Análisis Estadístico de Datos Direccionales. Aplicaciones Medioambientales

Proyecto Fin de Máster
Máster en Técnicas Estadísticas

María Leyenda Rodríguez

Resumen

En diversos campos surgen problemas estadísticos dónde los datos son dados mediante medidas angulares dando la orientación o angulos en el plano (datos circulares) o en el espacio (datos esféricos). Datos circulares es el caso más simple de esta categoría de datos llamada datos direccionales, dónde la única respuesta no es escalar, pero es angular o direccional. La suposición estadística básica es que los datos son una muestra aleatoria de una población de diecciones. Para trabajar con datos de esta naturaleza es necesario construir nuevos estadísticos pues los estadísticos usuales empleados para datos lineales son inapropiados en los datos direccionales, pues no tienen en cuenta la naturaleza circular de los datos direccionales.

En este trabajo vamos a analizar la dirección de viento recogida por la estación meteorológica A Mourela que pertenece a un Sistema de Control Suplementario de la Contaminación Atmosférica de la U.P.T. de As Pontes propiedad de Endesa Generación S.A. Debido a la naturaleza circular de esta variable comenzaremos describiendo los datos circulares según ([9]). Comenzaremos con la estimación no paramétrica tipo núcleo de la densidad circular haciendo un especial hincapié en la selección del parámetro de suavizado. Notemos que estos datos de dirección de viento son minutales y recogidos a lo largo del año 2010 lo que nos llevó a estudiar el modelo de Möebius de series de tiempo.

Debido al interés por estudiar la relación entre las concentraciones de SO_2 y la dirección de viento realizaremos la estimación no paramétrica tipo núcleo de la función de regresión dónde la selección del parámetro de suavizado también es de vital importancia. Notemos que para este apartado se va a trabajar con la dirección de viento recogida en las estaciones de medida automáticas B1, B2, C9, F2 Y G2 que también forman parte del Sistema de Control Suplementario de la Contaminación Atmosférica U.P.T de As Pontes; pues para realizar la estimación de la función de regresión necesitamos que tanto los datos de SO_2 y dirección de viento estén recogidos por la misma estación.

Contenidos

Capítulo 1

Introducción

Los datos circulares surgen de varias formas. Las dos principales ccorresponden a los principales instrumentos de medición circular, la brújula y el reloj. Entre las típicas observaciones medidas por la brújula están las direcciones de viento y las direcciones migratorias de los pájaros. Entre las típicas mediadas por el reloj están los tiempos de llegada (en un reloj de 24 horas) de los pacientes a una unidad de urgencias de un hospital. Conjuntos de datos similares surgen al considerar veces en una año (o veces en un mes) de los eventos correspondientes.

Las direcciones en el plano se pueden observar como vectores unitarios en el plano como puntos en el círculo unidad. Aunque hay otras dos formas muy útiles de observar las direcciones- como ángulos y como números complejos; escogiendo una direccción y orientación inicial (Esto es equivalente a escoger un sistema de cooredenadas ortogonal en el plano). Luego cada punto $\mathbf{x}$ en el círculo unidad puede ser representado por un ángulo θ o por un número complejo unitario z.

$$x = (\cos\theta, sen\theta)$$
$$z = e^{i\theta} = \cos\theta + isen\theta$$

En este capítulo, presentaremos la descriptiva de los datos circulares siguiendo ([9]). Comenzaremos por estudiar las medidas de localización, concentración y dispersión, secciones 1.1, 1.2., y finalmente nos centraremos en las principales distribuciones circulares, sección 1.3.

1.1 Medidas de localización

Sean $x_1, \ldots, x_n$ cuyos ángulos correspondientes son θ_i, $i = 1, \ldots, n$. La dirección media $\bar{\theta}$ de $\theta_1, \ldots, \theta_n$ es la dirección de $x_1 + \ldots + x_n$ que es el centro de masa $\bar{x}$ de $x_1, \ldots, x_n$.

Por tanto si las coordenadas cartesianas de x_j son $(\cos\theta_j, sen\theta_j)$, entonces $(\bar{C}, \bar{S})$ son las coordenadas cartesianas del centro de masas.

$$\bar{C} = \frac{1}{n}\sum_{j=1}^{n}\cos\theta_j;\ \bar{S} = \frac{1}{n}\sum_{j=1}^{n} sen\theta_j$$

Además $\bar{\theta}$ es la solución de las ecuaciones (1.1,1.2) y es la dirección de $x_1+\ldots+x_n$, denominada, dirección media

$$\bar{C} = \bar{R}cos\bar{\theta} \tag{1.1}$$
$$\bar{S} = \bar{R}sen\bar{\theta} \tag{1.2}$$

(supuesto $\bar{R} > 0$), dónde la longitud media resultante $\bar{R}$ es dada por

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2} \tag{1.3}$$

Notemos que $\bar{\theta}$ no está definida cuando $\bar{R} = 0$ y cuando $\bar{R} > 0$, $\bar{\theta}$ es dada por

$$\bar{\theta} = \begin{cases} \tan^{-1}(\bar{S}/\bar{C}) & \text{si} \quad \bar{C} \geq 0 \\ \\ \tan^{-1}(\bar{S}/\bar{C}) + \pi & \text{si} \quad \bar{C} < 0 \end{cases}$$

En el contexto de estadística circular $\bar{\theta}$ no es la media $(\theta_1 + \ldots + \theta_n)/n$. Sin embargo, la denominamos media muestral de la dirección y cumple las siguientes propiedades

- es equivariante bajo rotación.
- Diferentes estadísticos utilizando diferentes sistemas de coordenadas estarán de acuerdo en dónde está la media la muestral, a pesar de que puede usar números diferentes para describir su posición

Por otra, la mediana muestral de la dirección $\bar{\theta}$ de los ángulos $(\theta_1, \ldots, \theta_n)$ es algún ángulo ϕ tal que la mitad de los puntos se encuentran en el arco $[\phi, \phi + \pi)$ y la mayoría de los puntos están más cerca de ϕ que de $\phi + \pi$.

