FACULTAD
DE
CIENCIAS

Aplicación de técnicas de Análisis Cluster funcional para determinación de agrupamientos de pirámides de población

(Application of functional Cluster Analysis techniques to determinate groups of population pyramids)

Trabajo de fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autora: Julia Ruiz Salmón

Director: Juan Antonio Cuesta Albertos

Febrero-2017

Índice general

1 Resumen

1.1. Resumen

A través de este trabajo se pretende aplicar la técnica de las k-medias recortadas a pirámides de población, o más concretamente, a aproximaciones continuas de pirámides de población.

Con este fin, se intentan resolver dos cuestiones importantes relacionadas, por un lado, con el procedimiento de k-medias recortadas y, por el otro, con las pirámides de población. En el primer caso, se necesita extender el método de las k-medias recortadas desde $\Re^p$, espacio para el que fue desarrollada y donde normalmente se aplica, a un espacio funcional apropiado. En el segundo caso, se pretende obtener curvas suaves a partir de datos discretos, o dicho de otra forma, encontrar curvas suaves utilizando las pirámides de población como si fueran muestras (de tamaño del orden de millones o centenares de miles) obtenidas de funciones de densidad que hay que estimar. Para la estimación se emplea el estimador de Nadaraya-Watson.

Finalmente, el objetivo de este trabajo es aplicar los resultados anteriores para realizar una clasificación en grupos homogéneos de los países de América considerando sus pirámides de población. Con este fin, es necesario determinar el número de grupos. Para ello, se pretende utilizar una técnica (ver [1]) que permite realizar una clasificación de los datos disponibles en grupos pero sin tener en cuenta un cierto número de datos que producen anomalías en los resultados finales.

Además, dentro del objetivo final, se va a realizar una comparación de los resultados precisamente con los obtenidos en [1] donde el procedimiento para agrupar que utilizan está basado en el uso de funciones cuantiles asociadas a las pirámides y sin suavidad. Es por ello, que los resultados no tienen por qué ser iguales en ambos casos.

(El código utilizado en este trabajo está en el fichero `prueba_casopractico.R` del que se hizo entrega copia en la Secretaría de la Facultad de Ciencias).

1.2. Abstract

Through this work we intend to apply the trimmed k-means's technique to population pyramids, or particulary, to continuous approximations of population pyramids.

For this purpose, we try to solve two important questions related, on the one hand, to the trimmed k-means's procedure and, on the other hand, to population pyramids. First , it is necessary to extend the method of the trimmed k-means from $\Re^p$, space for which it was developed and where it is normally applied, to an appropriate functional space. Secondly, we intend to obtain smooth curves from discrete data, that is, to find smooth curves using the population pyramids like samples (of size of the order of millions or hundreds of thousands) which are obtained from density functions to be estimated. For this estimation, the Nadaraya-Watson estimator is used.

Finally, the objective of this work is to apply the previous results to make a classification in homogeneous groups of the America's countries considering their pyramids of population. For this purpose, it is necessary to determine the number of groups. To do this, we intend to use a technique (see [1]) that allows a classification of the data in groups but without considering a certain number of data that produce anomalies in the final results.

In addition, one of the final objectives is to make a comparison of our results with those obtained in [1]. They use a grouping procedure which is based on the use of quantile functions associated with the pyramids and without having any smoothness. That is the reason why the results don't have to be the same in both cases.

2

Introducción: Análisis Cluster

A lo largo de los años, el significado de la palabra 'Estadística' ha ido variando. Esta disciplina de las matemáticas comienza alrededor de 1749 y desde entonces se le han atribuido diferentes interpretaciones. En primer lugar, su significado fue el de dar información sobre los estados (en el ámbito de la política). Posteriormente, incluía dar información de cualquier tipo. Por último, y en la actualidad, para el análisis e interpretación de datos (ver [2]).

En esta última interpretación de la Estadística es en la que se incluye nuestro trabajo. En concreto, en la aplicación de un análisis, denominado Análisis Cluster, para conseguir agrupar datos y poder interpretar los resultados según los grupos obtenidos.

En la práctica, este análisis será utilizado para un problema concreto: la repartición en grupos homogéneos de diferentes países teniendo en cuenta la estructura de edad de su población, llamada pirámide de población.

La dificultad que presenta este análisis reside en que no existe un procedimiento que sea 100 % eficaz a la hora de buscar grupos homogéneos. De hecho, cuando se busca repartir un conjunto de datos en diferentes grupos, se hace a partir de uno o varios factores (en este caso será a partir de la edad de población). Estos factores, a su vez, pueden estar influenciados por las diversas interpretaciones que puedan tener. Por ejemplo, una persona podría ser considerada mayor a partir de los 60 años, o bien, a partir de los 70. Es decir, según el criterio o interpretación que se tenga, una persona de 60 años podría pertenecer a uno u otro grupo (en este caso, al grupo de mayor edad o a uno intermedio).

Un ejemplo claro que demuestra la inexistencia de un método totalmente eficaz para hacer agrupaciones es el caso de las tallas de camisetas: las camisetas se agrupan en seis tallas diferentes; XS, S, M, L, XL y XXL. Sin embargo, esas tallas tienen un alto grado de arbitrariedad. La talla S de una camiseta de una marca puede ser equivalente a la de la talla M de otra marca. De este modo, es imposible encontrar un método universal para buscar grupos homogéneos que resuelva este tipo de 'inconvenientes' que se plantean a la hora de realizar una agrupación de datos (de países en este caso y de personas en el caso de las tallas de camiseta).

2.1. Análisis Cluster

El Análisis Cluster es un procedimiento matemático para buscar agrupaciones homogéneas entre conjuntos de datos. En general, el Análisis Cluster consiste en dividir una muestra de n datos $\{x_1, \ldots, x_n\}$ en un cierto número de grupos. El criterio que se sigue para hacer las agrupaciones es intentar que datos similares entre sí pertenezcan al mismo grupo y datos diferentes entre sí se integren en diferentes grupos.

Existen dos familias de procedimientos para formar estas agrupaciones: por un lado las **clasificaciones jerárquicas**, que utilizan agrupaciones progresivas (o aglomerativas) y agrupaciones divisivas y, por otro lado, el método de las **k-medias**. Este último procedimiento utiliza una clasificación no jerárquica para agrupar un conjunto de datos en k grupos utilizando la media de esos grupos. La elección de k no está sujeta a ningún criterio, pero en este trabajo se aplica un procedimiento, que se verá posteriormente, con el que se intentará llegar a un número k de agrupaciones adecuado para cada problema. Además, no se utilizarán las k-medias clásicas, sino una variante denominada **k-medias recortadas**.

Este método será puesto en práctica para clasificar y repartir en k grupos, según la estructura de edad de la población, 36 países de América. Con ello el método de las k-medias tratará de darnos una descripción de las diferentes poblaciones de los países de América utilizando como objetos **estimadores de una función de densidad** desconocida.

A continuación se describen los procedimientos jerárquicos, dejando las k-medias para el siguiente capítulo.

Clasificación jerárquica

No vamos a entrar en los detalles de estos métodos de clasificación porque no se utilizan en este trabajo. No obstante, se indica que la clasificación jerárquica puede basarse en agrupaciones progresivas o divisivas. Además, hay que mencionar que para el cálculo de ambas agrupaciones existen programas con procedimientos muy rápidos. Sin embargo, las agrupaciones divisivas son menos utilizadas que las aglomerativas por su gran carga computacional. En el caso de las agrupaciones divisivas, únicamente en el primer paso ya se necesita tener en cuenta $2^{(n-1)} - 1$ posibles particiones de la muestra. Mientras que en las agrupaciones aglomerativas, en el paso inicial, sólo se necesita tener en consideración las posibles particiones en subconjuntos de dos elementos, $\frac{n(n-1)}{2}$ (puede verse información adicional de estos procedimientos en [3]).

Los procedimientos se describen grosso modo a continuación:

Agrupaciones progresivas:

El primer paso es coger cada dato como un grupo, es decir, se tienen tantos grupos como datos (n grupos). El siguiente paso consiste en reducir en uno el número de grupos uniendo los dos elementos más próximos entre sí (tendremos $n - 1$ grupos). En las sucesivas etapas, se disminuye, cada vez, en uno el número de grupos mediante la unión en un sólo grupo de los dos grupos más cercanos entre sí. Finalmente, se tendrá un único grupo incluyendo todos los datos.

Por lo tanto, este procedimiento jerárquico consiste en calcular, en primer lugar, distancias entre puntos (distancia euclídea, distancia de Mahalanobis, distancias L_q,...) y, después, dis-

tancias entre grupos (entorno próximo, entorno lejano, distancia promedio,...) para realizar las agrupaciones de los elementos de la muestra.

Agrupaciones divisivas:

Estas agrupaciones son lo contrario de las progresivas. El primer paso consiste en tener un sólo grupo con todos los elementos. El paso siguiente divide el grupo en dos grupos de tal manera que éstos sean lo más homogéneos posibles. En el siguiente paso, se divide uno de los dos grupos en dos, de modo que los tres grupos sean lo más homogéneos posibles. Este procedimiento sigue hasta que se tenga el mismo número de agrupaciones que de individuos.

3

k-medias y k-medias recortadas

3.1. k-medias

Como se ha señalado, el Análisis Cluster consiste en encontrar agrupaciones lo más homogéneas posibles. Hablando imprecisamente, en el caso de puntos en $\Re^p$, este análisis busca grupos de puntos similares entre sí, es decir, que las distancias entre puntos del mismo grupo sean pequeñas. Y, por lo tanto, deben ser pequeñas sus distancias a un punto situado, aproximadamente, en el centro del grupo.

El procedimiento de las **k-medias** se basa en el siguiente teorema:

Teorema 3.1.1

Dados $x_1,\ldots,x_n$, $z\epsilon\Re^p$, *se verifica que*

$$\sum_{i=1}^{n} d_E^2(x_i, z) \geq \sum_{i=1}^{n} d_E^2(x_i, \bar{x})$$

donde $\bar{x}=\frac{(x_1+\ldots+x_n)}{n}$, d_E *es la distancia euclídea y la igualdad sólo es cierta en el caso de que* $z=\bar{x}$.

No se incluirá la demostración de este resultado por una similitud con la del Teorema 3.1.3.

La interpretación de este teorema es que el punto que minimiza la suma de los cuadrados de las distancias euclídeas a los puntos de la muestra $\{x_1, \ldots, x_n\}$, es único y es la media de los puntos de la muestra: $\bar{x}$ (ver [3]).

Con la premisa anterior se puede hacer precisa la idea de hacer pequeñas las distancias (euclídeas) a un punto situado, más o menos, en el centro del grupo; haciendo que la suma de los cuadrados de dichas distancias sea lo más pequeña posible. Utilizando el Teorema 3.1.1, se escoge como *centro del grupo* la media $\bar{x}$, ya que hace mínima esta suma. El procedimiento denominado k-medias consiste en realizar k agrupaciones utilizando las medias como centros de dichos grupos. Dicho de otro modo:

Definición 3.1.2

Dado $X = \{x_1, ..., x_n\} \subset \Re^p$, se busca una partición de X en k subconjuntos disjuntos $X = X_1 \cup X_2 \cup ... \cup X_k$ de tal modo que si llamamos m_1, m_2,..., m_k a las respectivas medias de los puntos que forman cada uno de ellos, se cumpla que la cantidad

$$\sum_{x \epsilon X_1} d_E^2(x, m_1) + \sum_{x \epsilon X_2} d_E^2(x, m_2) + ... + \sum_{x \epsilon X_k} d_E^2(x, m_k)$$

sea lo menor posible, donde d_E es la distancia euclídea (ver [3]). Se llaman k-medias de X a las medias m_1, m_2,..., m_k.

El Teorema 3.1.1 y la Definición 3.1.2 anteriores están asociadas a n puntos de una muestra $\{x_1, \ldots, x_n\}$ en $\Re^p$. Este trabajo se centrará en coger como muestra n funciones reales y no n puntos en $\Re^p$. Estas funciones serán los estimadores núcleo de densidad (*kernel density estimations*) que explicaremos más adelante y serán los objetos utilizados para el procedimiento de k-medias recortadas.

A continuación, para introducir el resultado anterior para funciones, se explican los espacios de funciones L_p donde estas funciones van a ser definidas (en este caso se utilizará el espacio L_2).

Espacio de funciones: L_p

Sea (a, b) un intervalo en $\Re$ y $f : (a, b) \longmapsto [0, \infty)$ una función medible. Si se define la integral de f se comprueba que no depende de los valores que toma esta función en un conjunto de medida cero. Por tanto, se define una relación de equivalencia en el conjunto de las funciones medibles que las agrupa en clases, de manera que dos funciones medibles f y g se encuentran en la misma clase si y sólo si $f(t) = g(t)$ en c.t.p (en 'casi todo punto' significa que es válida para todos los puntos del dominio definido para las funciones, excepto como mucho en un conjunto de medida cero). Se denomina $M(a, b)$ a este conjunto de clases de funciones medibles y, finalmente, se definen los espacios L_p del siguiente modo:

$$L_p(a, b) = \{f \epsilon M(a, b) : ||f||_p < \infty\};\ 1 \leq p \leq \infty$$

donde $||f||_p := (\int |f(t)|^p dt)^{\frac{1}{p}}$ (ver [4]).

