USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Estimación de conjuntos de nivel para el COVID-2019

Laura Lueiro Fernández

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Estimación de conjuntos de nivel para el COVID-2019

Laura Lueiro Fernández

Septiembre, 2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística
Título: Estimación de conjuntos de nivel para el Covid-2019
Breve descrición do contido
En este trabajo nos ocuparemos del problema de la estimación no paramétrica de conjuntos de nivel mediante el método plug-in. Ilustraremos su funcionamiento en la práctica reconstruyendo los focos de mayor incidencia de contagios por COVID-19 en Estados Unidos durante 2020.
Recomendacións
Outras observacións

Índice general

Resumen

La pandemia producida por el COVID-19 está causando grandes estragos económicos, sociales y sanitarios a lo largo del mundo. En vista de la falta de recursos, encontrar aquellas zonas más perjudicadas permitiría enviar ayudas localizadas y frenar al virus más rápida y eficazmente. Este será el objetivo de este trabajo en el que se reconstruirán concretamente los focos de contagios por coronavirus en Estados Unidos mediante la estimación no paramétrica de conjuntos de nivel. En particular, examinaremos aquellos conjuntos que superen cierto umbral de probabilidad denominados regiones de alta densidad. Reconstruirlos supone estimar la función densidad, un problema bien conocido en la Estadística No Paramétrica. Para ello, el estimador más utilizado es el de tipo núcleo en cuya definición aparece una matriz ventana H que debe especificar el usuario y que influye notablemente en el resultado. Se revisará la teoría de cuatro métodos diferentes de selección de H: método de validación cruzada por mínimos cuadrados, método plug-in de Duong y Hazelton, método de validación cruzada sesgada y método de validación cruzada suavizado. Posteriormente, se aplicarán tres de estos métodos a nuestros datos y se compararán las soluciones obtenidas.

Abstract

The COVID-19 pandemic is wreaking great economic, social and health havoc across the world. Due to the lack of resources, targeting the worst-hit areas would allow us to send localized aid and stop the virus more quickly and effectively. This will be the objective of this work in which coronavirus outbreaks in the United States will be reconstruct using nonparametric estimation of level sets. Specifically, sets that exceed a certain probability threshold called high density regions will be examined. Recomposing them involves estimating the density function, a well-known problem in Nonparametric Statistics. For this purpose the most commonly used estimator is the kernel type, whose definition includes a

window matrix that must be specified by the user and which has a significant influence on the result. The theory of four different methods of selecting H will be reviewed: least squares cross validation method, Duong and Hazelton plug-in method, biased cross validation method and smoothed cross validation method. Three of these will then be applied to our data and the solutions obtained will be compared.

Capítulo 1

Introducción

En Diciembre de 2019 estalla en Wuhan (China) el primer brote de la enfermedad que se convertiría en una pandemia mundial, la COVID-19. Esta patología causada por los virus coronavirus afecta tanto a humanos como a animales provocando infecciones respiratorias que pueden ir desde el resfriado común hasta cuadros graves. Algunos de sus síntomas más frecuentes son tos seca, fiebre y cansancio mientras que los más preocupantes son dolor en el pecho, dificultad para respirar o incluso pérdida del habla o del movimiento. Esta afección se complica normalmente en personas mayores o que padecen dolencias médicas previas como hipertensión arterial, problemas cardiacos o pulmonares, diabetes o cáncer. El coronavirus se propaga fácilmente a través de las gotas de saliva o de secreciones nasales que creamos al toser o estornudar. Por ello ha sido necesaria una fabricación masiva de instrumental de protección. Hasta Febrero de 2021, se ha informado de más de 104.3 millones de contagios en 255 países en el mundo. Para más detalles ver [1].

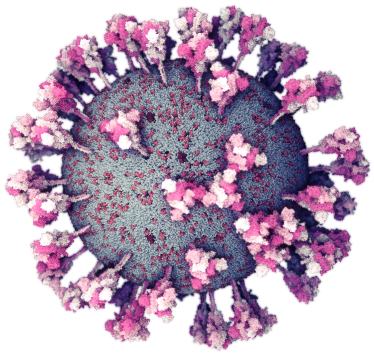

Figura 1.1: Primera imagen real en tres dimensiones del coronavirus SARS-CoV-2 en su membrana externa creada por Nanographics.

Uno de los principales problemas provocados por la epidemia es el colapso del sistema sanitario. La falta de camas, disponibilidad en las UCI (Unidad de Cuidados Intensivos), personal médico, pruebas de detección del virus, mascarillas u otros materiales inmovilizó estructuras hospitalarias completas. Esta situación no solo afectó a los contagiados por el virus sino a todo tipo de pacientes. La OMS (Organización Mundial de la Salud) informó de un aumento brusco de muertes por enfermedades tratables. Según [2], entre Marzo y Noviembre de 2020, se han producido más de 450.000 defunciones, por cualquier causa, con respecto al mismo período en los años 2016-2019 en la UE (Unión Europea).

El país con más fallecidos a causa del COVID-19 ha sido Estados Unidos. En [3] se han contabilizado un total de 26.722.300 infectados y 455.800 muertes. El Centro Nacional de Estadísticas de Salud informó de que entre Marzo y Julio de 2020 el número de defunciones había aumentado un 20 % sobre lo esperado. De acuerdo con [4], los estados más afectados fueron Nueva York, Nueva Jersey, Mississippi, Maryland y Luisiana entre otros.

De la misma forma que el virus provocó grandes problemas físicos en la población mundial también causó graves daños psicológicos. El miedo, el aislamiento, el duelo y la pérdida de ingresos generaron y agravaron los trastornos mentales. A esto hay que sumarle que la propia enfermedad puede traer consigo problemas cerebrovasculares, complicaciones neurológicas o estados delirantes. Las estadísticas muestran también un aumento notable del consumo de drogas y alcohol.

La pandemia paralizó los servicios de salud mental esenciales en el 93 % de países mientras a su vez aumentaba su demanda. Es bien sabido que los problemas en este campo son evidentes desde hace tiempo. La OMS emitió antes del comienzo del brote críticas acerca de cómo gestionan la mayoría de territorios los presupuestos nacionales de salud de los cuales solo se destina un 2 % a este ámbito. Esta misma organización llevó a cabo un estudio entre Junio y Agosto de 2020 que incluía 130 países con el objetivo de demostrar las alteraciones sufridas por este sector.

En China, Huang Jizheng y colaboradores evaluaron el estado psicológico del personal de diversos hospitales mediante un análisis descriptivo entre el 7 y el 14 de Febrero de 2020. Esta investigación contó con 246 participantes entre los que había médicos y enfermeros. Se les aplicó la escala de autoevaluación para la ansiedad (SAS) y la escala de autoevaluación para el trastorno de estrés postraumático (PTSD-SS). Como se puede comprobar en [5], los resultados fueron alarmantes: la tasa de ansiedad fue del 23,04 % mientras que la del estrés del 27,39 % .

A raíz de la catástrofe sanitaria tuvo lugar un impacto socioeconómico global causado por la detención del comercio, la hostelería y el turismo. Las restricciones provocaron un aumento del desempleo y de trabajadores ausentes. La economía mundial exhibió en 2020 una caída del PIB (Producto Interior Bruto) mayor a la observada en varias décadas.

En España en 2020, como se muestra en [6], el PIB descendió un 9,9 %, la pérdida de puestos de trabajo a tiempo completo fue de 962.000 y la renta nacional bajó un 7.7 % con respecto a 2019. A nivel de la Unión Europea la disminución de ingresos laborales medios estimada en [7] fue del 5,2 % comparando estos mismos años. Mundialmente, el desempleo aumentó de 33 millones a 220 en 2020 superando las cifras de la crisis financiera de 2009 y afectando mayormente a las mujeres que a los hombres con unas cifras del 5 % y 3.9 % respectivamente. De nuevo, una de las zonas con peores cifras es Estados Unidos donde se batió un récord histórico de solicitudes de ayudas por paro: 10 millones en tan solo dos semanas.

En otros países como China a pesar de la fuerte recesión se supo controlar la situación rápidamente. En los primeros tres meses de pandemia perseveró un desarrollo económico y social estable. Las exportaciones y la industria crecieron deprisa, el consumo se recuperó de forma constante y las cifras de empleo se mantuvieron invariantes. Si se desea tener más información ver [8] y [9].

En América Latina y el Caribe el COVID-19 impactó en un período difícil. La tasa de crecimiento del PIB había disminuido del 6 % al 0,2 % entre 2010 y 2019. En estas condiciones, la epidemia podría llegar a causar la mayor crisis económica y social del lugar desde el año 1900. El desplome de la actividad financiera mundial perjudicó a estos territorios especialmente a través del comercio de materias primas. El motivo es que la industria de la zona está inmersa en cadenas globales de valor en las que los Estados Unidos y China juegan un papel fundamental. Dada la nueva situación económica de los principales socios de la región y la devaluación de la exportación los precios disminuyeron un 8.8 %. A su vez los flujos de remesas (de las cuales un 90 % se emplean para necesidades básicas) hacia Latinoamérica y el Caribe se redujeron notablemente. Esto produjo graves efectos en la incidencia de la pobreza y en el consumo. Se pueden consultar estos datos en [10].

Contra estas y otras muchas complicaciones se enfrenta la población mundial con el objetivo común de detener la propagación del virus. Ubicar los territorios más afectados por el coronavirus y analizar cómo evolucionan estos en tiempo real ayudaría a repartir los recursos disponibles en función de su necesidad. Por ejemplo, se podrían enviar más

suministros sanitarios y profesionales médicos a las zonas colapsadas o administrar ayudas económicas a los lugares más dañados por la crisis. Es aquí donde juega un papel fundamental la Estadística.

La finalidad de este estudio será localizar los focos de contagios del COVID-19 en Estados Unidos, uno de los países más perjudicados por la pandemia. En Matemáticas, este problema puede resolverse a partir de la estimación no paramétrica de conjuntos de nivel la cual pertenece a la rama de estimación de conjuntos.

Esta teoría se ocupa de la reconstrucción de un conjunto desconocido o de alguna de sus características (volumen, borde...) a partir de una muestra aleatoria de puntos seleccionada dentro del conjunto. Aunque reconstruir el soporte de una distribución es un problema clásico de la estimación de conjuntos, aquí es más interesante estimar zonas de elevada concentración de probabilidad. En otras palabras, nos centraremos en la estimación de conjuntos de nivel para una función de densidad f.

