

XV Congreso Galego de Estatística e Investigación de Operacións  
Santiago de Compostela, 4, 5 e 6 de novembro de 2021

## Un novo test de unimodalidade para datos circulares

Diego Bolón<sup>1</sup>, Rosa M. Crujeiras<sup>1</sup> e Alberto Rodríguez-Casal<sup>1</sup>

<sup>1</sup> Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela.

### RESUMO

As modas dunha poboación son puntos de alta frecuencia ao redor dos cales se acumula a maior parte da probabilidade. Na literatura estatística existen varios procedementos non paramétricos para contrastar o número de modas en datos na recta real. Pero a metodoloxía empregada na maioría destes tests 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 circulares, procurando ademais que sexa facilmente extensible a outros contextos, como poden ser o toro ou a esfera. Para iso, introducimos a idea de pseudo-verosimilitude como un análogo da función de verosimilitude da estatística paramétrica. Isto permítenos formular o noso test coa mesma estrutura dos tests de razón de verosimilitudes paramétricos apoiándonos no concepto de xanela crítica. Ademais, a pseudo-verosimilitude pode 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. Unha vez proposto o novo test para contrastar o número de modas dunha poboación circular, comprobamos cal é o seu calibrado e potencia na práctica mediante un estudo de simulación. Para finalizar, ilustramos o funcionamento do test aplicándoo a un conxunto de datos reais.

**Palabras e frases chave:** Contraste; datos circulares; estimación tipo núcleo da densidade; multimodalidade; verosimilitude; xanela crítica.

### 1. INTRODUCCIÓN E MOTIVACIÓN

Unha moda dun ángulo aleatorio absolutamente continuo  $X$  é un punto  $x_0 \in (0, 2\pi]$  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*. Para tratar de realizar inferencia sobre o número de modas dunha poboación nacen os *tests de multimodalidade*. Dado  $X$  un ángulo aleatorio absolutamente continuo con  $j$  modas, un test de multimodalidade é un test estatístico que contrasta as hipóteses

$$H_0: j \leq k \text{ fronte a } H_1: j > k; \quad (1)$$

onde  $k$  é un número natural fixado de antemán. Neste traballo centrarémonos no test que contrasta unimodalidade fronte a multimodalidade, é dicir

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

Coa intención de motivar a necesidade deste tipo de contrastes introducimos os datos analizados por Ameijeiras-Alonso et al. (2018). Esta base de datos contén todos os incendios detectados polo sensor de imaxe 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, coñecer o número de modas, que neste contexto se corresponden coas tempadas de incencios ao longo do ano, cobra especial relevancia á hora de entender os seus patróns estacionais dos incendios forestais e así loitarmos mellor contra esta problemática.

