USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Distribucións circulares

Adrián Lago Balseiro

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Distribucións circulares

Adrián Lago Balseiro

Xullo, 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Distribucións circulares
Breve descrición do contido
Neste TFG proponse que o/a alumno/a realice unha revisión dos distintos modelos de distribucións circulares (modelo von Mises, cardiode, modelos enrolados, modelos proxectados, ...) e que, mediante estudos de simulación, explorase o funcionamento na práctica do método de máxima verosimilitude. Tamén se realizarán ilustracións con datos reais.

Índice xeral

Resumo

Este traballo consiste nunha introdución ás distribucións circulares do mesmo xeito que se faría nun curso de estatística. Primeiramente, definiranse os conceptos previos, como os momentos trigonométricos, a media ou a moda, que serán empregados posteriormente para estudar as principais distribucións circulares absolutamente continuas que existen, cuxas propiedades serán ilustradas. A maioría destas distribucións serán unimodais e simétricas, pero tamén se fará alusión a outras que non cumpran tales propiedades, facendo fincapé nas mesturas. A continuación describirase como se leva a cabo a estimación de parámetros dalgunhas destas distribucións mediante máxima verosimilitude e tamén coa axuda do algoritmo EM para as mesturas. Para comprobar empiricamente o funcionamento dos estimadores tratarase posteriormente a simulación de datos da von Mises e das súas mesturas. Finalmente, farase un estudo de datos reais, que serán modelados a partires das distribucións tratadas, seleccionando o mellor modelo mediante test de bondade de axuste e o criterio AIC.

Abstract

This work deals with an introduction to circular distributions the same way it would be developed in a statistics course. First of all, previous concepts, such as trigonometric moments, mean and mode, will be defined in order to be able to understand the distributions' propierties, which will be illustrated. Most of these distributions will be symmetric and unimodal, but some examples of other distributions will be shown, emphasizing the von Mises mixtures. After that, parameter estimation will be carried out by using the maximum maximum likelihood method with von Mises, wrapped normal and wrapped Cauchy and the von Mises mixtures, which will need the EM algorithm as well. Data simulation of von Mises distributions and von Mises mixtures will be used to check the empirical performance of the estimators and the deviation from the original values. Lastly, the previously

studied models will be fitted to real datasets, which will show their usefullness, selecting the distribution by goodness of fit tests and AIC criterion.

Introdución

As distribucións circulares xorden coa intención de estudar puntos no plano cartesiano dos cales só importe a dirección na que se atopan e non a distancia que teñan coa orixe, o que ocorre en moitas disciplinas científicas como a bioloxía, no estudo de movemento de animais; as ciencias ambientais, no estudo da dirección do vento; ou a física. Tamén é útil para estudar datos relacionados coas horas, como a chegada de pacientes á UCI ou horas de comezos de terremotos. Por suposto, isto pode xeneralizarse ao caso de esferas de dimensións superiores, en cuxo caso trataríase de datos direccionais, pero non será abordado neste traballo. No caso das distribucións circulares será importante determinar cal é a dirección de orixe e o sentido de rotación, o cal será necesario xa que os puntos na circunferencia non están ordenados, a diferencia dos puntos na recta real; isto supón unha gran diferencia entre as distribucións circulares e as definidas sobre $\mathbb{R}$. Haberá tamén outras diferencias significativas, como o sesgo ou o erro cadrático medio para o caso circular, que deberán ser definidos de xeito distinto ao caso das distribucións reais.

Unha importante precursora da estatística circular foi a enfermeira e matemática Florence Nightingale. De orixe británica, foi enviada á guerra como enfermeira polo seu país de orixe pola condición deste de aliado de Turquía ante Rusia. Entón creou o coñecido como gráfico de Florence Nightingale (ver a Figura 1), no que representaba os mortos na guerra no territorio que lle foi asignado divididos en tres grupos: mortos pola guerra, mortos por enfermidade e falecidos por outras causas. A idea desde gráfico é a mesma que se segue á hora de facer un diagrama de rosa, ou histograma circular: dividir unha circunferencia en ángulos iguais -neste caso 12, polos meses do ano- e dotar cada un dos sectores de maior ou menor altura en función da cantidade de mortos de cada tipo que houbese en cada mes.

O traballo dividirase en tres capítulos distintos. No primeiro deles, veranse os conceptos básicos, como a media, mediana ou os momentos trigonométricos, que serán empregados posteriormente neste mesmo capítulo na definición das distribucións circulares absolutamente continuas máis relevantes e no estudo das súas propiedades. A meirande parte destas distribucións cumprirán unha serie de propiedades comúns, como a simetría ou a unimo-

Figura 1: Florence Nightingale no panel da esquerda e gráfico de Florence Nightingale no panel da dereita. Imaxes baixo licenza Creative Commons.

dalidade, se ben se presentarán tamén algúns exemplos de distribucións que non posuirán tales características. No segundo capítulo abordarase a estimación de parámetros dalgunhas das distribucións mencionadas no capítulo anterior, dende un punto de vista teórico. Para a estimación empregarase o método de máxima verosimilitude que é, en esencia, o mesmo que no caso das distribucións definidas sobre a recta real; e tamén o algoritmo EM, que é de uso habitual cando existe un conxunto de datos que se descoñecen, pero a súa presenza facilitaría a resolución das ecuacións que dan lugar á estimación dos parámetros. No terceiro ocuparase a simulación de datos dunha distribución de von Mises e de mesturas de varias destas distribucións, co que se estudarán empíricamente as propiedades dos estimadores calculados no Capítulo 2 e tamén se estudará o sesgo, a varianza e o erro cadrático medio dos estimadores, para o que será preciso definilos no caso circular e cos que se obtén unha medida da precisión das estimacións. Para rematar o capítulo, modelaranse unha serie de datos reais mediante algunhas das distribucións descritas no primeiro capítulo. Para axustar as distribucións empregaranse test de bondade de axuste e o criterio AIC. Finalmente, inclúese un anexo con distribucións discretas para amosar que tamén existen no caso circular e poden ter certa relevancia, xunto cunha explicación teórica do test de bondade de axuste empregado no Capítulo 3.

A estrutura do traballo é similar á que se segue no grao en Matemáticas nas asignaturas

de estatística: primeiramente, defínense e estúdanse as medidas de localización, concentración e forma; a continuación defínense as distribucións máis importantes e estúdanse as súas propiedades vencelladas con estas medidas[1]; e, finalmente, lévase a cabo inferencia sobre os seus parámetros e estúdanse as propiedades dos mesmos dun xeito práctico[2].

Unha gran parte das distribucións que aparecen neste traballo teñen unha estreita relación coas definidas sobre a recta real ou sobre o plano: pola súa utilidade na práctica, a distribución de von Mises podería identifícarse coa distribución normal, de tal xeito que algunas referencias clásicas, como [11], refírense a ela como a normal circular, aínda que non sexa a súa denominación habitual. A distribución uniforme circular defínese do mesmo xeito que a uniforme sobre a recta real e, de feito, a función de densidade é a mesma variando o soporte. Ademais, pode obterse como caso particular de moitas das distribucións que se tratan no traballo. Pola súa parte, as distribucións proxectadas e as enroladas obtéñense a partir de distribucións sobre o plano ou a recta, respectivamente. É útil nalgúns casos a mestura de distribucións, ben de dúas distintas ou ben de varías do mesmo tipo pero con parámetros distintos para obter unha maior riqueza e un mellor axuste no modelado dun caso real. Con este obxectivo, introdúcense tamén as distribucións asimétricas e a Batschelet inversa, que permitirán facer fronte a situacións de asimetría.

Os principais documentos empregados para a realización do traballo son os libros [11] e [13] para a parte teórica xeral e [15] para a parte das mesturas de von Mises; e, finalmente, [16], para a parte práctica. Nesta empregarase o software `R` e, principalmente, os paquetes [2] e [14], que permiten, entre ambos, a representación gráfica de case todas as distribucións abordadas no traballo e tamén a estimación de parámetros da distribución de von Mises e das mesturas, coa axuda da libraría [10], que será fundamental no terceiro capítulo.

[1] A definición das medidas e das distribucións, así como as súas propiedades estúdanse nas asignaturas de Elementos de Probabilidade e Estatística do primeiro curso e en Probabilidade e Estatística do terceiro curso do Grao.

[2] A inferencia paramétrica abórdase na asignatura Inferencia Estatística do terceiro curso do Grao.

Capítulo 1

Introdución aos modelos de distribucións circulares

Neste primeiro capítulo lévase a cabo a introducción á estatística circular. Primeiramente, será preciso definir que é unha distribución circular e os conceptos de función de distribución e de densidade, esta última só para as distribucións absolutamente continuas, que serán as estudadas no traballo. Posteriormente, a partir da función característica, estudaranse os momentos trigonométricos que darán pé á definición de medidas de localización, concentración e forma no caso poboacional e, a partir deles, os seus análogos mostrais. Finalmente, definiranse as principais distribucións circulares e citaranse as súas propiedades máis relevantes, ilustrándoas na medida do posible. Estas serán, na maioría dos casos, a simetría e a unimodalidade, se ben introduciranse modificacións destas funcións e tamén mesturas delas, que permitirán obter funcións de densidade asimétricas e distribucións multimodais.

1.1. Funcións de distribución e de densidade

Unha distribución circular é aquela que acumula a probabilidade total na circunferencia unidade. Consideraranse, salvo que se diga o contrario, ángulos en radiáns e pertencentes ao intervalo $[0, 2\pi)$, de tal xeito que o ángulo θ será o mesmo que $(\theta + 2\pi)$. Dada unha variable aleatoria Θ, a súa función de distribución defínese, en analoxía coas distribucións sobre a recta real, do seguinte xeito:

$$F(\theta) = P(0 \leq \Theta \leq \theta), \quad \theta \in [0, 2\pi),$$

e satisfará unha condición adicional que non se cumpre nas distribucións reais como é

$$F(\theta + 2\pi) - F(\theta) = 1, \quad \forall\, \theta \in [0, 2\pi),$$

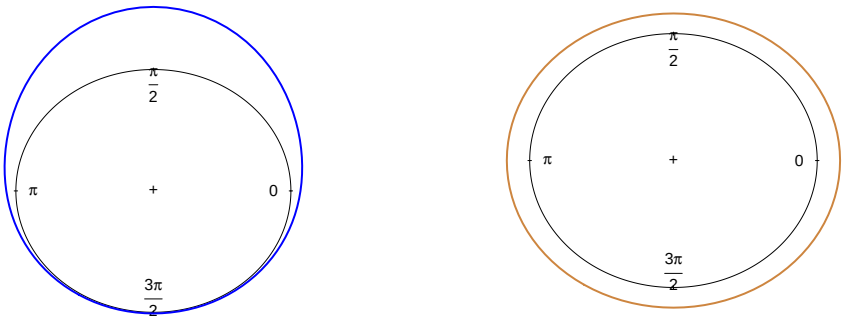


Figura 1.1: Representación de dous exemplos de distribucións circulares: a von Mises na esquerda e a circular uniforme na dereita.

co que se busca que, fixado calquera ángulo, dar un ciclo completo á circunferencia, tomando coma orixe ese propio ángulo, teña asociada unha probabilidade de 1. Outras propiedades que teñen estas funcións de distribución son:

$$\lim_{\theta\to\infty} F(\theta) = \infty \quad \text{e} \quad \lim_{\theta\to-\infty} F(\theta) = -\infty.$$

A función de distribución é, coma no caso real, continua pola dereita e, por definición, é claro que $F(0) = 0$ e $F(2\pi) = 1$. Por último, sexan $a \leq b < a + 2\pi$, tense que:

$$P(a < \Theta \leq b) = F(b) - F(a) = \int_a^b dF(\theta),$$

onde a integral é a de Lebesgue-Stieltjes. Nótese que a función de distribución depende do punto que sexa considerado como a orixe e tamén da dirección na que se pecorra a circunferencia, xa que non terá o mesmo valor a funcion de distribución, por exemplo, no cero se este é considerado coma orixe ou se esta é o ángulo $\theta = \pi$.

Consideraranse só distribucións absolutamente continuas, en cuxo caso poderase definir a función de densidade asociada á distribución, f, como aquela que cumpra as tres condicións seguintes:

1. $f(\theta) \geq 0, \quad \forall\, \theta \in [0, 2\pi)$,
2. $\int_0^{2\pi} f(\theta)d\theta = 1$ e
3. $f(\theta + 2k\pi) = f(\theta), \quad \forall\, k \in \mathbb{Z}$.

A condición de non negatividade tamén a satisfán as funcións de densidade sobre $\mathbb{R}^n$; a segunda é análoga ao caso real, posto que se require que a probabilidade total sexa 1; e na derradeira condición está a esixírselle á función de densidade que sexa $2\pi-$periódica.

1.2. Función característica e momentos trigonométricos

Do mesmo xeito que ocorre coas variables aleatorias sobre a recta real, defínese a función característica da variable aleatoria Θ do seguinte xeito:

$$\varphi_\Theta(t) = \mathbb{E}(e^{t\Theta i}), \tag{1.1}$$

sendo i a unidade imaxinaria e onde $\mathbb{E}$ representa a esperanza matemática. Ademais, en virtude da $2\pi-$periodicidade das variables aleatorias circulares, tense que:

$$\varphi_\Theta(t) = \mathbb{E}(e^{t\Theta i}) = \mathbb{E}(e^{t(\Theta+2\pi)i}) = e^{2\pi ti}\mathbb{E}(e^{t\Theta i}). \tag{1.2}$$

Das igualdades en (1.2), dedúcese que $\mathbb{E}(e^{t\Theta i}) = 0$ ou ben $e^{2\pi ti} = 1$. En virtude da igualdade $e^{2\pi ti} = \cos(2\pi t) + i\,\mathrm{sen}(2\pi t)$, necesariamente $t \in \mathbb{Z}$, polo que nas variables aleatorias circulares, a función característica estará definida só nos números enteiros (ver [11]). Así, defínese o momento trigonométrico p-ésimo da variable aleatoria Θ como o valor $\varphi_\Theta(p)$, e denotarase por φ_p. Como, ademais, só se consideran variables aleatorias continuas, (1.1) pode expresarse como:

$$\varphi_p = \mathbb{E}(e^{p\Theta i}) = \int_0^{2\pi} e^{p\theta i} dF(\theta) = \rho_p e^{\mu_p i}, \quad p \in \mathbb{Z}. \tag{1.3}$$

Nótese que foi empregada a fórmula módulo-exponencial dun número complexo, polo que ρ_p representa o módulo e μ_p o argumento de φ_p. En canto ao primeiro tense que:

$$\rho_p = |\varphi_p| = ||\mathbb{E}(e^{p\Theta i})|| \leq \mathbb{E}(||e^{p\Theta i}||) = \mathbb{E}(1) = 1. \tag{1.4}$$

Pola segunda propiedade da definición da función de densidade, $\varphi_0 = 1$. Así, como φ_p, $p \in \mathbb{Z}$, é un número complexo, poden definirse a súa parte real e imaxinaria, que se denotarán por $\alpha_p = \mathrm{Re}(\mathbb{E}(e^{p\Theta i})) = \mathbb{E}(\cos p\Theta)$ e $\beta_p = \mathrm{Im}(\mathbb{E}(e^{p\Theta i})) = \mathbb{E}(\mathrm{sen}\, p\Theta)$, $p \in \mathbb{Z}$, de tal xeito que $\varphi_p = \alpha_p + i\beta_p$. Tendo en conta que as funcións coseno e seno son par e impar respectivamente, tense que $\alpha_{-p} = \alpha_p$ e $\beta_{-p} = -\beta_p$, polo que $\bar{\varphi}_p = \varphi_{-p}$, onde $\bar{\varphi}$ representa o conxugado do número complexo φ. En virtude da definición do módulo e do argumento dun número complexo, tense, $\forall\, p \in \mathbb{Z}$, que:

$$\rho_p = \sqrt{\alpha_p^2 + \beta_p^2},$$

e

$$\mu_p = \operatorname{atan2}(\beta_p, \alpha_p) = \begin{cases} \arctan\left(\frac{\beta_p}{\alpha_p}\right), & \text{se} \quad \alpha_p > 0, \\ \arctan\left(\frac{\beta_p}{\alpha_p}\right) + \pi, & \text{se} \quad \alpha_p < 0, \beta_p \geq 0, \\ \arctan\left(\frac{\beta_p}{\alpha_p}\right) - \pi, & \text{se} \quad \alpha_p < 0, \beta_p < 0, \\ \frac{\pi}{2}, & \text{se} \quad \alpha_p = 0, \beta_p > 0, \\ \frac{-\pi}{2}, & \text{se} \quad \alpha_p = 0, \beta_p < 0, \\ \text{Indeterminado} & \text{se} \quad \alpha_p = 0 = \beta_p. \end{cases} \tag{1.5}$$

Nótese que esta función devolve valores no intervalo $(-\pi, \pi]$.

Dada unha dirección μ, defínese o momento trigonométrico p-ésimo da variable aleatoria Θ centrado con respecto de μ como:

$$\tilde{\varphi}_p = \mathbb{E}[e^{ip(\Theta-\mu)}] = \tilde{\alpha}_p + i\tilde{\beta}_p,$$

onde $\tilde{\alpha}_p = \mathbb{E}\left[\cos(\rho(\Theta - \mu))\right]$ e $\tilde{\beta}_p = \mathbb{E}\left[\operatorname{sen}(\rho(\Theta - \mu))\right]$. Dise que unha variable aleatoria circular, Θ, é simétrica con respecto dunha dirección μ se $f(\mu - \theta) = f(\mu + \theta), \quad \forall\ \theta \in [0, 2\pi)$, onde f é a función de densidade. Por exemplo, se a variable é simétrica respecto da dirección $\mu = 0$, a distribución sería a mesma que aquela obtida tras a transformación dos datos ao cambiar de signo aos mesmos. É salientable o feito de que se unha variable aleatoria é simétrica respecto dunha dirección μ, entón, a parte imaxinaria dos momentos trigonométricos é cero (ver [11], Capítulo 2), é dicir, $\tilde{\beta}_p = 0, \forall\ p \in \mathbb{Z}$ (de xeito análogo ao que ocorre coas variables aleatorias reais).

