Facultad de Ciencias

LA DISTANCIA DE WASSERSTEIN. ANÁLISIS DE LA RELACIÓN ENTRE LA TEMPERATURA DEL PLANETA Y EL NIVEL DE CO_2

WASSERSTEIN'S DISTANCE. ANALYSIS OF THE RELATIONSHIP BETWEEN GLOBAL TEMPERATURE AND THE LEVEL OF CO_2

Trabajo de Fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autora: Alba Diego Velarde

Director: Juan Antonio Cuesta Albertos

Septiembre-2022

Índice

Resumen

El problema del transporte óptimo fue propuesto por Monge y consiste en transformar una distribución en otra minimizando el coste de transporte. Kantorovich presentó una relajación del problema de Monge que permitió probar la existencia y unicidad de solución del problema bajo ciertas condiciones. Este problema introduce de manera natural una distancia entre distribuciones, la distancia de Wasserstein, que se define como el coste mínimo de transporte entre ellas. A su vez, esta distancia permite definir la media de Fréchet, que es una generalización a espacios métricos de la media usual. Con estas herramientas, en el presente trabajo se propone un modelo de regresión no paramétrica funcional basado en núcleos que relaciona las distribuciones anuales (de 1954 a 2021) de las diferencias entre la temperatura media diaria de la superficie terrestre y la temperatura media entre 1951 y 1980, con las concentraciones anuales de CO_2 en la atmósfera. Se estiman las distribuciones de probabilidad esperadas para cada año en base al nivel de CO_2 de ese año.

Palabras clave: transporte óptimo, distancia de Wasserstein, media de Fréchet, regresión no paramétrica, distribución de probabilidad, cuantil, temperatura, CO_2

Abstract

The optimal transportation problem was proposed by Monge and consists of transforming one distribution into another while minimizing the transportation cost. Kantorovich presented a relaxation of Monge's problem which made it possible to prove the existence and uniqueness of solution of the problem under certain conditions. This problem naturally introduces a distance between distributions, the Wasserstein distance, which is defined as the minimum transport cost between them. In turn, this distance allows us to define the Fréchet mean, which is a generalization to metric spaces of the usual mean. With these tools, in the present work we propose a functional nonparametric regression model based on kernels that relates the annual distributions (from 1954 to 2021) of the differences between the mean daily land surface temperature and the mean temperature between 1951 and 1980, with the annual concentrations of CO_2 in the atmosphere. Expected probability distributions are estimated for each year based on that year's CO_2 level.

Keywords: optimal transport, Wasserstein distance, Fréchet mean, nonparametric regression, probability distribution, quantile, temperature, CO_2.

1. Introducción

Una de las transformaciones más importantes que ha ocurrido durante los últimos años es el crecimiento exponencial de datos. Hasta hace poco tiempo, la información era escasa y de difícil acceso. Sin embargo, se esperaba que esa información fuese de calidad elevada y permitiese obtener conclusiones realistas. Actualmente, el valor de la información no se encuentra tanto en los datos concretos, sino en cómo se correlacionan entre sí. Por tanto, es cada vez más necesario asumir una actitud crítica en la elección y tratamiento de los datos [25].

Los seres humanos estamos continuamente generando y almacenando cantidades enormes de datos (mediciones, documentos, imágenes, sonidos...) relacionados con una multitud de campos: medio ambiente, salud, transporte, seguridad, biomedicina... El análisis de estos permite comprender muchos procesos que ocurren a nuestro alrededor y con ello, diseñar soluciones a distintos problemas. En este trabajo se analizan datos de calentamiento global, en concreto, relaciones entre las distribuciones anuales de temperaturas medias diarias de la superficie terrestre, más concretamente distribuciones de diferencias entre la temperatura diaria de la superficie terrestre y la temperatura media de la superficie terrestre entre 1951 y 1980, y las concentraciones de CO_2 en la atmósfera. Una comprensión de las variables que influyen en el aumento de la temperatura del planeta permitirá adoptar medidas para paliar los problemas que pueda acarrear este aumento: incremento de la temperatura oceánica, deshielo, subida del nivel del mar... cuyas consecuencias son muy negativas para la especie humana.

Para estudiar la relación entre las distribuciones de temperaturas y la concentración de CO_2 se ha planteado un modelo de regresión no paramétrica. Pero el hecho de que los datos sean distribuciones de probabilidad supone una complicación: se ha de trabajar en espacios funcionales.

Como punto de partida para tratar con datos en forma de distribuciones de probabilidad, se ha de introducir una distancia entre estas. Una manera intuitiva de definirla es hacerlo mediante el mínimo coste de transformar una en otra, donde el coste de la transformación se puede ver como el coste de transformar un "montón de arena", con una localización y una forma correspondientes a una de las distribuciones, en otro montón cuya distribución y forma están determinadas por la otra distribución. Esta es la denominada distancia de Wasserstein, definida como el coste mínimo de la transformación entre distribuciones en el problema del transporte óptimo de Kantorovich. En el capítulo 2 se presentará el problema del transporte óptimo que introduce de manera natural la distancia de Wasserstein. La teoría asociada a esta distancia se introducirá en el capítulo 3. Al disponer de esta distancia, se puede definir el concepto de media de una colección de distribuciones, la media de Fréchet. Este concepto se tratará en el capítulo 4.

En el capítulo 5 se presentará la teoría de la regresión no paramétrica aplicada a datos funcionales. Se trabajará con la regresión local por núcleos, en concreto, con el estimador de Nadaraya-Watson.

Finalmente, se aplicará esta teoría, usando la distancia de Wasserstein, para estimar las distribuciones de diferencias diarias de temperatura de los años entre 1954 y 2021 condicionadas a la concentración de CO_2 en la atmósfera del año correspondiente.

Es preciso mencionar que el presente trabajo es eminentemente teórico y su objetivo es el estudio del problema del transporte óptimo y la distancia de Wasserstein. El problema del transporte óptimo es un clásico en los campos de la teoría de la probabilidad, la optimización y la economía. De hecho, se han otorgado dos medallas Fields relacionadas con este tema: en 2010 para Cédric Villani y en 2019 para Alessio Figalli. Se pretende mostrar la gran aplicabilidad de este problema con el análisis de regresión que relaciona las distribuciones de temperaturas y el nivel de CO_2.

2. Transporte óptimo

En este capítulo se enuncia el problema del transporte óptimo, el cual introduce de manera natural una distancia entre distribuciones de probabilidad. Este problema fue abordado por primera vez por el matemático francés Gaspard Monge. Su enunciado genera distintas complicaciones que fueron resueltas por el matemático y economista Leonid Kantorovich al proponer una relajación del problema inicial. Las ideas generales del capítulo se han tomado de las referencias [4], [12], [21], [26], [27], [31], [35]. Se irán referenciando a lo largo del texto los resultados más concretos.

2.1. El problema de Monge

El problema del transporte óptimo, siguiendo la interpretación de Monge, desarrollada en 1781, se formula de la siguiente manera: dado un montón de arena y un hoyo con volumen igual al del montón de arena, se trata de encontrar la manera óptima de transportar la arena al hoyo, con el objetivo de minimizar el esfuerzo o coste de moverla. La arena se puede transportar y, por tanto distribuir, de muchas formas en el hoyo. Así, la solución del problema del transporte óptimo es encontrar el modo de ir trasladando la arena del montón al hoyo con el menor coste posible.

Sin pérdida de generalidad, se puede suponer que el volumen de arena es 1. Por tanto, el problema del transporte óptimo consiste en modificar una distribución de probabilidad conocida en otra que también se conoce, minimizando el coste de transporte. Por ejemplo, si hipotéticamente el montón de arena tuviese forma triangular, se identificaría con una distribución de probabilidad con función de densidad una función triangular. Si el hoyo tuviese forma cuadrada, se identificaría con una distribución uniforme. De esta manera, el problema del transporte óptimo consistiría en transformar una distribución con función de densidad triangular en una distribución uniforme.

Los elementos matemáticos que modelizan el problema del transporte óptimo son: el espacio donde se encuentra la arena, $\mathscr{X}$; el espacio donde está hoyo, $\mathscr{Y}$; la función de coste, $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, que determina cuánto cuesta mover una unidad de arena desde el punto $x \in \mathscr{X}$ a un punto $y \in \mathscr{Y}$ del hoyo; y la función de transporte, $T : \mathscr{X} \to \mathscr{Y}$, que indica cómo se transporta cada unidad de arena al hoyo. La distribución de la arena en el espacio $\mathscr{X}$ se asocia a una medida de probabilidad μ en $\mathscr{X}$ y la forma del hoyo se identifica con una medida de probabilidad ν en el espacio $\mathscr{Y}$.

Para ser rigurosos, al hablar de medidas hay que definir σ-álgebras. Para ello, se asume a lo largo de este trabajo, que $\mathscr{X}$ e $\mathscr{Y}$ son espacios métricos, completos y separables y se escogen las σ-álgebras de Borel en $\mathscr{X}$ e $\mathscr{Y}$, $\beta_{\mathscr{X}}$ y $\beta_{\mathscr{Y}}$ respectivamente, de manera que $\mu : \beta_{\mathscr{X}} \to [0,1]$ y $\nu : \beta_{\mathscr{Y}} \to [0,1]$. Como consecuencia, el espacio producto $\mathscr{X} \times \mathscr{Y}$, dotado de la topología producto, es completo y separable. La σ-álgebra de Borel de $\mathscr{X} \times \mathscr{Y}$ es el producto de σ-álgebras de $\mathscr{X}$ e $\mathscr{Y}$, $\beta_{\mathscr{X}\times\mathscr{Y}} = \beta_{\mathscr{X}} \times \beta_{\mathscr{Y}}$, por lo que cualquier función continua $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$ es medible. Así, parece riguroso asumir que c sea continua y no negativa. Finalmente, se supone que T es medible, por lo que $T^{-1}(B) \in \beta_{\mathscr{X}}$ para todo $B \in \beta_{\mathscr{Y}}$ y $\mu(T^{-1}(B))$ está bien definido.

Llamaremos β_d a la σ-álgebra de Borel en $\mathbb{R}^d$ con la distancia usual y $\mathcal{L}$ a la medida de Lebesgue en este espacio.

La principal característica de la formulación de Monge del problema del transporte es suponer que toda la masa de un punto $x \in \mathscr{X}$ se transporta a un único punto $y = T(x) \in \mathscr{Y}$. Es decir, toda la arena situada en la "columna" con base en x se envía al mismo punto y del hoyo. Como se verá posteriormente, la formulación de Kantorovich prescinde de esta restricción.

El coste total del transporte es, por tanto,

$$C(T) = \int_{\mathscr{X}} c(x, T(x)) d\mu(x). \tag{1}$$

Como la masa de arena se ha de conservar en el transporte, no toda aplicación $T : \mathscr{X} \to \mathscr{Y}$ es admisible. Toda región $B \subseteq \mathscr{Y}$ del hoyo con volumen $\nu(B)$ tiene que recibir un volumen $\nu(B)$ de arena. La arena que se transporta a B viene representada por el conjunto $\{x \in \mathscr{X} : T(x) \in B\} = T^{-1}(B)$ y su volumen viene dado por $\mu(T^{-1}(B))$. Por tanto, ha de cumplirse la igualdad

$$\mu(T^{-1}(B)) = \nu(B), \text{ para todo } B \in \beta_{\mathscr{Y}}. \tag{2}$$

Esta condición se representa como $T\sharp\mu = \nu$ y se lee como: ν es la medida *push-forward* de μ vía T.

Se presenta a continuación un teorema que caracteriza las medidas *push-forward* [21] y que se utilizará a lo largo del capítulo.

Teorema 1. *Sean $\mathscr{X}$ e $\mathscr{Y}$ espacios métricos completos y separables. Sean μ y ν medidas de probabilidad en $\mathscr{X}$ e $\mathscr{Y}$ y sea $T : \mathscr{X} \to \mathscr{Y}$ una función medible. Entonces, $T\sharp\mu = \nu$ si y solo si*

$$\int_{\mathscr{Y}} \varphi(y) d\nu(y) = \int_{\mathscr{X}} (\varphi \circ T)(x) d\mu(x), \text{ para toda } \varphi \in C_b(\mathscr{X})$$

El problema de Monge se trata, por tanto, de encontrar una función medible T_0 en la que se alcance el ínfimo del coste total de transporte (1):

$$C(T_0) = \inf_{T : T\sharp\mu = \nu} \int_{\mathscr{X}} c(x, T(x)) d\mu(x).$$

Sin embargo, este problema presenta diversas complicaciones.

Para comenzar, el conjunto de transportes admisibles, $\{T : T\sharp\mu = \nu\}$, puede ser vacío. Si $\mu = \delta_{x_0}$ la medida de Dirac[1] en $x_0 \in \mathscr{X}$, ν no es una medida de Dirac y $B = \{T(x_0)\}$, se tiene que $\mu(T^{-1}(B)) = 1 > \nu(B)$. Por tanto, no se cumple (2) y no existe ningún T admisible, es decir, $\{T : T\sharp\mu = \nu\} = \emptyset$.

Si $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$ y μ y ν son absolutamente continuas[2] con respecto a la medida de Lebesgue y $A \in \beta_d$, por el teorema de Radon-Nikodym[3] existen funciones f y g medibles tales que,

$$\mu(A) = \int_A f(x) dx \quad \text{y} \quad \nu(A) = \int_A g(x) dx. \tag{3}$$

[1] **Definición 1** (Medida de Dirac, [2]). *Sea $(\Omega, \mathcal{A})$ un espacio medible. Dado un conjunto $A \in \mathcal{A}$ y $x \in A$, se define la medida de Dirac en x, δ_x, como aquella que cumple $\delta_x(A) = 1$ si $x \in A$ y $\delta_x(A) = 0$ si $x \notin A$.*

[2] **Definición 2** (Medida absolutamente continua, [2]). *Sea $(\Omega, \mathcal{A}, \mu)$ un espacio de medida. Una medida ν definida en $\mathcal{A}$ es absolutamente continua con respecto a una medida μ, $\nu \ll \mu$, si $\mu(A) = 0$ implica que $\nu(A) = 0$ para todo $A \in \mathcal{A}$.*

[3] **Teorema 2** (Radon-Nikodym, [2]). *Sea $(\Omega, \mathcal{A}, \mu)$ un espacio de medida $\sigma - finito$ (Ω unión numerable de conjuntos medibles de medida finita) y sea $\nu \ll \mu$. Entonces existe una única función $f \geq 0$ medible tal que*

$$\nu(E) = \int_E f d\mu, \text{ para todo } E \in \Omega.$$

f es única en el sentido de que si existe otra función medible g que cumpla lo anterior, $f = g$ μ-c.s.

Si además T es un difeomorfismo, se tiene (ver [21]):

$$\nu(T(A)) = \int_{T(A)} g(x)dx = \int_A g(T(x))|J_T(x)|dx. \tag{4}$$

En la última igualdad se ha aplicado el teorema de cambio de variable.

Teniendo en cuenta (3) y (4), $T\sharp\mu = \nu$ es equivalente a la ecuación

$$f(x) = g(T(x))|J_T(x)|.$$

conocida como ecuación de Monge-Ampère. Se trata de una ecuación no lineal, lo que supone una gran dificultad para obtener el conjunto de funciones de transporte admisibles.

2.2. El problema de Kantorovich

La formulación del problema del transporte óptimo de Kantorovich es una relajación del problema de Monge, como se verá en la sección 2.3. La masa de un punto $x \in \mathscr{X}$ no ha de transportarse a un solo punto de $\mathscr{Y}$ como en el problema de Monge, sino que se puede repartir en varios puntos de $\mathscr{Y}$. Esta relajación elimina todas las dificultades mencionadas anteriormente.

El transporte de la arena pasa a ser descrito por una medida de probabilidad π definida en el espacio producto $\mathscr{X} \times \mathscr{Y}$. Así, $\pi(A \times B)$ es la cantidad de masa transportada desde $A \in \beta_{\mathscr{X}}$ a $B \in \beta_{\mathscr{Y}}$. La condición de conservación de la masa está ahora determinada por las ecuaciones:

$$\pi(A \times \mathscr{Y}) = \mu(A), \text{ para todo } A \in \beta_{\mathscr{X}}. \tag{5}$$

$$\pi(\mathscr{X} \times B) = \nu(B), \text{ para todo } B \in \beta_{\mathscr{Y}}. \tag{6}$$

Las ecuaciones (5) y (6) son equivalentes a las siguientes:

$$\pi_{\mathscr{X}}\sharp\pi = \mu. \tag{7}$$

$$\pi_{\mathscr{Y}}\sharp\pi = \nu. \tag{8}$$

donde $\pi_{\mathscr{X}}$ y $\pi_{\mathscr{Y}}$ son las proyecciones del espacio $\mathscr{X} \times \mathscr{Y}$ en los espacios $\mathscr{X}$ e $\mathscr{Y}$.

Las distribuciones π que satisfacen las ecuaciones (5) y (6) (o equivalentemente (7) y (8)) se denominan planes de transporte y el conjunto de estos viene representado por $\Pi(\mu, \nu)$. μ y ν son las distribuciones marginales de π.

El coste total de transporte asociado al plan π es:

$$C(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x, y)d\pi(x, y). \tag{9}$$

El problema de Kantorovich consiste, por tanto, en encontrar una plan de transporte π_0 para el que se alcance el ínfimo del coste total de transporte (9):

$$C(\pi_0) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{Y}} c(x, y)d\pi(x, y).$$

La formulación de Kantorovich presenta numerosas ventajas frente a la de Monge, las cuales se presentan

a continuación.

El conjunto $\Pi(\mu, \nu)$ nunca es vacío. La medida producto $\mu \otimes \nu$, definida como $[\mu \otimes \nu](A \times B) = \mu(A)\nu(B)$, siempre pertenece a $\Pi(\mu, \nu)$ ya que, trivialmente, cumple las condiciones (5) y (6).

Por otro lado, la función coste $C(\pi)$ (9) y las restricciones (5) y (6) son lineales en π, por lo que el problema de Kantorovich es un problema de programación lineal infinito-dimensional. Para ser rigurosos, para hablar de linealidad se tiene que dotar al espacio de medidas de una estructura lineal. Para ello, se define el espacio $M(\mathscr{X})$ de las medidas finitas con signo[4] de Borel, que es un espacio vectorial porque $(\mu_1 + \alpha\mu_2)(A) = \mu_1(A) + \alpha\mu_2(A)$ con $\alpha \in \mathbb{R}$, $\mu_1, \mu_2 \in M(\mathscr{X})$ y $A \in \beta_{\mathscr{X}}$.

En tercer lugar, la formulación de Kantorovich es simétrica en el sentido de que para toda distribución $\pi(\mu, \nu) \in \Pi(\mu, \nu)$ que transporta μ en ν, existe una distribución $\tilde{\pi}(\nu, \mu) \in \Pi(\nu, \mu)$ que transporta ν en μ. Esta distribución está caracterizada por $\tilde{\pi}(B \times A) = \pi(A \times B)$.

También existe una simetría en la función coste. Si se define $\tilde{c}(y, x) = c(x, y)$, entonces

$$C(\pi) = \int_{\mathscr{X} \times \mathscr{Y}} c(x, y) d\pi(x, y) = \int_{\mathscr{Y} \times \mathscr{X}} \tilde{c}(y, x) d\tilde{\pi}(y, x) = \tilde{C}(\tilde{\pi}).$$

El teorema 6 prueba la existencia de un minimizador para el problema de Kantorovich. Para la prueba de este teorema se han seguido las referencias [4], [21] y [31]. Se introduce previamente una serie de conceptos.