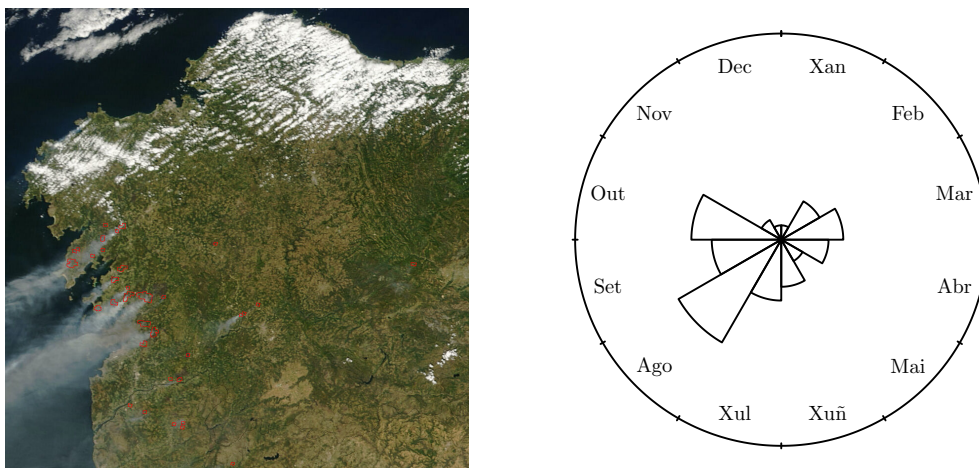


Figura 1: Á esquerda, os incendios (marcados en vermello) detectados polo satélite MODIS en Galicia o día 7 de agosto de 2006. Á dereita, o histograma circular (*rose diagram*) cos incendios detectados en Galicia dende o 10 de Xullo de 2002 ata o 9 de Xullo de 2012, detectados polo satélite MODIS da NASA.

Na Figura 1 pódense ver os datos anteriores representados nun histograma circular. Neste gráfico dividimos a circunferencia en doce rexións, cada unha correspondente a un mes do ano. Despois engadimos un sector circular a cada unha das seccións, de xeito que a área de cada sector sexa directamente proporcional a cantidade de incendios detectada nese mes. No histograma circular 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, en marzo. Estes dous picos de incendios están situados fóra da parte do ano con condicións meteorolóxicas máis favorables para a formación natural de incendios, polo que poden estar asociados a certos comportamentos humanos, como a queima preventiva ou a queima de restollos. Así, se tivesemos un test de multimodalidade que nos permita afirmar que hai probas estatisticamente significativas para a existencia de máis dunha moda con esta mostra estaríamos aportando evidencia de que a actividade humana condiciona a estrutura das vagas de incendios ao longo do ano.

Tendo en conta o anterior, o noso obxectivo é construír un test de multimodalidade para datos circulares que conte cunha metodoloxía facilmente adaptable a outros contextos, como poden ser datos direccionais, multidimensionais, no toro...

A estrutura deste traballo é a seguinte. Na Sección 2 comezamos introducindo brevemente as principais características do test de razón de verosimilitudes paramétrico, e apoiándonos nelas construímos a nosa proposta de test de multimodalidade para datos circulares. Na Sección 3 falamos das propiedades do selector de xanela  $h_{max}$ , que teñen consecuencias no comportamento do noso test. Na Sección 4 realizamos dous estudos de simulación para comprobar o calibrado e potencia do test na práctica e comentamos os resultados obtidos. Na Sección 5 aplicamos o test aos datos presentados nesta introdución e finalmente na Sección 6 comentamos as principais conclusións do traballo e como se podería adaptar esta nova proposta de test de multimodalidade a outros espazos que non sexan a circunferencia. A proba da Proposición 1, que precisamos para a construción do test, está recollida na Sección 7.

## 2. PROPOSTA DE TEST

Para abordar o problema (1), propoñemos un test de multimodalidade 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  $X_1, X_2, \dots, X_n$  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  $\Theta_0$  e  $\Theta_1$  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

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

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

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

A idea detrás da verosimilitude é que  $\mathcal{L}(\theta_0)$  representa a *factibilidade* de que o valor do parámetro descoñecido sexa  $\theta_0$  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  $\theta$  sexa igual a  $\theta_0$  que a  $\theta_1$  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  $\Theta_0$ , que é o que se corresponde a  $H_0$ . Polo tanto, un candidato a estatístico de contraste sería

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

que estará ben definido se a verosimilitude  $\mathcal{L}$  está limitada en todo o espazo de parámetros  $\Theta$ . Tendo en conta o razoamento anterior, o test de razón de verosimilitudes rexeitará a hipótese nula para valores pequenos do estatístico  $\lambda$ .

En vez de empregar  $\lambda$ , adóitase utilizar o estatístico equivalente

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

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,  $D$  converge 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. Baseando o novo test de multimodalidade no test de razón de verosimilitudes buscamos que herde estas boas características, así como conseguir unha metodoloxía adaptable para traballar en calquera espazo.