Formalmente, la estimación de conjuntos de nivel lidia con el problema de componer a partir de una muestra aleatoria $X_1, \ldots, X_n$ de una densidad f el conjunto

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}$$

con $t \geq 0$. Como veremos en el Capítulo 2, esta definición cuenta con numerosos inconvenientes en la práctica, por lo que es necesario modificarla. Los conjuntos de nivel que finalmente usaremos serán aquellos de menor volumen que contengan una cierta probabilidad mayor o igual a $1 - \alpha$. Es decir, en las condiciones anteriores serán las regiones

$$R(f_\alpha) = \{x : f(x) \geq f_\alpha\}$$

donde f_α es la mayor constante tal que $P_X(X \in R(f_\alpha)) \geq 1-\alpha$. Se trata de las denominadas regiones de alta densidad o *highest density regions* (HDR).

En el presente documento estimaremos determinadas HDR de la función densidad de la localización de infectados por el virus en Estados Unidos. En concreto, aquellas en las que la incidencia sea preocupablemente alta. A modo de ilustración, en la Figura 1.2 aparecen representados los contornos de algunas regiones para diferentes valores de α. Observar que cuanto mayor es el valor de α menor será la probabilidad contenida en $R(f_\alpha)$. Es decir, para encontrar los picos del virus se deberá trabajar con valores de α cercanos a 1.

El conjunto de datos que emplearemos proviene del repositorio [11] de GitHub2 de la Universidad de John Hopkins. Se trata de las cifras de infectados por COVID-19 en Estados

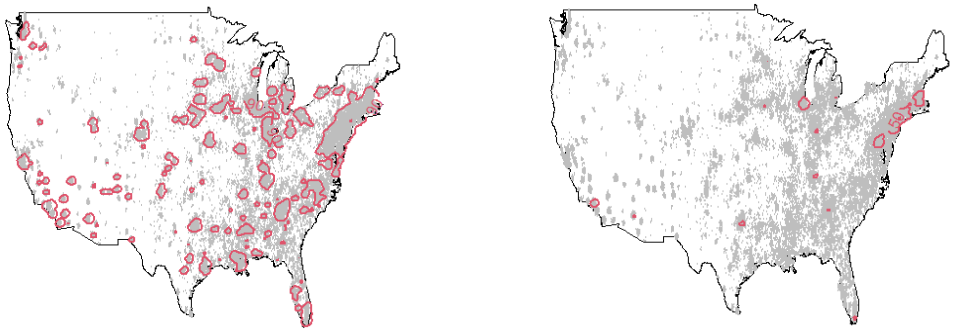

Figura 1.2: Contornos de las HDR´s para $\alpha = 0{,}10$ y $\alpha = 0{,}50$ respectivamente.

Unidos a nivel comarcal actualizadas a diario desde el 22 de Enero de 2020 hasta el 5 de Mayo de 2020. Esta información aparece estructurada en una tabla en la que podemos ver dicha cifra para cada estado americano y la ubicación de cada contagio en coordenadas geográficas. Hay que tener en cuenta que los casos confirmados incluyen, de acuerdo con las directrices de Centros para el Control y la Prevención de Enfermedades, los casos posibles y los casos positivos presuntivos. Para ver un ejemplo de cómo se distribuye la información en el mapa ver la Figura 1.3.

A principios del mes de Marzo se visualizan únicamente tres focos en Washington, Nueva York y California. Tan solo unos días después los infectados comienzan a aparecer en más estados como Michigan, Tennessee o Florida (sobre todo en el sur) y disminuyen en Washington. Desde medidados de Abril hasta principios de Mayo vemos cómo el virus se extiende mayormente por el sudeste del país. Aparecen multitud de nuevos brotes destacando el de la zona de Nueva York.

En lo que se refiere a la organización de este trabajo, será de la siguiente forma. En el Capítulo 1 se ha contextualizado brevemente la situación mundial desde la aparición del virus. Además, se ha introducido de manera concisa la teoría de la estimación de conjuntos y en concreto de conjuntos de nivel. Finalmente, se ha dado una idea del objetivo de este estudio y se han presentado e ilustrado los datos que se examinarán.

En el Capítulo 2 se profundizará en la estimación de conjuntos de nivel. Asimismo se dará una definición formal de las regiones de alta densidad mencionadas anteriormente que

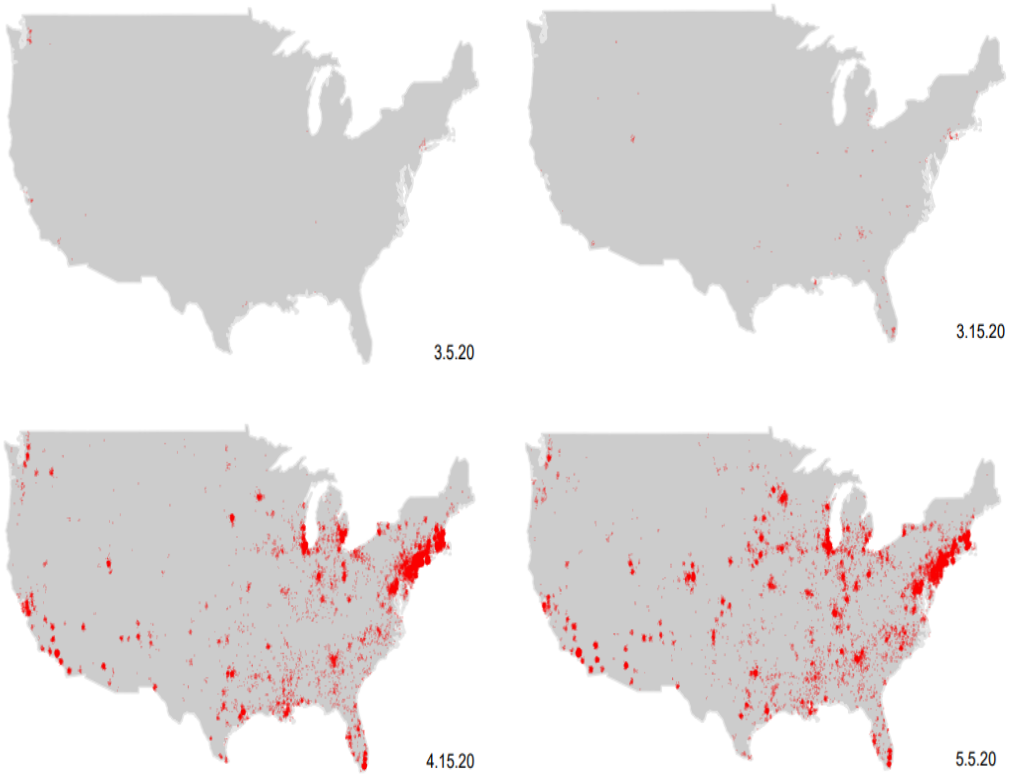

Figura 1.3: Coordenadas geográficas de los contagios en el mapa de Estados Unidos en los meses de Marzo, Abril y Mayo de 2020 respectivamente.

serán esenciales para el desarrollo de la investigación. Existen varias formas de reconstruir estas regiones pero en este caso hemos escogido el método plug-in por ser el que mejor se adapta a nuestros propósitos. Este método propone estimar la función densidad en la definición de los HDR. Para ello, se utiliza el conocido estimador de tipo núcleo el cual consta de diversos procedimientos de selección de la matriz ventana H. Cuatro de estos métodos selectores serán descritos en detalle: método plug-in de Duong y Hazelton y los criterios de validación cruzada sesgado (BCV), por mínimos cuadrados (UCV) y suavizado (SCV).Por último, compararemos sus rendimientos analíticamente estudiando sus tasas de convergencia.

En el Capítulo 3 pondremos en práctica la teoría expuesta en la sección anterior aplicándola a los datos comentados previamente en el Capítulo 1. A partir de las coordenadas geográficas de los contagios reconstruiremos la función densidad que los rige. Se utilizará el estimador tipo núcleo para el cual seleccionaremos la matriz ventana H de formas distintas.

Usaremos tres de los selectores de ventana vistos en el Capítulo 2 e ilustraremos las HDR obtenidas para cada uno sobre los mapas de Estados Unidos. Compararemos los resultados obtenidos para estudiar si existen diferencias entre ellos.

Finalizaremos este trabajo con el Capítulo 4 en el que se exponen las conclusiones finales a las que se llegan a partir de las soluciones alcanzadas.

Capítulo 2

Estimación de conjuntos de nivel

La estimación de conjuntos, examinada por primera vez en [12], es una rama de la estadística no paramétrica que aborda el problema de reconstruir un conjunto desconocido S en el espacio euclídeo $\mathbb{R}^d$ a partir de una muestra de puntos $X_1, \ldots, X_n$ seleccionados aleatoriamente en S. Una de sus propiedades más importantes es la fuerte motivación geométrica que esconde detrás. En esta teoría las formas de los conjuntos y las distancias que hay entre ellos juegan un papel fundamental. Por ello se considera como la contrapartida de la estimación funcional no paramétrica. Ver [13] para más detalles.

Destacan algunas de sus aplicaciones prácticas como el procesamiento de imágenes, la detección de un comportamiento anormal de un sistema estudiada en [14] o el análisis de clústers explicado en [15]. Si se desea ver una lista más amplia con sus respectivas referencias ver [16].

Tal y como se ha introducido en el Capítulo 1, $X_1, X_2 \ldots X_n$ denota una muestra aleatoria en $\mathbb{R}^d$ de una variable aleatoria X y nos referiremos a su distribución con P_X. Cuando X se asuma absolutamente continua, denotaremos por f su correspondiente densidad. También se designará por $\aleph n$ al conjunto de puntos de la muestra.

Uno de los ejercicios clásicos en estimación de conjuntos es reconstruir un soporte S de una distribución P_X en $\mathbb{R}^d$. Una de las dos manera de hacerlo es suponiendo restricciones en la forma de S (por ejemplo, limitarse a conjuntos convexos). En este contexto, Rényi y Sulanke señalaron en [17] y [18] que la envoltura convexa de la muestra $S_n = conv(< \aleph n >)$ resultaba ser un estimador bastante intuitivo. La otra forma de proceder es no restringir las propiedades geométricas de S definiendo estimadores más generales pero con peores velocidades de convergencia. Para este último caso Devroye y Wise proponen en [14] un estimador bastante natural. A partir de una muestra aleatoria $X_1, \ldots, X_n$ con distribución

P_X cuyo soporte es S, definen:

$$S_n = \bigcup_{n}^{i=1} B(X_i, \epsilon_n)$$

donde $B(X_i, \epsilon_n)$ denota la bola cerrada centrada en X_i con radio $\epsilon_n \geq 0$. De esta forma S_n no es más que una versión suavizada de la muestra. Algunos ejemplos ilustrados se pueden ver en la Figura 2.1 en [14].