Por último, os números reais α_p e β_p son os coeficientes de Fourier da serie asociada á función f (ver [13], Capitulo 3), de tal xeito que

$$dF(\theta) \approx \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \varphi_k e^{-k\theta i}.$$

Non obstante, a expresión anterior non garante que a serie sexa converxente nin, moito menos, que converxa á f. Pero se $\sum_{k=1}^{\infty} \alpha_k^2 + \beta_k^2 = \sum_{k=-\infty}^{\infty} |\varphi_k(\theta)|^2 < \infty$ e, como

$$\int_0^{2\pi} f^2(\theta) d\theta < \infty \Leftrightarrow \sum_{k=-\infty}^{\infty} |\varphi_k(\theta)|^2 < \infty,$$

(ver [13], Sección 4.2) entón, a variable Θ ten por función de densidade a f dada por:

$$f(\theta) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \varphi_k e^{-k\theta i} = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{k=1}^{\infty} (\alpha_k \cos(k\theta) + \beta_k \operatorname{sen}(k\theta)) \right\}, \tag{1.6}$$

sendo $f(\theta)$ converxente en L^2. Por último, a función de densidade dunha variable aleatoria circular simétrica admite, en virtude de (1.6) a seguinte serie de Fourier

$$f(\theta) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{k=1}^{\infty} \alpha_k \cos(k\theta) \right\}.$$

1.2.1. Momentos trigonométricos mostrais

Consideremos agora unha mostra de valores $\theta_1, \ldots, \theta_n$. En analoxía co caso poboacional, defínense

$$a_p = \frac{1}{n} \sum_{k=1}^{n} \cos(p\theta_k) \quad \text{e} \quad b_p = \frac{1}{n} \sum_{k=1}^{n} \operatorname{sen}(p\theta_k). \tag{1.7}$$

Deste xeito, defínese o momento trigonométrico mostral $p-$ésimo como

$$m_p = a_p + ib_p, \quad \forall\, p \in \mathbb{Z}.$$

O módulo deste momento trigonométrico mostral vén dado por

$$R_p = \sqrt{a_p^2 + b_p^2}$$

e denotarase por θ_p ao argumento de m_p, é dicir,

$$\theta_p = \operatorname{atan2}(b_p, a_p),$$

onde a función atan2 foi definida en (1.5). O que resultará de maior utilidade será o primeiro momento trigonométrico mostral, que se denotará por $\mathbf{R} = (a_1, b_1)$ e recibe o nome de vector resultante da suma das compoñentes mostrais, ou vector resultante. Denotarase por R ao módulo do vector $\mathbf{R}$, é dicir,

$$R = ||\mathbf{R}|| = \sqrt{a_1^2 + b_1^2}. \tag{1.8}$$

O argumento do número complexo m_1 será a media mostral no caso de que R sexa estritamente positivo e denotarase por $\bar{\theta}$, de tal xeito que:

$$\bar{\theta} = \arg(a_1 + ib_1), \tag{1.9}$$

é dicir, $\bar{\theta} \equiv \theta_1$. Noutro caso, dirase que non existe tal dirección media. A partir de (1.8) e (1.9) obtense que

$$\cos(\bar{\theta}) = \frac{a_1}{R}, \quad \operatorname{sen}(\bar{\theta}) = \frac{b_1}{R}. \tag{1.10}$$

De xeito análogo á media poboacional, o ángulo $\bar{\theta}$ toma os seguintes valores:

$$\bar{\theta} = \operatorname{atan2}(b_p, a_p) = \begin{cases} \arctan\left(\frac{b_p}{a_p}\right), & \text{se} \quad a_p > 0, \\ \arctan\left(\frac{b_p}{a_p}\right) + \pi, & \text{se} \quad a_p < 0, b_p \geq 0, \\ \arctan\left(\frac{b_p}{a_p}\right) - \pi, & \text{se} \quad a_p < 0, b_p < 0, \\ \frac{\pi}{2}, & \text{se} \quad a_p = 0, b_p > 0, \\ \frac{-\pi}{2}, & \text{se} \quad a_p = 0, b_p < 0, \\ \text{Indeterminado} & \text{se} \quad a_p = 0 = b_p. \end{cases} \tag{1.11}$$

En virtude de (1.10) é claro que $\sum_{k=1}^{n} \operatorname{sen}(\theta_k - \bar{\theta}) = 0$ e, consecuentemente, $\frac{1}{n}\sum_{k=1}^{n} \cos(\theta_k - \bar{\theta}) = R$. É importante tamén o feito de que se todos os datos son rotados un certo ángulo, o valor da dirección media variará nesa mesma cantidade. En efecto, sexa $\theta_1, \theta_2, \ldots, \theta_n$ unha mostra aleatoria con dirección media $\bar{\theta}$. Sexa $c \in [0, 2\pi)$, consideremos os valores $\theta_1 + c, \theta_2 + c, \ldots, \theta_n + c$ e vexamos que a dirección media dos valores anteriores é $\bar{\theta} + c$. Para iso, consideremos o par $\mathbf{R}' = \left(\frac{1}{n}\sum_{k=1}^{n} \cos(\theta_k + c), \frac{1}{n}\sum_{k=1}^{n} \operatorname{sen}(\theta_k + c)\right) = (a_1', b_1')$ no que se emprega a notación introducida anteriormente. Asi, a partir das expresións en (1.10), tense que

$$\begin{aligned} a_1' &= \frac{1}{n}\sum_{k=1}^{n} \cos(\theta_k + c) = \frac{1}{n}\sum_{k=1}^{n} [\cos(\theta_k)\cos(c) - \operatorname{sen}(\theta_k)\operatorname{sen}(c)] = a_1 \cos(c) - b_1 \operatorname{sen}(c) \\ &= R\cos(\bar{\theta})\cos(c) - R\operatorname{sen}(\bar{\theta})\operatorname{sen}(c) = R\cos(\bar{\theta} + c) \end{aligned}$$

e

$$\begin{aligned} b_1' &= \frac{1}{n}\sum_{k=1}^{n} \operatorname{sen}(\theta_k + c) = \frac{1}{n}\sum_{k=1}^{n} [\operatorname{sen}(\theta_k)\cos(c) + \cos(\theta_k)\operatorname{sen}(c)] = b_1 \cos(c) + a_1 \operatorname{sen}(c) \\ &= R\operatorname{sen}(\bar{\theta})\cos(c) + R\cos(\theta)\operatorname{sen}(c) = R\operatorname{sen}(\bar{\theta} + c). \end{aligned}$$

Deste xeito, $a_1'^2 + b_1'^2 = R^2\cos^2(\bar{\theta} + c) + R^2\operatorname{sen}^2(\bar{\theta} + c) = R^2$; polo que

$$R' = ||\mathbf{R}'|| = \sqrt{a_1'^2 + b_1'^2} = R.$$

Como consecuencia,

$$\frac{a_1'}{R'} = \frac{R\cos(\bar{\theta}+c)}{R} = \cos(\bar{\theta}+c), \quad \text{e} \quad \frac{b_1'}{R'} = \frac{R\,\text{sen}(\bar{\theta}+c)}{R} = \text{sen}(\bar{\theta}+c),$$

co que se conclúe que a media mostral circular é equivariante respecto á rotación dos datos un certo ángulo.

Por último, o momento trigonométrico poboacional $p-$ésimo centrado na media, $\bar{\theta}$, é

$$\tilde{m}_p = \tilde{a}_p + i\tilde{b}_p, \quad \forall\, p \in \mathbb{Z},$$

onde

$$\tilde{a}_p = \frac{1}{n}\sum_{k=1}^{n}\cos(p(\theta_k - \bar{\theta})) \quad \text{e} \quad \tilde{b}_p = \frac{1}{n}\sum_{k=1}^{n}\cos(p(\theta_k - \bar{\theta})).$$

1.3. Medidas de localización, concentración e forma

Na definición (1.3) é salientable o caso $p = 1$, xa que o valor $\mu_1 \equiv \mu$ correspóndese coa media teórica da distribución e esta será a principal medida de localización poboacional. No caso mostral, e como xa se definiu, a media será $\bar{\theta}$ no caso de que $R > 0$.

Outra medida é a mediana da variable aleatoria Θ, que se representará por $\tilde{\mu}$ e defínese no caso poboacional como:

$$\tilde{\mu} = \min_{\phi\in[0,2\pi)} \mathbb{E}\left[\pi - |\pi - |\phi - \Theta||\right].$$

No caso mostral, será un ángulo $\tilde{\theta}$ tal que a metade dos datos estean en $[\tilde{\theta}, \tilde{\theta}+2\pi]$ e a maioría estean máis próximos a $\tilde{\theta}$. É dicir, se hai una cantidade impar de datos, será o ángulo que deixa a metade dos datos 'á dereita' e a outra metade 'á esquerda', estando estes ordeados. Se é impar, será a media dos dous datos centrais da mostra ordeada.

A derradeira medida de localización será a moda, que é a dirección que fai máxima a función de densidade da distribución e o dato con maior frecuencia no caso mostral.

A principal medida de concentración no caso poboacional será $\rho_1 \equiv \rho$, obtido de (1.3) tomando $p = 1$. A maior valor de ρ, a distribución estará máis próxima ao ángulo μ e, pola contra, canto máis próximo a cero estea este valor, a distribución estará esparexida máis uniformemente. Hai dous casos dexenerados: $\rho = 0$ e $\rho = 1$. No primeiro, a distribución considerada sería uniforme na circunferencia e, no segundo, toda a probabilidade estaría vencellada a unha única dirección, μ. Lembremos que, por (1.4), $\rho \in [0, 1]$. O seu correspondente mostral será $||\mathbf{R}|| = R \in [0, 1]$, que analogamente, canto máis próximo a un sexa, máis concentrados estarán os datos ao redor da media, neste caso, mostral, $\bar{\theta}$.

Defínese a varianza circular poboacional como

$$\nu = 1 - \rho$$

e, do feito de que $\rho \in [0, 1]$, dedúcese que $\nu \in [0, 1]$. Cando $\rho = 1$, é dicir, a distribución está totalmente concentrada nun punto, $\nu = 0$. Tense que $\rho = 0$ para a distribución circular uniforme, que se verá na Sección 1.4.2, pero tamén se cumpre para calquera distribución ciclicamente simétrica, polo que de $\nu = 1$ non se pode deducir que a distribución estea uniformemente distribuída ao redor da circunferencia.

O análogo mostral de ν recibirá o nome de varianza circular mostral e defínese como

$$V = 1 - R,$$

onde R fai referencia á lonxitude resultante media definida en (1.8) e, xa que $R \in [0, 1]$, entón, $V \in [0, 1]$.

Como medida de dispersión destaca a desviación típica circular poblacional, que se define como

$$\sigma = \sqrt{-2\log(1-\nu)} = \sqrt{-2\log\rho}$$

O análogo mostral, a desviación típica circular mostral, vén dada por:

$$\widehat{\sigma} = \sqrt{-2\log(1-V)} = \sqrt{-2\log(R)}.$$

É importante precisar que, como $R \in [0, 1]$, entón $\log(R) \in (-\infty, 0]$ e, polo tanto, $\widehat{\sigma}$ está ben definido. Ademais, se R está moi próximo a un, é dicir, se os datos están moi concentrados ao redor da dirección media,

$$\widehat{\sigma} \approx (2V)^{1/2} = (2 - 2R)^{1/2}.$$

Ata agora, resultou de grande importancia só o momento de orde un, a partir do cal podiamos obter a media ou a concentración tanto poboacional como mostral. Pois ben, agora centrarémonos no momento de orde dous, que dará información sobre a asimetría e a curtose dos datos. Máis precisamente, a parte imaxinaria do momento mostral de orde dous centrado na media, $\tilde{b}_2$, que é a primeira parte imaxinaria dun momento mostral centrado na media non nulo e dá una medida da asimetría dos datos ao redor de tal media. Se a distribución é unimodal, $\tilde{b}_2$ estará próximo a cero en caso de que os datos estean distribuídos de xeito simétrico ao redor de $\bar{\theta}$; de valor grande no caso de que a mostra sexa asimétrica en sentido horario; e negativamente grande se a mostra é asimétrica en sentido antihorario (ver [16]). A cantidade estandarizada

$$\widehat{s} = \frac{\tilde{b}_2}{(1-R)^{3/2}},$$

ten a mesma interpretación que $\tilde{b}_2$, pero tomará valores maiores xa que $R \in (0, 1]$.

Por outra banda, na distribución normal enrolada, que se estuda na Sección 1.4.5, cúmprese que $\tilde{\alpha}_2 - \rho^4 = 0$, ver [13]. Isto leva a tomar o análogo mostral do valor anterior, é dicir, $\tilde{a}_2 - R^4$, como unha medida de curtose dos datos. Este será maior que cero se os datos teñen un apuntamento maior que o propio dunha normal enrolada. A cantidade estandarizada correspondente é

$$\widehat{k} = \frac{\tilde{a}_2 - R^4}{(1-R)^2}.$$

Os análogos poboacionais das medidas anteriores son $\tilde{\beta}_2$ para a asimetría e $\tilde{\alpha_2} - \rho^4$ para a curtose, ás que lles corresponden as cantidades estandarizadas:

$$s = \frac{\tilde{\beta}_2}{(1-\rho)^{3/2}} \quad \text{e} \quad k = \frac{\tilde{\alpha_2} - \rho^4}{(1-\rho)^2}.$$

1.4. Obtención de distribucións circulares

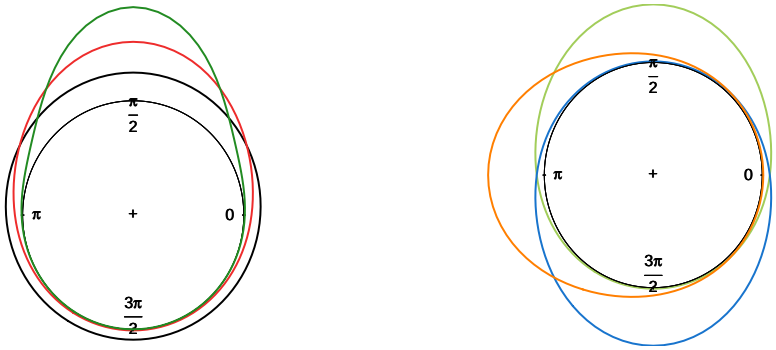
A continuación, presentaranse exemplos e características de distribucións circulares obtidas a partir doutras definidas sobre a recta real ou sobre o plano $\mathbb{R}^2$; outras serán xeralizacións a partir das características dalgunhas distribucións sobre a recta real; e mesmo algunha, como é o caso da cardioide, obtense a partir dunha curva. Ademais, co obxectivo de aumentar a riqueza das distribucións, estudaranse as modificacións e as mesturas das distribucións anteriores.

1.4.1. Distribución von Mises

Unha das máis importantes distribucións na recta real, por non decir a que máis, é a normal. Esta axuda a modelar unha gran cantidade de fenónemos ou, cando menos, aproxímase moito a facelo. Sería desexable que no caso circular atopásemos unha distribución que sexa moi empregada en moitos problemas aplicados. En 1918, von Mises introduciu un modelo que se coñece actualmente como distribución de von Mises -en honra á persoa que a empregou por primeira vez-. Defínese esta distribución como aquela que ten por función de densidade:

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \quad \theta, \mu \in [0, 2\pi), \ \kappa \geq 0, \tag{1.12}$$

onde μ e κ son parámetros. O primeiro deles fará referencia á media da distribución e o segundo empregarase como medida da concentración da mesma, de xeito análogo ao que ocorre con ρ. Así pois, denotarase esta distribución como $vM(\mu, \kappa)$. É oportuno detallar que $I_0(\kappa)$ refírese á función de Bessel modificada de primeiro tipo e de orde cero e é empregada



(a) $\mu = \frac{\pi}{2}$, $\kappa = 0.5$ en negro, $\kappa = 2$ en vermello;(b) $\kappa = 2$, $\mu = \frac{\pi}{2}$ na liña verde; π en laranxa; e e $\kappa = 4.5$ en verde. $\mu = \frac{3\pi}{2}$ en azul.

Figura 1.2: Exemplos da densidade da von Mises.

para normalizar a distribución. O valor $I_0(\kappa)$ calcúlase numericamente de acordo á seguinte expresión

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta}d\theta = \sum_{r=0}^{\infty}\left(\frac{\kappa}{2}\right)^{2r}\left(\frac{1}{r!}\right)^2.$$

A partir do desenvolvemento en serie de Fourier da función de densidade, dado por

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)}\left\{I_0(\kappa) + 2\sum_{k=1}^{\infty} I_k(\kappa)\cos[k(\theta-\mu)]\right\},$$

pode obterse a función de distribución mediante integración de xeito inmediato

$$F(\theta) = \frac{1}{2\pi I_0(\kappa)}\left\{I_0(\kappa)\theta + 2\sum_{k=1}^{\infty}\frac{I_k(\kappa)\,\mathrm{sen}[k(\theta-\mu)]}{k}\right\}, \quad \theta\in[0,2\pi),$$

onde $I_p(\kappa)$ fai referencia á funcion de Bessel modificada de primeiro tipo e de orde p. Lístanse agora unha serie de propiedades da distribución.

A densidade é unimodal e alcanza o seu valor máximo en $\theta = \mu$. Para comprobar isto é suficiente maximizar a función de densidade, f. Plantéxase entón a condición de punto crítico para f, que é $f'(\theta) = 0$:

$$f'(\theta) = \frac{-\kappa}{2\pi I_0(\kappa)}\,\mathrm{sen}(\theta-\mu)e^{\kappa\cos(\theta-\mu)} = 0.$$

Entón, ou ben $\frac{-\kappa}{2\pi I_0(\kappa)} = 0$ ou $\operatorname{sen}(\theta-\mu) = 0$. O primeiro caso levaría a que a distribución é a uniforme, que será estudada na Sección 1.4.2. Do segundo dedúcese que $\theta-\mu = k\pi, \quad k \in \mathbb{Z}$, pero como $\mu, \theta \in [0, 2\pi)$, necesariamente $\theta - \mu \in \{-\pi, 0, \pi\}$. Vexamos agora que $\mu = \theta$ é un máximo. A derivada segunda da función de densidade é

$$f''(\theta) = \frac{-\kappa e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)}\big[\cos(\theta-\mu) - \kappa\operatorname{sen}^2(\theta-\mu)\big],$$

polo tanto, $f''(\mu) = \frac{-\kappa e^{\kappa}}{2\pi I_0(\kappa)} < 0$, posto que $\kappa \geq 0$ e rexeitouse o caso $\kappa = 0$. Logo, μ dá un valor máximo á función de densidade. Máis concretamente

$$f(\mu) = \frac{e^{\kappa}}{2\pi I_0(\kappa)}. \tag{1.13}$$

Antimoda en $\mu \pm \pi$. Obtidos os puntos críticos da función de densidade e estudado o caso $\theta = \mu$, vexamos que ocorre cos outros dous valores. Neste caso, trátase de $\theta - \mu = \pi$ e $\theta - \mu = -\pi$. Avaliando $\theta = \mu + \pi$ e $\theta = \mu - \pi$ -que representan, en realidade, o mesmo ángulo- en f'', obtense o seguinte:

$$f''(\mu+\pi) = f''(\mu-\pi) = \frac{\kappa e^{-\kappa}}{2\pi I_0(\kappa)} > 0.$$

Logo, ambos dan valores de mínimo para f e o valor asociado é

$$f(\mu \pm \pi) = \frac{e^{-\kappa}}{2\pi I_0(\kappa)}. \tag{1.14}$$

O parámetro κ é un parámetro de concentración, é dicir, determina canto de próxima está o total da probabilidade con respecto da media μ. Isto obtense a partir do cociente de (1.13) e (1.14):

$$\frac{f(\mu)}{f(\mu \pm \pi)} = e^{2\kappa}.$$

Polo tanto, a maior valor do parámetro κ, maior será o cociente anterior, e maior será a diferencia dos valores entre o máximo e o mínimo e, polo tanto, maior será a concentración ao redor da media, μ. No panel da esquerda da Figura 1.2 pode apreciarse como a medida que aumenta o valor deste parámetro κ; a densidade concéntrase máis a redor do valor da media, que foi fixada no ángulo $\mu = \frac{\pi}{2}$, é dicir, ten unha función de atracción cara a media.

A densidade é simétrica respecto da dirección μ. Basta comprobar que $f(\mu-\theta) = f(\mu+\theta)$, $\forall\, \theta \in [0, 2\pi)$. En efecto, pola paridade da función coseno, tense que:

$$f(\mu-\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(-\theta)} = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos\theta} = f(\mu+\theta), \quad \forall\, \theta \in [0, 2\pi).$$

No panel da dereita da Figura 1.2 pode apreciarse como, efectivamente, a moda da distribución coincide precisamente co valor de μ en cada caso e a antimoda xustamente no

punto diametralmente oposto, é dicir, $\mu \pm \pi$. Por último, ilústrase na figura a simetría da densidade a partir da simetría da figura respecto da dirección μ.

Vexamos agora como se obteñen os momentos trigonométricos. Primeiramente, é preciso aclarar que

$$I_p(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \cos(p\theta)e^{\kappa\cos\theta}d\theta \quad \text{e} \quad \int_0^{2\pi} \operatorname{sen}(p\theta)e^{\kappa\cos\theta}d\theta = 0, \quad \forall\, p \in \mathbb{Z}. \tag{1.15}$$

Entón, aplicando a definición de momento de orde $p \in \mathbb{Z}$, tense o seguinte:

$$\begin{aligned}
\varphi_p &= \frac{1}{2\pi I_0(\kappa)}\int_0^{2\pi} e^{p\alpha i}e^{\kappa\cos(\alpha-\mu)}d\alpha = \frac{1}{2\pi I_0(\kappa)}\int_0^{2\pi} e^{p(\theta+\mu)i}e^{\kappa\cos\theta}d\theta \\
&= \frac{e^{p\mu i}}{2\pi I_0(\kappa)}\int_0^{2\pi} [\cos(p\theta) + i\operatorname{sen}(p\theta)]e^{\kappa\cos\theta}d\theta \\
&= \frac{e^{p\mu i}}{2\pi I_0(\kappa)}\int_0^{2\pi} \cos(p\theta)e^{\kappa\cos\theta}d\theta + \frac{ie^{\rho\mu i}}{2\pi I_0(\kappa)}\int_0^{2\pi} \operatorname{sen}(p\theta)e^{\kappa\cos\theta}d\theta \\
&= \frac{I_p(\kappa)e^{p\mu i}}{I_0(\kappa)}.
\end{aligned}$$

Na segunda igualdade empregouse o cambio de variable $\alpha = \theta + \mu$; na terceira, o feito de que $e^{i\beta} = \cos\beta + i\operatorname{sen}\beta, \forall\, \beta \in \mathbb{R}$; e na quinta, as aclaracións (1.15). Así, tendo en conta que $||e^{i\beta}|| = 1, \forall\, \beta \in \mathbb{R}$, entón $||\varphi_p|| = \frac{I_p(\kappa)}{I_0(\kappa)}$. En particular, o módulo do primeiro momento trigonométrico é $||\varphi_1|| = \frac{I_1(\kappa)}{I_0(\kappa)} \equiv A(\kappa)$ e depende de κ, que é o parámetro de concentración.