Definición 4 (Función semicontinua inferiormente, [31]). *Sea $\mathscr{X}$ un espacio métrico. Una función $f : \mathscr{X} \to \mathbb{R} \cup \{\pm\infty\}$ es semicontinua inferiormente si para toda sucesión $\{x_n\}_n$ tal que $x_n \to x$, se cumple $f(x) \leq \lim\inf_n f(x_n)$.*

Definición 5 (Espacio C_b, [26]). *Dado $\mathscr{X} \neq \emptyset$, el espacio $C_b(\mathscr{X})$ es el espacio formado por las funciones reales definidas en $\mathscr{X}$, continuas y acotadas.*

Definición 6 (Convergencia débil, [26]). *Sea $\mathscr{X}$ un espacio métrico. Una sucesión de medidas de probabilidad $\{\mu_n\}$ con $\mu_n \in M(\mathscr{X})$ converge débilmente a $\mu \in M(\mathscr{X})$ si para toda $f \in C_b(\mathscr{X})$, se cumple que $\int f d\mu_n \to \int f d\mu$.*

Definición 7 (Colección de medidas de probabilidad "tight", [31]). *Una colección de medidas de probabilidad $\mathscr{K}$ es "tight" si para todo $\epsilon > 0$, existe K compacto tal que $\inf_{\mu \in \mathscr{K}} \mu(K) > 1 - \varepsilon$.*

Definición 8 (Compacidad relativa, [23]). *Un conjunto de un espacio topológico es relativamente compacto si su clausura es compacta.*

También se enuncian los teoremas de Weirstrass [31], de Prokhorov [10] y un teorema sobre funciones semicontinuas inferiormente [30], que se utilizarán en la prueba del teorema 6. No se presentan sus demostraciones, pero se pueden encontrar en las referencias indicadas.

Teorema 3 (Weirstrass, [31]). *Sea $\mathscr{X}$ un espacio compacto y $f : \mathscr{X} \to \mathbb{R} \cup \{+\infty\}$ una función semicontinua inferiormente. Entonces existe $\bar{x} \in \mathscr{X}$ tal que $f(\bar{x}) = \min\{f(x) : x \in \mathscr{X}\}$.*

---

[4] **Definición 3** (Medida con signo, [2]). *Sea $(\Omega, \mathcal{A})$ un espacio medible. Una función $\mu : \mathcal{A} \to \mathbb{R} \cup \{\pm\infty\}$ es una medida con signo si*

1. *μ alcanza, como máximo, uno de los valores $+\infty$ o $-\infty$.*
2. *$\mu(\emptyset) = 0$.*
3. *Si $(E_i)_{i \in \mathbb{N}}$ es una sucesión de conjuntos medibles tales que $E_i \cap E_j = \emptyset$ para todo $i \neq j$, entonces $\mu\left(\cup_{i \in \mathbb{N}} E_i\right) = \sum_{i \in \mathbb{N}} \mu(E_i)$.*

Teorema 4 (Prokhorov, [10]). *Sea $\mathscr{X}$ un espacio completo y separable. Entonces, una familia $\mathscr{K} \subset P(\mathscr{X})$, con $P(\mathscr{X})$ el conjunto de probabilidades en el espacio $\mathscr{X}$, es relativamente compacta con respecto de la convergencia débil si y solo si es "tight".*

Teorema 5 (Caracterización de funciones semicontinuas inferiormente, [30]). *Sea $(\mathscr{X}, d)$ un espacio métrico y $f : \mathscr{X} \to [0, \infty]$ una función semicontinua inferiormente tal que $\text{ínf}_{p\in\mathscr{X}} f(p) < \infty$. Para cada $n \in \mathbb{N}$, se define*

$$g_n(x) = \inf_{p\in\mathscr{X}} \{f(p) + nd(x, p)\}.$$

Entonces, $\{g_n\}$ es una sucesión no decreciente de funciones no negativas y continuas en $\mathscr{X}$ tales que $f = \lim_n g_n$.

Teorema 6 (Existencia de solución del problema de Kantorovich). *El problema de Kantorovich admite solución cuando c es semicontinua inferiormente y no negativa y $\mathscr{X}$ e $\mathscr{Y}$ son espacios métricos completos y separables.*

Demostración. Hay que probar que $\Pi(\mu, \nu)$ es compacto en $P(\mathscr{X} \times \mathscr{Y})$ y que el funcional $\pi \to \int_{\mathscr{X}\times\mathscr{Y}} cd\pi$ es semicontinuo inferiormente. Así, aplicando el teorema de Weirstrass (teorema 3), quedará probado que el problema de Kantorovich admite un minimizador.

Por el teorema 4, $\{\mu\}$ y $\{\nu\}$ son "tight" porque $\mathscr{X}$ e $\mathscr{Y}$ son completos y separables. Veamos ahora que $\Pi(\mu, \nu)$ es "tight". Sean $\varepsilon > 0$ y $K_1 \subset \mathscr{X}$ y $K_2 \subset \mathscr{Y}$ compactos tales que $\mu(\mathscr{X} \setminus K_1) < \frac{\varepsilon}{2}$ y $\nu(\mathscr{Y} \setminus K_2) < \frac{\varepsilon}{2}$. Se tiene que

$$(\mathscr{X} \times \mathscr{Y}) \setminus (K_1 \times K_2) \subset [(\mathscr{X} \setminus K_1) \times \mathscr{Y}] \cup [\mathscr{X} \times (\mathscr{Y} \setminus K_2)].$$

Por tanto,

$$\pi [(\mathscr{X} \times \mathscr{Y}) \setminus (K_1 \times K_2)] \leq \pi [(\mathscr{X} \setminus K_1) \times \mathscr{Y}] + \pi [\mathscr{X} \times (\mathscr{Y} \setminus K_2)] = \mu(\mathscr{X} \setminus K_1) + \nu(\mathscr{Y} \setminus K_2) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

y queda probado que $\Pi(\mu, \nu)$ es "tight". Por el teorema de Prokhorov (teorema 4), $\Pi(\mu, \nu)$ es relativamente compacto, luego su adherencia, que denotamos $\overline{\Pi(\mu, \nu)}$, es compacta (definición 8).

Veamos ahora que $\Pi(\mu, \nu)$ es compacto. Para ello, veamos que $\Pi(\mu, \nu)$ es cerrado y, por tanto, $\overline{\Pi(\mu, \nu)} = \Pi(\mu, \nu)$. Usaremos la caracterización de conjuntos cerrados por sucesiones y la convergencia débil (definición 6). Sea $(\pi_n)_n \subset \Pi(\mu, \nu)$ una sucesión tal que $\pi_n \to \pi$. Veamos que $\pi \in \Pi(\mu, \nu)$. Sea $\varphi \in C_b(\mathscr{X})$. Con un abuso de notación, podemos considerar φ definida en $\mathscr{X} \times \mathscr{Y}$ tomando $\varphi(x, y) = \varphi(x)$.

$$\begin{aligned} \int \varphi(x) d\pi_{\mathscr{X}} \sharp \pi(x) &= \int \varphi(x, y) d\pi(x, y) = \lim_{n\to\infty} \int \varphi(x, y) d\pi_n(x, y) = \lim_{n\to\infty} \int \varphi(x) d\pi_{\mathscr{X}} \sharp \pi_n(x) \\ &= \int \varphi(x) d\mu(x), \end{aligned}$$

lo que implica, por el teorema 1, que $\pi_{\mathscr{X}} \sharp \pi = \mu$. La primera igualdad se da porque $\pi_{\mathscr{X}}$ es la proyección de $\mathscr{X} \times \mathscr{Y}$ en $\mathscr{X}$, la segunda porque $\pi_n \to \pi$ débilmente, la tercera porque $(\pi_n)_n \subset \Pi(\mu, \nu)$ y la última, de nuevo, por la definición de $\pi_{\mathscr{X}}$. Idénticamente, $\pi_{\mathscr{Y}} \times \pi = \nu$. Luego $\pi \in \Pi(\mu, \nu)$, pues se cumplen las condiciones (7) y (8). Por tanto, $\Pi(\mu, \nu)$ es cerrado, $\Pi(\mu, \nu) = \overline{\Pi(\mu, \nu)}$ y $\Pi(\mu, \nu)$ es compacto.

Veamos ahora que el funcional $\pi \to \int_{\mathscr{X}\times\mathscr{Y}} cd\pi$ es semicontinuo inferiormente.

Sea $(\pi_n)_n \subset \Pi(\mu,\nu)$ tal que $\pi_n \to \pi$. Como $\mathscr{X}\times\mathscr{Y}$ es un espacio métrico y $c:\mathscr{X}\times\mathscr{Y}\to[0,1]$ es semicontinua inferiormente y está acotada inferiormente (por 0), por el teorema 5, existe una sucesión no decreciente $\{c_k\}$ de funciones no negativas y continuas tales que $c = \lim_k c_k$. Como $\{c_k\}$ es no decreciente, $c = \lim_k c_k = \sup_k c_k$. Entonces,

$$\int_{\mathscr{X}\times\mathscr{Y}} c_k d\pi_n \leq \int_{\mathscr{X}\times\mathscr{Y}} c d\pi_n. \tag{10}$$

Tomando límites inferiores en (10),

$$\liminf_n \int_{\mathscr{X}\times\mathscr{Y}} c_k d\pi_n \leq \liminf_n \int_{\mathscr{X}\times\mathscr{Y}} c d\pi_n. \tag{11}$$

Como cada $c_k \in C_b(\mathscr{X}\times\mathscr{Y})$,

$$\liminf_n \int_{\mathscr{X}\times\mathscr{Y}} c_k d\pi_n = \int_{\mathscr{X}\times\mathscr{Y}} c_k d\pi.$$

Aplicando el teorema de la Convergencia Monótona,

$$\lim_k \int_{\mathscr{X}\times\mathscr{Y}} c_k d\pi = \int_{\mathscr{X}\times\mathscr{Y}} c d\pi.$$

Por tanto, sustituyendo en (11),

$$\int_{\mathscr{X}\times\mathscr{Y}} c d\pi \leq \liminf_n \int_{\mathscr{X}\times\mathscr{Y}} c d\pi_n,$$

que por la definición 4, se deduce que el funcional $\pi \to \int_{\mathscr{X}\times\mathscr{Y}} c d\pi$ es semicontinuo inferiormente.

□

2.3. Interpretación probabilística

Los problemas de Monge y Kantorovich se pueden interpretar de una manera probabilística. Para ello, hay que introducir los conceptos de elemento aleatorio y distribución de probabilidad de este.

Definición 9 (Elemento aleatorio, [26])**.** *Un elemento aleatorio en un espacio topológico $\mathscr{X}$ es una función $X:(\Omega,\mathscr{A},\mathbb{P})\to(\mathscr{X},\mathscr{B}_{\mathscr{X}})$ medible, con $(\Omega,\mathscr{A},\mathbb{P})$ un espacio de probabilidad y $\mathscr{B}_{\mathscr{X}}$ la σ-álgebra de Borel sobre el espacio $\mathscr{X}$.*

Definición 10 (Distribución de probabilidad de un elemento aleatorio, [26])**.** *Dado un elemento aleatorio X en un espacio topológico $\mathscr{X}$, se llama ley o distribución de probabilidad de X a la medida de probabilidad $\mu_X = X\sharp\mathbb{P}$ definida en el espacio $\mathscr{X}$, es decir, la medida de Borel que cumple $\mu_X(A) = \mathbb{P}[X^{-1}(A)] = \mathbb{P}(X\in A)$ para todo $A\in\beta_{\mathscr{X}}$.*

Dados $\mathscr{X}$, $\mathscr{Y}$, μ, ν y $c(x,y)$ como en la sección 2.1, el problema de Monge se puede modelizar con el elemento aleatorio $X:(\Omega,\mathscr{A},\mathbb{P})\to(\mathscr{X},\mathscr{B}_{\mathscr{X}})$ con $(\Omega,\mathscr{A},\mathbb{P})$ un espacio de probabilidad y cuya distribución de probabilidad es $\mu = X\sharp\mathbb{P}$. Se trata de encontrar una función medible $T:\mathscr{X}\to\mathscr{Y}$ tal que la distribución de $T(X)$ sea ν y que minimice el coste total:

$$C(T) = \int_{\mathscr{X}} c(x,T(x))d\mu(x) = \int_\Omega c[X(\omega),T(X(\omega))]d\mathbb{P}(\omega) = E[c(X,T(X))],$$

es decir, la función T que minimice la esperanza de la función coste c.

Por su parte, el problema de Kantorovich consiste en encontrar una distribución de probabilidad conjunta del elemento aleatorio bidimensional (X, Y), con $X : (\Omega, \mathscr{A}, \mathbb{P}) \to (\mathscr{X}, \mathscr{B}_{\mathscr{X}})$ e $Y : (\Omega, \mathscr{A}, \mathbb{P}) \to (\mathscr{Y}, \mathscr{B}_{\mathscr{Y}})$, cuyas marginales sean μ y ν y tal que $\pi = (X, Y)\sharp\mathbb{P}$ minimice el coste total, que es igual a la esperanza de la función coste:

$$C(\pi) = \int_{\mathscr{X}\times\mathscr{Y}} c(x, y)d\pi(x, y) = \int_{\Omega} c[X(\omega), Y(\omega)]d\mathbb{P}(\omega) = E[c(X, Y)].$$

π se denomina distribución conjunta o *coupling*.

El problema de Kantorovich es una relajación del problema de Monge porque a cada función de transporte T se le puede asociar una distribución π_T de manera que $C(T) = C(\pi_T)$. π_T es la distribución del elemento aleatorio $(X, Y) = (X, T(X))$ (la distribución de Y dado $X = x$ es $\delta_{T(x)}$) y se tiene:

$$C(\pi_T) = E[C(X, Y)] = E[E(c(X, Y)/X)] = E[c(X, T(X))] = C(T).$$

Por tanto,

$$C(\pi_T) = \int_{\mathscr{X}\times\mathscr{Y}} c(x, y)d\pi_T(x, y) = C(T) = \int_{\mathscr{X}} c(x, T(x))d\mu(x). \tag{12}$$

De (12) y el teorema 1, se deduce que π_T es la medida de probabilidad en el espacio $\mathscr{X} \times \mathscr{Y}$ dada por $\pi_T := (Id \times T)\sharp\mu$.

2.4. Caso discreto uniforme

En el caso de que las medidas μ y ν sean uniformes en n puntos, el problema de Kantorovich se reduce a una problema combinatorio finito. Se presenta este caso para introducir posteriormente el problema dual de una manera más intuitiva. Supongamos que

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \quad \text{y} \quad \nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}.$$

Los costes de transporte se pueden representar con una matriz $n \times n$, $C = (c_{ij})_{ij}$, con $c_{ij} = c(x_i, y_j)$. Cualquier función de transporte $T : \mathscr{X} \to \mathscr{Y}$ se puede asociar a una permutación en $S_n = \{1, ..., n\}$, de manera que, dada $\sigma \in S_n$, $T(x_i) = y_{\sigma(i)}$. Los planes de transporte se pueden representar por medio de matrices M, $n \times n$, $M = (M_{ij})_{ij}$ con $M_{ij} = \pi(\{(x_i, y_j)\})$, donde para que π sea un plan de transferencia, nM tiene que ser una matriz bistocástica[5].

En el caso discreto, el problema de Monge pasa a ser un problema de optimización combinatoria que se formula de la siguiente manera:

$$\inf_{\sigma\in S_n} C(\sigma) = \frac{1}{n} \inf_{\sigma\in S_n} \sum_{i=1}^{n} c_{i,\sigma(i)}.$$

[5]El conjunto de matrices bistocásticas, B_n, se define como el conjunto de matrices N, $n \times n$, que cumplen $\sum_{i=1}^{n} N_{ij} = 1$, $\sum_{j=1}^{n} N_{ij} = 1$ y $N_{ij} \geq 0$ [26].

El problema de Kantorovich es el problema lineal siguiente:

$$\inf_{M\in B_n/n} \sum_{i,j=1}^{n} c_{ij}M_{ij}.$$

2.5. Dualidad del problema de Kantorovich

Se ha visto en la sección 2.2 que el problema de Kantorovich es un problema de programación lineal. Por tanto, admite una formulación dual.

2.5.1. Caso discreto uniforme

La notación que no se introduzca en esta sección es la que se especificó en la sección 2.4.

Para mostrar la formulación dual del problema de Kantorovich, se utiliza la siguiente notación. La matriz $M = (\pi(\{(x_i, y_j)\}))_{ij}$ se representa por un vector en $\mathbb{R}^{n^2}$: $\vec{M} = (M_{11}, ..., M_{ij}, ..., M_{nn})^t$. Las $2n$ restricciones (se omite en aquí la restricción de que $M \geq 0$) para que $nM = n(\pi(\{(x_i, y_j\}))_{ij} \in B_n$ se representan con una matriz $2n \times n^2$, $A = \begin{pmatrix} \vec{1}_n & & & \\ & \vec{1}_n & & \\ & & \ddots & \\ & & & \vec{1}_n \\ I_n & I_n & \dots & I_n \end{pmatrix}$,

cumpliéndose:

$$A\vec{M} = \frac{1}{n}\vec{b} \in \mathbb{R}^{2n},$$

con $\vec{b} := 1^t_{2n}$, I_n la matriz identidad en $\mathbb{R}^n$ y $\vec{1}^t_m = (1, 1, ..., 1)^t \in \mathbb{R}^m$.

Así, el problema primal de Kantorovich se formula de la siguiente manera:

$$(P)\begin{cases} \text{mín}_{\vec{M}\in\mathbb{R}^{n^2}}\, C^t\vec{M} \\ \text{sujeto a } A\vec{M} = \frac{1}{n}\vec{b} \\ \qquad\qquad \vec{M} \geq 0 \end{cases} \tag{13}$$

Para obtener el problema dual de (13), por cada fila de la matriz A se introduce una variable. Estas variables se denominarán p_1,...,p_n,q_1,...,q_n. Llamamos $\vec{p} = (p_1, ..., p_n)^t$ y $\vec{q} = (q_1, ..., q_n)^t$. Se intercambian $(\vec{p}, \vec{q})^t$ y $\vec{M}$ y $\vec{b}$ y C y se busca el máximo en vez del mínimo. Así, la formulación dual es:

$$(D)\begin{cases} \text{máx}_{\vec{p}^t,\vec{q}^t\in\mathbb{R}^n}\, b^t \begin{pmatrix} \vec{p} \\ \vec{q} \end{pmatrix} \\ \text{sujeto a } A^t \begin{pmatrix} \vec{p} \\ \vec{q} \end{pmatrix} \leq C \end{cases} \tag{14}$$

2.5.2. Caso general

Los vectores $\vec{p}$ y $\vec{q}$ se pueden interpretar como restricciones de funciones $\varphi : \mathscr{X} \to \mathbb{R}$ y $\psi : \mathscr{Y} \to \mathbb{R}$, de manera que $p_i = \varphi(x_i)$ con $i = 1, ..., n$ y $q_j = \psi(y_j)$ con $j = 1, ..., n$. Las coordenadas del vector

$\vec{b} = (1, ..., 1)^t \in \mathbb{R}^{2n}$ se pueden ver como $b_i = \mu(\{x_i\})$ y $b_{j+n} = \nu(\{y_j\})$ con $i, j = 1, ..., n$.

De esta manera, el dual del caso discreto (14) se traduce en:

$$(D) \begin{cases} \sup_{(\varphi,\psi)\in L_1(\mu)\times L_1(\nu)} \left[\int_{\mathscr{X}} \varphi(x)d\mu(x) + \int_{\mathscr{Y}} \psi(y)d\nu(y)\right] \\ \text{sujeto a } (\varphi, \psi) \in \Phi_c \end{cases}$$

con $\Phi_c = \{(\varphi, \psi) \in L_1(\mu) \times L_1(\nu) : \varphi(x) + \psi(y) \leq c(x, y) \text{ para } (\mu \times \nu) - \text{casi todo } (x, y) \in \mathscr{X} \times \mathscr{Y}\}$.