- Cuando n es par, la mediana coincide con uno de ellos.
- Cuando n es impar, es conveniente tomar la mediana como el punto medio de dos puntos adyacentes adecuados.

1.2 Medidas de concentración y dispersión

La media de longitud resultante $\bar{R}$ introducida en (1.3) como longitud del centro de masas del vector $\bar{x}$ es dada por$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}$. Es la medida de dispersión más importante en datos direccionales.
Luego, si $x_1, \ldots, x_n$ son vectores unitarios, entonces $0 \leq \bar{R} \leq 1$.
Por tanto, podemos concluir que

- Si las direcciones $\theta_1, \ldots, \theta_n$ están estrechamente agrupadas luego $\bar{R} = 1$.
- Si $\theta_1, \ldots, \theta_n$ están muy dispersas luego $\bar{R}$ será prácticamente 0.

Además esta medida de concentración tiene las siguientes propiedades

- $\bar{R}$ es invariante bajo rotaciones.
- La longitud resultante $\bar{R}$ es la longitud del vector resultante $x_1 + \ldots + x_n$.
- $R = n\bar{R}$

A veces emplearemos otras medidas de dispersión de datos circulares, para comparar con datos en línea: la medida más simple es la varianza circular muestral,

$$V = 1 - \bar{R}$$

o la desviación circular estándard,

$$v = \sqrt{-2log\bar{R}} \tag{1.4}$$

También podemos considerar como medida de dispersión la distancia entre dos ángulos θ y ξ

$$min(\theta - \xi, 2\pi - (\theta - \xi)) = \pi - |\pi - |\theta - \xi||$$

o definir la medida de dispersión de los ángulos $\theta_1, \ldots, \theta_n$ sobre un ángulo dado α es

$$d_0(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(\pi - |\pi - |\theta_i - \alpha||)$$

la función d_0 toma el mínimo en la mediana muestral $\tilde{\theta}$. La desviación circular media es $d_0(\tilde{\theta})$

1.3 Distribuciones notables

Una forma de especificar una distribución en el círculo unidad es por medio de su función de distribución. Suponemos que ha sido escogida una dirección y orientación inicial en el círculo unidad. Luego la distribución puede ser considerada como que los ángulos aleatorios θ, y su función de distribución F es definida como la función en la recta real dada por

$$F(x) = Pr(0 < \theta \leq x),\ 0 \leq x \leq 2 * \pi$$

y

$$F(x + 2\pi) - F(x) = 1,\ -\infty < x < \infty \tag{1.5}$$

La ecuación (1.5) solo afirma que todo arco en el círculo unidad de longitud 2π tiene probabilidad uno (este arco es la totalidad de la circumferencia en el círculo). Para $\alpha \leq \beta \leq \alpha + 2\pi$,

$$Pr(\alpha < \theta \leq \beta) = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} dF(x) \tag{1.6}$$

dónde la integral es una integral de Lebesgue-Stieltjes. La función de distribución F es contínua por la derecha. En contraste con las distribuciones en la recta real,

$$lim_{x\to\infty}F(x) = \infty\ lim_{x\to-\infty}F(x) = -\infty$$

Por definición

$$F(0) = 0\ F(2\pi) = 1$$

Notemos que aunque la función F dependa de la elección de la elección inicial, (1.6) muestra que $F(\beta) - F(\alpha)$ es independiente de esta elección. por tanto cambiar la dirección inicial es simplemente añadir una constante a F.
Si la función de distribución F es absolutamente contínua tiene como función de densidad a f tal que

$$\int_{\alpha}^{\beta} f(x)dx = F(\beta) - F(\alpha),\ -\infty < \alpha \leq \beta < \infty$$

Una distribución circular es una distribución de probabilidad la cual está concentrada en la circunferencia de círculo unidad. Las distribuciones circulares son de dos tipos:

1. Discretas- asignan masas de probabilidad solo a un número de direcciones
2. Absolutamente contínuas

Una función f es la función de densidad de una distribución absolutamente contínua si y solo si

1. $f(\theta) \geq 0$ en casi todo $(-\infty, \infty)$
2. $\int_0^{2\pi} f(\theta)d\theta = 1$
3. $f(\theta) = f(\theta + 2\pi)$ en casi todo $(-\infty, \infty)$

1.3.1 Distribución lattice

La distribución lattice (1.7) es una distribución discreta que toma valores en los vértices de un polígono de m lados inscrito en el círculo unidad, $\frac{2\pi r}{m}$. Si los pesos (1.8) son $p_r = \frac{1}{m}$, entonces se le denomina distribución uniforme discretizada en m puntos.

$$Pr(\theta = v + \frac{2\pi r}{m}) = p_r \; r = 0, 1, \cdots, m-1 \tag{1.7}$$

$$p_r \geq 0, \; \sum_{r=0}^{m-1} p_r = 1 \tag{1.8}$$

1.3.2 Distribución uniforme

Es la distribución más básica en el círculo y a menudo es usada como modelo nulo. Esta es la única distribución en el círculo la cual es invariante bajo rotación y reflexión. Su función de densidad es

$$f(\theta) = \frac{1}{2\pi}$$

Por tanto, para $\alpha \leq \beta \leq \alpha + 2\pi$,

$$Pr(\alpha < \theta \leq \beta) = \frac{\beta - \alpha}{2\pi}$$

es decir, es proporcional a la longitud del arco.

