una justificación teórica de la metodología que se propone.

La idea general que está detrás de la metodología que empezamos a desarrollar en esta sección es que cuando dos muestras obtenidas de la misma distribución se recortan para hacerlas más similares, se produce un sobreajuste, en cierta forma se elimina parte de su aleatoriedad, y esto hace que estas muestras recortadas puedan distinguirse del resto de muestras de la misma distribución.

A continuación se describe de forma general el proceso de bootstrap que se sigue en cada paso.

Partimos de dos muestras de datos $\{x_1, x_2, \ldots, x_n\}$ e $\{y_1, y_2, \ldots, y_n\}$ obtenidas de dos realizaciones, $\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}$ e $\{Y_1(\omega), Y_2(\omega), \ldots, Y_n(\omega)\}$, de variables aleatorias i.i.d. tales que $\mathcal{L}(X_i) = P$ y $\mathcal{L}(Y_i) = Q$, $i = 1, \ldots, n$ y donde $P, Q \in \mathcal{P}_2(\mathbb{R})$. Como hasta ahora, estas variables aleatorias están definidas en el espacio (Ω, σ, ν). En lo que sigue haremos aparente la dependencia de $\omega \in \Omega$ para facilitar la explicación del proceso de bootstrap y ser precisos a la hora de obtener los resultados que lo justifican. Sean P_n^ω y Q_n^ω las respectivas medidas empíricas; y, $P_{n,\alpha}^\omega$ y $Q_{n,\alpha}^\omega$ los correspondientes recortes imparciales, i.e.,

$$\mathcal{W}_2(P_{n,\alpha}^\omega, Q_{n,\alpha}^\omega) = \min_{R \in \mathcal{R}_\alpha(P_n^\omega), S \in \mathcal{R}_\alpha(Q_n^\omega)} \mathcal{W}_2(R, S).$$

Paralelamente, P_α y Q_α serán tales que

$$\mathcal{W}_2(P_\alpha, Q_\alpha) = \min_{R \in \mathcal{R}_\alpha(P), S \in \mathcal{R}_\alpha(Q)} \mathcal{W}_2(R, S).$$

De forma general $P_{n,\alpha}^\omega$ es una medida sobre los puntos $\{X_1(\omega), \ldots, X_n(\omega)\}$ con pesos $\{w_1^X(\omega), \ldots, w_n^X(\omega)\}$. Y de forma similar para $Q_{n,\alpha}^\omega$. Puesto que los tamaños de las dos muestras coinciden, si $n(1-\alpha)$ es entero, los pesos $w_i^X(\omega)$ (resp. $w_j^Y(\omega)$) son 0 ó $\frac{1}{n(1-\alpha)}$ (ver Teorema 3.41), y por tanto $P_{n,\alpha}^\omega$ (resp. $Q_{n,\alpha}^\omega$) es una distribución concentrada sobre una submuestra de $\{X_1(\omega), \ldots, X_n(\omega)\}$ (resp. $\{Y_1(\omega), \ldots, Y_n(\omega)\}$). (Cuando $n(1-\alpha)$ no es entero lo que ocurre es que hay un punto en cada una de las dos muestras que se recorta parcialmente).

El siguiente paso consiste en juntar las dos muestras recortadas en una sola, es decir, formar $\{Z_1(\omega), Z_2(\omega), \ldots, Z_{2n}(\omega)\}$, donde

$$Z_i(\omega) = \begin{cases} X_i(\omega) & \text{si } 1 \leq i \leq n, \\ Y_{i-n}(\omega) & \text{si } n+1 \leq i \leq 2n. \end{cases}$$

Llamemos ahora $R^{\omega}_{n,\alpha} = \frac{1}{2}P^{\omega}_{n,\alpha} + \frac{1}{2}Q^{\omega}_{n,\alpha}$ a la medida asociada a la nueva muestra. Definimos los pesos,

$$w^Z_i(\omega) = \begin{cases} \frac{1}{2}w^X_i(\omega) & \text{si } 1 \leq i \leq n, \\ \frac{1}{2}w^Y_{i-n}(\omega) & \text{si } n+1 \leq i \leq 2n. \end{cases}$$

A continuación extraemos dos muestras bootstrap de tamaño n de la nueva muestra con el esquema de remuestreo dado por los pesos $\left\{w^Z_i(\omega)\right\}_i$, es decir, obtenemos dos nuevas muestras de datos $\left\{x^*_{n,1}, x^*_{n,2}, \ldots, x^*_{n,n}\right\}$ e $\left\{y^*_{n,1}, y^*_{n,2}, \ldots, y^*_{n,n}\right\}$. Estos datos pueden ser considerados como la realización de unas nuevas variables aleatorias i.i.d, $X^*_{n,1}, X^*_{n,2}, \ldots, X^*_{n,n}$ e $Y^*_{n,1}, Y^*_{n,2}, \ldots, Y^*_{n,n}$, cuya distribución condicional, dado que $X_1 = x_1, \ldots, X_n = x_n, Y_1 = y_1, \ldots, Y_n = y_n$, es $R^{\omega}_{n,\alpha}$. Podemos considerar que las muestras bootstrap se obtienen tras sortear un $\upsilon \in \Upsilon$ (en un nuevo espacio (Υ, γ)), de acuerdo con una probabilidad de transición $\eta(\omega, \cdot)$ para cada $\omega \in \Omega$. Así pues, consideraremos como espacio de referencia el espacio producto $(\Omega \times \Upsilon, \sigma \otimes \gamma)$ con la medida de probabilidad habitual, que llamaremos ρ, construida a partir de ν y las probabilidades de transición $\eta(\omega, \cdot)$. Llamemos ${P^*_n}^{\omega,\upsilon}$ y ${Q^*_n}^{\omega,\upsilon}$ a las respectivas medidas empíricas asociadas a las muestras bootstrap.

La base de nuestras inferencias radicará en la comparación entre $\mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$ y $\mathcal{W}_2(P^*_n, Q^*_n)$. De forma un tanto imprecisa, la idea intuitiva que pretendemos explotar es que si recortando una proporción α no conseguimos "igualar" las dos distribuciones, es decir, si $\mathcal{W}_2(P_\alpha, Q_\alpha) > 0$, entonces, por una parte P^*_n y Q^*_n vendrán de muestras de una misma distribución a la que se aproximan, y por otra $P_{n,\alpha}$ y $Q_{n,\alpha}$ lo serán de distribuciones distintas (en realidad, recortes de distribuciones diferentes), con lo que en general habremos de tener que $\mathcal{W}_2(P^*_n, Q^*_n) < \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$ para tamaños de n suficientemente grandes. En cambio, si recortando conseguimos "igualar" ambas distribuciones, i.e., $\mathcal{W}_2(P_\alpha, Q_\alpha) = 0$, entonces ambos pares de distribuciones se aproximarán a la misma distribución común, con la diferencia de que $P_{n,\alpha}$ y $Q_{n,\alpha}$ se han elegido para que estén "especialmente" próximas, por lo que en general $\mathcal{W}_2(P^*_n, Q^*_n) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})$.

Esta idea se refleja en el resultado del Teorema 5.31, para cuya demostración se requieren algunos resultados previos. Utilizando la misma notación que hasta ahora, si llamamos $F^*_n(\omega, \upsilon, t)$ a la función de distribución asociada a la medida ${P^*_n}^{\omega,\upsilon}$ y $G_{n,\alpha}(\omega, t)$ a la función de distribución de la medida $R^{\omega}_{n,\alpha}$, se tiene,

Proposición 5.27. *En las anteriores condiciones,*

$$\lim_{n\to\infty} \mathcal{W}_\infty(P_n^{*\,\omega,\upsilon}, R_{n,\alpha}^{\omega}) = \lim_{n\to\infty} \sup_{t\in\mathbb{R}} |F_n^*(\omega,\upsilon,t) - G_{n,\alpha}(\omega,t)| = 0 \qquad \rho\text{-casi seguro.}$$

Demostración. Sea $\varepsilon > 0$. Condicionalmente a la sucesión $\{Z_i\}_{i=1}^{\infty}$ las variables aleatorias $V_i := I_{(-\infty,t]}(X_{n,i}^*) - \sum_{j=1}^{2n} w_j^Z I_{(-\infty,t]}(Z_j)$ son independientes, centradas, y se tiene que $\mathrm{Var}(V_j) \leq 1$. Así pues, aplicando la desigualdad de Bernstein (ver Lema 2.19) a la probabilidad condicionada, tenemos,

$$\begin{aligned}
&\rho\left(|F_n^*(\cdot,\cdot,t) - G_{n,\alpha}(\cdot,t)| > \varepsilon / \{Z_i\}_{i=1}^{\infty}\right) \\
&= \rho\left(\left|\sum_{i=1}^{n} I_{(-\infty,t]}(X_{n,i}^*) - n\sum_{i=1}^{2n} w_i^Z I_{(-\infty,t]}(Z_i)\right| > \varepsilon n \,\middle/\, \{Z_i\}_{i=1}^{\infty}\right) \\
&= \rho\left(\left|\sum_{i=1}^{n}\left[I_{(-\infty,t]}(X_{n,i}^*) - \sum_{i=1}^{2n} w_i^Z I_{(-\infty,t]}(Z_i)\right]\right| > \varepsilon n \,\middle/\, \{Z_i\}_{i=1}^{\infty}\right) \\
&\leq 2\exp\left(\frac{-\frac{1}{2}\varepsilon^2 n^2}{n + \frac{1}{3}\varepsilon n}\right) = 2\exp\left(-H\right)^n,
\end{aligned}$$

donde $H = \frac{-\frac{1}{2}\varepsilon^2}{1+\frac{1}{3}\varepsilon} > 0$.