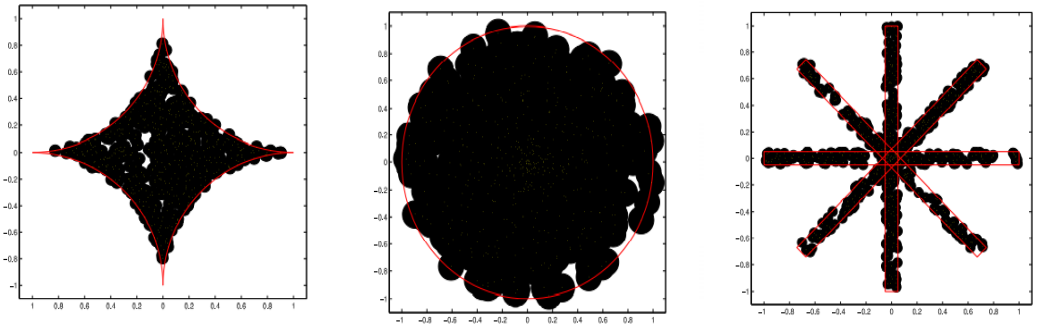

Figura 2.1: Algunos ejemplos ilustrados del estimador propuesto por Devroye y Wise.

De todos modos, en este trabajo no tiene interés estimar el soporte. El motivo es que en la mayoría de ocasiones hacerlo resulta nimio, pues gran parte de este es nulo desde el punto de vista de la probabilidad. Es decir, si f es la función de densidad subyacente, las partes de S en las que f tiene valores pequeños no son útiles en la práctica. Por ello nos centraremos en aquellos conjuntos de elevada concentración de probabilidad. Examinaremos la reconstrucción de conjuntos de nivel o HDR´s.

La estimación de conjuntos de nivel consiste en la reconstrucción, dada una muestra aleatoria de puntos $X_1, \ldots, X_n$ de una densidad f, del siguiente conjunto:

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\} \tag{2.1}$$

con $t > 0$. Advertir que el soporte de una función de densidad f es el conjunto de nivel $t = 0$, es decir, la estimación del soporte es un caso particular de la estimación de conjuntos de nivel.
Una importante aplicación práctica de este tipo de conjuntos se atribuye a John A. Hartigan quien relacionó en [15] las componentes de (2.1) con el concepto de clúster.

Algunos problemas de esta definición como su difícil interpretación en la praxis, que

está basada en un nivel t escogido por el usuario o que a priori no se conocen los posibles valores de $f(t)$, obligan a redefinir (2.1) en términos de probabilidades y no mediante un umbral de nivel. Una alternativa a la estimación de (2.1) será estimar regiones de "soporte substancial" que denominaremos regiones de alta densidad (HDR). En [20], Hyndman las define como sigue:

Definición 2.1. Dada una muestra aleatoria de puntos $X_1, \ldots, X_n$ de una densidad f se define la $100(1-\alpha)\,\%$ HDR como el subconjunto $R(f_\alpha)$ tal que:

$$R(f_\alpha) = \{x \in \mathbb{R}^d : f(x) \geq f_\alpha\}$$

donde f_α es la mayor constante tal que $P_X(X \in R(f_\alpha)) \geq 1-\alpha$.

Es decir, estimar HDR´s se basa en reconstruir los conjuntos de nivel que tienen una probabilidad contenida mayor o igual a $1-\alpha$. Esta será el concepto que emplearemos para localizar los focos del virus.

2.1. Método plug-in

Existen tres alternativas para calcular un estimador $\hat{R}(f_\alpha)$: método plug-in, método de exceso de masa y método de suavizado granulométrico. En este trabajo veremos únicamente el primero el cual se diferencia del resto en que no es necesario imponer restricciones geométricas sobre los estimadores. Dicho criterio es la respuesta más natural a esta cuestión. El método plug-in propone estimar $R(f_\alpha)$ de la forma:

$$\hat{R}(f_\alpha) = \{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\alpha\}$$

donde f_n denota el estimador tipo núcleo de f ,un concepto que veremos más adelante, y $\hat{f}_\alpha$ es el umbral estimado:

$$\hat{f}_\alpha = max\{t > 0 : P_X(\hat{G}(t)) \geq 1-\alpha\}$$

designando por $\hat{G}(t)$ a la reconstrucción plug-in del conjunto de nivel (2.1) dada por:

$$\hat{G}(t) = \{x \in \mathbb{R}^d : f_n(x) \geq t\}$$

El cálculo de $\hat{f}_\alpha$ ha sido desarrollado en [20] por Hydman quien indagó sobre la estimación de las HDR´s. Así como estas regiones para distribuciones discretas son simplemente el conjunto de elementos del espacio muestral con mayor probabilidad, para distribuciones

continuas resultan más complejas. Hydman intentó reconstruirlas mediante métodos numéricos de integración pero ninguno de ellos se extendía fácilmente al caso multivariante. Como alternativa, propuso una técnica de Monte Carlo que prescinde de la integración explícita y obtiene información sobre la probabilidad en ellas contenida .

Sea $Y = f(X)$ la variable aleatoria obtenida al transformar X mediante su propia función densidad. Se tiene que f_α es tal que $P_X(f(X)) \geq f_\alpha = 1 - \alpha$ y por tanto, es el cuantil α de Y. Su estimación se puede calcular a partir de un conjunto de variables aleatorias que comparten la misma distribución que Y. Veamos la forma de hacerlo.

Sea un conjunto de observaciones independientes $\{x_1, \ldots, x_n\}$ de la densidad $f(x)$. Luego, $\{f(x_1), \ldots, f(x_n)\}$ es otro conjunto del mismo tipo de la distribucion de Y. Sea $f_{(j)}$ el mayor de los $f_{(x_i)}$ tal que $f_{(j)}$ es el cuantil muestral $(\frac{j}{n})$ de Y. Entonces, se puede usar $f_{(j)}$ como estimador de f_α. Concretamente, se escoge $\hat{f}_\alpha = f_{(j)}$ cuando $j = |\alpha_n|$. Ya que de esta forma $\hat{f}_\alpha \to f_\alpha$ cuando $n \to \infty$ y por consiguiente $R(\hat{f}_\alpha) \to R(f_\alpha)$ cuando $n \to \infty$.

A continuación, introduciremos la definición del estimador tipo núcleo de la densidad. Empecemos recordando la definición de la función densidad que no es más que:

$$F'(x) = f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

Estimarla es una de las principales dificultades en la inferencia no paramétrica. Este problema dio lugar a las denominadas *técnicas de suavizado*. A pesar de existir multitud de métodos (ver por ejemplo [21]), sin duda el más conocido y utilizado es el estimador tipo núcleo. Su origen se remonta a las décadas de 1950 y 1960 cuando se empezó a investigar la estimación no paramétrica de la densidad para el caso univariante. Fix and Hodges fueron los primeros en examinar este problema en [22] llegando a un estimador similar al que más tarde fue propuesto en [23] por Rosenblatt:

$$f_h = \frac{F_n(x+h) - F_n(x-h)}{2h}$$

donde F_n es la distribución empírica y $h = h_n$ es una función de la muestra de tamaño n que se aproxima a 0 cuando $n \to \infty$. Señalar que F_n se define como:

$$\hat{F}(x) = F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X_i \leq x\}}$$

La generalización que hizo Parzen en [24] del estimador de Rosenblatt dio lugar al estimador tipo núcleo en el contexto univariante:

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$$

donde $K_h(z) = K(\frac{z}{h})\frac{1}{h}$, siendo K una función densidad simétrica llamada núcleo y $h > 0$ un parámetro ventana también llamado parámetro de suavizado o ancho de banda por algunos autores. Se podría interpretar este estimador como el conjunto de "protuberancias" situadas en las observaciones donde K determinaría sus formas y h sus asperezas. Esta idea recuerda al célebre histograma cuya estructura es muy similar solo que en lugar de "protuberancias" aparecen "cajas". A continuación, veremos que efectivamente estos conceptos están vinculados siendo el histograma un caso particular del estimador tipo núcleo.

El histograma fue el primer estimador no paramétrico de la función densidad. Aunque no se conoce con certeza su fecha de invención exacta, según diversos estudios lo más probable es que se remonte al siglo XVII. Demos una definición formal del mismo.
Sea $X_1, \ldots, X_n$ una muestra aleatoria con función densidad f contenida en un intervalo (a, b). Sea la partición t_k del intervalo en M intervalos $B_k = [t_{k-1}, t_k)$, tal que $a = t_0 < t_1 < \ldots < t_M = b$. Se denota el ancho de la caja B_k como $h_k = t_k - t_{k-1}$ y el número de X_i que se encuentran en B_k como v_k, tal que $\sum_{k=1}^{M} v_k = n$. Entonces el histograma viene dado por:

$$\hat{f}_H(x) = \frac{v_k}{n(t_k - t_{k-1})}, x \in B_k$$

y cero fuera de $[a, b)$.

El parámetro t_0 tiene un gran efecto visual sobre todo para tamaños de muestra pequeños. Para un cierto h existe un número ilimitado de posibles t_0 en el intervalo $(a - h, a]$ y su elección en algunos casos resulta problemática. Con motivo de mejorar la estimación y eliminar el impacto de este parámetro, Scott propuso en [25] la siguiente idea. Fijados un h y un m entero positivo se construye una colección de m histogramas cada uno con ancho de banda h pero con puntos iniciales $t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \ldots, \frac{(m-1)h}{m}$. La notación anterior para B_k y v_k se mantiene refiriéndonos ahora a estos nuevos histogramas. El promedio de todos ellos da como resultado un histograma suavizado con ancho de banda $\delta = \frac{h}{m}$ denominado histograma cambiado promedio (ASH):

$$\hat{f}_A(x) = \frac{1}{m}\sum_{j=1-m}^{m-1} \frac{(m - |j|v_{k+j})}{nh}$$

Cuando $m \to \infty$ y v_k vale o 0 o 1, el ASH se puede escribir de forma equivalente como:

$$\hat{f}_K(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

con $K(t) = [1 - [t]]_+$ donde $[x]_+$ denota la parte positiva de x o cero. Esto no es más que un caso particular del estimador tipo núcleo con una función K escogida específicamente.