1.3.3 Distribución von Mises

Desde el punto de la inferencia estadística, la distribución von Mises (1.9) es la más mejor distribución en el círculo. Pues esta es unimodal y es simétrica sobre $\theta = \mu$ (Figura 1.1). Además la moda se encuentra en $\theta = \mu$ y la antimoda en $\theta = \mu + \pi$. La relación de la moda de la densidad y la antimoda viene dada por $e^{2\kappa}$, así que cuanto mayor sea el valor de κ, mayor es el agrupamiento acerca de la moda.

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa cos(\theta - \mu)} \tag{1.9}$$
$$F(\theta; 0, \kappa) = \frac{1}{I_0(\kappa)} \int_0^{\theta} e^{\kappa cosu} du$$

- I_0 denota la función de Bessel modificada de primer tipo y orden 0,

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

- El parámetro μ es la media de las direcciones
- κ es el parámetro de concentración. En (Figura 1.1) se observa que cuanto más grande sea el parámetro de concentración su función de densidad von Mises estará más concentrada en su media circular.

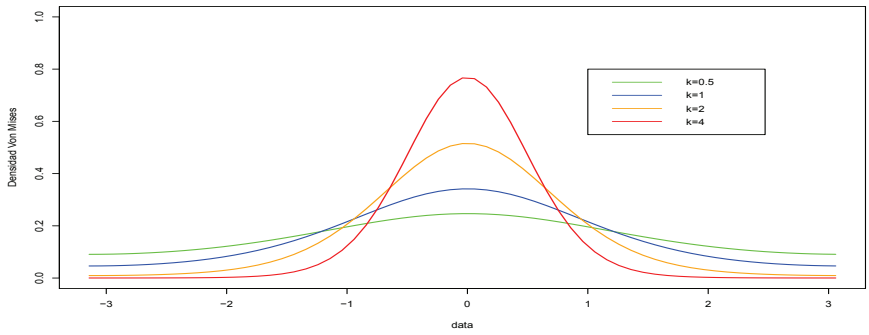

Figura 1.1: Representación de la función de densidad de Distribución von Mises vM(0,κ), κ=0.5,1,2,3,4.

Como el parámetro κ es una medida de concentración en el caso de que nuestra muestra siga una von Mises; nos hemos planteado explicar explicar la relación entre diferentes medidas de dispersión como son la media de longitud resultante

(1.3), la desviación circular estándard (1.4), la varianza definida de forma análoga que en el caso lineal(1.10) y el parámetro de concentración de la distribución von Mises κ ante una muestra que siga una distribución von Mises.
Para ello,

1. Generamos una muestra $vM\pi, \kappa)$ de tamaño 3000 para distintos valores del párametro κ (75 valores).

2. Para cada una de las muestras calculamos

   - La varianza (1.10) con respecto a la distancia euclídea (1.11) y a la distancia definida por la longitud de arco(1.12),

$$S_n^2(\theta) = \frac{1}{n}\sum_{i=1}^{n} d^2(\theta_i, \theta) \tag{1.10}$$

$$\theta = min_y S_n^2(y)$$

$$d^2(\theta_i, \theta j) = 2(1 - (cos(\theta_i - \theta_j))) \tag{1.11}$$

$$d^2(\theta_i, \theta j) = \pi - |\pi - |\theta_i - \theta_j|| \tag{1.12}$$

   - el parámetro de concentración $\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}$ y
   - la desviación circular estándard $\sqrt{-2log(\bar{R})}$.

3. Representamos el estimador local-lineal para estudiar las relaciones de estas medidas de concentración y dispersión, frente a diferentes valores de κ.

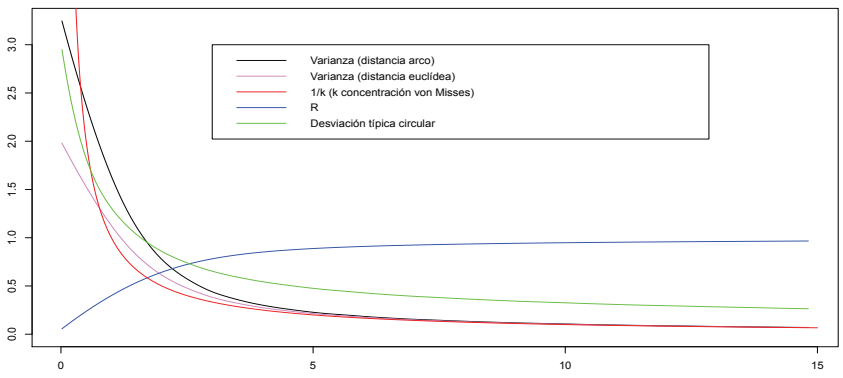

Figura 1.2: Relación entre$\bar{R}$,v,S_n^2 y κ

Mediante este estudio (Figura 1.2) se puede observar que v, S_n^2 (obtenida mediante cualquiera de las distancias) se comportan de forma similar. De modo que disminuye casi de modo exponencial a medida que κ aumenta. Mientras que $\bar{R}$ se comporta de modo diferente; aumenta a medida que κ aumenta pero se presenta una asíntota en 1.

