

Universidade de Vigo

Trabajo Fin de Máster

Test de multimodalidad: ventana crítica y exceso de masa

Jose Ameijeiras Alonso

Máster en Técnicas Estadísticas

2013-2014

Máster en Técnicas Estadísticas

Trabajo Fin de Máster

Test de multimodalidad: ventana crítica y exceso de masa

Jose Ameijeiras Alonso

Bajo la dirección de:

Rosa María Crujeiras Casais y Alberto Rodríguez Casal

Julio 2014

Índice general

1. Introducción	1
1.1. Herramientas exploratorias	4
1.2. Tipos de contrastes de multimodalidad	10
1.3. Detección de modas en densidades circulares	14
1.4. Estructura del trabajo	19
2. Herramientas gráficas	21
2.1. El árbol y el bosque de modas	21
2.1.1. El árbol de modas	22
2.1.2. El bosque de modas	26
2.2. SiZer	28
2.3. Herramientas gráficas para datos circulares	32
2.3.1. CircSiZer	32
3. Test de multimodalidad para datos lineales	37
3.1. Test basados en la ventana crítica	37
3.1.1. Test basado en la ventana crítica de Silverman	38
3.1.2. Calibrado de la ventana crítica de Silverman	38
3.1.3. Test basado en el estadístico de Cramér–von Mises	40
3.2. Test basado en el exceso de masa y <i>dip</i>	47

3.2.1. Test basado en el <i>dip</i>	48
3.2.2. Test de exceso de masa de Muller y Sawitzki	49
3.2.3. Calibrado de los test de exceso de masa y <i>dip</i>	50
3.2.4. Nueva propuesta: Calibrado de los test de exceso de masa y <i>dip</i> haciendo uso de la ventana crítica	52
3.3. Estudio de simulación	53
4. Test de multimodalidad para datos circulares	61
4.1. Test basados en la concentración crítica	61
4.1.1. Nueva propuesta: Test basado en la concentración crítica como estadístico . .	62
4.1.2. Test basado en el estadístico U^2 de Watson	63
4.2. Nueva propuesta: Test basado en el exceso de masa	64
4.3. Estudio de simulación para datos circulares	66
5. Aplicaciones a datos reales	71
5.1. Duración entre dos erupciones del géiser Old Faithful	72
5.2. La renta por hogar en Galicia	73
5.3. Dirección de viento en los episodios de contaminación por NO_x	75
5.4. Los incendios en la comarca de Vigo	78
A. Modelos empleados	85
B. Resumen de los test empleados.	89
Bibliografía	93

Capítulo 1

Introducción

Existen diversas causas que pueden dar lugar a incendios. Estos pueden ser provocados por factores climatológicos, como puede ser el exceso de calor solar, el cual induce la deshidratación de las plantas causando la emisión a la atmósfera de etileno, un compuesto químico altamente combustible, lo que provoca la posibilidad de que una simple chispa pueda causar un incendio forestal. Otro motivo que puede provocar incendios es la acción humana, ya sea con quemas intencionadas (con el objetivo de eliminar rastrojos o matorrales, regenerar pastos para el ganado, ...) o ocasionadas por negligencias o causas accidentales (una hoguera mal apagada, colillas, ...).

Relacionado con lo anterior, uno de los problemas al que se enfrentan en el campo de las ciencias forestales es el de conocer el número de temporadas de incendios en cada región del mundo. El determinar si, en una región, se producen más de una cantidad prefijada de estaciones de incendios es de gran importancia, pues permite conocer, por ejemplo, si algún factor antropogénico ha provocado que en dicha región hubiese más temporadas de incendios de las que deberían haber ocurrido por factores climatológicos. Además, el conocer el número de temporadas de incendios y cuándo ocurren permitiría una mejor gestión del régimen de incendios para reducir las emisiones de CO₂ causantes de, por ejemplo, los efectos invernadero. Para poder determinar el número de temporadas de incendios, no será necesario tener una caracterización completa de la variable aleatoria que estudia la fecha en la que se produce un incendio en una determinada región, si no que simplemente llega con saber qué valores de la variable aleatoria son los más probables.

En particular, se podría estudiar si en la comarca gallega de Vigo solo hay una temporada de incendios o más de una. Para ello, se debe tener en cuenta que este tipo de datos se repiten periódicamente, pues si se tiene, por ejemplo, una muestra de los incendios que ocurrieron en esta comarca entre el 2002 y el 2012, se tratará de igual forma a un incendio producido el 3 de Agosto de 2004, que a uno ocasionado el 3 de Agosto de 2010. Además, se espera un comportamiento similar a finales de año que a comienzos. En general, este tipo de muestras cíclicas se pueden representar en el círculo y se conocen como datos circulares. En la Figura 1.1 (izquierda) se muestra esta representación circular mediante el diagrama de rosa para una muestra de la fecha en la que se producen los incendios en la comarca de Vigo entre el 10 de Julio de 2002 y el 9 de Julio 2012 (a

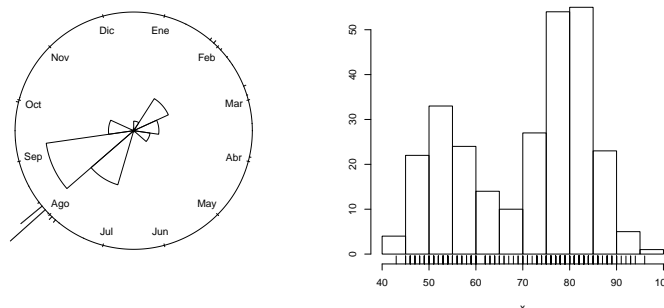


Figura 1.1: Izquierda: Diagrama de rosa para los incendios en la comarca de Vigo para el período que abarca desde el 10 de Julio de 2002 hasta el 31 de Diciembre de 2005 y del 1 de Enero del 2007 al 9 de Julio de 2012. Derecha: Histograma para los tiempos de espera entre dos erupciones del géiser Old Faithful situado en el Parque Nacional de Yellowstone en Wyoming, USA.

excepción de los ocasionados en el 2006). Este tipo de datos circulares presentan unas características especiales, que diferenciará su tratamiento estadístico del realizado en los datos lineales. Véase, por ejemplo, Jammalamadaka y Sengupta (2001) para una introducción al análisis estadístico de datos circulares.

El problema de determinar qué valores de la variable aleatoria son los más probables está presente en infinidad de situaciones. Así, por ejemplo, en el campo de la biología, también dentro del caso de medidas circulares, se puede encontrar el problema tratado por Schmidt–Koenig (1963) que es el de estudiar si una especie de pájaros vuela siempre en torno a una dirección, o en cambio, hay más de una dirección predominante en la dirección de vuelo.

En el caso de variables escalares, también se pueden encontrar numerosos ejemplos de esta situación. Así, en el campo de la geología, el determinar si el porcentaje de silicio de las condritas (un tipo de meteoritos) que cayeron en la Tierra se agrupan en torno a un valor o si, en cambio, el porcentaje de silicio de algunos de estos meteoritos se sitúa en torno a un valor y el de otros en torno a otros valores, es de gran utilidad, pues permitiría conocer si la Tierra fue formada por varios tipos de condritas y daría lugar a conocer más detalles sobre su origen (este problema ha sido tratado por Good y Gaskins, 1980). Otro ejemplo de esta situación es el tratado por Azzalini y Bowman (1990), donde se considera el tiempo que transcurre, en minutos, entre el comienzo de dos erupciones del géiser Old Faithful situado en el Parque Nacional de Yellowstone en Wyoming, USA (en la Figura 1.1, derecha, se ilustra el histograma de una muestra de estos tiempos), y se quiere determinar si lo más probable es que todos los tiempos de espera se agrupen en torno a los 80 minutos o si en cambio, hay dos tendencias, unos tiempos de espera largos, en torno a los 80 minutos, y otros cortos, en torno a los 55 minutos.

Todos estos ejemplos están relacionados con el problema de detección de modas de una variable

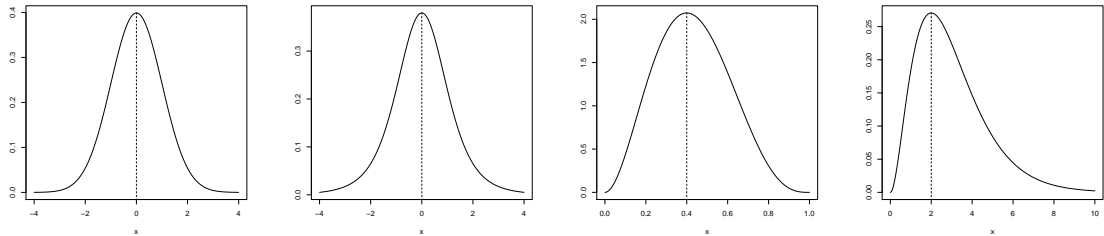


Figura 1.2: Ejemplos de funciones de densidad unimodales. De izquierda a derecha: densidad $N(0, 1)$, t_5 , $Beta(3,4)$ y $Gamma(3,1)$. Las líneas discontinuas muestran las localizaciones de las modas.

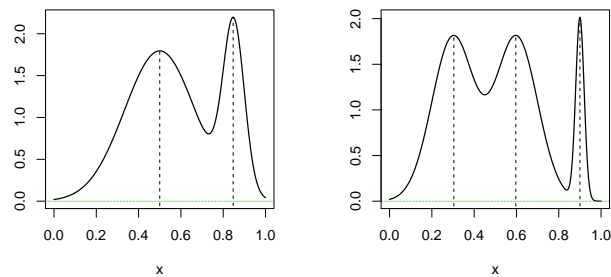


Figura 1.3: Ejemplos de funciones de densidad bimodales. Izquierda: mixtura bimodal de normales M5. Derecha: mixtura trimodal de normales M10 (ver Apéndice A). Las líneas discontinuas muestran las localizaciones de las modas.

aleatoria (v.a.) continua X . El concepto de moda fue introducido por Pearson (1895) y se definirá como el valor (o valores) en el que se alcanza un máximo local en la función de densidad. Si hay una sola moda, entonces la función de densidad (o la distribución asociada) es unimodal. En cambio, si la función de densidad presenta más de una moda se dirá que esta es multimodal (en particular, si la función de densidad tiene dos modas esta será bimodal; si tiene tres, será trimodal, ...).

En la Figura 1.2, se ilustran algunos ejemplos de funciones de densidad unimodales, como son la normal estándar, $N(0, 1)$, o la t de student con 5 grados de libertad, t_5 , ambas con moda en el punto 0; la $Beta(3,4)$ con una moda en el punto 0.4 o la $Gamma(3,1)$ con moda en el punto 2 (en Johnson *et al.* (1995) se pueden ver las expresiones de las funciones de densidad paramétricas empleadas a lo largo de este trabajo). En la Figura 1.3 se muestran algunos ejemplos de funciones de densidad multimodales, como son las mixturas de normales M5 y M10 descritas en el Apéndice A.

A lo largo de este Trabajo Fin de Máster (TFM) se tratará de resolver el problema de determinar

si en la comarca de Vigo hay más de una temporada de incendios, esto es, tratar de hacer inferencia sobre el número de modas de una distribución desconocida. En primer lugar, se puede pensar en abordar este problema de forma paramétrica, por ejemplo, a través del ajuste de una mixtura de densidades circulares. Pero esto puede conllevar al riesgo de cometer un error de especificación del modelo. Esto hace que se aborde este problema desde un punto de vista no paramétrico. Por lo tanto, será necesario analizar como estimar de forma no paramétrica la función de densidad y de distribución de Θ , siendo Θ una variable aleatoria circular.

Como lo que se pretende, en este TFM, es realizar un análisis exhaustivo de las distintas herramientas que permitan hacer inferencia sobre el número de modas en el caso circular ya que, como se mencionaba al principio de este capítulo, la fecha en la que se producen los incendios en la comarca de Vigo presenta una estructura cíclica. Se debe tener en cuenta que, en la literatura, el problema de analizar el número de modas ha sido estudiado en más profundidad para el caso lineal. Para poder realizar un estudio completo y aportar nuevas formas de tratar este problema, se analizará primero cómo abordar este problema para variables escalares y luego se tratará de extender las distintas ideas al caso circular.

Una forma de intentar determinar el número de modas es a través de herramientas exploratorias. Estas tienen la ventaja de que no solo permiten estimar su número si no que, además, permiten analizar su posición, esto es, en el caso de los incendios en la comarca de Vigo, permiten determinar el número de estaciones de incendios y su localización en el tiempo. Si bien, las herramientas exploratorias presentan el problema de que no proporcionan una manera formal de contrastar si la distribución de la muestra posee a lo sumo un número determinado de modas. Además, estas técnicas pueden no resultar adecuadas cuando se trata de realizar un análisis exhaustivo en grandes bases de datos. Estos dos inconvenientes de las herramientas gráficas quedan solventados con el uso de los test estadísticos para realizar el contraste de si la distribución de la muestra posee a lo sumo un número establecido de modas, lo que en particular, permitiría contrastar si en la comarca de Vigo hubo más de una temporada de incendios.

En lo que resta de capítulo, tanto para el caso lineal como para el circular, se analizarán las herramientas necesarias para poder realizar un análisis exploratorio que permita determinar el número de modas y su localización. También se introducirán los test formales para contrastar si la distribución de la muestra posee a lo sumo un número determinado de modas.

1.1. Herramientas exploratorias

Como se mencionó anteriormente, el problema de estudiar el número de modas surge cuando se dispone de una muestra aleatoria de datos y se desconoce la distribución de la que provienen. Una forma de tratar de determinar el número de modas sería estimar la densidad f de forma no paramétrica, ya que esta no exige conocer nada acerca de la distribución de la que procede la muestra. Para ello se puede emplear la estimación tipo núcleo (véase, por ejemplo, Wand y Jones (1995), Cap. 2). Dada una muestra aleatoria $\mathcal{X} = (X_1, \dots, X_n)$ de X se define el estimador tipo núcleo de f en el punto x como:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (1.1)$$

donde $h > 0$ es el parámetro ventana y K es una función de densidad unimodal simétrica respecto al origen, denominada núcleo. Un ejemplo de núcleo es la densidad gaussiana:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \text{ con } u \in \mathbb{R}. \quad (1.2)$$

En función del valor que tome el parámetro de ventana h , se pueden obtener distintas estimaciones de la densidad, que presentarán distinto número de modas. En la Figura 1.4 se muestran varias estimaciones no paramétrica de la densidad, utilizando el estimador (1.1), para distintos h . Se puede apreciar que, con el núcleo gaussiano, para una muestra de 600 datos obtenida a partir de la mixtura de normales M8 (véase Apéndice A), para valores bajos de h ($h = 0.17$ y $h = 0.23$) se tiene una estructura claramente bimodal, mientras que para valores altos de este parámetro ventana ($h = 0.4$ y $h = 0.6$) se observa una estructura unimodal. En particular, como se verá en el Teorema 3.1, con el núcleo (1.2) el número de modas será una función monótona decreciente de h . En este caso, si se usan como selectores de ventana la regla plug-in o el de validación cruzada insesgada (para obtener las expresiones de estos selectores véase Wand y Jones (1995), Cap. 3), se llegaría a la conclusión errónea de que la verdadera distribución es bimodal, pues estos devuelven parámetros de ventana próximos a 0.23, y se puede apreciar que la verdadera función de densidad de esta mixtura solo posee una moda.

Basándose en la idea de realizar la estimación tipo núcleo usando varias ventanas, existen diversas herramientas exploratorias como son el árbol de modas de Minnotte y Scott (1993), el bosque de modas de Minnotte *et al.* (1998) y el SiZer de Chaudhuri y Marron (1999) que permiten analizar cómo va variando el número de modas y su localización a medida que se modifica el valor del parámetro de ventana.

Con el fin de ilustrar el árbol y el bosque de modas, en la Figura 1.5 se muestra la estimación tipo núcleo (para parámetros de ventana $h = 1.25, h = 1.75, h = 4$ y $h = 8.5$), el árbol de modas (básico y completo) y el bosque de modas para una muestra de 272 observaciones de los tiempos de espera entre la ocurrencia de dos erupciones del géiser Old Faithful en el Parque Nacional de Yellowstone, Wyoming (USA). A continuación, se dará una idea intuitiva de cómo funcionan estas herramientas gráficas. Su desarrollo y construcción se presentarán en detalle en la Sección 2.1.

El árbol de modas básico (Figura 1.5, fila superior, derecha) permite, a través de las líneas continuas que se representan, estudiar la localización de las modas a medida que se va modificando en valor de h (eje vertical), mientras que las líneas discontinuas sirven para analizar como se pasa de tener k modas a $(k + 1)$ modas cuando disminuye el valor de h . En los puntos donde se “separan” los máximos relativos, se realiza un test basado en la masa de probabilidad (para obtener la definición de este concepto, véase Sección 2.1.1) para determinar qué modas son significativas, mostrándose a través de un círculo sólido las modas significativas a este nivel y con un círculo hueco las que no lo

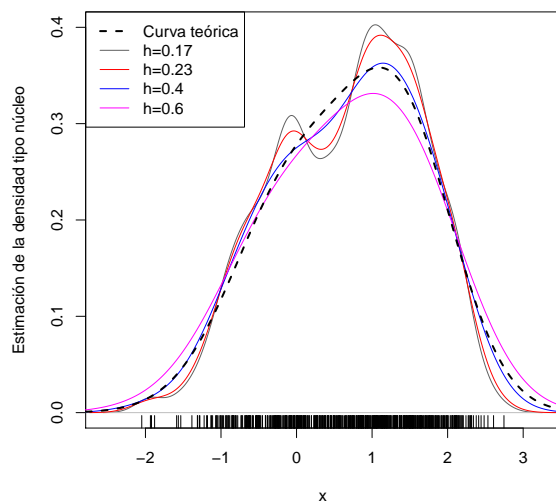


Figura 1.4: Densidad teórica y estimadores tipo núcleo, con núcleo gaussiano y parámetros ventana $h = 0.17$, $h = 0.23$, $h = 0.4$ y $h = 0.6$. Muestra de 600 datos de la mixtura de normales M8.

son. Observando la Figura 1.5, se puede ver que el árbol de modas básico detecta solo una moda significativa a este nivel, para los valores de h estudiados.

El árbol de modas completo (Figura 1.5, fila inferior, izquierda) permite, además de analizar de forma gráfica el número de modas, ver su masa de probabilidad asociada, ya que para los distintos valores de h estudiados, el grosor horizontal de cada moda (de color negro) será proporcional a la masa de probabilidad asociada a esa moda. Además, se añaden con líneas punteadas la localización de los mínimos relativos y de color gris los valores de x donde la segunda derivada de la estimación tipo núcleo es negativa indicando donde aparecen los “bultos” en la estimación tipo núcleo. Se muestran también la media y los cuartiles muestrales y el valor obtenido con un selector de ventana, Minnotte y Scott (1993) proponen emplear la obtenida por validación cruzada o la ventana sobresuavizada, h_{TS} , introducida por Terrell y Scott (1985).

Finalmente, el bosque de modas (Figura 1.5, fila inferior, derecha) permite estudiar, en función de la oscuridad de los píxeles representados, qué máximos relativos de la estimación tipo núcleo son realmente modas para distintos valores de h . Así, a la vista de la Figura 1.5, parece que para valores bajos de h (próximos a 1.5) se tendrían 6 modas, las más claras en torno a los valores 59, 78 y 82 (ya que es donde se tienen los píxeles más oscuros) y otras tres próximas a los valores 55, 62 y 89. Para el resto de valores de h (mayores a 2) se detectan dos modas, una en torno a 55 y otra en torno a 80.

El principal inconveniente que presenta el bosque de modas es que detecta como modas principales máximos relativos de la estimación tipo núcleo ocasionados por datos atípicos. Con el fin de solucionar

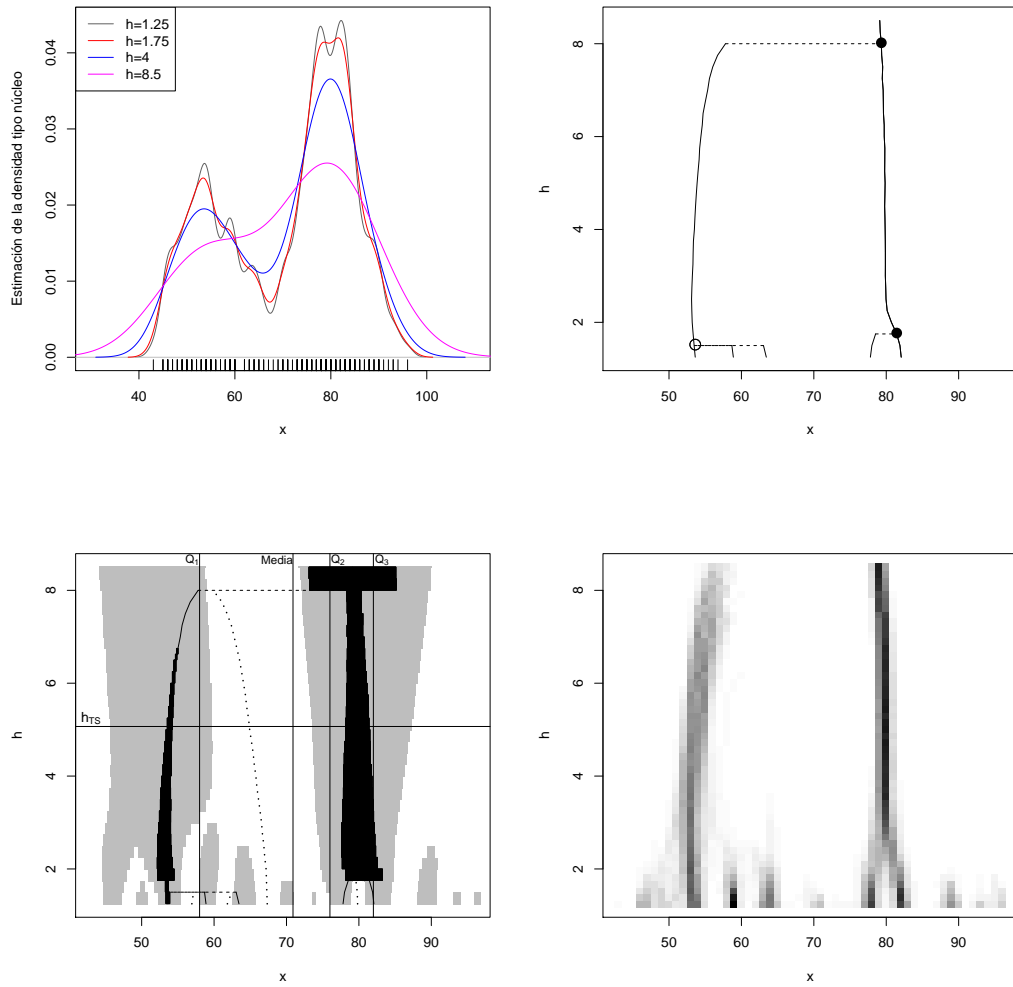


Figura 1.5: Fila superior: estimación tipo núcleo empleando núcleo gaussiano (izquierda) y árbol de modas básico (derecha). Fila inferior: árbol de modas completo (izquierda) y bosque de modas (derecha). Todos obtenidos a partir de la muestra de tiempos de espera entre la ocurrencia de dos erupciones de géiser para parámetros de ventana entre 1.25 y 8.5.

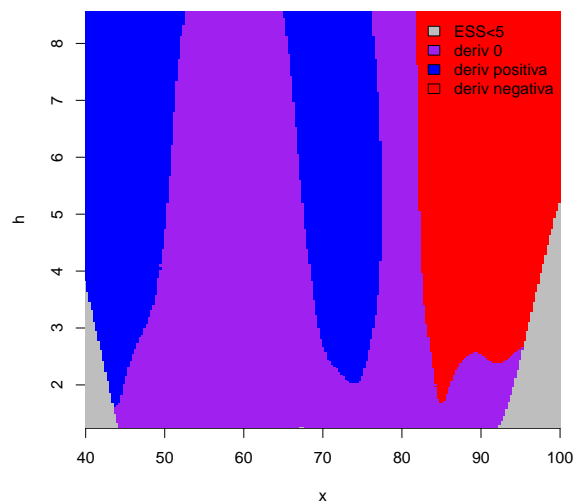


Figura 1.6: SiZer empleando parámetros de ventana h entre 1.25 y 8.5 (se presentan en el eje vertical) proveniente de la muestra de tiempos de espera entre la ocurrencia de dos erupciones de géiser.

este problema, se describirá en la Sección 2.2 el método SiZer¹ propuesto por Chaudhuri y Marron (1999). Este consiste en ir analizando cómo va variando la pendiente de la curva suavizada cuando se va modificando el valor del espacio (es decir, de x) y de la escala (esto es de h) con el fin de detectar cuando crece o decrece la función de densidad.

El SiZer para la muestra de tiempos de espera entre la ocurrencia de dos erupciones de géiser se muestra en la Figura 1.6. Para distintos valores de espacio y escala, el color azul indicará donde la pendiente de la curva suavizada es significativamente positiva, de color rojo donde es significativamente negativa, de color púrpura donde no es significativamente distinta de cero y el color gris indica donde no hay suficientes datos para determinar el comportamiento de la curva (con esto se evita que se muestren modas formadas por atípicos). En este caso, se llega a la conclusión de que solo hay una moda situada entorno al valor 80.

También existen herramientas gráficas que no utilizan la estimación tipo núcleo y que por tanto no son dependientes de h . Este es el caso de la herramienta exploratoria de Dümbgen y Walther (2008). Este método se basa en elaborar un contraste que permita determinar en qué regiones la densidad es significativamente creciente y en cuáles es significativamente decreciente (en sentido estricto) empleando un estadístico que depende únicamente de los valores de la muestra ordenados y del soporte de la variable aleatoria, que debe ser un intervalo continuo acotado, al menos, por uno de los extremos. En la Figura 1.7 (izquierda) se muestran los distintos intervalos de crecimiento y

¹Acrónimo de *Significant Zero*, que se puede traducir como Significativamente Cero, en referencia a cuando la pendiente es significativamente distinta de cero.

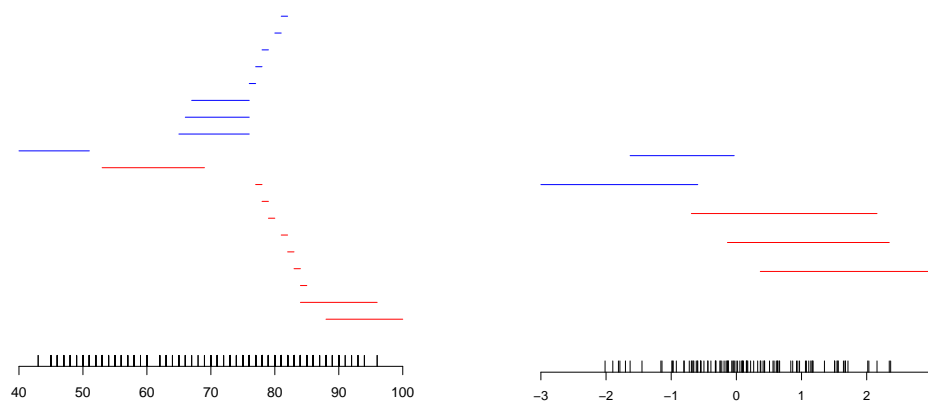


Figura 1.7: Herramienta exploratoria de Dümbgen y Walther (2008). Izquierda: muestra de tiempos de espera entre la ocurrencia de dos erupciones de géiser. Derecha: muestra de 100 observaciones provenientes de una distribución $N(0, 1)$

decrecimiento, obtenidos empleando la librería `modehunt` desarrollada por Rufibach y Walther (2013), para los datos de tiempo de espera entre erupciones. Cada una de las distintas líneas representan la localización de los intervalos de crecimiento, cuando estas son de color azul, y de decrecimiento, cuando estas son de color rojo. Si se supone que los tiempos de espera entre la ocurrencia de dos géiser deben de estar entre 40 y 100 minutos, esto es, que el soporte de esta variable aleatoria es $[40, 100]$, se puede apreciar en la Figura 1.7 que hay un intervalo de crecimiento empezando en el valor 40 y terminado en el valor 51 y que el resto de intervalos de crecimiento se sitúan entre el valor 70 y el 80. Los intervalos de decrecimiento para esta muestra serían uno situado entre los valores 53 y 69 y el resto situados entre los valores 80 y 100. Consecuentemente, a la vista de la Figura 1.7 se concluye que hay una moda en torno al punto 52 y otra en torno al punto 80.

En la Figura 1.7 (derecha) también se ilustran los intervalos significativamente crecientes y decrecientes para un nivel $\alpha = 0.01$ obtenidos a partir de una muestra de 100 datos provenientes de un $N(0, 1)$ cuando se toma como soporte el intervalo $[-3, 3]$. Este último ejemplo sirve para ilustrar el problema asociado a esta herramienta exploratoria. A veces, proporciona información inconsistente ya que se solapan los intervalos significativamente crecientes y decrecientes (en sentido estricto), con lo que no se puede determinar si la verdadera función de densidad es estrictamente creciente o decreciente en la zona donde se solapan los intervalos. Es por este motivo por el que no se analizará en más detalle esta herramienta exploratoria en el Capítulo 2.

1.2. Tipos de contrastes de multimodalidad

Si bien existen diversas herramientas exploratorias que permiten contabilizar el número de modas de una densidad, ninguna de ellas proporciona un test formal para contrastar si la distribución de la muestra posee a lo sumo un número determinado de modas y además, no son de utilidad cuando se quiere hacer un estudio sistemático sobre una cantidad elevada de muestras. Se revisará, en primer lugar, los test presentes en la literatura para realizar este contraste en el caso lineal, con el fin de tratar de obtener sus posibles adaptaciones al caso circular. Se verá que, en algunos casos, estos se pueden extender para contrastar si la verdadera distribución posee a lo sumo un número determinado de modas.

Formalmente, dada una muestra aleatoria simple de una variable aleatoria con función de densidad f , denotando por j al número de modas de f , se planteará el contraste de hipótesis:

$$H_0 : j \leq k, \text{ frente a } H_1 : j > k, \quad (1.3)$$

siendo $k \in \mathbb{Z}^+$ el número de modas que se quieren contrastar.

Con el objetivo de presentar los distintos métodos para realizar el contraste dado en (1.3), se han estudiado dos grandes bloques de test: los basados en la ventana crítica y los fundamentados en el exceso de masa o dip^2 .

Test basados en la ventana crítica

En la Sección 3.1, se analizarán los test basados en la ventana crítica para realizar el contraste (1.3), cuando $k \in \mathbb{Z}^+$. La ventana crítica, introducida por Silverman (1981), será la ventana h más pequeña para la cual se verifique que la estimación de la función de densidad dada en (1.1) posea a lo sumo k modas. A partir de esta ventana crítica, denotada por h_k , Silverman (1981) plantea usar este parámetro como estadístico de contraste, de forma que se rechazará la hipótesis nula cuando h_k sea muy “grande”, ya que esto significaría que es necesario sobresuavizar la estimación tipo núcleo para conseguir una estructura k -modal.

Con el fin de entender el concepto de ventana crítica, se puede emplear el árbol de modas básico que se muestra en la Figura 1.5. La ventana crítica, h_k , sería el último valor de h antes de que una de las k modas se “separe” dando lugar a una estimación con $(k + 1)$ modas. Así, por ejemplo, se podría ver en este gráfico, que para la muestra de tiempos de espera entre la ocurrencia de dos erupciones de géiser, la ventana crítica para una moda, h_1 , estará en torno al valor 8 (en concreto 8.065), mientras que la ventana crítica para dos modas, h_2 , estará en torno al valor 1.75 (exactamente 1.833). En la Figura 1.8 se muestran las estimaciones tipo núcleo cuando se toman como parámetros ventana h_1 y h_2 .

Con el fin de tratar de calibrar el test propuesto por Silverman (1981) cuando se quiere contrastar si la verdadera distribución es unimodal o multimodal (es decir $k = 1$), Hall y York (2001) obtienen la distribución de la ventana crítica, bajo H_0 , para una variable aleatoria con soporte conocido y

²Por su nombre en inglés, se puede traducir como test basados en la inmersión.

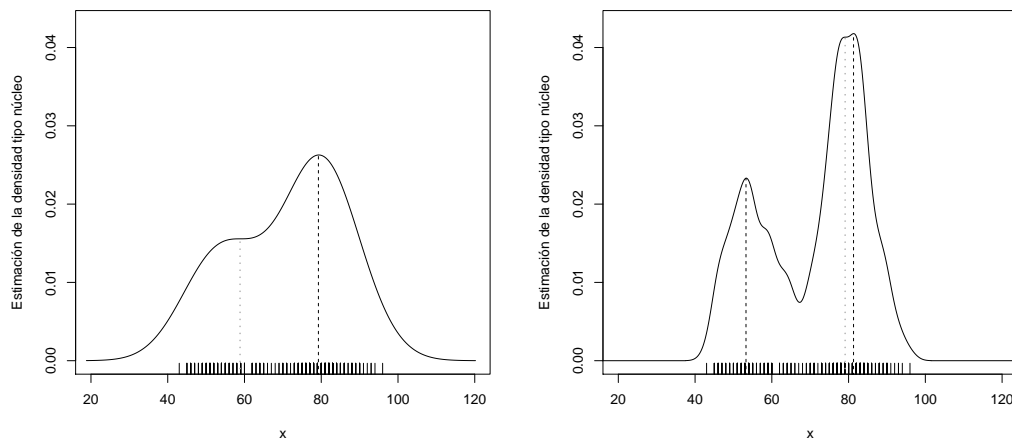


Figura 1.8: Estimación tipo núcleo empleando el núcleo gaussiano para la muestra de tiempos de espera entre la ocurrencia de dos erupciones de géiser. Izquierda: ventana crítica para $k = 1$ ($h_1 = 8.065$). Derecha: ventana crítica para $k = 2$ ($h_2 = 1.833$). Líneas de guiones (negro): localización de los máximos relativos. Trazo punteado (gris): localización donde surgirán las nuevas modas para parámetros de ventana menores que h_k .

acotado.

Por último, en este bloque de contrastes basados en la ventana crítica, se estudiará el test de Fisher y Marron (2001). Este consiste en realizar un contraste del tipo Cramér–von Mises, utilizando como estimación de la función de distribución bajo H_0 la integral de la estimación de la función de densidad, tomando como parámetro de suavizado a la ventana crítica.

Test basados en el *dip* o en el exceso de masa

Para poder analizar el segundo gran bloque de contrastes que se detallará en la Sección 3.2, será necesario introducir previamente la función de distribución empírica. Para ello, dada una muestra $\mathcal{X} = (X_1, \dots, X_n)$ se define la función de distribución empírica como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(X_i \leq x), \quad (1.4)$$

donde la función indicatriz $\mathcal{I}(A)$ toma los valores:

$$\mathcal{I}(A) = \begin{cases} 1 & \text{si se verifica } A, \\ 0 & \text{en otro caso.} \end{cases}$$

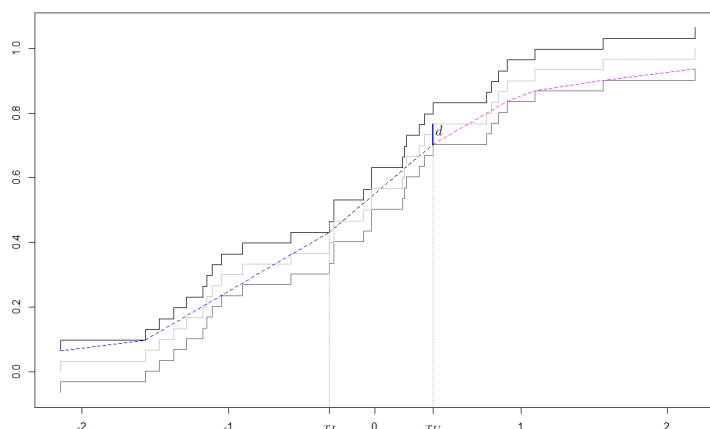


Figura 1.9: *Dip* para una muestra de 30 datos obtenida a partir de una $N(0, 1)$. Se muestra de color gris claro la distribución empírica de la muestra y de color más oscuro esta distribución cuando se le suma y resta el valor del *dip*. Con trazo discontinuo y de color azul, la parte donde la función monótona creciente es estrictamente convexa; de color rosa donde es estrictamente cóncava y de color negro donde la segunda derivada es cero. La moda se localizaría entre x_L y x_U y d es el valor que tomará el *dip*.

