Universidad de Valladolid

Facultad de Ciencias

Grado en matemáticas

TRABAJO DE FIN DE GRADO:

ANÁLISIS CLUSTER EN ESPACIOS ABSTRACTOS

AUTOR:
David Rodríguez Vítores

TUTOR:
Carlos Matrán Bea

Junio 2021

Índice

1. Introducción

Con la denominación de "análisis cluster", o análisis de conglomerados, hacemos referencia a un conjunto de técnicas diseñadas para encontrar diferentes patrones o comportamientos, "formas", dentro de un conjunto de datos. Esta descripción justifica sobradamente el interés del tema y su vigencia en el Análisis de Datos o en la Ciencia de los Datos, como ahora empieza a denominarse. El diseño de estas técnicas comporta problemas de concepto sobre las formas a buscar y especialmente de orden computacional incluso en las búsquedas de las formas conceptualmente más simples, como son las formas esféricas, que da lugar al algoritmo de k-medias. Los ingredientes presentes en el modelo de k-medias son las formas"esféricas", conceptualmente extensibles a espacios métricos generales, y las "medias", cuya característica de mejor aproximación por el criterio de mínimos cuadrados, dio lugar a las medias de Fréchet o baricentros en espacios abstractos. En este trabajo nos centraremos en el estudio de los baricentros y k-baricentros en un espacio de probabilidades, el espacio de Wasserstein, y mostraremos algunas aplicaciones, incluyendo posibles líneas de trabajo futuro en el Análisis de Datos.

Para ello, comenzamos introduciendo el **espacio y la distancia de Wasserstein**, y siguiendo [14], tratamos sus propiedades más importantes. El principal esfuerzo realizado en esta parte ha consistido en detallar las demostraciones de todos los resultados, incluyendo todos los argumentos necesarios para su comprensión al nivel de un estudiante del Grado en Matemáticas. A continuación, el trabajo se centra en el estudio de los **baricentros** en el espacio de Wasserstein. Replicando el desarrollo que aparece en [2], estudiamos sus principales propiedades y damos un algoritmo iterativo para el cálculo en determinadas situaciones. Posteriormente presentamos los **k-baricentros** como extensión natural de los baricentros. Debido a la complejidad del desarrollo teórico, en esta parte nos conformamos con mostrar algunos ejemplos de cálculo de los k-baricentros.

Seguidamente, mediante algunos ejemplos sencillos ilustramos diferentes aplicaciones de los k-baricentros, que nos muestran el potencial presente en este tipo técnicas para tratar problemas del Análisis de Datos, sobre todo relacionados con el análisis cluster. Por una parte, tratamos el problema de la **paralelización en el análisis cluster**: a partir de los k-baricentros desarrollamos un método que nos permite llevar a cabo un análisis cluster sobre un conjunto de datos mediante los algoritmos clásicos conocidos, como son las k-medias, dividiendo la tarea entre varios ordenadores. Este tipo de técnicas que nos permiten paralelizar el trabajo son fundamentales en la actualidad, puesto que la mayoría de problemas que aparecen hoy en día involucran grandes cantidades de datos, y puede ser muy costoso tratarlos en un único sistema. Por otra parte, de la misma forma que las k-medias nos permiten realizar una clasificación sobre un conjunto de datos, cuando tenemos un conjunto de probabilidades, podemos llevar a cabo una clasificación mediante los k-baricentros, logrando así un método para llevar a cabo un **análisis cluster sobre un conjunto de probabilidades**. Para ilustrar los dos métodos, hemos utilizado el mismo conjunto de datos, que contiene información de cervezas artesanales de diferentes tipos. De esta forma, haremos visible la diferencia que existe entre hacer un análisis cluster sobre un conjunto de datos de $\mathbb{R}^d$ (que serán cada una de las cervezas) o de probabilidades (que serán cada una de las tipologías de cervezas).

Finalmente, mostramos la utilidad del baricentro para la elección de unas **componentes principales comunes** que nos permitan estudiar un conjunto de datos en el cual aparecen individuos de diferentes clases. Las componentes principales comunes buscan unas direcciones en las que las estructuras de dispersión de todas las clases se expresen de forma sencilla. Este problema ya había sido tratado previamente, aunque en este trabajo proponemos un nuevo método para hallar las componentes a partir de los baricentros. Para el ejemplo de esta sección utilizamos los datos de las Iris, presentes en multitud de trabajos, y que nos permiten además comparar nuestras componentes principales comunes con las obtenidas a partir de otros métodos. A partir de este nuevo enfoque, presentamos las **componentes principales grupales** como la extensión natural de este concepto a partir de los k-baricentros. Consisten básicamente en agrupar nuestras clases de datos, y hallar para cada grupo unas componentes principales comunes en las que esas clases queden bien representadas. Esta técnica nos proporciona nuevamente un método para llevar a cabo un análisis cluster sobre un conjunto de probabilidades, que ilustramos en un ejemplo mediante la clasificación de diferentes estilos musicales.

Antes de comenzar, quiero mostrar mi agradecimiento al Departamento de Estadística e Investigación Operativa de la Universidad de Valladolid, puesto que la generación del software necesario para los distintos ejemplos no habría sido posible sin los programas que me han facilitado, que han sido desarrollados por los miembros de dicho departamento para trabajos previos.

2. Distancia de Wasserstein

En esta sección recurriremos repetidamente al estudio de probabilidades definidas en espacios producto. Siempre que tengamos dos espacios medibles (Ω_1, σ_1) y (Ω_2, σ_2), al considerar el espacio producto $\Omega_1 \times \Omega_2$, entenderemos que trabajamos en él con la σ-álgebra producto $\sigma_1 \otimes \sigma_2$, definida como la mínima σ-álgebra que contiene a todos los conjuntos $A \times B$, con $A \in \sigma_1$ y $B \in \sigma_2$. Como es habitual, denotaremos por (x, y) a los puntos de $\Omega_1 \times \Omega_2$. En particular, en el caso en que tengamos una función medible

$$f : \Omega_1 \times \Omega_2 \to \mathbb{R}$$

y una probabilidad π en $(\Omega_1 \times \Omega_2)$, denotaremos a la integral de f respecto de π por:

$$\int_{\Omega_1 \times \Omega_2} f d\pi \quad \text{o} \quad \int_{\Omega_1 \times \Omega_2} f(x, y) d\pi(x, y)$$

siempre que dicha integral exista, dependiendo si es necesario especificar las variables respecto de las que integramos o no.

Estos comentarios se extienden al caso en que tenemos un número finito de espacios probabilísticos $\{(\Omega_i, \sigma_i)\}_{i=1}^n$. En este caso consideraremos también siempre sobre el espacio producto $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$ la σ-álgebra producto $\sigma_1 \otimes \sigma_2 \otimes \cdots \otimes \sigma_n$, que es la menor que contiene a los conjuntos $A_1 \times A_2 \times \cdots \times A_n$, con $A_i \in \sigma_i \, \forall \, i = 1, ..., n$. Denotamos por $(x_1, x_2, ..., x_n)$ a los puntos de $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$, y si f es una función real medible y π una probabilidad en dicho espacio producto, entonces denotaremos a la integral de f respecto de π siempre que exista por:

$$\int_{\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n} f d\pi \quad \text{o} \quad \int_{\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n} f(x_1, x_2, ..., x_n) d\pi(x_1, x_2, ..., x_n)$$

De forma análoga se hará cuando estemos trabajando con una cantidad infinita numerable de espacios probabilísticos $\{(\Omega_i, \sigma_i)\}_{i=1}^\infty$ y consideremos el espacio producto $\prod_{i=1}^\infty \Omega_i$, teniendo en cuenta que en este caso la σ-álgebra producto es la menor de las que contienen a todos los conjuntos $A_1 \times \cdots \times A_n \times \prod_{i=n+1}^\infty \Omega_i$, con $A_i \in \sigma_i$, $\forall \, n \in \mathbb{N}, i = 1, ..., n$.

2.1. Definiciones y conceptos básicos

Sea B un espacio de Banach separable con una norma $\|\cdot\|$. Sea $1 \leq p < \infty$. Sea $\Gamma_p(B)$ el conjunto de probabilidades γ en β, la σ- álgebra de Borel de B, tales que

$$\int_B \|x\|^p d\gamma(x) < \infty$$

Esto es, si U es una variable aleatoria definida en algún espacio probabilístico (Ω, σ, P) que induce la probabilidad γ en B, se pide que U tenga momento finito de orden p, es decir $E\|U\|^p < \infty$.

Definición 1. *Dados $\mu, \nu \in \Gamma_p(B)$ se define la p-distancia de Wasserstein por:*

$$d_p(\mu, \nu) = \text{ínf}\{(E\|U - V\|^p)^{1/p} : \mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu\} \tag{1}$$

Es decir, el inferior se busca entre todas las variables aleatorias U, V cuyas leyes de probabilidad inducidas sobre B son μ, ν respectivamente. Además, para que $U - V$ tenga sentido como v.a., U, V han de estar definidas en un mismo espacio probabilístico (Ω, σ, P), es decir:

$$\Omega \xrightarrow{U} B \quad \text{tal que} \quad \mathscr{L}(U) = \mu$$

$$\Omega \xrightarrow{V} B \quad \text{tal que} \quad \mathscr{L}(V) = \nu$$

Y podemos definir entonces la variable aleatoria:

$$\Omega \xrightarrow{(U,V)} B \times B \quad \text{tal que} \quad \mathscr{L}(U,V) = \pi$$

Dado que las distribuciones marginales de U y de V no nos permiten conocer la distribución del vector aleatorio (U, V), no conocemos la probabilidad π que induce el vector aleatorio (U, V) en $B \times B$. Como la distribución de la variable aleatoria $U - V$ depende de la distribución conjunta de (U, V), si (U', V') es otro par de v.a. con las mismas propiedades, el hecho de que $U =_d U'$ y $V =_d V'$, no implica que

$$U - V =_d U' - V'$$

En consecuencia, tampoco se tiene necesariamente que las v.a. $\|U - V\|^p$ y $\|U' - V'\|^p$ estén igualmente distribuidas, y por tanto

$$E(\|U - V\|^p) \quad y \quad E(\|U' - V'\|^p)$$

pueden tomar valores distintos.

Vamos a ver que no es necesario precisar en cada caso cuáles son las v.a. U y V, ni fijarnos en qué espacio probabilístico están definidas. Sólo necesitaremos conocer la ley de probabilidad conjunta que inducen en $B \times B$. Para verlo, definimos las funciones coordenadas:

$$B \times B \xrightarrow{\Pi_X} B \qquad \Pi_X(a, b) = a$$

$$B \times B \xrightarrow{\Pi_Y} B \qquad \Pi_Y(a, b) = b$$

- Sean U, V dos variables aleatorias definidas en un espacio probabilístico (Ω, σ, P) tales que $\mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu$. Tenemos entonces un vector aleatorio:

  $$\Omega \xrightarrow{(U,V)} B \times B \quad \text{con ley de probabilidad conjunta} \quad \mathscr{L}(U,V) = \pi$$

  π es una probabilidad en $B \times B$, que además verifica:

  $$\pi \circ \Pi_X^{-1}(A) = \pi(A \times B) = P(U \in A) = P_U(A) = \mu(A) \quad \forall A \in \beta \quad \Rightarrow \pi \circ \Pi_X^{-1} = \mu$$

  $$\pi \circ \Pi_Y^{-1}(A) = \pi(B \times A) = P(V \in A) = P_V(A) = \nu(A) \quad \forall A \in \beta \quad \Rightarrow \pi \circ \Pi_Y^{-1} = \nu$$

  Es decir, cualquier par de variables aleatorias U, V definidas en un mismo espacio con $\mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu$, se pueden relacionar con la probabilidad π que inducen en $B \times B$, que cumple que

  $$\pi \circ \Pi_X^{-1} = \mu, \quad \pi \circ \Pi_Y^{-1} = \nu$$

y además:

$$E(\|U-V\|^p) = \int_\Omega \|U-V\|^p dP = \int_{B\times B} \|\Pi_X - \Pi_Y\|^p d\pi =$$
$$= \int_{B\times B} \|x-y\|^p d\pi(x,y)$$

- Recíprocamente, dada una probabilidad π en $B\times B$ que cumple que $\pi\circ\Pi_X^{-1}=\mu$ y $\pi\circ\Pi_Y^{-1}=\nu$, tomando las v.a. $U=\Pi_X, V=\Pi_Y$, tenemos dos v.a. definidas en un mismo espacio probabilístico $(B\times B, \beta\times\beta, \pi)$ y tales que

$$\mathscr{L}(U)=\mu, \quad \mathscr{L}(V)=\nu$$

y además:

$$E(\|U-V\|^p) = \int_\Omega \|U-V\|^p dP = \int_{B\times B} \|\Pi_X - \Pi_Y\|^p d\pi =$$
$$= \int_{B\times B} \|x-y\|^p d\pi(x,y)$$

Por tanto, a partir de esta relación, podemos escribir:

$$d_p(\mu,\nu) = \text{ínf}\left\{\left(\int_{B\times B}\|x-y\|^p d\pi(x,y)\right)^{1/p} : \pi \text{ prob. en } B\times B,\ \pi\circ\Pi_X^{-1}=\mu,\ \pi\circ\Pi_Y^{-1}=\nu\right\}$$

Esto nos permite relacionar la definición de $d_p(\mu,\nu)$ con el concepto de emparejamiento. Las definiciones y resultados de la próxima sección aparecen en [10, cap. 1].

2.2. Emparajamientos probabilísticos

Definición 2. *Sean (Ω_1,σ_1,μ) y (Ω_2,σ_2,ν) dos espacios probabilísticos. Emparejar μ y ν significa construir en un espacio probabilístico genérico (Ω,σ,P) dos variables aleatorias*

$$\Omega \xrightarrow{U} \Omega_1 \quad \text{y} \quad \Omega \xrightarrow{V} \Omega_2$$

tales que $\mathscr{L}(U)=\mu$ y $\mathscr{L}(V)=\nu$. La pareja (U,V) se dice que es un emparejamiento de (μ,ν). Abusando de lenguaje, la ley de probabilidad de (U,V) se llama también un emparejamiento de (U,V).

Si π es un emparejamiento de μ y ν,y además verifica que:

$$d_p(\mu,\nu) = \left(\int_{B\times B}\|x-y\|^p d\pi(x,y)\right)^{1/p}$$

entonces diremos que π es un emparejamiento óptimo de μ y ν para el coste dado por la p-distancia de Wasserstein.

Si μ y ν son las únicas leyes del problema, generalmente se toma $\Omega = \Omega_1 \times \Omega_2$. Emparejar μ y ν significa encontrar una probabilidad π en $\Omega_1 \times \Omega_2$ tal que admite μ y ν como probabilidades marginales sobre Ω_1 y Ω_2 respectivamente. Es decir, si Π_X, Π_Y son las proyecciones sobre Ω_1 y Ω_2:

$$\pi \circ \Pi_X^{-1} = \mu, \quad \pi \circ \Pi_Y^{-1} = \nu$$

Esto equivale a que para todos los conjuntos medibles $A \subset \Omega_1, B \subset \Omega_2$,

$$\pi(A \times \Omega_2) = \mu(A) \quad \pi(\Omega_1 \times B) = \nu(B)$$

Lema 1. *Siempre existen emparejamientos entre dos espacios probabilísticos* $(\Omega_1, \sigma_1, \mu)$ *y* $(\Omega_2, \sigma_2, \nu)$.

Demostración. Basta recurrir al emparejamiento trivial en el que las variables U, V son independientes. La construcción de dicho emparejamiento se hace a partir de las ideas presentes en teoria de la medida, al tratar las medidas producto. El desarrollo básico es el siguiente:

Consideramos el espacio producto $\Omega_1 \times \Omega_2$. Denotemos por σ a la σ-álgebra producto en dicho espacio. Para cada $D \in \sigma$ y para cada $x \in \Omega_1, y \in \Omega_2$, consideramos las secciones

$$D_x = \{y \in \Omega_2 : (x, y) \in D\} \subset \Omega_2$$

$$D_y = \{x \in \Omega_1 : (x, y) \in D\} \subset \Omega_1$$

Se puede comprobar que D_x es medible en Ω_2 para todo $x \in \Omega_1$, y D_y es medible en Ω_1 para todo $y \in \Omega_2$. Se cumple además que si definimos las funciones $\phi : \Omega_1 \to [0, \infty]$ y $\psi : \Omega_2 \to [0, \infty]$ por:

$$\phi(x) = \nu(D_x) \qquad \psi(y) = \mu(D_y)$$

entonces ϕ es medible en Ω_1 y ψ es medible en Ω_2. Definimos entonces la probabilidad producto de μ y ν como:

$$(\mu \times \nu)(D) = \int_{\Omega_1} \nu(D_x) d\mu(x) = \int_{\Omega_2} \mu(D_y) d\nu(y)$$

que se puede comprobar que está bien definida (ambas integrales dan el mismo valor para cualquier $D \in \sigma$), es una probabilidad en $\Omega_1 \times \Omega_2$ y además verifica que si A, B medibles con $A \subset \Omega_1$, $B \subset \Omega_2$ entonces:

$$P(A \times B) = \int_{\Omega_1} \nu((A \times B)_x) d\mu(x) = \int_A \nu(B) d\mu(x) = \mu(A)\nu(B)$$

Luego tomando como U, V las proyecciones de $\Omega_1 \times \Omega_2$ en Ω_1 y Ω_2 respectivamente, tenemos que (U, V) forman un emparejamiento de $(\Omega_1, \sigma_1, \mu)$ y $(\Omega_2, \sigma_2, \nu)$ y además son independientes. □

Presentamos ahora un concepto que tendrá importancia más adelante.

Definición 3. *Un emparejamiento (U,V) se dice que es determinista si existe una función medible* $T : \Omega_1 \longrightarrow \Omega_2$ *tal que* $V = T(U)$.

Esto es equivalente a:

- (U, V) es un emparejamiento de μ, ν cuya ley π está concentrada en el grafo de una función medible $T : \Omega_1 \longrightarrow \Omega_2$

- U tiene ley μ y V = T(U) tiene ley $\mu \circ T^{-1} = \nu$
- $\pi = \mu \circ (Id, T)^{-1}$

T usualmente se llama *aplicación de transporte*. Informalmente, T transporta la masa representada por la probabilidad μ a la masa representada por la probabilidad ν. Los emparejamientos deterministas no siempre existen. En el caso de que exista una aplicación de transporte T que lleve μ en ν, y que cumpla además que si U es una v.a. que induce probabilidad μ en B, la probabilidad π que induce $(U, T(U))$ en $B \times B$ sea un emparejamiento óptimo de (U, V) para el coste dado por la p-distancia de Wasserstein, decimos que T es una *aplicación de transporte óptimo para el coste dado por la p-distancia de Wasserstein.*

Vamos a probar ahora un lema que nos va a servir para asegurar la existencia de un tipo de emparejamientos que tendrán gran importancia más adelante. Para poder demostrarlo, se utilizará el concepto de las probabilidades de transición y su relación con las probabilidades en espacios producto. La explicación de estos conceptos y los resultados de esta teoría que vamos a utilizar están desarrollados de forma precisa en el apéndice 7.1.

Lema 2. ***(del pegado)*** *Sean $(\Omega_1, \sigma_1, \mu_1), (\Omega_2, \sigma_2, \mu_2), (\Omega_3, \sigma_3, \mu_3)$ tres espacios probabilísticos, tales que Ω_1, Ω_3 sean además espacios métricos completos y separables. Supongamos que P_{12} es una probabilidad en $\Omega_1 \times \Omega_2$ con marginales μ_1 y μ_2, y que P_{23} es una probabilidad en $\Omega_2 \times \Omega_3$ con marginales μ_2 y μ_3. Entonces se puede construir una probabilidad P en $\Omega_1 \times \Omega_2 \times \Omega_3$ tal que la distribución marginal sobre $\Omega_1 \times \Omega_2$ sea P_{12} y la distribución marginal sobre $\Omega_2 \times \Omega_3$ sea P_{23}.*

Demostración. Como se detalla en el apéndice 7.1, como Ω_1, Ω_3 son espacios métricos completos y separables, podemos descomponer las probabilidades P_{12} y P_{23} de la siguiente forma:

$$P_{12}(H) = \int_{\Omega_2} \nu_y^1(H_y) d\mu_2(y) \qquad \forall\, H \text{ medible } \subset \Omega_1 \times \Omega_2$$

$$P_{23}(J) = \int_{\Omega_2} \nu_y^3(J_y) d\mu_2(y) \qquad \forall\, J \text{ medible } \subset \Omega_2 \times \Omega_3$$

donde para cada $y \in \Omega_2$, ν_y^1 es una probabilidad de transición en Ω_1 y ν_y^3 una probabilidad de transición en Ω_3. Por tanto, para cada $y \in \Omega_2$, podemos considerar la probabilidad producto $\nu_y = \nu_y^1 \times \nu_y^3$ definida en $\Omega_1 \times \Omega_3$ por:

$$\nu_y(D) = \int_{\Omega_1} \nu_y^3(D_x) d\nu_y^1(x) = \int_{\Omega_3} \nu_y^1(D_z) d\nu_y^3(z) \qquad \forall\, D \text{ medible } \subset \Omega_1 \times \Omega_3$$

Por lo visto sobre medidas producto, sabemos que ν_y está bien definida. Además cumple que para todo $A \in \sigma_1$, $C \in \sigma_3$:

$$\nu_y(A \times C) = \nu_y^1(A) \nu_y^3(C)$$

Por tanto, para cada $y \in \Omega_2$, tenemos definida una probabilidad de transición ν_y en $\Omega_1 \times \Omega_3$. Definimos entonces la probabilidad P en $\Omega_1 \times \Omega_2 \times \Omega_3$ por:

$$P(E) = \int_{\Omega_2} \nu_y(E_y) d\mu_2(y) = \int_{\Omega_2} \left(\int_{\Omega_1} \nu_y^3((E_y)_x) d\nu_y^1(x) \right) d\mu_2(y) =$$

$$= \int_{\Omega_2} \left(\int_{\Omega_3} \nu_y^1((E_y)_z) d\nu_y^3(z) \right) d\mu_2(y)$$

para todo $E \subset \Omega_1 \times \Omega_2 \times \Omega_3$ medible. Veamos que esta probabilidad cumple lo que queríamos. Si H es medible en $\Omega_1 \times \Omega_2$, se tiene que:

$$P(H \times \Omega_3) = \int_{\Omega_2} \left(\int_{\Omega_1} \nu_y^3((H_y)_x) d\nu_y^1(x) \right) d\mu_2(y) =$$

$$= \int_{\Omega_2} \left(\int_{\Omega_1} I_{H_y}(x) d\nu_y^1(x) \right) d\mu_2(y) = \int_{\Omega_2} d\nu_y^1(H_y) d\mu_2(y) = P_{12}(H)$$

Por tanto, la probabilidad marginal de P sobre $\Omega_1 \times \Omega_2$ coincide con P_{12}. Análogamente, utilizando la otra representación de P obtenemos que la probabilidad marginal de P sobre $\Omega_2 \times \Omega_3$ es P_{23}, tal y como queríamos. □

Adaptando los correspondientes resultados a sus análogos cuando trabajamos con un número finito o infinito numerable de probabilidades, podemos generalizar este resultado:

Lema 3. ***(del pegado, generalización)***

- ***Para un número finito de factores:***
  Sean $(\Omega_j, \sigma_j, \mu_j), j = 1, 2, ..., n$, con $n \geq 2$, espacios probabilísticos, tales que $\Omega_2, \Omega_3, ..., \Omega_n$ sean además espacios métricos completos y separables.. Supongamos que P_{1j} es una probabilidad en $\Omega_1 \times \Omega_j$ con marginales μ_1 y μ_j para cada $j = 2, ..., n$. Entonces se puede construir una probabilidad P en $\Omega_1 \times \Omega_2 \times ... \times \Omega_n$ tal que la distribución marginal sobre $\Omega_1 \times \Omega_j$ sea P_{1j} para todo $j = 2, ..., n$.
- ***Para un número infinito numerable de factores:***
  Sean $(\Omega_j, \sigma_j, \mu_j), j \in \mathbb{N}$ espacios probabilísticos, tales que $\{\Omega_n\}_{n=2}^{\infty}$ sean además espacios métricos completos y separables.. Supongamos que P_{1j} es una probabilidad en $\Omega_1 \times \Omega_j$ con marginales μ_1 y μ_j para cada $j \in \mathbb{N}, j \geq 2$. Entonces se puede construir una probabilidad P en $\prod_{i=1}^{\infty} \Omega_i$ tal que la distribución marginal sobre $\Omega_1 \times \Omega_j$ sea P_{1j} para todo $j \in \mathbb{N}, j \geq 2$.

2.3. Relación con el problema de transporte óptimo

El problema de transporte óptimo es un problema clásico, que consiste en minimizar el coste de llevar una distribución de masa μ en un espacio normado B a otra distribución de masa ν en el mismo espacio B. Para que el problema tenga sentido, la masa inicial de la distribución μ ha de ser la misma que la masa final que hay en ν. Podemos suponer por tanto que la masa total es 1 y μ, ν son probabilidades en (B, σ). Suponemos además que tenemos una función de coste:

$$c(x, y) : B \times B \longrightarrow [0, +\infty)$$

que representa el coste de llevar una unidad de masa del punto x al punto y. En estas condiciones, la **formulación de Kantorovich del problema de transporte óptimo** es la siguiente:

Consideramos una probabilidad π en $B \times B$. Para cada $C \in \sigma, D \in \sigma$ pensemos que $\pi(C \times D)$ representa la cantidad de masa transportada de C a D. Para que π sea un plan de transporte válido, la cantidad de masa que sale de C ha de ser $\mu(C)$ y la cantidad de masa que llega a D ha de ser $\nu(D)$: es decir, se debe cumplir que:

$$\pi(C \times B) = \mu(C), \quad \pi(B \times D) = \nu(D) \quad \forall\, C \in \sigma, D \in \sigma$$

Es decir, π ha de ser una probabilidad en $B \times B$ que tenga probabilidades marginales μ, ν respectivamente. Denotamos el conjunto de todas estas probabilidades por $\Pi(\mu, \nu)$, que se corresponde a todos los posibles planes de transporte. $\Pi(\mu, \nu)$ es no vacío, ya que $\mu \times \nu \in \Pi(\mu, \nu)$. Podemos ahora definir el problema de transporte óptimo a partir de la formulación de Kantorovich:

Definición 4. *Dadas μ, ν probabilidades en B y una función de coste $c : B \times B \to \mathbb{R}$ medible, el problema de transporte óptimo consiste en hallar el mínimo de*

$$\mathbb{K}(\pi) = \int_{B \times B} c(x, y) d\pi(x, y)$$

entre todas las probabilidades $\pi \in \Pi(\mu, \nu)$.

Por tanto, en el caso en que μ, ν sean probabilidades en B tales que

$$\int \|x\|^p d\mu(x) < \infty \quad \int \|x\|^p d\nu(x) < \infty$$

tomando la función de coste $c(x, y) = \|x - y\|^p$, la solución del problema de transporte óptimo coincide exactamente con el valor de $d_p^p(\mu, \nu)$. Además, el inferior se alcanza para una probabilidad π en $B \times B$ que es un emparejamiento óptimo de μ y ν para el coste dado por la p-distancia de Wasserstein.

2.4. Resultados principales

Todos los resultados de esta sección se prueban en [14] en el caso general en que B es un espacio de Banach separable con una norma $\|\cdot\|$. Aquí nos centraremos en el caso particular en que $B = \mathbb{R}^d$, y denotaremos $\Gamma_p = \Gamma_p(\mathbb{R}^d)$.

Proposición 1. *Con las definiciones introducidas antes:*

1. *El inferior en la definición de $d_p(\mu, \nu)$ se alcanza*
2. *d_p es una distancia en Γ_p*

Demostración. Veamos en primer lugar que el inferior se alcanza:

Por los comentarios del principio, basta ver que existe una probabilidad π definida en $\mathbb{R}^d \times \mathbb{R}^d$ tal que:

1. Las probabilidades que inducen las funciones coordenadas Π_X y Π_Y son:

$$\pi \circ \Pi_X^{-1} = \mu \quad \pi \circ \Pi_Y^{-1} = \nu$$

2. $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y)$ sea mínima.

Hemos visto en el lema [1] que siempre existen probabilidades en $\mathbb{R}^d \times \mathbb{R}^d$ que cumplen 1). Luego el inferior se toma en un conjunto no vacío. Además, sabemos que para $1 \leq p < \infty$ existe una constante c_p tal que:

$$\|x+y\|^p \leq c_p(\|x\|^p + \|y\|^p) \quad \forall\, x, y \in \mathbb{R}^d$$

Por tanto:

$$0 \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} (c_p(\|x\|^p + \|y\|^p)) d\pi(x,y) =$$

$$= c_p \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x\|^p d\pi(x,y) + c_p \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y\|^p d\pi(x,y) =$$

$$= c_p \left(\int_{\mathbb{R}^d} \|x\|^p d\mu(x) + \int_{\mathbb{R}^d} \|x\|^p d\nu(x) \right) < \infty$$

Sabemos entonces que el inferior toma un valor finito, y por tanto existe una sucesión $\{\pi_n\}_{n=1}^{\infty}$ de probabilidades que cumplen que

$$\pi_n \circ \Pi_X^{-1} = \mu \quad \pi_n \circ \Pi_Y^{-1} = \nu$$

y tal que:

$$\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi_n(x,y) \right)^{1/p} \downarrow d_p(\mu,\nu)$$

Por tanto, si denotamos U_n, V_n a las variables aleatorias formadas al considerar las funciones coordenadas Π_X, Π_Y sobre el espacio probabilístico $(\mathbb{R}^d \times \mathbb{R}^d, \pi_n)$, tenemos que:

$$\mathscr{L}(U_n) = \mu \quad \mathscr{L}(V_n) = \nu \quad (E(\|U_n - V_n\|^p)^{1/p} \downarrow d_p(\mu,\nu)$$

Para el razonamiento posterior, utilizamos los conceptos y resultados que se detallan en el apéndice 7.2. Tenemos que:

$$\mathscr{L}(U_n) = \mu \quad \forall\, n \in \mathbb{N} \Rightarrow \{U_n\}_{n=1}^{\infty} \text{ ajustada}$$
$$\mathscr{L}(V_n) = \nu \quad \forall\, n \in \mathbb{N} \Rightarrow \{V_n\}_{n=1}^{\infty} \text{ ajustada}$$

Por tanto, la sucesión $\{(U_n, V_n)\}_{n=1}^{\infty}$ es ajustada por serlo sus coordenadas, luego $\{\mathscr{L}(U_n, V_n)\}_{n=1}^{\infty}$=$\{\pi_n\}_{n=1}^{\infty}$ es ajustada. Del teorema de Helly [4] deducimos que existe una probabilidad π y una subsucesión $\{\pi_{n_k}\}_{k=1}^{\infty}$ tal que

$$\pi_{n_k} \xrightarrow{k\to\infty}_d \pi$$

Si ahora denotamos U, V a las variables aleatorias formadas al considerar las funciones coordenadas sobre el espacio probabilístico $(\mathbb{R}^d \times \mathbb{R}^d, \pi)$, por la continuidad de dichas funciones tenemos que:

$$\mu = \pi_{n_k} \circ \Pi_X^{-1} \xrightarrow{k\to\infty}_d \pi \circ \Pi_X^{-1} \quad \Rightarrow \quad \mathscr{L}(U) = \pi \circ \Pi_X^{-1} = \mu$$

$$\nu = \pi_{n_k} \circ \Pi_Y^{-1} \xrightarrow{k\to\infty}_d \pi \circ \Pi_Y^{-1} \quad \Rightarrow \quad \mathscr{L}(V) = \pi \circ \Pi_Y^{-1} = \nu$$

Y además, como $\pi_{n_k} \xrightarrow{k\to\infty}_d \pi$, entonces por la continuidad de la función

$$f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$

definida por $f(x,y) = \|x-y\|^p$ sabemos que:

$$\|U_{n_k} - V_{n_k}\|^p \xrightarrow{k\to\infty}_d \|U-V\|^p$$

Por el teorema de Skorohod [6], sabemos además que existen variables aleatorias positivas $\{S_k\}_{k=1}^{\infty}, S$ definidas en algún espacio probabilístico Ω tales que

$$\mathscr{L}(S_k) = \mathscr{L}\left(\|U_{n_k} - V_{n_k}\|^p\right) \quad \mathscr{L}(S) = \mathscr{L}\left(\|U-V\|^p\right) \quad y \quad S_k \xrightarrow{k\to\infty}_{c.s.} S$$

Se tiene por tanto por el lema de Fatou que:

$$\left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi(x,y)\right)^{1/p} = E(\|U-V\|^p) = E(S) \leq \lim_{k\to\infty} E(S_k) = \lim_{k\to\infty} E(\|U_{n_k} - V_{n_k}\|)^p =$$

$$= \lim_{k\to\infty} \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi_{n_k}(x,y)\right)^{1/p} = d_p(\mu,\nu)$$

Como por definición de $d_p(\mu,\nu)$ se tiene que $d_p(\mu,\nu) \leq (\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi(x,y))^{1/p}$, entonces se tiene necesariamente que:

$$d_p(\mu,\nu) = \left(\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p d\pi(x,y)\right)^{1/p}$$

Vamos a comprobar ahora que $d_p : \Gamma_p \times \Gamma_p \longrightarrow \mathbb{R}$ es una distancia:

- $d(\mu,\nu) = d(\nu,\mu)$, por la simetría de la definición.
- $d(\mu,\nu) = \inf\{(E\|U-V\|^p)^{1/p} : \mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu\} \geq 0$
- $d(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$

  Si $d(\mu,\nu) = 0$, entonces existen v.a. U, V con $\mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu$ tales que $E(\|U-V\|^p) = 0$, y por tanto:

  $$\|U-V\|^p = 0\, c.s. \Rightarrow U = V\, c.s. \Rightarrow U =_d V \Rightarrow \mu = \mathscr{L}(U) = \mathscr{L}(V) = \nu$$

  Si $\mu = \nu$, si U es una v.a con $\mathscr{L}(U) = \mu = \nu$, entonces

  $$0 \leq d_p(\mu,\nu) \leq E(\|U-U\|^p)^{1/p} = 0 \Rightarrow d_p(\mu,\nu) = 0$$

- Desigualdad triangular

  Sean $\mu, \nu, \lambda \in \Gamma_p$. Utilizando el apartado 1, sabemos que existen una probabilidad π en $\mathbb{R}^d \times \mathbb{R}^d$ tal que si Π_X, Π_Y son las funciones coordenadas entonces:

$$\pi \circ \Pi_X^{-1} = \mu, \quad \pi \circ \Pi_Y^{-1} = \nu, \quad d(\mu, \nu) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) \right)^{1/p}$$

  De la misma forma, si ahora denotamos Π_Y, Π_Z a las funciones coordenadas en otro "plano" $\mathbb{R}^d \times \mathbb{R}^d$, sabemos que existe una probabilidad π' en $\mathbb{R}^d \times \mathbb{R}^d$ tal que:

$$\pi' \circ \Pi_Y^{-1} = \nu, \quad \pi' \circ \Pi_Z^{-1} = \lambda, \quad d(\nu, \lambda) = \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - z\|^p d\pi'(y, z) \right)^{1/p}$$

  Juntamos los dos planos sobre el eje Y para formar el espacio $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, y sean Π_X, Π_Y, Π_Z las funciones coordenadas en este espacio. De acuerdo con el lema del pegado [2], sabemos que existe una probabilidad π^* en $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ que verifica las siguientes propiedades:

  - La probabilidad marginal de π^* sobre los dos primeros factores es π, es decir,

$$\pi^* \circ (\Pi_X, \Pi_Y)^{-1} = \pi$$

  - La probabilidad marginal de π^* sobre los dos últimos factores es π', es decir,

$$\pi^* \circ (\Pi_Y, \Pi_Z)^{-1} = \pi'$$

  De aquí se deduce que:

$$\pi^* \circ \Pi_X^{-1} = \mu$$
$$\pi^* \circ \Pi_Y^{-1} = \nu$$
$$\pi^* \circ \Pi_Z^{-1} = \lambda$$

  Utilizando estas propiedades, aplicando la desigualdad de Minkowski en $L^p(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ a las variables aleatorias Π_X, Π_Y, Π_Z, obtenemos la siguiente cadena de desigualdades:

$$d_p(\mu, \lambda) \leq \left(\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|\Pi_X - \Pi_Z\|^p d\pi^* \right)^{1/p} \leq$$

$$\leq \left(\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|\Pi_X - \Pi_Y\|^p d\pi^* \right)^{1/p} + \left(\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|\Pi_Y - \Pi_Z\|^p d\pi^* \right)^{1/p} =$$

$$= \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_X - \Pi_Y\|^p d\pi \right)^{1/p} + \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_Y - \Pi_Z\|^p d\pi' \right)^{1/p} =$$

$$= \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\pi(x, y) \right)^{1/p} + \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|y - z\|^p d\pi'(y, z) \right)^{1/p} = d(\mu, \nu) + d(\nu, \lambda)$$

  □

En el caso en que B es la recta real, podemos calcular d_p de forma sencilla.

Proposición 2. *Si B es la recta real y $\mu, \nu \in \Gamma_p(\mathbb{R})$ vienen dadas por las funciones de distribución F,G, entonces considerando $\|x\| = |x|$ y denotando por F^{-1} y G^{-1} las correspondientes funciones cuantiles:*

$$d_p(\mu, \nu) = \left(\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt \right)^{1/p}$$

Demostración. La demostración que damos para este resultado es una adaptación de la que aparece en [8, Teor. 8.1] para una situación más general. En primer lugar, se hace la demostración en el caso en que las probabilidades determinadas por F y G están concentradas en un conjunto finito $X = \{x_0, x_1, ..., x_n\}$. El conjunto de probabilidades π en $X \times X$ tales que las probabilidades marginales sean μ, ν será por tanto finito, y sabemos entonces que habrá un número finito de v.a. U, V tales que la distribución de $U - V$ sea distinta. Por tanto, sabemos que existen U_0, V_0 tales que $E(|U_0 - V_0|^p)$ sea mínima. Denotamos entonces $p(x, y) = P(U_0 = x, V_0 = y)$.

Podemos asumir la siguiente propiedad: para todo $x_i > x_j$, $y_i > y_j$, siendo $x_i, x_j, y_i, y_j \in X$ se tiene que

$$\text{mín}[p(x_i, y_j), p(x_j, y_i)] = 0 \tag{2}$$

Si no fuera así, existiría algún $x_i > x_j, y_i > y_j$ tales que:

$$p = \text{mín}[p(x_i, y_j), p(x_j, y_i)] > 0$$

Definimos entonces $\tilde{U}_0, \tilde{V}_0$ nuevas variables aleatorias con distribución conjunta $\tilde{p}$ que viene dada por:

$$\begin{aligned}
\tilde{p}(x_i, y_i) &= p(x_i, y_i) + p \\
\tilde{p}(x_j, y_j) &= p(x_j, y_j) + p \\
\tilde{p}(x_i, y_j) &= p(x_i, y_j) - p \\
\tilde{p}(x_j, y_i) &= p(x_j, y_i) - p \\
\tilde{p}(x, y) = p(x, y) \quad &\text{si } x \neq x_i, x_j \text{ o } y \neq y_i, y_j
\end{aligned}$$

Se tiene por tanto que

$$\sum_{y \in X} \tilde{p}(x, y) = \sum_{y \in X} p(x, y) = \mu(x) \quad si\, x \neq x_i, x_j$$

$$\sum_{y \in X} \tilde{p}(x_i, y) = \Big(\sum_{y \neq y_i, y_j} p(x_i, y) \Big) + (p(x_i, y_i) + p) + (p(x_i, y_j) - p) = \sum_{y \in X} p(x_i, y) = \mu(x_i)$$

$$\sum_{y \in X} \tilde{p}(x_j, y) = \Big(\sum_{y \neq y_i, y_j} p(x_j, y) \Big) + (p(x_j, y_i) - p) + (p(x_j, y_j) + p) = \sum_{y \in X} p(x_j, y) = \mu(x_j)$$

Ls distribución marginal de $\tilde{U}_0$ es μ, $\mathscr{L}(\tilde{U}_0) = \mu$, y de forma análoga se ve que la distribución marginal de $\tilde{V}_0$ es ν, $\mathscr{L}(\tilde{V}_0) = \nu$.

De las relaciones:

$$x_j - y_i \leq x_j - y_j \leq x_i - y_j, \quad x_j - y_i \leq x_i - y_i \leq x_i - y_j$$

$$(x_i - y_i) + (x_j - y_j) = (x_i - y_j) + (x_j - y_i)$$

se deduce que existen $\lambda_1, \lambda_2 \in [0,1]$ tales que $\lambda_1 + \lambda_2 = 1$ y además:

$$x_i - y_i = (1-\lambda_1)(x_j - y_i) + \lambda_1(x_i - y_j) \qquad x_j - y_j = (1-\lambda_2)(x_j - y_i) + \lambda_2(x_i - y_j)$$

La función $f(x) = |x|^p$, siendo $p \geq 1$, es una función convexa. De las igualdades anteriores se deduce que:

$$|x_i - y_i|^p \leq (1-\lambda_1)|x_j - y_i|^p + \lambda_1|x_i - y_j|^p \qquad |x_i - y_i|^p \leq (1-\lambda_2)|x_j - y_i|^p + \lambda_2|x_i - y_j|^p$$

Sumando ambas desigualdades:

$$|x_i - y_i|^p + |x_j - y_j|^p \leq |x_i - y_j|^p + |x_j - y_i|^p$$

De aquí se deduce que:

$$E|\tilde{U}_0 - \tilde{V}_0|^p - E|U_0 - V_0|^p = p\left(|x_i - y_i|^p + |x_j - y_j|^p - |x_i - y_j|^p - |x_j - y_i|^p\right) \leq 0$$

$$\Rightarrow E|\tilde{U}_0 - \tilde{V}_0|^p \leq E|U_0 - V_0|^p$$

Por tanto, si la propiedad (2) que habíamos enunciado al principio no se cumple para U_0, V_0, se sustituyen por $\tilde{U}_0, \tilde{V}_0$. Si estas nuevas variables tampoco cumplen (2), podemos repetir el proceso y en un número finito de pasos obtendremos unas nuevas variables U^*, V^* que verifican (2). Además, el mínimo de las $E|U - V|^p$ cuando U, V varían entre las variables aleatorias construidas en cada paso se da para U^*, V^*.

