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

ANALIZANDO AS ESTRATEXIAS DE ESCAPE EN LARVAS DE PEIXE CEBRA MEDIANTE REGRESIÓN MULTIMODAL CIRCULAR

María Alonso-Pena[1] e Rosa M. Crujeiras[1]

[1]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela

RESUMO

Analizar a dirección de escape de animais en función de covariables é un problema que require de técnicas estatísticas alén dos métodos de regresión clásicos. Ademais da periodicidade do ángulo de escape, que esixe a utilización de métodos para variables circulares, os datos de escape animal adoitan requirir da exploración das direccións preferentes, en lugar da dirección media ou esperada. Neste traballo propoñemos o uso dun método non paramétrico para estimar as modas condicionais locais nun experimento con peixes cebra, dende a perspectiva dos modelos de regresión. Presentaremos os algoritmos de estimación e investigaremos o comportamento asintótico do estimador, así como o seu funcionamento con datos simulados. A nova metodoloxía é empleada para modelar o comportamento de escape dun grupo de larvas de peixe cebra cando fuxen dun depredador robot. De xeito máis xeral, o enfoque proposto neste traballo pode aplicarse a moitos outros problemas relativos ao comportamento animal ou a outros campos.

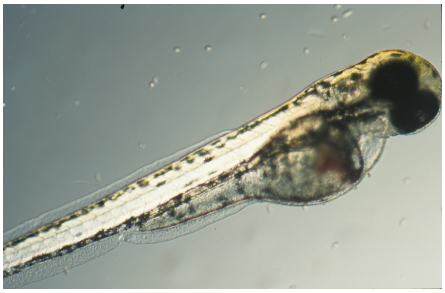
Palabras e frases chave: Regresión multimodal; Datos circulares; Estimación tipo núcleo; Escapoloxía animal; Regresión non paramétrica

1. INTRODUCIÓN

Existe unha ampla literatura no campo da bioloxía que trata o estudo da elección de orientación ou comportamento de escape en animais, e a influencia de covariables na resposta de escape (Scapini et al., 2002; Marchetti e Scapini, 2003; Card e Dickinson, 2008; Obleser et al., 2016; Sato et al., 2019). Un caso interesante é a análise das direccións de escape en animais cando fuxen dos seus depredadores e, aínda que a lóxica suxire que os animais deberían fuxir en dirección oposta aos seus depredadores, este non é sempre o caso, e cómpre analizar as respostas de escape. Máis especificamente, estamos interesados nas direccións de escape de larvas de peixes cebra (*Danio rerio*), como as que se mostran no panel esquedo da Figura 1, e como o ángulo no que o depredador se aproxima aos peixes afecta ao comportamento á hora de escapar. Os datos, obtidos de Nair et al. (2017a), conteñen as direccións de escape dun grupo de larvas de peixe cebra retidos nun acuario e os ángulos nos que un robot imitando un depredador se aproximaron a cada peixe. O panel dereito da Figura 1 mostra un esquema do experimento.

Cando se trata de analizar este tipo de datos con métodos de regresión clásicos, podemos atoparnos dous problemas: i) a periodicidade da variable determinando a dirección de escape, que é, por definición, unha variable circular e ii) a necesidade de estimar as direccións máis verosímiles, condicionadas a distintas covariables, en lugar da dirección media ou esperada (condicional).

A teoría clásica sobre datos circulares (observacións definidas sobre a circunferencia unidade) está presente dende fai décadas (Fisher, 1993; Mardia e Jupp, 2000; Jammalamadaka e SenGupta, 2001), pero o seu uso na práctica foi limitado debido á falta de observacións circulares precisas. Os avances tecnolóxicos fixeron posible o rexistro preciso deste tipo de datos, incrementando o interés no campo da estatística circular nos últimos anos (Pewsey et al., 2013; Ley e Verdebout, 2017). Ademais de en bioloxía, podemos atopar datos circulares en moitos outros ámbitos: xeoloxía (SenGupta e Rao, 1966), ciencias ambientais e oceanografía (Oliveira et al., 2013), medicina



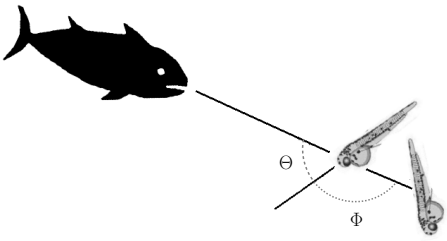


Figura 1: Esquerda: fotografía dunha larva de peixe cebra de Wikimedia Commons (2008). Dereita: esquema do experimento, onde Θ indica a dirección na que se aproxima o depredador e Φ a dirección de escape (elaboración propia a partir da imaxe de Wikimedia Commons (2014)).

(Mooney et al. 2003) ou ecoloxía (Ameijeiras-Alonso et al., 2019). A peculiar natureza deste tipo de observacións pon de manifesto a necesidade de crear ferramentas inferenciais específicas máis aló de aquelas pensadas para datos na recta real.

Como no caso que nos ocupa, no que temos outra variable influíndo na dirección de escape, podemos estudar os datos circulares dende a perspectiva da regresión. Dependendo da natureza da covariable, podemos distinguir dous escenarios distintos: se a covariable é escalar, podemos representar a curva de regresión na superficie dun cilindro, no que a altura do cilindro indica a magnitude da covariable e o ángulo representa o valor da resposta circular, como se representa nos paneis superiores da Figura 2. Por outra banda, se a natureza da covariable é tamén circular, a curva de regresión pódese representar na superficie dun toro, como se mostra nos paneis inferiores da Figura 2. Distintas propostas de modelos de regresión paramétricos involucrando variables circulares poden atoparse en Jammalamadaka e SenGupta (2001). Porén, estes modelos poden non ser o suficientemente flexibles para modelar relacións máis complexas entre as variables, polo que os modelos non paramétricos preséntanse como unha boa alternativa. Di Marzio et al. (2012) propuxeron un método tipo núcleo para regresión con resposta circular, no que a estimación vén dada polo suavizado das compoñentes seno e coseno da resposta.