Haciendo uso de la función de distribución empírica, se analizará el test basado en el *dip* propuesto por Hartigan y Hartigan (1985) para realizar el contraste (1.3) cuando $k = 1$. Este se basa en calcular la distancia de la función de distribución empírica a la función monótona creciente (no necesariamente de distribución) más próxima, con un único intervalo donde la función es cóncava seguido de uno donde es convexa.

En la Figura 1.9 se muestra como se obtiene el valor del *dip* para una muestra de 30 datos proveniente de una $N(0, 1)$. La idea para conseguir el estadístico es obtener la función monótona creciente que sea cóncava en un intervalo $(-\infty, x_L]$ (de forma que su derivada sería creciente en este intervalo), convexa en el intervalo $[x_U, \infty)$ (de forma que su derivada sería decreciente en este intervalo) y recta en el intervalo (x_L, x_U) (de forma que su derivada tenga pendiente 0 en este intervalo) que sea la más próxima a F_n . El *dip* se calculará como la distancia (se usa la norma del supremo) entre la función así creada y F_n . El *dip* así obtenido sirve para medir la discrepancia de la función de distribución empírica a la distribución unimodal más próxima, de forma que se rechazará la hipótesis nula cuando el valor del *dip* sea muy “grande”.

También se estudiará el test de exceso de masa de Müller y Sawitzki (1991) para realizar el contraste (1.3) cuando $k \in \mathbb{Z}^+$, y que utilizará el exceso de masa como estadístico de contraste. Dado un valor $\lambda > 0$, la idea del exceso de masa se basa en encontrar k conjuntos cerrados y conexos (en el caso lineal k intervalos cerrados), a los que se le conocerá como λ -conglomerados, de forma que se maximice la diferencia entre la probabilidad asociada a estos conjuntos y λ veces la medida de

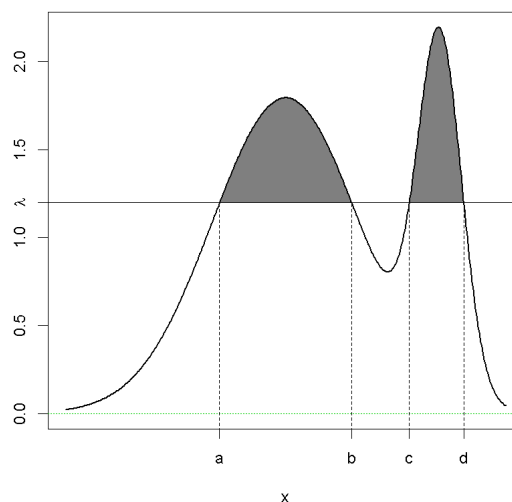


Figura 1.10: En gris se representa el exceso de masa para dos λ -conglomerados, los cuales son el intervalo $[a, b]$ y el $[c, d]$.

dichos conjuntos (en el caso lineal, la diferencia entre el extremo superior y el inferior del intervalo).

En la Figura 1.10 se muestra el exceso de masa poblacional para dos λ -conglomerados, que en este caso serían los intervalos $[a, b]$ y $[c, d]$. Estos dos intervalos son los que maximizan el valor del exceso de masa, que en este caso sería el área de la zona que se representa de color gris en la Figura 1.10, o lo que es lo mismo, $\mathbb{P}([a, b]) - \lambda(b - a) + \mathbb{P}([c, d]) - \lambda(d - c)$.

El estadístico de Müller y Sawitzki (1991) se obtendrá a partir de la máxima diferencia (cuando se hace variar el valor de λ) entre el exceso de masa de $(k + 1)$ y k λ -conglomerados. Se rechaza la hipótesis nula de que la función de densidad posee k modas cuando la diferencia entre tomar $(k + 1)$ conglomerados y k conglomerados sea muy “grande”. Además, se verá que esto se trata de una extensión del *dip*, ya que para $k = 1$, el valor del estadístico de Müller y Sawitzki (1991) es exactamente el doble del valor del *dip*. Por último, se ilustrará el calibrado de este test propuesto por Cheng y Hall (1998) para el caso de contrastar unimodalidad frente a multimodalidad.

Una de las aportaciones de este TFM será la de tratar de mejorar el calibrado de los test presentados hasta el momento. Con este fin, se ha realizado una nueva propuesta que consiste en aunar los dos grandes bloques de test de multimodalidad presentes en la literatura. Para ello, se obtendrá el valor del exceso de masa (o del *dip*) y se aproximará su distribución bajo H_0 empleando la metodología bootstrap (véase Sección 3.2). Esta metodología requerirá generar nuevas remuestras bajo H_0 . Para el remuestreo bootstrap, bajo H_0 , se propone usar la estimación tipo núcleo dada en (1.1), empleando como h la ventana crítica.

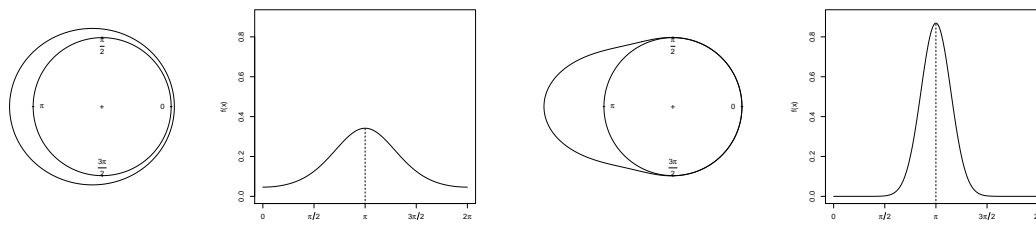


Figura 1.11: Ejemplos de distribuciones circulares unimodales. De izquierda a derecha: representación circular y lineal de la densidad circular $vM(\pi, 1)$ y representación circular y lineal de la densidad circular $WN(\pi, 0.9)$.

Como se verá en la siguiente sección, una ventaja de esta nueva propuesta es que presenta una adaptación que permitirá estudiar si hay una o más temporadas de incendios en la comarca de Vigo.

1.3. Detección de modas en densidades circulares

Como ya se ha mencionado, el objetivo final del presente trabajo será el de extender los test mostrados en el contexto de variable escalar al caso en el que la variable aleatoria es circular. Para ello, en primer lugar, se verá cómo se define la función de densidad, la de distribución y qué es una moda en el contexto circular.

Dada una variable aleatoria continua y circular Θ medida en radianes con soporte en el intervalo $[0, 2\pi)$, la función de densidad circular f^c , será una función verificando las siguientes condiciones (véase Jammalamadaka y Sengupta (2001), Cap. 2):

- $f^c(\theta) \geq 0, \forall \theta \in [0, 2\pi)$.
- $\int_0^{2\pi} f^c(\theta) d\theta = 1$.
- $f^c(\theta) = f^c(\theta + 2l\pi), \forall \theta \in [0, 2\pi)$ y $\forall l \in \mathbb{Z}$.

Un valor de Θ es una moda si en dicho punto, perteneciente al intervalo $[0, 2\pi)$, la función de densidad circular alcanza un máximo local. De nuevo, si hay una sola moda, la función de densidad circular (o la distribución asociada) será unimodal. En otro caso, la función de densidad circular (o la distribución asociada) se dirá multimodal (en particular, si la función de densidad circular posee dos modas será bimodal, etc.).

En la Figura 1.11, se ilustran algunos ejemplos de funciones de densidad circulares unimodales, como son la $vM(\pi, 1)$ (von Mises con media π y concentración 1) y la $WN(\pi, 0.9)$ (normal enrollada

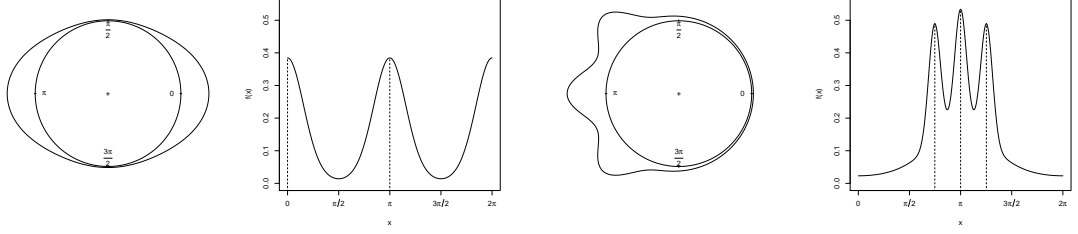


Figura 1.12: Ejemplos de distribuciones circulares multimodales. De izquierda a derecha: representación circular y lineal de la mixtura bimodal de von Mises M13 y representación circular y lineal de la mixtura trimodal de von Mises M15.

con media π y concentración 0.9) ambas con moda en el punto π (para obtener las expresiones de las distintas densidades circulares empleadas a lo largo de este trabajo véase Jammalamadaka y Sengupta (2001), Cap. 2). En la Figura 1.12, se ilustran ejemplos de distribuciones circulares multimodales, como son los las mixturas de von Mises M13 y M15 descritas en el Apéndice A.

A partir de la función de densidad circular f^c , se puede definir la función de distribución circular de una variable aleatoria continua, circular y con soporte en el intervalo $[0, 2\pi)$, como la función que verifica las siguientes propiedades (véase Di Marzio *et al.*, 2012):

- $F^c(\phi) = \int_0^\phi f^c(\theta)d\theta$ si se verifica que $\phi \in [0, 2\pi)^3$.
- $F^c(\phi + 2\pi) - F^c(\phi) = 1, \forall \phi \in \mathbb{R}$.

Es importante resaltar que, tal y como se puede observar en la Figura 1.13, en el caso circular, a diferencia de lo que ocurría en el caso lineal, se tiene que $\lim_{\phi \rightarrow -\infty} F^c(\phi) = -\infty$ y el $\lim_{\phi \rightarrow \infty} F^c(\phi) = \infty$.

Ante el problema de estimar el número de modas dada una muestra $\Theta = (\Theta_1, \dots, \Theta_n)$ de una variable aleatoria Θ con función de densidad circular f^c , descartando la alternativa paramétrica, una primera aproximación vendría dada por la estimación circular tipo núcleo de la función de densidad, \hat{f}_ν^c , que para un punto θ y concentración $\nu > 0$ se define como (véase, por ejemplo, Oliveira *et al.*, 2012):

$$\hat{f}_\nu^c(\theta) = \frac{1}{n} \sum_{i=1}^n K_\nu^c(\theta - \Theta_i), \quad (1.5)$$

³En general, para la función de distribución circular empezando en ϕ_0 , esta condición toma la siguiente forma: $F^c(\phi) = \int_{\phi_0}^\phi f^c(\theta)d\theta$ si se verifica que $\phi \in [\phi_0, \phi_0 + 2\pi)$.

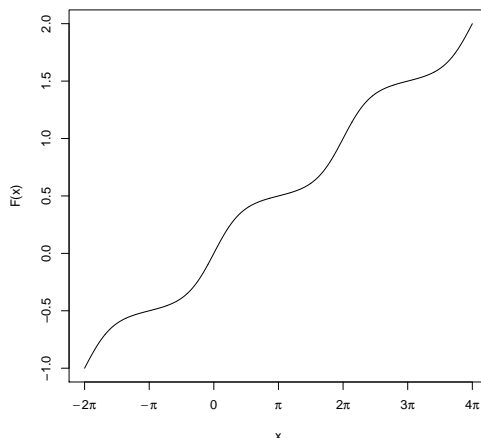


Figura 1.13: Función de distribución circular de una $vM(0, 1)$.

donde K_ν^c es el núcleo circular con parámetro de concentración $\nu > 0$. Un ejemplo de K_ν^c sería la función de densidad de una von Mises con media cero y concentración ν , es decir:

$$K_\nu^c(\varphi) = \frac{\exp(\nu \cos(\varphi))}{2\pi I_0(\nu)}, \quad (1.6)$$

donde I_0 es la función de Bessel modificada de primera especie y orden 0.

De forma análoga al caso lineal, en función del valor que tome el parámetro de concentración ν , se pueden llegar a distintas conclusiones acerca del número de modas que tendrá la verdadera función de densidad. Si bien, en este caso, el parámetro ν toma el papel inverso del parámetro de suavizado h , en el sentido de que, valores altos (bajos) de ν proporcionan estimaciones de la densidad con más (menos) modas. En la Figura 1.14 se muestra la estimación circular tipo núcleo dada en (1.5) para una muestra de 600 datos obtenidas a partir de una $C(\pi, 0.5)$ (Cardioide de media π y parámetro de concentración 0.5), empleando el núcleo circular von Mises dado en (1.6), para parámetros de concentración $\nu = 5$ (donde tiene una estructura unimodal), $\nu = 18.12$ y $\nu = 59.53$ (donde se tiene una estructura multimodal). Teniendo en cuenta que los dos últimos parámetros de concentración empleados son, respectivamente, el proporcionado por la regla plug-in (seleccionando el número de componentes de la mezcla de entre 2 y 5 von Mises por el criterio de información de Akaike) de Oliveira *et al.* (2012) y el obtenido por validación cruzada, empleando selectores automáticos del parámetro de concentración (véase Oliveira *et al.* (2012) para obtener las expresiones de ambos selectores) se llegaría a la conclusión errónea de que la verdadera distribución es multimodal.

Como se mencionó al principio de este capítulo, el tratamiento de los datos circulares exige un estudio especial, lo que hace que no sean válidas las herramientas empleadas en el contexto lineal. Si bien es cierto que, en muchos casos, sí que se pueden exportar alguna de las ideas de este contexto.

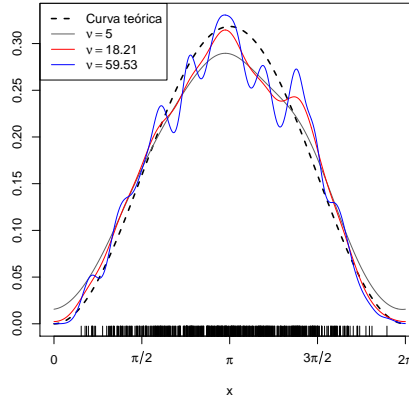


Figura 1.14: Curva teórica y estimadores circulares tipo núcleo de la función de densidad (con núcleo circular von Mises y parámetros de concentración $\nu = 5$, $\nu = 18.21$ y $\nu = 59.53$) para una muestra de 600 datos obtenidas a partir de una $C(\pi, 0.5)$.

Así, por ejemplo, en el caso de las herramientas gráficas se tiene el CircSiZer de Oliveira *et al.* (2012b) analizado en la Sección 2.3.1. Esta herramienta permite detectar el número de modas y su localización aproximada analizando el comportamiento de la derivada la curva suavizada cuando se va modificando el parámetro de localización, θ , y de concentración, ν .

En la Figura 1.15 se muestra el CircSiZer para una muestra de 155 observaciones de la fecha en el que se detectaron los incendios en la comarca de Vigo (estos datos se describen con más detalle al final del capítulo). De forma análoga a la realizada en el SiZer en el caso lineal, para distintos valores de θ y de ν , el color azul indicará las zonas donde la pendiente de la densidad suavizada es significativamente positiva, de color rojo donde es significativamente negativa, de color púrpura donde no es significativamente distinta de cero y el color gris indica donde no hay suficientes datos para determinar el comportamiento de la curva suavizada. Se toma el sentido marcado por la flecha (en este caso, el sentido horario) como el sentido positivo de la rotación. Los valores de $-\log_{10}(\nu)$ se representan en el radio de la circunferencia de manera que cada anillo se corresponde con un valor de ν , de menos a más suavizado (desde el centro al exterior). A partir de la Figura 1.15, para valores altos de ν (próximos a $-\log_{10}(\nu) = -1.40$) se detectan dos modas, una a finales de Febrero y otra a mediados de Agosto. Mientras que para valores bajos de ν (próximos a $-\log_{10}(\nu) = 0$) solo se observa una moda y esta se produce a mediados de Agosto.

La utilidad de estas herramientas gráficas es la que se observa en el gráfico de la derecha de la Figura 1.15 y es que, para una región, permite tanto estimar el número de estaciones de incendios como dar unas fechas aproximadas en la que se producen dichas temporadas. Las desventajas son las ya mencionadas anteriormente y es que no proporcionan una manera de determinar, controlando el nivel de significación, si esta densidad es unimodal o multimodal, y que no permiten realizar un análisis exhaustivo en grandes bases de datos.

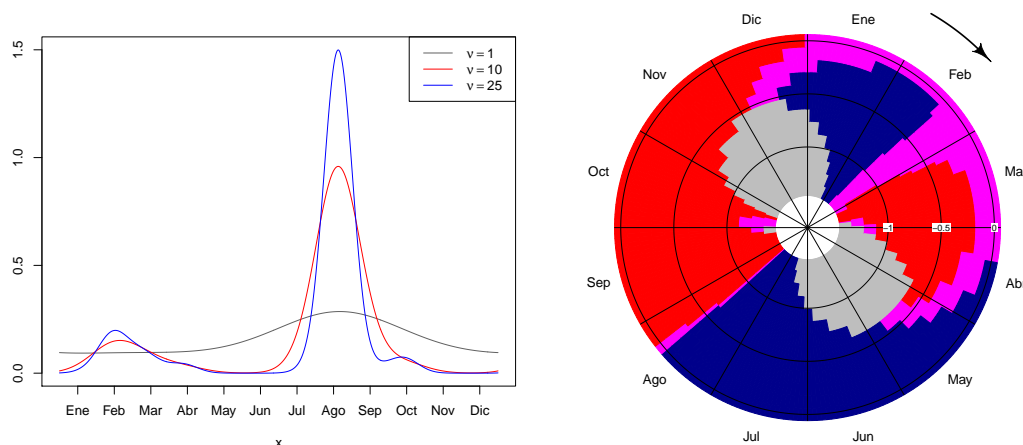


Figura 1.15: Estimación tipo núcleo (izquierda) y CircSiZer (derecha) empleando parámetros de concentración ν entre 25 y 1 (esto es, $-\log_{10}(\nu)$ entre -1.40 y 0) proveniente de la muestra de 155 observaciones de la fecha en el que se detectaron los incendios en la comarca de Vigo.

Con el objetivo de solventar estos inconvenientes en este contexto, se propone nuevamente introducir test formales para contrastar la hipótesis de que la verdadera distribución asociada a una muestra circular es unimodal y se verá que en algunos casos se podrá extender estos contrastes al caso de comprobar si la verdadera distribución posee a lo sumo k^c modas en $[0, 2\pi)$. Esto es, dada una muestra aleatoria simple $\Theta = (\Theta_1, \dots, \Theta_n)$ procedente de una variable aleatoria con función de densidad circular f^c , si el número de modas de la verdadera distribución circular es j^c , se planteará el contraste de hipótesis:

$$H_0 : j^c \leq k^c, \text{ frente a } H_1 : j^c > k^c, \quad (1.7)$$

siendo $k^c \in \mathbb{Z}^+$ el número de modas que se quieren contrastar.

Del mismo modo al realizado en el caso lineal, se estudiarán dos grandes bloques de test para realizar el contraste dado en (1.7), los basados en la concentración crítica y en el exceso de masa.

Test basados en la concentración crítica.

En la Sección 4.1 se estudiarán los test basados en la concentración crítica, cuyo fundamento es similar al de los basados en la ventana crítica, sin más que tener en cuenta que, en este caso, el número de modas de la estimación tipo núcleo circular dada en (1.5) es alto para valores grandes del parámetro de concentración, mientras que para valores pequeños de ν esta estimación se va aproximando cada vez más a una uniforme y consecuentemente, esta estimación va a tener un número pequeño de modas.

La obtención del parámetro de concentración crítica es similar al de la ventana crítica de Sil-

verman (1981): se trata de buscar el parámetro de concentración ν_{k^c} más grande de forma que la estimación $\hat{f}_{\nu_{k^c}}^c$ posea a lo sumo k^c modas en el intervalo $[0, 2\pi)$. Haciendo uso de este parámetro como estadístico, en el presente trabajo se propone un test para realizar el contraste (1.7). Para ello, se debe tener en cuenta que, en este caso, el sobreesuavizado se produce para parámetros de concentración bajos, por tanto, se rechazará la hipótesis nula cuando el estadístico de contraste sea muy “pequeño”. En particular, como nueva propuesta para este TFM, se adaptará la idea de Hall y York (2001) para contrastar unimodalidad frente a multimodalidad, ya que en este caso, se conoce el soporte de la variable aleatoria y está acotado.

Se revisará también el test de Fisher y Marron (2001) donde se utilizará este parámetro de concentración crítica para obtener la función de distribución circular suavizada con la que realizar un test del tipo U^2 de Watson (1961).

Test basado en el exceso de masa.

En la Sección 4.2 se realizará una adaptación del concepto de exceso de masa de Müller y Sawitzki (1991) al contexto circular. Para calibrar este test, en este TFM se propone emplear de nuevo la metodología bootstrap, utilizando para ello como densidad bajo H_0 la estimación tipo núcleo circular con parámetro ν igual a la concentración crítica.

1.4. Estructura del trabajo

En el Capítulo 2 se presentan con mayor detalle las herramientas gráficas para la estimación del número de modas, tanto para el caso lineal como para el contexto circular. Se comienza con una revisión en detalle del árbol y del bosque de modas para posteriormente analizar la construcción del SiZer y del CircSiZer.

En el Capítulo 3 se estudian distintos test para contrastar si la verdadera distribución asociada a una muestra de una variable escalar posee a lo sumo k modas. Se revisarán los distintos test existentes y se propone un nuevo test para contrastar unimodalidad frente a multimodalidad. El funcionamiento tanto del nuevo método como el de los existentes en la literatura se comprueba en un estudio de simulación, considerando una amplia clase de familias de densidad, tanto unimodales como multimodales.

En el Capítulo 4 se analizan distintos test para contrastar si la verdadera distribución asociada a una muestra de datos circulares posee a lo sumo k^c modas. Se revisará el test existente y se propondrá una adaptación, tanto del método de Hall y York (2001) como de la nueva propuesta realizada en el caso lineal, al contexto circular para contrastar unimodalidad frente a multimodalidad. El funcionamiento tanto de los nuevos métodos presentados como el del existente en la literatura se analiza en un estudio de simulación, considerando una amplia clase de familias de densidad circulares, tanto unimodales como multimodales.

En el Capítulo 5 se ilustrarán la aplicación de los diversos test introducidos en los capítulos anteriores mediante el análisis de dos conjuntos de datos en el contexto lineal y de otros dos en el circular.

La primera base de datos analizada será el de la muestra de 272 observaciones recogida entre el 1 y el 15 de Agosto de 1985 de los tiempos de espera entre la ocurrencia de dos erupciones del géiser Old Faithful situado en el Parque Nacional de Yellowstone en Wyoming, USA. Esta muestra está disponible en el conjunto de datos `faithful` del paquete `base` del software estadístico R Core Team (2013) y se describe en detalle en Azzalini y Bowman (1990). El objetivo de este estudio será el de analizar si todos los tiempos de espera se agrupan en torno a un valor, o si en cambio, hay al menos dos (unos tiempos de espera cortos y otros largos) tendencias.

El segundo conjunto de datos que se analizará es el de la renta por hogar en Galicia durante el año 2012. Para ello, a partir de la Encuesta de Condiciones de Vida de las Familias se han extraído de la web del Instituto Galego de Estatística (2012)⁴ la renta de 9190 hogares gallegos. El objetivo de este estudio será el de analizar si durante el año 2012 hubo solo una clase social, en la que todas las familias gallegas se situaron en torno a una única renta, o si en cambio hubo al menos dos (una más pobre y otra más rica).

La tercera base de datos servirá para ilustrar el funcionamiento de los test para el contexto circular y vendrá dada por la dirección media de la que procede el viento, recogida por un medidor situado en la localidad de Vilalba, cada vez que hay un exceso de NO_x en la atmósfera (esto es cada vez que el medidor detecta que hay más de $200\mu\text{g}/\text{m}^3$ de este contaminante), con el fin de analizar si todas las alarmas que se producen por exceso de NO_x podrían ser atribuidas a la central térmica de As Pontes de García Rodríguez o hay más fuentes de contaminación por NO_x en otras direcciones.

Por último, se muestra el ejemplo que ha motivado en gran parte este TFM y es el de las 155 observaciones acerca de cuándo se producen los incendios en la comarca de Vigo. Para ello, se han recogido datos del día en el que ocurren los incendios en esa zona durante 10 años, entre el 10 de Julio de 2002 y el 9 de Julio de 2012, eliminándose completamente el año 2006 por ser un año atípico en cuanto a número de incendios en Galicia. El objetivo, en este caso, es el de analizar si hay una única temporada de incendios o si, en cambio, hay al menos dos temporadas de incendios en dicha región. Agradecer al Profesor Dr. José Miguel Cardoso Pereira y a su grupo de Ecología Forestal de la Universidad de Lisboa la aportación de estos datos para la realización del presente TFM.

Finalmente, este trabajo incluye dos apéndices. En Apéndice A se describen los distintos modelos empleados a lo largo de este TFM. En Apéndice B se presenta un cronograma con el objetivo de representar cuándo fueron surgiendo las herramientas gráficas y los test analizados en los Capítulos 2, 3 y 4. Además, se muestra en el Apéndice B un pequeño resumen de las principales características de los test empleados en los Capítulos 3 y 4.

Todos los test presentados en los Capítulos 3 y 4, tanto los ya existentes como las nuevas aportaciones realizadas, han sido implementados en el software estadístico R Core Team (2013) para el presente TFM, a excepción del *dip* de Hartigan y Hartigan (1985), programado por Maechler (2013) en la librería `dip.test`.

⁴Estos datos han sido consultados el 01/04/2014 y se han extraído de la Edición del 2013 de la Encuesta de Condiciones de Vida de las Familias.

Capítulo 2

Herramientas gráficas

En este capítulo se analizarán en más profundidad las diversas herramientas exploratorias presentadas en las Secciones 1.1 y 1.3.

En el contexto lineal se comenzará estudiando el árbol y el bosque de modas. El árbol de modas, propuesto por Minnotte y Scott (1993), consiste en ir representando la localización de las modas que se obtienen cuando se utiliza la estimación tipo núcleo para diversos valores de h . Con el objetivo de ver qué modas, de las representadas, son significativas, Minnotte (1997) propone un test basado en la masa de probabilidad que tiene una moda antes de separarse. Finalmente, con el fin de tratar de eliminar la dependencia que tiene el árbol de modas de la muestra original, se presenta el bosque de modas de Minnotte *et al.* (1998).

La siguiente herramienta gráfica que se analizará será el SiZer, propuesto por Chaudhuri y Marron (1999). Esta consiste en crear intervalos de confianza para la derivada de la curva suavizada para distintos valores de espacio y escala. Esta herramienta servirá para determinar de forma exploratoria en qué puntos la curva suavizada es significativamente creciente y decreciente.

Finalmente, con el fin de determinar el número de modas en los datos circulares, se presenta el CircSiZer desarrollado por Oliveira *et al.* (2012b), una adaptación del SiZer al contexto circular.

2.1. El árbol y el bosque de modas

En esta sección se analizarán en detalle el árbol de modas propuesto por Minnotte y Scott (1993) y el bosque de modas introducido por Minnotte (1997) como herramientas gráficas para la detección del número de modas.

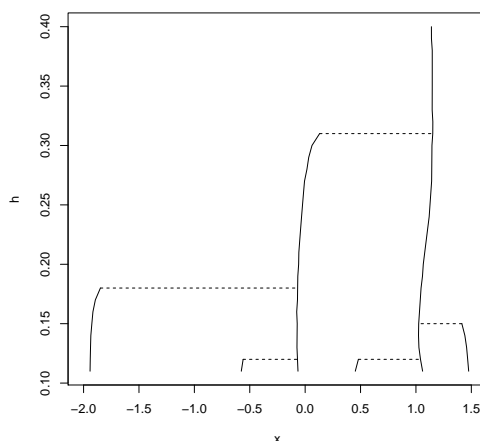


Figura 2.1: Árbol de modas básico empleando núcleo gaussiano con parámetros de ventana h entre 0.11 y 0.40 para la muestra de 600 datos obtenidas a partir de la mixtura de normales M8 que se empleó en la Figura 1.4. La líneas continuas representan la localización de las modas para cada h y las rectas horizontales discontinuas la “separación” de una moda.

2.1.1. El árbol de modas

Usando la estimación tipo núcleo dada en (1.1), la herramienta gráfica que proponen Minnotte y Scott (1993) consiste en, fijado el núcleo K , crear un árbol de modas representando la localización de los máximos relativos de \hat{f}_h para distintos valores de h . De este modo, se muestran las posiciones de las modas (en el eje de abscisas) para los distintos valores de h empleados (eje de ordenadas) creando un conjunto de líneas, los rastros de las modas. Con el fin de ver cómo se van “separando” o “juntando” las modas también se unen los rastros de las diferentes modas cuando estos desaparecen a partir de un cierto h . Teniendo en cuenta que cuando aparece una nueva moda, también surge un nuevo mínimo relativo, si la nueva moda está situada a la derecha del nuevo mínimo relativo, entonces significa que se ha separado de la moda de la izquierda, y por tanto se unirá el principio del rastro de esta nueva moda con una línea discontinua horizontal con la moda de la que se ha separado. Análogamente, si la nueva moda está situada a la izquierda del nuevo mínimo relativo, se unirá el principio del rastro de esta moda con la moda de su derecha. En la Figura 2.1 se muestra el árbol de modas, empleando núcleo gaussiano y valores de h entre 0.11 y 0.40, para la muestra de 600 datos obtenidas a partir de la mixtura de normales M8 (véase Apéndice A) que se empleó en la Figura 1.4.

Se debe tener en cuenta que, si el núcleo empleado no es gaussiano, puede no haber monotonía en el número de modas de \hat{f}_h , en el sentido de que siempre que aparece un máximo relativo, este no tiene porqué seguir apareciendo para valores más bajos del parámetro de ventana. En caso de emplear el núcleo gaussiano, como es el caso de la Figura 2.1, esta monotonía estará asegurada por

el Teorema 3.1 que se verá en la Sección 3.1.1, el cual garantiza que el número de modas de \hat{f}_h es una función monótona decreciente de h : todas las modas que aparecen para una determinada ventana se mantendrán para cualquier valor de h más pequeño.

Con el fin de representar más información en este árbol de modas, Minnotte y Scott (1993) proponen reemplazar las líneas de los rastros de las modas por regiones centradas alrededor de cada máximo relativo, cuyo ancho horizontal representará la masa de probabilidad asociada a dichos máximos. Esta masa de probabilidad se define de forma similar al exceso de masa de Müller y Sawitzki (1991) que se presenta en la Sección 3.2.2, estimando f a través de \hat{f}_h . Para dar la expresión que tiene dicha masa de probabilidad, si la estimación \hat{f}_h posee l máximos locales, se denotará como w_2, \dots, w_l a la localización (en orden creciente) de los mínimos relativos (si $l > 1$, ya que en otro caso \hat{f}_h no posee mínimos relativos), con $w_1 = -\infty$ y como $w_{l+1} = \infty$. A partir de estos w_i se define como masa de probabilidad de la moda i -ésima al siguiente valor:

$$M_i(h) = \lim_{\substack{t_1 \rightarrow w_i \\ t_2 \rightarrow w_{i+1}}} \int_{t_1}^{t_2} \{\hat{f}_h(x) - \max\{\hat{f}_h(t_1), \hat{f}_h(t_2)\}\}_+ dx, \quad (2.1)$$

donde $\{\cdot\}_+$ denota la parte positiva, ya que $\hat{f}_h(x) - \max\{\hat{f}_h(t_1), \hat{f}_h(t_2)\}$ podría tomar valores negativos. Así, el ancho horizontal de la moda i -ésima de \hat{f}_h (representado de color negro en el rastro de cada uno de los máximos relativos de \hat{f}_h) deberá ser proporcional al valor del $M_i(h)$.

Minnotte y Scott (1993) también proponen mostrar a través de líneas punteadas la localización de los mínimos relativos. Además, con el fin de detectar puntos de inflexión, se representará de color gris las zonas donde la segunda derivada de la estimación tipo núcleo es negativa.

Finalmente, se representarán también la media, los cuartiles muestrales y el valor de h obtenido con un selector de ventana. Minnotte y Scott (1993) proponen tomar o bien el de validación cruzada (véase Wand y Jones (1995), Cap. 3) o el sobreesuavizado introducido por Terrell y Scott (1985) cuya expresión es la siguiente:

$$h_{TS} = 3\hat{\sigma}(70\sqrt{\pi n})^{-1/5},$$

donde $\hat{\sigma}$ es la cuasidesviación típica muestral.

En la Figura 2.2 se representa el árbol de modas mejorado para la misma muestra que en la Figura 2.1. Para construir este árbol de modas se ha utilizado el núcleo gaussiano y parámetros de ventana entre 0.11 y 0.4. El ancho de la región oscura asociada a la moda i -ésima es proporcional al valor de $M_i(h)$, en particular, por cuestiones de visibilidad, se ha tomado un ancho igual a 0.8 veces $M_i(h)$.

Con el objetivo de contrastar cuales de los máximos relativos de \hat{f}_h son realmente modas, Minnotte (1997) propone un test similar al de Müller y Sawitzki (1991), que se revisará en la Sección 3.2.2. Denotando como v_1, \dots, v_l a la localización en orden creciente de las l modas de \hat{f}_h , de forma que si para la estimación \hat{f}_h , se aprecian l máximos locales, se tendría que:

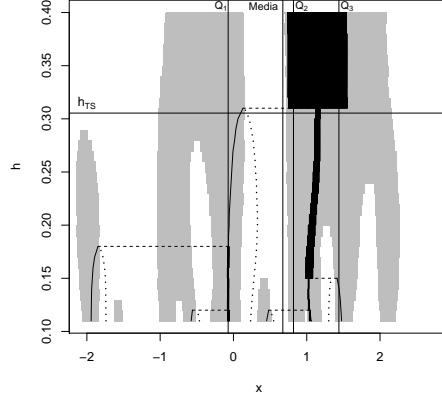


Figura 2.2: Árbol de modas mejorado empleando núcleo gaussiano con parámetros de ventana h entre 0.11 y 0.40 para la muestra de 600 datos obtenidas a partir de la mixtura de normales M8 que se utilizó en la Figura 1.4. El ancho de las regiones negras es igual a 0.8 veces $M_i(h)$. Las líneas punteadas representan la localización de los mínimos. Las rectas horizontales de guiones muestran la “separación” de una moda. Las regiones de color gris indican donde la segunda derivada de \hat{f}_h es negativa.

$$-\infty = w_1 < v_1 < w_2 < \dots < v_l < w_{l+1} = \infty.$$

Para realizar un contraste que permita rechazar o no la hipótesis nula de que el máximo relativo v_i no es realmente una moda, Minnotte (1997) propone usar como estadístico de contraste el $M_i(h)$, con h el parámetro de ventana más pequeño (dentro del rango de ventanas estudiadas en el árbol de modas) a partir del cual el máximo relativo v_i se “separa” formando dos máximos relativos. Notar que, aunque un máximo local se divida en dos para valores muy “pequeños” de h , solo se estudiarán los máximos relativos que se “separan” para el rango de valores h representados en el árbol de modas.