El siguiente lema permite ver que las restricciones del problema de Kantorovich ((5) y (6) o equivalentemente (7) y (8)) son equivalentes a restricciones funcionales.

Lema 1 (Restricciones funcionales, [26])**.** *Sean μ y ν medidas de probabilidad. $\pi \in \Pi(\mu, \nu)$ si y solo si para todas las funciones $\varphi \in C_b(\mathscr{X})$, $\psi \in C_b(\mathscr{Y})$,*

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

Demostración. $\pi \in \Pi(\mu, \nu)$ si y solo $\pi_{\mathscr{X}}\sharp\pi = \mu$ y $\pi_{\mathscr{Y}}\sharp\pi = \nu$, con $\pi_{\mathscr{X}}$ y $\pi_{\mathscr{Y}}$ las proyecciones de $\mathscr{X} \times \mathscr{Y}$ en $\mathscr{X}$ e $\mathscr{Y}$, respectivamente. Teniendo en cuenta el teorema 1, se tiene,

$$\pi_{\mathscr{X}}\sharp\pi = \mu \text{ si y solo si } \int_{\mathscr{X}} \varphi(x)d\mu(x) = \int_{\mathscr{X}\times\mathscr{Y}} \varphi(x)d\pi(x, y), \text{ para toda } \varphi \in C_b(\mathscr{X}). \tag{15}$$

$$\pi_{\mathscr{Y}}\sharp\pi = \nu \text{ si y solo si } \int_{\mathscr{Y}} \psi(y)d\nu(y) = \int_{\mathscr{X}\times\mathscr{Y}} \psi(y)d\pi(x, y), \text{ para toda } \psi \in C_b(\mathscr{Y}). \tag{16}$$

Sumando (15) y (16), se tiene

$$\begin{aligned} \int_{\mathscr{X}} \varphi(x)d\mu(x) + \int_{\mathscr{Y}} \psi(y)d\nu(y) &= \int_{\mathscr{X}\times\mathscr{Y}} \varphi(x)d\pi(x, y) + \int_{\mathscr{X}\times\mathscr{Y}} \psi(y)d\pi(x, y) \\ &= \int_{\mathscr{X}\times\mathscr{Y}} [\varphi(x) + \psi(y)]d\pi(x, y) \text{ para toda } \varphi \in C_b(\mathscr{X}), \text{ para toda } \psi \in C_b(\mathscr{Y}). \end{aligned}$$

□

Teorema 7 (Igualdad primal-dual, [35])**.** *Sean μ y ν medidas de probabilidad en espacios métricos completos y separables $\mathscr{X}$ e $\mathscr{Y}$ y sea $c : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}_+$ una función medible. Entonces,*

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{Y}} cd\pi = \sup_{(\varphi,\psi)\in\Phi_c} \left[\int_{\mathscr{X}} \varphi d\mu + \int_{\mathscr{Y}} \psi d\nu\right]. \tag{17}$$

Demostración. La demostración de este teorema se puede encontrar en [35]. □

2.6. Existencia y unicidad de solución para el problema de Kantorovich en $\mathbb{R}^d$ y para coste cuadrático

El objetivo final de esta sección es demostrar el teorema que caracteriza la solución del problema dual de Kantorovich con $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$ y para un coste cuadrático. Se presenta este caso porque es en el que los resultados son más simples.

Se introduce una serie de conceptos que aparecerán a lo largo de la sección.

Definición 11 (Función convexa, [26]). *Una función $f : \mathscr{X} \to \mathbb{R} \cup \{+\infty\}$ es convexa si su dominio, es un conjunto convexo y $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ para todo $x, y \in \mathscr{X}$, para todo $t \in [0,1]$.*

Definición 12 (Función propia, [21]). *Una función $f : \mathscr{X} \to \mathbb{R} \cup \{+\infty\}$ es propia si existe al menos un $x \in \mathscr{X}$ tal que $f(x) < +\infty$.*

Definición 13 (Subgradiente y subdiferencial, [11]). *Un subgradiente de una función convexa $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ en $x \in \mathbb{R}^d$ es un vector $v \in \mathbb{R}^d$ tal que*

$$f(y) \geq f(x) + v \cdot (y - x) \text{ para todo } y \in \mathbb{R}^d,$$

donde $\cdot$ denota el producto escalar usual.

El subdiferencial en un punto $x \in \mathbb{R}^d$, $\partial f(x)$, es el conjunto de todos los subgradientes de f en x.

Si f es convexa y $x \in \text{Int}(\text{Dom}(f))$, con $\text{Int}(\text{Dom}(f))$ el interior del dominio de f, el subdiferencial $\partial f(x)$ es no vacío [35].

Definición 14 (Función convexa conjugada, [35]). *La función convexa conjugada o transformada de Legendre de una función propia $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ en $x \in \mathbb{R}^d$ se define como:*

$$f^*(y) = \sup_{x \in \mathbb{R}^d} (x \cdot y - f(x)).$$

f^* está bien definida: por ser f propia, existe $x_0 \in \mathbb{R}^d$ tal que $f(x_0) < +\infty$. Por tanto,

$$f^*(y) = \sup_{x \in \mathbb{R}^d} (x \cdot y - f(x)) \geq x_0 \cdot y - f(x_0) > -\infty.$$

Proposición 1 (Caracterización del subdiferencial, [35]). *Sea f una función propia, convexa y semicontinua inferiormente en $\mathbb{R}^d$. Entonces, si $x, y \in \mathbb{R}^d$, se verifica*

$$x \cdot y = f(x) + f^*(y) \text{ si y solo si } y \in \partial f(x).$$

Demostración. Sean $x, y \in \mathbb{R}^d$. De la definición de función convexa conjugada, se tiene que $x \cdot y \leq f(x) + f^*(y)$. Por tanto, para que $x \cdot y = f(x) + f^*(y)$, se tiene que cumplir:

$$x \cdot y \geq f(x) + f^*(y),$$

que es equivalente a

$$x \cdot y \geq f(x) + z \cdot y - f(z), \text{ para todo } z \in \mathbb{R}^d,$$

que a su vez es equivalente a

$$f(z) \geq f(x) + y \cdot (z - x), \text{ para todo } z \in \mathbb{R}^d,$$

lo cual significa que $y \in \partial f(x)$.

□

Para simplificar las demostraciones, se divide el coste entre 2, de manera que $c(x, y) = \frac{\|x-y\|^2}{2}$. Se tiene, por tanto, que el coste total de transporte es:

$$C(\pi) := \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{\| x - y \|^2}{2} d\pi(x, y).$$

Se suponen μ y ν medidas de probabilidad de Borel con momentos de segundo orden finitos, por tanto,

$$M_2 := \int_{\mathbb{R}^d} \| x \|^2 \, d\mu(x) + \int_{\mathbb{R}^d} \| y \|^2 \, d\nu(x) < +\infty. \tag{18}$$

De aquí,

$$C(\pi) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{\| x - y \|^2}{2} d\pi(x, y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \| x \|^2 \, d\pi(x, y) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \| y \|^2 \, d\pi(x, y) = 2M_2 < +\infty.$$

La caracterización de la solución del problema de Kantorovich se va a hacer con el problema dual, por lo que se va a estudiar este en más detalle a continuación.

Se nombra la función objetivo del problema dual como:

$$J(\varphi, \psi) := \int_{\mathscr{X}} \varphi d\mu + \int_{\mathscr{Y}} \psi d\nu. \tag{19}$$

La restricción del problema dual, es decir, la condición para que $(\varphi, \psi) \in \Phi_c$, es:

$$\varphi(x) + \psi(y) \leq c(x, y) = \frac{\| x - y \|^2}{2}, \tag{20}$$

para $(\mu \times \nu)$-casi todo $(x, y) \in \mathscr{X} \times \mathscr{Y}$.

Desarrollando la parte derecha de (20) aplicando la desigualdad triangular, se tiene:

$$\varphi(x) + \psi(y) \leq c(x, y) = \frac{\| x - y \|^2}{2} = \frac{\| x \|^2}{2} + \frac{\| y \|^2}{2} - x \cdot y.$$

Por tanto,

$$x \cdot y \leq \left[\frac{\| x \|^2}{2} - \varphi(x) \right] + \left[\frac{\| y \|^2}{2} - \psi(y) \right]. \tag{21}$$

Se definen:

$$\tilde{\varphi}(x) := \frac{\| x \|^2}{2} - \varphi(x) \quad \text{y} \quad \tilde{\psi}(y) := \frac{\| y \|^2}{2} - \psi(y), \tag{22}$$

de manera que (21) queda:

$$x \cdot y \leq \tilde{\varphi}(x) + \tilde{\psi}(y).$$

Teniendo en cuenta (18),

$$\begin{aligned} \inf_{\pi \in \Pi(\mu,\nu)} C(\pi) &= \inf_{\mu \in P(\mathscr{X})} \int_{\mathbb{R}^d} \frac{\| x \|^2}{2} + \inf_{\nu \in P(\mathscr{Y})} \int_{\mathbb{R}^d} \frac{\| y \|^2}{2} + \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} (-x \cdot y) d\pi(x,y) \\ &= M_2 - \sup_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} x \cdot y d\pi(x,y). \end{aligned} \tag{23}$$

Pasando (23) a la formulación dual, teniendo en cuenta el teorema 7, se obtiene:

$$\sup_{(\varphi,\psi) \in \Phi_c} J(\varphi,\psi) = M_2 - \inf_{(\tilde{\varphi},\tilde{\psi}) \in \tilde{\Phi}_C} J(\tilde{\varphi},\tilde{\psi}), \tag{24}$$

donde $\tilde{\Phi}_C = \{(\tilde{\varphi},\tilde{\psi}) \in L^1(\mu) \times L^1(\nu) : x \cdot y \leq \tilde{\varphi}(x) + \tilde{\psi}(y)$ para $(\mu \times \nu)-$casi todo $(x,y) \in \mathscr{X} \times \mathscr{Y}\}$. Por tanto, de (24), se obtiene que es equivalente buscar un par óptimo $(\varphi,\psi) \in \Phi_c$ y buscar un par óptimo $(\tilde{\varphi},\tilde{\psi}) \in \tilde{\Phi}_C$. En particular, la ecuación (17) (teorema 7) queda:

$$\sup_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X} \times \mathscr{Y}} (x \cdot y) d\pi(x,y) = \inf_{(\tilde{\varphi},\tilde{\psi}) \in \tilde{\Phi}_C} J(\tilde{\varphi},\tilde{\psi}). \tag{25}$$

La función objetivo del problema dual (19) es creciente con φ y ψ, por lo que para alcanzar el supremo de J habrá que tomar funciones φ y ψ con valores lo más grandes posibles.

Supongamos que conocemos $\varphi \in L^1(\mu)$ siendo esta una buena candidata para alcanzar el supremo de J. Por tanto, la función ψ con valores más grandes posible será, cumpliendo $(\varphi,\psi) \in \Phi_c$,

$$\psi(y) = \inf_{x \in \mathbb{R}^d} \left[\frac{\| x - y \|^2}{2} - \varphi(x) \right] = \frac{\| y \|^2}{2} + \inf_{x \in \mathbb{R}^d} \left[\frac{\| x \|^2}{2} - x \cdot y - \varphi(x) \right] = \frac{\| y \|^2}{2} + \inf_{x \in \mathbb{R}^d} \left[\tilde{\varphi}(x) - x \cdot y \right].$$

Sustituyendo en la ecuación de $\tilde{\psi}(y)$ de (22),

$$\tilde{\psi}(y) = - \inf_{x \in \mathbb{R}^d} \left[\tilde{\varphi}(x) - x \cdot y \right] = \sup_{x \in \mathbb{R}^d} \left[\tilde{\varphi}(x) - x \cdot y \right] := \tilde{\varphi}^*(y),$$

donde $\tilde{\varphi}^*$ es la función convexa conjugada de $\tilde{\psi}$. Simétricamente, se puede tomar

$$\tilde{\varphi}(x) = \tilde{\psi}^*(x).$$

Por tanto, parece razonable que el par $(\tilde{\varphi},\tilde{\varphi}^*) \in \tilde{\Phi}_C$ sea un par óptimo para el problema de Kantorovich. El teorema 8 prueba esto además de la existencia de tal par. No se presenta su demostración pero se puede encontrar en [35].

Teorema 8 (Existencia de un par óptimo de funciones convexas conjugadas solución del problema dual de Kantorovich, [35])**.** *Sean μ y ν medidas de probabilidad en $\mathbb{R}^d$ con momentos de segundo orden finitos. Sea $\tilde{\Phi}_C = \{(\tilde{\varphi},\tilde{\psi}) \in L^1(\mu) \times L^1(\nu) : x \cdot y \leq \tilde{\varphi}(x) + \tilde{\psi}(y)$ para $(\mu \times \nu)-$casi todo $(x,y) \in \mathscr{X} \times \mathscr{Y}\}$. Entonces, existe un par $(\tilde{\varphi}_0,\tilde{\varphi}_0^*)$ de funciones propias convexas conjugadas y semicontinuas inferiormente en $\mathbb{R}^d$*

tales que

$$\inf_{(\tilde{\varphi},\tilde{\varphi}^*)\in\tilde{\Phi}_C} J = J(\tilde{\varphi}_0, \tilde{\varphi}_0^*).$$

Cabe mencionar que en el teorema 8 se puede sustituir $(\tilde{\varphi}, \tilde{\psi}) \in \tilde{\Phi}_C$ por $(\varphi, \psi) \in \Phi$ y entonces se tendría $\sup_{(\varphi,\psi)\in\Phi} J(\varphi, \psi) = J(\varphi_0, \varphi_0^*)$. Se recuerda que la solución del problema dual de Kantorovich es $\left(\frac{\|x\|^2}{2} - \tilde{\varphi}, \frac{\|y\|^2}{2} - \tilde{\varphi}^*\right)$ pero se trabaja con $(\tilde{\varphi}, \tilde{\varphi}^*)$ para simplificar los cálculos. Siempre que se ponga una tilde $\tilde{\cdot}$ sobre una función, se estará refiriendo a las definiciones de (22).

Una vez probada la existencia de un par óptimo para el problema de Kantorovich en el teorema 8, se presentan a continuación dos teoremas que caracterizan la solución del problema de Kantorovich [35].

Teorema 9 (Criterio de optimalidad de Knott y Smith, [35])**.** *Sean μ y ν medidas de probabilidad en $\mathbb{R}^d$ con momentos de segundo orden finitos. Sea $c(x, y) = \frac{\|x-y\|^2}{2}$ la función coste del problema de Kantorovich. Entonces, $\pi \in \Pi(\mu, \nu)$ es óptimo si y solo si existe una función φ_0 convexa y semicontinua inferiormente tal que*

$$\text{Soporte}(\pi) \subset \text{Grafo}(\partial\varphi_0).$$

donde Soporte(π) *es el conjunto* $\{x \in \mathbb{R}^2 : \pi[B(x, r)] > 0$ *para todo* $r > 0\}$.

Equivalentemente,

$$(\mu \times \nu)\{(x, y) : y \in \partial\varphi_0(x)\} = 1.$$

Además, $(\tilde{\varphi}_0, \tilde{\varphi}_0^*)$ es un minimizador de

$$\int_{\mathbb{R}^d} \tilde{\varphi}(x)d\mu(x) + \int_{\mathbb{R}^d} \tilde{\psi}(y)d\nu(y),$$

en el conjunto de las funciones $\tilde{\varphi}$, $\tilde{\psi}$ tales que $(\mu \times \nu)\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \tilde{\varphi}(x) + \tilde{\psi}(y) \geq x \cdot y\} = 1$.

Demostración. Se sabe por el teorema 6 que existe un π óptimo para el problema primal de Kantorovich y por el teorema 8 que existe un par óptimo para el problema dual.

Se prueba primero la condición necesaria. Sea $\pi \in \Pi(\mu, \nu)$ óptimo para el problema primal de Kantorovich y (φ, φ^*) para el problema dual. Sean $\tilde{\varphi}$ y $\tilde{\varphi}^*$ con $\tilde{\varphi}$ definida como en (22). Como π y $(\tilde{\varphi}, \tilde{\psi})$ son óptimos, alcanzan el supremo y el ínfimo, respectivamente, de (25), luego:

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} (x \cdot y)d\pi(x, y) = \int_{\mathbb{R}^d} \tilde{\varphi}(x)d\mu(x) + \int_{\mathbb{R}^d} \tilde{\varphi}^*(y)d\nu(y) = \int_{\mathbb{R}^d\times\mathbb{R}^d} [\tilde{\varphi}(x) + \tilde{\varphi}^*(y)]d\pi(x, y). \tag{26}$$

La última igualdad resulta de aplicar el lema 1. Despejando en (26),

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} [\tilde{\varphi}(x) + \tilde{\varphi}^*(y) - x \cdot y]d\pi(x, y) = 0. \tag{27}$$

Puesto que $(\tilde{\varphi}, \tilde{\varphi}^*) \in \tilde{\Phi}_C$ (teorema 8), $x \cdot y \leq \tilde{\varphi}(x) + \tilde{\varphi}^*(y)$ $(\mu \times \nu)$-casi seguro y el integrando de (27) es no negativo. Por tanto, (27) se cumple si y solo si $\tilde{\varphi}(x) + \tilde{\varphi}^*(y) = x \cdot y$. Aplicando la proposición 1, se deduce que $y \in \partial\tilde{\varphi}(x)$.

Ahora se prueba la condición suficiente. Sea $\pi \in \Pi(\mu, \nu)$ tal que $y \in \partial\tilde{\varphi}(x)$ para $(\mu \times \nu)$-casi todo (x, y) con $\tilde{\varphi}$ una función propia, convexa y semicontinua inferiormente. Por la proposición 1, $\tilde{\varphi}(x)+\tilde{\varphi}^*(y) = x \cdot y$.

Esto implica que:

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}[\tilde{\varphi}(x)+\tilde{\varphi}^*(y)-x\cdot y]d\pi(x,y)=0,$$

y de aquí se deduce:

$$\int_{\mathbb{R}^d\times\mathbb{R}^d}(x\cdot y)d\pi(x,y)=\int_{\mathbb{R}^d}\tilde{\varphi}(x)d\mu(x)+\int_{\mathbb{R}^d}\tilde{\varphi}^*(y)d\nu(y).$$

Y por tanto, π es óptimo para el problema de Kantorovich.

Se sigue del teorema 8 que $(\tilde{\varphi},\tilde{\varphi}^*)$ es un minimizador de

$$\int_{\mathbb{R}^d}\tilde{\varphi}(x)d\mu(x)+\int_{\mathbb{R}^d}\tilde{\psi}(y)d\nu(y),$$

donde $\tilde{\varphi}(x)+\tilde{\psi}(y)\geq x\cdot y$ para casi todo $x,y\in\mathbb{R}^d$. □

Teorema 10 (Caracterización de la solución del problema de Kantorovich en $\mathbb{R}^d$ para un coste cuadrático, [35])**.** *Sean μ y ν medidas de probabilidad en $\mathbb{R}^d$ con momentos de segundo orden finitos y μ absolutamente continua con respecto a la medida de Lebesgue y sea $c(x,y)=\frac{\|x-y\|^2}{2}$ la función coste del problema de Kantorovich. Entonces, la solución del problema del problema de Kantorovich es única μ-casi seguro y viene dada por*

$$\pi=(Id\times\nabla\tilde{\varphi})\sharp\mu,$$

donde $\tilde{\varphi}$ es una función convexa tal que $\nabla\tilde{\varphi}\sharp\mu=\nu$.

Demostración. Se divide la prueba en dos partes. En la primera parte se caracteriza el plan de transporte óptimo del problema de Kantorovich y en la segunda parte se prueba su unicidad.

