

NPCirc: técnicas nonparamétricas para a análise de datos circulares en R

María Oliveira, Rosa M. Crujeiras e Alberto Rodríguez-Casal

Departamento de Estatística e Investigación Operativa
Universidade de Santiago de Compostela

RESUMO

O paquete **NPCirc** para **R** inclúe diversos métodos nonparamétricos para a estimación da función de densidade e regresión con datos circulares. No contexto da densidade, implementa o estimador tipo núcleo da densidade circular, xunto con diferentes selectores do parámetro de suavizado. Para a regresión con resposta escalar e covariable circular, inclúense os estimadores de Nadaraya–Watson e Local Lineal adaptados á natureza circular da covariable, xunto coas respectivas regras de validación cruzada por mínimos cadrados para a selección do parámetro de suavizado. Ademais, a metodoloxía **CircSiZer**, para a determinación de características significativas na curva suavizada, tanto en densidade como en regresión, tamén está dispoñible na librería **NPCirc**. Finalmente, o paquete tamén inclúe funcións para xerar observacións e calcular a función de densidade de mesturas de distribucións circulares, así como algúns conxuntos de datos reais.

Palabras e frases chave: **CircSiZer**, datos circulares, función de densidade, función de regresión, métodos nonparamétricos.

1. INTRODUCCIÓN

Os datos circulares son un caso particular de datos direccionais onde as observacións son direccións en dúas dimensións que adoitan expresarse mediante ángulos. Debido á súa natureza circular, as técnicas estatísticas utilizadas para a análise de datos lineais non son válidas para a análise deste tipo de datos e polo tanto é necesario desenvolver métodos e software axeitados.

Con esta finalidade, créase a librería **NPCirc** (Oliveira *et al.*, 2013b) para o paquete estatístico **R** (R Development Core Team, 2012) cuxo principal obxectivo é dotar aos usuarios de **R** dun amplo conxunto de funcións para a estimación nonparamétrica da función de densidade e de regresión con datos circulares e así complementar os paquetes xa existentes para a análise deste tipo de datos. Concretamente, a librería **NPCirc** implementa:

- O estimador tipo núcleo circular para a estimación da función de densidade dunha mostra de datos circulares e varios métodos para a selección do parámetro de suavizado.
- No contexto da regresión, para o caso dunha variable resposta escalar e unha covariable circular, a librería inclúe os estimadores tipo núcleo de Nadaraya–Watson e Local–Lineal convenientemente adaptados á natureza circular da covariable e a regra de validación cruzada por mínimos cadrados para a selección do parámetro de suavizado.
- Unha técnica nonparamétrica para a análise exploratoria de datos circulares, denominada **CircSiZer**, que permite coñecer que características observadas na curva suavizada (densidade ou regresión), son estatisticamente significativas e cales se poden atribuír á variabilidade dos datos.
- Funcións para xerar mostras aleatoria e calcular a función de densidade de modelos circulares e mesturas.

2. DESCRICIÓN E ILUSTRACIÓN DA LIBRARÍA NPCirc

Nesta sección descríbense e ilústranse cada unha das funcións implementadas no paquete NPCirc. Na Táboa 1 preséntase a lista completa de funcións e conxuntos de datos incluídos na librería.

Conxuntos de datos	Descrición
cross.beds1	Datos de azimuths de estratos cruzados (I)
cross.beds2	Datos de azimuths de estratos cruzados (II)
cycle.changes	Datos relativos ás horas ás que a temperatura pasa de positiva a negativa e viceversa en Monte Alvear (Arxentina)
dragonfly	Datos de orientación de libélulas.
speed.wind, speed.wind2	Datos da dirección e velocidade do vento na costa atlántica de Galicia.
temp.wind	Datos da temperatura e dirección do vento en Vinciguerra (Tierra del Fuego, Arxentina)
Funcións	Descrición
circsizer.density	CircSiZer para densidade
circsizer.regression	CircSiZer para regresión
dcircmix	Función de densidade de mixturas de distribucións circulares
rcircmix	Xeración de mostras aleatorias de mixturas de distribucións circulares
kern.den.circ	Estimador tipo núcleo da función de densidade circular
kern.reg.circ	Estimadores de Nadaraya–Watson e Local–Lineal para covariable circular e resposta escalar
nu.CV	Reglas de validación cruzada por máxima verosimilitude e mínimos cadrados
nu.LSCV.reg	Regla de validación cruzada por mínimos cadrados para regresión.
nu.boot	Método bootstrap
nu.pi	Regla plug-in
nu.rt	Regla do pulgar

Táboa 1: Resumo dos contidos do paquete NPCirc: conxuntos de datos e funcións.