En este contexto, el Teorema 3.1.1 queda del siguiente modo:

Teorema 3.1.3

Dadas f_1, f_2,..., f_n, g funciones pertenecientes a L_2, se verifica que

$$\sum_{i=1}^{n} \|f_i - g\|^2 \geq \sum_{i=1}^{n} \|f_i - \bar{f}\|^2$$

donde $\bar{f} = \frac{(f_1 + ... + f_n)}{n} \epsilon L_2$ y la igualdad sólo es cierta en el caso de que $g = \bar{f}$.

(la demostración se ve en la sección 6.1 del apéndice.)

La Definición 3.1.2, traducida al campo funcional quedaría del modo siguiente:

Definición 3.1.4

Dado $F=\{f_1, f_2, \ldots, f_n\} \subset L_2$, buscamos una división de F en k subconjuntos disjuntos $F=F_1 \cup F_2 \cup .. \cup F_k$ de modo que si llamamos $\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_k$ a las medias de las funciones que forman cada uno de ellos, se cumple que la cantidad

$$\sum_{g \epsilon F_1} ||g - \bar{f}_1||^2 + \sum_{g \epsilon F_2} ||g - \bar{f}_2||^2 + ... + \sum_{g \epsilon F_k} ||g - \bar{f}_k||^2 \tag{3.1}$$

sea lo menor posible y donde $||\cdot||$ es la norma 2 en el espacio L_2. Se llaman k-medias de F a las funciones $\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_k$.

Procedimiento de cálculo de las k-medias

En términos prácticos, este procedimiento de cálculo es imposible y, aunque es sencillo comprobar que los k conjuntos que forman la partición deben ser convexos, no se conoce ningún algoritmo que permita encontrar una partición óptima. Para minimizar (3.1) se pueden calcular todas las posibles particiones de la muestra en subconjuntos convexos. El cálculo no es muy difícil en el caso real porque los subconjuntos son intervalos pero cuando, el tamaño n de la muestra y la dimensión p del espacio crecen (en el caso de puntos en $\Re^p$), las posibles particiones aumentan y computacionalmente el cálculo es muy costoso.
El problema es aún peor en el caso de funciones y puede darse el caso para dimensiones altas o funciones de que sea necesario analizar todas las posibles particiones de F en k subconjuntos.

Como consecuencia de las dificultades anteriores, se ha propuesto el siguiente algoritmo de cálculo basado en el Teorema 3.1.3 que directamente encuentra un mínimo local.

Algoritmo 3.1.5 *Algoritmo (caso k general)*

Dada una muestra de datos en $F=\{f_1, f_2, \ldots, f_n\}$, y el número k de grupos:

1. *Se eligen grupos iniciales al azar:*
   - *Para cada i=1,...,n, se llama g_i al resultado de elegir un elemento de $\{1, ..., k\}$ al azar.*
   - *Si se quedara algún grupo vacío, se elegería un elemento de $\{f_1, f_2, \ldots, f_n\}$ al azar para rellenar ese grupo y su media, en el paso 2., sería el propio elemento.*
2. *Se calculan las medias de los diferentes grupos:*

   $\bar{f}_j$ `=` $\texttt{mean}\{f_i : g_i = j\}$ `j=1,...,`k
3. *Para cada función $f_1, \ldots, f_n$:*

   a) *Se denomina d_i, i=1,...,n, al mínimo de las distancias entre la función f_i y las medias.*

   d_i `=` $\text{mín}_j \, ||f_i - \bar{f}_j||$

   b) *Se denomina g_i, i=1,...,n, al índice de esta media.*

$$||f_i - \bar{f}_{g_i}|| \;=\; \text{mín}_j \, ||f_i - \bar{f}_j||$$

- *Como en el paso 1., si hubiera grupos vacíos se asignaría un elemento al azar de* $\{f_1, f_2, \ldots, f_n\}$ *y su media sería el mismo elemento.*

4. Se itera desde el paso 2 al 3, hasta la convergencia.

El algoritmo sólo lleva a un mínimo local y, por ello, puede no conducir a la solución óptima. Para solventar esto, se recomienda probar con diferentes particiones iniciales y quedarse con la solución que obtenga el mínimo de $\sum_{g\epsilon F_1} ||g-\bar{f}_1||^2+\sum_{g\epsilon F_2}||g-\bar{f}_2||^2+...+\sum_{g\epsilon F_k}||g-\bar{f}_k||^2$.

Es importante destacar que, en los textos, estos enunciados y resultados están realizados para una muestra de puntos p-dimensionados. Solamente, se tienen conocimiento de un estudio donde se utiliza para funciones (ver [1]).

3.2. k-medias recortadas

En Estadística, la robustez es la resistencia de los procedimientos a la presencia de datos anómalos (ver [5]). Estos valores anómalos son denominados outliers. El procedimiento de las k-medias es muy poco robusto. Un ejemplo sencillo de ello para el caso $k = 2$ es:

Dado el conjunto de puntos $\{10, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1\}$, se calculan las 2-medias. Se observan dos grupos: por un lado, se encuentra el grupo formado por los 1's y, por el otro, el grupo formado por los 2's. Sin embargo, existe un dato, el 10, que está claramente aislado del resto. Este dato es lo que se denomina como outlier.

Como los datos son unidimensionales, si se llama

$$x_i = \begin{cases} 10 & \text{si}, i = 1 \\ 2 & \text{si}, i = 2, ..., 6 \\ 1 & \text{si}, i = 7, ..., 11 \end{cases}$$

está claro que puede restringirse el valor de n_1 al que minimice la expresión

$$\sum_{i=1}^{n_1} (x_i - m_1)^2 + \sum_{i=n_1+1}^{11} (x_i - m_2)^2$$

donde $m_1{:=}\frac{1}{n_1}\sum_{i=1}^{n_1} x_i$ y $m_2{:=}\frac{1}{11-n_1}\sum_{i=n_1+1}^{11} x_i$. Por otro lado, para n_1 está claro que sólo es necesario considerar los casos $n_1 = 6$ y $n_1 = 1$. En el primer caso se tiene que

$$(10 - 3{,}3)^2 + 5(2 - 3{,}3)^2 + 5(1 - 1)^2 \approx 53$$

y en el segundo

$$(10 - 10)^2 + 5(2 - 1{,}5)^2 + 5(1 - 1{,}5)^2 = 2{,}5$$

En conclusión, la segunda opción es la que proporciona el óptimo. Sin embargo, es evidente que hacer un grupo con los 1's y 2's y otro grupo con un único dato, no es una buena clasificación. Por lo tanto, se comprueba que el elemento 10 es un valor anómalo que trastorna el resultado de la agrupación. Estas anomalías pueden deberse a errores humanos que, por ejemplo, han

considerado el valor 1.0 por 10. No es razonable que un único error tenga un efecto tan grande.

Es por ello que se introduce el procedimiento de **k-medias recortadas**. Este método ya ha sido utilizado para funciones previamente en [1].

El procedimiento de k-medias recortadas consiste en realizar el método de k-medias pero, primero, fijando un número de puntos que no van a ser considerados en el procedimiento: los posibles outliers. Entonces, es necesario introducir una modificación en el Algoritmo 3.1.5.

Algoritmo 3.2.1 *Procedimiento de cálculo de k-medias recortadas: Algoritmo*

Dada una muestra de datos en $F=\{f_1, f_2, \ldots, f_n\}$*, el número* k *de grupos y* α *el número de datos que no se van a tener en cuenta:*

1. *Se eligen grupos iniciales al azar:*
   - *Para cada* $i=1,\ldots,n$ *se llama* g_i *al resultado de elegir un elemento de* $\{1, ..., k\}$ *al azar.*
   - *Si se quedara algún grupo vacío, se elegería un elemento de* $\{f_1, f_2, \ldots, f_n\}$ *al azar para rellenar ese grupo.*
2. *Se calculan las medias de los diferentes subconjuntos o agrupaciones:*

   $$\bar{f}_j = \texttt{mean}\{f_i : g_i = j\} \quad \texttt{j=1,...,k}$$

3. *Para cada función* $f_1,\ldots,f_n$*:*

   a) *Se denomina* d_i*,* $i=1,\ldots,n$*, al mínimo de las distancias entre la función* f_i *y las medias.*

   $$d_i = \min_j ||f_i - \bar{f}_j||$$

   b) *Se denomina* g_i*,* $i=1,\ldots,n$*, al índice de esta media.*

   $$||f_i - \bar{f}_{g_i}|| = \min_j ||f_i - \bar{f}_j||$$

   - *Como en el paso 1., si hubiera grupos vacíos se asignaría un elemento al azar de* $\{f_1, f_2, \ldots, f_n\}$*.*
4. *Se ordenan las distancias:* $d_{i_1} \geq d_{i_2} \geq \ldots \geq d_{i_n}$ *(*$d_i = d(f_i, \bar{f}_{g_i})$*).*
5. *Se hace* $g_{i_1}=g_{i_2}=\ldots=g_{i_\alpha}=0$*.*
   - *Si existieran grupos vacíos, se escogería un elemento al azar de los que han sido recortados* $(f_1, \ldots, f_\alpha)$ *y se adjudicaría al grupo vacío.*
6. *Se itera desde el paso 2 al 5, hasta la convergencia (ver [5]).*

3.2.1. Procedimiento para elegir k y α:

En la aplicación que se hace de este método surgen dos problemas para resolver simultáneamente:

1. Elegir el número k de agrupaciones para poder dar una buena descripción de las diferentes poblaciones de los países de América.

2. Determinar el número α de países que serán eliminados del conjunto inicial.

Como se ha mencionado previamente, para la elección de k no existe un procedimiento 100 % eficaz (es el caso de las tallas de camiseta) y, además, los algoritmos disponibles no siempre llevan a una solución óptima.

Se utilizará un procedimiento basado en las dos ideas siguientes:

Dado el conjunto de datos $X=\{x_1,\ldots, x_n\}$ (formado por vectores o funciones):

1. Si el punto x es un outlier en X, debería existir un nivel de recorte α_0 de tal manera que x debería ser recortado cuando se calculan k-medias recortadas, independientemente del k elegido, si recortamos un número mayor que α_0 puntos.

   En otras palabras, si x es un outlier debe de estar separado del resto de puntos y, por tanto, debe ser recortado independientemente del número de grupos. Además, si existen puntos más anómalos que x, éstos deben ser recortados antes que x.

2. Si se necesitan k medias para obtener una buena descripción del conjunto X, las k-medias deberían de ser estables cuando el valor de α varía.

   El punto 2 quiere decir que si α aumenta, los k diferentes centros (las medias) de cada grupo tienden a moverse, pero no demasiado. La razón es que cuando se recortan puntos, éstos son eliminados de la parte más externa de los grupos y, por lo tanto, el centro de cada grupo se mueve (es decir, la media de cada grupo varía), pero relativamente poco. En cambio, si se tienen más grupos de los necesarios, alguno es artificial y el nivel de recorte puede hacer variar la media asociada.

En la práctica, para aplicar estas ideas, se ha decidido realizar un análisis para obtener información cuando α=0, 1, 2, 3, 4, 5 o 6, es decir, cuando se recortan 0, 1, 2, 3, 4, 5 o 6 países y cuando se calculan k=2, 3, 4, 5, 6 medias o centros (ver [1]).

4

Estimadores de función de densidad

4.1. Estimador núcleo

El estimador núcleo de densidad $\hat{f}$ de una función de densidad f, (*kernel density estimator (KDE)* en inglés) es un estimador no paramétrico de la densidad que produjo una muestra. El término 'no paramétrico' se refiere a que, a priori no se conoce ninguna familia de pertenencia y son los datos que componen la muestra los que determinan completamente el estimador determinando el número de nodos, los intervalos de crecimiento,etc. La definición es la siguiente:

Definición 4.1.1

Se dispone de una muestra de variables aleatorias independientes e idénticamente distribuidas $\{x_1, \ldots, x_n\}$ *con una distribución de la que se desconoce su función de densidad,* f*. El objetivo es estimar esta función de densidad y su estimador núcleo (también denominado Nadaraya-Watson) es:*

$$f_{n,h_n}(x) := \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right), x \epsilon \Re \tag{4.1}$$

donde K *es una función simétrica, denominada 'núcleo' (kernel en inglés) y* h_n *es un parámetro de suavizado, denominado 'ancho de ventana'.*

La función K de la Definición 4.1.1 debe cumplir ciertas restricciones a la hora de su elección (ver [6]):

$$\int K(x)dx = 1$$

$$K(x) \geq 0$$

$$\int xK(x)dx = 0$$

$$\int x^2 K(x)dt = k_2 \neq 0$$

Además, K debe ser una función suave (frecuentemente se elige una función de clase C^∞). De esta manera, el estimador de densidad por núcleo hereda la condición de suavidad por ser suma de funciones de clase C^∞. Esta condición es necesaria puesto que, como se ha señalado anteriormente, el modelo que hay debajo de un estimador núcleo de densidad, supone que los datos se muestrean a partir de una distribución que tiene una función de densidad desconocida satisfaciendo la condición de suavidad.

En la práctica, se utilizará para K la función Gaussiana $K(t)= \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}t^2}$. De este modo, se garantizará el carácter suave del estimador núcleo de densidad ya que se prueba, de forma sencilla, que la función Gaussiana es de clase C^∞ (por serlo la función exponencial).