Veamos que la relación (2) determina completamente la distribución conjunta de (U, V). Supongamos que $\mathscr{L}(U) = \mu$ tiene soporte en $z_1, ..., z_m$, con $z_i < z_{i+1}$ y $\mu(z_i) > 0 \; \forall\, i = 1, ..., m$, y $\mathscr{L}(V) = \nu$ tiene soporte en $y_1, ..., y_p$, con $y_i < y_{i+1}$ y $\nu(y_i) > 0 \, \forall\, i = 1, ..., p$. La distribución conjunta de (U, V) viene dada por la matriz:

$$\begin{bmatrix} p(z_1, y_1) & p(z_1, y_2) & \cdots & p(z_1, y_p) \\ p(z_2, y_1) & p(z_2, y_2) & \cdots & p(z_2, y_p) \\ \vdots & \vdots & \ddots & \vdots \\ p(z_m, y_1) & p(z_m, y_2) & \cdots & p(z_m, y_p) \end{bmatrix}$$

Por la relación (2) la matriz necesariamente es de la forma:

$$\begin{bmatrix} p(z_1, y_1) & \cdots & p(z_1, y_{j_1}) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & p(z_2, y_{j_1}) & \cdots & p(z_2, y_{j_2}) & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & p(z_m, y_{j_m}) & \cdots & p(z_m, y_{j_m}) \end{bmatrix}$$

con $1 \leq j_1 \leq j_2 \leq j_3 \leq ... \leq j_m = p$.

- $p(z_1, y_1) > 0$. Si no fuera así, $p(z_1, y_1) = 0$, como $\mu(z_1) = \sum_{i=1}^{p} p(z_1, y_i) > 0$, existe $j \in \{2, ..., p\}$ tal que $p(z_1, y_j) > 0$. Entonces:

$$\min(p(z_1, y_j), p(z_k, y_1)) = 0 \quad \forall\, k = 2, ..., m \Rightarrow$$
$$\Rightarrow p(z_k, y_1) = 0 \quad \forall\, k = 1, 2, ..., m$$

  Por tanto $\nu(y_1) = \sum_{k=1}^{m} p(z_k, y_1) = 0$, absurdo.

- Si $p(z_1, y_j) > 0$ con $j > 2$, entonces $p(z_1, y_k) > 0 \quad \forall\, k = 2, ..., j-1$. Si no fuera así, razonando igual que antes se tendría que $\nu(y_k) = \sum_{i=1}^{m} p(z_i, y_k) = 0$, absurdo.

- Una vez que tenemos que la primera fila de la matriz es necesariamente de la forma

$$[\, p(z_1, y_1) \cdots p(z_1, y_{j_1})\, 0 \cdots 0\,]$$

  de la relación $\min(p(z_1, y_{j_1}), p(z_k, y_l)) = 0 \quad \forall\, l = 1, ..., j_1 - 1, k = 2, ..., m$ se deduce que $p(z_k, y_l) = 0 \quad \forall\, l = 1, ..., j_1 - 1, k = 2, ..., m$

- Para el resto de filas, hacemos un razonamiento análogo al que hemos hecho para la primera fila, y obtenemos el resultado.

Si sabemos que la matriz tiene esa forma, existe una única matriz que además cumple las condiciones

$$\nu(y_k) = \sum_{i=1}^{m} p(z_i, y_k) \quad \forall\, k = 1, 2, ..., p$$
$$\mu(z_k) = \sum_{i=1}^{p} p(z_k, y_i) \quad \forall\, k = 1, 2, ..., m$$

Por tanto, la propiedad (2) determina completamente la distribución conjunta. Si vemos que la transformación cuantil verifica dicha propiedad, habremos probado el resultado para este caso particular. Tenemos:

$$((0,1), \beta_{(0,1)}, \lambda) \xrightarrow{F^{-1}} \mathbb{R} \qquad ((0,1), \beta_{(0,1)}, \lambda) \xrightarrow{G^{-1}} \mathbb{R}$$

F^{-1}, G^{-1} son variables aleatorias con funciones de distribución F y G respectivamente. Si denotamos por π a la distribución conjunta de (F^{-1}, G^{-1}) se tiene que:

$$\pi(x_i, y_j) = P(F^{-1} = x_i, G^{-1} = y_j) = \lambda(\{\omega \in (0,1) : F^{-1}(\omega) = x_i, G^{-1}(\omega) = y_j\})$$
$$\pi(x_j, y_i) = P(F^{-1} = x_j, G^{-1} = y_i) = \lambda(\{\omega \in (0,1) : F^{-1}(\omega) = x_j, G^{-1}(\omega) = y_i\})$$

Supongamos que $\pi(x_i, y_j) > 0$. Entonces se tiene que la medida del conjunto

$$\{\omega \in (0,1) : F^{-1}(\omega) = x_i, G^{-1}(\omega) = y_j\} =$$
$$= \{\omega \in (0,1) : F^{-1}(\omega) = x_i\} \cap \{\omega \in (0,1) : G^{-1}(\omega) = y_j\}$$

es positiva. Necesariamente es intersección de dos conjuntos de medida no nula, que al ser una probabilidad discreta sabemos que son intervalos de la forma $(a_i, b_i]$. Luego como la intersección tiene medida positiva:

$$\{\omega \in (0,1) : F^{-1}(\omega) = x_i, G^{-1}(\omega) = y_j\} = (a, b] \quad 0 \le a < b \le 1$$

Veamos que entonces necesariamente $\pi(x_j, y_i) = 0$

- $w \in (a,b]$
  $\Rightarrow F^{-1}(\omega) = x_i, G^{-1}(\omega) = y_j$
  $\Rightarrow \omega \notin \{\omega \in (0,1) : F^{-1}(\omega) = x_j, G^{-1}(\omega) = y_i\}$

- $\omega > b$
  $\Rightarrow F^{-1}(\omega) \geq F^{-1}(b) = x_i > x_j$
  $\Rightarrow \omega \notin \{\omega \in (0,1) : F^{-1}(\omega) = x_j, G^{-1}(\omega) = y_i\}$

- $\omega \leq a$
  $\Rightarrow G^{-1}(\omega) \leq G^{-1}(b) = y_j < y_i$
  $\Rightarrow \omega \notin \{\omega \in (0,1) : F^{-1}(\omega) = x_j, G^{-1}(\omega) = y_i\}$

$\Rightarrow \pi(x_j, y_i) = \lambda(\{\omega \in (0,1) : F^{-1}(\omega) = x_j, G^{-1}(\omega) = y_i\}) = 0$. Si suponemos que $\pi(x_j, y_i) > 0$, un razonamiento análogo nos lleva a que $\pi(x_i, y_j) = 0$. La transformación cuantil cumple por tanto la propiedad (2).

Veamos que ocurre en el caso general en que $\mu, \nu \in \Gamma_p(\mathbb{R})$, con funciones de distribución F y G respectivamente. En este caso, sabemos que existen variables aleatorias U, V que verifican:

$$d_p^p(\mu,\nu) = \text{ínf}\{E|X-Y|^p : \mathscr{L}(X) = \mu, \mathscr{L}(Y) = \nu\} = E|U-V|^p$$

Para cada $n \in \mathbb{N}$ definimos las funciones $\phi_n : \mathbb{R}^+ \to \mathbb{R}^+$ por:

$$\phi_n(x) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{[(k-1)/2^n, k/2^n)}(x) + nI_{[n,\infty)}$$

y $\psi_n : \mathbb{R} \to \mathbb{R}$ por:

$$\psi_n(x) = \begin{cases} -\phi_n(-x) & si \quad x < 0 \\ \\ \phi_n(x) & si \quad x \geq 0 \end{cases}$$

Considerando entonces las variables $U_n = \psi_n(U)$ y $V_n = \psi_n(V)$, tenemos dos sucesiones de variables aleatorias simples que verifican que:

$$U_n \to_{c.s.} U \quad \text{y} \quad |U_n| \uparrow |U| \, c.s.$$

$$V_n \to_{c.s.} V \quad \text{y} \quad |V_n| \uparrow |V| \, c.s.$$

Por tanto, para cada $n \in \mathbb{N}$ se tiene que:

$$|U_n - V_n|^p \leq 2^p \left(|U_n|^p + |V_n|^p\right) \leq 2^p \left(|U|^p + |V|^p\right)$$

y además $|U_n - V_n|^p \to_{c.s.} |U-V|^p$. Por el teorema de la convergencia dominada:

$$E|U-V|^p = \lim_{n\to\infty} E|U_n - V_n|^p$$

Sean entonces para cada $n \in \mathbb{N}$, $\mu_n = \mathscr{L}(U_n)$ y $\nu_n = \mathscr{L}(V_n)$, que son probabilidades discretas, y sean F_n, G_n respectivamente sus funciones de distribución. Como $U_n \to_{c.s.} U$ y $V_n \to_{c.s.} V$,

entonces sabemos que $\mu_n \to_d \mu$ y $\nu_n \to_d \nu$, y por el teorema de Skorohod [6] podemos afirmar que:

$$F_n^{-1} \to_{c.s.} F^{-1} \qquad G_n^{-1} \to_{c.s.} G^{-1}$$

luego $|F_n^{-1} - G_n^{-1}|^p \to_{c.s.} |F^{-1} - G^{-1}|^p$, y entonces por el lema de Fatou:

$$\int_0^1 |F^{-1}(t) - G^{-1}(t)|^p dt \leq \liminf_{n\to\infty} \int_0^1 |F_n^{-1}(t) - G_n^{-1}(t)|^p dt$$

Además, por lo que hemos probado en el caso discreto sabemos que:

$$\int_{\mathbb{R}} |F_n^{-1}(t) - G_n^{-1}(t)|^p dt = d_p^p(\mu_n, \nu_n) \leq E|U_n - V_n|^p$$

Juntando todo lo que hemos visto tenemos la siguiente cadena de desigualdades:

$$\int_{\mathbb{R}} |F^{-1}(t) - G^{-1}(t)|^p dt \leq \liminf_{n\to\infty} \int_{\mathbb{R}} |F_n^{-1}(t) - G_n^{-1}(t)|^p dt \leq$$

$$\leq \liminf_{n\to\infty} E|U_n - V_n|^p = E|U - V|^p = d_p^p(\mu, \nu)$$

La otra desigualdad es consecuencia de que F^{-1}, G^{-1} son variables aleatorias con $\mathscr{L}(F^{-1}) = \mu, \mathscr{L}(G^{-1}) = \nu$, y por tanto:

$$d_p^p(\mu, \nu) = \text{ínf}\{E|X - Y|^p : \mathscr{L}(X) = \mu, \mathscr{L}(Y) = \nu\} \leq$$

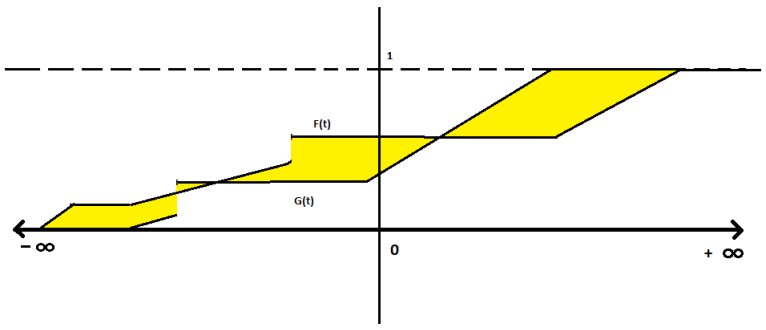
$$\leq E|F^{-1} - G^{-1}|^p = \int_{\mathbb{R}} |F^{-1}(t) - G^{-1}(t)|^p dt$$

□

Corolario 1. *Si B es la recta real, además p=1 y μ, ν probabilidades en $\mathbb{R}$ que vienen dadas por las funciones de distribución F,G, entonces considerando $\|x\| = |x|$,*

$$d_p(\mu, \nu) = \int_0^1 |F^{-1}(t) - G^{-1}(t)| dt = \int_{-\infty}^{\infty} |F(x) - G(x)| dx$$

Demostración. La primera igualdad es consecuencia directa de la proposición anterior. Para ver la segunda igualdad, basta ver ambas integrales representan el área entre los grafos de F y G.



- Tomando secciones verticales del conjunto de $\mathbb{R}^2$ delimitado por las gráficas de F y G, para cada $x \in \mathbb{R}$ la sección es un intervalo de longitud $|F(x) - G(x)|$, luego el área es:

$$\int_{-\infty}^{\infty} |F(x) - G(x)| dx$$

- Tomando secciones horizontales, para cada $t \in (0, 1)$ se tiene que la medida de la sección horizontal es $|F^{-1}(t) - G^{-1}(t)|$, excepto en un conjunto numerable y por tanto de medida nula, correspondientes a las discontinuidades de F^{-1}, G^{-1}. Por tanto, el área es:

$$\int_0^1 |F^{-1}(t) - G^{-1}(t)| dt$$

□

Proposición 3. *Sean $\mu_n, \mu \in \Gamma_p$. Son equivalentes las siguientes afirmaciones:*

a) $d_p(\mu_n, \mu) \xrightarrow{n\to\infty} 0$

b) $\mu_n \xrightarrow{n\to\infty} \mu$ *débilmente y* $\int \|x\|^p d\mu_n(x) \xrightarrow{n\to\infty} \int \|x\|^p d\mu(x)$

c) $\mu_n \xrightarrow{n\to\infty} \mu$ *débilmente y* $\|x\|^p$ *es uniformemente* μ_n*-integrable*

Demostración. Veamos las equivalencias:

- $a \Rightarrow b$
  Supongamos que U_n son v.a. con $\mathscr{L}(U_n) = \mu_n$ y U es una v.a. con $\mathscr{L}(U) = \mu$, tales que

$$d_p(\mu_n, \mu) = E(\|U_n - U\|^p)^{1/p}$$

  La generalización del lema del pegado a un número infinito numerable de factores nos permite asegurar la existencia de dichas variables aleatorias. Con la notación de dicho lema, basta tomar como factores $(\mathbb{R}^d, \beta^d, \mu_n)$ para todo $n = 0, 1, 2, ...$, siendo $\mu_0 = \mu$, y probabilidades P_{0n} que sean emparejamientos óptimos de μ y μ_n. Sabemos entonces que existe una probabilidad P en $\prod_{n=0}^{\infty} \mathbb{R}^d$ tal que sus probabilidades marginales sobre los factores 0 y n son P_{0n} para todo $n \geq 1$. Tomando entonces como variable U la proyección sobre el factor 0 y como variable U_n la proyección sobre el factor n, para cada $n \geq 1$, y se cumple lo que queríamos.

  Por la segunda desigualdad triangular de la norma $\|\cdot\|_p$:

$$|(\int \|x\|^p d\mu_n(x))^{1/p} - (\int \|x\|^p d\mu(x))^{1/p}| = |E(\|U_n\|^p)^{1/p} - E(\|U\|^p)^{1/p}| \leq$$

$$\leq E(\|U_n - U\|^p)^{1/p} = d_p(\mu_n, \mu) \xrightarrow{n\to\infty} 0$$

  y por tanto, por la continuidad de la función $f(x) = x^p$ se tiene que:

$$\int \|x\|^p d\mu_n(x) = f\left((\int \|x\|^p d\mu_n(x))^{1/p}\right) \xrightarrow{n\to\infty} f\left((\int \|x\|^p d\mu(x))^{1/p}\right) = \int \|x\|^p d\mu(x)$$

  Veamos ahora que la sucesión $\{U_n\}_{n=1}^{\infty}$ es ajustada. Sea $\epsilon > 0$

- Hemos visto que

$$E(\|U_n\|^p) = \int \|x\|^p d\mu_n(x) \xrightarrow{n\to\infty} \int \|x\|^p d\mu(x) = E(\|U\|^p)$$

  Por tanto, existe $n_0 \in \mathbb{N}$ tal que

$$E(\|U_n\|^p) \leq 2 \cdot E(\|U\|^p) < \infty \quad \forall\, n \geq n_0$$

- Por la desigualdad de Markov,

$$P(\|U_n\| > \alpha) \leq \frac{E(\|U_n\|^p)}{\alpha^p} \leq \frac{2 \cdot E(\|U\|^p)}{\alpha^p} \xrightarrow{\alpha\to\infty} 0$$

- Por tanto, dado $\epsilon > 0$, existe $R > 0$ tal que

$$P(\|U_n\| > R) < \epsilon \quad \forall\, n \geq n_0$$

  y por tanto la sucesión es ajustada.

Sea f una función Lipschiziana, $|f(x) - f(y)| \leq K\|x - y\|$. Se verifica que:

$$|\int f(x)d\mu_n(x) - \int f(x)d\mu(x)| = |E(f(U_n)) - E(f(U))| \leq E|f(U_n) - f(U)| \leq$$

$$\leq K \cdot E\|U_n - U\| \leq K \cdot (E\|U_n - U\|^p)^{1/p} \xrightarrow{n\to\infty} 0$$

donde la última desigualdad es consecuencia de la desigualdad de Jensen aplicada a la función $f(x) = x^p$. Si vemos que las funciones Lipschizianas son una clase separante, entonces por la proposición [19] se tendrá que $\mu_n \xrightarrow{n\to\infty} \mu$ débilmente. Dada la función

$$\phi(t) = \begin{cases} 1 & si \quad t \leq 0 \\ 1 - t & si \quad 0 \leq t \leq 1 \\ 0 & si \quad t \geq 1 \end{cases}$$

sabemos que el conjunto de funciones definidas por $f(x) = \phi(\frac{1}{\epsilon}d(x, C))$, donde C es un cerrado de $\mathbb{R}^d$ y $\epsilon > 0$, forman una clase separante. Estas funciones son Lipschizianas, ya que $\forall\, x, y \in \mathbb{R}^d$ se puede probar que se cumple la desigualdad

$$|f(x) - f(y)| \leq \frac{1}{\epsilon}\|x - y\|$$

Por tanto las funciones Lipschizianas también forman una clase separante.

Para la siguiente equivalencia vamos a utilizar las equivalencias de la proposición de **caracterización de la convergencia en los espacios** $L_p(\Omega)$ [20] que aparece en el apéndice 7.2.

- $b \Leftrightarrow c$
  En esta situación, sabemos por el teorema de Skorohod [6] que existe un espacio probabilístico (Ω, σ, P) y variables aleatorias U y U_n para cada $n \in \mathbb{N}$ tales que

  $$\mathscr{L}(U_n) = \mu_n, \mathscr{L}(U) = \mu \text{ y además } U_n \xrightarrow{c.s.} U$$

  La equivalencia se deduce entonces de las equivalencias 2) y 3) de la proposición [20] y de que:

  - La convergencia c.s. implica la covergencia en probabilidad.
  - La integrabilidad uniforme de la función $\|x\|^p$ respecto de las probabilidades μ_n es equivalente a la integrabilidad uniforme de la sucesión de variables aleatorias $\{\|U_n\|^p\}_{n=1}^{\infty}$
  - $E\|U_n\|^p = \int \|x\|^p d\mu_n(x)$ y $E\|U\|^p = \int \|x\|^p d\mu(x)$

- $b \Rightarrow a$
  Recurriendo a la misma representación que antes de las v.a. U y U_n, estas variables cumplen que $U_n \xrightarrow{c.s.} U$ y además:

  $$E\|U_n\|^p = \int \|x\|^p d\mu_n(x) \xrightarrow{n\to\infty} \int \|x\|^p d\mu(x) = E\|U\|^p$$

  Por la equivalencia entre las afirmaciones 1) y 2) de la proposición [20], sabemos que:

  $$U_n \xrightarrow{L_p(\Omega)} U \Rightarrow (E\|U_n - U\|^p)^{1/p} \xrightarrow{n\to\infty} 0$$

  Por tanto:

  $$d_p(\mu_n, \mu) = \text{ínf}\{(E\|X - Y\|^p)^{1/p} : \mathscr{L}(X) = \mu_n, \mathscr{L}(Y) = \mu\} \leq$$

  $$\leq (E\|U_n - U\|^p)^{1/p} \xrightarrow{n\to\infty} 0$$

□

Nota 1. La implicación $a \Rightarrow b$ también se puede hacer de forma sencilla a partir de la caracterización de la convergencia en $L_p(\Omega)$. Basta ver que las v.a. U_n, U que se construyeron en ese apartado cumplen que:

$$(E\|U_n - U\|^p)^{1/p} = d_p(\mu_n, \mu) \xrightarrow{n\to\infty} 0$$

y por tanto $U_n \xrightarrow{L_p(\Omega)} U$, y aplicar las equivalencias.

Corolario 2. *Sean $\{U_n\}_{n=1}^{\infty}$ variables aleatorias definidas en (Ω, σ, P) que toman valores en $\mathbb{R}^d$, independientes e igualmente distribuidas con distribución $\mu \in \Gamma_p$. Sea μ_n la distribución empírica de las variables $U_1, U_2, ..., U_n$. Entonces:*

$$d_p(\mu_n, \mu) \xrightarrow{n\to\infty} 0$$

Demostración. Denotamos para cada $\omega \in \Omega$ por μ_n^ω a la probabilidad que da peso $\frac{1}{n}$ a cada valor $U_1(\omega), U_2(\omega), ..., U_n(\omega)$. La función de distribución de esta probabilidad es:

$$F_n^\omega(x_1, x_2, ..., x_d) = \frac{1}{n} \sum_{i=1}^{n} I_{\prod_{j=1}^{d}(-\infty, x_j]}(U_i(\omega))$$

Las variables $I_{\prod_{j=1}^{d}(-\infty, x_j]}(U_i(\omega))$ son independientes e igualmente distribuidas. Por tanto, si F es la función de distribución asociada a μ, por la ley fuerte de los grandes números:

$$F_n^\omega(x_1, x_2, ..., x_d) \xrightarrow{n\to\infty} E\left(I_{\prod_{j=1}^{d}(-\infty, x_j]}(U_1)\right) = \mu\left(\prod_{j=1}^{d}(-\infty, x_j]\right) = F(x_1, x_2, ..., x_d)$$

para casi todo $\omega \in \Omega$. Sea Ω_x el conjunto de probabilidad 1 en el que se cumple dicha convergencia, que depende del valor de $x = (x_1, ..., x_d)$ escogido. En particular, para cada $q = (q_1, ..., q_d) \in \mathbb{Q}^d$, existe un conjunto Ω_q de probabilidad 1 tal que $F_n^\omega(q) \xrightarrow{n\to\infty} F(q)$ para cada $\omega \in \Omega_q$. Tomando $\Omega' = \bigcap_{q\in\mathbb{Q}^n} \Omega_q$, tenemos un conjunto de probabilidad 1 en el cual:

$$F_n^\omega(q) \xrightarrow{n\to\infty} F^\omega(q) \quad \forall\, q \in \mathbb{Q}^d$$

Por tanto, para cada $\omega \in \Omega'$ se tendrá que:

$$F_n^\omega(x) \xrightarrow{n\to\infty} F^\omega(x) \quad \forall\, x \in C(F)$$

siendo $C(F)$ el conjunto de puntos de continuidad de la función F, y por tanto $\mu_n^\omega \to_d \mu$ para todo $\omega \in \Omega'$. Por otra parte, de nuevo por la ley fuerte de los grandes números:

$$\int \|x\|^p d\mu_n^\omega(x) = \frac{\|U_1(\omega)\|^p + \|U_2(\omega)\|^p + ... + \|U_n(\omega)\|^p}{n} \xrightarrow{n\to\infty} E\|U_1\|^p = \int \|x\|^p d\mu(x)$$

para todo ω en un conjunto Ω'' de probabilidad 1 en Ω, ya que las variables $\{\|U_n\|^p\}_{n=1}^{\infty}$ también son independientes e igualmente distribuidas. Por tanto, si ω está en el conjunto de probabilidad 1 definido por $\Omega' \cap \Omega''$, se cumple que;

$$\mu_n^\omega \to_d \mu \text{ y } \int \|x\|^p d\mu_n^\omega(x) \xrightarrow{n\to\infty} \int \|x\|^p d\mu(x)$$

Y por la equivalencia $b \Rightarrow a$ de la proposición anterior, podemos concluir que $d_p(\mu_n^\omega, \mu) \to 0$ para todo $\omega \in \Omega' \cap \Omega''$, es decir, casi seguro. □

Definición 5. *En el caso en que tenemos dos variables aleatorias U, V que toman valores en $\mathbb{R}^d$, definimos $d_p(U, V)$ como la distancia entre las leyes de U y V, suponiendo que ambas están en Γ_p.*

Proposición 4. *Se cumplen las siguientes propiedades:*

1. *Si $a \in \mathbb{R}$, entonces $d_p(aU, aV) = |a| d_p(U, V)$*
2. *$d_p(LU, LV) \leq \|L\| d_p(U, V)$ para cualquier aplicación lineal $L : \mathbb{R}^d \to \mathbb{R}^d$*

Demostración. Veamos en primer lugar (1):
Sea $\mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu$, y sea π una probabilidad en $\mathbb{R}^d \times \mathbb{R}^d$ tal que si Π_X, Π_Y son las funciones coordenadas en $\mathbb{R}^d \times \mathbb{R}^d$, se cumple que:

$$\pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu$$

Se tiene entonces que si $a \neq 0$:

$$\mathscr{L}(\Pi_X) = \mathscr{L}(U) \quad \Leftrightarrow \quad \mathscr{L}(a\Pi_X) = \mathscr{L}(aU)$$
$$\mathscr{L}(\Pi_Y) = \mathscr{L}(V) \quad \Leftrightarrow \quad \mathscr{L}(a\Pi_Y) = \mathscr{L}(aV)$$

Por tanto:

$$d_p(aU, aV) =$$
$$= \text{ínf}\left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|a\Pi_X - a\Pi_Y\|^p d\pi \right)^{1/p} : \pi \text{ prob. en } \mathbb{R}^d \times \mathbb{R}^d,\ \pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu \right\} =$$
$$= \text{ínf}\left\{ |a| \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_X - \Pi_Y\|^p d\pi \right)^{1/p} : \pi \text{ prob. en } \mathbb{R}^d \times \mathbb{R}^d \quad \pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu \right\} =$$
$$= |a| d_p(U, V)$$

Si $a = 0$ la igualdad es obvia. Veamos (2):

Al igual que antes, sea $\mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu$, y sea π una probabilidad en $\mathbb{R}^d \times \mathbb{R}^d$ tal que si Π_X, Π_Y son las funciones coordenadas en $\mathbb{R}^d \times \mathbb{R}^d$, se cumple que:

$$\pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu$$

Se tiene entonces que:

$$\mathscr{L}(\Pi_X) = \mathscr{L}(U) \quad \Rightarrow \quad \mathscr{L}(L\Pi_X) = \mathscr{L}(LU)$$
$$\mathscr{L}(\Pi_Y) = \mathscr{L}(V) \quad \Rightarrow \quad \mathscr{L}(L\Pi_Y) = \mathscr{L}(LV)$$

Por tanto:

$$d_p(LU, LV) \leq$$
$$\leq \text{ínf}\left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|L\Pi_X - L\Pi_Y\|^p d\pi \right)^{1/p} : \pi \text{ prob. en } \mathbb{R}^d \times \mathbb{R}^d \quad \pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu \right\} =$$
$$= \text{ínf}\left\{ \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|L(\Pi_X - \Pi_Y)\|^p d\pi \right)^{1/p} : \pi \text{ prob. en } \mathbb{R}^d \times \mathbb{R}^d \quad \pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu \right\} \leq$$
$$\leq \text{ínf}\left\{ \|L\| \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\Pi_X - \Pi_Y\|^p d\pi \right)^{1/p} : \pi \text{ prob. en } \mathbb{R}^d \times \mathbb{R}^d \quad \pi \circ \Pi_X^{-1} = \mu,\ \pi \circ \Pi_Y^{-1} = \nu \right\} =$$
$$= \|L\| d_p(U, V)$$

□

En el siguiente resultado, utilizamos la notación $\|\cdot\|_{\mathbb{R}^d}$ y $\|\cdot\|_{\mathbb{R}^{d'}}$ para referirnos a la norma euclídea usual en en $\mathbb{R}^d$ y $\mathbb{R}^{d'}$ respectivamente, y poder así especificar en cada caso en qué espacio estamos trabajando.

Proposición 5. *Sean $d, d' \in \mathbb{N}$, y sea $1 \leq p, p' < \infty$.*
Supongamos que $\{U_n\}_{n=1}^{\infty}$ es una sucesión de v.a. con $\mathscr{L}(U_n) \in \Gamma_p(\mathbb{R}^d)$, y U es otra v.a. que cumple lo mismo y tal que

$$d_p(U_n, U) \xrightarrow{n\to\infty} 0$$

Supongamos que $\Phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ es una función continua tal que existe una constante K que cumple

$$\|\Phi(x)\|_{\mathbb{R}^{d'}}^{p'} \leq K(1 + \|x\|_{\mathbb{R}^d}^{p}) \quad \forall x \in \mathbb{R}^d$$

Entonces $d_{p'}(\Phi(U_n), \Phi(U)) \xrightarrow{n\to\infty} 0$

Demostración. Denotemos $\mu_n = \mathscr{L}(U_n), \mu = \mathscr{L}(U)$, $\nu_n = \mathscr{L}(\Phi(U_n))$ y $\nu = \mathscr{L}(\Phi(U))$
Por la proposición 3 sabemos que como $d_p(U_n, U) \xrightarrow{n\to\infty} 0$, entonces $\mu_n \xrightarrow{d} \mu$ y $\|x\|_{\mathbb{R}^d}^{p}$ es uniformemente μ_n-integrable.

Como Φ es continua, sabemos que $\nu_n = \mu_n \circ \Phi^{-1} \xrightarrow{d} \mu \circ \Phi^{-1} = \nu$. Veamos que $\|x\|_{\mathbb{R}^{d'}}^{p'}$ es uniformemente ν_n-integrable. Sabemos que:

$$\lim_{a\to\infty} \left(\sup_{n\in\mathbb{N}} \int_{\{\|x\|_{\mathbb{R}^d} > a\}} \|x\|_{\mathbb{R}^d}^{p} d\mu_n(x) \right) = 0$$

Por tanto:

$$\lim_{a\to\infty} \left(\sup_{n\in\mathbb{N}} \int_{\{\|x\|_{\mathbb{R}^{d'}} > a\}} \|x\|_{\mathbb{R}^{d'}}^{p'} d\nu_n(x) \right) = \lim_{a\to\infty} \left(\sup_{n\in\mathbb{N}} \int_{\{\|\Phi(x)\|_{\mathbb{R}^{d'}} > a\}} \|\Phi(x)\|_{\mathbb{R}^{d'}}^{p'} d\mu_n(x) \right) = [\boxtimes]$$

Si $\|\Phi(x)\|_{\mathbb{R}^{d'}} > a \Rightarrow K(1 + \|x\|_{\mathbb{R}^d}^{p}) \geq \|\Phi(x)\|_{\mathbb{R}^{d'}}^{p'} > a^{p'} \Rightarrow \|x\|_{\mathbb{R}^d}^{p} > \frac{a^{p'}-1}{K} \xrightarrow{a\to\infty} \infty$

$$\Rightarrow \|x\|_{\mathbb{R}^d} > (\frac{a^{p'}-1}{K})^{1/p}$$

Entonces:

$$[\boxtimes] \leq \lim_{a\to\infty} \left(\sup_{n\in\mathbb{N}} \int_{\{\|x\|_{\mathbb{R}^d} > (\frac{a^{p'}-1}{K})^{1/p}\}} K(1 + \|x\|_{\mathbb{R}^d}^{p}) d\mu_n(x) \right) =$$

$$= \lim_{b\to\infty} \left(\sup_{n\in\mathbb{N}} \int_{\{\|x\|_{\mathbb{R}^d} > b\}} K(1 + \|x\|_{\mathbb{R}^d}^{p}) d\mu_n(x) \right) \leq$$

$$\leq 2K \cdot \lim_{b\to\infty} \left(\sup_{n\in\mathbb{N}} \int_{\{\|x\|_{\mathbb{R}^d} > b\}} \|x\|_{\mathbb{R}^d}^{p} d\mu_n(x) \right) = 0$$

Aplicando la proposición 3 de nuevo, podemos concluir que:

$$d_{p'}(\Phi(U_n), \Phi(U)) \xrightarrow{n\to\infty} 0$$

□

En general, $d_{p'}(\Phi(U_n), \Phi(U))$ no se puede acotar por ninguna función de $d_p(U_n, U)$. Si tomamos $d = d' = 1, p = 2, p' = 1$, y $\Phi(x) = x^2$, entonces tomando las variables $U_n = u_n$ y $V_n = v_n$ siendo $\{u_n\}_{n=1}^{\infty}, \{v_n\}_{n=1}^{\infty}$ sucesiones de números reales con:

$$(u_n - v_n)^2 \xrightarrow{n\to\infty} 0 \quad y \quad |u_n^2 - v_n^2| \xrightarrow{n\to\infty} \infty$$

entonces se tiene que:

$$d_2(X_n, Y_n) \xrightarrow{n\to\infty} 0 \quad y \quad d_1(\Phi(X_n), \Phi(Y_n)) \xrightarrow{n\to\infty} \infty$$

Para conseguir una sucesión $\{u_n\}_{n=1}^{\infty}, \{v_n\}_{n=1}^{\infty}$ que cumpla la condición pedida, basta tomar $x_n = n^2, y_n = n^2 - 1/n$

$$x_n - y_n = \frac{1}{n} \quad \Rightarrow \quad (x_n - y_n)^2 = \frac{1}{n^2} \xrightarrow{n\to\infty} 0$$

$$|u_n^2 - v_n^2| = |n^4 - (n^4 + 2n + \frac{1}{n^2})| = 2n + \frac{1}{n^2} \xrightarrow{n\to\infty} \infty$$

Proposición 6. *Sean $\{U_j\}_{j=1}^m$ variables aleatorias independientes, lo mismo para $\{V_j\}_{j=1}^m$, y supongamos que las leyes de todas las variables aleatorias están en Γ_p. Entonces:*

$$d_p(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j) \leq \sum_{j=1}^m d_p(U_j, V_j)$$

Demostración. Asumimos sin pérdida de generalidad que las parejas $\{(U_j, V_j)\}_{j=1}^m$ son independientes y además:

$$d_p(U_j, V_j) = E(\|U_j - V_j\|^p)^{1/p}$$

Esto se puede hacer ya que:

- En primer lugar, al introducir el concepto de emparejamiento, vimos que dadas dos variables aleatorias U, V definidas en Ω y Ω' respectivamente, se pueden tomar (U, V) definida en $\Omega \times \Omega'$ tal que U, V sean independientes. Estamos repitiendo este proceso para las variables $W_1 = (U_1, V_1), W_2 = (U_2, V_2), ..., W_m = (U_m, V_m)$
- En la demostración de la proposición 1 vimos que se pueden tomar U_j, V_j tales que

$$d_p(U_j, V_j) = E(\|U_j - V_j\|^p)^{1/p}$$

Consideramos entonces la distribución generada en $\mathbb{R}^d \times \mathbb{R}^d$ por:

$$(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j) = \sum_{j=1}^m (U_j, V_j)$$

Cumple que las distribuciones marginales sobre $\mathbb{R}^d$ son las generadas por $\sum_{j=1}^m U_j$ y $\sum_{j=1}^m V_j$, siendo $\{U_j\}_{j=1}^m$ y $\{V_j\}_{j=1}^m$ variables independientes, ya que al ser $\{(U_j, V_j)\}_{j=1}^m$ independientes, también lo son las variables obtenidas al componer con las proyecciones.

Hemos obtenido por tanto una distribución concreta en $\mathbb{R}^d \times \mathbb{R}^d$ que cumple que las distribuciones marginales son $\sum_{j=1}^m U_j$ y $\sum_{j=1}^m V_j$. Por tanto, por la desigualdad de Minkowski:

$$d_p(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j) \leq E\left(\|\sum_{j=1}^m U_j - \sum_{j=1}^m V_j\|^p\right)^{1/p} =$$

$$= E\left(\|\sum_{j=1}^m (U_j - V_j)\|^p\right)^{1/p} \leq \sum_{j=1}^m (E(\|U_j - V_j\|^p))^{1/p} = \sum_{j=1}^m d_p(U_j, V_j)$$

□

2.5. 2-distancia de Wasserstein

Los resultados que expondremos a partir de ahora se centran en el caso más importante, que es cuando p=2. Las proposiciones 7, 8 y 9 aparecen en [14], en los lemas 8.7, 8.8 y 8.9.

El siguiente resultado es una mejora de la proposición anterior en la que aprovechamos el hecho de que la norma $\|\cdot\|_2$ en $\mathbb{R}^d$ sea la norma asociada al producto escalar habitual $<,>$. De hecho, las proposiciones 7 y 8 se pueden generalizar al caso en que B es un espacio de Hilbert cualquiera (así es como se enuncian en [14]).

Proposición 7. *Sean $\{U_j\}_{j=1}^m$ variables aleatorias independientes que toman valores en $\mathbb{R}^d$, lo mismo para $\{V_j\}_{j=1}^m$, y supongamos que las leyes de todas las variables aleatorias están en Γ_2. Supongamos además que $E(U_j) = E(V_j)$ para todo $j = 1, ..., m$. Entonces:*

$$d_2^2(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j) \leq \sum_{j=1}^m d_2^2(U_j, V_j)$$

Demostración. Hacemos la misma construcción que hicimos en la proposición anterior. Se tiene que:

- Si $k = j$

$$E(< U_j - V_j, U_k - V_k >) = E(\|U_j - V_j\|^2) = d_2(U_j, V_j)^2$$

- Si $k \neq j$

$$E(< U_j - V_j, U_k - V_k >) =$$
$$= E(< U_j, U_k > - < U_j, V_k > - < V_j, U_k > + < V_j, V_k >) =$$
$$= E(< U_j, U_k >) - E(< U_j, V_k >) - E(< V_j, U_k >) + E(< V_j, V_k >) =$$
$$=< E(U_j), E(U_k) > - < E(U_j), E(V_k) > - < E(V_j), E(U_k) > + < E(V_j), E(V_k) >=$$
$$=< E(U_j), E(U_k) > - < E(U_j), E(U_k) > - < E(U_j), E(U_k) > + < E(U_j), E(V_k) >= 0$$

Por tanto:

$$d_2^2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right) \leq E\left(\|\sum_{j=1}^m U_j - \sum_{j=1}^m V_j\|^2\right) =$$

$$= E\left(< \sum_{j=1}^m (U_j - V_j), \sum_{k=1}^m (U_k - V_k) >\right) = \sum_{k=1,j=1}^m E(< U_j - V_j, U_k - V_k >) =$$

$$= \sum_{j=1}^m E(< U_j - V_j, U_j - V_j >) = \sum_{j=1}^m d_2^2(U_j, V_j)$$

□

Proposición 8. *Sean U, V variables aleatorias que toman valores en $\mathbb{R}^d$ y tales que $U, V \in \Gamma_2$. Entonces:*

$$d_2^2(U, V) = d_2^2(U - E(U), V - E(V)) + \|E(U) - E(V)\|^2$$

Demostración. Sea $a = E(U), b = E(V)$. Sabemos que podemos escoger las v.a. U, V tales que:

$$E(\|U - V\|^2) = d_2^2(U, V)$$

Se tiene entonces que:

$$E(\|(U - a) - (V - b)\|^2) = E(\|(U - V) - (a - b)\|^2) =$$

$$= E(\|U-V\|^2+\|a-b\|^2-2 < U-V, a-b >) = E(\|U-V\|^2)+E(\|a-b\|^2)-2E(< U-V, a-b >) =$$

$$= E(\|(U - V)\|^2) - \|(a - b)\|^2$$

ya que como $a - b$ es constante, $U - V$ y $a - b$ son independientes y por tanto:

$$E(< U - V, a - b >) =< E(U - V), E(a - b) >=< a - b, a - b >= \|a - b\|^2$$

Tenemos por tanto la desigualdad:

$$d_2^2(U - a, V - b) \leq E(\|(U - a) - (V - b)\|^2) = d_2^2(U, V) - \|a - b\|^2$$

$$\Rightarrow d_2^2(U, V) \geq d_2^2(U - E(U), V - E(V)) + \|a - b\|^2$$

Para ver la otra desigualdad, sean U, V tales que

$$E(\|(U - a) - (V - b)\|^2) = d_2^2(U - a, V - b)$$

Se tiene entonces que:

$$d_2^2(U, V) \leq E(\|U-V\|^2) = E(\|(U-a)-(V-b)\|^2)+\|a-b\|^2 = d_2^2(U-E(U), V-E(V))+\|a-b\|^2$$

y tenemos por tanto la igualdad. □

Proposición 9. *Denotemos a la 2- distancia de Wasserstein en $\mathbb{R}^k$ por $d_{2,k}$, para cada $k \geq 1$. Sean $U_1, U_2, ..., U_n$ variables aleatorias independientes e igualmente distribuidas en $L_2(\mathbb{R})$, y sea $U = (U_1, U_2, ..., U_n)^T$. Lo mismo para $V_1, V_2, ..., V_n$ y $V = (V_1, V_2, ..., V_n)^T$. Supongamos además que $E(U_i) = E(V_i)$. Sea ahora A una matriz $m \times n$ de escalares. AU, AV son vectores aleatorios en $\mathbb{R}^m$. Entonces:*

$$d_{2,m}^2(AU, AV) \leq \mathrm{tr}(AA^T) d_{2,1}^2(U_i, V_i)$$

Demostración. Como hemos hecho habitualmente en las demostraciones anteriores, supongamos que las parejas (U_i, V_i) son independientes y que

$$E(|U_i - V_i|^2)^{1/2} = d_{2,1}(U_i, V_i)$$

Teniendo en cuenta que $\mathrm{tr}(CD) = \mathrm{tr}(DC)$ siempre que ambos productos tengan sentido, llegamos a:

$$d_{2,m}^2(AU, AV)^2 \leq E(\|AU - AV\|^2) = E(\|A(U - V)\|^2) =$$

$$= E(\mathrm{tr}(A(U-V)(U-V)^T A^T)) = E(\mathrm{tr}((U-V)(U-V)^T A^T A)) = \mathrm{tr}(E((U-V)(U-V)^T A^T A)) =$$

$$= \mathrm{tr}(E((U - V)(U - V)^T) A^T A) = [\boxtimes]$$

Se tiene además que:

$$E((U - V)(U - V)^T) =$$

$$= \begin{pmatrix} E((U_1 - V_1)^2) & E((U_1 - V_1)(U_2 - V_2)) & \cdots & E((U_1 - V_1)(U_n - V_n)) \\ E((U_2 - V_2)(U_1 - V_1)) & E((U_2 - V_2)^2) & \cdots & E((U_2 - V_2)(U_n - V_n)) \\ \vdots & \vdots & \ddots & \vdots \\ E((U_n - V_n)(U_1 - V_1)) & E((U_n - V_n)(U_2 - V_2)) & \cdots & E((U_n - V_n)^2) \end{pmatrix} =$$

$$= \begin{pmatrix} d_{2,1}^2(U_1, V_1) & 0 & \cdots & 0 \\ 0 & d_{2,1}^2(U_2, V_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{2,1}^2(U_n, V_n) \end{pmatrix} = d_{2,1}^2(U_1, V_1) I_{n \times n}$$

ya que vimos en la demostración de la proposición [7] que bajo estas condiciones,

$$E((U_i - V_i)(U_j - V_j)) = \begin{cases} E(|U_i - V_i|^2) = d_{2,1}^2(U_i, V_i)^2 & si \quad i = j \\ \\ 0 & si \quad i \neq j \end{cases}$$

y además $U_1, U_2, ..., U_n$ están igualmente distribuidas y lo mismo $V_1, V_2, ..., V_n$, luego $d_{2,1}(U_i, V_i) = d_{2,1}(U_j, V_j)$. Por tanto:

$$[\boxtimes] = \mathrm{tr}(d_{2,1}^2(U_1, V_1) I_{n \times n} A^T A) = d_{2,1}^2(U_1, V_1) \mathrm{tr}(A^T A)$$

□

La siguiente proposición es el producto final de muchos resultados parciales. La versión que incluimos de la proposición aparece en [17, Teor. 2.12], y nos permite caracterizar los emparejamientos óptimos y las aplicaciones de transporte óptimo para el coste dado por la 2-distancia de Wasserstein, a los que nos referiremos a partir de ahora simplemente como emparejamientos óptimos y aplicaciones de transporte óptimo. En este resultado juega un papel importante el análisis convexo, y necesitamos introducir en primer lugar algunas definiciones relacionadas con esta teoría. La primera generaliza el concepto de diferencial de una función en un punto, cuando estamos trabajando con funciones convexas.