El estimador tipo núcleo para el caso univariante es una excelente herramienta para la representación de distribuciones de datos. Por ello ha recibido en la literatura mucha más atención que el caso multivariante ya que visualizar funciones densidad de un elevado número de dimensiones es complicado. Ahora bien, particularmente el caso bivariante resulta llamativo pues posee las ventajas prácticas del univariante y las teóricas del multivariante que nos permiten comprender mejor los aspectos del suavizado de esta técnica. En este texto los datos que utilizaremos se encuentran en $\mathbb{R}^2$. Por esa razón, a partir de ahora nos centraremos en este estimador para el caso multivariante.
En la Definición 2.2 se define el estimador tipo núcleo para el caso general d-dimensional como sigue:

Definición 2.2. Sea $X_1, \cdots, X_n$ una muestra aleatoria d-dimensional con función de densidad f, el estimador tipo núcleo se define como:

$$f_{nH}(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)$$

donde $x = (x_1, x_2, \ldots, x_d)^T$ y $X_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T$, $i = 1, 2, \ldots, n$. Ahora el núcleo K es una función densidad simétrica y esférica, $K_H(z) = K(zH^{\frac{-1}{2}})|H|^{\frac{-1}{2}}$ y H es la matriz ancho de banda $d \times d$, simétrica y definida positiva.

La extensión del ASH a los datos multivariantes y en concreto a los bivariantes es sencilla. Se construye el número de histogramas 2-dimensionales o multidimensionales deseado desplazándolos todos una misma distancia a lo largo de los ejes coordenados y luego se promedian juntos. En la Figura (2.2) se puede contemplar un ejemplo.

2.2. Selección parámetro ventana

En los estimadores tipo núcleo, así como la elección de K no es demasiado influyente, elegir la matriz ventana es crucial respecto a la estimación obtenida. Dependiendo de los métodos de selección utilizados se pueden conseguir resultados muy diferentes. A modo de ejemplo, observar la Figura (2.3) en [26]. Las entradas de la matriz H tienen un claro significado estadístico: las diagonales proporcionan la varianza de cada una de las dos coordenadas de la distribución y el resto representan la covarianza que hay entre ellas.

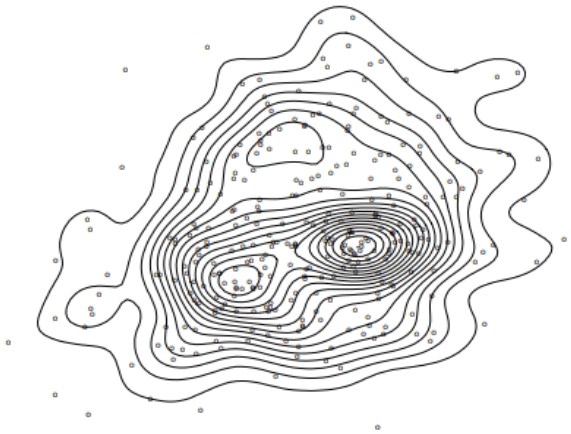

Figura 2.2: Ejemplo de ASH bivariante creado a partir de unos datos lipídicos.

Para el caso univariante existe un gran número de técnicas de selección del ancho de banda ver [27]. La idea original para extenderlas al caso multivariante de manera sencilla fue imponer condiciones sobre H. Inicialmente esta matriz se restringía a la clase $\mathcal{A} = \{h^2 I_d : h > 0\}$ de un escalar positivo multiplicando a la matriz identidad I_d o a la clase $\mathcal{D} = diag(h_i^2, \ldots, h_d^2 : h_1, \ldots, h_d > 0)$ de matrices diagonales definidas positivas. No obstante, se demostró que de esta forma la estimación en la mayoría de ocasiones empeoraba considerablemente y por ello fue necesario buscar métodos lo más generales posibles.

A día de hoy, la clase más amplia de matrices utilizada en este tipo de problemas, $\mathcal{F} = \{H \in \mathcal{M}_{d\times d} : H > 0, H = H^T\}$ fue introducida en [28] por Deheuvels en el año 1977 y no fue considerada hasta la década de los 90. En parte las causas fueron las limitaciones computacionales y las herramientas matemáticas disponibles en ese momento. Este hecho explica el lento progreso en este campo de la estadística. Hasta hace menos de una década, de los dos enfoques principales de selección, a saber, métodos plug-in y técnicas de validación cruzada, únicamente el primero había recibido atención bajo el contexto sin restricciones. Los pioneros en crear un método plug-in para cualquier tipo de matrices fueron Duong y Hazelton. No obstante, como se puede leer en [30], centraron la explicación del proceso aplicado a las matrices diagonales.

Diversos ensayos demuestran analíticamente el buen rendimiento de las matrices sin restricciones para una amplia variedad de datos experimentales. Por esta razón normalmente se recomienda su uso para la estimación de la función densidad. Más adelante en el Capítulo 3 ilustraremos las diferencias que surgen entre utilizar o no una H diagonal y

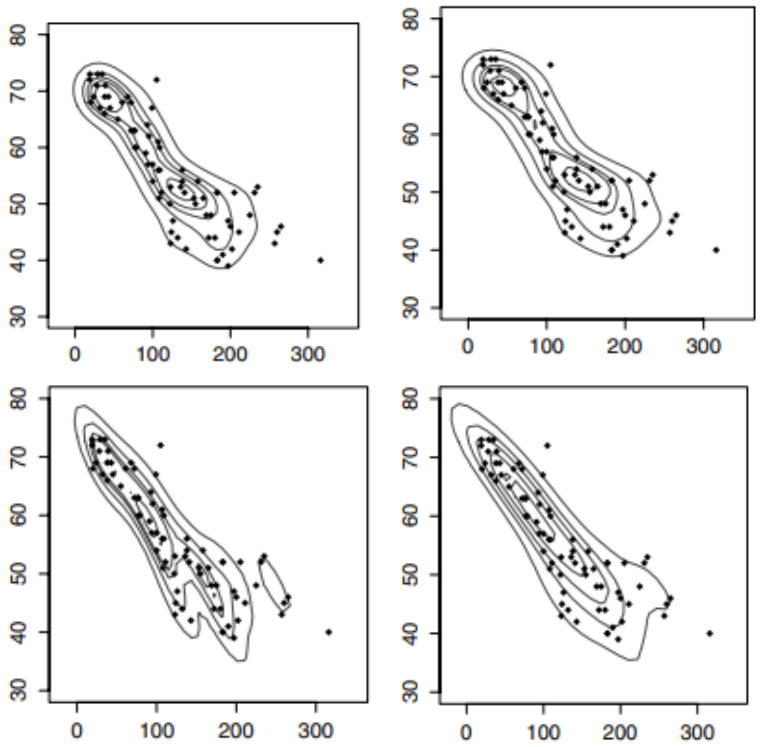

Figura 2.3: Estimaciones de la función densidad de los datos de mortalidad infantil obtenidos de UNICEF (Fondo de Naciones Unidas para la Infancia) a partir de diferentes matrices de ancho de banda. Las matrices se obtuvieron utilizando distintas variantes del método de validación cruzada.

comprobaremos lo inadecuadas que resultan para este trabajo.

Elegir la cantidad idónea de suavizado es en otras palabras, un problema de equilibrio entre sesgo y varianza. La falta de suavización da un sesgo bajo presentando así una alta variabilidad. Mientras que un exceso de suavizado provoca precisamente el efecto contrario.

El objetivo de estos criterios de selección consiste en encontrar un estimador fiable para una matriz ancho de banda óptima. Para ello, requerimos una medida de discrepancia. La forma más común de medir el rendimiento de f_{nH} evaluada en un punto fijo x es a través del error cuadrático medio MSE:

$$MSE(H) = MSE\{f_{nH}(x)\} = \mathbb{E}\{f_{nH}(x) - f(x)\}^2 dx = V(H) + SB(H)$$

donde $V(H)$ es la varianza y $SB(H)$ el sesgo al cuadrado.

El MSE es una medida local. En la estimación de la densidad, suele interesar el com-

portamiento global de f_{nH}. Para medir la eficiencia global, se integra MSE con respecto a x obteniendo así el error cuadrático integrado medio $MISE$. Todos los procedimientos tienen en común que buscan minimizar este concepto:

$$MISE(H) = MISE\{f_{nH}(x)\} = E\int_{\mathbb{R}}\{f_{nH}(x) - f(x)\}^2 dx \tag{2.2}$$

Con el fin de que el planteamiento tenga sentido, se supone que el núcleo multivariante K y la función densidad f son funciones de cuadrado integrable. Veamos esta definición:

Definición 2.3. Una función $f(x)$ de variable real con valores reales o complejos se dice de cuadrado integrable si la integral del cuadrado de su módulo converge.

$$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$$

La matriz ventana óptima verificará:

$$H_{MISE} = \underset{H\in\digamma}{\operatorname{argmin}}\, MISE(H)$$

donde recordemos que $\digamma$ es el conjunto de todas las matrices $d\times d$ simétricas y definidas positivas. No obstante, el óptimo H_{MISE} no tiene una forma cerrada y no es matemáticamente tratable. Una solución es emplear el análisis asintótico para encontrar una aproximación de $MISE$ que se pueda manejar, la cual se denomina $AMISE$. Seguidamente veremos cómo se llega a este nuevo concepto pero antes estableceremos algunas notaciones.