Como las dos funciones de distribución son funciones escalonadas crecientes, y dada la sucesión $\{Z_i\}_{i=1}^{\infty}$, tienen los saltos en los mismos $2n$ puntos, se tiene que

$$\rho\left(\sup_{t\in\mathbb{R}} |F_n^*(\cdot,\cdot,t) - G_{n,\alpha}(\cdot,t)| > \varepsilon \,\middle/\, \{Z_i\}_{i=1}^{\infty}\right) \leq 2(2n)\exp\left(-H\right)^n$$

Ahora, como la cota anterior es independiente de la sucesión, integrando, tenemos que

$$\rho\left(\sup_{t\in\mathbb{R}} |F_n^*(\cdot,\cdot,t) - G_{n,\alpha}(\cdot,t)| > \varepsilon\right) \leq 4n\exp\left(-H\right)^n,$$

y como $\exp(-H) < 1$ la serie de la derecha es convergente. Entonces la serie del lado izquierdo es convergente, y como consecuencia del lema de Borel-Cantelli, tenemos que para cada $\varepsilon > 0$,

$$\rho\left(\limsup\left\{\sup_{t\in\mathbb{R}} |F_n^*(\cdot,\cdot,t) - G_{n,\alpha}(\cdot,t)| > \varepsilon\right\}\right) = 0.$$

De lo que se deduce el resultado. ∎

Proposición 5.28. *Supongamos que $P, Q \in \mathcal{P}_2(\mathbb{R})$, tienen soporte en un compacto y además se tiene que $\mathcal{W}_2(P_\alpha, Q_\alpha) > 0$. Entonces,*

$$\mathcal{W}_2(P_n^{*\,\omega,\upsilon}, Q_n^{*\,\omega,\upsilon}) \to 0 \qquad \rho\text{-casi seguro.}$$

Demostración. Aplicando la desigualdad triangular tenemos que,

$$\mathcal{W}_2(P_n^{*\omega,\upsilon}, Q_n^{*\omega,\upsilon}) \leq \mathcal{W}_2(P_n^{*\omega,\upsilon}, R_{n,\alpha}^{\omega}) + \mathcal{W}_2(Q_n^{*\omega,\upsilon}, R_{n,\alpha}^{\omega}).$$

Así pues, basta con probar que $\mathcal{W}_2(P_n^{*\omega,\upsilon}, R_{n,\alpha}^{\omega}) \to 0$ (porque la misma demostración sirve para el otro término). Para ello, tenemos por un lado que x^2 es uniformemente integrable respecto de $\left\{P_n^{*\omega,\upsilon}\right\}_n$ y respecto de $\left\{R_{n,\alpha}^{\omega}\right\}_n$. La integrabilidad uniforme respecto de $\left\{R_{n,\alpha}^{\omega}\right\}_n$ se sigue de la acotación,

$$\begin{aligned}\int_{\{|x|>a\}} x^2 dR_{n,\alpha}^{\omega} &= \frac{1}{2}\left(\int_{\{|x|>a\}} x^2 dP_{n,\alpha}^{\omega} + \int_{\{|x|>a\}} x^2 dQ_{n,\alpha}^{\omega}\right) \\ &\leq \frac{2}{1-\alpha}\left(\int_{\{|x|>a\}} x^2 dP_n^{\omega} + \int_{\{|x|>a\}} x^2 dQ_n^{\omega}\right),\end{aligned}$$

y de la integrabilidad uniforme de x^2 respecto de $\{P_n^{\omega}\}_n$ y $\{Q_n^{\omega}\}_n$ (consecuencia a su vez de la Ley de los Grandes Números). La integrabilidad uniforme respecto de $\left\{P_n^{*\omega,\upsilon}\right\}_n$ se sigue del hecho de que P tiene soporte en un compacto.

Por otra parte, se verifican las condiciones para aplicar el Teorema 3.29 que asegura la unicidad del par (P_α, Q_α), y la parte (d) del Teorema 5.6 que en particular proporciona la consistencia débil ν-casi seguro de $\left\{P_{n,\alpha}^{\omega}\right\}_n$ y $\left\{Q_{n,\alpha}^{\omega}\right\}_n$ a P_α y Q_α respectivamente. Como consecuencia tenemos que

$$R_{n,\alpha}^{\omega} = \frac{P_{n,\alpha}^{\omega} + Q_{n,\alpha}^{\omega}}{2} \to_w R_\alpha := \frac{P_\alpha + Q_\alpha}{2} \quad \nu\text{-casi seguro} \tag{5.50}$$

Utilizando ahora la Proposición 5.27 y (5.50) concluimos que

$$P_n^{*\omega,\upsilon} \to_w R_\alpha \qquad \rho\text{-casi seguro} \tag{5.51}$$

Finalmente, combinando la integrabilidad uniforme de x^2 respecto de $\left\{P_n^{*\omega,\upsilon}\right\}_n$ y respecto de $\left\{R_{n,\alpha}^{\omega}\right\}_n$ con (5.50) y (5.51), tenemos que $\mathcal{W}_2(P_n^{*\omega,\upsilon}, R_{n,\alpha}^{\omega}) \to 0 \quad \rho$-casi seguro. Lo que concluye la demostración. ∎

Nota 5.29. Obviamente, la integrabilidad uniforme de x^2 respecto a $\left\{R_{n,\alpha}^{\omega}\right\}_n$ se sigue del hecho de que es un recorte de una combinación lineal de las medidas empíricas basadas en dos leyes, P y Q, que tienen soporte en un compacto. La razón por la que en la prueba anterior se ha abordado de forma diferente es meramente ilustrativa. De esta forma queda claro cuál es el principal escollo para suavizar la hipótesis relativa a los soportes de P y Q.

El resultado que se da a continuación es una generalización del apartado (a) del Teorema 2.1 de Bickel y Freedman (1981) en el que se prueba que el bootstrap para la media funciona cuando se muestrea de una empírica. En nuestro caso, muestrearemos (ver Teorema 5.31) de $R^{\omega}_{n,\alpha}$ que no es exactamente una empírica. La demostración sigue un esquema similar a la de la demostración del resultado de Bickel y Freedman (1981).

Proposición 5.30. *Sean $V_1, V_2 \ldots, V_n$ variables aleatorias independientes tales que $\mathcal{L}(V_i) = P_n$. Supongamos que $P_n \in \mathcal{P}_2(\mathbb{R})$ y que $\mathcal{W}_2(P_n, P) \to 0$, donde $P \in \mathcal{P}_2$. Entonces,*

$$\sqrt{n}(\bar{V}_n - \mu_n) \to_w N(0, \sigma^2),$$

donde $\bar{V}_n = \frac{1}{n}\sum_{i=1}^n V_i$, $\mu_n = \mathrm{E}(\bar{V}_n)$ y $\sigma^2 = \mathrm{Var}(P)$.

Demostración. Por una parte podemos escribir $S_n^V = \sqrt{n}(\bar{V}_n - \mu_n) = n^{-1/2}\sum_{i=1}^n (V_i - E(V_i))$. Sean ahora $U_1, U_2, \ldots, U_n$ variables aleatorias independientes tales que $\mathcal{L}(U_i) = P$. De igual forma que antes podemos escribir $S_n^U = \sqrt{n}(\bar{U}_n - E(U_1)) = n^{-1/2}\sum_{i=1}^n (U_i - E(U_i))$, y aplicando el Teorema Central del Límite usual ($P \in \mathcal{P}_2(\mathbb{R})$), tenemos que $S_n^U \to_w N(0, \sigma^2)$.

Por otra parte, aplicando sucesivamente los Lemas 2.4, 2.5 y 2.6, tenemos

$$\begin{aligned} \mathcal{W}_2^2(S_n^V, S_n^U) &\leq n^{-1}\sum_{i=1}^n \mathcal{W}_2^2(V_i - E(V_i), U_i - E(U_i)) = \mathcal{W}_2^2(V_1 - E(V_1), U_1 - E(U_1)) \\ &\leq \mathcal{W}_2^2(V_1, U_1) = \mathcal{W}_2^2(P_n, P). \end{aligned}$$

Finalmente, el resultado se sigue del hecho de que $\mathcal{W}_2(P_n, P) \to 0$.

■

Usando la misma notación que hasta ahora, tenemos el siguiente resultado

Teorema 5.31. *Sea $\alpha > 0$ y $P, Q \in \mathcal{P}_2(\mathbb{R})$. Supongamos que P y Q tienen funciones de densidad f y g, que tienen soporte en $[a, b]$, son continuas y estrictamente positivas,*

(a) *Si $\mathcal{W}_2(P_{\alpha'}, Q_{\alpha'}) = 0$ para algún $\alpha' \in (0, \alpha)$, entonces*

$$\rho\left(\mathcal{W}_2(P_n^*, Q_n^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\right) \to 1.$$

(b) *Si $\mathcal{W}_2(P_\alpha, Q_\alpha) > 0$, entonces $\rho\left(\mathcal{W}_2(P_n^*, Q_n^*) > \mathcal{W}_2(P_{n,\alpha}, Q_{n,\alpha})\right) \to 0$.*

Demostración.

Parte (a). Por un lado, en virtud del Teorema 3.30, si se tiene que la distancia recortada es cero sabemos que existe $R_0 \in \mathcal{R}_{\alpha_0}(P) \cap \mathcal{R}_{\alpha_0}(Q) \subset \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$, cuya función de densidad está dada por (3.30). Esto asegura que dicha función de densidad, que denotaremos por m_0, es continua y estrictamente positiva si lo son f y g. Además, puesto que el conjunto de las funciones con derivada continua es denso (en L_∞) en el conjunto de las funciones continuas y $\alpha' < \alpha$, podemos garantizar que existe una función de densidad m_1, próxima a m_0, que tiene derivada continua y corresponde a una medida de probabilidad $R_1 \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$. En estas condiciones tendremos que $n\mathcal{W}_2^2(P_{n,\alpha}^\omega, Q_{n,\alpha}^\omega) \to 0$ en ν-probabilidad. Para ello no hay más que utilizar la desigualdad triangular de la distancia $\mathcal{W}_2$ y aplicar el Teorema 5.26 a los dos últimos sumandos de la expresión

$$\begin{aligned} n\mathcal{W}_2^2(P_{n,\alpha}^\omega, Q_{n,\alpha}^\omega) &\leq n\mathcal{W}_2^2(\tilde{P}_{n,\alpha}^\omega, \tilde{Q}_{n,\alpha}^\omega) \\ &\leq 2\left(n\mathcal{W}_2^2(\tilde{P}_{n,\alpha}^\omega, R_1) + n\mathcal{W}_2^2(\tilde{Q}_{n,\alpha}^\omega, R_1)\right) \to 0 \text{ en } \nu\text{-probabilidad}, \quad (5.52) \end{aligned}$$

donde $\tilde{P}_{n,\alpha}^\omega = \arg\min_{R \in \mathcal{R}_\alpha(P_n^\omega)} \mathcal{W}_2(R, R_1)$ y $\tilde{Q}_{n,\alpha}^\omega = \arg\min_{R \in \mathcal{R}_\alpha(Q_n^\omega)} \mathcal{W}_2(R, R_1)$.

