Traballo Fin de Máster

Multimodalidade e verosimilitude

Diego Bolón Rodríguez

Máster en Técnicas Estatísticas

Curso 2020-2021

Proposta de Traballo Fin de Máster

Título en galego: Multimodalidade e verosimilitude
Título en español: Multimodalidad y verosimilitud
English title: Multimodality and Likelihood
Modalidade: Modalidade A
Autor: Diego Bolón Rodríguez, Universidade de Santiago de Compostela
Directores: Rosa María Crujeiras Casais, Universidade de Santiago de Compostela; Alberto Rodríguez Casal, Universidade de Santiago de Compostela
Breve resumo do traballo: Neste traballo construímos un novo test de multimodalidade baseado na pseudo-verosimilitude, buscando que sexa extensible a contextos fóra da recta real, e estudamos o seu comportamento na práctica tanto para datos lineais como circulares, comparándoo con outros tests presentes na literatura.

Dona Rosa María Crujeiras Casais, Titular de Universidade da Universidade de Santiago de Compostela, e don Alberto Rodríguez Casal, Titular de Universidade da Universidade de Santiago de Compostela, informan que o Traballo Fin de Máster titulado

Multimodalidade e verosimilitude

foi realizado baixo a súa dirección por don Diego Bolón Rodríguez para o Máster en Técnicas Estatísticas. Estimando que o traballo está terminado, dan a súa conformidade para a súa presentación e defensa ante un tribunal.

En Santiago de Compostela, a 22 de xuño de 2021.

A directora:

CRUJEIRAS CASAIS ROSA MARIA - 52931780V

Firmado digitalmente por CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-52931780V, givenName=ROSA MARIA, sn=CRUJEIRAS CASAIS, cn=CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Fecha: 2021.06.22 11:05:57 +02'00'

Dona Rosa María Crujeiras Casais

O director:

RODRIGUEZ CASAL ALBERTO - 34896829X

Firmado digitalmente por RODRIGUEZ CASAL ALBERTO - 34896829X
Fecha: 2021.06.22 11:35:21 +02'00'

Don Alberto Rodríguez Casal

O autor:

BOLON RODRIGUEZ DIEGO - 48110220Q

Firmado digitalmente por BOLON RODRIGUEZ DIEGO - 48110220Q
Fecha: 2021.06.22 11:42:35 +02'00'

Don Diego Bolón Rodríguez

Índice xeral

Resumo

Resumo en galego

As modas dunha poboación son un puntos de alta frecuencia ao redor do cales se acumula a maior parte da probabilidade. Na literatura estatística existen varios procedementos non paramétricos para contrastar o número de modas en datos lineais. Os máis relevantes son introducidos ao longo do presente traballo. Pero a metodoloxía empregada nos tests anteriores impide que sexan directamente extensibles a outros contextos, como poden ser datos multidimensionais ou datos circulares.

Partindo desta base, o principal obxectivo deste traballo é a construción dun novo test de multimodalidade para distribucións sobre a recta, procurando que sexa facilmente extensible a outros contextos. Para iso, introduciremos a idea de pseudo-verosimilitude como un análogo da función de verosimilitude da estatística paramétrica. Isto permitiranos formular o noso test coa mesma estrutura dos tests de razón de verosimilitudes paramétrico apoiándonos no concepto de xanela crítica. Ademais, a pseudo-verosimilitude poderá ser facilmente adaptada para facer inferencia en espazos de carácter xeral, pois para a súa construción só precisamos dun estimador non paramétrico da función de densidade da poboación. Para exemplificarmos a facilidade de extensión deste test, adaptarémolo para contrastar o número de modas para unha densidade circular.

Para finalizar, unha vez proposto o novo test para contrastar o número de modas dunha poboación, comprobaremos cal é o seu calibrado na práctica para datos lineais e circulares mediante cadanseu estudo de simulación. No caso lineal, tamén comparamos o seu comportamento cos principais tests de multimodalidade presentes na literatura estatística xa introducidos.

English abstract

The modes of a statistical population are high frequency points around which most of the probability mass is accumulated. We can find several non-parametric methods to test the number of modes in linear data literature. The most relevant ones are introduced throughout this work. But the previous tests are not easily adaptable to other contexts, as it may be multidimensional or circular data.

Starting from this point, the main goal of this work is testing multimodality for linear data with a methodology extensible beyond the real line. Keeping this in mind, we introduce the idea of pseudo-likelihood as an analogue of the parametric likelihood function. This allows us to formulate the new test with the same structure of the parametric likelihood ratio test using in the concept of critical bandwidth. Moreover, pseudo-likelihood can be easily adapted to perform inference in many abstract spaces, since we only need a non-parametric estimator of the density function to define it. To exemplify the ease of extension of this test, we adapt it to test the number of modes of circular data.

Once the new test is designed, we will check if it shows good calibration in practice for linear and circular data with two simulation studies. In the lineal case, we also compare its behavior with the main multimodality tests present in the literature already introduced.

Introdución e motivación

Unha moda dunha variable aleatoria absolutamente continua X é un punto x_0 onde a función de densidade de X, que denotaremos por f, ten un máximo local. Por tanto o concepto de moda fai referencia á idea de *concentración*: as modas son puntos de alta frecuencia ao redor dos cales se acumula a maior parte da probabilidade. Unha distribución (ou función de densidade) cunha soa moda denomínase *unimodal*. No caso de que teña máis dunha moda dise *multimodal*. Tamén hai nomes específicos para referirse ao número exacto de modas da poboación: *bimodal* se ten dúas modas, *trimodal* se ten tres, e, en xeral, *j-modal* se ten j modas distintas. O problema principal que trataremos neste traballo será o de realizar inferencia sobre o número de modas dunha densidade f a partir dunha mostra aleatoria da variable X.

Para tratar de motivar a necesidade dunha boa estimación do número de modas dunha poboación introduciremos os datos analizados por Choi et al. (2020). Esta base de datos contén distintas variables asociadas a 1990 falecementos de aves e morcegos por colisións contra aeroxeradores detectadas por 44 parques éolicos situados na parte nordeste dos Estados Unidos de América. Das 1990 mortes rexistradas, 975 correspóndense a falecementos de aves, mentres que as outras 1015 son de morcegos. Entre as variables recollidas atópanse o tipo de animal (paxaro ou morcego) e a distancia horizontal entre o cadáver do animal e o aeroxerador máis próximo. Na Figura 1 están representados os histogramas das distancias ao aeroxerador máis próximo para cada tipo de animal. Tal como salientan Choi et al. (2020), estimar o número de modas de cada distribución poboacional, así como a súa localización, permitiranos deseñar protocolos de busca máis efectivos e mellorar a estimación da mortalidade total, tendo así unha visión máis precisa de como os parques eólicos afectan á fauna silvestre.

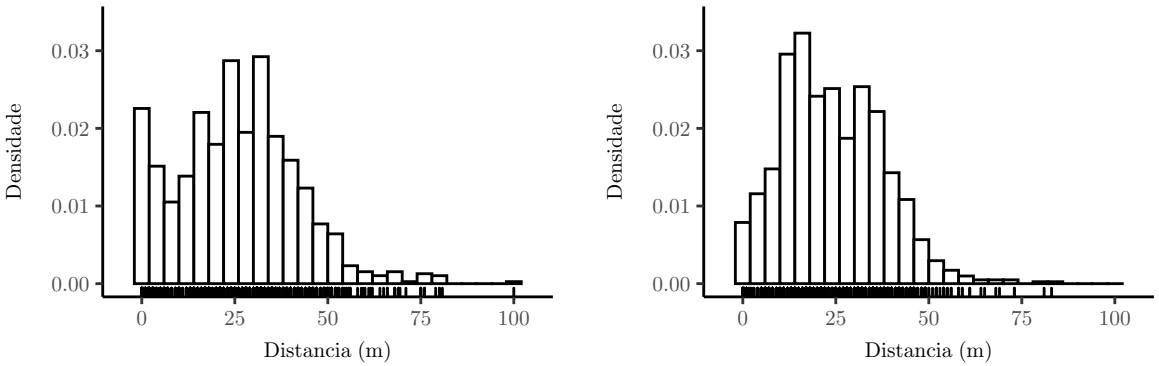


Figura 1: Histogramas cos datos dos falecementos por colisións contra aeroxeradores (Choi et al., 2020). A variable representada en ambos os histogramas é a distancia horizontal entre o cadáver atopado e o aeroxerador máis próximo, medida en metros. Á esquerda, o histograma cos datos dos falecementos de aves. Á dereita, os dos morcegos.

Como as modas son os máximos locais da función de densidade da variable aleatoria, o problema de estimar o número de modas dunha poboación e a súa localización enlaza directamente co de estimar

a súa densidade. Unha das principais técnicas non paramétricas para estimar a función de densidade dunha variable aleatoria é a estimación tipo núcleo (Wand and Jones, 1995, Capítulo 2). Dada unha mostra aleatoria simple $X_1, X_2, \ldots, X_n$ dunha variable aleatoria X absolutamente continua con función de densidade f, defínese a estimación tipo núcleo de f como

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

onde K é unha función de densidade unimodal e simétrica con media cero e h é un parámetro positivo. K denomínase habitualmente *función núcleo* ou *kernel*, metres que h recibe o nome de *xanela* ou *parámetro de suavización* (*bandwidth* en inglés). A escolla do kernel K non ten un grande impacto na estimación da función de densidade e resulta habitual escoller o denominado kernel gaussiano, é dicir, escoller como núcleo a normal estándar:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{2}$$

O que si ten un impacto significativo na estimación é o valor de h: distintos valores do parámetro de suavizado poden dar lugar a estimacións moi diferentes da función de densidade. Trasladando isto ao problema de estimación do número de modas, pode ser que dous valores distintos de h dean lugar a dúas estimacións de f cun número de modas distinto. É máis, Silverman (1981) proba que o número de modas de $\hat{f}_h$ é unha función decrecente de h, sempre que empregemos o kernel gaussiano na estimación. Tal comportamento vese reflectido na Figura 2, onde representamos a estimación tipo núcleo da densidade dos datos de Choi et al. (2020) para distintos valores do parámetro h.

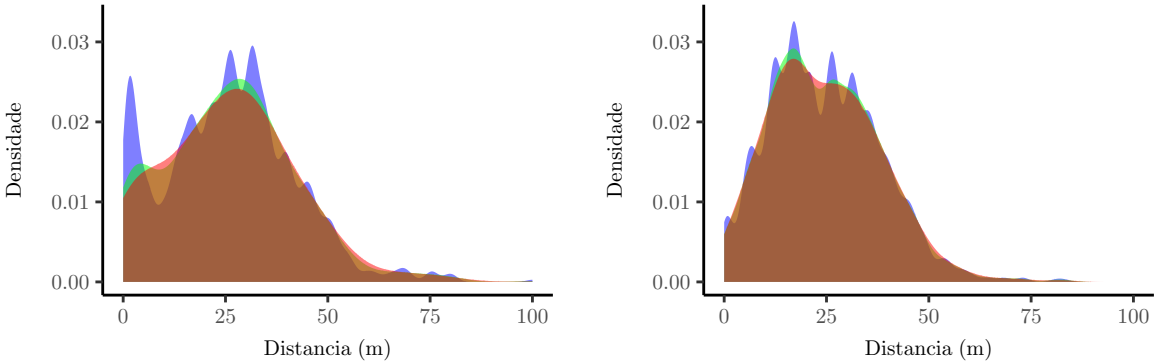


Figura 2: Estimación tipo núcleo da densidade con varios valores do parámetro de suavizado. Os datos empregados son os de Choi et al. (2020). A gráfica da esquerda refírese aos datos do grupo de aves, mentres que a da dereita trabállase coa de morcegos. Na gráfica da esquerda está representada a función $\hat{f}_h$ para os valores $h = 6$ (vermello), $h = 4.5$ (verde) e $h = 1.5$ (azul). Na gráfica da dereita os valores da xanela escollidos son $h = 3.5$ (vermello), $h = 2.5$ (verde) e $h = 1.5$ (azul).

Que o número de modas de $\hat{f}_h$ varíe en función de h é problemático, pois o noso obxectivo é precisamente estimar cantas modas hai na poboación orixinal. Fixémonos, por exemplo, na gráfica da esquerda da Figura 2, onde as tres estimacións tipo núcleo da función de densidade teñen un número distinto de modas. Parece claro o valor $h = 1.5$ (liña verde) dá lugar a unha estimación da densidade infrasuavizada, e por tanto a gran cantidade de modas detectada neste caso é artificial, especialmente as da cola dereita, que seguramente se deban a presenza de datos atípicos na mostra. En troques, as outras dúas estimacións da función de densidade non parecen ter ese problema de infrasuavización, pero o número de máximos locais é distinto nos dous casos: a estimación con $h = 6$ é unimodal, mentres que a correspondente ao valor $h = 4.5$ é bimodal. Daquela, cal será a nosa estimación do número de

modas da densidade poboacional a partir desta mostra? Con que argumentos decidimos entre unha ou outra?

Un primeiro intento de atallar o problema consiste en buscar un valor de h *óptimo*, de xeito que o erro global cometido ao aproximar f por $\hat{f}_h$ sexa o menor posible. Na literatura estatística hai varias técnicas para tratar de escoller de forma óptima este parámetro, denominadas selectores do parámetro de suavizado. Algúns dos selectores máis comúns son a regra do pulgar, validación cruzada sen nesgo ou a xanela plug-in e Sheater e Jones (Wand and Jones, 1995, Capítulo 3). Pero estes métodos só se preocupan de minimizar o erro global da estimación, non de que $\hat{f}_h$ reflicta correctamente as propiedades locais da poboación orixinal, como pode ser precisamente o número de modas. Por tanto esta non parece a maneira axeitada coa que enfocar o noso poblema.

Outra forma de intertar aproximar o número de modas de f consiste en empregar varios valores do parámetro de suavizado h en vez de centrarse só nun e explorar como se comporta o estimador $\hat{f}_h$ para cada un deles en canto ao número de modas. Se a poboación orixinal ten j modas, cabe esperar que $\hat{f}_h$ tamén teña j modas para unha gran cantidade de valores de h. Por tanto, poderíamos representar e estudar o comportamento de $\hat{f}_h$ para varios valores da xanela e empregar esta información para tratar de estimar o número de modas de f e a súa localización. Existen diversas ferramentas gráficas baseadas nesta idea, como son *mode tree* (Minnotte e Scott, 1993), *mode forest* (Minnotte et al., 1998), e *SiZeR* (Chaudhuri e Marron, 1999). Por exemplo, na Figura 3 represéntase os mapas SiZeR dos datos de Choi et al. (2020). Cada punto $(x, \log(h))$ do mapa SiZeR asóciase cunha cor distinta de tres posibles: azul, vermella e morada. O proceso de selección da cor consiste en empregar $\hat{f}'_h(x)$ para facer inferencia sobre o signo do seu valor esperado $\mathbb{E}(\hat{f}'_h(x))$, e escóllese a cor dependendo do resultado. Así, o punto $(x, \log(h))$ do mapa SiZeR será azul se temos probas estatísticas de que $\mathbb{E}(\hat{f}'_h(x))$ é significativamente maior que cero, vermello se $\mathbb{E}(\hat{f}'_h(x))$ é significativamente menor que cero, e morado se non se pode asegurar con suficiente significación estatística que sexa distinta de cero. Así, vendo os mapas SiZeR de esquerda a dereita, os cambios da cor azul á vermella indicarán a presenza de modas, mentres que as antimodas (os mínimos locais da función de densidade) daranse en cambios de vermello a azul. Ademais, para valores pequenos do parámetro de suavizado tamén se inclúe a cor gris. Unha rexión será de cor gris se cantidade de datos presente nesa rexión é insuficiente para poder establecer conclusións firmes sobre o signo da derivada.

Na parte esquerda da Figura 3 está representado o mapa SiZeR dos datos de Choi et al. (2020) correspondentes aos falecementos de aves. Nel observamos que, para valores grandes do parámetro de suavizado, a estimación tipo núcleo presenta só unha moda, situada no intervalo $[20, 30]$, que se reflicte no mapa SiZeR nunha rexión azul seguida doutra vermella. Reducindo o valor de h aparece unha segunda moda no intervalo $[0, 10]$, que se traduce nunha mancha vermella dentro da rexión azul anterior. Se seguimos diminuíndo o tamaño da xanela empezan a aparecer multiples modas, debido a estimacións infrasuavizadas da función de densidade. Por tanto, o mapa SiZeR apunta a que a distribución das distancias entre os cadáveres das aves e o aeroxerador máis próximo é bimodal, pois a rexión de valores de h que detectan dúas modas parece estar nunha zona de transición de estimacións da densidade sobresuavizadas (cunha moda moi clara) a estimacións infrasuavizadas (con moitas modas espurias).

O mapa SiZeR dos datos dos morcegos está na parte dereita da Figura 3. A lectura do mesmo é bastante similar a do anterior: valores grandes de h dan lugar a funcións $\hat{f}_h$ unimodais, acurtando un pouco a xanela aparece unha segunda moda á dereita da anterior, e conforme reducimos o valor de h vai xurdindo unha gran cantidade de modas debido a estimacións da función de densidade infrasuavizadas. Pero aquí non se presenta unha zona de transición tan clara entre as estimacións da densidade sobresuavizadas e infrasuavizadas: unha vez aparece a segunda moda rapidamente aparecen todas as demais. Por tanto, parece que esa segunda moda detectada tamén é un efecto da infrasuavización da estimación da densidade, e por tanto concluiríamos que a distibución das distancias nos morcegos é unimodal.

Nos dous parágrafos anteriores evidénciase o principal problema que presenta o SiZeR, e que comparte coas ferramentas exploratorias *mode tree* e *mode forest*: dependen fortemente do criterio e expe-

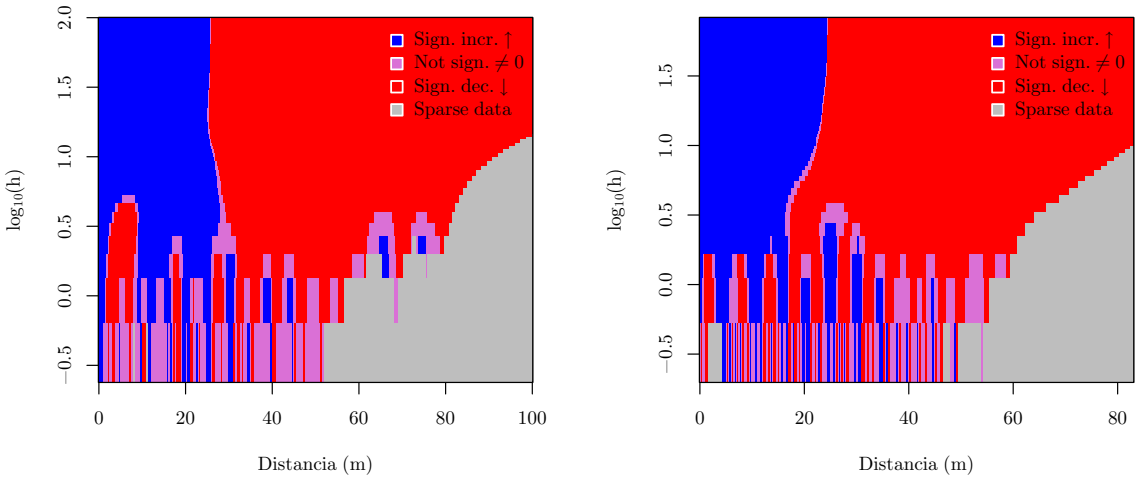


Figura 3: Mapas SiZeR asociados aos datos de falecementos de aves e morcegos por colisións contra aeroxeradores (Choi et al., 2020). O da esquerda correspóndese aos datos das aves. O da dereita, aos dos morcegos. Os mapas SiZeR obtiveronse en R empregando a librería `multimode` (Ameijeiras et al., 2021).

riencia de quen as está a usar. Isto é problemático, pois por un lado, é posible que dúas persoas distintas cheguen a conclusións diferentes sobre o número de modas a partir dos mesmos datos e empregando a mesma técnica, e por outro, resulta imposible aplicar estas ferramentas a un gran número de mostras nun tempo razoable. Habería que buscar, por tanto, un método sistemático e obxectivo para detectar o número de modas dunha distribución a partir dunha mostra aleatoria, isto é, un contraste estatístico.

Polo de agora, só estivemos falando do problema de detección do número de modas para os denominados *datos lineais*, onde cada observación da mostra é un número real. Pero este problema non se restrinxe unicamente ao caso no que a variable aleatoria a estudar teña por soporte a recta real. É posible definir o concepto de moda para datos multidimensionais, ou para datos dentro dunha variedade, como unha circunferencia, unha esfera ou un toro. Destes últimos, o caso máis estudado é o de *datos circulares*, onde cada observación da mostra se corresponde cun punto da circunferencia. Este tipo de datos xorden de xeito natural ao medir ángulos (a dirección do vento en meteoroloxía), ou ao contabilizar eventos ao longo do día (os tempos de ingreso dos pacientes nunha unidade de coidados intensivos) ou do ano (as datas dos incendios producidos nunha zona determinada). Por tanto, igual que no caso linear, precisaríamos dun contraste que nos permita detectar a cantidade de modas cando traballamos na circunferencia.