Os métodos anteriormente citados consideran a media condicional como a función a estimar. Non obstante, como se expuxo no punto ii), o enfoque clásico de *regresión á media* pode non ser axeitado en casos onde a densidade condicional é multimodal. A Figura 2 presenta datos simulados de modelos de regresión con resposta circular nos que a densidade condicional da resposta sobre a explicativa é bimodal. Nos casos correspondentes aos paneis esquerdos da Figura 2, a media condicional ou función de regresión *usual* non está nin sequera definida, dado que as densidades condicionais consideradas son bimodais e simétricas, non estando a media circular definida neste caso. No seu lugar, as modas condicionais locais preséntanse como unha mellor alternativa para modelar a relación entre as variables, resumindo os valores condicionais *máis probables* en lugar da esperanza condicional. Esta idea lévanos a chamada *regresión multimodal*, na que no lugar dunha función, o obxectivo é estimar unha multifunción ou función de avaliación múltiple.

A idea de estimar as modas locais da densidade condicional no contexto euclídeo foi primeiramente introducida por Scott (1992). Einbeck e Tutz (2006) propuxeron a versión condicional do algoritmo mean shift (Fukunaga e Hostetler, 1975; Cheng, 1995; Comaniciu e Meer, 2002) para estimar a multifunción de regresión baseándose nun estimador tipo núcleo da densidade condicional. As propiedades teóricas deste estimador foron estudadas por Chen et al. (2016) e estes modelos foron estendidos ao contexto de datos con erros de medición por Zhou e Huang (2016). Unha revisión recente da regresión multimodal con variables escalares pode atoparse en Chen (2018). O algoritmo mean shift foi xeneralizado ao caso de variables direccionais por Oba et al. (2005) no contexto do clustering non paramétrico, e foi tamén estudado por Kobayashi e Otsu (2010) nese mesmo contexto. A converxencia do algoritmo e outros resultados teóricos foron obtidos recentemente por Zhang e Chen (2020).

O obxectivo deste traballo é introducir o método da regresión multimodal non paramétrica

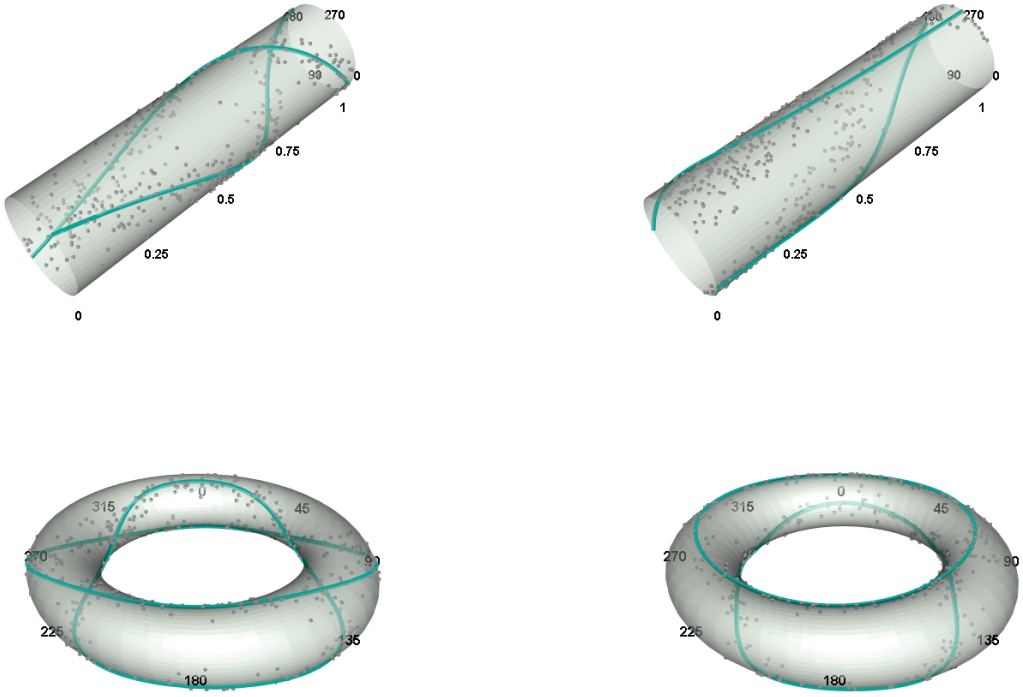


Figura 2: Representacións no cilindro e no toro de datos simulados e as verdadeiras multifuncións de regresión para os modelos models LC-1 (arriba e esquerda), LC-2 (arriba e dereita) con concentración $\tau = 8$; e modelos CC-1 abaixo e esquerda) e CC-2 (abaixo e dereita) con concentración $\tau = 12$. Os tamaños mostrais son $(n_1, n_2) = (200, 200)$ para todos os modelos.

circular para analizar situacións onde as modas locais condicionais son un mellor representante da relación entre as variables que a función de regresión clásica. A nosa meta é empregar esta técnica para estudar como a dirección de escape dos peixes cebra está influenciada polo ángulo no que se aproxima o depredador.

A estrutura do documento é a que segue: na Sección 2 presentamos o escenario da regresión multimodal circular e detallamos os algoritmos de estimación. Algúns resultados teóricos enúncianse na Sección 3, mentres que o problema da selección dos parámetros de suavizado estúdase na Sección 4. A Sección 5 mostra un estudo de simulación no que se analiza o comportamento dos estimadores na práctica. Finalmente, a Sección 6 contén a análise dos datos dos peixes cebra coa metodoloxía proposta.

2. O USO DO MEAN SHIFT CIRCULAR NA REGRESIÓN MULTIMODAL

O obxectivo da regresión multimodal non paramétrica é estimar a densidade condicional da variable resposta sobre a variable explicativa e, despois, computar as modas locais condicionais co algoritmo mean shift (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu e Meer, 2002). Nesta sección detallaremos o algoritmo de estimación deseñado para estimar a multifunción de regresión multimodal no contexto onde a variable resposta é de natureza circular e onde a variable explicativa pode ser tanto circular como escalar.