Sea $(\tilde{\varphi},\tilde{\varphi}^*)$ un par óptimo del problema dual de Kantorovich del teorema 8 . Sea μ una probabilidad absolutamente continua. Como $\tilde{\varphi}\in L^1(\mu)$ (teorema 8), entonces $\tilde{\varphi}$ es finita en μ-casi todo punto. Por tanto, $\mu[\mathrm{Dom}(\tilde{\varphi})]=1$, con $\mathrm{Dom}(\tilde{\varphi})=\{x\in\mathbb{R}^d:\tilde{\varphi}(x)\neq\pm\infty\}$. Como $\tilde{\varphi}$ es convexa, $\mathrm{Dom}(\tilde{\varphi})$ es un conjunto convexo y, por tanto, su frontera, $\mathrm{Front}(\mathrm{Dom}(\tilde{\varphi}))$, tiene medida de Lebesgue nula, $\mathscr{L}(\mathrm{Front}(\mathrm{Dom}(\tilde{\varphi})))=0$. Como μ es absolutamente continua con respecto a la medida de Lebesgue, $\mu(\mathrm{Front}(\mathrm{Dom}(\tilde{\varphi})))=0$. Por tanto, el interior de $\mathrm{Dom}(\tilde{\varphi})$ tiene medida μ igual a 1, $\mu(\mathrm{Int}(\mathrm{Dom}(\tilde{\varphi})))=1$.

Como $\tilde{\varphi}$ es una función propia convexa, $\tilde{\varphi}$ es diferenciable en casi todo punto de $\mathrm{Int}(\mathrm{Dom}(\tilde{\varphi}))$ (ver capítulo 9 [5]). Por tanto, el conjunto de puntos de $\mathrm{Int}(\mathrm{Dom}(\tilde{\varphi}))$ donde $\tilde{\varphi}$ no es diferenciable tiene medida de Lebesgue nula. Como μ es absolutamente continua, $\tilde{\varphi}$ es diferenciable en μ-casi seguro. Por tanto, $\partial\tilde{\varphi}(x)=\{\nabla\tilde{\varphi}(x)\}$ para μ-casi todo punto $x\in\mathbb{R}^d$. Como μ es la marginal de π en $\mathscr{X}=\mathbb{R}^d$, $\partial\tilde{\varphi}(x)=\{\nabla\tilde{\varphi}(x)\}$ para π-casi todo punto $(x,y)\in\mathbb{R}^d\times\mathbb{R}^d$. Puesto que $(\tilde{\varphi},\tilde{\varphi}^*)\in\tilde{\Phi}_C$ (teorema 8), se tiene que $\tilde{\varphi}(x)+\tilde{\varphi}^*(y)=x\cdot y$ (probado en la demostración del teorema 9) y por la proposición 1, se deduce que $y\in\partial\tilde{\varphi}(x)$. Por tanto, $y=\nabla\tilde{\varphi}(x)$ para π-casi todo $(x,y)\in\mathbb{R}^d\times\mathbb{R}^d$. Así, $\pi=(Id\times\nabla\tilde{\varphi})\sharp\mu$ con $\nabla\tilde{\varphi}\sharp\mu=\nu$. Por el teorema 6, existe un plan de transporte óptimo, por lo que ese plan de transporte deberá ser de la misma forma.

Ahora se prueba la unicidad del plan de transporte. Sea $\tilde{\phi}$ otra función convexa tal que $\nabla\tilde{\phi}\sharp\mu=\nu$. Hay que probar que $\tilde{\varphi}=\tilde{\phi}$ salvo en un conjunto de medida μ nula. Por el criterio de Knott-Smith, $(\tilde{\varphi},\tilde{\varphi}^*)$ y $(\tilde{\phi},\tilde{\phi}^*)$ son pares óptimos para el problema dual. Por tanto, en ambos se alcanza el ínfimo de (24).

$$\int_{\mathbb{R}^d}\tilde{\phi}d\mu+\int_{\mathbb{R}^d}\tilde{\phi}^*d\nu=\int_{\mathbb{R}^d}\tilde{\varphi}d\mu+\int_{\mathbb{R}^d}\tilde{\varphi}^*d\nu. \tag{28}$$

Suponiendo que π es el plan de transporte óptimo asociado a $(\tilde{\varphi}, \tilde{\varphi}^*)$, teniendo en cuenta el lema 1, (28) queda:

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} [\tilde{\phi}(x) + \tilde{\phi}^*(y)] d\pi(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} [\tilde{\varphi}(x) + \tilde{\varphi}^*(y)] d\pi(x,y).$$

De aquí y teniendo en cuenta (25), se obtiene:

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} [\tilde{\phi}(x) + \tilde{\phi}^*(y)] d\pi(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} [\tilde{\varphi}(x) + \tilde{\varphi}^*(y)] d\pi(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} (x \cdot y) d\pi(x,y),$$

y como $\pi = (Id \times \nabla\tilde{\varphi})\sharp\mu$, se tiene que:

$$\int_{\mathbb{R}^d} [\tilde{\phi}(x) + \tilde{\phi}^*(\nabla\tilde{\varphi}(x))] d\mu(x) = \int_{\mathbb{R}^d} (x \cdot \nabla\tilde{\varphi}(x)) d\mu(x),$$

o equivalentemente,

$$\int_{\mathbb{R}^d} [\tilde{\phi}(x) + \tilde{\phi}^*(\nabla\tilde{\varphi}(x)) - x \cdot \nabla\tilde{\varphi}(x)] d\mu(x) = 0. \tag{29}$$

El integrando de (29) es no negativo porque $(\tilde{\phi}, \tilde{\phi}^*) \in \tilde{\Phi}_C$, luego para que se dé la igualdad de (29), ha de ser $\tilde{\phi}(x) + \tilde{\phi}^*(\nabla\tilde{\varphi}(x)) = x \cdot \nabla\tilde{\varphi}(x)$ en μ-casi todo $x \in \mathbb{R}^d$. Por la proposición 1, $\nabla\tilde{\varphi}(x) \in \partial\tilde{\phi}(x)$ para μ-casi todo x. Como $\tilde{\phi}$ es diferenciable en μ-casi todo punto, $\partial\tilde{\phi}(x) = \{\nabla\tilde{\phi}(x)\}$, por lo que $\nabla\tilde{\varphi}(x) = \nabla\tilde{\phi}(x)$ para μ-casi todo x. Por tanto, $\tilde{\varphi}$ es única π-casi seguro y con ello el plan de transporte óptimo π.

□

2.7. Expresiones explícitas del coste total mínimo del problema de Kantorovich

El coste total mínimo del problema de Kantorovich típicamente no tiene una expresión explícita. Sin embargo, hay dos casos en los que sí se conoce. El primero de ellos es el caso en que $\mathscr{X} = \mathscr{Y} = \mathbb{R}$ y $c(x,y) = h(|x-y|)$ con h una función convexa y no negativa. En esta sección se prueba el resultado para coste cuadrático, $c(x,y) = \| x - y \|^2$. Cabe mencionar también el segundo caso, cuando $\mathscr{X} = \mathscr{Y} = \mathbb{R}^d$, $c(x,y) = \| x - y \|^2$ y μ y ν pertenecen a la siguiente familia de probabilidades [3]:

Definición 15 (Familia *localización-dispersión* de probabilidades, [3])**.** *Sea $\vec{X}$ un vector aleatorio con distribución de probabilidad $\mathbb{P}_X$, con $\mathbb{P}_X$ una probabilidad de Borel en $P(\mathbb{R}^d)$ con momento de orden 2 finito y absolutamente continua con respecto a la medida de Lebesgue. La familia $\mathcal{F}(\mathbb{P}_X) := \{\mathbb{P}(AX+m) : A \in \mathcal{M}_{d\times d}, m \in \mathbb{R}^d\}$ con $\mathbb{P}(AX+m)$ la distribución de probabilidad de $AX+m$ y $\mathcal{M}_{d\times d}$ el conjunto de matrices $d \times d$ simétricas definidas positivas, se denomina familia* localización-dispersión.

Si μ y ν son medidas de probabilidad de la familia $\mathcal{F}(\mathbb{P}_X)$, con medias m_μ y m_ν y matrices de covarianza Σ_μ y Σ_ν y $c(x,y) = \| x - y \|^2$ es la función coste del problema de Kantorovich, se tiene [3]:

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = \| m_\mu - m_\nu \|^2 + Tr\left(\Sigma_\mu + \Sigma_\nu - 2\left(\Sigma_\mu^{1/2}\Sigma_\nu\Sigma_\mu^{1/2}\right)^{1/2}\right).$$

Se retoma ahora el caso en el que $\mathscr{X} = \mathscr{Y} = \mathbb{R}$ y $c(x,y) = \| x - y \|^2$. Sean $\mu, \nu \in P(\mathbb{R})$ con funciones de distribución F y G.

Se presentan a continuación unas definiciones y resultados que se utilizarán en la obtención del transporte óptimo en $\mathbb{R}$ para coste cuadrático.

Definición 16 (Subconjunto monótono, [35]). *Un subconjunto $A \subseteq \mathbb{R}^2$ se dice monótono si dados $(x_1, x_2), (y_1, y_2) \in A$, se cumple que $[x_1 \leq x_2 \ \ y \ \ y_1 \leq y_2] \ \ o \ \ [x_1 \geq x_2 \ \ y \ \ y_2 \geq y_2]$.*

Equivalentemente, si se cumple que $(x_1 - x_2)(y_1 - y_2) \geq 0$.

Teorema 11 (Monotonía del subdiferencial, [14]). *Sea $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ una función propia semicontinua inferiormente. Son equivalentes:*

1. *∂f es un conjunto monótono.*
2. *f es convexa.*

Demostración. Una demostración de este teorema se puede encontrar en [14]. □

Definición 17 (Función cuantil, [35]). *Sea $F : \mathbb{R} \to [0, 1]$ una función de distribución. Se define la función cuantil asociada a F como*

$$F^{-1}(t) = \text{ínf}\{x : F(x) \geq t\}, \ t \in [0, 1].$$

Las funciones cuantiles son no decrecientes y continuas por la izquierda. Obviamente, F^{-1} puede tomar los valores $\pm\infty$.

La siguiente proposición recoge la relación existente entre funciones de distribución y funciones cuantil y permite obtener el corolario 1 que es clave en la construcción a desarrollar. Ambas demostraciones se pueden encontrar en [9].

Proposición 2. *Sea F una función de distribución de X y Q su función cuantil. Sean $u \in (0, 1)$ y $x \in \mathbb{R}$. Entonces,*

$$Q(u) \leq x \text{ si y solo si } u \leq F(x).$$

Demostración. Dado $u \in (0, 1)$, como F es continua por la derecha, el conjunto $\{x : u \leq F(x)\}$ está acotado por la izquierda. Además, como F es creciente, este conjunto es un intervalo no acotado por la derecha. Entonces, $\{x : u \leq F(x)\} = [Q(u), \infty)$ y el resultado queda demostrado.

□

Corolario 1. *Sea Q la función cuantil de la distribución $\mathbb{P}$ (la función cuantil asociada la función de distribución de $\mathbb{P}$) y sea $U \sim \mathcal{U}(0, 1)$, con $\mathcal{U}(0, 1)$ la distribución uniforme en el intervalo $(0, 1)$. Entonces $Q(U)$ es una variable aleatoria con distribución $\mathbb{P}$.*

Demostración. Veamos que $F = F_{Q(U)}$, con $F_{Q(U)}$ la función de distribución de la variable aleatoria $Q(U)$. Esto implica que $\mathbb{P} = \mathbb{P}_{Q(U)}$ con $P_{Q(U)}$ la distribución de la variable aleatoria $Q(U)$.

Sea $U \sim \mathcal{U}(0, 1)$. Sin pérdida de generalidad, podemos suponer que $U \in (0, 1)$. Por la proposición 2, tenemos que $F_{Q(U)}(x) = \mathbb{P}[Q(U) \leq t] = \mathbb{P}[U \leq F(x)] = F(x)$. □

Se enuncia a continuación el teorema que caracteriza la solución del problema de Kantorovich en términos de funciones cuantiles.

Teorema 12 (Caracterización de la solución del problema de Kantorovich en $\mathbb{R}$ para coste cuadrático, [35]). *Sean μ y ν medidas de probabilidad en $\mathbb{R}$ con funciones de distribución F y G, respectivamente. Sea π la medida de probabilidad en $\mathbb{R}^2$ con función de distribución $H(x,y) = \min(F(x), G(x))$. Entonces, $\pi \in \Pi(\mu,\nu)$ y es óptimo para el problema de transporte de Kantorovich con coste $c(x,y) = \| x - y \|^2$. Además,*

$$\inf_{\pi\in\Pi(\mu,\nu)} C(\pi) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt. \tag{30}$$

Demostración. Sean μ y ν medidas de probabilidad en $\mathbb{R}$ con funciones de distribución F y G. Vamos a denotar $F(x^-) := \lim_{z\nearrow x} F(z)$, que existe porque F es no decreciente y acotada. Como F es continua por la derecha, $F(x^+) = \lim_{z\searrow x} F(z) = F(x)$.

La demostración seguirá los siguientes pasos. Sea π la medida de probabilidad en $\mathbb{R}^2$ con función de distribución $H(x) = \min(F(x), G(x))$. Primero se probará que

$$\text{Soporte}(\pi) \subset \{(x,y) \in \mathbb{R}^2 : F(x^-) \leq G(y) \text{ y } G(y^-) \leq F(x)\}, \tag{31}$$

que se usará para probar que Soporte(π) es un conjunto monótono. Esto permitirá demostrar que, como $\pi \in \Pi(\mu,\nu)$, π es un plan de transporte óptimo del problema de Kantorovich. Finalmente se probará la igualdad (30).

En primer lugar se prueba que Soporte(π) $\subset \{(x,y) \in \mathbb{R}^2 : F(x^-) \leq G(y) \text{ y } G(y^-) \leq F(x)\}$. Supongamos, por reducción al absurdo, que existen $x, y \in \mathbb{R}$ tal que $(x,y) \in$ Soporte(π) y $F(x^-) > G(y)$. Sea x' en un entorno pequeño de x e y' en un entorno pequeño de y. Teniendo en cuenta que G es continua por la derecha y que F y G son no decrecientes, se tiene que $F(x') > G(y')$. Por tanto,

$$H(x',y') = \min[F(x'), G(y')] = G(y').$$

En consecuencia, en un entorno rectangular de (x,y), H no depende de x'. Así, dicho entorno tiene π-medida nula: $\pi[(x-\varepsilon, x+\varepsilon)\times(y-\varepsilon, y+\varepsilon)] = H(x+\varepsilon, y+\varepsilon) + H(x-\varepsilon, y-\varepsilon) - H(x-\varepsilon, y+\varepsilon) - H(x+\varepsilon, y-\varepsilon) = G(y+\varepsilon) + G(y-\varepsilon) - G(y+\varepsilon) - G(y-\varepsilon) = 0$. Como consecuencia, $(x,y) \notin$ Soporte(π), que supone una contradicción, pues se ha supuesto que $(x,y) \in$ Soporte(π). Idénticamente se prueba que si $(x,y) \in$ Soporte(π), entonces $G(y^-) \leq F(x)$. Por tanto, (31) está demostrado.

En segundo lugar, veamos que Soporte(π) es un conjunto monótono. Sean $(x_1,y_1), (x_2,y_2) \in$ Soporte(π) tal que $x_1 > x_2$. Hay que probar que $y_1 \geq y_2$. Teniendo en cuenta (31), se tiene

$$G(y_1) \geq F(x_1^-) \geq F(x_2) \geq G(y_2^-).$$

Si $G(y_1) > G(y_2^-)$, como G es no decreciente, se tiene que $y_1 \geq y_2$, y quedaría probado que Soporte(π) es monótono. Si no es así, necesariamente

$$G(y_1) = F(x_1^-) = F(x_2) = G(y_2^-). \tag{32}$$

Supongamos, por reducción al absurdo, que $y_2 > y_1$. Sea $\varepsilon > 0$ tal que $x_2 + \varepsilon < x_1$ e $y_1 < y_2 - \varepsilon$ y R_ε el rectángulo $(x_2-\varepsilon, x_2+\varepsilon)\times(y_2-\varepsilon, y_2+\varepsilon)$. Teniendo en cuenta (32), que implica que F es constante

en $[x_2, x_1)$ y G es constante en $[y_1, y_2)$ e igual a F, y teniendo en cuenta también que F y G son no decrecientes, se tiene:

$$\begin{aligned}
\pi[R_\varepsilon] &= H(x_2+\varepsilon, y_2+\varepsilon) + H(x_2-\varepsilon, y_2-\varepsilon) - H(x_2-\varepsilon, y_2+\varepsilon) - H(x_2+\varepsilon, y_2-\varepsilon) \\
&= \text{mín}[F(x_2+\varepsilon), G(y_2+\varepsilon)] + \text{mín}[F(x_2-\varepsilon), G(y_2-\varepsilon)] - \text{mín}[F(x_2-\varepsilon), G(y_2+\varepsilon)] \\
&\quad - \text{mín}[F(x_2+\varepsilon), G(y_2-\varepsilon)] \\
&= F(x_2+\varepsilon) + F(x_2-\varepsilon) - F(x_2-\varepsilon) - G(y_2-\varepsilon) \\
&= G(y_2-\varepsilon) + F(x_2-\varepsilon) - F(x_2-\varepsilon) - G(y_2-\varepsilon) = 0
\end{aligned}$$

Luego $(x_2, y_2) \notin \text{Soporte}(\pi)$, lo que supone una contradicción. Por tanto, $y_1 \geq y_2$ y queda probado que Soporte(π) es un conjunto monótono.

En consecuencia, Soporte(π) está contenido en un subconjunto monótono de $\mathbb{R}^2$ y, por el teorema 11, en el subdiferencial de una función convexa y semicontinua inferiormente. Por tanto, si probamos que $\pi \in \Pi(\mu, \nu)$, por el criterio de Knott-Smith (teorema 9), π es un plan de transporte óptimo.

Veamos ahora que $\pi = (F^{-1} \times G^{-1})\sharp\mathscr{L}_{[0,1]}$, con $\mathcal{L}_{[0,1]}$ la medida de Lebesgue en $[0,1]$. Las marginales de π son $F^{-1}\sharp\mathcal{L}_{(0,1)}$ y $G^{-1}\sharp\mathcal{L}_{(0,1}$ que, por el corolario 1, tienen distribuciones μ y ν, luego $\pi \in \Pi(\mu, \nu)$. Probemos que la función de distribución de $\pi = (F^{-1} \times G^{-1})\sharp\mathscr{L}_{[0,1]}$ es $H(x, y) = \text{mín}[F(x), G(y)]$.

Calculemos $\pi[R(x, y)] = \mathscr{L}_{[0,1]}\left[\{t \in \mathbb{R} : (F^{-1}(t), G^{-1}(t)) \in R(x, y)\}\right]$ con $R(x, y)$ el rectángulo $(-\infty, x] \times (-\infty, y]$:

$$\begin{aligned}
\pi[R(x,y)] &= \mathscr{L}_{[0,1]}\left[\{t \in \mathbb{R} : (F^{-1}(t), G^{-1}(t)) \in R(x,y)\}\right] \\
&= \mathscr{L}_{[0,1]}\left[\{t \in \mathbb{R} : F^{-1}(t) \leq x\} \bigcap \{t \in \mathbb{R} : G^{-1}(t) \leq y\}\right].
\end{aligned}$$

Por la proposición 2, $\{t \in \mathbb{R} : F^{-1}(t) \leq x\}$=$[0, F(x)]$ y $\{t \in \mathbb{R} : G^{-1}(t) \leq y\} = [0, G(y)]$. En consecuencia, $\{t \in \mathbb{R} : F^{-1}(t) \leq x\} \bigcap \{t \in \mathbb{R} : G^{-1}(t) \leq y\}$ es un intervalo con extremos 0 y $\text{mín}[F(x), G(y)] = H(x, y)$. Por tanto, $\mathscr{L}_{[0,1]}\left[\{t \in \mathbb{R} : F^{-1}(t) \leq x\} \bigcap \{t \in \mathbb{R} : G^{-1}(t) \leq y\}\right] = \text{mín}[F(x), G(y)] = H(x, y)$. Así, $\pi \in \Pi(\mu, \nu)$ y π es óptimo para el problema de Kantorovich.