O obxectivo principal deste traballo é a construción dun test de multimodalidade que sexa facilmente adaptable a espazos de carácter xeral, poñendo especial atención aos casos lineais e circular, e procurando tamén que sexa competitivo co resto dos tests presentes na literatura. O traballo está estuturado do seguinte xeito. No Capítulo 1 intoduciremos os principais tests estatísticos para contrastar a existencia dun número determinado de modas para datos lineais presentes na literatura, e realizaremos unha comparativa crítica dos tests presentados. Para rematar o capítulo, estudaremos se hai probas estatísticas significativas para rexeitar a unimodalidade dos datos presentados nesta introdución, empregando os únicos tres tests de multimodalidade que ofrecen un comportamento aceptable. No Capítulo 2 definiremos a función de pseudo-verosimilitude, e construiremos un novo test de multimodalidade para datos lineais baseado nela. Comentaremos como se podería adaptar este test a outros espazos distintos da recta real, e realizaremos un estudo de simulación para analizar o

seu comportamento na práctica, comparándoo cos tests de multimodalidade introducidos no Capítulo 1. Comezaremos o Capítulo 3 introducindo un exemplo con datos reais para motivar a necesidade dun test de multimodalidade para datos circulares. Continuaremos definindo brevemente os conceptos básicos necesarios para traballar con datos circulares, e extenderemos o test definido no Capítulo 2 a este contexto. Acabaremos o capítulo realizando un estudo de simulación para ver como se comporta o test de multimodalidade para datos circulares na práctica. A modo de conclusión, no Capítulo 4 realizaremos unha breve discusión sobre o camiño a seguir para continuar estudando o novo test de multimodalidade, así como as posibles problemáticas que poden xurdir á hora de extendelo a outros espazos.

Debo expresar a miña gratitude co Centro de Supercomputación de Galicia (CESGA), pois os seus recursos computacionais foron imprescindibles para a realización de todos os estudos de simulación deste traballo. Tamén debo dar as grazas a Jose Ameijeiras Alonso por facilitarnos os datos dos incendios empregados no inicio do Capítulo 3, así como os datos de Choi et al. (2020) presentados nesta introdución.

Capítulo 1

Compendio sobre tests de multimodalidade

Neste capítulo recolleremos os principais test de multimodalidade non paramétricos presentes na literatura estatística. Introduciremos cales son as ideas nas que se basea cada test para construír o estatístico de contraste, e explicaremos as técnicas que empregan para aproximar a distribución nula do mesmo. Ademais, faremos unha comparativa do comportamento na práctica dos mesmos, indicando as vantaxes e desvantaxes de cada un deles fronte ao resto.

Antes de comezar, precisamos concretar a que nos referimos cando falamos de *test de multimodalidade*. Dada X unha variable aleatoria absolutamente continua con j modas, un test de multimodalidade é un test estatístico que contrasta as hipóteses

$$H_0 : j \leq k \text{ fronte a } H_1 : j > k; \tag{1.1}$$

onde k é un número natural fixado de antemán. O máis común destes tests é o que contrasta unimodalidade fronte a multimodalidade, é dicir

$$H_0 : j = 1 \text{ fronte a } H_1 : j > 1. \tag{1.2}$$

Na literatura existen varios procedementos para realizar este tipo de tests. Algúns deles empregan técnicas paramétricas para construír o estatístico de contraste, por exemplo modelizando a poboación como unha mestura de distribucións, tal como se expón en McLachlan and Peel (2000). Con todo, esta forma de proceder depende moito de se o modelo paramétrico se axusta ben á mostra coa que se traballa ou non. Por iso nós imos centrarnos en contrastes non paramétricos, que non impoñen modelos a priori sobre os datos. Ao longo deste capítulo iremos vendo cinco tests de multimodalidade non paramétricos que se apoian en dúas ideas centrais: a xanela crítica e o exceso de masa.

1.1. Tests baseados na xanela crítica

Os tests baseados na xanela crítica teñen o seu xerme na proposta de Silverman (1981), quen introduce este concepto. A idea provén da propiedade de monotonía no número de modas da estimación tipo núcleo da densidade que xa comentamos na introdución: o número de modas do estimador $\hat{f}_h$ definido en (1) é unha función decrecente en h, sempre que o núcleo K empregado na estimación sexa o gaussiano. Esta monotonía permítelle a Silverman (1981) definir o concepto de xanela crítica, *critical bandwidth* en inglés. Así, a xanela crítica para k modas, h_k, defínese como o menor parámetro de suavización h tal que $\hat{f}_h$ ten k modas:

$$h_k = \text{mín}\{h : \hat{f}_h \text{ ten como máximo } k \text{ modas}\}. \tag{1.3}$$

Por tanto, h_k marca a fronteira entre as estimacións tipo núcleo con máis e menos de k modas: se h é menor que h_k, $\hat{f}_h$ ten máis de k modas, pero se h é maior que h_k, o número de modas de $\hat{f}_h$ é menor ou igual que k. Na Figura 1.1 están representadas as estimacións $\hat{f}_h$ para os datos de Choi et al. (2020) nas tres primeiras xanelas críticas, h_1, h_2 e h_3, correspondentes a unha, dúas e tres modas respectivamente.

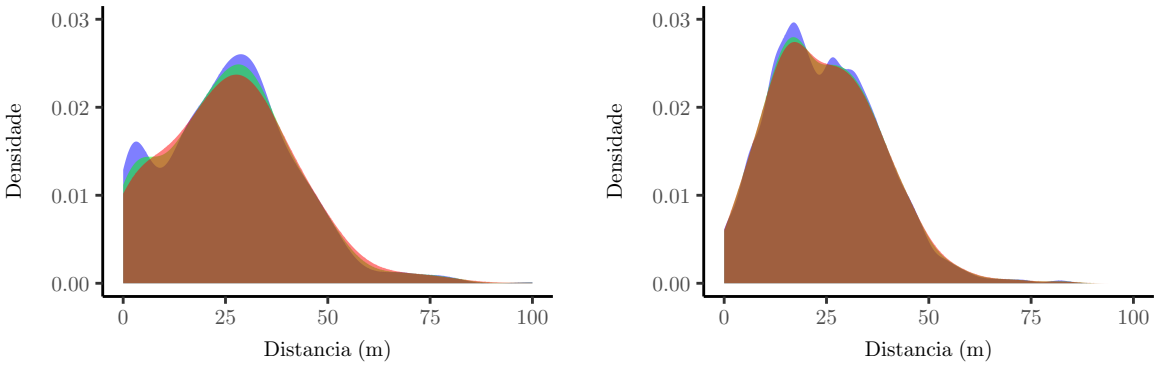


Figura 1.1: Xanelas críticas e estimadores tipo núcleo da densidade asociados para os datos de Choi et al. (2020). Na gráfica da esquerda están representados as estimacións tipo núcelo dos datos correspondentes aos falecementos de aves para as tres primeiras xanelas críticas. Os seus valores son: $h_1 = 6.53$ (vermello), $h_2 = 5.11$ (verde) e $h_3 = 3.69$ (azul). Na da dereita, os datos cos que se traballa son os do grupo de morcegos. Os valores das xanelas críticas neste caso son $h_1 = 3.93$ (vermello), $h_2 = 3.43$ (verde) e $h_3 = 2.24$ (azul). Todos os datos están aproximados ao segundo decimal. As xanelas críticas calculáronse en R empregando a librería `multimode` (Ameijeiras et al., 2021).

1.1.1. Test de Silverman

Silverman (1981), ademais de presentar a xanela crítica, tamén introduce un test de multimodalidade baseado nela. A súa idea consiste en empregar a xanela crítica como estatístico para contrastar a hipótese nula (1.1), rexeitándoa para valores grandes de h_k. O razoamento subxacente é que valores de h_k grandes indican que hai que sobresuavizar moito a estimación kernel da densidade para obter unha estimación con k modas, o que apunta a que o número de modas da poboación orixinal é maior.

O test calíbrase mediante bootstrap suavizado, xerando remostras da variable aleatoria con función de densidade g_0, onde g_0 é unha versión de $\hat{f}_{h_k}$ reescalada para que a súa varianza coincida coa varianza da mostra. Simular remostras a partir da densidade g_0 é doado, pois, tal e como indica Silverman (1981), as observacións independentes $X_1^*, X_2^*, \ldots, X_n^*$ de g_0 veñen dadas por

$$X_i^* := \left(1 + \frac{h_k^2}{s^2}\right)^{-1/2} (X_{\mathcal{I}(i)} + h_k \varepsilon_i), \quad i = 1, \ldots, n; \tag{1.4}$$

onde a escolla de índicies $\mathcal{I}(i)$ faise de forma independente e con reemprazamento de $\{1, \ldots, n\}$, os erros ε_i son independentes entre si e seguen unha distribución normal de media cero e varianza un, e s^2 é a varianza mostral.

A partir das remostras simuladas calcúlanse as réplicas do estatístico de contraste $h_k^{*,1}, h_k^{*,2}, \ldots, h_k^{*,B}$. Así, o p-valor asociado á mostra aproxímase pola proporción de réplicas maiores que a xanela crítica da mostra $X_1, X_2, \ldots, X_n$

$$\frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(h_k < h_k^{*,b}). \tag{1.5}$$

Por tanto, o test con nivel de significación α rexeitará a hipótese de que a poboación orixinal ten como máximo k modas se a cantidade anterior é menor que α.

Daquela, a implementación esquemática do test de Silverman (1981) con nivel de significación $\alpha \in [0,1]$ é:

1. Calculamos a xanela crítica h_k a partir da mostra orixinal $X_1, X_2, \ldots, X_n$. Tamén calculamos a varianza da mostra, $s^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

2. Simulamos a remostra $X_1^*, X_2^*, \ldots, X_n^*$ a partir da densidade $\hat{f}_{h_k}$ modificada mediante:

$$X_i^* := \left(1 + \frac{h_k^2}{s^2}\right)^{-1/2} (X_{\mathcal{I}(i)} + h_k \varepsilon_i), \quad i = 1, \ldots, n; \tag{1.6}$$

   onde os índices $\mathcal{I}(i)$ son escollidos uniformemente e con reemplazamento de $1, \ldots, n$ e os erros ε_i son independentes entre si e seguen unha distribución normal de media cero e varianza un.

3. A partir da remostra $X_1^*, X_2^*, \ldots, X_n^*$ calcúlase a réplica da xanela crítica h_k^*.

4. Repítense os dous pasos anteriores un número grande de veces B, obtendo así B réplicas do estatístico de contraste: $h_k^{*,1}, h_k^{*,2}, \ldots, h_k^{*,B}$.

5. O test rexeitará a hipótese nula se

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(h_k < h_k^{*,b}) < \alpha. \tag{1.7}$$

1.1.2. Test de Hall e York

O principal problema do test de Silverman (1981) é que a distribución de h_1 non depende só das modas da densidade orixinal f, senón que está moi influenciada polas colas da distribución. Isto débese a que $\hat{f}_{h_1}$ tende a detectar modas espúreas en datos atípicos, dando como resultado valores de h_1 demasiado grandes. Por exemplo, Hall e York (2001) comentan que se a densidade f se corresponde cunha t de Student con ν graos de liberdade, entón h_1 converxe a $+\infty$ ao aumentar o tamaño da mostra n, independentemente de cal sexa o valor de ν. Na Figura 1.2 represéntase o estimador $\hat{f}_{h_1}$ para dúas mostras dunha t de Student de 4 graos de liberdade de dous tamaños distintos. Como se pode observar, o aumento do tamaño da mostra produce un aumento no tamaño da xanela crítica, dando lugar a un estimador da densidade sobresuavizado.

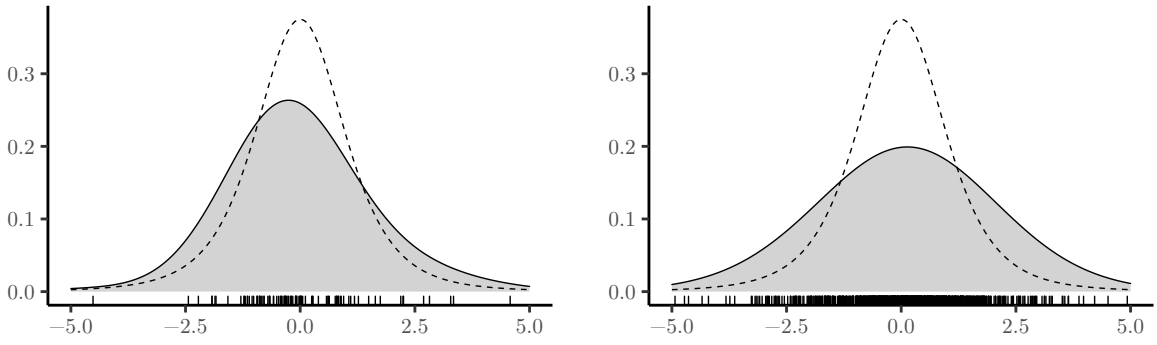
Para atallar este problema Hall e York (2001) propoñen variar lixeiramente a hipótese nula a contrastar. En vez de buscar modas en toda a recta real, concentrarémonos tan só nun intervalo compacto I. Con isto esperamos evitar o problema que tiña o test de Silverman (1981) para distribucións con colas pesadas. Así, a nova hipótese nula é:

$$H_0 : f \textit{ ten unha única moda dentro do intervalo compacto } I \textit{ e ningún mínimo local en } I; \tag{1.8}$$

e por tanto a xanela crítica redefínese para ser coherente coa nova hipótese nula:

$$h_{HY} = \text{mín}\{h : \hat{f}_h \text{ ten exactamente unha moda en } I\}. \tag{1.9}$$

Este cambio na hipótese nula podería dar problemas a hora de calcular a xanela crítica, pois o número de modas de $\hat{f}_h$ dentro do intervalo I non ten porque ser necesariamente unha función decrecente de h (incluso cando o kernel K empregado na estimación da densidade é o gaussiano), o que pode complicar o cálculo de h_{HY}. Pero, baixo condicións de regularidade pouco restrictivas, Hall e York (2001) proban que a probabilidade de que o número de modas de $\hat{f}_h$ dentro do intervalo I sexa decrecente en h converxe a 1 para o núcleo gaussiano. Tendo isto en conta, Hall e York (2001)



(a) $n = 100, h_1 = 1.063$; (b) $n = 1000, h_1 = 1.615$.

Figura 1.2: Represencación do estimador da densidade $\hat{f}_{h_1}$ para mostras t de Student de 4 graos de liberdade para 2 tamaños distintos. En trazo continuo está representada $\hat{f}_{h_1}$, e en trazo discontinuo a densidade orixinal da mostra. Debaixo de cada gráfica está representado o tamaño n da mostra e o valor de h_1 corespondente.

empregan h_{HY} como estatístico de contraste para un test de unimodalidade, rexeitando a hipótese nula H_0 para valores grandes do mesmo. Igual que na proposta de Silverman (1981), a distribución de h_{HY} baixo a nula aproxímase mediante bootstrap suavizado, remostrando a partir da función de densidade $\hat{f}_{h_{HY}}$ reescalada para que a súa varianza coincida coa da mostra.

Este método de calibrado ten un problema grave, e é que as remostras bootstrap h^*_{HY} do estatístico de contraste h_{HY} son dunha orde maior que o valor orixinal, polo que é preciso reescalar o estatístico orixinal mediante un factor de correción λ_α para poder realizar o contraste con nivel de significación α. Así, o test de Hall e York (2001) rexeitará a unimodalidade da mostra se

$$\mathbb{P}(\lambda_\alpha h_{HY} < h^*_{HY}) < \alpha. \tag{1.10}$$

Hai que salientar que este parámetro λ_α é o mesmo (asintoticamente) para todas as distribucións unimodais e non depende de ningún parámetro descoñecido. Con todo, e preciso estimar o seu valor na práctica para poder realizar o test. Hall e York (2001) propoñen dous métodos para logralo: un determinista, que pretende aproximar o valor asintótico de λ_α, e outro baseado en técnicas Monte Carlo. Aquí centrarémonos só no primeiro destes dous métodos, que consiste en buscar unha aproximación de λ_α como función de α mediante unha función racional. Para obter esa aproximación, Hall e York (2001) simularon mostras de tamaño $n = 10000$ dunha normal estándar. É importante que o tamaño da mostra sexa tan grande, pois o factor de corrección λ_α é un valor asintótico. Por outra banda, que a distribución escollida sexa a normal estándar ou outra non ten un grande impacto, pois, como xa se dixo, λ_α é o mesmo para todas as distribucións unimodais. En total simularon 5000 mostras normais estándar. Despois aplicáronlle o test a cada mostra realizando 5000 remostras, e calcularon o valor $\hat{\lambda}_\alpha$ que conseguía que o test esté correctamente calibrado para os niveis de significación $\alpha = 0.001, 0.002, \ldots, 0.999$. Con eses 999 valores, Hall e York (2001) axustaron unha curva racional da forma

$$\lambda_\alpha = \frac{a_1\alpha^3 + a_2\alpha^2 + a_3\alpha + a_4}{\alpha^3 + a_5\alpha^2 + a_6\alpha + a_7} \tag{1.11}$$

para conseguir a aproximación de λ_α desexada en función de $\alpha \in (0, 1)$. Os valores $a_1, \ldots, a_7$ obtidos por Hall e York (2001) están recollidos na Táboa 1.1. A Figura 1.3 representa a aproximación do factor de corrección dada pola expresión (1.11). Aí observamos que λ_α toma valores maiores que un na maior parte dos casos (salvo cando α é próximo a un), o que nos indica que as réplicas bootstrap da xanela crítica tenden a ser máis grandes que o valor orixinal.

O test anterior pódese extender para contrastar a existencia de k modas en I, tal como amosan

a_1	a_2	a_3	a_4	a_5	a_6	a_7
0.94029	-1.59914	0.17695	0.48971	-1.77793	0.36162	0.42423

Táboa 1.1: Coeficientes estimados por Hall e York (2001) para o calculo de λ_α.

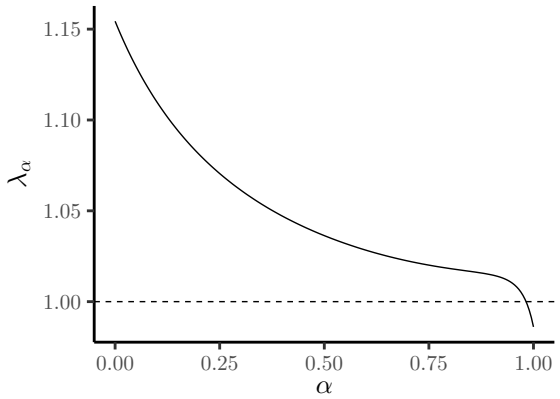


Figura 1.3: Aproximación do factor de corrección λ_α en función de α dada pola expresión (1.11).

Hall e York (2001). A idea consiste en empregar a xanela crítica

$$h_{HY,k} = \text{mín}\{h : \hat{f}_h \text{ ten como máximo } k \text{ en } I\} \tag{1.12}$$

como estatístico de contraste e proceder de xeito análogo. Mais se $k \geq 2$, entón Hall e York (2001) proban que a distribución asintótica de $h_{HY,k}$ depende dos parámetros descoñecidos $c_l = f(t_l)^{1/5}/\left|f''(t_l)\right|^{2/5}$ con $l = 1, \ldots, 2k-1$, onde $t_1, \ldots, t_{2k-1}$ son os puntos críticos de f no intervalo I. Isto fai que o calibrado mediante bootstrap suavizado non aproxime correctamente a distribución nula de $h_{HY,k}$ cando $k \geq 2$.

A modo de resumo, rematamos esta sección dando unha implementación esquemática do test de Hall e York (2001) para contrastar a unimodalidade dos datos con nivel de significación α:

1. Calculamos a xanela crítica $h_{HY,1}$ a partir da mostra orixinal $X_1, X_2, \ldots, X_n$. Tamén calculamos a varianza da mostra, $s^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

2. Simulamos a remostra $X_1^*, X_2^*, \ldots, X_n^*$ a partir da densidade $\hat{f}_{h_{HY,1}}$ modificada mediante:

$$X_i^* := \left(1 + \frac{h_{HY,1}^2}{s^2}\right)^{-1/2} (X_{\mathcal{I}(i)} + h_{HY,1}\varepsilon_i), \quad i = 1, \ldots, n; \tag{1.13}$$

   onde os índices $\mathcal{I}(i)$ son escollidos uniformemente e con reemplazamento de $\{1, \ldots, n\}$, os erros ε_i son independentes entre si e seguen unha distribución normal de media cero e varianza un, e s^2 é a varianza da mostra orixinal.

3. A partir da remostra $X_1^*, X_2^*, \ldots, X_n^*$ calcúlase a réplica da xanela crítica $h_{HY,1}^*$.

4. Repítense os dous pasos anteriores un número grande de veces B, obtendo así B réplicas do estatístico de contraste: $h_{HY,1}^{*,1}, \ldots, h_{HY,1}^{*,B}$.