Considérese unha variable resposta circular, Φ, con soporte na circunferencia unidade, $\mathbb{T} = (-\pi, \pi]$ e unha variable explicativa que pode ser tanto unha variable escalar X con soporte en

$\Omega \subset \mathbb{R}$ ou unha explicativa circular con soporte en $\mathbb{T}$. Denotaremos por Δ unha variable explicativa xenérica, con soporte na recta real ou na circunferencia unidade. Sexa $\{(\Delta_j, \Phi_j)\}_{j=1}^n$ unha mostra aleatoria de (Δ, Φ). Para modelar a relación entre a explicativa e a resposta, consideramos a multifunción de regresión modal (Einbeck e Tutz, 2006) que, para cada δ está definida como o conxunto de modas locais (ou máximos locais) da función de densidade condicional:

$$M(\delta) = \left\{ \phi : \frac{\partial}{\partial \phi} f(\phi|\delta) = 0, \quad \frac{\partial^2}{\partial \phi^2} f(\phi|\delta) < 0 \right\}, \tag{1}$$

onde $f(\phi|\delta)$ é a densidade condicional de Φ dado o valor de Δ. A estimación de M lévase a cabo mediante un enfoque indirecto: primeiramente, estimamos a densidade condicional e, despois, calcúlanse as modas locais condicionaiss. Para a estimación da densidade condicional $f(\phi|\delta)$ utilizaremos un estimador tipo núcleo (Di Marzio et al, 2016). Se a variable explicativa é escalar ($\Delta = X$), o estimador vén dado por

$$\hat{f}(\phi|x) = \frac{\sum_{j=1}^n L_h\left(x - X_j\right) K_\kappa\left(\phi - \Phi_j\right)}{\sum_{j=1}^n L_h\left(x - X_j\right)}.$$

Neste caso $L_h(\cdot)$ é unha función núcleo *linear* ou usual, con ancho de banda h e $K_\kappa(\cdot)$ é unha función núcleo circular con parámetro de concentración κ. Se a explicativa é circular ($\Delta = \Theta$), estimaremos a densidade condicional como

$$\hat{f}(\phi|\theta) = \frac{\sum_{j=1}^n K_\nu\left(\theta - \Theta_j\right) K_\kappa\left(\phi - \Phi_j\right)}{\sum_{j=1}^n K_\nu\left(\theta - \Theta_j\right)},$$

onde o núcleo asociado á explicativa ten ν como parámetro de concentración e o núcleo asociado a Φ ten concentración κ. De agora en adiante denotaremos os pesos correspondentes á variable explicativa (X ou Θ) no punto δ como $w_\delta(\Delta_j)$, $j = 1, ..., n$. Nótese que estes pesos dependen dun parámetro de concentración h ou ν (dependendo da natureza de Δ). Así,

$$\hat{f}(\phi|\delta) = \frac{1}{n\hat{f}(\delta)} \sum_{j=1}^n w_\delta(\Delta_j) K_\kappa(\phi - \Phi_j), \quad \hat{f}(\delta) = \frac{1}{n} \sum_{j=1}^n w_\delta(\Delta_j).$$

En consecuencia, o estimador da multifunción de regresión modal (1) vén dado por

$$\widehat{M}(\delta) = \left\{ \phi : \frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = 0, \quad \frac{\partial^2}{\partial \phi^2} \hat{f}(\phi|\delta) < 0 \right\}. \tag{2}$$

Asumiremos que a función núcleo circular asociada á variable resposta satisfai

$$K_\kappa(\cdot) = c_\kappa K[\kappa(1 - \cos(\cdot))], \tag{3}$$

onde c_κ é unha constante normalizadora que depende de κ. A densidade de von Mises é un exemplo de núcleo circular satisfacendo esta condición. Para obter os máximos locais de $\hat{f}(\delta|\phi)$, establecemos a condición necesaria de punto crítico: $\frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = 0$. Polo tanto, se aplicamos (3), temos que

$$\frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = \frac{\kappa c_\kappa}{n\hat{f}(\delta)} \sum_{j=1}^n w_\delta(\Delta_j) K'[\kappa(1 - \cos(\phi - \Phi_j))] \sin(\phi - \Phi_j). \tag{4}$$

Por conseguinte, a derivada do estimador da densidade condicional con respecto a ϕ é unha suma poderada dos senos das diferencias de cada observación ao punto ϕ. Neste caso, a función seno utilízase para medir a variación ou diferenzas entre as observacións e o punto ϕ. Isto é bastante intuitivo dado que se $\phi = \Phi_j$, entón $\sin(\phi - \Phi_j) = 0$. Ademais, se a diferenza $\phi - \Phi_j$ é moi pequena, entón $\sin(\phi - \Phi_j) \approx \phi - \Phi_j$. Expandindo o último factor no lado dereito de (4), obtemos

$$\frac{\partial}{\partial \phi} \hat{f}(\phi|\delta) = \frac{\kappa c_\kappa}{n\hat{f}(\delta)} \sum_{j=1}^n w_\delta(\Delta_j) K'[\kappa(1 - \cos(\phi - \Phi_j))](\sin\phi\cos\Phi_j - \cos\phi\sin\Phi_j),$$

e igualándoo a cero, temos

$$\sin\phi\sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\cos\Phi_j = \cos\phi\sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\sin\Phi_j,$$

onde $T(\cdot) = c_T K'[\kappa(1-\cos(\cdot))]$. Polo tanto, se denotamos

$$S_\delta(\phi) = \sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\sin\Phi_j \quad \text{e} \quad C_\delta(\phi) = \sum_{j=1}^{n} w_\delta(\Delta_j)T(\phi-\Phi_j)\cos\Phi_j,$$

temos que se $S_\delta(\phi) \neq 0$ ou $C_\delta(\phi) \neq 0$, entón $\phi = \text{atan2}\left(S_\delta(\phi), C_\delta(\phi)\right)$, onde o operador $\text{atan2}(a,b)$ devolve o ángulo entre o eixo das x e o vector que vai da orixe a (b,a) (see Jammalamadaka e SenGupta, 2001, Capítulo 1). Deste xeito, obtemos que o estimador modal $\phi_m \equiv \phi_m(\delta)$ vén dado por

$$\phi_m = \tilde{\omega}(\phi_m) = \text{atan2}\left(S_\delta(\phi_m), C_\delta(\phi_m)\right).$$

Nótese que a función $\tilde{\omega}(\phi)$ devolve unha media circular ponderada das observacións, dado que $S_\delta(\phi)$ é unha suma ponderada de $\sin\Phi_j$ e $C_\delta(\phi)$ é unha suma ponderada de $\cos\Phi_j$ (onde os pesos dependen do punto (δ,ϕ)). Como na anterior expresión temos unha ecuación de punto fixo, utilizamos un algoritmo tipo mean shift para obter o estimador da moda condicional. Definimos a función mean shift circular como

$$\tilde{m}(\phi) = \sin(\tilde{\omega}(\phi) - \phi).$$

Como se comentou anteriormente, dado que para valores pequenos de $\tilde{\omega}(\phi)-\phi$ temos $\sin(\tilde{\omega}(\phi)-\phi) \approx \tilde{\omega}(\phi)-\phi$, a función seno utilízase para medir a variación de ϕ a $\tilde{\omega}(\phi)$. Ademais, para a moda local do estimador da densidade condicional, a función mean shift circular toma o valor cero. En consecuencia, a multifunción de regresión estimada (2) obtense utilizando o procedemento mean shift circular, que se describe no Algoritmo 1. Nótese que o número de puntos iniciais, p, pode ser diferente para cada valor de δ, e para inicializar o algoritmo en rexións próximas aos datos, recomendamos unha inicialización local onde, para cada valor δ, os valores iniciais son os cuartís circulares (véxase Fisher, 1993) das observacións da varriable resposta máis próximas a δ.