Para construír un análogo non paramétrico da función de verosimilitude apoiámonos no estimador tipo núcleo da densidade. Sexa  $X_1, \dots, X_n$  unha mostra aleatoria simple do ángulo aleatorio  $X$ . O estimador tipo núcleo da función de densidade de  $X$  é a función

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i); \quad (2)$$

onde  $h \in \Theta = (0, +\infty)$  é un número real positivo, e  $K_h$  é unha familia de densidades circulares simétricas e centradas no ángulo cero indexadas polo parámetro  $h$ . As funcións  $K_h$  denomínanse

*funcións núcleo* ou *kernels*, mentres que  $h$  recibe o nome de *xanela* ou *parámetro de suavizado* (*bandwidth* en inglés). Unha escolla habitual da función núcleo é a normal enrolada  $WN(0, h^2)$ :

$$K_h(x) = \frac{1}{\sqrt{2\pi h^2}} \sum_{k=-\infty}^{+\infty} \exp\left(\frac{-(x + 2\pi k)^2}{2h^2}\right). \quad (3)$$

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

$$\mathcal{L}(h) = \prod_{i=1}^n \hat{f}_h(X_i), \quad (4)$$

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 directamente por eles.

Para poder formular o test de multimodalidade coa linguaxe dos tests de razón de verosimilitudes paramétricos temos traducir as hipóteses a contrastar definidas en (1) nunha división do espazo de parámetros  $\Theta = (0, +\infty)$ . Isto lograrémolo apoiándonos de novo na estimación de núcleo da densidade. Así, os valores de  $h$  que ofrezan unha estimación tipo núcleo con  $k$  modas como máximo serán os asociados a hipótese nula e os demais estarán asociados a hipótese alternativa. Esta división do espazo de parámetros pódese simplificar tendo en conta que o número de modas do estimador tipo núcleo  $\hat{f}_h$  é unha función decrecente en  $h$  sempre que empreguemos como núcleo a normal enrolada (véxase Huckemann et al., 2016). Isto permítenos definir o concepto de xanela crítica. A xanela crítica para  $k$  modas,  $h_k$ , non é máis que o menor valor do parámetro de suavizado para o cal a estimación tipo núcleo da densidade ten  $k$  modas como máximo. É dicir:

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

Así,  $h_k$  representa a fronteira entre as estimacións tipo núcleo da densidade con máis e menos de  $k$  modas: se o parámetro de suavizado  $h$  é menor que  $h_k$ , entón  $\hat{f}_h$  ten máis de  $k$  modas, e se  $h$  é maior que  $h_k$ , entón o número de modas de  $\hat{f}_h$  non é maior que  $k$ . Pero esta é precisamente a división entre a hipótese nula e a hipótese alternativa que definimos en (1). Por tanto, a xanela crítica facilítanos enormemente a división do espazo de parámetros  $\Theta$ :

$$\Theta_0 = \{h > 0: \hat{f}_h \text{ ten } k \text{ modas como máximo}\} = [h_k, +\infty);$$

$$\Theta_1 = \{h > 0: \hat{f}_h \text{ ten máis de } k \text{ modas}\} = (0, h_k).$$

Xa temos introducidas todas as ferramentas para poder construír o noso test de multimodalidade. O estatístico de contraste, será:

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

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. Resulta sinxelo ver que  $\lim_{h \rightarrow 0} \mathcal{L}(h) = +\infty$ , polo que  $\mathcal{L}$  non é unha función limitada e por tanto  $\sup_{h>0} \ell(h)$  non é un número real. Un xeito de solventar este problema é redefinir a función  $\mathcal{L}$  dada por (4) mediante validación cruzada. Para iso, imos apoiarnos nas funcións auxiliares:

$$\hat{f}_h^{-i}(x) = \frac{1}{n-1} \sum_{j=1, j \neq i}^n K_h(x - X_j).$$

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

$$\mathcal{L}_{CV}(h) = \prod_{i=1}^n \hat{f}_h^{-i}(X_i), \quad h > 0. \quad (5)$$

Con esta modificación logramos que a función  $\mathcal{L}_{CV}$  si que está limitada superiormente, tal como nos garante o seguinte resultado.