Por otro lado, utilizando el Lema 2.6, tenemos que

$$\begin{aligned} n\mathcal{W}_2^2(P_n^{*\omega,\cdot}, Q_n^{*\omega,\cdot}) &\geq n\left(\bar{X}_n^*(\cdot,\omega) - \bar{Y}_n^*(\cdot,\omega)\right)^2 \\ &= \left(\sqrt{n}(\bar{X}_n^*(\cdot,\omega) - \mu_n(\omega)) - \sqrt{n}(\bar{Y}_n^*(\cdot,\omega) - \mu_n(\omega))\right)^2, \quad (5.53) \end{aligned}$$

donde $\mu_n(\omega) = \mathrm{E}\bar{X}_n^*(\cdot,\omega)$ y $\bar{X}_n^*(\cdot,\omega) = \frac{1}{n}\sum_{i=1}^n X_{n,i}^*(\cdot,\omega)$ (de forma equivalente $\bar{Y}_n^*(\cdot,\omega)$).

Sea ahora $\Gamma \subset \Omega$ el conjunto de probabilidad 1 de $\omega' s$ en el que $P_n^\omega \to P$ y $Q_n^\omega \to Q$. Fijando $\omega \in \Gamma$, es claro que de cada subsucesión de $\left\{P_{n,\alpha}^\omega\right\}_n$ (resp. $\left\{Q_{n,\alpha}^\omega\right\}_n$), como es *tight*, podemos extraer una subsucesión convergente. Sean $\left\{P_{n_k,\alpha}^\omega\right\}_{n_k}$ y $\left\{Q_{n_k,\alpha}^\omega\right\}_{n_k}$ dos de esas subsucesiones convergentes. Tendremos entonces que,

$$\begin{aligned} P_{n_k,\alpha}^\omega &\to_w P_\alpha^\omega \\ Q_{n_k,\alpha}^\omega &\to_w Q_\alpha^\omega, \end{aligned}$$

donde, en virtud del Teorema 3.7, tenemos que $P_\alpha^\omega \in \mathcal{R}_\alpha(P)$ y $Q_\alpha^\omega \in \mathcal{R}_\alpha(Q)$. En consecuencia, $R_{n_k,\alpha}^\omega \to_w R_\alpha^\omega = \frac{P_\alpha^\omega + Q_\alpha^\omega}{2} \in \mathcal{R}_\alpha(P) \cap \mathcal{R}_\alpha(Q)$. Esta convergencia, junto con la integrabilidad uniforme de x^2 respecto de $\left\{R_{n_k,\alpha}^\omega\right\}_{n_k}$, proporciona la convergencia $\mathcal{W}_2(R_{n_k,\alpha}^\omega, R_\alpha^\omega) \to 0$. Ahora, utilizando la Proposición 5.30 tenemos que $\sqrt{n_k}\left(\bar{X}_{n_k}^*(\cdot,\omega) - \mu_{n_k}(\omega)\right) \to_w N(0, \sigma^2(\omega))$

y $\sqrt{n_k}\left(\bar{Y}^*_{n_k}(\cdot,\omega)-\mu_{n_k}(\omega)\right) \to_w N(0,\sigma^2(\omega))$, donde $\sigma^2(\omega)=\text{Var}(R^\omega_\alpha)$. Y, como condicionalmente a $\{X_1(\omega),\ldots,X_n(\omega)\}$ y a $\{Y_1(\omega),\ldots,Y_n(\omega)\}$ (i.e., a ω), $\bar{X}^*_n(\cdot,\omega)$ y $\bar{Y}^*_n(\cdot,\omega)$ son independientes, entonces

$$n\left(\bar{X}^*_{n_k}(\cdot,\omega)-\bar{Y}^*_{n_k}(\cdot,\omega)\right)^2 \to_{\mathcal{L}} 4\sigma^2(\omega)Z,$$

donde $\mathcal{L}(Z)=\chi^2_1$. Por otra parte, $\mathcal{R}_\alpha(P)\cap\mathcal{R}_\alpha(Q)$ es compacto y como $P,Q\ll\ell$ no puede existir un recorte degenerado con lo que,

$$\min_{R^\omega_\alpha\in\mathcal{R}_\alpha(P)\cap\mathcal{R}_\alpha(Q)} \text{Var}\,(R^\omega_\alpha)>0,$$

y por tanto $\exists c>0$ tal que, $4\sigma^2(\omega)Z\geq cZ$ para todo ω. Ahora, para cada $\varepsilon>0$ puedo encontrar un $\delta>0$ tal que $\rho(cZ<\delta)<\frac{\varepsilon}{2}$ por una parte. De lo que, utilizando (5.53), tenemos que

$$1-\frac{\varepsilon}{2}\leq\rho\left(cZ>\delta\right)\leq\liminf\rho\left(n(\bar{X}^*_n-\bar{Y}^*_n)^2>\delta\right)\leq\liminf\rho\left(n\mathcal{W}_2^2(P^*_n,Q^*_n)>\delta\right). \quad (5.54)$$

Por otra parte, en virtud de (5.52), para dichos $\varepsilon,\delta>0$, existe n_0 tal que para todo $n\geq n_0$ se tiene que $\rho\left(n\mathcal{W}_2^2(P_{n,\alpha},Q_{n,\alpha})>\delta\right)<\frac{\varepsilon}{2}$. Combinando este resultado con (5.54) tenemos

$$1-\varepsilon\leq\liminf\rho\left(n\mathcal{W}_2^2(P^*_n,Q^*_n)>n\mathcal{W}_2^2(P_{n,\alpha},Q_{n,\alpha})\right),$$

para cada $\varepsilon>0$ y por tanto el resultado.

Parte (b)

Con lo visto hasta aquí es inmediato. Por una parte tenemos que se verifican las condiciones del Teorema 5.6 y por ello

$$\mathcal{W}_2(P^\omega_{n,\alpha},Q^\omega_{n,\alpha})\to\mathcal{W}_2(P_\alpha,Q_\alpha)>0 \qquad \nu\text{-casi seguro},$$

pero por otra parte, aplicando la Proposición 5.28, tenemos que

$$\mathcal{W}_2(P^*_n,Q^*_n)\to 0 \qquad \rho\text{-casi seguro}.$$

■

Capítulo 6

Otras aplicaciones.

En este Capítulo se presentan brevemente algunas aplicaciones derivadas de la metodología bootstrap descrita en la Sección 5.6. En la primera sección se ilustra el funcionamiento de los tests bootstrap que se pueden realizar a partir del Teorema 5.31. En la Sección 6.1 utilizamos esta metodología para comparar una muestra con una distribución de referencia y para comparar dos muestras. En cambio, en la Sección 6.2, se emplea esta metodología para buscar el patrón o núcleo común a varias distribuciones.

6.1. Tests bootstrap.

A partir del Teorema 5.31 podemos pensar en diseñar un test para contrastar que dos distribuciones P y Q son similares a nivel $\alpha \in (0,1)$, o más concretamente, para contrastar la hipótesis $H_0 : \exists \alpha' \in (0,\alpha)$ tal que $\mathcal{W}_2(P_{\alpha'}, Q_{\alpha'}) = 0$ vs $H_a : \mathcal{W}_2(P_\alpha, Q_\alpha) > 0$. Si bien el Teorema 5.31 cubre específicamente el caso de dos muestras, de forma completamente similar se puede probar el resultado equivalente para el caso de una muestra. A continuación incluimos tres pequeños estudios de simulación en los que se compara una muestra con una distribución de referencia, y dos muestras entre sí.

6.1.1. Problemas de una muestra.

El primer estudio consiste en comparar una muestra de una normal estándar con diversos grados de contaminación con una N(0,1) poblacional (i.e., un problema de una muestra). Para ello se toman n_1 puntos de una N(0,1) que se contaminan con n_2 puntos de una N(d, s).

Con esto se tiene una muestra de n_1+n_2 puntos que se compara contra la N(0,1) poblacional (que no se recorta). De la muestra recortada más próxima a la N(0,1) se obtienen réplicas bootstrap, calculándose la frecuencia de veces que la distancia $\mathcal{W}_2$ entre la muestra bootstrap y la N(0,1) es mayor que la distancia entre la recortada inicial y la N(0,1). Esa frecuencia, que en adelante llamaremos p-valor, se ha obtenido fijando $n_1 = 100$, para varios tamaños de recorte (1 %, 5 %, 10 %, 20 % y 50 %), varios valores de n_2 (2, 5, 10, 20), y varios valores de s (0.5, 1, 2) y d (2, 5, 10). En cada caso se obtiene el p-valor basado en 1000 réplicas bootstrap y se repite el proceso 100 veces, calculando al final el p-valor medio y el número medio de puntos acertados al recortar (número de puntos de la distribución contaminante que son recortados). Los resultados aparecen en la Tabla 6.1, donde en cada celda aparecen, en la primera fila el número medio de puntos acertados para $\alpha = 0.01, 0.05, 0.10, 0.20$ y 0.50 respectivamente. Y justo debajo el p-valor medio correspondiente.