Definición 6. *Dada una función convexa $\varphi : \mathbb{R}^d \to \mathbb{R}$, se define su subdiferencial en el punto x como*

$$\partial\varphi(x) = \{y \in \mathbb{R}^d : \varphi(z) \geq \varphi(x) + < y, x - z > \forall\, z \in \mathbb{R}^d\}$$

Si φ es diferenciable en el punto x, entonces $\partial\varphi(x)$ está formado por un único punto que coincide con $\nabla\varphi(x)$.

A partir de la definición, podemos definir una aplicación $\partial\varphi$ que envía a cada $x \in \mathbb{R}^d$ en el conjunto $\partial\varphi(x)$. A partir de ahora identificaremos a la aplicación $\partial\varphi$ con el conjunto dado por su grafo, que denotaremos $Gr(\partial\varphi) \subset \mathbb{R}^d \times \mathbb{R}^d$. Introducimos otro concepto que aparecerá en la proposición:

Definición 7. *Dada una función semicontinua inferior φ, se define su función conjugada por:*

$$\varphi^*(y) = sup\{x \cdot y - \varphi(x) :\ x \in \mathbb{R}^d\}$$

Estamos ya en condiciones de ver el resultado:

Proposición 10. *Sean $\mu, \nu \in \Gamma_p$ y sea π la distribución conjunta de un par de variables aleatorias (U,V) que toman valores en $\mathbb{R}^d$ y tales que $\mathscr{L}(U) = \mu, \mathscr{L}(V) = \nu$.*

(a) La distribución de π es un emparejamiento óptimo de μ y ν si y sólo si existe una función convexa semicontinua inferior φ tal que $V \in \partial\varphi(U)$ c.s.

(b) Si asumimos que μ es una probabilidad que desaparece en conjuntos de dimensión $d-1$, entonces existe un único emparejamiento óptimo de μ y ν, que viene caracterizado por una aplicación de transporte óptimo T. Esto es, $\pi = \mu \circ (Id, T)^{-1}$ (o $V = T(U)$ c.s. para alguna v.a. U con $\mathscr{L}(U) = \mu$) y además:

$$d_2^2(\mu,\nu) = \int \|x - T(x)\|^2 \mu(dx) =$$

$$= \text{ínf}\{\int \|x - S(x)\|^2 d\mu(x) :\ \text{S cumple } \nu = \mu \circ S^{-1}\}$$

Dicha aplicación se caracteriza por $T = \nabla\varphi$ $\mu-$c.s., la única $\mu-$c.s. función que lleva μ en ν y que es el gradiente de una función semicontinua inferior φ. Además, si ν tampoco da masa a conjuntos de dimensión $d-1$, entonces para $\mu-$c.s. x, $\nu-$c.s. y, se tiene que:

$$\nabla\varphi^* \circ \nabla\varphi(x) = x \qquad \nabla\varphi \circ \nabla\varphi^*(y) = y$$

y φ^* es la única función $\nu-$c.s. de transporte óptimo que lleva ν en μ y que es el gradiente de una función convexa y semicontinua inferior.

La demostración de la proposición utiliza argumentos matemáticos que se salen del nivel de este trabajo. En [10] aparecen varias formas de probar el resultado. Mientras que las primeras se basan en teorema de dualidad de Kantorovich, la última demostración que se hace utiliza un enfoque geométrico del que es interesante comentar algún aspecto. Lo que se intenta en este caso es generalizar la idea de que las funciones de transporte óptimo en $\mathbb{R}$ son exactamente las funciones crecientes, a partir del siguiente concepto:

Definición 8. *Un subconjunto $A \subset \mathbb{R}^d \times \mathbb{R}^d$ es cíclicamente monótono si cumple que para todo $m \geq 1$ y para todos $(x_1, y_1), ..., (x_m, y_m) \in A$,*

$$\sum_{i=1}^{m} \|x_i - y_i\|^2 \leq \sum_{i=1}^{m} \|x_i - y_{i-1}\|^2$$

fijando la notación $y_0 = y_m$.

La demostración se basa entonces en probar que los dos siguientes resultados:

1. Si π es un emparejamiento óptimo de μ y ν para el coste dado por la 2-distancia de Wasserstein, entonces el soporte de π es cíclicamente monótono.
2. Un conjunto $A \subset \mathbb{R}^d \times \mathbb{R}^d$ es cíclicamente monótono si y sólo si está incluido en el subdiferencial de una función convexa semicontinua inferior. De hecho, los conjuntos maximales (para la relación de inclusión) cíclicamente monótonos son exactamente los subdiferenciales de funciones convexas semicontinuas inferiores.

De este razonamiento se pueden sacar además algunos hechos interesantes, que nos permiten ver que existen aplicaciones que bajo ciertas condiciones siempre son transportes óptimos. Los siguientes resultados aparecen en [1, cap. 2]:

- Entendiendo que una función $T : \mathbb{R}^d \to \mathbb{R}^d$ es cíclicamente monótona si lo es su grafo, tenemos que si $U \in \Gamma_2(\mathbb{R}^d)$ y T es cíclicamente monótona, entonces $(U, T(U))$ es un emparejamiento de $\mathscr{L}(U)$ y $\mathscr{L}(T(U))$ para el que existe una función convexa semicontinua inferior φ que verifica $T(U) \in \partial\varphi(U)$. Por tanto, $(U, T(U))$ es un emparejamiento óptimo de $\mathscr{L}(U)$ y $\mathscr{L}(T(U))$, sea cual sea la variable aleatoria $U \in \Gamma_2$.
- Se puede ver entonces que si $T : \mathbb{R}^d \to \mathbb{R}^d$ es medible e inyectiva, y $\frac{\partial T}{\partial x}$ definida positiva, entonces $(U, T(U))$ es un emparejamiento óptimo de $\mathscr{L}(U)$ y $\mathscr{L}(T(U))$ para cualquier variable aleatoria $U \in \Gamma_2(\mathbb{R}^d)$ que tenga función de densidad.
- Se puede probar además que si T es una aplicación lineal con matriz A definida positiva, entonces T es una aplicación de transporte óptimo entre $\mathscr{L}(U)$ y $\mathscr{L}(T(U))$ para cualquier $U \in \Gamma_2$.

Los siguientes corolarios también son consecuencia de la proposición [10]:

Corolario 3. *Si $\pi_j = \mathscr{L}(U, V_j), j = 1, 2, ..., k$ son emparejamientos óptimos de μ y ν_j, para $j = 1, 2, ..., k$, entonces dados unos pesos $\lambda_j, j = 1, 2, ..., k$, tales que $\sum_{j=1}^{k} \lambda_j = 1$, la probabilidad*

$$\pi = \mathscr{L}(U, \sum_{j=1}^{k} \lambda_j V_j)$$

es un emparejamiento óptimo de μ y $\mathscr{L}(\sum_{j=1}^{k} \lambda_j V_j)$.

Demostración. Por (a), para cada $j = 1, 2, ..., k$, existe una función semicontinua inferior φ_j tal que $V_j \in \partial\varphi_j(U)$ c.s.. Tomando la función

$$\varphi = \sum_{j=1}^{k} \lambda_j \varphi_j$$

tenemos una función semicontinua inferior y tal que $\sum_{j=1}^{k} \lambda_j V_j \in \partial\varphi(U)$ c.s., y de nuevo por (a), esto es equivalente a que $\pi = \mathscr{L}(U, \sum_{j=1}^{k} \lambda_j V_j)$ sea un emparejamiento óptimo de μ y $\mathscr{L}(\sum_{j=1}^{k} \lambda_j V_j)$. □

Corolario 4. *La optimalidad de una aplicación es una característica que no depende de las medidas transportadas. En otras palabras, si T es una aplicación de transporte óptimo que lleva μ en ν, y $\mu^*, \nu^* \in \Gamma_2$ son tales que $\nu^* = \mu^* \circ T^{-1}$ y el soporte de μ^* está contenido en el de μ, entonces T es también una aplicación de transporte óptimo de μ^* a ν^*, es decir:*

$$d_2(\mu^*, \nu^*) = \left(\int \|x - T(x)\|^2 d\mu^*(x) \right)^{1/2}$$

Demostración. Sabemos que T será la única $\mu-c.s.$ función que lleva μ en ν y que es el gradiente de una función semicontinua inferior φ. Como T lleva μ^* en ν^* y es gradiente de una función semicontinua inferior φ, será la función de transporte óptimo de μ^* en ν^*. □

Este hecho nos permite calcular distancias de Wasserstein entre algunas probabilidades cuando sabemos que están relacionadas por alguna función de transporte óptimo.

Se puede ver que en general la composición de funciones de transporte óptimo no preserva la optimalidad: Si T_1 es una aplicación de transporte óptimo que lleva μ_1 en μ_2, y T_2 es una aplicación de transporte óptimo que lleva μ_2 en μ_3, entonces no tiene por qué ocurrir que $T_2 \circ T_1$ sea una aplicación de transporte óptimo que lleva μ_1 en μ_3. Sin embargo, tanto las combinaciones lineales positivas como los límites puntuales de funciones óptimas si que conservan la optimalidad. Esta última propiedad es sencilla de a partir del concepto de función cíclicamente monónona, viendo que las aplicaciones de transporte óptimo son cíclicamente monótonas y el límite puntual de funciones cíclicamente monótonas es cíclicamente monótono, y por tanto es una aplicación de transporte óptimo.

La siguiente proposición afirma que la distancia de Wasserstein entre dos probabilidades no depende de la base en la que estamos expresando dichas probabilidades.

Proposición 11. *Sean $\mu, \nu \in \Gamma_2$. Si A es una matriz unitaria, y T es la transformación lineal dada por $T(x) = Ax$, se cumple que:*

$$d_2(\mu, \nu) = d_2(\mu \circ T^{-1}, \nu \circ T^{-1})$$

Demostración. Sean U, V variables aleatorias definidas en un espacio probabilístico común y tales que:

$$\mathscr{L}(U) = \mu \quad \mathscr{L}(V) = \nu \quad \text{y} \quad d_2(\mu, \nu) = (E\|U - V\|^2)^{1/2}$$

Por las propiedades de las matrices unitarias,

$$d_2^2(\mu,\nu) = E\|U-V\|^2 = E\|AU-AV\|^2 \geq d_2^2(\mu\circ T^{-1}, \nu\circ T^{-1})$$

Análogamente, existen U^*, V^* son variables aleatorias definidas en un espacio probabilístico común tales que:

$$\mathscr{L}(U^*) = \mu\circ T^{-1} \quad \mathscr{L}(V^*) = \nu\circ T^{-1} \text{ y } d_2(\mu\circ T^{-1}, \nu\circ T^{-1}) = (E\|U^*-V^*\|^2)^{1/2}$$

entonces $\mathscr{L}(A'U^*) = \mu\circ T^{-1}\circ T = \mu$ y $\mathscr{L}(A'V^*) = \nu\circ T^{-1}\circ T = \nu$. Por tanto:

$$d_2^2(\mu\circ T^{-1}, \nu\circ T^{-1}) = E\|U^*-V^*\|^2 = E\|A'U^*-A'V^*\|^2 \geq d_2^2(\mu,\nu)$$

□

Antes de enunciar el siguiente resultado, recordemos que si $U = (U_1, U_2, ..., U_d)$ es un vector aleatorio con valores en $\mathbb{R}^d$, entonces su media es un vector $m = (m_1, m_2, ..., m_d) \in \mathbb{R}^d$ y su matriz de covarianzas es una matriz Σ de tamaño $d\times d$, que viene dada por:

$$\Sigma = Var(U) = \begin{pmatrix} Var(U_1) & Cov(U_1,U_2) & \cdots & Cov(U_1,U_d) \\ Cov(U_1,U_2) & Var(U_2) & \cdots & Cov(U_2,U_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(U_1,U_d) & Cov(U_2,U_d) & \cdots & Var(U_d) \end{pmatrix}$$

Es una matriz simétrica, con todos los elementos de la diagonal no negativos, y que además será semidefinida positiva ya que si $a\in\mathbb{R}^d$, entonces a^TU es una variable aleatoria real que cumple que

$$Var(a^TU) = a^TVar(U)a = a^T\Sigma a \geq 0$$

Esto lo utilizaremos en la demostración del siguiente teorema, que es un caso particular del teorema 2.1 en [7].

Teorema 1. *Sean μ,ν probabilidades en Γ_2 con medias m_μ, m_ν, y matrices de covarianza Σ_μ, Σ_ν. Si Σ_μ suponemos que es regular, entonces:*

$$d_2^2(\mu,\nu) \geq \|m_\mu - m_\nu\|^2 + tr(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}) \tag{3}$$

Además, la igualdad se da si y sólo si la aplicación $T(x) = (m_\nu - m_\mu) + Ax$ transporta μ en ν, siendo A la matriz semidefinida positiva dada por

$$A = \Sigma_\mu^{-1/2}(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}\Sigma_\mu^{-1/2}$$

Demostración. Hemos visto que si μ^* y ν^* son las probabilidades que quedan al centrar μ y ν, entonces:

$$d_2^2(\mu,\nu) = \|m_\mu - m_\nu\|^2 + d_2^2(\mu^*,\nu^*)$$

Podemos suponer entonces sin pérdida de generalidad $m_\mu = m_\nu = 0$.

A es una matriz simétrica y semidefinida positiva. Sabemos entonces que existe una base

ortonormal $\{e_k\}_{k=1}^d$ formada por autovectores de la matriz A. Sean $\{\lambda_k\}_{k=1}^d$ los autovalores asociados a dichos autovectores. En primer lugar, se tiene que:

$$\mathrm{tr}\left((\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}\right) = \mathrm{tr}(\Sigma_\mu^{1/2}A\Sigma_\mu^{1/2}) = \mathrm{tr}(A\Sigma_\mu) = \sum_{i=1}^d e_i^T(A\Sigma_\mu)e_i =$$

$$= \sum_{i=1}^d (\Sigma_\mu Ae_i)^T \cdot e_i = \sum_{i=1}^d \lambda_i(\Sigma_\mu e_i)^T \cdot e_i = \sum_{i=1}^d \lambda_i(e_i^T\Sigma_\mu e_i) \tag{4}$$

Sean U, V dos vectores aleatorios en $\mathbb{R}^d$ tales que:

$$\mathscr{L}(U) = \mu \quad \mathscr{L}(V) = \nu \quad y \quad d_2^2(\mu,\nu) = E\|U-V\|^2$$

Si denotamos $U_i = U^Te_i, V_i = V^Te_i$ para $i = 1, 2, ..., d$ tenemos que:

$$d_2^2(\mu,\nu) = E\|U-V\|^2 = E\left(\sum_{i=1}^d |U_i - V_i|^2\right) = \sum_{i=1}^d E|U_i - V_i|^2$$

Para cada $i = 1, 2, ..., d$ sean $\bar{U}_i, \bar{V}_i$ variables aleatorias reales tales que:

$$\bar{U}_i =_d U_i \quad \bar{V}_i =_d V_i \quad y \quad d_2^2(\mathscr{L}(U_i), \mathscr{L}(V_i)) = E\|\bar{U}_i - \bar{V}_i\|^2$$

Si denotamos $\sigma(U_i), \sigma(V_i)$ a las desviaciones típicas de las variables U_i, V_i, entonces tenemos las siguientes desigualdades:

$$d_2^2(\mu,\nu) = \sum_{i=1}^d E|U_i - V_i|^2 \geq \sum_{i=1}^d d_2^2(\mathscr{L}(U_i), \mathscr{L}(V_i)) = \sum_{i=1}^d E|\bar{U}_i - \bar{V}_i|^2 \geq \tag{5}$$

$$\geq \sum_{i=1}^d \left(\left(E(\bar{U}_i^2\right)^{1/2} - \left(E(\bar{V}_i^2\right)^{1/2}\right)^2 = \sum_{i=1}^d \left(\left(E(U_i^2\right)^{1/2} - \left(E(V_i^2\right)^{1/2}\right)^2 =$$

$$= \sum_{i=1}^d (\sigma(U_i) - \sigma(V_i))^2 = \sum_{i=1}^d \left((e_i^T\Sigma_\mu e_i)^{1/2} - (e_i^T\Sigma_\nu e_i)^{1/2}\right)^2 =$$

$$= \sum_{i=1}^d \left((e_i^T\Sigma_\mu e_i) + (e_i^T\Sigma_\nu e_i) - 2(e_i^T\Sigma_\mu e_i)^{1/2}(e_i^T\Sigma_\nu e_i)^{1/2}\right)$$

Se tiene además que $A\Sigma_\mu A = \Sigma_\nu$, por tanto:

$$e_i^T\Sigma_\nu e_i = e_i^TA\Sigma_\mu Ae_i = (Ae_i)^T\Sigma_\nu(Ae_i) = \lambda_i^2(e_i^T\Sigma_\mu e_i)$$

Hemos probado entonces que:

$$d_2^2(\mu,\nu) \geq \sum_{i=1}^d (e_i^T\Sigma_\mu e_i) + \sum_{i=1}^d (e_i^T\Sigma_\nu e_i) - 2\sum_{i=1}^d \lambda_i(e_i^T\Sigma_\mu e_i) =$$

$$= \text{tr}(\Sigma_\mu + \text{tr}(\Sigma_\nu) - 2\text{tr}\left((\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2})^{1/2}\right)$$

donde la última igualdad es consecuencia de la igualdad (4).

Veamos la segunda parte de la proposición. Para ello, notemos que la primera desigualdad en (5) es una igualdad si y sólo si

$$d_2^2(\mathscr{L}(U_i), \mathscr{L}(U_i)) = E|U_i - V_i|^2 \quad \forall i \in \{1, 2, ..., d\}$$

Es decir, si y sólo si podemos asumir que $\bar{U}_i =_{c.s.} U_i$ y $\bar{V}_i =_{c.s.} V_i$. La segunda desigualdad es una igualdad si y sólo si $\bar{U}_i, \bar{V}_i$ están relacionadas linealmente: $\bar{V}_i = \alpha_i \bar{U}_i$ para alguna constante $\alpha_i > 0$. En ese caso,

$$\begin{cases} \text{Var}(\bar{Y}_i) = \text{Var}(\alpha_i \bar{U}_i) = \alpha_i^2 \text{Var}(\bar{U}_i) = \alpha_i^2 (e_i^T \Sigma_\mu e_i) \\ \\ \text{Var}(\bar{Y}_i) = e_i^T \Sigma_\nu e_i = e_i^T A \Sigma_\mu A e_i = (Ae_i)^T \Sigma_\mu (Ae_i) = \lambda_i^2 (e_i^T \Sigma_\mu e_i) \end{cases}$$

Luego necesariamente $\alpha_i = \lambda_i$ y por tanto si se dan las dos igualdades,

$$V^T \cdot e_i = V_i = \lambda_i U_i = \lambda_i U^T \cdot e_i = U^T A e_i \quad \forall\, i = 1, 2, ..., d$$

y se tiene por tanto que $V^T = U^T A \Rightarrow V = AU$, es decir, la aplicación lineal de matriz A lleva la probabilidad μ en ν.

Recíprocamente, si la aplicación lineal de matriz A lleva la probabilidad μ en ν, si $\mathscr{L}(U) = \mu$ se tiene que tomando $V = AU$, entonces $\mathscr{L}(V) = \nu$ y se tiene que para cada $i = 1, ..., n$:

$$V_i = V^T e_i = (AU)^T e_i = U^T (Ae_i) = \lambda_i U^T e_i = \lambda_i U_i$$

Luego V_i, U_i están relacionadas linealmente y la segunda desigualdad es por tanto una igualdad. El hecho de que la primera desigualdad sea una igualdad es consecuencia de que los vectores $U, V{=}AU$ tienen la misma estructura de dependencia. Incidiremos sobre esto en la nota posterior a la demostración. □

Nota 2. De la demostración del teorema anterior se pueden deducir los siguientes hechos: Si U, V son dos vectores aleatorios en $\mathbb{R}^d$ con distribución μ, ν centradas y $\{e_k\}_{k=1}^d$ es una base ortonormal de $\mathbb{R}^d$ y $U_i = U^T \cdot e_i, V_i = V^T \cdot e_i$, se tienen las desigualdades

$$d_2^2(\mu, \nu) \geq \sum_{i=1}^d d_2^2(\mathscr{L}(U_i), \mathscr{L}(U_i)) \geq \sum_{i=1}^d (\sigma(U_i) - \sigma(V_i))^2$$

Por una parte, la cota que nos da la segunda desigualdad tan sólo depende de las matrices de covarianzas Σ_μ y Σ_ν. El mejor valor para la segunda de las cotas se obtiene de hecho en el caso en que como base $\{e_k\}_{k=1}^d$ se toma la base de autovectores de la matriz

$$A = \Sigma_\mu^{-1/2} (\Sigma_\mu^{1/2} \Sigma_\nu \Sigma_\mu^{1/2})^{1/2} \Sigma_\mu^{-1/2}$$

ya que en ese caso se puede ver que la cota se alcanza. Si tenemos dos distribuciones Gaussianas con matrices de covarianzas Σ_μ y Σ_ν, en ese caso la aplicación T del teorema lleva una

distribución en la otra, y de la demostración de esa proposición se puede ver que en ese caso todas las desigualdades serán igualdades.

Por otra parte, en [9, cap. 2] se prueba que la cota dada por la primera desigualdad:

$$d_2^2(\mu,\nu) \geq \sum_{i=1}^{d} d_2^2(\mathscr{L}(U_i),\mathscr{L}(U_i))$$

es una igualdad si y sólo si los vectores μ,ν tienen la misma estructura de dependencia. Esto ocurre en particular en el caso en que estamos trabajando con distribuciones normales y $\{e_k\}_{k=1}^d$ es una base ortonormal en la cual las matrices de covarianzas Σ_μ, Σ_ν son diagonales. Si además μ,ν son centradas, entonces se tiene que:

$$d_2^2(U_i,V_i) = (\sigma(U_i) - \sigma(V_i))^2$$

ya que U_i, V_i son distribuciones normales centradas con varianzas $\sigma(U_i)^2, \sigma(V_i)^2$, y entonces sus funciones cuantiles asociadas son

$$F^{-1}(t) = \sigma(U_i)\phi^{-1}(t)$$

$$G^{-1}(t) = \sigma(V_i)\phi^{-1}(t)$$

siendo ϕ^{-1} la función cuantil asociada a una N(0,1). Por la caracterización que vimos de la distancia de Wasserstein en el caso unidimensional,

$$d_2^2(U_i,V_i) = \int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt = \int_0^1 (\sigma(U_i) - \sigma(V_i))^2 \phi^{-1}(t)^2 dt =$$

$$= (\sigma(U_i) - \sigma(V_i))^2 \int_0^1 \phi^{-1}(t)^2 dt = (\sigma(U_i) - \sigma(V_i))^2$$

Resumiendo, si tenemos distribuciones normales y $\{e_k\}_{k=1}^d$ es una base ortonormal en la cual las matrices de covarianzas Σ_μ, Σ_ν son diagonales, entonces

$$d_2^2(\mu,\nu) = \sum_{i=1}^{d} d_2^2(\mathscr{L}(U_i),\mathscr{L}(U_i)) = \sum_{i=1}^{d} (\sigma(U_i) - \sigma(V_i))^2$$

Nota 3. Con el mismo argumento de la demostración, se puede probar que si descomponemos el espacio $\mathbb{R}^d$ como suma de dos subespacios ortogonales,

$$\mathbb{R}^d = H \oplus H^\perp$$

entonces si denotamos U_1, V_1 a las proyecciones de U y V sobre H, y U_2, V_2 a las proyecciones de U y V sobre $H^\perp$, entonces:

$$d_2^2(\mu,\nu) \geq d_2^2(\mathscr{L}(U_1),\mathscr{L}(V_1)) + d_2^2(\mathscr{L}(U_2),\mathscr{L}(V_2))$$

Al igual que antes, sabemos que la desigualdad será una igualdad en el caso en que μ, ν tengan la misma estructura de dependencia. En particular, en el caso en que μ, ν son normales, si existe una base ortonormal $\{e_k\}_{k=1}^d$ en la cual las matrices de covarianzas de μ, ν tienen la forma:

$$\Sigma_\mu = \left[\begin{array}{c|c} A_\mu & 0 \\ \hline 0 & B_\mu \end{array}\right] \qquad \Sigma_\nu = \left[\begin{array}{c|c} A_\nu & 0 \\ \hline 0 & B_\nu \end{array}\right]$$

con A_i matriz de tamaño $l \times l$, B_i matriz de tamaño $(d-l) \times (d-l)$, $l \in \{1, ..., d\}$, entonces tomando como H el subespacio generado por $\{e_k\}_{k=1}^l$, siguiendo la notación del principio tendremos que:

$$d_2^2(\mu, \nu) = d_2^2(\mathscr{L}(U_1), \mathscr{L}(V_1)) + d_2^2(\mathscr{L}(U_2), \mathscr{L}(V_2))$$

3. Baricentros en espacios de Wasserstein

Comenzamos introduciendo el concepto de media Frechet. La idea de la media Frechet consiste en dar un representante central de un conjunto de puntos de un espacio métrico. Más formalmente:

Definición 9. *Sea (M,d) un espacio métrico, y sean $x_1, x_2, ..., x_n \in M$. Para cada $p \in M$, definimos:*

$$\Phi(p) = \sum_{i=1}^n d^2(p, x_i)$$

Si existe algún $m \in M$ tal que $\Phi(m)$ sea mínimo, entonces decimos que:

$$m = \underset{p \in M}{\operatorname{argmin}}\ \Phi(p) = \underset{p \in M}{\operatorname{argmin}}\ \sum_{i=1}^n d^2(p, x_i)$$

es una media Frechet (o un baricentro) de $x_1, x_2, ..., x_n$.

La media Frechet de un conjunto de puntos puede no ser única. Hay muchas medias Frechet en distintos espacios métricos y con distintas distancias. Vamos a ver algún ejemplo:

1. Tomamos $M = \mathbb{R}$ con la norma euclídea usual: $d(x, y) = |x - y|$. Sean $x_1, x_2, ..., x_n \in \mathbb{R}$. La media Frechet será entonces el mínimo de la función:

   $$\Phi(p) = \sum_{i=1}^n (p - x_i)^2$$

   Hallamos el mínimo de dicha función, que es C^∞.

   $$\Phi'(p) = \sum_{i=1}^n 2(p - x_i) = 2(\sum_{i=1}^n p - \sum_{i=1}^n x_i) = 2(np - \sum_{i=1}^n x_i)$$

   $$\Phi'(p) = 0 \Leftrightarrow p = \frac{\sum_{i=1}^n x_i}{n}$$

   $$\Phi''(p) = \sum_{i=1}^n 2 = 2n > 0$$

   $\Rightarrow \frac{\sum_{i=1}^n x_i}{n}$ es un mínimo de $\Phi(p)$, luego es una media Frechet de $x_1, x_2, ..., x_n$.

2. Tomamos $M = \mathbb{R}^d$ con la distancia habitual: $d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + ... + (x_d - y_d)^2}$. Sean $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in \mathbb{R}^d$, siendo $\mathbf{x}_i = (x_i^1, ..., x_i^d)$. La media Frechet será entonces el mínimo de la función:

$$\Phi(\mathbf{p}) = \sum_{i=1}^{n} (p_1 - x_i^1)^2 + ... + (p_d - x_i^d)^2$$

Hallamos el mínimo de dicha función, que es C^∞.

$$\frac{\partial \Phi}{\partial p_j}(\mathbf{p}) = \sum_{i=1}^{n} 2(p_j - x_i^j) = 2(\sum_{i=1}^{n} p_j - \sum_{i=1}^{n} x_i^j)$$

$$\frac{\partial \Phi}{\partial p_j}(\mathbf{p}) = 0 \Leftrightarrow p_j = \frac{\sum_{i=1}^{n} x_i^j}{n}$$

$$\nabla \Phi(\mathbf{p}) = 0 \Leftrightarrow \mathbf{p} = \left(\frac{\sum_{i=1}^{n} x_i^1}{n}, ..., \frac{\sum_{i=1}^{n} x_d^j}{n} \right) = \frac{\mathbf{x}_1 + \mathbf{x}_2 + ... + \mathbf{x}_n}{n}$$

Y además la matriz Hessiana en ese punto es:

$$H\Phi(p) = \begin{pmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2 \end{pmatrix}$$

que es definida positiva, ya que todos los autovalores son positivos.

$\Rightarrow \dfrac{\mathbf{x}_1 + \mathbf{x}_2 + ... + \mathbf{x}_n}{n}$ es un mínimo, es la media Frechet en este caso.

3. Cuando $M = (0, \infty)$ y $d(x, y) = |log(x) - log(y)|$, obtenemos la denominada media geométrica. Sean $x_1, x_2, ..., x_n \in M$. Buscamos $p > 0$ que minimice la función:

$$\Phi(p) = \sum_{i=1}^{n} (log(p) - log(x_i))^2$$

Hallamos el mínimo de dicha función:

$$\Phi'(p) = \sum_{i=1}^{n} 2(log(p) - log(x_i)) \frac{1}{p} = \frac{2}{p}(nlog(p) - \sum_{i=1}^{n} log(x_i))$$

$$\Phi'(p) = 0 \Leftrightarrow p = e^{log(p)} = e^{\frac{\sum_{i=1}^{n} log(x_i)}{n}} = (x_1 x_2 \cdots x_n)^{1/n}$$

Es un mínimo ya que es único extremo relativo de Φ, que es C^∞ y cumple que

$$\lim_{p \to \infty} \Phi(p) = \lim_{p \to 0} \Phi(p) = \infty$$

$\Rightarrow \sqrt{x_1 x_2 ... x_n}$ es la media Frechet en este caso y se conoce como media geométrica.

Definición 10. *Dadas probabilidades $\nu_1, \nu_2..., \nu_k \in \Gamma_2$, se dice que $\bar{\mu} \in \Gamma_2$ es un baricentro de dichas probabilidades si es una media Frechet de dichos puntos en el espacio Γ_2 utilizando la distancia de Wasserstein d_2.*

Es decir, el baricentro de $\nu_1, \nu_2..., \nu_k \in \Gamma_2$, si existe, será una probabilidad $\bar{\mu}$ que cumpla que:

$$\sum_{i=1}^{k} d_2^2(\bar{\mu}, \nu_i) = \text{ínf}\{\sum_{i=1}^{k} d_2^2(\mu, \nu_i) : \mu \in \Gamma_2\}$$

$\bar{\mu}$ nos servirá como "representante"de las probabilidades $\nu_1, \nu_2..., \nu_k$. En muchos casos, diferentes motivos darán lugar a que tenga sentido dar más importancia a alguna probabilidad ν_i y menos a alguna probabilidad ν_j. Para esto estudiamos el problema en el que si $0 < \lambda_i \leq 1, i \in \{1, 2, ..., k\}$ y $\sum_{i=1}^{k} \lambda_i = 1$, queremos hallar $\bar{\mu} \in \Gamma_2$ que minimice:

$$V(\mu) = \sum_{i=1}^{k} \lambda_i d_2^2(\mu, \nu_i), \quad \mu \in \Gamma_2$$

Es decir, buscamos $\bar{\mu} \in \Gamma_2$ que cumpla que:

$$V(\bar{\mu}) = \text{ínf}\{V(\mu) : \mu \in \Gamma_2\}$$

En este caso, si existe $\bar{\mu}$ será el baricentro de las probabilidades $\nu_1, \nu_2..., \nu_k$ con pesos $\{\lambda_i\}_{i=1}^{k}$. Generalmente nos referiremos a él simplemente como baricentro de las probabilidades $\nu_1, \nu_2..., \nu_k$, dando por hecho que estamos considerando en cada caso los pesos $\{\lambda_i\}_{i=1}^{k}$ correspondientes.

En esta sección vamos a estudiar propiedades de los baricentros, con el objetivo de dar un método iterativo para aproximar el baricentro de un conjunto de probabilidades. Esto nos va a permitir calcular el baricentro de forma sencilla cuando estamos trabajando con familias de localización y escala. Todos los resultados de esta sección relativos a los baricentros, junto con sus correspondientes demostraciones, aparecen en [2].

Proposición 12. *Dadas probabilidades $\nu_1, \nu_2..., \nu_k \in \Gamma_2$:*

- *Existe un baricentro de dichas probabilidades.*
- *Si además alguna de las probabilidades $\nu_1, \nu_2..., \nu_k$ da probabilidad nula a los conjuntos de dimensión $d-1$, entonces el baricentro es único.*

Demostración. Dadas probabilidades $\nu_1, \nu_2..., \nu_k \in \Gamma_2$ y probabilidades conjuntas π_j en $\mathbb{R}^d \times \mathbb{R}^d$ tales que las marginales sean μ y ν_j, por el lema del pegado sabemos que existe una probabilidad $\bar{\pi}$ en $(\mathbb{R}^d)^{k+1}$ tal que la probabilidad marginal sobre los factores 1 y j+1 coincida con π_j, para cada $j = 1, 2, ..., k$. Tenemos así que:

$$\text{ínf}\{V(\mu) : \mu \in \Gamma_2\} =$$

$$= \text{ínf}\left\{\int_{(\mathbb{R}^d)^{k+1}} \sum_{j=1}^{k} \lambda_j \|x - x_j\|^2 d\pi(x, x_1, ..., x_k) : \pi \in \Pi(\cdot, \nu_1, \nu_2, ..., \nu_k)\right\}$$

siendo $\Pi(\cdot, \nu_1, \nu_2, ..., \nu_k)$ el conjunto de probabilidades en $(\mathbb{R}^d)^{k+1}$ cuyas n últimas distribuciones marginales son $\nu_1, \nu_2 ..., \nu_k$. Dados $x_1, x_2, ..., x_k$ fijos, se tiene la siguiente desigualdad para cada $x \in \mathbb{R}^d$:

$$\sum_{j=1}^{k} \lambda_j \|x - x_j\|^2 \geq \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2$$

siendo $\bar{x} = \sum_{j=1}^{k} \lambda_j x_j$. Para ver esta desigualdad, basta ver que en $\bar{x}$ se alcanza un mínimo de dicha función. Esto significa que entre todas las funciones $\bar{\pi}$ con la misma distribución marginal sobre los k últimos factores, la función $\int_{(\mathbb{R}^d)^{k+1}} \sum_{j=1}^{k} \lambda_j \|x - x_j\|^2 d\bar{\pi}(x, x_1, ..., x_k)$ se minimiza cuando $\bar{\pi}$ se concentra en el conjunto $\{(x, x_1, x_2, ..., x_k) : x = \sum_{j=1}^{k} \lambda_j x_j\}$. Esto significa que:

$$\text{ínf}\, \{V(\mu) : \mu \in \Gamma_2\} =$$

$$= \text{ínf} \left\{ \int_{(\mathbb{R}^d)^k} \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 d\pi(x_1, ..., x_k) : \pi \in \Pi(\nu_1, \nu_2, ..., \nu_k) \right\}$$

siendo $\Pi(\nu_1, \nu_2, ..., \nu_k)$ el conjunto de probabilidades en $(\mathbb{R}^d)^k$ cuyas distribuciones marginales son $\nu_1, \nu_2 ..., \nu_k$. El baricentro será entonces la ley inducida por $\bar{\pi}$, un mínimo en la segunda parte de la igualdad, a través de la aplicación

$$(x_1, ..., x_k) \mapsto \bar{x} = \sum_{j=1}^{k} \lambda_j x_j$$

Veamos que existe un mínimo $\bar{\pi}$. Sea

$$\bar{V} = \text{ínf} \left\{ \int_{(\mathbb{R}^d)^k} \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 d\bar{\pi}(x_1, ..., x_k) : \bar{\pi} \in \Pi(\nu_1, \nu_2, ..., \nu_k) \right\}$$

y supongamos que $\{\pi_n\}_{n=1}^{\infty}$ es una sucesión de elementos de $\Pi(\nu_1, \nu_2, ..., \nu_k)$ tal que:

$$\int_{(\mathbb{R}^d)^k} \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 d\pi_n(x_1, ..., x_k) \xrightarrow{n \to \infty} \bar{V}$$

Como las probabilidades marginales de π_n son $(\pi_n)_1 = \nu_1, (\pi_n)_2 = \nu_2, ..., (\pi_n)_k = \nu_k$ para todo $n \in \mathbb{N}$, entonces sabemos que la sucesión $\{\pi_n\}_{n=1}^{\infty}$ es ajustada. Tomando subsucesiones si fuera necesario, podemos suponer que $\{\pi_n\}_{n=1}^{\infty}$ converge débilmente a $\bar{\pi} \in \Pi(\nu_1, \nu_2, ..., \nu_k)$, que tendrá probabilidades marginales $(\bar{\pi})_1 = \nu_1, (\bar{\pi})_2 = \nu_2, ..., (\bar{\pi})_k = \nu_k$. Aplicando primero el terorema de Skorohod [6] y luego el lema de Fatou, al igual que hicimos en la demostración de la proposición [1], obtenemos que :

$$\int_{(\mathbb{R}^d)^k} \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 d\bar{\pi}(x_1, ..., x_k) \leq \lim_{n \to \infty} \int_{(\mathbb{R}^d)^k} \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 d\pi_n(x_1, ..., x_k) = \bar{V}$$

Como $\bar{\pi} \in \Pi(\nu_1, \nu_2, ..., \nu_k)$ por definición sabemos que $\bar{V} \leq \int_{(\mathbb{R}^d)^k} \sum_{j=1}^k \lambda_j \|\bar{x} - x_j\|^2 d\bar{\pi}(x_1, ..., x_k)$ y entonces necesariamente:

$$\bar{V} = \int_{(\mathbb{R}^d)^k} \sum_{j=1}^k \lambda_j \|\bar{x} - x_j\|^2 d\bar{\pi}(x_1, ..., x_k)$$

La unicidad del baricentro se puede ver que se cumple en la situación del enunciado, tal y como se hace en [3, prop. 3.5.] trabajando con duales. Nosotros vamos a dar una idea de la demostración en el caso en que una de las distribuciones $\nu_1, \nu_2, ..., \nu_k$ tiene densidad:

Si ν tiene densidad, entonces se puede probar que la función $\mu \mapsto d_2^2(\mu, \nu)$ es estrictamente convexa. De aquí se deduce que si una de las distribuciones $\nu_1, \nu_2, ..., \nu_k$ tiene densidad, entonces la función $V(\mu)$ es estrictamente convexa. Esto nos da la unicidad del baricentro, ya que la función solo podrá tener un mínimo.

□

3.1. Aproximación de punto fijo al baricentro

El cálculo del baricentro presenta en muchos casos grandes dificultades. En esta sección vamos a tratar el problema de encontrar un el baricentro de un conjunto finito de probabilidades $\nu_1, \nu_2, ..., \nu_k \in \Gamma_{2,ac}$, siendo $\Gamma_{2,ac}$ el conjunto de probabilidades sobre $(\mathbb{R}^d, \beta^d)$ con momento de orden 2 finito y absolutamente continuas.

La idea será la siguiente: Se introducirá un operador G sobre $\Gamma_{2,ac}$ y se probará que, bajo condiciones muy generales, el baricentro será un punto fijo de dicho operador. Finalmente, daremos un proceso iterativo para aproximar dicho baricentro. Esto nos proporcionará un método de cálculo efectivo del baricentro en familias de localización y escala, como la familia de las distribuciones Gaussianas.

Vamos a ver en primer lugar como definimos el operador:

$$G : \Gamma_{2,ac} \longrightarrow \Gamma_{2,ac}$$

De la proposición [10], deducimos que si $\mu \in \Gamma_{2,ac}$, existen aplicaciones de transporte óptimo $T_1, T_2, ..., T_k$ únicas $\mu - c.s.$ y tales que $\mathscr{L}(T_j(U)) = \nu_j$ para cada $j = 1, 2, ..., k$, siendo U un vector aleatorio tal que $\mathscr{L}(U) = \mu$. Es decir, para cada $j = 1, 2, ..., k$,

$$d_2^2(\mu, \nu_j) = E\|U - T_j(U)\|^2$$

Definimos entonces:

$$G(\mu) = \mathscr{L}\left(\sum_{j=1}^k \lambda_j T_j(U)\right)$$

Tenemos entonces el siguiente resultado, que no vamos a demostrar:

Proposición 13. *Si ν_j tiene densidad para cada $j = 1, 2, ..., k$, entonces*

- *G va de $\Gamma_{2,ac}$ en $\Gamma_{2,ac}$, es decir, lleva una probabilidad μ absolutamente continua y con momento de orden 2 finito en otra probabilidad $G(\mu)$ con las mismas propiedades.*
- *G es una función continua para la distancia d_2.*

Nota 4. Si asumimos que $\nu_1, \nu_2, ..., \nu_k \in \Gamma_{2,ac}$ y al menos una de las distribuciones tiene densidad acotada, entonces podemos afirmar los siguiente:

- Bajo estas condiciones hay un único baricentro que tiene función de densidad acotada.
- La conclusión del teorema anterior puede mejorarse bajo estas condiciones. Si por ejemplo ν_1 tiene densidad acotada f_1, entonces $G(\mu)$ tiene densidad acotada g que cumple:

$$\|g\|_\infty \leq \lambda_1^{-d} \|f_1\|_\infty$$

Proposición 14. *Si $\mu \in \Gamma_{2,ac}$ entonces se tiene que:*

$$V(\mu) \geq V(G(\mu)) + d_2^2(\mu, G(\mu))$$

Como consecuencia, $V(\mu) \geq V(G(\mu))$, siendo la desigualdad estricta si $\mu \neq G(\mu)$. En particular, si μ es un baricentro, entonces sabemos que

$$V(\mu) = \text{ínf}\{V(\mu') : \mu' \in \Gamma_{2,ac}\} \leq V(G(\mu))$$

y por tanto se tiene necesariamente que $\mu = G(\mu)$, es decir, μ es un punto fijo del operador G.