Proposición 1.1. *Cando o parámetro de concentración tende a infinito, a distribución pode aproximarse por unha distribución normal sobre a recta real. É dicir,* $\beta = \sqrt{\kappa}(\theta-\mu) \xrightarrow{d} N(0,1)$*, se* $\kappa \to \infty$*, onde* $\xrightarrow{d}$ *representa a converxencia en distribución.*

Demostración. É preciso citar primeiro unha aproximación que será empregada na demostración e é a seguinte

$$I_0(\kappa) \approx \frac{e^\kappa}{\sqrt{2\pi\kappa}}. \tag{1.16}$$

Consideremos a función de densidade da distribución $vM(\mu,\kappa)$, que ten a seguinte expresión $f(\theta) = \frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)}$ e o polinomio de Taylor de orde 1 aproximando á función $\cos(\theta)$, $\cos(\theta) \approx 1 - \frac{\theta^2}{2}$. Sexa agora $\beta = \sqrt{\kappa}(\theta-\mu)$ dada no enunciado da proposición. Para valores pequenos de $\frac{\beta}{\sqrt{\kappa}}$ tense que, $\cos(\theta-\mu) = \cos\frac{\beta}{\sqrt{\kappa}} \simeq 1 - \frac{\beta^2}{2\kappa}$. Logo, facendo o cambio de variable $\theta - \mu = \frac{\beta}{\sqrt{\kappa}}$ na función de densidade da von Mises, tense que:

$$f(\beta) = \frac{1}{2\pi I_0(\kappa)\sqrt{\kappa}}e^{\kappa\cos(\frac{\beta}{\sqrt{\kappa}})} \approx \frac{e^{\kappa\cos\frac{\beta}{\sqrt{\kappa}}}}{2\pi\frac{e^\kappa}{\sqrt{2\pi\kappa}}\sqrt{\kappa}} \approx \frac{e^{\kappa(1-\frac{\beta^2}{2\kappa})}}{e^\kappa\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}e^{\frac{-\beta^2}{2}},$$

onde na primeira aproximación empregouse a aclaración (1.16). □

1.4.2. Distribución uniforme

Esta distribución reflicte a non preferencia dunha dirección sobre as demais, como ocorre coa homónima real. Isto pode apreciarse no panel da dereita da Figura 1.1. A definición en ambos casos é formalmente a mesma, posto que a función de densidade é

$$f(\theta) = \frac{1}{2\pi}, \quad \theta \in [0, 2\pi).$$

Así, pola definición de función de distribución, se $\alpha \leq \beta < \alpha + 2\pi$, entón $P(\alpha < \Theta \leq \beta) = \frac{\beta - \alpha}{2\pi}$ polo que a probabilidade é proporcional ao arco de α ata β. Xa que

$$\varphi_p = \int_0^{2\pi} \frac{e^{p\theta i} d\theta}{2\pi} = \frac{1}{2\pi} \int_0^{2\pi} \cos p\theta d\theta + \frac{i}{2\pi} \int_0^{2\pi} \operatorname{sen} p\theta d\theta = 0, \quad \forall\, p \in \mathbb{Z},\ p \neq 0,$$

todos os momentos trigonométricos son nulos, agás $\varphi_0 = 1$. Polo tanto, non hai unha dirección media definida, o que concorda con que ningunha dirección destaca sobre o resto. Cúmprese, ademais, que a lonxitude da media mostral R e a dirección $\bar{\theta}$ son independentes se, e só se, a mostra provén dunha distribución uniforme, ver [11]. Ademais, se Θ_1 segue unha distribución uniforme e Θ_2 ten unha distribución calquera, sendo estas independentes, entón a función característica de $(\Theta_1 + \Theta_2)$ coincide coa función característica da distribución uniforme, é dicir, ten todos os momentos nulos agás o de orde 0, que toma o valor 1 e, polo tanto, $(\Theta_1 + \Theta_2)$ segue unha distribución uniforme, xa que a función de disitribución queda totalmente determinada pola función característica (ver [13], Capítulo 4).

1.4.3. Distribucións proxectadas

Pártese agora dunha densidade no plano $\mathbb{R}^2$. A idea neste caso é realizar unha proxección radial da distribución á circunferencia. Alxebricamente, consideremos o vector bivariante (X, Y), e transformemos o vector anterior a cordenadas polares, é dicir, substituíndo $X = S\cos(\Theta)$ e $Y = S\operatorname{sen}(\Theta)$. Logo, para acumular a probabilidade dunha dirección, basta integrar a función de densidade de (X, Y), g, sobre todas as distancias posibles fixado un ángulo. Deste xeito, a nova función de densidade, f, que dependerá só da dirección, será

$$f(\theta) = \int_0^{\infty} sg(s\cos(\theta), s\operatorname{sen}(\theta)) ds.$$

É destacable a posibilidade de obter a distribución de von Mises a partir da proxección dunha distribución no plano sobre a circunferencia. Consideremos as variables $X \in N(\cos\mu, \frac{1}{\kappa})$ e $Y \in N(\operatorname{sen}\mu, \frac{1}{\kappa})$ independentes. Tendo en conta o cambio de variable a coordenadas polares $X = S\cos\Theta$ e $Y = S\operatorname{sen}\Theta$, a función de densidade conxunta das

(a) $\boldsymbol{\mu} = (0, 0)$; azul, $\tau = 0.3$; e verde, $\tau = 0.9$.

(b) $\tau = 0.50$; lila, $\boldsymbol{\mu} = (-2, 0)$; e laranxa, $\boldsymbol{\mu} = (2.5, -1.5)$.

Figura 1.3: Exemplos da densidade da distribución normal proxectada.

variables (S, Θ) é a seguinte:

$$f(s,\theta) = \frac{\kappa s}{2\pi} \exp\left\{\frac{-\kappa}{2}\left(s^2 - 2s\cos(\theta-\mu)+1\right)\right\}, \quad s \in [0,\infty), \quad \theta \in [0, 2\pi).$$

Polo tanto, se consideramos a función de densidade anterior condicionada a que $S = 1$, é dicir, a densidade da restrición aos puntos da circunferencia, obtense a función de densidade da distribución de von Mises, $vM(\mu, \kappa)$, ver [11].

Unha das distribucións máis importantes no plano é a normal bivariante, $\phi(x, y; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, onde $\boldsymbol{\mu} = (\mu, \nu)'$ é o vector de medias e $\boldsymbol{\Sigma}$ a matriz de varianzas-covarianzas da distribución. A partir da distribución anterior pode obterse a normal proxectada. Denotemos por τ a correlación entre as variables X e Y; e σ_1^2 e σ_2^2 as súas varianzas respectivas. Entón, a función de densidade da distribución vén dada por

$$f(\theta) = \frac{1}{C(\theta)}\left\{\phi(\mu,\nu;0,\boldsymbol{\Sigma}) + aD(\theta)\Phi[D(\theta)]\phi\left[\frac{a(\mu\,\mathrm{sen}(\theta) - \nu\cos(\theta))}{\sqrt{C(\theta)}}\right]\right\},$$

sendo

$$a = \frac{1}{\sigma_1\sigma_2\sqrt{1-\tau^2}}, \quad C(\theta) = a^2\left(\sigma_2^2\cos^2(\theta) - \tau\sigma_1\sigma_2\,\mathrm{sen}(2\theta) + \sigma_1^2\,\mathrm{sen}^2(\theta)\right),$$

$$D(\theta) = \frac{a^2}{\sqrt{C(\theta)}}\left[\mu\sigma_2(\sigma_2\cos(\theta) - \tau\sigma_1\,\mathrm{sen}(\theta)) + \nu\sigma_1(\sigma_1\,\mathrm{sen}(\theta) - \tau\sigma_2\cos(\theta))\right]$$

e ϕ e Φ son, respectivamente, as funcións de densidade e de distribución de $N(0,1)$. Exemplos desta distribución poden verse na Figura 1.3, onde pode apreciarse graficamente o papel que xogan o vector de medias e o coeficiente de correlación. Simplifícase o caso anterior cando $\boldsymbol{\mu} = \mathbf{0}$ e $\tau = 0$, co que se obtén a función de densidade

$$f(\theta) = \frac{\sqrt{1-b^2}}{2\pi(1 - b\cos(2\theta))},$$

sendo $b = \frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$. Outro exemplo, neste caso con $\boldsymbol{\mu} = 0$ e $\sigma_1 = 1 = \sigma_2$, leva á función de densidade

$$f(\theta) = \frac{\sqrt{1-\tau^2}}{2\pi(1 - \tau\,\mathrm{sen}(2\theta))}.$$

Deste caso séguese que se as variables son incorreladas, é dicir, $\tau = 0$, a distribución sería unha uniforme.

1.4.4. Distribucións enroladas

Agora consideremos unha variable aleatoria definida sobre $\mathbb{R}$, con funcións de densidade g e de distribución G. Para envolver a distribución sobre a circunferencia é preciso reducir a modulo 2π todos os valores nos que estea definida a densidade. Para isto, consideremos a relación

$$\Theta = X \pmod{2\pi}. \tag{1.17}$$

Así, a cada $\theta \in [0, 2\pi)$ asócianselle todos os puntos da forma $x = \theta + 2\pi k$, para $k \in \mathbb{Z}$ e a función de distribución desta nova variable vén dada por

$$F(\theta) = \sum_{k\in\mathbb{Z}} [G(\theta + 2\pi k) - G(2k\pi)], \quad \theta \in [0, 2\pi),$$

e a función de densidade da nova variable, denotada por f, definirase como

$$f(\theta) = \sum_{k\in\mathbb{Z}} g(\theta + 2\pi k), \quad \theta \in [0, 2\pi). \tag{1.18}$$

Enúncianse a continuación unha serie de propiedades deste tipo de distribucións:

I. Se $\Theta_1 = X_1 \pmod{2\pi}$ e $\Theta_2 = X_2 \pmod{2\pi}$, entón $(\Theta_1 + \Theta_2) = (X_1 + X_2) \pmod{2\pi}$.

II. O momento trigonométrico p-ésimo da variable enrolada coincide coa avaliación do enteiro p na función característica da variable orixinal. En efecto, sexan G e F as funcións de distribución de X e Θ respectivamente. Entón, por definición

$$\varphi_p = \int_0^{2\pi} e^{p\theta i} dF(\theta).$$

Facendo uso da relación (1.17), obtense que:

$$\varphi_p = \sum_{k\in\mathbb{Z}} \int_{2k\pi}^{2\pi(k+1)} e^{p\theta i} dG(x) = \int_{-\infty}^{\infty} e^{p\theta i} dG(x) = \phi_X(p),$$

onde ϕ_X representa a función característica da variable aleatoria definida sobre a recta real, X.

III. Se a función característica da variable aleatoria sobre a recta real, ϕ_X, é integrable, entón a variable Θ ten función de densidade e, ademais, admite a seguinte expresión como serie de Fourier

$$f(\theta) = \sum_{k\in\mathbb{Z}} g(\theta + 2k\pi) = \frac{1}{2\pi}\left\{1 + 2\sum_{k=1}^{\infty}\left[\alpha_k \cos(k\theta) + \beta_k \operatorname{sen}(k\theta)\right]\right\}. \tag{1.19}$$

IV. Hai unha cantidade non finita de distribucións na recta real que poden ser enroladas ao redor da circunferencia, ver [13] para máis detalles.

A continuación expóñense dous exemplos importantes de variables obtidas envolvendo a orixinal sobre a circunferencia.

1.4.5. Distribución normal enrolada

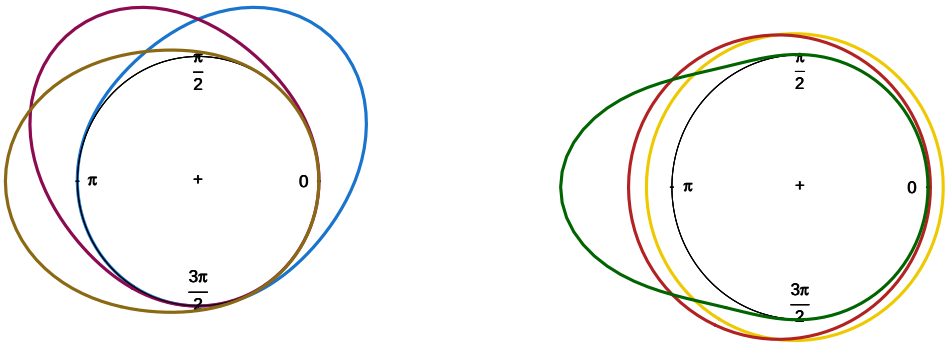
Esta distribución obtense tras envolver unha $N(\mu, \sigma^2)$ ao redor da circunferencia. Así, en virtude de (1.18), a función de densidade vén dada por:

$$f(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k\in\mathbb{Z}} \exp\left[\frac{-(\theta - \mu - 2\pi k)^2}{2\sigma^2}\right]. \tag{1.20}$$

Outra expresión para a función de densidade é a seguinte:

$$f(\theta) = \frac{1}{2\pi}\left[1 + 2\sum_{k=1}^{\infty} \rho^{k^2} \cos\left(k(\theta - \mu)\right)\right], \tag{1.21}$$

onde $\rho = e^{\frac{-\sigma^2}{2}} \in (0, 1]$ é o parámetro de concentración, que se interpreta do mesmo xeito que o módulo do primeiro momento trigonométrico. Como pode verse no panel da dereita da Figura 1.4, a medida que o valor de ρ aumenta, a distribución está, en conxunto, máis próxima ao redor de valor de μ. Denotarase esta distribución por $WN(\mu, \rho)$. É preciso apuntar que a densidade pode ser aproximada polos tres primeiros termos de (1.21) se $\sigma^2 > 2\pi$ e polo termo correspondente a $k = 0$ de (1.20) se $\sigma^2 \leq 2\pi$. Esta distribución é unimodal sendo μ a moda. Isto pode apreciarse no panel esquerdo da Figura 1.4, na que



(a) $\rho = 0.80$, $\mu = \frac{\pi}{4}$ en azul, $\frac{3\pi}{4}$ en morado; e π en marrón.

(b) $\mu = \pi$, $\rho = 0.125$ en amarelo; $\rho = 0.5$ en marrón; e $\rho = 0.9$ en verde.

Figura 1.4: Exemplos da densidade da distribución normal enrolada.

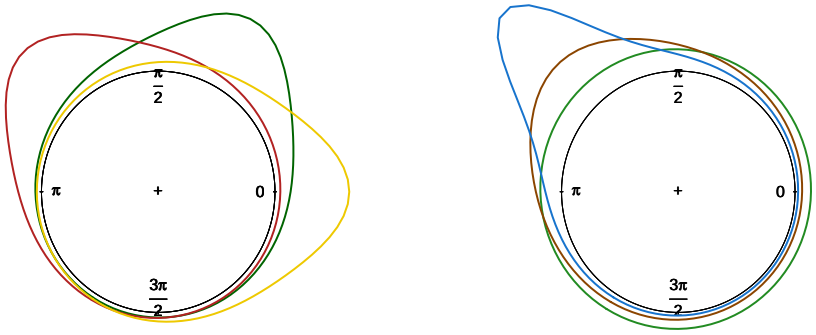
a moda coincide co valor que tome o parámetro μ. Ademais, é, por definición, simétrica respecto da dirección μ, como pode verse na Figura 1.4. En efecto,

$$f(\mu+\theta) = \frac{1}{2\pi}\left[1 + 2\sum_{k=1}^{\infty}\rho^{k^2}\cos\left(k\theta\right)\right] = \frac{1}{2\pi}\left[1 + 2\sum_{k=1}^{\infty}\rho^{k^2}\cos\left(k(-\theta)\right)\right] = f(\mu-\theta),$$

debido á paridade da función coseno.

Ademais, esta distribución posúe a propiedade de reprodutividade, xa que se $\Theta_1 \in WN(\mu_1, \rho_1)$ e $\Theta_2 \in WN(\mu_2, \rho_2)$, onde Θ_k foi obtida a partir de $X_k \in N(\mu_k, \sigma_k^2),\ k \in \{1, 2\}$ e, ademais, son independentes, entón $(\Theta_1 + \Theta_2) \in WN(\mu_1 + \mu_2, \rho_1\rho_2)$. En efecto, por ser independentes e $X_k \in N(\mu_k, \sigma_k^2), \quad k \in \{1, 2\}$, entón $(X_1 + X_2) \in N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Por outra banda, é coñecido que $(\Theta_1 + \Theta_2) = (X_1 + X_2) \pmod{2\pi}$. Por tanto, pode concluírse que $(\Theta_1 + \Theta_2) \in WN(\mu_1 + \mu_2, \rho)$ onde $\rho = \exp(-\frac{\sigma_1^2+\sigma_2^2}{2}) = \exp(-\frac{\sigma_1^2}{2})\exp(-\frac{\sigma_2^2}{2}) = \rho_1\rho_2$ é o parámetro de concentración.

Por último, cabe destacar que esta distribución aproxímase á uniforme circular cando $\rho \to 0$, como se pode ver no panel da dereita da Figura 1.4, e fai o mesmo con respecto á distibución que concentra toda a súa densidade nunha única dirección, concretamente μ, cando $\rho \to 1$, o que pode verse no panel da esquerda na Figura 1.4.



(a) $\rho = 0.60$, $\mu = 0$ amarelo, $\mu = \frac{\pi}{3}$ verde; e $\mu = \frac{3\pi}{4}$, vermello.

(b) $\mu = \frac{3\pi}{4}$, $\rho = 0.1$, verde; $\rho = 0.5$, marrón; e $\rho = 0.75$ azul.

Figura 1.5: Exemplos da densidade da distribución de Cauchy enrolada.

1.4.6. Distribución de Cauchy enrolada

Consideremos agora a función de densidade da distribución de Cauchy sobre a recta real:

$$g(x) = \frac{\pi^{-1}\sigma}{\sigma^2 + (x-\mu)^2}, \quad x \in \mathbb{R}$$

e a súa función característica $\phi_X(t) = e^{-\sigma|t| + t\mu i}, \quad t \in \mathbb{R}$. A partir de (1.19) e empregando a función característica anterior, pode obterse a función de densidade da distribución circular que se está a tratar, que ten a seguinte expresión (ver [13]):

$$f(\theta) = \sum_{k\in\mathbb{Z}} g(\theta + 2k\pi) = \frac{1}{2\pi}\left\{1 + 2\sum_{k=1}^{\infty}\rho^k \cos k(\theta-\mu)\right\}, \tag{1.22}$$

onde $\rho = e^{-\sigma}$. Así pois, farase referencia a esta distribución mediante $WC(\mu, \rho)$. Por outra banda, tendo en conta que

$$\sum_{k=1}^{\infty}\rho^k e^{-ik(\theta-\mu)} = \sum_{k=1}^{\infty}\rho^k \cos k(\mu-\theta) + i\sum_{k=1}^{\infty}\rho^k \operatorname{sen} k(\mu-\theta),$$

entón a expresión (1.22) pode escribirse como

$$f(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}.$$

Ademais, a función característica é $\varphi_p = \rho^{|p|}e^{p\mu i}$, onde $\alpha_p = \rho^{|p|}\cos(p\mu)$ e $\beta_p = \rho^{|p|}\operatorname{sen}(p\mu)$. Así, a dirección media é μ e o parámetro de concentración da distribución será ρ, como pode apreciarse no panel da dereita da Figura 1.5. Nela, foi fixado un valor para o parámetro μ e varíanse os valores de ρ, de tal xeito que pode observarse que canto maior é o valor deste último parámetro, a distribución está máis concentrada en torno á dirección μ. A distribución é unimodal, sendo μ a moda, o cal pode verse no panel da esquerda da Figura 1.5, na que se fixou o valor de ρ e a moda da distribución coincide co valor que foi asociado a μ. Ademais, é simétrica respecto da dirección media, μ, como está ilustrado na Figura 1.5; e cando o parámetro de concentración tende a cero, a distribución tende á uniforme circular, mentres que se tende a 1, aproxímase á distribución que acumula toda a probabilidade na dirección μ, tamén apreciable nestas mesmas imaxes.

1.4.7. Distribución cardioide

A partir da perturbación da densidade da uniforme mediante o coseno pode obterse a seguinte función de densidade

$$f(\theta) = \frac{1}{2\pi}\left\{1 + 2\rho\cos(\theta - \mu)\right\}, \quad \theta, \mu \in [0, 2\pi), \quad \rho \in \left(\frac{-1}{2}, \frac{1}{2}\right). \tag{1.23}$$

Polo tanto, se $\rho = 0$, a distribución redúcese a unha uniforme. Representarase por $C(\mu, \rho)$ e o nome da distribución é debido a que a curva que representa $r = f(\theta)$ é a cardioide. A distribución é unimodal con moda μ e antimoda $\mu \pm \pi$. En efecto, a ecuación $f'(\theta) = 0$, onde $f'(\theta) = \frac{-\rho}{\pi}\operatorname{sen}(\theta - \mu)$, ten como solucións os valores $\theta - \mu = 0$, $\theta - \mu = \pi$ e $\theta - \mu = -\pi$. Así, como $f''(\theta) = \frac{-\rho}{\pi}\cos(\theta - \mu)$ e ademais $f''(\mu) = \frac{-\rho}{\pi}$ e $f''(\mu \pm \pi) = \frac{\rho}{\pi}$, conclúese que $\theta = \mu$ dá valor máximo e $\theta = \mu \pm \pi$ mínimo á función de densidade. Isto pode apreciarse na Figura 1.6, onde a dirección de máxima probabilidade coincide co valor que toma μ e, ademais, a dirección de menor probabilidade é precisamente $\mu \pm \pi$. Ademais, é, por definición, simétrica con respecto da dirección μ. En efecto,

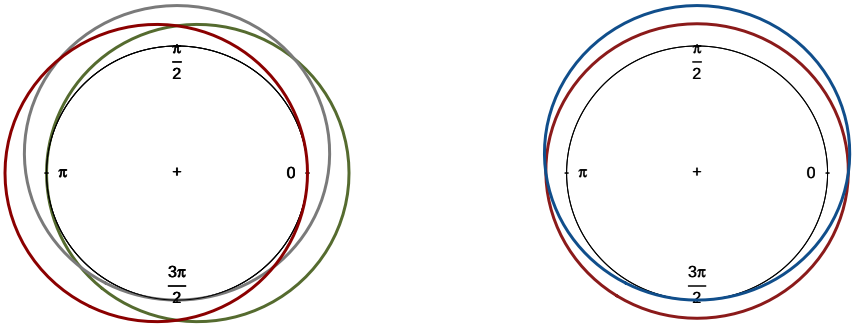
$$f(\mu + \theta) = \frac{1}{2\pi}\left\{1 + 2\rho\cos(\theta)\right\} = \frac{1}{2\pi}\left\{1 + 2\rho\cos(-\theta)\right\} = f(\mu - \theta),$$

en virtude da paridade do coseno.