La distribución von Mises juega un papel cruciial en el campo de los datos direccionales de hecho, juega el mismo papel que la distribución gausiana en los datos lineales. Debido a esto nos va a interesar aproximar, relacionar la distribución von Mises con otras distribuciones. Así pues,

- Si $\kappa = 0$, entonces $vM\mu, \kappa)$ es la distribución uniforme. La aproximación $exp(x) \cong 1 + x$, muestra que para κ pequeños $vM\mu, \kappa) \cong C(\mu, \kappa/2)$, dónde $C(\mu, \kappa/2)$ denota una distribución Cardioide. Por tanto, una distribución von Mises con parámetro de concentración pequeño puede ser aproximado por una distribución Cardioide con la misma dirección media y media de longitud resultante.

- Cuando $\kappa \to \infty$ la distribución vMμ,κ) está concentrada en el punto $\theta = \mu$. Si κ es grande, $\xi = \kappa^{1/2}(\theta - \mu)$. Luego, la función de densidad(1.9) de ξ es proporcional a

$$exp{-}\kappa[1 - cos(\kappa^{-1/2}\xi)] \tag{1.13}$$

  Para κ grande,

$$1 - cos(\kappa^{-1/2} * \xi) = \frac{1}{2}\kappa^{-1}\xi^2 + O(\kappa^{-2})$$

  así pues, de (1.13), $\xi \sim N(0, 1)$. Luego para grades valores de κ,

$$\theta \sim vM\mu, \kappa) \Rightarrow \kappa^{-1/2}(\theta - \mu) \sim N(0, 1),\ \kappa \to \infty \tag{1.14}$$

- De foma más general, cualquier von Mises puede ser aproximada por una distribución normal wrapped.

$$vM\mu, \kappa) \cong WN(\mu, A(\kappa)),\ \kappa \to \infty$$
$$A_\kappa = I_1(\kappa)/I_0(\kappa)$$

Respecto a la convolución, notemos que la convolución de dos von Mises no es una distribución von Mises. En cambio, la convolución de dos distribuciones normal wrapped, $WN(\mu_1, A(\kappa_1))$ y $WN(\mu_2, A(\kappa_2))$ es la distribución wrapped

normal $WN(\mu_1 + \mu_2, A(\kappa_1)A(\kappa_2))$. La cual puede ser aproximada por$vM\mu_1 + \mu_2, A^{-1}(A(\kappa_1)A(\kappa_2)))$

$$\theta_1 + \theta_2 \sim vM\mu_1 + \mu_2, A^{-1}(A(\kappa_1)A(\kappa_2)))$$

1.3.4 Distribución Cardioide

La perturbación de la densidad uniforme por la función coseno da lugar a una distribución Cardioide $C(\mu, \rho)$, cuya función de densidad es,

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

- La media de longitud resultante es ρ
- La media de la dirección es μ
- La distribución es simétrica y unimodal en μ (si $\rho > 0$) (Figura 1.3).
- Si $\rho = 0$ la distribución se reduce a la distribución uniforme (Figura 1.3).
- El principal uso de estas distribuciones es como aproximaciones de poca concentración a las distribuciones von Mises

$$\theta_i \sim C(\mu_i, \rho_i)(i = 1, 2) \Rightarrow \theta_1 + \theta_2 \sim C(\mu_1 + \mu_2, \rho_2\rho_2)$$

Notemos que el conjunto de distribuciones Cardioides es cerrado bajo convolución.

1.3.5 Distribución Normal proyectada

Las distribuciones en el círculo pueden ser obtenidas mediante proyección radial de la distribución en el plano. Sea x un vector aleatorio bidimensional, Pr(x=0)=0. Luego $||x||^{-1}x$ es un punto aleatorio sobre el círculo unidad.

$$p(\theta; \mu, \sum) =$$

$$\frac{\phi(\theta; 0, \sum) + |\sum|^{-1/2} D(\theta)\phi(D(\theta))\phi(|\sum|^{-1/2}(x^T \sum^{-1} x)^{-1/2}\mu \wedge x)}{x^T \sum^{-1} x}$$

$\phi(\cdot; 0, \sum)$ denota la función de densidad de $\mathbf{N}_2(0, \sum)$, ϕ y Φ denotan la función densidad y la función de densidad acumulada de N(0,1), $x = (cos\theta, sen\theta)^T$

$$D(\theta) = \frac{\mu^T \sum^{-1} x}{(x^T \sum^{-1} x)^{1/2}}$$

$$\mu \wedge x = \mu_1 sen\theta - \mu_2 cos\theta \ \mu = (\mu_1, \mu_2)^T$$

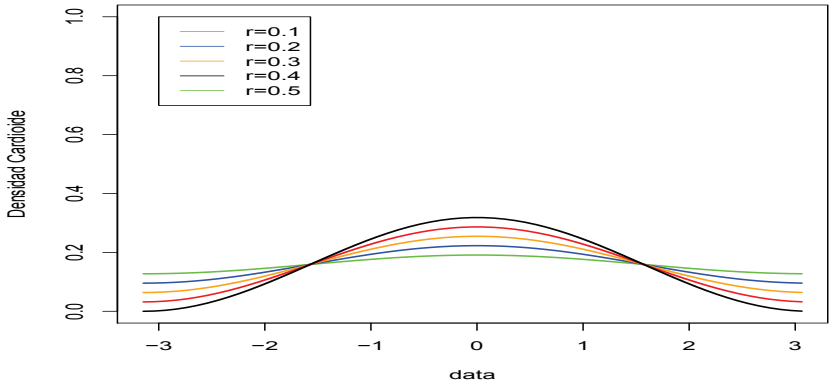

Figura 1.3: Representación de distribuciones cardiodes C(0,ρ), ρ=(0.1,0.2,0.3,0.4,0.5)