5. Calcúlase o factor de corrección λ_α mediante a fórmula:

$$\lambda_\alpha = \frac{a_1\alpha^3 + a_2\alpha^2 + a_3\alpha + a_4}{\alpha^3 + a_5\alpha^2 + a_6\alpha + a_7} \tag{1.14}$$

   onde os coeficientes $a_1, \ldots, a_7$ son os dados na Táboa 1.1.

6. O test rexeita a unimodalidade dos datos se

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\lambda_\alpha h_{HY,1} < h^{*,b}_{HY,1}) < \alpha. \tag{1.15}$$

1.2. Tests baseados no exceso de masa

Outra perspectiva coa que enfocar o contraste de multimodalidade é empregando o denominado exceso de masa. Detrás deste concepto está a idea de que en cada moda a función de densidade ten *bultos*, rexións onde a función de desidade se eleva de forma considerable. Así, co exceso de masa estamos a medir o volume deses bultos cortándoos a partir de certa altura.

O exceso de masa dunha densidade f, tal e como o definen Müller e Sawitzki (1991), é a función real non negativa

$$E(\lambda) = \int_{C_\lambda} f(x)dx - \lambda\|C_\lambda\| = \mathbb{P}_f(C_\lambda) - \lambda\|C_\lambda\|, \tag{1.16}$$

onde $C_\lambda = \{x : f(x) \geq \lambda\}$ e $\|C_\lambda\|$ é a medida do conxunto C_λ. O que esta a medir $E(\lambda)$ é a area da rexión $\{(x, y) : \lambda < y < f(x)\}$ tal como se ilustra na Figura 1.4.

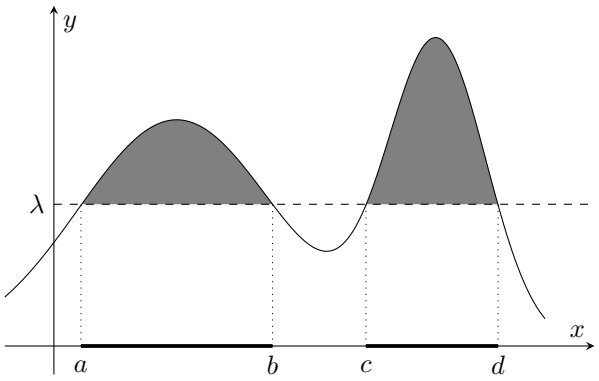


Figura 1.4: Exceso de masa dunha función de densidade f a nivel λ. $E(\lambda)$ mide a área da rexión sombreada. Neste caso, o conxunto C_λ está formado pola unión dos dous intervalos disxuntos (a, b) e (c, d) resaltados na gráfica, cada un asociado a unha moda de f.

Que f teña k modas significa que f ten k bultos, cada un asociado a cadansúa moda. Polo tanto, dado $\lambda > 0$, o conxunto C_λ estará formado por un máximo de k compoñentes conexas distintas, onde cada unha delas se corresponde a un bulto ou moda de f. Daquela, se supoñemos que f é k-modal, o exceso de masa pódese expresar como:

$$E(\lambda) = \sup_{C_1,\ldots,C_k} \left\{ \sum_{c=1}^{k} \mathbb{P}_f(C_c) - \lambda\|C_c\| \right\}; \tag{1.17}$$

onde o supremo anterior se calcula en todas as familias de k conxuntos conexos disxuntos dous a dous, $\{C_1, \ldots, C_k\}$. Pero a cantidade do lado dereito de (1.17) pódese estimar empiricamente, simplemente

hai que intercambiar a probabilidade de f pola probabilidade asociada a distribución empírica. Así, dada unha mostra, o estimador natural de $E(\lambda)$ sería:

$$E_{n,k}(\lambda) = \sup_{C_1,\ldots,C_k} \left\{ \sum_{c=1}^{k} \mathbb{P}_n(C_c) - \lambda \|C_c\| \right\}; \tag{1.18}$$

onde $\mathbb{P}_n$ é a probabilidade asociada a distribución empírica e, como antes, o supremo cálculase en todas as familias de k conxuntos conexos disxuntos dous a dous, $\{C_1, \ldots, C_k\}$.

Pero a cantidade $E_{n,k}(\lambda)$ só será un estimador consistente de $E(\lambda)$ cando a densidade f teña k modas como máximo. No caso de que f teña máis de k modas $E_{n,k}(\lambda)$ estará infraestimando o valor $E(\lambda)$, pois estamos a supoñer que o conxunto C_λ ten menos compoñentes conexas das que realmente ten. Daquela, se o número de modas de f é menor ou igual que k, entón tanto $E_{n,k+1}(\lambda)$ como $E_{n,k}(\lambda)$ son bos estimadores de $E(\lambda)$ e deberían tomar valores parecidos. Se, pola contra, f ten máis de k modas, entón $E_{n,k}(\lambda)$ é máis pequeno que $E(\lambda)$, e engadir un conxunto máis no supremo de (1.18) debería de mellorar a estimación. Así, a diferenza entre o exceso de masa estimado con k conxuntos e con $k+1$, $D_{n,k+1}(\lambda) = E_{n,k+1}(\lambda) - E_{n,k}(\lambda)$, expresa canto se afasta f de cumprir a hipótese nula (1.1).

Tendo en conta o anterior, os test de multimodalidade baseados no exceso de masa empregan como estatístico de contraste a cantidade:

$$\Delta_{n,k+1} = \max_{\lambda>0} D_{n,k+1}(\lambda) \tag{1.19}$$

e rexeitan a hipótese nula (1.1) para valores grandes de $\Delta_{n,k+1}$. No que segue, imos ver tres propostas deste tipo: o test de Müller e Sawitzki (1991), o test de Cheng e Hall (1998) e o test Ameijeiras et al. (2019). A diferenza entre estes tres contrastes está no método de remostraxe que empregan para aproximar a distribución baixo a hipótese nula do estatístico de exceso de masa. A proposta de Müller e Sawitzki (1991) calibra o test supoñendo que os datos proveñen dunha distribución uniforme. O test de Cheng e Hall (1998) emprega un método de remostraxe paramétrico, mentres que o de Ameijeiras et al. (2019) emprega bootstrap suavizado tomando como referencia o estimador tipo núcleo con xanela crítica.

1.2.1. Test de Müller e Sawitzki

O calibrado que propoñen Müller e Sawitzki (1991) para o test de multimodalidade baseado no exceso de masa é extremadamente simple: estiman os valores críticos da distribución nula de $\Delta_{n,k+1}$ mediante Monte Carlo, supoñendo que os datos proveñen dunha distribución uniforme en $[0,1]$. É dicir, simulan un número grande B de mostras uniformes do mesmo tamaño que a mostra orixinal, e a partir delas calculan B réplicas do estatístico, $\Delta^{*,1}_{n,k+1}, \Delta^{*,2}_{n,k+1}, \ldots, \Delta^{*,B}_{n,k+1}$. Despois estiman o valor crítico do test con nivel de significación α a partir das réplicas, escollendo κ_α tal que

$$\mathbb{P}(\Delta^*_{n,k+1} > \kappa_\alpha) = \alpha. \tag{1.20}$$

Müller e Sawitzki (1991) comentan que a desigualdade $\mathbb{P}(\Delta_{n,k+1} > \kappa_\alpha | F) < \alpha$ non se verifica para todas as distribucións unimodais F, especialmente cando o tamaño da mostra n é pequeno. Pero argumentan que a desigualdade si se cumpre para tamaños grandes apoiándose nos seus resultados teóricos e experimentais, e polo tanto o test estaría ben calibrado asintóticamente.

Así, a implementación do test de Müller e Sawitzki (1991) de nivel α, contada de forma esquemática, sería:

1. Dada a mostra orixinal $X_1, X_2, \ldots, X_n$, calcúlase o estatístico de exceso de masa: $\Delta_{n,k+1}$.
2. Simúlase a remostra $X^*_1, X^*_2, \ldots, X^*_n$ dunha densidade uniforme en $[0,1]$, e a partir dela calcúlase a réplica do estatístico $\Delta^*_{n,k+1}$.

3. Repítese o paso anterior un número grande B de veces, obtendo así B réplicas do estatístico do exceso de masa: $\Delta_{n,k+1}^{*,1}, \Delta_{n,k+1}^{*,2}, \ldots, \Delta_{n,k+1}^{*,B}$.

4. O test rexeita a hipótese nula se

$$\frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(\Delta_{n,k+1} < \Delta_{n,k+1}^{*,b}) < \alpha. \tag{1.21}$$

1.2.2. Test de Cheng e Hall

O calibrado do test de exceso de masa descrito por Müller e Sawitzki (1991) dá como resultado un contraste demasiado conservador, rexeitando a hipótese nula menos veces do que lle correspondería polo seu nivel de singificación. Para tratar de aliviar este problema Cheng e Hall (1998) introduciron un novo método de calibrado que está baseado no seguinte resultado: baixo a hipótese de que f é unimodal con moda en x_0 (e outras condicións de regularidade pouco restrictivas sobre f), a distribución asintótica de $\Delta_{n,2}$ só depende do factor

$$d = |f''(x_0)| / f^3(x_0). \tag{1.22}$$

Tendo en conta este feito, Cheng e Hall (1998) aproximaron a distribución de $\Delta_{n,2}$ mediante réplicas Monte Carlo, $\Delta_{n,2}^*$, que calculan a partir de mostras xeradas dunha distribución unimodal cun valor de d similar ao da distribución dos datos orixinais. Como o valor d dos datos orixinais é un parámetro descoñecido, primeiro témolo que estimar a partir da mostra. Isto pódese facer facilmente mediante estimación tipo núcleo. Se $\hat{x}_0$ denota a moda principal de $\hat{f}_h$, o estimador de d é:

$$\hat{d} = \frac{\left|\hat{f}_{h'}''(\hat{x}_0)\right|}{\hat{f}_h^3(\hat{x}_0)}; \tag{1.23}$$

onde $\hat{f}_h$ $\hat{f}_{h'}''$ son os estimadores tipo kernel de f e f'' con núcleo gausiano, e h e h' son os parámetros de suavizado globalmente óptimos, substituíndo todas as cantidades que dependen da densidade teórica descoñecida polas correspondentes á distribución normal $N(0, s^2)$ onde s^2 é a varianza da mostra (é dicir, h e h' son as xanelas da regra do pulgar para f e f'' respectivamente).

Unha vez calculado estimador $\hat{d}$ escóllese unha distribución unimodal con $d = \hat{d}$ para facer remostraxe e calcúlanse as réplicas do estatístico de exceso de masa, $\Delta_{n,2}^*$, a partir das remostras xeradas. Así, o test con nivel de significación α rexeitará a unimodalidade dos datos se $\mathbb{P}(\Delta_{n,2}^* > \Delta_{n,2}) < \alpha$.

En principio hai liberdade para escoller a distribución coa que se remostra, sempre que o seu valor de d coincida co estimado a partir da mostra orixinal $\hat{d}$. Cheng e Hall (1998) empregan as seguintes distribucións, que abranguen todo o rango de valores posibles para $\hat{d}$:

1. Se $\hat{d} < 2\pi$, empregan unha distribución Beta(β, β), onde escollen o parámetro β para que o seu valor de d coincida con $\hat{d}$.

2. Se $\hat{d} = 2\pi$, lanzan as remostras dunha distribución normal calquera.

3. Se $\hat{d} > 2\pi$, utilizan unha distribución t de Student reescalada para facer a remostraxe. É dicir, simulan as remostras dunha variable aleatoria con función de densidade

$$g_\beta(x) = \frac{1}{\mathrm{B}(\beta - \frac{1}{2}, \frac{1}{2})}\frac{1}{(1+x^2)^\beta}; \tag{1.24}$$

onde

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

e o parámetro β se escolle para que o valor de d da distribución coincida co estimado a partir da mostra, $\hat{d}$.

Por último, Cheng e Hall (1998) comentan que este calibrado se podería adaptar para contrastar a hipótese máis xeral de que f ten k modas fronte a que f ten máis de k modas. Simplemente hai que calcular os estimadores $\hat{d}_l$ dos valores

$$d_l = |f''(x_l)| / f^3(x_l), \tag{1.25}$$

con $1 \leq l \leq 2k-1$, onde x_l denota o l-ésimo punto crítico da densidade f. Despois xéranse as remostras dunha distribución con $2k-1$ puntos críticos que verifique $d_l = \hat{d}_l$ para $1 \leq l \leq 2k-1$. Finalmente calcúlanse as remostras $\Delta^*_{n,k+1}$ para aproximar a distribución baixo a hipótese nula de $\Delta_{n,k+1}$, e o test rexeita a hipótese nula se

$$\mathbb{P}(\Delta^*_{n,k+1} > \Delta_{n,k+1}) < \alpha, \tag{1.26}$$

onde α é o nivel nominal do test. Con todo, a procura de distribucións paramétricas que cumpran o requerido para poder calibrar o test é excesivamente complexa, facendo que na práctica só se empregue este test para contrastar unimodalidade.

Como resumo, damos unha implementación do test de Cheng e Hall (1998) de nivel α para contrastar unimodalidade fronte multimodalidade.

1. Dada a mostra orixinal $X_1, X_2, \ldots, X_n$, calcúlase o estatístico de exceso de masa: $\Delta_{n,2}$.

2. Calcúlase o valor $\hat{d}$ da mostra orixinal mediante

$$\hat{d} = \frac{\left|\hat{f}''_{h'}(\hat{x}_0)\right|}{\hat{f}^3_h(\hat{x}_0)}; \tag{1.27}$$

   onde $\hat{f}_h$ $\hat{f}''_{h'}$ son os estimadores tipo kernel de f e f'', h e h' son os parámetros de suavizado baseados na regra do pulgar en cada caso, e $\hat{x}_0$ é a moda principal de $\hat{f}_h$.

3. Escóllese a distribución da que remostrar en función de $\hat{d}$.

4. Simúlase a remostra $X^*_1, X^*_2, \ldots, X^*_n$ da distribución escollida, e a partir dela calcúlase a réplica do estatístico $\Delta^*_{n,2}$.

5. Repítese o paso anterior un número grande B de veces, obtendo así B réplicas do estatístico do exceso de masa: $\Delta^{*,1}_{n,2}, \ldots, \Delta^{*,B}_{n,2}$.

6. O test rexeita a unimodalidade se

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\Delta_{n,2} < \Delta^{*,b}_{n,2}) < \alpha. \tag{1.28}$$

1.2.3. Test de Ameijeiras et al.

Como xa comentamos, o test de Cheng e Hall (1998) está limitado a contrastar só unimodalidade, pois atopar distribucións que verifiquen a igualdade dos coeficientes definidos en (1.25) cos estimados a partir da mostra para así poder calibrar o test resulta inviable na práctica cando $k > 1$, a pesar das recomendacións dos autores. Ademais, o test ten outro inconveniente que paga a pena resaltar: non é directamente extensible a contextos nos que os datos non proveñan da recta real. Por un lado, pode ser que d definido en (1.22) non teña un análogo claro cando traballamos con outro tipo de datos, como datos multidimensionais. Por outro, para cada novo espazo no que queiramos traballar teremos que especificar unha nova familia de distribucións que abranga todos os valores posibles de d pola propia estrutura do test.

O último test que imos ver, que é o proposto por Ameijeiras et al. (2019), trata de atallar estes dous problemas xuntando o concepto de exceso de masa co de xanela crítica. En principio o test segue as

ideas de Cheng e Hall (1998): ambos empregan o exceso de masa $\Delta_{n,k+1}$ como estatístico de contraste, e, para tratar de aproximar mellor a súa distribución baixo a nula, os dous estiman os valores d_l, con $l = 1, \ldots, 2k-1$ a partir da mostra. A diferenza principal entre os dous contrastes está no método de calibrado. Mentres que Cheng e Hall (1998) lanzan as remostras dunhas distribucións fixas con k modas de xeito que os valores d_l coincidan cos da mostra, Ameijeiras et al. (2019) propón usar un bootstrap suavizado, remostrando dunha versión de $\hat{f}_{h_k}$ modificada lixeiramente para que os seus valores d_l definidos por (1.25) sexan iguais aos estimados a partir da mostra. Isto permítenos aplicar o test para contrastar a existencia de $k > 1$ modas, porque se está a evitar o problema de ter que buscar unha familia de distribucións que cubra todo o rango posible de valores d_l, pois a distribución de remostraxe vén dada directamente polos datos a través dunha versión modificada de $\hat{f}_h$. Ademais, os conceptos de exceso de masa e xanela crítica son xeralizables a datos circulares, e por tanto o test de Ameijeiras et al. (2019) é directamente extensible para contrastar multimodalidade no círculo. Pola contra, e igual que sucede co test de Cheng e Hall (1998), as ideas do test non se poden trasladar ao caso multidimensional de forma directa, pois os valores d_l non teñen un análogo claro nese contexto.

A estimación dos parámetros d_l ofrecida por Ameijeiras et al. (2019) é lixeiramente diferente a de Cheng e Hall (1998). Ameijeiras et al. (2019) emprega a función $\hat{f}_{h_k}$ na estimación dos d_l, pois é un bo candidato á hora de estimar a localización das modas e antimodas da poboación orixinial, así como a densidade neses puntos. Ademais, baixo condicións pouco restrictivas pódese probar que a orde da xanela crítica h_k é a óptima para estimar a densidade orixinal f (ver Mammen et al., 1992). Pola contra, h_k non é unha boa opción a hora de tratar de estimar f'', pois os parámetros de suavizado óptimos son máis grandes neste caso. Tendo isto en conta, a estimación que Ameijeiras et al. (2019) propón para os parámetros d_l é:

$$\hat{d}_l = \frac{\left|\hat{f}''_{h'}(\hat{x}_l)\right|}{\hat{f}^3_{h_k}(\hat{x}_l)}, \quad l = 1, \ldots, 2k-1; \tag{1.29}$$

onde $\hat{f}_{h_k}$ e $\hat{f}''_{h'}$ son os estimadores tipo núcleo de f e f'', h_k é a xanela crítica para k modas, h' é a xanela plug-in de Sheather e Jones para estimar a segunda derivada de f, e $\hat{x}_1, \ldots, \hat{x}_{2k-1}$ son as modas e antimodas de $\hat{f}_{h_k}$.

Unha vez calculados os $\hat{d}_l$, temos que modificar a densidade $\hat{f}_{h_k}$ para que os seus valores d_l sexan iguais aos estimados. Así, a nova densidade g_0 coincidirá con $\hat{f}_{h_k}$ salvo en pequenos intervalos arredor das $2k-1$ modas e antimodas. Esta modificación realízase de xeito que as modas e antimodas de $\hat{f}_{h_k}$ coincidan coas modas e antimodas de g_0, e ademais se verifique que

$$\hat{d}_l = \frac{|g_0''(\hat{x}_l)|}{g_0^3(\hat{x}_l)}, \quad \text{e} \quad \hat{f}_{h_k}(\hat{x}_l) = g_0(\hat{x}_l), \quad l = 1, \ldots, 2k-1 \tag{1.30}$$

para todas elas. Para máis detalles sobre estas modificacións, véxase Ameijeiras et al. (2019).

Unha vez obtida a densidade modificada g_0, o proceso de calibrado é similar aos demais tests. Simúlase un número grande de remostras a partir de g_0 e con elas calcúlanse as réplicas do estatístico de contraste $\Delta^*_{n,k+1}$. Rexeitaremos a hipótese nula de que f ten polo como máximo k modas con nivel de significación α se a proporción de réplicas maiores que o estatístico orixinal é menor que α, é dicir, se $\mathbb{P}(\Delta_{n,k+1} < \Delta^*_{n,k+1}) < \alpha$.

Para finalizar a sección, e a modo de resumo, presentamos unha implementación esquemática do test de Ameijeiras et al. (2019) con nivel de significación α:

1. Dada a mostra orixinal $X_1, X_2, \ldots, X_n$, calcúlase o estatístico de exceso de masa: $\Delta_{n,k+1}$.

2. Calcúlase os valores $\hat{d}_l$ da mostra orixinal mediante

$$\hat{d}_l = \frac{\left|\hat{f}''_{h'}(\hat{x}_l)\right|}{\hat{f}^3_{h_k}(\hat{x}_l)}, \quad l = 1, \ldots, 2k-1; \tag{1.31}$$

onde $\hat{f}''_{h'}$ é estimador tipo kernel de f'', h' é os parámetro de suavizado plug-in de Sheather e Jones, e $\hat{x}_1, \ldots, \hat{x}_{2k-1}$ son as modas e antimodas de $\hat{f}_{h_k}$.

3. A partir de $\hat{f}_{h_k}$ e dos valores $\hat{d}_l$, calcúlase a densidade modificada g_0.

4. Simúlase a remostra $X_1^*, X_2^*, \ldots, X_n^*$ da densidade g_0, e a partir dela calcúlase a réplica do estatístico $\Delta_{n,k+1}^*$.

5. Repítese o paso anterior un número grande B de veces, obtendo así B réplicas do estatístico do exceso de masa: $\Delta_{n,k+1}^{*,1}, \Delta_{n,k+1}^{*,2}, \ldots, \Delta_{n,k+1}^{*,B}$.