Algoritmo 1: Mean shift condicional circular

Datos: Mostra $\{(\Delta_i,\Phi_i\}_{i=1}^n$, parámetros de suavizado κ e h/ν.

1. Inicializar puntos da malla $\mathcal{S} \subset \Omega$ se $\Delta = X$ ou $\mathcal{T} \subset \mathbb{T}$ se $\Delta = \Theta$.
2. Para cada $\delta \in \mathcal{S}$ (ou $\delta \in \mathcal{T}$), seleccionar valores iniciales $\phi_0^{(1)}(\delta), ..., \phi_0^{(p)}(\delta)$.
3. Para $k = 1, ..., p$ iterar ata alcanzar converxencia:

$$\phi_{l+1}^{(k)} = \text{atan2}\left(\sum_{i=1}^{n} w_\delta(\Delta_i)T(\phi_l^{(k)}-\Phi_i)\sin\Phi_i, \sum_{i=1}^{n} w_\delta(\Delta_i)T(\phi_l^{(k)}-\Phi_i)\cos\Phi_i\right),$$

con $l = 0, 1, ...$

3. ALGUNHAS CONSIDERACIÓNS TEÓRICAS

O obxectivo desta sección é dar taxas de converxencia asintótica do estimador non paramétrico para a regresión multimodal introducido na Sección 2. Nótese que as métricas de erro usuais en regresión tipo núcleo (como o Erro Cadrático Medio Integrado ou o Erro Cadrático Integrado) non son axeitadas para medir a calidade do estimador (2) dado que, no contexto da regresión multimodal, o noso obxectivo é estimar a multifunción (1) e, polo tanto, para cada valor da variable explicativa hai, posiblemente, un conxunto de valores da variable resposta. En consecuencia, seguiremos o traballo de Chen et al. (2016) e consideraremos erros puntuais e globais baseados nunha distancia entre conxuntos.

En primeiro lugar, presentamos a distancia de Hausdorff que, para dous conxuntos $A, B \subset \mathbb{R}^q$ defínese como

$$\text{Haus}(A,B) = \max\left\{\sup_{x\in A} d(x,B), \sup_{x\in B} d(x,A)\right\}, \tag{5}$$

onde $d(x, A) = \inf_{z\in A} ||x - z||$. Esta distancia mide como de próximos están dous conxuntos definidos en espazos euclídeos. Na regresión multimodal para variables escalares, a distancia de Hausdorff utilízase para medir a distancia da verdadeira multifunción á súa versión estimada (para un punto fixado da variable explcativa). Porén, no caso que nos ocupa, $M(\delta)$ e $\widehat{M}(\delta)$ son subconxuntos de $\mathbb{T}$ e, polo tanto, é necesario xeneralizar esta distancia considerando, para $A, B \subset \mathbb{T}$,

$$\widetilde{\text{Haus}}(A, B) = \max\left\{ \sup_{x\in A} \tilde{d}(x, B), \sup_{x\in B} \tilde{d}(x, A) \right\}, \tag{6}$$

con $\tilde{d}(x, A) = \inf_{z\in A} 1 - \cos(x - z)$. Utilizaremos a distancia definida en (6) para construír unha medida de erro puntual na regresión multimodal circular. Definimos o erro puntual como

$$\tilde{\Lambda}(\delta) = \widetilde{\text{Haus}}(M(\delta), \widehat{M}(\delta)). \tag{7}$$

O erro puntual mide como de próximas están a multifunción de regresión verdadeira e a estimada para cada posible valor δ da variable explicativa. Para obter unha medida de erro global, introducimos o Erro Medio Circular Integrado modal (EMCI_m), definido como

$$\text{EMCI}_m(\widehat{M}) = \mathbb{E}\left[\int_{\delta\in\text{Supp}(\Delta)} \tilde{\Lambda}(\delta) d\delta \right],$$

onde $\text{Supp}(\Delta)$ denota o soporte de Δ que, recordemos pode denotar tanto unha variable escalar como unha variable circular. Esta medida é o valor esperado do erro puntual integrado, e está baseado na versión integrada do chamado *Erro Cadrático Medio Circular*, introducido po Kim e SenGupta (2017) como o análogo circular do Erro Cadrático Medio en variables escalares. As seguintes proposicións mostran a consistencia dos estimadores.

Proposición 1. *Sexa X unha variable explicativa escalar e Φ unha variable resposta circular. Considérese o estimador da regresión multimodal en (2). Baixo certas condicións de regularidade, tense*

$$\tilde{\Lambda}(x) = O(h^2 + \kappa^{-1}) + O_P\left(\sqrt{\frac{\kappa^{3/2}}{nh}}\right)$$

e

$$EMCI_m(\widehat{M}) = O(h^2 + \kappa^{-1}) + O\left(\sqrt{\frac{\kappa^{3/2}}{nh}}\right),$$

cando $\kappa \to \infty$, $h \to 0$ e $nh\kappa^{-(1+2r)/2} (\log n)^{-1} \to \infty$.

Proposición 2. *Sexa Θ unha variable explicativa circular e Φ unha variable resposta circular. Considérese o estimador da regresión multimodal en (2). Baixo certas condicións de regularidade, tense*

$$\tilde{\Lambda}(\theta) = O(\nu^{-1} + \kappa^{-1}) + O_P\left(\sqrt{\frac{\kappa^{3/2}\nu^{1/2}}{n}}\right)$$

e

$$EMCI_m(\widehat{M}) = O(\nu^{-1} + \kappa^{-1}) + O\left(\sqrt{\frac{\kappa^{3/2}\nu^{1/2}}{n}}\right),$$

cando $\kappa \to \infty$, $\nu \to \infty$ e $n\kappa^{-(1+2r)/2}\nu^{-1/2} (\log n)^{-1} \to \infty$.

Os resultados previos mostran que os erro puntuais $\tilde{\Delta}(x)$ e $\tilde{\Delta}(\theta)$ converxen a cero coa mesma taxa de converxencia que a primeira derivada parcial, con respecto a variable resposta, dos estimadores tipo núcleo da densidade conxunta, $\frac{\partial}{\partial\phi}\hat{f}(x, \phi)$ e $\frac{\partial}{\partial\phi}\hat{f}(\theta, \phi)$.