Demostración. Sean $x_1, x_2, ..., x_k \in \mathbb{R}^d$, y escribamos $\bar{x} = \sum_{j=1}^k \lambda_j x_j$. Tenemos que:

$$\sum_{j=1}^k \lambda_j \|x - x_j\|^2 = \sum_{j=1}^k \lambda_j \|\bar{x} - x_j\|^2 + \|x - \bar{x}\|^2 \qquad \forall\, x \in \mathbb{R}^d \tag{6}$$

Veamos de dónde viene dicha igualdad:

$$\sum_{j=1}^k \lambda_j \|x - x_j\|^2 = \sum_{j=1}^k \lambda_j \|(x - \bar{x}) + (\bar{x} - x_j)\|^2 =$$

$$= \sum_{j=1}^k \lambda_j \left(\|x - \bar{x}\|^2 + \|\bar{x} - x_j\|^2 + 2 < x - \bar{x}, \bar{x} - x_j > \right) =$$

$$= (\sum_{j=1}^k \lambda_j) \|x - \bar{x}\|^2 + \sum_{j=1}^k \lambda_j \|\bar{x} - x_j\|^2 + 2 \sum_{j=1}^k \lambda_j < x - \bar{x}, \bar{x} - x_j > = [\boxtimes]$$

Sabemos que $\sum_{j=1}^k \lambda_j = 1$. Además se tiene que:

$$\sum_{j=1}^k \lambda_j < x - \bar{x}, \bar{x} - x_j > = \sum_{j=1}^k \lambda_j \left(< x - \bar{x}, \bar{x} > - < x - \bar{x}, x_j > \right) =$$

$$= (\sum_{j=1}^{k} \lambda_j) < x - \bar{x}, \bar{x} > - \sum_{j=1}^{k} \lambda_j < x - \bar{x}, x_j >=$$

$$=< x - \bar{x}, \bar{x} > - < x - \bar{x}, \sum_{j=1}^{k} \lambda_j x_j >=< x - \bar{x}, \bar{x} > - < x - \bar{x}, \bar{x} >= 0$$

Y por tanto de la igualdad de arriba se deduce que:

$$[\boxtimes] = \sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 \ + \ \|x - \bar{x}\|^2$$

Denotando entonces por T_j la aplicación de transporte óptimo que lleva μ en ν_j, para cada $j = 1, 2, ..., k$, y $\bar{T}(x) = \sum_{j=1}^{k} \lambda_j T_j(x)$ se tiene que:

$$V(\mu) = \sum_{j=1}^{k} \lambda_j \int \|x - T_j(x)\|^2 d\mu(x) = \int \sum_{j=1}^{k} \lambda_j \|x - T_j(x)\|^2 d\mu(x) =$$

$$= \int \sum_{j=1}^{k} \lambda_j \|\bar{T}(x) - T_j(x)\|^2 d\mu(x) + \int \|\bar{T}(x) - x\|^2 d\mu(x) \tag{7}$$

donde el último paso es consecuencia de la igualdad que hemos probado al principio.
Por el corolario 3, sabemos que $\bar{T}$ es una aplicación de transporte óptimo que lleva μ en $G(\mu)$, y por tanto:

$$\int \|\bar{T}(x) - x\|^2 d\mu(x) = d_2^2(\mu, G(\mu)) \tag{8}$$

Finalmente, si $\bar{\pi}_j$ es la probabilidad inducida por la aplicación $(T_j, \bar{T})$, vemos que $\bar{\pi}_j$ es un emparejamiento de ν_j y $G(\mu)$ y por tanto:

$$\int \sum_{j=1}^{k} \lambda_j \|\bar{T}(x) - T_j(x)\|^2 d\mu(x) = \sum_{j=1}^{k} \lambda_j \int \|\bar{T}(x) - T_j(x)\|^2 d\mu(x) \geq$$

$$\geq \sum_{j=1}^{k} \lambda_j d_2^2(\nu_j, G(\mu) = V(G(\mu)) \tag{9}$$

Combinando (7),(8),(9) obtenemos lo que queríamos:

$$V(\mu) \geq V(G(\mu)) + d_2^2(\mu, G(\mu))$$

□

Nota 5. La proposición anterior sigue siendo cierta bajo condiciones más generales:

Supongamos únicamente que $\mu, \nu_1, ..., \nu_k \in \Gamma_2$. Asumamos que π_j es un emparejamiento óptimo de μ y ν_j para cada $j = 1, 2, ..., k$. Por lo visto al hablar de emparejamientos, podemos asumir

que existe existe algún espacio Ω en el que existan variables aleatorias $U, V_1, V_2, ..., V_k$ tales que $\pi_j = \mathscr{L}(U, V_j)$ (La diferencia con el caso anterior es que ahora no podemos afirmar que existan funciones de transporte óptimo T_j tales que $V_j = T_j(U)$).

En general, existen muchas distribuciones conjuntas de $(U, V_1, V_2, ..., V_k)$ que son compatibles con la construcción anterior. Para cada una de ellas, consideramos $\bar{V} = \sum_{j=1}^{k} \lambda_j V_j$. La distribución de $(U, \bar{V})$ será por el corolario [3] un emparejamiento óptimo de μ y $\mathscr{L}(\bar{V})$. Definiendo ahora $\tilde{G}(\mu) = \mathscr{L}(\bar{V})$, podemos replicar el argumento de la demostración anterior para obtener la desigualdad:

$$V(\mu) \geq V(\tilde{G}(\mu)) + d_2^2(\mu, \tilde{G}(\mu))$$

Hay que notar que en este caso $\tilde{G}(\mu)$ no está definida de forma única, depende de la distribución conjunta de $(U, V_1, V_2, ..., V_k)$.

Corolario 5. *Si $\nu_1, ..., \nu_k \in \Gamma_{2,ac}$ y al menos una de ellas tiene densidad acotada, si $\bar{\mu}$ es el único baricentro de $\nu_1, ..., \nu_k$, entonces $G(\bar{\mu}) = \bar{\mu}$.*

Demostración. Es consecuencia directa de la proposición anterior y de la nota 4, que asegura que bajo estas condiciones el baricentro tiene densidad y por tanto pertenece a $\Gamma_{2,ac}$. □

Este corolario es la base sobre la que se asienta el siguiente algoritmo iterativo para el cálculo del baricentro. Comenzamos tomando $\mu_0 \in \Gamma_{2,ac}$ y consideramos la sucesión:

$$\mu_{n+1} = G(\mu_n) \qquad n \geq 0$$

Por la proposición 13, sabemos que el algoritmo está bien definido. Si $\mu_n \in \Gamma_{2,ac}$, entonces $\mu_{n+1} \in \Gamma_{2,ac}$. Veamos ahora la consistencia de la iteración.

Proposición 15. *La sucesión $\{\mu_n\}_{n=0}^{\infty}$ definida anteriormente es ajustada. Si además suponemos que una de las probabilidades $\nu_1, ..., \nu_k \in \Gamma_{2,ac}$ tiene densidad acotada, entonces toda subsucesión de $\{\mu_n\}_{n=0}^{\infty}$ que converje débilmente, converge en la distancia d_2 a una probabilidad de $\Gamma_{2,ac}$ que es un punto fijo de G. Por tanto, bajo la misma suposición que antes, si G tiene un único punto fijo $\bar{\mu}$, entonces $\bar{\mu}$ es el único baricentro de $\nu_1, ..., \nu_k$ y además*

$$d_2(\mu_n, \bar{\mu}) \xrightarrow{n \to \infty} 0$$

Demostración. Para cada $n \geq 0$, consideramos un vector aleatorio U_n con $\mathscr{L}(U_n) = \mu_n$, y sea $T_{n,j}$ la aplicación de transporte óptimo que lleva μ_n en ν_j para cada $j = 1, 2, ..., k$. Sea $V_{n,j} = T_{n,j}(U_n)$.

La sucesión $\{(V_{n,1}, V_{n,2}, ..., V_{n,k})\}_{n=0}^{\infty}$ tiene las probabilidades marginales constantes, son $\nu_1, ..., \nu_k$. Podemos afirmar por lo tanto que es ajustada. Por continuidad, la sucesión $\{\sum_{j=1}^{k} \lambda_j V_{n,j}\}_{n=0}^{\infty}$ es también ajustada. Se tiene que $\mu_{n+1} = \mathscr{L}(\sum_{j=1}^{k} \lambda_j V_{n,j})$, luego $\{\mu_n\}_{n=0}^{\infty}$ es una sucesión ajustada.

El hecho de que las familias $\{\|V_{n,j}\|^2\}_{n=0}^{\infty}$ sean uniformemente integrables, puesto que la ley de todas ellas es ν_j, implica que también es uniformemente integrable la sucesión $\{\|\sum_{j=1}^{k} \lambda_j V_{n,j}\|^2\}_{n=0}^{\infty}$. Por tanto se tiene que $\|x\|^2$ es uniformemente μ_n-integrable, y por la proposición [3] sabemos

entonces que cualquier subsucesión que converja débilmente converge en la distancia d_2.

Si además suponemos que una de las probabilidades $\nu_1, ..., \nu_k \in \Gamma_{2,ac}$ tiene densidad acotada (podemos suponer que es ν_1 con densidad f_1), entonces de la nota 4 tenemos que si μ_n tiene densidad g_n,

$$\|g_n\|_\infty \leq \lambda_1^{-d} \|f_1\|_\infty$$

Se tiene por tanto que si $A \subset \mathbb{R}^d$ es medible, entonces:

$$\mu_n(A) = \int_A g_n(x)dx \leq \int_A \lambda_1^{-d}\|f_1\|_\infty dx \leq \lambda_1^{-d}\|f_1\|_\infty l_d(A)$$

siendo $l_d(A)$ la medida de Lebesgue de A en $\mathbb{R}^d$. Por tanto, si $\bar{\mu}$ es el límite para la convergencia débil de una subsucesión $\{\mu_{n_m}\}_{m=0}^\infty$, entonces

$$\bar{\mu}(A) \leq \liminf_{m\to\infty} \mu_{n_m}(A) \leq \lambda_1^{-d}\|f_1\|_\infty l_d(A)$$

De aquí deducimos que $\bar{\mu}$ tiene densidad, ya que para cada $x \in \mathbb{R}^d$,

$$\bar{\mu}(\{x\}) \leq \lambda_1^{-d}\|f_1\|_\infty l_d(\{x\}) = \lambda_1^{-d}\|f_1\|_\infty 0 = 0$$

Además la densidad de $\bar{\mu}$ está acotada superiormente por $\lambda_1^{-d}\|f_1\|_\infty$, ya que si no fuera así, en el conjunto $B = \{x \in \mathbb{R}^d : \text{ densidad de } \bar{\mu} \text{ en } x > \lambda_1^{-d}\|f_1\|_\infty\}$ no se cumpliría la desigualdad:

$$\bar{\mu}(B) \leq \lambda_1^{-d}\|f_1\|_\infty l_d(B)$$

a no ser que $l_d(B) = 0$, en cuyo caso B sería de medida nula y se podría asignar valor 0 a la función en los puntos de B, y tendríamos una función igual c.s. a la función de densidad original y que cumpliría lo dicho.
Como tenemos que $d_2(\mu_{n_m}, \bar{\mu}) \xrightarrow{m\to\infty} 0$, por la continuidad que vimos en la proposición [13] de la función G respecto de la distancia d_2 tenemos que:

$$d_2(G(\mu_{n_m}), G(\bar{\mu})) = d_2(\mu_{n_m+1}, G(\bar{\mu})) \xrightarrow{m\to\infty} 0$$

Por otra parte, se tiene también que la función:

$$V(\mu) = \sum_{j=1}^{k} \lambda_j d_2^2(\mu, \nu_j)$$

es una función continua respecto de la distancia d_2. Esto se deduce por ejemplo de la desigualdad:

$$|V(\mu_1)^{1/2} - V(\mu_2)^{1/2}| \leq d_2(\mu_1, \mu_2)$$

Para ver esta desigualdad, sea $\Omega = \{1, ..., n\}$ y P una probabilidad en Ω tal que $P(k) = \lambda_k$ para $k = 1, ..., n$. Sean U, V dos v.a. definidas en Ω tales que $U(k) = d_2(\mu_1, \nu_k), V(k) = d_2(\mu_2, \nu_k)$ para $k = 1, 2, ..., n$. Se tiene entonces que:

$$|V(\mu_1)^{1/2} - V(\mu_2)^{1/2}| = |E(U^2)^{1/2} - E(V^2)^{1/2}| = |\|U\|_2 - \|V\|_2| \leq$$

$$\leq \|U - V\|_2 = \left(\sum_{j=1}^{n} \lambda_j |d_2(\mu_1, \nu_j) - d_2(\mu_2, \nu_j)|^2 \right)^{1/2} \leq$$

$$\leq \left(\sum_{j=1}^{n} \lambda_j d_2^2(\mu_1, \mu_2) \right)^{1/2} = d_2(\mu_1, \mu_2)$$

Como consecuencia,

$$V(\mu_{n_m}) \xrightarrow{m\to\infty} V(\bar{\mu}) \quad y \quad V(\mu_{n_{m+1}}) \xrightarrow{m\to\infty} V(G(\bar{\mu}))$$

Por la proposición [14], $\{V(\mu_n)\}_{n=0}^{\infty}$ es una sucesión decreciente de números reales no negativos, luego converge. Por tanto:

$$V(\bar{\mu}) = \lim_{n\to\infty} V(\mu_{n_m}) = \lim_{n\to\infty} V(\mu_{n_{m+1}}) = V(G(\bar{\mu}))$$

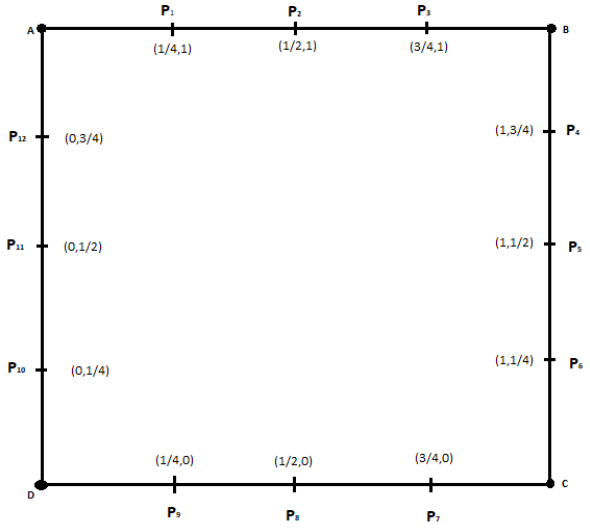
y de nuevo por la proposición [14], tenemos que $\bar{\mu} = G(\bar{\mu})$.

Por tanto hemos visto que bajo la suposición de que una de las ν_j tiene densidad acotada, el único baricentro $\bar{\mu}$ es un punto fijo de G. Si G tiene un único punto fijo, entonces como $\{\mu_n\}_{n=0}^{\infty}$ es ajustada, por el teorema de Helly [4] cualquier subsucesión de $\{\mu_n\}_{n=0}^{\infty}$ tiene otra subsucesión que converge débilmente, y como debe hacerlo hacia un punto fijo, converge hacia $\bar{\mu}$. Por lo tanto μ_n converge hacia $\bar{\mu}$ débilmente y además $d_2(\mu_n, \bar{\mu}) \xrightarrow{n\to\infty} 0$. □

Ejemplo. Vamos a ver mediante un ejemplo que el operador G puede tener más de un punto fijo. Consideramos el conjunto $S = \{A, B, C, D\}$, siendo

$$A = (0,1) \quad B = (1,1) \quad C = (1,0) \quad D = (0,0)$$

Para $s \in S$, sea ν_s la distribución uniforme en los puntos de $S \setminus \{s\}$, y sea μ_1 la distribución uniforme discreta con soporte en los 12 puntos P_i siguientes alrededor del cuadrado unidad:



Las aplicaciones $T_s, s \in S$ descritas en la siguiente tabla son las únicas aplicaciones de transporte óptimo entre μ_1 y $\nu_s, s \in S$. Representamos en cada caso los puntos yendo a cada punto de S para cada una de las cuatro aplicaciones de transporte óptimo.

Aplicación	A	B	C	D
T_A	-	P_1, P_2, P_3, P_4	P_5, P_6, P_7, P_8	$P_9, P_{10}, P_{11}, P_{12}$
T_B	P_{12}, P_1, P_2, P_3	-	P_4, P_5, P_6, P_7	P_8, P_9, P_{10}, P_{11}
T_C	P_{11}, P_{12}, P_1, P_2	P_3, P_4, P_5, P_6	-	P_7, P_8, P_9, P_{10}
T_D	$P_{10}, P_{11}, P_{12}, P_1$	P_2, P_3, P_4, P_5	P_6, P_7, P_8, P_9	-

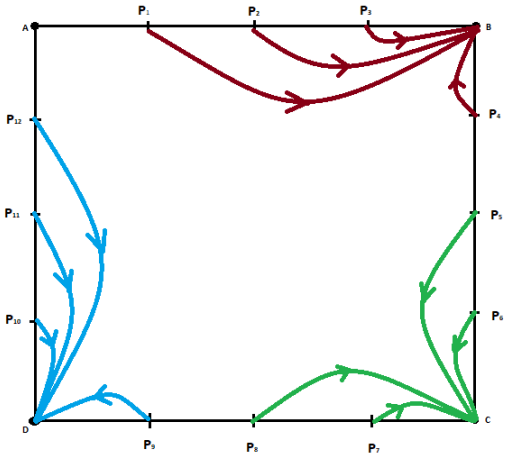
Tabla 1: Aplicaciones de transporte óptimo entre μ_1 y $\nu_s, s \in S$

Por ejemplo, si U es una variable aleatoria con $\mathscr{L}(U) = \mu_1$, el vector aleatorio $(U, T_A(U))$ tendrá ley de probabilidad conjunta:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
A	-	-	-	-	-	-	-	-	-	-	-	-
B	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	-	-	-	-	-	-	-	-
C	-	-	-	-	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	-	-	-	-
D	-	-	-	-	-	-	-	-	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Tabla 2: Distribución conjunta de $(U, T_A(U))$

Claramente $\mathscr{L}(T_A(U)) = \nu_A$. Además, fijándonos en el siguiente dibujo:



se ve que básicamente lo que hace la aplicación T_A es llevar la masa concentrada en los puntos de probabilidad positiva de μ_1 en los vértices de S distintos de A más cercanos a cada punto, con lo que en este caso queda

$$E\|U - T_A(U)\|^2 = \sum_{i=1}^{4} \frac{1}{12}\|P_i - B\|^2 + \sum_{i=5}^{8} \frac{1}{12}\|P_i - C\|^2 + \sum_{i=9}^{12} \frac{1}{12}\|P_i - D\|^2$$

Por tanto, esta distribución minimiza el valor de $E\|U - V\|^2$, luego T_A es una aplicación de transporte óptimo.

Dados $\delta > 0, \gamma > 0$, consideramos la distribución ν_s^δ, uniforme en la unión de las tres bolas de radio δ y centro en los puntos de $S \setminus \{s\}, s \in S$, y μ_1^γ uniforme en la unión de las 12 bolas de radio γ y centros en los puntos $P_1, P_2, ..., P_{12}$. Denotemos $T_s^{\gamma,\delta}$ a la aplicación de transporte óptimo entre μ_1^γ y ν_s^δ

Con la misma idea que en el dibujo de arriba se ve que si tomamos $\delta > 0, \gamma > 0$ suficientemente pequeños, entonces se tendrá que

$$T_s^{\gamma,\delta}(B(P_i, \gamma)) \subset B(T_s(P_i), \delta) \quad i = 1, 2, ..., 12$$

También podemos tomar $\gamma > \delta$ de forma que para todo $x \in \mathbb{R}^2$, la bola $B(x, \gamma)$ contiene al cuadrado de centro x y de lado δ. Consideramos entonces la aplicación:

$$x \longmapsto \bar{T}^{\gamma,\delta}(x) = \frac{1}{4}\sum_{s \in S} T_s^{\gamma,\delta}(x) := (t_1, t_2)$$

Si tenemos por ejemplo que $x \in B(P_1, \gamma)$ entonces de acuerdo con la tabla 1, tendremos que :

$$T_A^{\gamma,\delta}(x) \in B(B, \delta)$$

$$T_s^{\gamma,\delta}(x) \in B(A, \delta) \quad \forall s \in S \setminus \{A\}$$

Tenemos así que:

$$0{,}25 - \delta \leq t_1 \leq 0{,}25 + \delta$$

$$1 - \delta \leq t_2 \leq 1 + \delta$$

En consecuencia, $\bar{T}^{\gamma,\delta}(x)$ pertenece al cuadrado de lado δ y centro P_1, y por tanto:

$$\bar{T}^{\gamma,\delta}(B(P_1, \gamma)) \subset B(P_1, \gamma)$$

Esto mismo ocurre para el resto de puntos del soporte de μ_1^γ, y podemos concluir que:

$$\bar{T}^{\gamma,\delta}(B(P_i, \gamma)) \subset B(P_i, \gamma) \quad \forall\, i = 1, 2, ..., 12$$

Podemos además iterar el procedimiento para definir μ_1^* como el límite (a través de una subsucesión convergente) de la sucesión que se forma al tomar

$$\mu_{n+1}^\gamma = G_\delta(\mu_n^\gamma)$$

siendo G_δ el operador G que definimos anteriormente, asociado a la familia $\{\nu_s^\delta, s \in S\}$ y con pesos uniformes. Es decir:

$$\mu_2^\gamma = G_\delta(\mu_1^\gamma) = \mathscr{L}(\bar{T}^{\gamma,\delta}(U))$$

siendo U una v.a. con $\mathscr{L}(U) = \mu_1^\gamma$. Por lo que hemos visto antes, sabemos que μ_2^γ será una probabilidad concentrada en la unión de las bolas de centro P_i y radio γ, tal que:

$$\mu_2^\gamma(B(P_i,\gamma)) = \mu_1^\gamma(B(P_i,\gamma)) \quad \forall\, i = 1,2,...,12$$

Con un razonamiento inductivo, se puede ver que:

$$\mu_{n+1}^\gamma(B(P_i,\gamma)) = \mu_n^\gamma(B(P_i,\gamma)) \quad \forall\, n \in \mathbb{N},\, i = 1,2,...,12$$

y por tanto, si μ_1^* es el límite a través de una subsucesión convergente de $\{\mu_n\}_{n=1}^\infty$, se tendrá que:

$$\mu_1^*(B(P_i,\gamma)) = \mu_1^\gamma(B(P_i,\gamma)) \quad i = 1,2,...,12$$

y además por la proposición [15], $\mu_1^* \in \Gamma_{2,ac}$ y es un punto fijo de G_δ.

Busquemos ahora otro punto fijo de la función G_δ. Para ello, consideramos μ_2, la probabilidad con soporte en los puntos $P_1, P_2, ..., P_{13}$ siendo $P_1, P_2, ..., P_{12}$ los mismos que antes y $P_{13} = (1/2, 1/2)$. Sea $l = 3^{-1/2}$, y sea:

$$p = \frac{1}{2}(1-l)^2 \qquad q = (2l-1)(1-l) \qquad r = (2l-1)^2$$

y definimos:

$$\mu_2(P_i) = \begin{cases} p & si \quad i = 1,3,4,6,7,9,10,12 \\ q & si \quad i = 2,5,8,11 \\ r & si \quad i = 13 \end{cases}$$

Se tiene que realmente define una probabilidad ya que:

$$8p + 4q + r = 4(1-l)^2 + 4(2l-1)(1-l) + (2l-1)^2 =$$

$$= (2(1-l) + (21-l))^2 = (2 - 2l + 2l - 1)^2 = 1$$

Las funciones de transporte óptimo $T_s^*, s \in S$, entre μ_2 y cada una de las probabilidades $\nu_s, s \in S$, vienen descritas en la siguiente tabla:

Aplicación	A	B	C	D
T_A^*	-	P_1, P_2, P_3, P_4	$P_5, P_6, P_7, P_8, P_{13}$	$P_9, P_{10}, P_{11}, P_{12}$
T_B^*	P_{12}, P_1, P_2, P_3	-	P_4, P_5, P_6, P_7	$P_8, P_9, P_{10}, P_{11}, P_{13}$
T_C^*	$P_{11}, P_{12}, P_1, P_2, P_{13}$	P_3, P_4, P_5, P_6	-	P_7, P_8, P_9, P_{10}
T_D^*	$P_{10}, P_{11}, P_{12}, P_1$	$P_2, P_3, P_4, P_5, P_{13}$	P_6, P_7, P_8, P_9	-

Tabla 3: Aplicaciones de transporte óptimo entre μ_2 y $\nu_s, s \in S$

Las sumas que estamos haciendo todo el tiempo son las siguientes:

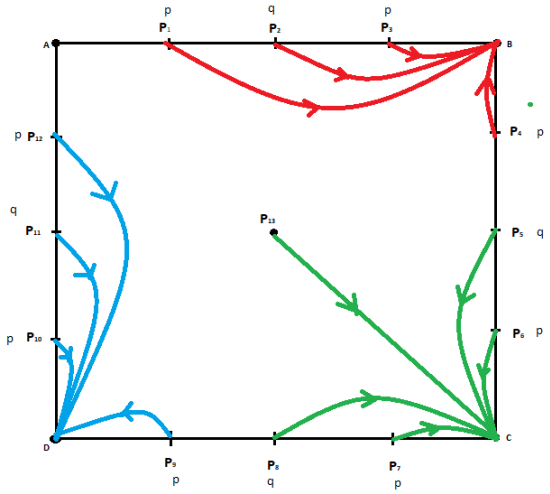
$$3p + q = 3(\frac{1}{2}(1-l)^2) + (2l-1)(1-l) = (1-l)(\frac{3}{2}(1-l) + (2l-1) =$$

$$= (1-l)(\frac{1}{2} + \frac{1}{2}l) = \frac{1}{2}(1+l)(1-l) = \frac{1}{2}(1-\frac{1}{3}) = \frac{1}{3}$$

$$2p + 2q + r = (1-l)^2 + 2(2l-1) + (2l-1)^2 = ((1-l) + (2l-1))^2 =$$

$$= l^2 = \frac{1}{3}$$

Luego efectivamente en cada caso la aplicación T_s^* lleva μ_2 en ν_s, para cada $s \in S$. Gráficamente, lo que hace por ejemplo T_A^* es:



Al igual que antes, se puede ver que si U es una variable aleatoria que sigue una distribución μ_2, entonces $(U, T_A^*(U))$ minimiza el valor de $E\|U-V\|^2$ entre todas las variables U,V con momento de orden 2 finito y tal que $\mathscr{L}(U) = \mu_2$ y $\mathscr{L}(V) = \nu_A$.
Se puede repetir un procedimiento análogo al que hicimos para hallar μ_1^*, para obtener a partir de μ_2 y de unos valores adecuados de γ, δ, una distribución μ_2^* absolutamente continua, que verifica:

$$\mu_2^*(B(P_i, \gamma)) = \mu_2(B(P_i, \gamma)) \quad i = 1, 2, ..., 13$$

y que además es un punto fijo de G_δ.

Tomando δ, γ los valores más bajos que hemos necesitado para que ambas construcciones funcionen, obtenemos dos puntos fijos distintos para el operador G_δ, ya que $\mu_1^* \neq \mu_2^*$.

3.2. Baricentros en familias de localización y escala

Recordemos que si $U = (U_1, U_2, ..., U_d)$ es un vector aleatorio con valores en $\mathbb{R}^d$, con media $m = (m_1, m_2, ..., m_d) \in \mathbb{R}^d$ y matriz de covarianzas Σ, entonces si $Z = AU + p$, donde A es una matriz de tamaño $d \times d$ y $p \in \mathbb{R}^d$, se cumplen las siguientes propiedades:

$$E(Z) = AE(U) + p \qquad Var(Z) = A\Sigma A^T$$

Sea $\mathscr{M}_{d\times d}^+$ el conjunto de las matrices reales, simétricas y definidas positivas de tamaño $d \times d$.

Definición 11. *Sea U_0 un vector aleatorio con ley de probabilidad $\mu_0 \in \Gamma_{2,ac}(\mathbb{R}^d)$. La familia*

$$\mathscr{F}(\mu_0) = \{\mathscr{L}(AU_0 + p) : A \in \mathscr{M}_{d\times d}^+, p \in \mathbb{R}^d\}$$

de leyes de probabilidad definidas por transformaciones definidas positivas de U_0 se dice que es una familia de localización y escala.

Por supuesto, una familia de localización y escala $\mathscr{F}(\mu_0)$ se puede parametrizar mediante los parámetros p y A. Notemos también que si m_0 y A_0 son la media y la matriz de covarianzas de U_0, entonces la familia puede definirse también a partir de la probabilidad

$$\mu_0^* = \mathscr{L}(A_0^{-1/2}(U_0 - m_0))$$

que por los comentarios anteriores sabemos que tiene media 0 y matriz de covarianzas Id. Esto nos permite parametrizar a la familia a través de la media y la matriz de covarianzas de las distribuciones de la familia, algo que asumiremos a partir de ahora. Con esta suposición, una ley de probabilidad de la familia $\mathscr{F}(\mu_0)$ se denotará en términos de su media m y su matriz de covarianzas Σ por $\mu_{m,\Sigma}$.

La aplicación dada por $T(x) = (m_2 - m_1) + Ax$, siendo

$$A = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}$$

cumple que si U es un vector aleatorio con $\mathscr{L}(U) = \mu_{m_1,\Sigma_1}$ entonces $E(T(U)) = m_2$ y además:

$$\mathrm{Var}(T(U)) = \left(\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}\right)\Sigma_1\left(\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}\right)^T =$$

$$= \left(\Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)\left(\Sigma_1^{-1/2}\Sigma_1\Sigma_1^{-1/2}\right)\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}\right) =$$

$$= \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2} = \Sigma_1^{-1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})\Sigma_1^{-1/2} = \Sigma_2$$

Por tanto, T transporta la probabilidad $\mathscr{L}(U) = \mu_{m_1,\Sigma_1}$ en la probabilidad $\mathscr{L}(U) = \mu_{m_2,\Sigma_2}$, luego por el teorema [1] sabemos que:

$$d_2^2(\mu_{m_1,\Sigma_1}, \mu_{m_2,\Sigma_2}) = \|m_1 - m_2\|^2 + \mathrm{tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right) =$$

$$= d_2^2(N(m_1, \Sigma_1), N(m_2, \Sigma_2)) \tag{10}$$

La última igualdad se deduce del hecho de que la distancia sólo depende de los vectores de medias y las matrices de covarianzas de las distribuciones de la familia de localización y escala en general, y de que sabemos que la familia de distribuciones normales d-dimensionales es un caso particular de familia de localización y escala.

Vamos a tratar ahora el problema de calcular el baricentro de una familia de distribuciones normales multivariantes no singulares, y por lo visto anteriormente, podremos generalizarlo a familias de distribuciones pertenecientes a la misma familia de localización y escala. En concreto, vamos a comenzar tratando el caso en que $\nu_1, \nu_2, ..., \nu_k$ son **distribuciones normales centradas y no degeneradas**, $\nu_j = N(0, \Sigma_j)$ para $j = 1, 2, ..., k$. Bajo estas condiciones, sabemos que el baricentro es único. Por otra parte, tenemos la siguiente igualdad, que es un caso particular de la igualdad (6), considerando $x = 0$,

$$\sum_{j=1}^{k} \lambda_j \|\bar{x} - x_j\|^2 = \sum_{j=1}^{k} \lambda_j \|x_j\|^2 - \|\bar{x}\|^2$$

Además, vimos al hablar de la existencia del baricentro que para hallar el baricentro teníamos que encontrar un mínimo entre todos los vectores aleatorios $(U_1, U_2, ..., U_k)$ definidos en algún espacio probabilístico (Ω, σ, P), con $\mathscr{L}(U_j) = N(0, \Sigma_j)$ para todo $j = 1, 2, ..., k$, del valor de:

$$\int_\Omega \sum_{j=1}^{k} \lambda_j \|\bar{U} - U_j\|^2 dP = \int_\Omega \sum_{j=1}^{k} \lambda_j \|U_j\|^2 dP - \int_\Omega \|\bar{U}\|^2 dP = [\boxtimes]$$

siendo $\bar{U} = \sum_{j=1}^{k} \lambda_j U_j$, y en ese caso el baricentro será $\bar{\mu} = \mathscr{L}(\bar{U})$. Por tanto, se trata de encontrar un mínimo de:

$$[\boxtimes] = \sum_{j=1}^{k} \lambda_j E\|U_j\|^2 - E\|\lambda_1 U_1 + ... + \lambda_k U_k\|^2$$

Como el primer sumando es fijo, ya que solo depende de las distribuciones marginales de las variables U_j, deducimos que el mínimo de este valor se da cuando se alcanza el valor máximo de

$$E\|\lambda_1 U_1 + ... + \lambda_k U_k\|^2 = \sum_{j=1}^{k} \lambda_j^2 E\|U_j\|^2 + 2 \sum_{1 \le j < l < \le k} \lambda_j \lambda_l E(U_j \cdot U_l)$$

De esta última expresión se ve que $E\|\lambda_1 U_1 + ... + \lambda_k U_k\|^2$ sólo depende de la estructura de covarianzas del vector $(k \times d)$-dimensional $(U_1, U_2, ..., U_k)$.

Dada cualquier estructura de covarianzas, se puede encontrar un vector aleatorio normal centrado con dichas estructura de covarianzas. Así, tenemos la existencia de un mínimo para el problema de encontrar una distribución de $(U_1, U_2, ..., U_k)$ con $\mathscr{L}(U_j) = N(0, \Sigma_j)$ para todo

$j = 1, 2, ..., k$, y por tanto existe un baricentro para $\nu_1, \nu_2, ..., \nu_k$ que será una distribución normal, ya que el baricentro es la ley de $\bar{U} = \sum_{j=1}^{k} \lambda_j U_j$, y este vector aleatorio en $\mathbb{R}^d$ es normal por ser combinación lineal de las componentes del vector aleatorio normal $(k \times d)$-dimensional $(U_1, U_2, ..., U_k)$, debido a las propiedades de la normal multivariante. Además, la distribución será centrada, por la linealidad de la esperanza, y no degenerada, ya que sabemos que tiene función de densidad, y si fuera degenerada estaría concentrada en un conjunto de medida nula de $\mathbb{R}^d$. Por tanto, el baricentro de una familia de distribuciones normales centradas no degeneradas centradas sigue una distribución normal que será además no degenerada y centrada.

Con esto en mente, abusando de notación escribimos $V(\Sigma)$ para referirnos a $V(N(0, \Sigma))$. El problema de hallar el baricentro consiste ahora en minimizar $V(\Sigma)$. Habíamos visto que:

$$d_2^2(\mu_{m_1,\Sigma_1}, \mu_{m_2,\Sigma_2}) = d_2^2(N(m_1, \Sigma_1), N(m_2, \Sigma_2)) =$$

$$= \|m_1 - m_2\|^2 + \operatorname{tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)$$

Por tanto:

$$V(\Sigma) = \sum_{j=1}^{k} \lambda_j \, d_2^2(N(0, \Sigma), N(0, \Sigma_j)) =$$

$$= \sum_{j=1}^{k} \lambda_j \operatorname{tr}\left(\Sigma + \Sigma_j - 2(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\right) =$$

$$= \operatorname{tr}(\Sigma) + \sum_{j=1}^{k} \lambda_j \operatorname{tr}(\Sigma_j) - 2\sum_{j=1}^{k} \lambda_j \operatorname{tr}((\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2})$$

Por los comentarios posteriores a las proposición [10], la aplicación biyectiva y definida positiva dada por:

$$T(x) = \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\Sigma^{-1/2}x$$

es la aplicación de transporte óptimo entre $N(0, \Sigma)$ y $N(0, \Sigma_j)$. Por tanto, si U es un vector aleatorio con distribución $N(0, \Sigma)$ el operador G que definimos en la sección anterior vendrá dado por:

$$G(N(0, \Sigma)) = \mathscr{L}\left((\sum_{j=1}^{k} \lambda_j \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\Sigma^{-1/2})U\right)$$

Que será una una normal centrada y con matriz de covarianzas dada por:

$$\left(\sum_{j=1}^{k} \lambda_j \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\Sigma^{-1/2}\right) \Sigma \left(\sum_{j=1}^{k} \lambda_j \Sigma^{-1/2}(\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2}\Sigma^{-1/2}\right)^T =$$

$$= \left(\Sigma^{-1/2}(\sum_{j=1}^{k} \lambda_j (\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2})\Sigma^{-1/2}\right) \Sigma \left(\Sigma^{-1/2}(\sum_{j=1}^{k} \lambda_j (\Sigma^{1/2}\Sigma_j\Sigma^{1/2})^{1/2})\Sigma^{-1/2}\right) =$$

$$= \Sigma^{-1/2} \left(\sum_{j=1}^{k} \lambda_j \, (\Sigma^{1/2} \Sigma_j \Sigma^{1/2})^{1/2} \right)^2 \Sigma^{-1/2}$$

Denotemos entonces en esta sección por G a la aplicación que asocia a cada matriz de covarianzas Σ la matriz de covarianzas de $G(N(0,\Sigma))$, es decir:

$$\Sigma \longmapsto \Sigma^{-1/2} \left(\sum_{j=1}^{k} \lambda_j \, (\Sigma^{1/2} \Sigma_j \Sigma^{1/2})^{1/2} \right)^2 \Sigma^{-1/2}$$

Proposición 16. *Con esta notación, si* $\Sigma \in \mathscr{M}^+_{d\times d}$*:*

$$V(\Sigma) - V(G(\Sigma)) \geq \operatorname{tr}\left((\Sigma^{1/2}(Id - H(\Sigma))^2 \Sigma^{1/2})^{1/2} \right)$$

y para cualquier $\Sigma' \in \mathscr{M}^+_{d\times d}$*:*

$$V(\Sigma') - V(\Sigma) \geq \operatorname{tr}\left((Id - H(\Sigma))(\Sigma - \Sigma') \right)$$

donde $H(\Sigma) = \sum_{j=1}^{k} \lambda_j \, \Sigma^{-1/2} (\Sigma^{1/2} \Sigma_j \Sigma^{1/2})^{1/2} \Sigma^{-1/2}$

De la primera desigualdad observamos que una condición necesaria para que $N(0,\Sigma)$ sea un baricentro es que $H(\Sigma) = Id$, ya que sólo en este caso el lado derecho de la primera desigualdad es 0.

De la segunda desigualdad, vemos además que la condición $H(\Sigma) = Id$ es suficiente para que $N(0,\Sigma)$ sea un baricentro, ya que si se cumple dicha condición, entonces para cualquier $\Sigma' \in \mathscr{M}^+_{d\times d}$,se tiene que $V(\Sigma) \leq V(\Sigma')$

Por tanto, como sabemos que el baricentro de normales no degeneradas existe, es único y además es una normal no degenerada, entonces lo que acabamos de probar es que el baricentro será $N(0,\Sigma)$ siendo Σ la única matriz solución de $H(S) = Id$, que es equivalente a la ecuación matricial:

$$S = \sum_{j=1}^{k} \lambda_j \, (S^{1/2} \Sigma_j S^{1/2})^{1/2}$$

Sin embargo, estamos lejos aun así de conseguir un método efectivo para el cálculo del baricentro. A través de la iteración introducida en la sección anterior, vamos a poder obtener un algoritmo consistente para aproximar el baricentro de distribuciones normales centradas, que nos servirá además para aproximar la solución $\Sigma \in \mathscr{M}^+_{d\times d}$ de la ecuación $H(\Sigma) = Id$. Antes de ver esto vamos a probar un lema:

Lema 4.

$$d_2^2(N(m_n,\Sigma_n), N(m,\Sigma)) \xrightarrow{n\to\infty} 0 \quad \Leftrightarrow \quad \|\Sigma_n - \Sigma\| \xrightarrow{n\to\infty} 0 \quad y \quad \|m_n - m\| \xrightarrow{n\to\infty} 0$$

Demostración. Para la demostración vamos a utilizar varios resultados conocidos:

- Cuando estamos trabajando con variables aleatorias normales unidimensionales, si P_n sigue una distribución $N(m_n, \sigma_n^2)$ para cada $n \in \mathbb{N}$, y P es una distribución $N(m, \sigma^2)$, entonces:
$$P_n \longrightarrow_d P \Leftrightarrow m_n \to m \text{ y } \sigma_n^2 \to \sigma^2$$
- Si $\{U_n\}_{n=1}^{\infty}$ son variables aleatorias para las que existe $\delta > 0$ tal que
$$\sup_{n\in\mathbb{N}} E\|U_n\|^{2+\delta} < \infty$$
entonces $\{\|U_n\|^2\}_{n=1}^{\infty}$ es uniformemente integrable.
- $\{\tilde{U}_n\}_{n=1}^{\infty}$, $\tilde{U}$ son vectores aleatorios en $\mathbb{R}^d$, entonces:
$$\tilde{U}_n \longrightarrow_d \tilde{U} \Leftrightarrow \tilde{\lambda}\tilde{U_n} \longrightarrow_d \tilde{\lambda}\tilde{U} \quad \forall\, \tilde{\lambda} \in \mathbb{R}^d$$

Sean $\tilde{P}$, $\tilde{P}_n$ distribuciones normales d-dimensionales con medias $\tilde{m}$ y $\tilde{m}_n$ y matrices de covarianzas Σ, Σ_n para cada $n \in \mathbb{N}$, y sean $\tilde{U}_n, \tilde{U}$ v.a. con $\mathscr{L}(\tilde{U}_n) = \tilde{P}_n, \mathscr{L}(\tilde{U}) = \tilde{P}$. Se tiene entonces que:

Si $d_2^2(\tilde{P}_n, \tilde{P}) \xrightarrow{n\to\infty} 0$, sabemos que $\tilde{U_n} \to_d \tilde{U}$. Por tanto:
$$\tilde{\lambda}\tilde{U_n} \longrightarrow_d \tilde{\lambda}\tilde{U} \quad \forall\, \tilde{\lambda} \in \mathbb{R}^d$$
y como $\tilde{\lambda}\tilde{U}_n, \tilde{\lambda}\tilde{U}$ son distribuciones normales con medias $\tilde{\lambda}\tilde{m}_n, \tilde{\lambda}\tilde{m}$ y varianzas $\tilde{\lambda}^T\Sigma_n\tilde{\lambda}, \tilde{\lambda}^T\Sigma\tilde{\lambda}$, se tiene que:
$$\tilde{\lambda}\tilde{m}_n \to \tilde{\lambda}\tilde{m} \qquad \tilde{\lambda}^T\Sigma_n\tilde{\lambda} \to \tilde{\lambda}\Sigma\tilde{\lambda} \quad \forall\, \tilde{\lambda} \in \mathbb{R}^d$$
y por tanto necesariamente
$$\tilde{m}_n \to \tilde{m} \qquad \Sigma_n \to \Sigma$$
Recíprocamente, si $\tilde{m}_n \to \tilde{m}$ y $\Sigma_n \to \Sigma$, entonces
$$\tilde{\lambda}\tilde{m}_n \to \tilde{\lambda}\tilde{m} \qquad \tilde{\lambda}^t\Sigma_n\tilde{\lambda} \to \tilde{\lambda}\Sigma\tilde{\lambda} \quad \forall\, \tilde{\lambda} \in \mathbb{R}^d$$
luego $\tilde{\lambda}\tilde{U}_n \longrightarrow_d \tilde{\lambda}\tilde{U} \quad \forall\tilde{\lambda} \in \mathbb{R}^d$ y por tanto $\tilde{U}_n \to_d \tilde{U}$. Además, como $\{\tilde{m}_n\}_{n=1}^{\infty}$ y $\{\Sigma_n\}_{n=1}^{\infty}$ son convergentes, para cada $\delta > 0$
$$\sup_{n\in\mathbb{N}} E\|\tilde{U}_n\|^{2+p} < \infty$$
y entonces $\{\|\tilde{U}_n\|^2\}_{n=1}^{\infty}$ es uniformemente integrable. Por las equivalencias de la proposición [3] se tiene entonces que $d_2^2(\tilde{P}_n, \tilde{P}) \xrightarrow{n\to\infty} 0$ □

Teorema 2. *Supongamos que $\Sigma_1, \Sigma_2, ..., \Sigma_k$ son matrices simétricas, semidefinidas positivas de tamaño $d \times d$ y con alguna de ellas definida positiva. Consideramos alguna matriz $S_0 \in \mathscr{M}_{d\times d}^{+}$ y definimos:*
$$S_{n+1} = S_n^{-1/2}\left(\sum_{j=1}^{k}\lambda_j\,(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}\right)^2 S_n^{-1/2}, \qquad n \geq 0$$

Si $N(0,\Sigma_0)$ es el baricentro de $N(0,\Sigma_1), N(0,\Sigma_2), ..., N(0,\Sigma_k)$, entonces

$$d_2(N(0,S_n), N(0,\Sigma_0)) \xrightarrow{n\to\infty} 0$$

Además, la matriz de covarianzas del baricentro cumple que:

$$det(\Sigma_0)^{1/2d} \geq \sum_{j=1}^{k} \lambda_j det(\Sigma_j)^{1/2d}$$

En particular, Σ_0 es definida positiva y es la única solución definida positiva de la ecuación

$$S = \sum_{j=1}^{k} \lambda_j \, (S^{1/2}\Sigma_j S^{1/2})^{1/2}$$

y también cumple que:

$$tr(S_n) \leq tr(S_{n+1}) \leq tr(\Sigma_0) \leq \sum_{j=1}^{k} tr(\Sigma_j)$$

dándose la igualdad en la última desigualdad si y sólo si $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$.