6. O test rexeita a unimodalidade se

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\Delta_{n,k+1} < \Delta_{n,k+1}^{*,b}) < \alpha. \tag{1.32}$$

1.3. Comparativa crítica dos tests

Para rematar este capítulo comentaremos o comportamento na práctica de cada un dos test presentados, tanto á hora de calibrar o test baixo a hipótese nula como de detectar a alternativa. Esta comparativa está baseada na tese de doutoramento de Ameijeiras (2017), onde realiza un amplo estudo de simulación para comparar o calibrado e potencia dos cinco test aquí expostos xunto co de Fisher e Marron (2001). A razón pola que nós non recollemos este último test neste traballo é precisamente o mal calibrado que presenta nese estudo.

No estudo de simulación proposto, Ameijeiras (2017) emprega unha ampla variedade de distribucións. En total, considera once distribucións unimodais, dez bimodais e cinco trimodais. De cada unha delas simula 500 mostras, e para cada mostra realiza 500 remostras do mesmo tamaño para poder calibrar os distintos tests. Despois recolle a proporción de rexeitamentos de cada test, xunto coa súa desviación típica estimada. Fai isto para tres tamaños de mostra distintos, $n = 50$, $n = 200$ e $n = 1000$ ($n = 100$ en vez de $n = 1000$ nas simulacións nas que se busca comparar a potencia), e para os tres niveis de significación máis usuais, $\alpha = 0.01$, $\alpha = 0.05$ e $\alpha = 0.1$. Nesta sección só recolleremos as conclusións principais do devandito estudo de simulación. Se se queren consultar as táboas con todas as proporcións de rexeitamentos calculadas, remítese á tese de Ameijeiras (2017).

Comezemos comentando o calibrado dos cinco test a hora de contrastar a hipótese nula de unimodalidade, é dicir:

$$H_0 : f \textit{ ten unha única moda.} \tag{1.33}$$

Neste caso os tests de Silverman (1989) e de Müller e Sawitzki (1998) son excesivamente conservadores, obténdose con eles proporicións de rexeitamento moi próximas a cero en todos os escenarios, incluso para nivel de significación $\alpha = 0.1$ e tamaño de mostra $n = 1000$. Os tres tests restantes amosan un mellor comportamento. O test de Hall e York (2001) presenta proporcións de rexeitamento próximas ao nivel de significación, aínda que tende a rexeitar lixeiramente por debaixo do debido para tamaños de mostra pequenos. O test de Cheng e Hall (1998) tamén ofrece proporcións de rexeitamento similares ao nivel de significación na gran maioría dos supostos. Pero, tal e como predeciron os propios autores, amosa certos problemas cando a distribución dos datos é asimétrica debido ao propio método de calibrado do test. Nestes casos, este test dá proporcións de rexeitamento por debaixo do nivel de significación α. Por último, o propio test de Ameijeiras et al. (2019) ten un comportamento aceptable en todos os casos, aínda que tende a rexeitar lixeiramente por debaixo do nivel nominal.

Pasamos a falar da potencia dos tres tests que ofreceron un calibrado aceptable: os de Hall e York (2001), Cheng e Hall (1998) e Ameijeiras et al. (2019). Aquí atopamos diferenzas claras entre o test baseado na xanela crítica e os baseados no exceso de masa. Os segundos (Cheng e Hall, 1998; e Ameijeiras et al. (2019)) son incapaces de detectar a alternativa de bimodalidade cando a segunda moda ten un exceso de masa asociado moi pequeno. Pola contra, o test de Hall e York (1998) non sufre

por este feito, acadando proporcións de rexeitamento bastante altas incluso para tamaños de mostra pequenos ($n = 50$). No resto dos casos, o test máis potente resulta ser o de Cheng e Hall (1998) para todos os tamaños e niveis de significación. Os outros dous non teñen unha potencia comparable, sendo máis potente o test de Ameijeiras et al. (2019) cando o tamaño de mostra é pequeno ($n = 50$), e o de Hall e York (2001) cando o tamaño é moderado ou grande ($n = 100$ e $n = 200$).

Finalmente, no seu estudo de simulación Ameijeiras (2017) tamén investiga o comportamento dos test á hora de contrastar a hipótese nula

$$H_0 : f \textit{ ten como máximo dúas modas.} \tag{1.34}$$

Aquí so se consideran os únicos dous tests aplicables neste contexto dos cinco considerados: o de Silverman (1989) e o de Ameijeiras et al. (2019). O test de Ameijeiras et al. (2019) consigue un calibrado aceptable neste caso, tanto se a distribución dos datos é unimodal ou bimodal, mentres que o de Silverman (1989) sigue ofrecendo proporcións de rexeitamento moi por debaixo do nivel de significación α para todas as distribucións consideradas. No que respecta á potencia, só falaremos do test de Ameijeiras et al. (2019) por ser o único cun calibrado aceptable. Do estudo de simulación dedúcese que este test detecta satisfactoriamente a alternativa cando a distribución dos datos é trimodal salvo para tamaños de mostra pequenos ($n = 50$), e a potencia do mesmo aumenta ao aumentar o tamaño da mostra.

1.4. Análise de datos reais

Na sección anterior comprobamos que, de todos os tests de multimodalidade expostos, os únicos que presentan un bo calibrado na práctica son o test de Hall e York (2001), o test de Cheng e Hall (1998) e o test de Ameijeiras et al. (2019). Polo tanto, xa temos tres métodos sistemáticos e obxectivos para contrastar o número de modas dunha distribución a partir dunha mostra aleatoria dada, tal e como buscabamos na introducción deste traballo. Entón, podemos empregar estes tres tests para estudar a existencia de máis dunha moda nos datos de Choi et al. (2020) que presentamos na Introdución. Se recordamos, estes datos estudaban os falecementos de animais voadores por impactos contra aeroxeradores, para así poder deseñar protocolos de busca máis efectivos e mellorar o noso comprendemento de como os parques eólicos afectan a fauna silveste. A base de datos recolle a distancia horizontal entre os cadáveres atopados e a base do aeroxerador máis próximo, que se separan en dúas mostras: unha de morcegos, con 1015 observacións, e outra de aves, con 975 observacións. Imos contrastar a presenza dunha única moda nas dúas mostras, fronte a alternativa de que hai máis dunha moda.

Aplicamos primeiro o test de Hall e York (2001) a estes datos. Na mostra de paxaros, o valor da xanela crítica calculado é $h_1 = 5.1046$, que ofrece un p-valor de 0.035. Por tanto, o test de Hall e York (2001) rexeita a unimodalidade dos datos das aves ao 5 % de significación. Para a mostra de morcegos, o valor da xanela crítica é $h_1 = 3.4315$, cun p-valor de 0.174. O p-valor é maior que os niveis de significación máis usuais, polo que neste caso o test non detecta probas estatísticas o suficientemente significativas para rexeitar a unimodalidade.

Pasamos a aplicarlle aos datos de Choi et al. (2020) os dous tests baseados no exceso de masa, o de Cheng e Hall (1998) e o de Ameijeiras et al. (2019). O valor do estatístico de exceso de masa para a mostra das aves é $\Delta_{975,2} = 0.062054$. Ambos os dous test dan p-valores moi pequenos a esta mostra, menores que 10^{-10}, polo que os dous contrastes rexeitan a unimodalidade dos datos das aves para calquera nivel de significación habitual. No caso dos datos dos morcegos, o exceso de masa toma o valor de $\Delta_{1015,2} = 0.053185$, e os dous test ofrecen un p-valor moi próximo a cero, menor que 10^{-10}. Daquela, e ao contrario que o test de Hall e York (2001), os tests de Cheng e Hall (1998) e Ameijeiras et al. (2019) rexeitan a unimodalidade da mostra dos morcegos para os niveis de significación habitual.

A disparidade dos resultados para a mostra de morcegos non debería de sorprendernos. Por un lado, cando comentamos os mapas SiZeR dos datos de Choi et al. (2020) xa vimos que o mapa SiZeR dos morcegos era máis complexo que o das aves, dificultando dar unha resposta clara sobre o número de modas da súa función de densidade. Por outro, na sección anterior vimos o test de Cheng e Hall

(1998) é máis potente que o de Hall e York (2001) na maior parte dos escenarios considerados, polo que non resulta extraño que os tests baseados no exceso de masa detecten certa estrutura multimodal nos datos que o test de Hall e York (2001) non é capaz de captar.

Os p-valores tan pequenos obtidos ao aplicar os tests de Cheng e Hall (1998) e Ameijeiras et al. (2019) fannos sospeitar de que sería posible a existencia de incluso máis de dúas modas, tanto para a mostra dos morcegos como a das aves. Daquela, resultaría interesante contrastar a hipótese nula de que existen como máximo dúas modas fronte a alternativa de que existen máis de dúas modas para estes datos. O único test dos presentados que nos permite contrastar estas hipóteses é o test de Ameijeiras et al. (2019). Se o aplicamos á mostra dos paxaros obtemos o valor do estatístico de contraste $\Delta_{975,3} = 0.038902$, cun p-valor asociado moi pequeno, menor que 10^{-10}. Os resultados para a mostra dos morcegos é similar, obtendo un estatístico de exceso de masa $\Delta_{1015,3} = 0.045922$, e un p-valor tamén menor que 10^{-10}. Por tanto, o test rexeita a hipótese de que o número de modas é menor ou igual que dous para as dúas mostras para calquera nivel de significación usual.

Capítulo 2

Unha nova proposta de test de multimodalidade

O novo test de multimodalidade vai estar baseado no test de razón de verosimilitudes da estatística paramétrica. O test de razón de verosimilitudes, introducido por primeira vez por Neyman e Pearson no ano 1936, é unha familia de contrastes estatísticos que se empregan en inferencia paramétrica para contrastar a situación do parámetro descoñecido dentro espazo de parámetros. Supoñamos que $X_1, X_2, \ldots, X_n$ é unha mostra aleatoria simple dunha variable aleatoria absolutamente continua con función de densidade f pertencente a familia paramétrica $\{f_\theta : \theta \in \Theta\}$, onde $\Theta \subset \mathbb{R}^m$. Dividamos o espazo de parámetros en dous subconxuntos disxuntos, escollendo Θ_0 e Θ_1 tales que $\Theta_0 \cap \Theta_1 = \emptyset$ e $\Theta_0 \cup \Theta_1 = \Theta$. O test de razón de verosimilitudes contrasta hipóteses da forma

$$H_0 : \theta \in \Theta_0 \text{ fronte a } H_1 : \theta \in \Theta_1. \tag{2.1}$$

Aínda que non o pareza a primeira vista, a maior parte das hipóteses a contrastar en estatística paramétrica son expresables da forma anterior. Por exemplo, supoñamos que a mostra $X_1, \ldots, X_n$ provén dunha densidade normal $N(\mu, \sigma^2)$ onde $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ son parámetros descoñecidos, e queremos contrastar se a súa media é igual a un valor prefixado μ_0. Pois este contraste é do tipo anterior, onde agora o espazo total de parámetros é $\Theta = \mathbb{R} \times \mathbb{R}^+$, mentres que o subespazo asociados ás hipóteses nula e alternativa son $\Theta_0 = \{(\mu, \sigma^2) \in \Theta : \mu = \mu_0\} = \{\mu_0\} \times \mathbb{R}^+$ e $\Theta_1 = \{(\mu, \sigma^2) \in \Theta : \mu \neq \mu_0\}$.

A construción do estatístico de contraste neste tipo de tests emprega a función de verosimilitude $\mathcal{L}$, que é a función real positiva definida como

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_\theta(X_i). \tag{2.2}$$

A idea detrás da verosimilitude é que $\mathcal{L}(\theta_0)$ representa a *factibilidade* de que o valor do parámetro descoñecido sexa θ_0 unha vez observada a mostra. Así, dada unha mostra, que $\mathcal{L}(\theta_0)$ sexa maior que $\mathcal{L}(\theta_1)$ significa que é máis *verosímil* que o parámetro descoñecido θ sexa igual a θ_0 que a θ_1 á vista dos datos observados. Entón, no caso de que a hipótese nula sexa certa, a función de verosimilitude debería tomar valores grandes dentro do conxunto Θ_0, que é o que se corresponde a H_0. Polo tanto, un cadidato a estatístico de contraste sería

$$\lambda = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \in [0, 1]; \tag{2.3}$$

que estará ben definido se a verosimilitude $\mathcal{L}$ está limitada en todo o espazo de parámetros Θ. Tendo en conta o razoamento anterior, se H_0 é certa, $\mathcal{L}$ acadará o seu máximo no conxunto Θ_0 ou preto del, e daquela λ tomará un valor próximo a 1. Se, pola contra, H_0 é falsa, $\mathcal{L}$ tomará valores pequenos

no conxunto Θ_0, e λ estará preto de 0. Tendo en conta o razoamento anterior, o test de razón de verosimilitudes rexeitará a hipótese nula para valores pequenos do estatístico λ.

En vez de empregar o estatístico λ, a maior parte das veces utilízase o estatístico equivalente

$$D = -2\log(\lambda) = 2\left[\sup_{\theta\in\Theta}\ell(\theta) - \sup_{\theta\in\Theta_0}\ell(\theta)\right] \geq 0, \tag{2.4}$$

onde $\ell(\theta) = \log\mathcal{L}(\theta)$, rexeitando agora a hipótese nula para valores grandes de D. Isto permítenos obter un test asintótico, pois o Teorema de Wilks (Wilks, 1938) asegura que, baixo a hipótese nula, D converxe en distribución a unha chi cadrado baixo condicións de regularidade bastante laxas. Ademais, os contrastes de razón de verosimilitudes resultan ser os tests uniformemente máis potentes en varios escenarios, tal e como garanten resultados como o Lema de Neyman-Person (Neyman e Pearson, 1933), o Teorema de Karlin-Rubin ou o Teorema de Lehmann (Karlin, 1957). Todo o anterior explica a gran popularidade do test de razón de verosimilitudes dentro da estatística paramétrica.

Neste capítulo imos ver como trasladar a idea do test de razón de verosimilitudes a un contexto non paramétrico de contraste do número de modas dunha poboación, tratando de conseguir así un test facilmente adaptable a situacións con datos non lineais, como poden ser datos circulares ou multidimensionais.

2.1. A pseudo-verosimilitude

Sexa $X_1, X_2, \ldots, X_n$ unha mostra aleatoria simple dunha variable aleatoria absolutamente continua X. Como xa vimos, un estimador non paramétrico da función de densidade de X é o estimador tipo núcleo $\hat{f}_h$ que definimos en (1). Unha vez obtida unha estimación da densidade de X podemos derivar desta un análogo da función de verosimilitude paramétrica. Así, definimos a *pseudo-verosimilitude* da mostra como a función real positiva

$$\mathcal{L}(h) = \prod_{i=1}^{n}\hat{f}_h(X_i), \tag{2.5}$$

onde $h > 0$. Poderíamos pensar na función $\mathcal{L}$ como nunha verosimilitude paramétrica onde a familia paramétrica de densidades que estamos a supoñer é $\{\hat{f}_h : h > 0\}$, aínda que esta familia non é independente dos datos observados, senón que vén determinada directemente por eles.

Unha forma de contrastar multimodalidade mediante a pseudo-verosimilitude definida en (2.5) consiste en combinar a función $\mathcal{L}$ coa idea de xanela crítica de Silverman (1989). Se supoñemos que a pseudo-verosimilitude se comporta de maneira similar á verosimilitude paramétrica, entón $\mathcal{L}(h_k)$ debería tomar valores grandes cando a densidade f ten k modas, e pequenos en caso contrario. Por tanto, poderíamos pensar en empregar como estatístico de contraste o valor

$$E_k = \ell(h_{k+1}) - \ell(h_k); \tag{2.6}$$

onde $\ell(h) = \log\mathcal{L}(h)$, o núcleo empregado para construír a función $\mathcal{L}$ é o gaussiano, e h_k e h_{k+1} son as xanelas críticas correspondentes a k e $k+1$ modas. Rexeitaremos a hipótese nula para valores grandes do estatístico de contraste E_k. O test calíbrase mediante bootstrap suavizado, remostrando da densidade $\hat{f}_{h_k}$.

Esta forma de proceder para contrastar multimodalidade ten dous problemas. O primeiro é que, se j denota o número de modas da densidade f, con este estatístico estamos contrastando as hipóteses

$$H_0 : j = k \text{ fronte a } H_1 : j = k+1; \tag{2.7}$$

polo que en principio sería un contraste máis restrictivo que o resto dos tests de multimodalidade vistos ata o de agora. Con todo, isto non é unha gran limitación a hora de aplicar o test na práctica.

Porén, o segundo problema impide o seu uso na práctica, pois este test está mal calibrado segundo os resultados do noso estudo de simulación. Na Táboa 2.1 recollénse as proporcións de rexeitamento

obtidas ao aplicar o devandito test á 1000 mostras de tamaño $n = 100$ de distintas distribucións unimodais. Como se pode ver, o test chega a ter unhas proporcións de rexeitamento moito máis altas que o nivel de significación para a distribución chi cadrado, chegando a rexeitar o 19 % das veces para o nivel do 5 % e o 30 % para o 10 %.

Táboa 2.1: Proporcións de rexeitamentos ao aplicarlle o test anterior a 1000 mostras procedentes de diversas distribucións unimodais. As distribucións escollidas son unha normal estándar, unha distribución chi cadrado con 5 graos de liberdade, unha Beta(3, 2) e unha Beta(6, 2). O tamaño das mostras é o mesmo en todos os casos: $n = 100$. Para calibrar o test empregáronse $B = 500$ remostras mediante bootstrap suavizado. As proporcións de rexeitamento do test están calculadas para os tres niveis de significación máis usuais: 1 %, 5 % e 10 %. Ao lado de cada proporción de rexeitemento aparece, entre parénteses, a súa desviación típica estimada multiplicada por 1.96. Todas os datos estan arredondados ao 3 decimal.

	1 %	5 %	10 %
Normal estándar	0.000(0.000)	0.015(0.008)	0.040(0.012)
Chi cadrado	0.066(0.015)	0.191(0.024)	0.309(0.029)
Beta(3,2)	0.009(0.006)	0.055(0.014)	0.121(0.020)
Beta(6,2)	0.007(0.005)	0.044(0.013)	0.088(0.018)

2.2. Proposta do novo test de multimodalidade

Para intentarmos resolver os problemas que presenta o estatístico E_k definido en (2.6) trataremos de trasladar de forma directa a estrutura do test de razón de verosimilitudes paramétrico ao noso contexto. É dicir, supoñendo que f e j-modal, queremos contrastar

$$H_0 : j \leq k \text{ fronte a } H_1 : j > k; \tag{2.8}$$

mediante un test de razón de verosimilitudes empregando a función de verosimilitude $\mathcal{L}$ definida en (2.5). Para logralo temos que traducir as hipóteses H_0 e H_1 nunha partición do espazo de parámetros $(0, +\infty)$, algo doado tendo en conta o concepto de xanela crítica. Se h_k é a xanela crítica para k modas, entón a estimación tipo núcleo da densidade $\hat{f}_h$ terá como máximo k modas se e só se $h \geq h_k$. Así, o conxunto asociado a hipótese nula H_0 é $[h_k, +\infty)$, e o estatístico de razón de verosimilitudes buscado vén dado por

$$D_k = 2\left[\max_{h>0} \ell(h) - \max_{h \geq h_k} \ell(h)\right], \tag{2.9}$$

onde $\ell(h) = \log \mathcal{L}(h)$. Igual que no test de razón de verosimilitudes usual, rexeitamos a hipótese nula para valores grandes de D_k. Pero, tal como está pensado, o estatístico de contraste D_k non está ben definido. Non resulta moi dificil demostrar que $\lim_{h\to 0} \mathcal{L}(h) = +\infty$, e por tanto $\max_{h>0} \ell(h)$ non é un número real.

Un xeito de solventar este problema é redefinir a función $\mathcal{L}$ dada por (2.5) mediante validación cruzada. Para iso, imos apoiarnos nas funcións:

$$\hat{f}_h^{-i}(x) = \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h}\right). \tag{2.10}$$

Daquela, definimos a *pseudo-verosimilitude por validación cruzada* da mostra $X_1, \ldots, X_n$ como

$$\mathcal{L}_{CV}(h) = \prod_{i=1}^{n} \hat{f}_h^{-i}(X_i), h > 0. \tag{2.11}$$

Con isto conseguimos que, ao contrario de $\mathcal{L}$, $\mathcal{L}_{CV}$ si teña un máximo global.

Proposición 1. *Sexa $X_1, \ldots, X_n$ unha mostra aleatoria simple dunha variable aleatoria absolutamente continua X. Sexa $\mathcal{L}_{CV}$ a función de pseudo-verosimilitude por validación cruzada definida en* (2.11)*, onde a función núcleo K verifica:*

a) K é unha función limitada.

b) $\lim_{h\to 0} \frac{1}{h} K\left(\frac{1}{h}\right) = \lim_{h\to 0} \frac{1}{h} K\left(\frac{-1}{h}\right) = 0$.

Entón a función $\mathcal{L}_{CV}(h)$ ten un máximo no intervalo $(0, +\infty)$ con probabilidade 1.