Sean A y B dos matrices cualesquiera. Se denota por A^T a la matriz traspuesta de A. El producto de Kronecker se escribirá como $A \otimes B$. Del mismo modo se expresará $A^{\otimes r} = A \otimes \cdots \otimes A$. Por otro lado, para cualquier función $f : \mathbb{R}^d \to \mathbb{R}$ denotaremos su vector gradiente por

$$Df = \left(\frac{\partial f}{\partial x_1}, \dots \frac{\partial f}{\partial x_d}\right) \in \mathbb{R}^d$$

Para las derivadas de f de orden superior en lugar de organizarlas como una matriz lo haremos en un vector. Escribiremos

$$D^{\otimes^r} f = (Df)^{\otimes^r} = \frac{\partial^r f}{(\partial x)^{\otimes r}} \in \mathbb{R}^{d^r}$$

para el vector que contiene todas las derivadas parciales de orden r.
También se indicará por $vecA$ y $vechA$ los operadores y medios vectoriales respectivamente aplicados a una matriz simétrica A. El operador vec toma las columnas de una matriz $d \times d$ y construye con ellas un solo vector de longitud d^2. Mientras que $vech$ apila la mitad

triangular inferior de una matriz $d \times d$ en un solo vector de longitud $\frac{d(d+1)}{2}$. Por ejemplo, denotando la matriz Hessiana de una función f por

$$Hf = \frac{\partial^2 f}{(\partial x \partial x^T)} \in \mathcal{M}_{d\times d}$$

se tiene que $vecHf = D^{\otimes^2} f$. Además, para una función $f : \mathbb{R}^b \to \mathbb{R}^p$ se introduce la notación

$$R(f) = \int_{\mathbb{R}^d} f(x) f(x)^T dx \in \mathcal{M}_{p\times p}$$

que es una matriz simétrica y definida positiva tal que $vecR(f) = \int_{\mathbb{R}^d} f(x)^{\otimes^2} dx \in \mathbb{R}^{p^2}$. Por último, falta añadir que para la función núcleo K en el caso multivariante existe momento de orden dos finito $m_2(K) \in \mathbb{R}$ tal que $m_2(K) I_d = \int_{\mathbb{R}^d} xx^T K(x) dx$ donde I_d es la matriz identidad de dimensión $d \times d$.

Es bien sabido que se puede descomponer $MISE(H) = ISB(H) + IV(H)$ donde los términos ISB (sesgo cuadrado integrado) e IV (la varianza integrada) se describen como:

$$ISB(H) = \int_{\mathbb{R}} \mathbb{E}\{f_{nH}(x) - f(x)\}^2 dx$$

$$IV(H) = \int_{\mathbb{R}} Var\{f_{nH}(x)\} dx$$

Chacón, Duong y Wand expandieron en [29] la expresión de $MISE$ extendiendo cada uno de sus términos y bajo ciertas hipótesis de suavidad obtuvieron su aproximación asintótica:

$$AMISE(H) = n^{-1}|H|^{\frac{-1}{2}} R(K) + \frac{m_2(K)^2}{4}(vech^T H)\psi_4(vechH)$$

donde ψ_4 es la matriz de dimensión $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ dada por:

$$\psi_4 = \int_{\mathbb{R}} vech\{2Hf(x) - dgHf(x)\} vech^T\{2Hf(x) - dgHf(x)\} dx = D_d^T R(D^{\otimes 2} f) D_d$$

aquí $dgHf$ se refiere a la matriz diagonal construida reemplazando todas las entradas de Hf fuera de la diagonal por ceros. También aparece D_d designando la matriz de duplicación de orden d que transforma $vec(A)$ en $vech(A)$ para cualquier matriz A simétrica de dimensión $d \times d$.

Lo relevante es que la tractabilidad de $AMISE$ nos permite encontrar el óptimo que sustituirá a H_{MISE}:

$$H_{AMISE} = \underset{H\in F}{\operatorname{argmin}}\, AMISE(H)$$

Seguidamente se profundizará en estas técnicas de selección. En concreto, se expone la teoría para matrices sin restricciones del método plug-in de Duong y Hazelton y de tres variantes del método de validación cruzada que son el sesgado, el suavizado y el insesgado o por mínimos cuadrados. Debido a que los anchos de banda óptimos H_{MISE} y H_{AMISE} dependen de la densidad desconocida f, la forma de proceder de todos los métodos será estimar $AMISE$ o $MISE$ y buscar una matriz H que minimice esta estimación.

2.2.1. Método plug-in de Duong y Hazelton

Los métodos plug-in son una clase importante de selectores que deriva de esta nueva expansión $AMISE$. Sus múltiples versiones nacen de las diferentes formas de aproximar la matriz desconocida ψ_4. Duong y Hazelton proponen en [30] un enfoque para seleccionar un H sin restricciones aunque una parte de su explicación se limita a las matrices diagonales. En cierto modo, este método es la variante perfeccionada del criterio de Wand y Jones que aparece en [31].

Wand y Jones se percataron de que cada componente de la matriz ψ_4 se podía escribir de la forma:

$$\psi_r = \int_{\mathbb{R}^d} f^{(r)}(x) f(x) dx$$

para algún $r = (r_1, \ldots, r_d) \in \mathbb{N}^d$ tal que $|r| = \sum_{i=1}^{d} r_i = 4$, donde se está denotando

$$f^{(r)} = \frac{\partial^{|r|} f}{\partial x_1^{r_1}, \ldots, \partial x_d^{r_d}}$$

En consecuencia, se les ocurrió reconstruir mediante estimadores tipo núcleo cada uno de estos elementos por separado:

$$\hat{\psi}_r(G) = \frac{1}{n^2} \sum_{i,j=1}^{n} (L_G)^{(r)}(X_i - X_j)$$

donde la ventana G y el núcleo L pueden ser diferentes a H y K, respectivamente.
Para medir el error de la aproximación emplearon el error cuadrático medio:

$$MSE_r(G) = \mathbb{E}[\{\hat{\psi}_r(G) - \psi_r\}^2]$$

Con todo, esta forma de proceder tiene el inconveniente de que al estimar individualmente estas componentes con diferentes G la matriz resultante no puede ser definida positiva. En consecuencia, la estimación de $AMISE$ no tendría un mínimo global finito y por tanto, no sería posible obtener H.

Duong y Hazelton solucionaron este problema escogiendo la misma ventana G para todas los elementos $\hat{\psi}_r(G)$. Esto es equivalente a aproximar ψ_4 mediante:

$$\hat{\psi}_4 = D_d^T R(D^{\otimes 2} f_{nG}) D_d$$

Usar un único ancho de banda garantiza que la estimación resultante sea siempre definida positiva. Aquella matriz que minimiza la suma de los $MSE_r s$ correspondientes a todos los ψ_r es la G óptima buscada.

A pesar de ser un procedimiento en el que H no tiene restricciones sí hay que imponer sobre G la condición $G = g^2 I_d$ con $g > 0$. En cambio, existen otros métodos en los que no hay limitaciones respecto al tipo de matrices que en ellos se utilizan, como por ejemplo el método plug-in multietapa que aparece descrito en [32].

2.2.2. Método de validación cruzada insesgado o por mínimos cuadrados

Los métodos de validación cruzada utilizan tanto la estimación de $MISE$ como la de $AMISE$, y algunos combinan ambos y minimizan la función resultante como se verá más adelante. El más conocido aunque no el más eficiente de este tipo de procedimientos es el método de validación cruzada por mínimos cuadrados propuesto tanto en [33] como en [34] por los autores Rudemo y Bouman respectivamente. Su importancia reside en que se usa como referencia de comparación respecto al resto de criterios. Este enfoque está motivado por las definiciones de ISE y $MISE$.

En el caso univariante, para evaluar el rendimiento de f_n se busca minimizar lo que se denomina el error cuadrático integrado:

$$ISE(h) = ISE\{f_n(x)\} = \int \{f_n(x) - f(x)\}^2 dx = \int f_n^2(x)dx - 2\mathbb{E}\{f_n(x)\} + \int f^2(x)dx$$

El primero de los términos de esta expresión es completamente conocido, el segundo se deja en función del valor de $f_n(x)$, el tercero es constante y no depende de h, por tanto puede ser ignorado. Advertir que estimar $\mathbb{E}\{f_n(X)\}$ mediante $\frac{1}{n}\sum_{i=1}^n f_n(X_i)$ donde X es una variable aleatoria independiente de $X_1, \ldots, X_n$ no es adecuado debido a que f_n depende de X_i. Sin embargo, prescindiendo del término i-ésimo se obtiene la siguiente aproximación:

$$LSCVh = \int f_n^2(x)dx - \frac{2}{n}\sum_{i=1}^n f_{n,-i}(X_i)$$

Donde n es el número de observaciones y $f_{n,-i}$ es la estimación de la función de densidad

sin usar el dato X_i denominada *jack-knife*:

$$f_{n,-i}(x) = (n-1)^{-1} \sum_{j=1, j\neq i} K_h(x - X_j)$$

De manera que el h buscado será aquel que minimice $LSCVh$.

La generalización al caso multivariante se realiza fácilmente utilizando la definición del estimador tipo núcleo en el contexto d-dimensional. Así el $LSCVh$ se define en este caso tal que:

$$UCV(H) = \int_{\mathbb{R}^d} f_{nH}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_{nH,-i}(X_i)$$

donde

$$f_{nH,-i}(x) = (n-1)^{-1} \sum_{j=1, j\neq i} K_H(x - X_j)$$

Ahora la matriz H óptima será aquella que minimice $UCV(H)$.

La función $UCV(H)$ es insesgada en el sentido de que $\mathbb{E}\{UCV(H)\} = MISE\{f_{nH}\} - R(f)$ ignorando la constante $R(f)$ que no depende del parámetro H, de ahí el nombre del método.

2.2.3. Método de validación cruzada sesgado

Esta técnica fue propuesta por primera vez para el caso univariante en [35] por Scott y Terrel. Realmente se trata de un híbrido entre los dos métodos anteriores en el sentido de que la función se minimiza mediante la validación cruzada de mínimos cuadrados pero haciendo uso de algunas ideas del método Duong y Hazelton. Al igual que los plug-in, el BCV (método validación cruzada sesgado) también busca minimizar una estimación del $AMISE$. Este criterio se basa en reemplazar en (2.2) la f desconocida por su estimador tipo núcleo, resultando así

$$BCV(H) = n^{-1}|H|^{\frac{-1}{2}} R(K) + \frac{m_2(K)^2}{4}(vec^T R(D^{\otimes^2} f_{nH})(vecH)^{\otimes^2}$$

De esta forma el ancho de banda buscado verifica $\hat{H}_{BCV} = argmin_{H\in F} BCV(H)$.

Este método fue creado con la intención inicial de reducir la variabilidad de UCV (método validación cruzada insesgado) y en consecuencia disminuir globalmente el $MISE$ de f_n. Algunos trabajos posteriores destacaron la importancia de utilizar un ancho de banda diferente para estimar la matriz de curvatura $R(D^{\otimes^2} f)$ a pesar de la carga computacional que esto supondría. Por ello el BCV no ha atraído suficiente interés.

2.2.4. Método de validación cruzada suavizado

El método de validación cruzada suavizado (SCV) para el caso univariante fue introducido en [36] tras los primeros estudios sobre este procedimiento realizados en [37]. Recientemente se ha desarollado el caso multivariante en [26].