1.3.6 Distribución Wrapped

Dada una distribución en la línea, se puede envolver alrededor de la circunferencia del círculo de radio uno. Si x es una variable aleatoria en la línea, la variable aleatoria correspondiente x_w de la distribución wrapped viene dada por

$$x_w = x(mod 2\pi)$$

Si el círculo es identificado con el conjunto de números complejos de módulo la unidad luego el mapa wraping $x \to x_w$ pueden ser escrito como

$$x \to e^{2\pi i x}$$

Si x tiene como función de distribución F $\Rightarrow$ la función de distribución F_w de x_w viene dada por

$$F_w(\theta) = \sum_{\kappa=-\infty}^{\infty} \{F(\theta + 2\pi\kappa) - F(2\pi\kappa)\}$$

Si x tiene como función de masa de probilidad f $\Rightarrow$ la función de densidad f_w de x_w es

$$f_w(\theta) = \sum_{\kappa=-\infty}^{\infty} f(\theta + 2\pi\kappa)$$

Las distribuciones Wrapped tienen las siguientes propiedades:

1. $(x + y)_w = x_w + y_w$

2. Si la función característica de x es ϕ entonces la función característica $\{\phi_p : p = 0, \pm 1, \ldots\}$ de x_w es dada por $\phi_p = \phi(p)$

3. Si ϕ es integrable entonces x tiene función una densidad y

$$f_w(\theta) = \sum_{\kappa=-\infty}^{\infty} f(\theta + 2\pi\kappa) = \frac{1}{2\pi}\left[1 + 2\sum_{p=1}^{\infty}(\alpha_p cosp\theta + \beta_p senp\theta)\right]$$

$$\Phi(p) = \alpha_p + i\beta_p$$

4. Si x es infinitamente divisible luego x_w es infinitamente divisible.

Hay muchas distribuciones lineales las cuales pueden ser envueltas en cualquier distribución dada en el círculo. Sea g la función de densidad de una distribución en el círculo y define una función de densidad en la línea por

$$f(x) = p_r g(x)$$
$$2\pi r < x \leq 2\pi(r+1)\ r = 0, \pm 1, \pm 2, \ldots,$$

p_r son números no negativos tales que, $\sum_{r=-\infty}^{\infty} p_r = 1$. Luego, $f_w = g$.

Capítulo 2

Función de densidad circular

En este capítulo, descibiremos una estimación no paramétrica tipo núcleo de la densidad circular, sección 2.1. Al igual que en el caso lineal, la selección del parámetro de suavizado será crucial. Se estudiarán tres métodos de selección uno usando técnica plug-in y dos mediante validación cruzada que surgen de minimizar la función de pérdidida dada por minimizar el error cuadrártico medio (L2) o minimizar la función de pérdida dada por Kullback-Leibler.

2.1 Estimación no paramétrica tipo núcleo de la densidad circular

En el caso elemental de la estimación no paramétrica tipo núcleo de la densidad univariante f(x) usuando observaciones con valores reales $x_1, \ldots, x_n$, con núcleo K viene dada por el estimador de Parzen-Rosemblat

$$\hat{f}(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right)$$

dónde

- h es el parámetro de suavizado o ventana.
- Usualmente se supone que la función núcleo es una función de masa de probabilidad simétrica, por ejemplo, la densidad Gausiana.
- La etimación de la densidad tipo núcleo se extiende facilmente a datos circulares, aunque se debe tener cuidado en la selección de la función núcleo.

Sin embargo, cuando usamos datos en el círculo, no podemos emplear la distancia en el espacio Euclídeo, así que las diferencias $\theta - \theta_i$ serán remplazadas por el ángulo dado por la diferencia entre dos vectores,

$$d_i = ||\theta - \theta_i|| = min(|\theta - \theta_i|, 2\pi - |\theta - \theta_i|)$$

Esto también puede escribirse como $d_i = cos^{-1}(x^T x_i)$, dónde $x^T = (cos\theta, sen\theta)$ es un vector unitario. Una elección natural de la función núcleo es usar una función de densidad circular como la dada por la distribución de von Mises. Esto permite una representación alternativa de la estimación no paramétrica tipo núcleo de la densidad,

$$\hat{f}(x) = \frac{1}{nh} \sum K(\frac{1 - x^T x_i}{h})$$

Dada una muestra aleatoria de ángulos $\theta_1, \ldots, \theta_n \in [0, 2\pi]$. Si consideraremos que el estimador no parametrico de la densidad con función núcleo la distribución von Mises (2.1), viene dado por

$$\hat{f}(\theta, \nu) = \frac{c_0(\nu)}{n} \sum_{i=1}^{n} L(\nu \cos(\theta - \theta_i)) \tag{2.1}$$

$$c_0(\nu) = \frac{1}{2\pi I_0(\nu)} \; L(x) = e^x \tag{2.2}$$

- Dónde $I_r(\nu)$ es la función de Bessel modificada de orden r.
- El parámetro de concentración ν ahora ha asumido el papel de la inversa del parámetro de suavizado.

Para cualquier θ y entero j, los ángulos θ y $\theta + 2\pi j$ son idénticos una función de masa de probabilidad puede ser necesaria para el núcleo K. Una posible opción planteada por [9] es la distribución von Mises (1.9). El parámetro de concentración controla el grado de suavidad en estimación no paramétrica tipo núcleo de la densidad circular y es análogo al parámetro de suavizado excepto que valores altos de ν proporcionan menos suavidad y pequeños valores de ν conducen a mayor suavidad.