**Proposición 1.** *Sexa  $X_1, \dots, X_n$  unha mostra aleatoria simple dun ángulo aleatorio absolutamente continuo  $X$ . Sexa  $\mathcal{L}_{CV}$  a función de pseudo-verosimilitude por validación cruzada definida en (5), onde  $K_h$  é a normal enrolada definida en (3). Entón a función  $\mathcal{L}_{CV}(h)$  está limitada superiormente no intervalo  $(0, +\infty)$  con probabilidade 1.*

A proba deste resultado recóllese na Sección 7. Polo tanto, o estatístico

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

onde  $\ell_{CV}(h) = \log \mathcal{L}_{CV}(h)$ , está ben definido pola Proposición 1. Este será finalmente 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  $D_k$ .

Para o calibrado empregamos bootstrap suavizado, xerando remostras da densidade suavizada  $\hat{f}_{h_k}$ , onde  $k$  é precisamente o número de modas que queremos contrastar. A partir destas remostras calculamos réplicas do estatístico  $D_k$  e estimaremos o p-valor do test calculando a proporción de réplicas maiores que o valor do estatístico sobre a mostra orixinal.

Así, de xeito esquemático, o test de multimodalidade con nivel de significación  $\alpha \in (0, 1)$  para datos circulares é:

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

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

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:

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

2. Obtemos a remostra  $X_1^*, \dots, X_n^*$  da densidade suavizada  $\hat{f}_{h_k}$  e calculamos o valor do estatístico para a remostra, que denotamos por  $D_k^*$ .
3. Repetimos  $B$  veces o paso 2, conseguindo así  $B$  réplicas do estatístico:  $D_k^{*,1}, D_k^{*,2}, \dots, D_k^{*,B}$ .
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. O SELECTOR DE XANELA $h_{max}$

Tamén podemos empregar a función de pseudo-verosimilitude para seleccionar un parámetro de suavizado  $h$  á hora de construír o estimador da densidade  $\hat{f}_h$  que definimos en (2). Dada a mostra circular  $X_1, \dots, X_n$ , poderíase escoller como xanela o valor  $h_{max} > 0$  que verifique que

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

que existe por ser a función  $\mathcal{L}_{CV}$  limitada superiormente.

As principais propiedades deste selector da xanela foron estudadas por Hall et al. (1987), que probaron que  $h_{max}$  é un valor *ótimo* 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.

Ademais, as condicións necesarias por Hall et al. (1987) para a converxencia asintótica de  $\hat{f}_{h_{max}}$  adiántannos 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.

#### 4. ESTUDO DE SIMULACIÓN E RESULTADOS

Nesta sección estudaremos o comportamento na práctica da nova proposta de test de multimodalidade mediante dous estudos de simulación. No primeiro deles buscamos comprobar o calibrado do novo test á hora de contrastar a unimodalidade dos datos, e no segundo estudaremos a potencia do mesmo ante unha alternativa de bimodalidade.

En ambos os estudos de simulación xeráronse  $M = 1000$  mostras de tamaño  $n$  de diversas distribucións. A cada unha destas mostras aplicámoslle o novo test, vendo se rexeita ou non a unimodalidade dos datos para varios niveis de significación  $\alpha$ . Finalmente, calculamos a proporción de mostras rexeitadas do total. Imos realizar o anterior para dous tamaños de mostra distintos,  $n = 100$  e  $n = 500$ , tanto para estudar potencia como calibrado. Os niveis de significación considerados son os tres máis usuais:  $\alpha = 0.01$ ,  $\alpha = 0.05$  e  $\alpha = 0.1$ , e os p-valores aproxímanse mediante  $B = 500$  remostras bootstrap en todos os casos. Para datos comprobar o calibrado empréganse cinco distribucións unimodais distintas, mentres para o estudo de potencia considéranse catro distribucións bimodais diferentes. As distribucións unimodais consideradas foron:

- **Modelo 1 (M1):** unha distribución de von Mises:  $vM(0, 1)$ .
- **Modelo 2 (M2):** unha mixtura de von Mises:  $0.2 \cdot vM(2\pi/3, 3) + 0.6 \cdot vM(\pi, 1.4) + 0.2 \cdot vM(4\pi/3, 3)$ .
- **Modelo 3 (M3):** unha mixtura de von Mises:  $0.05 \cdot vM(2\pi/3, 7) + 0.9 \cdot vM(\pi, 1) + 0.05 \cdot vM(4\pi/3, 7)$ .
- **Modelo 4 (M4):** unha von Mises *sine-skewed*:  $ssvM(\pi, 1, -0.9)$ .
- **Modelo 5 (M5):** unha distribución beta modificada para que o soporte sexa o intervalo  $[\pi/2, 3\pi/2]$  e enrolada:  $\exp[i(\pi \cdot \text{Beta}(3, 2) + \pi/2)]$ .

Estes modelos buscan representar a unha gran variedade de situacións, empezando por casos simples (unha von Mises, Modelo 1), pasando por modas planas (Modelo 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).

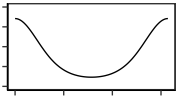
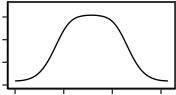
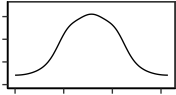
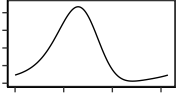
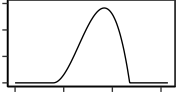
Por outro lado, as distribucións circulares con dúas modas escollidas para o estudo de potencia son:

- **Modelo 6 (M6):** unha mixtura de dúas von Mises:  $0.5 \cdot vM(2, 5) + 0.5 \cdot vM(4, 5)$ .
- **Modelo 7 (M7):** unha mixtura de von Mises:  $0.9 \cdot vM(\pi/2, 2) + 0.1 \cdot vM(3\pi/2, 3)$ .
- **Modelo 8 (M8):** unha mixtura de von Mises:  $0.5 \cdot vM(\pi - 1, 1.5) + 0.5 \cdot vM(\pi + 1, 1.5)$ .
- **Modelo 9 (M9):** unha mixtura de tres von Mises:  $0.3 \cdot vM(\pi/2, 6) + 0.5 \cdot vM(3\pi/4, 2) + 0.2 \cdot vM(7\pi/4, 4)$ .

Igual que cos modelos unimodais, estas distribucións buscan representar unha gran variedade de situacións: modas separadas (Modelo 6), modas xuntas (Modelo 8), unha moda secundaria moito menor que a principal (Modelos 7 e 9), e ata un modelo con certa asimetría (Modelo 9).

Para a realización do novo test de multimodalidade empregouse 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).

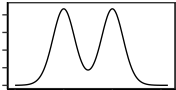
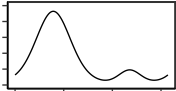
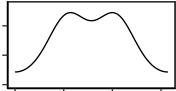
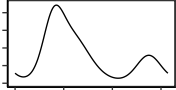
Táboa 1: Proporcións de rexeitamento para o test de multimodalidade baseado en pseudo-verosimilitude con nivel de significación ao 1%, 5% e 10% calculadas a partir de  $M = 1000$  mostras. Ao lado de cada proporción de rexeitamento, e entre paréntese, aparece a súa desviación típica estimada multiplicada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

Modelo	Densidade	Tamaño	Nivel de significación		
			1 %	5 %	10 %
M1		$n = 100$	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)
M2		$n = 100$	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)
M3		$n = 100$	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)
M4		$n = 100$	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)
M5		$n = 100$	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)

Comezamos comprobando o calibrado do test a hora de detectar unimodalidade. Na Táboa 1 aparecen os resultados deste estudo de simulación, e pódese observar que o novo test parece estar ben calibrado para tamaños de mostra grandes. O contraste só presenta proporcións de rexeitamento superiores ao nivel de significación para o Modelo 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 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 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 4 con  $\alpha = 0.1$ , onde a proporción de rexeitamento está lixeiramente por debaixo do nivel. Por último, para o Modelo 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.