La hipótesis nula de que el máximo relativo v_i no es realmente una moda se rechazará para valores altos del estadístico $M_i(h)$ dado en (2.1). Para conseguir la distribución del estadístico bajo H_0 , Minnotte (1997) propone emplear la metodología bootstrap (véase Efron, 1979), obteniendo valores del estadístico para distintas remuestras generadas bajo H_0 y con estos valores del estadístico aproximar la distribución de $M_i(h)$.

Para poder generar las réplicas bajo H_0 , será necesario que el punto v_i no sea un máximo relativo. Si $h_{MS}(i)$ es la ventana empleada en el cálculo del estadístico de contraste, para generar las remuestras bootstrap, Minnotte (1997) propone hacer uso de la distribución cuya función de densidad asociada es la estimación $\hat{f}_{h_{MS}(i)}$ modificada para que esta no posea un máximo relativo entre v_{i-1} y v_{i+1} (entre $-\infty$ y v_2 si $i = 1$ o entre v_{l-1} y ∞ si $i = l$).

Con el fin de encontrar la densidad $\tilde{f}_{h_{MS}(i)}$ más “parecida” a $\hat{f}_{h_{MS}(i)}$ verificando que no posee

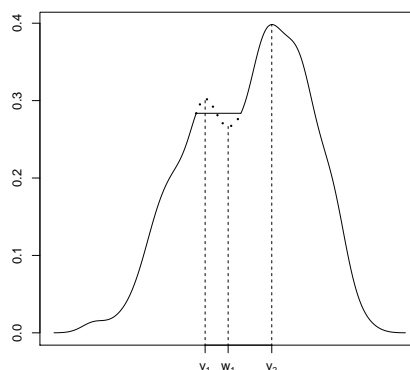


Figura 2.3: Ejemplo de función de densidad asociada a la distribución a partir de la cual se generarán las remuestras bootstrap, cuando $h = 0.19$ e $i = 1$. Se representa con trazo continuo la función de densidad que será empleada, $\hat{f}_{0.19}$. En trazo punteado se muestra la estimación original de la función de densidad, $\hat{f}_{0.19}$.

ninguna moda en el intervalo (v_{i-1}, v_{i+1}) (en el $(-\infty, v_2)$ si $i = 1$ o en el (v_{l-1}, ∞) si $i = l$), Minnotte (1997) propone realizar un procedimiento que consistirá en desplazar parte de la masa de probabilidad del máximo relativo v_i a uno de los dos “valles” (o a los dos si fuera necesario) donde se encuentran los mínimos relativos asociados a v_i hasta que v_i deje de ser máximo relativo. El “valle” en el que se empezará a desplazar la masa de probabilidad se escogerá en función de cual sea la moda más pronunciada de las dos que se encuentran a cada uno de los lados de v_i . Una vez relleno este “valle” si v_i sigue siendo una moda, entonces se desplazará la masa de probabilidad sobrante al otro “valle”. Obviamente, en caso de solo haya un “valle” entonces se desplazará toda la masa de probabilidad a este, hasta que v_i no sea moda.

En la Figura 2.3 se ilustra un ejemplo de $\tilde{f}_{h_{MS}(i)}$ tomando la muestra de 600 datos procedente de la distribución M8. Para ello, se ha escogido el primer máximo relativo, v_1 , que aparece en la Figura 2.1 para valores de h entre 0.19 y 0.31. Como el parámetro de ventana más pequeño de los estudiados para el cual se sigue teniendo solo esta moda es 0.19, ya que para valores más bajos de h el rastro de esta moda se divide en dos, se tomará el estadístico de contraste dado en (2.1) cuando $h = 0.19$, esto es, $M_1(0.19)$. Como en este caso, la moda que se debe “eliminar” de $\hat{f}_{0.19}$ es la primera, para obtener $\tilde{f}_{0.19}$ se desplazará toda la masa de probabilidad necesaria de v_1 al “valle” de w_1 hasta que se consiga que $\tilde{f}_{0.19}$ sea unimodal.

A partir de la distribución asociada a $\tilde{f}_{h_{MS}(i)}$ se generarán B remuestras bootstrap, cuyo tamaño muestral será el mismo que el de la muestra original. Para cada una de estas réplicas se calcula la estimación tipo núcleo de la función de densidad tomando como parámetro de ventana $h_{MS}(i)$, esto es, $\hat{f}_{h_{MS}(i)}^{*b}$ con $b = 1, \dots, B$. Haciendo uso de cada una de estas estimaciones, se buscarán los máximos

relativos de $\hat{f}_{h_{\text{MS}}(i)}^{*b}$ (a los que se denotará como $v_1^{*b} < \dots < v_{l_b}^{*b}$) más próximos a los de $\hat{f}_{h_{\text{MS}}(i)}$ a través del algoritmo de emparejamiento proporcionado en el Apéndice del artículo de Minnotte y Scott (1993).

Tras realizar este proceso de emparejamiento, se tomará aquella moda de $\hat{f}_{h_{\text{MS}}(i)}^{*b}$ que proporcione una mayor masa de probabilidad en el intervalo de interés. Este intervalo estará acotado por los máximos relativos emparejados con v_{i-1} y v_{i+1} o, como podría ocurrir que no todas las modas estuviesen emparejadas, en caso de no estarlo, por la propia localización de las modas (se tomará como acotación inferior $-\infty$ cuando $i = 1$ y como acotación superior ∞ cuando $i = l$)¹. Para la moda así obtenida, se aplica el proceso de disminuir el valor de la h hasta que esta moda se “separe” en dos (denotando como $h_{\text{MS}}^{*b}(i)$ a esta ventana), y a partir de esta ventana se calcula el exceso de masa de dicha moda, al que se le denotará como $M_i(h_{\text{MS}}^{*b}(i))^{*b}$. A partir de los B valores obtenidos de $M_i(h_{\text{MS}}^{*b}(i))^{*b}$ se puede aproximar la distribución de la masa de probabilidad bajo H_0 . De esta forma, para un nivel de significación α , se rechazará la hipótesis nula de que v_i no es moda si $\mathbb{P}(M_i(h_{\text{MS}}^*(i))^* \leq M_i(h_{\text{MS}}(i)) | \mathcal{X}) \geq 1 - \alpha$, siendo \mathcal{X} la muestra original.

Una vez realizado este contraste para todas las modas que se “separan” en el árbol de modas mostrado en la Figura 2.1 se procederá a representar donde se rechaza o no la hipótesis nula en dicho árbol. Para ello, cuando es el máximo relativo v_i el que se separa en $h_{\text{MS}}(i)$, si se rechaza la hipótesis nula, esto es, si la moda es significativa a nivel α , se situará un círculo relleno en el árbol de modas básico en el punto $(v_i, h_{\text{MS}}(i))$. Por el contrario, se representará con un punto hueco si no hay evidencias significativas para rechazar la hipótesis nula. Ante lo conservador que es este test, Minnotte (1997) propone emplear un nivel de significación $\alpha = 0.15$.

En la Figura 2.4 se muestra el árbol de modas representado en la Figura 2.1 incluyendo los contrastes realizados para los distintos $M_i(h)$, donde i es el índice de las distintas modas que se separan. Además, aunque este test está pensado para cuando hay al menos dos modas, imitando las representaciones que realiza Minnotte (1997) se incluye un punto relleno en la primera “separación” que se produce. Obviamente, esta moda siempre va a ser significativa para cualquier valor de significación α , pues en este caso se tiene que $M_1(h_{\text{MS}}(1)) = 1$, y como $M_1(h_{\text{MS}}^{*b}(1))^{*b} \leq 1$, con $b = 1, \dots, B$, se verifica que $\mathbb{P}(M_1(h_{\text{MS}}^*(1))^* \leq M_1(h_{\text{MS}}(1)) | \mathcal{X}) = 1 \geq 1 - \alpha, \forall \alpha \in [0, 1]$.

2.1.2. El bosque de modas

Ante la dependencia que presenta el árbol de modas con la muestra empleada, Minnotte *et al.* (1998) proponen representar varios árboles de modas para la misma muestra. Con este objetivo, se generan varias remuestras bootstrap de la muestra original empleando la distribución asociada a la función de distribución empírica dada en (1.4), esto es, obtener nuevas remuestras de tamaño n obtenidas a partir de la selección al azar con reemplazamiento de los individuos de la muestra original $\mathcal{X} = (X_1, \dots, X_n)$.

Para facilitar la visualización, en vez de representar todos los árboles de modas, Minnotte *et*

¹En el artículo original de Minnotte (1997) no se especifica nada acerca de que pasos seguir cuando este intervalo no contiene a ningún máximo relativo de $\hat{f}_{h_{\text{MS}}(i)}^{*b}$.

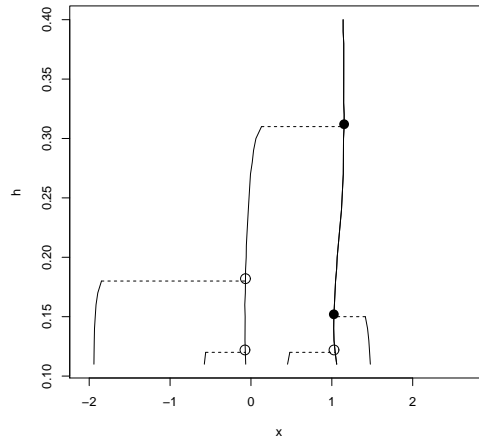


Figura 2.4: Árbol de modas básico empleando núcleo gaussiano con parámetros de ventana h entre 0.11 y 0.40 para la muestra de 600 datos obtenidas a partir de la mixtura de normales M8 que se empleó en la Figura 1.4. Los círculos rellenos son las modas significativas a nivel $\alpha = 0.15$. Los círculos vacíos representan los máximos relativos para los cuales no hay evidencias significativas para rechazar la hipótesis de que v_i no es moda para este nivel.

al. (1998) proponen escoger un ancho de píxel fijo en el gráfico del árbol de modas y contar el número de veces que cae en ese píxel algún rastro de moda, de forma que, entre la gama de grises, se representará con un color gris más oscuro a aquellos píxeles donde caigan un mayor número de rastros de modas. En la Figura 2.5 se representa el bosque de modas para la muestra de 600 datos provenientes de la mixtura M8 que se empleó en la Figura 2.1 y de 99 réplicas del mismo tamaño muestral generadas a partir del sorteo con reemplazamiento de la muestra original, tomando el ancho de píxel proporcional a 0.05 veces la unidad tomada en el eje de las abscisas.

Como mencionan Minnotte *et al.* (1998), esta representación presenta dos grandes problemas: el primero es la dependencia que tiene con el ancho del píxel seleccionado y el segundo es que algunas modas artificiales causadas por atípicos aparecen de color gris muy oscuro, como es el caso del punto -2 en la Figura 2.5, ya que al tratarse de un solo punto aislado este siempre que aparezca, creará una moda artificial en dicho punto para valores bajos del parámetro ventana.

Ante esta problemática, Minnotte *et al.* (1998) proponen centrarse en la segunda derivada y usar esta idea para los puntos donde la segunda derivada es negativa. Los autores sugieren generar varias réplicas bootstrap de la muestra original y representar de color gris más oscuro aquellos píxeles donde la segunda derivada de la estimación tipo núcleo sea negativa con más frecuencia. Sin embargo, el problema de emplear la segunda derivada es el que ya se apreciaba en la Figura 2.2 y es que aunque siempre que hay una moda la segunda derivada es negativa, la otra implicación no siempre es cierta. Por tanto, aunque el estudiar las segundas derivadas sirve para dar una acotación del número de modas, este análisis no va a permitir dar una estimación acerca del verdadero número de modas.

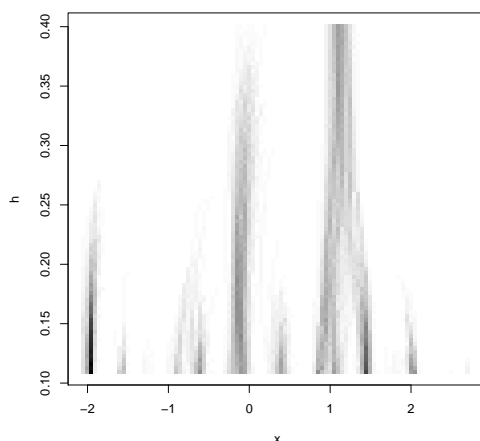


Figura 2.5: Bosque de modas empleando núcleo gaussiano con parámetros de ventana h entre 0.11 y 0.40 proveniente de la muestra de 600 datos obtenidas a partir de la mixtura de normales M8 y de 99 remuestras bootstrap.

2.2. SiZer

El principal problema que presenta el árbol y el bosque de modas es que no son capaces de determinar qué modas han sido formadas por atípicos, es decir, cuales de los máximos relativos de la estimación tipo núcleo han sido creados de forma artificial. Una de las herramientas gráficas que solventa este problema es la propuesta por Chaudhuri y Marron (1999), conocida como el SiZer. Esta herramienta permite detectar qué características de la estimación tipo núcleo están realmente presentes y cuáles son artificios de la muestra desde una perspectiva de espacio–escala, esto es, para los diversos valores que puede tomar la variable aleatoria X y para un rango de parámetros ventana.

Con el fin de ilustrar la idea detrás de esta herramienta gráfica, se muestra en la Figura 2.6 (derecha) el SiZer para la muestra de 600 observaciones extraídas de la mixtura unimodal de normales M8 que se empleó en la Figura 1.4. Con el objetivo de encontrar donde se encuentran las modas y los “valles”, se muestra de color azul las zonas donde la pendiente de la densidad suavizada es significativamente positiva, de color rojo donde es significativamente negativa, el color púrpura se usa para indicar donde la curva suavizada no es significativamente creciente ni decreciente y el gris es usado para indicar donde no hay suficientes datos para determinar el comportamiento de la densidad suavizada. En el eje vertical se muestran los distintos valores de h estudiados. Consecuentemente, fijado un parámetro de ventana, se identificará una moda en la curva suavizada cuando una región de color azul esté seguida por una de color rojo. Así, se puede ver como en este caso, para valores de h mayores que 0.17, el SiZer identifica correctamente una única moda situada entre el 0 y el 1.5 (como se puede ver en la Figura 2.6, izquierda, la moda de la distribución M8 se encuentra en el punto 1.1).

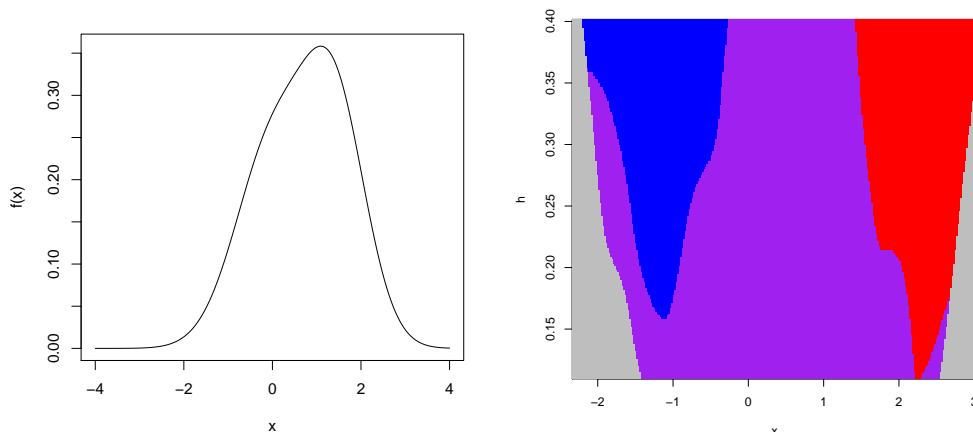


Figura 2.6: Izquierda: densidad de la mixtura de normales M8. Derecha: SiZer con parámetros de ventana h entre 0.11 y 0.40 (se presentan en el eje vertical) proveniente de la muestra de 600 datos obtenidas a partir de la mixtura de normales M8.

La idea detrás del SiZer de Chaudhuri y Marron (1999) es que aunque la curva suavizada $f_h(x) \equiv \mathbb{E}(\hat{f}_h(x))$ no es igual a $f(x)$, sí que se verifica que la estructura de modas y “valles” se conserva para un rango apropiado de valores de h como se puede observar en la Figura 2.7. Ante esta situación Chaudhuri y Marron (1999) proponen hacer inferencia sobre la curva suavizada f_h , en vez de sobre f , para un rango de valores de espacio y de escala.

Como el interés reside en encontrar las modas, Chaudhuri y Marron (1999) proponen crear intervalos de confianza, en espacio y escala, para $f'_h(x) \equiv \mathbb{E}(\hat{f}'_h(x))$, ya que la derivada de la estimación tipo núcleo:

$$\hat{f}'_h(x) = \frac{1}{nh^2} \sum_{i=1}^n K' \left(\frac{x - X_i}{h} \right), \quad (2.2)$$

es un estimador insesgado para esta curva en cada localización x y en cada escala h . Chaudhuri y Marron (1999) fijan como K el núcleo gaussiano dado en (1.2), ya que por el Teorema 3.1 que se verá en la Sección 3.1.1 se tiene que el número de veces que se alcanza el valor cero en (2.2) es una función monótona decreciente en h .

Los límites del intervalo de confianza para $f'_h(x)$ vendrán dados de la siguiente forma:

$$\hat{f}'_h(x) \pm q(\alpha) \cdot \hat{\text{dt}} \left(\hat{f}'_h(x) \right), \quad (2.3)$$

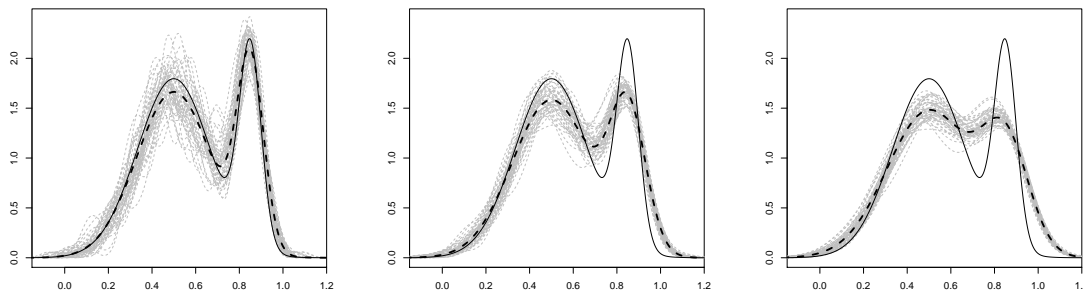


Figura 2.7: Se presenta en trazo sólido y negro la mezcla bimodal de normales M5 (véase Apéndice A). Con trazo negro y discontinuo se muestra $f_h(x)$. Con trazo gris y discontinuo se representan las estimaciones \hat{f}_h obtenidas a partir de 50 muestras de tamaño $n = 200$. Izquierda: $h = 0.035$. Centro: $h = 0.065$. Derecha: $h = 0.095$.

donde $q(\alpha)$ será un cuantil apropiado y $\hat{\text{dt}}(\hat{f}'_h(x))$ será la estimación de la desviación típica de $\hat{f}'_h(x)$.

Para un nivel de significación α dado, se analizará en espacio y escala, los límites de confianza presentados en (2.3). Si ambos límites están por encima de cero entonces se representará de color azul en el mapa SiZer. Análogamente, si ambos son menores que cero entonces se representará de color rojo en el mapa SiZer. Si uno está por encima y otro por debajo, se representará de color púrpura.

Como el estimador de la derivada de f es una ponderación de las derivadas de K evaluadas en un punto, Chaudhuri y Marron (1999) proponen estimar la desviación típica de $f'_h(x)$ como la raíz cuadrada de:

$$\widehat{\text{var}}(\hat{f}'_h(x)) = \frac{1}{nh^4} S^2 \left(K' \left(\frac{x - X_1}{h} \right), \dots, K' \left(\frac{x - X_n}{h} \right) \right),$$

donde S^2 es la varianza muestral de los n valores que toman $K'((x - X_i)/h)$, con $i = 1, \dots, n$.

Para calcular el cuantil $q(\alpha)$, Chaudhuri y Marron (1999) proponen dos aproximaciones basadas en la distribución normal y dos basadas en las técnicas bootstrap.

Entre las aproximaciones gaussianas, el método más simple denotado como $q_1(\alpha)$ consiste en aproximar $q(\alpha)$ de la siguiente forma:

$$q_1(\alpha) = \Phi^{-1} \left(1 - \frac{\alpha}{2} \right),$$

donde $\Phi^{-1}(z)$ la función cuantil de la normal estándar. Tal y como señalan Chaudhuri y Marron

(1999), esta aproximación hace que aparezcan modas formadas por datos atípicos en el mapa SiZer.

La segunda aproximación gaussiana, $q_2(\alpha)$, se basa en el hecho de que si dos valores x e y , con $x \neq y$, están suficientemente “alejados”², las estimaciones $\hat{f}'_h(x)$ y $\hat{f}'_h(y)$ pueden ser consideradas como independientes, mientras que estas están claramente correladas cuando x e y están muy próximas entre sí. Así, la idea sería crear $m(h)$ bloques de datos y haciendo uso de estos bloques, crear intervalos de confianza independientes. Este $m(h)$ será una función que depende del “tamaño efectivo de muestra” (ESS , por sus siglas en inglés), el cual se define del siguiente modo:

$$ESS(x, h) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)}{K(0)}, \quad (2.4)$$

a partir del ESS se escogerá como $m(h)$ al siguiente valor:

$$m(h) = \frac{n}{\overline{ESS}(x, h)}, \quad (2.5)$$

donde $\overline{ESS}(x, h)$ representa la media muestral en x de los valores de $ESS(x, h)$ dados en (2.4).

Asumiendo la independencia de los $m(h)$ bloques de datos, la aproximación del cuantil simultáneo en x , $q_2(\alpha)$, vendría dada por:

$$q_2(\alpha; h) = \Phi^{-1} \left(\frac{1 + (1 - \alpha)^{1/m(h)}}{2} \right).$$

El valor del ESS dado en (2.4) también es de utilidad para conocer en qué regiones la aproximación normal en la que se basa el cálculo de q_1 y q_2 puede ser inadecuada. Así, debido a la deficiente aproximación por la falta de datos, las regiones donde se verifica que $ESS(x, h) < n_0$ se representarán de color gris el mapa SiZer (Chaudhuri y Marron (1999) recomiendan tomar $n_0 = 5$). Además, también se modificará el valor de $m(h)$ para evitar problemas con las cantidades bajas de ESS . De esta forma, para la expresión del $m(h)$ que se daba en (2.5) en vez de calcular la media muestral de los valores de $ESS(x, h)$ en el rango de valores de espacio escogidos para realizar el mapa SiZer, se realizará únicamente en los $x \in D_h$, donde el conjunto D_h es:

$$D_h = \{x : ESS(x, h) \geq n_0\}.$$

Las otras dos aproximaciones de $q(\alpha)$ que proponen Chaudhuri y Marron (1999) están basadas en la metodología bootstrap y tienen por objetivo eliminar la aproximación gaussiana del cálculo de los cuantiles $q(\alpha)$. Para ello se generan B réplicas bootstrap, \mathcal{X}^{*b} , obtenidas a partir de la distribución empírica. Para cada réplica bootstrap, se calcularía la estimación tipo núcleo de la derivada de f dada en (2.2), $\hat{f}'_h(x)^{*b}$, y su versión estandarizada:

²Donde el concepto de “proximidad” va a depender de la escala h .

$$Z(x, h)^{*b} = \frac{\hat{f}'_h(x)^{*b} - \hat{f}'_h(x)}{dt(\hat{f}'_h(x))}, \text{ con } b = 1, \dots, B.$$

A partir de estos $Z(x, h)^{*b}$ el tercer método que proponen Chaudhuri y Marron (1999) para aproximar el cuantil $q(\alpha)$, también simultáneo en x , es el de tomar $q_3(\alpha; h)$ como el cuantil empírico $(1 - \alpha/2)$ de los $\max_{x \in D_h} |Z(x, h)^*|$ calculados sobre las réplicas bootstrap.

Cabe mencionar que aunque el objetivo de Chaudhuri y Marron (1999) es el de crear intervalos de confianza en espacio y escala para $f'_h(x)$, las aproximaciones dadas hasta el momento crean intervalos de confianza que son aproximadamente simultáneos sobre x , pero no sobre h . Con el fin de solucionar este problema, la cuarta aproximación que proponen Chaudhuri y Marron (1999) es la de tomar el cuantil $q_4(\alpha)$ como el cuantil empírico $(1 - \alpha/2)$ de los $\max_h \max_{x \in D_h} |Z(x, h)^*|$ calculados sobre las réplicas bootstrap. De esta forma, los cuantiles $q_4(\alpha)$ serán simultáneos sobre x y sobre h .

En la librería `feature` desarrollada por Duong y Wand (2011) está programado una adaptación del método q_1 teniendo en cuenta las localizaciones donde $ESS(x, h) < 5$. Se ha empleado esta librería para obtener el mapa SiZer que se mostraba en la Figura 2.6. Para este ejemplo, se ha utilizado la aproximación q_1 y se han representado los valores para los que se tiene que $ESS < 5$ (de color gris) para la muestra de 600 datos procedentes de la distribución M8 (véase Apéndice A) que se empleó en la Figura 1.4, tomando para ello los parámetros de ventana entre 0.11 y 0.4 y un nivel de significación $\alpha = 0.05$.

2.3. Herramientas gráficas para datos circulares

Como se mencionó en la Introducción, el objetivo final del presente TFM es el de estimar el número de modas cuando se tiene la fecha en la que se produjeron los incendios en la comarca de Vigo. Con el fin de estimar el número de modas cuando se tienen datos circulares, se analizará en profundidad la herramienta exploratoria presente en la literatura para la detección de modas en el contexto circular, conocida como CircSiZer y desarrollada por Oliveira *et al.* (2012b).

2.3.1. CircSiZer

El CircSiZer de Oliveira *et al.* (2012b) es la adaptación al contexto circular del SiZer de Chaudhuri y Marron (1999). En el CircSiZer, para un rango de valores θ situados en el intervalo $[0, 2\pi)$ y de ν , se estudiará el comportamiento de la derivada de la curva suavizada, $f'_\nu^c(\theta) \equiv \mathbb{E}(\hat{f}'_\nu^c(\theta))$, para el cual la derivada de la estimación circular tipo núcleo dada en (1.5), $\hat{f}'_\nu^c(\theta)$, es un estimador insesgado para cada θ y ν . Para este caso, Oliveira *et al.* (2012b) propone tomar como K_ν^c el núcleo circular von Mises dado en (1.6), de manera que, para valores de $\theta \in [0, 2\pi)$, $\hat{f}'_\nu^c(\theta)$ tiene la siguiente forma:

$$\hat{f}_\nu^{c'}(\theta) = \frac{-\nu \text{sen}(\theta - \Theta_i)}{n2\pi I_0(\nu)} \sum_{i=1}^n \exp(\nu \cos(\theta - \Theta_i)), \text{ con } 0 \leq \theta < 2\pi. \quad (2.6)$$

Los límites del intervalo de confianza para $f_\nu^{c'}(\theta)$ vendrán dados de la siguiente forma:

$$\hat{f}_\nu^{c'}(\theta) \pm q^c(\alpha) \cdot \widehat{\text{dt}}\left(\hat{f}_\nu^{c'}(\theta)\right), \quad (2.7)$$

donde $q^c(\alpha)$ serán cuantiles apropiados y $\widehat{\text{dt}}\left(\hat{f}_\nu^{c'}(\theta)\right)$ será la estimación de la desviación típica de $\hat{f}_\nu^{c'}(\theta)$.

El código de colores en el mapa CircSiZer será el mismo que el empleado en el mapa SiZer, de forma que cuando ambos límites de confianza estén por encima de cero para un θ y ν dados se representará ese punto en color rojo en el mapa del CircSiZer, cuando ambos estén por debajo de cero se representará de color azul y en otro caso de color púrpura.

Siguiendo las ideas de Chaudhuri y Marron (1999), Oliveira *et al.* (2012b) propone estimar la desviación típica de $\hat{f}_\nu^{c'}(\theta)$ como la raíz cuadrada de:

$$\widehat{\text{var}}\left(\hat{f}_\nu^{c'}(\theta)\right) = \frac{1}{n} S^2\left(K_\nu^{c'}(\theta - \Theta_1), \dots, K_\nu^{c'}(\theta - \Theta_n)\right),$$

donde $K_\nu^{c'}(\theta - \Theta_i)$ es la derivada del núcleo circular von Mises dado en (1.6) en el punto $(\theta - \Theta_i)$, con $i = 1, \dots, n$.

Para calcular los cuantiles $q^c(\alpha)$ de (2.7), Oliveira (2013)³ propone cuatro métodos, dos basados en aproximaciones gaussianas y dos basados en la metodología bootstrap.

El primer método, de forma análoga a la propuesta por Chaudhuri y Marron (1999), consistiría en crear un intervalo de confianza usando una aproximación normal, de forma que, denotando a este método como q_1^c , se tendría que $q_1^c(\alpha) = \Phi^{-1}(1 - \alpha/2)$.

El segundo método consiste en tomar los cuantiles q_2^c de la siguiente forma:

$$q_2^c(\alpha; \nu) = \Phi^{-1}\left(\frac{1 + (1 - \alpha)^{1/m^c(\nu)}}{2}\right),$$

donde el número de bloques de datos, $m^c(\nu)$, será:

$$m^c(\nu) = \frac{n}{\overline{ESS^c}(\theta, \nu)},$$

siendo $\overline{ESS^c}(\theta, \nu)$ la media muestral de los $ESS^c(\theta, \nu)$ en el conjunto $D_\nu^c = \{\theta : ESS^c(\theta, \nu) \geq n_0\}$. El tamaño de muestra efectivo para el caso circular, ESS^c , se define como:

³Se toma esta referencia, ya que en Oliveira *et al.* (2012b) solo aparece la tercera aproximación presentada.

$$ESS^c(\theta, \nu) = \frac{\sum_{i=1}^n K_\nu^c(\theta - \Theta_i)}{K_\nu^c(0)}. \quad (2.8)$$

De nuevo, para el caso circular, Oliveira (2013) toma como $n_0 = 5$ y representan de color gris los valores de (θ, ν) donde $ESS^c(\theta, \nu) < n_0$.

Con el fin de eliminar la aproximación gaussiana de los cuantiles, Oliveira (2013) propone dos aproximaciones bootstrap. La primera, esta relacionada con q_1^c en el sentido de que se crean intervalos de confianza puntuales. Para obtener esta aproximación se generan B nuevas remuestras, Θ^{*b} , de tamaño n obtenidas a partir de la selección al azar con reemplazamiento de los valores de la muestra original $\Theta = (\Theta_1, \dots, \Theta_n)$. Para cada réplica bootstrap, se calcula el valor de $Z^c(\theta, \nu)^{*b}$, a partir de la estimación circular tipo núcleo de la derivada de f^c dada en (2.6), $\hat{f}_\nu^{c'}(\theta)^{*b}$, de la siguiente forma:

$$Z^c(\theta, \nu)^{*b} = \frac{\hat{f}_\nu^{c'}(\theta)^{*b} - \hat{f}_\nu^{c'}(\theta)}{\hat{\text{dt}}(\hat{f}_\nu^{c'}(\theta)^{*b})}, \text{ con } b = 1, \dots, B.$$

donde, en este caso, a diferencia de lo que hacen Chaudhuri y Marron (1999), para calcular los $Z^c(\theta, \nu)^{*b}$ se toma como denominador la estimación de la desviación típica de $\hat{f}_\nu^{c'}(\theta)^{*b}$.

Para aproximar los cuantiles, a diferencia de los métodos q_1^c y q_2^c donde se tomaban los límites de confianza dados en (2.7), para la tercera aproximación, Oliveira (2013) propone tomar el siguiente intervalo de confianza para $f_\nu^{c'}(\theta)$:

$$\left(\hat{f}_\nu^{c'}(\theta) - q_{3,-}^c(\alpha; \theta, \nu) \cdot \hat{\text{dt}}\left(\hat{f}_\nu^{c'}(\theta)\right), \hat{f}_\nu^{c'}(\theta) - q_{3,+}^c(\alpha; \theta, \nu) \cdot \hat{\text{dt}}\left(\hat{f}_\nu^{c'}(\theta)\right) \right),$$

donde $q_{3,-}^c(\alpha; \theta, \nu)$ es el cuantil muestral $(1 - \alpha/2)$ de $Z^c(\theta, \nu)^*$ y $q_{3,+}^c(\alpha; \theta, \nu)$ es el cuantil muestral $(\alpha/2)$ de $Z^c(\theta, \nu)^*$.

El segundo método bootstrap, al igual que q_2^c , proporciona un intervalo de confianza simultáneo sobre θ y para ello, se usarán los valores:

$$\begin{aligned} Z_{\text{inf}}^c(\nu)^{*b} &= \inf_{\theta \in D_\nu^{c*}} Z^c(\theta, \nu)^{*b}, \\ Z_{\text{sup}}^c(\nu)^{*b} &= \sup_{\theta \in D_\nu^{c*}} Z^c(\theta, \nu)^{*b}, \end{aligned}$$

donde D_ν^{c*} es la versión bootstrap de D_ν^c ⁴, esto es, $D_\nu^{c*} = \{\theta : ESS^c(\theta, \nu)^* \geq n_0\}$, con $ESS^c(\theta, \nu)^*$ tomando el siguiente valor:

⁴En este TFM, a la hora de calcular los valores de $Z_{\text{inf}}^c(\nu)^{*b}$ y los $Z_{\text{sup}}^c(\nu)^{*b}$, para ser consistente con que los ínfimos y los supremos se calculan sobre cada una de las réplicas bootstrap, se propone hallar estos valores en los $\theta \in D_\nu^{c*}$, a diferencia de la propuesta de Oliveira (2013) donde se calculan sobre los $\theta \in D_\nu^c$.

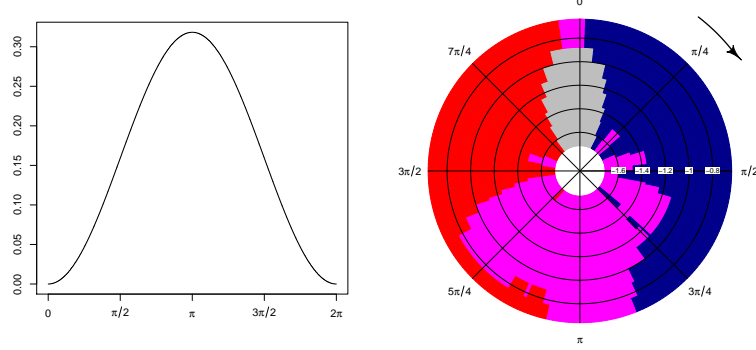


Figura 2.8: Izquierda: densidad de la $C(\pi, 0.5)$. Derecha: CircSiZer empleando la aproximación q_3^c con parámetros de concentración ν entre 45 y 5 (esto es, $-\log_{10}(\nu)$ entre -1.65 y -0.70) proveniente de la muestra de 600 datos obtenidas a partir de la distribución $C(\pi, 0.5)$. En azul: se representa los valores de (θ, ν) para los cuales la derivada de $f_\nu^c(\theta)$ es significativamente positiva para un nivel $\alpha = 0.05$. En rojo: los valores para los cuales la derivada es significativamente negativa para un nivel $\alpha = 0.05$. En púrpura: los valores para los cuales no es significativamente distinta de cero para un nivel $\alpha = 0.05$. En gris: los valores donde $ESS^c(\theta, \nu) < 5$.

$$ESS^c(\theta, \nu)^* = \frac{\sum_{i=1}^n K_\nu^c(\theta - \Theta_i^*)}{K_\nu^c(0)}$$

En este caso, el intervalo de confianza vendría dado de la siguiente forma:

$$\left(\hat{f}_\nu^c(\theta) - q_{4,-}^c(\alpha; \nu) \cdot \hat{dt} \left(\hat{f}_\nu^c(\theta) \right), \hat{f}_\nu^c(\theta) - q_{4,+}^c(\alpha; \nu) \cdot \hat{dt} \left(\hat{f}_\nu^c(\theta) \right) \right),$$

donde $q_{4,-}^c(\alpha; \nu)$ es el cuantil muestral $(1 - \alpha/2)$ de $Z_{\text{sup}}^c(\nu)^*$ y $q_{4,+}^c(\alpha; \nu)$ es el cuantil muestral $(\alpha/2)$ de $Z_{\text{inf}}^c(\nu)^*$.