4. A SELECCIÓN DOS PARÁMETROS DE SUAVIZADO

Como adoita suceder en estimación tipo núcleo, a selección dos parámetros de suavizado resulta crucial en regresión multimodal. No contexto da regresión tipo núcleo clásica na recta real (Fan e Gijbels, 1996), un valor alto da fiestra h produce un estimador sobresuvizado, mentres que un valor pequeno de h implica unha estimación infrasuavizada da función de regresión. Pola contra, o comportamento do parámetro de concentración presente na regresión tipo núcleo circular (Di Marzio et al. 2009, 2012) é inverso: cando a concentración κ e grande, obtemos un estimador infrasuavizado, mentres que un valor pequeno de κ produce unha versión sobresuavizada do estimador.

Así e todo, no contexto da regresión multimodal son necesarios dous parámetros de suavizado: un asociado á variable explicativa e outro asociado á variable resposta. O rol que desempeñan estes parámetros na estimación da multifunción de regresión é moi diferente. O parámetro asociado á variable explicativa controla o suavizado do estimador da multifunción, xogando un papel similar ao do parámetro de suavizado na regresión tipo núcleo clásica. Pola contra, o parámetro asociado á variable resposta inflúe no número de modas estimadas. A razón detrás deste comportamento é que un estimador infrasuavizado da densidade condicional dará lugar a moitas modas locais estimadas, producindo un alto número de *ramas* estimadas da multifunción.

A literatura estatística sobre a selección dos parámetros de suavizado para regresión multimodal no contexto de variables escalares é relativamente escasa. Dado que a estimación se leva a cabo mediante a obtención dos máximos locais da densidade condicional, Einbeck e Tutz (2006) recomendan utilizar métodos deseñados para a estimación da densidade condicional. Non obstante, estes métodos poden non ser idóneos na práctica, dado que a estimación da moda está relacionada pero non é equivalente á estimación da densidade. Como apuntan Casa et al. (2020), unha estimación da densidade pode estar próxima á verdadeira densidade en termos do Erro Cadrático Integrádo, pero ter moitas modas esrimadas que, no caso da regresión multimodal, daría lugar a moitas ramas na multifunción estimada. Zhou e Huang (2019) propuxeron dous métodos diferentes para obter parámetros de suavizado na práctica no contexto da regresión multimodal con variables escalares. O primeiro, coñecido como validación cruzada modal, aspira a equilibrar o número de modas locais estimados e a distancia da multifunción estimada aos datos. Aínda que este método mostra un bo comportamento na práctica, nada asegura que minimizar a función de validación cruzada modal minimizará o Erro Cadrático Medio Integrado modal do estimador ou calquera outro criterio de erro. O segundo procedemento proposto por Zhou e Huang (2019) é minimizar o Erro Cadrático Integrado modal do estimador (a versión integrada da distancia de Hausdorff entre o estimador e a verdadeira multifunción) utilizando un método de remostraxe baseado nunha mestura de regresións parámetricas. Outro criterio, proposto por Chen et al. (2016), consiste en construír unha banda de predicción para a multifunción de regresión e posteriormente seleccionar os parámetros que minimicen unha función de perda definida como o volume de dita banda. Porén, os autores asumen que o parámetro de suavizado é o mesmo para ambas variables, o cal non está xustificado, especialmente tendo en conta o diferente papel que xogan os dous parámetros. Ademais, a selección do parámetro depende do nivel de predicción previamente fixado.

Validación cruzada modal para regresión circular. Unha das vantaxes da validación cruzada modal de Zhou e Huang (2019) é que a súa adaptación a casos máis complexos como o da regresión circular é case inmediata. Aínda que as propiedades teóricas deste procedemento non están ben estudadas, o comportamento deste método na práctica resulta satisfactorio. Para o escenario presentado na Sección 2, onde a variable resposta é circular, a validación cruzada modal consiste en seleccionar os parámetros g e κ (onde g representa tanto h ou ν, dependendo da natureza da variable explicativa) mediante a minimización de

$$CV(g,\kappa) = \frac{1}{n}\sum_{i=1}^{n} \tilde{d}(\widehat{M}_{-i}^{g,\kappa}(\Delta_i), Y_i) N_{-i}(\Delta_i), \tag{8}$$

onde $\tilde{d}(x, A) = \inf_{z\in A} 1 - \cos(x - z)$, $\widehat{M}_{-i}^{g,\kappa}$ é o estimador de M utilizando os datos $\{(\Delta_j, \Phi_j) : j \neq i\}$ construído cos parámetros g e κ e $N_{-i}(\Delta_i)$ denota o número de modas locais estimadas cando $\Delta = \Delta_i$. Este método non se basea en fundamentacións teóricas e precísanse de ensaios computacionais para avaliar a súa eficacia.

5. ESTUDO DE SIMULACIÓN

Nesta sección analizamos o comportamento do estimador tanto no caso no que a variable explicativa é escalar como no caso onde está presenta unha natureza circular. En primeiro lugar presentamos os escenarios de simulación e de seguido, amosamos os resultados obtidos.

Escenarios de simulación. Nos nosos exemplos simulados, a mostra está dividida en dous grupos, con cada grupo correspondéndose cunha rama da multifunción obxectivo. Para cada observación utilizamos o subíndice ji, onde j denota o grupo ou número de rama e i denota o número de observación dentro de cada grupo. Ademais, n_j denota o tamaño mostral para o j-ésimo grupo. Nótese que estes grupos subxacentes non son coñecidos na práctica e non dispoñemos de información sobre eles. Os modelos simulados móstranse na Tabla 1.

O primeiro modelo en cada escenario correspóndese con dúas curvas paralelas ou, visto dende outra perspectiva, a únha curva de regresión cun erro bimodal. No que se refire ao segundo modelo de cada escenario, as dúas curvas non son paralelas. En todos os casos, os tamaños mostrais son $(n_1, n_2) \in \{(100,100),(100,200),(200,200),(200,300),(300,300)\}$. Exemplos de datos simulados de todos os modelos poden verse na Figura 2 xunto coas multifuncións verdadeiras.