Demostración. Tendo en conta de que o núcleo K é unha función limitada, deducimos que

$$\lim_{h\to+\infty} \hat{f}_h^{-i}(x) = \lim_{h\to+\infty} \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} K\left(\frac{x - X_j}{h}\right) = 0 \tag{2.12}$$

para todo $x \in \mathbb{R}$ e $i = 1, \ldots, n$. Daquela tamén se verifica que

$$\lim_{h\to+\infty} \mathcal{L}_{CV}(h) = \lim_{h\to+\infty} \prod_{i=1}^{n} \hat{f}_h^{-i}(X_i) = 0. \tag{2.13}$$

Supoñamos agora que todos os valores $X_1, \ldots, X_n$ son todos distintos entre si. Daquela, pola Condición b), verifícase que

$$\lim_{h\to 0} \frac{1}{(n-1)h} K\left(\frac{X_i - X_j}{h}\right) = 0. \tag{2.14}$$

para todo $i \neq j$. Disto dedúcese que

$$\lim_{h\to 0} \hat{f}_h^{-i}(X_i) = \lim_{h\to 0} \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) = 0, \tag{2.15}$$

para todo $i \in \{1, \ldots, n\}$ e por tanto $\lim_{h\to 0} \mathcal{L}_{CV}(h) = 0$. Como $\mathcal{L}_{CV}(h)$ é continua en $(0, +\infty)$, entón temos que $\mathcal{L}_{CV}(h)$ ten un máximo no intervalo $(0, +\infty)$ sempre que os valores $X_1, \ldots, X_n$ sexan todos distintos entre si. Como X é unha variable aleatoria absolutamente continua, isto sucede con probabilidade 1. □

O núcleo gaussiano é limitado, e ademais pódese ver que tamén verifica a Condición b) empregando a regra de L'Hôpital. Polo tanto a Proposición 1 garántenos que o estatístico

$$D_k = 2\left[\max_{h>0} \ell_{CV}(h) - \max_{h\geq h_k} \ell_{CV}(h)\right], \tag{2.16}$$

onde $\ell_{CV}(h) = \log \mathcal{L}_{CV}(h)$, está ben definido. Este será o estatístico de contraste que empreguemos no noso test de multimodalidade. Igual que no test de razón de verosimilitudes usual, rexeitaremos a hipótese nula para valores grandes de D_k. O calibrado do test realizarase por bootstrap suavizado, lanzando remostras da función de densidade $\hat{f}_{h_k}$.

Para estudar o comportamento da función de pseudo-verosimilitude por validación cruzada na práctica representamos a función $\mathcal{L}_{CV}$ para varias mostras, e en todos os casos considerados resultaba ser decrecente no intervalo $[h_{max}, +\infty)$, onde h_{max} é tal que $\mathcal{L}_{CV}(h_{max}) = \max_{h>0} \mathcal{L}_{CV}(h)$. Así, as probas empíricas apuntan a que o estatístico

$$D'_k = \begin{cases} 2\left[\ell_{CV}(h_{max}) - \ell_{CV}(h_k)\right], & h_{max} < h_k; \\ 0, & h_{max} \geq h_k; \end{cases} \tag{2.17}$$

coincide con D_k. Isto permitiríanos reducir lixeiramente a complexidade computacional do test, pois non precisaríamos calcular o máximo da función ℓ_{CV} no intervalo $[h_k, +\infty)$. Nas simulacións realizadas o test baseado no estatístico D'_k ofrece resultados moi similares ao outro test, con pequenas diferenzas debidas seguramente a erros numéricos á hora de aproximar os máximos da función ℓ_{CV}. Con todo, como non probamos que os estatísticos D_k e D'_k coinciden, centrarémonos só no primeiro deles.

Por último, debemos comentar que, ao empregar a xanela crítica de Silverman (1981) no test, herdamos con ela todos os problemas que leva asociados e que xa comentamos no capítulo anterior. Isto pódenos levar a un mal calibrado, especialmente para densidades con colas pesadas. Por exemplo na Táboa 2.2 están recollidas as proporcións de rexeitamento obtidas ao aplicar o test baseado no estatístico D'_k a mostras procedentes do Modelo 5 definido na Sección 2.5.1. Como se pode ver, a formulación actual do test é excesivamente liberal, con proporcións de rexeitamento preto do 40 % para $\alpha = 0,1$ e tamaños de mostra $n = 500$.

Táboa 2.2: Proporicións de rexeitamento do test baseado no estatístico D_k con xanela crítica de Silverman (1981) para mostras de varios tamaños do Modelo 5. Aplicouse o test a 1000 mostras do Modelo 5 de diversos tamaños. Para cada mostra realizáronse 500 remostras para aproximar o p-valor do test. Os tamaños de mostra escollidos foron $n = 100$, $n = 250$ e $n = 500$. As proporcións de rexeitamento do test están calculadas para os tres niveis de significación máis usuais: 1 %, 5 % e 10 %. Ao lado de cada proporción de rexeitamento, e entre parénteses, está a súa desviación típica estimada multiplicada por 1.96. Todos os datos están arredondados ao 3 decimal.

	1 %	5 %	10 %
$n = 100$	0.034(0.011)	0.096(0.018)	0.157(0.023)
$n = 250$	0.066(0.015)	0.178(0.024)	0.289(0.028)
$n = 500$	0.080(0.017)	0.245(0.027)	0.373(0.030)

Para tratarmos de remediar estes problemas de sobresuavizado que presenta a xanela crítitica de Silverman (1981), substituirémola pola de Hall e York (2001) tanto na definición do estatístico D_k como na densidade de remostraxe. É dicir, o estatístico de contraste será agora da forma

$$D_k = 2\left[\max_{h>0} \ell_{CV}(h) - \max_{h \geq h_{HY,k}} \ell_{CV}(h)\right], \tag{2.18}$$

e remostraremos da densidade $\hat{f}_{h_{HY,k}}$ para aproximar o p-valor asociado á mostra. Con estes cambios, a hipótese nula que estaremos contrastando con este test será

$$H_0 : f \text{ } \textit{ten } j \textit{ modas no intervalo } I; \tag{2.19}$$

onde I é un intervalo compacto escollido a priori. Así a todo, isto non supón unha gran diferencia a hora de aplicar o test na práctica.

Así, de forma resumida, o noso test de multimodalidade con nivel de significación α é:

1. A partir da mostra $X_1, \ldots, X_n$, obtemos a xanela crítica

   $$h_{HY,k} = \text{mín}\{h > 0 : \hat{f}_h \text{ ten } k \text{ modas en } I\}$$

   e as xanelas que maximizan a verosimilitude por validación cruzada baixo a hipótese nula e a alternativa:

   $$\mathcal{L}_{CV}(h_{max}) = \max_{h>0}\{\mathcal{L}_{CV}(h)\}; \quad \mathcal{L}_{CV}(h_{H_0}) = \max_{h \geq h_{HY,k}} \{\mathcal{L}_{CV}(h)\}.$$

   E con elas calculamos o estatístico de contraste:

   $$D_k = 2\left[\ell_{CV}(h_{max}) - \ell_{CV}(h_{H_0})\right].$$

2. Obtemos a remostra $X_1^*, \ldots, X_n^*$ da densidade suavizada $\hat{f}_{h_{HY,k}}$ e calculamos o valor do estatístico para esa mostra: D_k^*.

3. Repetimos B veces o paso 2, conseguindo así B réplicas do estatístico: $D_k^{*,1}, D_k^{*,2}, \ldots, D_k^{*,B}$.

4. O test rexeitará a hipótese nula de que f ten como máximo k modas se

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(D_k^{*,b} > D_k) < \alpha.$$

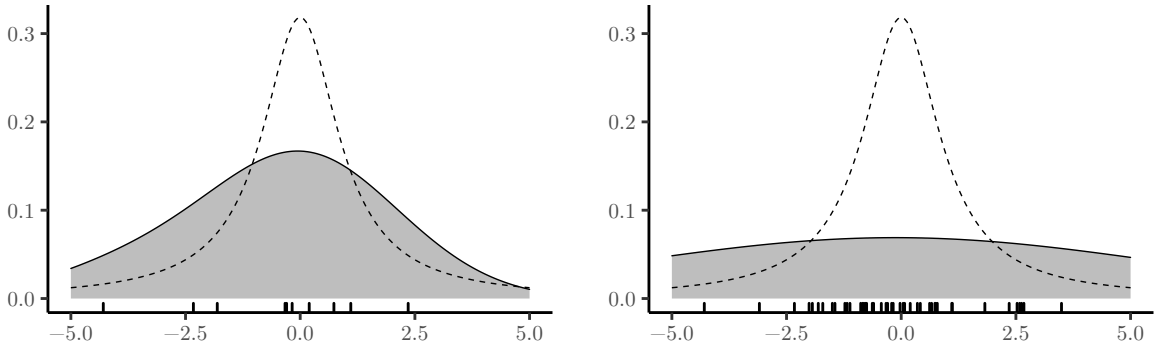
2.3. O parámetro de suavizado h_{max}

A idea de pseudo-verosimilitude tamén nos permite construír un selector do parámetro de suavizado h. Así, dada a mostra $X_1, \ldots, X_n$, poderíamos escoller como parámetro de suavización o valor $h_{max} > 0$ que cumpra que

$$\mathcal{L}_{CV}(h_{max}) = \max_{h>0} \mathcal{L}_{CV}(h); \tag{2.20}$$

que existirá sempre que o núcleo K que usemos verifique as condicións da Proposición 1, o que inclúe a maioría dos núcleos empregados na estimación tipo núcleo da densidade. Esta forma de escoller o parámetro de suavizado foi proposta por Duin (1976).

Hall (1987) estudou o comportamento deste selector do parámetro de suavizado, chegando a conclusión de que as propiedades de consistencia e converxencia desta xanela dependen das colas da densidade orixinal f e do núcleo K empregado na estimación. Se as colas da densidade orixinal dos datos f son máis pesadas que as do núcleo K, entón h_{max} diverxe a $+\infty$ ao aumentar n. Pódese dar unha interpretación sinxela deste feito a partir da demostración da Proposición 1. Para que h_{max} estea ben definido, necesitamos que as colas do núcleo K decrezan o suficientemente rápido para que a Condición b) da Proposción 1 se verifique. O problema está en que se o fan *demasiado rápido*, a función $\mathcal{L}_{CV}$ converxerá a 0 cando $h \to 0$ tan rápido que non recollerá fielmente as propiedades da mostra. A Figura 2.1 ilustra este comportamento. Nela temos representados os estimadores $\hat{f}_{h_{max}}$ para dúas mostras dunha distribución de Cauchy estándar de tamaños distintos. Pódese ver como o aumento do tamaño da mostra leva asociado un aumento da xanela h_{max}, dando lugar a estimadores da densidade moi sobresuavizados.



(a) $n = 10, h_{max} = 1.679$; (b) $n = 50, h_{max} = 5.562$;

Figura 2.1: Representación do estimador da densidade $\hat{f}_{h_{max}}$ para mostras dunha distribución de Cauchy estándar para catro tamaños distintos. En trazo continuo está representada $\hat{f}_{h_{max}}$, e en trazo discontinuo a densidade orixinal da mostra. Debaixo de cada gráfica está representado o tamaño n da mostra e o valor de h_{max} corespondente.

Así que, para garantir que a xanela h_{max} sexa consistente, teremos que seleccionar un núcleo con colas pesadas na estimación tipo núcleo da densidade, como pode ser unha dobre exponencial ou unha

t de Student. Pero a escolla de algúns destes núcleos para a construcción do test leva asociada a perda de monotonía do número de modas do estimador tipo núcleo da densidade $\hat{f}_h$. Isto complicaría a partición do espazo de parámetros en función das hipóteses nula e alternativa, dificultando o cálculo do estatístico de contraste e a aplicación do test na práctica. Nós empregamos sempre o núcleo gausiano na práctica, priorizando así a facilidade de computación do test por enriba do seu comportamento con distribucións con colas pesadas.

2.4. Adaptabilidade do test a outros contextos

Como xa comentamos na introdución deste capítulo, o noso obxectivo é a construción dun test de multimodalidade facilmente adaptable a outros contextos onde os datos non fosen lineais, como poden ser datos na circunferencia, na esfera, no toro... Pero, o test presentado na Sección 2.2 anterior cumpre este obxectivo?

A ferramenta básica que empregamos para definir o estatístico de contraste do test é a función de pseudo-verosimilitude $\mathcal{L}_{CV}$, construída a partir do estimador da densidade $\hat{f}_h$. O estimador tipo núcleo da densidade é facilmente adaptable a situacións onde os datos pertenzan a espazos distintos da recta real, como poden ser datos na esfera (Mardia e Jupp, 2000, Cap. 12). Así que tamén poderemos extender a idea pseudo-verosimilitude a estes espazos, definíndoa mediante unha expresión análoga a (2.11). Unha vez definida a función $\mathcal{L}_{CV}$ teríamos demostrar que está limitada superiormente, algo que parece bastante plausible. Se ese é o caso, poderemos construír o estatístico de contraste D_k seguindo o mesmo razoamento que no caso linear: calculamos o subconxunto do espazo de parámetros no que $\hat{f}_h$ verifica a hipótese nula, computamos o máximo da función $\ell_{CV}(h) = \log(\mathcal{L}_{CV}(h))$ nese conxunto e o seu máximo global, e restamos os dous valores.

Por todo o anterior o test de multimodalidade basado na pseudo-verosimilitude cumpre o obxectivo de ser directamente extensible a outros contextos fóra da recta real. De feito, no Capítulo 3 deste traballo centrarémonos en adaptar o test a datos circulares, seguindo o camiño marcado no parágrafo anterior. Ademais, como comentaremos nese mesmo capítulo, a pseudo-verosimilitude e a xanela h_{max} presentan un mellor comportamento no caso circular que no caso linear, o que apunta a que o test ofrecerá mellores resultados traballando con datos na circunferencia que con datos lineais.

2.5. Estudo de simulación

Unha vez introducido o novo test, a seguintes preguntas a responder son obvias. Como funciona na práctica? Está ben calibrado? É competitivo frente ao resto de test de multimodalidade presentes na literatura? Cómpre dar unha resposta a estas preguntas pois non demos en ningún momento resultados teóricos que fundamenten o novo test, senón que este foi construído intuitivamente mediante unha xeralización do test de razón de verosimilitudes paramétrico. É máis, os resultados de Hall (1987) sobre o parámetro de suavizado h_{max} que comentamos na sección anterior apuntan a que o novo test seguramente teña problemas ante distribucións con colas pesadas.

Nesta sección trataremos de responder as preguntas anteriores mediante un estudo de simulación. Con el buscamos comprobar o calibrado do novo test á hora de contrastar a unimodalidade dos datos. Ademais, compararemos os seus resultados cos doutros tres tests de multimodalidade cun calibrado aceptable según a tese de Ameijeiras (2017): o test de Hall e York (2001), o test de Cheng e Hall (1998) e o test de Ameijeiras et al. (2019). Simulamos $M = 1000$ mostras de tamaño n de diversas distribucións unimodais. A cada unha destas mostras aplicámoslle os catro contrastes, vendo se rexeitan ou non a unimodalidade dos datos para varios niveis de significación α. Finalmente, calulamos a proporción de mostras rexeitadas do total para cada test e nivel de significación. Como só consideramos distribucións unimodais (é dicir, baixo a hipótese nula) se os tests están ben calibrados as proporcións de rexeitamentos non deberían de exceder o nivel de significación α.

Imos realizar o anterior para dous tamaños de mostra distintos, $n = 100$ e $n = 500$, e seis distribucións unimodais distintas. Os niveis de significación considerados son os tres máis usuais: $\alpha = 0.01$,

$\alpha = 0.05$ e $\alpha = 0.1$.

Para a realización dos tests de Hall e York (2001), de Cheng e Hall (1998) e de Ameijeiras et al. (2019) empregouse o código da libraría de R `multimode` (Ameijeiras, 2021). O test baseado en pseudo-verosimilitude programouse empregando código propio deseñado ex professo para este estudo de simulación. Debido ao seu alto custo computacional, todas as simulacións foron realizadas mediante os recursos computacionais do Centro de Supercomputación de Galicia (CESGA).

2.5.1. Distribucións consideradas

As seis distribucións unimodais consideradas no estudo de simulación son as seguintes:

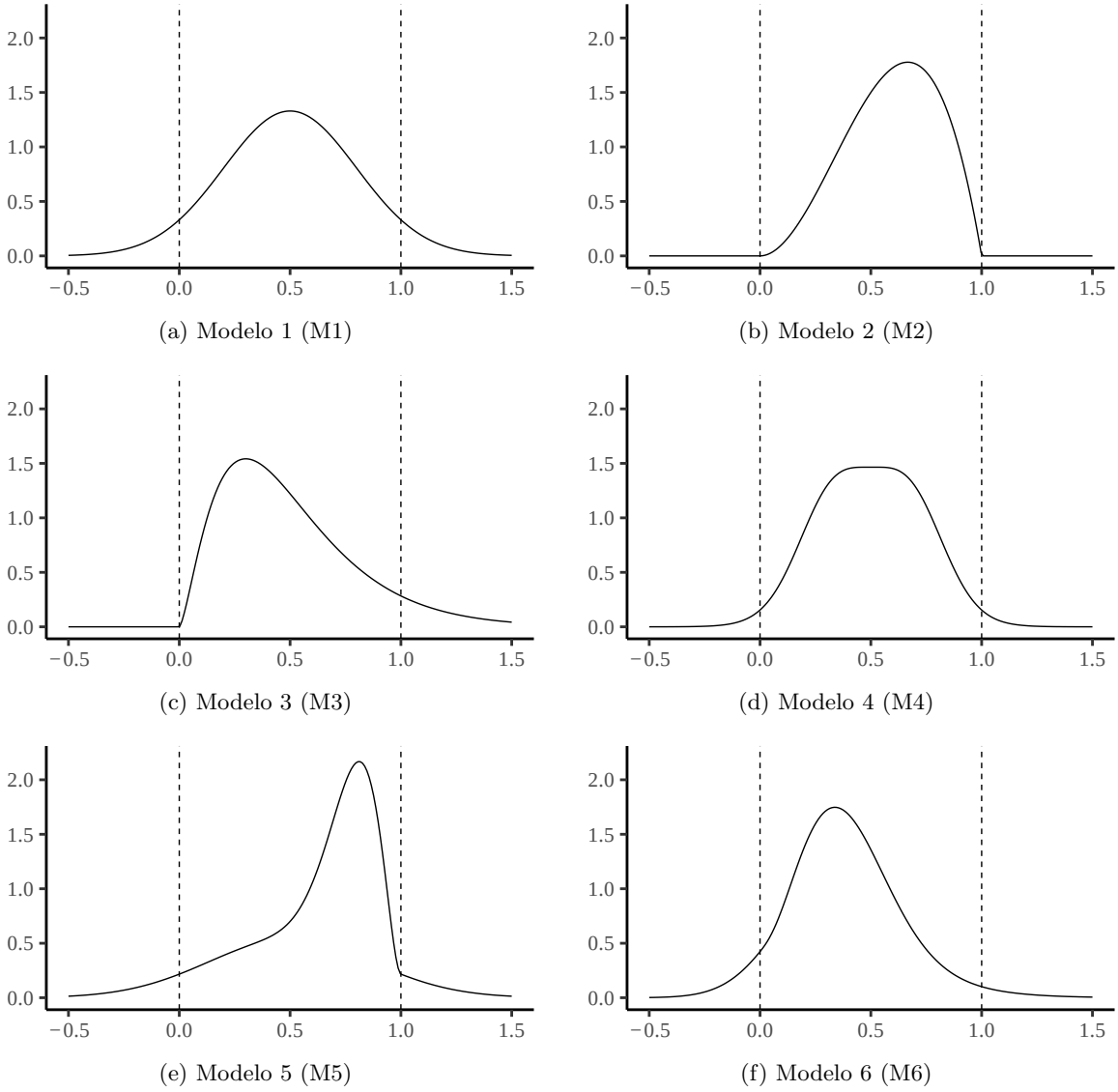
- **Modelo 1 (M1)**: unha normal de media 0.5 e varianza 0.09.
- **Modelo 2 (M2)**: unha Beta(3, 2).
- **Modelo 3 (M3)**: unha chi cadrado de 5 graos de liberdade reescalada: $0.1 \cdot \chi_5$.
- **Modelo 4 (M4)**: unha mixtura de normales: $0.6 \cdot N(0.5, 0.0502) + 0.2 \cdot N(0.3, 0.02) + 0.2 \cdot N(0.7, 0.02)$.
- **Modelo 5 (M5)**: unha mixtura dunha normal cunha beta: $0.5 \cdot \text{Beta}(10, 3) + 0.5 \cdot \text{N}(0.5, 0.137)$.
- **Modelo 6 (M6)**: unha mixtura dunha normal cunha gamma: $0.6 \cdot \text{N}(0.307, 0.0518) + 0.4 \cdot \text{Gamma}(4, 8)$.

As densidades destas distribucións están representadas na Figura 2.2. Están escollidas de xeito que acumulen aproximadamente o 90 % da probabilidade no intervalo $I = [0, 1]$ e teñan a súa moda nese mesmo intervalo. Isto facilita a aplicación tanto do test de Hall e York como do test baseado na verosimilitude empírica. Ademais, estás distribucións buscan representar a unha gran variedade de situacións, empezando pola simple normalidade (Modelo 1), pasando por distribucións con soporte compacto (Modelo 2), distintos graos de asimetría (Modelos 3, 5 e 6) e modas planas (Modelo 4).