Como $\pi = (F^{-1} \times G^{-1})\sharp\mathscr{L}_{[0,1]}$, para cualquier función f medible no negativa definida en $\mathbb{R}^2$, se tiene,

$$\int_{\mathbb{R}^2} f(x, y) d\pi(x, y) = \int_0^1 f(F^{-1}(t), G^{-1}(t)) dt.$$

En concreto, para $c(x, y) = \| x - y \|^2$,

$$\int_{\mathbb{R}^2} c(x, y) d\pi(x, y) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt.$$

Como $\pi = (F^{-1} \times G^{-1})\sharp\mathscr{L}_{[0,1]}$ es óptimo,

$$\inf_{\pi \in \Pi(\mu,\nu)} C(\pi) = C((F^{-1} \times G^{-1})\sharp\mathscr{L}_{[0,1]}) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt.$$

□

Observación 1. *El teorema 12 se cumple también con* $c(x, y) = h(|x-y|)$ *con* h *convexa y no negativa.*

3. Distancia de Wasserstein

El problema del transporte óptimo permite cuantificar el coste de transformar una distribución en otra y busca minimizarlo. Este coste mínimo se puede emplear para definir una distancia en el espacio de probabilidades $P(\mathscr{X})$ sobre un espacio $\mathscr{X}$. Esta distancia, conocida como distancia de Wasserstein, define un espacio métrico, subespacio de $P(\mathscr{X})$, que se conoce como espacio de Wasserstein. Este capítulo se ha desarrollado a partir de las referencias [4], [12], [21], [26] y [27].

Definición 18 (p-espacio de Wasserstein, [35])**.** *Sea $(\mathscr{X}, d)$ un espacio métrico completo y separable. Se define el p-espacio de Wasserstein, con $p \geq 1$, como:*

$$\mathscr{W}_p(\mathscr{X}) = \left\{ \mu \in P(\mathscr{X}) : \int_{\mathscr{X}} d(x, x_0)^p d\mu(x) < \infty, \text{ para algún } x_0 \in \mathscr{X} \text{ fijo} \right\}.$$

Observación 2. *La desigualdad triangular garantiza que si*

$$\int d(x, x_0)^p d\mu(x) < \infty$$

para algún $x_0 \in \mathscr{X}$, entonces lo es para todo punto de $\mathscr{X}$.

Claramente, si d es acotada, el p-espacio de Wasserstein coincide con $P(\mathscr{X})$.

Definición 19 (Distancia de Wasserstein, [35])**.** *Sean μ y ν dos medidas de probabilidad pertenecientes al p-espacio de Wasserstein, $\mathscr{W}_p(\mathscr{X})$, con $p \geq 1$. Se define la p-distancia de Wasserstein entre μ y ν, $W_p(\mu, \nu)$, como la raíz p-ésima del mínimo coste de transporte entre μ y ν en el problema de Kantorovich con respecto a la función coste $c(x, y) = d(x, y)^p$:*

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} C_p(\pi) \right)^{1/p} = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathscr{X} \times \mathscr{X}} d(x_1, x_2)^p d\pi(x_1, x_2) \right)^{1/p}.$$

Teorema 13 (Distancia de Wasserstein, [35])**.** *La distancia de Wasserstein W_p es una distancia en el p-espacio de Wasserstein para todo $p \in [1, \infty)$.*

Para la prueba de este teorema se utilizará el siguiente lema. Una demostración de este resultado se puede encontrar en [4]:

Lema 2 (Del Pegado, [4])**.** *Sean $\mathscr{X}_1, \mathscr{X}_2, \mathscr{X}_3$ tres espacios completos y separables. Sea μ_i una medida de probabilidad, con soporte en $\mathscr{X}_i$, $i = 1, 2, 3$. Sean $\pi_{12} \in \Pi(\mu_1, \mu_2)$ y $\pi_{23} \in \Pi(\mu_2, \mu_3)$ planes de transporte. Entonces, existe una medida de probabilidad $\pi \in P(\mathscr{X}_1 \times \mathscr{X}_2 \times \mathscr{X}_3)$ con marginales π_{12} sobre $\mathscr{X}_1 \times \mathscr{X}_2$ y π_{23} sobre $\mathscr{X}_2 \times \mathscr{X}_3$.*

Demostración del teorema 13. Para probar que W_p es una distancia, se tiene que probar que dadas $\mu,\ \nu,\ \lambda \in \mathscr{W}_p(\mathscr{X})$:

1. $W_p(\mu, \nu) < \infty$.
2. $W_p(\mu, \nu) \geq 0$.
3. $W_p(\mu, \nu) = 0$ si y solo si $\mu = \nu$.
4. $W_p(\mu, \nu) = W_p(\nu, \mu)$.
5. $W_p(\mu, \nu) \leq W_p(\mu, \lambda) + W_p(\lambda, \nu)$.

Se demuestran a continuación cada uno de estos puntos.

1. Sea $\pi \in \Pi(\mu, \nu)$. Se tiene que:

$$\begin{aligned} W_p(\mu,\nu) &\leq \left(\int_{\mathscr{X}\times\mathscr{X}} d(x_1,x_2)^p d\pi(x_1,x_2)\right)^{1/p} \\ &\leq \left(\int_{\mathscr{X}\times\mathscr{X}} (d(x_1,z_0)+d(z_0,x_2))^p d\pi(x_1,x_2)\right)^{1/p} \\ &\leq \left(\int_{\mathscr{X}} d(x_1,z_0)^p d\mu(x_1)\right)^{1/p} + \left(\int_{\mathscr{X}} d(z_0,x_2)^p d\nu(x_2)\right)^{1/p} < \infty. \end{aligned}$$

   La primera desigualdad se tiene porque π es un plan de transporte admisible, pero no tiene por qué ser el óptimo. La segunda viene de aplicar la desigualdad triangular a d y la tercera de aplicar la desigualdad de Minkowski. Finalmente, aplicando la observación 2, se tiene que $W_p(\mu, \nu)$ es finita.

2. Es trivial porque d es no negativa.

3. Supongamos que $\mu = \nu$. Sea X un elemento aleatorio con distribución μ y consideremos el elemento aleatorio bidimensional (X, X), cuya distribución obviamente pertenece a $\Pi(\mu, \mu)$. Se tiene:

$$W_p(\mu,\mu) = \left(\inf_{\pi\in\Pi(\mu,\mu)} E_\pi[c(X,Y)]\right)^{1/p} \leq E[c(X,X)] = 0.$$

   Como W_p es no negativa, se concluye que $W_p(\mu, \mu) = 0$.

   Se prueba ahora la condición necesaria. Sean $\mu, \nu \in \mathscr{W}_p(\mathscr{X})$ tales que $W(\mu, \nu) = 0$ y veamos que $\mu = \nu$. Sea π_0 un plan de transporte óptimo (que por el teorema 6 sabemos que existe). Si $\left(\text{ínf}_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} d(x_1,x_2)^p d\pi(x_1,x_2)\right)^{1/p} = 0$, entonces $d(x_1, x_2)^p = 0$ π_0-casi seguro. Por lo tanto, ha de ser $x_1 = x_2$ π_0-casi seguro, por lo que las dos marginales de π_0 coinciden.

4. En la sección 2.2, se vio que para toda $\pi \in \Pi(\mu, \nu)$ existe $\tilde{\pi} \in \Pi(\nu, \mu)$ con $\tilde{\pi}(B \times A) = \pi(A \times B)$, luego

$$\begin{aligned} W_p(\mu,\nu) &= \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} d(x_1,x_2)^p d\pi(x_1,x_2)\right)^{1/p} \\ &= \left(\inf_{\tilde{\pi}\in\Pi(\nu,\mu)} \int_{\mathscr{X}\times\mathscr{X}} d(x_1,x_2)^p d\tilde{\pi}(x_2,x_1)\right)^{1/p} = W_p(\nu,\mu). \end{aligned}$$

5. Sean $\mu, \nu, \gamma \in \mathscr{W}_p(\mathscr{X})$ con soportes $\mathscr{X}$,$\mathscr{Y}$ y $\mathscr{Z}$, respectivamente. Sean $\pi_{\mu\nu}$ un plan de transporte óptimo entre μ y ν y $\pi_{\nu\gamma}$ un plan de transporte óptimo entre ν y γ. Se define π como en el lema de Pegado (lema 2), es decir $\pi \in P(\mathscr{X} \times \mathscr{Y} \times \mathscr{Z})$ con marginales $\pi_{\mu\nu}$ sobre $\mathscr{X} \times \mathscr{Y}$ y $\pi_{\nu\gamma}$ sobre $\mathscr{Y} \times \mathscr{Z}$. Se llama $\pi_{\mu\gamma}$ a la marginal de π en $\mathscr{X} \times \mathscr{Z}$. Se tiene,

$$
\begin{aligned}
W_p(\mu,\gamma) &\leq \left(\int_{\mathscr{X}\times\mathscr{Z}} d(x,z)^p d\pi_{\mu\gamma}(x,z)\right)^{1/p} = \left(\int_{\mathscr{X}\times\mathscr{Y}\times\mathscr{Z}} d(x,z)^p d\pi(x,y,z)\right)^{1/p} \\
&\leq \left(\int_{\mathscr{X}\times\mathscr{Y}\times\mathscr{Z}} [d(x,y)+d(y,z)]^p d\pi(x,y,z)\right)^{1/p} \\
&\leq \left(\int_{\mathscr{X}\times\mathscr{Y}\times\mathscr{Z}} d(x,y)^p d\pi(x,y,z)\right)^{1/p} + \left(\int_{\mathscr{X}\times\mathscr{Y}\times\mathscr{Z}} d(y,z)^p d\pi(x,y,z)\right)^{1/p} \\
&= \left(\int_{\mathscr{X}\times\mathscr{Y}} d(x,y)^p d\pi_{\mu\nu}(x,y)\right)^{1/p} + \left(\int_{\mathscr{Y}\times\mathscr{Z}} d(y,z)^p d\pi_{\nu\gamma}(y,z)\right)^{1/p} \\
&= W_p(\mu,\nu) + W_p(\nu,\gamma).
\end{aligned}
$$

La primera desigualdad se tiene porque $\pi_{\mu\gamma}$ es un plan de transporte, no tiene por qué ser el óptimo. La segunda desigualdad viene de la desigualdad triangular aplicada a d. Por último, la tercera desigualdad se obtiene al aplicar la desigualdad de Minkowski.

□

3.1. Caso $\mathscr{X} = \mathbb{R}$ y coste cuadrático

En esta sección se toma $\mathscr{X} = \mathbb{R}$. Utilizando la distancia de Wasserstein asociada al coste cuadrático $c(x,y) = \| x-y \|^2$, se tiene

$$
W_2(\mu,\nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^2} \| x-y \|^2 \, d\pi(x,y)\right)^{1/2}.
$$

El correspondiente espacio de Wasserstein coincide con el espacio de probabilidades en $\mathbb{R}$ con momento de orden dos finito:

$$
\mathscr{W}_2(\mathbb{R}) = \left\{\mu \in P(\mathbb{R}) : \int_{\mathbb{R}} \| x \|^2 < \infty\right\}.
$$

Por el teorema 12, suponiendo que $\mu,\nu \in \mathscr{W}_2(\mathbb{R})$ con funciones cuantil F^{-1} y G^{-1}, se tiene que la distancia de Wasserstein asociada al coste cuadrático cuando $\mathscr{X} = \mathbb{R}$, se puede calcular de la siguiente manera:

$$
W_2(\mu,\nu) = \left(\int_0^1 |F^{-1}(t) - G^{-1}(t)|^2 dt\right)^{1/2}. \tag{33}
$$

4. Media de Fréchet o baricentro de Wasserstein en el espacio $\mathscr{W}_2(\mathscr{X})$

En este capítulo se introduce una generalización del concepto de media en el 2-espacio de Wasserstein $\mathscr{W}_2(\mathscr{X})$. Se presentan ciertos resultados sobre su existencia y unicidad y se aporta una expresión cuando $\mathscr{X} = \mathbb{R}$. Esta idea de media se puede generalizar a cualquier p-espacio de Wasserstein $\mathscr{W}_p(\mathscr{X})$. Se han seguido las referencias [1], [22], [26] y [35].

4.1. Funcional y media de Fréchet

La media es uno de los parámetros que más interés tiene en el campo de la estadística. En un espacio de Hilbert $\mathscr{H}$, la media de un conjunto de datos se puede definir como el único elemento de $\mathscr{H}$ que minimiza el funcional:

$$H(x) = \sum_{i=1}^{n} \| x - x_i \|^2,$$

que resulta ser $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Una noción más general de la media es la media ponderada, que se define de la misma manera, salvo que las distancias entre puntos van ponderadas por pesos. La media ponderada de un conjunto de datos $\{x_1, ..., x_n\}$ se define como el elemento que minimiza el funcional:

$$\bar{H}(x) = \sum_{i=1}^{n} w_i \| x - x_i \|^2,$$

donde w_i es el peso asociado al valor x_i y que de nuevo tiene la expresión $\bar{x}_w = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i x_i$.

Este concepto se puede generalizar a cualquier espacio métrico considerando la distancia correspondiente en dicho espacio métrico, aunque es posible que no exista una expresión explícita de la media. En particular, esto es posible en el 2-espacio de Wasserstein, $\mathscr{W}_2(\mathscr{X})$, ya que según el teorema 13, dotado de la distancia de Wasserstein W_2, es un espacio métrico. El baricentro en este espacio, denominado baricentro de Wasserstein o media de Fréchet, es aquel que minimiza el denominado funcional de Fréchet en el espacio $\mathscr{W}_2(\mathscr{X})$.

Definición 20 (Funcional y funcional pesado de Fréchet, [26])**.** *El funcional de Fréchet asociado a las medidas $\mu_1, ..., \mu_n \in \mathscr{W}_2(\mathscr{X})$ es:*

$$H(\gamma) = \sum_{i=1}^{n} W_2^2(\gamma, \mu_i), \quad \gamma \in \mathscr{W}_2(\mathscr{X}).$$

De manera más general, el funcional de Fréchet pesado se define como:

$$H(\gamma) = \sum_{i=1}^{n} w_i W_2^2(\gamma, \mu_i), \quad \gamma \in \mathscr{W}_2(\mathscr{X}).$$

Los conceptos de funcional y funcional pesado de Fréchet se pueden extender a cualquier p-espacio de Wasserstein $\mathscr{W}_p(\mathscr{X})$ sin más que cambiar $W_2^2(\mu, \nu)$ por $W_p^p(\mu, \nu)$.

Definición 21 (Media de Fréchet, [26])**.** *Sean $\mu_1, ..., \mu_n$ medidas en $\mathscr{W}_2(\mathscr{X})$. Una media de Fréchet de $\{\mu_1, ..., \mu_n\}$ es, si existe, un minimizador del funcional de Fréchet H definido en $\mathscr{W}_2(\mathscr{X})$.*

La noción de media de Fréchet se puede extender también a cualquier p-espacio de Wasserstein, definiéndose como el minimizador del funcional de Fréchet definido en $\mathscr{W}_p(\mathscr{X})$.

4.1.1. Formulación multimarginal, existencia y unicidad

Gangbo y Swiech [22] formularon un problema del transporte multimarginal cuya solución es equivalente a encontrar la media de Fréchet de las distribuciones involucradas.

El problema es el siguiente: sean $\mu_1, ..., \mu_n$ medidas en $\mathscr{W}_2(\mathbb{R}^d)$ y sea $\Pi(\mu_1, ..., \mu_n)$ el conjunto de medidas de probabilidad sobre $\mathbb{R}^{d \cdot n}$ que tienen como distribuciones marginales a $\mu_1, ..., \mu_n$. Los elementos de $\Pi(\mu_1, ..., \mu_n)$ reciben el nombre de *multicouplings* de $\mu_1, ...\mu_n$. El problema consiste en hallar $\pi \in \Pi(\mu_1, ..., \mu_n)$ que minimice el funcional

$$G(\pi) = \int_{\mathbb{R}^{d \cdot n}} \sum_{i<j} \| x_i - x_k \|^2 \, d\pi(x_1, ..., x_n). \tag{34}$$

El siguiente teorema prueba la equivalencia entre el problema del transporte óptimo multimarginal y la media de Fréchet. En [22] se prueba la existencia de solución para el problema del transporte óptimo multimarginal. Por tanto, el teorema 14 prueba también la existencia de media de Fréchet para una colección de medidas.

Teorema 14 (Equivalencia del problema del transporte óptimo multimarginal y la media de Fréchet, [26])**.** *Si $\mu_1, ..., \mu_n \in \mathscr{W}(\mathbb{R}^d)$, se cumple que μ es una media de Fréchet de $\{\mu_1, ..., \mu_n\}$ si y solo si existe un multicoupling $\pi \in \mathscr{W}(\mathbb{R}^{d \cdot n})$ de $\{\mu_1, ..., \mu_n\}$ tal que $\mu = M\sharp\pi$, con $M(x) = M(x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. Además, $H(\mu) = G(\pi)$ con H el funcional de Fréchet y G definido como en (34).*

Demostración. La demostración de este teorema se puede encontrar en [1]. □

Agueh y Carlier probaron la unicidad de la media de Frechet en $\mathscr{W}_2(\mathbb{R}^d)$ cuando una de las medidas es absolutamente continua [1].

4.1.2. Caso $\mathscr{X} = \mathbb{R}$

Cuando $\mathscr{X} = \mathbb{R}$, la 2-distancia de Wasserstein tiene una expresión explícita. Esta expresión es la que se obtuvo para el coste total mínimo del problema de Kantorovich en $\mathbb{R}$ con coste cuadrático (ecuación (30) del teorema 12).

Sean $\mu_1, ..., \mu_n$ medidas en $\mathscr{W}_2(\mathbb{R})$. La media de Fréchet de $\mu_1, ..., \mu_n$, que existe por el teorema 14, es aquella medida μ que minimiza el funcional de Fréchet (definición 20). Cuando $\mathscr{X} = \mathbb{R}$, el funcional de Fréchet tiene la expresión:

$$H(\gamma) = \sum_{i=1}^n W_2^2(\gamma, \mu_i) = \sum_{i=1}^n \int_0^1 (F_{\mu_i}^{-1}(t) - F_\gamma^{-1}(t))^2 dt = \int_0^1 \sum_{i=1}^n (F_{\mu_i}^{-1}(t) - F_\gamma^{-1}(t))^2 dt.$$

Buscamos la función cuantil de la media de Fréchet. Para que se minimice H, es suficiente minimizar $\sum_{i=1}^n (F_{\mu_i}^{-1}(t) - F_\gamma^{-1}(t))^2$ para cada $t \in \mathbb{R}$. Por tanto, dado $t \in \mathbb{R}$, vamos a minimizar la función real de variable real $h \to \psi(h) := \sum_{i=1}^n (F_{\mu_i}^{-1}(t) - h)^2$. Derivando se tiene:

$$\frac{d}{dh}\psi(h) = \frac{d}{dh}\sum_{i=1}^{n}(F_{\mu_i}^{-1}(t) - h)^2 = 2\sum_{i=1}^{n}(F_{\mu_i}^{-1}(t) - h). \tag{35}$$

Obviamente esta derivada se anula si y solo si $h = \frac{1}{n}\sum_{i=1}^{n} F_{\mu_i}^{-1}(t)$. Como la función

$$t \to \frac{1}{n}\sum_{i=1}^{n} F_{\mu_i}^{-1}(t)$$

es una función cuantil, es la función cuantil de la media de Fréchet. Más precisamente, la media de Fréchet de $\mu_1, ...\mu_n$ es la probabilidad con función cuantil dada por (35).