Antes de comentar los resultados concretos obtenidos en esta simulación es importante realizar la siguiente observación de carácter general. En los resultados obtenidos en el Teorema 5.31 se da el comportamiento asintótico de los p-valores arriba mencionados. Sin embargo, en la aplicación práctica de esta metodología, el tamaño muestral estará fijo. De esta forma, en la medida en que un p-valor correspondiente a un tamaño de recorte esté próximo a 0 ó a 1 podremos decir que para ese α la distancia entre los recortes poblacionales es mayor que cero o no. Mientras que, obviamente, si el p-valor no está de forma clara próximo a 0 ó 1, lo prudente será afirmar que no podemos decir qué pasa.

De forma general observamos que cuando la distribución contaminante está alejada ($d = 5, 10$) y el tamaño de recorte es igual o superior a la fracción de contaminación, el procedimiento acierta siempre con los puntos que no pertenecen a la $N(0,1)$ y el p-valor es prácticamente siempre igual a 1. En esta misma situación si el tamaño de recorte no llega a la fracción de contaminación, el número medio de puntos acertados es el máximo que puede ser mientras que el p-valor es próximo a 0, indicando que la muestra recortada no puede provenir de la $N(0,1)$. Cuando $d = 2$, la distribución contaminante está más cerca de la $N(0,1)$ y acertar con los puntos que provienen exactamente de la contaminación es más difícil. Sin embargo, cuando el tamaño de recorte es superior a la fracción de contaminación el p-valor siempre es próximo a 1. Incluso cuando el tamaño de recorte es algo menor que la contaminación, si hay bastante superposición (por ejemplo cuando $s = 2$), se alcanzan

$n_1 = 100$		d																	
n_2	s	2					5					10							
		1 %	5 %	10 %	20 %	50 %	1 %	5 %	10 %	20 %	50 %	1 %	5 %	10 %	20 %	50 %			
2	0.5	0.23	0.52	0.60	0.75	1.22	1.01	2	2	2	2	1.02	2	2	2	2			
		0.840	0.947	0.984	0.999	0.999	0.483	0.872	0.949	0.987	0.998	0.406	0.881	0.947	0.983	0.998			
	1	0.40	0.67	0.79	0.93	1.38	1.01	1.99	1.99	2	2	1.02	2	2	2	2			
		0.781	0.920	0.969	0.993	0.999	0.517	0.907	0.963	0.988	0.999	0.406	0.860	0.927	0.976	0.998			
	2	0.59	0.85	0.92	1.01	1.34	0.97	1.79	1.81	1.84	1.90	1.02	2	2	2	2			
		0.785	0.919	0.967	0.990	0.999	0.582	0.893	0.956	0.988	0.998	0.405	0.885	0.963	0.992	0.999			
5	0.5	0.40	1.61	2.07	2.48	3.44	1.05	5	5	5	5	1.05	5	5	5	5			
		0.664	0.897	0.966	0.989	0.999	0.073	0.786	0.937	0.987	0.999	0.056	0.791	0.937	0.984	0.999			
	1	0.69	1.66	1.95	2.28	3.27	1.05	4.88	4.93	4.97	4.99	1.05	5	5	5	5			
		0.710	0.919	0.976	0.995	0.999	0.084	0.792	0.938	0.981	0.996	0.052	0.765	0.918	0.976	0.997			
	2	0.95	2.16	2.40	2.74	3.58	1.04	4.25	4.35	4.40	4.65	1.05	5	5	5	5			
		0.559	0.889	0.964	0.990	0.999	0.152	0.834	0.935	0.983	0.997	0.054	0.733	0.905	0.981	0.999			
10	0.5	0.76	3.46	5.02	5.88	7.52	1.10	5.50	9.99	10	10	1.10	5.50	10	10	10			
		0.331	0.661	0.887	0.978	0.999	0.006	0.053	0.821	0.963	0.998	0.004	0.036	0.834	0.974	0.995			
	1	1.04	3.76	4.83	5.52	7.25	1.10	5.50	9.90	9.96	9.98	1.10	5.50	10	10	10			
		0.334	0.723	0.892	0.981	0.999	0.006	0.067	0.835	0.989	0.999	0.004	0.037	0.840	0.980	0.999			
	2	1.07	3.93	4.76	5.39	7.16	1.10	5.50	8.79	9.01	9.40	1.10	5.50	10	10	10			
		0.296	0.826	0.957	0.989	0.999	0.015	0.200	0.898	0.984	0.999	0.005	0.041	0.822	0.971	0.997			
20	0.5	0.99	4.99	9.36	13.56	16.16	1.20	6	12	20	20	1.20	6	12	20	20			
		0.036	0.164	0.516	0.952	0.998	0.000	0.001	0.009	0.912	0.996	0.000	0.000	0.006	0.890	0.998			
	1	1.18	5.41	9.07	11.63	14.89	1.20	6	12	19.83	19.92	1.20	6	12	20	20			
		0.077	0.347	0.724	0.976	0.999	0.000	0.001	0.011	0.891	0.998	0.000	0.000	0.006	0.901	0.997			
	2	1.20	5.77	9.39	11.22	14.55	1.20	6	12	18.14	18.88	1.20	6	12	20	20			
		0.049	0.397	0.830	0.980	0.999	0.000	0.002	0.051	0.936	0.999	0.000	0.000	0.005	0.867	0.997			

Tabla 6.1: Número medio de puntos de la distribución contaminante que son recortados y p-valor medio bootstrap.

valores del p-valor por encima de 0.8 (ver por ejemplo, $n_2 = 10$ y $n_2 = 20$ con $\alpha = 0.05$ y $\alpha = 0.1$, que muestran 0.826 y 0.830 respectivamente).

6.1.2. Problemas de dos muestras.

En esta subsección se muestra el resultado de 2 pequeños estudios de simulación. En el primero de ellos se comparan muestras de 5 poblaciones normales (A $\sim$ N(0,1), B $\sim$ N(0,1), C $\sim$ N(1,1), D $\sim$ N(1,2) y E $\sim$ N(2,1)) para dos tamaños muestrales ($n = 30, 100$). En este estudio, se obtiene en primer lugar una muestra de cada una de las distribuciones. Para algunas de las posibles parejas de muestras, y niveles de recorte que van desde 0 % (incluido como referencia) hasta 30 %, se calculan las parejas de muestras recortadas más próximas y, una vez obtenidas dichas parejas, se toman 1000 parejas de muestras bootstrap siguiendo la metodología descrita en la Sección 5.6. A continuación se calcula la frecuencia de veces que la distancia $\mathcal{W}_2$ entre la pareja de muestras bootstrap es mayor que la distancia entre la pareja de recortadas. Los resultados aparecen en las tablas 6.2 y 6.3.

	0 %	1 %	5 %	10 %	20 %	30 %
A vs B	0.669	0.706	0.855	0.968	0.996	1.000
A vs C	0.002	0.000	0.004	0.011	0.146	0.897
A vs D	0.002	0.005	0.013	0.032	0.164	0.683
A vs E	0.000	0.000	0.000	0.000	0.000	0.000
C vs D	0.006	0.016	0.018	0.033	0.163	0.433

Tabla 6.2: p-valores cuando $n = 30$.

	0 %	1 %	5 %	10 %	20 %	30 %
A vs B	0.964	0.974	0.998	1.000	1.000	1.000
A vs C	0.000	0.000	0.000	0.000	0.004	0.371
A vs D	0.000	0.000	0.000	0.000	0.091	1.000
A vs E	0.000	0.000	0.000	0.000	0.000	0.000
C vs D	0.000	0.000	0.000	0.001	0.134	0.905

Tabla 6.3: p-valores cuando $n = 100$.

En la primera fila de las tablas 6.2 y 6.3 tenemos dos muestras de la misma distribución. El resultado es claro, los p-valores son altos en general desde tamaños de recorte bajos. El sobreajuste al que hacíamos referencia en la sección anterior se muestra también al aumentar

el tamaño de recorte. Cuanto mayor es el valor de α menor parece el valor de n para que el p-valor esté próximo a 1.

En la segunda comparación, si exceptuamos el caso en el que $\alpha = 0.3$, los p-valores son bajos, lo que sugiere que la distancia entre los correspondientes recortes poblacionales no es cero. Cuando $\alpha = 0.3$ y $n = 30$, el p-valor es en cambio próximo a 1 (p-valor=0.897), lo que podría indicar que si recortamos un 30 % en la N(0,1) y un 30 % en la N(1,1), conseguimos hacerlas iguales. En cambio en la Tabla 6.3, vemos que el correspondiente p-valor ha disminuido, es 0.371. Lo que apuntaría más bien a que la distancia entre los recortes poblacionales no es 0. La conclusión entonces, es que estamos ante una situación en la que no podemos afirmar claramente que la distancia entre los recortes poblacionales sea 0 o no.

Los p-valores observados en la tercera comparación de la Tabla 6.2 son claramente próximos a 0 si $\alpha \leq 10\,\%$. En cambio, si $\alpha = 0.3$, el p-valor está más cerca de 1. El correspondiente p-valor cuando $n = 100$ es de hecho 1. Esto apunta a que el nivel de recorte para igualar la N(0,1) con la N(1,2), es inferior al que se necesita para igualar la N(0,1) a la N(1,1), lo que parece razonable.

Los p-valores correspondientes a la cuarta comparación (N(0,1) vs N(2,1)) indican claramente que ni aún recortando el 30 % en ambas conseguimos igualarlas.

Finalmente, cuando comparamos C vs D (misma localización y diferente varianza), la combinación de los resultados de ambas tablas parece indicar que con valores de α hasta 0.1 no alcanzamos a igualarlas, pero con $\alpha = 0.3$ tal vez podría ser.

Este primer estudio lo complementamos con un segundo en el que introducimos una muestra contaminada. Comparamos así una muestra de tamaño $n = 100$ de una N(0,1) con una muestra que contiene 95 valores de una N(0,1) y 5 de una N(10,0.5). Los p-valores de la Tabla 6.4 muestran que hay que recortar alrededor de un 2-3 % (un poco menos que la fracción de contaminación) para conseguir p-valores no significativos a los niveles habituales.