Como se ha mencionado anteriormente, aplicando el método k-medias recortadas se pretende agrupar los diferentes países del continente americano. El método se quiere realizar a partir de curvas suaves que aproximen los datos de cada país mejor que las pirámides de población. Estas curvas serán estimadores que se aproximen a la función de densidad real, en este caso, desconocida. En particular, se utilizarán una clase de estimadores: estimadores núcleo de densidad.

De este modo, se introduce un resultado de consistencia (ver [6]) que, dado que se desconoce totalmente la función de densidad f, garantiza la convergencia del estimador f_{n,h_n} a la función f bajo ciertas condiciones de continuidad en f y una elección apropiada de h_n.

Proposición 4.1.2

Supongamos que f es uniformemente continua en (-∞,+∞) y que el ancho de ventana h_n *satisface que*

$$h_n \longrightarrow 0 \ y\ nh_n \log(n)^{-1} \longrightarrow \infty \ cuando\ n \longrightarrow \infty$$

entonces

$$\sup_x \mid f_{n,h_n}(x) - f(x) \mid \longrightarrow 0 \ cuando\ n \longrightarrow \infty \ (ver\ [6])$$

siendo esta convergencia 'en probabilidad' (ver sección 6.2 y [7]).

Para garantizar una buena aproximación de una función estimador de densidad a la función de densidad real, es necesario que la diferencia de ambas funciones sea cero, o en su caso, que la distancia de la función estimador de densidad a la función de densidad converja a cero.

Para el método de k-medias se utilizarán distancias entre estimadores núcleo de densidad. Es importante la convergencia del supremo de la diferencia entre la función estimador y la función de densidad ya que se puede demostrar, de forma sencilla (ver Teorema 6.3.1 en apéndice), que la convergencia del supremo garantiza que $d(f_{n,h_n}, f) = \int_a^b |f_{n,h_n} - f(x)|^2 dt$ converja a cero donde (a, b) es el intervalo en el que están comprendidos los datos (en este caso práctico, el intervalo estará formado por las diferentes edades que tiene la población de los países de América, desde 0 hasta 105 años).

Por otro lado, la elección del ancho de ventana, h_n, resulta trascendente. Su importancia determina al igual que la elección de K, la suavidad de la curva. Si se selecciona un ancho de ventana muy pequeño o muy grande, las curvas son muy puntiagudas (lo que se conoce como 'undersmoothing') o se produce una superposición de las curvas (conocido como 'oversmoothing'), respectivamente. De este modo, un ancho de ventana razonable es un ancho intermedio

(ver [8]). Además, como se ha señalado en la Proposición 4.1.2, h_n debe tender a cero pero a una velocidad lenta.
Veamos un ejemplo práctico de 'undersmoothing' y 'oversmoothing' realizado con el software R:

Ejemplo 4.1.3 *Se crea un vector X de tamaño 2000 de una mezcla de dos distribuciones normales de tamaño 1000. La primera distribución normal tiene media 0 y desviación típica 1 y la segunda distribución normal tiene media 3.5 y desviación típica 2. Además, se fija un conjunto I para el eje de abscisas.*

Se representan las funciones estimador y la función de densidad real (curva roja) con diferentes anchos de ventana: 0.1, 0.5 y 1.

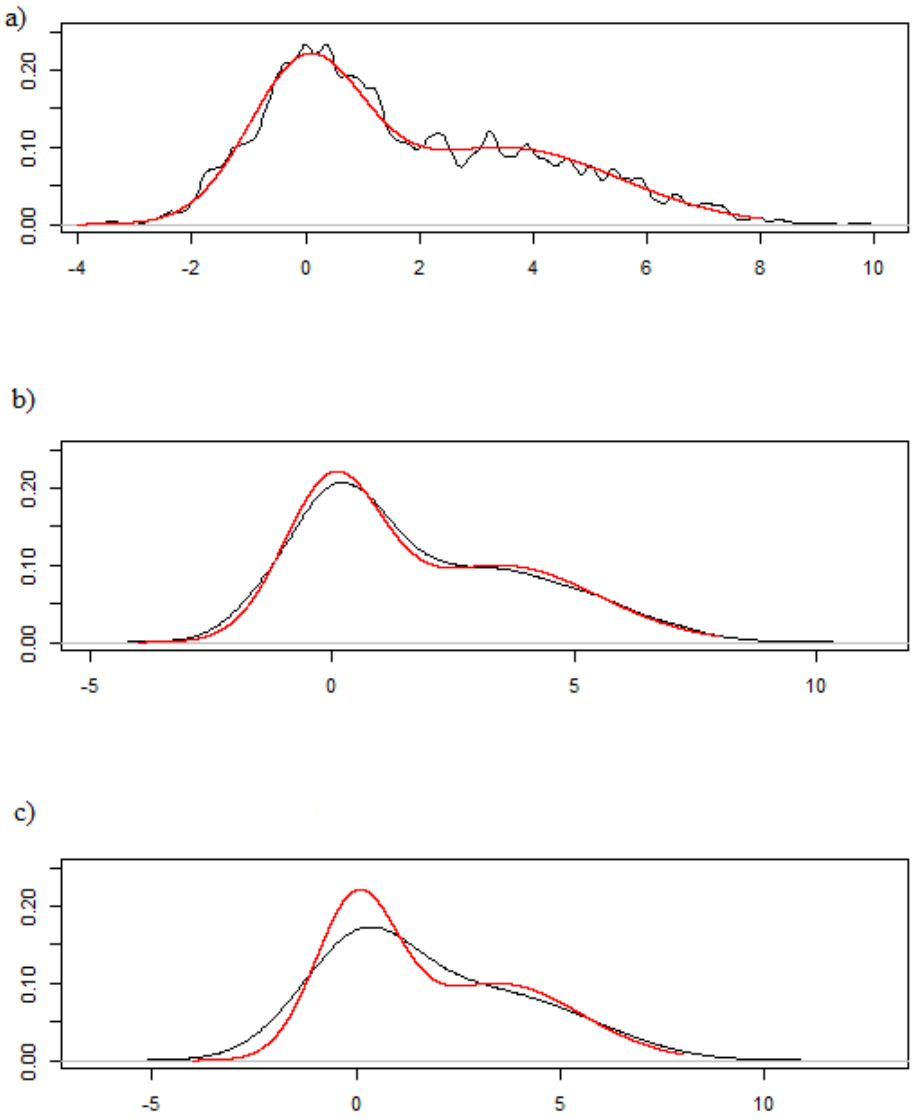

Figura 4.1: a) Estimador núcleo de densidad Gaussiano con un ancho de ventana h_n=0.1 ('undersmoothing'). b) Estimador núcleo de densidad Gaussiano con un ancho de ventana h_n=0.5. c) Estimador núcleo de densidad Gaussiano con un ancho de ventana h_n=1 ('oversmoothing'). Curva roja: función de densidad del histograma de datos.

Finalmente, se dibuja el histograma del vector X.

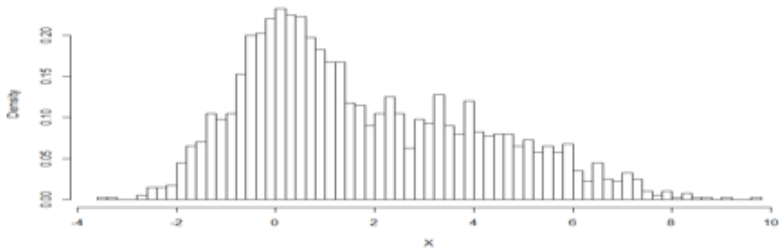

Figura 4.2: Histograma de los datos.

Claramente, la función que mejor se aproxima a la función de densidad del histograma es la correspondiente a h_n=0.5, es decir, un ancho de ventana intermedio.

4.1.1. Elección de la ventana

Para la elección del parámetro de suavizado, frecuentemente se recurre a la minimización del Error Cuadrático Medio Integrado del estimador ($MISE$, por las siglas en inglés: *Mean Integrated Square Error*).

El $MISE$ de un estimador es la integral del promedio de los errores al cuadrado, es decir, es la integral del cuadrado de la diferencia entre el estimador y lo que se pretende estimar. En este caso, el estimador es el núcleo de densidad f_{n,h_n} y lo que se quiere estimar es una función de densidad f. Este método depende de la varianza y del sesgo del estimador y su expresión es la integral de:

$$MISE\left(\hat{f}(x)\right) = \int E\left[\left(\hat{f}(x) - f(x)\right)^2\right] dx = \int Var(\hat{f}(x))dx + \int Bias_h^2(x)dx, \tag{4.2}$$

$$Bias_h(x) \quad = \quad \int K(t)f(x-ht)dt - f(x) \tag{4.3}$$

$$Var(\hat{f}(x)) \quad = \quad E\left[\hat{f}(x)^2\right] - \left(E[\hat{f}(x)]\right)^2 \tag{4.4}$$

(ver [6] para la última igualdad en 4.2)

Un cálculo relativamente largo, pero no excesivamente costoso (ver apéndice 6.5) permite comprobar que, bajo las hipótesis adecuadas (ver [9]), el $MISE$ se minimiza tomando

$$h_{n,opt} = k_2^{-\frac{2}{5}} \left(\int K(t)^2 dt\right)^{\frac{1}{5}} \left(\int f''(x)^2 dt\right)^{-\frac{1}{5}} n^{-\frac{1}{5}} \tag{4.5}$$

donde $k_2 = \int t^2 K(t)dt$ [6].

5

Análisis de pirámides de población

5.1. Introducción del análisis: tratamiento previo de los datos

El objetivo de este trabajo era analizar un caso práctico: agrupar los países de América según la estructura de edad de su población. Para la realización de este análisis se aplicará la teoría explicada en los capítulos anteriores. Además, los resultados que se obtengan van a ser comparados con los obtenidos en [1], con el objetivo de comentar las relaciones y discrepancias (si existen) de utilizar funciones cuantiles, en su caso, y estimadores kernel, en el nuestro.

Los datos a analizar son los mismos de [1] (para facilitar la comparación de los resultados obtenidos). Estos datos fueron descargados de `http://www.census.gov/population/international/data/idb/informationGateway.php` el 13 de Agosto de 2015. Este archivo contiene bastantes variables de carácter demográfico pero en este trabajo se considerará, para cada país, la situación que esta página proporciona del total de la población, el número de personas con 0, 1, 2,..., 84 años de edad, y el número de personas con edades entre 85 y 90, entre 91 y 95, entre 96 y 100 y entre 101 y 105 que son adjudicados a las edades 87, 92, 97 y 102 (ver explicación en la sección 5.2).

Además, para evitar países con poblaciones extremadamente pequeñas, se eliminaron del análisis los países con población inferior a los 100000 habitantes.

Como conclusión, el número de países de América para clasificar será 36. A continuación, se muestra la lista de países.

Países	Población
Canadá	35099836
EE.UU	321368864
Aruba	112162
Bahamas	324597
Barbados	290604
Cuba	11031433
Curacao	148406
República Dominicana	10478756
Grenada	110694
Haití	10110019
Jamaica	2950210
Puerto Rico	3598357
St Lucía	163922
St Vicente y Grenadines	102627
Trinidad y Tobago	1222363
Islas Vírgenes	103574
Belize	347369
Costa Rica	4814144

Tabla 5.1: Países con su número de población.

Países	Población
El Salvador	6141350
Guatemala	14918999
Honduras	8746673
Mexico	121736809
Nicaragua	5907881
Panamá	3657024
Argentina	43431886
Bolivia	10800882
Brasil	204259812
Chile	17508260
Colombia	46736728
Ecuador	15868396
Guyana	735222
Paraguay	6783272
Perú	30444999
Suriname	579633
Uruguay	3341893
Venezuela	29275460

Tabla 5.2: Países con su número de población.

5.2. Explicación del análisis del caso práctico

En primer lugar y previo a poner en práctica el análisis, se observan las pirámides de población. La Figura 5.1 es un ejemplo de pirámide de población:

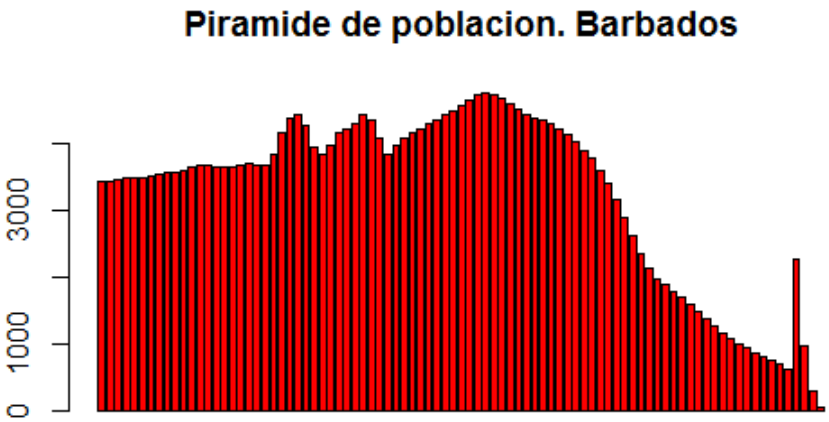

Figura 5.1: Pirámide de población de Barbados.