El SCV propone una estimación de $MISE$ mediante la sustitución de la varianza integrada IV exacta en la ecuación (2.2) por su aproximación asintótica y manteniendo la forma exacta del sesgo cuadrado integrado ISB. Es decir, en lugar de estimar el ISB asintótico, se reemplaza la densidad f por su estimador núcleo

$$f_{nG}(x) = \frac{1}{n}\sum_{i=1}^{n} L_G(x - X_i)$$

obteniendo:

$$I\hat{S}B(H,G) = \int_{\mathbb{R}^d} [\{K_H * f_{nG}\}(x) - f_{nG}]^2 dx = n^{-2}\sum_{i,j=1}^{n} (\hat{K}_H * \hat{L}_G - 2K_H * \hat{L}_G + \hat{L}_G)(X_i - X_j)$$

donde $\hat{K} = K * K$, $\hat{L} = L * L$ y $*$ denota el producto de convolución. Añadiendo el término IV se consigue la siguiente función:

$$SCV(H,G) = n^{-1}|H|^{\frac{-1}{2}} R(K) + n^{-2}\sum_{i,j=1}^{n} (\hat{K}_H * \hat{L}_G - 2K_H * \hat{L}_G + \hat{L}_G)(X_i - X_j) \quad (2.3)$$

El óptimo de este método verifica $\hat{H}_{SCV} = argmin_{H \in F} SCV(H,G)$.

A diferencia de la aproximación de $AMISE$ en el método plug-in, no se requiere ningún estimador de Ψ_4 en la ecuación (2.3), con la compensación de las sumas dobles más intensivas computacionalmente. Se ha demostrado que este método es más prometedor que el BCV en cuanto a reducir la variabilidad de UCV.

2.3. Comparación teórica de los métodos

Una medida de referencia para comparar analíticamente el rendimiento asintótico entre los diferentes métodos de selección es la tasa relativa de convergencia. Ciñéndonos a [30], se dice que el selector $\hat{H}$ converge a H_{AMISE} a una tasa relativa $n^{-\alpha}$, para $\alpha > 0$, cuando

$$vech(\hat{H} - H_{AMISE}) = O_p(J_{d'} n^{-\alpha}) vech H_{AMISE}$$

donde $J_{d'}$ es la matriz con 1´s en todas sus entradas y dimensiones $d' \times d'$ con $d' = \frac{1}{2}d(d+1)$. La extensión del orden asintótico a las matrices se ha hecho de forma que para secuencias

(A_n) y (B_n) se tiene que $A_n = o(B_n)$ si $a_{ij} = o(b_{ij})$ para todos los elementos a_{ij} y b_{ij} de A_n y B_n respectivamente. El motivo por el cual en lugar de utilizar $O_p(I_d)$,donde I es la matriz identidad, se utiliza $O_d(J'_d)$ es enmendar los casos en los que H_{AMISE} es diagonal y el selector $\hat{H}$ no tiene restricciones. De este modo el $J_{d'}$ asegura que la ecuación y en consecuencia la tasa de convergencia siga estando bien definida.

Observar que la norma al cuadrado del término de la izquierda de la ecuación, $\|vec(\hat{H} - H_{AMISE})\|^2$, es en efecto $MSE(\hat{H})$. Esta estrecha relación es la base para el cálculo de estas tasas.

Las tasas de todos los selectores para diferentes tipos de matrices se resumen en el Cuadro 2.1. Recordemos que $\mathcal{A} = \{h^2 I_d : h > 0\}$, $\mathcal{F} = \{H \in \mathcal{M}_{d\times d} : H > 0, H = H^T\}$ y $\mathcal{D} = \{diag(h_1^2, \ldots, h_d^2) : h_1, \ldots, h_d > 0\}$. En general, el rendimiento de cada criterio disminuye a medida que la dimensión crece.

Selector	Clase	Radio de convergencia a H_{AMISE}
$\hat{H}_{UCV}$	$\mathcal{A}, \mathcal{D}, \mathcal{F}$	$n^{\frac{-min\{d,4\}}{(2d+8)}}$
$\hat{H}_{BCV}$	$\mathcal{A}, \mathcal{D}, \mathcal{F}$	$n^{\frac{-min\{d,4\}}{(2d+8)}}$
$\hat{H}_{SCV}$	$d = 1$	$n^{\frac{-5}{14}}$
$\hat{H}_{SCV}$	$d > 1, \mathcal{A}, \mathcal{D}, \mathcal{F}$	$n^{\frac{-2}{(d+6)}}$
$\hat{H}_{PI}$	$d = 1$ o $d > 1$, $\mathcal{A}$, $\mathcal{D}$	$n^{\frac{-min\{8,(d+4)\}}{(2d+12)}}$
$\hat{H}_{PI}$	$d > 1, \mathcal{F}$	$n^{\frac{-2}{(d+6)}}$

Cuadro 2.1: Comparación de las tasas de convergencia relativas para H_{AMISE} de los métodos de selección de matriz de ancho de banda.

En lo que respecta a los métodos sin restricciones, el plug-in de Duong y Hazelton y el SCV tienen las mejores tasas de convergencia exceptuando el caso bivariante en el que ambos empeoran ligeramente. La causa es que solo es posible la aniquilación de sesgos para $d = 1$. En el caso de $d > 1$ con matrices sin restricciones el sesgo solo llega a minimizarse produciéndose así un índice de convergencia de $n^{\frac{-2}{(d+6)}}$.

Observar que los dos tienen la misma tasa relativa y esto es así porque comparten la estrategia de estimar y minimizar la forma asintótica $AMISE$.

Para $d = 2, 3$ los selectores BCV y UCV convergen más lentamente que el SCV y el plug-in con un coeficiente de $n^{\frac{-min\{d,4\}}{(2d+8)}}$. Sin embargo, para dimensiones $d \geq 4$ las

propiedades asintóticas de estos métodos con respecto al resto mejoran paradójicamente. Como resultado, para una muestra de gran tamaño se comportan mejor que SCV en las dimensiones más altas. Si bien estos contrastes son interesantes desde el punto de vista teórico, las implicaciones en la práctica no son nada claras.

En el caso de las matrices restringidas las tasas para los métodos de validación cruzada UCV, BCV, SCV no varían. Los selectores plug-in con $\hat{H}_{PI,\mathcal{D}}$ o $\hat{H}_{PI,\mathcal{A}}$ tienen una tasa de convergencia más rápida $n^{\frac{-min8,(d+4)}{(2d+12)}}$, de nuevo debido a la aniquilación de sesgos- Esto indica que determinar los elementos fuera de diagonal de la matriz de ancho de banda, que determinan la orientación del núcleo, es el aspecto más complejo del método de plug-in sin restricciones.

Capítulo 3

Análisis de los datos

En el capítulo expuesto a continuación se reconstruirán distintos conjuntos de nivel para los datos de contagios en Estados Unidos presentados en el Capítulo 1. Concretamente, se aplicarán para la selección de la matriz H tres de los métodos explicados en la sección previa y se analizará la evolución en tiempo real de las HDR´s estimadas ilustrándolas en el mapa. Se trata de examinar las diferencias entre los resultados obtenidos. Las técnicas utilizadas serán el criterio plug-in y los métodos de validación cruzada por mínimos cuadrados y suavizado.

A priori, investigar cuál es el mejor procedimiento para unos datos específicos no es viable puesto que las definiciones de $MISE$ y $AMISE$ dependen de la función densidad estimada de la cual se desconoce su valor. Lo que sí es posible es evaluar el rendimiento de los estimadores realizando simulaciones para distribuciones concretas y muestras aleatorias. Experimentos de esta clase se recogen en multitud de documentos para todo tipo de métodos y tamaños muestrales. Si bien no se recomienda extrapolar conclusiones generales, sus resultados sirven de guía en la mayor parte de situaciones. Por ello, aunque en este trabajo no hemos realizado ningún estudio de simulación, compararemos sus soluciones con las que obtengamos en este caso particular.

Nuestra información está estructurada en un archivo de Microsoft Excel donde aparecen una serie de códigos referentes a Estados Unidos y a cada usuario infectado (UID, FIPS, code3, ...), la región afectada y sus coordenadas (latitud y longitud), el estado americano en el que se encuentra y el número de contagios que aparecen en ella a diario desde el 22 de Enero de 2020 hasta el 5 de Mayo de 2020.

Para estimar las HDR´s los días serán agrupados en conjuntos de 6. Cada uno de ellos denominará semana independientemente de que no esté formado exactamente por 7 días.

Esto dará lugar a 10 muestras, correspondientes a las 10 semanas, cuyos tamaños son respectivamente 89, 1007, 8007, 58673, 149814, 221738, 214264, 204644, 209343 y 167116.

El objetivo es reconstruir los focos de COVID-19 con mayor incidencia, en consecuencia prestaremos especial atención a aquellas semanas con un número de contagios elevado. No obstante, interesa revisar el resto de días pues este tipo de estudios sirven tanto para comprobar si el virus se expande como si disminuye en alguna zona. Conocer esta evolución ayuda a una repartición equitativa de recursos. Por ello, para cada método resumiremos brevemente el desarrollo de los clústers desde Enero hasta Mayo seleccionando específicamente las semanas 2, 6, 8 y 10.

En cuanto a la probabilidad contenida en los conjuntos de nivel, como se indicó en la introducción las zonas realmente preocupantes se dan para aquellos valores de α más próximos a 1. Para este análisis se han escogido los valores $\alpha = 0,90$, $\alpha = 0,75$ y $\alpha = 0,50$.

Empecemos con el caso más interesante según nuestros propósitos pues es el caso con mayor α, $\alpha = 0,90$. En la Figura 3.1 se expone la evolución de las HDR´s para el método plug-in sin restricciones. Se observa que en la primera semana aparecen ya dos pequeños clústers, uno en Nueva York y otro en Seattle. Para la 6 vemos cómo solo se mantiene el de Nueva York siendo este el único foco hasta la semana 10 en la que aparece uno nuevo en Chicago. En principio este resultado tiene lógica pues Nueva York y Chicago están entre las tres ciudades más pobladas de Estados Unidos y por consiguiente es probable que en ellas haya mayor número de contagios. Además, Nueva York es la capital financiera y cultural del país. En ella destaca su extenso y muy ramificado transporte público y el movimiento constante de visitantes y turistas tanto domésticos como extranjeros.