	0 %	1 %	2 %	3 %	4 %	5 %	6 %	7 %	8 %	9 %	10 %
(1)	0.017	0.031	0.045	0.101	0.137	0.224	0.316	0.457	0.618	0.782	0.902
(2)	0.018	0.022	0.052	0.087	0.165	0.181	0.280	0.308	0.452	0.585	0.742

Tabla 6.4: A $\sim$ N(0,1), B $\sim$ 0.95*N(0,1)+0.5*N(10,0.5). (1): A vs B recortando las dos, (2): A vs B recortando sólo B.

6.2. Búsqueda del núcleo común a varias distribuciones.

En esta parte de la memoria se presenta una metodología que hasta donde llega nuestro conocimiento es nueva en la búsqueda de un patrón común a varias muestras (resp. poblaciones). Cuando hablamos de patrón nos referimos a un núcleo común en las muestras (resp. poblaciones) o dicho de otra forma, a un subgrupo de individuos con un comportamiento homogéneo en lo que respecta a la variable de interés. La diferencia con respecto a la propuesta en secciones anteriores es que ahora no se trata de comparar dos muestras entre sí (o una muestra contra una distribución de referencia), buscamos la comparación o tratamiento simultáneo de más de dos muestras.

En este contexto, es fácil prever que si hay una parte homogénea en todas las muestras, las partes o individuos heterogéneos pueden ser muy dispares de muestra a muestra. Algo que el investigador no puede determinar a priori. Y es aquí donde los recortes imparciales pueden mostrar toda su capacidad para detectar estas zonas dentro de cada muestra, así como las muestras más dispares con respecto al resto.

Dado que la forma que tenemos de comparar muestras es a través del uso de métricas probabilísticas que proporcionan la distancia o grado de similitud entre dos de ellas, no tenemos otro remedio que diseñar un procedimiento en el que llevemos nuestro grupo de muestras a este terreno. Se abren varias posibilidades, una de ellas sería buscar los recortes imparciales de cada muestra que minimizasen una combinación lineal de términos que incluyesen todos los pares posibles de muestras. Algo que conectaría con la programación multiobjetivo y que sería factible desde el punto de vista computacional. Sin embargo, y al igual que ocurre con otras técnicas estadísticas como el análisis discriminante o los métodos de selección de variables en regresión, optaremos por un proceso secuencial. En cada paso de este proceso secuencial cada muestra es comparada con la unión de todas las demás. Esto nos permite identificar por una parte las zonas de cada muestra más heterogéneas respecto del total, y por otra, las muestras que son más diferentes del resto. En cada paso de este proceso secuencial se eliminará del conjunto la muestra más heterogénea respecto del resto, iterando este proceso hasta obtener un conjunto de muestras que salvo en un cierto porcentaje, el de recorte, se puedan considerar homogéneas. Al final de este proceso comprobaremos si alguna de las muestras que en algún paso fue considerada heterogénea y por tanto excluida, puede volver a ser incluida. Es claro, que al igual que otros procesos secuenciales empleados

en Estadística, no podemos asegurar que el conjunto de muestras que obtenemos al final como homogéneas sea el "óptimo" (en un sentido amplio). Por una parte puede haber otros subconjuntos de muestras homogéneas entre sí. Por otra, y dependiendo de como hayamos manejado los niveles de recorte, puede haber subconjuntos de muestras con un nivel más alto de homogeneidad.

A continuación aplicamos estas ideas a un conjunto de datos reales. El conjunto de datos para este ejemplo consiste en las calificaciones del examen de selectividad obtenidas en una determinada asignatura del distrito asociado a la Universidad de Valladolid. Los 1550 exámenes recibidos por el tribunal calificador se repartieron entre 10 correctores de forma que cada uno recibió entre 152 y 156 (ver Tabla 6.5).

Corrector	1	2	3	4	5	6	7	8	9	10
Nº de exámenes	155	152	155	156	156	156	156	154	156	154

Tabla 6.5: Número de exámenes por corrector.

En la Figura 6.1 se muestra el diagrama de cajas múltiple con las calificaciones de cada cor

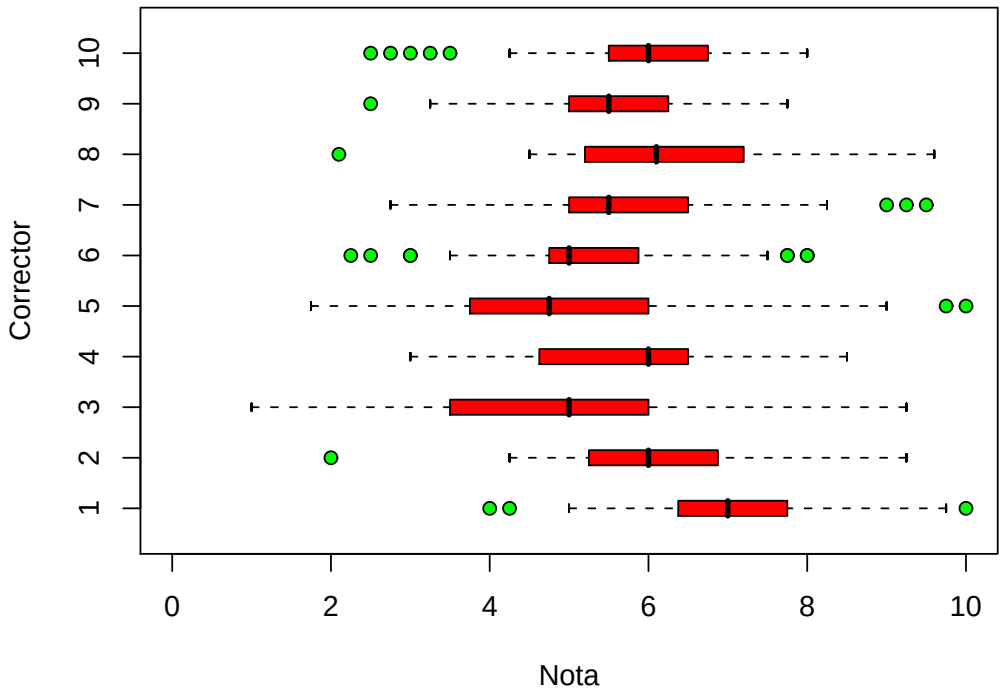

Figura 6.1: Calificaciones de los 10 correctores

La cuestión que se plantea es decidir si los correctores están calificando de forma homogénea. Y en caso de no ser así, identificar qué correctores son los que corrigen de forma más diferente (e identificar, si lo hay, el subgrupo de los que corrigen de forma homogénea).

El proceso secuencial descrito al principio se traduce en este ejemplo en lo siguiente: En cada paso hay un grupo de correctores y se comparan las calificaciones de cada corrector con las de todos los demás del grupo. Se busca si hay algún corrector cuyas correcciones no sean homogéneas con el resto de correctores y si lo hay se extrae del grupo. En caso de haber varios se excluye el más dispar. Además, y puesto que la procedencia de los alumnos es variada, se admite la posibilidad de que haya entre los alumnos de cada corrector un subgrupo con un comportamiento heterogéneo con respecto al resto.

Los resultados obtenidos para este ejemplo, aplicando la metodología derivada del Teorema 5.31, con 1000 réplicas bootstrap y recortando un 1 %, 5 %, 10 % y 20 % de los exámenes recibidos por cada corrector, respectivamente, son los que se observan en las Tablas 6.6-6.12.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
1	0.000	0.000	0.000	0.000
2	0.000	0.134	0.996	1.000
3	0.000	0.000	0.000	0.036
4	0.079	0.497	0.951	1.000
5	0.000	0.000	0.001	0.572
6	0.000	0.000	0.003	0.642
7	0.406	0.997	1.000	1.000
8	0.000	0.000	0.450	1.000
9	0.000	0.001	0.058	1.000
10	0.000	0.000	0.003	0.591

Tabla 6.6: p-valores cuando todos los correctores están en el grupo.

A la vista de la sucesión de p-valores de la Tabla 6.6, es claro que la distribución de calificaciones del corrector 1 no puede considerarse similar a la del resto. Ni siquiera recortando un 20 % logramos que el p-valor se aleje del 0. En esta misma tabla observamos ya algunos correctores que salvo por una pequeña proporción de calificaciones, muestran un comportamiento homogéneo con el resto. Son los correctores 2, 4 y 7 si recortamos un 5 %.

Y se añadiría al grupo el 8, si recortamos hasta un 10 %. Mirando también a la evolución con el tamaño de recorte de los p-valores de otros correctores se puede adivinar que otros tienen una corrección diferente al resto. Un buen candidato es el número 3. La necesidad de un proceso secuencial viene dada por el hecho de que la distribución con la que comparamos las calificaciones de cada corrector puede estar formada por correctores con un comportamiento no homogéneo. Así pues, cada vez que detectemos un corrector heterogéneo lo sacamos del grupo y volvemos a repetir el análisis.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
2	0.000	0.001	0.740	1.000
3	0.000	0.000	0.000	0.124
4	0.083	0.466	0.949	1.000
5	0.000	0.000	0.006	0.766
6	0.000	0.000	0.031	0.996
7	0.254	0.996	1.000	1.000
8	0.000	0.000	0.036	1.000
9	0.000	0.007	0.399	1.000
10	0.000	0.000	0.000	0.367

Tabla 6.7: p-valores con el corrector 1 fuera del grupo.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
2	0.000	0.089	0.992	1.000
4	0.173	0.518	0.966	1.000
5	0.000	0.000	0.000	0.146
6	0.000	0.000	0.037	0.993
7	0.920	0.999	1.000	1.000
8	0.000	0.000	0.299	1.000
9	0.000	0.055	0.931	1.000
10	0.000	0.000	0.010	0.652

Tabla 6.8: p-valores con los correctores 1 y 3 fuera del grupo.