En Figura 5.1, se aprecian picos en la cola de la gráfica. Esto es debido a que los datos contienen únicamente estimaciones del número de personas por año de edad, del 0 a 84 y, a partir de aquí, las estimaciones de los intervalos 85-89, 90-94, 95-99 y 100 en adelante se adjudican a los puntos 87, 92, 97 y 102. Por esta razón, se divide la cantidad de personas de las edades $a = 87$, 92, 97 y 102 entre las edades $a-1$, $a-2$, a, $a+1$, $a+2$ con una proporción de pesos dada por el vector $S = (2'3, 2'0125, 1'725, 1'4375, 1)$. Estos fueron los pesos elegidos, un tanto arbitrariamente en [1]. Aquí se mantiene esta elección para facilitar la comparación de resultados.

Con esta adjudicación desaparecen los picos de la cola, como puede comprobarse en la Figura 5.2.

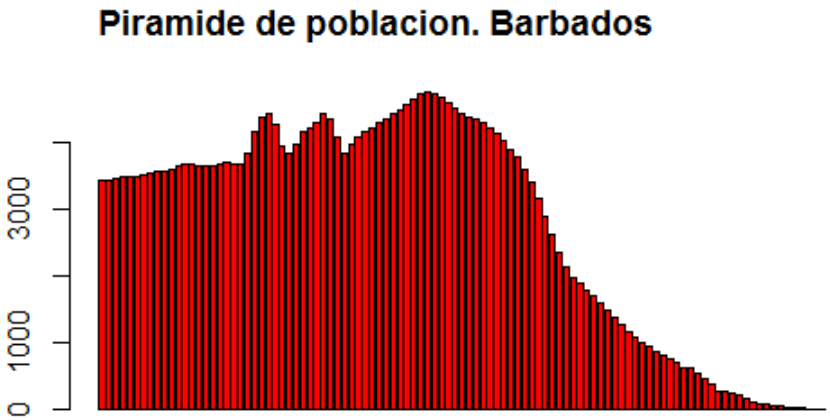

Figura 5.2: Pirámide de población de Barbados sin picos en la cola.

Es importante comentar que, el tamaño del vector de edades comprendidas entre 0 y 105 es pequeño (tamaño de 106 elementos) en comparación con el gran número de datos (del orden de millones). Por esta razón, en el momento de estimar las funciones de densidad (ver Sección 4.1) se obtiene un h_n demasiado pequeño que produce en la gráfica un caso de 'undersmoothing'. La Figura 5.3 muestra el caso de Trinidad y Tobago. Aparecen picos similares con todos los países analizados.

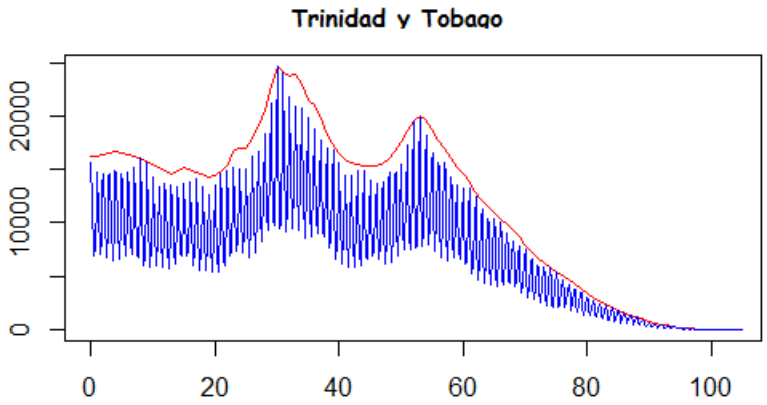

Figura 5.3: Curva roja: Pirámide de población de Trinidad y Tobago. Curva azul: Función estimador de densidad de la pirámide de población de Trinidad y Tobago con h_n=0.2536779. Se produce una curva 'undersmoothing'.

Este problema se puede solventar teniendo en cuenta que, en realidad, las edades son los partes enteras de las edades reales. Por ello, cada intervalo se ha dividido en 100 y se ha repartido las personas uniformemente entre estos valores. Por ejemplo, en Canadá hay 359312 con edad 0 que se han repartido entre las edades 0, 0.01,...,0.99 a partes iguales. Análogamente para el resto de años, excepto para el año 105 que es donde se para. De esta manera, el vector de edades ya no es de tamaño 106, sino de tamaño 10501 (=105·100+1). Una vez que se consigue este tamaño se obtiene una ventana h_n razonable acorde a la gran cantidad de datos. El resultado se ve en

la siguiente imagen.

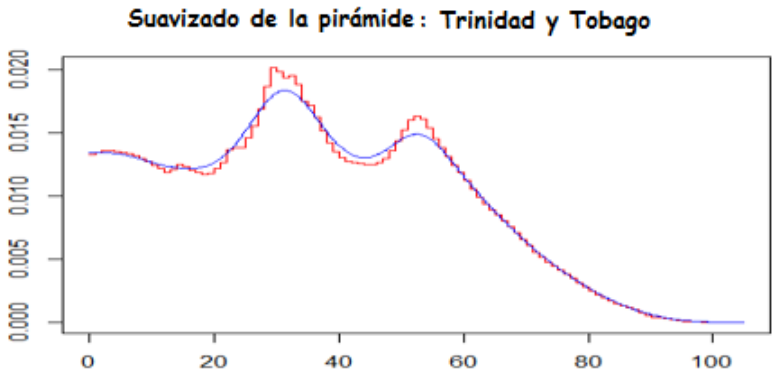

Figura 5.4: Curva roja: Pirámide de población de Trinidad y Tobago. Curva azul: Función estimador de densidad de la pirámide de población de Trinidad y Tobago con h_n=3.164746.

5.3. Análisis de las pirámides

Como se ha mencionado anteriormente, el análisis que se va a utilizar es el Análisis Cluster. Concretamente, el método de k-medias recortadas. Se quiere aplicar este procedimiento a aproximaciones continuas de las pirámides de población de cada país de América. El primer paso del análisis es obtener las aproximaciones continuas a partir de las pirámides de población, es decir, obtener unas curvas suaves a partir de datos discretos. Para ello, se emplean las pirámides de población de cada país como muestras (de enorme tamaño) obtenidas de funciones de densidad que hay que estimar. Por tanto, se pretende utilizar estimadores de densidad como curvas suaves que se aproximen a las funciones de densidad.

En resumen, el método de k-medias recortadas se aplicará a una muestra de estimadores de densidad con el fin de realizar un agrupamiento de los países de América atendiendo a sus pirámides de población.

El primer paso es calcular los estimadores de densidad $\hat{f}_1$, $\hat{f}_2$,..., $\hat{f}_{36}$, uno por cada país. El modelo de estimador elegido es el de Nadaraya-Watson denominado estimador por núcleo de densidad (4.1.1). Para el núcleo K se ha elegido la función Gaussiana y para calcular el ancho de ventana, h_n, se aplica el método de Sheater & Jones. Este método se describe en [10] y se basa en realizar una estimación preliminar de la expresión $\int f''(x)^2 dx$ en (4.5).

Como resultado de lo anterior, se tienen los estimadores $\hat{f}_1$, $\hat{f}_2$,..., $\hat{f}_{36}$, cada uno de ellos calculado en 512 puntos equiespaciados $0 = t_1 < t_2 < ... < t_{512} = 105$. Un ejemplo de la función estimada puede verse en la Figura 5.4. Además, se utilizará la siguiente aproximación del producto escalar en $L_2[0, 105]$; dadas $f, g \epsilon L_2[0, 105]$

$$< f, g > \approx \frac{105}{512} \sum_{i=1}^{512} f(t_i) g(t_i)$$

A partir de este momento únicamente hay que aplicar el Algoritmo 3.2.1 de k-medias recortadas tal y como se presenta en la Sección 3.2. Sin embargo, está pendiente la selección de los valores apropiados de k y α. Estos problemas se analizan en las siguientes subsecciones.

5.3.1. Nivel de recorte

Como se ha mencionado en la subsección 3.2.1, se deben tener en cuenta dos ideas por las que el procedimiento que se utiliza para la selección del nivel de recorte es adecuado. La primera razón es que si existen datos anómalos estos deben ser recortados independientemente del número de agrupaciones. La segunda razón es que si el nivel de recorte aumenta, las k-medias no varían demasiado y el resultado final apenas varía.

Para la elección de α, se han calculado las diferentes k-medias ($k = 2, ..., 6$), haciendo un recorte, en cada caso, de 0 hasta 6 países. Por tanto, cada país tiene la posibilidad de ser recortado un máximo de 30 veces. En la Tabla 5.3 aparecen los países que fueron recortados, al menos una vez, junto con el número de veces que fueron recortados.

País	Número de veces recortado
Islas Vírgenes	30
Cuba	19
Trinidad y Tobago	12
Guatemala	6
Canadá	5
Belize	5
Jamaica	4
Barbados	4
Haití	4
Honduras	3
Nicaragua	3
Uruguay	2
Curacao	1
Argentina	1
Bolivia	1
Guyana	1

Tabla 5.3: Número de veces que son recortados los países, que son recortados al menos una vez.

Se observa en la Tabla 5.3 que, hasta Guatemala incluído, la diferencia en el número de veces que un país ha sido recortado y el siguiente es uno o ninguno. Estas diferencias aumentan drásticamente a partir de Trinidad y Tobago.
Se puede concluir, por tanto, que todos los países excepto Islas Vírgenes, Cuba y Trinidad y Tobago están más o menos agrupados y que los únicos outliers son estos tres.

El efecto comentado se observa más claro en la Figura 5.5 donde se aprecia como los países recortados, menos Islas Vírgenes, Cuba y Trinidad y Tobago, se agrupan alrededor de una recta y estos últimos alrededor de otra de pendiente muy superior.

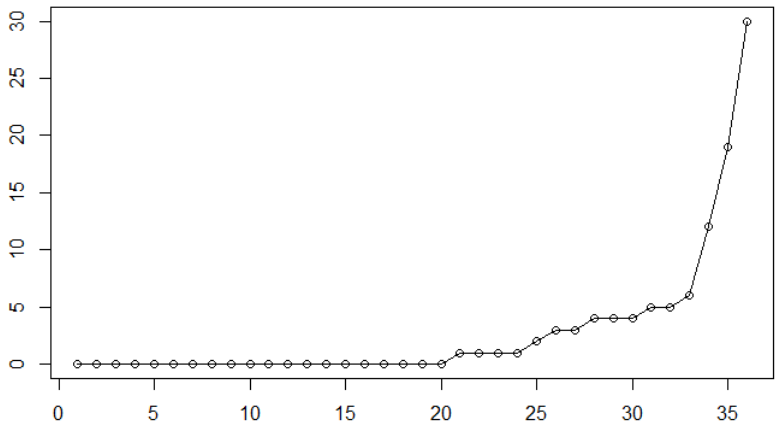

Figura 5.5: Gráfica de los países (eje X, en orden de recorte de decreciente a creciente) y del número de veces (eje Y) que ha sido recortado cada país.

Por esta razón, se considera recortar tres países, es decir, $\alpha = 3$. Este nivel de recorte es, finalmente, el elegido para este trabajo.

5.3.2. Valor óptimo de k

El segundo problema que hay que analizar es la selección del número de agrupaciones, k.

El método que se sigue para encontrar un k razonable es analizar los k grupos resultantes cuando $k = 2, 3, 4, 5, 6$ y el nivel de recorte $\alpha = 0, 1, 2, 3, 4, 5, 6$. Una vez realizado el análisis, se acepta que existen al menos k grupos si las variaciones que se producen en las k-medias, cuando α cambia, son pequeñas. Finalmente, si se encuentran más grupos de los necesarios, alguno será artificial y el nivel de recorte hará que las k-medias varíen demasiado. En este último caso, se concluye que el número de agrupaciones no es adecuado y que el número razonable es el último valor de k que encuentra estabilidad en los resultados.

Se comienza analizando los centros de las agrupaciones obtenidas para cada $k = 2, 3, 4, 5, 6$ para $\alpha = 0, 1, 2, 3, 4, 5, 6$. Más precisamente: dado un valor de k, se tienen k centros de cada uno de los grupos obtenidos para cada valor de α. Estos centros son funciones que se agrupan entre sí por una similitud y se representan gráficamente. Por ejemplo, si se fija $k = 2$ y se toma $\alpha = 0, 1, 2, 3, 4, 5, 6$ se tendrán siete parejas de funciones: (f_1^i, f_2^i), $i = 0, .., 6$.
Se toma como referencia la pareja (f_1^0, f_2^0) y, a continuación, se reordenan las parejas (f_1^i, f_2^i), $i = 1, .., 6$ de modo que

$$||f_1^0 - f_1^i|| < ||f_1^0 - f_2^i||, i = 1, ..., 6$$

Posteriormente se realiza la Figura 5.6 cuyo gráfico izquierdo contiene a las funciones $f_1^0, f_1^1, ..., f_1^6$ y el derecho a $f_1^0, f_2^1, ..., f_2^6$.