4.2. Funcional y media poblacionales de Fréchet

El concepto de media de Fréchet se puede generalizar al nivel poblacional. De esta manera, las medidas $\mu_1, ..., \mu_n$ de la sección 4.1 serían una muestra de una medida de probabilidad aleatoria Λ.

Definición 22 (Medida aleatoria, [26]). *Una medida aleatoria Λ es una función medible $\Lambda : (\Omega, \mathscr{A}, \mathbb{P}) \to (\mathscr{W}_p(\mathscr{X}), \beta_{\mathscr{W}_p(\mathscr{X})})$.*

Definición 23 (Funcional poblacional de Fréchet, [26]). *Sea Λ una medida aleatoria en $\mathscr{W}_2(\mathscr{X})$. El funcional de Fréchet asociado a Λ se define como*

$$H(\gamma) = E[W_2^2(\gamma, \Lambda)], \quad \gamma \in \mathscr{W}_2(\mathscr{X}).$$

Definición 24 (Media poblacional de Fréchet, [26]). *La media de Fréchet de una medida aleatoria Λ en $\mathscr{W}_2(\mathscr{X})$ es, si existe y es única, el minimizador del funcional de Fréchet H asociado a Λ en $\mathscr{W}_2(\mathscr{X})$.*

Estos conceptos, al igual que en el caso de un conjunto finito de medidas, se pueden extender a cualquier p-espacio de Wasserstein.

4.2.1. Existencia y unicidad

Los siguientes teoremas prueban la existencia y unicidad de la media de Fréchet bajo ciertas condiciones. No se presentan sus demostraciones pero se pueden encontrar en [26].

Teorema 15 (Existencia de la media de Fréchet para medidas aleatorias en $\mathscr{W}_2(\mathbb{R}^d)$, [26]). *El funcional de Fréchet $H(\gamma)$ asociado a cualquier medida aleatoria $\Lambda : (\Omega, \mathscr{A}, \mathbb{P}) \to \mathscr{W}_2(\mathbb{R}^d)$ admite un minimizador.*

Teorema 16 (Existencia de la media de Fréchet para medidas aleatorias con $\mathscr{X}$ un espacio de Hilbert infinito-dimensional, [26]). *Sea $\mathscr{X}$ un espacio de Hilbert infinito-dimensional. Sea $\Lambda : (\Omega, \mathscr{A}, \mathbb{P}) \to \mathscr{W}_2(\mathscr{X})$. Si existe un conjunto $K \subseteq \mathscr{X}$ convexo y cerrado tal que $\mathbb{P}[\Lambda(K) = 1] = 1$, entonces el funcional de Fréchet $H(\gamma)$ admite un minimizador con soporte en K.*

Teorema 17 (Unicidad de la media de Fréchet, [26]). *Sea Λ una medida aleatoria en $\mathscr{W}_2(\mathscr{X})$ con un funcional de Fréchet asociado finito. Si Λ es absolutamente continua con probabilidad positiva, entonces la media de Fréchet de Λ, si existe, es única.*

5. Regresión no paramétrica funcional

Este capítulo se ha desarrollado a partir de las referencias [6], [13], [17], [18], [19], [20], [33], [34] y [36].

5.1. Regresión no paramétrica

La idea básica de la inferencia no paramétrica es realizar la menor cantidad de suposiciones sobre los datos disponibles.

Un problema de regresión consiste en estimar relaciones entre dos variables cuantitativas. Más concretamente, usualmente modeliza la dependencia de la media de una variable Y (variable dependiente) con una variable X (variable independiente o explicativa). Las relaciones vienen dadas por una función r denominada función de regresión. La regresión no paramétrica consiste en una colección de técnicas para el ajuste de funciones de regresión cuando no se tiene mucha información acerca de sus formas. Además, normalmente se pretenden obtener funciones suaves por lo que estas técnicas reciben el nombre de técnicas de suavizado.

El modelo de regresión no paramétrica se plantea de la siguiente manera:

$$Y = r(X) + \varepsilon,$$

con ε una variable aleatoria independiente de X y con $E[\varepsilon] = 0$. r es la denominada función de regresión.

Dado un conjunto de observaciones $(x_1, y_1), ..., (x_n, y_n)$, el problema consiste en estimar $r(x) = E[Y/X = x]$ para x en el conjunto imagen de X. El estimador de $r(x)$ se denota por $\hat{r}_n(x)$.

Para obtener una estimación suavizada de la función de regresión, se hace que intervengan en la estimación de la función r en un punto x las observaciones en ese punto y algunos cercanos (aquellas dentro de una ventana) y se pesan dichos valores con un peso que depende de la distancia entre la observación y x.

Existen distintas técnicas de suavizado: regresión por núcleos, regresión local polinómica, árboles de regresión, *splines* suavizadores...

Nos limitaremos a desarrollar la teoría de la regresión por núcleos ya que es la que se empleará para el caso práctico.

5.1.1. Regresión no paramétrica con variable independiente funcional y variable dependiente escalar

Los modelos de regresión que involucran datos funcionales se pueden agrupar en tres tipos: aquellos donde la variable dependiente y la independiente son funcionales, aquellos donde la variable dependiente es funcional y la variable independiente es un escalar o un vector y aquellos donde la variable dependiente es un escalar o un vector y la independiente funcional. Este último tipo es con el que se trata en este trabajo.

Existe abundante bibliografía de los dos primeros tipos, mientras que del último solo se han encontrado las referencias [6], [13] y [17].

Se plantea a continuación el modelo de regresión no paramétrica con una variable independiente real y una variable dependiente funcional.

Sean $(x_1, \mathcal{Y}_1), ..., (x_n, \mathcal{Y}_n)$ con $x_i \in \mathbb{R}$ e $\mathcal{Y}_i$ funciones. El funcional de regresión $r : \mathbb{R} \to E$ se define como:

$$r(x) = E[\mathcal{Y}/X = x]. \tag{36}$$

Con (36) se está suponiendo implícitamente la existencia de la probabilidad condicionada de $\mathcal{Y}$ a x. Queda fuera de los objetivos de este trabajo probar la existencia de esta probabilidad.

En la práctica, se dispone de valores de cada función $\mathcal{Y}_i$. Por tanto, la manera de obtener $\hat{r}_n$ consiste en obtener valores de la función, $\hat{r}_n(x)$, y aplicar técnicas de interpolación.

Habría que obtener condiciones bajo las cuales se puede obtener un estimador de r a partir de la muestra disponible con buenas propiedades asintóticas. Sin embargo, no se ha encontrado bibliografía acerca de esto para el caso de regresión con variable independiente escalar y variable dependiente funcional.

5.2. Regresión local por núcleos

Una de las técnicas de suavizado más empleadas es la regresión local por núcleos. Consiste en estimar $r(x)$ como una media ponderada donde los pesos vienen dados por ciertas funciones llamadas núcleos. Antes de presentar los estimadores, se incluyen unas nociones acerca de los núcleos y el pesado local por núcleos.

Definición 25 (Núcleo)**.** *Un núcleo es una función $K : \mathbb{R} \to \mathbb{R}^+$ integrable. En muchas aplicaciones, se incluyen también las siguientes condiciones:*

1. $\int_{-\infty}^{\infty} K(u)du = 1$.
2. $K(-u) = K(u)$.

En general, se emplean núcleos tales que $\text{argmax}_{u\in\mathbb{R}} K(u) = 0$.

Algunos de los núcleos más usados son:

- Núcleo *box*: $K(u) = \frac{1}{2}1_{[-1,+1]}(u)$.
- Núcleo triángulo: $K(u) = (u+1)1_{[-1,0]}(u) + (1-u)1_{[0,1]}(u)$.
- Núcleo cuadrático: $K(u) = \frac{3}{4}(1-u^2)1_{[-1,+1]}(u)$.
- Núcleo gaussiano: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$.

con $1_I(u) = \begin{cases} 1 & \text{si } u \in I \\ 0 & \text{en otro caso} \end{cases}$.

Generalmente, $K(u)$ disminuye a medida que $|u|$ aumenta.

Sean $x_1, ..., x_n$ valores de una muestra aleatoria simple tomada de la distribución de la variable aleatoria real X. El pesado por núcleos en un punto x consiste en asociar a cada una de estos valores un peso que depende de la distancia entre x y x_i y que tiende a disminuir con la distancia.

Matemáticamente, el pesado por núcleos consiste en transformar los valores $x_1, ..., x_n$ de una variable aleatoria real en otros, $\tilde{x}_1, ...\tilde{x}_n$, de la siguiente manera:

$$\tilde{x}_i = \tilde{x}_i(x, h, K) = \frac{1}{h} K\left(\frac{x - x_i}{h}\right),$$

donde K es una función núcleo y h recibe el nombre de parámetro de suavizado.

Volviendo a la regresión local por núcleos, a continuación se presenta el estimador de Nadaraya-Watson para $r(x)$.

Definición 26 (Estimador Nadaraya-Watson)**.** *Sean $(x_1, y_1), ..., (x_n, y_n)$ observaciones de unas variables X e Y reales. Sea $h > 0$. El estimador de Nadaraya-Watson se define como:*

$$\hat{r}_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}. \tag{37}$$

Por tanto, el peso asociado a cada valor y_i, $i = 1, ..., n$, de la variable dependiente es:

$$w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)},$$

y (37) se puede reescribir como:

$$\hat{r}_n(x) = \sum_{i=1}^{n} w_i(x) y_i.$$

Obsérvese que los pesos, y por tanto $\hat{r}_n$, dependen de h. Sin embargo, no hacemos explícita esta dependencia en la notación.

Para comprender mejor el método de la estimación por núcleos, considérese un núcleo K tal que $\{u \in \mathbb{R} : K(u) \neq 0\} = [0, 1]$. Por tanto, si $|x - x_i| > h$, entonces $w_i = 0$ y en consecuencia, en la estimación de $r(x)$ solo intervienen los y_i's para los cuales el correspondiente x_i dista de x menos que h. Por tanto, h juega un papel muy importante en la estimación: cuanto menor es h, menos valores intervienen en la estimación y cuanto mayor es, más valores intervienen. De ahí la denominación de h como "ventana".

Cabe mencionar que si K es el núcleo gaussiano, en la estimación $r_n(x)$ realmente intervienen todos los valores de y_i, pero el peso asociado a y_i decrece exponencialmente con el cuadrado de la distancia entre x_i y x.

Existe bibliografía sobre la regresión por núcleos función-función [19], [20]. Sin embargo, para el caso que nos ocupa, para la regresión escalar-función, no se ha encontrado nada. Se define, por analogía al caso de regresión función-función, el estimador de Nadaraya-Watson cuando la variable dependiente es una función y la variable independiente un escalar.

Definición 27 (Estimador funcional de Nadaraya-Watson)**.** *Sea $(x_1, \mathcal{Y}_1), ..., (x_n, \mathcal{Y}_n)$ un conjunto de observaciones de una variable real X y una variable funcional $\mathcal{Y}$. Sea $h > 0$. El estimador funcional de Nadaraya-Watson se define como:*

$$\hat{r}_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \mathcal{Y}_i}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}. \tag{38}$$

Para la consistencia del estimador de Nadaraya-Watson (37), se requiere que la función de regresión r sea continua. Extrapolando esto al caso de aplicación de este trabajo, con $\mathcal{Y}$ funcional, se impone la continuidad del funcional de regresión r (36) para que (38) sea consistente.

5.2.1. Elección del parámetro de suavizado

La elección del núcleo no es tan determinante como la elección del parámetro h. Primero se explica el procedimiento de elección para el caso en el que la variable Y es real y posteriormente se generaliza al caso en el que $\mathcal{Y}$ es una variable funcional.

Suponemos $r : \mathbb{R} \to \mathbb{R}$ una función de regresión. Si $\hat{r}_n(x)$ es la estimación de $r(x)$, el error cuadrático medio viene dado por:

$$R(h) = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_n(x_i) - r(x_i)\right)^2\right].$$

Por tanto, para obtener la mejor estimación de r se tendrá que escoger el parámetro h que minimice $R(h)$. Pero r es desconocida. Una estrategia para solucionar este problema es minimizar una estimación $\hat{R}(h)$ de $R(h)$. Se podría pensar en minimizar $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{r}_n(x_i))$, pero de esta manera se estarían utilizando los datos dos veces: para estimar la función y para estimar el $R(h)$, lo que podría resultar en un sobreajuste.

Una manera de solucionar esto es estimar $R(h)$ con la técnica *leave-one-out cross-validation*.

Definición 28 (Puntuación *leave-one-out cross-validation*)**.** *Se define la puntuación leave-one-out cross-validation como*

$$CV = \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{r}_{(-i)}(x_i)\right)^2, \tag{39}$$

donde $\hat{r}_{-i}$ es la estimación de r obtenida omitiendo el i-ésimo par, es decir, $\hat{r}_{-i}(x) = \sum_{j=1}^{n} w_{j,-i}(x)y_j$, donde

$$w_{j,-i}(x) = \begin{cases} 0 & \text{si } j = i \\ \frac{K\left(\frac{x-x_j}{h}\right)}{\sum_{k\neq i,k=1}^{n} K\left(\frac{x-x_k}{h}\right)} & \text{si } j \neq i \end{cases}.$$

Por tanto, se trata de escoger el h que minimice la puntuación *leave-one-out cross validation*.

El método *cross-validation* elimina el problema del sobreajuste puesto que ya no interviene y_i en la estimación de $r(x_i)$. La justificación de este método se encuentra en lo siguiente:

$$\begin{aligned}
E\left[\left(y_i - \hat{r}_{(-i)}(x_i)\right)^2\right] &= E\left[(y_i - r(x_i) + r(x_i) - \hat{r}_{(-i)}(x_i))^2\right] \\
&= E\left[(y_i - r(x_i))^2 + (r(x_i) - \hat{r}_{(-i)}(x_i))^2 + 2(y_i - r(x_i))(r(x_i) - \hat{r}_{(-i)}(x_i))\right] \\
&= Var(Y) + 2E\left[(y_i - r(x_i))(r(x_i) - \hat{r}_{(-i)}(x_i))\right] + E\left[(r(x_i) - \hat{r}_{(-i)}(x_i))^2\right] \\
&= Var(Y) + 2(r(x_i) - \hat{r}_{(-i)}(x_i))E\left[y_i - r(x_i)\right] + E\left[(r(x_i) - \hat{r}_{(-i)}(x_i))^2\right] \\
&= Var(Y) + 2(r(x_i) - \hat{r}_{(-i)}(x_i))(E[y_i] - E[r(x_i)]) + E\left[(r(x_i) - \hat{r}_{(-i)}(x_i))^2\right] \\
&= Var(Y) + E\left[(r(x_i) - \hat{r}_{(-i)}(x_i))^2\right] \\
&\approx Var(Y) + E\left[(r(x_i) - \hat{r}_n(x_i))^2\right].
\end{aligned}$$

Por tanto,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{r}_{(-i)}(x_i)\right)^2\right] \approx Var(Y) + E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_n(x_i) - r(x_i)\right)^2\right].$$

Luego,

$$E[\hat{R} - R] \approx Var(Y).$$

Por lo que la puntuación *leave-one-out cross-validation* es aproximadamente un estimador insesgado.

De nuevo, se generaliza esto a un espacio de funciones. Sea $(x_1, \mathcal{Y}_1), ..., (x_n, \mathcal{Y}_n)$ un conjunto de observaciones de una variable real X y una variable funcional $\mathcal{Y}$. (39) se transforma en:

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} d(\mathcal{Y}_i, \hat{r}_{(-i)}(x_i)), \tag{40}$$

donde $\hat{r}_{-i}$ es la estimación obtenida omitiendo el i-ésimo par, es decir,

$$\hat{r}_{-i}(x) = \sum_{j=1}^{n} w_{j,-i}(x)\mathcal{Y}_j, \tag{41}$$

donde

$$w_{j,-i}(x) = \begin{cases} 0 & \text{si } j = i \\ \frac{K\left(\frac{x-x_j}{h}\right)}{\sum_{k\neq i, k=1}^{n} K\left(\frac{x-x_k}{h}\right)} & \text{si } j \neq i \end{cases}. \tag{42}$$

En este trabajo, $\mathcal{Y}_1, ...\mathcal{Y}_n$ son funciones cuantil, por tanto distribuciones de probabilidad en el espacio de Wasserstein $\mathscr{W}_2(\mathbb{R})$, y por ello se utiliza la distancia de Wasserstein W_2^2. Así, (40) es:

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mathcal{Y}_i, \hat{r}_{(-i)}(x_i)). \tag{43}$$

5.3. Estimación de funciones de densidad

Sea $X_1, ..., X_n$ una muestra de una variable aleatoria real X con función de densidad f. Se presentan a continuación algunas técnicas para estimar f.

5.3.1. Histograma

El estimador clásico de la función de densidad es el histograma. La estimación de la función de densidad consiste en dividir el intervalo donde X toma valores en subintervalos de la misma longitud y contar el número de X_i's en cada subintervalo. Sin pérdida de generalidad, suponemos que la función de densidad cumple que $\{x : f(x) \neq 0\} = [0, 1]$. Sea $m \in \mathbb{Z}$ y se definen los intervalos como:

$$B_1 = \left[0, \frac{1}{m}\right), \; B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), ..., B_m = \left[\frac{m-1}{m}, 1\right].$$

Se definen el ancho de subintervalo como $h = 1/m$ y el número de observaciones en B_j como n_j.

De esta manera, dada la muestra $X_1, ..., X_n$, el estimador de la función densidad es:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{m} \frac{n_j}{nh} 1_{B_j}(x).$$

El principal problema del histograma es que la estimación $\hat{f}$ es una función escalonada aunque f sea continua.

5.3.2. Histograma móvil

Este método suaviza la estimación obtenida con el histograma. La diferencia con este reside en que en vez de dividir el intervalo en el que X toma valores en subintervalos fijos, se considera un ancho de subintervalo fijo, $2h$, y en la estimación de la función de densidad en un punto se toma el subintervalo de anchura $2h$ centrado en ese punto:

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} 1_{(x-h,x-h]}(X_i). \tag{44}$$

Esta estimación también es escalonad, pero es más suave que la obtenida con el histograma. Además, es el punto de partida de la estimación por núcleos de la función de densidad. Para ver esto, reescribimos (44) de la siguiente manera:

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} 1_{(x-h,x+h]}(X_i) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{2} 1_{(-1,1]} \left(\frac{x - X_i}{h} \right) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - X_i}{h} \right), \tag{45}$$

con $K\left(\frac{x-X_i}{h}\right) = \frac{1}{2} 1_{(-1,1]}\left(\frac{x-X_i}{h}\right)$.

5.3.3. Estimación por núcleos

Debido a que (45) sigue siendo continua a trozos, surge de manera intuitiva la idea sustituir K por una función continua. Esto da lugar a la estimación por núcleos. Así, la estimación por núcleos de la función de densidad f en un punto x viene dada por:

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left(\frac{x - X_i}{h} \right), \tag{46}$$

con K un núcleo como los definidos en la sección 5.2. El núcleo *box* da lugar al histograma móvil.

La elección del parámetro h se hace de la misma manera que la explicada en la sección 5.2.1.

Tomando una secuencia de parámetros h, $\{h_n\}$ tal que $h_n \to 0$ y asumiendo que f es continua, se tiene que $\hat{f}_{n,h}(x) \xrightarrow{c.s.} f(x)$ [16].

Teniendo en cuenta (46), se observa que (37) se puede escribir como:

$$\hat{r}_n(x) = \frac{\hat{\phi}_n(x)}{\hat{f}_{n,h}(x)}. \tag{47}$$

El numerador de (47) es una estimación por núcleos de $\phi(x) = \int y f(x,y) dy$, con $f(x,y)$ la distribución conjunta de (X,Y) y el denominador es una estimación de la densidad marginal $f(x)$ de X.