Pasamos logo a estudar a potencia do test á hora de detectar a alternativa de bimodalidade. Os resultados deste segundo estudo de simulación aparecen recollidos na Táboa 2. O test parece detectar satisfactoriamente a alternativa en todos os escenarios considerados. Ademais, os resultados son coherentes co esperado tendo en conta a forma das distribucións empregadas. O test logra proporcións de rexeitamento altas para os modelos con modas claramente separadas, e ademais as proporcións de rexeitamento ván diminuíndo ao diminuír o tamaño da moda secundaria (por orde, de maior a menor, Modelo 6, Modelo 9 e Modelo 7). O test logra proporcións de rexeitamento moito máis pequenas para a distribución con modas pegadas (Modelo 8), pero sempre maiores que o nivel de significación. Por outra parte, as proporcións de rexeitamento crecen ao aumentar o tamaño da mostra en todos os casos.

Táboa 2: Proporções de rexeitamentos para o test baseado na 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.

Modelo	Densidade	Tamaño	Nivel de significación		
			1 %	5 %	10 %
M6		$n = 100$	0.911(0.018)	0.996(0.004)	0.997(0.003)
		$n = 500$	1.000(0.000)	1.000(0.000)	1.000(0.000)
M7		$n = 100$	0.295(0.028)	0.575(0.031)	0.714(0.028)
		$n = 500$	0.985(0.008)	0.995(0.004)	0.996(0.004)
M8		$n = 100$	0.033(0.011)	0.095(0.018)	0.131(0.021)
		$n = 500$	0.106(0.019)	0.256(0.027)	0.369(0.030)
M9		$n = 100$	0.373(0.030)	0.684(0.029)	0.820(0.024)
		$n = 500$	1.000(0.000)	1.000(0.000)	1.000(0.000)

## 5. APLICACIÓN A DATOS REAIS

Unha vez comprobado o bo calibrado do novo test de multimodalidade para datos circulares, podemos aplicalo aos datos de incendios de Ameijeiras-Alonso et al. (2018) presentados na introdución para ver se podemos afirmar que existe máis dunha moda neste caso. Calculamos o estatístico de contraste para a mostra dos incendios e 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 tempada de incendios ao longo do ano, algo que, como indicamos na introdución, podería verse explicado pola influencia da actividade humana no patrón natural de incendios en Galicia.

## 6. CONCLUSIÓNS

A principal conclusión deste traballo é clara. O novo test de multimodalidade para datos circulares baseado na pseudo-verosimilitude está ben calibrado e detecta satisfactoriamente a hipótese alterativa. Ademais, para a construción do mesmo só precisamos do estimador tipo núcleo da función de densidade  $\hat{f}_h$ . Este tipo de estimadores son facilmente adaptables a distintos tipos de espazos, como poden ser a esfera (véxase Mardia e Jupp, 2000, Cáp. 12), o cilindro ou o toro. Por tanto, o noso test de multimodalidade será extensible de forma directa a todos estes espazos onde o estimador tipo núcleo está ben definido. Iso si, a necesidade de que as funcións de densidade estén limitadas fóra de cero apunta a que o test só vaia ter un bo comportamento en variedades compactas (esfera, toro...), pois en variedades non compactas esta restricción non se pode verificar. A principal problemática de adaptar o test a estes espazos está 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.