Demostración. Comencemos probando la desigualdad

$$\det(S_n)^{1/2d} \geq \sum_{j=1}^{k} \lambda_j \det(\Sigma_j)^{1/2d}$$

que en particular nos da el hecho de que S_n no es singular para ningún n, ya que

$$\sum_{j=1}^{k} \lambda_j \det(\Sigma_j)^{1/2d} > 0$$

puesto que $\det(\Sigma_j) \geq 0$ para todo $j = 1, 2, ..., k$, por ser todas las matrices definidas positivas, y además para algún j sabemos que la matriz Σ_j es definida positiva y por tanto la desigualdad es estricta. Por tanto, la secuencia está bien definida. Esta desigualdad es una consecuencia directa de la desigualdad de Minkowski para el determinante [6, Cor. II.3.21], que afirma que si A y B son matrices semidefinidas positivas, entonces:

$$(\det(A+B))^{1/n} \geq (\det(A))^{1/n} + (\det(B))^{1/n}$$

Por tanto, aplicando este resultado k veces obtenemos que

$$\left(\det(\sum_{j=1}^{k} \lambda_j (S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2})\right)^{1/d} \geq \sum_{j=1}^{k} \lambda_j \left(\det(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}\right)^{1/d}$$

Se tiene además que:

$$\left(\det(\sum_{j=1}^{k}\lambda_j(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2})\right)^{1/d} = \left(\det((S_n^{1/2}S_{n+1}S_n^{1/2})^{1/2})\right)^{1/d} = (\det(S_n))^{1/2d}(\det(S_{n+1}))^{1/2d}$$

$$\sum_{j=1}^{k}\lambda_j\left(\det(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}\right)^{1/2d} = (\det(S_n))^{1/2d}\sum_{j=1}^{k}\lambda_j\det((\Sigma_j))^{1/2d}$$

De estas igualdades y de la desigualdad de arriba se deduce lo que queríamos. Por otra parte, de la proposición [15] y de la relación existente entre la función G definida en esa sección y la definición de la aplicación G que hemos dado en esta sección, se deduce que la sucesión $\{N(0,S_n)\}_{n=1}^{\infty}$ es ajustada, lo cual implica que $\{S_n\}_{n=1}^{\infty}$ está acotada (lo están todas sus entradas). Tomamos una subsucesión convergente $S_{n_m} \xrightarrow{n\to\infty} \Sigma$. Por continuidad,

$$\det(\Sigma)^{1/2d} \geq \sum_{j=1}^{k}\det(\Sigma_j)^{1/2d} > 0$$

Esto muestra que $\Sigma \in \mathscr{M}_{d\times d}^{+}$.

La aplicación V es continua en $\mathscr{M}_{d\times d}^{+}$ ya que podemos escribir $V(S)$ en términos de la traza de S, luego

$$V(S_{n_m}) \xrightarrow{n\to\infty} V(\Sigma)$$

Sabemos que G es continua respecto de d_2, y por el lema anterior es continua en $\mathscr{M}_{d\times d}^{+}$. Por tanto:

$$V(G(S_{n_m})) = V(S_{n_m+1}) \xrightarrow{n\to\infty} V(G(\Sigma))$$

Como $\{V(S_n)\}_{n=1}^{\infty}$ es una sucesión no negativa y decreciente, sabemos que es convergente y por tanto:

$$V(\Sigma) = V(G(\Sigma))$$

En vista de la primera desigualdad de la proposición anterior, esto sólo puede ocurrir si

$$\Sigma^{1/2}(Id - H(\Sigma))^2\Sigma^{1/2} = 0 \Rightarrow H(\Sigma) = Id$$

ya que $\Sigma \in \mathscr{M}_{d\times d}^{+}$, luego $\Sigma^{1/2} \in \mathscr{M}_{d\times d}^{+}$ y por tanto es invertible.
Como habíamos visto que la única solución de $H(S) = Id, S \in \mathscr{M}_{d\times d}^{+}$, era la matriz de covarianzas del baricentro de $\nu_1, \nu_2, ..., \nu_k$, entonces $\Sigma = \Sigma_0$. De nuevo por la proposición [15] tenemos que:

$$d_2^2(N(0,S_n), N(0,\Sigma_0)) \xrightarrow{n\to\infty} 0$$

Por último, veamos la cadena de desigualdades:

$$\mathrm{tr}(S_n) \leq \mathrm{tr}(S_{n+1}) \leq \mathrm{tr}(\Sigma_0) \leq \sum_{j=1}^{k}\mathrm{tr}(\Sigma_j)$$

- Para la primera desigualdad, escribamos H_n a la matriz de la aplicación de transporte óptimo de $N(0, S_n)$ a $N(0, S_{n+1})$.

$$H_n = S_n^{-1/2}(S_n^{1/2}S_{n+1}S_n^{1/2})^{1/2}S_n^{-1/2} =$$

$$= S_n^{-1/2}\left(\sum_{j=1}^{k}\lambda_j\,(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}\right)S_n^{-1/2}$$

Se tiene que $S_{n+1} = H_n S_n H_n$ y en consecuencia:

$$(S_n^{1/2}S_{n+1}S_n^{1/2})^{1/2} = (S_n^{1/2}H_nS_nH_nS_n^{1/2})^{1/2} =$$

$$= \left(\left(\sum_{j=1}^{k}\lambda_j\,(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}\right)^2\right)^{1/2} = \sum_{j=1}^{k}\lambda_j\,(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2} = S_n^{1/2}H_nS_n^{1/2} \qquad (11)$$

De aquí concluimos que

$$d_2^2(N(0,S_n),N(0,S_{n+1})) = \mathrm{tr}(S_n) + \mathrm{tr}(S_{n+1}) - 2\mathrm{tr}((S_n^{1/2}S_{n+1}S_n^{1/2})^{1/2}) =$$

$$= \mathrm{tr}(S_n) + \mathrm{tr}(S_{n+1}) - 2\mathrm{tr}(S_n^{1/2}H_nS_n^{1/2}) = \mathrm{tr}(S_n) + \mathrm{tr}(S_{n+1}) - 2\mathrm{tr}(S_nH_n)$$

por la conmutatividad de la traza. Por otra parte:

$$V(S_n) = \mathrm{tr}(S_n) + \sum_{j=1}^{k}\lambda_j\,\mathrm{tr}(\Sigma_j) - 2\sum_{j=1}^{k}\lambda_j\,\mathrm{tr}((S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}) =$$

$$= \mathrm{tr}(S_n) + \sum_{j=1}^{k}\lambda_j\,\mathrm{tr}(\Sigma_j) - 2\mathrm{tr}\left(\sum_{j=1}^{k}\lambda_j\,(S_n^{1/2}\Sigma_j S_n^{1/2})^{1/2}\right) =$$

$$= \mathrm{tr}(S_n) + \sum_{j=1}^{k}\lambda_j\,\mathrm{tr}(\Sigma_j) - 2\mathrm{tr}(S_nH_n)$$

Luego se tiene que :

$$V(S_n) - V(S_{n+1}) =$$

$$= \left(\mathrm{tr}(S_n) + \sum_{j=1}^{k}\lambda_j\,\mathrm{tr}(\Sigma_j) - 2\mathrm{tr}(S_nH_n)\right) - \left(\mathrm{tr}(S_{n+1}) + \sum_{j=1}^{k}\lambda_j\,\mathrm{tr}(\Sigma_j) - 2\mathrm{tr}(S_{n+1}H_{n+1})\right) =$$

$$= (\mathrm{tr}(S_n) - 2\mathrm{tr}(S_nH_n)) - (\mathrm{tr}(S_{n+1}) - 2\mathrm{tr}(S_{n+1}H_{n+1})) \qquad (12)$$

Combinando las igualdades (11) y (12) con la desigualdad de la proposición [14] obtenemos que:

$$\mathrm{tr}(S_n) - 2\mathrm{tr}(S_nH_n) - \mathrm{tr}(S_{n+1}) + 2\mathrm{tr}(S_{n+1}H_{n+1}) = V(S_n) - V(S_{n+1}) \geq$$

$$\geq d_2^2(N(0,S_n),N(0,S_{n+1})) = \mathrm{tr}(S_n) + \mathrm{tr}(S_{n+1}) - 2\mathrm{tr}(S_nH_n)$$

lo cual implica que $2(\text{tr}(S_{n+1}H_{n+1}) - \text{tr}(S_{n+1})) \geq 0$, y por tanto

$$\text{tr}(S_{n+1}) \leq \text{tr}(S_{n+1}H_{n+1})$$

De aquí se deduce que:

$$0 \leq d_2^2(N(0,S_{n+1}),N(0,S_{n+2})) = (\text{tr}(S_{n+1}) - \text{tr}(S_{n+1}H_{n+1}))+(\text{tr}(S_{n+2}) - \text{tr}(S_{n+1}H_{n+1})) \leq$$

$$\leq \text{tr}(S_{n+2}) - \text{tr}(S_{n+1}H_{n+1})$$

lo que implica que $\text{tr}(S_{n+2}) \geq \text{tr}(S_{n+1}H_{n+1})$ y por tanto:

$$\text{tr}(S_{n+1}) \leq \text{tr}(S_{n+2})$$

- Veamos ahora que $\text{tr}(S_n) \leq \sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j)$. Notemos que:

$$0 \leq V(S_{n+1}) = (\text{tr}(S_{n+1}) - \text{tr}(S_{n+1}H_{n+1})) + \left(\sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j) - \text{tr}(S_{n+1}H_{n+1})\right)$$

Como hemos probado que $\text{tr}(S_{n+1}) \leq \text{tr}(S_{n+1}H_{n+1})$, entonces necesariamente:

$$\text{tr}(S_n) \leq \text{tr}(S_{n+1}H_{n+1}) \leq \sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j)$$

- La última de las desigualdades que hay que probar se deduce por continuidad. También se puede ver del hecho de que como $\Sigma_0 = \sum_{j=1}^k \lambda_j \, (\Sigma_0^{1/2}\Sigma_j\Sigma_0^{1/2})^{1/2}$, entonces

$$\text{tr}(\Sigma_0) = \sum_{j=1}^k \lambda_j \left(\text{tr}(\Sigma_0^{1/2}\Sigma_j\Sigma_0^{1/2})\right)^{1/2}$$

y por tanto

$$0 \leq V(\Sigma_0) = \text{tr}(\Sigma_0) + \sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j) - 2\sum_{j=1}^k \lambda_j \left(\text{tr}(\Sigma_0^{1/2}\Sigma_j\Sigma_0^{1/2})\right)^{1/2} =$$

$$= \text{tr}(\Sigma_0) + \sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j) - 2\text{tr}(\Sigma_0) = \sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j) - \text{tr}(\Sigma_0)$$

de donde deducimos que $\sum_{j=1}^k \lambda_j \text{tr}(\Sigma_j) \geq \text{tr}(\Sigma_0)$ y además la igualdad se da únicamente si $V(\Sigma_0) = 0$, es decir, si todas las distribuciones son iguales y por tanto:

$$\Sigma_1 = \Sigma_2 = ... = \Sigma_k$$

□

Nota 6. En algunos casos, las iteraciones convergen en un único paso al baricentro. Esto ocurre por ejemplo cuando estamos en dimensión 1, donde $\Sigma_0 \in \mathbb{R}$ es la única solución de la ecuación:

$$\Sigma_0 = \sum_{i=1}^{k} \lambda_i (\Sigma_0^{1/2} \Sigma_i \Sigma_0^{1/2})^{1/2}$$

$$\Sigma_0 = \sum_{i=1}^{k} \lambda_i \Sigma_0^{1/4} \Sigma_i^{1/2} \Sigma_0^{1/4}$$

$$\Sigma_0^{1/2} = \sum_{i=1}^{k} \lambda_i \Sigma_i^{1/2} \Rightarrow \Sigma_0 = \left(\sum_{i=1}^{k} \lambda_i \Sigma_i^{1/2} \right)^2$$

Más generalmente, si tenemos que $\Sigma_i \Sigma_j = \Sigma_j \Sigma_i$ para todo $i, j \in \{1, 2, ..., k\}$, entonces $\Sigma_i = U \Lambda_i U^T$ para alguna matriz ortogonal U y alguna matriz diagonal Λ_i, para cada $i = 1, ..., k$. Es este caso, se puede ver que la matriz del baricentro en la base dada por las filas de U es $\Lambda = \left(\sum_{i=1}^{k} \lambda_i \Lambda_i^{1/2} \right)^2$ y por tanto la matriz del baricentro en la base original es:

$$\Sigma_0 = U \Lambda U^T = U \left(\sum_{i=1}^{k} \lambda_i \Lambda_i^{1/2} \right)^2 U^T = U \left(\sum_{i=1}^{k} \lambda_i \Lambda_i^{1/2} \right) U^T U \left(\sum_{i=1}^{k} \lambda_i \Lambda_i^{1/2} \right) U^T =$$

$$= \left(\sum_{i=1}^{k} \lambda_i U \Lambda_i^{1/2} U^t \right)^2 = \left(\sum_{i=1}^{k} \lambda_i \Sigma_i^{1/2} \right)^2$$

En ambos casos, si comenzamos la iteración con $S_0 = Id$, en un paso tenemos que

$$S_1 = S_0^{-1/2} \left(\sum_{j=1}^{k} \lambda_j (S_0^{-1/2} \Sigma_j S_0^{1/2})^{1/2} \right)^2 S_0^{-1/2} = \left(\sum_{i=1}^{k} \lambda_i \Sigma_i^{1/2} \right)^2 = \Sigma_0$$

Por otra parte, el teorema [2] proporciona además una demostración de que dadas $\Sigma_1, \Sigma_2, ..., \Sigma_k \in \mathscr{M}_{d\times d}^{+}$, existe una única solución $\bar{S} \in \mathscr{M}_{d\times d}^{+}$ tal que

$$\bar{S} = \sum_{i=1}^{k} \lambda_i \left(\bar{S}^{1/2} \Sigma_i^{1/2} \bar{S}^{1/2} \right)^{1/2}$$

Esta ecuación matricial está estrechamente ligada al cálculo del baricentro. Sin embargo, también podemos aprovechar el algoritmo que hemos construido para aproximar el baricentro para obtener aproximaciones de la única solución en $\mathscr{M}_{d\times d}^{+}$ de esta ecuación.
La conclusión que obtenemos es que dada cualquier matriz $S_0 \in \mathscr{M}_{d\times d}^{+}$, si definimos una sucesión de matrices $\{S_n\}_{n=1}^{\infty}$ como hicimos en el teorema [2], entonces

$$\lim_{n\to\infty} S_n = \bar{S}$$

Así tenemos un método iterativo consistente para hallar la solución de la ecuación.

Veamos ahora a hallar el **baricentro de distribuciones normales que no están centradas** y con matrices de covarianzas no singulares, $N(m_1, \Sigma_1), ..., N(m_k, \Sigma_k)$. Por la proposición [8] sabemos que:

$$d_2^2(N(m,\Sigma), N(m_j,\Sigma_j)) = \|m - m_j\|^2 + d_2^2(N(0,\Sigma), N(0,\Sigma_j))$$

Deducimos por tanto que el baricentro en general es el baricentro de las correspondientes distribuciones centradas desplazado por $\sum_{j=1}^k \lambda_j m_j$, ya que el baricentro es el minimizador de:

$$\sum_{j=1}^k \lambda_j d_2^2(N(m,\Sigma), N(m_j,\Sigma_j)) = \sum_{j=1}^k \lambda_j \left(\|m - m_j\|^2 + d_2^2(N(0,\Sigma), N(0,\Sigma_j))\right) =$$

$$= \sum_{j=1}^k \lambda_j \|m - m_j\|^2 + \sum_{j=1}^k \lambda_j d_2^2(N(0,\Sigma), N(0,\Sigma_j))$$

y el mínimo del sumando de la izquierda se alcanza cuando $m = \sum_{j=1}^k \lambda_j m_j$, y el del sumando de la derecha se alcanza en el baricentro de las probabilidades centradas $N(0, \Sigma_1), ..., N(0, \Sigma_k)$, es decir, cuando Σ es la única solución definida positiva de la ecuación:

$$S = \sum_{j=1}^k \lambda_j \left(S^{1/2} \Sigma_j S^{1/2}\right)^{1/2}$$

Con todo esto en mente, vamos ahora a centrarnos en el problema de hallar el baricentro $\nu_1, \nu_2, ..., \nu_k$, siendo éstas probabilidades de una **familia de localización y escala general.**

Corolario 6. *Si $\nu_j = \mu_{m_j,\Sigma_j} \in \mathscr{F}(\mu_0), j = 1, 2, ..., k$ donde μ_0 tiene densidad, está centrada y su matriz de covarianzas es Id, entonces el baricentro de $\nu_1, \nu_2, ..., \nu_k$ con pesos $\{\lambda_j\}_{j=1}^k$ es μ_{m_0,Σ_0} siendo $m_0 = \sum_{j=1}^k \lambda_j m_j$ y Σ_0 la única solución definida positiva de la ecuación*

$$S = \sum_{j=1}^k \lambda_j (S^{1/2} \Sigma_j S^{1/2})^{1/2}$$

Además, si $S_0 \in \mathscr{M}_{d\times d}^+$ y definimos S_n por

$$S_{n+1} = S_n^{-1/2} \left(\sum_{j=1}^k \lambda_j \left(S_n^{1/2} \Sigma_j S_n^{1/2}\right)^{1/2} \right)^2 S_n^{-1/2}, \qquad n \geq 0$$

entonces $\|S_n - \Sigma_0\| \to 0$ y se cumple que

$$det(\Sigma_0)^{1/2d} \geq \sum_{j=1}^k \lambda_j det(\Sigma_j)^{1/2d}$$

$$tr(S_n) \leq tr(S_{n+1}) \leq tr(\Sigma_0) \leq \sum_{j=1}^{k} \lambda_j tr(\Sigma_j)$$

dándose la igualdad en la última desigualdad si y sólo si $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$.

Demostración. Para ver que el baricentro de $\mu_{m_1,\Sigma_1}, ..., \mu_{m_1,\Sigma_1}$ es μ_{m_0,Σ_0}, tenemos que ver que para cada $P \in \Gamma_2$ se cumple que:

$$\sum_{j=1}^{k} \lambda_j d_2^2(P, \mu_{m_j,\Sigma_j}) \geq \sum_{j=1}^{k} \lambda_j d_2^2(\mu_{m_0,\Sigma_0}, \mu_{m_j,\Sigma_j})$$

Supongamos que $P \in \Gamma_2$ tiene media m y matriz de covarianzas Σ. Teniendo en cuenta (10), la desigualdad (3) del teorema [1] y que sabemos que $N(m_0, \Sigma_0)$ es el baricentro de las probabilidades $N(m_1, \Sigma_1), ..., N(m_k, \Sigma_k)$ con pesos $\{\lambda_j\}_{j=1}^{k}$, tenemos que:

$$\sum_{j=1}^{k} \lambda_j d_2^2(P, \mu_{m_j,\Sigma_j}) \geq \sum_{j=1}^{k} \lambda_j \left(\|m - m_j\|^2 + \text{tr}(\Sigma + \Sigma_j - 2(\Sigma_j^{1/2} \Sigma \Sigma_j^{1/2})^{1/2} \right) =$$

$$= \sum_{j=1}^{k} \lambda_j d_2^2(N(m, \Sigma), N(m_j, \Sigma_j)) \geq \sum_{j=1}^{k} \lambda_j d_2^2(N(m_0, \Sigma_0), N(m_j, \Sigma_j)) =$$

$$= \sum_{j=1}^{k} \lambda_j d_2^2(\mu_{m_0,\Sigma_0}, \mu_{m_j,\Sigma_j})$$

que es lo que queríamos ver. El resto del corolario es consecuencia directa del teorema [2]. □

Este corolario puede aplicarse por ejemplo al caso de distribuciones uniformes sobre elipsoides. Es decir, el caso en que μ_0 es la distribución uniforme en $B(0, \sqrt{d+2}) \subset \mathbb{R}^d$, que se puede ver que sigue una distribución centrada y con matriz de covarianzas Id. Ahora, $\mu_{m,\Sigma}$ es la distribución uniforme sobre el elipsoide

$$E_d(m, \Sigma) = \{x \in \mathbb{R}^d : (x - m)^T \Sigma (x - m) \leq \sqrt{d+2}\}$$

y el corolario anterior admite la siguiente interpretación. En la familia de los conjuntos compactos y convexos con interior no vacío de $\mathbb{R}^d$, podemos considerar la métrica:

$$w(C, D) = d_2(U_C, U_D)$$

donde U_C denota la distribución uniforme en C. Dados $C_1, ..., C_k$ en dicha familia, el baricentro de dichos conjuntos será el conjunto compacto y convexo de interior no vacío C que minimiza

$$\sum_{j=1}^{k} \lambda_j w^2(C, C_j)$$

en el caso en que exista. En este contexto, el corolario anterior implica que el baricentro de una familia de elipsoides $E_d(m_1, \Sigma_1), E_d(m_2, \Sigma_2), ..., E_d(m_k, \Sigma_k)$ es el elipsoide $E_d(m_0, \Sigma_0)$ siendo $\mu_0 = \sum_{j=1}^{k} \lambda_j m_j$ y Σ_0 la única solución definida positiva de la ecuación:

$$S = \sum_{j=1}^{k} \lambda_j (S^{1/2} \Sigma_j S^{1/2})^{1/2}$$

Además, el corolario nos da un procedimiento iterativo para el cálculo de este baricentro de elipsoides.

3.3. Ejemplos de cálculo del baricentro

Veamos en primer lugar que cada distribución Gaussiana se puede relacionar a partir de su vector de medias y su matriz de covarianzas con un elipsoide. Sabemos que si una variable sigue una distribución multivariante en $\mathbb{R}^d$ con vector de medias m y matriz de covarianzas Σ, entonces su función de densidad es:

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{\frac{-1}{2}(x-m)' \Sigma^{-1} (x-m)}$$

Definición 12. *Dada una distribución en $\mathbb{R}^d$ con vector de medias μ y matriz de covarianzas Σ, se define la distancia de Mahalanobis de un punto $x \in \mathbb{R}^d$ al centro de la distribución μ como:*

$$d(x; \mu, \Sigma) = \sqrt{(x-m)' \Sigma^{-1} (x-m)}$$

Esta distancia lo que hace es tener en cuenta la "estructura de dispersión"dada por Σ que tiene la distribución. De la definición de la función de densidad para la normal multivariante, se deduce que cuanto mayor sea la distancia de Mahalanobis de un punto al centro de la distribución, menos valdrá la función de densidad en ese punto. Además, los conjuntos de equidensidad son los formados por aquellos puntos que tienen la misma distancia de Mahalanobis al centro de la distribución, es decir, son de la forma:

$$\{x : (x-m)' \Sigma^{-1} (x-m) = cte\}$$

que son elipsoides en $\mathbb{R}^d$.

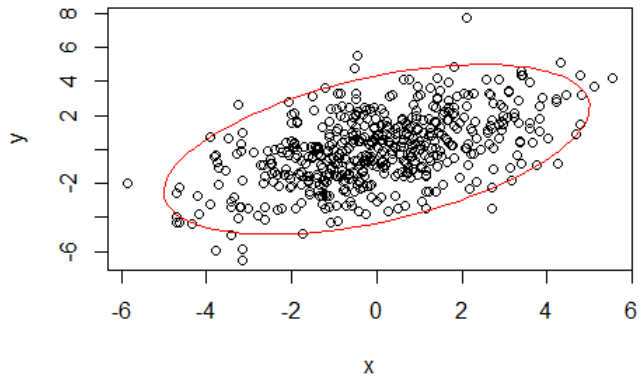
Esto hace que al tomar una muestra aleatoria de una distribución normal, los valores que se observan tomen la forma de los elipsoides que definen su vector de medias y su matriz de covarianzas. Por ejemplo, si tomamos en $\mathbb{R}^2$ la distribución normal centrada y con matriz de covarianzas

$$\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$$

y generamos una muestra aleatoria de 200 observaciones de esta distribución, al representarla junto al conjunto

$$\{(x, y) \in \mathbb{R}^2 : (x, y)' \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}^{-1} (x, y) = \frac{5}{2}\}$$

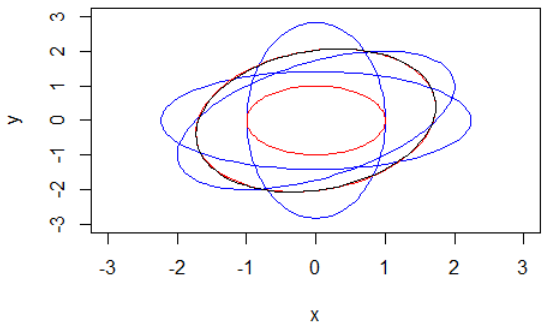
nos queda el siguiente gráfico:



Veamos ahora un ejemplo de computación del baricentro para distribuciones Gaussianas centradas en $\mathbb{R}^2$, con matrices de covarianzas

$$\Sigma_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 8 \end{pmatrix} \quad \Sigma_3 = \begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix}$$

y con pesos uniformes. De acuerdo con la explicación previa, vamos a representar cada distribución a partir de su elipsoide asociado. Vamos a aplicar el procedimiento iterativo que hemos explicado en la sección anterior hasta que la diferencia $V(S_{n+1}) - V(S_n)$ sea menor que una cierta tolerancia, que hemos tomado igual a 10^{-10}. La gráfica siguiente muestra los elipsoides asociados a las distribuciones normales que aparecen. En azul, los correspondientes a $\Sigma_1, \Sigma_2, \Sigma_3$. En rojo, los resultados que nos van dando las distintas iteraciones. Por último, en negro queda representado el elipsoide asociado al baricentro.



Al ejecutar el programa, se ve además que el número de iteraciones necesarias para alcanzar el baricentro en esta situación tan sólo ha sido 7.

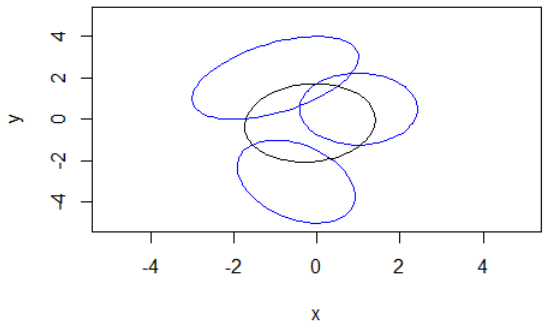
Veamos ahora un ejemplo de computación del baricentro para distribuciones Gaussianas en $\mathbb{R}^2$, con pesos uniformes, vectores de medias dados por:

$$m_1 = \begin{pmatrix} -1 \\ 2 \end{pmatrix} \quad m_2 = \begin{pmatrix} -1/2 \\ -3 \end{pmatrix} \quad m_3 = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}$$

y con matrices de covarianzas:

$$\Sigma_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix} \quad \Sigma_3 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

En este caso, tras 6 iteraciones el algoritmo llega al baricentro. La gráfica muestra en azul los elipsoides asociados a las tres distribuciones normales de las que partíamos, y en negro el elipsoide asociado al baricentro.



4. k-baricentros y k-baricentros recortados en espacios de Wasserstein

El baricentro de un conjunto de probabilidades $\nu_1, ..., \nu_n \in \Gamma_2(\mathbb{R}^d)$ se ha definido como una media Fréchet en dicho espacio y nos sirve como representante de las probabilidades $\nu_1, ..., \nu_n$. La media Fréchet admite una generalización en muchas ocasiones que nos sirve para resolver el problema de buscar k representantes en un conjunto de elementos de un espacio métrico. El caso más sencillo es cuando tenemos un conjunto de puntos $x_1, x_2, ..., x_n \in \mathbb{R}^d$. En ese caso se definen las **k-medias** de $x_1, x_2, ..., x_n$ como los k centros óptimos $m_1^*, ..., m_k^*$ que verifican:

$$\{m_1^*, ..., m_k^*\} = \arg\min_{m_1,...,m_k} \sum_{i=1}^{n} \min_{j=1,...,n} \|x_i - m_j\|^2$$

Se puede probar la existencia de estos k centros óptimos $m_1^*, ..., m_k^*$. Aunque no existen condiciones sencillas para garantizar la unicidad, se suele asumir que ésta se verifica.

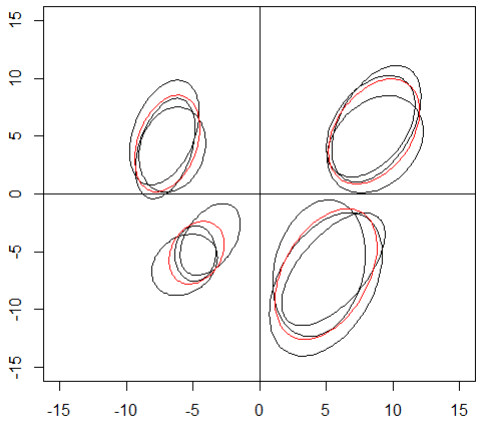
Con las mismas ideas podemos generalizar el baricentro de un conjunto de probabilidades $\nu_1, ..., \nu_n$. El concepto de los k-baricentros nos va a permitir obtener k representantes de las probabilidades $\nu_1, ..., \nu_n$.

Definición 13. *Un k-baricentro de las probabilidades $\nu_1, ..., \nu_n$ con pesos $\{\lambda_i\}_{i=1}^n$ es cualquier conjunto de k probabilidades $\{\bar{\mu}_1, ..., \bar{\mu}_k\} \subset \Gamma_2(\mathbb{R}^d)$ que verifique que para cada conjunto $\{\mu_1, ..., \mu_k\} \subset \Gamma_2(\mathbb{R}^d)$:*

$$\sum_{i=1}^{n} \lambda_i \min_{j=1,...,k} d_2^2(\nu_i, \bar{\mu}_j) \leq \sum_{i=1}^{n} \lambda_i \min_{j=1,...,k} d_2^2(\nu_i, \mu_j)$$

La existencia de los k-baricentros se puede probar bajo condiciones generales. Al igual que ocurre para las k-medias, no hay condiciones sencillas disponibles para garantizar la unicidad, pero se suele asumir que esta condición se cumple.

El cálculo de los k-baricentros en familias de localización y escala se puede llevar a cabo computacionalmente a partir del algoritmo que se desarrolla en [13]. El siguiente ejemplo muestra el 4-baricentro que hemos obtenido a partir de un conjunto de 12 distribuciones normales 2-dimensionales, en el que todas las distribuciones vienen representadas gráficamente a partir de sus elipsoides correspondientes.



En [13] se introduce además una extensión de este concepto mediante la consideración de los k-baricentros recortados. Estos siguen la misma idea que los k-baricentros, pero aquí mediante un proceso de reasignación de los pesos, se permite descartar una parte de los datos que tenemos. La definición formal es:

Definición 14. *Un k-baricentro recortado de nivel α de las probabilidades $\nu_1, ..., \nu_n$ con pesos $\{\lambda_i\}_{i=1}^n$ es un conjunto de k probabilidades $\{\bar{\mu}_1, ..., \bar{\mu}_k\} \subset \Gamma_2(\mathbb{R}^d)$ para el que existen unos pesos $\{\bar{\lambda}_i\}_{i=1}^n \in C_{\{\lambda_i\}_{i=1}^n}$, donde*

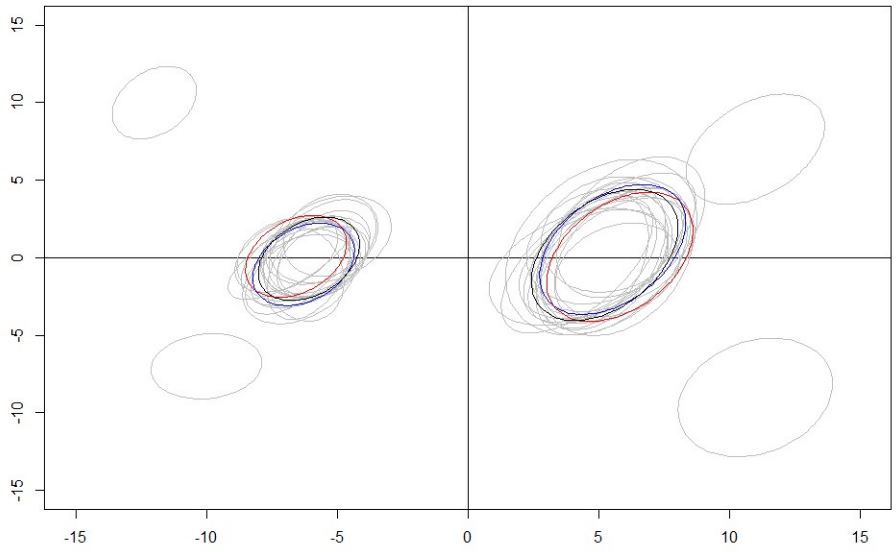
$$C_{\{\lambda_i\}_{i=1}^n} = \left\{ \{\lambda_i^*\}_{i=1}^n : 0 \leq \lambda_i^* \leq \frac{\lambda_i}{1-\alpha}, \quad \sum_{i=1}^n \lambda_i^* = 1 \right\}$$

tal que para cada conjunto $\{\mu_1, ..., \mu_k\} \subset \Gamma_2(\mathbb{R}^d)$ y para cada $\{\lambda_i^*\}_{i=1}^n \in C_{\{\lambda_i\}_{i=1}^n}$ se verifica

$$\sum_{i=1}^n \bar{\lambda}_i \min_{j=1,...,k} d_2^2(\nu_i, \bar{\mu}_j) \leq \sum_{i=1}^n \lambda_i^* \min_{j=1,...,k} d_2^2(\nu_i, \mu_j)$$

Esta definición generaliza el concepto de baricentro recortado que aparece en [18], que se correspondería con el caso $k = 1$, y no es más que una extensión del baricentro en la que permitimos descartar una proporción α de los datos. En [18] se estudian las principales propiedades del baricentro recortado, y se da un algoritmo para su computación.

El cálculo de los k-baricentros recortados se hace mediante el algoritmo que aparece en [13], que generaliza el anterior. La principal ventaja de los k-baricentros recortados frente a los k-baricentros es que poseen una mayor estabilidad cuando aparecen valores atípicos. En la sección siguiente, cuando trabajemos con ejemplos de aplicación de estas técnicas, veremos la importancia que tiene esta estabilidad. De momento, vamos a comparar los resultados que se obtienen cuando trabajamos con distintos niveles de α, teniendo en cuenta que el caso $\alpha = 0$ coincide con los k-baricentros que ya habíamos definido antes. En la siguiente gráfica, el elipsoide rojo representa el 2-baricentro de las 20 gaussianas con pesos uniformes correspondientes a los elipsoides grises, el azul el 2-baricentro recortado de nivel $\alpha = 0{,}05$ y el negro el 2-baricentro recortado de nivel $\alpha = 0{,}1$.



Los resultados de la gráfica se corresponden a los siguientes hechos:

- El 2-baricentro no modifica los pesos, luego da pesos uniformes a todas las probabilidades. El resultado que nos queda son 2 baricentros, de los cuales el de la izquierda es el baricentro con pesos uniformes de todas las probabilidades que aparecen a la izquierda del eje de ordenadas y el de la derecha es el baricentro con pesos uniformes de todas las probabilidades que aparecen a la derecha del eje de ordenadas. Esto hace que las 2 probabilidades que aparecen solas a la derecha desvíen el baricentro de la derecha un poco hacia la derecha del centro de los elipsoides sobre los que está situado. Para el baricentro de la izquierda, ocurre algo análogo.
- Al hacer el 2-baricentro recortado de nivel $\alpha = 0{,}05$, el resultado del algoritmo da peso 0 a los probabilidades correspondientes a los elipsoides de arriba a la izquierda y de abajo a la derecha, y uniforme al resto, y nos da los baricentros con pesos uniformes del resto de elipsoides de la derecha y de la izquierda independientemente. Así perdemos parte de la desviación de los baricentros sobre el centro de los elipsoides, pero no toda.
- Cuando hacemos el 2-baricentro recortado de nivel $\alpha = 0{,}1$, el resultado del algoritmo da peso 0 a las probabilidades correspondientes a los 4 elipsoides de las esquinas, y uniformes al resto. Los dos baricentros que nos da son los baricentros de las probabilidades sobre las que se encuentran cada uno de sus elipsoides. De esta forma, hemos perdido la desviación debida a las 4 probabilidades que están en las esquinas.

5. Aplicaciones: Análisis cluster

El análisis cluster es un procedimiento multivariante que busca las clasificación de n individuos en grupos (clusters), de forma que los individuos de cada grupo sean lo más homogéneos posibles entre sí y lo más diferentes posibles a los individuos de los otros grupos.

La clasificación es uno de los objetivos principales de la ciencia, y el análisis cluster nos proporciona un método efectivo para llevarla a cabo. Esto justifica la importancia de esta técnica, que presenta aplicaciones prácticas en numerosos campos:

- En taxonomía, para agrupar animales o plantas en diferentes especies.
- En medicina tiene múltiples utilidades. Por ejemplo, para definir diferentes categorías de tumores por sus propiedades, y poder aplicar posteriormente en cada caso el tratamiento más adecuado.

- En astronomía, se utiliza para agrupar estrellas o galaxias atendiendo a sus propiedades.
- En fotografía, agrupando píxeles para buscar diferentes objetos en imágenes. Los reproductores de video también utilizan análisis cluster para buscar escenas en una película.
- Las grandes empresas agrupan a clientes con patrones de consumo parecido.

Para llevar a cabo el análisis cluster, partimos de un conjunto de n individuos de los que se han medido d variables. Cada individuo puede verse como un punto de $\mathbb{R}^d$. Los métodos que se utilizan pueden dividirse en dos:

- Métodos jerárquicos: Dada una distancia d en $\mathbb{R}^k$, llamada índice de disimilaridad, se calculan las distancias entre todos los individuos. En cada paso se van uniendo individuos, o grupos, de acuerdo con un determinado criterio, que se llama índice de agregación. Éstos métodos tienen la ventaja de que no nos exigen fijar el número de grupos a buscar. Sin embargo, su elevado coste computacional hacen que sean poco útiles cuando estamos con muestras de gran tamaño.

- Métodos no jerárquicos: Fijado previamente un número k, estos métodos buscan agrupar los individuos en k grupos optimizando algún criterio. Son más eficientes computacionalmente, y son los que más se usan. Vamos a explicar algunos de los más importantes:

1. Método de las k-medias:
   Vimos que este método nos permite obtener k centros óptimos $m_1^*, ..., m_k^*$ (que serán los k puntos centrales en las nubes de puntos de cada uno de los grupos) tales que:

$$\{m_1^*, ..., m_k^*\} = \underset{m_1,...,m_k}{\arg\min} \sum_{i=1}^{n} \underset{j=1,...,n}{\text{mín}} \|x_i - m_j\|^2$$

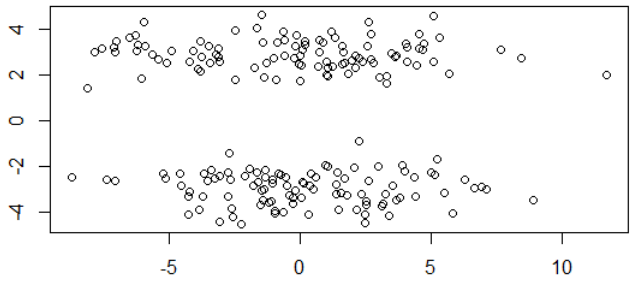
   Los clusters se obtienen agrupando mediante criterios de proximidad con las k-medias:

$$\mathsf{cluster}J = \{x_i : \|x_i - m_J^*\|^2 \leq \|x_i - m_j^*\|^2 \quad \forall\, j \neq J\}$$

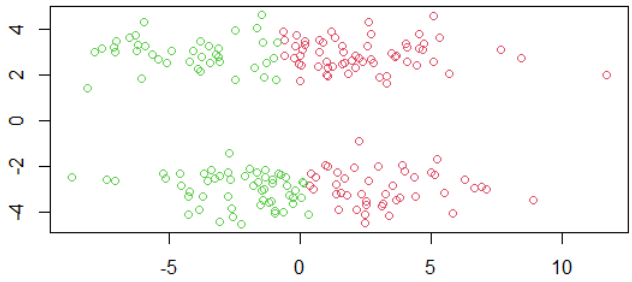
   Es equivalente a buscar k centros $m_1, ..., m_k$ y una partición $\{R_1, ..., R_k\}$ de $\{1, ..., n\}$ minimizando

$$\sum_{j=1}^{k} \sum_{i \in R_j} \|x_i - m_j\|^2$$

   Este método se puede llevar a cabo mediante un algoritmo sencillo que nos permite obtener las k-medias en pocas iteraciones. La principal desventaja de este método es que busca clusters esféricos, y con dispersiones parecidas. En el siguiente ejemplo, aparecen dos grupos que a simple vista están bien diferenciados.