El CircSiZer usando el método q_3^c ha sido programado en la librería NPCirc por Oliveira *et al.* (2013). En la Figura 2.8 (derecha) se muestra el mapa CircSiZer haciendo uso de la aproximación q_3^c para la muestra de 600 observaciones procedente de la distribución $C(\pi, 0.5)$ que se empleó en la Figura 1.14. Se han tomado parámetros de concentración ν entre 5 y 45 y un nivel de significación $\alpha = 0.05$.

En este caso la representación del CircSiZer es circular y los ángulos θ entre $[0, 2\pi)$ se deben leer en el sentido que marca la flecha (en este ejemplo, el de las agujas del reloj), mientras que con el objetivo de representar, como en el caso del SiZer, la estimación más infrasuavizada próxima al 0, se muestran los valores de $-\log_{10}(\nu)$ en el radio asociado a $\theta = \pi/2$, los cuales van aumentando a medida que crece el radio de la circunferencia. Así, para el ejemplo mostrado en la Figura 2.8, se

tendría que para los valores de ν estudiados solo hay una moda y esta se encuentra entorno al valor π (se puede ver en la Figura 2.8, izquierda, que la única moda se sitúa en el valor π).

Capítulo 3

Test de multimodalidad para datos lineales

En este capítulo, se revisarán los distintos test que permitan analizar en el contexto de variables escalares si se puede rechazar la hipótesis de que la verdadera distribución asociada a una muestra aleatoria simple es unimodal, pudiendo adaptar, en algunos casos, estos test al contraste de que la verdadera distribución posee a lo sumo k modas. Además, se aportará una nueva propuesta con un doble objetivo, el de mejorar los test presentes en la literatura para variables escalares y el de proporcionar un test que permita ser adaptado al contexto circular.

En las dos primeras secciones de este capítulo se estudiarán dos grandes bloques de test, los basados en la ventana crítica y los basados en el *dip* o en el exceso de masa. Todos estos test han sido programados en el software estadístico R Core Team (2013), a excepción del test de Hartigan y Hartigan (1985), programado por Maechler (2013) en la librería `dip.test`. En la tercera y última sección de este capítulo se realizará un estudio de simulación que permita comparar los distintos test analizados.

3.1. Test basados en la ventana crítica

En esta sección, se analizarán los test basados en la ventana crítica, que se fundamentan en encontrar la ventana h_k más pequeña de forma que la estimación \hat{f}_{h_k} para ese parámetro de ventana tenga k modas. Silverman (1981) propone definir la ventana crítica, denotada como h_k , del siguiente modo:

$$h_k = \min\{h : \hat{f}_h \text{ tiene a lo sumo } k \text{ modas}\}. \quad (3.1)$$

Entre los test basados en la ventana crítica, haciendo simplemente uso de h_k , se analizará el test

de Silverman (1981) para el contraste (1.3), cuando $k \in \mathbb{Z}^+$. Hall y York (2001) muestran que el test de Silverman (1981) no está correctamente calibrado y proponen un método para resolver este problema para el caso en que $k = 1$, trabajando con un soporte acotado y conocido. Por último, se revisará el test de Fisher y Marron (2001) donde se utilizará la ventana crítica para poder obtener la función de distribución suavizada con la que realizar un test del tipo Cramér–von Mises. Para calibrar este test, Fisher y Marron (2001) proponen un mecanismo de remuestreo que se basa en eliminar todas las modas artificiales que se forman en la estimación de la función de densidad cuando se toma como ventana h_k .

3.1.1. Test basado en la ventana crítica de Silverman

La idea que propone Silverman (1981) es la de considerar como estadístico de contraste para (1.3) cuando $k \in \mathbb{Z}^+$ la ventana crítica definida en (3.1). Para la obtención de la ventana crítica es de utilidad el siguiente resultado:

Teorema 3.1. (Silverman, 1981) *El número de modas de \hat{f}_h , cuando se emplea el núcleo gaussiano dado en (1.2), es una función monótona decreciente de h . Consecuentemente, dada la ventana crítica definida en (3.1), \hat{f}_h tiene más de k modas si y solo si $h < h_k$.*

Este teorema muestra que h_k puede ser obtenida a través de un algoritmo de búsqueda dicotómica. El test de Silverman (1981) usará como estadístico a la ventana crítica definida en (3.1) para contrastar la hipótesis nula de que la verdadera distribución posee a lo sumo k modas frente a la hipótesis alternativa de que la verdadera distribución posee más de k modas, rechazándose esta hipótesis nula para valores “altos” de h_k .

El problema que surge con el uso de este estadístico es que se desconoce su distribución bajo la hipótesis nula. Con el objetivo de saber cuándo rechazar la hipótesis nula, como se tiene un algoritmo computacionalmente rápido para la búsqueda de la ventana crítica, Silverman (1981) propone recurrir a la metodología de remuestreo bootstrap para aproximar la distribución de h_k bajo la hipótesis nula.

Para ello, se genera la remuestra bootstrap $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$ a partir de la función de densidad \hat{f}_{h_k} , que se obtiene como el estimador tipo núcleo introducido en (1.1) tomando como ventana h_k . A partir de las B réplicas bootstrap, \mathcal{X}^{*b} con $b = 1, \dots, B$, calculando sus valores h_k^{*b} asociados se puede aproximar la distribución de la ventana crítica bajo H_0 . Para un nivel de significación α , se rechazará H_0 si $\mathbb{P}(h_k^* \leq h_k | \mathcal{X}) \geq 1 - \alpha$.

Los resultados de las simulaciones que se presentarán en la Sección 3.3, justificarán la necesidad de un nuevo calibrado, como el propuesto por Hall y York (2001) que se detallará a continuación.

3.1.2. Calibrado de la ventana crítica de Silverman

Hall y York (2001) demuestran que el test de Silverman (1981) no está correctamente calibrado. Para un tamaño muestral grande (cuando $n \rightarrow \infty$), bajo la hipótesis nula, debiera ocurrir que el

p-valor $U_n = \mathbb{P}(h_k^* \leq h_k | \mathcal{X})$ se aproximase a una uniforme en el intervalo $(0, 1)$, y esto no es cierto para el caso del test de Silverman (1981). Este hecho justificaría que los porcentajes de rechazo empíricos y los niveles de significación teóricos fuesen considerablemente distintos, como ocurre en los ejemplos de simulación que se analizan en la Sección 3.3. El problema, como muestran Hall y York (2001), es que el calibrado del test depende de la verdadera distribución de donde proceda la muestra, que es desconocida. Para el caso $k = 1$, acotando a priori el soporte de la variable aleatoria, Hall y York (2001) prueban que se puede calibrar el test de Silverman (1981) independientemente de la distribución que tenga la variable de interés. Aún así, Hall y York (2001) demuestran que será necesario corregir el test original ya que, como se verá, los valores de h_k^* son más pequeños que los de h_k bajo H_0 .

Como se ha mencionado, para poder calibrar el test de Silverman (1981), se necesitará acotar el soporte de la variable aleatoria, lo cual conllevará a modificar el contraste de hipótesis que se estaba planteando hasta el momento. De esta forma, dada la muestra $\mathcal{X} = (X_1, \dots, X_n)$, si j es el verdadero número de modas en un intervalo cerrado I , se definirá el contraste de hipótesis como sigue:

$$H_0 : j = 1 \text{ y no tiene mínimos locales en el intervalo } I. \quad (3.2)$$

Esto también obligará a redefinir la expresión de la ventana crítica de la siguiente forma:

$$h_{\text{HY}} = \text{mín}\{h : \hat{f}_h \text{ tiene exactamente una moda en } I\}. \quad (3.3)$$

Esta restricción en el soporte de la variable viene justificada por el hecho de que en caso de no realizar dicha acotación, las propiedades de h_1 (siendo h_1 la ventana crítica cuando se toma $k = 1$) vendrían dadas por los valores atípicos de la muestra, puesto que son estos los que crean habitualmente modas artificiales en las colas de la estimación de la función de densidad.

El problema asociado a esta nueva ventana crítica es que cuando se usa la expresión de h_{HY} dada en (3.3), en vez de la expresión h_1 que se daba en (3.1), no se tiene porqué verificar necesariamente la tesis del Teorema 3.1. Este problema lo solucionan Hall y York (2001) demostrando que la probabilidad de que el número de modas de \hat{f}_h sea una función monótona en h converge a uno cuando se usa el núcleo gaussiano para la estimación \hat{f}_h . Este resultado permite seguir empleando un algoritmo dicotómico para la obtención de la ventana crítica h_{HY} .

Hall y York (2001) demuestran que dado un nivel de significación α , para un valor apropiado de λ_α (independiente de la verdadera función de densidad unimodal f), se rechazará, con un nivel de significación α , la hipótesis nula H_0 definida en (3.2), si se verifica que $\mathbb{P}(h_{\text{HY}}^* \leq \lambda_\alpha h_{\text{HY}} | \mathcal{X}) \geq 1 - \alpha$. En esta expresión h_{HY}^* denota la versión bootstrap de h_{HY} , esto es:

$$h_{\text{HY}}^* = \text{mín}\{h : \hat{f}_h^* \text{ tiene exactamente una moda en } I\},$$

donde para obtener la función \hat{f}_h^* , cada réplica $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$ es generada a partir de \hat{f}_{HY} .

Para poder determinar los valores de λ_α , Hall y York (2001) proponen usar dos métodos. El primero es el de utilizar la corrección asintótica basada en la distribución límite, con la que se

α	τ_α	α	τ_α	α	τ_α	α	τ_α
0.01	0	0.11	0.039	0.21	0.119	0.35	0.261
0.02	0.001	0.12	0.047	0.22	0.127	0.40	0.313
0.03	0.003	0.13	0.054	0.23	0.136	0.45	0.364
0.04	0.007	0.14	0.061	0.24	0.146	0.50	0.424
0.05	0.011	0.15	0.067	0.25	0.155	0.55	0.483
0.06	0.015	0.16	0.077	0.26	0.164	0.60	0.543
0.07	0.019	0.17	0.084	0.27	0.174	0.65	0.605
0.08	0.022	0.18	0.092	0.28	0.185	0.75	0.721
0.09	0.028	0.19	0.100	0.29	0.198	0.85	0.838
0.10	0.033	0.20	0.110	0.30	0.208	0.95	0.945

Tabla 3.1: Aproximación de los valores de τ_α para distintos niveles de significación α . Obtenidos tras realizar 1000 réplicas en cada una de las 10000 muestras de tamaño $n = 10000$ procedentes de una distribución $N(0, 1)$ empleando el intervalo $I = [-1.5, 1.5]$.

obtiene que el valor que toma λ_α es:

$$\lambda_\alpha = \frac{0.94029\alpha^3 - 1.59914\alpha^2 + 0.17695\alpha + 0.48971}{\alpha^3 - 1.77793\alpha^2 + 0.36162\alpha + 0.42423}. \quad (3.4)$$

El segundo método vendría dado por la estimación de λ_α usando técnicas de Monte Carlo, ya que la distribución de este parámetro no depende, asintóticamente, de la verdadera función de densidad de la variable aleatoria. Para poder aplicar este método, será necesario encontrar el τ_α , de forma que el porcentaje de rechazos de H_0 sea α cuando se toman muestras de tamaño suficientemente grandes procedentes de una distribución unimodal, para la regla que rechaza H_0 si se verifica que $\mathbb{P}(h_{\text{HY}}^* \leq h_{\text{HY}} | \mathcal{X}) \geq 1 - \tau_\alpha$. Así, se han obtenido los valores de τ_α , utilizando la metodología de remuestreo bootstrap descrita en la Sección 3.1.1, simulando 10000 muestras de tamaño $n = 10000$ (realizando en cada una $B=1000$ réplicas) generadas a partir de la distribución $N(0, 1)$ y tomando el intervalo $I = [-1.5, 1.5]$. En la Tabla 3.1 se muestran los valores de τ_α para distintos niveles de significación α tomando valores entre 0.01 y 0.95. De esta forma, para un nivel de significación $\alpha = 0.05$, se rechazaría H_0 si se verifica que $\mathbb{P}(h_{\text{HY}}^* \leq h_{\text{HY}} | \mathcal{X}) \geq 1 - \tau_{0.05} = 1 - 0.0109$.

Observando ambos métodos queda claro que es necesario corregir el test original, ya que tomar $\lambda_\alpha = 1$ como propone Silverman (1981) daría lugar a un test claramente conservador.

3.1.3. Test basado en el estadístico de Cramér–von Mises

El problema del test de Hall y York (2001) es que no se puede aplicar el contraste (1.3) cuando $k > 1$. Si se toma la ventana crítica para este contraste en el intervalo I , esto es:

$$h_{\text{HY},k} = \min\{h : \hat{f}_h \text{ tiene a los sumo } k \text{ modas en } I\},$$

se tiene que la distribución asintótica de $h_{\text{HY},k}$ depende del valor que tomen los $(2k - 1)$ valores de $c_j = f(t_j)^{1/5} / |f''(t_j)|^{2/5}$, siendo los t_j los extremos relativos de f en I para $j = 1, \dots, 2k - 1$ (véase Hall y York, 2001).

Fisher y Marron (2001) proponen extender este contraste al caso en que H_0 establece que la verdadera distribución posee a lo sumo k modas, con $k \in \mathbb{Z}^+$. Para ello, en vez de usar como estadístico de contraste la ventana crítica h_k , se propone usar la estimación tipo núcleo de la función de distribución, $\hat{F}_{h_k}(x) = \int_{-\infty}^x \hat{f}_{h_k}(t) dt$, para construir un estadístico del tipo Cramér–von Mises.

Dada una muestra $\mathcal{X} = (X_1, \dots, X_n)$, el criterio de Cramér–von Mises permite realizar el siguiente contraste de hipótesis:

$$H_0 : F(x) = F_0(x) \forall x \in \mathbb{R}, \text{ frente a } H_1 : F(x) \neq F_0(x) \text{ para algún } x, \quad (3.5)$$

donde F_0 es una función de distribución dada. El estadístico de Cramér–von Mises para el contraste (3.5) viene dado por la siguiente expresión:

$$\begin{aligned} T &= n \int_{-\infty}^{\infty} [F_n(x) - F_0(x)]^2 dF_0(x) = \\ &= n \underbrace{\int_{-\infty}^{\infty} F_n^2(x) dF_0(x)}_{(a)} - 2n \underbrace{\int_{-\infty}^{\infty} F_n(x) F_0(x) dF_0(x)}_{(b)} + n \underbrace{\int_{-\infty}^{\infty} F_0^2(x) dF_0(x)}_{(c)}, \end{aligned} \quad (3.6)$$

donde F_n es la función de distribución empírica de la muestra que se definió en (1.4).

Si F_0 es continua y realizando el cambio de variable $u = F_0(x)$ en el sumando (c) de (3.6), se puede ver que:

$$(c) = n \int_0^1 u^2(x) du = \frac{n}{3}.$$

Si $X_{(1)} < \dots < X_{(n)}$ denota a la muestra ordenada se tiene que:

$$\begin{aligned} (a) &= n \left(\sum_{i=1}^{n-1} \frac{i^2}{n^2} \int_{X_{(i)}}^{X_{(i+1)}} dF_0(x) + \int_{X_{(n)}}^{\infty} dF_0(x) \right) = \\ &= \sum_{i=1}^{n-1} \frac{i^2}{n} (F_0(X_{(i+1)}) - F_0(X_{(i)})) + n (1 - F_0(X_{(n)})) = \\ &= -\frac{1}{n} F_0(X_{(1)}) + \sum_{i=2}^n \frac{(i-1)^2 - i^2}{n} F_0(X_{(i)}) + n = n - \sum_{i=1}^n \frac{2i-1}{n} F_0(X_{(i)}). \end{aligned}$$

Haciendo nuevamente el cambio $u = F_0(x)$, el sumando (b) de la expresión dada en (3.6) será:

$$\begin{aligned}
(b) &= -2n \left(\sum_{i=1}^{n-1} \frac{i}{n} \int_{X_{(i)}}^{X_{(i+1)}} F_0(x) dF_0(x) + \int_{X_{(n)}}^{\infty} F_0(x) dF_0(x) \right) = \\
&= -2 \left(\sum_{i=1}^{n-1} i \left(\frac{F_0^2(X_{(i+1)})}{2} - \frac{F_0^2(X_{(i)})}{2} \right) + n \left(\frac{1}{2} - \frac{F_0^2(X_{(n)})}{2} \right) \right) = \\
&= \sum_{i=1}^n F_0^2(X_{(i)}) - n
\end{aligned}$$

Además, por otro lado, se puede ver que:

$$\sum_{i=1}^n \left(\frac{2i-1}{2n} \right)^2 = \frac{n}{3} - \frac{1}{12n}. \quad (3.7)$$

Haciendo uso de la igualdad dada en (3.7) sobre la expresión de T dada en (3.6) se puede concluir lo siguiente:

$$\begin{aligned}
T &= (a) + (b) + (c) - \frac{n}{3} + \frac{1}{12n} + \frac{n}{3} - \frac{1}{12n} \\
&= n - \sum_{i=1}^n \frac{2i-1}{n} F_0(X_{(i)}) + \sum_{i=1}^n F_0^2(X_{(i)}) - n + \frac{n}{3} - \frac{n}{3} + \frac{1}{12n} + \frac{n}{3} - \frac{1}{12n} = \\
&= - \sum_{i=1}^n \frac{2i-1}{n} F_0(X_{(i)}) + \sum_{i=1}^n F_0^2(X_{(i)}) + \frac{1}{12n} + \sum_{i=1}^n \left(\frac{2i-1}{2n} \right)^2 = \\
&= \sum_{i=1}^n \left(F_0(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}.
\end{aligned}$$

Tomando como $F_0 = \hat{F}_{h_k}$, se tendría que la expresión del estadístico de Cramér-von Mises para este caso viene dado de la siguiente forma:

$$T_k = \sum_{i=1}^n \left(\hat{F}_{h_k}(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}. \quad (3.8)$$

Para conocer la distribución del estadístico T_k dado en (3.8), bajo la hipótesis nula de que la verdadera distribución posee a lo sumo k modas, Fisher y Marron (2001) proponen usar nuevamente la metodología bootstrap. Para ello, se generan B réplicas bootstrap, $\mathcal{X}^{*b} = (X_{(1)}^{*b}, \dots, X_{(n)}^{*b})$, con $b = 1, \dots, B$, a partir de la función de densidad \hat{f}_{h_k} . Usando estas réplicas, se construye la versión bootstrap del estadístico de Cramér-von Mises T_k^{*b} , sin más que sustituir a $(X_{(1)}, \dots, X_{(n)})$ por

$(X_{(1)}^{*b}, \dots, X_{(n)}^{*b})$ en la expresión (3.8). Para un nivel de significación α , se rechazará H_0 si $\mathbb{P}(T_k^{*b} \leq T_k | \mathcal{X}) \geq 1 - \alpha$.

En la Sección 3.3 se muestra el deficiente comportamiento de este test. Fisher y Marron (2001) indican que un mal calibrado vendría justificado por dos motivos, uno es por la aparición de modas artificiales en la estimación de la función de densidad, asociadas a valores atípicos, y el otro sería que la ventana crítica de Silverman (1981) no tiene en cuenta la cantidad de masa de probabilidad asociada a cada moda.

Para eliminar las modas artificiales asociadas a datos atípicos, Fisher y Marron (2001) proponen no tener en cuenta a aquellas modas cuya función de densidad asociada no supere una altura mínima, que se denota como λ_0 . Para solucionar el problema de que la ventana crítica no tiene en cuenta la cantidad de masa de probabilidad asociada a cada moda, Fisher y Marron (2001) plantean que para determinar si un máximo relativo de la estimación de la función de densidad es moda o no, este debe tener una cantidad mínima de masa de probabilidad asociada, m_0 , por encima de una altura mínima.

Con el fin de obtener la expresión de la ventana crítica dada por Fisher y Marron (2001), a partir de la estimación tipo núcleo de la función de densidad dada en (1.1) para una ventana h , al igual que en el árbol de modas, se denotará como v_i a la localización de los máximos locales (en orden creciente en el soporte de la variable aleatoria X). Si la estimación \hat{f}_h posee l máximos locales, se denotará como w_i a la localización (en orden creciente) de los mínimos locales con $i = 2, \dots, l$, con $w_1 = -\infty$ y como $w_{l+1} = \infty$.

Para no rechazar la hipótesis nula a causa de las modas causadas por datos atípicos, se considerarán solo las modas que superen una altura mínima λ_0 , esto es, se tendrán en cuenta solo las modas asociadas a este conjunto:

$$I_{\text{FM}} = \{i = 1, \dots, l : \hat{f}_h(v_i) > \lambda_0\}.$$

Para cada $i \in I_{\text{FM}}$, se define el nivel de masa asociada a un máximo relativo como:

$$\lambda_i = \max \left\{ \lim_{t_1 \rightarrow w_i} \hat{f}_h(t_1), \lim_{t_2 \rightarrow w_{i+1}} \hat{f}_h(t_2), \lambda_0 \right\}.$$

Por último, para poder analizar el exceso de masa asociado a cada máximo relativo, antes es necesario ver cuales serán los puntos de corte asociados a estos excesos para cada v_i con $i \in I_{\text{FM}}$. Estos son:

$$a_i = \begin{cases} w_i & \text{si } \lambda_i = \lim_{t_1 \rightarrow w_i} \hat{f}_h(t_1), \\ \hat{f}_h^{-1}(w_i, v_i)(\lambda_i) & \text{en otro caso,} \end{cases} \quad b_i = \begin{cases} w_{i+1} & \text{si } \lambda_i = \lim_{t_2 \rightarrow w_{i+1}} \hat{f}_h(t_2), \\ \hat{f}_h^{-1}(v_i, w_{i+1})(\lambda_i) & \text{en otro caso,} \end{cases}$$

donde $\hat{f}_h^{-1}(w_i, v_i)(\lambda_i)$ representa al primer punto mayor que w_i tal que \hat{f}_h para ese punto toma el valor λ_i .

Se puede ver que los $a_i \leq v_i \leq b_i$ están bien definidos para todo $i \in I_{\text{FM}}$. Si $\lambda_i \neq \lim_{t_1 \rightarrow w_i} \hat{f}_h(t_1)$ entonces, por definición, se cumple que $\lambda_i > \lim_{t_1 \rightarrow w_i} \hat{f}_h(t_1)$. Además, se verifica que $\hat{f}_h(v_i) > \lambda_0$ por ser $i \in I_{\text{FM}}$ y también se cumple que $\hat{f}_h(v_i) > \lim_{t_2 \rightarrow w_{i+1}} \hat{f}_h(t_2)$ por ser w_{i+1} el primer mínimo relativo después de v_i si $i = 1, \dots, l-1$ o por ser $\hat{f}_h(v_i) > 0 = \lim_{t_2 \rightarrow \infty} \hat{f}_h(t_2)$ si $i = l$. Si \hat{f}_h es continua y se cumple que $\lim_{t_1 \rightarrow w_i} \hat{f}_h(t_1) < \lambda_i < \hat{f}_h(v_i)$ existirá al menos un punto z_i en el intervalo (w_i, v_i) verificando que $\hat{f}_h(z_i) = \lambda_i$. Se puede razonar de forma análoga para ver que los b_i están bien definidos.

Una vez introducida la notación, ya se puede calcular el exceso de masa asociado a cada máximo local v_i , con $i \in I_{\text{FM}}$, que se denotará como $E_i(h)$ y que se define como:

$$E_i(h) = \int_{a_i}^{b_i} (\hat{f}_h(x) - \lambda_i) dx. \quad (3.9)$$

Dado un exceso de masa mínimo, m_0 , para eliminar las modas asociadas a un exceso de masa pequeño, se considerarán solo a aquellos máximos relativos tales que $E_i(h) > m_0$. Para los $i \in I_{\text{FM}}$ verificando que $E_i(h) \leq m_0$, se realizará el siguiente algoritmo en el que en cada paso, se considera el índice m asociado al $E_m(h)$ más pequeño. Si el correspondiente λ_m toma el valor λ_0 , entonces se eliminará al índice m del conjunto I_{FM} . En cambio, cuando $\lambda_m > \lambda_0$ se realizará alguna de las siguientes acciones:

- Si $m = 1$ se combinará al máximo relativo v_1 con el que se encuentre a su derecha (es decir, b_1 pasaría a ser igual a b_2 y se eliminaría el índice 1 de I_{FM})
- Cuando $1 < m < l$, si se verifica que $a_m = w_m$, se combinará al máximo relativo v_m con el que se encuentre a su izquierda (es decir, a_m pasaría a ser igual a a_{m-1} y se eliminaría el índice m de I_{FM}).
- Si $a_m \neq w_m$ y $b_m = w_{m+1}$, con $1 < m < l$, se combinará al máximo relativo v_m con el que se encuentre a su derecha (es decir, b_m pasaría a ser igual a b_{m+1} y se eliminaría el índice m de I_{FM}).
- Si $m = l$, se combinará al máximo relativo v_l con el que se encuentre a su izquierda (es decir, a_l pasaría a ser igual a a_{l-1} y se eliminaría el índice l de I_{FM}).

A partir de estos nuevos puntos de corte se redefine la expresión dada en (3.9) de la siguiente forma:

$$E'_m(h) = \int_{a_m}^{b_m} \{\hat{f}_h(x) - \lambda_m\}_+ dx. \quad (3.10)$$

Una vez finalizado este paso, se repetirá este proceso hasta emparejar a todos los posibles v_i con masa asociada menor que m_0 .

Cuando se hayan combinado a todos los posibles máximos relativos, se denotará como $E'_{(i)}(h)$ a los valores ordenados de mayor a menor de $E'_i(h)$, con $i \in I'_{FM}$, donde I'_{FM} son los índices de I_{FM} que quedan tras realizar el proceso anterior. Esto es, si l' es el cardinal del conjunto I'_{FM} , se tendría que $E'_{(1)}(h) \geq \dots \geq E'_{(l')}(h)$.

Si se define como moda de Fisher y Marron (2001) a aquellos máximos relativos que tienen una masa de probabilidad asociada superior a m_0 y que están por encima de λ_0 , se denotará como j_{FM} al número de modas de Fisher y Marron (2001) de la verdadera f y se planteará el contraste de hipótesis:

$$H_0 : j_{FM} \leq k_{FM}, \text{ frente a } H_1 : j_{FM} > k_{FM}, \quad (3.11)$$

siendo $k_{FM} \in \mathbb{Z}^+$ el número de modas de Fisher y Marron (2001) que se quieren contrastar.

Para obtener el estadístico que se empleará para este contraste, se define:

$$S_{k_{FM}}(h) = \begin{cases} \sum_{i=k+1}^{l'} E'_{(i)}(h) & \text{si } k_{FM} < l', \\ 0 & \text{en otro caso.} \end{cases} \quad (3.12)$$

Con el objetivo de comprender un poco mejor el proceso de obtención de los $E'_i(h)$, se detallan a continuación los pasos a seguir para la obtención del número de modas de Fisher y Marron (2001) si se tuviese como estimación tipo núcleo la curva que se presenta en la Figura 3.1.

1. Observando la Figura 3.1 se puede ver que hay 6 máximos locales a los que se les denotará v_1, \dots, v_6 y 5 mínimos locales a los que se les denotará como w_2, \dots, w_6 . Además, se añadirán los puntos $w_1 = -\infty$ y $w_7 = \infty$
2. Como para el sexto máximo local se cumple que $\hat{f}_h(v_6) \leq \lambda_0$, este máximo local no será candidato a ser moda para esta estimación de la función de densidad. Así, en este momento, se tiene que $I_{FM} = \{1, 2, 3, 4, 5\}$.
3.
 - a) El nivel de masa de v_1 sería $\lambda_1 = \max \left\{ \lim_{t_1 \rightarrow -\infty} \hat{f}_h(t_1), \hat{f}_h(w_2), \lambda_0 \right\} = \hat{f}_h(w_2)$.
 - b) Como se verifica que $\lambda_1 = \hat{f}_h(w_2)$, uno de los puntos de corte asociados al exceso de masa de v_1 sería $b_1 = w_2$, mientras que a_1 coincidirá con el primer punto de la recta real (mayor que $w_1 = -\infty$) verificando que $\hat{f}_h(a_1) = \lambda_1$.
 - c) A partir de estos puntos de corte se puede calcular el exceso de masa asociado a v_1 como $E_1(h) = \int_{a_1}^{w_2} (\hat{f}_h(x) - \hat{f}_h(w_2)) dx$. Los valores de $E_2(h), E_3(h), E_4(h)$ y $E_5(h)$ se obtendrían de forma análoga.
4.
 - a) Si se verifica que $E_5(h) \leq m_0$, como $a_5 \neq w_5$ y $b_5 \neq w_6$, entonces se eliminaría este candidato moda, de forma que se tendría que $I'_{FM} = \{1, 2, 3, 4\}$.

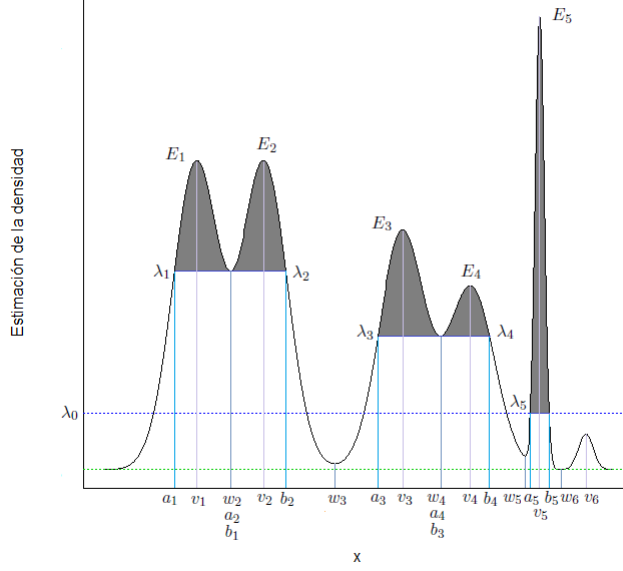


Figura 3.1: Ejemplo del cálculo de exceso de masa para el test de Fisher y Marron.

- b) Si $E_1(h) > m_0$ y $E_2(h) > m_0$ se tendría que v_1 y v_2 son dos modas de Fisher y Marron (2001).
- c) Si $E_3(h) \leq m_0$ y $E_4(h) \leq m_0$, como $E_4(h) < E_3(h)$, entonces se redefinirían los puntos de corte asociados al índice 4, de forma que se combinarían v_3 con v_4 dejando solo un posible candidato a moda, el índice 3, cuyo exceso de masa asociado tendría los puntos de corte a_3 y b_4 , así, se tiene que $E'_3(h) = \int_{a_3}^{b_4} \{\hat{f}_h(x) - \lambda_4\}_+ dx = \int_{a_3}^{b_4} \{\hat{f}_h(x) - \hat{f}_h(w_4)\}_+ dx$.
- d) Si se verifica que $E'_3(h) > m_0$, entonces v_3 sería una moda y como se ha eliminado v_4 para que v_3 alcanzase el exceso de masa mínimo, se tendría que $I'_{FM} = \{1, 2, 3\}$.
5. Una vez se han combinado todos los posibles v_i , como $I'_{FM} = \{1, 2, 3\}$, para esta estimación de la función de densidad se verifica que $S_{k_{FM}}(h) > 0$ si $k_{FM} < 3$ y $S_{k_{FM}}(h) = 0$ en otro caso.

A partir de $S_{k_{FM}}(h)$, ya se puede obtener la expresión de la ventana crítica presentada por Fisher y Marron (2001). Así, dados m_0 y λ_0 , la ventana crítica sería:

$$h_{FM} = \sup \{h : S_{k_{FM}}(h) > 0\}. \quad (3.13)$$

Se puede ver que para un valor suficientemente grande del parámetro de ventana (se denotará

a este parámetro como h_a), siempre se verifica que \hat{f}_{h_a} tiene una sola moda y que, si se tienen al menos k datos de la muestra distintos, siempre que m_0 sea lo suficientemente pequeño, existe un h_b tal que $S_{k_{FM}}(h_b) > 0$. Por tanto, una forma de encontrar el valor de h_{FM} , dado un λ_0 y un m_0 (suficientemente pequeño), sería empezar en un h_c suficientemente grande para que \hat{f}_{h_c} tenga una sola moda e ir reduciendo el valor de h hasta $S_{k_{FM}} > 0$ para la estimación tipo núcleo dada por \hat{f}_h y tomar como h_{FM} a este h .

La nueva ventana dada en (3.13) servirá para realizar el contraste (3.11), para ello, se redefine el estadístico de Cramér-von Mises que se vio en (3.8) de la siguiente forma:

$$T_{FM} = \sum_{i=1}^n \left(\hat{F}_{h_{FM}}(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \quad (3.14)$$

donde $\hat{F}_{h_{FM}}(x) = \int_{-\infty}^x \hat{f}_{h_{FM}}(t) dt$.

El estadístico (3.14) se usará como estadístico de contraste para (3.11). Su distribución bajo la H_0 se podría obtener usando una metodología de remuestreo bootstrap de forma análoga a la realizada con el estadístico T_k en (3.8).

El problema asociado al calibrado de este test es que a diferencia de lo que ocurría con el calibrado de Hall y York (2001) donde bastaba con conocer un soporte acotado de la variable aleatoria, en este se necesitaría disponer de algún criterio para seleccionar λ_0 y m_0 , ya que Fisher y Marron (2001) no especifican cómo obtener los valores de estos parámetros.

3.2. Test basado en el exceso de masa y *dip*

A lo largo de esta sección se analizará el test basado en el *dip* propuesto por Hartigan y Hartigan (1985) para realizar el contraste (1.3) con $k = 1$. El estadístico de contraste para este test se obtendrá a partir de la distancia entre la función de distribución empírica F_n definida en (1.4) y la función unimodal (función con un único máximo, monótona creciente y acotada, aunque no necesariamente de distribución) más próxima.

También se estudiará el test de exceso de masa propuesto por Müller y Sawitzki (1991) para contrastar la hipótesis nula de que la verdadera distribución posee a lo sumo k modas, frente a la alternativa de que la distribución posee más de k modas. En este caso, el estadístico de contraste vendrá determinado por la diferencia entre el exceso de masa de probabilidad (obtenido a través de la distribución empírica) que se tiene en $(k + 1)$ modas y el presentado en k modas. Se presentan conjuntamente estos dos test ya que se puede demostrar que en el caso unidimensional, para contrastar unimodalidad frente a multimodalidad de la distribución, estos son equivalentes, en el sentido de que el valor del estadístico dado en el test de exceso de masa es exactamente dos veces el valor del estadístico del test *dip*. Esta equidad y que el exceso de masa sea el estadístico que se puede adaptar al contexto circular justificarán que se de una descripción más detallada del test presentado por Müller y Sawitzki (1991).