Análogamente, ocurre con $k = 3, 4, 5, 6$. Las gráficas se exponen a continuación:

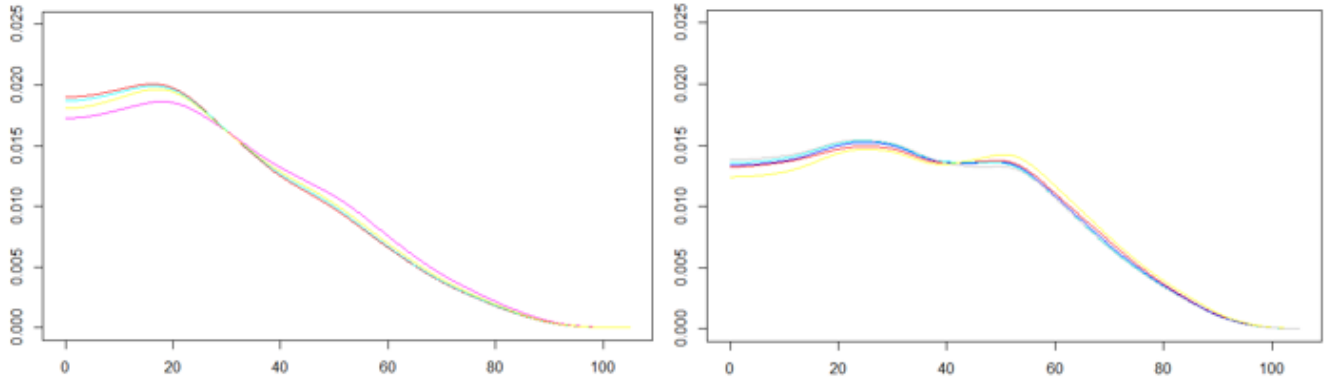

Figura 5.6: 2-medias cuando $\alpha = 0, 1, 2, 3, 4, 5, 6$. Los colores se refieren a los diferentes valores de α.

Observando la Figura 5.6, están claros los dos grupos que se obtienen: en la gráfica de la izquierda se encuentran los países con población más joven, mientras que en la de la derecha están los países cuya población es más envejecida. La homogeneidad que se observa en las funciones representadas sugiere que $k \geq 2$.

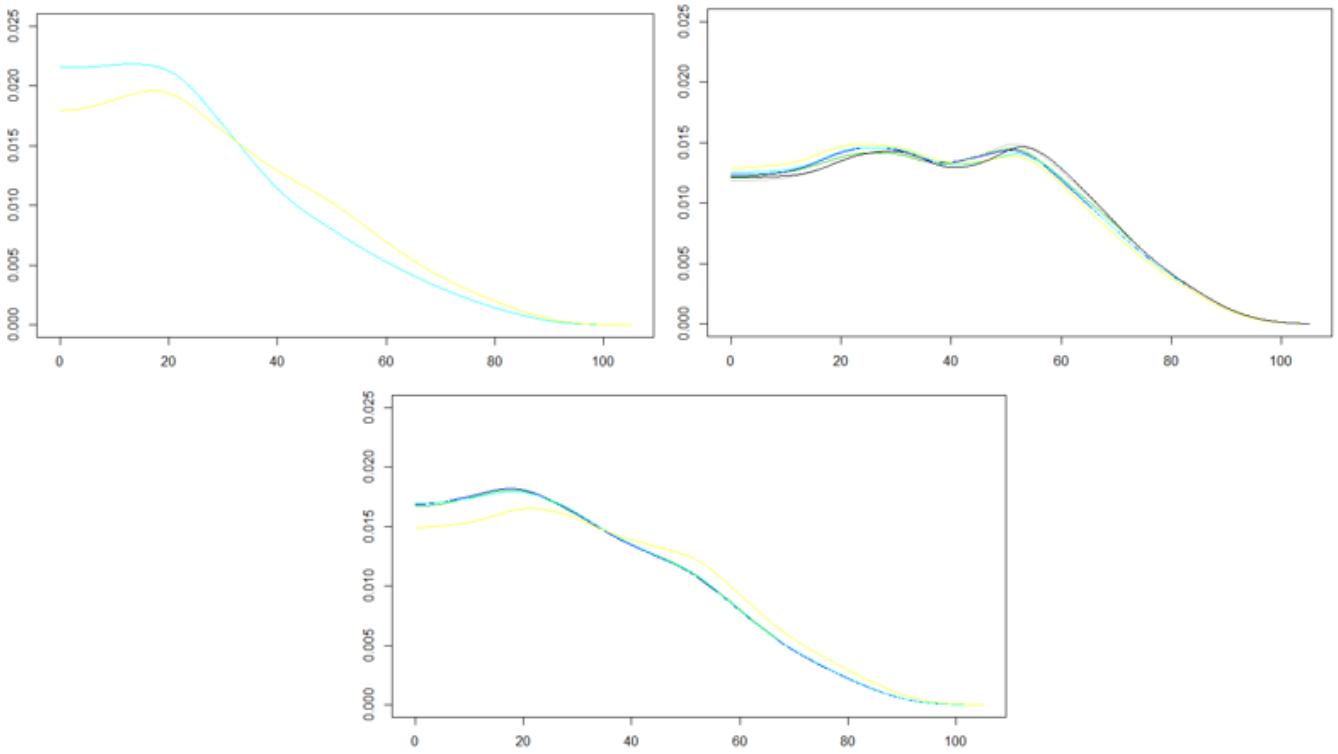

Figura 5.7: 3-medias cuando $\alpha = 0, 1, 2, 3, 4, 5, 6$. Los colores se refieren a los diferentes valores de α.

En la Figura 5.7 se representan las 3-medias obtenidas siguiendo el criterio mencionado. En la gráfica superior izquierda se encuentran los países que tienen el mayor número de habitantes con 0 años de edad, en la superior derecha están los países que tienen el mayor número de personas con edad en torno al 25 y al 55 de edad (ofrece dos máximos locales) y, por último, en la tercera gráfica aparecen los países cuyo mayor número de habitantes tiene en torno a 20 años.

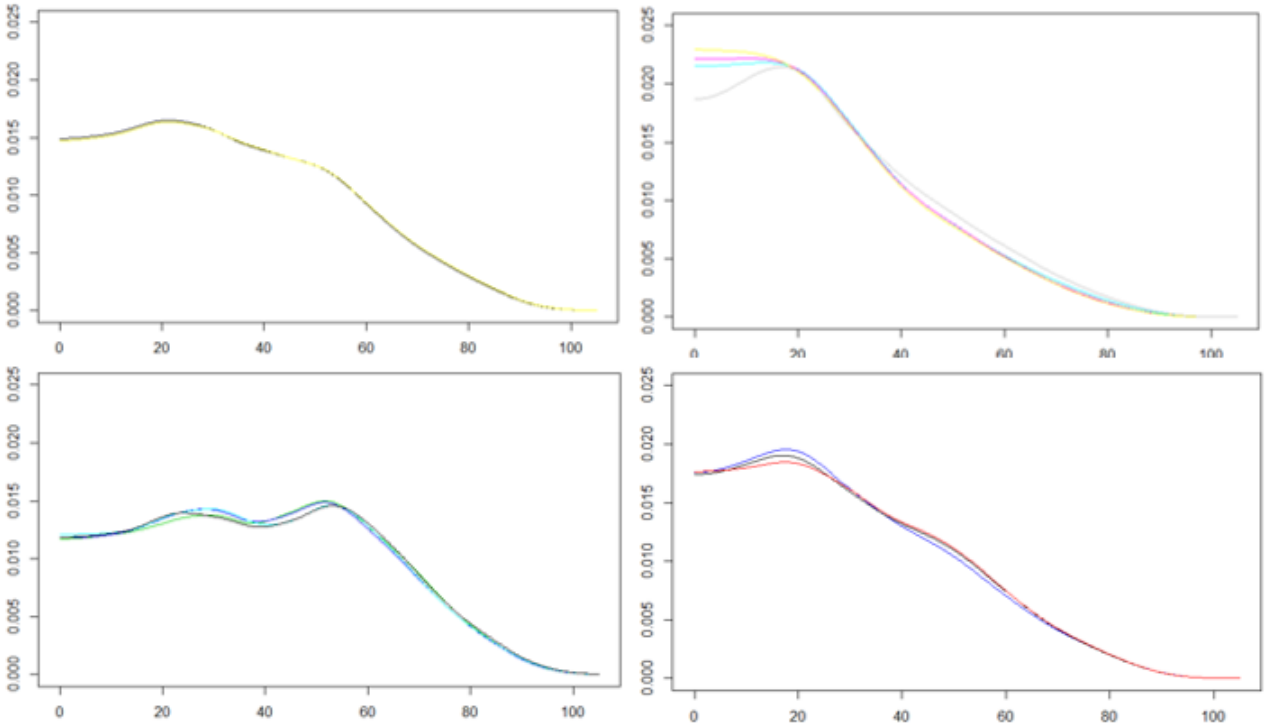

Figura 5.8: 4-medias cuando $\alpha = 0, 1, 2, 3, 4, 5, 6$. Los colores se refieren a los diferentes valores de α.

El caso de $k = 4$, como se observa en la Figura 5.8, es similar al caso $k = 3$. Únicamente se aprecia una diferencia entre ambos casos. En el grupo nuevo, la mayoría de población también tiene en torno a 20 años de edad pero con una mayor proporción de personas a diferencia del otro grupo con esta misma característica (explicación: en uno de los grupos el máximo de habitantes está por encima de 0.017 en el intervalo de edad [0,20], mientras que en el otro, está por debajo).

Las agrupaciones obtenidas en los casos $k = 3$ y $k = 4$ (ver Figuras 5.7 y 5.8) también son bastante homogéneas. Por lo tanto, se puede concluir que $k \geq 4$.

Por otro lado, en los casos $k = 5, 6$, aparece una inestabilidad en la representación gráfica que justifica la decisión de no poder tomar cinco o seis agrupaciones para estos datos (ver Figuras 5.9 y 5.10). En el primer caso, tres de los cinco grupos son estables mientras que dos de ellos son inestables (en uno se aprecia más esa inestabilidad que en el otro). En el segundo caso, ocurre lo mismo que en el caso anterior: aparecen dos grupos inestables de los seis.

De acuerdo con el criterio supuesto se puede concluir que $k = 4$ es la elección adecuada para el número de grupos.

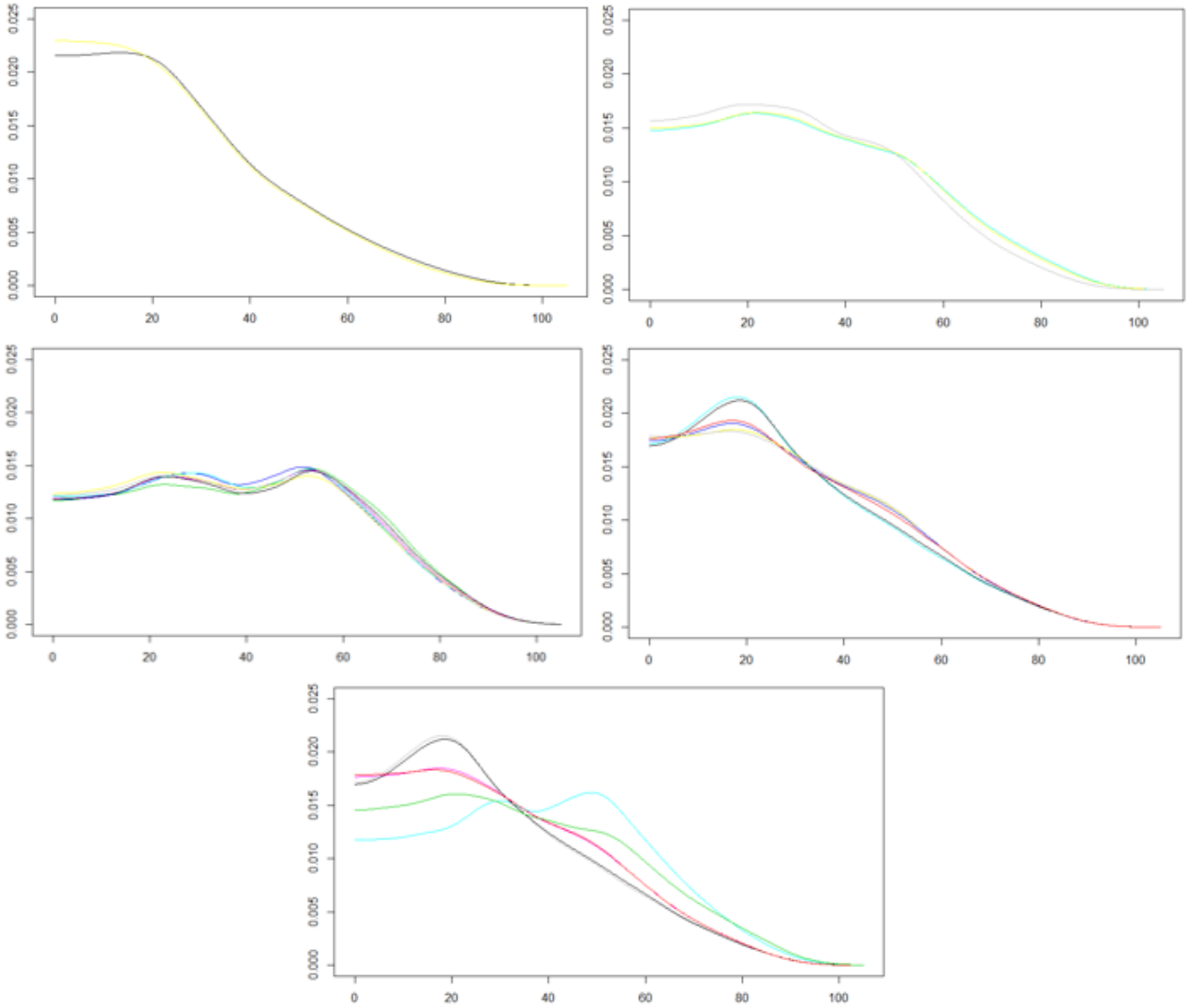

Figura 5.9: 5-medias cuando $\alpha = 0, 1, 2, 3, 4, 5, 6$. Los colores se refieren a los diferentes valores de α.