Sin embargo, al hacer las k-medias con R a través de la orden *kmeans*, nos da una separación en grupos distinta de la esperada:



Si queremos tener más libertad a la hora de hacer grupos con distintas formas, tenemos que buscar otro tipo de métodos más versátiles:

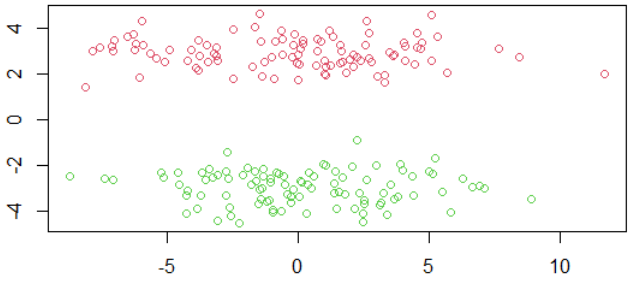
2. Métodos basados en maximizar la verosimilitud de la clasificación:
En este método se buscan k centros $m_1, ..., m_k$, k matrices de covarianzas $S_1, ..., S_k$ y una partición $\{R_1, ..., R_k\}$ de $\{1, ..., n\}$ maximizando

$$\sum_{j=1}^{k} \sum_{i \in R_j} \phi(x_i; m_j, S_j)$$

donde $\phi(x_i; m_j, S_j)$ es la densidad de una distribución normal d-dimensional, con centro m_j y matriz de covarianzas S_j. De esta forma permitimos que los grupos tengan estructuras de dispersión distintas, que vienen dadas por las matrices de covarianzas $S_1, ..., S_k$.

Las k-medias son un caso particular de estos métodos. Obtenemos las k-medias cuando maximizamos la verosimilitud imponiendo $S_1 = ... = S_k = s \cdot I_d$, con $s \in \mathbb{R}^+$.

El programa *tclust* (ver [16] para más detalles) en R nos permite llevar a cabo este procedimiento. El resultado que nos da para el conjunto de datos que teníamos antes ahora sí coincide con lo esperado:



La orden *tclust* nos permite además despreciar una proporción α de los datos, con lo que se mejora la estabilidad del método frente a los posibles valores atípicos que puedan aparecer. Esto lo utilizaremos en los ejemplos de la próxima sección.

5.1. Paralelización en el análisis cluster

En la actualidad, muchos de los problemas presentes en la computación involucran grandes cantidades de datos. Las limitaciones computacionales provocan que la resolución de estos problemas no se pueda llevar a cabo en un sólo ordenador en un tiempo razonable. Por ello es muy importante encontrar métodos que nos permitan paralelizar estos problemas, para poder encontrar su solución trabajando con varios ordenadores a la vez.

El ejemplo más sencillo de paralelización ocurre cuando estamos buscando la media de un conjunto de n datos $\{x_1, ..., x_n\} \subset \mathbb{R}^d$. Sea $\{R_1, ..., R_k\}$ una partición de $\{1, ..., n\}$ en k grupos tal que $|R_j| = n_j$, $n_1 + ... + n_k = n$. Entonces si calculamos por separado la media de los elementos de cada R_j:

$$m_j = \frac{\sum_{i \in R_j} x_i}{n_j}$$

entonces podemos calcular la media de $\{x_1, ..., x_n\}$ de forma sencilla:

$$m = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{n_1 \frac{\sum_{i \in R_1} x_i}{n_1} + ... + n_k \frac{\sum_{i \in R_k} x_i}{n_k}}{n} = \frac{n_1 m_1 + ... + n_k m_k}{n}$$

Este método tiene además la ventaja de que si nos añaden otros p datos cuando ya hemos calculado la media de los n datos, haciendo lo mismo que antes, podemos calcular la media de

los $n + p$ datos de forma sencilla.

Para llevar a cabo la **paralelización en el análisis cluster**, nos gustaría hacer algo parecido con las k-medias, o con las soluciones del modelo de maxima verosimilitud. Este razonamiento sencillo que hemos utilizado para el caso de la media no lo vamos a poder reproducir. La solución que vamos a dar para este problema será a partir de los k-baricentros.

Supongamos que tenemos un conjunto de n datos. Al aplicar la orden *tclust* para hallar k grupos, nos devuelve k centros y k matrices de covarianzas, que son los maximizadores del modelo de máxima verosimilitud. Es decir, nos da los parámetros de k distribuciones normales $P_1, ..., P_k$. Conociendo éstos, podemos llevar a cabo la partición en clusters del conjunto de datos, relacionando cada punto con la distribución para la cual la distancia de Mahalanobis del punto al centro sea menor.

Si dividimos nuestros conjunto de datos en m grupos con n_j individuos cada uno, podemos aplicar la orden *tclust* en cada uno de ellos para hallar la partición de esos datos en k grupos. Al igual que antes, tenemos así k distribuciones normales asociadas a dicha partición, $P_1^j, ..., P_k^j$, para cada $j = 1, ..., m$. Si la partición de los datos es razonable, cada una de las probabilidades representará uno de los grupos que habíamos hallado al principio, y por tanto será similar a una de las $P_1, ..., P_k$. Nuestro objetivo ahora sería recuperar las distribuciones $P_1, ..., P_k$ a partir de las $P_1^j, ..., P_k^j$, para cada $j = 1, ..., m$. Para ello, lo que haremos será calcular el k-baricentro de dichas probabilidades con pesos $\frac{n_j}{k \cdot n}$ para cada una de las probabilidades $P_i^j, i = 1, ..., k$, $j = 1, ..., m$, obteniendo así unas distribuciones $P_1^*, ..., P_k^*$. La clasificación en grupos de nuestros datos se hará entonces a partir del mínimo de las distancias de Mahalanobis de cada uno de los datos a las distribuciones $P_1^*, ..., P_k^*$.

Para una mayor estabilidad del método, lo que se hace generalmente es permitir a la orden *tclust* despreciar una proporción α de los datos en cada conjunto de datos de la partición, y posteriormente calcular los k-baricentros recortados de nivel β. Además, cuando hallamos todas las distancias de Mahalanobis, podemos estudiar cuál es el porcentaje γ que está peor representado por esos grupos. Estos serán los puntos para los que el mínimo de las distancias de Mahalanobis a los centros de las distribuciones $P_1^*, ..., P_k^*$ sea mayor, que los podemos dejar sin clasificar. Así es como procederemos en los ejemplos siguientes.

5.1.1. Ejemplo artificial de paralelización

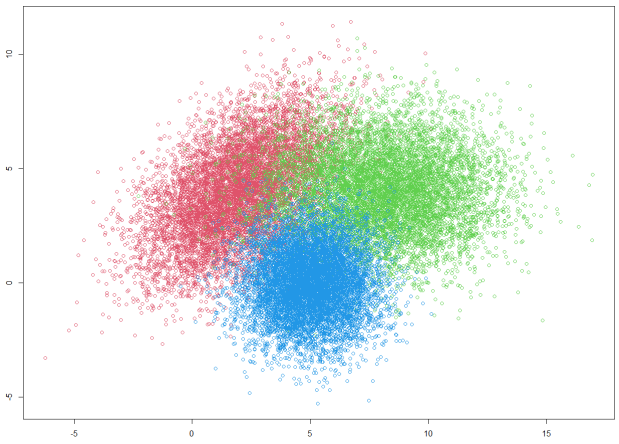
Para entender mejor el método de paralelización en el análisis cluster, vamos a llevarlo a cabo en un ejemplo sencillo en dos dimensiones, para poder visualizarlo sobre el plano. Para ello, consideramos una muestra de tamaño $m = 25000$, de los cuales un 33 % de los datos provienen de una normal de media m_1 y matriz de covarianzas Σ_1, un 33 % de los datos provienen de una normal de media m_2 y matriz de covarianzas Σ_2 y un 34 % de los datos provienen de una

normal 2-dimensional de media m_3 y matriz de covarianzas Σ_3, siendo:

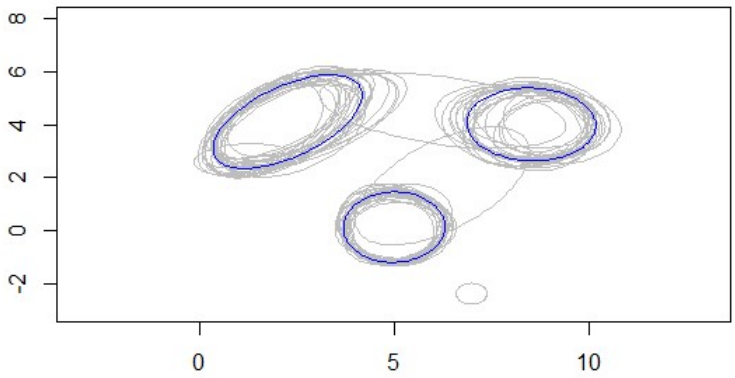
$$m_1 = (2,4) \quad m_2 = (8,4) \quad m_3 = (5,0)$$

$$\Sigma_1 = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} \quad \Sigma_2 = \begin{pmatrix} 6 & 0 \\ 0 & 3 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

La muestra nos queda:

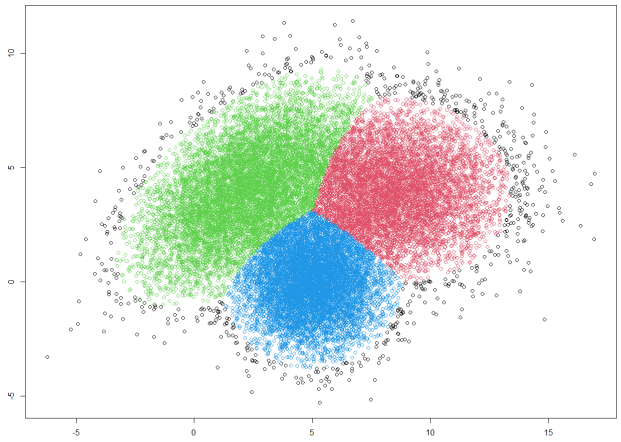


Buscaremos entonces separar la muestra en 3 grupos. Nuestro tamaño de paralelización será $u = 25$. Es decir, tomaremos una partición de la muestra de m datos en 25 submuestras de $\frac{m}{u} = 1000$ datos, y en cada una de las paralelizaciones calcularemos el resultado de la orden *tclust* sobre esa submuestra, permitiendo que se desprecie en cada caso una proporción $\alpha = 0{,}05$ de los datos. Cada paralelización nos da como resultado 3 elipsoides, y al calcular el 3-baricentro recortado de nivel $\beta = 0{,}1$ de los $3 \cdot 25 = 75$ elipsoides obtenemos las 3 probabilidades que buscábamos:

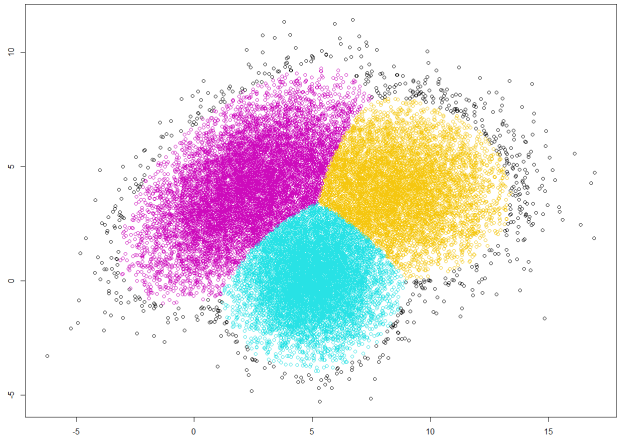


Vemos que hay elipsoides que son significativamente diferentes que los del resto de grupos. En las paralelizaciones asociadas a esos grupos, la submuestra asociada tendrá una forma distinta a la que tiene la muestra total. Esto es lógico que ocurra, ya que las submuestras se han cogido al azar, luego es razonable que alguna de ellas no sea un buen representante de la muestra total. La ventaja de utilizar el k-baricentro recortado es que permite que los resultados de esas paralelizaciones no influyan en el resultado final, obteniendo así una mayor estabilidad.

Realizando la asignación de cada uno de los datos de la muestra a cada grupo a partir de las distancias de Mahalanobis a los 3-baricentros, y dejando el 2.5 % peor representado sin clasificar, nos quedan los siguientes grupos:



Comparando estos resultados con los que nos da la orden *tclust* al ejecutarla sobre la muestra total, vemos que la separación en grupos es muy similar



Cruzando los grupos dados por cada uno de los métodos nos queda la tabla siguiente, en la que la fila y la columna S.C. representan los elementos sin clasificar:

Grupo	S.C.	1	2	3
S.C.	589	0	17	19
a	30	8700	151	318
b	1	0	8838	41
c	5	0	1	6290

Es decir, tan sólo se clasifican de forma distinta el

$$\frac{19+17+30+318+151+41+1+5+1}{25000}\cdot 100\,\% = 2{,}332\,\%$$

de los datos.

5.1.2. Ejemplo cervezas artesanales

Veamos un ejemplo real de la utilidad que pueden tener este tipo de técnicas para la solución de problemas reales, como el de la clasificación de cervezas artesanales. En los últimos años ha habido un incremento notable del interés por este tipo de cervezas, y en consecuencia han aparecido muchos pequeños productores que han desarrollado su propia cerveza artesanal con unas características únicas. La clasificación de dichas cervezas es un problema complejo, y para la gente no muy entendida en esta materia, es complicado distinguir entre los cientos de tipos de cervezas artesanales existentes. Además, las clases de cerveza existentes se deben en muchos casos a dos factores principales: el tipo de fermentación, y la ciudad en que ha sido fabricada. Generalmente, luego se agrupan en subclases atendiendo a sus principales propiedades: sabor, composición, graduación alcohólica y color.

Nuestro objetivo ahora va a consistir en llevar a cabo una clasificación en 5 grupos de las cervezas artesanales, atendiendo únicamente a su sabor, graduación alcohólica y color. Para ello, supongamos que hemos realizado un estudio sobre 70865 cervezas artesanales, todas ellas con una graduación superior a 4^{o}, de 176 variedades distintas de acuerdo a la clasificación formal existente de las cervezas artesanales. Supongamos también que el estudio se realiza simultáneamente en 5 laboratorios. El total de las cervezas se reparte entre los laboratorios en muestras de 12000, 13000, 14000, 16000 y 15865 cervezas artesanales distintas, que han sido repartidas de forma totalmente aleatoria entre los 5 laboratorios, y de las cuales se miden las siguientes 4 variables:

1. FG (Final Gravity) - Densidad relativa tras el proceso de fermentación. Da una idea de la cantidad de azúcar que tiene la cerveza.

2. ABV (Alcohol by Volume) - Proporción de alcohol que tiene la cerveza respecto a su volumen total.

3. IBU (International Bitterness Unit) - Es una medida del amargor de la cerveza. Se mide en una escala de 0 a 100, aunque hay cervezas que tienen valores superiores a 100. Sin embargo, los valores superiores a 100 no somos capaces de distinguirlos.

4. Color - Color de la cerveza, medido en una escala de 0 a 100.

Tras haber centrado y escalado los datos obtenidos, para que todas las variables tengan la misma importancia, cada laboratorio nos da como resultado de su experimento una clasificación en 5 grupos de las cervezas que ha evaluado. La clasificación se hace mediante la orden *tclust*, permitiendo despreciar $\alpha = 0{,}01$ del total de los datos. Cada una de estas clasificaciones tiene asociados 5 centros y 5 matrices de covarianzas. Para obtener entonces una clasificación global de todas las cervezas artesanales que se han medido en todos los laboratorios, hacemos el 5-baricentro recortado de nivel $\beta = 0{,}2$, que nos permite obtener una clasificación en 5 grupos de todas las cervezas artesanales medidas en los 5 laboratorios. Dejaremos el 1 % de las cervezas peor representado sin asignar a ningún grupo, ya que pueden ser cervezas más raras que no se parezcan a ninguno de los grupos.

El resultado que nos queda son 5 grupos, tales que el tamaño de cada uno es

Grupo	N.C.	1	2	3	4	5
Tamaño	709	3212	6282	10077	11091	39494

sus dispersiones son muy similares en cuanto al tamaño (los autovalores de sus matrices de covarianza son parecidos), y sus centros vienen dados por:

Grupo	FG	ABV	IBU	Color
1	1.021791	9.223349	67.251830	43.063985
2	1.018206	8.869378	33.419762	13.321936
3	1.015562	7.203439	94.845669	8.633675
4	1.015475	5.908167	37.889226	33.841978
5	1.012431	5.564761	33.608852	7.688540

De esta forma hemos obtenido una clasificación de las 70865 cervezas. Podemos sacar alguna conclusión de los datos obtenidos.

- Cada uno de los centros obtenidos nos sirve como representante de los grupos de cervezas obtenidos. Así, una cerveza con FG=1.021, ABV=9.22º, IBU=67.25, Color=43.06 es un representante del primer grupo. Las cervezas que estén en este grupo serán más parecidas a esta que a las cervezas con los parámetros de el resto de centros. Por tanto, es posible que si al probar entre las cervezas dadas por los 5 centros la que más te guste sea la del centro j, preferirás las cervezas del grupo j, que son más similares a esa.

- Por el tamaño de los grupos, se deduce que la mayoría de cervezas se parecen a la del centro del grupo 5, con parámetros: FG=1.012, ABV=5.56º, IBU=33.61, Color=7.69. En este grupo se incluirán por tanto las cervezas más normales. Veamos la distribución en los grupos que hemos creado de algunas de las cervezas más fabricadas y consumidas:

Grupo	S.C.	1	2	3	4	5
German Pils	0	0	1	2	7	188
German Pilsner (Pils)	0	1	9	4	4	439
International Amber Lager	0	0	0	0	3	79
Oktoberfest/Märzen	2	0	10	2	7	335
Ordinary Bitter	0	0	0	4	5	108

- Mediante la tabla que cruza los grupos creados con los 176 grupos iniciales, podemos analizar la relación que existe entre ambas clasificaciones. Por ejemplo, las cervezas tipo "stout"se corresponden claramente con el grupo 1 que hemos creado:

Grupo	S.C.	1	2	3	4	5
Imperial Stout	14	618	5	0	35	1
Russian Imperial Stout	24	858	9	2	33	1

  o las cervezas tipo IPA más generales se corresponden sobre todo con el grupo 3:

Grupo	S.C.	1	2	3	4	5
Double IPA	12	7	139	683	2	19
Imperial IPA	22	64	187	1153	21	27
Specialty IPA: Belgian IPA	6	1	29	133	3	47

- También existen otros tipos de cervezas menos específicos en los cuáles es más difícil establecer una relación con un determinado grupo:

Grupo	S.C.	1	2	3	4	5
British Strong Ale	3	5	49	17	34	86
Experimental Beer	5	21	69	30	43	229

- Para realizar un análisis más detallado entre los grupos creados y los tipos de cerveza conocidos, se podría realizar un análisis de correspondencias simples.

Nota 7. Los datos se han descargado de https://www.kaggle.com/jtrofe/beer-recipes, donde aparecen datos de 23 variables para 75000 cervezas, la mayoría. Para llevar a cabo el análisis, hemos seleccionado la variable que nos da el tipo de cerveza y las variables FG, ABV, IBU, Color. Además, nos hemos quedado sólo con cervezas con una graduación alcohólica superior a $4^{\underline{o}}$.

5.2. Análisis cluster sobre conjuntos de probabilidades

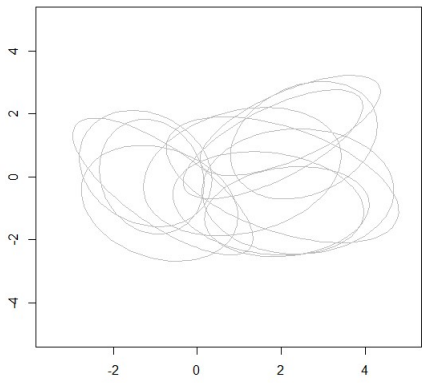
Hemos visto que las k-medias nos dan un método efectivo para llevar a cabo un análisis cluster sobre un conjunto de datos de $\mathbb{R}^d$. Del mismo modo, los k-baricentros nos van a permitir realizar un análisis cluster sobre un conjunto de probabilidades $\nu_1, ..., \nu_n \in \Gamma_2(\mathbb{R}^d)$. Este análisis se realiza de forma análoga al que hacíamos con las k-medias:

- Hallando los k-baricentros con pesos uniformes, obtenemos k centros óptimos $\bar{\mu}_1, ..., \bar{\mu}_k$ tales que:
$$\{\bar{\mu}_1, ..., \bar{\mu}_k\} = \underset{\mu_1,...,\mu_k}{\arg\min} \sum_{i=1}^{n} \underset{j=1,...,k}{\text{mín}} d_2^2(\nu_i, \mu_j)$$
- Los clusters se obtienen luego agrupando mediante criterios de proximidad con los k-baricentros:
$$\mathsf{cluster}J = \{\nu_i : d_2^2(\nu_i, \bar{\mu}_J) \leq d_2^2(\nu_i, \bar{\mu}_j) \quad \forall\, j \neq J\}$$
- Si en vez de utilizar los k-baricentros llevamos a cabo los k-baricentros recortados de nivel α, lo que hacemos es despreciar una proporción α de las probabilidades $\nu_1, ..., \nu_n$. Trabajando así podemos conseguir una mayor estabilidad en el método.

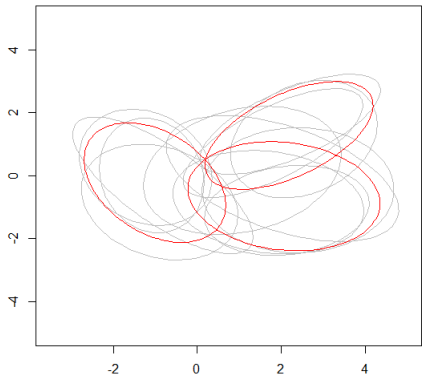
Veamos ahora algunos ejemplos de utilización de esta técnica:

5.2.1. Ejemplo artificial

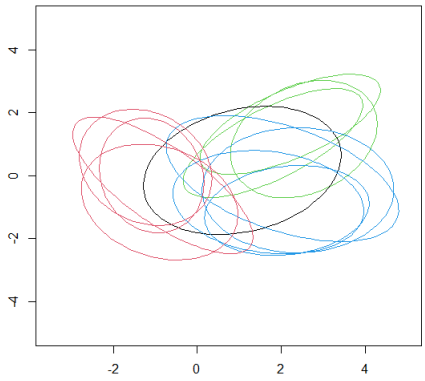
En este ejemplo vamos realizar un análisis cluster sobre el conjunto de 12 distribuciones normales 2-dimensionales con distintos vectores de medias y matrices de covarianzas. Las 12 distribuciones están representadas en el siguiente gráfico a partir de sus elipsoides asociados:



Vamos a buscar 3 grupos, y además vamos a tomar $\alpha = 0{,}1$, permitiendo así que se pueda despreciar una de las distribuciones en el cálculo de los 3-baricentros. El algoritmo nos da como resultado que los 3-baricentros son son las distribuciones normales asociados a los elipsoides rojos de la siguiente gráfica:



Realizamos ahora la asignación en clusters de las probabilidades. La gráfica siguiente muestra en los colores rojo, azul y verde los clusters creados, y en negro queda el elipsoide asociado a la distribución normal que se ha despreciado en el cálculo del 3-baricentro.



5.2.2. Ejemplo cervezas artesanales

Para este ejemplo vamos a utilizar de nuevo el conjunto de datos de cervezas artesanales sobre el que habíamos desarrollado el ejemplo de paralelización en el análisis cluster. Este conjunto de datos estaba formado por observaciones de 70865 cervezas artesanales sobre 4 variables: FG, ABV, IBU y Color, que hemos escalado para que todas tengan la misma importancia. Las

cervezas se agrupaban originalmente en 176 grupos, atendiendo a la clasificación habitual.

Recordemos que en el ejemplo de paralelización, nuestro objetivo había sido lograr una clasificación de las 70865 cervezas atendiendo únicamente a las propiedades medidas sobre ellas: FG, ABV, IBU y Color. Ahora vamos a plantear un problema muy distinto, el objetivo ahora va a ser realizar una clasificación de los distintos tipos de cerveza conocidos. Para este ejemplo, vamos a quedarnos sólo con los 20 tipos para los que aparecen más cervezas en el conjunto de datos, para tener una muestra suficientemente representativa de cada uno de ellos. Estos son los que vamos a estudiar:

American Amber Ale	American Brown Ale	American IPA
American Light Lager	American Pale Ale	American Stout
Blonde Ale	California Common Beer	Double IPA
Imperial IPA	Irish Red Ale	Kölsch
Oatmeal Stout	Robust Porter	Russian Imperial Stout
Saison	Sweet Stout	Weissbier
Weizen/Weissbier	Witbier	

Para cada tipo de cerveza i, tenemos datos de las 4 variables medidas sobre una cantidad n_i cervezas de ese tipo. Calculando la media m_i y la matriz de covarianzas muestral Σ_i de esos datos, podemos relacionar a cada uno de los tipos con una distribución:

$$\nu_i = N(m_i, \Sigma_i) \in \Gamma_2$$

Podemos entonces realizar un análisis cluster sobre las probabilidades ν_i, a partir de los k-baricentros. El resultado de este análisis nos dará una clasificación de los diferentes tipos en k grupos, agrupando a aquellos tipos en los que tanto sus centros como sus estructuras de dispersión (es decir, sus medias y matrices de covarianzas) sean más similares. Las siguientes tablas muestran los distintos resultados obtenidos para distintos valores de k:

Para $k = 2$:

Grupo 1	Grupo 2
American Amber Ale	American Stout
American Brown Ale	Oatmeal Stout
American IPA	Robust Porter
American Light Lager	Russian Imperial Stout
American Pale Ale	Sweet Stout
Blonde Ale	
California Common Beer	
Double IPA	
Imperial IPA	
Irish Red Ale	
Kölsch	
Saison	
Weissbier	
Weizen/Weissbier	
Witbier	

Para $k = 3$:

Grupo 1	Grupo 2	Grupo 3
American Amber Ale	American Stout	American IPA
American Brown Ale	Oatmeal Stout	Double IPA
American Light Lager	Robust Porter	Imperial IPA
American Pale Ale	Russian Imperial Stout	
California Common Beer	Sweet Stout	
Irish Red Ale		
Kölsch		
Saison		
Blonde Ale		
Weissbier		
Weizen/Weissbier		
Witbier		

Para $k = 4$:

Grupo 1	Grupo 2	Grupo 3	Grupo 4
American Amber Ale	American Stout	Double IPA	Blonde Ale
American Brown Ale	Oatmeal Stout	Imperial IPA	Kölsch
American IPA	Robust Porter		Saison
American Light Lager	Russian Imperial Stout		Weissbier
American Pale Ale	Sweet Stout		Weizen/Weissbier
California Common Beer			Witbier
Irish Red Ale			

A la vista de las tablas, las clasificaciones que se van obteniendo para los distintos valores de k parecen bastante razonables, ya que va agrupando en los distintos niveles los tipos de cerveza que podríamos pensar que son más parecidos, ya que forman parte de una clase común. Por ejemplo, se pueden observar los siguientes hechos:

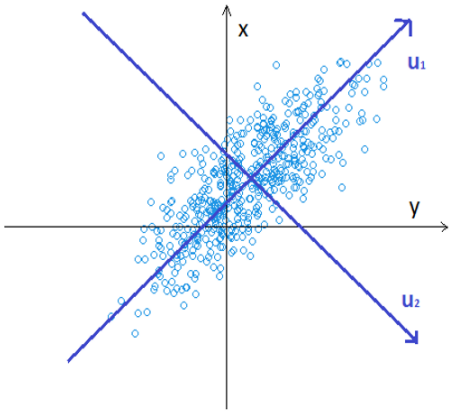
- Para $k = 2$, la clasificación obtenida crea un grupo con las cervezas de tipo "Stout", y las separa resto. Este grupo se mantiene sin cambios cuando $k = 3, 4$, lo que nos hace pensar que las distribuciones de este tipo de cervezas son muy similares entre sí y se diferencian muy bien del resto.

- El grupo creado para $k = 2$ que englobaba al resto de los tipos de cerveza, se divide en el caso $k = 3$ en dos subgrupos, de forma que ahora separa a las tres cervezas de tipo IPA del resto.

- Cuando hacemos $k = 4$, el grupo 1 obtenido en el caso $k = 2$ se divide esta vez en 3 subgrupos. En el primero, se agrupan la mayoría de los tipos de cerveza "American". El segundo está formado por las de tipo IPA (excepto la American IPA, que pertenece ahora al primer grupo). En el tercer grupo, la mayoría de los tipos de cervezas que aparecen son de estilo europeo, generalmente de Alemania, y son de tipo Ale (Blonde Ale, Kölsch, Saison) o de trigo (Weissbier, Weizen/Weissbier, Witbier).

6. Aplicaciones: Componentes principales

El Análisis de Componentes Principales (ACP) es una técnica de descripción estadística para la visualización de la información contenida en un conjunto de datos, formado por observaciones de d variables de una muestra de n individuos. Es decir, tenemos n puntos de $\mathbb{R}^d$. El objetivo fundamental es reducir la dimensión del conjunto de datos, minimizando la pérdida de información. Para ello, el ACP busca las direcciones que recojan la mayor cantidad de información posible en términos de la cantidad de varianza que explica dicha dirección. En el siguiente ejemplo, tenemos un conjunto de datos en dos dimensiones, que puede ser representado en una dimensión a partir de la dirección u_1 sin mucha pérdida de información.



Técnicamente, el ACP busca construir nuevas variables, a partir de las existentes, en las cuales se maximice la suma de distancias al cuadrado de las proyecciones sobre el subespacio generado por las k primeras, y la proyección de la media de los datos sobre dicho subespacio, para cada $k = 1, , ..., d$. Si Σ es la matriz de covarianzas muestrales asociada al conjunto de datos, se puede probar entonces que dichas direcciones son las correspondientes a los autovectores ortogonales $\{u_i\}_{i=1}^d$ de la matriz Σ con autovalores asociados $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0$, que sabemos que existen en virtud del teorema de descomposición en valores singulares. Además, se tiene que la variabilidad explicada por las direcciones de $\{u_i : i \in S\}$ para $S \subset \{1, ..., d\}$ es

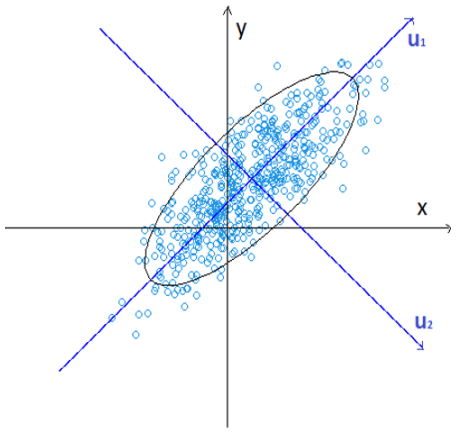
$$100 \cdot \frac{\sum_{i \in S} \lambda_i}{\sum_{j=1}^{d} \lambda_j} \, \%$$

Las principales ventajas de ACP son:

- Las primeras componentes principales son las que mejor expliquen la variabilidad en los datos. Quedándonos con ellas, logramos reducir la dimensión del conjunto de datos minimizando la pérdida de información, ya que nos estaremos quedando con el subespacio de dimensión k que verifica que las proyecciones de los datos sobre ese subespacio tienen la mayor variabilidad posible.

- Las proyecciones de nuestros datos sobre el subespacio generado por las componentes principales van a ser ortogonales, su matriz de covarianzas es diagonal. Esto hace que toda la variabilidad del conjunto de datos se explique a partir de las varianzas de cada una de las componentes, y por lo tanto la estructura de dispersión del conjunto de datos queda representada de una forma mucho más sencilla.

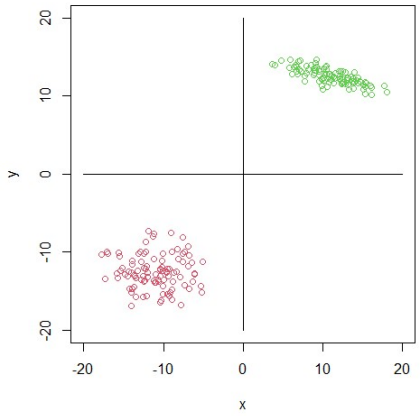
Por otra parte, dado un conjunto de datos como el de antes, podemos modelizarlo y tratarlo como si fueran muestras de una distribución Gaussiana, cuya media es la media de todos los individuos, y su matriz de covarianzas es la matriz de covarianzas muestral. Se cumple entonces que las direcciones de los ejes principales del elipsoide asociado a dicha distribución coinciden con los autovectores de la matriz Σ. Es decir, los ejes principales del elipsoide son exactamente las componentes principales que habíamos hallado antes, como se ve en la siguiente imagen:



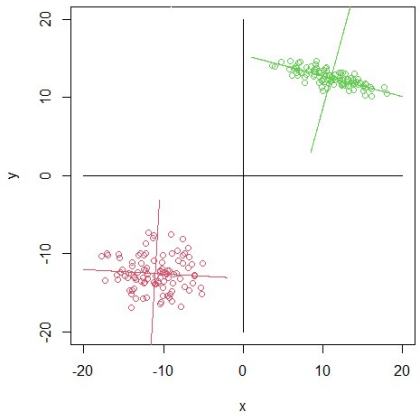
De esta forma podemos establecer una relación entre las componentes principales de un conjunto de datos y una determinada probabilidad, lo cuál nos será útil más adelante.

6.1. Componentes Principales Comunes

Supongamos ahora que existe una partición de nuestro conjunto de datos en k clases, de forma que los individuos pertenecientes a cada clase toman valores parecidos en las variables. Por ejemplo, si tenemos el conjunto de dos dimensiones centrado, dado por:



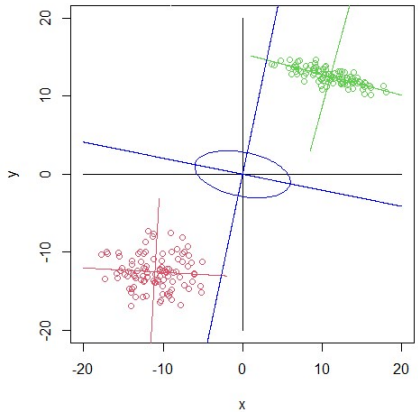
En esta situación, podemos plantearnos hallar por separado las componentes principales para cada una de las k clases. De esta forma, para cada clase tenemos unas componentes principales que nos permiten describir de forma sencilla la estructura de dispersión de dicha clase, es decir, en las que su matriz de covarianzas es diagonal. En nuestro ejemplo anterior, las componentes principales para cada uno de las clases quedarían:



El objetivo de las **componentes principales comunes (CPC)** es hallar, cuando sea razonable, unas nuevas componentes en las cuales las diferencias entre las estructuras de dispersión de unas y otras clases se expliquen de forma sencilla. Si suponemos que las k clases tienen las mismas componentes principales, aunque no necesariamente en el mismo orden (la dirección de la i-ésima componente principal de una clase puede ser la dirección de la j-ésima componente principal para otra, con $j \neq i$), entonces tomando como componentes principales comunes esas componentes, logramos que las matrices de covarianzas de todos las clases en las nuevas variables creadas sean diagonales. En ese caso, la dispersión en cada una de las clases se va a poder estudiar a partir de la varianza en cada una de las componentes, lo que nos va a permitir llevar a cabo comparaciones sencillas entre unas clases y otras.

En la práctica, cuando trabajamos con datos de varias clases, las componentes principales de todos ellos nunca van a coincidir exactamente. Aun así, es razonable pensar que en muchas situaciones van a existir relaciones entre las componentes de unas y otras clases. En ese caso, lo que se busca mediante las componentes principales comunes es determinar si las componentes principales de cada una de las clases son similares entre sí, y en el caso de que lo sean, buscar unas direcciones en las cuáles las matrices de covarianzas de cada una de las clases sean prácticamente diagonales. En esta situación, la dispersión de cada clase puede explicarse en gran medida a partir de las varianzas de estas nuevas componentes, y podremos comparar la variabilidad de las clases fijándonos sólo en estos valores sin perder casi nada de información.

Nuestro objetivo a la hora de buscar unas CPC será hallar una especie de media o representante de las componentes principales de cada una de las clases. El concepto de las componentes principales comunes ha sido tratado anteriormente. Flury da una buena explicación de este concepto en [4], así como un método para hallar unas CPC. En este trabajo se propone una forma alternativa para escoger las componentes principales comunes, que consiste en tomar las direcciones de los ejes principales del elipsoide asociado al baricentro de las distribuciones Gaussianas asociadas a cada uno de los grupos. En el ejemplo sencillo que estábamos tratando, las componentes principales globales son las direcciones que aparecen en azul:

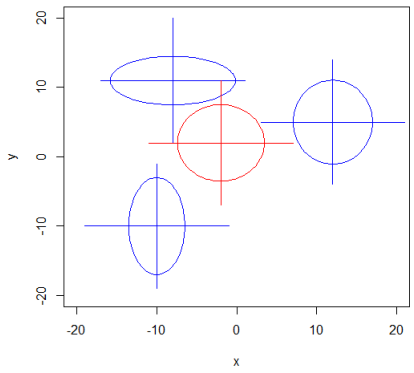


Vemos que cada una de las direcciones que nos dan los ejes principales del baricentro son similares a una componente principal de cada una de las clases (que no tiene por qué ser la misma). Las CPC nos dan en este caso unas componentes similares a las componentes principales que obtendríamos trabajando de forma individual con cada una de las clases. La elección de las direcciones dadas por los ejes del elipsoide asociado al baricentro como CPC parece bastante razonable. El siguiente resultado, que nos asegura que en el caso idóneo en que todos las clases tengan las mismas componentes principales, el baricentro también tendrá esas mismas componentes, justifica esta elección de las CPC.

Teorema 3. *Sean $P_1, ..., P_n$ probabilidades Gaussianas en $\mathbb{R}^d$ con matrices de covarianzas $\Sigma_1, ..., \Sigma_n$, y sea $\bar{P}$ el baricentro asociado a dichas probabilidades con pesos $\{\lambda_i\}_{i=1}^n$, que es una distribución Gaussiana con matriz de covarianzas $\bar{\Sigma}$. Entonces, si $\{e_k\}_{k=1}^d$ es una base*

ortonormal en la cual las matrices de covarianzas de $\Sigma_1, ..., \Sigma_n$ *son diagonales, entonces la matriz de covarianzas de* $\bar{P}$ *en esa base también es diagonal.*

El resultado del teorema se puede apreciar en el siguiente gráfico, donde el elipsoide rojo es el asociado al baricentro de las otras 3 probabilidades:



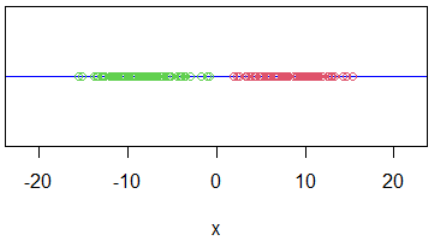
Por tanto, en el caso en que todas las clases tengan las mismas componentes principales, las CPC nos van a dar una base de $\mathbb{R}^d$ en la cual las matrices de covarianzas de todos los grupos son diagonales. Cuando las componentes principales de todos los grupos son similares, las propiedades geométricas de los baricentros van a hacer que las matrices de covarianza de cada uno de los grupos en dichas componentes sean "casi diagonales". Si nos fijamos en el ejemplo sencillo que hemos estado tratando, las matrices de covariazas en las CPC halladas son:

$$\Sigma_1 = \begin{pmatrix} 8{,}845 & 0{,}622 \\ 0{,}622 & 4{,}699 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 9{,}971 & -0{,}542 \\ -0{,}542 & 0{,}457 \end{pmatrix}$$

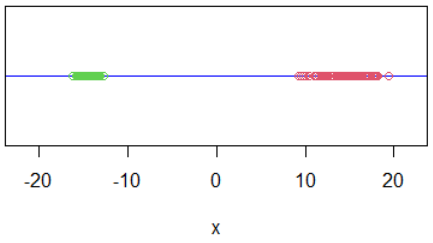
Aunque las matrices de covarianza parecen estar "cerca de la diagonalización", esto no se puede ver a simple vista. Necesitaremos justificar este hecho con más precisión. Sin embargo, vamos a suponer en este caso que la diagonalización obtenida es buena, para poder presentar algunas utilidades de las CPC antes de comenzar con detalles técnicos relativos a la diagonalización de matrices.

La principal ventaja de tener matrices "casi diagonales" es que vamos a poder llevar a cabo razonamientos para la reducción de la dimensión del conjunto de datos sin fijarnos más que en los elementos diagonales de las matrices de covarianzas. En situaciones como ésta, es interesante notar que ya no estamos en disposición de afirmar tal y como hacíamos en ACP que una componente principal común sea más importante que otra, ya que puede ocurrir que una componente principal común sea muy similar a la primera componente principal de una clase, y a la última de otra. La elección de unas componentes u otras se hace dependiendo del propósito que se persiga. Vamos a intentar ilustrar este hecho mediante la elección de una componente principal en nuestro ejemplo sencillo, atendiendo a distintos criterios:

- Si buscamos la dirección que más variabilidad recoge dentro de cada clase, la mejor opción será tomar la componente en la cual las varianzas de ambas clases sean mayores. Tomamos entonces la primera, para la cual la varianza es 8.845 para la primera clase y 9.971 para la segunda. En este caso la proyección sobre esta componente quedará:



- Si buscamos la dirección en la cual la diferencia en la variabilidad entre una clase y otra se haga más visible, la mejor opción será tomar la segunda componente, en la cuál la varianza para la primera clase es 4.699 y para la segunda 0.457 . En este caso, la proyección quedará:



En las proyecciones que hemos obtenido ahora, se ha tenido en cuenta únicamente las diferencias existentes entre la dispersión de una clase y otra. Sin embargo, a la hora de elegir las componentes principales globales que más nos interesan, también podemos hacerlo teniendo en cuenta las diferencias que existen en la localización de ambas clases de datos:

- Supongamos que buscamos la componente principal común que separa mejor la media de cada una de las clases. Calculando las medias de cada clase sobre las componentes principales globales, nos queda que son:

$$m_1 = (8{,}286, 14{,}495) \qquad m_2 = (-8{,}286, -14{,}495)$$

  Por tanto la CPC que mejor separa las medias es claramente la segunda, tal y como se deducía de los gráficos de las proyecciones de arriba.