Modelo	Xeneración da mostra	Multifunción de regresión
LC-1	$\Phi_{1i} = (6\tan^{-1}(2.5X_{1i} - 3) + \varepsilon_{1i})(\text{mod } 2\pi)$ $\Phi_{2i} = (\pi + 6\tan^{-1}(2.5X_{1i} - 3) + \varepsilon_{1i})(\text{mod } 2\pi)$ $X_1, X_2 \sim U(0,1)$	$M(x) = \{6\tan^{-1}(2.5x - 3),$ $\pi + 6\tan^{-1}(2.5x - 3)\}(\text{mod } 2\pi)$
LC-2	$\Phi_{1i} = (\text{atan2}(\sin(3X_{1i}^2), \cos(3X_{1i}^2)) + \varepsilon_{1i})(\text{mod } 2\pi)$ $\Phi_{2i} = (\pi/2 + 2\tan^{-1}(10X_{2i} - 1/2) + \varepsilon_{2i})(\text{mod } 2\pi)$ $X_1, X_2 \sim U(0,1)$	$M(x) = \{\text{atan2}(\sin(3x^2), \cos(3x^2)),$ $\pi/2 + 2\tan^{-1}(10x + 1/2)\}(\text{mod } 2\pi)$
CC-1	$\Phi_{1i} = (2\cos\Theta_{1i} + \varepsilon_{1i})(\text{mod } 2\pi)$ $\Phi_{2i} = (3\pi/4 + 2\cos\Theta_{1i} + \varepsilon_{2i})(\text{mod } 2\pi)$ $\Theta_1, \Theta_2 \sim$ Circular Uniform	$M(\theta) = \{2\cos\theta, 3\pi/4 + 2\cos\theta\}(\text{mod } 2\pi)$
CC-2	$\Phi_{1i} = (3/4\cos\Theta_{1i} - \pi/2 + \varepsilon_{1i})(\text{mod } 2\pi)$ $\Phi_{2i} = (\pi/2 - \cos\Theta_{1i} + \varepsilon_{2i})(\text{mod } 2\pi)$ $\Theta_1, \Theta_2 \sim$ Circular Uniform	$M(\theta) = \{3/4\cos\theta - \pi/2, \pi/2 - 2\cos\theta\}(\text{mod } 2\pi)$

Táboa 1: Modelos simulados. LC denota explicativa escalar e resposta circular e CC denota explicativa circular e resposta circular. (mod 2π) denota módulo 2π. En todos os modelos $\varepsilon_{1i}, \varepsilon_{2i} \sim vM(0,\tau)$ with $\tau \in \{6,8,10\}$.

Para medir o comportamento dos estimadores, o Erro Medio Circular Integrado modal (EMCI_m) foi aproximado tras xerar 500 réplicas Monte Carlo de datos simulados e calculando a media do Erro Integrado Circular modal (EIC_m) do estimador multimodal:

$$\text{EIC}_m(\widehat{M}) = \int_{\delta \in \text{Supp}(\Delta)} \widetilde{\Lambda}(\delta) d\delta,$$

onde as integrais foron aproximadas numericamente mediante a regra de Simpson e $\widetilde{\Lambda}$ está definida en (7). Nos experimentos de simulación, os parámetros de suavizado foron seleccionados mediante validación cruzada modal. Os resultados compáranse cos obtidos tras utilizar os parámetros que minimizan o EIC_m, que se toman como referencia.

Resultados. O EMCI_m estimado para todos os modelos pode atoparse na Táboa 2. Como se esperaba, o valor estimado do EMCI_m xeralmente diminúe ao incrementar o tamaño mostral. Os poucos casos nos que non se observa este comportamento son cando os tamaños mostrais en cada grupo non son iguais, é dicir$(n_1, n_2) \in \{(100,200),(200,300)\}$. Ademais, un valor grande da concentración do erro tamén da lugar a unha menor estimación do EMCI_m. O desempeño do criterio de validación cruzada modal é tamén máis que aceptable, dado que para valores grandes do tamaño mostral, os valores estimados do EMCI_m cando os parámetros de suavizado se seleccionan mediante este criterio están moi preto dos valores obtidos cos parámetros *óptimos*. O peor rendemento

obtense co modelo LC-2 con $\kappa = 6$. Como se mostra no panel de arriba e esquerda da Figura 2, cando x toma valores próximos a 1, as dúas ramas da multifunción están moi próximas, o que fai difícil discernir entre os dous grupos cando a concentración do erro é baixa.

		$\tau = 6$		$\tau = 8$		$\tau = 10$	
Model	(n_1, n_2)	B	CV	B	CV	B	CV
LC-1	$(100, 100)$	0.021	0.034	0.016	0.023	0.013	0.018
	$(100, 200)$	0.016	0.039	0.012	0.024	0.011	0.018
	$(200, 200)$	0.011	0.019	0.008	0.012	0.007	0.010
	$(200, 300)$	0.010	0.016	0.007	0.011	0.006	0.009
	$(300, 300)$	0.007	0.012	0.006	0.008	0.005	0.006
LC-2	$(100, 100)$	0.016	0.035	0.012	0.025	0.010	0.018
	$(100, 200)$	0.016	0.070	0.010	0.039	0.008	0.024
	$(200, 200)$	0.010	0.023	0.007	0.012	0.005	0.009
	$(200, 300)$	0.009	0.038	0.006	0.017	0.005	0.011
	$(300, 300)$	0.007	0.015	0.005	0.009	0.004	0.007
CC-1	$(100, 100)$	0.144	0.395	0.117	0.213	0.102	0.148
	$(100, 200)$	0.120	0.182	0.097	0.147	0.087	0.120
	$(200, 200)$	0.066	0.089	0.054	0.066	0.047	0.057
	$(200, 300)$	0.058	0.068	0.046	0.054	0.040	0.045
	$(300, 300)$	0.042	0.056	0.035	0.045	0.030	0.038
CC-2	$(100, 100)$	0.116	0.227	0.092	0.173	0.080	0.150
	$(100, 200)$	0.085	0.232	0.068	0.159	0.057	0.120
	$(200, 200)$	0.055	0.091	0.045	0.069	0.039	0.059
	$(200, 300)$	0.046	0.092	0.037	0.066	0.032	0.049
	$(300, 300)$	0.037	0.055	0.030	0.043	0.026	0.033

Táboa 2: Estimacións por Monte Carlo do EMCI_m para os modelos LC-1, LC-2, CC-1 e CC-2 con diferentes valores da concentración τ e tamaños mostrais. Resultados baixo B (benchmark) corresponden aos obtidos con parámetros de suavizado minimizando o EIC_m e resultados baixo CV refírense aos obtidos con parámetros seleccionados mediante validación cruzada modal.