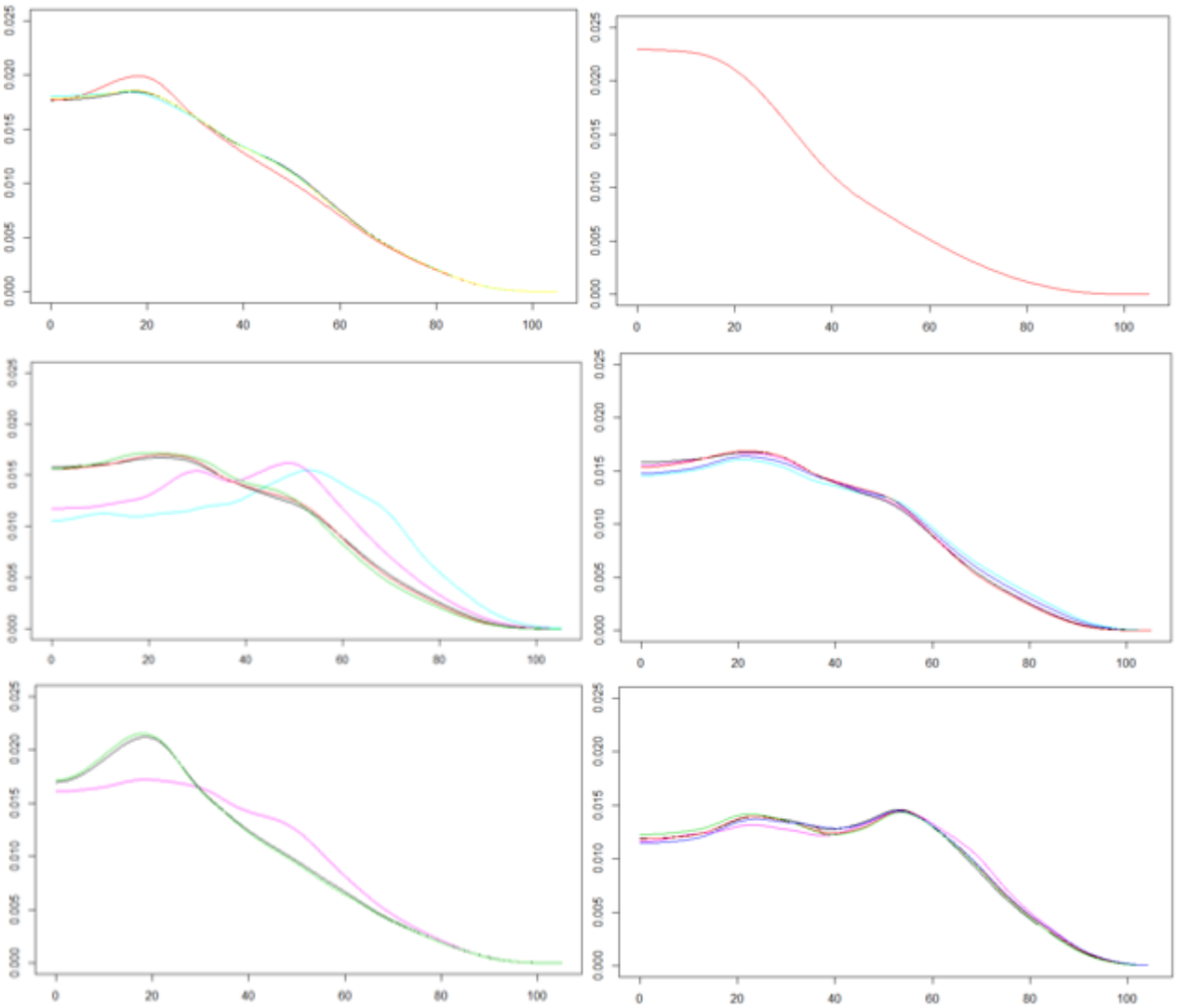

Figura 5.10: 6-medias cuando $\alpha = 0, 1, 2, 3, 4, 5, 6$.Los colores se refieren a los diferentes valores de α.

5.4. Análisis de los resultados

En esta sección se van a analizar los resultados obtenidos al aplicar el procedimiento de k-medias recortadas (supuesto en [1]) a las pirámides de población de los países de América.

Por las razones expuestas en las subsecciones 5.3.1 y 5.3.2, se tiene $k = 4$ y $\alpha = 3$.

La sección se divide en dos apartados. En el primero se componen las k-medias entre sí y los países recortados y, en la segunda, se analizan los grupos obtenidos.

5.4.1. k-medias y países recortados

La Figura 5.11 contiene la representación gráfica de las funciones de las 4-medias obtenidas y de los países que, finalmente, han sido recortados: Islas Vírgenes, Cuba y Trinidad y Tobago (en gris).

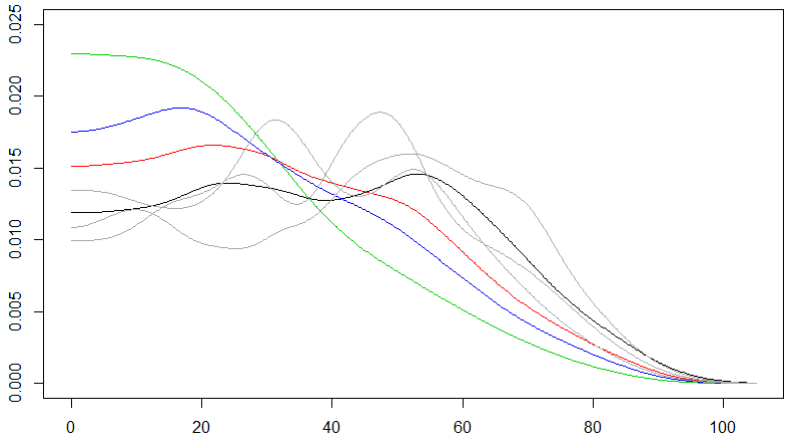

Figura 5.11: Las 4-medias cuando $\alpha = 3$. Los colores se refieren a la media de cada grupo y las curvas grises son los tres países recortados: Islas Vírgenes, Cuba y Trinidad y Tobago.

Como se ha señalado, los grupos obtenidos identifican cuatro grados diferentes de envejecimiento de la población: la curva verde representa a países de población joven, con una proporción alta de habitantes en el intervalo de edad [0,20] y la curva negra representa una población envejecida con una proporción alta de personas de alrededor de 55 años. Las curvas azul y roja representan estados intermedios.

Además, se puede observar en la Figura 5.11, como las funciones estimador de los tres países recortados no se asemeja a ninguna de las 4-medias. Por tanto, es razonable el recorte de esos tres países.

5.4.2. Grupos obtenidos

La Figura 5.12 incluye una gráfica por cada uno de los cuatro grupos resultantes. En estas gráficas se muestra la estimación de las funciones de los países pertenecientes a cada grupo y su correspondiente 4-media.

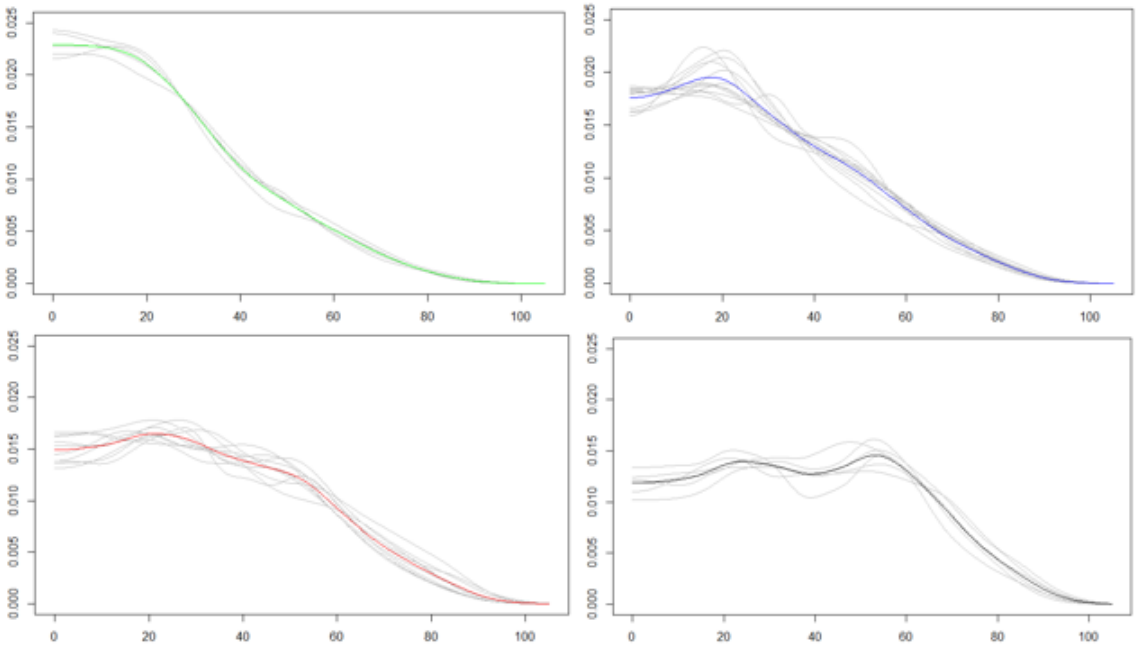

Figura 5.12: Las 4-medias (arriba: curvas verde y azul; abajo: roja y negra) y los países asociados a cada una de ellas (curvas grises).

En la Figura 5.12 se observa homogeneidad en los grupos obtenidos. En ningún caso, los miembros de cada grupo se alejan tanto de su k-media como lo haría un outlier (Islas Vírgenes, Cuba y Trinidad y Tobago) si se metiera en cualquiera de los grupos.

A continuación, se muestra una tabla de clasificación de cada país en su grupo correspondiente, así como de los países que han sido recortados.

Grupo I	Belize, Bolivia, Guatemala, Haití, Honduras
Grupo II	Nicaragua, República Dominicana, Ecuador, El Salvador, Guyana, Jamaica, México, Panamá, Paraguay, Perú, Suriname, Venezuela
Grupo III	Argentina, Bahamas, Brasil, Chile, Costa Rica, Grenada, St Lucia, St Vicent y Grenadines, Uruguay, Colombia
Grupo IV	Aruba, Barbados, Curacao, Puerto Rico, EE.UU, Canadá
Recortados	Islas Vírgenes, Cuba, Trinidad y Tobago

Tabla 5.4: Miembros de los grupos. $k = 4$ y $\alpha = 3$. Metodología de este trabajo

5.5. Comentarios finales

5.5.1. Comparación con los resultados de [1]

Como se ha señalado en la introducción, el objetivo de este trabajo es comparar los resultados obtenidos con la metodología presentada, con los obtenidos en [1].

Ambas metodologías tienen en común el utilizar un espacio L_2 para el cálculo de las k-medias, y difieren en que en [1] se emplean funciones cuantiles (por su interés desde el punto de vista de las métricas de Wasserstein, ver [1]) y aquí se han empleado suavizaciones de las pirámides que, por tanto, están más relacionados con los datos iniciales.

A continuación, se incluyen las gráficas de los resultados obtenidos utilizando funciones cuantiles para poder comparar los resultados de ambos métodos.

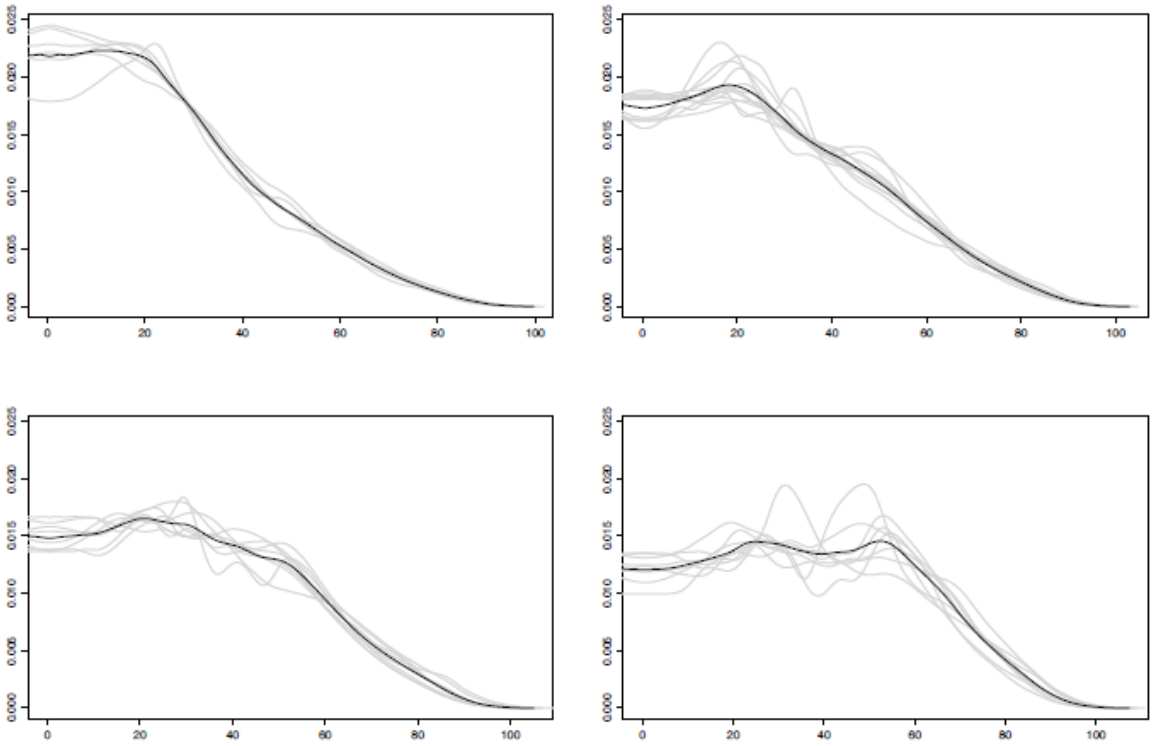

Figura 5.13: Los 4-baricentros (curvas negras) y los países asociados a cada una de ellas (curvas grises). Gráfica tomada de [1].