Por último, tense que $\varphi_1 = |\rho|e^{i\mu}$, polo tanto, $|\varphi_1| = |\rho|$; así o parámetro ρ fai referencia á concentración e μ é a dirección media. Que ρ é o parámetro de concentración pode verse no panel da dereita da Figura 1.6, onde a medida que aumenta o valor de ρ, a distribución está máis próxima do valor de μ. Ademais, cúmprese que $\varphi_p = 0, \forall\ p > 1$.

Integrando (1.23) respecto de θ, obtense que a función de distribución vén dada por:

$$F(\theta) = \frac{1}{2\pi}\left\{\theta + 2\rho\operatorname{sen}(\theta - \mu)\right\}, \quad \forall\ \theta \in [0, 2\pi).$$



(a) $\rho = 0.50$, $\mu = 0$, en verde; $\mu = \frac{\pi}{2}$, en azul; e(b) $\mu = \frac{\pi}{2}$, $\rho = 0.05$ en marrón; e $\rho = 0.5$ en azul. $\mu = \pi$, en marrón.

Figura 1.6: Exemplos da densidade da distribución carioide.

1.4.8. Mesturas de distribucións

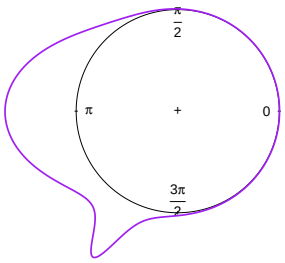
Ata o de agora, só se citaron disitribucións circulares unimodais. De querer xerar outras con dúas ou máis modas, é útil a mestura dalgunhas distribucións que xa foron mencionadas previamente, o que se leva a cabo do seguinte xeito: sexan $f_1(\theta), f_2(\theta), \dots, f_M(\theta)$ funcións de densidade dunha mesma ou de diferentes distribucións, sen ter, necesariamente, todas elas os mesmos parámetros. A partir destas pode obterse a función de densidade seguinte:

$$g(\theta) = \sum_{k=1}^{M} p_k f_k(\theta), \tag{1.24}$$

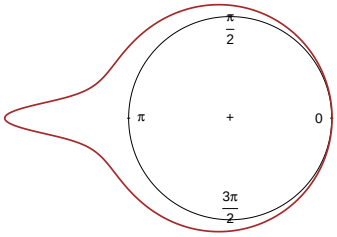
onde os p_k chámanse parámetros de mestura e $\sum_{k=1}^{M} p_k = 1$, sendo $p_k \geq 0$, é dicir, obtense unha combinación convexa das funcións de densidade. Como exemplo, considéranse as distribucións de von Mises seguintes con diferentes parámetros, $vM(\mu_k, \kappa_k), \quad k \in \{1, \dots, M\}$ a partir das cales obtense a seguinte función de densidade

$$g(\theta) = \sum_{k=1}^{M} p_k \frac{e^{\kappa_k \cos(\theta - \mu_k)}}{2\pi I_0(\kappa)}. \tag{1.25}$$

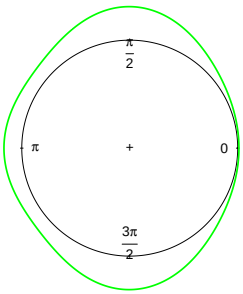
Poden verse exemplos dalgunhas funcións de densidade coa forma (1.24) e do caso particular de (1.25) na Figura 1.7. Nestas imaxes pode comprobarse como ao mesturar densidades unimoidais pode obterse unha densidade con varias modas e tamén como a mestura de densidades simétricas pode dar lugar a densidades que non teñan esta propiedade.



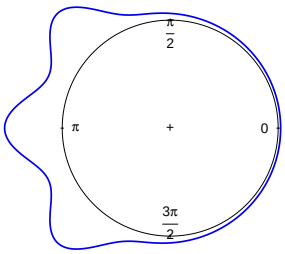
(a) $\frac{4}{5}vM(\pi, 5) + \frac{1}{5}WC(\frac{4\pi}{3}, 0.9)$.



(b) $\frac{2}{3}C(\pi, 0.5) + \frac{1}{3}WC(\pi, 0.9)$.



(c) $\frac{2}{5}vM(\frac{\pi}{2}, 4) + \frac{1}{5}vM(\pi, 5) + \frac{2}{5}vM(\frac{3\pi}{2}, 4)$.



(d) $\frac{1}{2}vM(\pi, 1) + \frac{1}{6}vM(\pi - 0.8, 30) + \frac{1}{6}vM(\pi, 30) + \frac{1}{6}vM(\pi + 0.8, 30)$.

Figura 1.7: Algúns exemplos de densidades de mesturas de distribucións.

1.4.9. Distribucións asimétricas

As mesturas de distintas distribucións permiten obter densidades asimétricas e/ou multimodais. Introdúcese agora un método xeral (ver [12]) co que poder obter máis distribucións con estas características e para iso, considérase unha función de densidade simétrica ao redor de μ, f_0. A partir dela, defínese a función

$$f(\theta) = f_0(\theta - \mu)(1 + \lambda \operatorname{sen}(\theta - \mu)), \tag{1.26}$$

onde $\theta \in [0, 2\pi)$ e $\lambda \in (-1, 1)$ chámase parámetro de asimetría. En virtude da definición, é claro que f segue sendo simétrica cando $\lambda = 0$. Por outra banda, se $\lambda > 0$, é asimétrica en sentido antihorario e se $\lambda < 0$, en sentido horario. Ademais, a partir desta función de densidade pode obterse a función de distribución da nova variable mediante integración do seguinte xeito

$$F(\theta) = \int_0^\theta [f_0(\varphi) + \lambda f_0(\varphi) \operatorname{sen}(\varphi)]\, d\varphi = F_0(\theta) + \lambda \int_0^\theta f_0(\varphi) \operatorname{sen}(\varphi) d\varphi,$$

onde $F_0(\theta)$ é a función de distribución asociada a $f_0(\theta)$ e se supuxo, sen perda de xeralidade, que $\mu = 0$.

En canto aos momentos trigonométricos, compre introducir a seguinte notación: chamemos $\alpha_{0,p}$ á parte real do momento trigonométrico p-ésimo da variable aleatoria orixinal, que ten por función de densidade $f_0(\theta)$. Así, os momentos trigonométricos da nova distribución serán, como se expón en [1],

$$\alpha_p = \mathbb{E}[\cos p\theta] = \alpha_{0,p} \quad \text{e} \quad \beta_p = \mathbb{E}[\operatorname{sen} p\theta] = \lambda \frac{\alpha_{0,p-1} - \alpha_{0,p+1}}{2}, \quad p \in \mathbb{Z}.$$

Polo tanto, a parte real do momento trigonométrico da nova variable coincide co da simétrica e, así, $\forall\, p \in \mathbb{Z}$, tense que:

$$\rho_p = \sqrt{(\alpha_{0,p})^2 + \lambda^2 \frac{(\alpha_{0,p-1} - \alpha_{0,p+1})^2}{4}} \quad \text{e} \quad \mu_p = \operatorname{atan2}\left(\lambda \frac{\alpha_{0,p-1} - \alpha_{0,p+1}}{2}, \alpha_{0,p}\right).$$

En particular, o parámetro de concetración e a dirección media son

$$\rho = \sqrt{(\alpha_{0,1})^2 + \lambda^2 \frac{(1 - \alpha_{0,2})^2}{4}} \quad \text{e} \quad \mu = \operatorname{atan2}\left(\lambda \frac{1 - \alpha_{0,2}}{2}, \alpha_{0,1}\right),$$

respectivamente.

Estas novas densidades tamén son, nalgúns casos, bimodais se f_0 é unha función de densidade axeitada. Por exemplo, no caso da Cauchy enrolada, é estritamente unimodal. A continuación póñense exemplos de funcións de densidade coa forma de (1.26), sendo estes exemplos e as súas propiedades os expostos en [1]. Primeiramente, consideremos a función

de densidade da distribución de von Mises, dada por (1.12), tomando como a media $\mu = 0$. Así, por (1.26), a nova función sería

$$f(\theta) = \frac{e^{\kappa\cos(\theta)}}{2\pi I_0(\kappa)}\left(1 + \lambda\,\mathrm{sen}(\theta)\right), \quad \theta \in [0, 2\pi).$$

De xeito inmediato, pode obterse a función de distribución como

$$\begin{aligned} F(\theta) &= \int_0^\theta \left(\frac{e^{\kappa\cos(\varphi)}}{2\pi I_0(\varphi)}\left(1 + \lambda\,\mathrm{sen}(\varphi)\right)\right) d\varphi = \int_0^\theta \frac{e^{\kappa\cos(\varphi)}}{2\pi I_0(\varphi)} d\varphi + \int_0^\theta \frac{\mathrm{sen}(\varphi) e^{\kappa\cos(\varphi)}}{2\pi I_0(\varphi)} d\varphi = \\ &= F_0(\theta) - \frac{\lambda}{2\pi I_0(\kappa)}\left[e^\kappa - e^{\kappa\cos(\theta)}\right], \quad \theta \in [0, 2\pi). \end{aligned}$$

Os momentos trigonométricos desta nova distribución son

$$\alpha_p = \frac{I_p(\kappa)}{I_0(\kappa)} \quad \text{e} \quad \beta_p = \lambda\frac{I_{p-1}(\kappa) - I_{p+1}(\kappa)}{2I_0(\kappa)} = \frac{p\lambda I_p(\kappa)}{\kappa I_0(\kappa)} = \frac{p\lambda\alpha_p}{\kappa},$$

onde se empregou na segunda igualdade da parte imaxinaria que $I_{p-1}(\kappa) - I_{p+1}(\kappa) = \frac{2pI_p(\kappa)}{\kappa}$. Así,

$$\rho_p = \frac{I_p(\kappa)}{\kappa I_0(\kappa)}\sqrt{\kappa^2 + p^2\lambda^2}, \quad \text{e} \quad \mu_p = \mathrm{atan2}(p\lambda, \kappa)$$

e, en particular, o parámetro de concentración e a media serán

$$\rho = \frac{A(\kappa)}{\kappa}\sqrt{\kappa^2 + \lambda^2} \quad \text{e} \quad \mu = \mathrm{atan2}(\lambda, \kappa).$$

Con respecto á modalidade, ao denotar $D = (1-\lambda^2)\kappa^6 + \lambda^2(16\lambda^4 - 20\lambda^2 + 3)\kappa^4 + \lambda^4(8\lambda^2 + 3)\kappa^2 + \lambda^6$, a distribución é bimodal cando $D \geq 0$ e, se $D < 0$, unimodal, como se expón en [1], onde tamén se indica que na maioría dos casos é bimodal.

Outro exemplo pode obterse tomando como distribución inicial a cardioide, é dicir, tomar $f_0(\theta) = \frac{1}{2\pi}\{1 + 2\rho\cos(\theta - \mu)\}$, co que a función de densidade resultante sería

$$f(\theta) = \frac{1}{2\pi}\{1 + 2\rho\cos(\theta)\}\left(1 + \lambda\,\mathrm{sen}(\theta)\right), \quad \theta \in [0, 2\pi),$$

onde $0 \leq \rho \leq 1$ e $0 \leq \lambda \leq 1$. Facendo os cambios $\cos\xi = \frac{\kappa}{\sqrt{\kappa^2+\lambda^2}}$ e $\mathrm{sen}\,\xi = \frac{\lambda}{\sqrt{\kappa^2+\lambda^2}}$, a densidade pode ser escrita como segue

$$f(\theta) = \frac{1}{2\pi}\left\{1 + \sqrt{\lambda^2 + \kappa^2}\cos(\theta - \xi) + \frac{\kappa\lambda}{2}\cos\left[2\left(\theta - \frac{\pi}{4}\right)\right]\right\}, \quad \theta \in [0, 2\pi).$$

A partir desta función, pode obterse a función de distribución inmediatamente, mediante integración desta respecto de θ, resultando o seguinte:

$$F(\theta) = \frac{1}{2\pi}\left\{\theta + \pi + \lambda(1 + \cos(\theta))\left[-1 + \frac{\kappa}{2}(1 - \cos\theta)\right]\right\}, \quad \theta \in [0, 2\pi).$$

Por ser todos os momentos cero na cardioide, agás os de orde cero e un, os momentos da nova distribución son

$$\alpha_1 = \frac{\rho}{2} \quad \beta_1 = \frac{\lambda}{2} \quad \text{e } \beta_2 = \frac{\rho\lambda}{4},$$

sendo $\alpha_p = 0,\ \forall\ p \geq 2$ e $\beta_p = 0,\ \forall\ p \geq 3$. Do mesmo xeito que coa distribución anterior, existen casos nos que a densidade é unimodal e outros nos que é bimodal. Ao denotar por $D = (1-\lambda^2)(1+8\lambda^2)^2\kappa^6 + 3\lambda^2(1-26\lambda^2+16\lambda^4)\kappa^4 + 3\lambda^4(1+5\lambda^2)\kappa^2 + \lambda^6$, se $D \geq 0$, a distribución é unimodal, e será bimodal se $D < 0$. Do mesmo xeito que ocorría coa von Mises, en [1] móstrase que na maioría dos casos será bimodal.

1.4.10. Distribución Batschelet inversa

Neste caso non se trata dunha distribución en particular, senón que a partir de certa transformación pode obterse unha distribución con dous parámetros máis que aquela que se tome como base, aumentando así a riqueza das distribucións que xa se teñan definido e permitindo modelar máis situacións. Trátase da transformación de escala, que se estuda en profundidade en [16]. Para definilas, é preciso introducir primeiramente as funcións

$$t_{1,\nu}(\theta) = \theta - \nu(1+\cos\theta) \quad \text{e} \quad s_\lambda(\theta) = \theta - \frac{1}{2}(1+\lambda)\operatorname{sen}\theta,$$

coa que se define :

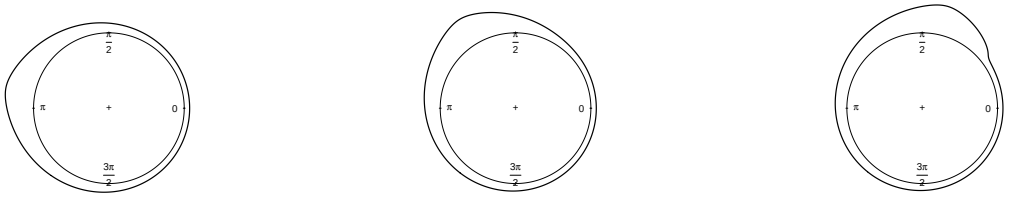
$$t_\lambda(\theta) = \begin{cases} \frac{1-\lambda}{1+\lambda}\theta + \frac{2\lambda}{1+\lambda}s_\lambda^{-1}(\theta) & \text{se } \lambda \in (-1,1], \\ \theta - \operatorname{sen}\theta & \text{se } \lambda = -1, \end{cases}$$

onde $s_\lambda^{-1}(\theta)$ fai referencia á función inversa de $s_\lambda(\theta)$ e $\nu,\ \lambda \in [-1,1]$ son parámetros de asimetría e apuntamento da función respectivamente. Tomando como base a von Mises, é dicir, a partir da composición da función anterior e da densidade da von Mises, pode obterse a seguinte función de densidade

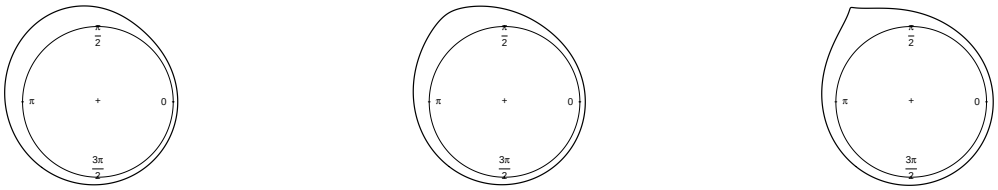
$$f(\theta) = \frac{1}{2\pi I_0(\kappa)K_{\kappa,\lambda}}\exp\left\{\kappa\cos\left(\frac{1-\lambda}{1+\lambda}t_\nu(\theta-\mu) + \frac{2\lambda}{1+\lambda}s_\lambda^{-1}(t_\nu(\theta-\mu))\right)\right\}, \qquad (1.27)$$

onde $t_\nu(\theta) = t_{1,\nu}^{-1}(\theta)$, μ é o parámetro de localización, $\kappa \geq 0$ é o parámetro de concentración e

$$K_{\kappa,\lambda} = \begin{cases} \frac{1-\lambda}{1+\lambda} - \frac{2\lambda}{(1-\lambda)2\pi I_0(\kappa))}\int_0^{2\pi}\exp\{\kappa\cos(\theta - \frac{1}{2}(1-\lambda)\operatorname{sen}\theta)\}d\theta & \text{se} \quad \lambda \in [-1,1), \\ 1 - A_1(\kappa) & \text{se} \qquad \lambda = 1. \end{cases}$$



(a) $\lambda = 0.5$ e $\nu = 0.1$. (b) $\lambda = 0.5$ e $\nu = 0.5$. (c) $\lambda = 0.5$ e $\nu = 0.9$.



(d) $\nu = 0.5$ e $\lambda = 0.1$. (e) $\nu = 0.5$ e $\lambda = 0.5$. (f) $\nu = 0.5$ e $\lambda = 0.9$.

Figura 1.8: Exemplos de densidade da distribución de Batschelet inversa con base a von Mises de parámetros $\mu = \pi$ e $\kappa = 1.25$.

Con respecto a ν, pode verse na primeira fila da Figura 1.8 como, fixados os valores dos demais parámetros, a medida que este aumenta, a densidade é máis asimétrica respecto da moda, que tamén varía, aínda que esta é sempre única. Na segunda fila da mesma figura, pode apreciarse como, canto maior sexa o valor do parámetro λ, a densidade presenta un maior apuntamento na moda e canto menor é este, o trazado é máis suave. Cabe sinalar que se tomase os mesmos valores pero negativos, as representacións obtidas serían simétricas con respecto da dirección $\mu = \pi$ ás representadas co parámetro positivo. Cúmprese que a media é $(\mu - 2\nu)$ (mód 2π) cando $\kappa > 0$ e, en particular, se $\nu = 0$, esta resulta ser μ, o cal explica a variación da moda nos paneis da primeira fila de 1.8.

Neste capítulo citáronse algunhas distribucións circulares importantes. Non obstante, existen moitas máis e poden ser consultadas en [11] ou en [16], por exemplo as envoltas estables, as beta circulares ou a familia de distribucións Jones-Pewsey, á que pertencen algunhas das distribucións expostas, como a von Mises, a normal enrolada ou a Cauchy enrolada. A continuación veremos como estimar os parámetros dalgunhas das distribucións máis salientables das que foron expostas neste capítulo, empregando o método de máxima

verosimilitude e coa axuda do algoritmo EM para as mesturas de von Mises.