2.2 Selección del parámetro de suavizado.

A continuación, nos centramos en la elección del parámetro de suavizado. Existen varios métodos para seleccionar el parámetro de suavizaso subjetivamente, regla plug-in escala von Mises, o automáticamente, validación cruzada.

2.2.1 Regla plug-in escala von Mises

Si suponemos que $f(\cdot)$ es von Mises con concentración κ y (sin pérdida de generalidad) media circular $\mu = 0$. El error asintótico cuadrático medio integrado es

$$AMISE(\nu) = 3\kappa^2 I_2(2\kappa)/\left\{32\pi\nu^2 I_0(\kappa)^2\right\}$$

El error asintótico cuadrático medio integrado es de la forma $a\nu^{-2} + b\nu^{1/2}$ la cual puede ser minimizada diferenciando respecto de ν e igualando a cero. Esto permite una "Regla plug-in escala von Mises" para el parámetro de suavizado ν basado en la estimación de κ ([10].

$$\nu = [3n\hat{\kappa}^2 I_2(2\hat{\kappa})\left\{4\pi^{1/2} I_0(\hat{\kappa})^2\right\}^{-1}]^{2/5} \tag{2.3}$$

Por tanto el parámetro de suavizado, así estimado, tiene una expresión de la forma $\nu = Cn^{2/5}$ dónde

$$C = [3\hat{\kappa}^2 I_2(2\hat{\kappa})\left\{4\pi^{1/2} I_0(\hat{\kappa})^2\right\}^{-1}]^{2/5}$$

2.2.2 Validación cruzada

Introducimos la validación cruzada para minimizar la función pérdida dada por error cuadrático medio y la función de pérdida dada por Kullback-Leibler. Consideremos la estimación no paramétrica tipo núcleo de la densidad construída dejando fuera el valor θ_j de la muestra.

$$\hat{f}_j(\theta, \nu) = \frac{c_0(\nu)}{n} \sum_{i \neq j} L(\nu \cos(\theta - \theta_i))$$

$$c_0(\nu) = \frac{1}{2\pi I_0(\nu)} \; L(x) = e^x$$

Sea

$$cv_2(\nu) = 2n^{-1} \sum_{i=1}^{n} \hat{f}_i(\theta_i, \nu) - \int \hat{f}^2(\theta, \nu) d\theta$$

$$cv_{KL} = n^{-1} \sum_{i=1}^{n} \hat{f}_i(\theta_i, \nu)$$

Luego $-cv_2(\nu) + \int f^2$ y $-cv_{KL}(\nu) + \int flog(f)$ son estimaciones insesgadas de la perdida del error cuadrático $L_2(\nu)$ y la pérdida dada por Kullback-Leibler ([6]), respectivamente.

$$v_2 = argmin_{\nu \geq 0} - cv_2(\nu) \tag{2.4}$$

$$v_{KL} = argmin_{\nu \geq 0} - cv_{KL}(\nu) \tag{2.5}$$

$$\tag{2.6}$$

De estas dos expresiones obtenemos dos posibles valores para el parámetro de suavizado.

2.2.3 Ilustración

Con el fin de estudiar como se comportan los tres parámetros de suavizado, obtenidos en este capítulo, para la estimación no paramétrica tipo núcleo de la densidad (2.1). Hemos simulado una muestra de tamaño 100 que sigue una mixtura de von Mises y se ha representado su distribución teórica junto a las tres estimaciones de la densidad obtenidas tras seleccionar diferentes parámetros de suavizado (2.3, 2.4, 2.5).

En (Figura 2.1) se observa como las estimaciones no parámetricas tipo núcleo de la densidad (2.1) obtenidas tras estimar el parámetro de validacíón cruzada (2.4, 2.5) son mejores que la dada por el parámetro de suavizado obtenido mediante plug-in (2.3). Pues ambas estimaciones se parecen notablemente a la teórica.

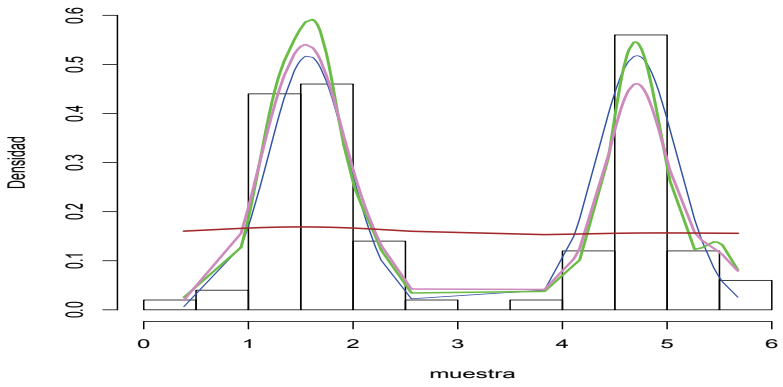

Figura 2.1: Repesentación de la densidad teorica de una mixtura de von Mises vM(π/2,15),vM(3π/2,15) y la estimación no paramétrica de la densidad seleccónando el parámetro de suavizado mediante las técnicas estuadiadas. L2 (verde), KL (violeta), plug-in (marrón)

Capítulo 3

Función de regresión circular-lineal

En este capítulo descibiremos una estimación no paramétrica tipo núcleo de la regresión circular-lineal, sección 3.1, que se puede utilizar para investigar la relación entre una variable explicativa direccional y una variable respuesta lineal. Mediante validación cruzada obtendremos razonables parámetros de suavizado, sección 3.2.