En primer lugar, para la elección de k, en ambos trabajos se concluye que la mejor agrupación de los 36 países de América es hacer una clasificación en cuatro grupos. Además, en la Figura 5.12 y la Figura 5.13, se puede observar que las k-medias de los grupos son muy similares en ambos trabajos. Sin embargo, la repartición de los países en los cuatro grupos difiere, pero muy poco. Además, en [1] se concluye que sólo hay dos países anómalos: Canadá e Islas Vírgenes. A continuación, se presentan las dos clasificaciones resultantes de los países en los cuatro grupos: por un lado, con funciones cuantiles y, por otro, con funciones estimador.

Grupo I	Belize, Bolivia, Guatemala, Haití, Honduras, **Nicaragua**
Grupo II	**Colombia**, República Dominicana, Ecuador, El Salvador, Guyana, Jamaica, México, Panamá, Paraguay, Perú, Suriname, Venezuela
Grupo III	Argentina, Bahamas, Brasil, Chile, Costa Rica, Grenada, St Lucia, St Vicent y Grenadines
Grupo IV	Aruba, Barbados, **Cuba**, Canadá, Curacao, Puerto Rico, **Trinidad y Tobago**, EE.UU, **Uruguay**
Recortados	**Canadá**, Islas Vírgenes

Tabla 5.5: Miembros de los grupos. $k = 4$ y $\alpha = 2$ Tomada de [1].

Grupo I	Belize, Bolivia, Guatemala, Haití, Honduras
Grupo II	**Nicaragua**, República Dominicana, Ecuador, El Salvador, Guyana, Jamaica, México, Panamá, Paraguay, Perú, Suriname, Venezuela
Grupo III	Argentina, Bahamas, Brasil, Chile, Costa Rica, Grenada, St Lucia, St Vicent y Grenadines, **Uruguay**, **Colombia**
Grupo IV	Aruba, Barbados, Curacao, Puerto Rico, EE.UU, **Canadá**
Recortados	Islas Vírgenes, **Cuba**, **Trinidad y Tobago**

Tabla 5.6: Miembros de los grupos. $k = 4$ y $\alpha = 3$. Metodología de este trabajo

En total, en las Tablas 5.5 y 5.6 se pueden observar 6 diferencias (marcadas en negrita) entre los países que pertenecen a uno u otro grupo en ambos casos.

Las Figuras 5.14, 5.15 y 5.16 analizan la situación de los tres países que cambian de grupo en nuestro trabajo con respecto de [1]: Nicaragua, Colombia y Uruguay. En estas figuras se representa en rojo la k-media del grupo adjudicado en [1] y en azul la del grupo adjudicado aquí. A la vista de la similitud entre las k-medias obtenidas, las dos k-medias representadas son las obtenidas con la metodología presentada aquí.

La primera diferencia es que Nicaragua figura en el Grupo I de la Tabla 5.5 y en el Grupo II de la Tabla 5.6. Por tanto, mientras que en [1] se considera que Nicaragua tiene un mayor número de población en torno a 0 años de edad, en este caso, se considera que el mayor número de habitantes de ese país tiene en torno a 20 años de edad.

En la Figura 5.14 se representan la curva correspondiente a Nicaragua y las medias de los grupos I y II.

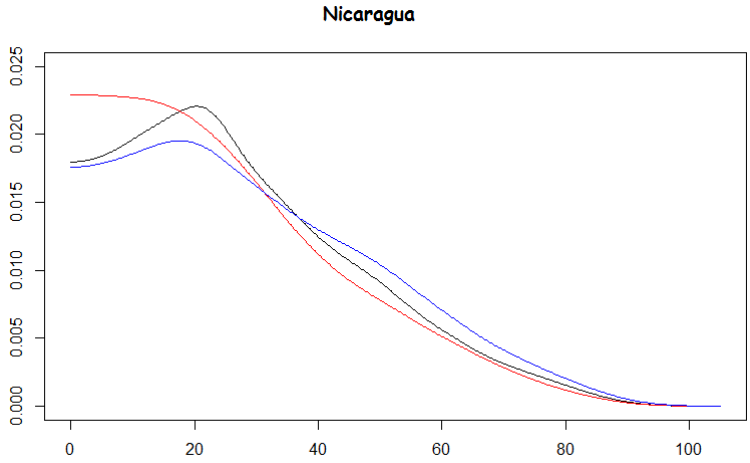

Figura 5.14: Función de Nicaragua (curva negra), media del Grupo I (curva roja) y media del Grupo II (curva azul).

La Figura 5.14 muestra una mayor similitud entre las curvas de Nicaragua y la curva (azul) correspondiente al Grupo II. Además, la distancia L_2 de la función de Nicaragua a la curva azul es 0.000715 y a la curva roja es 0.00134.

Visualmente, la curva de Colombia también se parece más a la curva azul que a la roja y, de hecho, la distancia a esta última (0.000668) es casi el doble que la distancia a la azul (0.000393).

La situación de Uruguay, a la vista de la Figura 5.16, no es tan clara. Sin embargo, su distancia L_2 a la curva azul (0.000727) es más de un orden de magnitud inferior que su distancia a la roja (0.012945).

Por lo tanto, se puede concluir que las clasificaciones obtenidas en este trabajo son más correctas las de [1].

En las Figuras 5.15 y 5.16 se presenta el resultado anterior.

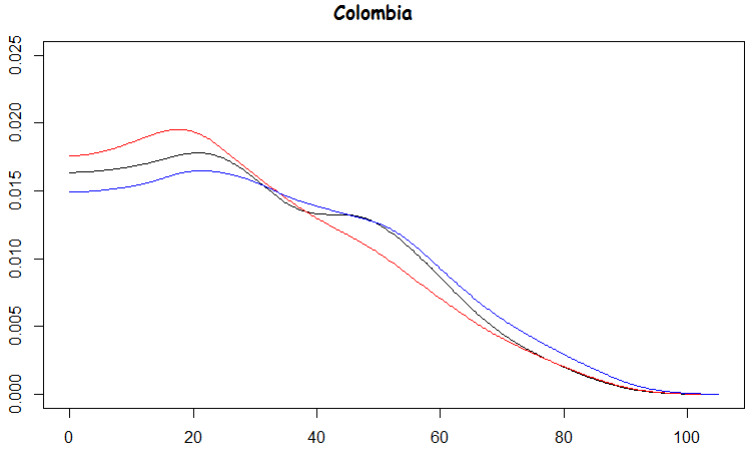

Figura 5.15: Función de Colombia (curva negra), media del Grupo II (curva roja) y media del Grupo III (curva azul).

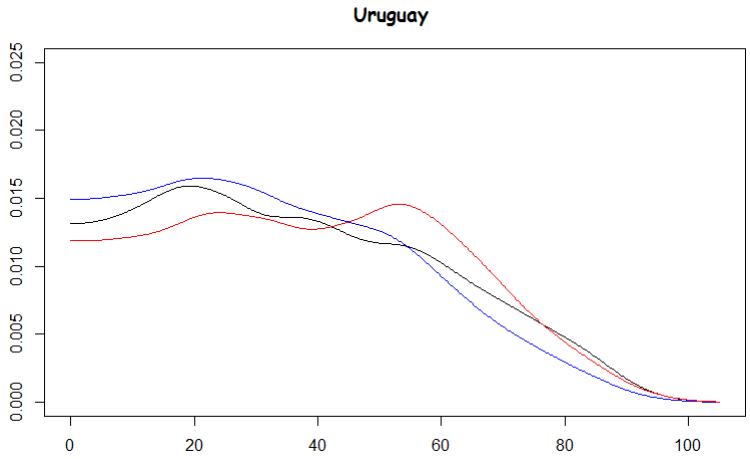

Figura 5.16: Función de Uruguay (curva negra), media del Grupo IV (curva roja) y media del Grupo III (curva azul).

Por último, las tres últimas diferencias tienen relación con países que han sido recortados utilizando ambas metodologías: Canadá (recortada en [1]), Cuba y Trinidad y Tobago (con nuestro trabajo). En los tres casos, cuando el país no está recortado, se le adjudica al Grupo IV. Por ello, se calcula la distancia de cada uno de estos a la media de este grupo y se representa en la Figura 5.17 la media del Grupo IV (curva negra), las curvas de Canadá (magenta), Cuba (naranja) y Trinidad y Tobago (marrón) junto con las del resto de países pertenecientes a este grupo (en gris).

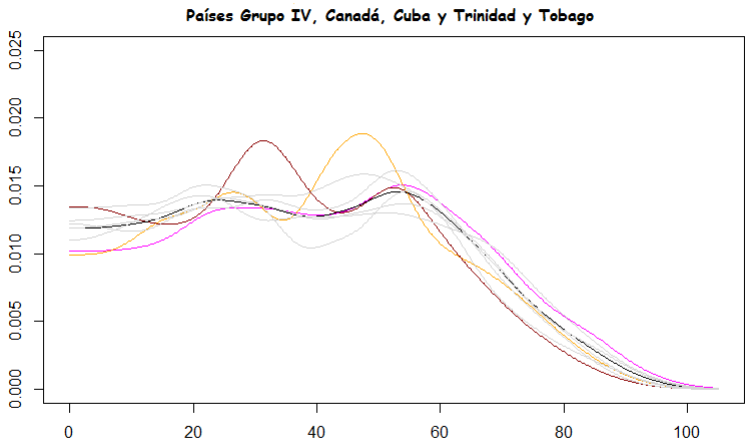

Figura 5.17: Media del Grupo IV (curva negra) y curvas de los países Canadá, Cuba y Trinidad y Tobago (curvas magenta, naranja y marrón) junto con las curvas del resto de países del Grupo IV.

Visualmente, da la impresión de que Cuba y Trinidad y Tobago se alejan claramente del grupo por la posición de los máximos, mientras que Canadá se parece bastante más.
Respecto de las distancias a la media (Canadá: 0.000465, Cuba: 0.00161 y Trinidad y Tobago: 0.00157) se observa que Canadá esta a algo menos de un tercio que las otras dos. Las distancias del resto de países del Grupo IV se muestran en la Tabla 5.7.

Orden	País	Distancia
1	EE.UU	0.000070
2	Aruba	0.000200
3	Puerto Rico	0.000332
4	Canadá	0.000465
5	Curacao	0.000540
6	Barbados	0.000617

Tabla 5.7: Distancias a la media de los países integrantes del Grupo IV.

Se comprueba que Canadá ocupa el lugar 4 en la Tabla 5.7. En cambio, Cuba y Trinidad y Tobago están claramente por encima de la mayor de las distancias restantes. Por ello, parece razonable el recorte de sólo estos dos últimos.

Por lo tanto, parece que la metodología empleada en este trabajo permite observar mejores resultados que la basada en funciones cuantiles usada en [1]. Las ideas se exponen a continuación:

1. *Elección del número de agrupaciones*: en los dos trabajos se ha encontrado el mismo número de agrupaciones para la clasificación de los países, $k = 4$.

2. *Elección del número de países recortados y comparación de los grupos*: en [1] se han recortado dos países, mientras que, en este, se han eliminado tres. Además, en el resultado de la clasificación de los países en cada grupo se han encontrado diferencias utilizando ambas metodologías de trabajo. Finalmente, justificando cada discrepancia obtenida, se ha concluido que las agrupaciones obtenidas han sido mejores que en [1].

6 Apéndice

6.1. Demostración del Teorema 3.1.3

Sean $g \epsilon L_2$ y $f_1,\ldots,f_n \epsilon\ L_2$. Entonces

$$\sum_{i=1}^{n} ||f_i - g||^2 = \sum_{i=1}^{n} \left(\int |f_i(t) - g(t)|^2 dt \right) = \int \left(\sum_{i=1}^{n} |f_i(t) - g(t)|^2 \right) dt$$

Sea $t_0 \epsilon\, [0, 105]$ y consideramos la función $h(a)=\sum_{i=1}^{n} |f_i(t_0)-a|^2, a \epsilon \Re$. Se deriva y para buscar el mínimo se iguala a cero:

$$\frac{d}{da} h(a) = -2 \sum_{i=1}^{n} (f_i(t_0) - a) = 0$$

Un simple cálculo permite concluir que esta función sólo se anula en $a_0=\frac{1}{n} \sum_{i=1}^{n} f_i(t_0)$.

Por otro lado, queda por demostrar que la función donde se anula la primera derivada, es un mínimo.