En la Tabla 6.7 se muestran los resultados después de haber eliminado al corrector 1 del grupo. Asumiendo que en caso de que un corrector tenga un subgrupo heterogéneo éste no superará el 10 %, observamos que hay varios correctores candidatos a salir del grupo: el 3, el 10, el 5 y el 8, por este orden. Mirando a la columna de p-valores del 20 %, elegimos el 3.

Los p-valores que se obtienen cuando volvemos a repetir el proceso sin los correctores 1 y 3 en el grupo se muestran en la Tabla 6.8. En este caso, son los correctores 5, 10 y 6 los más diferentes. Se observa además que el corrector 8 que en el paso anterior podía ser considerado diferente con un 10 % de recorte, no lo es aquí. La diferencia está en el grupo de referencia. Sacamos del grupo al corrector 5.

A la vista de los resultados en la Tabla 6.9, el corrector 6 sale del grupo.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
2	0.006	0.619	1.000	1.000
4	0.058	0.349	0.888	1.000
6	0.000	0.000	0.001	0.769
7	0.712	0.998	1.000	1.000
8	0.000	0.005	0.709	1.000
9	0.000	0.056	0.975	1.000
10	0.000	0.001	0.038	0.879

Tabla 6.9: p-valores con los correctores 1, 3 y 5 fuera del grupo.

Ya en la Tabla 6.10, con los correctores 1, 3, 5 y 6 fuera del grupo, observamos que los correctores más diferentes son el 9, 10 y 8, por este orden, si miramos a la columna del 5 %. Si combinamos estos resultados con los de la columna correspondiente a un recorte del 10 %, entonces el candidato a salir será claramente el 10. Como la distancia recortada es de 0.33 puntos cuando recortamos el 5 % y de 0.25 puntos cuando recortamos el 10 %, optamos por eliminar el corrector 10.

En la Tabla 6.11 se muestran los p-valores después de haber eliminado los correctores 1, 3, 5, 6 y 10. A nivel del 10 % no aparece ningún corrector claramente diferente (aunque sí a nivel 5 %) por lo que optamos por parar. Podemos ahora repetir el procedimiento para ver si alguno de los correctores que se ha ido quedando fuera, pudiera volver a considerarse

homogéneo con los que ahora tenemos en el grupo. A la vista de la Tabla 6.12 es claro que no.

En la Figura 6.2 se muestra el diagrama de cajas con las calificaciones del grupo que corrige de forma homogéna y los correctores dispares.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
2	0.067	0.992	1.000	1.000
4	0.013	0.196	0.825	1.000
7	0.322	0.914	1.000	1.000
8	0.001	0.020	0.865	1.000
9	0.000	0.004	0.338	1.000
10	0.000	0.005	0.120	0.979

Tabla 6.10: p-valores con los correctores 1, 3, 5 y 6 fuera del grupo.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
2	0.027	0.857	1.000	1.000
4	0.017	0.165	0.684	1.000
7	0.771	0.999	1.000	1.000
8	0.000	0.017	0.819	1.000
9	0.000	0.004	0.352	1.000

Tabla 6.11: p-valores con los correctores 1, 3, 5, 6 y 10 fuera del grupo.

	Tamaño del recorte			
Corrector	1 %	5 %	10 %	20 %
1	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.008
5	0.000	0.000	0.000	0.001
6	0.000	0.000	0.000	0.000

Tabla 6.12: p-valores para examinar si entra alguno de los correctores que está fuera.

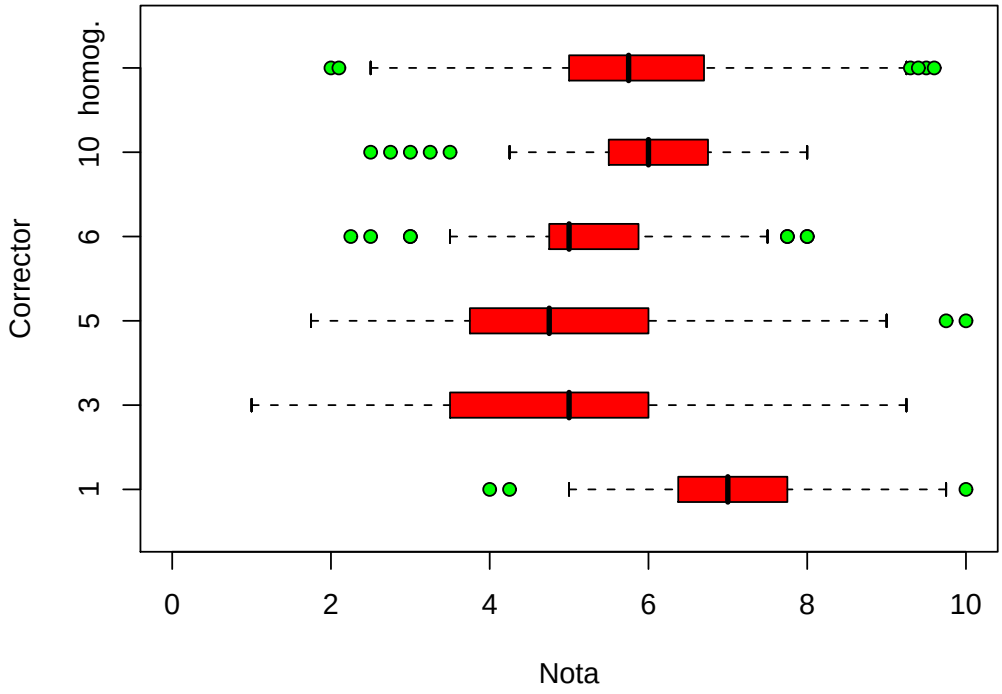

Figura 6.2: Calificaciones de los correctores eliminados y del grupo obtenido como homogéneo.

Apéndice A

Códigos de los programas informáticos utilizados.

A.1. Problemas de una muestra

A.1.1. Función en R para el cálculo del recorte de una empírica que mejor aproxima la U[0,1].

Fichero `onesample.R`

```
### Funcion en R para calcular el recorte imparcial exacto cuando
### recortamos en una muestra y comparamos con la U[0,1] con la metrica W2
### Paquetes: quadprog

### Devuelve una lista con los siguientes elementos
### sample: los valores de la muestra,
### dist1: la distancia W2 al cuadrado entre la U[0,1] y la muestra recortada
### dist2: la distancia W2 al cuadrado entre la U[0,1] y la muestra sin recortar
### weights: las probabilidades de los puntos de la muestra despues de recortar

OneSample.iTrim <- function(sample, alpha=0.1) {

n<-length(sample)
sample<-sort(sample)

### building the matrices and vectors needed by quadprog package
# initialization of d vector needed by quadprog
d<-rep(0,n-1)
diagonal<-rep(0,n-1)
A<-NULL
for (i in 1:(n-1)) {
```

```
d[i]<- sample[i+1]^2-sample[i]^2
diagonal[i]<- 2*(sample[i+1]-sample[i])
A<-cbind(A,c(rep(0,i-1),-1,1,rep(0,(n-i-1))))
}

for (i in 1:(n-2)) {A<-rbind(A,c(rep(0,i-1),-1,1,rep(0,(n-i-2)) ))}
A<-rbind(A,c(1,rep(0,n-2)),c(rep(0,n-2),-1))

# The A matrix, the D matrix and the b0 vector needed by quadprog
A<-t(A)
D<-diag(diagonal)
b0<-c(rep(-1/((1-alpha)*n),n-1),1-1/((1-alpha)*n),rep(0,n-2),0,-1)

### load quadprog package and solve
library(quadprog)
result<-solve.QP(D,d,A,b0)

### Computing outputs
distance1<-1/3 + sample[n]^2-sample[n] - crossprod(d,result$solution) +
               0.5*((result$solution)%*%D%*%(result$solution))

# The L2 Wasserstein distance between the U[0,1] and the original sample
c<-(1:(n-1))/n
distance2<-1/3 + sample[n]^2-sample[n] - crossprod(d,c) + 0.5*(c%*%D%*%c)

# The probability mass function of the impartially trimmed sample
weights<-diff(c(0,result$solution,1))
output<-list(sample=sample, dist1=distance1, dist2=distance2, weights=weights)
return(output)
}
```

A.1.2. Código en AMPL+MINOS para el cálculo del recorte de una empírica que mejor aproxima la U[0,1] (programación cuadrática).

Fichero `prog_cuadratica.dat` (se proporcionan los valores de la muestra x)

```
param n:=10;
param alpha:=0.1;
param x:=
1 0.1
2 0.12
3 0.21
4 0.22
5 0.3
6 0.5
```

```
7 0.6
8 0.7
9 0.75
10 0.8;
```

Fichero `prog_cuadratica.run`

```
#include prog_cuadratica.run;
reset;
model prog_cuadratica.mod;
data prog_cuadratica.dat;
let{i in puntos:i<n}v[i]:=x[i+1]^2-x[i]^2;
let{i in puntos:i<n}w[i]:=x[i+1]-x[i];
option solver minos;
solve;
display suma_gi;
display c;
```

Fichero `prog_cuadratica.mod` (contiene el modelo)

```
param n;
param alpha;
set puntos:=1..n;
param x{i in puntos}; # tiene que venir ordenada
param v{i in puntos:i<n};
param w{i in puntos:i<n};
var c{i in puntos:i<n}>=0;
maximize suma_gi:
sum{i in puntos:i<n}(v[i]*c[i]-w[i]*c[i]^2);
subject to orden1{i in puntos:i<n-1}:
0 <= c[i+1]-c[i];
subject to orden2{i in puntos:i<n-1}:
c[i+1]-c[i] <= 1/(n*(1-alpha));
subject to borde1:
1-c[n-1]<=1/(n*(1-alpha));
subject to borde2:
c[1]<=1/(n*(1-alpha));
```