2.5.2. Resultados

Na Táboas 2.3 e 2.4 están recollidos os resultados do estudo de simulación realizado. A principal conclusión é clara: o novo test de multimodalidade non está ben calibrado para todos os escenarios considerados. As proporcións de rexeitamento obtidas por este test varían moito entre os distintos modelos. Só logramos proporcións de rexeitamento próximas ao nivel de significación no Modelo 3 para mostras grandes ($n = 500$) e no Modelo 4 en todos os tamaños. Para os Modelos 1 e 6 o test é conservador, con proporcións de rexeitamento moito menores que o nivel de significación para todos os tamaños. O mesmo sucede coas mostras de tamaño $n = 100$ do Modelo 3. Porén, nos Modelos 2 e 5 os resultados son os contrarios: as proporcións de rexeitamento son moito maiores que o nivel en todos os casos. Ademais, nestes dous casos aumentar do tamaño de mostra de $n = 100$ a $n = 500$ provoca que as proporcións de rexeitamento aumenten, afastándose aínda máis do nivel de singificación α.

No resto dos tests os resultados son similares aos do estudo de simulación da tese de doutoramento de Ameijeiras (2017). O test de Ameijeiras et al. (2019) é o que mellor comportamento presenta. Só presenta proporcións de rexeitamento maiores que o nivel de significación para o Modelo 4. No Modelo 5 as proporcións de rexeitamento son menores que o nivel de significación, pero ao aumentar o tamaño da mostra a $n = 500$ as proporcións vanse achegando ao nivel nos casos nos que $\alpha = 0.05$ e $\alpha = 0.1$. Co Modelo 1 sucede algo similar. Neste caso con mostras con $n = 100$ observacións o test é conservador, con proporcións de rexeitamento menores que o nivel. Porén, as proporcións de rexeitamento obtidas para mostras con tamaño $n = 500$ son aceptables neste modelo. Nos Modelos 2, 3 e 6 as proporcións de rexeitamento están moi próximas ao nivel de significación, salvo en casos puntuais onde o test rexeita por baixo do nivel (Modelo 3: $n = 100$ e $\alpha = 0.5$; Modelo 6: $n = 100$ e $\alpha = 0.1$, $n = 500$ e $\alpha = 0.01$).



(a) Modelo 1 (M1)

(b) Modelo 2 (M2)

(c) Modelo 3 (M3)

(d) Modelo 4 (M4)

(e) Modelo 5 (M5)

(f) Modelo 6 (M6)

Figura 2.2: Representación das funcións de densidade das seis distribucións unimodais consideradas no estudo de simulación. As liñas verticais discontinuas marcan o intervalo $I = [0, 1]$, que é o que imos empregar para calcular as xanelas críticas de Hall e York tanto para o test de Hall e York como no novo test baseado en verosimilitude empírica.

O test de Cheng e Hall (1998) obtén proporcións de rexeitamento próximas ao nivel de significación para os Modelos 1 e 3. Nos Modelos 2 e 4 o comportamento do test é liberal, con proporcións de rexeitamento maiores que o nivel en todos os casos. Nos Modelos 5 e 6 o comportamento é o contrario, rexeitando por debaixo do nivel de significación para os dous tamaños de mostra considerados. Ademais, nos Modelos 4 e 5, pasar de tamaños de mostra $n = 100$ a $n = 500$ fai que as proporcións de rexeitamento se distancien aínda máis do nivel de significación.

Por último, o test de Hall e York (2001) ofrece proporcións de rexeitamento maiores que o nivel de significación para os Modelos 4 e 5 para os dous tamaños de mostra e todos os valores de α considerados, exceptuando os casos con $n = 100$ e $\alpha = 0.01$, que parecen respetar o nivel. Ademais, aumentar o

tamaño de mostra a $n = 500$ aumenta aínda máis as proporcións de rexeitamento, afastándoas do nivel. Cos Modelos 1 e 6 o test ofrece proporcións de rexeitamento por debaixo do nivel con mostras de tamaño $n = 100$, mais ao aumentar o tamaño a $n = 500$ fai que as proporcións de rexeitamento sexan similares ao nivel de significación nos dous modelos (quitando o caso do Modelo 1 con $\alpha = 0.01$, que ofrece unha proporción de rexeitamento por debaixo do nivel). No Modelo 2 o comportamento do test de Hall e York (2001), é mais complexo. O test parece respetar o nivel nos casos con $n = 100$ e $\alpha = 0.1$, $n = 500$ e $\alpha = 0.01$, e $n = 500$ e $\alpha = 0.05$. Porén, o resto das proporcións de rexeitamento están afastadas do nivel de significación, ben porque son menores ca el ($n = 100$ e $\alpha = 0.01$, $n = 100$ e $\alpha = 0.05$) ou ben porque son maiores ($n = 500$). Finalmente, co Modelo 3 obtivéronse proporcións de rexeitamento próximas ao nivel, salvo para tamaños de mostra e niveis de significación pequenos ($n = 100$ e $\alpha = 0.01$, $n = 100$ e $\alpha = 0.05$), onde as proporcións están por debaixo do nivel nominal do test.

Debido ao mal comportamento observado no estudo de simulación, non hai lugar para a análise da potencia do test de multimodalidade baseado en pseudo-verosimilitude. Tampouco cabe, por tanto, a aplicación do test nos datos de Choi et al. (2020), como si que fixemos cos demais test de multimodalidade.

Táboa 2.3: Proporcións de rexeitamentos para os test de Ameijeiras et al. (Test ACR) e de Cheng e Hall (Test CH) con nivel de significación ao 1 %, 5 % e 10 % calculadas a partir de $M = 1000$ mostras. Ao lado de cada proporción de rexeitamento, entre parénteses, aparece a seu erro estándar estimado multiplicado por 1.96. Todos os test calibráronse empregando $B = 500$ remostras. Cada fila correspóndese cunha distribución e tamaño de mostra distinto.

		Test ACR			Test CH		
		1 %	5 %	10 %	1 %	5 %	10 %
M1	$n = 100$	0.006(0.005)	0.037(0.012)	0.073(0.016)	0.008(0.006)	0.040(0.012)	0.087(0.017)
	$n = 500$	0.007(0.005)	0.044(0.013)	0.100(0.019)	0.010(0.006)	0.049(0.013)	0.098(0.018)
M2	$n = 100$	0.011(0.006)	0.046(0.013)	0.103(0.019)	0.023(0.009)	0.085(0.017)	0.154(0.022)
	$n = 500$	0.010(0.006)	0.053(0.014)	0.095(0.018)	0.029(0.010)	0.085(0.017)	0.145(0.022)
M3	$n = 100$	0.010(0.006)	0.037(0.012)	0.094(0.018)	0.007(0.005)	0.045(0.013)	0.086(0.017)
	$n = 500$	0.009(0.006)	0.062(0.015)	0.114(0.020)	0.011(0.006)	0.053(0.014)	0.099(0.019)
M4	$n = 100$	0.014(0.007)	0.071(0.016)	0.129(0.021)	0.034(0.011)	0.120(0.020)	0.185(0.024)
	$n = 500$	0.017(0.008)	0.060(0.015)	0.122(0.020)	0.044(0.013)	0.106(0.019)	0.199(0.025)
M5	$n = 100$	0.008(0.006)	0.031(0.011)	0.068(0.016)	0.004(0.004)	0.018(0.008)	0.044(0.013)
	$n = 500$	0.005(0.004)	0.039(0.012)	0.083(0.017)	0.003(0.003)	0.014(0.007)	0.031(0.011)
M6	$n = 100$	0.009(0.006)	0.045(0.013)	0.076(0.016)	0.005(0.004)	0.041(0.012)	0.071(0.016)
	$n = 500$	0.004(0.004)	0.044(0.013)	0.098(0.018)	0.004(0.004)	0.027(0.010)	0.077(0.017)

Táboa 2.4: Proporcións de rexeitamentos para o test de Hall e York (Test HY) e o novo test (Test D) con nivel de significación ao 1 %, 5 % e 10 % calculadas a partir de $M = 1000$ mostras. Ao lado de cada proporción de rexeitamento, entre parénteses, aparece a seu erro estándar estimado multiplicado por 1.96. Os dous tests calibráronse empregando $B = 500$ remostras. Cada fila correspóndese cunha distribución e tamaño de mostra distinto. En todos os casos o intervalo para calcular as xanelas críticas de Hall e York é o mesmo: $I = [0, 1]$.

		Test HY			Test D		
		1 %	5 %	10 %	1 %	5 %	10 %
M1	$n = 100$	0.001(0.002)	0.017(0.008)	0.064(0.015)	0.001(0.002)	0.011(0.006)	0.039(0.012)
	$n = 500$	0.005(0.004)	0.047(0.013)	0.107(0.019)	0.002(0.003)	0.019(0.008)	0.034(0.011)
M2	$n = 100$	0.003(0.003)	0.036(0.012)	0.100(0.019)	0.018(0.008)	0.080(0.017)	0.141(0.022)
	$n = 500$	0.012(0.007)	0.064(0.015)	0.143(0.022)	0.050(0.014)	0.147(0.022)	0.238(0.026)
M3	$n = 100$	0.001(0.002)	0.029(0.010)	0.091(0.018)	0.008(0.006)	0.035(0.011)	0.079(0.017)
	$n = 500$	0.009(0.006)	0.044(0.013)	0.119(0.020)	0.011(0.006)	0.052(0.014)	0.100(0.019)
M4	$n = 100$	0.006(0.005)	0.071(0.016)	0.168(0.023)	0.006(0.005)	0.045(0.013)	0.094(0.018)
	$n = 500$	0.019(0.008)	0.106(0.019)	0.202(0.025)	0.008(0.006)	0.050(0.014)	0.092(0.018)
M5	$n = 100$	0.006(0.005)	0.063(0.015)	0.154(0.022)	0.018(0.008)	0.075(0.016)	0.139(0.021)
	$n = 500$	0.018(0.008)	0.085(0.017)	0.171(0.023)	0.040(0.012)	0.106(0.019)	0.152(0.022)
M6	$n = 100$	0.001(0.002)	0.024(0.009)	0.074(0.016)	0.002(0.003)	0.007(0.005)	0.018(0.008)
	$n = 500$	0.007(0.005)	0.042(0.012)	0.107(0.019)	0.003(0.003)	0.011(0.006)	0.023(0.009)

Capítulo 3

Adaptación do test ao caso circular

Na Introdución comentamos que, igual que acontece con datos lineais, era importante construír un test de multimodalidade para datos circulares, pero non motivamos de xeito específico a súa necesidade. Para facelo contaremos cos datos analizados por Ameijeiras et al. (2018). Esta base de datos contén todos os incendios detectados polo satélite MODIS (acrónimo de *MODerate resolution Imaging Spectroradiometer*) da NASA (*National Aeronautics and Space Administration*) en Galicia, dende o 10 de Xullo de 2002 ata o 9 de Xullo de 2012. En total rexistráronse 6804 incendios durante ese período, e a cada un deles asignóuselle un número do 1 ao 366 en función do día do ano no que comezou. Aquí, a detetección de modas, que neste contexto se denominan *tempadas de incencios*, cobra especial relevancia á hora de entender os seus patróns estacionais e así loitarmos mellor contra a problemática dos incendios forestais.

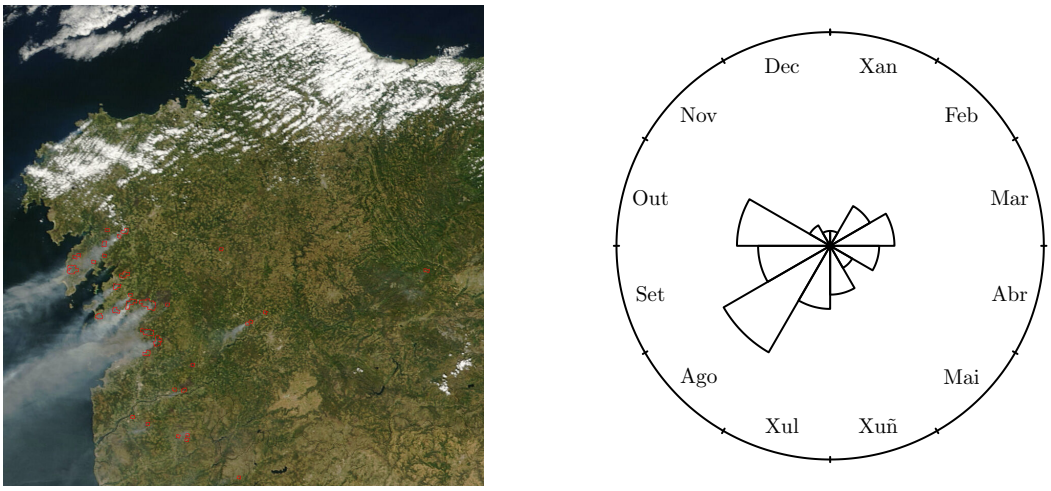


Figura 3.1: Á esquerda, os incendios (marcados en vermello) detectados polo satélite MODIS en Galicia o día 7 de agosto de 2006. Á dereita, o histograma circular (*rose diagram*) cos incendios detectados en Galicia dende o 10 de Xullo de 2002 ata o 9 de Xullo de 2012, detectados polo satélite MODIS da NASA.

Na Figura 3.1 pódense ver os datos anteriores representados nun histograma circular. Neste gráfico, cada sector circular se corresponde cun mes do ano, e a área de cada sector determina a cantidade de incendios detectados nese mes: a maior área, maior é o número de incendios detectados. No histograma vemos que, como era de esperar, a maior parte dos incendios detectados sucederon ao longo do verán, entre os meses de xullo e setembro. Pero tamén parece haber outros dous períodos con alta frecuencia

de incendios; un a principios do outono, principalmente en outubro, e outro a finais do inverno, ao redor de marzo. Estes dous picos de incendios están situados fóra da parte do ano onde as condicións meteorolóxicas son máis propicias para a aparicion de incendios de maneira natural, polo que poden estar asociados a certos comportamentos humanos, como a queima preventiva ou a queima de restrollos. Así, neste contexto, detectar a existencia de máis dunha moda proporcionaría evidencia de que a actividade humana condiciona a estrutura das vagas de incendios ao longo do ano.

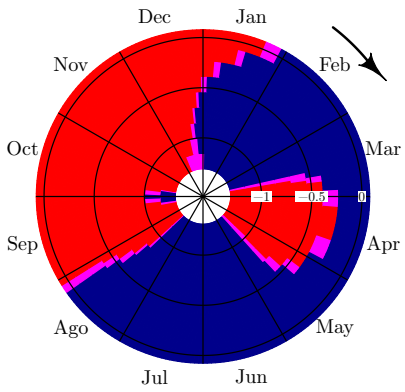


Figura 3.2: Mapa SiZeR circular dos incendios detectados en Galicia dende o 10 de Xullo de 2002 ata o 9 de Xullo de 2012, detectados polo satélite MODIS da NASA. Os nomes dos meses do ano están en inglés. O mapa SiZeR circular foi calculado empregando a libraría de R `NPCirc` (Oliveira et al., 2014).

Igual que no caso linear, existen diversas técnicas exploratorias co obxectivo de identificar o número de modas dunha mostra circular e a súa localización, como pode ser o SiZeR circular. O funcionamento e a interpretación do SiZeR circular son totalmente análogas ás do SiZeR para datos lineais: en cada punto realízase inferencia sobre o signo da derivada dun estimador tipo núcleo da densidade, e a cada rexión asóciaselle un valor dependendo do signo detectado. A cor vermella indica que hai probas estatisticamente significativas para afirmar que o signo da derivada da densidade suavizada é negativo, a cor azul apunta que a derivada é significativamente maior que cero, e a cor morada indica que non temos probas significativas para afirmar que a derivada é distinta de 0.

Se observamos agora o mapa SiZeR circular dos datos dos incendios de Ameijeiras et al. (2018) representado na Figura 3.2, vemos que os datos teñen unha moda clara a finais do mes de agosto, marcada por un cambio de cor azul a vermella para todos os valores do parámetro de suavizado considerados. Ao ir reducindo o parámetro de suavizado (é dicir, ao ir avanzando cara o interior do círculo) aparece unha nova moda no mes de marzo. Se reducimos máis aínda a xanela, aparece unha pequena rexión azul no mes de outubro, indicando a presenza dunha terceira moda nese mes.

De forma totalmente paralela ao que acontecía co SiZeR para datos lineais, afirmar a existencia dun número determinado modas só coa axuda dun mapa SiZeR circular é algo subxectivo e debatible. Así pois, é posible que dúas persoas distintas, vendo o mesmo mapa SiZeR da mesma mostra circular, dean aproximacións distintas sobre o número de modas dos datos. Por exemplo, á vista do SiZeR dos datos dos incendios de Ameijeiras et al. (2018) da Figura 3.2, cantas modas hai nos datos? Poderíase dicir que tres, pero tamén sería posible argumentar que a moda detectada no mes de outubro é espuria, un artefacto creado pola mostra, e que por tanto a densidade orixinal dos datos só ten dúas modas. O mesmo argumento se podería extender para defender a existencia de unha única moda nos datos. Para solucionar este problema, aparece a necesidade de construír un contraste estatístico para detectar o número de modas dunha densidade circular a partir dunha mostra, igual que no caso linear.

Neste capítulo intentaremos adaptar o test de multimodalidade definido no capítulo anterior a

datos circulares. Primeiro introduciremos os conceptos básicos necesarios para traballarmos con datos circulares. Despois pasaremos a definir os conceptos de estimador tipo núcleo da densidade e pseudo-verosimilitude para datos circulares, que nos permitirán extender o test do capítulo anterior a este contexto. Para acabar, realizaremos un estudo de simulación para tratar de comprobar se o novo test está ben calibrado.

3.1. Conceptos básicos

Antes de comezar, debemos comentar que na literatura estatística existen dous enfoques distintos a hora de traballar con datos circulares. O primeiro deles consiste traballar directamente coas cordenadas cartesianas da circunferencia unidade $S^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$. Con esta visión, unha distribución direccional non é máis que unha distribución no plano $\mathbb{R}^2$ que concentra toda a súa probabilidade na circunferencia. A outra opción é a de identificar cada vector de S^1 co ángulo que forma co semieixo positivo das x, medido en radiáns. Así asociamos cada punto de circunferencia cun punto do intervalo $(0, 2\pi]$, polo que podemos entender as distribucións circulares estudando as distribucións en $\mathbb{R}$ con soporte este intervalo. O primeiro destes enfoques permite xeneralizar facilmente os conceptos empregados a dimensións superiores. Como aquí só imos traballar con distribuicións na circunferencia unidimensional, empregaremos o segundo enfoque por simplicidade.

Tendo en conta o anterior, para nós un ángulo aleatorio Θ é unha aplicación Borel-medible entre un espazo de probabilidade Ω e o intervalo $(0, 2\pi]$:

$$\Theta : \Omega \to (0, 2\pi]. \tag{3.1}$$

Así, dado un ángulo aleatorio Θ, podemos definir a súa *función de distribución* F como a función de variable real definida no intervalo $(0, 2\pi]$ como

$$F(\theta) = \mathbb{P}(0 < \Theta \leq \theta), \quad 0 < \theta \leq 2\pi \tag{3.2}$$

e que extendemos ao resto da recta real mediante

$$F(\theta + 2\pi) - F(\theta) = 1, \quad \forall \theta \in \mathbb{R}. \tag{3.3}$$

A ecuación (3.3) garántenos que arco completo de circunferencia teña probabilidade 1. Ademais, permítenos calcular as probabilidades sen forzarnos a traballar con ángulos medidos entre 0 e 2π radiáns. Así, dados $\phi \leq \psi \leq \phi + 2\pi$, temos que

$$F(\psi) - F(\phi) = \mathbb{P}(\phi < \Theta \leq \psi). \tag{3.4}$$

Polo que poderíamos traballar con valores negativos, que se corresponden con ángulos medidos en sentido horario, ou con valores alén 2π, que se corresponden con dar varias voltas á circunferencia.

As funcións de distribución de ángulos aleatorios comparten varias caracterísitcas coas funcións de distribución dunha variable aleatoria real. Por exemplo, F tamén é unha función crecente e continua pola dereita en todo punto da recta real $\mathbb{R}$. Porén, as funcións de distribución de ángulos aleatorios non están limitadas ao intervalo $[0, 1]$, senón que verifican que

$$\lim_{\theta \to -\infty} F(\theta) = -\infty, \quad \lim_{\theta \to +\infty} F(\theta) = +\infty. \tag{3.5}$$

Ademais, F cumpre

$$F(0) = 0, \quad F(2\pi) = 1. \tag{3.6}$$

Diremos que o ángulo aleatorio Θ é *absolutemante continuo* cando a súa función de distribución F sexa unha función absolutamente continua en $\mathbb{R}$. Nese caso, existe unha función f cumprindo

$$\int_{\phi}^{\psi} f(\theta) d\theta = F(\psi) - F(\phi), \quad \forall \phi, \psi \in \mathbb{R}. \tag{3.7}$$

Tal función denomínase *función de densidade* do ángulo aleatorio Θ. Igual que as funcións de densidade das variables aleatorias, $f(\theta) \geq 0$ para case todo punto $\theta \in \mathbb{R}$. Como avanzar 2π radiáns dende calquera punto correspóndese con dar unha volta completa á circunferencia, temos que

$$\int_{\phi}^{\phi+2\pi} f(\theta)d\theta = 1, \quad \forall \phi \in \mathbb{R}. \tag{3.8}$$

Ademais, a ecuación (3.3) provoca que a f sexa unha función 2π-periódica, é dicir:

$$f(\theta) = f(\theta + 2\pi), \tag{3.9}$$

para case todo punto $\theta \in \mathbb{R}$.