As funcións dcircmix e rcircmix

A función `dcircmix` permite calcular a función de densidade das distribucións circulares: *von Mises*, *Cardioid*, *wrapped Cauchy*, *wrapped normal* e *wrapped skew-normal* e de mixturas destas distribucións (ver Jammalamadaka e SenGupta, 2001 e Pewsey, 2000). A función `rcircmix` permite xerar mostras aleatorias de calquera destas distribucións. Ambas funcións teñen un argumento chamado `model` que permite especificar un modelo entre os modelos definidos en Oliveira *et al.* (2012). Por exemplo, a función de densidade do modelo 20 nun grella de 500 puntos entre 0 e 2π pode obterse co código:

```
R> t <- seq(0, 2*pi, length=500)
R> f20 <- dcircmix(x=t, model=20, dist=NULL, param=NULL)
```

e unha mostra aleatoria de tamaño 100 deste modelo pode obterse mediante:

```
R> data20 <- rcircmix(n=100, model=20, dist=NULL, param=NULL)
```

Utilizando funcións do paquete `circular` (Lund e Agostinelli, 2011), coas liñas de código

```
R> plot(circular(data20), shrink=1.2)
R> lines(circular(t), f20)
```

pódese representar a función de densidade e a mostra xerada, resultando a Figura 1 (panel esquerdo).

Para calcular a función de densidade e xerar mostras aleatorias dun modelo diferente aos pre-definidos basta con especificar os argumentos `dist` e `param` destas funcións. Por exemplo, a función de densidade dunha mixtura na mesma proporción dunha distribución *wrapped normal* $WN(\pi/2, 0.7)$ e unha distribución *wrapped skew-normal* $WSN(\pi, 1, 10)$ obtense co código:

```
R> fmix <- dcircmix(x=t, model=NULL, dist=c("wn", "wsn"),
+ param=list(p=c(0.5, 0.5), mu=c(pi/2, pi), con=c(0.7, 1), sk=c(0, 10)))
```

e unha mostra aleatoria deste modelo mediante o código

```
R> datamix <- rcircmix(n=100, model=NULL, dist=c("wn", "wsn"),
+ param=list(p=c(0.5, 0.5), mu=c(pi/2, pi), con=c(0.7, 1), sk=c(0, 10)))
```

A correspondente función de densidade e a mostra aleatoria están representadas na Figura 1 (panel dereito).

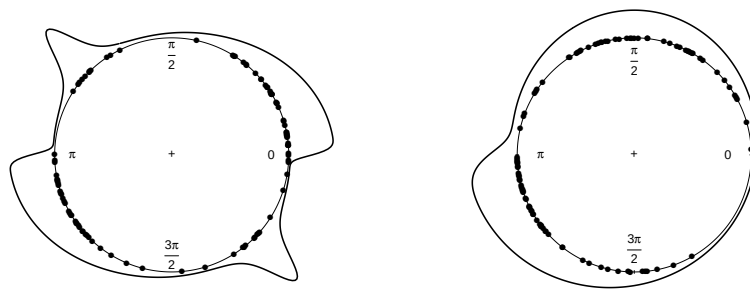


Figura 1: Panel esquerdo: Mostra aleatoria de tamaño $n = 100$ (puntos sobre o círculo) xerada do modelo 20 de Oliveira *et al.* (2012) e función de densidade de dito modelo. Panel dereito: Mostra aleatoria de tamaño $n = 100$ (puntos sobre o círculo) xerada a partir dunha mixtura na mesma proporción dunha distribución *wrapped normal* $WN(\pi/2, 0.7)$ e unha distribución *wrapped skew-normal* $WSN(\pi, 1, 10)$ e función de densidade correspondente a este modelo.

Estimación da densidade

A función `kern.den.circ` implementa o estimador tipo núcleo circular da función de densidade con núcleo *von Mises* (ver por exemplo Oliveira *et al.*, 2012). Esta función devolve o valor do estimador avaliado nos puntos especificados polo argumento `t`, construído a partir da mostra de ángulos (entre 0 e 2π) especificada polo argumento `x` e utilizando como valor do parámetro de suavizado o valor indicadado polo argumento `nu`. Para a mostra xerada anteriormente, o estimador tipo núcleo da función da densidade, cun valor do parámetro de suavizado igual a 40 e avaliado nos puntos da mostra, obtense como segue:

```
R> est20 <- kern.den.circ(x=data20, t=data20, nu=40)
```

A Figura 2 (panel dereito) amosa o resultado. O valor do parámetro `nu` na función `kern.den.circ` pode ser fixado polo usuario ou ben seleccionado mediante algún dos métodos implementados (ver Oliveira *et al.*, 2012 para unha comparación das distintas técnicas):

`nu.boot` Implementa o procedemento bootstrap proposto por Di Marzio *et al.* (2011).

```
R> nu.boot(x=data20)
```

```
[1] 0.06544481
```

`nu.CV` Implementa as regras de validación cruzada por mínimos cadrados e máxima verosimilitude propostas por Hall *et al.* (1987). A regra de validación cruzada a usar, especifícase co argumento `method`.