- Si nuestro objetivo va a ser buscar CPC adecuadas para realizar un análisis discriminante posteriormente, entonces está claro que también vamos a tener en cuenta la localización de las clases para elegir dichas CPC. El principal objetivo del análisis discriminante es buscar subespacios en los cuales la variabilidad entre grupos sea lo mayor posible, y dentro de cada grupo lo menor posible. En este ejemplo sencillo, si buscamos la componente que mejor discrimina entre las dos clases, nos interesará entonces que las medias estén lo más separadas posibles, y que la variabilidad de los datos de cada clase en dicha componente sea la menor posible. La segunda CPC cumple ambas condiciones, luego nos quedaríamos con ella para realizar después un análisis discriminante.

Generalmente, se pueden hacer razonamientos análogos cuando el conjunto de datos tiene dimensión $d > 2$, y buscamos subespacios de dimensión $1 < l < d$ siguiendo un determinado criterio. Sin embargo, en ocasiones vamos a tener que recurrir a métodos más complejos para la elección de las componentes principales comunes adecuadas. Por ejemplo, si buscamos subespacios dimensión mayor que 2 para realizar un análisis discriminante, necesitaremos criterios que nos permitan comparar la eficacia del algoritmo de discriminación que queremos aplicar en cada uno de los subespacios posibles que nos dan las componentes principales comunes.

Por otra parte, en el caso en que en las CPC calculadas las matrices de covarianzas estén "lejos de la diagonalización", podemos quedarnos con un subespacio generado por algunas componentes principales comunes en las que estemos más próximos a dicha diagonalización, siempre que sean suficientemente explicativas. De esta forma logramos también reducir la dimensión del conjunto de datos, atendiendo al criterio de intentar lograr una explicación sencilla de nuestros datos en algunas componentes.

En el siguiente apartado, nos vamos a centrar en el estudio de diferentes formas de medir lo cerca que estamos de la diagonalización en cada una de las clases tras elegir las CPC.

6.1.1. Medidas de diagonalización

Sean $\Sigma_1, \Sigma_2, ..., \Sigma_k$ las matrices de covarianzas asociadas a cada una de las clases, y sea $\bar{\Sigma}$ la matriz de covarianzas del baricentro. Denotemos por $C_1, C_2, ..., C_k$ y $\bar{C}$ a las matrices de covarianzas de cada una de las clases y del baricentro, escritas en la base dada por las CPC halladas a partir del baricentro. Sabemos que en esa base $\bar{C}$ es diagonal. Queremos estudiar entonces si $C_1, C_2, ..., C_k$ están próximas a la diagonalización en dicha base. Sea para cada $i = 1, 2, ..., k$:

$$C_i = \begin{pmatrix} (\sigma_i^1)^2 & * & \cdots & * \\ * & (\sigma_i^2)^2 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & (\sigma_i^d)^2 \end{pmatrix}$$

La diagonalización para cada matriz C_i se puede estudiar atendiendo a distintos criterios:

1. Mediante el cociente de la suma de cuadrados de los elementos de la diagonal entre la suma de cuadrados de todos los elementos de la matriz (la norma Frobenius al cuadrado

$\|C_i\|_F^2$).

$$m_1(C_i) = \frac{\sum_{j=1}^{d}(\sigma_j^1)^4}{\|C_i\|_F^2}$$

Se cumple que $0 \leq m_1(C_i) \leq 1$, siendo este valor igual a 1 si y sólo si C_i es diagonal.

2. Mediante la comparación de los autovectores $\{u_j^i\}_{j=1}^d$ de la matriz Σ_i con los autovectores $\{\bar{u}_j\}_{j=1}^d$ de $\bar{\Sigma}$, que suponemos normalizados. Es decir, estamos comparando las componentes principales de Σ_i con las del baricentro $\bar{\Sigma}$. La comparación se realiza estudiando los cosenos que forman unos autovectores con otros, relacionando en cada paso los autovectores $u_j^i, \bar{u}_l$ para los cuales

$$|\cos(u_j^i, \bar{u}_l)| = |< u_j^i, \bar{u}_l >|$$

sea mayor, hasta tener así todos los autovectores relacionados. Si $f : \{1, ..., d\} \to \{1, ..., d\}$ es la biyección que nos da la relación entre unos y otros, entonces definimos:

$$m_2(C_i) = \sum_{j=1}^{d} |< u_j^i, \bar{u}_{f(j)} >|$$

Se cumple que $0 \leq m_2(C_i) \leq d$, siendo este valor igual a d si y sólo los autovectores de Σ_i son los mismos que los de $\bar{\Sigma}$, es decir, si y solo si C_i es diagonal.

3. La siguiente medida que introducimos es una simplificación de la introducida por Flury en [4]. En dicho trabajo, presenta la siguiente medida para la diagonalización de k grupos simultáneamente, en la que n_i es el número de datos presentes el grupo i, para cada $i = 1, 2, ..., k$.

$$\sum_{i=1}^{k} n_i \log\left(\frac{|\text{diag}(C_i)|}{|C_i|}\right)$$

siendo $|\cdot|$ el determinante de la matriz, y $\text{diag}(C_i)$ la matriz que queda al sustituir todos los elementos no diagonales de C_i por 0. Aquí vamos a simplificar esta medida al cálculo de la de una única matriz C_i, sin aplicar el logaritmo y sin tener en cuenta el número n_i de datos presentes en cada grupo. De esta forma perdemos las propiedades de convergencia que tiene las suma utilizada por Flury, pero seguimos teniendo que matrices cercanas a la diagonalización para la suma anterior, también estarán cercanas a la diagonalización para la suma anterior, también estarán cercanas a la diagonalización para:

$$m_3(C_i) = \frac{|\text{diag}(C_i)|}{|C_i|}$$

Se cumple que $m_3(C_i) \geq 1$, y se da la igualdad si y sólo si C_i es diagonal.

Vamos ahora a proponer una nueva medida para la diagonalización basada en la distancia de Wasserstein y los conceptos que se han ido exponiendo en este trabajo. La idea para definir esta medida se basa en las desigualdades vistas en la demostración del teorema [1] y explicadas en las notas posteriores. Presentamos ahora todo el razonamiento que hemos llevado a cabo

hasta llegar a el resultado obtenido:

Sea P una matriz diagonal y definida positiva, es decir, una matriz diagonal con todos sus elementos diagonales estrictamente mayores que 0, de la forma siguiente:

$$P = \begin{pmatrix} (p^1)^2 & 0 & \cdots & 0 \\ 0 & (p^2)^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (p^d)^2 \end{pmatrix}$$

Si denotamos $d_2^2(C_i, P) = d_2^2(N(0, C_i), N(0, P))$, por la nota 2 sabemos que si $U = (U_1, ..., U_d)$ y $V = (V_1, ..., V_d)$ son dos vectores aleatorios con $\mathscr{L}(U) = N(0, C_i)$ y $\mathscr{L}(V) = N(0, P)$

$$d_2^2(C_i, P) \geq \sum_{j=1}^{d} d_2^2(\mathscr{L}(U_j), \mathscr{L}(V_j)) = \sum_{j=1}^{d} (\sigma_i^j - p^j)^2$$

Sabemos además que en el caso en que la matriz C_i sea diagonal, la estructura de dependencia de ambas distribuciones hace que la desigualdad se convierta en una igualdad:

$$d_2^2(C_i, P) = \sum_{j=1}^{d} (\sigma_i^j - p^j)^2$$

Tenemos por tanto que:

$$0 \leq \frac{\sum_{j=1}^{d} (\sigma_i^j - p^j)^2}{d_2^2(C_i, P)} \leq 1$$

Es razonable pensar entonces que si C_i está próxima a la diagonalización, el cociente tomará valores próximos a 1, sea cual sea la matriz P diagonal y definida positiva. Sin embargo, al estudiar esta medida para ejemplos concretos, se han observado distintos comportamientos dependiendo de la matriz P escogida.

- Cuando las matrices C_i y P son prácticamente iguales, los autovectores de C_i son muy similares a los de P. Sin embargo, los cocientes en esta situación tienen una gran variabilidad, y no están necesariamente cerca de 1. Esto hace que tomando $P = \bar{C}$, si todas las C_i son muy similares, nos quedan resultados inestables.
- Cuando los autovalores de la matriz P son en módulo mucho más grandes o mucho más pequeños que los de C_i, el cociente toma valores muy próximos a 1, a pesar de que los autovectores de la matriz C_i sean muy distintos a los de P. Por este motivo, será interesante utilizar una P_i distinta para cada $i = 1, 2, ..., k$, de forma que las diferencias entre los autovalores de P_i y C_i no sean muy grandes. Además, parece lógico buscar P_i que no dependa de la base en que está escrita C_i.

Para encontrar una medida de diagonalización más adecuada, vamos a centrarnos en buscar entre las matrices P isotrópicas, es decir, diagonales y con todos los elementos de la diagonal

iguales, de la forma,

$$P = \begin{pmatrix} p^2 & 0 & \cdots & 0 \\ 0 & p^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p^2 \end{pmatrix} = p^2 \cdot Id$$

Este tipo de matrices tienen la propiedad de que escritas en cualquier base, nos queda la misma matriz. Si A es una matriz unitaria, la matriz P escrita en la base dada por las columnas de A es:

$$A'PA = p^2 \cdot A'IdA = p^2 A'A = p^2 \cdot Id = P$$

Si A es la matriz unitaria cuyas columnas son los autovectores de las direcciones principales dadas por el baricentro, entonces se tiene además que $C_i = A'\Sigma_i A$. Por la proposición [11], si T es la aplicación lineal de matriz A', se tiene que:

$$d_2(N(0,\Sigma_i), N(0,P)) = d_2(N(0,\Sigma_i) \circ T^{-1}, N(0,P) \circ T^{-1}) =$$

$$= d_2(N(0, A'\Sigma_i A), N(0, A'PA)) = d_2(N(0,C_i), N(0,P)) \tag{13}$$

Por tanto, el cociente

$$\frac{\sum_{j=1}^{d}(\sigma_i^j - p)^2}{d_2^2(C_i, P)}$$

tiene denominador constante para todas las bases ortonormales en las que expresamos la matriz C_i, y por tanto se maximiza para la base en la cuál $\sum_{j=1}^{d}(\sigma_i^j - p)^2$ sea máximo. Queremos ahora buscar un valor de p adecuado. Para cada $i = 1, 2, ..., k$, vamos a utilizar $p = \bar{\sigma}_i = \frac{1}{d}\sum_{j=1}^{d}\bar{\sigma}_i^j$, siendo $\{\bar{\sigma}_i^j\}_{j=1}^{d}$ las desviaciones típicas de las componentes de Σ_i escrita en la base en la que dicha matriz diagonaliza, que no depende de la base en la que está escrita C_i. Además, por la definición de $\bar{\sigma}_i$, es lógico pensar que las diferencias entre los autovalores de C_i y $\bar{\sigma}_i^2$ no van a ser muy grandes. Por estos motivos, es razonable pensar que la matriz:

$$P_i = \begin{pmatrix} \bar{\sigma}_i^2 & 0 & \cdots & 0 \\ 0 & \bar{\sigma}_i^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{\sigma}_i^2 \end{pmatrix}$$

va a funcionar bien cuando medimos la diagonalización de C_i comparando con ella. De esta forma, el cociente se maximiza en la base en la cual la suma:

$$\sum_{j=1}^{d}\left(\sigma_i^j - \bar{\sigma}_i\right)^2 = \sum_{j=1}^{d}\left(\sigma_i^j - \frac{1}{d}\sum_{j=1}^{d}\bar{\sigma}_i^j\right)^2$$

sea máxima. Aunque no hemos logrado dar una justificación teórica que nos permita relacionar directamente estas sumas con la diagonalización de una matriz, si hemos comprobado experimentalmente que este cociente no nos da los problemas que aparecían cuando utilizábamos como matriz P el baricentro o la identidad. Además, podemos dar una caracterización de $\bar{\sigma}_i$,

que puede ayudar en futuros trabajos a estudiar con más detalle estos cocientes. Básicamente, se puede probar que $\bar{\sigma}_i$ es el valor de p para el cual $d_2(\Sigma_i, p^2 Id)$ alcanza el mínimo, ya que escribiendo Σ_i en la base en la que es diagonal, por (13) se tiene que:

$$d_2(\Sigma_i, p^2 Id) = \sum_{j=1}^{d} (\bar{\sigma}_i^j - p)^2$$

En resumen, en este trabajo utilizaremos :

$$m_4 = \frac{\sum_{j=1}^{d} (\sigma_i^j - \bar{\sigma}_i)^2}{d_2^2(C_i, P_i)} = \frac{\sum_{j=1}^{d} (\sigma_i^j - \frac{1}{d}\sum_{j=1}^{d} \bar{\sigma}_i^j)^2}{d_2^2(C_i, P_i)}$$

como medida de diagonalización, asumiendo que funciona bien.

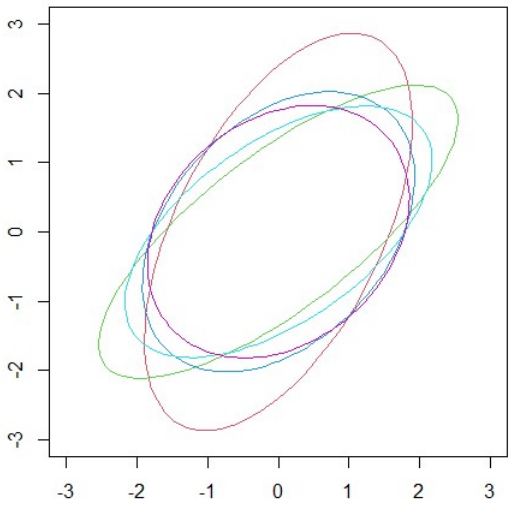
6.1.2. Ejemplo artificial para la comparación de medidas

Vamos a presentar ahora un ejemplo sencillo en dimensión 2 para poder ilustrar todos los conceptos mencionados anteriormente, y poder llevar a cabo una comparación de las 4 medidas de la diagonalización en un ejemplo sencillo. Para ello, supongamos que tenemos datos procedentes de 5 clases distintas, con matrices de covarianzas:

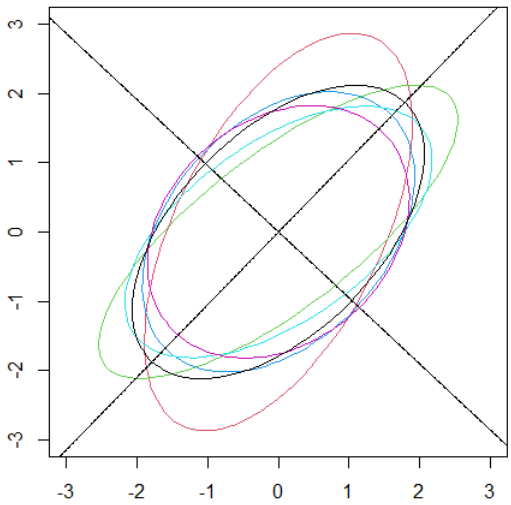
$$\Sigma_1 = \begin{pmatrix} 3{,}59 & 2{,}97 \\ 2{,}97 & 8{,}23 \end{pmatrix} \qquad \Sigma_2 = \begin{pmatrix} 6{,}45 & 4{,}13 \\ 4{,}13 & 4{,}48 \end{pmatrix} \qquad \Sigma_3 = \begin{pmatrix} 3{,}71 & 1{,}49 \\ 1{,}49 & 4{,}09 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 4{,}71 & 2{,}23 \\ 2{,}23 & 3{,}29 \end{pmatrix} \qquad \Sigma_5 = \begin{pmatrix} 3{,}42 & 0{,}86 \\ 0{,}86 & 3{,}31 \end{pmatrix}$$

En el siguiente gráfico, representamos cada una de las matrices con el elipsoide asociado a la distribución $N(0, \Sigma_i)$, en los colores rojo, verde, azul, cian y magenta para i=1,2,3,4,5 respectivamente.



Calculamos entonces el baricentro de dichas probabilidades. En el siguiente gráfico aparece el elipsoide asociado al baricentro en color negro. También aparecen los ejes principales asociados al baricentro, que son las CPC que vamos a tomar.



Queremos dar una idea ahora de lo cerca que estamos de la diagonalización para cada una de las matrices. Para ello, vamos a utilizar las medidas m_1, m_2, m_3 y m_4 de las sección anterior. Los resultados obtenidos aparecen en la siguiente tabla. Para que sea más visual, nos referimos a cada una de las clases por el color del elipsoide asociado.

	Rojo	Verde	Azul	Cian	Magenta
m_1	0.9034291	0.9708986	0.9991391	0.9692874	0.9992556
m_2	1.906031	1.980173	1.998330	1.968751	1.997000
m_3	1.229587	1.117520	1.001161	1.062825	1.000850
m_4	0.6407040	0.9032392	0.9930594	0.8682647	0.9878395

En este caso sencillo, se ve que todas las medidas dan resultados muy similares, que coinciden con lo que visualmente podíamos esperar. Todas las medidas coinciden en que las clases asociadas a los elipsoides azul y magenta son las que mejor se representan en estas CPC, seguidas de las clases asociadas a los elipsoides verde y cián, y que la peor representada es la asociada al elipsoide rojo. Sin embargo, para algunas medidas, el elipsoide azul está mejor representado que el magenta, y para otras ocurre lo contrario. Lo mismo pasa con el verde y el cián. Esto es lógico que ocurra a veces, ya que en cada caso el criterio para medir la diagonalización es distinto. Lo importante es que se puede ver que en general todas las medidas nos dan criterios similares para medir la diagonalización, que coinciden con la idea geométrica de las componentes principales comunes que estamos buscando.

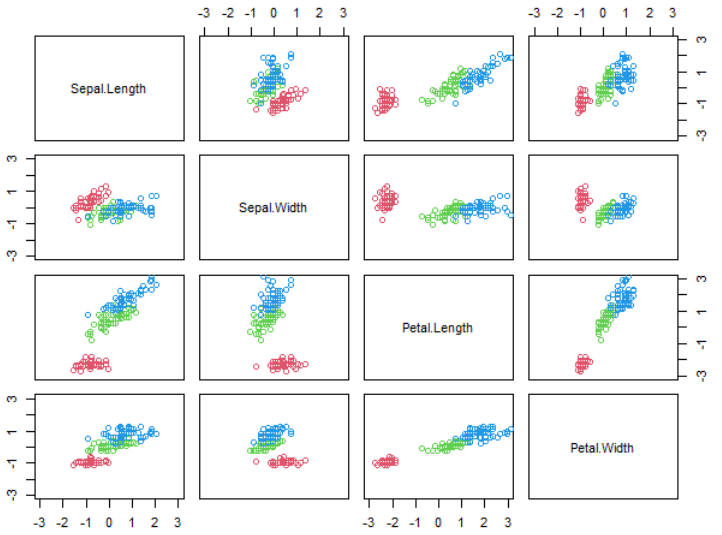
En el resto de ejemplos que vamos a estudiar, vamos a tomar como medida de diagonalización a la medida m_4. Además, cuando queramos medir la diagonalización de las matrices de covarianzas de k grupos en unas CPC, si las matrices de covarianzas de esas k clases en la base son $C_1, C_2, ..., C_k$, siguiendo la notación anterior utilizaremos como medida el valor de:

$$\frac{1}{k}\sum_{i=1}^{k} m_4(C_i) = \frac{1}{k}\sum_{i=1}^{k} \frac{\sum_{j=1}^{d}(\sigma_i^j - \bar{\sigma}_i)^2}{d_2^2(C_i, P_i)} \quad \in [0,1]$$

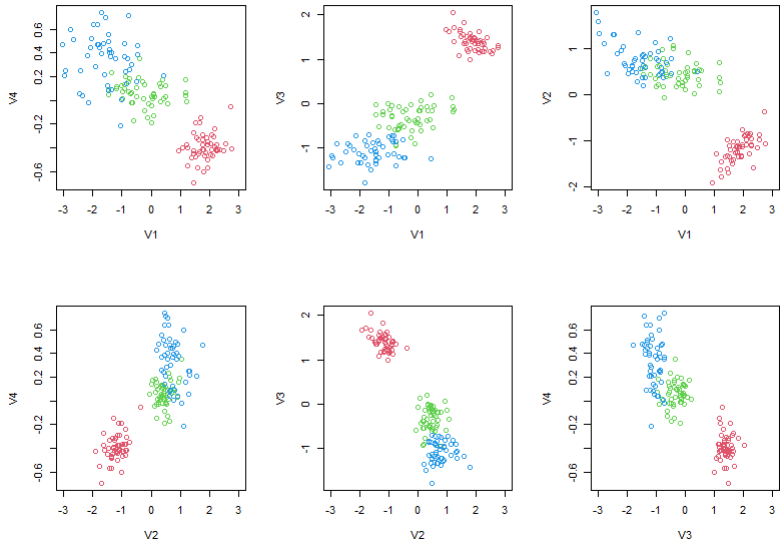
siendo este valor cercano a 1 cuando estamos cerca de la diagonalización en todos las clases.

6.1.3. Ejemplo con los datos de las iris

Vamos a aplicar el razonamiento anterior al conjunto de datos iris. Este conjunto está formado por observaciones tomadas para un total de 150 flores, de las cuáles 50 son de la especie setosa, 50 de la especie versicolor y 50 de la especie virginica. De cada una de las flores se tienen datos de 4 variables, correspondientes a la medida del ancho y el largo de pétalos y sépalos. Una vez centrados los datos, los podemos representar mediante el siguiente conjunto de gráficas. Los individuos correspondientes a la especie setosa aparecen en color rojo, los correspondientes a la especie versicolor en color verde, y los de la especie virginica en color azul.



Vamos ahora a calcular las componentes principales comunes para los tres grupos. Para ello, calculamos las matrices de covarianzas muestrales asociadas a cada una de las especies, a partir de los 50 datos que tenemos de cada una, y hallamos el baricentro asociado a las distribuciones normales centradas y con las matrices de covarianzas calculadas. Calculamos los ejes del elipsoide asociado al baricentro para utilizarlos como componentes principales comunes. Represemos gráficamente los datos sobre las CPC obtenidas:



Las matrices de covarianzas de cada una de las especies en las CPC halladas son:

$$C_{setosa} = \begin{pmatrix} 0{,}1739 & 0{,}0803 & -0{,}0461 & 0{,}0104 \\ 0{,}0803 & 0{,}0818 & -0{,}0277 & 0{,}0135 \\ -0{,}0461 & -0{,}0277 & 0{,}0404 & 0{,}0001 \\ 0{,}0104 & 0{,}0135 & 0{,}0001 & 0{,}0129 \end{pmatrix}$$

$$C_{versicolor} = \begin{pmatrix} 0{,}4798 & -0{,}029 & 0{,}0478 & -0{,}0149 \\ -0{,}0299 & 0{,}0622 & 0{,}0045 & 0{,}0054 \\ 0{,}0478 & 0{,}0045 & 0{,}0719 & 0{,}0024 \\ -0{,}0149 & 0{,}0054 & 0{,}0024 & 0{,}0108 \end{pmatrix}$$

$$C_{virginica} = \begin{pmatrix} 0{,}6725 & -0{,}1117 & 0{,}0274 & 0{,}0035 \\ -0{,}1117 & 0{,}1157 & 0{,}0071 & -0{,}0227 \\ 0{,}0274 & 0{,}0071 & 0{,}0561 & -0{,}0098 \\ 0{,}0035 & -0{,}0227 & -0{,}0098 & 0{,}0439 \end{pmatrix}$$

Vamos ahora a ver si la estamos cerca de la diagonalización en estas componentes. Para ello, utilizamos la medida de diagonalización m_4 para cada uno de los grupos. Nos quedan los siguientes resultados.

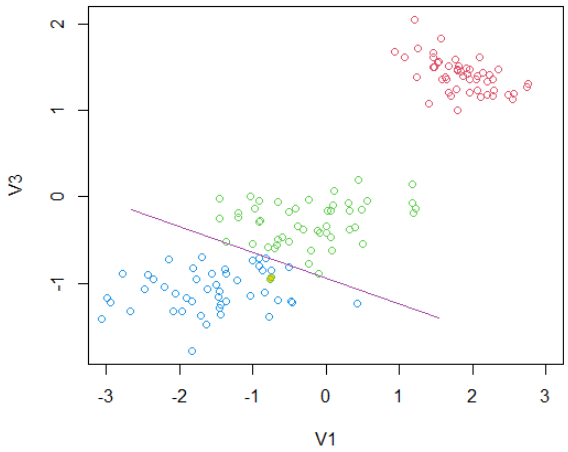
Setosa	Versicolor	Virginica
0.5770966	0.9550782	0.9040537

Fijándonos individualmente en cada una de las especies, podemos ver que la especie Setosa está mucho peor representada en estos ejes que las especies Virginica y Versicolor. Si queremos la medida de diagonalización para los tres grupos a la vez, estudiamos el valor de:

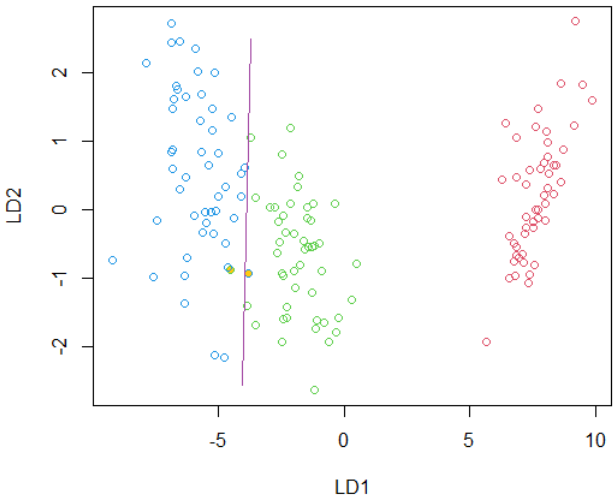
$$\frac{1}{3}\sum_{i=1}^{k} m_4(C_i) = 0{,}812$$

No nos queda un valor muy alto, luego la diagonalización no es muy buena. Por tanto, si estudiamos las diferencias entre las estructuras de dispersión de las distintas especies únicamente a partir de las varianzas en estas componentes principales comunes, estamos perdiendo bastante información.

Sin embargo, estas componentes halladas son interesantes si nuestro objetivo va a ser realizar después un análisis discriminante. Utilizando el gráfico por pares, se ve que las variables 1 y 3 van a ser las que más nos interesan, ya que se puede establecer una separación de los grupos mediante hiperplanos, salvo por dos puntos:



que son el mismo número de puntos que no se pueden separar mediante hiperplanos cuando utilizamos las direcciones que nos da el algoritmo de discriminación lineal:



Nuestro objetivo ahora va a ser intentar buscar un subespacio, formado por algunas de las CPC halladas, en el que estemos más cerca de la diagonalización.

- En primer lugar, buscamos el subespacio H de dimensión 3 generado por 3 de las CPC en el cual estemos más próximos a la diagonalización. Denotando C_i^H a la matriz que se obtiene al restringir C_i a las filas y columnas de H, estudiamos los valores de

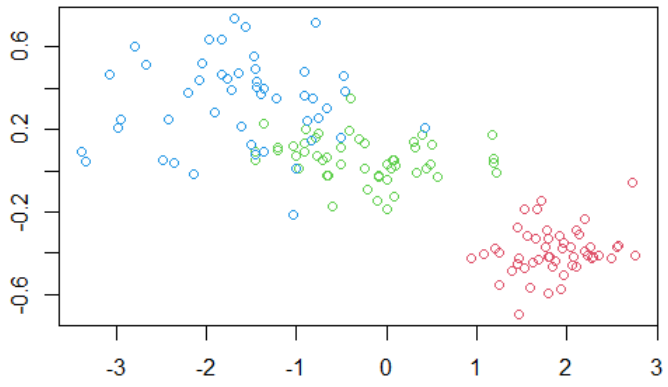
$$\frac{1}{3}\sum_{i=1}^{k} m_4(C_i^H)$$

  El valor más alto lo alcanza cuando H es el subespacio generado por las CPC primera, tercera y cuarta, y este valor es 0.9854631.

- Podemos hacer el razonamiento análogo para buscar un subespacio de 2 dimensiones. En este caso, el valor más alto del cociente nos queda cuando $H = \{1, 4\}$, y el valor es 0.9997832. Las matrices de los grupos las CPG 1 y 4 son:

$$\Sigma_{set} = \begin{pmatrix} 0{,}1739 & 0{,}0104 \\ 0{,}0104 & 0{,}0129 \end{pmatrix} \quad \Sigma_{vers} = \begin{pmatrix} 0{,}4798 & -0{,}0149 \\ -0{,}0149 & 0{,}0108 \end{pmatrix} \quad \Sigma_{virg} = \begin{pmatrix} 0{,}6725 & 0{,}0035 \\ 0{,}0035 & 0{,}0439 \end{pmatrix}$$

  En estas dos componentes, prácticamente toda la variabilidad se explica a partir de la varianza de cada componente. Si observamos la gráfica que nos queda al representar la gráfica en estas dos dimensiones, vemos que las direcciones de crecimiento de la dispersión de todas las especies son similares a los ejes. Es decir, las CPC que hemos hallado son parecidas a las componentes principales de cada una de las especies.



Si queremos reducir la dimensionalidad del conjunto, podríamos intentar buscar las CPC que mejor expliquen las diferencias en la dispersión de las distintas variables. Para ello, denotemos

$\tilde{\sigma}^j = \frac{1}{3}\sum_{i=1}^{3}\sigma_i^j$ para $j = 1, 2, 3, 4$, siendo σ_i^j la desviación típica de la componente principal j en la especie i, y estudiamos las sumas: Además, estudiando las sumas

$$\sum_{i=1}^{n}(\sigma_i^j - \tilde{\sigma}^j)^2, \quad j = 1, ..., d$$

nos quedan los siguientes valores:

CPC 1	CPC 2	CPC 3	CPC 4
0.0848	0.0041	0.0022	0.0068

Por tanto, las componentes 1 y 4 serán en principio aquellas que expliquen mejor las diferencias entre la dispersión de los diferentes grupos a través de las varianzas de las componentes. Como además son las componentes en las que estamos más cerca de la diagonalización, parece adecuado representar nuestros datos sobre estas dos componentes si queremos ver las diferencias existentes entre las estructuras de dispersión de los diferentes grupos.

La principal ventaja de este razonamiento es que se puede llevar a cabo de forma análoga para conjuntos de dimensión arbitrariamente grande. En casos como el del ejemplo, en los que la dimensión no es muy alta, y el número de combinaciones posibles entre dos variables es pequeño, podemos utilizar el gráfico por pares para elegir cuáles son las variables que más nos interesan en cada situación. En dimensiones altas, es necesario disponer de técnicas computacionales que nos permitan elegir las CPC que más nos interesan en cada caso.

Vamos ahora a comparar las CPC que hemos obtenido a partir del baricentro con las que halla Flury en [4], para ver si hemos llegado a resultados parecidos. Las componentes principales comunes en cada método, expresadas por columas, son las siguientes:

$$\text{Comp. baricentro} = \begin{pmatrix} -0{,}7433 & 0{,}2018 & 0{,}6121 & 0{,}1791 \\ -0{,}3701 & -0{,}8786 & -0{,}0742 & -0{,}2921 \\ -0{,}5342 & 0{,}3859 & -0{,}6853 & -0{,}3097 \\ -0{,}1584 & -0{,}1954 & -0{,}3874 & 0{,}8868 \end{pmatrix}$$

$$\text{Comp. Flury} = \begin{pmatrix} 0{,}7367 & -0{,}1640 & -0{,}6471 & 0{,}1804 \\ 0{,}2468 & -0{,}8346 & 0{,}4645 & -0{,}1607 \\ 0{,}6047 & 0{,}5221 & 0{,}5003 & -0{,}3338 \\ 0{,}1753 & 0{,}0628 & 0{,}3382 & 0{,}9925 \end{pmatrix}$$

Los ángulos entre las componentes obtenidas por los dos métodos, comparándolas por orden, son los siguientes:

CPC 1	CPC 2	CPC 3	CPC 4
8,20°	27,19°	25,24°	8,90°

Por tanto, se obtiene dos componenetes principales comunes muy similares, la primera y la cuarta, ya que los ángulos entre ellas son menores que 10°. Para las otras dos, los ángulos son mayores, aunque ninguno supera los 30°. Por tanto, las CPC nos dan una buena alternativa

a las halladas por Flury, y además tienen la ventaja de ser computacionalmente mucho más eficientes.

Por último, vamos a resaltar un hecho que nos va a servir de motivación para la siguiente sección. Habíamos visto que los grupos Versicolor y Virginica se representaban mucho mejor que el grupo Setosa en las CPC halladas. Podemos entonces preguntarnos si las componentes principales de estas dos especies son similares entre sí, y comprobar si podemos obtener unas componentes en las que se expliquen bien las dispersiones de estas dos especies. Calculamos entonces las CPC para todas las parejas posibles formadas por dos especies, y representamos en la siguiente tabla la medida para la diagonalización de cada pareja de especies en las CPC halladas para esa pareja:

Setosa/Versicolor	Setosa/Virginica	Versicolor/Virginica
0.8173	0.749326	0.9686

A las vista de los resultados, podemos afirmas que las especies Versicolor y Virginica tienen componentes principales muy similares entre sí, y significativamente distintas de las de la especie setosa. En problemas como este, es razonable pensar en tomar varias componentes principales, de forma que podamos explicar bien la dispersión de cada clase en una de dichas componentes. Esto es lo que buscamos en la siguiente sección.

6.2. Componentes Principales Grupales

En esta sección, vamos a introducir las componentes principales grupales (CPG) como una generalización de las componentes principales comunes. Supongamos al igual que antes que tenemos una partición del conjunto de datos en k clases, y que ahora no parece existir una relación entre las componentes principales de cada una de las clases. Es decir, hemos calculado las CPC y las medidas de diagonalización nos dicen que estamos lejos de la diagonalización de todas las clases. Podemos entonces plantearnos hacer una subdivisión en m grupos de las k clases de datos, de forma que podamos dar en cada grupo unas componentes en las cuáles las matrices de covarianzas de todas las clases estén próximas a la diagonalización. Al conjunto formado por las m componentes halladas lo denominaremos **componentes principales grupales**. Vamos a dar ahora un procedimiento para encontrar unas CPG adecuadas. Sean $\Sigma_1, \Sigma_2, ..., \Sigma_k$ las matrices de covarianzas de las k clases, y sean P_i probabilidades normales centradas y con matriz de covarianzas Σ_i para $i = 1, 2, ..., k$. Si buscamos m grupos:

- En primer lugar, tomamos las componentes principales asociadas a los m-baricentros $\{\bar{P}_j^0\}_{j=1}^m$ con pesos uniformes de las probabilidades $\{P_i\}_{i=1}^k$, que nos dan una partición del conjunto de índices $\{S_1^0, ..., S_k^0\}$, siendo $S_j^0 \subset \{1, ..., m\}$ el conjunto de índices asociados al baricentro $\bar{P}_j^0$.

- Asignamos cada clase i a uno de los m-baricentros $\{\bar{P}_j^0\}_{j=1}^m$, estudiando el mínimo de las distancias

  $$m_4(C_i^j) \qquad j = 1, ..., m$$

  donde C_i^j es la matriz de covarianzas de la clase i escrita en la base de autovectores dada por el baricentro $\bar{P}_j^0$ (si queremos medir la diagonalización con otro criterio, será

conveniente llevar a cabo la partición atendiendo a dicho criterio). Esto nos dará una partición del conjunto de índices $\{S_1^1, ..., S_k^1\}$, siendo $S_j^1 \subset \{1, ..., m\}$ el conjunto de índices para los cuales se alcanza el mínimo en el baricentro $\bar{P}_i^0$.

- Si los grupos $\{S_1^1, ..., S_k^1\}$ obtenidos coinciden con los dados por los m-baricentros, $\{S_1^0, ..., S_k^0\}$, las componentes principales halladas para cada grupo $\{P_i\}_{i \in S_j^1}, j = 1, 2, ..., m$, coinciden con las del baricentro de dichas probabilidades. Si no es así, podemos tomar en cada grupo j las componentes principales asociadas al baricentro de $\{P_i\}_{i \in S_j^1}$, obteniendo así m baricentros $\{\bar{P}_j^1\}_{j=1}^m$ que nos dan m componentes principales nuevas.

- A partir de estas nuevas m componentes principales asignadas a los baricentros $\{\bar{P}_j^1\}_{j=1}^m$, podemos asignar cada clase i a uno de los m-baricentros estudiando el mínimo de las distancias

$$m_4(C_i^j) \qquad i = 1, ..., m$$

  donde C_i^j es la matriz de covarianzas de la clase i escrita en la base de autovectores dada por el baricentro $\bar{P}_j^1$. Esto nos dará una partición del conjunto de índices $\{S_1^2, ..., S_k^2\}$, siendo $S_j^2 \subset \{1, ..., m\}$ el conjunto de índices para los cuales se alcanza el mínimo en el baricentro $\bar{P}_i^1$.

- Si la partición $\{S_1^2, ..., S_k^2\}$ es distinta a la partición $\{S_1^1, ..., S_k^1\}$, hallamos de nuevo los m baricentros $\{\bar{P}_j^2\}_{j=1}^m$ asociados a cada grupo $\{P_i\}_{i \in S_j^2}, j = 1, 2, ..., m$, y obtenemos así m componentes principales nuevas que nos permiten hacer una nueva partición $\{S_1^3, ..., S_k^3\}$.

- Repetimos este proceso hasta llegar a un $n \in \mathbb{N}$ que verifique que la partición $\{S_1^n, ...S_k^n\}$ sea la misma que $\{S_1^{n+1}, ..., S_k^{n+1}\}$. La partición en m grupos que tomaremos será:

$$\{S_1, ..., S_k\} = \{S_1^n, ..., S_k^n\}$$

  y las componentes principales grupales serán las m componentes principales de los m baricentros $\{\bar{P}_j^n\}_{j=1}^m$.

- Si para ningún $n \in \mathbb{N}$ se de la igualdad de las particiones que habíamos mencionado en el punto anterior, pararemos el proceso en la iteración N, para $N \in \mathbb{N}$ fijado. Las componentes principales grupales halladas de esta forma pierden la propiedad de ser en cada grupo el baricentro de las probabilidades asociadas a ese grupo, pero van a seguir siendo una buena forma de elegir m componentes principales en las que las matrices de covarianza de todos los grupos sean "lo más diagonales posibles".

Una vez hecho esto, ya tenemos la división de los grupos. Podemos ver si las direcciones de crecimiento en cada división son similares estudiando las sumas:

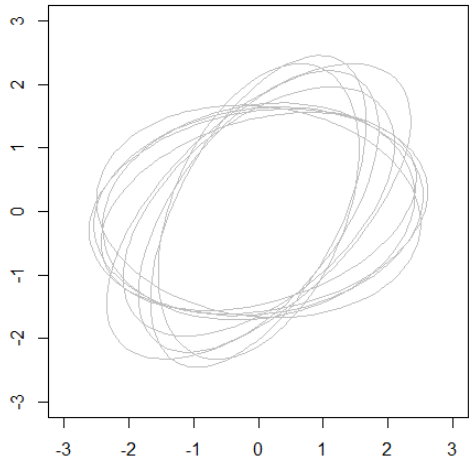
$$\frac{1}{|S_j|} \sum_{i \in S_j} m_4(C_i^j) \quad j = 1, ..., m \text{ si } S_j \neq \emptyset$$

Si el cociente está próximo a 1 en todos los grupos, habremos conseguido que en cada S_j las matrices de covarianzas casi diagonalicen en las CPG dadas por $\bar{P}_i$. Sino, podemos probar a refinar la partición, aumentando el valor de m.

Otra posibilidad para el cálculo de las componentes principales globales es permitir que se desprecie unas proporción α de las probabilidades. Una forma de hacer esto es comenzar calculando los m-baricentros recortados de nivel α, y despreciando la proporción α de las probabilidades que nos indica el k-baricentro. De esta forma se logrará una mayor estabilidad del método, y en muchas ocasiones evitaremos que aparezcan grupos formados por una única clase.

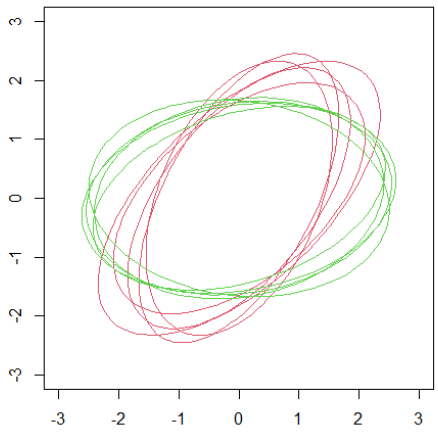
6.2.1. Ejemplo artificial

Vamos a ilustrar mediante un ejemplo sencillo en dimensión 2 la utilización y el cálculo de las CPG. Para ello, supongamos que tenemos datos de 10 clases distintas, cuyas matrices de covarianzas se corresponden con los siguientes elipsoides:



Si calculamos directamente las CPC, la medida de diagonalización de todas las clases, calculada como la media de la medida de diagonalización m_4 en cada clase, nos queda un valor de 0.4821. Estamos por tanto muy lejos de conseguir una diagonalización adecuada para todas las clases.