6. Aplicación

El Sol es la fuente de radiación que calienta la Tierra. Además del calentamiento solar directo de la superficie terrestre, también se produce un calentamiento indirecto debido a que parte de la radiación absorbida por la superficie terrestre es reemitida con longitudes de onda en el rango infrarrojo, absorbida por la atmósfera y reemitida en todas direcciones, tanto hacia fuera como hacia dentro de la atmósfera. Este proceso radiativo mantiene estable la temperatura del planeta y recibe el nombre de efecto invernadero. Fue descubierto por Joseph Fourier en 1924, comprobado experimentalmente por John Tyndall en 1863 y cuantificado por Svante Arrhenius en 1894 [24].

La reemisión de la radiación en la atmósfera se debe a las moléculas de los denominados gases de efecto invernadero: vapor de agua (H_2O), dióxido de carbono (CO_2), metano (CH_4), óxido de nitrógeno (N_2O), ozono (O_3) y los clorofluorocarburos (CFC). Los cambios en el nivel de vapor de agua se consideran una respuesta a los cambios en los demás gases de efecto invernadero.

A lo largo del tiempo, se ha mantenido un equilibrio entre la radiación recogida del sol y la radiación infrarroja reemitida al espacio: aproximadamente, la radiación recogida era igual a la reemitida. Sin embargo, en las últimas décadas este equilibrio se ha roto y la reemisión de radiación es menor que la absorción. Este desequilibrio se debe a los cambios en los niveles de gases de efecto invernadero, principalmente el CO_2 [24]. Como resultado del desequilibrio, se producen cambios en el sistema climático, entre ellos, el aumento de la temperatura media de la superficie terrestre.

Dos proyectos, el Proyecto Europeo de Extracción de Hielo en la Antártida (EPICA) y el proyecto colaborativo de perforación de hielo entre Rusia, Estados Unidos y Francia en la estación rusa Vostok, han permitido reconstruir las concentraciones de CO_2 en la atmósfera desde hace unos 800.000 años. Estos datos se representan en la figura 1. Se pueden observar periodos regulares de descenso y aumento de las concentraciones de CO_2, donde el inicio de los periodos de bajada coincide con el comienzo de periodos de glaciaciones [7]. Sin embargo, en los últimos años se ha dado el incremento más brusco de la serie, periodo en el que se han alcanzado concentraciones récord. Este aumento comenzó con la Revolución Industrial y es debido a la quema de combustibles fósiles.

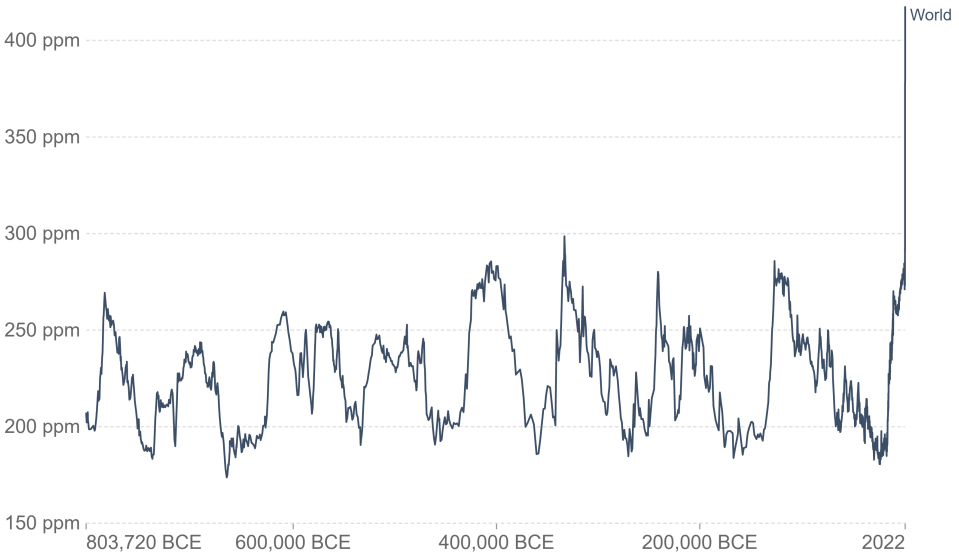

Figura 1: Concentración de dióxido de carbono (en partes por millón: número de moléculas de CO_2 en un millón de moléculas de aire) en la atmósfera a lo largo de los años. Imagen tomada de [28].

Es, por tanto, de interés, estudiar la relación entre la temperatura media en la superficie terrestre y nivel de CO_2 en la atmósfera.

Todas las representaciones y cálculos que se presentan a continuación han sido realizados con R.

6.1. Presentación de los datos

En este trabajo se dispone de dos conjuntos de datos, uno relacionado con temperaturas medias diarias de la superficie del planeta y otro con niveles de CO_2 en la atmósfera.

Los datos de temperaturas se han obtenido de la base de datos Berkeley Earth daily TAVG [8], de la organización Berkeley Earth, una organización estadounidense independiente, sin ánimo de lucro, centrada en la ciencia de datos ambientales. Esta base de datos contiene las diferencias entre la temperatura media diaria de la superficie terrestre y la temperatura media de la superficie terrestre entre enero de 1951 y diciembre de 1980 (8.90 ± 0.06 °C), entre 1880 y 2022.

Se asocia una distribución de probabilidad a las diferencias de temperatura de cada año. Para ello, para cada año y se construye la distribución de probabilidad $\mathbb{P}_T^y$, cuya función de distribución se define de la siguiente manera:

$$F_T^y(t) = \frac{\sharp\{d \in D_y : T(d) \leq t\}}{\sharp D_y},\ t \in \mathbb{R} \tag{48}$$

donde D_y es el conjunto de días del año y y $T(d)$ es la diferencia entre la temperatura media de la superficie terrestre del día d y la temperatura media entre enero de 1951 y diciembre de 1980. Por tanto, $F_T^y(t)$ es la proporción de días del año y con una diferencia de temperatura menor que t.

Por otro lado, se dispone de las concentraciones de CO_2 en la atmósfera desde 1954 hasta 2021, en partes por millón (ppm). Estas han sido descargadas de la publicación en línea Our World in Data [28].

Se utilizan los datos del periodo de 1954 a 2021 por ser el intervalo mayor del que se disponen de datos de las dos variables.

Para representar los datos se ha empleado la función *geom_density_ridges_gradient*. Esta función permite representar varias funciones de densidad. Las estimaciones de las funciones de densidad se realizan estimando valores de cada función de densidad mediante estimación por núcleos (ver sección 5.3). El núcleo empleado por defecto es el núcleo gaussiano. Para la representación gráfica de las funciones de densidad, la función *geom_density_ridges_gradient* interpola linealmente los valores estimados. Esta es la representación utilizada en la figura 2: en el eje x se representan las diferencias, en el eje z los valores de la función de densidad y en el eje y los años. Se ha establecido una escala de colores con las concentraciones de CO_2 de manera que se ha coloreado cada distribución con el color correspondiente al nivel de CO_2 de ese año.

Cabe mencionar que las distribuciones con las que se trabaja no tienen función de densidad porque están soportadas en, como mucho, 366 puntos. Por lo tanto, para cada año, la función de densidad que se representa en la figura 2 es de un suavizado de la distribución con función de densidad dada por (48).

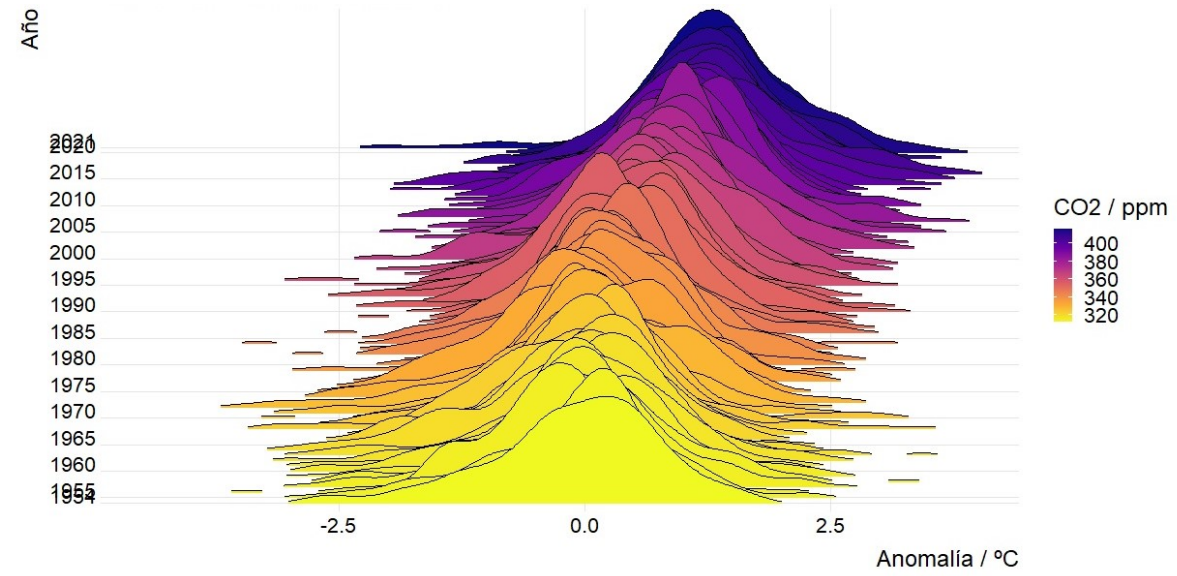

Figura 2: Funciones de densidad de las distribuciones de diferencias entre la temperatura media diaria de la superficie terrestre y la temperatura media entre enero de 1951 y diciembre de 1980, de los años entre 1954 y 2021. Las funciones de distribución de estas probabilidades vienen dadas por (48). Las distribuciones están coloreadas usando una escala de colores determinada por las concentraciones de CO_2, de manera que el color de cada distribución se corresponde con el color asociado a la concentración de CO_2 de dicho año.

En la figura 3 se representan las concentraciones de CO_2 de cada año. Se aprecia claramente que la concentración de CO_2 ha ido aumentando todos los años en el intervalo de tiempo en el que se tienen datos.

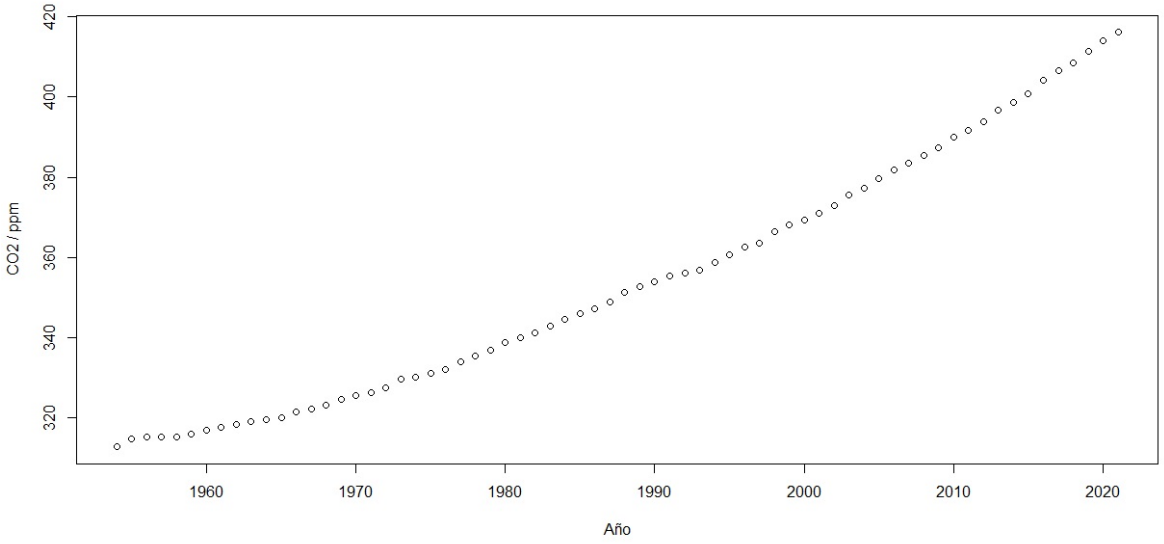

Figura 3: Concentración de CO_2 en la atmósfera de cada año, desde 1954 hasta 2021.

6.2. Modelo de regresión no paramétrica

Para estudiar la relación entre las distribuciones de diferencias en la temperatura y las concentraciones de CO_2, se ha planteado el siguiente modelo de regresión no paramétrica.

Sea $Y = \{1954, ..., 2021\}$ el conjunto de años de los que se disponen datos. Sea $\mathbb{P}_T : Y \longrightarrow \mathscr{W}_2(\mathbb{R})$ la variable aleatoria que a cada año le asigna la distribución de diferencias de temperatura de ese año (definida en (48)). Se denomina $\mathbb{P}_T^y := \mathbb{P}_T(y)$. Sea c_y la concentración de CO_2 en la atmósfera del año y. Suponemos que la concentración de CO_2 está representada por una variable aleatoria $C : Y \to \mathbb{R}$ tal que $c_y := C(y)$. Sabemos que el principal factor que afecta a la temperatura de la superficie es la concentración de CO_2. Por tanto, se plantea el modelo

$$\mathbb{P}_T = r(C) + \varepsilon, \tag{49}$$

donde $r : \mathbb{R} \to \mathscr{W}_2(\mathbb{R})$ es la denominada función de regresión y ε es una variable aleatoria en $\mathscr{W}_2(\mathbb{R})$, independiente de C y con $E[\varepsilon] = 0$. Por lo tanto, en el año y,

$$\mathbb{P}_T^y = \mathbb{P}_T^{c_y} + \varepsilon,$$

donde $\mathbb{P}_T^{c_y} = r(c_y)$. Por construcción, se tiene que $\mathbb{P}_T^{c_y} = E[\mathbb{P}_T^y / C = c_y] \in \mathscr{W}_2(\mathbb{R})$. Se pretenden estimar las curvas $\mathbb{P}_T^{c_y}$, es decir, las curvas anuales de las diferencias de temperatura esperadas en función de la concentración de CO_2 en la atmósfera.

Puesto que $\mathbb{P}_T^{c_y}$ es una distribución de probabilidad, se puede hacer la identificación $\mathbb{P}_T^{c_y} \longleftrightarrow F^{-1}_{\mathbb{P}_T^{c_y}}$ y, en consecuencia, es suficiente estimar las funciones cuantiles $F^{-1}_{\mathbb{P}_T^{c_y}}$ para cada $y \in Y$.

6.2.1. Estimaciones de las funciones cuantiles

Suponemos que la función $r : \mathbb{R} \to \mathscr{W}_2(\mathbb{R})$ es continua considerando en $\mathbb{R}$ la métrica usual y en $\mathscr{W}_2(\mathbb{R})$ la convergencia débil. Como la aplicación $\mathbb{P} \to F^{-1}_{\mathbb{P}}$ es continua con respecto a la convergencia débil, resulta que la aplicación $c \to P_{\mathbb{P}_T^c}$ lo es también. Así queda justificado el uso de estimadores núcleo.

Se dispone de los datos $\mathbb{P}_T^{1954}, ..., \mathbb{P}_T^{2021}$. Para estimar $F^{-1}_{\mathbb{P}_T^{c_y}}$ se emplean esos datos y el estimador de Nadaraya-Watson (38):

$$\widehat{F^{-1}_{\mathbb{P}_T^{c_y}}}(p) = \frac{\sum_{j=1954}^{2021} K\left(\frac{|c_y - c_j|}{h}\right) F^{-1}_{\mathbb{P}_T^j}(p)}{\sum_{j=1954}^{2021} K\left(\frac{|c_y - c_j|}{h}\right)}, \quad p \in [0,1].$$

En la aplicación práctica del procedimiento se han calculado 250 cuantiles de cada año, es decir, $\widehat{F^{-1}_{\mathbb{P}_T^{c_y}}}(p)$ para 250 valores de p y en las representaciones gráficas se ha realizado una interpolación lineal de estos valores.

Siguiendo una práctica habitual, los 250 valores de p se han tomado equiespaciados y entre $1/251$ y $250/251$ para no calcular los cuantiles extremos 0 y 1, pues se comete más error.

6.2.2. Elección del parámetro h

Para la elección del parámetro h, se ha tomado aquel valor que minimiza la puntuación *leave one out cross-validation* (43). En concreto, teniendo en cuenta la expresión de W_2 en $\mathbb{R}$ (33), se ha tomado el h que minimiza la cantidad

$$\sum_{y=1954}^{2021} W_2^2\left(\mathbb{P}_T^y, \widehat{\mathbb{P}_{T_{-y}}^{c_y}}\right) = \sum_{y=1954}^{2021}\left(\int_0^1 |F_{\mathbb{P}_T^y}^{-1}(p) - \widehat{F_{\mathbb{P}_{T_{-y}}^{c_y}}^{-1}}(p)|^2 dp\right), \tag{50}$$

donde, teniendo en cuenta (41) y (42), para el año y, $\widehat{F_{\mathbb{P}_{T_{-y}}^{c_y}}^{-1}}$ tiene la expresión:

$$\widehat{F_{\mathbb{P}_{T_{-y}}^{c_y}}^{-1}}(p) = \frac{\sum_{j\neq y, j=1954}^{2021} K\left(\frac{|c_y - c_j|}{h}\right) F_{\mathbb{P}_T^j}^{-1}(p)}{\sum_{k\neq y, k=1954}^{2021} K\left(\frac{|c_y - c_k|}{h}\right)}, \quad p \in [0,1].$$

En la práctica, como se ha explicado anteriormente, se ha trabajado con 250 valores de cada $F_{\mathbb{P}_T^y}^{-1}$ y $\widehat{F_{\mathbb{P}_{T_{-y}}^{c_y}}^{-1}}$. Se denota por $P = \{\frac{i}{251} : i = 1, \cdots, 250\}$ el conjunto de órdenes de cuantiles empleados. Por tanto, las integrales de (50) se aproximan por el siguiente sumatorio:

$$\sum_{p\in P} |F_{\mathbb{P}_T^y}^{-1}(p) - \widehat{F_{\mathbb{P}_{T_{-y}}^{c_y}}^{-1}}(p)|^2 \frac{1}{251}.$$

Por tanto, h se ha escogido como aquel que minimiza la siguiente cantidad (se omite el factor $1/251$ ya que no afecta a la minimización):

$$\sum_{y=1954}^{2021}\left(\sum_{p\in P} |F_{\mathbb{P}_T^y}^{-1}(p) - \widehat{F_{\mathbb{P}_{T_{-y}}^{c_y}}^{-1}}(p)|^2\right). \tag{51}$$

A continuación se explica cómo se han escogido los valores de h para los cuales se ha calculado (51).

Se han calculado las distancias entre las concentraciones de CO_2 de todos los años, $d_{ij} = |c_i - c_j|$, $i = 1954, ..., 2021$, $j = 1954, ..., 2021$ e $i \neq j$. Se ha definido $h_m = \text{máx}_i(\text{mín}_{j\neq i}\, d_{ij})$ y h_M a la mitad del rango de los valores d_{ij}. Se han tomado 10 valores equidistantes entre h_m y h_M.

El sentido de esta manera de elegir h se entiende fácilmente con el núcleo *box*. Con este núcleo, solo participan en la estimación de $F_{\mathbb{P}_T^{c_y}}$ los años cuya concentración de CO_2 dista de c_y menos que el valor de la ventana h. Así pues, con esta manera de elegir h, se asegura que intervengan entre uno, si $h = h_m$, y alrededor de la mitad de los años disponibles, si $h = h_M$, en la estimación de cada $F_{\mathbb{P}_T^{c_y}}^{-1}(p)$.

6.3. Resultados

Para comenzar, se presentan los resultados de la elección del parámetro h.

Con los datos de CO_2 disponibles, los valores de h_m y h_M han sido: $h_m = 2.526$ y $h_M = 51.689$. En la figura 4 se representa el valor de la expresión (51) para cada valor candidato de h.