Capítulo 2

Estimación de parámetros

Neste capítulo veremos como se poden estimar os parámetros de tres das máis relevantes das distribucións no Capítulo 1, como son a von Mises, a normal enrolada e a Cauchy enrolada; así como da mestura de von Mises. Para isto empregarase o método de máxima verosimilitude que, nalgúns casos, proporcionará unha expresión explícita da solución e noutros terá que ser calculada con métodos numéricos. Este método consiste no seguinte: considérase unha mostra aleatoria simple de datos $\theta_1, \theta_2, \dots, \theta_n$, é dicir, un conxunto de observacións independientes e identicamente distribuídas. Denotemos as densidades por $f_\gamma(\theta_1), f_\gamma(\theta_2), \dots, f_\gamma(\theta_n)$, sendo γ o parámetro ou conxunto de parámetros presentes na función de densidade. Defínese a función de densidade conxunta da mostra como

$$f(\theta_1, \dots, \theta_n; \gamma)$$

e, en virtude da independencia dos datos, tense que:

$$\mathcal{L} = f(\theta_1, \dots, \theta_n; \gamma) = f_\gamma(\theta_1) \dots f_\gamma(\theta_n) = \prod_{k=1}^{n} f_\gamma(\theta_k).$$

Diremos que $\mathcal{L}$ é a función de verosimilitude e será importante $l := \log(\mathcal{L})$, que recibe o nome de log-verosimilitude, onde log denota o logaritmo neperiano. O método de máxima verosimilitude busca maximizar a función de verosimilitude $\mathcal{L}$ con respecto de γ, ou, en virtude da monotonía estrita da funcion logaritmo, maximizar l, para obter cal é o valor máis probable do parámetro (ou parámetros) γ, levándose a cabo do mesmo xeito que con datos sobre a recta real[1].

[1]Este método introdúcese na asignatura de Inferencia Estatística do terceiro curso do Grao en Matemáticas.

2.1. Estimación da distribución von Mises

En virtude de (1.12), a función de densidade da von Mises é

$$f(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)};\quad \theta\in[0,2\pi),$$

polo que, neste caso, os parámetros a estimar son μ e κ. Consideremos entón unha mostra aleatoria simple $\theta_1,\theta_2,\dots,\theta_n$ dunha von Mises con media μ e concentración κ. Así, a función de verosimilitude é

$$\mathcal{L}=\prod_{k=1}^{n}f_{(\mu,\kappa)}(\theta_k)=(2\pi I_0(\kappa))^{-n}e^{\kappa\sum_{k=1}^{n}\cos(\theta_k-\mu)}.$$

Para seguir o método de máxima verosimilitude é preciso maximizar a función anterior. Non obstante, como a función logarítmo é estritamente crecente, o máximo da función $\mathcal{L}$ pode ser obtido a partir da log-verosimilitude:

$$l=-n\log(2\pi I_0(\kappa))+\kappa\sum_{k=1}^{n}\cos(\theta_k-\mu). \tag{2.1}$$

A derivada parcial desta log-verosimilitude respecto de μ é

$$\frac{\partial l}{\partial\mu}=\kappa\sum_{k=1}^{n}\operatorname{sen}(\theta_k-\mu).$$

Ao plantexar a condición de punto crítico, $\frac{\partial l}{\partial\mu}=0$, tense o seguinte

$$\sum_{k=1}^{n}\operatorname{sen}(\theta_k-\mu)=0$$

e, en virtude da fórmula do seno do ángulo suma,

$$\sum_{k=1}^{n}\left[\operatorname{sen}(\theta_k)\cos(\mu)\right]=\sum_{k=1}^{n}\left[\cos(\theta_k)\operatorname{sen}(\mu)\right].$$

Sacando factor común en ambos membros e empregando as definicións (1.7), resulta que

$$b_1\cos(\mu)=a_1\operatorname{sen}(\mu),$$

concluíndo finalmente que

$$\widehat{\mu}=\operatorname{atan2}(b_1,a_1)=\bar{\theta}.$$

Agora, a derivada parcial de (2.1) con respecto a κ resulta ser

$$\frac{\partial l}{\partial\kappa}=-n\frac{I_1(\kappa)}{I_0(\kappa)}+\sum_{k=1}^{n}\cos(\theta_k-\mu),$$

onde foi empregado que $\frac{d}{d\kappa}I_0(\kappa) = I_1(\kappa)$ (ver [11]). A partir das expresións en (1.10) e tendo en conta que $A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$, a igualdade $0 = \frac{\partial l}{\partial \kappa}$ é equivalente a

$$0 = n[-A(\kappa) + R\cos(\bar{\theta} - \widehat{\mu})].$$

Finalmente, tendo en conta que $\widehat{\mu} = \bar{\theta}$, obtense que

$$A(\widehat{\kappa}) = R \Longrightarrow \widehat{\kappa} = A^{-1}(R). \tag{2.2}$$

É preciso agora citar as seguints propiedades da funcion $A(\kappa)$:

1. $0 \leq A(\kappa) \leq 1, \quad \forall\, \kappa \geq 0.$
2. $A(\kappa) \to 1$ cando $\kappa \to \infty$.
3. $A'(\kappa) = 1 - A(\kappa)^2 - \frac{A(\kappa)}{\kappa} \geq 0, \quad \forall\, \kappa \geq 0.$

A existencia de $A^{-1}(R)$ pode asegurarse en virtude da continuidade da función $A(\kappa)$, da propiedade 1 anterior e tendo en conta que $R \in (0, 1]$. A terceira propiedade permite asegurar que a funcion $A(\kappa)$ é monótona polo tanto a solución a (2.2) é única (ver [11]). Ademais, poden obterse aproximacións do valor de $\widehat{\kappa}$ mediante as seguintes expresións

$$\widehat{\kappa} \approx 2R + R^3 + \frac{5}{6}R^5, \quad \widehat{\kappa} \approx -0.4 + 1.39R + \frac{0.43}{1-R} \qquad \text{e} \quad \widehat{\kappa} \approx \frac{1}{2-2R},$$

sendo a primeira útil cando $R < 0.53$, a segunda cando $0.53 \leq R \leq 0.85$ e a terceira noutro caso, sendo R o valor definido en (1.8). Por último, para comprobar que $\widehat{\mu}$ e $\widehat{\kappa}$ son, efectivamente, máximos, é preciso calcular a matriz hessiana da función de log-verosimilitude, l, que ten a seguinte expresión:

$$H = \begin{pmatrix} -n\widehat{\kappa}R & 0 \\ 0 & -nA'(\widehat{\kappa}) \end{pmatrix}$$

que é definida negativa, polo tanto, o par $(\widehat{\mu}, \widehat{\kappa})$ dá un máximo para a función l.

Acaba de verse que na distribución de von Mises a media pode ser estimada a través da media dunha mostra, do mesmo xeito ao que ocorre na distribución normal sobre a recta real. Ademais, cúmprese o recíproco.

Proposición 2.1. *A única distribución circular que cumpre que a dirección media se estima con máxima probabilidade pola media mostral é a distribución de von Mises.*

Demostración. Supoñamos unha mostra aleatoria simple $\theta_1, \theta_2, \ldots, \theta_n$ e consideremos a función de densidade da forma $f(\theta - \mu)$, onde $\mu \in [0, 2\pi)$ é o parámetro da media. Logo, a verosimilitude da mostra será $\mathcal{L} = \prod_{k=1}^{n} f(\theta_k - \mu)$. En consecuencia, $l = \sum_{k=1}^{n} \log f(\theta_k - \mu)$

e a ecuación de verosimilitude para a media é

$$0 = \frac{\partial l}{\partial \mu} = c \sum_{k=1}^{n} \frac{f'(\theta_k - \mu)}{f(\theta_k - \mu)}, \tag{2.3}$$

onde $c \in \mathbb{R}$ é unha constante. Por outra banda, no caso de que $\bar{\theta}$ estimase a μ, debería cumprirse que

$$\sum_{k=1}^{n} \operatorname{sen}(\theta_k - \mu) = 0. \tag{2.4}$$

As igualdades (2.3) e (2.4) han de cumprirse termo a termo, polo que $\forall\, \theta \in \{\theta_1, \dots, \theta_n\}$,

$$\frac{f'(\theta - \mu)}{f(\theta - \mu)} = \kappa \operatorname{sen}(\theta - \mu),$$

para κ constante. Resolvendo a ecuación diferencial anterior, obtense que:

$$f(\theta - \mu) = C e^{\kappa \cos(\theta - \mu)},$$

que é a densidade da distribución da von Mises, xa que a condición $\int_0^{2\pi} f(\theta - \mu) d\theta = 1$ implica que $C = \frac{1}{2\pi I_0(\kappa)}$. □

2.2. Estimación da distribución normal enrolada

Consideremos a función de densidade da distribución normal enrolada dada por (1.20) e unha mostra aleatoria simple dunha distribución normal enrolada, $WN(\mu, \sigma)$, $\theta_1, \theta_2, \dots, \theta_n$. Así, a función de verosimilitude vén dada por:

$$\mathcal{L} = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{\sigma^n} \left\{ \sum_{k \in \mathbb{Z}} \exp\left[\frac{-(\theta_1 - \mu - 2\pi k)^2}{2\sigma^2} \right] \dots \sum_{k \in \mathbb{Z}} \exp\left[\frac{-(\theta_n - \mu - 2\pi k)^2}{2\sigma^2} \right] \right\},$$

polo tanto, a función de log-verosimilitude é a seguinte:

$$l = \frac{-n}{2} \log(2\pi) - n \log(\sigma) + \sum_{i=1}^{n} \log \left[\sum_{k \in \mathbb{Z}} \exp\left[\frac{-(\theta_i - \mu - 2\pi k)^2}{2\sigma^2} \right] \right].$$

As ecuacións de verosimilitude resultan entón

$$0 = \frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \frac{\sum_{k \in \mathbb{Z}} \frac{\theta_i - \mu - 2\pi k}{\sigma^2} \exp\left[\frac{-(\theta_i - \mu - 2\pi k)^2}{2\sigma^2} \right]}{\sum_{k \in \mathbb{Z}} \exp\left[\frac{-(\theta_i - \mu - 2\pi k)^2}{2\sigma^2} \right]} \quad \text{e} \tag{2.5}$$

$$0 = \frac{\partial l}{\partial \sigma} = \frac{-n}{\sigma} + \sum_{i=1}^{n} \frac{\sum_{k \in \mathbb{Z}} \frac{(\theta_i - \mu - 2\pi k)^2}{\sigma^3} \exp\left[\frac{-(\theta_i - \mu - 2\pi k)^2}{2\sigma^2} \right]}{\sum_{k \in \mathbb{Z}} \exp\left[\frac{-(\theta_i - \mu - 2\pi k)^2}{2\sigma^2} \right]}. \tag{2.6}$$

Xorde entón un problema, posto que os parámetros están na forma $\mu = u(\mu, \sigma)$ e $\sigma = v(\mu, \sigma)$, onde u e v son funcións nas que se atopan as sumas infinitas das ecuacións de verosimilitude. Máis concretamente, despexando de (2.5) e (2.6), tense que

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k\in\mathbb{Z}}(\theta_i - 2k\pi)\exp\left[\frac{-(\theta_i-\mu-2k\pi)^2}{2\sigma^2}\right]}{\sum_{k\in\mathbb{Z}}\exp\left[\frac{-(\theta_i-\mu-2k\pi)^2}{2\sigma^2}\right]} \quad \text{e} \tag{2.7}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k\in\mathbb{Z}}(\theta_i - \mu - 2k\pi)^2\exp\left[\frac{-(\theta_i-\mu-2k\pi)^2}{2\sigma^2}\right]}{\sum_{k\in\mathbb{Z}}\exp\left[\frac{-(\theta_i-\mu-2k\pi)^2}{2\sigma^2}\right]}}, \tag{2.8}$$

logo, tales funcións $u(\mu, \sigma)$ e $v(\mu, \sigma)$ son os lados dereitos de (2.7) e (2.8), respectivamente. Deste xeito, plantéxase un problema, posto que non se poden computar estas sumas infinitas. Non obstante, como xa se comentou na Sección 1.4.5, a densidade pode ser aproximada por unha cantidade finita de termos, de tal xeito que u e v poden representarse por un número finito de sumandos. Finalmente, ao non poder obter unha expresión explícita da solución, é preciso definir un método iterativo do seguinte xeito: dados dous valores iniciais $\mu_0 \in [0, 2\pi)$ e $\sigma_0 > 0$, defínense μ_{j+1} e σ_{j+1}, para $j = 0, 1, 2, \ldots$ do seguinte xeito

$$\mu_{j+1} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=-t}^{t}(\theta_i - 2k\pi)\exp\left[\frac{-(\theta_i-\mu_j-2k\pi)^2}{2\sigma_j^2}\right]}{\sum_{k=-t}^{t}\exp\left[\frac{-(\theta_i-\mu_j-2k\pi)^2}{2\sigma_j^2}\right]} \quad \text{e}$$

$$\sigma_{j+1} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=-t}^{t}(\theta_i - \mu_j - 2k\pi)^2\exp\left[\frac{-(\theta_i-\mu_j-2k\pi)^2}{2\sigma_j^2}\right]}{\sum_{k=-t}^{t}\exp\left[\frac{-(\theta_i-\mu_j-2k\pi)^2}{2\sigma_j^2}\right]}},$$

para algún enteiro t. Este proceso continúa ata que se obteña a converxencia.

2.3. Estimación da distribución Cauchy enrolada

Para a estimación dos parámetros da Cauchy enrolada é preciso facer unha transformación na súa función de densidade, expresada en (1.22). Ao denotar por $\mathbf{x} = (\cos\theta, \operatorname{sen}\theta)'$ e $\boldsymbol{\mu} = \frac{2\rho}{1+\rho^2}(\cos\mu, \operatorname{sen}\mu)'$, (1.22) pode expresarse como

$$f(\theta; \mu, \rho) = \frac{\sqrt{1 - ||\boldsymbol{\mu}||^2}}{2\pi(1 - \boldsymbol{\mu}'\mathbf{x})}. \tag{2.9}$$

Nótese que $||\boldsymbol{\mu}|| = \frac{2\rho}{1+\rho^2} < 1, \quad \forall\rho \geq 0$, polo que (2.9) está ben definido. Así, dada unha mostra aleatoria simple, $\theta_1, \theta_2, \ldots, \theta_n$, a función de verosimilitude é

$$\mathcal{L} = \prod_{k=1}^{n} f_{(\mu,\rho)}(\theta_k) = (1 - ||\boldsymbol{\mu}||^2)^{\frac{n}{2}} \frac{1}{(2\pi)^n} \prod_{k=1}^{n} \frac{1}{1 - \boldsymbol{\mu}'\mathbf{x}_k},$$

onde $\mathbf{x}_k = (\cos\theta_k, \operatorname{sen}\theta_k)$. Así, a log-verosimilitude vén dada por:

$$l = \frac{n}{2}\log(1 - ||\boldsymbol{\mu}||^2) - n\log(2\pi) - \sum_{k=1}^{n}\log(1 - \boldsymbol{\mu}' x_k).$$

As ecuacións de verosimilitude están dadas por $\frac{\partial l}{\partial \boldsymbol{\mu}} = 0$ e resultan ser

$$\sum_{k=1}^{n} \omega_k (x_k - \boldsymbol{\mu}) = 0,$$

onde $\omega_k = \frac{1}{1-\boldsymbol{\mu}'\mathbf{x}_k}$. Así, o estimador $\widehat{\boldsymbol{\mu}}$ pode obterse mediante o seguinte método iterativo. Consideremos un par inicial $\boldsymbol{\mu}_0$ cumprindo que $||\boldsymbol{\mu}_0|| < 1$. Defínese entón $\boldsymbol{\mu}_{\nu+1}$ como

$$\boldsymbol{\mu}_{\nu+1} = \frac{\sum_{k=1}^{n} \omega_{k,\nu}\mathbf{x}_k}{\sum_{k=1}^{n} \omega_{k,\nu}}, \quad \nu = 0, 1, 2, \ldots,$$

onde os pesos $\omega_{k,\nu+1}$ están definidos en cada paso por

$$\omega_{k,\nu+1} = \frac{1}{1 - \boldsymbol{\mu}_k'\mathbf{x}_k}.$$

Este algoritmo converxe a unha única solución das ecuacións de verosimilitude, poñamos $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \widehat{\mu}_2)'$. Así, os estimadores de máxima verosimilitude dos parámetros μ e ρ da distribución de Cauchy enrolada están dados por:

$$\widehat{\mu} = \operatorname{atan2}(\widehat{\mu}_2, \widehat{\mu}_1) \quad \text{e} \quad \widehat{\rho} = \frac{1 - \sqrt{1 - ||\widehat{\boldsymbol{\mu}}||^2}}{||\widehat{\boldsymbol{\mu}}||}.$$

É preciso especificar que para valores pequenos de ρ, os resultados deste algoritmo son pouco fiables (ver [11]).

2.4. Estimación da mestura de von Mises

Consideremos a función de densidade da mestura de von Mises, dada en (1.25). Para a estimación dos parámetros é preciso obter a función de verosimilitude. Para iso, consideremos unha mostra aleatoria simple, $\theta_1, \theta_2, \ldots, \theta_n$, procedentes dunha mestura de von Mises. Deste xeito, a función de verosimilitude vén dada por

$$\mathcal{L} = \prod_{i=1}^{n} f_{(\boldsymbol{\mu},\boldsymbol{\kappa},\mathbf{p})}(\theta_i) = \prod_{i=1}^{n} \sum_{m=1}^{M} p_m f(\theta_i | \mu_m, \kappa_m, p_m),$$

onde $\Theta = (\theta_1, \dots, \theta_n)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_M)$ é un vector que contén todas as medias, $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_M)$, os parámetros de concentración e, finalmente, $\mathbf{p} = (p_1, \dots, p_M)$ os parámetros de mestura involucrados nas funcións de densidade anteriores. Así, a log-verosimilitude é

$$l = \log \mathcal{L} = \sum_{i=1}^{n} \log \left(\sum_{m=1}^{M} p_m f(\theta_i | \mu_m, \kappa_m, p_m) \right). \tag{2.10}$$

Non obstante, existe un problema na estimación, posto que non se sabe de que mestura procede cada un dos θ_i. Deste xeito, sería apropiado atopar unha forma de, cando sexa posible, expresar (2.10) de manera que en cada sumando todos os ángulos procedan da mesma mestura. Para isto, [15] propón considerar un vector $\mathbf{Z} = (Z_1, \dots, Z_n)$ cumprindo que Z_i toma o valor m cando θ_i está xerado pola mestura $m-$ésima. Deste xeito, pode escribirse (2.10) como

$$l = \sum_{i:Z_i=1} \log(p_1 f_1(\theta_i | \mu_1, \kappa_1, p_1)) + \dots + \sum_{i:Z_i=M} \log(p_M f_M(\theta_i | \mu_M, \kappa_M, p_M)). \tag{2.11}$$

Nótese que en cada sumando, o valores de μ_i e κ_i son diferentes, polo que a estimación destes parámetros lévase a cabo do seguinte xeito: sexa $j \in \{1, \dots, M\}$, as ecuacións de verosimilitude para os parámetros μ_j e κ_j están dadas por:

$$\begin{aligned}
0 &= \sum_{i:Z_i=j} \frac{\partial}{\partial \mu_j} \log(p_j f_j(\theta_i | \mu_j, \kappa_j, p_m)) = \sum_{i:Z_i=j} \frac{\frac{\partial}{\partial \mu_j} f_j(\theta_i | \mu_j, \kappa_j, p_m)}{f_j(\theta_i | \mu_j, \kappa_j, p_m)} \\
&= \kappa_j \sum_{i:Z_i=j} \operatorname{sen}(\theta_i - \widehat{\mu}_j) \quad \text{e} \\
0 &= \sum_{i:Z_i=j} \frac{\partial}{\partial \kappa_j} \log(p_j f_j(\theta_i | \mu_j, \kappa_j, p_m)) = \sum_{i:Z_i=j} \frac{\frac{\partial}{\partial \kappa_j} f_j(\theta_i | \mu_j, \kappa_j, p_m)}{f_j(\theta_i | \mu_j, \kappa_j, p_m)} \\
&= \sum_{i:Z_i=j} [\cos(\theta_i - \widehat{\mu}_j) - A(\widehat{\kappa}_j)].
\end{aligned}$$

Así, as estimacións dos parámetros da media e da concentración son, para todo $j \in \{1, \dots, M\}$,

$$\begin{aligned}
\frac{\operatorname{sen} \widehat{\mu}_j}{\cos \widehat{\mu}_j} &= \frac{\sum_{i:Z_i=j} \operatorname{sen}(\theta_i)}{\sum_{i:Z_i=j} \cos(\theta_i)} \Longleftrightarrow \widehat{\mu}_j = \operatorname{atan2}\left(\sum_{i:Z_i=j} \operatorname{sen}(\theta_i), \sum_{i:Z_i=j} \cos(\theta_i) \right) \quad \text{e} \\
A(\widehat{\kappa}_j) &= \frac{1}{n_j} \sum_{i:Z_i=j} \cos(\theta_i - \widehat{\mu}_j) \Longleftrightarrow \widehat{\kappa}_j = A^{-1}\left(\frac{1}{n_j} \sum_{i:Z_i=j} \cos(\theta_i - \widehat{\mu}_j) \right),
\end{aligned}$$

supoñendo que $\sum_{i:Z_i=j}\cos(\theta_i)\neq 0$ e onde n_j é o cardinal do conxunto $\{i : Z_i = j\}$ e a función atan2 foi definida en (1.5). Nótese que ningún dos estimadores dependen do valor do parámetro de mestura. Para a estimación de tales parámetros verase o caso $M = 2$. Por definición, tense que $p_1 + p_2 = 1$, é dicir, pode denotarse $p \equiv p_1$ e $p_2 = 1 - p$, polo que a log-verosimilitude pode escribirse como:

$$l = \sum_{i:Z_i=1} \log\left(pf_1(\theta_i|\mu_1,\kappa_1,p_1)\right) + \sum_{i:Z_i=2} \log\left((1-p)f_2(\theta_i|\mu_2,\kappa_2,p_2)\right).$$

Polo tanto, derivando parcialmente respecto de p, plantexando a condición de punto crítico, $\frac{\partial l}{\partial p} = 0$, e tendo en conta que $n_1 + n_2 = n$, obtense que:

$$\sum_{i:Z_i=1}\frac{1}{p} + \sum_{i:Z_i=2}\frac{-1}{1-p} = 0 \Longleftrightarrow \frac{n_1}{p} = \frac{n_2}{1-p} \Longleftrightarrow \widehat{p} = \frac{n_1}{n}.$$

Como $\frac{\partial^2 l}{\partial p^2} = -\left[\frac{n_1}{p^2} + \frac{n_2}{(1-p)^2}\right] < 0, \quad \forall\, p \in [0,1]$, $\widehat{p}$ é un máximo.