Unha vez definida a función de densidade dun ángulo aleatorio, o concepto de moda xorde de forma análoga ao caso linear. Daquela, unha *moda* dun ángulo aleatorio Θ é un punto do intervalo $(0, 2\pi]$ onde a súa función de densidade f ten un máximo local. De forma paralela, unha *antimoda* de Θ é un punto do intervalo $(0, 2\pi]$ onde f ten un mínimo local.

Estes serán os conceptos basicos de estatístca circular que precisaremos neste traballo. Se se desexa profundizar máis no estudo dos datos circulares, recomendamos consultar os libros de Mardia e Jupp (2000) e Pewsey et al. (2013). Ambos os dous libros foron empregados para a realización desta sección.

3.2. Exemplos de distribucións circulares

Antes de pasar a traballar cos conceptos que nos van permitir construír un test de multimodalidade para datos circulares, presentaremos algúns exemplos de distribucións de ángulos aleatorios notables. Imos ver tres familias distintas: a distribución de von Mises, a normal enrolada e a distribución de von Mises *sine-skewed*.

3.2.1. A distribución de von Mises

Unha *distribución de von Mises* con dirección media μ e concentración $\kappa \geq 0$, que denotaremos por $\text{vM}(\mu, \kappa)$, é unha distribución dun ángulo aleatorio con función de densidade

$$f_{(\mu,\kappa)}(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}; \tag{3.10}$$

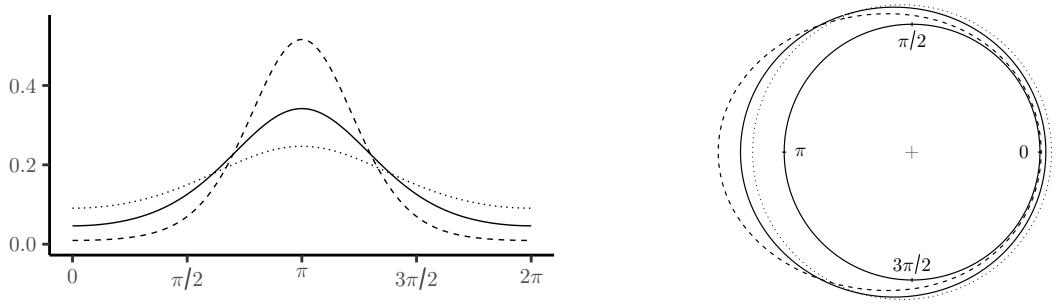
onde I_0 é a función de Bessel de primeira especie e orde 0 definida por

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos \theta}. \tag{3.11}$$

Todas as distribucións de von Mises son unimodais con moda en $\theta = \mu$ e antimoda en $\theta = 2\pi - \mu$. Ademais son sempre simétricas respecto a μ, é dicir:

$$f_{(\mu,\kappa)}(\mu + \theta) = f_{(\mu,\kappa)}(\mu - \theta), \quad \forall \theta \in \mathbb{R}. \tag{3.12}$$

Por outro lado, canto maior é o valor de concentración κ, máis se acumula a probalidade ao redor da moda. No caso caso especial no que $\kappa = 0$, a distribución de von Mises coincide coa distribución uniforme na circunferencia. Por outro lado, cando $\kappa \to +\infty$, a distribución $\text{vM}(\mu, \kappa)$ ten por límite a distribución dexenerada no ángulo $\theta = \mu$ (ver Mardia e Jupp, 2000, Cap. 12). Na Figura 3.3 pódense ver representadas as funcións de densidade dunha distribución de von Mises para varios valores do parámetro de concentración.



(a) (b)

Figura 3.3: Representación de funcións de densidades von Mises con dirección media $\mu = \pi$ e varios valores do parámetro κ. En trazo continuo aparece representada a función de densidade dunha $\text{vM}(\pi, 1)$; en trazo discontinuo aparece a densidade dunha $\text{vM}(\pi, 2)$; e en liña de puntos, a dunha $\text{VM}(\pi, 0.5)$. Están representadas tanto sobre a recta real (a), como sobre a circunferencia (b).

3.2.2. A distribución normal enrolada

Existen varias formas sinxelas de obter distribucións na circunferencia a partir de distribucións de variables aleatorias reais. Unha das máis comúns consiste en empregar a aplicación continua

$$\begin{array}{rcl} G : \mathbb{R} & \longrightarrow & S^1 \\ x & \longmapsto & (\cos(x), \sin(x)); \end{array} \tag{3.13}$$

que leva a recta real na circunferecia *enrolándoa* sobre si mesma. Así, se X é unha variable aleatoria real, entón $\Theta = G(X)$ é un ángulo aleatorio. Ademais, as funcións de distribución e densidade de X e de $\Theta = G(X)$ están relacionadas mediante as ecuacións

$$F_w(\theta) = \sum_{k \in \mathbb{Z}} F(\theta + 2\pi k) - F(2\pi k), \tag{3.14}$$

$$f_w(\theta) = \sum_{k \in \mathbb{Z}} f(\theta + 2\pi k); \tag{3.15}$$

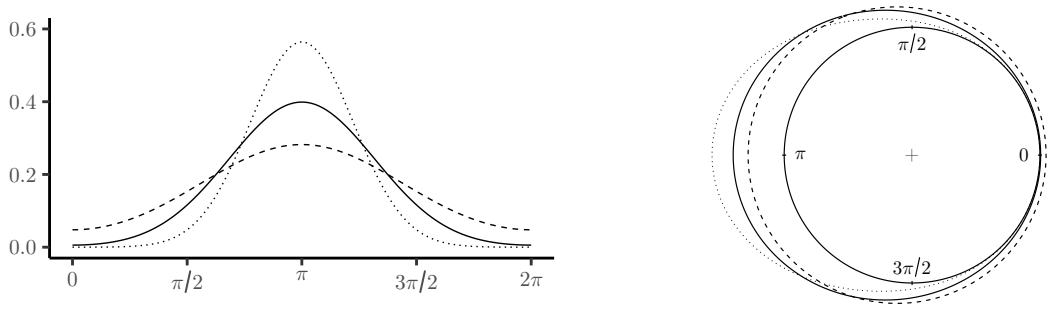
onde F e f son as funcións de distribución e densidade de X e F_w e f_w son as funcións de distribución e densidade de $\Theta = G(X)$.

O exemplo máis habitual deste tipo de distribucións na circunferencia é a *distribución normal enrolada* $\text{WN}(\mu, \sigma^2)$, que se constrúe enrolando a distribución normal $\text{N}(\mu, \sigma^2)$ mediante a aplicación G anterior. Así, tendo en conta a ecuación (3.15), unha normal enrolada $\text{WN}(\mu, \sigma^2)$ terá función de densidade

$$f_{(\mu,\sigma^2)}(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=-\infty}^{+\infty} \exp\left(\frac{-(\theta - \mu + 2\pi k)^2}{2\sigma^2}\right). \tag{3.16}$$

As propiedades da normal enrolada son similares as da distribución de von Mises. É tamén unha distribución unimodal con moda en $\theta = \mu$ e antimoda en $\theta = 2\pi - \mu$, e é simétrica respecto a μ. σ^2 ten un papel recíproco ao de κ: canto menor sexa o valor de σ^2, maior será a concentración de probabilidade ao redor da moda. Así, se $\sigma^2 \to 0$ a distribución límite de $\text{WN}(\mu, \sigma^2)$ é a dexenerada no ángulo μ, mentres que se $\sigma^2 \to +\infty$ a distribución límite é a distribución uniforme na circunferencia (consultar Mardia e Jupp, 2000, Cap. 12).

Na Figura 3.4 pódense ver representadas as densidades de distribucións normais enroladas para varios valores de σ^2.



(a) (b)

Figura 3.4: Representación de funcións de densidades normais enroladas con dirección media $\mu = \pi$ e varios valores do parámetro σ^2. En trazo continuo aparece representada a función de densidade dunha $\text{WN}(\pi, 1)$; en trazo discontinuo aparece a densidade dunha $\text{WN}(\pi, 2)$; e en liña de puntos, a dunha $\text{WN}(\pi, 0.5)$. Están representadas tanto sobre a recta real (a), como sobre a circunferencia (b).

3.2.3. A distribución de von Mises *sine-skewed*

As dúas distribución circulares consideradas ata o de agora son simétricas respecto a súa moda. Non obstante é posible que adiquiran certo grao de asimetría modificando lixeiramente a súa función de densidade, tal como mostran Umbach e Jammalamadaka (2009). Esta idea permite definir a distribución de von Mises *sine-skewed* $\text{ssvM}(\mu, \kappa, \lambda)$, onde $-1 \leq \lambda \leq 1$, como a distribución na circunferencia con función de densidade:

$$f_{(\mu,\kappa,\lambda)}(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta-\mu)}[1 + \lambda \sin(\theta - \mu)]; \tag{3.17}$$

onde a función I_0 é a función de Bessel de primeira especie e orde 0 definida en (3.11).

Como se pode ver, a desidade dunha distribución $\text{ssvM}(\mu, \kappa, \lambda)$ é a densidade dunha distribución de von Mises $\text{vM}(\mu, \kappa)$ multiplicada polo factor $1 + \lambda \sin(\theta - \mu)$, permitindo así dotala de certa asimetría. O parámetro λ controla o grao de asimetría da distribución. Canto máis próximo a 1 sexa o valor de $|\lambda|$ maior será a súa asimetría. Ademais, o signo de λ indica a dirección da asimetría: antihoraria se o signo é positivo, horaria se é negativo.

Na Figura 3.5 aparecen as funcións de densidade de distribución de von Mises *sine-skewed* para varios valores do parámetro λ.

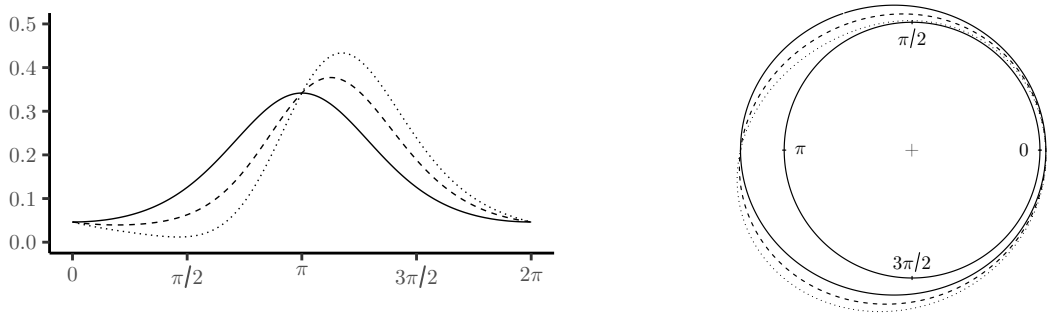
3.3. O novo test de multimodalidade para datos circulares

Unha vez introducidas as ferramentas básicas necesarias para traballar con datos circulares, podemos empezar a extender o noso test de multimodalidade a este novo tipo de datos. Primeiro concretemos cal é o noso obxectivo neste contexto. Sexa Θ un ángulo aleatorio, sexa f a súa función de densidade e j o seu número de modas. O que queremos é construír un test estatístico que nos permita contrastar as hipóteses

$$H_0 : j \leq k, \quad H_1 : j > k; \tag{3.18}$$

onde k é un número natural dado.

Seguindo as mesmas ideas que no caso linear, o primeiro que precisamos é un estimador da función de densidade. Dada $\Theta_1, \ldots, \Theta_n$ unha mostra aleatoria simple do ángulo aleatorio absolutamente continuo Θ, o estimador da función de densidade de Θ análogo ao estimador tipo núcleo do caso linear



(a) (b)

Figura 3.5: Representación de funcións de densidades von Mises *sine-skewed* con $\mu = \pi$, $\kappa = 1$ e varios valores do parámetro de asimetría λ. En trazo continuo aparece representada a función de densidade dunha ssvM$(\pi, 1, 0)$; en trazo discontinuo aparece a densidade dunha ssvM$(\pi, 1, 0.5)$; e en liña de puntos, a dunha ssvM$(\pi, 1, 0.9)$. Están representadas tanto sobre a recta real (a), como sobre a circunferencia (b).

é

$$\hat{f}_h(\theta) = \frac{1}{n}\sum_{i=1}^{n} K_h(\theta - \Theta_i) \tag{3.19}$$

onde $h > 0$ é o parámetro de suavizado ou xanela, e K_h é a densidade dunha normal enrolada WN$(0, h^2)$.

Igual que sucede cando traballamos con datos na recta real, o valor do parámetro de suavizado ten un grande impacto no estimador $\hat{f}_h$, sendo posible que dous valores distintos de h den lugar a dous estimadores de f cun número de modas distinto. De feito, tal e como proban Huckemann et al. (2016), o uso da normal enrolada en (3.19) garante que o número de modas de modas de $\hat{f}_h$ é monótono respecto a h: canto menor sexa o valor de h, maior será o número de modas de $\hat{f}_h$. Esta propiedade permítenos definir a xanela crítica en datos circulares de forma análoga ao caso linear. Así, dado k un número natural, a *xanela crítica para k modas*, h_k, defínese como:

$$h_k = \text{mín}\{h : \hat{f}_h \text{ ten como máximo } k \text{ modas}\}. \tag{3.20}$$

O estimador da función de densidade dun ángulo aleatorio definido en (3.19) tamén nos permite construír unha función de pseudo-verosimilitude para datos circulares seguindo o mesmo razoamento que no caso linear. Así, para cada $i = 1, \ldots, n$, definimos as funcións auxiliares

$$\hat{f}_h^{-i}(\theta) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} K_h\left(\theta - \Theta_j\right). \tag{3.21}$$

A partir das funcións $\hat{f}_h^{-i}$ podemos definir a *pseudo-verosimilitude por validación cruzada* da mostra $\Theta_1, \ldots, \Theta_n$ mediante

$$\mathcal{L}_{CV}(h) = \prod_{i=1}^{n} \hat{f}_h^{-i}(\Theta_i), \quad h > 0. \tag{3.22}$$

Unha vez temos definidos os conceptos de xanela crítica e pseudo-verosimilitude para mostras circulares, a xeralización do test de multimodalidade do capítulo anterior a este novo contexto xorde de forma inmediata. O noso estatístico de contraste será, por tanto:

$$D_k = 2\left[\sup_{h>0} \ell_{CV}(h) - \sup_{h\geq h_k} \ell_{CV}(h)\right], \tag{3.23}$$

onde $\ell_{CV}(h) = \log \mathcal{L}_{CV}(h)$. Rexeitaremos a hipótese nula, é dicir, que o ángulo aleatorio Θ ten como máximo k modas, para valores grandes de D_k. Igual que no caso linear, o calibrado do test realizarase por bootstrap suavizado, lanzando varias remostras da función de densidade $\hat{f}_{h_k}$ para tratar de aproximar a distribución baixo a nula do estatístico D_k.

Se reparamos na definición do estatístico D_k veremos que en principio podería non estar ben definido. Se a función $\mathcal{L}_{CV}$ non estivese limitada superiormente, algún dos dous supremos presentes en (3.23) podería ser $+\infty$ e por tanto o estatístico de contraste non se podería calcular. A seguinte proposición, análoga á Proposición 1 probada no caso linear, garántenos que a función $\mathcal{L}_{CV}$ está limitada superiormente con probabilidade 1, e por tanto D_k é sempre calculable na práctica.

Proposición 2. *Sexa $\Theta_1, \ldots, \Theta_n$ unha mostra aleatoria simple dun ángulo aleatorio absolutamente continuo Θ. Sexa $\mathcal{L}_{CV}$ a función de pseudo-verosimilitude por validación cruzada definida en* (3.22)*. Entón a función $\mathcal{L}_{CV}(h)$ está limitada superiormente no intervalo $(0, +\infty)$ con probabilidade 1.*

Demostración. Sexa K_h a función de densidade dunha normal enrolada $\mathrm{WN}(0, h^2)$ definida en (3.16). Tendo en conta que $\lim_{h\to+\infty} K_h(\theta) = 1/2\pi$ para todo $\theta \in \mathbb{R}$, temos que:

$$\lim_{h\to+\infty} \hat{f}_h^{-i}(\theta) = \lim_{h\to+\infty} \frac{1}{(n-1)} \sum_{j=1, j\neq i}^{n} K_h\left(\theta - \Theta_j\right) = \frac{1}{2\pi} \tag{3.24}$$

para todo $\theta \in \mathbb{R}$ e $i = 1, \ldots, n$. Daquela tamén se verifica que

$$\lim_{h\to+\infty} \mathcal{L}_{CV}(h) = \lim_{h\to+\infty} \prod_{i=1}^{n} \hat{f}_h^{-i}(\Theta_i) = (2\pi)^{-n}. \tag{3.25}$$

Por outro lado, temos que

$$K_h(\theta) = \frac{1}{\sqrt{2\pi h^2}} \sum_{k\in\mathbb{Z}} \exp\left(-\frac{(\theta - 2\pi k)^2}{2h^2}\right) \tag{3.26}$$

Imos estudar a serie en (3.26) separando os termos positivos dos negativos. Daquela, para os positivos temos que:

$$\begin{aligned}\sum_{k=0}^{+\infty} \exp\left(-\frac{(\theta + 2\pi k)^2}{2h^2}\right) &\leq \sum_{k=0}^{+\infty} \exp\left(-\frac{\theta^2}{2h^2} - \frac{2\pi^2 k^2}{h^2}\right) \leq \\ &\leq \exp\left(-\frac{\theta^2}{2h^2}\right) \sum_{k=0}^{+\infty} \exp\left(-\frac{2\pi^2 k}{h^2}\right) = \exp\left(-\frac{\theta^2}{2h^2}\right) \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1}. \end{aligned} \tag{3.27}$$

A primeira desigualdade de (3.27) dedúcese de que $(a+b)^2 \geq a^2 + b^2$ se $a, b \geq 0$. A última desigualdade é inmediata, pois estamoslle a sumar máis términos á serie (hai máis números naturais que cadrados perfectos).

Para os termos negativos, temos que:

$$\begin{aligned}\sum_{k=1}^{+\infty} \exp\left(-\frac{(\theta - 2\pi k)^2}{2h^2}\right) &= \sum_{k=1}^{+\infty} \exp\left(-\frac{[(\theta - 2\pi) - 2\pi(k-1)]^2}{2h^2}\right) = \\ &= \sum_{k=0}^{+\infty} \exp\left(-\frac{[(2\pi - \theta) + 2\pi k]^2}{2h^2}\right) \leq \exp\left(-\frac{(2\pi - \theta)^2}{2h^2}\right) \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1}. \end{aligned} \tag{3.28}$$

Onde a última desigualdade se deduce de aplicar (3.27).

Do anterior deducemos que:

$$K_h(\theta) \leq \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1} \left[\frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(2\pi-\theta)^2}{2h^2}\right) + \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{\theta^2}{2h^2}\right)\right]; \tag{3.29}$$

e por tanto $\lim_{h\to 0} K_h(\theta) = 0$ para todo $\theta \in (0, 2\pi)$.

Supoñamos agora que todos os valores $\Theta_1, \ldots, \Theta_n$ son todos distintos entre si. Entón, o anterior implica

$$\lim_{h\to 0} \hat{f}_h^{-i}(\Theta_i) = \lim_{h\to 0} \frac{1}{(n-1)} \sum_{j=1, j\neq i}^{n} K_h\left(\Theta_i - \Theta_j\right) = 0, \tag{3.30}$$

para todo $i \in \{1, \ldots, n\}$ e por tanto $\lim_{h\to 0} \mathcal{L}_{CV}(h) = 0$. Como $\mathcal{L}_{CV}(h)$ é continua en $(0, +\infty)$, entón temos que $\mathcal{L}_{CV}(h)$ está limitada no intervalo $(0, +\infty)$ sempre que os valores $\Theta_1, \ldots, \Theta_n$ sexan todos distintos entre si. Como Θ é un ángulo aleatorio absolutamente continuo, isto sucede con probabilidade 1. □

Así, de xeito esquemático, o test de multimodalidade para datos circulares é:

1. A partir da mostra $\Theta_1, \ldots, \Theta_n$, obtemos a xanela crítica

   $$h_k = \text{mín}\{h > 0 : \hat{f}_h \text{ ten } k \text{ modas }\}$$

   e as xanelas que maximizan a verosimilitude por validación cruzada baixo a hipótese nula e a alternativa:

   $$\mathcal{L}_{CV}(h_{max}) = \max_{h>0}\{\mathcal{L}_{CV}(h)\}; \quad \mathcal{L}_{CV}(h_{H_0}) = \max_{h\geq h_k}\{\mathcal{L}_{CV}(h)\}.$$

   E con elas calculamos o estatístico de contraste:

   $$D_k = 2\left[\ell_{CV}(h_{max}) - \ell_{CV}(h_{H_0})\right].$$

2. Obtemos a remostra $\Theta_1^*, \ldots, \Theta_n^*$ da densidade suavizada $\hat{f}_{h_k}$ e calculamos o valor do estatístico para esa mostra: D_k^*.