## 7. PROBA DA PROPOSICIÓN 1

*Demostración da Proposición 1.* Sexa  $K_h$  a función de densidade dunha normal enrolada  $WN(0, h^2)$  definida en (3). Tendo en conta que  $\lim_{h \rightarrow +\infty} K_h(x) = 1/2\pi$  para todo  $x \in \mathbb{R}$ , temos que:

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

Imos ver un resultado similar pero cando  $h \rightarrow 0$ . Para iso temos que traballar coa definición da normal enrolada  $WN(0, h^2)$ :

$$K_h(x) = \frac{1}{\sqrt{2\pi h^2}} \sum_{k \in \mathbb{Z}} \exp\left(-\frac{(x - 2\pi k)^2}{2h^2}\right) \quad (6)$$

Imos estudar a serie en (6) separando os termos positivos dos negativos. Daquela, para os positivos temos que:

$$\begin{aligned} \sum_{k=0}^{+\infty} \exp\left(-\frac{(x + 2\pi k)^2}{2h^2}\right) &\leq \sum_{k=0}^{+\infty} \exp\left(-\frac{x^2}{2h^2} - \frac{2\pi^2 k^2}{h^2}\right) \leq \\ &\leq \exp\left(-\frac{x^2}{2h^2}\right) \sum_{k=0}^{+\infty} \exp\left(-\frac{2\pi^2 k^2}{h^2}\right) = \exp\left(-\frac{x^2}{2h^2}\right) \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1}. \end{aligned} \quad (7)$$

A primeira desigualdade de (7) dedúcese de que  $(a+b)^2 \geq a^2 + b^2$  se  $a, b \geq 0$ . A última desigualdade é inmediata, pois estámolles a sumar máis términos á serie (hai máis números naturais que cadrados perfectos).

Aplicando un razonamento similar para os termos negativos, temos que:

$$\sum_{k=1}^{+\infty} \exp\left(-\frac{(x - 2\pi k)^2}{2h^2}\right) \leq \exp\left(-\frac{(2\pi - x)^2}{2h^2}\right) \left[1 - \exp\left(-\frac{2\pi^2}{h^2}\right)\right]^{-1}.$$

De todo o anterior deducimos que:

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

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

Supoñamos agora que todos os valores  $X_1, \dots, X_n$  son todos distintos entre si. Entón, o anterior implica

$$\lim_{h \rightarrow 0} \hat{f}_h^{-i}(X_i) = \lim_{h \rightarrow 0} \frac{1}{(n-1)} \sum_{j=1, j \neq i}^n K_h(X_i - X_j) = 0,$$

para todo  $i \in \{1, \dots, n\}$  e por tanto  $\lim_{h \rightarrow 0} \mathcal{L}_{CV}(h) = 0$ . 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  $X_1, \dots, X_n$  sexan todos distintos entre si. Como  $X$  é un ángulo aleatorio absolutamente continuo, isto sucede con probabilidade 1.  $\square$

## AGRADECEMENTOS

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 presentados na introdución.

Este traballo foi realizado grazas ao apoio económico do Proxecto MTM2016-76969-P (INN-PAR2D) da Axencia Estatal de Investigación (AEI) cofinanciado polo Fondo Europeo de Desenvolvemento Rexional (ERDF).

---

**REFERENCIAS**

- Ameijeiras-Alonso, J., Crujeiras, R. M., e Rodríguez-Casal, A. (2018). Directional statistics for wildfires. En *Applied Directional Statistics*, páx. 203–226. Chapman and Hall/CRC.
- Hall, P., Watson, G., e Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, 74(4):751–762.
- Huckemann, S., Kim, K.-R., Munk, A., Rehfeldt, F., Sommerfeld, M., Weickert, J., Wollnik, C., et al. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli*, 22(4):2113–2142.
- Karlin, S. (1957). Pólya type distributions, II. *The Annals of Mathematical Statistics*, 28(2):281–308.
- Mardia, K. V. e Jupp, P. E. (2000). *Directional Statistics*. John Wiley & Sons.
- 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(694-706):289–337.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.