A.1.3. Programa en R para calcular el estadístico $D_{n,k}$

Fichero `normasupremo.R`

```
## Programa en R para calcular el recorte que hace minima la distancia del
## supremo a una funcion de distribucion poblacional dada, F

## Funciones auxiliares
## Funcion para el calculo del superior de D+, D- en un punto dado
```

```
superior<- function(x,i,m) {
a<-max(abs(x-i/m),abs(x-(i-1)/m))
return(a)
}

## Funcion principal: Devuelve el valor del estadistico Dnk
## Argumentos: la funcion de distribucion, F, de P, evaluada en la
## muestra y el numero de puntos que se recortan, k
dnk<- function(Fsample,k) {
Fsample<-sort(Fsample)
n<-length(Fsample)
m<-n-k
matrizA<-array(0,dim=c(n-m+1,n))
for (i in 1:n) {
h2<-min(i,n-m+1)
h1<-max(1,i-m+1)
for (j in h1:h2){
matrizA[j,i]<-superior(datos[i,2],i-j+1,m)
}
}

matrizB<-array(0,dim=c(n-m,n))
for (i in 1:n) {
h2<-min(i,n-m)
h1<-max(1,i-m)
for (j in h1:h2){
matrizB[j,i]<-abs(datos[i,2]- (i-j)/m)
}
}

matrizA2<-matrizA
matrizB2<-matrizB

for (i in 2:n) {
h2<-min(i,n-m+1)
h1<-max(2,i-m+1)
# Para la primera fila el valor se calcula de forma distinta
matrizA2[1,i]<-max(matrizA2[1,i],matrizA2[1,i-1])
for (j in h1:h2){
matrizA2[j,i]<-min(max(matrizA2[j,i], matrizA2[j,i-1]),
    max(matrizA2[j,i], matrizB2[j-1,i-1]))
}
```

```
h2<-min(i,n-m)
h1<-max(2,i-m)
# Para la primera fila el valor se calcula de forma distinta
matrizB2[1,i]<-max(matrizB2[1,i],matrizA2[1,i-1])
for (j in h1:h2){
matrizB2[j,i]<-min(max(matrizB2[j,i] ,matrizA2[j,i-1]),
    max(matrizB2[j,i], matrizB2[j-1,i-1]))
}

}
minimo<-min(matrizA2[n-m+1,n],matrizB2[n-m,n])
return(minimo)
}
```

A.2. Problemas de dos muestras

A.2.1. Funcion en R para calcular los recortes en dos muestras

Fichero `twosample.R`

```
### Funcion en R para calcular el recorte imparcial entre dos muestras
### usando la metrica W_p
### Se necesita el paquete lpSolve
### Devuelve una lista con los siguientes elementos
### sample1: los valores de la primera muestra, sample2: los de la segunda,
### dist: la distancia W_p elevada a p
### mass1: la masa de probabilidad sobre cada punto de la muestra 1
### despues de recortar imparcialmente, mass2: idem para la muestra 2

TwoSample.iTrim <- function(Sample1=rnorm(6),Sample2=rnorm(9),
               Alpha1=0.1, Alpha2=0.1, p=2)
{
# Determining sample sizes
n<-length(Sample1)
m<-length(Sample2)
# Initialization and building of matrices and vectors needed by lpSolve
objective.in<-rep(0,n*m)
# coefficients of the objective function
for (i in 1:n) {
for (j in 1:m) objective.in[(i-1)*m+j]<-(Sample1[i]-Sample2[j])^p
}
# design matrix : dimensions n+m+1 x n*m
```

```
M1<-NULL
M2<-NULL
aux2<-diag(m)
for (i in 1:n){
aux1<-matrix(0,n,m)
aux1[i,]<-rep(1,m)
M1<-cbind(M1,aux1)
M2<-cbind(M2,aux2)
}
const.mat<-rbind(M1,M2,rep(1,n*m))
# direction of constraints
const.dir<-c(rep("<=",n+m),"=")
# right hand side
const.rhs<-c(rep(1/((1-Alpha1)*n),n),rep(1/((1-Alpha2)*m),m),1)

### Load library lpSolve and solve
library(lpSolve)
a<-lp(direction="min",objective.in,const.mat,const.dir,const.rhs)

### Computing outputs
distance<-a$objval
iTrim.Sample1<-crossprod(t(M1),a$solution)
iTrim.Sample2<-crossprod(t(M2),a$solution)

output<-list(sample1=Sample1,sample2=Sample2, dist=distance,
                    mass1=iTrim.Sample1, mass2=iTrim.Sample2)
return(output)
}
```

A.2.2. Código en AMPL+CPLEX cuando recortamos en dos muestras

Fichero `twosample.mod`

```
param N;
param M;
param a {1..N};
param b {1..M};
param alpha1>0, default 0.1;
param alpha2>0, default 0.1;

var z {1..N,1..M}>=0;

minimize dist_total:
sum{i in 1..N, j in 1..M}(a[i]-b[j])**2*z[i,j];
```

```
subject to Fila {i in 1..N}:
sum{j in 1..M} z[i,j] <= 1/(N*(1-alpha1));

subject to Columna {j in 1..M}:
sum{i in 1..N} z[i,j] <= 1/(M*(1-alpha2));

subject to recortex:
sum{i in 1..N,j in 1..M} z[i,j]= 1;
```

Fichero `twosample.run`

```
reset;
model twosample.mod;
data twosample.dat;
option solver cplexamp;
solve;
display dist_total;
```

Fichero `twosample.dat`

```
param N:=1000;
param M:=1000;
param a:=
1 0.004634
2 0.006411
........
1000 0.998761;
param b:=
1 -3.7632
2 -3.0999
........
1000 3.1321;
```

A.3. Recorte con el mismo patrón

A.3.1. Programa en R para la comparación de una y dos muestras

Fichero `unaydosmuestras_mismopatron.R`

```
## Programa en R para obtener el recorte con el mismo patron entre
## una muestra y una de referencia o entre dos muestras

## Funciones auxiliares

t.area<-function(integrando, corte) {
k<-length(integrando)
```

```
area<-(1/k)*sum(integrando*(integrando<=corte))
return(area)
}

## Funcion que devuelve el valor del estadistico T_{n,\alpha} para una muestra
## Argumentos: la muestra y el tamano de recorte
## Atencion a esta funcion que incorpora la funcion cuantil de la distribucion de
## referencia siempre que R la tenga implementada. En el ejemplo, una beta.

Tna<-function(sample,alpha) {
k<-10000
Grid<-seq(0.0001,0.9999,1/(k-1))
Fn_1<-quantile(sample1,probs=Grid,names=FALSE,type=1)
G_1<-qbeta(Grid,1,1)
integrando<-(Fn_1-G_1)^2
corte<-quantile(integrando,1-alpha,names=FALSE,type=1)
area<-t.area(integrando,corte)
return(area)
}

## Funcion que devuelve el valor del estadistico T_{n,m,\alpha} para dos muestras
## Argumentos: las dos muestras y el tamano de recorte

Tnma<-function(sample1,sample2,alpha) {
k<-10000
Grid<-seq(0.0001,0.9999,1/(k-1))
Fn_1<-quantile(sample1,probs=Grid,names=FALSE,type=1)
Gm_1<-quantile(sample2,probs=Grid,names=FALSE,type=1)
integrando<-(Fn_1-Gm_1)^2
corte<-quantile(integrando,1-alpha,names=FALSE,type=1)
area<-t.area(integrando,corte)
return(area)
}


## Funcion que devuelve la estimacion de la var asintotica para una muestra
## Argumentos: muestra y el vector de (i/n)-cuantiles de la de referencia, mas
## el tamano de recorte

sna<-function(sample,cuantiles,alpha) {
sample<-sort(sample)
n<-length(sample)
spacings<-diff(sample)
midranks<-(sample[1:(n-1)]+sample[2:n])/2
```

```
corte<-quantile(abs(sample-cuantiles,1-alpha,names=FALSE,type=1)
ani<-spacings*(midranks-cuantiles[1:(n-1)])*
                    (abs(sample[1:(n-1)]-cuantiles[1:(n-1)])<=corte)
lni<-cumsum(ani)
s2<-(4/(1-alpha)^2)*(sum(lni^2)/n-(sum(lni)/n)^2))
return(s2)
}

## Funcion que devuelve la estimacion de la var asintotica para dos muestras
## Argumentos: las dos muestras y el tamano de recorte

snma<-function(sample1,sample2,alpha) {
n<-length(sample1)
m<-length(sample2)
cuan1<-quantile(sample1,probs=(1:(m-1))/m,names=FALSE,type=1)
cuan2<-quantile(sample2,probs=(1:(n-1))/n,names=FALSE,type=1)
s2<-(m*sna(sample1,cuan2,alpha)+n*sna(sample2,cuan1,alpha))/(n+m)
return(s2)
}
```

A.3.2. Programa en R para buscar la mejor aproximación normal

Fichero `casinormalidad.R`

```
## Funciones auxiliares

t.area<-function(integrando, corte) {
area<-mean(integrando*(integrando<=corte))
return(area)
}

# n=num. de filas
v2m<-function(k,n) {return(list(i=(k%%n)+n*(k%%n==0),j=k%/%n+(k%%n>0) )) }

## Funcion que busca en una malla los valores de la media y la desviacion tipica
## de la normal mas cercana

# Esta funcion devuelve la media y la desv. tipica de la normal mas proxima,
# y la distancia al cuadrado (el valor del estad. T_{n,\alpha})

grid_search<-function(x,alpha,mu0,mu1,mu_step,s0,s1,s_step) {
mu<-seq(mu0,mu1,by=mu_step)
sigma<-seq(s0,s1,by=s_step)
n<-length(mu)
```

```
m<-length(sigma)
area<-array(dim=c(n,m))
t<-seq(0.00001,0.99999,by=0.00001)
for (i in 1:n){
for (j in 1:m){
integrando<-(quantile(x,t,names=FALSE,type=1)-mu[i]-sigma[j]*qnorm(t))^2
corte<-quantile(integrando,1-alpha,names=FALSE,type=1)
area[i,j]<-t.area(integrando,corte)
}
}

return(list(mu=mu[v2m(which.min(area),n)$i],
sigma=sigma[v2m(which.min(area),n)$j], dist=min(area)))
}

## Funcion que estima la varianza asintotica

var_est<-function(x,alpha,mu_opt,s_opt) { # x debe estar ordenado
#calculo ln(t)
n<-length(x)
aux<-(x[1:(n-1)]-mu_opt-s_opt*qnorm((1:(n-1))/n))^2
corte<-quantile(aux , 1-alpha, names=FALSE,type=1)
a<-(diff(x[1:n]^2)/2
               - diff(x[1:n])*(s_opt*qnorm((1:(n-1))/n)+mu_opt))*(aux<=corte)
ln<-cumsum(a)
int_ln<-(1/n)*sum(ln)
int_ln2<-(1/n)*sum(ln^2)
var<-(4/(1-alpha)^2)*(int_ln2-int_ln^2)
return(var)
}
```