3. Repetimos B veces o paso 2, conseguindo así B réplicas do estatístico: $D_k^{*,1}, D_k^{*,2}, \ldots, D_k^{*,B}$.

4. O test rexeitará a hipótese nula de que f ten como máximo k modas se

   $$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(D_k^{*,b} > D_k) < \alpha.$$

3.4. A xanela h_{max} con datos circulares

Continuando cos paralelismos entre os casos linear e circular, tamén podemos empregar a pseudo-verosimilitude circular para seleccionar un parámetro de suavizado h á hora de construír o estimador da densidade $\hat{f}_h$ que definimos en (3.19). Dada a mostra circular $\Theta_1, \ldots, \Theta_n$, poderíase escoller como xanela o valor $h_{max} > 0$ que verifique que

$$\mathcal{L}_{CV}(h_{max}) = \max_{h>0} \mathcal{L}_{CV}(h); \tag{3.31}$$

que existirá sempre que o núcleo K_h que usemos sexa a normal enrolada, tal como nos garante a Proposición 2.

As principais propiedades deste selector da xanela foron estudadas por Hall et al. (1987). Ao contrario que todo o anterior, o comportamento de h_{max} en datos circulares non é análogo ao caso linear. Para datos circulares h_{max} é un valor *óptimo* do parámetro de suavizado, no sentido de que

o estimador tipo núcleo $\hat{f}_{h_{max}}$ é un estimador asintóticamente consistente da verdadeira función de densidade f, sempre que f sexa o suficientemente regular e esté limitada fóra de cero (para máis información sobre as condicións de regularidade necesarias, véxase Hall et al., 1987). h_{max} é, polo tanto, unha escolla habitual do parámetro de suavizado á hora de estimar a función de densidade dun ángulo aleatorio mediante unha estimación tipo núcleo. Isto explica por que decidimos extender o test baseado en pseudo-verosimilitude á circunferencia a pesar do seu mal comportamento no estudo de simulación con datos lineais: sabíamos que, a priori, as ideas nas que se fundamenta este contraste teñen un comportamento aceptable con datos circulares, contrariamente ao que acontece na recta real. Esperamos logo que o test imite esta dinámica, e que teña un comportamento aceptable na práctica. Ademais, a ausencia de colas ao traballar con datos circulares fai que non se precise realizar unha modificación da xanela crítica similar a feita por Hall e York (2001) no caso linear, facilitando a formulación e aplicación do test.

Por outra parte, observando o parágrafo anterior pódese adiantar unha das posibles fraquezas do novo contraste de multimodalidade para datos circulares. Como precisamos que f esté limitada fóra de 0 para garantir a consistencia do estimador $\hat{f}_{h_{max}}$, é esperable que o test teña problemas ao detectar o verdadeiro número de modas dunha distribución cando a densidade desta se anula, ben nun punto, ben nun arco de circunferencia de medida positiva. Daquela, deberemos de prestarlle especial atención a este tipo de situacións á hora de comprobar o rendemento de test na práctica.

3.5. Estudo de simulación

Igual que aconteceu no caso linear, o test de multimodalidade para datos circulares introducido na sección anterior non está cimentado en resultados teóricos, senón que xorde da nosa intuición a hora de extender o test de razón de verosimilitudes paramétrico a este novo contexto. Daquela, para comprobar se as nosas intuicións eran certas e o test está ben calibrado, realizamos o estudo de simulación que presentamos nesta sección. Nel simulamos $M = 1000$ mostras de tamaño n de diversas distribucións unimodais circulares. Aplicamos o test baseado en pseudo-verosimilitude a cada unha das mostras, vendo se rexeitan ou non a unimodalidade dos datos para varios niveis de significación α. Finalmente, calculamos a proporción de mostras rexeitadas do total para cada nivel de significación. Como só consideramos distribucións unimodais (é dicir, baixo a hipótese nula) se os tests están ben calibrados as proporcións de rexeitamentos deberían de ser similares ao nivel de significación α.

Imos realizar o anterior para dous tamaños de mostra distintos, $n = 100$ e $n = 500$, e cinco distribucións circulares unimodais distintas. Os niveis de significación considerados son os tres máis usuais: $\alpha = 0.01$, $\alpha = 0.05$ e $\alpha = 0.1$.

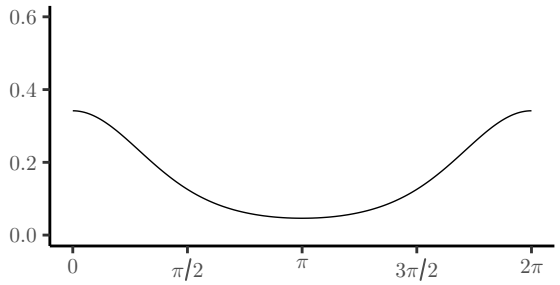
Para a realización do test empregouse código R escrito expresamente para a realización deste estudo de simulación. Todas as simulacións realizáronse empregando os recursos computacionais do Centro de Supercomputación de Galicia (CESGA).

3.5.1. Distribucións consideradas

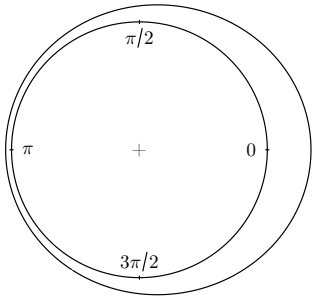
As cinco distribucións direccionais escollidas no estudo de simulación foron as que seguen.

- **Modelo Circular 1 (MC1)**: unha distribución de von Mises $\text{vM}(0, 1)$.
- **Modelo Circular 2 (MC2)**: unha mixtura de von Mises: $0.2 \cdot \text{vM}(2\pi/3, 3) + 0.6 \cdot \text{vM}(\pi, 1.4) + 0.2 \cdot \text{vM}(4\pi/3, 3)$.
- **Modelo Circular 3 (MC3)**: unha mixtura de von Mises: $0.05 \cdot \text{vM}(2\pi/3, 7) + 0.9 \cdot \text{vM}(\pi, 1) + 0.05 \cdot \text{vM}(4\pi/3, 7)$.
- **Modelo Circular 4 (MC4)**: unha von Mises *sine-skewed*: $\text{ssvM}(\pi, 1, -0.9)$.
- **Modelo Circular 5 (MC5)**: unha distribución beta modificada para que o soporte sexa o intervalo $[\pi/2, 3\pi/2]$ e enrolada: $G\left(\pi \cdot \text{Beta}(3, 2) + \pi/2\right)$.

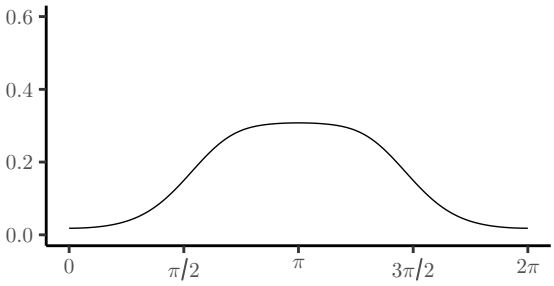
As densidades destas distribucións están representadas nas Figuras 3.6 e 3.7. Buscan representar a unha gran variedade de situacións, empezando por casos simples (unha von Mises, Modelo Circular 1), pasando por modas planas (Modelo Circular 2), distintos graos de asimetría (Modelos 4 e 5) e finalizado cunha distribución sectorial, onde todos os datos están concentrados na semicircunferencia esquerda (Modelo 5).



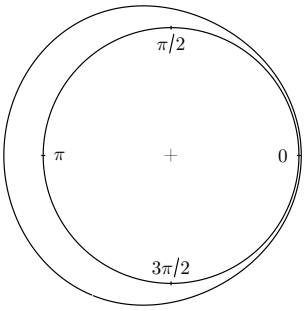
(a) Modelo Circular 1 (MC1).



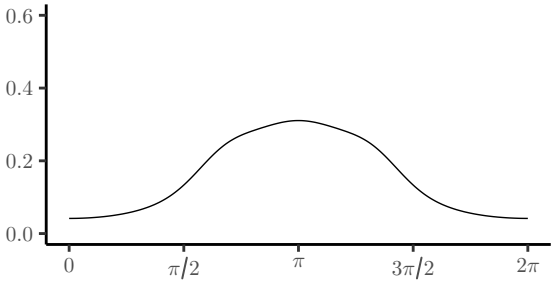
(b) Modelo Circular 1 (MC1) representado sobre a circunferencia.



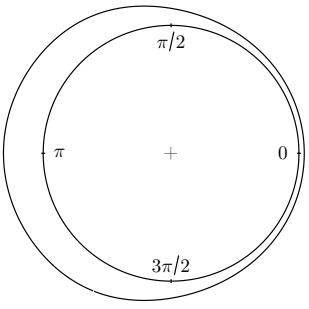
(c) Modelo Circular 2 (MC2).



(d) Modelo Circular 2 (MC2) representado sobre a circunferencia.

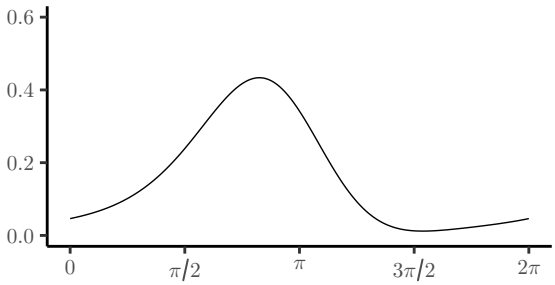


(e) Modelo Circular 3 (MC3).

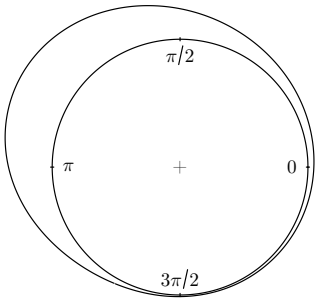


(f) Modelo Circular 3 (MC3) representado sobre a circunferencia.

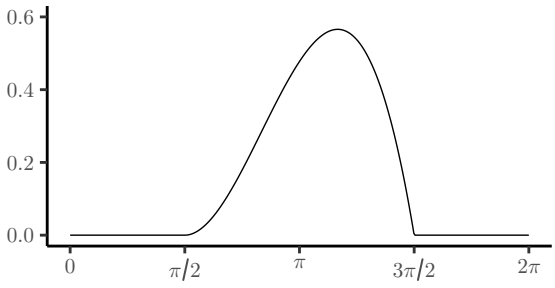
Figura 3.6: Representación das funcións de densidade dos Modelos Circulares 1, 2 e 3 sobre a recta real (columna esquerda) e a circunferencia (columna dereita).



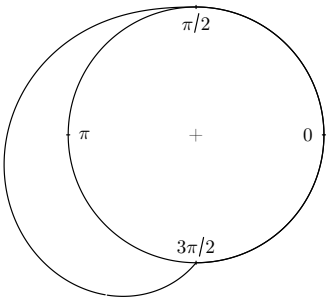
(a) Modelo Circular 4 (MC4).



(b) Modelo Circular 4 (MC4) representado sobre a circunferencia.



(c) Modelo Circular 5 (MC5).



(d) Modelo Circular 5 (MC5) representado sobre a circunferencia.

Figura 3.7: Representación das funcións de densidade dos Modelos Circulares 4 e 5 sobre a recta real (columna esquerda) e a circunferencia (columna dereita).

3.5.2. Resultados

Na Táboa 3.1 aparecen os resultados do estudo de simulación realizado, e pódese observar que o test baseado na pseudo-verosimilitude presenta un mellor comportamento para datos circulares que para datos lineais. O test só presenta proporcións de rexeitamento superiores ao nivel de significación para o Modelo Circular 5. Corrobórase por tanto a sospeita que expuxemos na sección anterior e o test ten problemas á hora de detectar a hipótese nula cando a función de densidade dos datos se anula nun sector circular de lonxitude positiva.

Para os Modelos Circulares 2, 3 e 4 o test é conservador para mostras pequenas ($n = 100$) na gran maioría dos casos, con proporcións de rexeitamento menores ao nivel de significación. As únicas excepcións están no Modelo Circular 4 con $\alpha = 0.01$ e $\alpha = 0.05$, onde as proporcións de rexeitamento están moi próximas ao nivel nos dous casos. Pola contra, para mostras de tamaño $n = 500$ o test ofrece proporcións de rexeitamento similares ao nivel nestes tres modelos, salvo no caso do Modelo Circular 4 con $\alpha = 0.1$, onde a proporción de rexeitamento está lixeiramente por debaixo do nivel. Por último, para o Modelo Circular 1 o test é conservador para todos os tamaños e niveis de significación, é dicir, as proporcións de rexeitamento están por debaixo do nivel para todos os casos.

3.6. Análise de datos reais

Unha vez comprobado o bo calibrado do novo test de multimodalidade con datos circulares, empregarémolo para responder a pregunta formulada ao inicio deste capítulo. A estrutura das vagas de

Táboa 3.1: Proporcións de rexeitamento para o test de multimodalidade baseado en pseudo-verosimilitude con nivel de significación ao 1 %, 5 % e 10 % calculadas a partir de $M = 1000$ mostras. Ao lado de cada proporción de rexeitamento, e entre paréntese, aparece a súa desviación típica estimada multiplicada por 1.96. O test calibrouse empregando $B = 500$ remostras. Cada fila correspóndese cunha combinación de distribución e tamaño de mostra distinta.

		1 %	5 %	10 %
MC1	$n = 100$	0.001(0.002)	0.013(0.007)	0.038(0.012)
	$n = 500$	0.005(0.004)	0.033(0.011)	0.065(0.015)
MC2	$n = 100$	0.003(0.003)	0.036(0.012)	0.072(0.016)
	$n = 500$	0.015(0.008)	0.053(0.014)	0.101(0.019)
MC3	$n = 100$	0.004(0.004)	0.027(0.010)	0.057(0.014)
	$n = 500$	0.014(0.007)	0.049(0.013)	0.095(0.018)
MC4	$n = 100$	0.011(0.006)	0.047(0.013)	0.075(0.016)
	$n = 500$	0.009(0.006)	0.041(0.012)	0.081(0.017)
MC5	$n = 100$	0.011(0.006)	0.063(0.015)	0.133(0.021)
	$n = 500$	0.030(0.011)	0.138(0.021)	0.232(0.026)

incendios vese condicionada pola actividade humana en Galicia? É dicir, existen probas estatisticamente significativas para afirmar que existe máis dunha moda nos datos de Ameijeiras et al. (2018)? Calculamos o estatístico de contraste para a mostra dos incendios é obtemos o valor $D_1 = 8142.012$. Estimamos o p-valor asociado a mostra mediante $B = 500$ remostras, e a aproximación conseguida do p-valor é 0. Polo tanto, o contraste rexeita a unimodalidade dos datos baixo calquera nivel de significación usual. Ou o que é o mesmo, hai probas estatisticamente significativas para afirmar que hai máis dunha moda na mostra, algo que, como indicamos ao principio deste capítulo, podería verse explicado pola actividade humana en Galicia.

Capítulo 4

Discusión

Os bos resultados observados no estudo de simulación do capítulo anterior deberían animarnos a continuar investigando o comportamento do novo test de multimodalidade para datos circulares. Os primeiros pasos a seguir son evidentes: cómpre analizar a potencia do novo novo test a hora de detectar a multimodalidade en datos circulares, así como comparar o seu comportamento na práctica con outros test de multimodalidade na circunferencia, como pode ser o test de Ameijeiras et al. (2019). Tamén habería que pasar das intuicións as realidades, é dicir, fundamentar teoricamente o test. Por exemplo, teríase que probar baixo que condicións de regularidade sobre a función de densidade f o contraste está asintóticamente ben calibrado. Se, tal como parece, a condición de que f esté limitada fóra de cero é fundamental para o bo calibrado asíntotico do test, sería interesante estudar se hai algunha forma de contrastar se se verifica esa condición a partir unha mostra aleatoria. Por outra parte, tamén sería de interese estudar a distribución límite do estatístico de contraste D_k, no caso de que esta exista, para así comprobar se hai un análogo ao Teorema de Wilks neste contexto.

A outra liña de investigación a seguir centraríase na adaptabilidade do test que tanto enfatizamos ao longo do traballo. Á vista dos malos resultados obtidos con datos lineais, non parece que o test vaia ter un comportamento satisfactorio en espazos como $\mathbb{R}^m$ ou o cilindro. Neses espazos, a imposibilidade de que as funcións de densidade estén limitadas fóra de cero seguramente se reflicta nun mal comportamento da función de pseudo-verosimilitude de xeito similar ao que acontece na recta real, imposiblitando a obtención dun test ben calibrado. Por tanto, as extensións deste test terán máis sentido en variedades compactas (como a esfera ou o toro) onde a esixencia de que f esté limitada fóra de cero si que resulta factible. Nestes contextos, as ferramentas básicas do contraste, que son o estimador tipo núcleo da función de densidade e a pseudo-verosimilitude, si que teñen un traslado directo, análogo ao razoamento feito para pasar de datos lineais a circulares. A principal problemática de adaptar o test a estes espazos seguramente esté no calibrado do mesmo, pois en máis dunha dimensión non hai polo de agora un concepto paralelo á xanela crítica unidimensional. Por tanto, o noso calibrado mediante bootstrap suavizado, que se apoia plenamente na xanela crítica, debe de ser reformulado.

Bibliografía

[1] Ameijeiras Alonso, J. (2017). *Assessing Simplifying Hypotheses in Density Estimation*. Tese de Doutoramento, Universidade de Santiago de Compostela.

[2] Ameijeiras–Alonso, J., Crujeiras, R. M., Rodríguez-Casal, A. (2018). Directional statistics for wildfires. En: *Applied directional statistics*. Chapman and Hall/CRC, 187-210.

[3] Ameijeiras-Alonso, J., Crujeiras, R. M., e Rodríguez-Casal, A. (2019). Mode testing, critical bandwidth and excess mass. *Test*, 28, 900-919.

[4] Ameijeiras-Alonso, J., Benali, A., Crujeiras, R. M., Rodríguez-Casal, A., e Pereira, J. M. (2019). Fire seasonality identification with multimodality tests. *Annals of Applied Statistics*, 13, 2120-2139.

[5] Ameijeiras-Alonso, J., Crujeiras, R.M. e Rodríguez-Casal, A. (2021). multimode: an R package for mode assessment. *Journal of Statistical Software*, 97, 1-32

[6] Chaudhuri, P., e Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94, 807-823.

[7] Cheng, M. Y., e Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society*. Series B, 60, 579-589.

[8] Choi, D. Y., Wittig, T. W., e Kluever, B. M. (2020). An evaluation of bird and bat mortality at wind turbines in the Northeastern United States. *PloS one*, 15, e0238034.

[9] Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, 25, 1175-1179.

[10] Fisher, N. I., e Marron, J. S. (2001). Mode testing via the excess mass estimate. Biometrika, 88, 499-517.

[11] Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15, 1491-1519.

[12] Hall, P., Watson, G. S., e Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, 74, 751-762.

[13] Hall, P., e York, M. (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica*, 11, 515-536.

[14] Huckemann, S., Kim, K. R., Munk, A., Rehfeldt, F., Sommerfeld, M., Weickert, J., e Wollnik, C. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli*, 22, 2113-2142.

[15] Karlin, S. (1957). Pólya type distributions, II. *The Annals of Mathematical Statistics*, 28, 281-308.

[16] Neyman, J., e Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289-337.

[17] Mammen, E., Marron, J. S., e Fisher, N. I. (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probability Theory and Related Fields*, 91, 115-132.

[18] Mardia, K. V., e Jupp, P. E. (2000). *Directional Statistics*. John Wiley & Sons.

[19] McLachlan, G., e Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

[20] Minnotte, M. C., e Scott, D. W. (1993). The mode tree: A tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics*, 2, 51–68.

[21] Minnotte, M. C., Marchette, D. J., e Wegman, E. J. (1998). The bumpy road to the mode forest. *Journal of Computational and Graphical Statistics*, 7, 239–251.

[22] Müller, D. W., e Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Annals of Statistics*, 86, 738-746.

[23] Pewsey, A., Neuhäuser, M., e Ruxton, G. D. (2013). *Circular Statistics in R*. Oxford University Press.

[24] Oliveira, M., Crujeiras, R. M., e Rodriguez-Casal A. (2014). NPCirc: An R Package for Nonparametric Circular Methods. *Journal of Statistical Software*, 61, 1-26

[25] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society*. Series B, 43, 97-99.

[26] Umbach, D., e Jammalamadaka, S. R. (2009). Building asymmetry into circular distributions. *Statistics & probability letters*, 79, 659-663.

[27] Wand, M. P., e Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall.

[28] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9, 60-62.