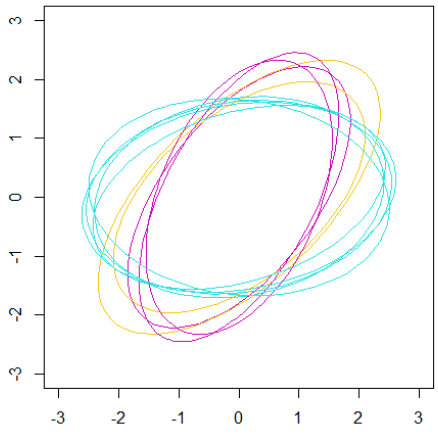
Vamos a intentar hallar unas componentes principales grupales en las que queden bien representadas todas las clases. Si buscamos 2 componentes principales, la asignación en grupos que nos queda es la siguiente:



La medida de diagonalización en cada grupo es:

Grupo Rojo	Grupo Verde
0.8738	0.9203

Por tanto, utilizando las CPG que hemos hallado estamos cerca de la diagonalización en los dos grupos que hemos creado, sobre todo en el verde. Podemos intentar mejorar la diagonalización buscando separar las clases en 3 grupos mediante las CPG. En este caso, la separación en grupos que nos dan las clases es:



y la medida de diagonalización dentro de cada grupo es:

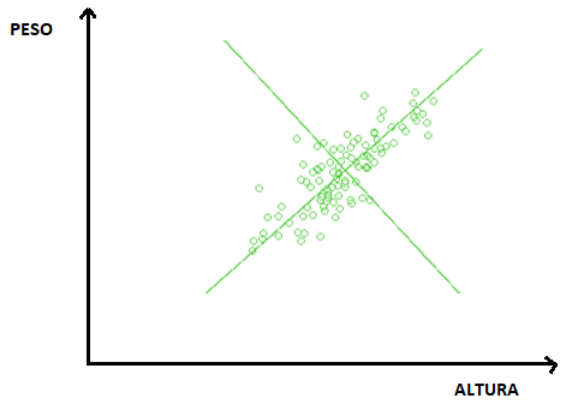
Grupo Cian	Grupo Morado	Grupo Amarillo
0.9203	0.9553	0.9968

Por tanto, hemos mejorado los resultados obtenidos cuando considerábamos dos grupos, y podría ser interesante considerar esta separación en 3 grupos en vez de la separación 2. Básicamente, lo que se ha hecho es dividir el grupo rojo en los grupos morado y amarillo, y hallar unas nuevas componentes en las que cada subgrupo esté más cerca de la diagonalización. El grupo cian es el mismo que el verde que teníamos antes. La elección de una cantidad de grupos u otra, dependerá en cada caso concreto del uso que vayamos a hacer de los datos, para ver si es preferible tener menos grupos y perder un poco más de información al representar las dispersiones en términos de las varianzas de las componentes, o si es preferible aumentar el número de grupos para representar mejor las dispersiones en términos de las varianzas.

6.2.2. Aplicación a la clasificación

Cuando tenemos problemas reales en los que medimos las mismas variables sobre distintos individuos de distintas clases, aunque las matrices de covarianza varíen entre unas clases y otras, en ocasiones es razonable pensar que presentan una estructura básica común entre algunas de las clases. Muchas veces esta estructura básica común se debe a que las variables que aparecen en el problema tienen direcciones de crecimiento parecidas en dichas clases.

Estas relaciones aparecen habitualmente en taxonomía. Pensemos para ello en el caso más sencillo posible, cuando estamos midiendo la altura y el peso de individuos de distintas especies. Si representamos la nube de puntos de los individuos de una determinada especie veremos una clara dirección de crecimiento, correspondiente al aumento del peso de los individuos cuando aumenta su altura.



La dirección de crecimiento que representa el aumento del peso cuando aumenta la altura se corresponde con la primera componente principal del grupo. Sin embargo, aunque en todas las especies vaya a aparecer esa dirección de crecimiento, puede cambiar significativamente de unas especies a otras:

- Por ejemplo, si la especie que estuviéramos tratando fueran hipopótamos, su complexión física hará que el aumento del peso sea mucho más significativo que el aumento de la altura, y la dirección de crecimiento que estemos estudiando tenga mucha más pendiente.
- Si estudiamos lo que ocurre con los humanos, sucederá al revés. El hecho de que caminemos erguidos hace que tengamos una gran altura en relación con el peso, y por tanto los cambios en la altura serán más significativos que los cambios en el peso. La pendiente de la dirección de crecimiento será menor.

En esta situación, se podría llevar a cabo una clasificación de las diferentes especies en grupos, pidiendo únicamente que en los grupos formados las direcciones de crecimiento sean parecidas.

En taxonomía, cuando trabajamos con variables más complejas, es frecuente que sigan apareciendo relaciones (que ya no vamos a poder adivinar a simple vista) entre las direcciones de crecimiento de las variables en las distintas especies estudiadas, que nos permitan establecer asociaciones entre ellos. Algunos ejemplos concretos de estas relaciones aparecen en [5], [4]. En el ejemplo con el conjunto de datos Iris, también se podía observar una relación fuerte entre las direcciones de crecimiento de las especies Virginica y Versicolor. Este tipo de relaciones pueden aparecer también en muchos otros campos, por ejemplo en la música, como veremos a continuación.

La teoría que hemos tratado anteriormente sobre las componentes principales grupales, nos permite desarrollar un método para estudiar dichas relaciones entre las direcciones de crecimiento de las variables entre diferentes grupos, y posteriormente dar una clasificación de los grupos atendiendo a estas propiedades. En el ejemplo siguiente, vamos a dar una clasificación alternativa para datos de 10 géneros musicales, que nos va a permitir ver si existen relaciones entre las componentes principales de unos géneros y otros.

6.2.3. Clasificación de géneros de música

En este ejemplo, el conjunto de datos a tratar contiene información sobre canciones de 10 estilos de música distintos: blues, clásica, country, disco, hiphop, jazz, metal, pop, reggae y rock. De cada uno de los géneros, se tienen datos medidos sobre un intervalos de 30 segundos de 100 canciones de ese género. Las variables que aparecen son:

- **Tempo**: Velocidad a la que se mide el paso de la música.
- **Beats**: Unidad del ritmo en música.
- **chroma stft**: Transformada de Fourier de tiempo reducido.
- **rmse**: Raíz cuadrada del error cuadrático medio.
- **spectral centroid**: Indica el centro de masa del espectro.
- **spectral bandwidth**: Intervalo de longitud de onda en el cual la cantidad de espectro radiado es mayor o igual que la mitad de su valor máximo.
- **rolloff**: Mide lo escarpado que es la función de transmisión con frecuencia.

- **zero crossing rate**: Tasa a la que la señal cambia de positiva a negativa o al revés.

En primer lugar escalamos el conjunto de datos, para que todas las variables tengan la misma importancia. Para cada género musical i tenemos datos de esas 8 variables medidas sobre 100 canciones. Calculando la matriz de covarianzas muestral Σ_i de esos datos, podemos plantearnos ahora el problema de hallar las componentes principales grupales para clasificar los 10 géneros musicales. Hay que resaltar que para la clasificación que vamos a dar ahora, lo único que nos va a importar son las direcciones de crecimiento de las variables en las matrices de covarianza de cada uno de los géneros. La clasificación se hace buscando en qué géneros musicales dichas componentes son similares, es decir, agrupando a los géneros para los que existan unas CPG en las cuales las matrices de covarianzas de todos ellos "casi diagonalicen".

- En primer lugar, calculamos las componentes principales comunes asociadas al baricentro de las 10 matrices de covarianzas. El cociente asociado nos queda 0.9436. Esto nos quiere decir que las componentes principales de todos los géneros son similares, las variables tienen direcciones de crecimiento parecidas en todos los géneros. Si nuestro objetivo fuera encontrar unas componentes adecuadas en las que representar todos los géneros, utilizar las CPC sería una buena elección. Sin embargo, como nuestro objetivo es llevar a cabo una clasificación, podemos calcular las CPG para m grupos e ir viendo qué géneros tienen componentes más parecidas.

- Mostramos ahora los resultados que nos dan las CPG para distintos valores de m. En cada caso mostramos la clasificación obtenida, así como el valor de la diagonalización en cada grupo:

  Para m=2:

Grupo	Medida diagonalización en el grupo	Géneros que forman el grupo
Grupo 1	0.9655	blues, country,disco, jazz, pop, reggae
Grupo 2	0.9446	classical, hiphop, metal, rock

  Para m=3:

Grupo	Medida diagonalización en el grupo	Géneros que forman el grupo
Grupo 1	0.9655	blues, country, disco, jazz, pop, reggae
Grupo 2	1	classical
Grupo 3	0.9552	hiphop, metal, rock

  Para m=4:

Grupo	Medida diagonalización en el grupo	Géneros que forman el grupo
Grupo 1	0.9631	blues, country, jazz, pop, reggae
Grupo 2	1	classical
Grupo 3	0.9552	hiphop, metal, rock
Grupo 4	1	disco

Comentamos los resultados obtenidos:

- Para m=2 agrupa a la música clásica con estilos como el metal, el rock y el hiphop, en contra de lo que cabía esperar. Sin embargo, la diagonalización en este grupo es peor que en los anteriores, luego las direcciones de crecimiento de las variables en esos estilos no tienen por qué ser tan similares entre sí como en el otro grupo. De hecho, cuando hacemos m=3 o m=4, el grupo se divide, dejando por un lado al metal, el rock y el hiphop y por otra a la música clásica, que forma un grupo ella sola. Esto es coherente con lo que cabía esperar: la música clásica es bastante distinta al resto de géneros.
- Casi todos los estilos que aparecen en el grupo 1 formado cuando $m = 2$, son estilos de música que parecen "más tranquilos"que el rock, el heavy y el hiphop, intuitivamente hablando. El único estilo que no encaja del todo en esta descripción de "música tranquila"podría ser la música disco. Cuando hacemos $m = 4$, la clasificación obtenida lo separa del resto

Por tanto, lo que hemos podido apreciar es que detrás los estilos de música que muchas veces nos suenan parecidos, tienen detrás una estructura bastante similar en cuanto a las relaciones y las direcciones de crecimiento existentes entre las variables que estamos estudiando.

Podemos intentar dar un enfoque distinto al problema, permitiendo un recorte de nivel α. Si nos fijamos en los resultados obtenidos, básicamente se han formado dos grupos al hacer m=2, y al aumentar el valor de m se han ido sacando de esos grupos a los estilos más diferentes, que han ido formando grupos con un único género musical. En esta situación, parece razonable intentar buscar únicamente 2 grupos, permitiendo que se desprecie una proporción α de los géneros musicales. Los resultados que obtenemos al trabajar de esta forma son:

Para $\alpha = 0{,}1$:

Grupo	Medida diagonalización en el grupo	Géneros que forman el grupo
Grupo 1	0.9655	blues, country, disco, jazz, pop, reggae
Grupo 2	0.9552	hiphop, metal, rock

Para $\alpha = 0{,}2$:

Grupo	Medida diagonalización en el grupo	Géneros que forman el grupo
Grupo 1	0.9631	blues, country, jazz, pop, reggae
Grupo 2	0.9552	hiphop, metal, rock

Llegamos a los mismos resultados que habíamos obtenido trabajando sin recortes. En ejemplos como este en los que no aparecen muchas probabilidades, los resultados son sencillos de interpretar con cualquiera de los métodos que hemos utilizado. Sin embargo, en ejemplos más complejos en los que aparezcan muchas probabilidades, el uso de los k-baricentros recortados va a ser mucho más recomendable, ya que simplemente despreciará las probabilidades que más difieran del resto.

Nota 8. Los datos han sido descargados de https://www.kaggle.com/insiyeah/musicfeatures. De las 30 variables presentes en el conjunto de datos, para realizar el analisis nos hemos quedado con el estilo musical y las 8 variables explicadas arriba.

7. Apéndice

7.1. Probabilidades de transición

Comencemos explicando este concepto el **caso discreto**. Supongamos que tenemos un conjunto de causas $\{c_1, ..., c_n\}$ y un conjunto de efectos $\{e_1, ..., e_m\}$. Cada causa c_i tiene una probabilidad $p_i > 0$ y además tenemos unas probabilidades de transición P_{c_i} para cada $i = 1, ..., n$, definidas en $\{e_1, ..., e_m\}$, tal que $P_{c_i}(e_j)$ nos da la probabilidad del efecto c_j si sabemos que ha ocurrido la causa c_i. Queremos entonces definir una probabilidad en

$$\{c_1, ..., c_n\} \times \{e_1, ..., e_m\}$$

que sea razonable, es decir, que la probabilidad marginal de la causa c_i sea p_i, y que la probabilidad del efecto j condicionada a la causa i sea $P_{c_i}(e_j)$. Definiendo $P(c_i, e_j) = p_i P_{c_i}(e_j)$, se cumplen ambas propiedades:

$$P(c_i) = \sum_{j=1}^{m} P(c_i, e_j) = \sum_{j=1}^{m} p_i P_{c_i}(e_j) = p_i \sum_{j=1}^{m} P_{c_i}(e_j) = p_i$$

$$P(e_j/c_i) = \frac{P(c_i, e_j)}{P(c_i)} = \frac{p_i P_{c_i}(e_j)}{p_i} = P_{c_i}(e_j)$$

Queremos ahora extender estas ideas al **caso de probabilidades no discretas**. Los resultados que vamos a exponer aparecen en [12, cap. 6]. Supongamos que tengamos dos espacios medibles (Ω_1, σ_1) y (Ω_2, σ_2), una probabilidad μ sobre (Ω_1, σ_1) y para cada $x \in \Omega_1$ una probabilidad de transición ν_x sobre (Ω_2, σ_2). Supongamos además que para cada $B \in \sigma_2$, la función $x \mapsto \nu_x(B)$ es σ_1-medible. Se define entonces en $(\Omega_1 \times \Omega_2, \sigma)$, siendo σ la mínima σ-álgebra que contiene a conjuntos $A \times B$ con A y B medibles, la siguiente probabilidad:

$$P(D) = \int_{\Omega_1} \nu_x(D_x) d\mu(x) \qquad \forall\, D \in \sigma \tag{14}$$

Por el teorema de las medidas producto (ver [12, teor. 2.6.2]) sabemos que P es realmente una probabilidad en $\Omega_1 \times \Omega_2$, y además es la única que satisface:

$$P(A \times B) = \int_A \nu_x(B) d\mu(x)$$

para todo $A \in \sigma_1, B \in \sigma_2$. De aquí se deduce que la probabilidad marginal de P sobre Ω_1 coincide con μ:

$$P(A \times \Omega_2) = \int_A \nu_x(\Omega_2) d\mu(x) = \int_A 1 d\mu(x) = \mu(A)$$

Denotemos (x, y) a los puntos de $\Omega_1 \times \Omega_2$. Si suponemos que para un cierto $x_0 \in \Omega_1$, se cumple que $\mu(x_0) > 0$, entonces la probabilidad de y condicionada a $x = x_0$ coincide con la probabilidad de transición ν_{x_0}:

$$P(y \in B/x = x_0) = \frac{P(\{x_0\} \times B)}{P(x = x_0)} = \frac{\int_{\{x_0\}} \nu_x(B) d\mu(x)}{\mu(x_0)} = \frac{\mu(x_0)\nu_{x_0}(B)}{\mu_(x_0)} = \nu_{x_0}(B)$$

para todo $B \in \sigma_2$. De esta forma, las probabilidades de transición se pueden entender como una generalización de las probabilidades condicionadas de P, definidas para todos los $x \in \Omega_1$, incluso cuando $\mu(x) = 0$.

Nuestro objetivo ahora va a ser realizar este razonamiento en sentido contrario. Partiendo de una probabilidad P sobre $\Omega_1 \times \Omega_2$, queremos encontrar unas probabilidades de transición que cumplan la relación (14). La existencia de dichas probabilidades no siempre está garantizada, necesitaremos imponer condiciones topológicas sobre Ω_2 para asegurar su existencia. Nosotros nos centraremos en demostrar que existen en el caso $\Omega_2 = \mathbb{R}^d$, aunque se puede probar su existencia en condiciones más generales, cuando Ω_2 es un espacio métrico completo y separable.

Proposición 17. *Sean (Ω_1, σ_1) y $(\Omega_2, \sigma_2) = (\mathbb{R}^d, \beta^d)$ dos espacios medibles, y P una probabilidad definida en el espacio producto $(\Omega_1 \times \Omega_2, \sigma)$, siendo σ la σ-álgebra producto. Denotemos por μ a la probabilidad marginal de P sobre Ω_1. Entonces existen probabilidades de transición $\{\nu_x\}_{x \in \Omega_1}$ en (Ω_2, σ_2) tales que para cada $D \in \sigma$,*

$$P(D) = \int_{\Omega_1} \nu_x(D_x) d\mu(x)$$

Demostración. Para cada $B \in \beta^d$ fijo, consideramos la medida:

$$\lambda_B(A) = P(A \times B)$$

λ_B es una medida positiva y finita en Ω_1, y además es absolutamente continua respecto a μ, ya que:

$$\mu(A) = 0 \Rightarrow \lambda_B(A) = P(A \times B) \leq P(A \times \Omega_2) = \mu(A) = 0 \Rightarrow \lambda_B(A) = 0$$

Por el teorema de Lebesgue-Radon-Nikodym, existe una función g_B medible en (Ω_1, σ_1) y única μ-c.s. tal que:

$$\lambda_B(A) = P(A \times B) = \int_A g_B(x) d\mu(x) \quad \forall\, A \in \sigma_1$$

Sea entonces para cada $q = (q_1, ..., q_d) \in \mathbb{Q}^d$ fijo, el conjunto:

$$I_q = \prod_{i=1}^{d} (-\infty, q_1] \in \beta^d$$

Para simplificar la notación, denotemos $\lambda_q = \lambda_{I_q}$ y $g_q = g_{I_q}$. Sea entonces $F_x(q) = g_q(x)$ para cada $x \in \Omega_1, q \in \mathbb{Q}^d$. Vamos a ver propiedades que nos van a permitir extender de forma razonable F_x a una función de distribución de una probabilidad en $(\mathbb{R}^d, \beta^d)$ para casi todo $x \in \Omega_1$.

1. $F_x(q) \geq 0$ para todo $q \in \mathbb{Q}^d$ μ-c.s.

   Para cada $q \in \mathbb{Q}^d$, λ_q es una medida positiva y por tanto $g_q \geq 0$ μ-c.s. Por tanto, para cada $q \in \mathbb{Q}^d$, existe un conjunto A_q con $\mu(A_q) = 1$ tal que $F_x(q) = g_q(x) \geq 0$ para todo

$x \in A_q$. Tomando $A = \cap_{q \in \mathbb{Q}^d} A_q$ tenemos un conjunto de medida 1, por ser intersección numerable de conjuntos de medida 1, y en el que para todo $x \in A$ se verifica que:

$$F_x(q) \geq 0 \text{ para todo } q \in \mathbb{Q}^d$$

2. $F_x(q) \leq 1$ para todo $q \in \mathbb{Q}^d$ μ-c.s.

   Veamos en primer lugar que dado $q \in \mathbb{Q}^d$, el conjunto $B_q = \{x \in \Omega_1 : g_q(x) \leq 1\}$ tiene probabilidad 1 en Ω_1. Razonamos por reducción al absurdo. Si no fuera así, el conjunto

   $$B_q^C = \{x \in \Omega_1 : g_q(x) > 1\}$$

   tendría probabilidad positiva, y por tanto se tendría que:

   $$\mu(B_q^C) = P(B_q^C \times \mathbb{R}^d) \geq P(B_q^C \times I_q) =$$

   $$= \int_{\{x \in \Omega_1 : g_q(x) > 1\}} g_q(x) d\mu(x) > \int_{\{x \in \Omega_1 : g_q(x) > 1\}} d\mu(x) = \mu(B_q^c)$$

   y llegamos por tanto a un absurdo. Por tanto, $\mu(B_q) = 1$ y si tomamos $B = \cap_{q \in \mathbb{Q}^d} B_q$, tenemos un conjunto de medida 1 en Ω_1 por ser intersección numerable de conjuntos de medida 1, y en el que para todo $x \in B$ se verifica que:

   $$F_x(q) \leq 1 \text{ para todo } q \in \mathbb{Q}^d$$

3. $\Delta_{r,s} F_x = \sum_{y \in V} sign(y) F_x(y) \geq 0$ para todos $r = (r_1, ..., r_d), s = (s_1, ..., s_d), \in \mathbb{Q}^d$ con $r \leq s$ (componente a componente) μ-c.s., siendo V el conjunto de 2^d vértices de $(r_1, s_1] \times ... \times (r_d, s_d]$ y $sign(y)$ 1 o -1, según que el número de componentes de y que coinciden con las de r es par o impar.

   Dados $r \leq s \in \mathbb{Q}^d$, veamos que el conjunto

   $$C_{r,s} = \{x \in \Omega_1 : \sum_{y \in V} sign(y) F_x(y) \geq 0\} = \{x \in \Omega_1 : \sum_{y \in V} sign(y) g_y(x) \geq 0\}$$

   cumple que $\mu(C_{r,s}) = 1$. Si no fuera así, el conjunto

   $$C_{r,s}^C = \{x \in \Omega_1 : \sum_{y \in V} sign(y) g_y(x) < 0\}$$

   tendría probabilidad positiva, y por tanto se tendría que:

   $$P\left(C_{r,s}^C \times ((r_1, s_1] \times ... \times (r_d, s_d])\right) = \sum_{y \in V} sign(y) P(C_{r,s}^C \times I_y) =$$

   $$= \sum_{y \in V} sign(y) \int_{C_{r,s}^C} g_y(x) d\mu(x) = \int_{C_{r,s}^C} \left(\sum_{y \in V} sign(y) g_y(x) \right) d\mu(x) < 0$$

Hemos llegado a un absurdo, luego $\mu(C_{r,s}) = 1$. Tomando entonces

$$C = \bigcap_{r,s \in \mathbb{Q}^d : r \leq s} C_{r,s}$$

tenemos un conjunto de medida 1 en Ω_1 y que cumple que para todo $x \in C$:

$$\Delta_{r,s} F_x = \sum_{y \in V} sign(y) F_x(y) \geq 0 \quad \forall\, r, s \in \mathbb{Q}^d \text{ con } r \leq s$$

4. $\lim_{n\to\infty} F_x(q_n) = 1$ y $\lim_{n\to\infty} F_x(q_{-n}) = 0$ μ-c.s., siendo $q_n = (n, n, ..., n) \in \mathbb{Q}^d$.

   Consideramos el conjunto $D_1 = \{x \in \Omega_1 : \lim_{n\to\infty} F_x(q_n) = 1\}$. Veamos que $\mu(D_1) = 1$. Si no fuera así, entonces $D_1^C = \{x \in \Omega_1 : \lim_{n\to\infty} F_x(q_n) \neq 1\}$ tendría medida positiva. Por los apartados anteriores, $\{F_x(q_n)\}_{n=1}^{\infty}$ acotada por 1 c.s., y además es creciente c.s. Esto es consecuencia de que para cada $n \in \mathbb{N}$, el conjunto

   $$R_n = \{x \in \Omega_1 : F_x(q_n) > F_x(q_{n+1})\} = \{x \in \Omega_1 : g_{q_n}(x) > g_{q_{n+1}}(x)\}$$

   es de medida nula, ya que si no fuera así:

   $$P(R_n \times I_{q_n}) = \int_{R_n} g_{q_n}(x) d\mu(x) > \int_{R_n} g_{q_{n+1}}(x) d\mu(x) = P(R_n \times I_{q_{n+1}})$$

   lo cuál no puede ocurrir ya que $R_n \times I_{q_n} \subset R_n \times I_{q_{n+1}}$. La sucesión es creciente fuera de $\bigcup_{n=1}^{\infty} R_n$, que es de medida nula en Ω_1
   Por tanto, para casi todo $x \in D_1^C$, la sucesión $\{F_x(q_n)\}_{n=1}^{\infty}$ converge hacia algún valor $l < 1$, y se tiene entonces que si $\mu(D_1^C) > 0$:

   $$\mu(D_1^C) = P(D_1^C \times \mathbb{R}^d) = \lim_{n\to\infty} P(D_1^C \times I_{q_n}) = \lim_{n\to\infty} \int_{D_1^C} g_{q_n}(x) d\mu(x) =$$

   $$= \int_{D_1^C} (\lim_{n\to\infty} g_{q_n})(x) d\mu(x) = \int_{D_1^C} (\lim_{n\to\infty} F_x(q_n)) d\mu(x) < \mu(D_1^C)$$

   donde la penúltima igualdad es consecuencia del teorema de la convergencia dominada. Hemos llegado a una contradicción, y por tanto $\mu(D_1^C) = 0 \Rightarrow \mu(D_1) = 1$.

   De forma análoga se prueba que existe un conjunto D_2 con $\mu(D_2) = 1$ en el cual $\lim_{n\to\infty} F_x(q_{-n}) = 0$ para todo $x \in D_2$. Tomando $D = D_1 \cap D_2$ tenemos un conjunto de medida 1 en Ω_1 en el que para cada $x \in D$ se cumple que:

   $$\lim_{n\to\infty} F_x(q_n) = 1 \text{ y } \lim_{n\to\infty} F_x(q_{-n}) = 0$$

5. $\lim_{n\to\infty} F_x(q + s_n) = F_x(q)$ para todo $q \in \mathbb{Q}^d$ μ-c.s., siendo $s_n = (\frac{1}{n}, ..., \frac{1}{n}) \in \mathbb{Q}^d$.

   Para cada $q \in \mathbb{Q}^d$, consideramos el conjunto:

   $$E_q = \{x \in \Omega_1 : \lim_{n\to\infty} F_x(q + s_n) = F_x(q)\} = \{x \in \Omega_1 : \lim_{n\to\infty} g_{q+s_n}(x) = g_q(x)\}$$

Veamos que $\mu(E_q) = 1$. Para ello veamos que $\lim_{n\to\infty} g_{q+s_n}(x) = g_q(x)$ μ-c.s. El límite sabemos que existe c.s., puesto que es c.s una sucesión decreciente de funciones (razonando igual que en el apartado anterior para ver que era creciente). Para cada $K \in \sigma_1$ se cumple que:

$$P(K \times I_q) = \lim_{n\to\infty} P(K \times I_{q+s_n}) = \lim_{n\to\infty} \int_K g_{q+s_n}(x)d\mu(x) =$$

$$= \int_K (\lim_{n\to\infty} g_{q+s_n})(x)d\mu(x)$$

y por la unicidad c.s. de la función g_q se tiene que $\lim_{n\to\infty} g_{q+s_n} = g_q$ μ-c.s. Por tanto tenemos que $\mu(E_q) = 1$, y si tomamos

$$E = \bigcap_{q\in\mathbb{Q}^d} E_q$$

tenemos un conjunto de medida 1 en Ω_1 y en el que si $x \in E$, entonces verifica que $\lim_{n\to\infty} F_x(q + s_n) = F_x(q)$ para todo $q \in \mathbb{Q}^d$

Por tanto, tomando $F = A \cap B \cap C \cap D \cap E$, conjunto de medida 1 en Ω_1, para cada $x \in F$ se cumplen las propiedades 1), 2), 3), 4) y 5). De las 4 primeras propiedades se deduce que podemos definir para cada $x \in F$ una función de distribución de la siguiente forma:

$$G_x(y) = \lim_{q\to y^+} F_x(q) \quad \forall\, y \in \mathbb{R}^d$$

Y de la propiedad 5) deducimos además que para cada $r \in \mathbb{Q}^d$ y para cada $x \in F$ se cumple que

$$G_x(r) = \lim_{q\to r^+} F_x(q) = F_x(r) = g_r(x)$$

Para cada $x \in F$, sea ν_x la probabilidad definida por la función de distribución G_x, y para cada $x \notin F$ sea ν_x otra probabilidad cualquiera en $\mathbb{R}^d$. Consideramos la familia de conjuntos:

$$\mathscr{C} = \{H \in \beta^d : \nu_x(H) = g_H(x)\, c.s.\}$$

Para cada $I_q = (-\infty, q_1] \times ... \times (-\infty, q_d]$ con $q = (q_1, ..., q_d) \in \mathbb{Q}^d$ se tiene que si $x \in F$, entonces:

$$\nu_x(I_q) = G_x(q) = g_q(x) = g_{I_q}(x)$$

Se tiene por tanto que $\mathscr{E} = \{I_q : q \in \mathbb{Q}^d\} \subset \mathscr{C}$, y veamos que de hecho también contiene a los productos de intervalos $(s_1, r_1] \times ... \times (s_d, r_d]$ con extremos racionales, que forman un álgebra de conjuntos. Para ver esto, basta ver que si $K, H \in \mathscr{C}$ y además $K \subset H$, entonces $H \setminus K \in \mathscr{C}$.

Si $\nu_x(H) = g_H(x)$ c.s. y $\nu_x(K) = g_K(x)$ c.s., entonces para casi todo x se verifican las dos igualdades. Además se tiene que para todo $J \in \sigma_1$:

$$P(J\times(L\setminus K)) = P(J\times L) - P(J\times K) = \int_J g_H(x)d\mu(x) - \int_J g_K(x)d\mu(x) = \int_J (g_H(x) - g_K(x))d\mu(x)$$

y por unicidad, $g_{H\setminus K} = g_H - g_K$ c.s. Por tanto se tienen c.s. las siguientes igualdades:

$$\nu_x(H \setminus K) = \nu_x(H) - \nu_x(K) = g_H(x) - g_K(x) = g_{H\setminus K}(x)$$

y por tanto $H \setminus K \in \mathscr{C}$.

Veamos que $\mathscr{C}$ es una clase monótona. Sea $\{H_j\}_{j=1}^{\infty}$ una sucesión creciente de conjuntos de $\mathscr{C}$. En primer lugar, se tiene que para cada $J \in \Omega_1$:

$$P(J \times \cup_{j=1}^{\infty} H_j) = \lim_{j\to\infty} P(J \times H_j) = \lim_{j\to\infty} \int_J g_{H_j}(x)d\mu(x) =$$

$$= \int_J (\lim_{j\to\infty} g_{H_j}(x))d\mu(x)$$

y por tanto por la unicidad, $\lim_{j\to\infty} g_{H_j} = g_{\cup_{j=1}^{\infty} H_j}$ c.s., puesto que el límite sabemos que existe c.s. por ser g_{H_j} una sucesión creciente de funciones casi seguro (esto se ve igual que en el apartado 4) y acotadas c.s. por 1. Luego se cumple μ-c.s. que:

$$\nu_x(\cup_{j=1}^{\infty} H_j) = \lim_{j\to\infty} \nu_x(H_j) = \lim_{j\to\infty} g_{H_j}(x) = g_{\cup_{j=1}^{\infty} H_j}(x)$$

Por tanto $\cup_{j=1}^{\infty} H_j \in \mathscr{C}$. Como $\mathscr{C}$ es una clase monótona que contiene al álgebra formado por los productos de intervalos con extremos racionales, entonces necesariamente contiene a β^d, y por tanto hemos probado que:

$$\text{Para todo } H \in \beta^d \text{ se cumple que } \nu_x(H) = g_H(x)\, c.s.$$

Veamos que las probabilidades cumplen lo que queríamos. Dados $A \in \sigma_1, B \in \sigma_2$, sabemos que $\nu_x(B) = g_B(x)$ fuera de un conjunto S con $\mu(S) = 0$. Por tanto:

$$P(A \times B) = \lambda_B(A) = \int_A g_B(x)d\mu(x) = \int_{A\setminus S} g_B(x)d\mu(x) =$$

$$= \int_{A\setminus S} \nu_x(B)d\mu(x) = \int_A \nu_x(B)d\mu(x)$$

Si definimos ahora para cada $D \in \sigma$,

$$Q(D) = \int_{\Omega_1} \nu_x(D_x)d\mu(x)$$

sabemos que Q es una probabilidad en $(\Omega_1 \times \Omega_2, \sigma)$ y además coincide con P en los conjuntos $A \times B$ con $A \in \sigma_1, B \in \sigma_2$, que sabemos que generan la σ-álgebra producto σ. Por tanto necesariamente coinciden en σ, luego:

$$P(D) = \int_{\Omega_1} \nu_x(D_x)d\mu(x) \qquad \forall\, D \in \sigma$$

□

7.2. Convergencia débil: Definiciones y principales resultados.

A lo largo de todo el trabajo se utilizan argumentos relacionados con la convergencia débil para la prueba de numerosos resultados. Todas esta teoría aparece desarrollada en [11]. En esta parte, vamos a enunciar los conceptos y los resultados más importantes para todo el desarrollo teórico expuesto a lo largo del trabajo.

Definición 15. *Sean $\{\mu_n\}_{n=1}^{\infty}$ y μ probabilidades en $(\mathbb{R}^d, \beta^d)$, y sean respectivamente $\{F_n\}_{n=1}^{\infty}$ y F sus funciones de distribución asociadas. Se dice que $\{\mu_n\}_{n=1}^{\infty}$ converge débilmente hacia μ si se cumple que:*

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ para todo punto de continuidad } x \text{ de } F$$

y en ese caso se representa por $\mu_n \to_d \mu$.

Si $\{U_n\}_{n=1}^{\infty}$ y U son variables aleatorias que toman valores en $\mathbb{R}^d$ (que pueden estar definidos en diferentes espacios probabilísticos) se dice que U_n converge débilmente hacia U si

$$\mathscr{L}(U_n) \to_d \mathscr{L}(U)$$

y en ese caso se representa por $U_n \to_d U$.

En primer lugar, vamos a relacionar la convergencia débil con los principales tipos de convergencia de variables aleatorias. Si $\{U_n\}_{n=1}^{\infty}$ y U son variables aleatorias definidas en un mismo espacio probabilístico (Ω, σ, P) que toman valores en $\mathbb{R}^d$, entonces recordemos que:

- U_n converge *casi seguro* hacia U si $\lim_{n\to\infty} U_n(\omega) = U(\omega)$ para casi todo $\omega \in \Omega$. En ese caso se escribe $U_n \to_{c.s.} U$.
- U_n converge *en probabilidad* hacia U si para todo $\epsilon > 0$, $\lim_{n\to\infty} P(\|U_n - U\| > \epsilon) = 0$. En ese caso se escribe $U_n \to_p U$.

Se tiene la siguiente relación:

Proposición 18. *Sean $\{U_n\}_{n=1}^{\infty}$ y U son variables aleatorias definidas en un mismo espacio probabilístico (Ω, σ, P) que toman variables en $\mathbb{R}^d$. Entonces:*

1. *Si $U_n \to_{c.s.} U \quad \Rightarrow \quad U_n \to_p U$*
2. *Si $U_n \to_p U \quad \Rightarrow \quad U_n \to_d U$*

Para la siguiente definición utilizaremos el concepto de *rectángulo acotado*. Un rectángulo acotado es un conjunto $A \subset \mathbb{R}^d$ tal que:

$$A = I_1 \times I_2 \times \cdots \times I_d$$

siendo $I_j, j = 1, 2, ..., d$ intervalos acotados en $\mathbb{R}$.

Definición 16. *Se dice que una sucesión de probabilidades $\{\mu_n\}_{n=1}^{\infty}$ en $(\mathbb{R}^d, \beta^d)$ es ajustada si para todo $\epsilon > 0$, existe un rectángulo acotado A tal que*

$$\mu_n(A) > 1 - \epsilon \quad \text{para todo } n \in \mathbb{N}$$

Si $\{U_n\}_{n=1}^{\infty}$ son variables aleatorias que toman valores en $\mathbb{R}^d$, se dice que $\{U_n\}_{n=1}^{\infty}$ es ajustada si lo es la sucesión $\{\mathscr{L}(U_n)\}_{n=1}^{\infty}$ de sus leyes de probabilidad.

El siguiente teorema es de gran importancia, ya que se utiliza para probar la mayoría de resultados relacionados con la convergencia débil.

Teorema 4. *(**de Helly**): Sea $\{\mu_n\}_{n=1}^{\infty}$ es una sucesión de probabilidades en $(\mathbb{R}^d, \beta^d)$. Entonces $\{\mu_n\}_{n=1}^{\infty}$ es ajustada si y sólo si para todo subsucesión $\{\mu_{n_m}\}_{m=1}^{\infty}$, existe una nueva subsucesión $\{\mu_{n_{m_k}}\}_{k=1}^{\infty}$ y una probabilidad μ en $(\mathbb{R}^d, \beta^d)$ tales que $\mu_{n_{m_k}} \to_d \mu$*

Vamos ahora a estudiar algunas propiedades que nos permiten caracterizar la convergencia débil. Para ello, definamos primero el concepto de *clase separante*. Un conjunto de funciones $\eta \subset \mathscr{C}(\mathbb{R}^d)$ es una clase separante si dadas μ, ν probabilidades en $(\mathbb{R}^d, \beta^d)$, la condición:

$$\int_{\mathbb{R}^d} f(x)d\mu(x) = \int_{\mathbb{R}^d} f(x)d\nu(x) \quad \text{para toda } f \in \eta$$

implica que $\mu = \nu$.

Proposición 19. *Si $\{\mu_n\}_{n=1}^{\infty}$ es una sucesión ajustada de probabilidades en $(\mathbb{R}^d, \beta^d)$, y η es una clase separante, entonces:*

$$\mu_n \to_d \mu \Leftrightarrow \lim_{n\to\infty} \int_{\mathbb{R}^d} f(x)d\mu_n(x) = \int_{\mathbb{R}^d} f(x)d\mu(x) \quad \text{para toda } f \in \eta$$

El siguiente teorema es de gran utilidad, ya que resume las principales equivalencias existentes entre la convergencia débil de una sucesión de probabilidades y otras convergencias de integrales de funciones o de probabilidades de determinados conjuntos.

Teorema 5. *(**Portmanteau**): Si $\{\mu_n\}_{n=1}^{\infty}$ y μ son probabilidades en $(\mathbb{R}^d, \beta^d)$, son equivalentes las siguientes condiciones:*

1. $\mu_n \to_d \mu$
2. $\lim_{n\to\infty} \int_{\mathbb{R}^d} f(x)d\mu_n(x) = \int_{\mathbb{R}^d} f(x)d\mu(x)$ para toda $f \in \mathscr{C}(\mathbb{R}^d)$
3. $\limsup_{n\to\infty} \mu_n(A) \leq \mu(A)$ *para todo A cerrado, $A \subset \mathbb{R}^d$*
4. $\liminf_{n\to\infty} \mu_n(A) \geq \mu(B)$ *para todo B abierto, $B \subset \mathbb{R}^d$*
5. $\lim_{n\to\infty} \mu_n(A) = \mu(A)$ *para todo conjunto A de μ-continuidad (A es de μ-continuidad si verifica que $\mu(\bar{A} \setminus \mathring{A}) = 0$)*

El teorema que vamos a enunciar ahora nos permite relacionar la convergencia débil con la convergencia casi seguro. La importancia de relacionar estos dos conceptos se encuentra principalmente en que vamos a poder aplicar los teoremas de paso al límite ampliamente conocidos para la convergencia casi seguro: el teorema de la convergencia monótona, el teorema de la convergencia dominada y el lema de Fatou.

Teorema 6. *(**de Skorohod**): Si $\{\mu_n\}_{n=1}^{\infty}$ y μ son probabilidades en $(\mathbb{R}^d, \beta^d)$, entonces $\mu_n \to_d \mu$ si y sólo si existe un espacio probabilístico (Ω, σ, P) y variables aleatorias U_n y U definidas en Ω, con valores en $\mathbb{R}^d$ y con $\mathscr{L}(U_n) = \mu_n \, \forall \, n \in \mathbb{N}$ y $\mathscr{L}(U) = \mu$, tales que $\mu_n \to_{c.s.} \mu$.*

En el caso particular en que $d = 1$, si $\{F_n\}_{n=1}^{\infty}$ y F son las funciones de distribución asociadas a $\{\mu_n\}_{n=1}^{\infty}$ y μ, se tiene que:

$$\mu_n \to_d \mu \Leftrightarrow F_n^{-1}(y) \to_{c.s.} F^{-1}(y)$$

Por último, vamos a explicar un resultado que nos permite relacionar la convergencia en los espacios L_p con la convergencia en probabilidad, y que combinado con los resultados anteriores, nos permite relacionar la convergencia en los espacios L_p con la convergencia débil . Dado un espacio probabilístico (Ω, σ, P), entendemos que el espacio $L^p(\Omega)$, con $1 \leq p < \infty$, es el espacio de los vectores aleatorios $U : \Omega \to \mathbb{R}^d$ medibles y tales que $\|U\|^p$ es integrable. Este espacio es un espacio normado con las norma:

$$\|U\|_p = \left(\int_\Omega \|U\|^p dP \right)^{1/p} = E\left(\|U\|^p \right)^{1/p}$$

Antes de dar el resultado, recordemos que una sucesión de variables aleatorias reales $\{U_n\}_{n=1}^\infty$ se dice que es *uniformemente integrable* si

$$\lim_{a \to \infty} \left(\sup_{n \in \mathbb{N}} \int_{\{\|U_n\| > a\}} \|U_n\| dP \right) = 0$$

Se tiene entonces que:

Proposición 20. ***(caracterización de la convergencia en los espacios*** $L_p(\Omega)$***).*** *Si* $\{U_n\}_{n=1}^\infty$ *y* U *son v.a. en* $L_p(\Omega)$*, son equivalentes las siguientes afirmaciones:*

1) $U_n \xrightarrow{L_p(\Omega)} U$

2) $U_n \longrightarrow_p U$ *y* $E\|U_n\|^p \xrightarrow{n \to \infty} E\|U\|^p$

3) $U_n \longrightarrow_p U$ *y* $\{\|U_n\|^p\}_{n=1}^\infty$ *es uniformemente integrable.*

Referencias

[1] L. Ruschendorf, S.T. Rachev (1990). *A Characterization of Random Variables with Minimum L^2-Distance*. J. Mult. Anal. 32: 48-54

[2] P. C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos, C. Matrán (2016). *A fixed-point approach to barycenters in Wasserstein space*. J. Math. Anal. Appl. 441: 744–762

[3] M. Agueh, G. Carlier (2011). *Barycenters in the Wasserstein space*. SIAM J. Math. Anal. 43(2): 904–924

[4] B. N. Flury (1984). *Common principal components in K groups*. J. Amer. Stat. Assoc. Vol. 79, 388: 892-898

[5] L.P. Lefkovitch (1993). *Consensus Principal Components*. Biom J 35.5: 567-580

[6] R. Bathia (1997). *Matrix Analysis*. Springer.

[7] J. A. Cuesta-Albertos, C. Matrán-Bea, A. Tuero-Diaz (1996). *On lower bounds for the L^2-Wasserstein metric in a Hilbert space*. J. Theor. Prob., Vol. 9, 2

[8] P. Major (1978). *On the invariance principle for sums of independent identically distributed random variables* . J. Mult. Anal. 8: 487-517

[9] J. A. Cuesta-Albertos, L. Ruschendorf, A. Tuero-Diaz (1993). *Optimal coupling of multivariate distributions and stochastic processes*. J. Mult. Anal. 46: 335-361

[10] C. Villani (2006). *Optimal transport, old and new*. Springer.

[11] P. Billingsley (1995). *Probability and Measure, Third Edition*. Wiley Series in Probability and Statistics.

[12] R. B. Ash (1972). *Real Analysis and Probability*. Academic Press.

[13] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, A. Mayo-Íscar (2019). *Robust clustering tools based on optimal transportation*. Statistics and Computing 29: 139–160

[14] P. J. Bickel, D. A. Freedman (1981). *Some asymptotic theory for the bootstrap*. Ann. Statist. 9: 1196-1217

[15] V. M. Panaretos, Y. Zemel (2019). *Statistical aspects of Wasserstein distances*. Annu. Rev. Stat. Appl, 6: 405–31

[16] H. Fritz, L.A. García-Escudero, A. Mayo-Iscar (2012). *tclust: an R package for a trimming approach to cluster analysis*. J. Stat. Softw. 47(12)

[17] C. Villani (2003). *Topics in Optimal Transportation*. American Mathematical Society.

[18] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, C. Matrán (2018). *Wide consensus aggregation in the Wasserstein space. Application to location-scatter families.* Bernoulli 24(4A): 3147–3179