6. ANÁLISE DO ESCAPE DAS LARVAS DE PEIXE CEBRA

Nesta sección analizamos os datos de larvas de peixes cebra presentados na Introdución. Todos os detalles do experimento poden atoparse en Nair et al. (2017b). Os datos represéntanse na Figura 3, tanto no toro coma no plano. O noso obxectivo é estudar como o ángulo no que se aproxima o robot depredador (variable explicativa circular) inflúe á dirección de escape dos peixes (variable resposta circular). Dado que a variable explicativa só cubre unha parte da circunferencia, poderíase argumentar que esta variable pode ser considerada como escalar en lugar de circular. Porén, o soporte da variable explicativa vén dado polo rango de visión de cada animal. Dado que existen animais, como os camaleóns, cun rango de visión de 360 graos, é importante tratar esta variable como circular para que o método sexa extendible a outros animais.

Neste experimento, unha dirección de escape no intervalo $[-\pi, 0)$ indica un escape contralateral, mentres que unha dirección de escape en $[0, \pi)$ clasifícase como ipsilateral. Ademais, valores da dirección de estímulo máis cara a esquerda no panel dereito da Figura 3 indica que o robot se aproximou á larva en cuestión do lado rostral (próximo á parte dianteira do corpo). Por outra banda, direccións de estímulo máis cara a dereita no panel dereito da Figura 3 mostran que o robot se aproximou á larva dende o lado caudal (preto da cola). Aproximarse aos peixes dende os lados rostral ou caudal significa que o robot apareceu na visión periférica dos peixes.

O estimador tipo núcleo da regresión para explicativas e respostas circulares, proposto por Di Marzio et al. (2012), foi aplicado aos datos, onde o parámetro de concentración foi obtido mediante validación cruzada. Dito estimador está representado en verde e con trazo contínuo na Figura 3. De acordo con este estimador, a dirección de escape media é ipsilateral cando o robot se aproxima aos peixes polo lado rostral, contralateral cando se achega ás larvas do seu

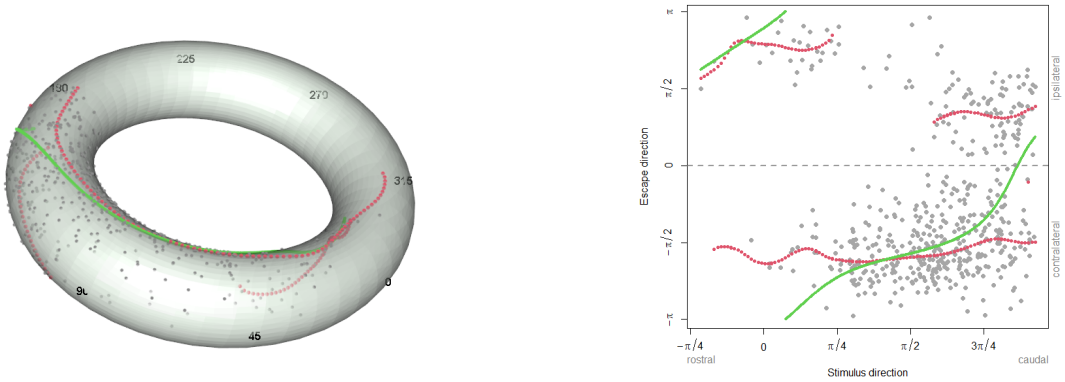


Figura 3: Representacións no toro e no plano dos datos de peixes cebra coa multifunción de regresión estimada (vermello, trazo punteado) e a estimación da función de regresión usual (verde, trazo continuo).

lado esquerdo ou dereito e arredor de cero cando se aproxima polo lado caudal. Co fin de obter máis coñecemento sobre a relación entre a dirección de escape e o ángulo de estímulo, aplicamos o estimador da regresión multimodal, representado en vermello e trazo punteado na Figura 3. Os parámetros de suavizados foron seleccionados mediante o criterio de validación cruzada modal. Podemos observar que, cando o robot aparece na visión periférica dos peixes (tanto o lado rostral como caudal) existen dúas direccións de escape prederidas, unha ipsilateral e outra contralateral. Por outra banda, só se estima unha moda cando o robot aparece dentro do campo de visión frontal dos peixes, indicando un escape contralateral neste caso.

En conclusión, o estimador da regresión multimodal permítenos estimar as dúas direccións de escape prefiridas cando as larvas detectan ao seu depredador mediante a súa visión periférica, é dicir, cando o depredador se achega dende a parte frontal ou traseira. Pola contra, se este se aproxima aos peixes dende os lados dereito ou esquedo dos seus corpos (onde se atopan os ollos dos peixes), a dirección de escape resulta oposta á dirección de ameaza.

7. CONCLUSIÓNS

A literatura existente sobre a regresión con variables circulares enfócase predominantemente na regresión *á media*. Con todo, como mostra a nosa análise das direccións de escape de larvas de peixe cebra na Sección 6, cando a densidade condicional dos datos presenta unha estructura multimodal, a estimación da media condicional pode levar a resultados enganosos.

Neste traballo introducimos un método non paramétrico para estudar as modas locais dunha variable resposta condicionadas a unha variable explicativa, cando a resposta e/ou a explicativa son variables circulares. O noso método de regresión multimodal permite estimar os valores máis probables da resposta dada a explicativa, proporcionando un coñecemento máis axeitado da relación entre as variables máis aló da regresión en media. As propiedades do estimador foron estudadas de xeito teórico e tamén mediante un estudo de simulación. O enfoque de suavizado tipo núcleo que se toma neste traballo require da selección de parámetros de suavizado, tarefa que tamén foi realizada.

Dende un punto de vista práctico, na nosa análise dos datos de peixes cebra utilizando a regresión multimoda, vimos que as larvas teñen unha dirección de escape prefirida cando son atacadas dende unha dirección lateral aos seus corpos. Emporiso, cando o depredador se achega aos peixes dende os lados rostral ou caudal, existen dúas direccións de escape prefiridas. Ademais, o estimador multimodal permítenos percibir graficamente as diferenzas entre o comoportamento de escape dos animais cando son perseguidos dende os lados caudal e rostral.

Finalmente, a metodoloxía presentada neste traballo establece as bases para ferramentas inferenciais para regresión con variables circulares, construídas dende unha perspectiva multimodal.

Por exemplo, o estimador introducido na Sección 2 pode utilizarse para construír tanto bandas de confianza como bandas de predicción que, dada a estrutura multimodal dos datos, resultaría en bandas moito máis estreitas que as construídas mediante os estimadores de regresión en media.