Por último, debido al deficiente calibrado que tienen los test anteriores, se presentan el calibrado de Cheng y Hall (1998) y una nueva propuesta para el caso de contrastar unimodalidad frente a multimodalidad, que constituye una de las aportaciones de este TFM.

3.2.1. Test basado en el *dip*

La idea de la obtención del *dip* se basa en que si F es una función de distribución unimodal con una moda en m será convexa en $(-\infty, m]$, ya que su derivada (la función de densidad) es creciente en $(-\infty, m)$ y será cóncava en $[m, \infty)$, ya que su derivada es decreciente en (m, ∞) .

Con el fin de poder calcular el estadístico *dip* con el que se contrastará la hipótesis nula de que la verdadera distribución es unimodal frente a la hipótesis alternativa de que la verdadera distribución es multimodal, dadas dos funciones acotadas, F y G , se introducirá a la distancia $\rho(F, G) = \sup_x \{|F(x) - G(x)|\}$. Denotando como \mathcal{U} a la clase de funciones de distribución unimodales, se tiene que el *dip* de una función de distribución F , quedaría definido de la siguiente forma:

$$D(F) = \varrho(F, \mathcal{U}) = \inf_{G \in \mathcal{U}} \{\rho(F, G)\}. \quad (3.15)$$

Además, se puede comprobar fácilmente que el *dip* verifica que $D(F_1) \leq D(F_2) + \rho(F_1, F_2)$ y que $D(F) = 0$ si y solo si $F \in \mathcal{U}$. Por tanto, valores “grandes” del *dip* se traducirán en una gran discrepancia entre una distribución dada y una unimodal.

Con el objetivo de poder calcular en la práctica el *dip* de una función de distribución, se extenderá su definición a las funciones acotadas que son constantes en $(-\infty, 0]$ y en $[1, \infty)$ y que para algún $m \in [0, 1]$, son convexas en $[0, m]$ y cóncavas en $[m, 1]$ (se denotará a esta clase de funciones como \mathcal{V}). Para esta clase de funciones, se define el *dip* como sigue:

$$D(F) = \varrho(F, \mathcal{V}) = \inf_{G \in \mathcal{V}} \{\rho(F, G)\}. \quad (3.16)$$

El siguiente resultado, probado por Hartigan y Hartigan (1985), demuestra que para funciones de distribución la definición del *dip* dada en (3.16) es consistente con la dada en (3.15).

Teorema 3.2. (Hartigan y Hartigan, 1985) *Sea F una función de distribución verificando que $F(0) = 0$ y que $F(1) = 1$. Entonces se tiene que $\varrho(F, \mathcal{U}) = \varrho(F, \mathcal{V})$.*

A la vista del Teorema 3.2, bastará con calcular el valor de $\varrho(F, \mathcal{V})$ para obtener el valor del *dip* con respecto a la función de distribución unimodal más próxima.

Para estimar, a partir de la muestra, el valor del estadístico *dip* de la verdadera distribución se puede hacer uso del Teorema de Glivenko–Cantelli. Este teorema garantiza que $\rho(F_n, F) \rightarrow 0$ de forma casi segura. Además, es fácil ver que si $\rho(F_n, F) \rightarrow 0$, entonces se verifica que $D(F_n) \rightarrow D(F)$ de forma casi segura. Con esto se obtiene que, con probabilidad uno, para un tamaño muestral

suficientemente grande, el *dip* obtenido a través de la función empírica se aproximará al *dip* de la verdadera función de distribución.

Consecuentemente, basta con calcular $g(F_n, \mathcal{V})$ para estimar el valor del *dip*. La construcción de la función de la clase \mathcal{V} más próxima a F_n se ilustró en la Figura 1.9 mostrada en la Sección 1.2 y se detalla en la Sección 4 de Hartigan y Hartigan (1985).

Como estadístico de contraste, Hartigan y Hartigan (1985) emplean al valor de $\sqrt{n}D(F_n)$, para obtener su distribución bajo H_0 , Hartigan y Hartigan (1985) proponen utilizar la distribución del *dip* de la distribución $U(0, 1)$ obtenida a través de la metodología de Monte Carlo. Para ello, se generan M muestras de tamaño n procedentes de una distribución uniforme y se calcula el *dip* asociado a cada una de estas muestras. A partir de ellos se generarán los valores críticos κ_α , de forma que dada una muestra \mathcal{X} , de tamaño n , para un nivel de significación α , se rechazará H_0 si $\mathbb{P}(\kappa_\alpha \leq D(F_n)) \geq 1 - \alpha$.

Los resultados de las simulaciones que se presentarán en la Sección 3.3, justificarán la necesidad de un nuevo calibrado, como el propuesto por Cheng y Hall (1998) o la nueva propuesta que se presenta en este Trabajo Fin de Máster.

3.2.2. Test de exceso de masa de Muller y Sawitzki

El estadístico de Müller y Sawitzki (1991) es una extensión del estadístico de Hartigan y Hartigan (1985) para el caso en el que se quiera contrastar la hipótesis nula de que la verdadera distribución posee a lo sumo k modas frente a la hipótesis alternativa de que la verdadera distribución posee más de k modas. Se basa en la idea de que una moda estará presente donde haya un exceso de masa de probabilidad concentrado.

Para introducir el estadístico de contraste de este test, dada la función continua de densidad f y un $\lambda \in \mathbb{R}^+$, se define la función exceso de masa como:

$$E(f, \lambda) = \int \{f(x) - \lambda\}_+ dx = \int_{C_\lambda} f(x) - \lambda |C_\lambda| = \mathbb{P}(C_\lambda) - \lambda |C_\lambda| = \sup_C (\mathbb{P}(C) - \lambda |C|),$$

donde $C_\lambda = \{x : f(x) \geq \lambda\}$, $|C_\lambda|$ es la medida del conjunto C_λ y C es un conjunto de Borel.

Denotando como λ -conglomerados a las componentes conexas de C_λ , se sabe que una distribución con j modas, tendrá a lo sumo j λ -conglomerados. Si se denota como C_1, \dots, C_j a los j λ -conglomerados que tiene f , se obtiene que $C_\lambda = \bigcup_{i=1}^j C_i$ y que $E(f, \lambda) = \sum_{i=1}^j \mathbb{P}(C_i) - \lambda |C_i|$. Así, se define la función exceso de masa para j modas para el caso unidimensional como:

$$E_j(f, \lambda) = \sup_{C_1, \dots, C_j} \left\{ \sum_{i=1}^j \int_{C_i} (f(x) - \lambda) dx \right\} = \sup_{C_1, \dots, C_j} \left\{ \sum_{i=1}^j \mathbb{P}(C_i) - \lambda |C_i| \right\}, \quad (3.17)$$

donde el supremo se toma sobre todos los posibles C_i , con $i = 1, \dots, j$, siendo estos C_i conjuntos cerrados y conexos, es decir, intervalos cerrados. Como los C_i son intervalos del tipo $C_i = [a_i, b_i]$, se tiene que, en este caso, $\|C_i\| = b_i - a_i$.

A partir de (3.17), dada una muestra $\mathcal{X} = (X_1, \dots, X_n)$, se define el exceso de masa empírico para k modas del siguiente modo:

$$E_{n,k}(\mathbb{P}_n, \lambda) = \sup_{C_1, \dots, C_k} \left\{ \sum_{l=1}^k \mathbb{P}_n(C_l) - \lambda \|C_l\| \right\}, \quad (3.18)$$

con $\mathbb{P}_n(C_l) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(X_i \in C_l)$.

Con el fin de obtener un estadístico que permita realizar el contraste (1.3), para un λ dado, se denota como $D_{n,k+1}(\lambda) = E_{n,k+1}(\mathbb{P}_n, \lambda) - E_{n,k}(\mathbb{P}_n, \lambda)$. A partir de este $D_{n,k+1}(\lambda)$ se define el estadístico $\Delta_{n,k+1}$ del siguiente modo:

$$\Delta_{n,k+1} = \max_{\lambda} \{D_{n,k+1}(\lambda)\}. \quad (3.19)$$

Müller y Sawitzki (1991) muestran que para el caso $k = 1$, $\Delta_{n,k+1}$ es igual a dos veces el valor del estadístico *dip*.

La hipótesis de que la verdadera distribución posee a lo sumo k modas es rechazada a favor de que posee al menos $(k + 1)$ modas para valores grandes de $\Delta_{n,k+1}$. El problema, tal y como afirman Müller y Sawitzki (1991), es que la distribución de este estadístico, depende de la verdadera distribución de la que proviene la muestra observada.

Para el caso $k = 1$, Müller y Sawitzki (1991), al igual que hacían Hartigan y Hartigan (1985), proponen emplear la distribución que tiene $\Delta_{n,2}$ para una $U(0, 1)$ como distribución del estadístico bajo la hipótesis nula. La elección de esta distribución es debida a que, tal y como prueban Müller y Sawitzki (1991), para tamaños muestrales suficientemente grandes, para la regla que dada una muestra \mathcal{X} , de tamaño n , para un nivel de significación α , rechaza H_0 si $\mathbb{P}(\kappa_{\alpha} \leq D(F_n)) \geq 1 - \alpha$, se tiene que el porcentaje de rechazos para una muestra generada bajo H_0 siempre será menor o igual que el nivel nominal α .

3.2.3. Calibrado de los test de exceso de masa y *dip*

A la vista del deficiente calibrado de los test propuestos por Hartigan y Hartigan (1985) y Müller y Sawitzki (1991) para $\Delta_{n,2}$, Cheng y Hall (1998) proponen un calibrado de este estadístico basándose en que para muestras de tamaño grande, bajo la hipótesis nula de que la distribución es unimodal, la distribución de $\Delta_{n,2}$ es independiente de la distribución desconocida excepto por el factor:

$$c = \left(\frac{(f(x_0))^3}{|f''(x_0)|} \right)^{1/5}, \quad (3.20)$$

donde x_0 denota la localización de la única moda de f .

Como la distribución de $\Delta_{n,2}$ es independiente de la verdadera distribución de la variable aleatoria salvo el parámetro c , para calibrar la distribución de $\Delta_{n,2}$, Cheng y Hall (1998) sugieren utilizar la distribución que tiene dicho estadístico cuando la muestra sigue alguna de las tres clases de funciones de densidad que se ilustrarán a continuación y que cubren todo el rango de posibles valores de $d = c^{-5} = |f''(x_0)|/f^3(x_0)$. La elección se hará estimando d a partir de la muestra. Teniendo en cuenta que $d \in [0, \infty)$, las clases propuestas por Cheng y Hall (1998) son las siguientes:

- Si $d < 2\pi$, entonces se emplea para calibrar la distribución Beta(b, b), cuya función de densidad es:

$$\psi(x; b) = \frac{1}{\mathbf{B}(b, b)} [x(1-x)]^{b-1},$$

para $x \in (0, 1)$ y $b > 1$, siendo $\mathbf{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$ la función beta.

Como, en este caso, independientemente de b , la única moda se alcanza en el punto $x_0 = 1/2$, se tomará como b al parámetro que verifique la siguiente igualdad $d = |\psi''(0.5; b)|/\psi^3(0.5; b) = \mathbf{B}(b, b)^2 \cdot 2^{4b-1}(b-1)$.

- Si $d = 2\pi$, se usa como referencia cualquier distribución normal, ya que independientemente de los parámetros μ y σ^2 , denotando como $\psi(\mu)$ al valor de la función de densidad de la $N(\mu, \sigma^2)$ en su moda, es decir en μ , se tiene que $d = |\psi''(\mu)|/\psi^3(\mu) = 2\pi$.
- Si $d > 2\pi$, se emplea como distribución de referencia una t de Student reescalada, cuya función de densidad es la siguiente:

$$\psi(x; b) = \frac{1}{\mathbf{B}(b-0.5, 0.5)(1+x^2)^b},$$

donde $x \in \mathbb{R}$ y $b > 0.5$.

Como independientemente del parámetro b , la única moda de esta función de densidad se encuentra en el punto $x_0 = 0$, b será el parámetro que verifique la siguiente igualdad $d = |\psi''(0; b)|/\psi^3(0; b) = 2b \cdot \mathbf{B}(b-0.5, 0.5)^2$.

Se toman estas tres distribuciones como referencia al igual que hacen Cheng y Hall (1998), si bien estas pueden ser sustituidas por cualquier otra clase de referencia que goce de una relación biyectiva con el conjunto $[0, \infty)$ de posibles valores de d .

El problema será que, como se desconoce la verdadera función de densidad f , para saber qué distribución se debe usar, es necesario estimar la expresión de d , lo cual repercutirá a su vez en

que se tenga que estimar la expresión de b . Para realizar esta estimación se denotará como \hat{f}_g'' a la estimación tipo núcleo de f'' con parámetro de ventana g , esto es:

$$\hat{f}_g''(x) = \frac{1}{ng^3} \sum_{i=1}^n K'' \left(\frac{x - X_i}{g} \right).$$

En particular, para estimar $f''(x_0)$, Cheng y Hall (1998) proponen utilizar el estimador anterior con núcleo gaussiano y el selector de g óptimo basado en la hipótesis de que f sigue una normal. Así, el valor de g se obtiene a través de la siguiente expresión (véase Scott (1992), Sec. 6.2):

$$\hat{g} = 0.94n^{-1/9} \hat{\sigma},$$

donde $\hat{\sigma}$ será la cuasidesviación típica muestral.

Para estimar $f(x_0)$, Cheng y Hall (1998) proponen utilizar \hat{f}_h con un núcleo gaussiano, y para seleccionar el valor de h emplean el selector de ventana normal, esto es, \hat{h} tendrá la siguiente expresión (véase Scott (1992), Sec. 6.2):

$$\hat{h} = \left(\frac{4}{3} \right)^{1/5} n^{-1/5} \hat{\sigma},$$

donde $\hat{\sigma}$ será nuevamente la cuasidesviación típica muestral.

Denotando como \hat{x}_0 a la posición donde se encuentra la moda predominante para \hat{f}_h , esto es, $\hat{x}_0 = \arg \max \{\hat{f}_h\}$, se tiene que la estimación de d vendrá dada por la siguiente expresión:

$$\hat{d} = \frac{|\hat{f}_g''(\hat{x}_0)|}{\left(\hat{f}_h(\hat{x}_0) \right)^3}. \quad (3.21)$$

Para construir el test de Cheng y Hall (1998) para una muestra $\mathcal{X} = (X_1, \dots, X_n)$ dada, es necesario obtener el valor de $\Delta_{n,2}$ y del \hat{d} de (3.21) asociados a dicha muestra. Haciendo uso del valor de \hat{d} obtenido, se estima \hat{b} y se escoge la función de densidad $\psi(x; \hat{b})$ asociada. A partir de la función de densidad $\psi(x; \hat{b})$, se construyen las muestras bootstrap $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$ y, utilizando estas remuestras, se calcularán sus $\Delta_{n,2}^*$ asociados. Se aproximará la distribución de $\Delta_{n,2}$ de forma que para un nivel de significación α se rechazará la hipótesis nula de que la verdadera función de densidad f es unimodal si $\mathbb{P}(\Delta_{n,2}^* \leq \Delta_{n,2} | \mathcal{X}) \geq 1 - \alpha$.

3.2.4. Nueva propuesta: Calibrado de los test de exceso de masa y *dip* haciendo uso de la ventana crítica

Con el objetivo de mejorar el calibrado del test de Müller y Sawitzki (1991) para contrastar la hipótesis nula de que la verdadera distribución posee una moda frente a la alternativa de que la

verdadera distribución posee más de una moda, en el presente TFM se introduce otro calibrado del estadístico $\Delta_{n,2}$ basado en la ventana crítica.

Este test no solo presenta la ventaja de mejorar el calibrado del desarrollado por Müller y Sawitzki (1991), si no que además, a diferencia del test de Cheng y Hall (1998), permite su extensión a datos circulares, posibilitando, en particular, realizar un test para contrastar si en la comarca de Vigo hay una única temporada de incendios o más de una.

Los pasos a seguir para realizar esta nueva propuesta son los siguientes:

1. Dada la muestra $\mathcal{X} = (X_1, \dots, X_n)$, obtener el valor del estadístico $\Delta_{n,2}$ a partir de la ecuación (3.19) cuando $k = 1$.
2. Calcular el valor de la ventana crítica dada en (3.1) cuando $k = 1$. Se denotará a esta ventana como h_1 .
3. Haciendo uso de h_1 se tomará como función de densidad, bajo la hipótesis nula, la estimación tipo núcleo \hat{f}_{h_1} para generar B réplicas bootstrap $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$.
4. A partir de cada una de las B réplicas bootstrap \mathcal{X}^{*b} , con $b = 1, \dots, B$; se obtienen los estadísticos del test de exceso de masa asociados a cada una de estas réplicas, $\Delta_{n,2}^{*b}$.
5. Para un nivel de significación α se rechazará la hipótesis nula si $\mathbb{P}(\Delta_{n,2}^* \leq \Delta_{n,2} | \mathcal{X}) \geq 1 - \alpha$.

La ventaja que presenta este test, que ya mostraba el de Cheng y Hall (1998), sobre la propuesta de Hall y York (2001) es que no se necesita conocer el soporte de la verdadera distribución de la muestra. Si bien, cuando este sea conocido, nuestra propuesta sería emplear, en el paso 2, la ventana crítica de Hall y York (2001) dada en (3.3) en vez de la ventana crítica de Silverman (1981).

3.3. Estudio de simulación

A continuación se presenta un estudio de simulación para los test descritos en este capítulo, los cuales han sido programados en el software estadístico R Core Team (2013). En todos los estudios realizados, se han tomado los niveles de significación $\alpha = 0.01, 0.05, 0.1$, tamaños muestrales $n = 50$, $n = 200$ y $n = 1000$ ($n = 100$ en lugar de $n = 1000$ en los estudios de potencia), se han usado 500 muestras distintas de cada distribución y se han realizado $B = 500$ réplicas bootstrap para obtener el p-valor asociado a cada muestra.

Con el fin de analizar el calibrado de los test presentados, en las Tablas 3.2 y 3.3, para el contraste dado en (1.3) cuando $k = 1$, se muestra el porcentaje de rechazos de los test de Silverman (1981) (Sección 3.1.1), la versión sencilla del de Fisher y Marron (2001), esto es, con $\lambda_0 = 0$ y $m_0 = 0$ ya que no se tiene ningún criterio para seleccionar estos parámetros (Sección 3.1.3), el de Cheng y Hall (1998) (Sección 3.2.3) y la nueva propuesta presentada para este trabajo utilizando la ventana crítica (Sección 3.2.4), en las siguientes distribuciones unimodales (véase Johnson *et al.*, 1995): Beta(3,4),

Gamma(3,1), t_5 , χ_3^2 , Weibull(5,1), la normal M1 (modelos sencillos) y las mixturas de normales M2, M3, M8 y M9 (modelos complejos) descritas en el Apéndice A.

Además, en las Tablas 3.2 y 3.3 se ha realizado un estudio de simulación para el contraste con soporte conocido dado en (3.2) para el Método 1 propuesto por Hall y York (2001) (Sección 3.1.2), el que rechaza H_0 si $\mathbb{P}(h_{HY}^* \leq \lambda_\alpha h_{HY} | \mathcal{X}) \geq 1 - \alpha$, donde los λ_α son los dados en (3.4). También se han calculado los porcentajes de rechazo para la nueva propuesta presentada para este trabajo utilizando la ventana de Hall y York (2001) (Sección 3.2.4). Se muestran en estas tablas el porcentaje de rechazos en las siguientes distribuciones unimodales: Beta(3,4) en el intervalo $[0, 1]$, Gamma(3,1) en el intervalo $[0.5, 5]$, t_5 en el intervalo $[-1.5, 1.5]$, χ_3^2 en el intervalo $[0, 5]$, Weibull(5,1) en el intervalo $[0.5, 1.3]$ y de las mixturas de normales (descritas en el Apéndice A) M1 en el intervalo $[0, 1]$, M2 en el intervalo $[0, 1]$, M3 en el intervalo $[0, 1]$, M8 en el intervalo $[-1.5, 2.5]$ y M9 en el intervalo $[-5, 5]$.

No se muestran las tablas del test propuesto por Hartigan y Hartigan (1985), empleado también por Müller y Sawitzki (1991), ya que, tal y como señalan Cheng y Hall (1998), se ha obtenido un test demasiado conservador. Para comprobar este hecho, se han generado 1000 muestras, de 1000 datos cada una, de una distribución $N(0,1)$, de una Beta(3,4), de una Gamma(3,1) y de una t_5 . Los valores críticos κ_α han sido generados a partir de los *dip* obtenidos para 10000 muestras, con 1000 datos cada una, procedentes de una distribución $U(0,1)$. Con el fin de obtener el valor del *dip* para cada muestra se ha empleado la función programada en R en la librería `dipTest` por Maechler (2013). Para los distintos estudios de simulación realizados bajo la hipótesis nula, se ha obtenido una probabilidad de rechazo de la hipótesis nula menor o igual a 0.005, para un nivel de significación α del 0.2.

En la Tabla 3.4 se ha realizado un estudio de potencia de los distintos test. Se ha calculado el porcentaje de rechazos para las siguientes distribuciones, descritas en el Apéndice A, con dos modas en el intervalo $[0, 1]$: M4, M5, M6 y M7. En el caso del test propuesto por Hall y York (2001) y de la nueva propuesta utilizando la ventana de Hall y York (2001) se ha empleado como intervalo $I = [0, 1]$ en las cuatro distribuciones estudiadas.

Test basados en la ventana crítica

En primer lugar, se analizarán los dos test que permiten realizar el contraste (1.3) cuando $k \in \mathbb{Z}^+$.

Para el test de Silverman (1981) presentado en la Sección 3.1.1, se tiene que cuando se observan los modelos estudiados bajo H_0 , el porcentaje de rechazos está, incluso para tamaños muestrales elevados, significativamente (a un nivel del 5%) por debajo del nivel de significación nominal, a excepción de los modelos t_5 (para $\alpha = 0.1$ y $n = 1000$), M8 (para $\alpha = 0.1$ y $n = 200$ o $n = 1000$) y M3 (para $\alpha = 0.05$ o $\alpha = 0.1$ y $n = 200$ o $n = 1000$). Este hecho permite concluir que este test será claramente conservador.

En el caso del test de Fisher y Marron (2001) presentado en la Sección 3.1.3, aunque en la mayoría de los modelos estudiados bajo H_0 se observa que el porcentaje de rechazos está muy por encima del nivel de significación nominal. Se tiene que en ciertos modelos como son la Beta(3,4), la Weibull(5,1), el M1 y el M9 (para $n = 50$) el porcentaje de rechazos está significativamente (a un nivel del 5%) por debajo del nivel teórico α .

A la vista de estos resultados, se puede concluir que ninguno de los test presentados que permiten

realizar el contraste (1.3) cuando $k > 1$ va a tener un calibrado correcto.

De los dos métodos presentados por Hall y York (2001) en la Sección 3.1.2 para realizar el contraste (3.2) se muestra solo los resultados del Método 1 ya que, en general, el calibrado obtenido con este método es ligeramente mejor para estos modelos que el obtenido con la otra propuesta de Hall y York (2001). Para este test se tiene que el calibrado es bastante deficiente cuando $n = 50$, mejorando claramente con el tamaño muestral. Así, para los tamaños $n = 200$ y $n = 1000$ en los modelos sencillos se obtiene que el porcentaje de rechazos está próximo al nivel de significación impuesto. Analizando los modelos complejos, parece que este test no consigue un buen calibrado incluso para tamaño muestral grande, ya que en general (para $n = 1000$), el porcentaje de rechazos obtenidos está significativamente (a un nivel del 5%) por encima del nivel nominal α .

Test basados en el *dip* o en el exceso de masa

En cuanto al test de Cheng y Hall (1998) presentado en la Sección 3.2.3, se puede observar en las Tablas 3.2 y 3.3 que funciona bastante bien en los modelos sencillos, incluso desde tamaños muestrales muy pequeños ($n = 50$), a excepción del modelo t_5 y de la normal M1 (para $n = 1000$), donde ya el porcentaje de rechazos (para todos los tamaños muestrales estudiados) era significativamente (a un nivel del 5%) inferior a α usando el Método 1 de Hall y York (2001). En los modelos complejos, en general, el calibrado obtenido no es correcto, a excepción del modelo M9, donde el porcentaje de rechazos no es significativamente (a un nivel del 5%) distinto del nivel nominal α .

Para la nueva propuesta presentada para este TFM en la Sección 3.2.4 cuando se usa la ventana crítica, se tiene que el calibrado es bastante bueno, incluso desde tamaños muestrales muy pequeños ($n = 50$), a excepción de los modelos t_5 y M1 (para $n = 1000$) donde el porcentaje de rechazos de H_0 es significativamente (a un nivel del 5%) inferior a α y del modelo complejo M9, donde el porcentaje de rechazos es significativamente superior al nivel teórico α .

En cuanto a la modificación empleando la ventana de Hall y York (2001) para la nueva propuesta presentada en la Sección 3.2.4, en todos los modelos estudiados bajo H_0 , el porcentaje de rechazos conseguido para los distintos niveles de significación ha sido similar al obtenido con la nueva propuesta empleando la ventana crítica de Silverman (1981), a excepción del modelo M9 donde se consigue que, para tamaño muestral alto, el porcentaje de rechazos no sea significativamente (a un nivel del 5%) distinto del nivel nominal α .

En cuanto a los resultados de potencia de los test que calibran correctamente, estos son el test de Cheng y Hall (1998) y de la nueva propuesta presentada en la Sección 3.2.4, se puede observar en la Tabla 3.4 que ninguno de los dos es más potente ya que mientras que en la mixtura M7 (para $n = 50$) y en la M4 (para $n = 200$) la nueva propuesta (tanto como con la ventana crítica como con la de Hall y York, 2001) es claramente más potente que el test de Cheng y Hall (1998), en las mixturas M5 y M6 se obtienen resultados contrarios. Además, se observa que la potencia de ambos test es bastante buena, en general, ya que mejora con el tamaño muestral aproximándose a 1. En cuanto a la potencia, el problema más grave que se ha observado es que tanto la nueva propuesta (para tamaños $n = 50$ y $n = 100$) como el test de Cheng y Hall (1998) (para tamaños $n = 50, n = 100$ y $n = 200$) no son capaces de detectar que el modelo M4 no es unimodal.

Ante todos estos hechos, entre los test presentados, la recomendación para contrastar si una

distribución es unimodal sería emplear el test de Cheng y Hall (1998) o la nueva propuesta con la ventana crítica de Silverman (1981) cuando se desconoce el soporte de la variable aleatoria o la versión modificada con la ventana de Hall y York (2001) cuando el soporte es conocido.

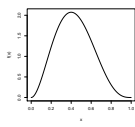
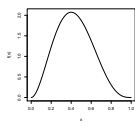
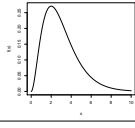
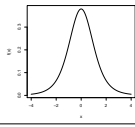
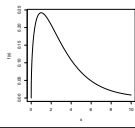
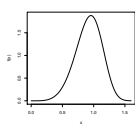
		I	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
				Silverman (1981)	Fisher y Marron (2001)	Método 1 de Hall y York (2001)						
Beta(3,4)		[0,1]	n=50 n=200 n=1000	Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
				0 (0)	0 (0)	0.006 (0.007)	0 (0)	0.024 (0.013)	0.054 (0.020)	0 (0)	0.018 (0.012)	0.046 (0.018)
				0 (0)	0.006 (0.007)	0.032 (0.015)	0.002 (0.004)	0.022 (0.013)	0.070 (0.022)	0.006 (0.007)	0.056 (0.020)	0.102 (0.027)
			n=50 n=200 n=1000	Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
				0.024 (0.013)	0.078 (0.024)	0.130 (0.029)	0.014 (0.010)	0.048 (0.019)	0.092 (0.025)	0.014(0.010)	0.048(0.019)	0.092(0.025)
				0.020 (0.012)	0.082 (0.024)	0.126 (0.029)	0.002 (0.004)	0.052 (0.019)	0.098 (0.026)	0.002(0.004)	0.052(0.019)	0.098(0.026)
0.024 (0.013)	0.064 (0.021)	0.110 (0.027)	0.010 (0.009)	0.048 (0.019)	0.084 (0.024)	0.010(0.009)	0.048(0.019)	0.084(0.024)				
Gamma(3,1)		[0.5,5]	n=50 n=200 n=1000	Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
				0 (0)	0 (0)	0.002 (0.004)	0.030 (0.015)	0.118 (0.028)	0.178 (0.034)	0 (0)	0.020 (0.012)	0.074 (0.023)
				0 (0)	0 (0)	0.004 (0.006)	0.044 (0.018)	0.116 (0.028)	0.188 (0.034)	0.006 (0.007)	0.034 (0.016)	0.078 (0.024)
			n=50 n=200 n=1000	Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
				0.020 (0.012)	0.054 (0.020)	0.098 (0.026)	0.016 (0.011)	0.048 (0.019)	0.102 (0.027)	0.016(0.011)	0.046(0.018)	0.082(0.024)
				0.008 (0.008)	0.054 (0.020)	0.086 (0.025)	0.012 (0.010)	0.044 (0.018)	0.104 (0.027)	0.010(0.009)	0.038(0.017)	0.086(0.025)
0.002 (0.004)	0.042 (0.018)	0.096 (0.026)	0.004 (0.006)	0.058 (0.020)	0.108 (0.027)	0.004(0.006)	0.026(0.014)	0.084(0.024)				
t_5		[-1.5,1.5]	n=50 n=200 n=1000	Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
				0 (0)	0 (0)	0.002 (0.004)	0.054 (0.020)	0.138 (0.030)	0.236 (0.037)	0 (0)	0.004 (0.006)	0.030 (0.015)
				0 (0)	0 (0)	0.034 (0.016)	0.148 (0.031)	0.294 (0.040)	0.390 (0.043)	0.002 (0.004)	0.014 (0.010)	0.032 (0.015)
			n=50 n=200 n=1000	Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
				0.004 (0.006)	0.026 (0.014)	0.062 (0.021)	0 (0)	0.038 (0.017)	0.078 (0.024)	0(0)	0.032(0.015)	0.062(0.021)
				0 (0)	0.016 (0.011)	0.054 (0.020)	0.006 (0.007)	0.034 (0.016)	0.074 (0.023)	0.004(0.006)	0.034(0.016)	0.072(0.023)
0 (0)	0.020 (0.012)	0.052 (0.019)	0.006 (0.007)	0.026 (0.014)	0.072 (0.023)	0.002(0.004)	0.024(0.013)	0.062(0.021)				
χ_3^2		[0,5]	n=50 n=200 n=1000	Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
				0 (0)	0 (0)	0 (0)	0.076 (0.023)	0.242 (0.038)	0.366 (0.042)	0.002 (0.004)	0.018 (0.012)	0.080 (0.024)
				0 (0)	0 (0)	0.004 (0.006)	0.172 (0.033)	0.320 (0.041)	0.466 (0.044)	0.018 (0.012)	0.060 (0.021)	0.114 (0.028)
			n=50 n=200 n=1000	Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
				0.008 (0.008)	0.060 (0.021)	0.098 (0.026)	0.010 (0.009)	0.062 (0.021)	0.122 (0.029)	0.006(0.007)	0.048(0.019)	0.086(0.025)
				0.018 (0.012)	0.070 (0.022)	0.116 (0.028)	0.022 (0.013)	0.090 (0.025)	0.138 (0.030)	0.014(0.010)	0.066(0.022)	0.122(0.029)
0.008 (0.008)	0.048 (0.019)	0.096 (0.026)	0.012 (0.010)	0.066 (0.022)	0.126 (0.029)	0(0)	0.038(0.017)	0.084(0.024)				
Weibull(5,1)		[0.5,1.3]	n=50 n=200 n=1000	Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
				0 (0)	0 (0)	0.002 (0.004)	0.002 (0.004)	0.012 (0.010)	0.032 (0.015)	0 (0)	0.010 (0.009)	0.056 (0.020)
				0 (0)	0.002 (0.004)	0.012 (0.010)	0.004 (0.006)	0.018 (0.012)	0.042 (0.018)	0.002 (0.004)	0.042 (0.018)	0.076 (0.023)
			n=50 n=200 n=1000	Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
				0.004 (0.006)	0.010 (0.009)	0 (0)	0.004 (0.006)	0.006 (0.007)	0.024 (0.013)	0.006 (0.007)	0.046 (0.018)	0.094 (0.026)
				0.008 (0.008)	0.052 (0.019)	0.100 (0.026)	0.002 (0.004)	0.042 (0.018)	0.076 (0.023)	0.002(0.004)	0.038(0.017)	0.076(0.023)
0.014 (0.010)	0.056 (0.020)	0.098 (0.026)	0.004 (0.006)	0.044 (0.018)	0.100 (0.026)	0.004(0.006)	0.042(0.018)	0.096(0.026)				
0.002 (0.004)	0.042 (0.018)	0.082 (0.024)	0 (0)	0.036 (0.016)	0.084 (0.024)	0(0)	0.034(0.016)	0.084(0.024)				

Tabla 3.2: Porcentajes de rechazo (entre paréntesis aparece 1.96 veces su desviación típica aproximada), con niveles de significación $\alpha = 0.01, 0.05, 0.1$, para los test de unimodalidad de: Silverman (1981), Fisher y Marron (2001), el Método 1 de Hall y York (2001) para los intervalos especificados en la columna I , Cheng y Hall (1998) y para la nueva propuesta presentada en la Sección 3.2.4 usando la ventana crítica y la ventana de Hall y York (2001) empleando los intervalos especificados en la columna I . Estos han sido obtenidos tras realizar 500 réplicas bootstrap en 500 muestras de tamaño $n = 50$, $n = 200$ y $n = 1000$ procedentes de las distribuciones: Beta(3,4), Gamma(3,1), t_5 , χ_3^2 y Weibull(5,1).