Non obstante, non sempre se coñecerán as compoñentes do vector $\mathbf{Z}$, polo que se denomina variable non observada. Nese caso, será preciso introducir o coñecido como algoritmo EM, que é empregado en situacións deste tipo. Antes de presentar o algoritmo, é preciso saber o número de mesturas, M, do que se compón cada función de densidade, o cal se deu por suposto ata agora. Para iso, introdúcese o criterio de información de Akaike, ou AIC[2], que selecciona o modelo que mellor se axusta á mostra tendo en conta dous factores: recompensa a bondade de axuste e penaliza o número de parámetros. De este xeito, o criterio defínese, como

$$\mathrm{AIC} = 2d - 2\log(L), \tag{2.12}$$

onde d é o número de parámetros e L é o máximo da función de verosimilitude do modelo estimado. O criterio de Akaike considera considera mellor o modelo que teña un menor valor do AIC. Pode verse en [15] como se o tamaño mostral non é sufucientemente grande con respecto ao número de desidades da von Mises que forman a mestura, o criterio non estima ben o número M.

O algoritmo pode inicializarse de dous xeitos diferentes (ver [15]). O que se empregará neste traballo consiste en asociar probabilidades de pertenza dos datos da mostra a uns grupos que estes determinan, chamados centroides. Consideremos unha mostra aleatoria simple, $\theta_1, \theta_2, \ldots, \theta_n$, estando cada un destes estes θ_i xerado por unha mestura de M von Mises. Defínese a semellanza entre dous ángulos calesquera γ_1 e γ_2 como $\mathrm{Sem}(\gamma_1,\gamma_2) = \cos(\gamma_1-\gamma_2)$. Nótese que $\mathrm{Sem}(\gamma_1,\gamma_2) \in [-1,1], \quad \forall\, \gamma_1,\gamma_2 \in [0,2\pi)$. Primeiramente, calcúlase

[2]Este criterio é empregado na asignatura de Modelos de Regresión e Análise Multivariante para a selección de modelos.

a media circular dos ángulos da mostra, a cal será o primeiro centroide da mostra, Ce_1. Como se trata da mestura de M densidades da von Mises, calcularanse os $M-1$ centroides restantes. Sexa $m \in \{2, \dots, M\}$, calcúlase a semellanza entre cada un dos datos da mostra, $\theta_1, \dots, \theta_n$, e os $(m-1)$ centroides xa calculados; de aí a necesidade de definir dun xeito diferente un primeiro centroide. Entón, para cada un dos ángulos, θ_j, se toma o máximo das semellanzas entre θ_j e os $(m-1)$ centroides. O ángulo que devolva o menor dos valores anteriores será o centroide $m-$ésimo, Ce_m.

É preciso agora definir a disimilitude entre dous ángulos γ_1 e γ_2 como $\mathrm{DSem}(\gamma_1, \gamma_2) = 1 - \mathrm{Sem}(\gamma_1, \gamma_2) = 1 - \cos(\gamma_1 - \gamma_2)$, e calcular as disimilitudes entre todo elemento da mostra, θ_j, e os centroides, é dicir,

$$\mathrm{DSem}(\theta_j, Ce_i) = 1 - \cos(\theta_j - Ce_i), \quad \forall\, j \in \{1, \dots, n\} \text{ e } \quad \forall\, i \in \{1, \dots, M\}.$$

Do resultado anterior distinguiranse dous casos: $\mathrm{DSem}(\theta_j, Ce_i) = 0$ e $\mathrm{DSem}(\theta_j, Ce_i) \neq 0$. No primeiro destes casos, o punto da mostra coincide co centroide e se lle asigna probabilidade un a tal centroide e cero a todos os demais. Noutro caso, a probabilidade de pertencer a un grupo representado por un centroide é inversamente proporcional á disimilitude do ángulo co centroide.

Vexamos agora como funciona o algoritmo en si. Lembremos que no caso que se ocupa agora non se coñecen as compoñentes do vector $\mathbf{Z}$. A primeira etapa consiste no cálculo da esperanza da log-verosimilitude na que se supuxo coñecido o vector $\mathbf{Z}$ en función da distribución condicionada da variable non observada aos valores de $\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}$ e $\mathbf{p}$, que suporemos coñecidos a partir da inicialización do algoritmo. Denotemos por $p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p})$ a tal distribución, que vén dada por:

$$p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p}) = \frac{p_m f_m(\theta_i, \mu_m, \kappa_m)}{\sum_{l=1}^{M} p_l f_l(\theta_i, \mu_l, \kappa_l)}. \tag{2.13}$$

En [15] obtense a esperanza de (2.11), que é

$$\sum_{m=1}^{M}\sum_{i=1}^{n} p\left(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p}\right) \log(p_m) + \sum_{m=1}^{M}\sum_{i=1}^{n} p\left(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p}\right) \log f_m(\theta_i|\mu_m, \kappa_m).$$

O seguinte paso consiste na maximización da cantidade anterior en función dos parámetros $\boldsymbol{\mu}, \boldsymbol{\kappa}$ e $\mathbf{p}$. Este algoritmo continúa ata que se obteña a converxencia, que está garantida en [15]. Tal converxencia alcánzase ben cando as diferencias entre o valor maximizado que toma a función de log-verosimilitude en dúas iteracións consecutivas é menor que unha cantidade prefixada, ou ben cando as variacións entre os parámetros en dúas iteracións consecutivas sexan menores que un valor determinado. En [3] probouse que os estimadores,

tras a aplicación do algoritmo EM, para os parámetros μ_m, κ_m e p_m, $\quad \forall\, m \in \{1, \ldots, M\}$, son os seguintes:

$$\begin{aligned}
\widehat{\mu}_m &= \operatorname{atan2}\left(\sum_{i=1}^{n} \operatorname{sen}\theta_i p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p}), \sum_{i=1}^{n} \cos\theta_i p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p})\right), \\
\widehat{\kappa}_m &= A^{-1}\left(\frac{\sum_{i=1}^{n} \cos(\theta_i - \widehat{\mu}_m) p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p})}{\sum_{i=1}^{n} p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p})}\right) \quad \text{e} \\
\widehat{p}_m &= \frac{1}{n}\sum_{i=1}^{n} p(m|\theta_i, \boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{p}).
\end{aligned}$$

A verosimilitude seguindo este proceso é non decrecente, é dicir, unha iteración non empeora o resultado obtido na anterior, como se pode comprobar en [3]. Existen variacións deste algoritmo que poden ser atopadas en [3].

Capítulo 3

Estudo de simulación e ilustración con datos reais

Na primeira parte deste capítulo presentarase un estudo de simulación para a von Mises e as mesturas de varias destas distribucións. O obxectivo é ver as propiedades dos estimadores e comprobar a súa precisión con diferentes medidas, o que permitirá tamén verificar se os resultados teóricos da Sección 2.1 e da Sección 2.4 son acertados ou non. Finalmente, farase un estudo de datos reais, no que se axustará a distribución máis axeitada de entre as expostas no Capítulo 1 a uns conxuntos de datos facendo uso de tests de bondade de axuste e do criterio AIC.

3.1. Simulación de von Mises

Para a simulación no caso dunha única von Mises, obteranse mostras aleatorias de direccións prodecentes dunha $vM(0, \kappa_j)$, onde κ_j pode tomar os valores $\kappa_1 = 0.5$, $\kappa_2 = 1$ ou $\kappa_3 = 2$ e se fixou a media como $\mu = 0$. Farase mediante un proceso iterativo, no que en cada etapa se xerarán n_i valores, onde n_i pode ser $n_1 = 50$, $n_2 = 100$ ou $n_3 = 1000$. Por prover todos os datos dunha distribución von Mises, a estimación da media é, por máxima verosimilitude, a media mostral e a estimación do parámetro de concentración κ, obtense mediante $A^{-1}(R)$, onde R é a lonxitude do vector resultante, -como pode verse na Sección 2.1- valor que pode ser obtido numericamente. Para cada valor de κ_i e de n_i diferentes, este proceso levarase a cabo $B = 500$ veces, de tal xeito que se obterían dous vectores de lonxitude 500, un que almacenaría os valores das estimacións da media e outro que faría o mesmo coas estimacións do parámetro de concentración. Así, denotarase por $\mathbf{m}^{i,j}$ ao primeiro vector, onde i fai referencia ao n_i e j ao valor de κ_j correspondentes; e $\mathbf{c}^{i,j}$ ao segundo, onde se emprega a notación do mesmo xeito que no vector anterior. Desta

forma, as estimacións correspondentes son as seguintes: no caso da media, a estimación $\widehat{\mu}_{i,j}$ será a media circular, definida en (1.11), dos elementos do vector $\mathbf{m}^{i,j}$ e, no caso da concentración,

$$\widehat{\kappa}_{i,j} = \frac{1}{B}\sum_{k=1}^{B} c_k^{i,j},$$

onde $c_k^{i,j}$ fai referencia ao elemento $k-$ésimo do vector $\mathbf{c}^{i,j}$. Os resultados das estimacións de μ e κ preséntanse nas Táboas 3.1 e 3.2, respectivamente. Con respecto ao primeiro parámetro e, en xeral, a medida que aumenta o tamaño mostral, n, a estimación achégase cada vez máis ao valor orixinal do parámetro. Non obstante, para $\kappa = 0.50$ e $\kappa = 2$ co tamaño mostral $n = 100$ parece haber unha anomalía, posto que as estimacións son peores que no caso $n = 50$. Doutra banda, para o parámetro κ, si que se cumpre o esperado: o valor da estimación achégase cada vez máis ao fixado orixinalmente para o parámetro a medida que o valor de n aumenta, independentemente do valor de κ.

$\widehat{\mu}_{(i,j)}$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.0071	0.0134	-0.0010
$n = 100$	0.0149	0.0084	-0.0019
$n = 1000$	0.0015	0.0028	0.0007

Cadro 3.1: Estimacións por Monte Carlo de μ con tamaños mostrais e valores dos parámetros de concentración indicados.

Na Figura 3.1 represéntanse graficamente algúns dos datos xerados, da densidade da von Mises cos parámetros estimados en cada caso e da comparación desta coa densidade teórica correspondente en cada caso, o que nos permite ver canto de próximas están entre elas. Ademais, empréganse os diagramas de rosa, tamén coñecidos como histogramas circulares, que toman unha maior altura cantos máis datos estean no arco correspondente. Neste caso, dividiuse en 15 arcos da mesma lonxitude e, como se pode comprobar nos nove paneis, a maior concentración de datos se produce a redor da dirección cero, o que concorda con que a dirección fixada sexa $\mu = 0$. Ademais, a medida que aumenta o valor do parámetro de concentración, unha maior cantidade de ángulos se situan próximos á media, polo que as diferencias de altura entre as barras próximas a cero e as diametralmente opostas é maior, de tal xeito que algunhas chegan a desaparecer con $\kappa = 2$.

Nestas representacións pode observarse como, a medida que aumenta o número de observacións e fixado calquera valor de κ, a estimación da densidade está máis próxima da densidade teórica, o que suxire que o aumento do valor de n efectivamente mellora a esti-

mación de ambos parámetros, μ e κ. Non obstante, o caso $n = 50$ xa dá boas aproximacións da densidade orixinal.

$\widehat{\kappa}_{(i,j)}$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.5526	1.0568	2.0895
$n = 100$	0.5302	1.0209	2.0360
$n = 1000$	0.5005	1.0001	1.9962

Cadro 3.2: Estimacións por Monte Carlo de κ con tamaños mostrais e valores dos parámetros de concentración indicados.

Preséntanse os diagramas de caixa das compoñentes dos vectores que conteñen as estimacións da media e da concentración dos datos xerados aleatoriamente, en cada un dos nove casos considerados. Os diagramas de caixa correspondentes á estimación da media están representados na Figura 3.2; todos eles teñen unha mediana moi próxima a cero, que é o valor orixinal do parámetro. Ademais, a medida que aumenta o valor de n, diminúen o número de datos atípicos e o rango intercuartílico, polo que os datos están máis próximos, o que se produce porque as estimacións son mellores a medida que aumenta o tamaño mostral. En relación á concentración, os diagramas de caixa represéntanse na Figura 3.3. Neste caso, tamén pode observarse como, a medida que aumenta o valor de n para un valor de κ fixado, existen, en xeral, un menor número de datos atípicos e a mediana está tamén próxima ao valor orixinal do parámetro.

Agora é preciso definir os conceptos de sesgo, varianza e erro cadrático medio circular, introducidos en [19], xa que cambian con respecto ao caso euclídeo.

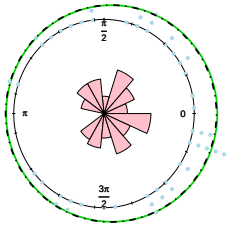
Definición 3.1. Dado un estimador $\widehat{\mu}_0$ dun parámetro μ_0, defínese o erro cadrático medio circular, ECMC, como:

$$\text{ECMC}(\widehat{\mu}_0) = \mathbb{E}(1 - \cos(\widehat{\mu}_0 - \mu_0)).$$

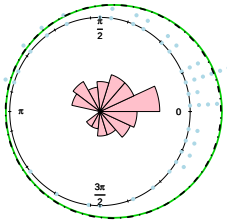
O erro cadrático medio será cero cando $\widehat{\mu_0} = \mu_0$ e toma un valor máximo de dous cando $\widehat{\mu_0} - \mu_0 = \pm\pi$. Aproximarase empíricamente do seguinte xeito:

$$\widehat{\text{ECMC}}(\widehat{\mu}_{i,j}) = \frac{1}{B}\sum_{k=1}^{B}[1 - \cos(m_k^{i,j})],$$

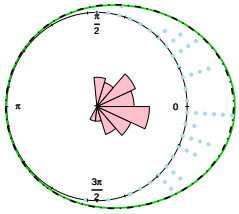
onde $m_k^{i,j}$ fai referencia ao elemento $k-$ésimo do vector $\mathbf{m}^{i,j}$. Como se pode ver na Táboa 3.3, a medida que aumenta o valor do número de observacións, n, o erro cadrático medio



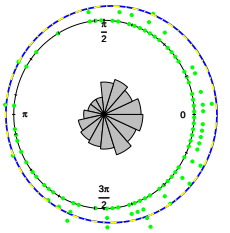
(a) $n = 50$ e $\kappa = 0.50$.



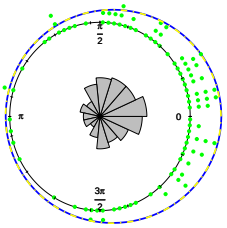
(b) $n = 50$ e $\kappa = 1$.



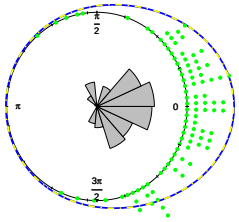
(c) $n = 50$ e $\kappa = 2$.



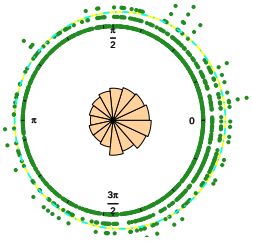
(d) $n = 100$ e $\kappa = 0.50$.



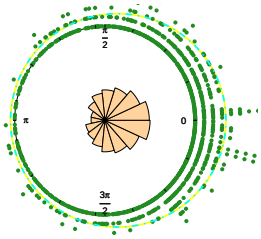
(e) $n = 100$ e $\kappa = 1$.



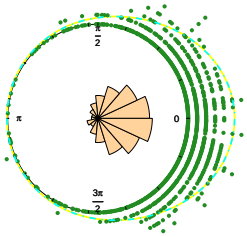
(f) $n = 100$ e $\kappa = 2$.



(g) $n = 1000$ e $\kappa = 0.50$.

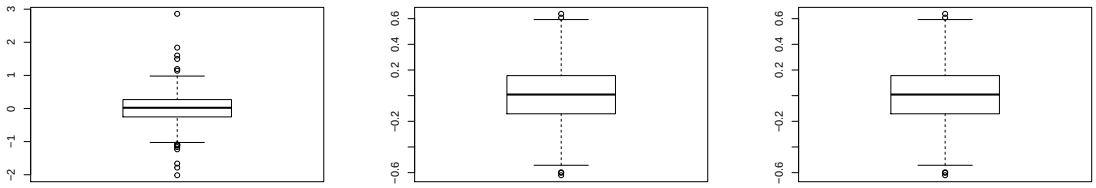


(h) $n = 1000$ e $\kappa = 1$.



(i) $n = 1000$ e $\kappa = 2$.

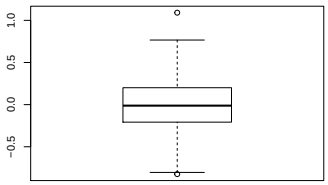
Figura 3.1: Representacións dalgúns conxuntos de datos xerados a partir dunha von Mises coas condicións indicadas baixo elas. Ademais, represéntanse diagramas de rosa nos centros, a liña sólida representa a von Mises cos parámetros estimados en cada caso e a liña de puntos e raias a densidade a partir da cal se obtiveron os datos.



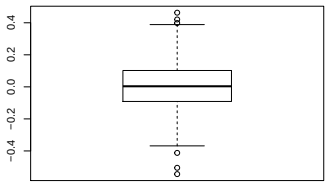
(a) $n = 50$ e $\kappa = 0.50$.

(b) $n = 50$ e $\kappa = 1$.

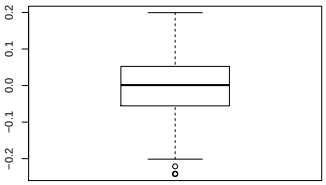
(c) $n = 50$ e $\kappa = 2$.



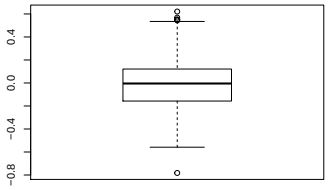
(d) $n = 100$ e $\kappa = 0.50$.



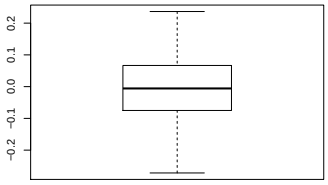
(e) $n = 100$ e $\kappa = 1$.



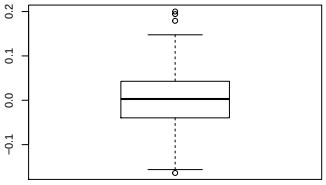
(f) $n = 100$ e $\kappa = 2$.



(g) $n = 1000$ e $\kappa = 0.50$.



(h) $n = 1000$ e $\kappa = 1$.



(i) $n = 1000$ e $\kappa = 2$.

Figura 3.2: Diagramas de caixa das estimacións do parámetro μ en cada un dos casos especificados baixo cada panel.

circular achégase cada vez máis a cero para calquera valor de κ, o que dá unha idea de que as estimacións son cada vez mellores a medida que aumenta o valor de n.

$\widehat{\text{ECMC}}(\widehat{\mu}_{i,j})$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.0981	0.0238	0.0075
$n = 100$	0.0414	0.0125	0.0036
$n = 1000$	0.0039	0.0012	0.0003

Cadro 3.3: Estimacións do erro cadrático medio circular do estimador da media con tamaños mostrais e valores dos parámetros de concentración indicados.