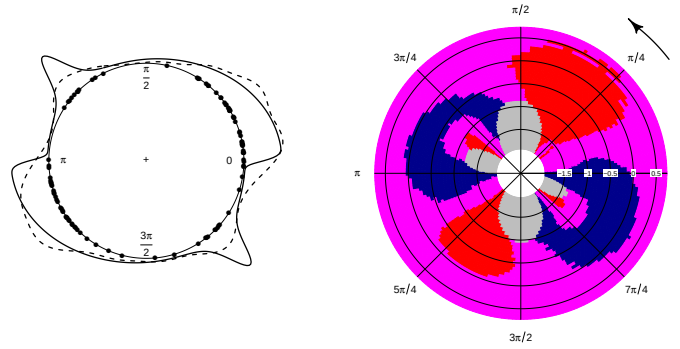


Figura 2: Panel esquerdo: Mostra aleatoria de tamaño $n = 100$ (puntos sobre o círculo) xerada do modelo 20 de Oliveira *et al.* (2012), función de densidade de dito modelo (liña continua) e estimador da función de densidade da mostra (liña discontinua). Panel dereito: CircSiZer map para densidade.

```
R> nu.CV(x=data20, method="LCV")
```

```
[1] 31.03298
```

```
R> nu.CV(x=data20, method="LSCV")
```

```
[1] 43.03189
```

nu.pi Implementa a regra plug-in proposta por Oliveira *et al.* (2012).

```
R> nu.pi(x=data20)
```

```
[1] 46.03157
```

nu.rt Por defecto implementa a regra do pulgar proposta por Taylor (2008).

```
R> nu.rt(x=data20)
```

```
[1] 0.2017836
```

A función `circsizer.density` implementa o método CircSiZer para o contexto da densidade (ver Oliveira *et al.*, 2013c). Os principais argumentos desta función son: **x**, a mostra de ángulos e **NU**, unha grella de valores para o parámetro de suavizado. Esta función devolve un gráfico, coñecido como CircSiZer map, que por medio de diferentes cores permite identificar rexións de crecemento/decrecemento significativo da curva suavizada. Por exemplo, para a mostra xerada a partir do modelo 20, o correspondente CircSiZer map (ver panel dereito da Figura 2) obtense co código:

```
R> circsizer.density(x=data20, NU=seq(0,100,by=5), ngrid=250, alpha=0.05, B=500,
type=3, raw.data=FALSE, log.scale=TRUE, zero=0, clockwise=FALSE, title=NULL,
labels=NULL, label.pos=NULL, rad.pos=NULL)
```

Cada anel do CircSiZer map indica o comportamento da curva suavizada no intervalo $[0, 2\pi)$ para cada valor do parámetro de suavizado considerado (indicado ao longo do radio en escala $-\log_{10}$). Dado un valor do parámetro de suavizado, a cor azul (vermella) indica os puntos do intervalo $[0, 2\pi)$ onde a correspondente curva suavizada é significativamente crecente (decrecente) e a cor rosa indica onde non se pode determinar crecemento/decrecemento significativo da curva suavizada. Aqueles puntos onde non hai suficientes datos para establecer significación son pintados de cor gris. Deste xeito, unha pico significativo pode indentificarse mediante o patrón azul-vermello e un val mediante o patrón vermello-azul, tomando como sentido de rotación o indicado pola frecha.

Así, na Figura 2, obsérvase que para valores pequenos do parámetro de suavizado só dúas modas son identificadas como significativas mentres que para valores máis grandes as catro modas que presenta o modelo son significativas.

Estimación da función de regresión

A función `kern.reg.circ` implementa os estimadores Local linear e Nadaraya–Watson para a estimación da función de regresión no caso dunha covariable circular e resposta escalar (ver Di Marzio *et al.*, 2009 e Oliveira *et al.*, 2013a). Os argumentos desta función son: `x`, a mostra de ángulos en radians (entre 0 e 2π) para a covariable circular; `y`, os valores para a variable dependente (lineal); `t`, o vector de ángulos (en radians) onde evaluar o estimador; `nu`, o valor do parámetro de suavizado a usar; `method`, indica o estimador a usar ("LL" par o estimador local linear e "NW" para o estimador Nadaraya–Watson).

O valor do parámetro de suavizado `nu` pode ser fixado polo usuario ou ben obtido chamando a función `nu.LSCV.reg`, que implementa a regra de validación cruzada por mínimos cadrados. Os argumentos `x`, `y` e `method` desta función teñen o mesmo significado que para a función `kern.reg.circ`.