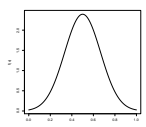
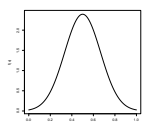
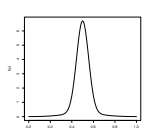
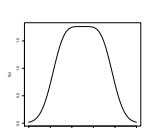
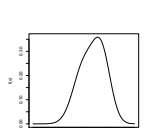
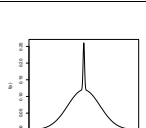
		I	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
				Silverman (1981)	Fisher y Marron (2001)	Método 1 de Hall y York (2001)						
M1		[0,1]	n=50	0 (0)	0 (0)	0.002 (0.004)	0 (0)	0.016 (0.011)	0.060 (0.021)	0 (0)	0.020 (0.012)	0.052 (0.019)
			n=200	0 (0)	0 (0)	0.004 (0.006)	0 (0)	0.010 (0.009)	0.042 (0.018)	0.006 (0.007)	0.038 (0.017)	0.074 (0.023)
			n=1000	0 (0)	0.004 (0.006)	0.008 (0.008)	0 (0)	0.008 (0.008)	0.042 (0.018)	0.008 (0.008)	0.066 (0.022)	0.102 (0.027)
			Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))			
			n=50	0.016 (0.011)	0.062 (0.021)	0.116 (0.028)	0.004 (0.006)	0.044 (0.018)	0.098 (0.026)	0.004(0.006)	0.044(0.018)	0.094(0.026)
			n=200	0.008 (0.008)	0.044 (0.018)	0.086 (0.025)	0.004 (0.006)	0.040 (0.017)	0.072 (0.023)	0.004(0.006)	0.036(0.016)	0.070(0.022)
n=1000	0.004 (0.006)	0.032 (0.015)	0.064 (0.021)	0.002 (0.004)	0.022 (0.013)	0.060 (0.021)	0.002(0.004)	0.022(0.013)	0.058(0.020)			
M2		[0,1]	n=50	0 (0)	0 (0)	0.004 (0.006)	0.158 (0.032)	0.340 (0.042)	0.458 (0.044)	0.038 (0.017)	0.074 (0.023)	0.200 (0.035)
			n=200	0 (0)	0 (0)	0.024 (0.013)	0.240 (0.037)	0.452 (0.044)	0.602 (0.043)	0.068 (0.022)	0.148 (0.031)	0.252 (0.038)
			n=1000	0 (0)	0.004 (0.006)	0.022 (0.013)	0.184 (0.034)	0.350 (0.042)	0.464 (0.044)	0.058 (0.020)	0.124 (0.029)	0.208 (0.036)
			Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))			
			n=50	0.002 (0.004)	0.046 (0.018)	0.090 (0.025)	0.006 (0.007)	0.060 (0.021)	0.102 (0.027)	0.006(0.007)	0.060(0.021)	0.102(0.027)
			n=200	0.002 (0.004)	0.032 (0.015)	0.062 (0.021)	0.014 (0.010)	0.052 (0.019)	0.092 (0.025)	0.014(0.010)	0.052(0.019)	0.092(0.025)
n=1000	0.004 (0.006)	0.030 (0.015)	0.074 (0.023)	0.004 (0.006)	0.048 (0.019)	0.110 (0.027)	0.004(0.006)	0.050(0.019)	0.108(0.027)			
M3		[0,1]	n=50	0 (0)	0.002 (0.004)	0.016 (0.011)	0.014 (0.010)	0.052 (0.019)	0.098 (0.026)	0.002 (0.004)	0.052 (0.019)	0.104 (0.027)
			n=200	0 (0)	0.034 (0.016)	0.116 (0.028)	0.028 (0.014)	0.128 (0.029)	0.216 (0.036)	0.052 (0.019)	0.174 (0.033)	0.274 (0.039)
			n=1000	0.006 (0.007)	0.054 (0.020)	0.108 (0.027)	0.026 (0.014)	0.090 (0.025)	0.154 (0.032)	0.072 (0.023)	0.168 (0.033)	0.242 (0.038)
			Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))			
			n=50	0.034 (0.016)	0.128 (0.029)	0.196 (0.035)	0.018 (0.012)	0.070 (0.022)	0.128 (0.029)	0.018(0.012)	0.070(0.022)	0.128(0.029)
			n=200	0.092 (0.025)	0.208 (0.036)	0.280 (0.039)	0.034 (0.016)	0.094 (0.026)	0.176 (0.033)	0.034(0.016)	0.094(0.026)	0.176(0.033)
n=1000	0.066 (0.022)	0.178 (0.034)	0.258 (0.038)	0.018 (0.012)	0.068 (0.022)	0.128 (0.029)	0.018(0.012)	0.064(0.021)	0.128(0.029)			
M8		[-1.5,2.5]	n=50	0 (0)	0.002 (0.004)	0.008 (0.008)	0.012 (0.010)	0.088 (0.025)	0.146 (0.031)	0.002 (0.004)	0.026 (0.014)	0.090 (0.025)
			n=200	0 (0)	0.020 (0.012)	0.080 (0.024)	0.024 (0.013)	0.134 (0.030)	0.232 (0.037)	0.020 (0.012)	0.124 (0.029)	0.222 (0.036)
			n=1000	0 (0)	0.026 (0.014)	0.086 (0.025)	0.020 (0.012)	0.104 (0.027)	0.176 (0.033)	0.054 (0.020)	0.172 (0.033)	0.248 (0.038)
			Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))			
			n=50	0.050 (0.019)	0.140 (0.030)	0.214 (0.036)	0.024 (0.013)	0.066 (0.022)	0.134 (0.030)	0.024(0.013)	0.066(0.022)	0.132(0.030)
			n=200	0.098 (0.026)	0.224 (0.037)	0.314 (0.041)	0.022 (0.013)	0.102 (0.027)	0.186 (0.034)	0.022(0.013)	0.102(0.027)	0.184(0.034)
n=1000	0.048 (0.019)	0.140 (0.030)	0.216 (0.036)	0.010 (0.009)	0.056 (0.020)	0.108 (0.027)	0.010(0.009)	0.056(0.020)	0.102(0.027)			
M9		[-5,5]	n=50	0 (0)	0 (0)	0.002 (0.004)	0.002 (0.004)	0.012 (0.010)	0.040 (0.017)	0 (0)	0.008 (0.008)	0.032 (0.015)
			n=200	0 (0)	0 (0)	0 (0)	0.006 (0.007)	0.064 (0.021)	0.192 (0.035)	0.002 (0.004)	0.020 (0.012)	0.062 (0.021)
			n=1000	0 (0)	0.002 (0.004)	0.004 (0.006)	0.020 (0.012)	0.232 (0.037)	0.510 (0.044)	0.006 (0.007)	0.070 (0.022)	0.140 (0.030)
			Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))			
			n=50	0.012 (0.010)	0.048 (0.019)	0.088 (0.025)	0.004 (0.006)	0.034 (0.016)	0.064 (0.021)	0(0)	0.004(0.006)	0.006(0.007)
			n=200	0.012 (0.010)	0.066 (0.022)	0.106 (0.027)	0.012 (0.010)	0.074 (0.023)	0.148 (0.031)	0.004(0.006)	0.032(0.015)	0.060(0.021)
n=1000	0.016 (0.011)	0.050 (0.019)	0.094 (0.026)	0.022 (0.013)	0.098 (0.026)	0.180 (0.034)	0.016(0.011)	0.058(0.020)	0.110(0.027)			

Tabla 3.3: Porcentajes de rechazo (entre paréntesis aparece 1.96 veces su desviación típica aproximada), con niveles de significación $\alpha = 0.01, 0.05, 0.1$, para los test de unimodalidad de: Silverman (1981), Fisher y Marron (2001), el Método 1 de Hall y York (2001) para los intervalos especificados en la columna I , Cheng y Hall (1998) y para la nueva propuesta presentada en la Sección 3.2.4 usando la ventana crítica y la ventana de Hall y York (2001) empleando los intervalos especificados en la columna I . Estos han sido obtenidos tras realizar 500 réplicas bootstrap en 500 muestras de tamaño $n = 50$, $n = 200$ y $n = 1000$ procedentes de las distribuciones: M1, M2, M3, M8 y M9.

		I	α	0.01	0.05	0.10	0.01	0.05	0.10	0.01	0.05	0.10
M4		[0,1]		Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
			n=50	0 (0)	0 (0)	0 (0)	0.158 (0.032)	0.668 (0.041)	0.868 (0.030)	0.002 (0.004)	0.096 (0.026)	0.550 (0.044)
			n=100	0 (0)	0.350 (0.042)	0.978 (0.013)	0.976 (0.013)	1 (0)	1 (0)	0.780 (0.036)	1 (0)	1 (0)
			n=200	0.914 (0.025)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
				Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
			n=50	0.012 (0.010)	0.052 (0.019)	0.090 (0.025)	0.016 (0.011)	0.060 (0.021)	0.116 (0.028)	0.016(0.011)	0.060(0.021)	0.116(0.028)
M5		[0,1]		Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
			n=50	0 (0)	0.004 (0.006)	0.050 (0.019)	0.086 (0.025)	0.256 (0.038)	0.406 (0.043)	0.006 (0.007)	0.108 (0.027)	0.308 (0.040)
			n=100	0.006 (0.007)	0.188 (0.034)	0.582 (0.043)	0.560 (0.044)	0.848 (0.031)	0.942 (0.020)	0.224 (0.037)	0.754 (0.038)	0.902 (0.026)
			n=200	0.070 (0.022)	0.720 (0.039)	0.936 (0.021)	0.924 (0.023)	0.984 (0.011)	0.994 (0.007)	0.768 (0.037)	0.972 (0.014)	0.986 (0.010)
				Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
			n=50	0.240 (0.037)	0.462 (0.044)	0.578 (0.043)	0.116 (0.028)	0.316 (0.041)	0.448 (0.044)	0.116(0.028)	0.316(0.041)	0.448(0.044)
M6		[0,1]		Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
			n=50	0 (0)	0.044 (0.018)	0.192 (0.035)	0.134 (0.030)	0.290 (0.040)	0.418 (0.043)	0.032 (0.015)	0.280 (0.039)	0.444 (0.044)
			n=100	0.004 (0.006)	0.120 (0.028)	0.298 (0.040)	0.188 (0.034)	0.414 (0.043)	0.536 (0.044)	0.120 (0.028)	0.386 (0.043)	0.560 (0.044)
			n=200	0.082 (0.024)	0.408 (0.043)	0.624 (0.042)	0.464 (0.044)	0.700 (0.040)	0.788 (0.036)	0.422 (0.043)	0.700 (0.040)	0.820 (0.034)
				Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
			n=50	0.230 (0.037)	0.398 (0.043)	0.536 (0.044)	0.112 (0.028)	0.254 (0.038)	0.360 (0.042)	0.102(0.027)	0.250(0.038)	0.360(0.042)
M7		[0,1]		Silverman (1981)			Fisher y Marron (2001)			Método 1 de Hall y York (2001)		
			n=50	0 (0)	0.012 (0.010)	0.302 (0.040)	0.192 (0.035)	0.628 (0.042)	0.844 (0.032)	0.058 (0.020)	0.648 (0.042)	0.856 (0.031)
			n=100	0.058 (0.020)	0.896 (0.027)	0.990 (0.009)	0.884 (0.028)	0.994 (0.007)	0.998 (0.004)	0.930 (0.022)	1 (0)	1 (0)
			n=200	0.978 (0.013)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
				Cheng y Hall (1998)			Nueva propuesta (ventana crítica)			Nueva propuesta (ventana de Hall y York (2001))		
			n=50	0.014 (0.010)	0.180 (0.034)	0.436 (0.043)	0.030 (0.015)	0.250 (0.038)	0.542 (0.044)	0.030(0.015)	0.254(0.038)	0.526(0.044)

Tabla 3.4: Porcentajes de rechazo (entre paréntesis aparece 1.96 veces su desviación típica aproximada), con niveles de significación $\alpha = 0.01, 0.05, 0.1$, para los test de unimodalidad de: Silverman (1981), Fisher y Marron (2001), el Método 1 de Hall y York (2001) para los intervalos especificados en la columna I , Cheng y Hall (1998) y para la nueva propuesta presentada en la Sección 3.2.4 usando la ventana crítica y la ventana de Hall y York (2001) empleando los intervalos especificados en la columna I . Estos han sido obtenidos tras realizar 500 réplicas bootstrap en 500 muestras de tamaño $n = 50$, $n = 100$ y $n = 200$ procedentes de las distribuciones bimodales: M4, M5, M6 y M7.

Capítulo 4

Test de multimodalidad para datos circulares

Una vez revisados los distintos test presentes en la literatura para el caso lineal es el momento de abordar el objetivo final de este TFM que es el de poder determinar si en la comarca de Vigo hay una única temporada de incendios o más de una. Para ello, en este capítulo se revisarán y se propondrán adaptaciones al caso circular de distintos test que permitan contrastar si la verdadera distribución circular asociada a una muestra aleatoria simple es unimodal. Con este fin, al igual que en el contexto lineal, en las dos primeras secciones se estudiarán dos grandes bloques de test, los basados en la concentración crítica y los basados en el exceso de masa. Todos estos test han sido programados en el software estadístico R Core Team (2013). En la última sección de este capítulo se presentará un estudio de simulación para comparar los distintos test analizados en las dos secciones anteriores de este capítulo.

4.1. Test basados en la concentración crítica

En esta sección, se analizarán los test basados en el parámetro de concentración crítica para realizar el contraste (1.7) cuando $k^c \in \mathbb{Z}^+$. Este parámetro de concentración crítica, tal y como se mencionó en la Introducción, es el ν más grande, de forma que la estimación de la función de densidad circular dada en (1.5) posee a lo sumo k^c modas en el intervalo $[0, 2\pi)$. Esto es, se define la concentración crítica como:

$$\nu_{k^c} = \sup\{\nu : \hat{f}_\nu^c \text{ tiene a lo sumo } k^c \text{ modas en el intervalo } [0, 2\pi)\}. \quad (4.1)$$

Aunque se puede usar el parámetro ν_{k^c} como estadístico para realizar el contraste dado en (1.7) cuando $k^c \in \mathbb{Z}^+$, extendiendo las ideas de Hall y York (2001) solo se estudiará el caso $k^c = 1$.

También se revisará la adaptación al caso circular del test de Fisher y Marron (2001), donde, en lugar de emplear el estadístico de Cramér–von Mises, se empleará un estadístico similar, en el sentido de que permite contrastar si la verdadera distribución circular es igual a una dada (el estadístico U^2 propuesto por Watson, 1961).

4.1.1. Nueva propuesta: Test basado en la concentración crítica como estadístico

En el caso circular al igual que ocurría en el caso lineal, se tiene que el número de modas empleando \hat{f}_ν^c va a depender del parámetro de concentración ν que se utilice. Mencionar que, aunque se tiene que valores altos (bajos) de ν proporcionan estimaciones de la densidad con más (menos) modas, la monotonía en el número de modas en función de ν solo está garantizada en el caso de emplear el núcleo de la normal enrollada (véase Huckemann et al, 2014)¹. La idea será la de emplear el parámetro de concentración crítica dado en (4.2) como estadístico de contraste para (1.7). Si bien, ante el mal calibrado que se vio en la Sección 3.3 para el test de Silverman (1981) y sabiendo que en el caso circular se trabaja siempre en el soporte acotado $[0, 2\pi)$ parece más apropiado extender las ideas de Hall y York (2001) a este contexto.

Teniendo en cuenta que, en este caso, se rechazará H_0 para valores bajos del parámetro de concentración, para realizar el contraste (1.7) cuando $k^c = 1$, en el presente trabajo, se propone realizar el test que, para un nivel de significación α , rechace H_0 si se verifica que $\mathbb{P}(\nu_1^* \geq \lambda_\alpha^c \nu_1 | \Theta) \geq (1 - \alpha)$, donde ν_1^* denotará la versión bootstrap de la ν_1 dada en (4.2), esto es:

$$\nu_1^* = \sup\{\nu : \hat{f}_\nu^{c*} \text{ tiene una moda en el intervalo } [0, 2\pi)\}, \quad (4.2)$$

siendo \hat{f}_ν^{c*} la estimación tipo núcleo circular dada en (1.5) para una muestra bootstrap $\Theta^* = (\Theta_1^*, \dots, \Theta_n^*)$ generada de la distribución con densidad $\hat{f}_{\nu_1}^c$.

Para poder determinar los valores de los λ_α^c , el método que se propone es el de aproximar su distribución por técnicas de Monte Carlo, de forma análoga a la que se realizaba en una de las propuestas de Hall y York (2001) para el caso lineal (Sección 3.1.2). Para ello, será necesario encontrar el τ_α^c , de forma que para la regla que rechaza H_0 si $\mathbb{P}(\nu_1^* \geq \nu_1 | \Theta) \geq 1 - \tau_\alpha^c$, se tenga que el porcentaje de rechazos de H_0 sea α cuando se toman muestras de tamaño suficientemente grande. De esta forma, tomando, por ejemplo, 1000 muestras de tamaño $n = 1000$ procedentes de una distribución $vM(0, 1)$ y realizando en cada una de ellas 500 réplicas bootstrap se han obtenido los valores de τ_α^c que se muestran en la Tabla 4.1 para niveles de significación α tomando valores entre 0.01 y 0.95.

¹Aunque la monotonía no está garantizada cuando se emplea el núcleo von Mises dado en (1.6), por motivos computacionales, para calcular ν_{k^c} se realizará un algoritmo de búsqueda dicotómica empleando este núcleo.

α	τ_α^c	α	τ_α^c	α	τ_α^c	α	τ_α^c
0.01	0.002	0.11	0.050	0.21	0.112	0.35	0.264
0.02	0.002	0.12	0.064	0.22	0.124	0.40	0.306
0.03	0.002	0.13	0.066	0.23	0.134	0.45	0.352
0.04	0.008	0.14	0.080	0.24	0.144	0.50	0.412
0.05	0.008	0.15	0.084	0.25	0.152	0.55	0.448
0.06	0.016	0.16	0.088	0.26	0.156	0.60	0.514
0.07	0.022	0.17	0.096	0.27	0.166	0.65	0.584
0.08	0.028	0.18	0.104	0.28	0.178	0.75	0.678
0.09	0.038	0.19	0.106	0.29	0.188	0.85	0.822
0.10	0.042	0.20	0.112	0.30	0.194	0.95	0.940

Tabla 4.1: Aproximación de los valores de τ_α^c para distintos niveles de significación α . Obtenidos tras realizar 1000 réplicas en cada una de las 1000 muestras de tamaño $n = 1000$ procedentes de una distribución $vM(0, 1)$.

4.1.2. Test basado en el estadístico U^2 de Watson

Fisher y Marron (2001) proponen en su artículo la extensión al caso circular del test basado en el estadístico de Cramér–von Mises para contrastar la hipótesis nula de que la verdadera distribución posee a lo sumo k^c modas frente a la hipótesis alternativa de que posee más de k^c modas en el círculo.

Análogamente a lo que ocurría con el criterio de Cramér–von Mises, para una muestra $\Theta = (\Theta_1, \dots, \Theta_n)$, el criterio de Watson (1961) permite realizar el contraste de hipótesis:

$$H_0 : F(\theta) = F_0(\theta) \forall \theta \in [0, 2\pi), \text{ frente a } H_1 : F(\theta) \neq F_0(\theta) \text{ para algún } \theta \in [0, 2\pi), \quad (4.3)$$

donde F_0 es una función de distribución circular dada. El estadístico U^2 de Watson (1961) para el contraste (4.3) viene dado a través de la siguiente expresión:

$$U^2 = \sum_{i=1}^n \left(u_{(i)} - \left(\bar{u} - \frac{1}{2} \right) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \quad (4.4)$$

donde $u_{(i)}$ son los valores ordenados de $u_i = F_0(\Theta_i)$, con $i \in \{1, \dots, n\}$ y \bar{u} su promedio.

Para poder obtener los valores de u_i cuando se quiere contrastar que la verdadera distribución posee a lo sumo k^c modas, se propone tomar $F_0 = \hat{F}_{\nu_{k^c}}^c$, donde la estimación de la función de

²Aunque aquí se ha decidido tomar como parámetro de concentración ν_{k^c} para la estimación de la función de distribución circular por seguir las ideas que presentan Fisher y Marron (2001) en el caso lineal, en su artículo original no se especifica nada acerca de como tomar el parámetro de concentración para obtener F_0 .

distribución circular es la función verificando que $\hat{F}_{\nu_{k^c}}^c(\phi) = \int_0^\phi \hat{f}_{\nu_{k^c}}^c(\theta) d\theta$, y además, $\hat{F}_{\nu_{k^c}}^c(\phi + 2\pi) - \hat{F}_{\nu_{k^c}}^c(\phi) = 1, \forall \phi \in \mathbb{R}$. Siendo $\hat{f}_{\nu_{k^c}}^c$ la estimación de la función de densidad circular con parámetro de concentración ν_{k^c} .

Así, tomando $F_0 = \hat{F}_{\nu_{k^c}}^c$, se tiene que la expresión del estadístico U^2 de Watson (1961) dado en (4.4) para este caso es:

$$U_{k^c}^2 = \sum_{i=1}^n \left(u_{(i)} - \left(\bar{u} - \frac{1}{2} \right) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}, \quad (4.5)$$

donde $u_{(i)}$ son los valores ordenados de $u_i = \hat{F}_{\nu_{k^c}}^c(\Theta_i)$, con $i \in \{1, \dots, n\}$.

Para obtener la distribución del estadístico $U_{k^c}^2$ bajo H_0 se propone usar la metodología bootstrap³. Para un nivel de significación α , se rechazará H_0 si $\mathbb{P}(U_{k^c}^{2*} \leq U_{k^c}^2 | \Theta) \geq 1 - \alpha$, donde $U_{k^c}^{2*}$ denota el valor que toma la U^2 de Watson (1961) dada en (4.5) para la muestra bootstrap $\Theta^* = (\Theta_1^*, \dots, \Theta_n^*)$ generada a partir de la distribución que tiene como función de densidad $\hat{f}_{\nu_{k^c}}^c$.

4.2. Nueva propuesta: Test basado en el exceso de masa

En este caso, la idea es la de extender el test basado en el exceso de masa de Müller y Sawitzki (1991) que se presentó en la Sección 3.2.4 para el caso circular. Para ello, será preciso saber que es un λ -conglomerado en el caso circular. Estos serán todos los intervalos cerrados disjuntos $[\theta_1, \theta_2] \subset [0, 2\pi)$ verificando las siguientes condiciones para $0 < t < t_0$, con t_0 suficientemente pequeño:

- $f^c(\theta) \geq \lambda, \forall \theta \in [\theta_1, \theta_2]$.
- $\lim_{t \rightarrow 0} f^c(\theta_1 - t) < \lambda$.
- $\lim_{t \rightarrow 0} f^c(\theta_2 + t) < \lambda$.

Además, por la periodicidad de f^c también será λ -conglomerado el conjunto $\{[0, \theta_3] \cup [\theta_4, 2\pi)\}$ si, para $0 < t < t_0$, con t_0 suficientemente pequeño, se verifica que:

- $0 \leq \theta_3 \leq \theta_4 < 2\pi$.
- $f^c(\theta) \geq \lambda, \forall \theta \in \{[0, \theta_3] \cup [\theta_4, 2\pi)\}$.
- $\lim_{t \rightarrow 0} f^c(\theta_4 - t) < \lambda$.
- $\lim_{t \rightarrow 0} f^c(\theta_3 + t) < \lambda$.

³De nuevo, aunque en su artículo no se especifica nada de como obtener la distribución del estadístico, se están siguiendo las ideas que emplean Fisher y Marron (2001) para el caso lineal.

De nuevo, si una distribución circular posee j^c modas en el intervalo $[0, 2\pi)$, entonces tiene j^c λ -conglomerados, denotando a cada uno de ellos como $C_1^c, \dots, C_{j^c}^c$, acudiendo a (3.17) se tiene que la función exceso de masa circular para j^c modas es:

$$E_{j^c}^c(f, \lambda) = \sup_{C_1^c, \dots, C_{j^c}^c} \left\{ \sum_{i=1}^{j^c} \mathbb{P}(C_i^c) - \lambda \|C_i^c\| \right\},$$

donde $\|C_i^c\|$ toma el valor $(\theta_2 - \theta_1)$ para los C_i^c de la forma $[\theta_1, \theta_2]$ y el valor $(2\pi - \theta_4 + \theta_3)$ para el λ -conglomerado $\{[0, \theta_3] \cup [\theta_4, 2\pi)\}$.

El exceso de masa empírico dado en (3.18) para una muestra $\Theta = (\Theta_1, \dots, \Theta_n)$, vendría dado de la siguiente forma:

$$E_{n, k^c}^c(\mathbb{P}_n^c, \lambda) = \sup_{C_1, \dots, C_{k^c}} \left\{ \sum_{l=1}^{k^c} \mathbb{P}_n^c(C_l^c) - \lambda \|C_l^c\| \right\},$$

siendo k^c un número máximo de modas en $[0, 2\pi)$ prefijado y con $\mathbb{P}_n^c(C_l^c) = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(\Theta_i \in C_l^c)$.

De nuevo, con el fin de obtener un estadístico para realizar el contraste (1.7) cuando $k^c \in \mathbb{Z}^+$, dado un valor de λ , se define como $D_{n, k^c+1}^c(\lambda) = E_{n, k^c+1}^c(\mathbb{P}_n^c, \lambda) - E_{n, k^c}^c(\mathbb{P}_n^c, \lambda)$. El estadístico que se empleará para este contraste será:

$$\Delta_{n, k^c+1}^c = \max_{\lambda} \{D_{n, k^c+1}^c(\lambda)\}. \quad (4.6)$$

Como el test de Müller y Sawitzki (1991) proporcionaba un calibrado demasiado conservador en el caso lineal, lo que se hará es extender la nueva propuesta presentada en el caso lineal (Sección 3.2.4) para el contraste de unimodalidad frente a multimodalidad al caso circular. Para ello, se realizarán los siguientes pasos:

1. Dada la muestra $\Theta = (\Theta_1, \dots, \Theta_n)$, obtener el valor del estadístico $\Delta_{n, 2}^c$ a partir de la ecuación (4.6) cuando $k^c = 1$.
2. Calcular el parámetro de concentración crítica dado en (4.2) cuando $k^c = 1$. Se denotará a esta concentración como ν_1 .
3. A partir de este ν_1 se tomará como función de densidad, bajo la hipótesis nula, la estimación tipo núcleo $\hat{f}_{\nu_1}^c$ para generar B réplicas bootstrap $\Theta^* = (\Theta_1^*, \dots, \Theta_n^*)$.
4. A partir de cada una de las B réplicas bootstrap Θ^{*b} , con $b = 1, \dots, B$; se generan los estadísticos basados en el exceso de masa asociados a cada una de estas réplicas, $\Delta_{n, 2}^{c*}$.
5. Para un nivel de significación α se rechazará la hipótesis nula si $\mathbb{P}(\Delta_{n, 2}^{c*} \leq \Delta_{n, 2}^c | \Theta) \geq 1 - \alpha$.

4.3. Estudio de simulación para datos circulares

A continuación se presenta un estudio de simulación para los test descritos en este capítulo, los cuales han sido programados en el software estadístico R Core Team (2013). En todos los estudios realizados, se han tomado los niveles de significación $\alpha = 0.01, 0.05, 0.1$, tamaños muestrales $n = 50$, $n = 200$ y $n = 1000$ ($n = 100$ en lugar de $n = 1000$ en los estudios de potencia), se han usado 500 muestras distintas de cada distribución y se han realizado $B = 500$ réplicas bootstrap para obtener el p-valor asociado a cada muestra.

Con el fin de analizar el calibrado de los test presentados, en las Tablas 4.2 y 4.3, para el contraste dado en (1.7) cuando $k^c = 1$, se muestra el porcentaje de rechazos de los test de Fisher y Marron (2001) (Sección 4.1.2), de la nueva propuesta basada en la concentración crítica (Sección 4.1.1) y de la nueva propuesta basada en el exceso de masa (Sección 4.2), en las siguientes distribuciones circulares unimodales (para obtener sus expresiones véase Jammalamadaka y Sengupta (2001), Cap. 2): $vM(\pi, 10)$, $WN(\pi, 0.9)$, $WC(\pi, 0.8)$, $C(\pi, 0.5)$, Triangular(0.3) y las mixturas de von Mises M11 y M12 descritas en el Apéndice A.

En la Tabla 4.4 se ha realizado un estudio de potencia de los distintos test. Se ha calculado el porcentaje de rechazos para las siguientes distribuciones multimodales (véase Apéndice A para obtener sus expresiones): M13 (bimodal), M14 (bimodal), M16 (bimodal) y M15 (trimodal).

En las Tablas 4.2 y 4.3, se puede ver que el test basado en la U^2 de Watson propuesto por Fisher y Marron (2001) no está correctamente calibrado ya que, al igual que ocurría en el caso lineal con el test basado en el estadístico de Cramér-von Mises, se observa que en ciertos modelos, como son la $WC(\pi, 0.8)$ o la mixtura M12, el porcentaje de rechazos está muy por encima del nivel de significación nominal incluso para tamaño muestral $n = 1000$. Mientras que en otros modelos, como son la $WN(\pi, 0.9)$, la $C(\pi, 0.5)$ o el Triangular(0.3), el porcentaje de rechazos está por debajo del nivel teórico.

El test basado en la concentración crítica presentado en la Sección 4.1.1, tampoco presenta un buen calibrado, especialmente para tamaños muestrales bajos ($n = 50$ y $n = 200$). Para $n = 1000$ el calibrado de este test mejora, pero aún así, se tiene que tanto en la $WC(\pi, 0.8)$ como en la mixtura M12 el porcentaje de rechazos está significativamente (a un nivel del 5%) por encima del nivel nominal. Además, para la $WN(\pi, 0.9)$ y la Triangular(0.3) este porcentaje está significativamente por debajo del nivel teórico.

Finalmente, para la nueva propuesta basada en el exceso de masa presentada en la Sección 4.2, se puede observar que, incluso para tamaños muestrales bajos ($n = 50$) el calibrado es bastante bueno, obteniéndose un test ligeramente conservador, ya que para algunos modelos, como son la $WN(\pi, 0.9)$, la $WC(\pi, 0.8)$, el Triangular(0.3) o la mixtura M11 el porcentaje de rechazos está ligeramente por debajo del nivel nominal. El principal inconveniente en el calibrado es que no se logra mejorar, en el sentido de que el porcentaje de rechazos se vaya aproximando cada vez más al nivel teórico α , a medida que se aumenta el tamaño muestral.

Consecuentemente, como solo la nueva propuesta basada en el exceso de masa presenta un calibrado bastante bueno, la recomendación para realizar el contraste (1.7) cuando $k^c = 1$ es la de emplear la propuesta presentada en la Sección 4.2.

En cuanto a la potencia, se puede observar en la Tabla 4.4, que la nueva propuesta basada en el exceso de masa también funciona bien, en el sentido de que esta mejora con el tamaño muestral aproximándose a uno cuando $n = 200$. El único caso en el que se han encontrado problemas, en cuanto a potencia, es en el modelo M16 ya que para $n = 50$ no es capaz de detectar que esta distribución es unimodal.

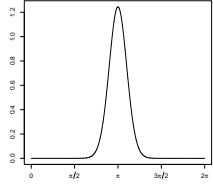
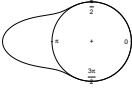
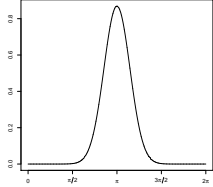
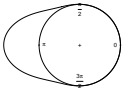
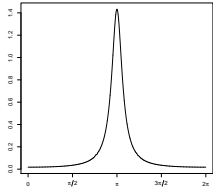
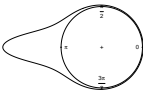
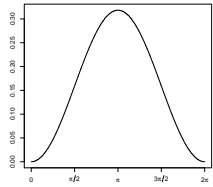

			α	0.01	0.05	0.10
$vM(\pi, 10)$				Concentración crítica		
			n=50	0(0)	0(0)	0(0)
			n=200	0(0)	0(0)	0.004(0.006)
			n=1000	0.038(0.017)	0.094(0.026)	0.108(0.027)
				U^2 de Watson		
			n=50	0.002(0.004)	0.016(0.011)	0.066(0.022)
			n=200	0(0)	0.016(0.011)	0.052(0.019)
			n=1000	0.016(0.011)	0.040(0.017)	0.078(0.024)
				Exceso de masa		
			n=50	0.008(0.008)	0.032(0.015)	0.084(0.024)
			n=200	0.002(0.004)	0.024(0.013)	0.092(0.025)
			n=1000	0.012(0.010)	0.044(0.018)	0.086(0.025)
$WN(\pi, 0.9)$				Concentración crítica		
			n=50	0(0)	0(0)	0.004(0.006)
			n=200	0(0)	0.004(0.006)	0.050(0.019)
			n=1000	0(0)	0.032(0.015)	0.056(0.020)
				U^2 de Watson		
			n=50	0.004(0.006)	0.022(0.013)	0.062(0.021)
			n=200	0(0)	0.018(0.012)	0.050(0.019)
			n=1000	0(0)	0.012(0.010)	0.046(0.018)
				Exceso de masa		
			n=50	0.004(0.006)	0.048(0.019)	0.106(0.027)
			n=200	0.010(0.009)	0.032(0.015)	0.068(0.022)
			n=1000	0.002(0.004)	0.026(0.014)	0.062(0.021)
$WC(\pi, 0.8)$				Concentración crítica		
			n=50	0.052(0.019)	0.176(0.033)	0.288(0.040)
			n=200	0.076(0.023)	0.176(0.033)	0.306(0.040)
			n=1000	0.092(0.025)	0.176(0.033)	0.248(0.038)
				U^2 de Watson		
			n=50	0.648(0.042)	0.814(0.034)	0.880(0.028)
			n=200	0.696(0.040)	0.840(0.032)	0.886(0.028)
			n=1000	0.568(0.043)	0.708(0.040)	0.798(0.035)
				Exceso de masa		
			n=50	0.006(0.007)	0.028(0.014)	0.046(0.018)
			n=200	0(0)	0.012(0.010)	0.036(0.016)
			n=1000	0.002(0.004)	0.018(0.012)	0.042(0.018)
$C(\pi, 0.5)$				Concentración crítica		
			n=50	0(0)	0(0)	0.038(0.017)
			n=200	0(0)	0.030(0.015)	0.070(0.022)
			n=1000	0.030(0.015)	0.050(0.019)	0.088(0.025)
				U^2 de Watson		
			n=50	0.006(0.007)	0.020(0.012)	0.040(0.017)
			n=200	0.002(0.004)	0.020(0.012)	0.042(0.018)
			n=1000	0.006(0.007)	0.024(0.013)	0.048(0.019)
				Exceso de masa		
			n=50	0.004(0.006)	0.042(0.018)	0.088(0.025)
			n=200	0.008(0.008)	0.042(0.018)	0.080(0.024)
			n=1000	0.004(0.006)	0.042(0.018)	0.078(0.024)

Tabla 4.2: Porcentajes de rechazo (entre paréntesis aparece 1.96 veces su desviación típica aproximada), con niveles de significación $\alpha = 0.01, 0.05, 0.1$, para el test basado en el estadístico U^2 de Watson y para las nuevas propuestas presentadas basadas en la concentración crítica y en el exceso de masa. Estos han sido obtenidos tras realizar 500 réplicas bootstrap en 500 muestras de tamaño $n = 50$, $n = 200$ y $n = 1000$ procedentes de las distribuciones: $vM(0, 10)$, $WN(\pi, 0.9)$, $WC(\pi, 0.8)$ y $C(0, 0.5)$.

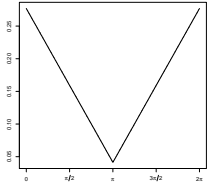
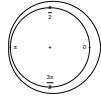
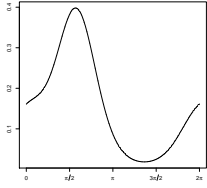
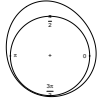
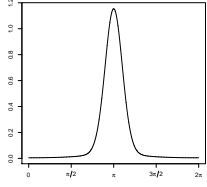
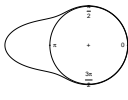
			α	0.01	0.05	0.10
Triangular(0.3)			Concentración crítica			
			n=50	0(0)	0(0)	0.030(0.015)
			n=200	0(0)	0(0)	0.004(0.006)
			n=1000	0(0)	0.030(0.015)	0.054(0.020)
			U^2 de Watson			
			n=50	0.012(0.010)	0.036(0.016)	0.066(0.022)
			n=200	0(0)	0.008(0.008)	0.024(0.013)
			n=1000	0(0)	0.010(0.009)	0.028(0.014)
			Exceso de masa			
n=50	0.010(0.009)	0.034(0.016)	0.062(0.021)			
n=200	0(0)	0.020(0.012)	0.064(0.021)			
n=1000	0.004(0.006)	0.028(0.014)	0.056(0.020)			
M11			Concentración crítica			
			n=50	0(0)	0(0)	0.030(0.015)
			n=200	0(0)	0.004(0.006)	0.070(0.022)
			n=1000	0(0)	0.054(0.020)	0.108(0.027)
			U^2 de Watson			
			n=50	0.012(0.010)	0.034(0.016)	0.092(0.025)
			n=200	0.006(0.007)	0.040(0.017)	0.114(0.028)
			n=1000	0.004(0.006)	0.036(0.016)	0.092(0.025)
			Exceso de masa			
n=50	0.004(0.006)	0.024(0.013)	0.074(0.023)			
n=200	0.002(0.004)	0.030(0.015)	0.072(0.023)			
n=1000	0(0)	0.034(0.016)	0.062(0.021)			
M12			Concentración crítica			
			n=50	0.062(0.021)	0.140(0.030)	0.234(0.037)
			n=200	0.114(0.028)	0.234(0.037)	0.336(0.041)
			n=1000	0.132(0.010)	0.232(0.037)	0.314(0.041)
			U^2 de Watson			
			n=50	0.350(0.042)	0.512(0.044)	0.604(0.043)
			n=200	0.636(0.042)	0.772(0.037)	0.854(0.031)
			n=1000	0.516(0.044)	0.642(0.042)	0.710(0.040)
			Exceso de masa			
n=50	0.010(0.009)	0.044(0.018)	0.102(0.027)			
n=200	0.012(0.010)	0.042(0.018)	0.080(0.024)			
n=1000	0.010(0.009)	0.042(0.018)	0.080(0.024)			

Tabla 4.3: Porcentajes de rechazo (entre paréntesis aparece 1.96 veces su desviación típica aproximada), con niveles de significación $\alpha = 0.01, 0.05, 0.1$, para el test basado en el estadístico U^2 de Watson y para las nuevas propuestas presentadas basadas en la concentración crítica y en el exceso de masa. Estos han sido obtenidos tras realizar 500 réplicas bootstrap en 500 muestras de tamaño $n = 50, n = 200$ y $n = 1000$ procedentes de las distribuciones: Triangular(0.3), M11 y M12.