Bibliografía

Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. y Matrán, C. (2008a). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, **103**, 697-704. 14

Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. y Matrán, C. (2008b). Similarity of probability measures through trimming. Sometido. 14

Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. y Matrán, C. (2008c). Assessing when a sample is mostly normal. Sometido. 14

Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. y Matrán, C. (2008d). Assessing when a sample is mostly normal. *First Workshop of the ERCIM Working Group on Computing & Statistics (ERCIM'08). 18-21 de Junio de 2008. Neuchâtel, Suiza.* 14

Alvarez-Esteban, P.C.; Cuesta-Albertos, J.A. y Matrán, C. (2004). Trimming and goodness-of-fit. *IX CLAPEM. Congreso Latinoamearicano de Probabilidad y Estadística Matemática. 22-26 de Marzo de 2004. Punta del Este, Uruguay.* 14

Araujo, A. y Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables.* John Wiley and Sons, Inc. Nueva York, EEUU. 17, 29, 31

Ash, R.B (1972). *Real Analysis and Probability.* Academic Press, Inc. Londrés, Reino Unido. 49

del Barrio, E.; Cuesta-Albertos, J.A. y Matrán, C. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, **9**, 1–96. 18

DEL BARRIO, E.; CUESTA-ALBERTOS, J.A.; MATRÁN, C. y RODRÍGUEZ RODRÍGUEZ, J. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.,* **27**, 1230–1239. 18, 115

DEL BARRIO, E.; DEHEUVELS, P. y VAN DE GEER, S. (2007). *Lectures on Empirical Processes. Theory and Statistical Applications.* EMS Series of Lectures in Mathematics. EMS Publishing House. Zürich, Suiza. 18

DEL BARRIO, E.; GINÉ, E. y UTZET, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, **11**, 131–189. 18, 106, 134, 142

BAZARAA, M.S.; SHERALI, H.D. y SHETTY, C.M. (1993). *Nonlinear Programming. Theory and Algorithms. Second Edition.* John Wiley and Sons, Inc. Nueva York, EEUU 63

BICKEL, P.J. y FREEDMAN, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.,* **9**, 1196–1217. 16, 17, 49, 149

BILLINGSLEY, P. (1995). *Probability and Measure, Third Edition.* John Wiley and Sons, Inc. Nueva York, EEUU. 30

CASCOS, I. y LÓPEZ-DÍAZ, M. (2008). Consistency of the α-trimming of a probability. Applications to central regions. *Bernoulli,* **14**, 580-592. 9

CHOW S. y LIU J. (1992). *Design and Analysis of Bioavailability and Bioequivalence Studies.* Marcel Dekker, Inc. Nueva York, EEUU. 40

CSÖRGŐ, M. y HORVÁTH, L. (1993). *Weighted Approximations in Probability and Statistics.* Wiley. Nueva York, EEUU. 101, 143

CSÖRGŐ, M. y RÉVÉSZ, P. (1978). Strong approximations of the quantile process. *Ann. Statist.,* **6**, 882–894. 104

CUESTA ALBERTOS, J.A.; GARCÍA ESCUDERO, L.A. y GORDALIZA, A. (2002). On the asymptotics for trimmed best k-nets. *J. Multivariate Anal.,* **82**, 486–516. 9

CUESTA ALBERTOS, J.A.; GORDALIZA, A. y MATRÁN, C. (1997a). Trimmed k-means: An attempt to robustify quantizers. *Ann. Statist.,* **25**, 553–576. 9

Cuesta Albertos, J.A.; Gordaliza, A. y Matrán, C. (1998). Trimmed best k-nets: A robustified versión of an L_∞-based clustering method. *Statist. Probab. Lett.,* **36**, 401–413. 9

Cuesta Albertos, J.A. y Matrán, C. (1986). Strong Laws of Large Numbers in Abstract Spaces via Skorohod's Representation Theorem. *Sankhyā,* series A, **48**, 98–103. 95

Cuesta Albertos, J.A. y Matrán, C. (1989). Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.,* **17**, 1264–1276. 17

Cuesta Albertos, J.A.; Matrán, C. y Mayo, A. (2008). Estimators based in adaptively trimming cells in the mixture model. *J. R. Stat. Soc. Ser. B,* **70**, 779–802. 9

Cuesta Albertos, J.A.; Matrán, C. y Tuero Díaz, A. (1996). Properties of optimal maps for the L_2-Monge-Kantorovich transportations problem. *Manuscrito no publicado.* 18

Cuesta Albertos, J.A.; Matrán, C. y Tuero Díaz, A. (1997b). Optimal transportation plans and convergence in distribution. *J. Multivariate Anal.,* **60**, 72–83. 20

Cuesta-Albertos, J. A.; Matrán, C. y Tuero Díaz, A. (1997c). On the monotonicity of optimal transportation plans. *J. Math. Anal. Appl.,* **215**, 86–94. 20, 93

Dudley, R.M. (1989). *Real Analysis and Probability.* Wadsworth & Brook/Cole. Pacific Grove, CA, EEUU. 106

Evans, L. C. y Gariepy, R. F. (1992). *Measure Theory and Fine Properties of Functions.* Studies in Advanced Mathematics. CRC Press. Boca Raton, FL, EEUU. 54

Fleischer, P. (1964). Sufficient conditions for achieving minimum distorsion in a quantizer. *IEEE Int. Conv. Rec.*, 104–111. 12

Fourer, R.; Gay, D.M. y Kernighan, B.W. (2003). *AMPL. A Modeling Language for Mathematical Programming. Second Edition.* Brooks/Cole. Pacific Grove, CA, EEUU. 14

García Escudero, L.A. y Gordaliza, A. (1999). Robutness properties of k means and trimmed k means. *J. Amer. Statist. Assoc.,* **94**, No. 447, 956–969. 9

García Escudero, L.A. y Gordaliza, A. (2005). Generalized radius processes for elliptically contoured distributions. *J. Amer. Statist. Assoc.,* **100**, 1036–1045. 9

García Escudero, L.A.; Gordaliza, A. y Matrán, C. (1999a). A central limit theorem for multivariate generalized trimmed k-means. *Ann. Statist.* **27**, 1061–1079. 9, 50

García Escudero, L.A.; Gordaliza, A. y Matrán, C. (1999b). Asymptotics for trimmed k-means and associated tolerance zones. *J. Statist. Plann. Inference,* **77**, 247–262. 9

García Escudero, L.A.; Gordaliza, A. y Matrán, C. (2003). Trimming tools in exploratory data analysis. *J. Comput. Graph. Statist.,* **12**, 434–449. 9

Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J. Approx. Theory,* **64**, 162–180. 9

Grubbs, F.E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Statist.,* **21**, 27–58. 9

Hand, D.J.; Daly, F.; Lunn, A.D.; McConway, K.J. y Ostrowski, E. *A Handbook of Small Data Sets.* Chapman & Hall. Londres, Reino Unido. 126

Hartigan, J.A. (1978). Asymptotic distribution for clustering criteria. *Ann. Statist.,* **6**, 117–131. 12

Li, L. y Flury, B. (1995). Uniqueness of principal points for univariate distributions. *Statist. Probab. Lett.,* **25**, 323–327. 12

Massart, P. (2007). *Concentration Inequalities and Model Selection.* Lecture Notes in Mathematics, Vol. 1896. Springer Verlag. Berlin, Alemania. 24

McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.,* **80**, 309–323. 19, 20

Moore, D.S. y McCabe, G.P. (2003). *Introduction to the Practice of Statistics, Fourth Edition.* W.H. Freeman and Company. Nueva York, EEUU. 118

Munk, A. y Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B,* **60**, 223–241. 10, 23, 47, 117

National Institute of Standards and Technology (2007). *NIST/SEMATECH e-Handbook of Statistical Methods.* *http://www.itl.nist.gov/div898/handbook/eda/section4/eda42a.htm.* Gaithersburg, MD, EEUU.

Papadimitriou, C.H. y Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity.* Dover Publications. Mineola, NY, EEUU. 21

Parzen, E. (1979). Nonparametric statistical data modeling (with discussion). *J. Amer. Statist. Assoc.*, **74**, 105–121. 104

Pollard, D. (1981). Strong consistency of k-means clustering. *Ann. Statist.,* **9**, 135–140. 12

Pollard, D. (1982). A central limit theorem for k-means clustering. *Ann. Probab.,***10**, 919–926. 12

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing. http://www.R-project.org.* Viena, Austria. 14

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems. (2 vol.)* Springer Series in Statistics. Probability and its Applications. Springer Verlag. Nueva York, EEUU. 17, 19, 20

Resnick, S. I. (1987). *Extreme Value, Regular Variation, and Point Processes.* Springer Verlag. New York, EEUU. 98

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, y W. Werz (Eds.), *in Mathematical Statistics and Applications, Volume B.* Reidel, Dordrecht, Alemania. 9

Stute, W. y Zhu, L.X. (1995). Asymptotics of k-means clustering based on projection pursuit. *Sankhyā,* **57**, 462–471. 12

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes.* Springer Verlag. New York, EEUU. 114

Villani, C. (2003). *Topics in Optimal Transportation.* Graduate Studies in Mathematics Vol. 58, Amer. Math. Soc. Providence, RI, EEUU. 17, 20

Villani, C. (2009). *Optimal Transport: Old and New.* Springer Verlag. New York, EEUU. 17