La estimación no paramétrica tipo núcleo de la función de regresión es comunmente usada como una forma de resumir la relación entre dos variables sin requerir el supuesto de una forma paramétrica para describir esta relación.

3.1 Estimación no paramétrica tipo núcleo de la regresión circular-lineal

Dadas n observaciones de una variable direccional explicativa $\theta_1, \ldots, \theta_n$ y una variable respuesta lineal $y_1, \ldots, y_n$, suponemos que $y_i = m(\theta_i) + \epsilon i$, dónde *epsiloni* son variables con media cero, independientes e identicamente distribuídos. Luego, la estimación no paramétrica tipo núcleo de la regresión circular-lineal, $m(\theta)$, viene dada por ([5])

$$\hat{m}(\theta; \kappa) = \frac{\sum_{i=1}^{n} y_i g(\theta - \theta_i, 0, \kappa)}{\sum_{i=1}^{n} g(\theta - \theta_i, 0, \kappa)} \tag{3.1}$$

La cual es una estimación (3.1) análoga al estimador de Nadaraya-Watson para caso lineal.

3.2 Selección del parámetro de suavizado

La elección del grado de suavizado es crucial para la estimación no paramétrica tipo núcleo de la regresión, asi como para la estimación no paramétrica tipo núcleo de la densidad. Existen varios métodos para seleccionar el parámetro de suavizado subjetivamente o automáticamente y pueden ser extendidos al caso de la estimación no paramétrica tipo núcleo de la regresión.

Por ejemplo, cuando el núcleo de suavizado evaluado sobre datos lineales es un núcleo Gausiano, Silverman recomienda un parámetro de suavizado de $0.9\hat{\sigma}n^{-1/5}$ y el valor $h_s = 1.06\hat{\sigma}n^{-1/5}$ es sugerido por Scott. (En ambos casos $\hat{\sigma}$ es tipicamente dado por el mínimo de la desviación típica muestral y el rango intercuartílico dividido por 1.34). Ambas pueden estar conectadas con el parámetro κ en la distribución de von Mises usando el siguiente enlace conocido entre la distribución de von Mises y la distribución Gausiana propiedad de la distribución von Mises [9].
Si θ es distribuída de acuerdo con la distribución von Mises centrada en μ y con parámetro κ, luego

$$\kappa^{-1/2}(\theta - \mu) \rightarrow_D N(0,1) \ \kappa \rightarrow \infty$$

Esto sugiere la elección del parámetro de concentración κ mediante

$$\kappa = \frac{1}{h_s^2} \tag{3.2}$$

Otra opción es seleccionar κ por validación cruzada de mínimos cuadráticos (LCV). Como en el caso lineal, escogemos κ de modo que minimize a la función dónde son las estimaciones sin el dato j-ésimo.

$$CV(\kappa) = n^{-1} \sum_{j=1}^{n} [y_j - \hat{m}^{-j}(\theta_j; \kappa)]^2 \tag{3.3}$$

dónde

$$\hat{m}^{-j}(\theta_j; \kappa) = \frac{\sum_{i \neq j}^{n} y_i g(\theta_j - \theta_i, 0, \kappa)}{\sum_{i \neq j}^{n} g(\theta - \theta_i, 0, \kappa)}$$

La estimación no paramétrica tipo núcleo seleccionando el parámetro de suavizado mediante LCV puede ser inconsistente bajo una variedad de circunstancias. En particiular, para datos discretos con múltiples valores repetidos, validación cruzada tiende a sugerir parámetros de suavizado que suavizan muy poco, es decir, κ tiende a ser muy grande. Por tanto valores de κ escogidos de acuerdo (3.2) serán preferidos en este caso.

Capítulo 4

El modelo de Möbius de series de tiempo

En este capítulo un modelo de regresión circular propuesto en [3] es adaptado al contexto de series de tiempo siguiendo la metodología propuesta en ([7]), sección 4.1. La distribución de θ_t dado θ_{t-1} es modelada usando una distribución von Mises particular. En la sección 4.1 comentaremos el modelo de regresión estudiado por [3] y lo adaptaremos a una serie de tiempo. La función de máxima verosimilitud (condicionada a la primera observación) es obtenida en la sección 4.2 y la estimación de los parámetros.

4.1 El modelo Möbius de series de tiempo

La componente determinística del modelo de regresión estudiado por [3] une a la variable angular dependiente v con la variable angular independiente u mediante

$$tan\frac{1}{2}(v-\beta) = \omega \tan\frac{1}{2}(u-\alpha) \tag{4.1}$$

dónde $\omega \in [-1,1]$ es el parámetro que determina la pendiente y $-\pi \leq \alpha, \beta < \pi$ son parámetros que determinan la localización angular. La ecuación (4.1) nos conduce a

$$v = \beta + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(u-\alpha)\right\} \tag{4.2}$$

Luego, aplicaremos la ecuación (4.2) a las series de datos temporales, reemplazando el ángulo v por θ_t y el ángulo u por θ_{t-1}, $t = 2, \ldots, n$. Esta sustitución nos sugiere la existencia de un único parámetro de localización, $\alpha = \beta$, obteniendo así

$$\tan\frac{1}{2}(\theta_t-\alpha) = \omega \tan\frac{1}{2}(\theta_{t-1}-\alpha) \tag{4.3}$$

y

$$\theta_t = \alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\} \tag{4.4}$$

ecuaciones análogas a (4.1) y (4.2).