Para demostrar que $\bar{f}$ es un mínimo, se calcula la segunda derivada de h, h'', y se sustituye el valor de la media $\bar{f}$. Se tiene que

$$h''(a) = 2n > 0, \forall\ n > 0$$

Por lo tanto, $\bar{f}$ es el mínimo que estamos buscando. Además, como $f_i \epsilon L_2$, i=1,. . . ,n, entonces $\sum_{i=1}^{n} f_i \epsilon L_2$ y, por tanto, $\bar{f} \epsilon L_2$.

6.2. Convergencia en probabilidad: Definición

Sea $\{T_n\}$ una serie de estimadores de θ y $\{X_1, ..., X_n\}$ una muestra de variables aleatorias. T_n es consistente en probabilidad si

$$T_n(X_1, .., X_n) \rightarrow_{c.p} \theta$$

o lo que es lo mismo,

$$T_n(X_1, .., X_n) - \theta \rightarrow_{c.p} 0$$

6.3. Demostración del Teorema 6.3.1

Teorema 6.3.1 *Si* $\sup|\hat{f}_{n,h_n}(t) - f(t)| \longrightarrow_{c.p} 0$ *(convergencia en probabilidad) cuando* $n \longrightarrow \infty$ *y* $a, b \epsilon \Re$ *con* $a \leq b$ *entonces* $d(\hat{f}_{n,h_n}, f) = \int_a^b |\hat{f}_{n,h_n}(t) - f(t)|^2 dt \longrightarrow_{c.p} 0$ *(convergencia en probabilidad) cuando* $n \longrightarrow \infty$.

Demostración: Sea $C_n = \sup|\hat{f}_{n,h_n}(t) - f(t)|$. Entonces, por hipótesis, $C_n^2 = \sup|\hat{f}_{n,h_n}(t) - f(t)|^2 \longrightarrow 0$. Además,

$$\int_a^b |\hat{f}_{n,h_n}(t) - f(t)|^2 dt \leq \int_a^b \sup|\hat{f}_{n,h_n}(t) - f(t)|^2 dt = C_n^2(b-a) \longrightarrow 0$$

porque $C_n^2 \longrightarrow 0$ y $(b-a)\epsilon\Re$.

6.4. Instrucciones en R del Ejemplo 4.1.3:

♯Datos a representar: una muestra de tamaño 2000 de una mezcla de dos distribuciones normales de media 0, desviación tipica 1 y de media 3.5, desviacion típica 2, respectivamente, de 1000 observaciones cada una.

```
X=c(rnorm(1000),rnorm(1000,m= 3.5,sd= 2)) ♯Se generan los datos.
```

```
I=seq(from=-4,to=8,by=.001) ♯I es conjunto del eje de abscisas en el que se van a representar los datos.
```

```
FI=.5*dnorm(I)+.5*dnorm(I,m=3.5,sd=2) ♯Es la función de densidad de la distribución que generó la muestra en los puntos I.
```

```
lines(I,FI) ♯Dibuja la función FI en los puntos de I (ver Figura 4.1: curva roja).
```

♯Se dibujan las funciones estimador de la función real con diferentes anchos de ventana (ver Figura 4.1.

```
plot(density(X,bw=0.1,kernel='gaussian')) ♯Ancho de ventana: 0.1
```

```
plot(density(X,bw=0.5,kernel='gaussian')) ♯Ancho de ventana: 0.5
```

```
plot(density(X,bw=1,kernel='gaussian')) ♯Ancho de ventana: 1
```

♯Finalmente, se dibuja el histograma de los datos (ver Figura 4.2).

```
hist(X,freq=FALSE,breaks=70)
```

6.5. Obtención de la expresión del ancho de ventana óptimo

Si se supone que se cumplen las hipótesis adecuadas (en 2.5. Asymptotic MSE and MISE approximations, [9]), entonces el $MISE$ *se minimiza tomando*

$$h_{n,opt} = k_2^{-\frac{2}{5}} \left(\int K(t)^2 dt \right)^{\frac{1}{5}} \left(\int f''(x)^2 dt \right)^{-\frac{1}{5}} n^{-\frac{1}{5}}$$

Demostración:

Para esta demostración se seguirá, esencialmente, el desarrollo utilizado en [9] (sección *2.5. Asymptotic MSE and MISE approximations*).

Se utilizarán las expresiones 4.2, 4.3 y 4.4 y, para ello, dado $y \epsilon \Re$, se necesita calcular $E[\hat{f}(y)]$ y $E[\hat{f}(y)^2]$. Aplicando la definición de $\hat{f}$ (ver 4.1.1) se tiene que:

$$\begin{aligned} E[\hat{f}(y)] &= E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y-X_i}{h}\right)\right] \\ &= \frac{1}{nh}\sum_{i=1}^{n} E\left[K\left(\frac{y-X_i}{h}\right)\right] \\ &= \frac{1}{h}\int K\left(\frac{y-x}{h}\right) f(x)dx \qquad (6.1) \\ E[\hat{f}(y)^2] &= \frac{1}{n^2h^2} E\left[\sum_{i=1}^{n}\left(\sum_{j=1}^{n} K\left(\frac{y-X_i}{h}\right) K\left(\frac{y-X_j}{h}\right)\right)\right] \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n} E\left[K\left(\frac{y-X_i}{h}\right)^2\right] \\ &\quad + \frac{1}{n^2h^2}\sum_{i\neq j} E\left[K\left(\frac{y-X_i}{h}\right) K\left(\frac{y-X_j}{h}\right)\right] \qquad (6.2) \\ &= \frac{1}{nh^2}\int K\left(\frac{y-x}{h}\right)^2 f(x)dx \\ &\quad + \frac{1}{n^2h^2} C_n E\left[K\left(\frac{y-X_1}{h}\right) K\left(\frac{y-X_2}{h}\right)\right] \qquad (6.3) \end{aligned}$$

donde el paso de (6.2) a (6.3) es justificado por ser X_i variables independientes e igualmente distribuidas. Como X_1 y X_2 son variables independientes, también lo son las variables $K\left(\frac{y-X_1}{h}\right)$ y $K\left(\frac{y-X_2}{h}\right)$ y, por lo tanto, la esperanza de su producto es igual al producto de sus esperanzas. Por otro lado, X_1 y X_2 también son variables igualmente distribuidas, por ello, también lo son las variables $K\left(\frac{y-X_1}{h}\right)$ y $K\left(\frac{y-X_2}{h}\right)$, y por tanto, la esperanza de las variables $K\left(\frac{y-X_1}{h}\right)$ y $K\left(\frac{y-X_2}{h}\right)$ es la misma.

$$E\left[K\left(\frac{y-X_1}{h}\right) K\left(\frac{y-X_2}{h}\right)\right] = \left(\int K\left(\frac{y-x}{h}\right) f(x)dx\right)^2 \qquad (6.4)$$

Además, $C_n{=}n(n-1)$ es el número de sumandos que hay en (6.2) y de (6.4) se tiene que

$$\begin{aligned} E[\hat{f}(y)^2] &= \frac{1}{nh^2}\int K\left(\frac{y-x}{h}\right)^2 f(x)dx \\ &\quad + \frac{1}{n^2h^2} n(n-1)\left(\int K\left(\frac{y-x}{h}\right) f(x)dx\right)^2 \qquad (6.5) \end{aligned}$$

Por tanto, con (6.1) en (4.3) del Capítulo 3, se tiene que

$$Bias_h(x) = E\left[\hat{f}(x)\right] - f(x) = \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy - f(x) \qquad (6.6)$$

Por otro lado, (6.1) y (6.5) junto con (4.4) dan

$$\begin{aligned} nVar(\hat{f}(x)) &= n\left(E\left[\hat{f}(x)^2\right] - \left(E[\hat{f}(x)]\right)^2\right) \\ &= \int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left(\int \frac{1}{h}K\left(\frac{x-y}{h}\right) f(y)dy\right)^2 \\ &= \int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - [f(x) + Bias_h(x)]^2 \end{aligned}$$

y, de aquí,

$$Var(\hat{f}(x)) = \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \frac{1}{n}[f(x) + Bias_h(x)]^2 \tag{6.7}$$

Ahora, se hace el cambio de variable $y = x - ht$ en (6.6) y (6.7) y se utilizan las series de Taylor hasta el orden 4 para

$$f(x-ht) = f(x) - htf'(x) + \frac{1}{2}h^2t^2f''(x) - \frac{1}{6}h^3t^3f'''(x) + \frac{1}{24}h^4t^4f^{iv)}(x_t)$$

donde $x_t \epsilon (x, x-ht)$. Se tiene, entonces

$$\begin{aligned} Bias_h(x) &= \int K(t)f(x-ht)dt - f(x) \\ &= \int K(t)\left[f(x-ht) - f(x)\right]dt \\ &= \int K(t)\left(-htf'(x) + \frac{1}{2}h^2t^2f''(x) - \frac{1}{6}h^3t^3f'''(x) + \frac{1}{24}h^4t^4f^{iv)}(x_t)\right)dt \\ &= \frac{1}{2}h^2\left(\int t^2K(t)dt\right)f''(x) + \frac{1}{24}h^4\left(\int t^4K(t)dt\right)f^{iv)}(x_t) \\ &= \frac{1}{2}h^2\int K(t)t^2dtf''(x) + o(h^2) \end{aligned} \tag{6.8}$$

donde la segunda igualdad viene de que $\int K(t)dt = 1$, la cuarta viene de que K es una función simétrica con respecto de 0 [(4.1), Capítulo 3] y la última viene de que, por hipótesis, f^{iv} está acotada y $\int t^4K(t)dt < \infty$ (ver [9]).

Como $\lim_{n\to\infty} h_n = 0$ y se había denominado $h = h_n$, se tiene que, de (6.8) y (6.7) con un razonamiento similar al desarrolado en (6.8),

$$\begin{aligned} Var(\hat{f}(x)) &= \frac{1}{nh}\int K(t)^2 f(x-ht)dt - \frac{1}{n}[f(x) + O(h^2)]^2 \\ &= \frac{1}{nh}\int K(t)^2[f(x) - htf'(x) + ...]dt + O(n^{-1}) \\ &= \frac{1}{nh}f(x)\int K(t)^2dt + O(n^{-1}) \end{aligned}$$

Como f es una función de densidad de probabilidad, la integración sobre x de la varianza da la aproximación siguiente y bajo las condiciones del *Theorem 2.1.7* en [11] se tiene que

$$\int Var(\hat{f}(x))dx \approx \frac{1}{nh}\int K(t)^2dt \tag{6.9}$$

y, de aquí, siguiendo el razonamiento del *Theorem 2.1.7* de [11] se obtiene la expresión del $MISE$ donde k_2=$\int t^2 K(t)dt$.

$$\begin{aligned} MISE(\hat{f}(x)) &= \int Var(\hat{f}(x))dx + \int Bias_h^2(x)dx \\ &\approx \frac{1}{nh}\int K(t)^2dt + \frac{1}{4}h^4k_2^2\int f''(x)^2dx \end{aligned} \tag{6.10}$$

Se busca el h que minimiza (6.10) derivando los dos términos principales respecto de h e igualando a cero:

$$-\frac{1}{nh^2}\int K(t)^2dt + h^3k_2^2\int f''(x)^2dx = 0$$

de donde se obtiene

$$h_{n,opt} = k_2^{-\frac{2}{5}}\left(\int K(t)^2dt\right)^{\frac{1}{5}}\left(\int f''(x)^2dt\right)^{-\frac{1}{5}}n^{-\frac{1}{5}}$$

Por lo tanto, la ventana óptima es de orden $n^{-\frac{1}{5}}$ y depende de $\int K(t)^2dt$ (que no plantea problemas porque K es elegido por el estadístico) y de la cantidad desconocida $\int f''(x)^2dx$. Este problema se resuelve en la práctica haciendo una situación preliminar de f que permita estimar esta cantidad.

Bibliografía

[1] E. del Barrio, J.A. Cuesta-Albertos, C. Matrán, and A. Mayo-Íscar. Robust clustering tools based on optimal transportation. *arXiv preprint arXiv:1607.01179*, pages 14–19, 2016.

[2] Wikipedia. Historia de la estadística. https://es.wikipedia.org/wiki/Historia_de_la_estad‡C3‡ADstica.

[3] J.A. Cuesta Albertos. *Análisis Multivariante*. Universidad de Cantabria, 2015.

[4] Manuel Gónzalez. *Análisis Funcional*. Universidad de Cantabria, 2015.

[5] J.A. Cuesta Albertos. Una técnica estadística para inferencia en paralelo. *Introducción al Python para Big Data*, 2016.

[6] Bernard W Silverman. *Density Estimation for Statistics and Data Analysis*. CRC press, 1986.

[7] Alicia Nieto Reyes. *Statistical Inference*. Universidad de Cantabria, 2015.

[8] Jenq-Neng Hwang, Shyh-Rong Lay, and Alan Lippman. Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, 42(10):2795–2810, 1994.

[9] Jones M-C Wand, M.P. *Kernel Smoothing*. CRC press, 1995.

[10] Mathias Bourel. Comparación en la elección de una ventana óptima para algunos estimadores de densidad. *Memoria Investigaciones en Ingeniería*, (11), 2013.

[11] B.L.S. Prakasa Rao. *Nonparametric Functional Estimation:Teorema 2.1.7*. SIAM, 1983.