Definición 3.2. A varianza circular, VC, dun estimador $\widehat{\mu}_0$ con dirección media μ defínese como:

$$\text{VC}(\widehat{\mu}_0) = \mathbb{E}(1 - \cos(\widehat{\mu}_0 - \mu)).$$

Aproximarase este valor empíricamente como

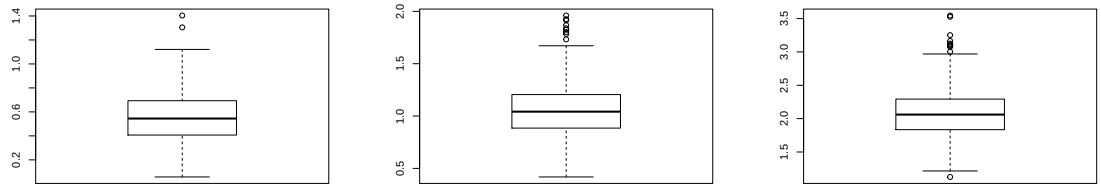
$$\widehat{\text{VC}}(\widehat{\mu}_{i,j}) = \frac{1}{B}\sum_{k=1}^{B}[1 - \cos(m_k^{i,j} - \widehat{\mu}_{i,j})].$$

A medida que aumenta o valor do tamaño mostral, n, e, para calquera valor de κ, o valor da varianza circular achégase a cero, como se pode observar na Táboa 3.4. É preciso aclarar que, aínda que varios dos valores nas Táboas 3.3 e 3.4 coincidan non se trata de que estes sexan iguais, senón que a aproximación a catro decimais coincide no caso do erro cadrático medio circular e da varianza circular. En canto á varianza do estimador κ, que xa é a varianza usual por tratarse dun parámetro real non negativo, pode verse na Táboa 3.5 como se aproxima a cero a medida que aumenta o valor de n, independentemente do valor que tome κ. Esta varianza estimarase do seguinte xeito:

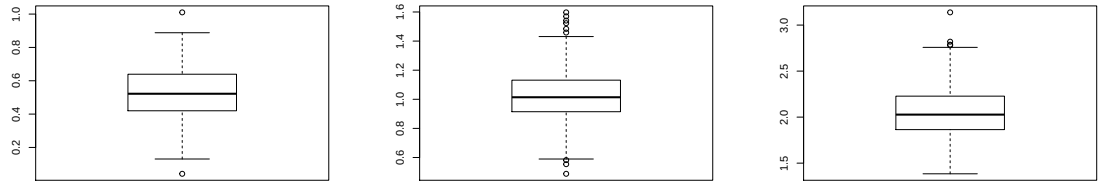
$$\widehat{\text{Var}}(\widehat{\kappa}_{(i,j)}) = \frac{1}{B}\sum_{k=1}^{B}(c_k^{i,j} - \widehat{\kappa}_{i,j})^2.$$

Definición 3.3. Pola súa parte, o sesgo circular dun estimador $\widehat{\mu}_0$ dun parámetro μ_0 defínese como segue:

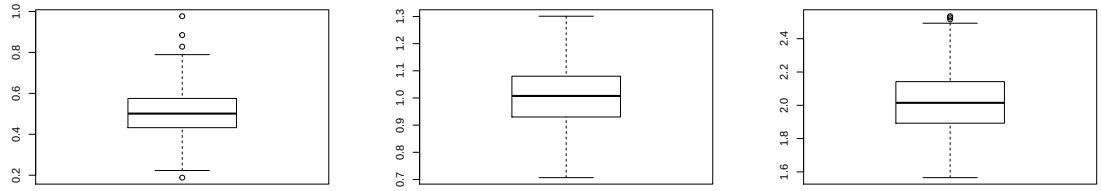
$$SC(\widehat{\mu}_0) = \mathbb{E}(\text{sen}(\widehat{\mu}_0 - \mu_0)).$$



(a) $n = 50$ e $\kappa = 0.50$. (b) $n = 50$ e $\kappa = 1$. (c) $n = 50$ e $\kappa = 2$.



(d) $n = 100$ e $\kappa = 0.50$. (e) $n = 100$ e $\kappa = 1$. (f) $n = 100$ e $\kappa = 2$.



(g) $n = 1000$ e $\kappa = 0.50$. (h) $n = 1000$ e $\kappa = 1$. (i) $n = 1000$ e $\kappa = 2$.

Figura 3.3: Diagramas de caixa das estimacións do parámetro κ en cada un dos casos especificados baixo cada panel.

$\widehat{\mathrm{VC}}(\widehat{\mu}_{(i,j)})$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.0981	0.0237	0.0075
$n = 100$	0.0413	0.0125	0.0036
$n = 1000$	0.0038	0.0012	0.0003

Cadro 3.4: Estimacións da varianza circular do estimador da media, μ, con tamaños mostrais e valores dos parámetros de concentración indicados.

$\widehat{\text{Var}}(\widehat{\kappa}_{(i,j)})$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.0440	0.0652	0.1345
$n = 100$	0.0227	0.0272	0.0646
$n = 1000$	0.0022	0.0030	0.0063

Cadro 3.5: Estimacións da varianza dos estimadores de κ con tamaños mostrais e valores dos parámetros de concentración indicados.

A pesar de cambiar as definicións con respecto ao caso sobre $\mathbb{R}$, en ambos casos cúmprese a seguinte relación:

$$\text{ECMC}(\widehat{\mu_0}) = VC(\widehat{\mu_0}) + (SC(\widehat{\mu_0}))^2.$$

Para o caso que nos ocupa, estimarase o sesgo circular de $\widehat{\mu}_{i,j}$ do seguinte xeito:

$$\widehat{\text{SC}}(\widehat{\mu}_{i,j}) = \frac{1}{B}\sum_{k=1}^{B} \text{sen}(m_k^{i,j}).$$

Os resultados móstranse na Táboa 3.6. Neste caso, para os valores de $\kappa = 1$ e $\kappa = 2$, a medida que aumenta o tamaño mostral, n, o sesgo achégase cada vez máis a cero. Non obstante, para o caso $\kappa = 0.5$, o valor do sesgo é maior cando $n = 100$ que cando $n = 50$, o que parece ser unha anomalía, posto que ocorre precisamente o contrario ao que cabe esperar. En canto ao sesgo do estimador $\widehat{\kappa}$, que de novo resulta ser o sesgo usual, posto que se trata dun parámetro real non negativo, os resultados son como cabería esperarse. Nótese tamén que se o valor do parámetro aumenta, o sesgo é menor. Doutra banda, como pode verse na Táboa 3.7, a medida que aumenta o valor do parámetro n, o sesgo achégase cada vez máis a cero en valor absoluto e xa non aparecen resultados anómalos, coma no caso do estimador da media circular. Este sesgo estímase do seguinte xeito

$$\widehat{\text{sesgo}}(\widehat{\kappa}_{i,j}) = \widehat{\kappa}_{i,j} - \kappa_{i,j}.$$

$\widehat{\text{Sesgo}}(\widehat{\mu}_{i,j})$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.0048	0.0130	-0.0010
$n = 100$	0.0143	0.0083	0.0007
$n = 1000$	0.0015	0.0028	0.0006

Cadro 3.6: Estimacións do sesgo circular do estimador do parámetro μ con tamaños mostrais e valores dos parámetros de concentración indicados.

$\widehat{\text{Sesgo}}(\widehat{\kappa}_{i,j})$	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$
$n = 50$	0.0526	0.0568	0.0895
$n = 100$	0.0302	0.0209	0.0360
$n = 1000$	0.0005	0.0001	-0.0038

Cadro 3.7: Estimacións do sesgo dos estimadores de κ con tamaños mostrais e valores dos parámetros de concentración indicados.

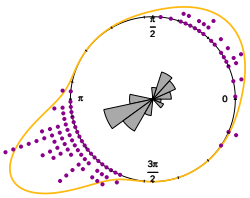
Nos cadros están as estimacións dos valores anterioremente definidos. Máis concretamente, as estimacións da media e da varianza nas Táboas 3.1 e 3.2, respectivamente; o erro cadrático medio circular de $\widehat{\mu}$ na Táboa 3.3; a varianza circular de $\widehat{\mu}$ e $\widehat{\kappa}$ nas Táboas 3.4 e 3.5, respectivamente; e o sesgo circular de $\widehat{\mu}$ e sesgo de $\widehat{\kappa}$, nas Táboas 3.6 e 3.7, respectivamente.

3.2. Simulación de mesturas de von Mises

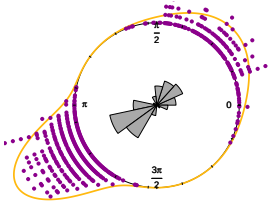
Ocuparase agora a simulación de datos de mesturas de von Mises. Para cada distribución e cada valor $n_1 = 100$, $n_2 = 500$ ou $n_3 = 1000$, fixados, repetirase $B = 500$ veces o seguinte proceso: simúlanse n_i valores de ángulos procedentes da distribución a tratar e, coñecido o número de funcións que interveñen na mestura, estímanse os parámetros a partir deses datos. Unha vez feito isto, estimaranse o parámetro da media como a media circular das estimacións obtidas e os parámetros de concentración e de mestura como a media das estimacións obtidas en cada etapa. Este proceso é análogo ao empregado na Sección 3.1 cunha única von Mises. As distribucións que se tratarán serán as seguintes:

$$
\begin{aligned}
D_1 &= 0.65vM\left(\frac{7\pi}{6}, 14\right) + 0.35vM\left(\frac{\pi}{4}, 5\right), \\
D_2 &= 0.55vM\left(\pi, 8\right) + 0.45vM\left(0, 12\right), \\
D_3 &= 0.60vM\left(\frac{7\pi}{4}, 12\right) + 0.30vM\left(\frac{2\pi}{3}, 9\right) + 0.10vM\left(\frac{\pi}{6}, 5\right) \quad \text{e} \\
D_4 &= 0.35vM(5, 15) + 0.30vM(4, 10) + 0.2vM(\pi, 7.5) + 0.15vM(2, 5).
\end{aligned}
$$

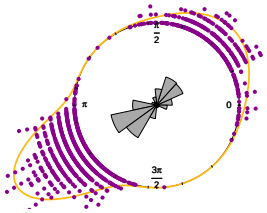
As estimacións dos parámetros da media preséntanse na Táboa 3.8, as da concentración na Táboa 3.9 e as dos parámetros de mestura na Táboa 3.10. Nelas apréciase como a medida que aumenta o número de compoñentes que interveñen na mestura, a precisión das estimación é peor. Isto faise notar na estimación $\hat{\boldsymbol{\kappa}}$ para a distribución D_4, na que os valores



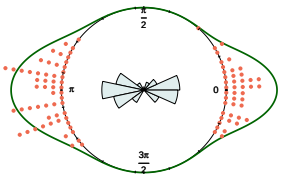
(a) $n = 100$ para D_1.



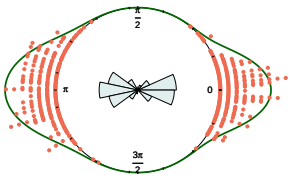
(b) $n = 500$ para D_1.



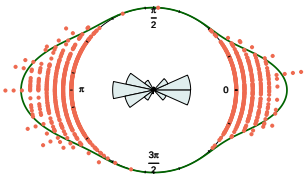
(c) $n = 1000$ para D_1.



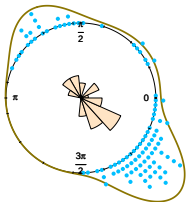
(d) $n = 100$ para D_2.



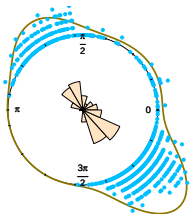
(e) $n = 500$ para D_2.



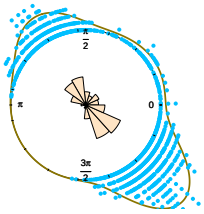
(f) $n = 1000$ para D_2.



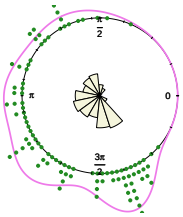
(g) $n = 100$ para D_3.



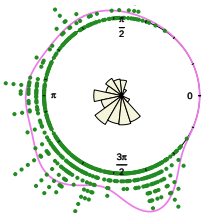
(h) $n = 500$ para D_3.



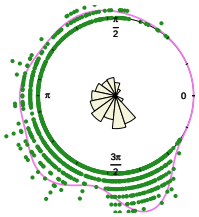
(i) $n = 1000$ para D_3.



(l) $n = 100$ para D_4.



(m) $n = 500$ para D_4.



(n) $n = 1000$ para D_4.

Figura 3.4: Representación dalgúns datos obtidos por simulación da mestura de von Mises, xunto coa súa densidade teórica e un diagrama de rosa.

estimados distan considerablemente do valor orixinal do parámetro debido á complexidade do modelo. Tamén, como era de prever, canto maior é o tamaño da mostra que se obtén, mellores son, en xeral, as estimacións. Finalmente, cabe destacar que na estimación dos parámetros de mestura xa se obteñen valores próximos ao orixinal do parámetro para as mesturas con dúas compoñentes cando o número de valores xerados aleatoriamente é $n = 100$, o cal non ocorre con D_3 nin D_4. Na primeira delas, para $n = 1000$, as estimacións son bastante precisas, a diferencia de D_4, cuxas estimacións desvíanse máis dunha centésima en case todas as compoñentes. Algúns exemplos de datos xerados represéntanse na Figura 3.4.

$\widehat{\boldsymbol{\mu}}$	$n = 100$	$n = 500$	$n = 1000$
D_1	(0.6665,0.7911)	(3.6655,0.7870)	(3.6660,0.7864)
D_2	(3.1416,0.0019)	(3.1428,0.0001)	(3.1426,0.0003)
D_3	(5.5627, 2.5618, 0.7762)	(5.5025, 2.1240, 0.5214)	(5.4979, 2.0987, 0.5233)
D_4	(5.0254, 4.1664, 3.3324, 2.0614)	(5.0141, 4.1054, 3.3043, 2.1699)	(5.0155, 4.1031, 3.3314, 2.0299)

Cadro 3.8: Estimacións por Monte Carlo de $\boldsymbol{\mu}$ nos catro modelos considerados e cos tamaños mostrais indicados. En cada caso obtense un número de estimacións da media igual ao número de compoñentes do modelo.

$\widehat{\boldsymbol{\kappa}}$	$n = 100$	$n = 500$	$n = 1000$
D_1	(14.5675,5.4138)	(14.1548,5.1003)	(14.0519,5.0432)
D_2	(12.7709,8.3479)	(12.1275,8.0492)	(12.0343,8.0273)
D_3	(27.4549, 11.1133, 5.4450)	(12.7073, 9.6984, 5.3505)	(12.1933, 9.3807, 5.1938)
D_4	(53.9771, 16.9006, 8.4516, 4.4301)	(19.1311, 11.1025, 6.2177, 4.1243)	(17.8402, 10.8059, 5.7296, 4.0576)

Cadro 3.9: Estimacións por Monte Carlo de $\boldsymbol{\kappa}$ nos catro modelos considerados e cos tamaños mostrais indicados. En cada caso obtense un número de estimacións da media igual ao número de compoñentes do modelo.

$\widehat{\boldsymbol{p}}$	$n = 100$	$n = 500$	$n = 1000$
D_1	(0.6495,0.3505)	(0.6491, 0.3509)	(0.6494,0.3506)
D_2	(0.4378,0.5622)	(0.4501,0.5499)	(0.4502,0.5498)
D_3	(0.5768, 0.2965, 0.1267)	(0.5935, 0.2989, 0.1076)	(0.5976, 0.2978, 0.1046)
D_4	(0.3875, 0.2808, 0.2012, 0.1306)	(0.3684, 0.2887, 0.2096, 0.1333)	(0.3633, 0.2904, 0.2101, 0.1363)

Cadro 3.10: Estimacións por Monte Carlo dos parámetros de mestura nos catro modelos considerados e cos tamaños mostrais indicados. En cada caso obtense un número de estimacións da media igual ao número de compoñentes do modelo.

$\widehat{\mathrm{ECMC}}(\hat{\boldsymbol{\mu}})$	$n = 100$	$n = 500$	$n = 1000$
D_1	(0.0005, 0.0033)	(0.0001, 0.0006)	$(5.5178 \times 10^{-5}, 3.2493 \times 10^{-4})$
D_2	(1.9988, 1.9990)	(1.9998, 1.9998)	$(9.4186 \times 10^{-5}, 1.1916 \times 10^{-4})$
D_3	(0.0179, 0.2749, 0.1597)	(0.0016, 0.016, 0.0159)	$(9.7500 \times 10^{-5}, 2.6810 \times 10^{-4}$, 0.0045)
D_4	(0.1215, 1.3244, 0.0823, 0.0579)	(0.0008, 0.0337, 0.0507, 0.0172)	(0.0006, 0.0318, 0.0509, 0.0157)

Cadro 3.11: Estimacións por Monte Carlo do erro cadrático medio circular do estimador $\hat{\boldsymbol{\mu}}$ nos catro modelos considerados e cos tamaños mostrais indicados. En cada caso obtense un número de estimacións da media igual ao número de compoñentes do modelo.

Na Táboa 3.11 preséntase as estimacións do erro cadrático medio circular do estimador $\hat{\boldsymbol{\mu}}$. Para todas as distribucións cúmprese que a maior número de valores xerados, menor é o erro cadrático medio circular, o dá unha idea de que canto maior é o valor de n mellor será a estimación da media de cada distribución que forma parte da mestura, o cal se pode comprobar directamente na Táboa 3.9. Non obstante, é preciso apreciar que a maior número de elementos que conforman a mestura, peor é o resultado obtido, xa que os modelos son máis complicados.

$\widehat{\text{ECM}}(\hat{\boldsymbol{\kappa}})$	$n = 100$	$n = 500$	$n = 1000$
D_1	(7.5183, 1.7938)	(1.2772, 0.3211)	(0.6003, 0.1329)
D_2	(7.9015, 2.2584)	(1.3096, 0.4189)	(0.6559, 0.2112)
D_3	(10766.61, 15.1785, 8.4133)	(5.2760, 2.2688, 5.0590)	(0.9036 1.0386, 3.6103)
D_4	(17888.47, 598.21, 18.8977, 4.5422)	(48.0237, 22.1213, 8.3215, 2.2062)	(25.8394 22.3378, 8.6066, 2.0110)

Cadro 3.12: Estimacións por Monte Carlo do erro cadrático medio do estimador $\hat{\boldsymbol{\kappa}}$ nos catro modelos considerados e cos tamaños mostrais indicados. En cada caso obtense un número de estimacións da media igual ao número de compoñentes do modelo.

Na Táboa 3.12 reflíctense os resultados da estimación do erro cadrático medio do estimador de κ para cada compoñente de cada mestura. De novo, a maior valor de n, menor é o valor do erro cadrático medio. Non obstante, para as distribucións D_3 e, sobre todo, D_4, os valores que se obteñen non son satisfactorios. Por unha banda, os resultados para $n = 100$ son esaxeradamente grandes e, por outra parte, os relativos a $n = 1000$ non son tan bos como cabería esperar, a pesar de mellorar considerablemente os obtidos para $n = 100$.

$n = 100$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
D_1	0	391	85	19	5
D_2	0	416	60	17	7
D_3	0	15	402	57	26
D_4	0	56	325	127	23

Cadro 3.13: Cantidade de veces que foi seleccionado, polo método Monte Carlo e segundo o criterio AIC, o número de compoñentes que forman a mestura para as distribucións indicadas, con tamaño mostral $n = 100$ en cada simulación.

$n = 1000$	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
D_1	0	463	27	8	2
D_2	0	466	27	5	2
D_3	0	0	474	20	6
D_4	0	0	64	361	75

Cadro 3.14: Cantidade de veces que foi seleccionado, polo método Monte Carlo e segundo o criterio AIC, o número de compoñentes que forman a mestura para as distribucións indicadas, con tamaño mostral $n = 1000$ en cada simulación.

Nas estimacións anteriores era coñecido o número de distribucións involucradas nas mesturas. Non obstante, pode determinarse o número de mesturas que se axustan mellor a un conxunto de datos. Para iso, fixada unha das distribucións, D_1, D_2, D_3 ou D_4, levaranse a cabo 500 simulacións de datos e seleccionarase, mediante o criterio AIC o número, M, de compoñentes que interveñen nas mesturas, en cada simulación. Nas Táboas 3.13 e 3.14, preséntanse o número de veces que se seleccionou cada número de mesturas; na primeira para o tamaño mostral $n = 100$ e na segunda para $n = 1000$. Para os modelos D_1 e D_2, que son ambas mesturas de dúas compoñentes, os resultados son bastante claros tanto para o tamaño mostral $n = 100$, no que se selecciona o número correcto nun 78.2 % e 83.2 % dos casos, respectivamente; como para $n = 1000$, que devolve o resultado correcto nun 92.6 % e 93.2 % das veces. Ocorre o mesmo para D_3, cuxos resultados son acertados no 80.4 % dos casos para $n = 100$ e 94.8 % para o tamaño mostral $n = 1000$, polo que son concluíntes. Finalmente, para D_4 e $n = 100$ non se obtén o agardado nun principio, xa que selecciona 3 compoñentes nun 65 % dos casos, cando realmente a forman 4. Isto débese á pouca cantidade de datos xerados en cada iteración en comparación co número de

compoñentes que forman a mestura, o que contrasta co caso $n = 1000$, no que os resultados se corresponden coa realidade nun 72.2 % das veces.

3.3. Ilustración con datos reais

Nesta sección búscase axustarlle a un conxunto de datos algunha das distribucións das estudadas no Capítulo 1. Estes conxuntos de datos foron obtidos do software `R`, aínda que foron presentados orixinalmente en diferentes publicacións, as cales se especificarán na descrición de cada un deles.