AGRADECEMENTOS

Este traballo foi financiado polo Proxecto MTM2016–76969–P e co-financiado pola European Regional Development Fund (ERDF), os Grupos de Referencia Competitivos 2017–2020 (ED431C 2017/38) da Xunta de Galicia a través da ERDF. O traballo de M. Alonso-Pena foi financiado pola Xunta de Galicia a través do contrato predoutoral con referencia ED481A-2019/139 da Consellería de Educación, Universidade e Formación Profesional. As autoras agradecen tamén ao Centro de Supercomputación de Galicia (CESGA) polos recursos computacionais e a Alejandra López Pérez pola axuda gráfica.

REFERENCIAS

Ameijeiras-Alonso, J., Lagona, F., Ranalli, M. e Crujeiras, R.M. (2019). A circular nonhomogeneous hidden Markov field for the spatial segmentation of wild fire occurrences. Environmetrics, 30, e2501.

Card, G. e Dickinson, M.H. (2008). Visually mediated motor planning in the escape response of Drosophila. Current Biology, 18, 1300–1307.

Casa, A., Chacón, J.E. e Menardi, G. (2020). Modal clustering asymptotics with applications to bandwidth selection. Electronic Journal of Statistics, 14, 835–856.

Chen, Y.-C. (2018). Modal regression using kernel density estimation: A review. Wiley Interdisciplinary Reviews: Computational Statistics, 10, e1431.

Chen, Y.-C., Genovese, C.R., Tibshirani, R.J. e Wasserman, L. (2016). Nonparametric modal regression. Annals of Statistics, 44, 489–514.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 790–799.

Comaniciu, D. e Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 603–619.

Di Marzio, M. and Panzera, A. e Taylor, C.C. (2009). Local polynomial regression for circular predictors. Statistics & Probability Letters, 798, 2066–2075.

Di Marzio, M., Fensore, S., Panzera, A. e Taylor, C.C. (2016). A note on nonparametric estimation of circular conditional densities. Journal of Statistical Computation and Simulation, 86, 2573–2582.

Di Marzio, M., Panzera, A. e Taylor, C.C. (2012). Non-parametric regression for circular responses. Scandinavian Journal of Statistics, 40, 238–255.

Einbeck, J. e Tutz, G. (2006). Modelling beyond regression functions: An application of multimodal regression to speedflow data. Journal of the Royal Statistical Society, Series C, Applied Statistics, 55, 461–475.

Fan, J. e Gijbels, I. (1996). Local Polynomial Modelling and its Applications. Chapman and Hall, London.

Fisher, N.I. (1993). Statistical Analysis of Circular Data. Cambridge University Press, Cambridge.

Fukunaga, K. e Hostetler, L. (1975). Regression models for angular responses. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory, 21, 32–40.

Jammalamadaka, S.R. e SenGupta, A. (2001). Topics in Circular Statistics. World Scientific, Singapore.

Kim, S. e SenGupta, A. (2017). Multivariate-multiple circular regression. Journal of Statistical Computation and Simulation, 87, 1277–1291.

Kobayashi, T. e Otsu, N. (2010). Von mises-fisher mean shift for clustering on a hypersphere. In Proceedings of the 20th international conference on pattern recognition, pages 2130–2133. IEEE.

Ley, C. e Verdebout, T. (2017). Modern Directional Statistics. Chapman & Hall, Boca Raton.

Marchetti, M. e Scapini, F. (2003). Use of multiple regression models in the study of sandhopper orientation under natural conditions. Estuarine, Coastal and Shelf Science, 58, 207–215.

Mardia, K.V. e Jupp, P.E. (2000). Directional Statistics. John Wiley & Sons, Inc., New York.

Mooney, J.A., Helms, P.J. e Jollife, I.T. (2003). Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. Computational Statistics and Data Analysis, 41, 505–513.

Nair, A., Changsing, K., Stewart, W.J. e McHenry, M.J. (2017a). Data from: Fish prey change strategy with the direction of a threat. https://doi.org/10.5061/dryad.47mq9.

Nair, A., Changsing, K., Stewart, W.J. e McHenry, M.J. (2017b). Fish prey change strategy with the direction of a threat. Proceedings of the Royal Society B, 284, 20170393.

Oba, S., Kato, K. e Ishii, S. (2005). Multi-scale clustering for gene expression profiling data. In 5th IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05), pages 210–217. IEEE.

Obleser, P., Hart, V. Malkemper, E.P. et al. (2016). Compass-controlled escape behavior in roe deer. Behavioral Ecology and Sociobiology, 70, 1345–1355.

Oliveira, M., Crujeiras, R.M. e Rodríguez-Casal, A. (2013). Nonparametric circular methods for exploring environmental data. Environmental and Ecological Statistics, 20, 1–17.

Pewsey, A., Neuhuser, M. e Ruxton, G.D. (2013). Directional Statistics. Oxford University Press, Oxford.

Sato, N., Shidara, H. e Ogawa, H. (2019). Trade-off between motor performance and behavioural flexibility in the action selection of cricket escape behaviour. Scientific Reports, 9, 18112.

Scapini, F., Aloia, A., Bouslama, M. et al. (2002). Multiple regression analysis of the sources of variation in orientation of two sympatric sandhoppers, talitrus saltator and talorchestia brito, from an exposed mediterranean beach. Behavioral Ecology, 51, 403–414.

Scott, D.W. (1992). Multivariate Density Estimation. Wiley, New York

SenGupta, S. e Rao, J.S. (1966). Statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranhita: Godavari Valley. Sankhya: The Indian Journal of Statistics, Series B, 28, 165–174.

Wikimedia Commons (2008). Larval zebrafish. Author: CSIRO. https://commons.wikimedia.org/wiki/File:CSIRO_ScienceImage_7598_larval_zebra.jpg [Online; accessed July 27th, 2021].

Wikimedia Commons (2014). Life cycle of Zebrafish. https://commons.wikimedia.org/wiki/File:Zebrafish-model-4-638.jpg [Online; accessed July 27th, 2021].

Zhang, Y. e Chen, Y.-C. (2020). Kernel smoothing, mean shift, and their learning theory with directional data. arXiv e-prints, page arXiv:2010.13523.

Zhou, H. e Huang, X. (2016). Nonparametric modal regression in the presence of measurement error. Electronic Journal of Statistics, 10, 3579–3620.

Zhou, H. e Huang, X. (2019). Bandwidth selection for nonparametric modal regression. Communications in Statistics - Simulation and Computation, 48, 968–984.