Estos dos focos que como veremos a lo largo de esta parte predominan en todo momento, son los más alarmantes. Señalar que contienen una probabilidad de al menos el 10 %. Analizando estas mismas semanas para el mismo valor de α vemos que el resto de procedimientos detecta exactamente los mismos clústers con unos perímetros muy parecidos. Esto se debe a que estos métodos son más precisos cuanto mayor es α.

Veamos qué ocurre ahora si $\alpha = 0,75$. En la Figura 3.2 aparece representado en esta ocasión el desarrollo de los clústers obtenidos mediante el criterio SCV. Aquí se percibe una pequeña evolución más clara de cómo el virus se extiende hacia otras zonas del mapa. Inicialmente se muestran de nuevo los brotes de Nueva York y Seattle, esta vez contenidos en la región estimada $\hat{R}(f_{0,75})$.

Figura 3.1: HDR´s con $\alpha = 0,90$ en las semanas 2, 6, 8 y 10 respectivamente para el método plug-in sin restricciones.

En la semana siguiente se puede ver al igual que en la Figura 3.1 que se conserva el contorno de Nueva York y desaparece el de Seattle. A diferencia con el caso anterior, la HDR de Chicago surge ya en la penúltima semana. Esto quiere decir que esta zona en esos días tiene entre el 10 % y 25 % de los contagios por coronavirus. En la 10 se mantienen Chicago y Nueva York como focos peligrosos y aparecen unos nuevos en Boston, Filadelfia, Washintong DC y los Ángeles. Un resultado coherente con el hecho de que estas ciudades tienen densidades de población bastante elevadas con respecto al resto.

Comparando con los demás selectores observamos que todos coinciden en la localización de los clústers. Sin embargo, advertimos que así como las formas de las regiones utilizando las matrices H_{PI} y H_{SCV} son similares, las del caso H_{UCV} difieren de las anteriores. Hemos representado estas desigualdades para la semana 10 en la Figura 3.3.

Figura 3.2: HDR´s con $\alpha = 0,75$ en las semanas 2, 6, 8 y 10 respectivamente para el criterio SCV.

Por último, veamos qué sucede esta vez para el valor de α más bajo, $\alpha = 0,50$. Una síntesis del avance del COVID-19 para el criterio de validación cruzada suavizado se puede observar en la Figura 3.4. Aquí encontramos más discrepancias entre un método y otro. Empecemos por la primera semana en la que de nuevo aparecen los dos focos de Seattle y Nueva York pero esta vez ampliados. Señalar que se distingue un clúster de menor tamaño cerca de Nueva York en la ciudad de Boston.

En la siguiente semana el de Seattle vuelve a desaparecer y surgen otros nuevos: Nueva Orleans, los Ángeles, Chicago, Detroit, Miami y Filadelfia. Miami que hasta ahora no había sido señalado como foco, también tiene una densidad de población muy alta. En la semana 8 Nueva Orleans deja de ser una zona alarmante y aparece un clúster en Columbus. Final-

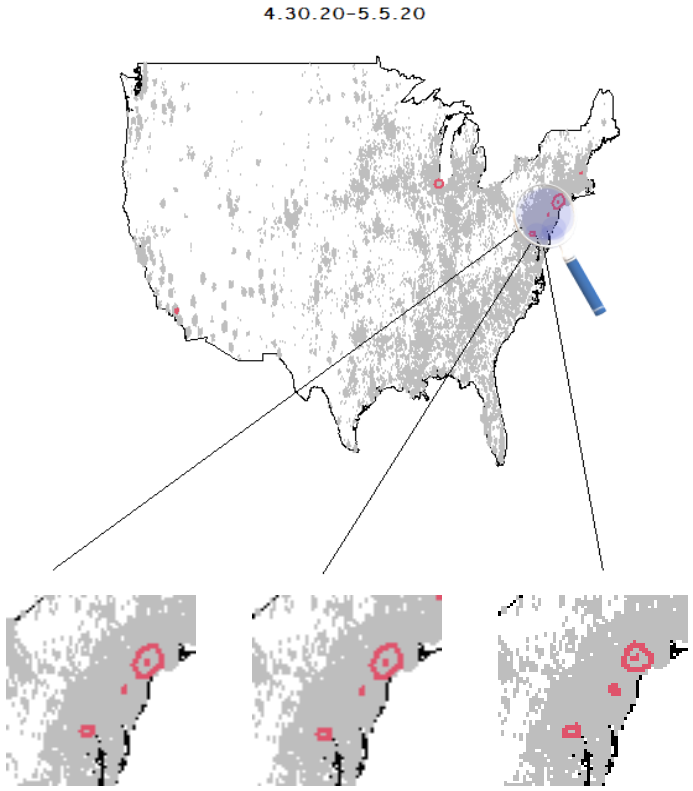

Figura 3.3: HDR´s con $\alpha = 0,75$ en la semana 10 respectivamente para H_{PI}, H_{SCV}, H_{UCV}.

mente, se preservan los focos de Nueva York, Boston, Wahingston, los Ángeles, Chicago y Detroit y se forman nuevos en Atlanta, Dallas, Phoenix, Indianapolis y en las zonas de Tenesse y Iowa. Claramente los contagios se han propagado hacia el sur de Estados Unidos que es la región con menos restricciones del país. En algunos de estos lugares ni siquiera era obligatorio quedarse confinado en casa pues en cada estado decidía el propio gobernador qué medidas tomar.

Hagamos una comparación de los tres criterios para cada semana. En la primera mientras el enfoque plug-in sin restricciones y SCV son prácticamente iguales, UCV llama la atención. Este método manifiesta dos HDR´s claramente distinguidas en Washington y otras dos en Nueva York y Boston más separadas que antes. Además detecta Los Ángeles como zona problemática ya en esta semana. Observar la Figura 3.5.

En cuanto a la semana 6 una vez más el UCV discrepa del resto. Por lo que se refiere a las estimaciones mediante H_{PI} y H_{SCV} siguen siendo prácticamente idénticas. A medida

Figura 3.4: HDR´s con $\alpha = 0,50$ en las semanas 2, 6, 8 y 10 respectivamente para el método plug-in sin restricciones.

que aumenta el tamaño de la muestra se acentúan estas desigualdades. En consecuencia, en la semana 8 UCV ofrece un mapa en el que encontramos dos focos en Ohio al tiempo que el resto de criterios señalan un foco en Michigan y otro en Ohio. Esto implica una diferencia importante entre una selección y otra pues quizá se estuviese pasando por alto una zona que realmente necesita ayuda. Se puede comprobar en la Figura 3.6.

Para la semana 10 ocurre exactamente lo mismo. En la estimación mediante H_{UCV} dada en la Figura 3.7 aparece un clúster nuevo respecto a los otros métodos, el de Saint Paul.

De todo lo anterior se deduce que la estimación mediante H_{UCV} también está influenciada

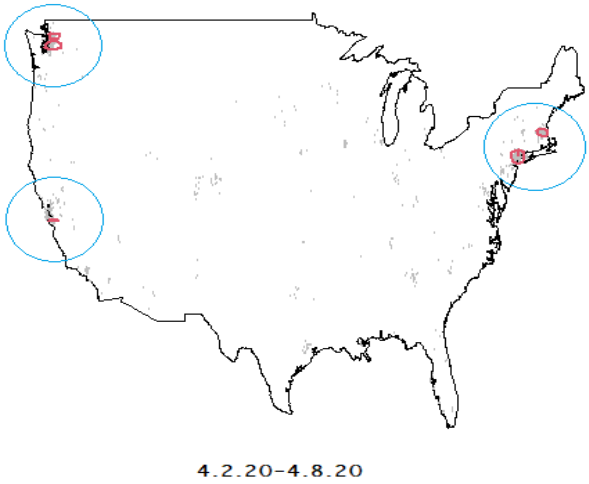

Figura 3.5: HDR´s con $\alpha = 0,50$ en la semana 2 para el método UCV.

por el valor de α pues es todavía peor a medida que este aumenta.

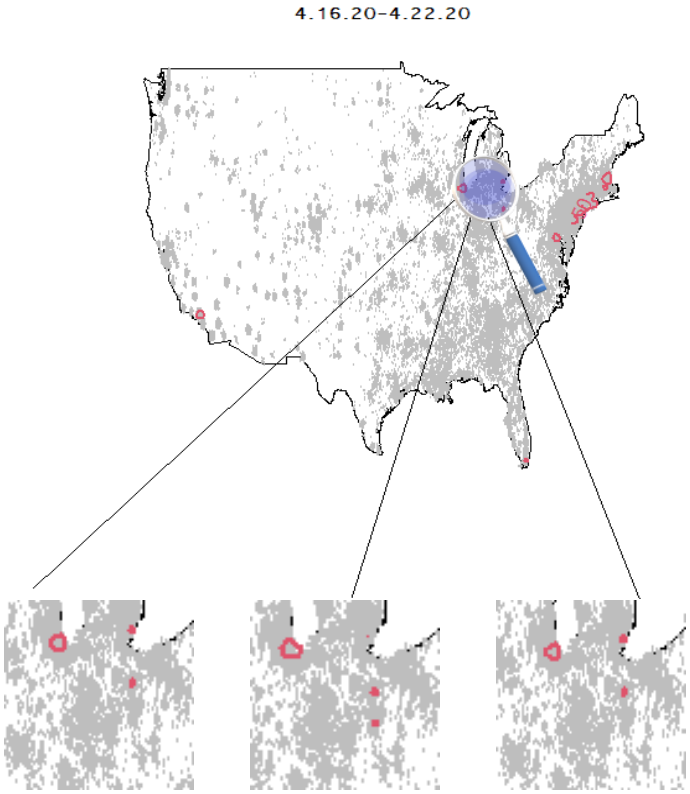

Figura 3.6: Contornos de las regiones dentro de Ohio y Michigan para las matrices H_{PI}, H_{UCV} y H_{SCV} respectivamente.