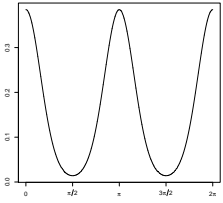

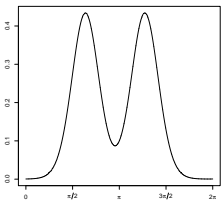
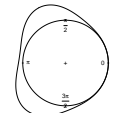
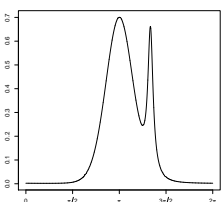
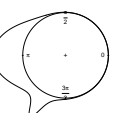
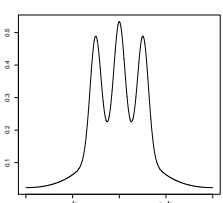
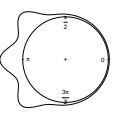
			α	0.01	0.05	0.10
M13				Concentración crítica		
			n=50	0.272(0.039)	0.668(0.041)	0.866(0.030)
			n=100	0.396(0.043)	0.860(0.030)	0.986(0.010)
			n=200	0.590(0.043)	0.954(0.018)	0.998(0.004)
				U^2 de Watson		
			n=50	1(0)	1(0)	1(0)
			n=100	1(0)	1(0)	1(0)
			n=200	1(0)	1(0)	1(0)
				Exceso de masa		
n=50	0.952(0.019)	0.990(0.009)	0.998(0.004)			
n=100	1(0)	1(0)	1(0)			
n=200	1(0)	1(0)	1(0)			
M14				Concentración crítica		
			n=50	0(0)	0.104(0.027)	0.512(0.044)
			n=100	0.070(0.022)	0.826(0.033)	0.964(0.016)
			n=200	0.748(0.038)	0.988(0.010)	0.998(0.004)
				U^2 de Watson		
			n=50	0.776(0.037)	0.920(0.024)	0.954(0.018)
			n=100	0.966(0.016)	0.996(0.006)	0.998(0.004)
			n=200	1(0)	1(0)	1(0)
				Exceso de masa		
n=50	0.566(0.043)	0.758(0.038)	0.860(0.030)			
n=100	0.912(0.025)	0.976(0.013)	0.984(0.011)			
n=200	0.998(0.004)	1(0)	1(0)			
M16				Concentración crítica		
			n=50	0.038(0.017)	0.094(0.026)	0.168(0.033)
			n=100	0.062(0.021)	0.132(0.030)	0.226(0.037)
			n=200	0.116(0.028)	0.284(0.040)	0.418(0.043)
				U^2 de Watson		
			n=50	0.090(0.025)	0.220(0.036)	0.340(0.042)
			n=100	0.134(0.030)	0.330(0.041)	0.510(0.044)
			n=200	0.284(0.040)	0.536(0.044)	0.690(0.041)
				Exceso de masa		
n=50	0.010(0.009)	0.054(0.020)	0.122(0.029)			
n=100	0.022(0.013)	0.122(0.029)	0.232(0.037)			
n=200	0.240(0.037)	0.478(0.044)	0.614(0.043)			
M15				Concentración crítica		
			n=50	0(0)	0.032(0.015)	0.094(0.026)
			n=100	0.030(0.015)	0.106(0.027)	0.182(0.034)
			n=200	0.050(0.019)	0.128(0.029)	0.218(0.036)
				U^2 de Watson		
			n=50	0.020(0.012)	0.120(0.028)	0.248(0.038)
			n=100	0.064(0.021)	0.248(0.038)	0.448(0.044)
			n=200	0.148(0.031)	0.430(0.043)	0.612(0.043)
				Exceso de masa		
n=50	0.062(0.021)	0.184(0.034)	0.294(0.040)			
n=100	0.112(0.028)	0.372(0.042)	0.542(0.044)			
n=200	0.404(0.043)	0.734(0.039)	0.844(0.032)			

Tabla 4.4: Porcentajes de rechazo (entre paréntesis aparece 1.96 veces su desviación típica aproximada), con niveles de significación $\alpha = 0.01, 0.05, 0.1$, para el test basado en el estadístico U^2 de Watson y para las nuevas propuestas presentadas basadas en la concentración crítica y en el exceso de masa. Estos han sido obtenidos tras realizar 500 réplicas bootstrap en 500 muestras de tamaño $n = 50, n = 100$ y $n = 200$ procedentes de las distribuciones: M13 (bimodal), M14 (bimodal), M16 (bimodal) y M15 (trimodal).

Capítulo 5

Aplicaciones a datos reales

A lo largo de este capítulo se analizarán distintos casos reales en los cuales es de interés determinar si los datos presentan una estructura unimodal o multimodal. Aunque el ejemplo final, que ha motivado gran parte de este estudio, es el de determinar si solo hay una temporada de incendios o más de una, se han encontrado numerosos ejemplos en los que determinar si los datos presentan una estructura unimodal o multimodal es de gran interés. A continuación, se estudiarán algunos de estos casos.

En primer lugar, en el contexto lineal, se analizará el tiempo que transcurre entre el comienzo de dos erupciones del géiser Old Faithful situado en el Parque Nacional de Yellowstone en Wyoming, USA, con el objetivo de determinar si solo hay una duración principal, donde lo más probable es que todos los tiempos de espera se agrupen en torno a los 80 minutos o si en cambio, estos tiempos se agrupan en torno a, al menos, dos valores distintos.

En segundo lugar, también en el contexto lineal, se estudiará una muestra de los ingresos por hogar en Galicia. El objetivo será el de determinar si las rentas se agrupan en torno a un valor o si, en cambio, estas se agrupan, al menos, en torno a dos valores y, por tanto, ninguno de los modelos paramétricos unimodales clásicos empleados en el campo de la economía son válidos para modelizar la situación actual de la renta en Galicia.

Como tercer ejemplo, ya en el contexto circular, se estudiará la dirección media de la que procede el viento, dada por un medidor situado en la localidad de Vilalba, cada vez que hay un exceso de NO_x en la atmósfera, con el fin de determinar si todas las alarmas que se producen por exceso de este contaminante podrían ser relacionadas con la actividad de la central térmica de As Pontes de García Rodríguez.

Como ejemplo final, se analiza la muestra que ha motivado gran parte de este TFM, que es la de la fecha en la que se producen los incendios en la comarca de Vigo, con el fin de determinar si en esta región solo hay una temporada de incendios o más de una.

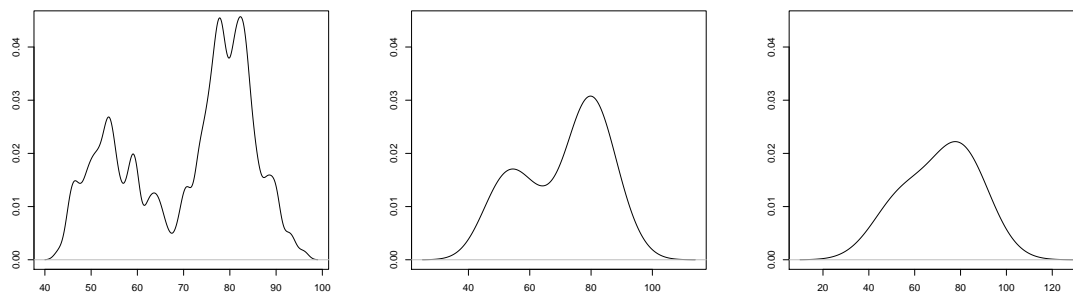


Figura 5.1: Estimación de la densidad con núcleo gaussiano empleando la muestra del tiempo que transcurre, en minutos, entre el comienzo de dos erupciones del géiser Old Faithful. Izquierda: $h = 1$. Centro: $h = 6$. Derecha: $h = 11$.

5.1. Duración entre dos erupciones del géiser Old Faithful

El primer conjunto de datos que se estudiará es el de la muestra del tiempo que transcurre, en minutos, entre el comienzo de dos erupciones del géiser Old Faithful situado en el Parque Nacional de Yellowstone en Wyoming, USA. Esta muestra de 272 observaciones se puede encontrar en el paquete `base` de R Core Team (2013) y es parte de la muestra de 299 observaciones, recogidas entre el 1 y el 15 de Agosto de 1985, tratadas en detalle por Azzalini y Bowman (1990).

Este conjunto de datos ha sido empleado tradicionalmente para mostrar un ejemplo donde la estadística no paramétrica es de utilidad, ya que parece que ninguna de las distribuciones paramétricas clásicas, como puede ser la normal, sirven para modelizar estos tiempos de espera. En la Figura 5.1 se muestra la estimación de la densidad, con núcleo gaussiano, obtenida a partir de la muestra del tiempo que transcurre, en minutos, entre el comienzo de dos erupciones en el géiser Old Faithful. En el Capítulo 1 se mostró la aplicación de distintas herramientas gráficas sobre esta base de datos (véanse Figuras 1.1, 1.5, 1.6 y 1.7).

Observando tanto las estimaciones tipo núcleo, como las distintas herramientas exploratorias presentadas en el Capítulo 1 parece claro que hay una moda en torno a los 80 minutos. Lo que no parece concluyente es si podría haber otro máximo relativo en torno al valor 55.

Para poder responder si todos los tiempos se agrupan en torno a una duración o si, en cambio, se agrupan al menos en torno a dos y por tanto, los modelos paramétricos unimodales clásicos no son correctos para modelizar esta variable, se puede hacer uso de los test vistos en el Capítulo 3 de este TFM. El problema que surge a la hora de realizar el test de Hall y York (2001) es que al desconocer en qué intervalo se sitúan los tiempos transcurridos entre dos erupciones, no se puede establecer cuál es el intervalo cerrado I empleado en el cálculo de la ventana dada en (3.3). Con el fin de obtener un p-valor para el Método 1 de Hall y York (2001) se ha escogido el intervalo $[45, 90]$, ya que este

posee todos los valores obtenidos de la muestra.

Se han calculado los p -valores obtenidos con los test de Silverman (1981) (Sección 3.1.1), el Método 1 de Hall y York (2001) (Sección 3.1.2)¹, la versión sencilla de Fisher y Marron (2001) (Sección 3.1.3), el de Cheng y Hall (1998) (Sección 3.2.3) y la nueva propuesta presentada para este trabajo utilizando la ventana crítica (Sección 3.2.4). En todos los casos el p -valor obtenido ha sido 0, excepto para el test de Silverman (1981) donde el p -valor asociado a este contraste es 0.004. A la vista de estos resultados, con todos los test estudiados, incluso con el test conservador de Silverman (1981), para un nivel de significación del 5%, se rechazaría la hipótesis de que la muestra procede de una distribución unimodal. Cabe destacar que el SiZer, ver Figura 1.6, no detecta esta segunda moda.

5.2. La renta por hogar en Galicia

En el campo de la economía es habitual querer modelizar la distribución de la renta en una región durante un período de tiempo. Para ello, se han considerado numerosas distribuciones paramétricas entre las que estarían la lognormal, la logt de Student, la gamma, la Singh–Maddala o la Dagum Tipo I entre otras (véase Kleiber y Kotz (2003) para obtener sus expresiones). Así, por ejemplo, en Bandourian *et al.* (2002) se realiza un estudio para ver cuál de entre estas y otras distribuciones paramétricas se ajusta mejor a la renta por vivienda de 23 países durante 5 periodos de tiempo comprendidos ente el año 1971 y el 1997. En Díez (2004) se realiza un estudio para ver cuál de estas distribuciones se adapta mejor a cada una de las 17 comunidades autónomas de España para los ingresos por hogar con los datos recogidos entre Abril de 1990 y Marzo de 1991 para la Encuesta de Presupuestos Familiares. En este caso, se llega a la conclusión de que las distribuciones que se podían emplear para modelizar las rentas en Galicia durante este periodo era una logt de Student, una Singh–Maddala o una Dagum Tipo I.

El inconveniente de todos estos modelos paramétricos es que no admiten más de una moda, con lo que se está asumiendo siempre que todas las rentas por hogar se agrupan en torno a un valor. El objetivo de este estudio sería el de determinar si en Galicia, tras la crisis, todas las rentas por hogar se siguen situando en torno a un valor o si, en cambio, estas se agrupan en torno a más de una cantidad y las distribuciones paramétricas empleadas hasta el momento para modelizar la distribución de la renta por hogar no son válidas.

Para determinar si durante el año 2012 en Galicia hubo solo una tendencia de ingresos o más de una, se analizarán los datos obtenidos por el Instituto Galego de Estatística (2012) en la Encuesta de Condiciones de Vida de las Familias. En esta encuesta, entre otros datos, se recoge la renta anual que tiene cada uno de los individuos que conforma un hogar en función de la fuente de ingresos (por trabajo por cuenta ajena o propia, por prestaciones, por subsidios de desempleo, por rentas o por otra fuente de ingreso). A partir de esta encuesta, se han extraído las rentas anuales, independientemente de la fuente de ingresos, de 9190 viviendas gallegas.

Con el objetivo de imitar los estudios realizados para modelizar la distribución de la renta por

¹En el caso del Método 1 de Hall y York (2001), tanto en esta base de datos como en la empleada en la Sección 5.2, se ha calculado el porcentaje de veces que se tiene $(\lambda_{0.05} h_{HY} < h_{HY}^*)$.

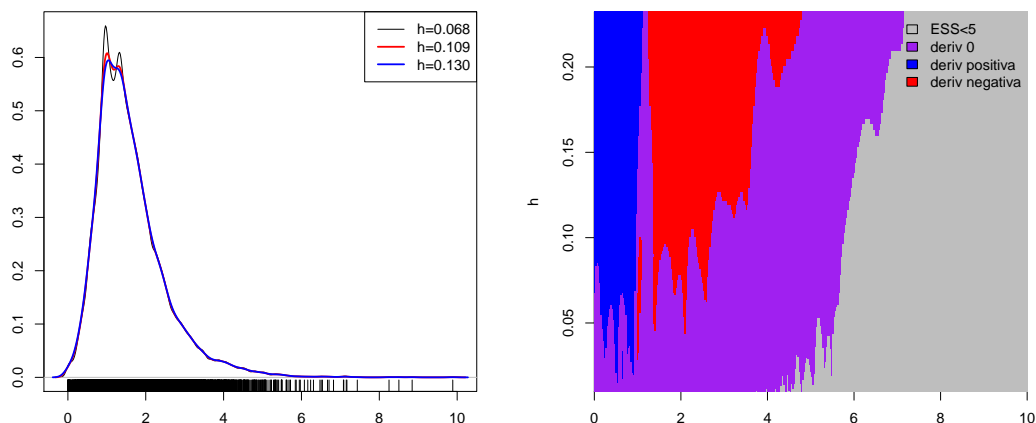


Figura 5.2: Izquierda: estimación de la densidad con núcleo gaussiano y distintas ventanas obtenidas a través de selectores automáticos (véase Wand y Jones (1995), Cap. 3): $h = 0.068$ (regla plug-in), $h = 0.109$ (regla del pulgar) y $h = 0.130$ (validación cruzada). Derecha: SiZer empleando parámetros de ventana h entre 0.01 y 0.23 (se presenta en el eje vertical). Ambos empleando la muestra del ingreso equivalente per capita normalizado por hogar en Galicia.

hogar (véase Díez, 2004), se eliminarán a aquellas viviendas con ingresos nulos (en este caso no hay ningún hogar en esta situación) y se empleará el ingreso equivalente per capita normalizado por hogar, esto es, si X_i es la renta del hogar i -ésimo, se tomará la muestra formada por los $\tilde{X}_i = X_i/(\sqrt{N_i}\bar{X})$, con $i = 1, \dots, 9190$, donde \bar{X} es la media muestral y N_i el número de miembros del hogar i .

En la Figura 5.2 se muestra la estimación tipo núcleo y el SiZer empleando la aproximación q_1 que se analizó en la Sección 2.2 (mostrando de color gris los valores de espacio y escala para los que $ESS(x, h) < 5$) para la muestra de los ingresos equivalentes per capita normalizados de los 9190 hogares gallegos. Observando, tanto el SiZer como la estimación tipo núcleo en la Figura 5.2, parece claro que hay una moda entre el valor 1 y el 1.2. Lo que no parece tan claro es si el segundo máximo relativo que aparece en las estimaciones tipo núcleo (con parámetro $h = 0.05$ y $h = 0.10$) en torno al valor 1.3 será realmente una moda. De hecho, observando el gráfico obtenido con el SiZer (Figura 5.2, derecha) se llegaría a la conclusión de que esta muestra es unimodal. Si bien, se debe recordar que esta herramienta exploratoria no proporciona una manera formal de determinar si esta muestra es realmente unimodal o multimodal.

Con el fin de determinar, de manera formal, si hay una moda o más de una moda, se han calculado los p -valores asociados a los distintos test de unimodalidad analizados en el Capítulo 3. Como se desconoce el intervalo en el que se deberían situar las rentas por hogar en Galicia, para realizar el

Método 1 propuesto por Hall y York (2001), se ha escogido el intervalo $[0, 4]$ ya que es donde parece que se sitúa gran parte de la muestra (en particular es donde se sitúan el 97.58 % de los valores de la muestra) y, aunque en esta ocasión la estimación tipo núcleo permite ver que en los hogares con rentas altas no se forma una nueva moda, la elección de este intervalo evitaría determinar que hay una segunda moda ocasionada por un pequeño sector de la población con rentas muy altas.

Se puede ver que, en este caso, en función del test empleado se llega a distintas conclusiones acerca de si se debe rechazar o no la hipótesis nula de unimodalidad. Así, para un nivel de significación del 5 % se rechazaría la hipótesis nula con los test de Cheng y Hall (1998) (p-valor igual a 0) y con la nueva propuesta empleando la ventana crítica (p-valor igual a 0). Mientras que no habría evidencias para rechazarla con el test de Silverman (1981) (p-valor 0.372), con el Método 1 de Hall y York (2001) (p-valor igual a 0.074) y con la versión sencilla de Fisher y Marron (2001) (p-valor 0.280). Tal y como se analizó en la Sección 3.3, los únicos test que parecían que podrían calibrar correctamente, son los test de Cheng y Hall (1998) y la nueva propuesta presentada. Como con ambos test se ha obtenido un p-valor igual a 0, para un nivel de significación del 5 %, se rechaza la hipótesis nula de que la muestra procede de una distribución unimodal. Consecuentemente, hay evidencias significativas de que existen al menos dos tendencias en los ingresos equivalentes per capita normalizados por hogar en Galicia y, por tanto, ninguna de las distribuciones unimodales empleadas habitualmente para modelizar esta variable es válida.

5.3. Dirección de viento en los episodios de contaminación por NO_X

Uno de los problemas a los que se enfrentan hoy en día las empresas es el de respetar el medio ambiente, controlando la cantidad nociva de productos contaminantes que envían a la atmósfera. Con este fin, las empresas están obligadas a colocar medidores para detectar cuándo se sobrepasan de los límites que se imponen desde los órganos administrativos. Un ejemplo de este caso, se tiene con las partículas de NO_X que se pueden expulsar a la atmósfera. El Real decreto 102/2011 relativo a la mejora de la calidad del aire y la directiva europea 50/2008/CE establecen que los valores límite de protección de la salud para el NO_X es de $200\mu\text{g}/\text{m}^3$, que no podrá superarse más de 18 horas al año.

El problema surge en que, por ejemplo, el NO_X no solo se produce por las actividades industriales. También se puede generar por actividades agrícolas, por la emisión de los vehículos de gasolina y diésel o por otras causas. Una de las formas de determinar si las alarmas por el exceso de NO_X no son solo producidas por la actividad de una empresa sería el estudiar la dirección media de la que procede el viento en el tiempo en que se produce una alarma por exceso de NO_X en el aire. Así, si hay más de una moda en la dirección de la que viene el viento cada vez que se produce una alarma por exceso de este contaminante, significará que en ocasiones este contaminante llega de la empresa, pero que esta no es el único factor a considerar.

Un ejemplo de este caso, se tiene en la tercera base de datos analizada, donde se ha estudiado la dirección de la que procedía el viento en cada una de las 151 alarmas que se recogieron en un medidor situado en la localidad de Vilalba en los datos recogidos minutalmente durante dos años,

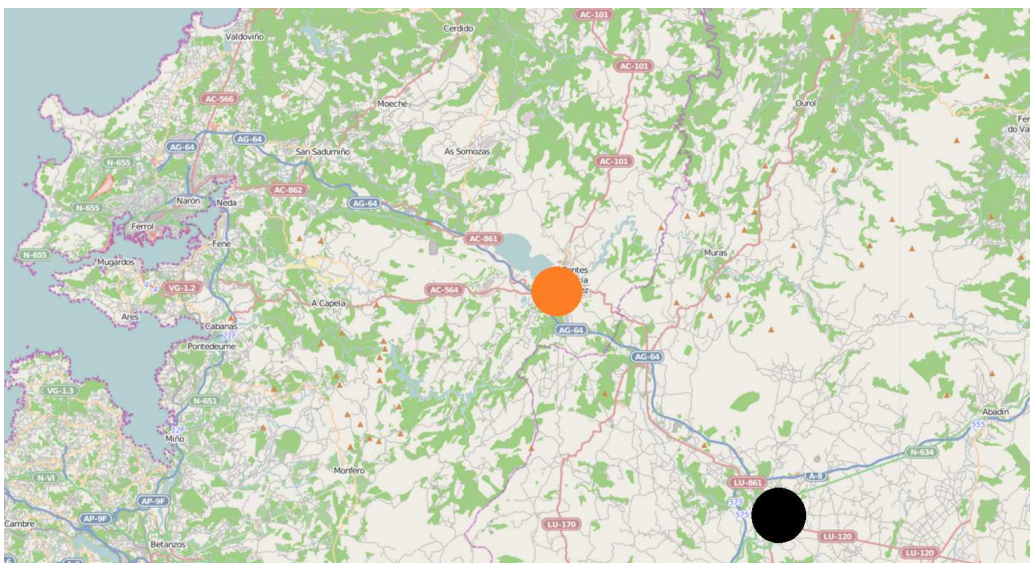


Figura 5.3: Localización del medidor de NO_x (círculo negro) y de la central térmica de As Pontes (círculo naranja).

entre el 1 de Julio del 2011 y el 31 de Junio del 2013, para ver si toda la contaminación de NO_x se puede atribuir a la central térmica de As Pontes de García Rodríguez (se muestra en la Figura 5.3 la localización del medidor y de la central térmica²). Si todos los excesos de NO_x fuesen provocados por la central térmica, cada vez que surgiese una alarma, el viento debería soplar procedente del noroeste³, esto es, debería haber una única tendencia en la dirección de la que procede el viento.

Dentro de los dos años estudiados, ya sea por falta de datos o por estar soplando el viento de forma constante durante todo el día⁴, se han eliminado 33 días de los 731 estudiados. Para los 698 días restantes, el medidor de Vilalba ha recogido minutalmente la dirección de la que procedía el viento y la cantidad de NO_x en el aire. Entendiéndose que, mientras no se deje de sobrepasar el umbral de $200\mu\text{g}/\text{m}^3$, todas las observaciones recogidas formarán parte de una misma alarma, esto es, se supone que nunca se producen dos alarmas distintas seguidas en el tiempo. Para aquellas alarmas que duraron más de un minuto se ha tomado la dirección media de la que procedía el viento mientras se producía dicha alarma.

El tiempo total en el que se ha sobrepasado el umbral ha sido de 187 minutos durante estos 698 días (muy por debajo del límite legal establecido). La duración en minutos de cada una de las alarmas producidas se muestra en el diagrama de barras de la Figura 5.4 (izquierda). Se representa el diagrama de rosa para la hora media a la que se produjeron cada una de las 151 alarmas por

²Se ha obtenido este mapa y el que se muestra en la Figura 5.8 de la Sección 5.4 de OpenStreetMap (2014).

³Debido a la poca distancia que hay entre la central y el medidor (aproximadamente 22 km), se está suponiendo que las emisiones de la central llegan a través de las corrientes de aire procedentes del noroeste.

⁴Se entiende que durante esos días hubo un error de medición del aparato y por eso se eliminan.

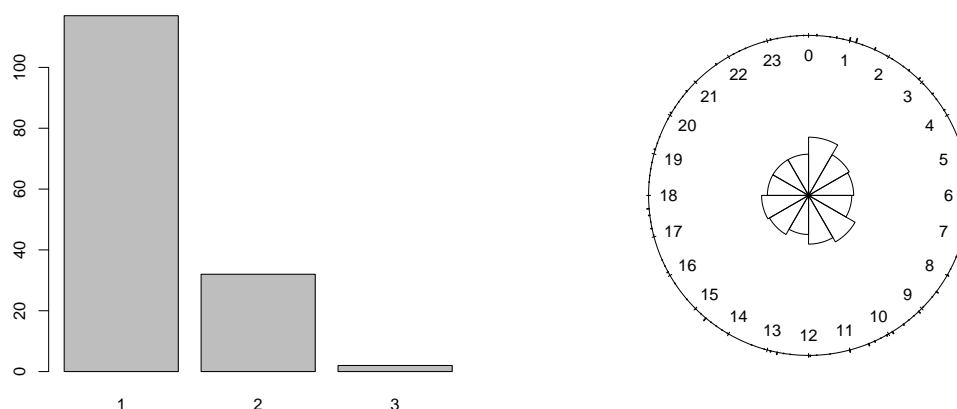


Figura 5.4: Izquierda: diagrama de barras para la duración en minutos de cada alarma por exceso de NO_x en el aire. Derecha: diagrama de rosa para la hora media en la que se produjeron cada una de las alarmas por exceso de NO_x.

exceso de NO_x en la Figura 5.4 (derecha).

Observando la Figura 5.4 (derecha) parece que las alarmas se producen de forma uniforme a lo largo de todo el día. De hecho, aplicando el test de uniformidad en el círculo de Kuiper (véase Jammalamadaka y Sengupta (2001), Sección 7.2) para un nivel de significación del 5% no existen evidencias significativas para rechazar la hipótesis nula de uniformidad. Esto permitiría descartar, por ejemplo, que una gran parte de las alarmas que se producen por exceso de NO_x sean provocadas por el tráfico rodado durante las horas puntas. Mientras que no descartaría que estas sean ocasionadas por la central térmica, ya que esta produce NO_x a lo largo de todo el día.

Continuando con el objetivo de analizar si gran parte de las alarmas son provocadas por la central térmica de As Pontes, a partir de la muestra de la dirección media de la que procede el viento cada vez que se produce una alarma, en la Figura 5.5 se ha realizado la estimación circular tipo núcleo (izquierda) y se representa el CircSiZer empleando la aproximación q_3^c que se analizó en la Sección 2.3.1 (derecha).

Este sería un ejemplo donde las herramientas exploratorias son de gran utilidad, ya que observando la Figura 5.5 se puede ver que lo más probable es que haya dos modas una entre el oeste y el noroeste, que es la dirección de la que debería provenir el viento si la alarma es causada por la central térmica, y otra en torno a la dirección este. Haciendo uso de la estimación tipo núcleo para parámetros de concentración bajos ($\nu = 0.15$), parece que si solo hubiese una moda esta se debería situar en torno a la dirección este. A partir de estos hechos, parece claro que no todo el exceso de

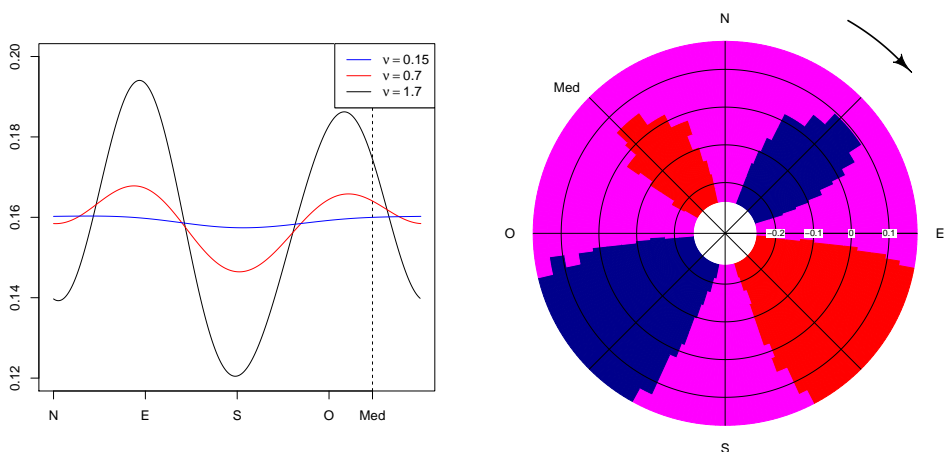


Figura 5.5: Izquierda: estimación circular de la densidad con núcleo von Mises y parámetros de concentración $\nu = 0.15$, $\nu = 0.7$ y $\nu = 1.7$. Derecha: CircSiZer empleando parámetros de concentración ν entre 0.7 y 1.7 (se presenta en el radio los valores de $-\log_{10}(\nu)$). Ambos empleando la muestra de la dirección media de la que procede el viento cada vez que se produce una alarma. Con “Med” se señala de donde debería proceder el viento para que la alarma fuese provocada por la central térmica.

NO_x en aire puede ser atribuido a la actividad de la central térmica.

Con el fin de comprobar, de manera formal, que hay más de una dirección principal de la que procede el viento cada vez que se produce una alarma, se han realizado los test estudiados en el Capítulo 4 para comprobar, a un nivel del 5%, si se rechaza la hipótesis nula de unimodalidad. Aunque en este caso, para un nivel de significación del 5%, se llega a la misma conclusión con el test de Fisher y Marron (2001) (Sección 4.1.2), con la nueva propuesta basada en la concentración crítica (Sección 4.1.1) y con la nueva propuesta basada en el exceso de masa (Sección 4.2), ya que los p -valores obtenidos son, respectivamente, 0, 0.002 y 0. Se debe recordar que el único test válido para realizar este contraste es el del test basado en el exceso de masa, ya que es el único que presentaba un calibrado correcto para el caso circular. En este caso, de acuerdo con las herramientas exploratorias, a la vista del p -valor asociado al test basado en el exceso de masa, parece claro que no todas las alarmas por exceso de NO_x pueden ser atribuidas a la central térmica.

5.4. Los incendios en la comarca de Vigo

El último ejemplo que se va a mostrar, y que ha motivado la realización de gran parte de este TFM, es el de los incendios producidos en la comarca de Vigo con el fin de determinar si en esta

comarca hay una temporada de incendios o más de una.

La muestra de donde se produjeron los incendios y en que fecha, han sido proporcionados por J. M. C. Pereira y su grupo de investigación (para un análisis de estos datos véase Oom y Pereira (2013) o Le Page *et al.* (2010), por ejemplo) tras procesar y depurar los datos obtenidos a partir del Espectroradiómetro de Imágenes de Media Resolución (MODIS por sus siglas en inglés, *Moderate Resolution Imaging Spectroradiometer*), situado en la órbita terrestre, a 705 km del suelo, por la NASA, a bordo de los satélites Terra (EOS AM) y Aqua (EOS PM). Este instrumento científico permite cartografiar toda la superficie del planeta, cada dos días, detectando incendios a través de un algoritmo implementado en el MODIS cuando estos se están produciendo en el momento que pasa alguno de los satélites (estos pasan tanto por la mañana como por la tarde). El MODIS es capaz de detectar incendios pequeños, en cuadrículas de 50 m², bajo buenas condiciones (con poca nubosidad, poco humo, ...) y en superficies de 900 m² bajo peores condiciones. Se muestra en la Figura 5.6 una de las imágenes obtenidas por el MODIS en el satélite Terra de la NASA el día 7 de Agosto de 2006, donde se detectaron fuegos activos en la comarca de Vigo⁵. Para más información acerca del instrumento o del algoritmo empleado para detectar incendios véase MODIS (2014).

Con el fin de determinar si hay una temporada de incendios o más de una en la comarca de Vigo, se han recogido todos los incendios que se detectaron en dicha región durante 10 años, entre el 10 de Julio de 2002 y el 9 de Julio de 2012. De estos 10 años, se ha decidido eliminar el año 2006 ya que, aunque a la comarca de Vigo no pareció afectarle en exceso, se sabe que este año fue atípico en Galicia, en el sentido de que se produjeron una cantidad muy elevada de incendios. Observando el gráfico de barras de la Figura 5.7, se puede ver el exceso de incendios que hubo en Galicia durante el 2006. Mientras que en los años 2003, 2004, 2005 y 2011 se detectaron alrededor de 1000 incendios y del 2007 al 2010 en torno a los 250, en el 2006 se detectaron más de 3500 incendios.

En el caso de la comarca de Vigo no parece tan claro el determinar si este año resultó atípico, en cuanto a la cantidad de incendios, ya que se detectaron casi los mismos incendios que durante el 2005. Aún así, se ha decidido eliminar a este año del estudio ya que se tiene que, mientras que en el resto de años los incendios están repartidos desde Enero hasta Octubre, en el 2006 todos los incendios se detectaron entre el 5 y el 10 de Agosto. Se obtienen así, un total de 155 incendios ocasionados entre los días 10 de Julio de 2002 y el 31 de Diciembre de 2005 y el período que va desde el 1 de Enero del 2007 al 9 de Julio de 2012.

En la Figura 5.8 se representan las localizaciones de los incendios detectados en la comarca de Vigo desde el 10 de Julio de 2002 al 9 de Julio de 2012. No se representan los incendios ocurridos en Noviembre y Diciembre, ya que durante estos 10 años no se recogió ningún incendio durante estos meses en la comarca de Vigo. En la Figura 5.8 los incendios en color amarillo y marrón corresponderían a la época en la que, normalmente, los organismos públicos permiten las quemas controladas y de residuos forestales, que es fuera de lo que la Consellería de Medio Rural e do Mar (2014) denomina época de máximo peligro de incendios, comprendida entre el 15 de Junio y el 30 de Septiembre (aunque en ocasiones, tampoco se permite realizar quemas controladas en otros días por alto riesgo de incendios). La distinción realizada en la Figura 5.8 entre los incendios ocasionados

⁵Esta imagen ha sido realizada por Schmaltz (2006) para Visible Earth, el catálogo de imágenes y animaciones de la NASA.