Para ilustrar o funcionamento destas funcións considérase o conxunto de datos `wind.speed2` (descargados de <http://www.puertos.es>) que contén medidas da dirección e velocidade do vento na costa atlántica de Galicia durante os meses de Novembro–Febreiro correspondentes aos anos 2003–2012. Na Figura 3 (panel esquerdo) amósanse os estimadores de Nadaraya–Watson (curva vermella) e Local Linear (curva verde) para o modelo de regresión da velocidade do vento sobre a dirección do vento, obtidos co código:

```
R> data("speed.wind2")
R> dir <- rad(speed.wind2$Direction)
R> speed <- speed.wind2$Speed
R> t <- seq(0, 2*pi, length=200)
R> nuNW <- nu.LSCV.reg(x=dir, y=speed, method="NW")
R> nuLL <- nu.LSCV.reg(x=dir, y=speed, method="LL")
R> estLL <- kern.reg.circ(x=dir, y=speed, t=t, nu=nuLL, method="LL")
R> estNW <- kern.reg.circ(x=dir, y=speed, t=t, nu=nuNW, method="NW")
```

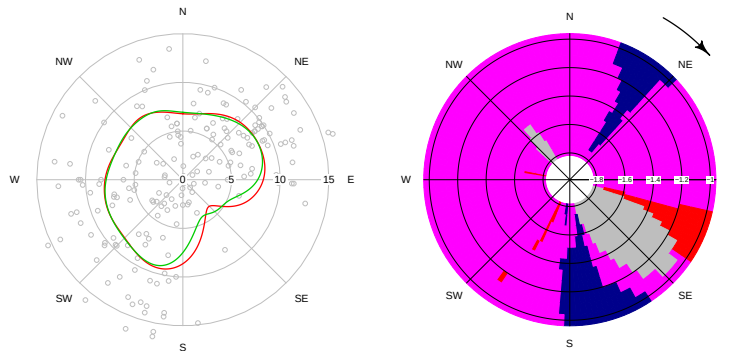


Figura 3: Panel esquerdo: Estimadores de Nadaraya–Watson (curva vermella) e Local Lineal (curva verde). Panel dereito: CircSiZer map para a regresión.

A función `circsizer.regression` proporciona o CircSiZer map para a regresión (ver Oliveira *et al.*, 2013c). Os principais argumentos desta función son: `x`, a mostra de ángulos en radians (entre 0 e 2π) para a covariable circular; `y`, a mostra de ángulo para a variable dependente `y` e `NU`, unha grella de valores para o parámetro de suavizado. A Figura 3 (panel dereito) mostra o correspondente CircSiZer map para os datos de dirección e velocidade do vento, obtido co código:

```
R> circsizer.regression(dir, vel, NU=seq(10,60,by=5), type=1)
```

Neste caso, as rexións azuis do CircSiZer map mostran que a velocidade do vento aumenta cos ventos do NE e S–SW. A zona gris indica que os ventos do SE son pouco frecuentes.

AGRADECIMENTOS

Este traballo foi financiado polo Ministerio de Ciencia e Innovación, a través do proxecto MTM2008-03010 e pola rede IAP StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences), do goberno belga. Os autores agradecen aos profesores A. Pewsey e A. Pérez-Alberti polos datos facilitados.

REFERENCIAS

- Lund, U. and Agostinelli, R. (2011) `circular`: Circular Statistics. R package version 0.4-3. URL <http://www.R-project.org/package=circular>.
- Di Marzio, M., Panzera A. and Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters*, **79**, 2066–2075.
- Hall, P., Watson, G. P. and Cabrera, J. (1987). Kernel density estimation for spherical data. *Biometrika*, **74**, 751–762.
- Jammalamadaka, S. R. and SenGupta, A. (2001) *Topics in Circular Statistics*. World Scientific, Singapore.
- Oliveira, M., Crujeiras, R. M. and Rodríguez-Casal, A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis*, **56**, 3898–3908.
- Oliveira, M., Crujeiras, R. M. and Rodríguez-Casal, A. (2013a). Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.
- Oliveira, M., Crujeiras, R.M. and Rodríguez-Casal, A. (2013b) `NPCirc`: Nonparametric circular methods. R package version 1.0.0. URL <http://www.R-project.org/package=NPCCirc>.
- Oliveira, M., Crujeiras, R. M. and Rodríguez-Casal, A. (2013c) `CircSiZer`: an exploratory tool for circular data. *Journal of Environmental and Ecological Statistics*. DOI 10.1007/s10651-013-0249-0.
- Pewsey, A. (2000) The wrapped skew-normal distribution on the circle. *Communications in Statistics - Theory and Methods*, **29**, 2459–2472.
- R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493–3500.