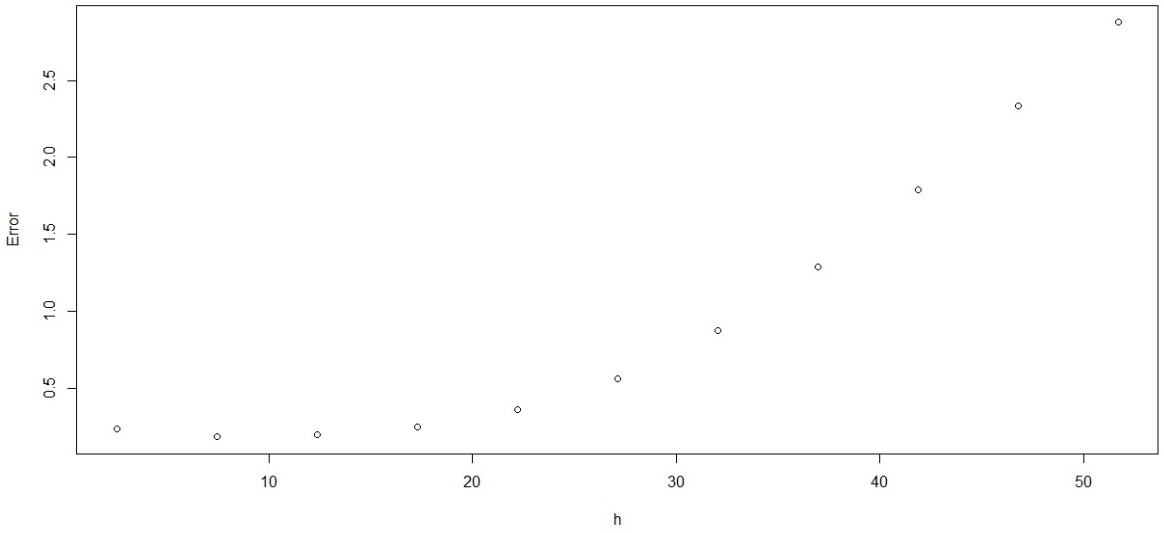

Figura 4: Valor del error (50) para cada valor de h.

Como se puede observar en la figura 4, el valor óptimo para h es el segundo más pequeño, que se corresponde con $h = 7.442$. Este valor es el que se ha empleado para las estimaciones de los valores de las funciones cuantil

Se muestran en la figura 5 las funciones cuantil $\hat{F}^{-1}_{\mathbb{P}^{c_y}_T}$ con $y = 1954, ..., 2021$. Se ha empleado la función *geom_line* que interpola linealmente los puntos.

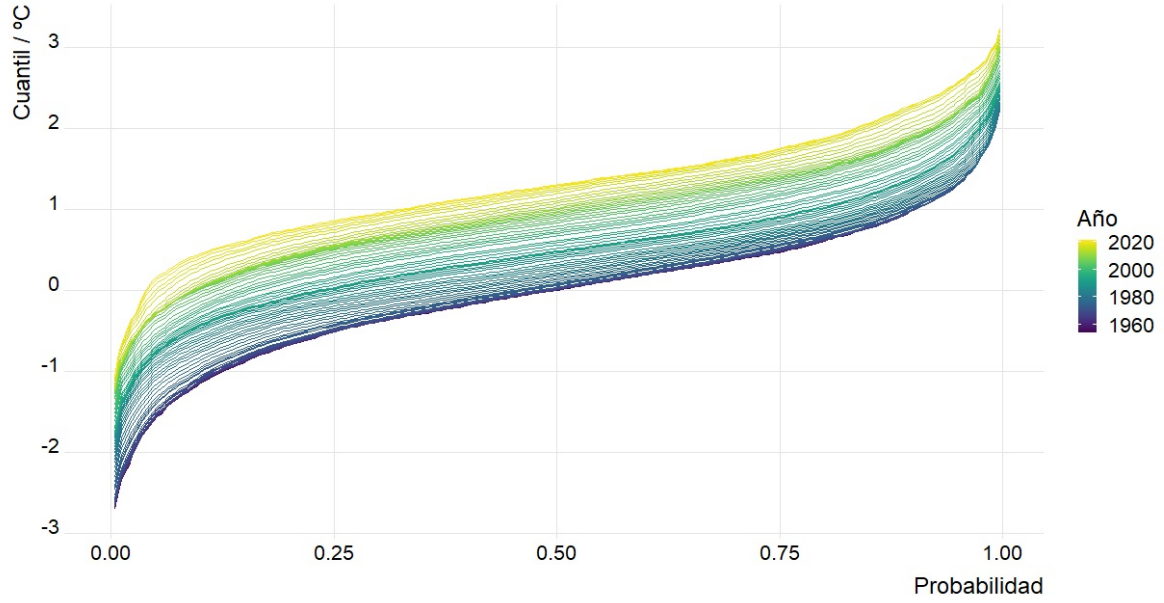

Figura 5: Estimaciones de las funciones cuantil $\hat{F}^{-1}_{\mathbb{P}^{c_y}_T}$ para cada año, desde 1954 a 2021.

Para entender mejor el comportamiento de las funciones cuantil con el paso del tiempo y, por tanto, con el aumento de la concentración de CO_2 en la atmósfera, se han representado en la figura 6 los percentiles 5, 10, 25, 50, 75 y 90 a lo largo de los años. El percentil $100p$ hace referencia a la diferencia

de temperatura t tal que la proporción de días del año y en el que la diferencia de temperatura (de la distribución condicionada al nivel de CO_2 de ese año) es menor o igual que t es p.

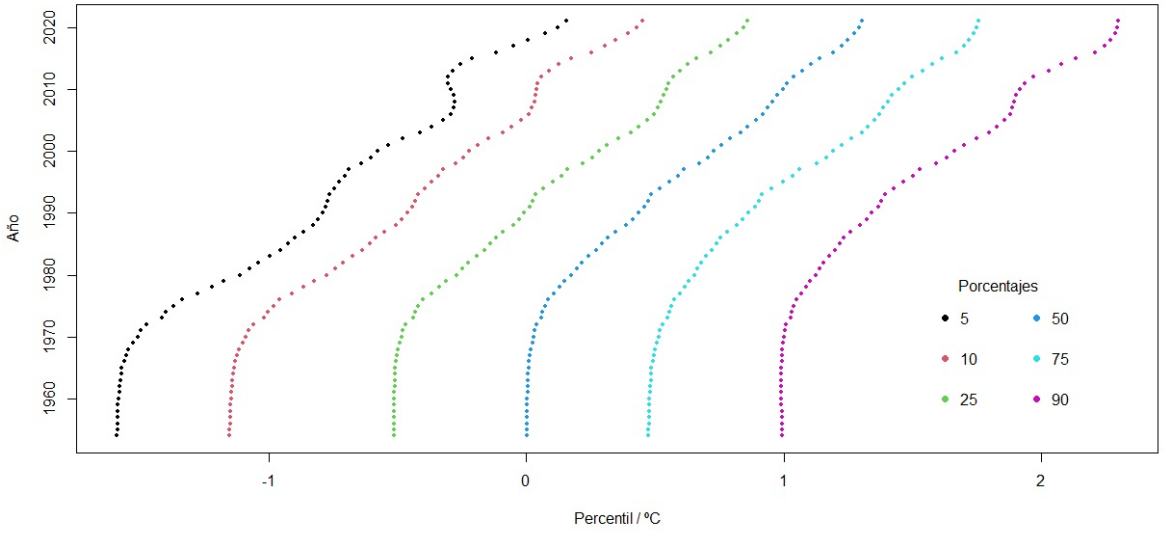

Figura 6: Representación gráfica de los percentiles 5, 10, 25, 50, 75 y 90 a lo largo de los años.

En la figura 6 se observa que el ritmo de desplazamiento hacia diferencias mayores hasta aproximadamente 1970 es muy pequeño, manteniéndose las temperaturas prácticamente constantes. Sin embargo, a partir de 1970 se observa un claro desplazamiento hacia diferencias mayores. Únicamente el percentil 5 entre los años 2008 y 2011 decrece.

Para los percentiles extremos se observan, a partir de 1970, cambios de ritmo en el desplazamiento hacia diferencias mayores. Por ejemplo, el ritmo de desplazamiento del percentil 5 es aproximadamente constante entre 1980 y 1990 y más alto que entre 1990 y 1994. También se observa un clara disminución del ritmo de desplazamiento del percentil 10 entre 2005 y 2010. Sin embargo, en los percentiles más centrales, estos cambios de ritmos son prácticamente despreciables. También se observa que en los últimos años los percentiles más altos (75 y 90) se han mantenido aproximadamente constantes, mientras que los demás percentiles han seguido desplazándose hacia valores mayores. Esto significa, por tanto, que los días más cálidos han sido igual de cálidos en los últimos años.

6.4. Conclusiones

Para concluir, se exponen una serie de reflexiones acerca del modelo propuesto y de los resultados obtenidos.

En la figura 3 se observa que la concentración de CO_2 en la atmósfera ha ido aumentando paulatinamente en los últimos años. Además, en la figura 1 se puede comprobar que el aumento en el periodo considerado ha sido más rápido que nunca. En la figura 2 se observa un desplazamiento de las distribuciones hacia diferencias, y por tanto, temperaturas mayores. Así pues, es razonable plantear la hipótesis de que la distribución de temperaturas de un año dependa de la concentración de CO_2 en la atmósfera de dicho año, y con ello un modelo de regresión como (49).

En la figura 6 se han representado ciertos percentiles que permiten estudiar el comportamiento a lo largo del tiempo de las distribuciones de temperatura condicionadas al nivel de CO_2. Se observa un desplazamiento de todos los percentiles hacia temperaturas mayores con el paso del tiempo, salvo del percentil 5, entre 2008 y 2011. También se observan cambios de ritmo en los desplazamientos, siendo estos mucho más acusados en los percentiles extremos, como se ha comentado previamente.

La estimación de percentiles extremos, es decir, más cercanos a 0 y 1, presenta más error, por lo que hay que ser cautelosos a la hora de extraer conclusiones a partir de estos. Así, nos centramos en los percentiles centrales. Fijándose en ellos, se pueden extraer dos conclusiones claras: un aumento en la concentración de CO_2 en la atmósfera se traduce en un aumento de la temperatura y un aumento en el ritmo de acumulación de CO_2 en la atmósfera se traduce en un aumento del ritmo del desplazamiento de las distribuciones de temperatura hacia temperaturas mayores.

La primera de las conclusiones se deduce de que se observa que el CO_2 ha aumentado a lo largo de los años entre 1954 y 2021 y las estimaciones de las funciones cuantil de las distribuciones de diferencias de temperaturas, luego de las distribuciones de temperaturas de cada año, condicionadas al nivel de CO_2, se han desplazado hacia diferencias mayores. A medida que aumenta la concentración de CO_2, las funciones cuantil de la figura 5 se desplazan hacia temperaturas mayores. Esto supone que si $t \in \mathbb{R}$ y un año la proporción de días con una diferencia menor que t es p, el año siguiente, la diferencia asociada a la misma proporción p será mayor. Por tanto, las distribuciones de temperatura $\mathbb{P}_T^{c_y}$ se desplazan hacia diferencias mayores y la función r se puede decir "creciente" en el sentido de que, con el aumento de CO_2 a lo largo del tiempo, las funciones cuantil se desplazan hacia temperaturas mayores.

La segunda conclusión se obtiene del hecho que hasta 1970 el ritmo de aumento de la concentración de CO_2 en la atmósfera era mucho menor que a partir de este año y en la figura 6 se observa también este cambio de ritmo en el desplazamiento de las distribuciones de temperatura hacia diferencias mayores. Sin embargo, para un estudio riguroso del ritmo de desplazamiento de las distribuciones, se tendría que trabajar con derivadas. Una línea de trabajo futura podría ser esta: cómo afecta el cambio de ritmo de acumulación de CO_2 al ritmo de desplazamiento de las distribuciones de temperatura.

La naturaleza no paramétrica del modelo impide su uso para realizar estimaciones sobre las distribuciones de probabilidad de las diferencias en puntos fuera del rango de los datos, es decir, en años fuera del rango de 1954 a 2021. Aunque queda pendiente para un posible futuro trabajo, este modelo, junto con el *bootstrap*, permite contrastar la hipótesis nula de no-calentamiento o dar conjuntos de confianza para las distribuciones de diferencias de temperaturas para niveles de CO_2 dentro del rango de los valores disponibles.

Por último, usualmente se evalúa el crecimiento de las temperaturas utilizando un parámetro como puede ser la media. Sin embargo, la metodología utilizada, junto con las ideas desarrolladas en [15], permite analizar la evolución de todas las temperaturas del año. Esta sería otra posible vía de continuación del trabajo.

En resumen, un modelo simple como es el de regresión, ha permitido observar claramente la relación entre la temperatura y la concentración CO_2 en la atmósfera.

Referencias

[1] Agueh, M., & Carlier, G. (2011). Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, *43*(2), 904-924.https://doi.org/10.1137/100805741

[2] Alegría, P. (s. f.). *Teoría de la medida*. https://www.ehu.eus/~mtpalezp/mundo/teomed/apuntes

[3] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J., & Matrán, C. (2016). A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, *441*(2), 744–762. https://doi.org/10.1016/j.jmaa.2016.04.045

[4] Ambrosio, L., Bressan, A., & Helbing, D. (2009). A User's Guide to Optimal Transport. En *Modelling and Optimisation of Flows on Networks* (pp. 1–156). Springer. https://doi.org/10.1007/978-3-642-32160-3

[5] Arutyunov, A. V., & Obukhovskii, V. (2016). *Convex and Set-Valued Analysis*. De Gruyter. https://doi.org/10.1515/9783110460308

[6] Bauer, A., Scheipl, F., Küchenhoff, H.,& Gabriel, A. A. (2018). An introduction to semiparametric function-on-scalar regression. *Statistical Modelling*, *18*(3–4), 346–364. https://doi.org/10.1177/1471082x17748034

[7] Bereiter, B., Eggleston, S., Schmitt, J., Nehrbass-Ahles, C., Stocker, T. F., Fischer, H., Kipfstuhl, S., & Chappellaz, J. (2015). Revision of the EPICA Dome C CO2 record from 800 to 600 kyr before present. *Geophysical Research Letters*, *42*(2), 542–549. https://doi.org/10.1002/2014gl061957

[8] Berkeley Earth. (1880–2022). *Berkeley Earth daily TAVG* [Conjunto de datos]. Berkeley Earth. Recuperado 1 de agosto de 2022, de http://berkeleyearth.lbl.gov/auto/Global/Complete_TAVG_daily.txt

[9] Billingsley, P. (1995). *Probability and Measure* (3 ed.). Wiley-Interscience.

[10] Billingsley, P. (1999). *Convergence of Probability Measures* (2.a ed.). Wiley-Interscience. https://doi.org/10.1002/9780470316962

[11] Boyd, S., Duchi, J., Pilanci, M., & Vandenberghe, L. (2022). *Subgradients*. Standford University. Recuperado 15 de agosto de 2022, de https://web.stanford.edu/class/ee364b/lectures/subgradients_notes.pdf

[12] Chehebar, N. G. (2021). *Transporte Óptimo y Baricentro con distancia de Fermat*. [Tesis de licenciatura]. Universidad de Buenos Aires.

[13] Chen, X., Li, H., Liang, H., & Lin, X. (2018). Functional response regression analysis. *Journal of Multivariate Analysis*, *169*, 218-233. https://doi.org/10.1016/j.jmva.2018.09.009

[14] Correa, R., Jofré, A., & Thibault, L. (1995). Subdifferential characterization of convexity. *Recent advances in nonsmooth optimization*, 18-23.

[15] del Barrio, E., Cuesta-Albertos, J. A., & Matran, C. (pendiente de publicar). Invariant measures of disagreement with stochastic dominance. https://arxiv.org/abs/1804.02905

[16] Einmahl, U., & Mason, D. M. (2005, 1 junio). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, *33*(3). https://doi.org/10.1214/009053605000000129

[17] Faraway, J. J. (1997). Regression Analysis for a Functional Response. *Technometrics*, *39*(3), 254–261. https://doi.org/10.1080/00401706.1997.10485118

[18] Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer. https://doi.org/10.1007/0-387-36620-2

[19] Ferraty, F., Goia, A., & Vieu, P. (2002). Régression non-paramétrique pour des variables aléatoires fonctionnelles mélangeantes. *Comptes Rendus Mathematique*, *334*(3), 217–220. https://doi.org/10.1016/s1631-073x(02)02248-3

[20] Ferraty, F., Laksaci, A., Tadj, A., & Vieu, P. (2011). Kernel regression with functional response. *Electronic Journal of Statistics*, *5*. https://doi.org/10.1214/11-ejs600

[21] Frungillo, M. O. (2016). *Transporte Óptimo y Aplicaciones*. [Tesis de licenciatura]. Universidad de Buenos Aires.

[22] Gangbo, W., & Swiech, A. (1998). Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on Pure and Applied Mathematics*, *51*(1), 23–45. https://www.math.ucla.edu/~wgangbo/publications/mmk11.pdf

[23] Khatskevich, V., & Shoiykhet, D. (2012). Preliminaries. En *Differentiable Operators and Nonlinear Equations* (Vol. 66, pp. 1–34). Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-6-8512

[24] Lacis, A. A., Schmidt, G. A., Rind, D., & Ruedy, R. A. (2010). Atmospheric CO2: Principal Control Knob Governing Earth's Temperature. *Science*, *330*(6002), 356–359. https://doi.org/10.1126/science.1190653

[25] Lemus-Delgado, D., & Pérez Navarro, R. (2020). Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos. *Colombia Internacional*, *102*, 41–62. https://doi.org/10.7440/colombiaint102.2020.03

[26] Panaretos, V. M., & Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer Publishing. https://doi.org/10.1007/978-3-030-38438-8

[27] Panaretos, V. M., & Zemel, Y. (2019). Statistical Aspects of Wasserstein Distances. *Annual review of statistics and its application*, *6*, 405-431. https://doi.org/10.1146/annurev-statistics-030718-104938

[28] Ritchie, H. (1954–2021). *CO2 and GHG Emissions* [Conjunto de datos]. Our World In Data. Recuperado 1 de agosto de 2022, de https://ourworldindata.org/explorers/climate-change?facet=none&hideControls=true&Metric=CO%E2%82%82+concentrations&Long-run+series%3F=true&country=~OWID_WRL

[29] Ritchie, H., & Roser, M. (2020). *Global Atmospheric CO2 Concentration* [Gráfico]. Our Worl in Data. https://ourworldindata.org/explorers/climate-change?facet=none&hideControls=true&Metric=CO%E2%82%82+concentrations&Long-run+series%3F=true&country=~OWID_WRL

[30] Sam, S. V. (2022). *Solutions to Real and Complex Analysis*. dokumen.tips. Recuperado 20 de agosto de 2022, de https://dokumen.tips/documents/solutions-to-real-and-complex-analysis-by-walter-rudin-mathematic87blogfacom-5659ca27e1cae.html?page=1

[31] Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, Pdes, and Modeling*: *87* (Vol. 87). Birkhauser. https://doi.org/10.1007/978-3-319-20828-2

[32] Schilling, R. L. (2005). *Measures, Integrals and Martingales* (1.a ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511810886

[33] Sheather, S. J. (2004). Density estimation. *Statistical Science*, *19*(4), 588–597. https://www.jstor.org/stable/4144429

[34] Silverman, B. W. (1998). *Density estimation for statistics and data analysis* (1.a ed.). Routledge. https://doi.org/10.1201/9781315140919

[35] Villani, C. (2021). *Topics in Optimal Transportation* (Vol. 58). American Mathematical Society.

[36] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer. https://doi.org/10.1007/0-387-30623-4