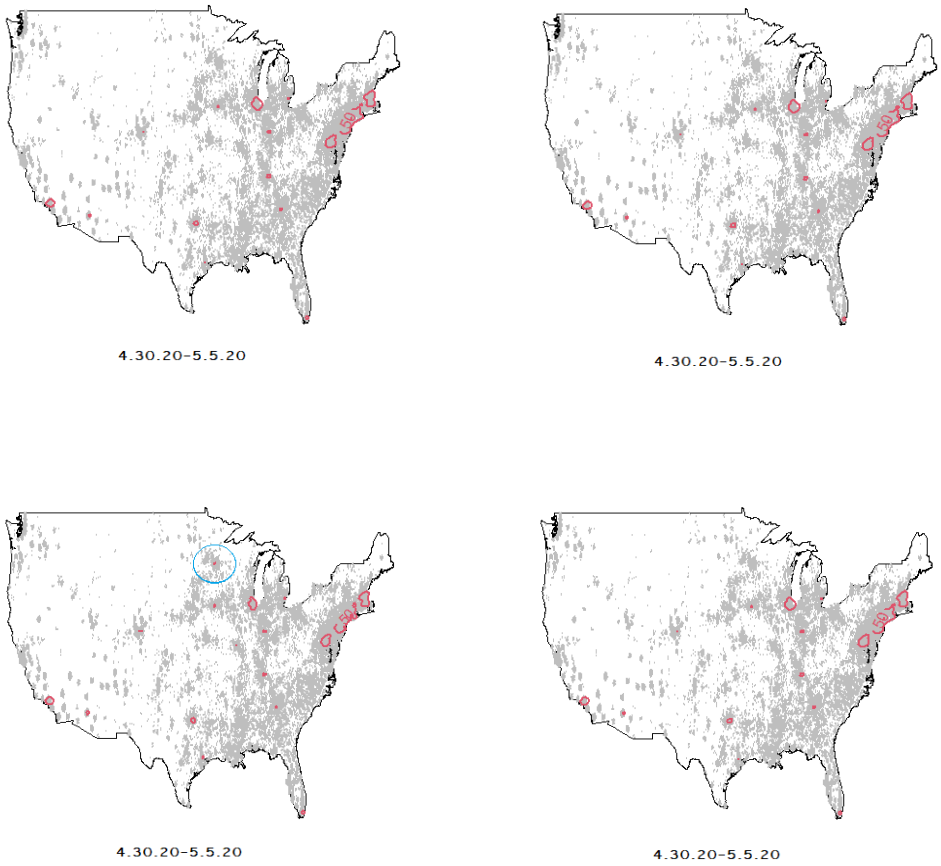

Figura 3.7: HDR´s con $\alpha = 0,50$ en la semana 10 para los métodos plug-in con H diagonal, plug-in sin restricciones, UCV y SCV respectivamente.

Capítulo 4

Conclusiones

Al comienzo de este trabajo, en el Capítulo 1, se planteó el problema de la reconstrucción de determinados conjuntos de nivel a partir de una muestra aleatoria simple de distribución desconocida. También fueron presentados los datos de la localización de los contagios por COVID-19 en Estados Unidos que utilizaremos de ejemplo. En el Capítulo 2 se formuló formalmente el problema de la reconstrucción de las HDR´s. En particular, nos hemos interesado por aquellas técnicas que no asumían restricciones geométricas sobre los conjuntos como puede ser el método plug-in. También hemos expuesto el problema de la estimación de la matriz H y descrito los métodos para seleccionarla que llevaremos a la práctica. Finalmente, se evaluaron sus órdenes de convergencia para comparar sus rendimientos teórico.

Para terminar este documento, hemos mostrado un ejemplo de aplicación práctica sobre los datos de los contagios, estudiando la evolución espacio-temporal de los clústers del virus. Para ello, hemos implementado los distintos métodos en el software R para ilustrar las HDR´s en el mapa.

En las imágenes hemos podido observar cómo rápidamente se han ido desplazando los focos de COVID-19 de norte a sur y casi siempre localizados en las ciudades con más habitantes o mayor densidad de población. Lo que es congruente con la realidad teniendo en cuenta que los estados con las restricciones menos severas en aquel momento se encuentran en esta zona. Hemos obtenido diversos clústers a lo largo del país, pero sin duda las dos regiones más preocupantes son Nueva York y Chicago. Estos focos han estado siempre presentes y han sido claramente detectados por todos los criterios aplicados. Por tanto, es aquí donde se debería actuar con más rapidez a la hora de enviar ayuda localizada.

Por otra parte, hemos comprobado que realmente los mejores métodos para estimar la

función densidad en casos similares a este son el método plug-in o el SCV. Preferiblemente se aconseja usar el primero dada su mayor velocidad computacional. Ambos nos han aportado soluciones muy similares en las que nos podríamos apoyar para controlar los efectos de la pandemia.

Sin duda, el menos recomendable es el criterio UCV el cual a lo largo del análisis discrepó constantemente del resto. Este método tiene un comportamiento paradójico. Teóricamente su tasa de convergencia mejora a medida que aumentan la dimensión y el tamaño muestral, pero en la práctica se ha comprobado que bajo estas condiciones no funciona adecuadamente. Por ejemplo, la estimación mediante H_{UCV} detectaba contornos con formas muy diferentes a las estimadas por los otros métodos e incluso las localizaciones de sus clústers no coincidían con las del resto.

Bibliografía

[1] *Preguntas y respuestas sobre la enfermedad por coronavirus (COVID-19).* (2020) Organización Mundial de la Salud. https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses

[2] Eurostat. (2021, 16 Febrero) *Excess mortality in 2020 reached its peak in November.* https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20210216-2

[3] *Coronavirus Disease 2019 (COVID-19).* (2020, 11 Febrero) Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2Findex.html

[4] Woolf, S. H., Chapman, D. A. and Sabo, R. T. (2020, 12 Octubre) Excess Deaths From COVID-19 and Other Causes, March-July 2020. *JAMA*, **324**(15), 1562-1564. https://jamanetwork.com/journals/jama/fullarticle/2771761?guestAccessKey=0c3da4cd-caaf-48e1-bf7b-2cc3c20a94b7&utm_source=silverchair&utm_medium=email&utm_campaign=article_alert-jama&utm_content=olf&utm_term=101220

[5] Lozano-Vargas, A. (2020, Enero) Impacto de la epidemia del Coronavirus (COVID-19) en la salud mental del personal de salud y en la población general de China. *Revista de Neuro-Psiquiatría*, **83**(1). http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S0034-85972020000100051

[6] Instituto Nacional de Estadística (INE). (2020, 26 Marzo) Contabilidad Nacional Trimestral de España: principales agregados. https://www.ine.es/daco/daco42/daco4214/cntr0420.pdf

[7] Eurostat. (2020, 10 Diciembre) COVID-19 impact on employment income. https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/ddn-20201210-2

[8] National Bureau of Statistics of China. (2020, 17 Abril) Decline of Major Economic Indicators Significantly Narrowed Down in March. `http://www.stats.gov.cn/english/PressRelease/202004/t20200417_1739339.html`

[9] National Bureau of Statistics of China. (2020, 16 Marzo) National Economy Withstood the Impact of COVID-19 in the First Two Months. `http://www.stats.gov.cn/english/PressRelease/202003/t20200316_1732244.html`

[10] Comisión Económica para América Latina (CEPAL). (2020, Mayo) Informe sobre el impacto económico en América Latina y el Caribe de la enfermedad por coronavirus (COVID-19).

[11] *GitHub-CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE.* GitHub. `https://github.com/CSSEGISandData/COVID-19`

[12] Geffroy, J. (1964) Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique des Universites de Paris*, **13**, 191–210.

[13] Simonoff, J. S. (1996) *Smoothing Methods in Statistics.* Springer-Verlag, New York.

[14] Devroye, L. and Wise, G. (1980) Detection of abnormal behavior via non-parametric, estimation of the support. *SIAMJ (Sociaty for Industrial and Applied Mathematics) Journal on Applied Mathematics*, **3**, 480-488.

[15] Hartigan, J. A. (1975) *Clustering Algorithms.* John Wiley and Sons Inc.

[16] Korostelev, A. P. and Tsybakov, A. B. (1993) Estimation of the density support and its functionals. *Problems of Information Transmission*, **29**, 1–15.

[17] Rényi, A. and Sulanke, R. (1963) Über die konvexe Hülle von n zufällig gewählten Punkten. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **2**, 75–84.

[18] Rényi, A. and Sulanke, R. (1964) Über die konvexe Hülle von n zufällig gewählten Punkten.(II) *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **3**, 138–147.

[19] Aaron, C. and Bodart, O. (2016) Local convex hull support and boundary estimation. *Journal of Multivariate Analysis*, **147**, 82-101.

[20] Hyndman, R. J. (1996) Computing and Graphing Highest Density Regions. *The American Statistician*, **50**, 120-126.

[21] Silverman, B. W. (1985) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

[22] Fix, E. and Hodges, J. L. (1951) *An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation.* Reprinted by Silverman, B. W. and Jones, M. C. (1989) *International Statistical Review*, **57**, 233-247.

[23] Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832-837.

[24] Parzen, E. (1962) On the estimation of a probabilty density function and the mode. *Annals of Mathematical Statistics*, **33**, 1065-1076.

[25] Scott, D.W. (1985b) Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *Annals of Mathematical Statistics*, **13**, 1024–1040.

[26] Duong, T. and Hazelton, M. L. (2005) Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation. *Scandinavian Journal of Statistics*, **32**(3), 485–506.

[27] Jones, M., Marron, J. and Shealther, S. (1996) Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* **11**, 337-381.

[28] Deheuvels, P. (1977) Estimation non paramétrique de la densité par histogrammes généralisés. II, *Publications de l´Institut de Statistique de l´Universite de Paris* **22**, 1–23.

[29] Chacón, J. E., Duong, T. and Wand, M. P. (2011) Asymptotics for general multivariate kernel density derivate estimators. *Statistica Sinica*, **21**, 807-840.

[30] Duong, T. and Hazelton, M. L (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *Journal of Nonparametric Statistics*, **15**, 17-30.

[31] Wand, M. P and Jones, M. C. (1994) Multivariate plug-in bandwidth selection. *Computational Statistics* **9**, 97-117.

[32] Chacón, J. E. and Duong, T. (2010) Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, **19**, 375–398.

[33] Rudemo, M. (1982) Empirical choice of histogram and kernel density estimators, *Scandinavian Journal of Statistics*, **9**, 65-78.

[34] Bowman, A. W (1984, Agosto) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**(2), 353–360.

[35] Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**, 1131–1146.

[36] Hall, P., Marron, J. S. and Park, B. U. (1992) Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1–20.

[37] Jones, M. C., Marron, J. S. and Park, B. U. (1991) A simple root n bandwidth selector. *Annals of Statistics*, **19**, 1919–1932.