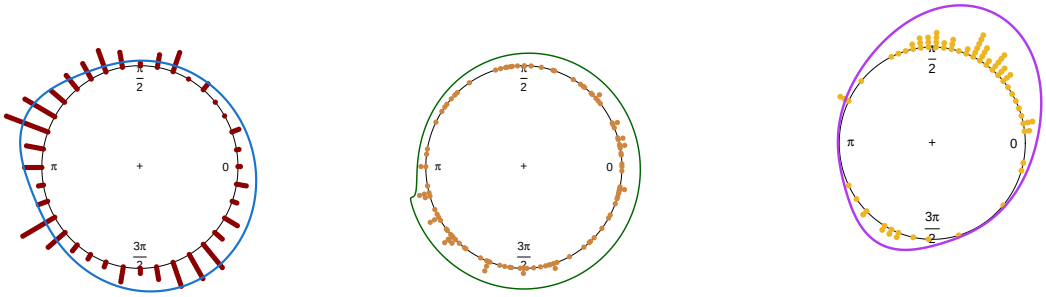
As estimacións do número de compoñentes que interveñen na mestura e dos parámetros destas permiten apreciar como engadir só unha distribución máis á mestura fai que o modelo sexa moito máis complexo. Isto pode comprobarse nas Táboas 3.11 e 3.12, onde os erros cometidos nas estimacións son moito maiores nas distribucións D_3 e D_4, sobre todo nesta última. A situación é aínda peor cando o número de datos é baixo para a cantidade de compoñentes da mestura, como se puido ver na Táboa 3.13. Por iso, aínda que un maior número de compoñentes poida axustarse mellor aos datos, o número de parámetros penaliza moito a elección destas distribucións se este é elevado. Por ese motivo, nesta sección empregarase o criterio AIC, que escolle o modelo que teña maior equilibrio entre o axuste e o número de parámetros.

Para determinar se un conxunto de datos pode ser modelado por unha distribución pode empregarse o test de bondade de axuste dunha distribución, neste caso, a von Mises, a normal enrolada e a uniforme, o cal se detalla no Apéndice B. Os $p-$valores para os test de bondade de axuste para as diferentes distribucións e os conxuntos de datos figuran na Táboa 3.16. A primeira compoñente fai referencia ao test de Rayleigh e a segunda ao test de Kuiper. Non obstante, pode darse que un conxunto de datos poidan ser modelados por diferentes distribucións, o cal dificultará cal elixir. Para iso, farase uso do AIC, cuxos valores para algunhas das distribucións descritas ao longo do traballo e para os distintos conxuntos de datos tratados están presentes na Táboa 3.16. Nela, na columna referida á mestura de von Mises, representarase o AIC da mestura que mellor modele os datos e a columna relativa á Batschelet fai referencia á mestura da distribución de Batschelet inversa con base a von Mises, cuxa densidade vén dada en (1.27). Todos os datos que se estudan, xunto coa distribución que mellor se axusta a eles están representados na Figura 3.5.

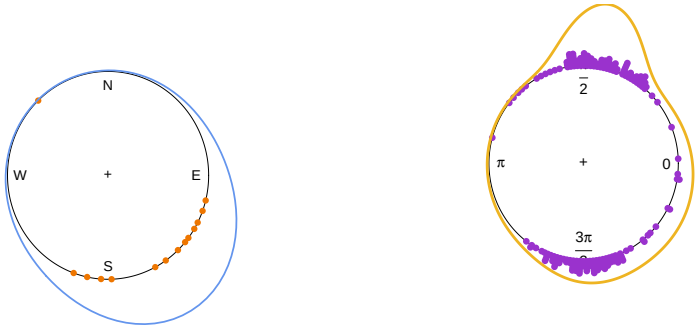
O primeiro conxunto, o referido como Dirección do vento da Figura 3.5, fai referencia a 239 medicións da dirección do vento na illa de Saturna, Canadá, entre os dias 1 e 10

Test de bondade de axuste	Uniforme	vM	wN
Dir. vento	$(< 0.01, < 0.01)$	$(< 0.01, < 0.01)$	$(< 0.01, < 0.01)$
Andoriñas	$(0.1314, < 0.05)$	$(0.11, 0.16)$	$(0.13, 0.16)$
Tartarugas	$(< 0.01, < 0.01)$	$(0.001, 0.006)$	$(< 0.01, < 0.01)$
Ras	$(0.0002, < 0.01)$	$(0.375, 0.510)$	$(0.12, 0.13)$
Libélulas	$(0.052, < 0.01)$	$(0.02, 0.02)$	$(0.02, 0.02)$

Cadro 3.15: $p-$valores dos test de bondade de axuste.



(a) Dirección do vento (b) Andoriñas (c) Tartarugas



(d) Ras (e) Libélulas

Figura 3.5: Representación dos datos tratados xunto cos modelos que, segundo o criterio AIC, mellor axustan os datos.

AIC	Mestura de vM	Uniforme	wN	wC	Batschelet
Dir. Vento	**836.24**	878.51	895.77	862.53	852.31
Andoriñas	418.96	419.04	585.52	419.45	**416.39**
Tartarugas	**221.20**	279.36	302.88	230.50	234.30
Ras	**37.85**	51.46	41.69	68.06	39.15
Libélulas	**509.22**	786.61	1587.52	786.33	705.86

Cadro 3.16: Valores de AIC obtidos para os distintos conxuntos de datos e as distribucións indicadas. Resáltase en negriña o menor valor para cada conxunto de datos.

de outubro de 2016. Por isto, os datos teñen dependencia temporal, aínda que non se terá en conta porque se busca que sirvan de ilustración. Estes datos foron presentados en [9] e están dispoñibles na libraría [5]. Os $p-$valores do test de bondade de axuste figuran na primeira fila da Táboa 3.15. Estes valores levan a rexeitar as tres distribucións que se presentan aos niveis de significación usuais. Para ver cal se emprega faise uso do criterio AIC, que devolve os datos da primeira fila da Táboa 3.16. Segundo este criterio, o modelo que mellor se axusta dos que aquí se consideran é unha mestura de dúas von Mises, en particular, $0.20vM(2.76, 7.89) + 0.80vM(5.33, 0.94)$, seleccionándose o número de compoñentes da mestura mediante o criterio AIC.

O segundo conxunto de datos, chamado andoriñas, fai referencia ás direccións de migración de 114 andoriñas no outono. Estes datos aparecen orixinalmente en [8] e foron obtidos da libraría [2]. Os $p-$valores dos test de bondade de axuste para estes datos son os presentes na segunda fila da Táboa 3.15. Estes datos non deixan claro os resultados sobre a uniformidade e non hai evidencias significativas para rexeitar que os datos proveñan dunha von Mises ou dunha normal enrolada para os niveis de significación usuais. Segundo o criterio AIC, cuxos valores para os modelos considerados están presentes na segunda fila da Táboa 3.16, o modelo máis axeitado é unha Batschelet inversa con base a distribución de von Mises con parámetros $\widehat{\mu} = -0.32$, $\widehat{\kappa} = 0.48$, $\widehat{\nu} = 1$ e $\widehat{\lambda} = 0.84$.

O terceiro conxunto de datos consiste en 76 observacións da dirección de movemento de tartarugas despois de someterse a un tratamento. Estes datos foron presentados orixinalmente en [18] e están dispoñibles na libraría [7]. Os $p-$valores do test de bondade de axuste preséntanse na terceira fila da Táboa 3.15, a partir dos cales se rexeitan as tres distribucións consideradas e os niveis de significación usuais, é dicir, 0.10, 0.05 e 0.01. Os valores do AIC para algunhas distribucións que se estudaron no traballo están presentados na terceira fila da Táboa 3.16. Entón, o modelo que mellor representa os datos sería unha mestura de von Mises, máis concretamente, $0.16vM(4.21, 8.44)+0.84vM(1.11, 2.62)$, sendo

o número de mesturas seleccionado mediante o criterio AIC.

O cuarto conxunto de datos consiste en 14 direccións de movemento de ras que foron postas en liberdade tras estar 30 horas encerradas. Estes datos foron obtidos da libraría [2] e preséntanse orixinalmente en [6]. Os $p-$valores do test de bondade de axuste están representados na cuarta fila da Táboa 3.15. A partir deles, rexéitase a uniformidade e pero non hai evidencias significativas para rexeitar a von Mises nin a normal enrolada. Para decidir cal distribución empregar, faise uso do AIC. Os resultados están presentes na cuarta fila da Táboa 3.16. Así, a densidade que mellor se axusta a este conxunto de datos é $vM(2.55, 2.17)$.

O derradeiro conxunto de datos consiste na orientación de 214 libélulas con respecto ao acimut. Estes datos preséntanse orixinalmente en [4] e están dispoñibles na libraría [14]. Os $p-$valores do test de bondade de axuste para as distribucións consideradas reflíctense na última fila da Táboa 3.15, a partir dos cales rexéitase que os datos poidan modelarse por unha uniforme, unha única von Mises ou unha normal enrolada, cun nivel de significación do 10 %. Segundo o criterio AIC, que para este conxunto de datos pode verse na derradeira fila da Táboa 3.16, o modelo máis axeitado corresponde á mestura de tres von Mises; máis concretamente: $0.43vM(1.41, 13.90)+0.48vM(4.73, 8.04)+0.09vM(5.63, 1.32)$, seleccionándose o número de compoñentes da mestura mediante o criterio AIC.

Capítulo 4

Apéndices

A. Algúns modelos discretos

Ao longo do traballo foron só tratadas as distribucións absolutamente continuas sobre a circunferencia. Non obstante, tanto a distribución uniforme discreta como a distribución de Poisson enrolada son suficientemente importantes como para ser mencionadas. Así, descríbense ambas a continuación e alguhas das súas propiedades máis importantes.

A.1. Distribución uniforme discreta

Consideremos unha distribución con

$$P_r\left(\theta = \nu + \frac{2\pi r}{m}\right) = p_r, \quad r \in \{1, 2, \dots, m-1\},$$

onde $p_r \geq 0$ e $\sum_{r=0}^{m-1} p_r = 1$. Todos os puntos da forma $\nu + \frac{2\pi r}{m}$ están sobre unha circunferencia de raio 1, logo pode asociárselle a cada un unha probabilidade. Se $p_i = p_k, \forall\, i, j \in \{1, 2, \dots, m-1\}$, entón, pola condición de que a suma dos pesos terá que ser un, é claro que $p_r = \frac{1}{m}$. Se $\nu = 0$, entón a circunferencia sobre a que están os puntos está centrada na orixe de coordenadas do plano. Neste caso particular, a función característica é a seguinte:

$$\varphi_p = \sum_{r=0}^{m-1} p_r e^{\frac{2rp\pi i}{m}}.$$

Deste xeito, $\varphi_0 = 1$ se $p = 0 \pmod m$. A función característica cumpre que $\varphi_p = 1$ se $p = 0 \pmod m$ e $\varphi_p = 0$ noutro caso (ver [13]).

A.2. Poisson enrolada

Esta distribución pertence ao grupo das distribucións obtidas a partir do envolvemento sobre a circunferencia dalgunha distribución definida na recta real, que foron introducidas

na Sección 1.4.4. Nelas, considerouse un agrupamento dos datos en módulo 2π. Logo, se $m \in \mathbb{Z}$, as reducións módulo $2\pi m$, transforman todos os valores a elementos do grupo alxébrico das raíces m-ésimas da unidade, que pode ser pensado como un subgrupo da circunferencia de radio un, posto que o módulo de todos eses números complexos será un. En analoxia con (1.17), defínese $\theta = 2\pi x \pmod{2\pi m}$. A partir da definición de raíz m-ésima dun número complexo, correspóndense cos seguintes ángulos $\{\frac{2\pi r}{m} | r = 0, \dots, m-1\}$. A función masa de probabilidade vén dada por:

$$P_r\left(\theta = \frac{2\pi r}{m}\right) = \sum_{k\in\mathbb{Z}} p(r + km),$$

onde p representa á función masa de probabilidade da variable sobre a recta real. Así, se esta variable é unha Poisson de media λ, a expresión anterior sería a seguinte:

$$P_r\left(\theta = \frac{2\pi r}{m}\right) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{r+km}}{(r+km)!}, \quad r \in \{0, \dots, m-1\}.$$

A expresión anterior admite a seguinte representación finita (ver [16]):

$$P_r\left(\theta = \frac{2\pi r}{m}\right) = \frac{1}{m} \sum_{k=0}^{m} e^{\varphi^k \lambda} \varphi^{-rk},$$

onde φ é unha raíz $m-$ésima da unidade. En virtude da segunda propiedade das variables enroladas, sendo coñecido que a función característica da Poisson é $\varphi(t) = \exp\{\lambda(e^{it} - 1)\}$ e que no caso circular só toma valores no conxunto $\{\frac{2\pi r}{m} | r = 0, \dots, m-1\}$, entón a función característica será

$$\varphi_p = \exp\{\lambda(1 - e^{\frac{2\pi i p}{m}})\}.$$

Por último, é coñecido que, no caso sobre a recta real, a distribución de Poisson é reprodutiva con respecto á súa media, λ. Ocorre o mesmo coa Poisson circular grazas á propiedade 1 das variables enroladas.

B. Test de bondade de axuste

O obxectivo deste test é saber se hai evidencias significativas que indiquen que os datos non proceden da distribución que se sospeite. Centrarase a explicación para a von Mises, pero o procedemento que se empregará pode ser estendido a outras distribucións, como a Normal enrolada. Para iso, será preciso coñecer primeiramente algún test para a uniformidade da distribución, posto que os da von Mises se derivan directamente deles. Os que serán empregados neste caso serán o test de Rayleigh e o test Kuiper, pero outros, como o test U^2 de Wilson ou o test de espaciamento de Rao, poden consultarse en [13].

Primeiramente, o test de Rayleigh baséase na lonxitude do vector resultante, definida en (1.8). Se este módulo é maior que un valor limiar, a distribución considérase demasiado concentrada para ser uniforme, polo que se rexeita a uniformidade. Máis concretamente, en [13] móstrase que, para valores grandes de n,

$$2nR^2 \sim \chi^2_2,$$

onde a notación $\sim$ fai referencia que $2nR^2$ segue unha distribución χ^2_2. Polo tanto, se o valor observado de $2nR^2$ supera o cuantil da χ^2_2 para o nivel de significación elixido, rexéitase a uniformidade.

Por outra banda, o test de Kuiper está baseado no test de Kolmogorov-Smirnov para distribucións sobre a recta real. Será preciso entón definir a función de distribución empírica para datos circulares. Consideremos unha mostra ordeada $0 \leq \theta_{(1)}, \theta_{(2)}, \dots, \theta_{(n)} < 2\pi$, entón, a función de distribución empírica defínese como

$$S_n(\theta) = \frac{i}{n}, \quad \text{se } \theta_{(i)} \leq \theta < \theta_{(i+1)}.$$

Deste xeito, tanto $S_n(\theta)$ coma $F(\theta)$, que fai referencia á función de distribución da uniforme, dada por $F(\theta) = \frac{\theta}{2\pi}$, dependen do punto tomado coma orixe e da orientación escollida. A partir do test de Kolmogorov-Smirnov, defínense

$$D_n^+ = \sup_\theta \{S_n(\theta) - F(\theta)\} \quad \text{e} \quad D_n^- = \inf_\theta \{S_n(\theta) - F(\theta)\}.$$

Rexeitarase entón que os datos proveñan dunha distribución uniforme cando

$$\max_n \{D_n^+, D_n^-\}$$

toma un valor grande. Non obstante, sería apropiado atopar un xeito de contrastar a uniformidade no que non se dependa da orixe e direccións fixadas. Definindo

$$V_n = D_n^+ + D_n^-$$

satisfaise as condicións desexadas, como se pode ver en [13]. A hipótese nula de uniformidade rexéitase cando o valor do estatístico V_n toma valores grandes. Tamén se cumpre que é invariante ante rotacións, o cal se proba en [13].

Como xa se comentou, estes test darán lugar, de xeito case inmediato, a test para a comprobación da procedencia dunha von Mises dos datos que se manexen. Cómpre mencionar antes de introducir estes test que existen xeitos gráficos de ter unha idea do resultado que poden devolver, pero coma no caso de variables aleatorias reais, os métodos gráficos son certamente subxectivos e non proporcionan un criterio apropiado. Estes gráficos son o $P-P$ plot e o $Q-Q$ plot[1]. No primeiro deles, represéntase o valor da función de distribución empírica fronte á distribución da von Mises nos puntos considerados. Por outra parte, no segundo represéntanse os cuantiles empíricos fronte aos cuantiles da distribución da von Mises.

En [16] introdúcese o seguinte método: primeiramente, axústase un modelo de von Mises aos datos. A partir deste axuste, os valores dos estatísticos dos test que se levan a cabo, o de Rayleigh e o de Kuiper, son calculados para os valores de $2\pi\widehat{F}(\theta_1), 2\pi\widehat{F}(\theta_2), \dots, 2\pi\widehat{F}(\theta_n)$, onde $\widehat{F}$ é a función de distribución da von Mises cos parámetros estimados no paso anterior. Posteriormente, xéranse unha cantidade de datos prodecentes da von Mises cos parámetros estimados igual ao número orixinal de valores e se repite o proceso anterior B veces, sendo este valor escollido antes de que comece o proceso; canto maior sexa este, máis preciso será o resultado obtido. Finalmente, o $p-$valor de cada test estímase a partir da proporción dos $(B+1)$ estatísticos maiores ou iguais que o que corresponde ao conxunto inicial de datos. Esta idea pode estenderse a calquera distribución que se desexe contrastar, sendo necesario poder estimar os parámetros para aplicar o procedemento.

[1] Este método é empregado nas asignaturas de Inferencia Estatística e Modelos de Regresión e Análise Multivariante para a diagnose dun modelo de regresión, en particular, para contrastar, tamén dun xeito pouco rigoroso, a normalidade dos erros dun modelo de regresión.

Bibliografía

[1] Abe, T. e Pewsey, A., Sine-skewed circular distributions, *Statistical Papers*, **52** (2011), 683–707.

[2] Agostinelli, C. e Lund, U., `R` package `circular`: Circular Statistics (version 0.4-93), 2017, URL: https://r-forge.r-project.org/projects/circular/.

[3] Banerjee, A., Dhillon, I.S., Ghosh, J. e Sra, S., Clustering on the unit hypersphere usin von Mises-Fisher distributions, *Journal of Machine Learning Research*, **6** (2005), 1345-1382.

[4] Batschelet, E., *Circular Statistics in Biology*, Academic Press, New York, 1981.

[5] Chakraborty, S. e Wong, S.W.K., (2019) BAMBI: Bivariate Angular Mixture Models. R package version 2.2.0. URL: https://CRAN.R-project.org/package=BAMBI.

[6] Collett, D., Outliers in Circular Data, *Applied Statistics*, **29** (1980), 50-57.

[7] Fernández Durán, J.J. e Gregorio Domínguez, M.M., CircNNTSR: An R Package for the Statistical Analysis of Circular, Multivariate Circular, and Spherical Data Using Nonnegative Trigonometric Sums, *Journal of Statistical Software*, **70(6)** (2016), 1-19.

[8] Giunchi, D. e Baldaccini N., Orientation of juvenile barn swallows (Hirundo rustica) tested in Emlen funnels during autumn migration, *Behavioral Ecology and Sociobiology*, **56** (2004), 124-131.

[9] Histórico do tempo atmosférico do Goberno de Canadá.

[10] Hornik K. e Grün, B., movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions, *Journal of Statistical Software*, **58(10)**, 2014, 1-31, URL: https://doi.org/10.18637/jss.v058.i10.

[11] Jammalamadaka, S.R. e SenGupta, A., *Topics in Circular Statistics*, World Scientific, Singapore, 2001.

[12] Ley, C. e Verdebout, T. *Modern Directional Statistics*, Chapman and Hall/CRC Press, Boca Raton, 2017.

[13] Mardia, K.V. e Jupp, P.E., *Directional Statistics*, Wiley , New York, 1999.

[14] Maria Oliveira, Rosa M. Crujeiras e Alberto Rodriguez Casal, NPCirc: An R Package for Nonparametric Circular Methods. *Journal of Statistical Software*, **61(9)**, 2014, 1-26. URL: http://www.jstatsoft.org/v61/i09/.

[15] Oliveira Pérez, M., *Nonparametric Circular Methods for Density and Regression*. Universidade de Santiago de Compostela, Santiago de Compostela, 2013.

[16] Pewsey, A., Neuhäuser, M. e Ruxton, G.D., *Circular Statistics in R*, Oxford University Press, New York, 2013.

[17] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020, URL: http://www.r-project.org/.

[18] Stephens, M.A., Techniques for directional data, *Technical Report 150*, 1969.

[19] Sungsu K. e SenGupta, A., Multivariate-multiple Circular Regression, *Journal of Statistical Computation and Simulation*, **87** (2017), 1277-1291.