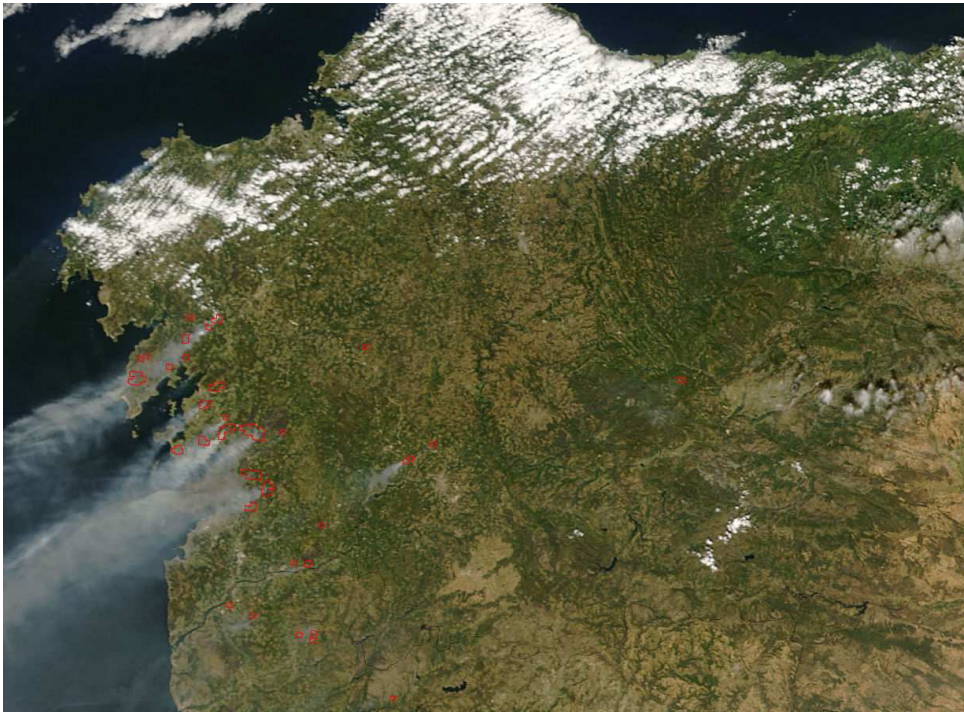


Figura 5.6: Imagen del noroeste de la Península Ibérica obtenida por el MODIS en el satélite Terra de la NASA el día 7 de Agosto de 2006. Se muestran los lugares donde el MODIS detectó incendios activos en el interior del contorno rojo.

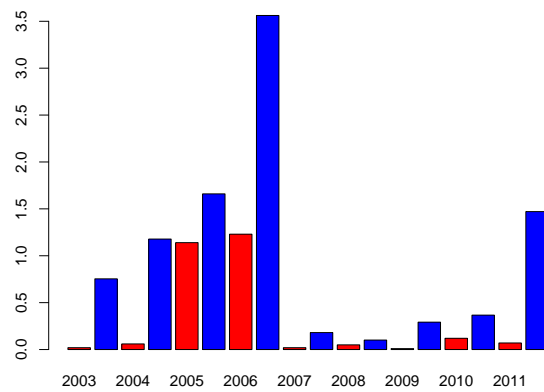


Figura 5.7: Gráfico de barras para el número de incendios por año. En rojo: incendios ocasionados en Vigo (divididos por 100). En azul: incendios producidos en Galicia (divididos por 1000).

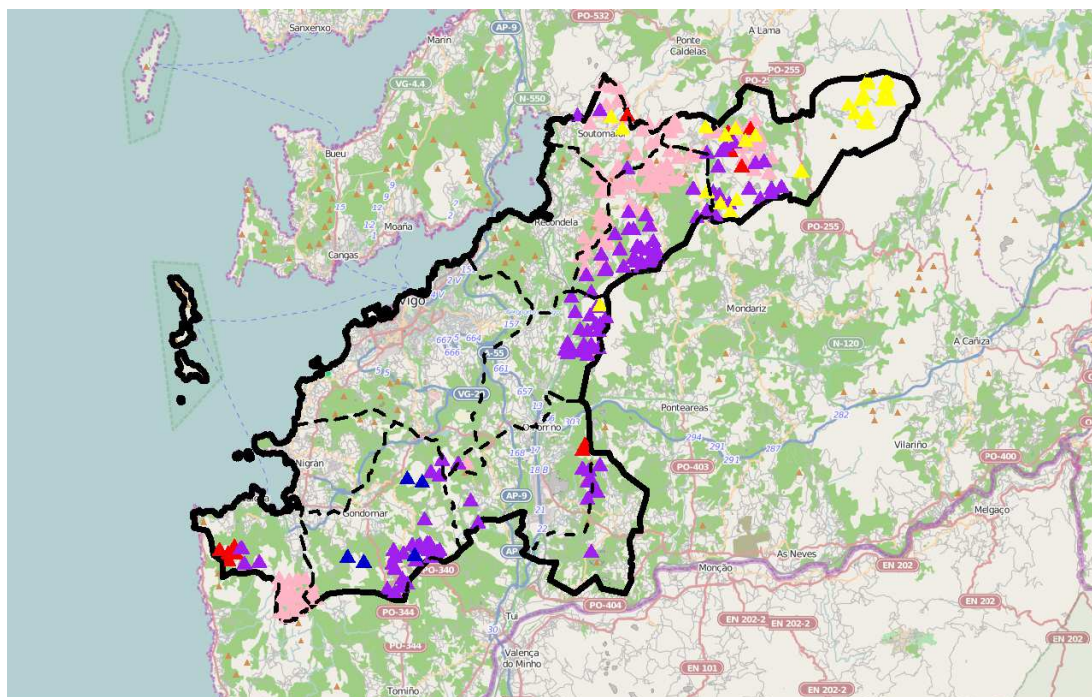


Figura 5.8: Interior del contorno continuo grueso negro: comarca de Vigo. Interior del contorno discontinuo negro: delimitación de los ayuntamientos de la comarca de Vigo. Triángulos rosas: incendios del 2006. Triángulos amarillos: incendios producidos entre el 1 de Enero y el 14 de Junio. Triángulos rojos: incendios ocasionados entre el 15 de Junio y el 15 de Agosto. Triángulos púrpuras: incendios originados entre el 16 de Agosto y el 30 de Septiembre. Triángulos azules: incendios originados entre el 1 de Octubre y el 31 de Octubre.

en Octubre y los producidos en el resto de meses situados fuera de la época de máximo peligro de incendios viene dada ya que se ha observado, en las distintas comarcas de Galicia, que la estimación de la temporada de incendios en la época estival suele ir desde finales de Junio hasta finales de Octubre. Este hecho se puede observar en el gráfico de barras de la Figura 5.9 (tanto para Galicia como para la comarca de Vigo), y también en la estimación circular tipo núcleo y en el CircSiZer realizados para la muestra de 155 observaciones de la fecha en la que se detectaron los incendios en la comarca de Vigo en la Figura 1.15.

Sabiendo que la ciudad de Vigo y sus alrededores se dedican principalmente a la actividad industrial y pesquera, puede resultar extraño que se produzca una estación de incendios ocasionada por las quemadas controladas y de residuos forestales asociada a la actividad agrícola. Con el fin de analizar si la moda que se estimó (para parámetros de concentración altos), en la Figura 1.15 del Capítulo 1, entre los meses de Enero y Abril, fue ocasionada por un año atípico en el que se

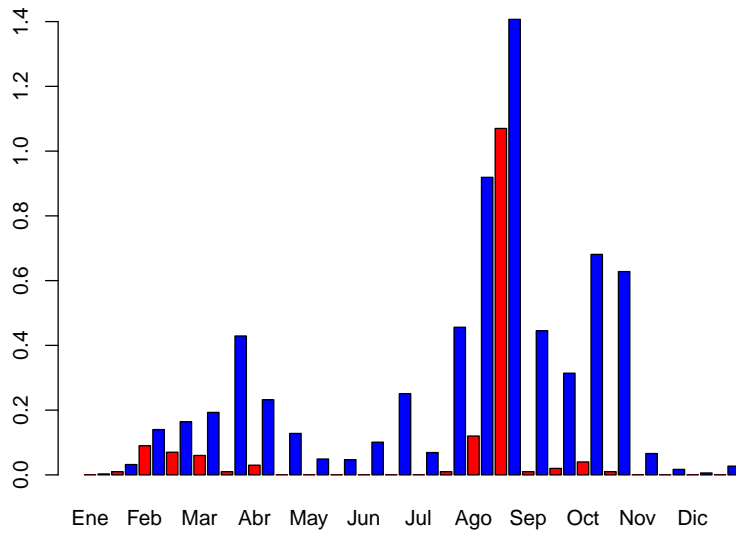


Figura 5.9: Gráfico de barras para el número de incendios divididos por quincenas. En rojo: incendios ocasionados en Vigo (divididos por 100). En azul: incendios producidos en Galicia (divididos por 1000).

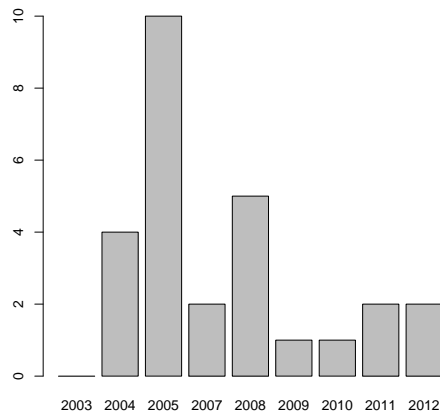


Figura 5.10: Gráfico de barras para el número de incendios detectados, cada año, entre el 1 de Enero y el 30 de Abril.

produjeron un exceso de quemas durante este periodo, se ha representado en el gráfico de barras de la Figura 5.10 el número de incendios detectados, cada año, entre el 1 de Enero y el 30 de Abril. Observando la Figura 5.10, parece claro que esta posible moda no es un efecto provocado por un año atípico. La explicación de esta posible moda se puede encontrar en la Figura 5.8, donde se puede observar que los incendios producidos entre el 1 de Enero y el 14 de Junio no son ocasionados ni en la ciudad de Vigo ni en sus alrededores, si no que casi todos (a excepción de cuatro) se producen en el ayuntamiento agrícola de Fornelos de Montes (ayuntamiento situado más al noreste de la comarca).

A través de las herramientas exploratorias no queda claro si la moda que aparece a finales de Febrero es realmente una moda y por tanto se tienen dos modas o si, en cambio, solo hay una temporada de incendios en la comarca de Vigo y esta se produce en torno a mediados de Agosto. Con el objetivo de determinar si solo hay una moda en la época estival o, en cambio, hay más de una temporada de incendios en la comarca de Vigo, se pueden emplear los test de multimodalidad vistos en el Capítulo 4. Como los datos obtenidos son una discretización de los datos originales (ya que solo pueden tomar 366 valores), con el fin de recuperar la estructura continua, se va a suponer que estos incendios pudieron ocurrir en cualquier momento del día. Para ello, si en la muestra original hay un incendio ocasionado el día a , se generará un número aleatorio de la distribución uniforme en $[2\pi(a - 1)/366, 2\pi a/366)$.

Utilizando los 155 valores generados a partir de la muestra original se han realizado los distintos test presentados en el Capítulo 4. Para los tres test se ha obtenido un p -valor igual a 0. Aunque al igual que ocurría con el ejemplo de la Sección 5.3, a un nivel del 5%, la conclusión sería la misma con los tres test, el único válido para realizar el contraste de unimodalidad frente a multimodalidad es el presentado en la Sección 4.2. Observando el p -valor obtenido para el test basado en el exceso de masa, a un nivel del 5%, se puede concluir que se rechaza la hipótesis nula de que en la comarca de Vigo solo hay una temporada de incendios.

Como no queda muy claro si 2006 ha sido un año atípico en la comarca de Vigo, también se ha realizado el test basado en el exceso de masa para los 246 datos generados a partir de la muestra de los incendios detectados en esta comarca desde el 10 de Julio de 2002 al 9 de Julio de 2012. En este caso, el p -valor obtenido es igual a 0, por tanto se seguiría rechazando la hipótesis nula de unimodalidad.

Otra de las dudas que pueden surgir es qué ocurre cuando se considera toda la comarca de Vigo sin el ayuntamiento de Fornelos de Montes. Se puede ver que, aplicando el test basado en el exceso de masa sobre los 112 valores generados a partir de la muestra de los incendios que se detectaron en los ayuntamientos de la comarca de Vigo, excluyendo el de Fornelos de Montes, el p -valor obtenido es igual a 0.136. Por tanto, a un nivel del 5%, no hay evidencias significativas para rechazar la hipótesis nula de que solo se produzca una temporada de incendios en la comarca de Vigo cuando se elimina el ayuntamiento de Fornelos de Montes.

Apéndice A

Modelos empleados

A continuación, se muestran las expresiones de las mezclas de normales empleadas a lo largo de este TFM, para ello se ha usado la notación $\sum_{i=1}^l p_i \cdot N(\mu_i, \sigma_i^2)$, donde cada $N(\mu_i, \sigma_i^2)$ representará cada una de las componentes de la mezcla y los p_i serán los pesos asociados a cada componente y verificarán que $\sum_{i=1}^l p_i = 1$.

- M1: $N(0.5, 1/36)$.
- M2: $0.9 \cdot N(0.5, 1/324) + 0.1 \cdot N(0.5, 1/36)$.
- M3: $0.2 \cdot N(0.3, 0.01) + 0.6 \cdot N(0.5, 0.024) + 0.2 \cdot N(0.7, 0.01)$.
- M4: $0.95 \cdot N(0.3, 0.01) + 0.05 \cdot N(0.8, 0.0004)$.
- M5: $0.75 \cdot N(0.5, 1/36) + 0.25 \cdot N(0.85, 0.0025)$.
- M6: $0.5 \cdot N(0.3, 0.01) + 0.5 \cdot N(0.6, 0.01)$.
- M7: $0.86 \cdot N(0.4, 0.001) + 0.14 \cdot N(0.74, 0.0004)$.
- M8: $0.5 \cdot N(0, 0.49) + 0.5 \cdot N(1.4, 0.25)$.
- M9: $0.95 \cdot N(0, 10) + 0.05 \cdot N(0, 0.02)$.
- M10: $0.45 \cdot N(0.3, 0.01) + 0.45 \cdot N(0.6, 0.01) + 0.1 \cdot N(0.9, 0.0004)$.

Las expresiones de las mezclas de von Mises empleadas a lo largo de este TFM, usando la notación $\sum_{i=1}^l p_i \cdot vM(\mu_i, \kappa_i)$, con $\sum_{i=1}^l p_i = 1$, serían las que siguen:

- M11: $0.25 \cdot vM(0, 2) + 0.75 \cdot vM(\pi/\sqrt{3}, 2)$.
- M12: $0.9 \cdot vM(\pi, 10) + 0.1 \cdot vM(\pi, 1)$.
- M13: $0.5 \cdot vM(0, 4) + 0.5 \cdot vM(\pi, 4)$.
- M14: $0.5 \cdot vM(2, 5) + 0.5 \cdot vM(4, 5)$.
- M15: $1/6 \cdot vM(\pi - 0.8, 30) + 1/2 \cdot vM(\pi, 1) + 1/6 \cdot vM(\pi, 30) + 1/6 \cdot vM(\pi + 0.8, 30)$.

Las expresión de la mixtura de modelos circulares empleada en este TFM es la que sigue:

- M16: Mixtura de von Mises y Cauchy enrollada: $0.8 \cdot vM(\pi, 5) + 0.2 \cdot WC(4\pi/3, 0.9)$.

Se muestra la representación lineal de estos 16 modelos en la Figura A.1.

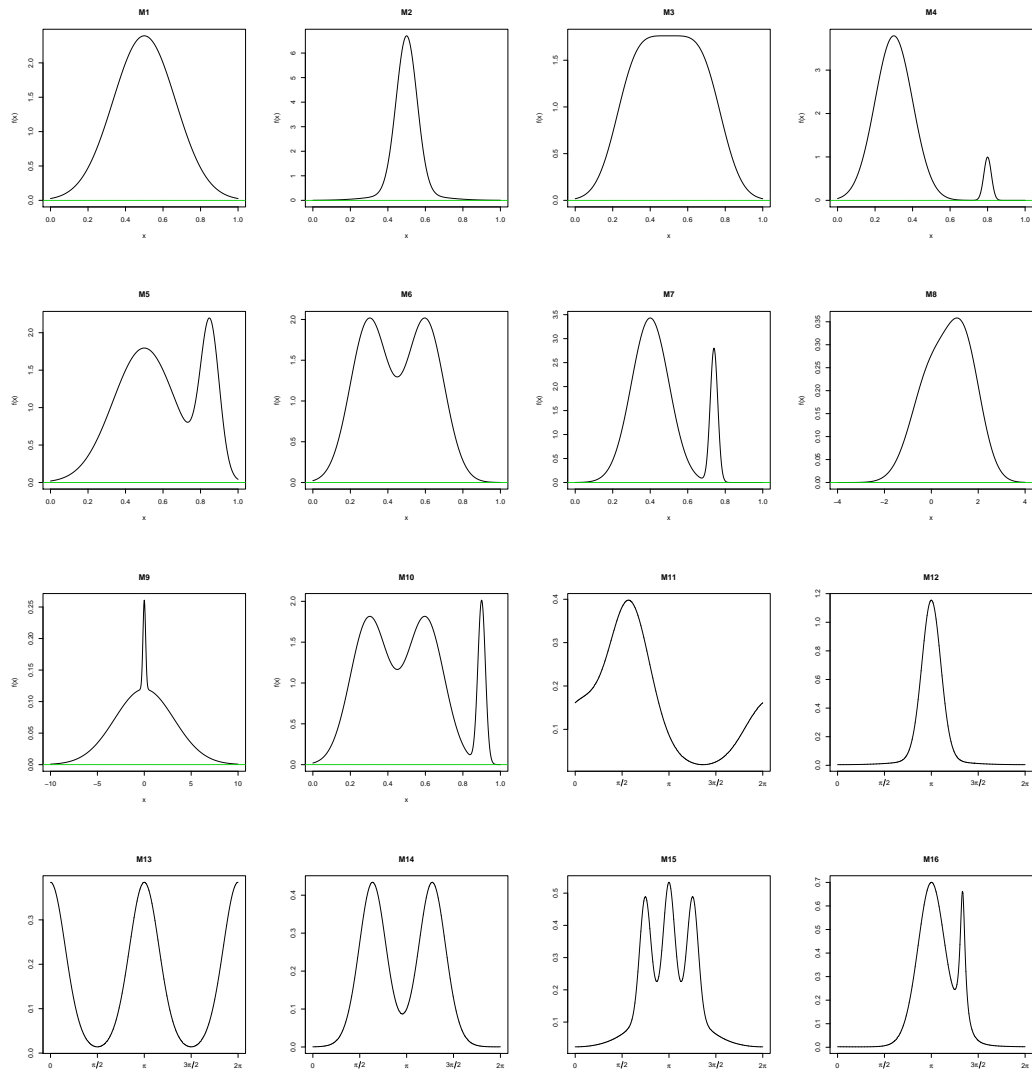


Figura A.1: Modelos utilizados en este TFM. M1–M10: funciones de densidad lineales. M11–M16: funciones de densidad circulares.

Apéndice B

Resumen de los test empleados.

En este anexo, en las Tablas B.1 y B.2, se presenta un resumen de todos los test presentados a lo largo de este Trabajo Fin de Máster, destacando los aspectos más importantes de cada test, esto es, el contraste que realizan, el estadístico usado para ello, el calibrado propuesto para tratar de obtener un test que funcione correctamente (si lo hay) y por quién o quiénes fueron propuestos dichos test y calibrados.

Además, en la Figura B.1, se presenta un diagrama cronológico para analizar como fueron surgiendo en la literatura las distintas herramientas gráficas y los distintos test. Se verá como surge en el año 1981 el primer test de multimodalidad por la vía de la ventana crítica desarrollado por Silverman (1981) y en 1985 el primer test por la vía del *dip* o el exceso de masa desarrollado por Hartigan y Hartigan (1985). A partir de ese momento y hasta el año 2001, se desarrollarán estas dos tendencias para tratar de proporcionar un test de multimodalidad que calibre correctamente. La estimación del número de modas y de su localización a través de las herramientas gráficas comienza, en el contexto lineal, en el año 1993 con Minnotte y Scott (1993) estudiando directamente la localización de las modas a través del árbol de modas. En el año 2001 con Fisher y Marron (2001) surgirá el primer test de multimodalidad para el contexto circular y no es hasta el 2012 cuando aparece la primera herramienta exploratoria en el contexto circular presentada por Oliveira *et al.* (2012b).

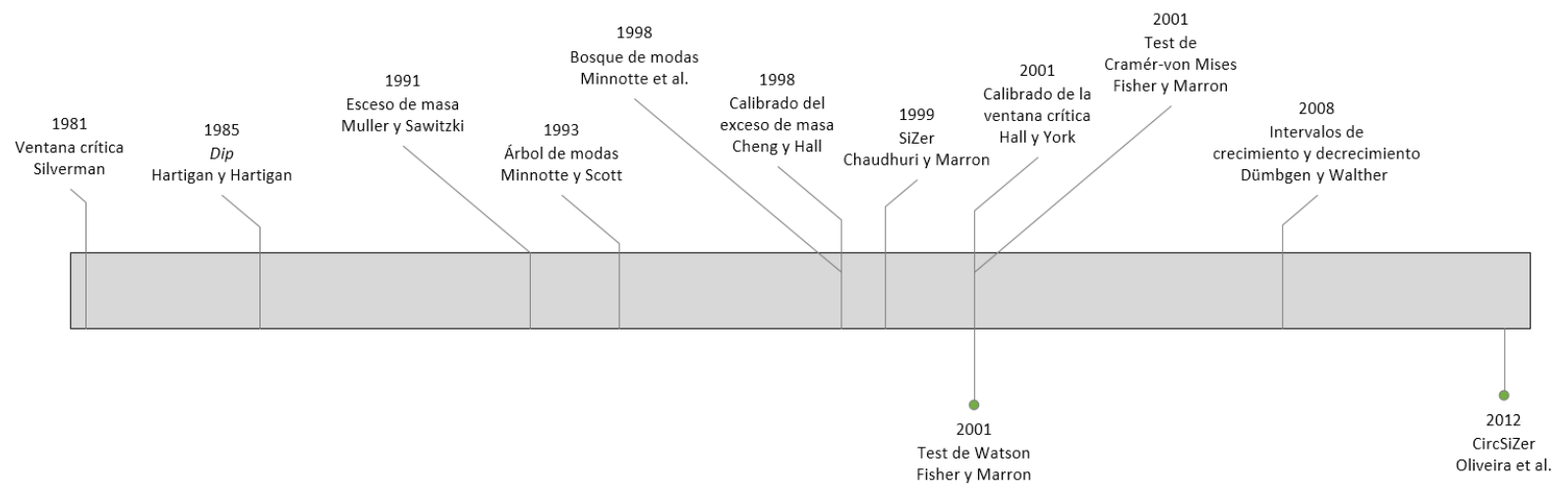


Figura B.1: Diagrama cronológico de los distintos test de multimodalidad y herramientas gráficas para la detección de modas.

Nombre	Contraste	Estadístico	Calibrado	Propuesto por
Ventana crítica	$H_0 : j \leq k$ $H_1 : j > k$	$h_k = \min\{h : \hat{f}_h \text{ tiene a lo sumo } k \text{ modas}\}$ Se rechazará H_0 si $\mathbb{P}(\hat{h}_k^* \leq h_k \mathcal{X}) \geq 1 - \alpha$		Silverman (1981)
	$H_0 : j = 1$ y no tiene mínimos locales en el intervalo I $H_1 : j > 1$ en I		$h_{HY} = \min\{h : \hat{f}_h \text{ tiene exactamente una moda en } I\}$ Método 1: Se rechazará H_0 si $\mathbb{P}(h_{HY}^* \leq \lambda_\alpha \mathcal{X}) \geq 1 - \alpha$, donde los λ_α son los dados en (3.4) Método 2: Se rechazará H_0 si $\mathbb{P}(h_{HY}^* \leq h_{HY} \mathcal{X}) \geq 1 - \tau_\alpha$, los τ_α se obtienen a partir de las h_{HY} para una $N(0, 1)$ en el intervalo $[-1.5, 1.5]$	Hall y York (2001)
Cramér-von Mises	$H_0 : j \leq k$ $H_1 : j > k$	$T_k = \sum_{i=1}^n \left(\hat{F}_{h_k}(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$ $\hat{F}_{h_k}(x) = \int_{-\infty}^x \hat{f}_{h_k}(t) dt$ Se rechazará H_0 si $\mathbb{P}(T_k^* \leq T_k \mathcal{X}) \geq 1 - \alpha$		Fisher y Marron (2001)
	$H_0 : j_{FM} \leq k_{FM}$ $H_1 : j_{FM} > k_{FM}$		$T_{FM} = \sum_{i=1}^n \left(\hat{F}_{h_{FM}}(X_{(i)}) - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$ $h_{FM} = \sup \{h / S_{k_{FM}}(h) > 0\}$ $\hat{F}_{h_{FM}}(x) = \int_{-\infty}^x \hat{f}_{h_{FM}}(t) dt$ Se rechazará H_0 si $\mathbb{P}(T_{FM}^* \leq T_{FM} \mathcal{X}) \geq 1 - \alpha$	
<i>Dip</i>	$H_0 : j = 1$ $H_1 : j > 1$	$D(F_n) = \inf_{G \in \mathcal{V}} \left(\sup_x F_n(x) - G(x) \right)$ Se rechazará H_0 si $\mathbb{P}(\kappa_\alpha \leq D(F_n)) \geq 1 - \alpha$, los κ_α se obtienen a partir de los <i>dip</i> para una $U(0, 1)$		Hartigan y Hartigan (1985)
Exceso de masa	$H_0 : j \leq k$ $H_1 : j > k$	$\Delta_{n,k+1} = \max_\lambda \{E_{n,k+1}(F_n, \lambda) - E_{n,k}(F_n, \lambda)\}$ $E_{n,k}(F_n, \lambda) = \sup_{C_1, \dots, C_k} \left\{ \sum_{l=1}^k \mathbb{P}_n(C_l) - \lambda C_l \right\}$		Müller y Sawitzki (1991)
	$H_0 : j = 1$ $H_1 : j > 1$		$\Delta_{n,2} = 2 \cdot D(F)$ La distribución se aproxima por Monte Carlo a partir de $\psi(\hat{b})$ Se rechazará H_0 si $\mathbb{P}(\Delta_{n,2}^* \leq \Delta_{n,2} \mathcal{X}) \geq 1 - \alpha$. Se toma \mathcal{X}^* siguiendo una v.a. con densidad \hat{f}_{h_1}	Cheng y Hall (1998) Nueva propuesta (2014)

Tabla B.1: Resumen de los test empleados para el caso lineal.

Nombre	Contraste	Estadístico	Propuesto por
Concentración crítica	$H_0 : j^c = 1$ $H_1 : j^c > 1$	$\nu_1^* = \sup\{\nu : \hat{f}_{\nu}^{c*} \text{ tiene una moda en el intervalo } [0, 2\pi)\}$ Se rechazará H_0 si $\mathbb{P}(\nu_1^* \geq \nu_1 \Theta) \geq 1 - \tau_{\alpha}^c$, los τ_{α}^c se obtienen a partir de los ν_1 para una $vM(0, 1)$	Nueva propuesta (2014)
U^2 de Watson	$H_0 : j^c \leq k^c$ $H_1 : H_1 : j^c > k^c$	$U_{k^c}^2 = \sum_{i=1}^n (u_{(i)} - (\bar{u} - \frac{1}{2}) - \frac{2i-1}{2n})^2 + \frac{1}{12n}$ $u_i = \hat{F}_{\nu_{k^c}^c}(\Theta_i) = \int_0^{\Theta_i} \hat{f}_{\nu_{k^c}^c}(\theta) d\theta$ Se rechazará H_0 si $\mathbb{P}(U_{k^c}^2 \leq U_{k^c}^2 \Theta) \geq 1 - \alpha$	Fisher y Marron (2001)
Exceso de masa	$H_0 : j^c = 1$ $H_1 : j^c > 1$	$\Delta_{n,2}^c = \max_{\lambda} \{E_{n,2}^c(\mathbb{P}_n^c, \lambda) - E_{n,1}^c(\mathbb{P}_n^c, \lambda)\}$ $E_{n,k^c}^c(\mathbb{P}_n^c, \lambda) = \sup_{C_1, \dots, C_{k^c}} \left\{ \sum_{l=1}^{k^c} \mathbb{P}_n^c(C_l^c) - \lambda \ C_l^c\ \right\}$ Se rechazará H_0 si $\mathbb{P}(\Delta_{n,2}^{c*} \leq \Delta_{n,2}^c \Theta) \geq 1 - \alpha$. Se toma Θ^* siguiendo una v.a. con densidad $\hat{f}_{\nu_1}^c$	Nueva propuesta (2014)

Tabla B.2: Resumen de los test empleados para el caso circular.

Bibliografía

- Azzalini, A. y Bowman, A. W. (1990) A look at some data on the Old Faithful geyser. *Applied Statistics*, **39**, 357-365.
- Bandourian, R., McDonald, J. B. y Turley, R. S. (2002). *A comparison of parametric models of income distribution across countries and over time*. Maxwell School of Citizenship and Public Affairs. Syracuse University.
- Chaudhuri, P. y Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association*, **94**, 807–823.
- Cheng, M. Y. y Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society. Series B*, **60**, 579–589.
- Consellería de Medio Rural e do Mar (2014). *Permisos para la realización de quemas*. Xunta de Galicia, Galicia, España. [fecha de consulta 21/05/2014]. Disponible en Web: http://www.medioruralemar.xunta.es/es/areas/forestal/produccion_e_industrias/autorizaciones/#a8.
- Di Marzio, M., Panzera, A. y Taylor, C. C. (2012). Smooth estimation of circular cumulative distribution functions and quantiles. *Journal of Nonparametric Statistics*, **24**, 935–949.
- Díez, S. Á. (2004). Contrastes no paramétricos de bondad de ajuste de la distribución de la renta: una aplicación con datos de España y de Italia. *Estadística Española*, **46**, 77–94.
- Dümbgen, L. y Walther, G. (2008). Multiscale Inference about a density. *The Annals of Statistics*, **36**, 1758–1785.
- Duong, T. y Wand, M. (2011). *feature: Feature significance for multivariate kernel density estimation*. R package version 1.2.8. [fecha de consulta 03/01/2014]. Disponible en Web: <http://CRAN.R-project.org/package=feature>.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.
- Fisher, N.I. y Marron, J. S. (2001). Mode testing via the excess mass estimate. *Biometrika*, **88**, 419–517.

- Good, I. J. y Gaskins, R. A. (1980), Density Estimation and Bump–Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data. *Journal of the American Statistical Association*, **75**, 42–73.
- Hall, P. y York, M. (2001). On the calibration of Silverman’s test for multimodality. *Statistica Sinica*, **11**, 515–536.
- Hartigan, J. A. y Hartigan, P. M. (1985). The Dip Test of Unimodality. *Journal of the American Statistical Association*, **86**, 738–746.
- Huckemann, S. F., Kim, K. R., Munk, A., Rehfeld, F., Sommerfeld, M., Weickert, J. y Wollnik, C. (2014). The circular SiZer, inferred persistence of shape parameters and application to stem cell stress fibre structures. *arXiv preprint arXiv:1404.3300*.
- Instituto Galego de Estatística (2012). *Enquisa de condicións de vida das familias. Gasto dos fogares. Edición 2013*. Cifras IGE. [fecha de consulta 01/04/2014]. Disponible en Web: http://www.ige.eu/web/mostrar_paxina.jsp?paxina=004002003001.
- Jammalamadaka, S. R. y Sengupta, A. (2001). *Topics in circular statistics* (Vol. 5). World Scientific. Singapore.
- Johnson, N. L., Kotz, S. y Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 1 y Vol. 2). Wiley Series in Probability and Statistics. New York.
- Kleiber, C. y Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley Series in Probability and Statistics. New Jersey.
- Le Page, Y., Oom, D., Silva, J., Jönsson, P. y Pereira, J. (2010). Seasonality of vegetation fires as modified by human action: observing the deviation from eco–climatic fire regimes. *Global Ecology and Biogeography*, **19**, 575–588.
- Maechler, M. (2013). *Diptest: Hartigan’s dip test statistic for unimodality – corrected code*. R package version 0.75–5 [fecha de consulta 03/01/2014]. Disponible en Web: <http://CRAN.R-project.org/package=diptest>.
- Minnotte, M. C. (1997). Nonparametric testing of the existence of modes. *The Annals of Statistics*, **25**, 1646–1660.
- Minnotte M. C., Marchette D. J. y Wegman, E. J. (1998). The Bumpy Road to the Mode Forest, *Journal of Computational and Graphical Statistics*, **7**, 239–251.
- Minnotte, M. C. y Scott, D. W. (1993). The mode tree: A tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, **2**, 51–68.
- MODIS (2014). *The Moderate-resolution Imaging Spectroradiometer Website*. The National Aeronautics and Space Administration, Washington, D.C., United States. [fecha de consulta 21/05/2014]. Disponible en Web: <http://modis.gsfc.nasa.gov/>.
- Müller, D. W. y Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *The Annals of Statistics*, **13**, 70–84.

- Oliveira, M., Crujeiras, R. M. y Rodríguez-Casal, A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics & Data Analysis*, **56**, 3898–3908.
- Oliveira, M., Crujeiras, R. M. y Rodríguez-Casal, A. (2012b). CircSiZer: an exploratory tool for circular data. *Environmental and Ecological Statistics*, **56**, 1–17.
- Oliveira, M. (2013). *Nonparametric circular methods for density and regression*. Dirigida por Crujeiras, R. M. y Rodríguez-Casal, A. Tesis doctoral. Universidade de Santiago de Compostela, Departamento de Estatística e Investigación Operativa.
- Oliveira, M., Crujeiras, R. M. y Rodríguez-Casal, A. (2013). *NPCirc: Nonparametric Circular Methods*. R package version 2.0.0. [fecha de consulta 04/01/2014]. Disponible en Web: <http://CRAN.R-project.org/package=NPCirc>.
- Oom, D. y Pereira, J. M. C. (2013). Exploratory spatial data analysis of global MODIS active fire data. *International Journal of Applied Earth Observation and Geoinformation*, **21**, 326–340.
- Open Street Map Community (2014). *The Free Wiki World Map*. Open Street Map Foundation. [fecha de consulta 16/04/2014]. Disponible en Web: <http://www.openstreetmap.org/>.
- Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, **186**, 343–414.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. [fecha de consulta 03/01/2014]. Disponible en Web: <http://www.R-project.org/>.
- Rufibach K. y Walther G. (2013). *Multiscale Analysis for Density Functions*. R package version 1.0.6. [fecha de consulta 03/01/2014]. Disponible en Web: <http://CRAN.R-project.org/package=modehunt>.
- Schmaltz, J. (2006). *MODIS Rapid Response Team*. Goddard Space Flight Center, Greenbelt, Maryland, United States. [fecha de consulta 21/05/2014]. Disponible en Web: <http://visibleearth.nasa.gov/view.php?id=17136>.
- Schmidt-Koenig, K. (1963). On the role of loft, the distance and site of release in pigeon homing (the “cross-loft experiment”). *Biological Bulletin*, **125**, 154–164.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. New York.
- Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B*, **43**, 97–99.
- Terrell, G. R. y Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association*, **80**, 209–214.
- Wand M. P. y Jones M. C. (1995). *Kernel Smoothing*. Chapman and Hall. Great Britain.
- Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, **48**, 109–114.