Para el modelo de serie de tiempo dado en (4.4) se asume que $\theta_t|\theta_{t-1}$ sigue una distribución von Mises

$$\theta_t|\theta_{t-1} \sim M(\alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\}, k) \tag{4.5}$$

Por tanto el modelo de serie de tiempo se convierte en

$$\theta_t = \alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\} + \epsilon_t \tag{4.6}$$

dónde $\epsilon_t \sim M(0, k)$. Nos referiremos a la distribución condicionada de la media direccional de θ_t dado θ_{t-1} como μ_t. Esto es

$$\mu_t = \alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\} \tag{4.7}$$

El valor de ω se restringe al intervalo [-1,1] así que α es único de identificación. Esto es, restamos π al valor de α en la ecuación (4.3). Luego

$$\tan\frac{1}{2}(\theta_t - \alpha + pi) = \omega \tan\frac{1}{2}(\theta_{t-1} - \alpha + pi) \tag{4.8}$$

Ahora tengamos en cuenta que $\frac{1}{2}(\phi + \pi) = -cot\frac{1}{2}\phi$, (4.8) es equivalente a

$$-cot\frac{1}{2}(\theta_t - \alpha) = -\omega cot\frac{1}{2}(\theta_{t-1} - \alpha)$$

de lo que se obtiene

$$\tan\frac{1}{2}(\theta_t - \alpha) = \frac{1}{\omega}\tan\frac{1}{2}(\theta_{t-1} - \alpha) \tag{4.9}$$

La equivalencia de (4.8) y (4.9) muestra que, si ω no estuviese restringido al intervalo [-1,1], y si $(\hat{\alpha}, \hat{\omega})$ es una solución de (refmodel3), luego $(\hat{\alpha} - \pi, \frac{1}{\hat{\omega}})$ sería una solución equivalente.

4.2 Función de máxima verosimilitud

Como

$$\theta_t|\theta_{t-1} \sim M(\alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1}-\alpha)\right\}, k)$$

se obtiene que

$$f(\theta_t|\theta_{t-1}) = 2\pi I_0(\kappa)^{-1} exp\kappa\cos(\theta_t - \mu_t)$$

dónde μ_t es dada por (4.7)
Luego la función de masa de probabilidad de $\theta_2, \ldots, \theta_n$ dado θ_1 es

$$\begin{aligned} f(\theta_2,\ldots,\theta_n|\theta_1) &= f(\theta_2\theta_1)f(\theta_3,\ldots,\theta_n|\theta_1,\theta_2) \\ &= f(\theta_2|\theta_1)f(\theta_3|\theta_1,\theta_2)f(\theta_4,\ldots,\theta_n|\theta_1,\theta_2,\theta_3) \\ = \ldots = f(\theta_2|\theta_1)f(\theta_3|\theta_1,\theta_2)f(\theta_4|\theta_1,\theta_2,\theta_3)\cdots f(\theta_n|\theta_1,\ldots,\theta_{n-1}) \end{aligned}$$

Pero en ecuación (4.5) el valor de θ_t depende solo del valor θ_{t-1}. Además,

$$f(\theta_t|\theta_1,\ldots,\theta_{t-1}) = f(\theta_t|\theta_{t-1},$$

$\forall t = 2,\ldots,n$. La función de máxima verosimilitud condicionada es

$$L_C(\alpha,\omega,\kappa) = \{2\pi I_0(\kappa)\}^{-(n-1)} exp\left\{\kappa\sum_{t=2}^{n} cos\left[\theta_t - \alpha - 2\tan^{-1}\left\{\omega\tan\frac{1}{2}(\theta_{t-1}-\alpha)\right\}\right]\right\}$$

dando la verosimilitud

$$l_C(\alpha,\omega,\kappa) = const.-(n-1)logI_0\kappa+\kappa\sum_{t=2}^{n}\cos\left[\theta_t - \alpha - 2\tan^{-1}\left\{\omega\tan\frac{1}{2}(\theta_{t-1}-\alpha)\right\}\right].$$

Esto se máximiza respècto los parámetros α y ω desconocidos

$$l_C(\alpha,\omega) = \sum_{t=2}^{n} cos\left[\theta_t - \alpha - 2\tan^{-1}\left\{\omega\tan\frac{1}{2}(\theta_{t-1}-\alpha)\right\}\right]. \tag{4.10}$$

La maximización de (4.10) la haremos usando la función *nlm* de R pero minimizando la función $-l_C(\alpha,\omega)$ usando un algoritmo tipo Newton. Una vez que $l(\alpha,\omega)$ ha sido máximizada con respecto a α y ω, una aproximación mediante verosimilitud puede ser usada para obtener una estimación de máxima verosimilitud de κ, maximizando

$$l_C(\hat{\alpha},\hat{\omega},\kappa) = const. - (n-1)logI_0\kappa + \kappa l_{\hat{\alpha},\hat{\omega}} \tag{4.11}$$

respecto de κ. Diferenciando (4.11) respecto de κ y teniendo en cuenta que $d(I_0(\kappa))/d\kappa = I_1(\kappa)$, la función de Bessel de primer tipo y de orden uno, obtenemos

$$\frac{\partial}{\partial\kappa}[l_C(\hat{\alpha},\hat{\omega},\kappa)] = -(n-1)\frac{I_1(\kappa)}{I_0(\kappa)} + l_{\hat{\alpha},\hat{\omega}},$$

así que κ es la solución de

$$\frac{I_1(\hat{\kappa})}{I_0(\kappa)} = \frac{l_{\hat{\alpha},\hat{\omega}}}{n-1}.$$