Assessment of significant features in nonparametric curve estimates with circular data

Rosa M. Crujeiras, María Oliveira and Alberto Rodríguez–Casal

Department of Statistics and Operations Research
University of Santiago de Compostela

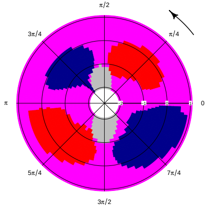
Department of Mathematical Sciences
Durham University

Assessment of significant features in nonparametric curve estimates with circular data

Rosa M. Crujeiras, María Oliveira and Alberto Rodríguez–Casal

Department of Statistics and Operations Research
University of Santiago de Compostela

Department of Mathematical Sciences
Durham University



Circular kernel density estimator

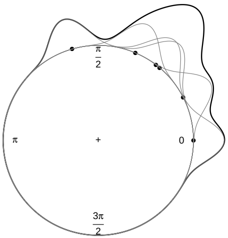
Given a random sample of angles $\Theta_1, \ldots, \Theta_n \in [0, 2\pi)$ from some unknown circular density f, the circular KDE is given by:

$$\hat{f}(\theta; \nu) = \frac{1}{n} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i)$$

K_ν is a circular kernel function with concentration parameter $\nu > 0$.

Taking the von Mises density as kernel:

$$\hat{f}(\theta; \nu) = \frac{1}{n 2\pi I_0(\nu)} \sum_{i=1}^{n} e^{\nu \cos(\theta - \Theta_i)}$$



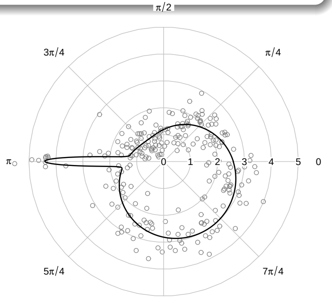
Circular–linear regression model

Let $\{(\Theta_i, Y_i),\ i = 1, \ldots, n\}$ be a random sample from (Θ, Y) a circular and a linear random variables, respectively. The relation between these variables can be modeled by

$$Y_i = f(\Theta_i) + \sigma(\Theta_i)\varepsilon_i, \quad i = 1, \ldots, n,$$

where f denotes the regression function.

- Kernel smoother
- Spline smoother



Local linear estimator

The local linear regression estimate for $f(\theta)$ and $f'(\theta)$ at an angle θ are given by $\hat{f}(\theta;\nu) = \hat{a}$ and $\hat{f}'(\theta;\nu) = \hat{b}$, where

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_{i=1}^{n} K_\nu(\theta - \Theta_i) \left[Y_i - (a + b\sin(\theta - \Theta_i))\right]^2$$

K_ν is a circular kernel function with concentration parameter ν .

Di Marzio, M., Panzera A. and Taylor, C.C. (2009)
Local polynomial regression for circular predictors.
Statistics & Probability Letters, **79**, 2066–2075.

Periodic smoothing spline estimator

The periodic smoothing spline estimator is given by the smooth function $\hat{f}_\lambda$ that minimizes the penalized least squares criterion

$$S(g) = \sum_{i=1}^{n} [Y_i - g(\Theta_i)]^2 + \lambda \int_0^T [g''(\theta)]^2 \, d\theta$$

over the class of twice c.d. periodic functions with period $T = 2\pi$.

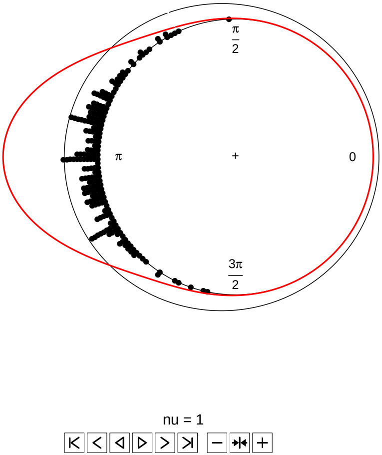
- It is shown that $\hat{f}_\lambda$ is a periodic cubic spline on $[\Theta_1, \Theta_{n+1}]$ with knots at the points Θ_i, $i = 1, \ldots, n+1$, where $\Theta_{n+1} = \Theta_1 + T$.
- The parameter λ plays the role of the smoothing parameter.

Cogburn, R. and Davis, H.T. (1974)
Periodic splines and spectral estimation.
Annals of Statistics, **2**, 1108–1126.

The Holy Grail of smoothing

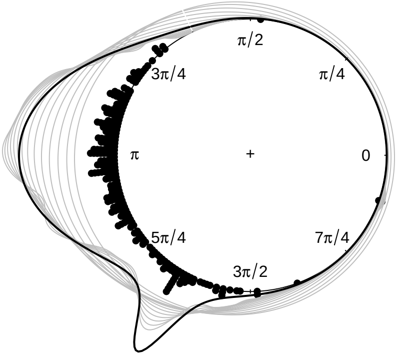
Finding a *suitable bandwidth* for smoothing a density or a regression curve:

- Plug–in rules
- Cross–validation



Beyond *bandwidth* selection...

Forget about recovering the original curve... and try to identify significant underlying structures, such as peaks and valleys in the density or regression.



Family of smoothers for a density model

The idea of CircSiZer method

- CircSiZer is an adaptation to circular data of the original SiZer proposed by Chaudhuri and Marron (1999) for linear data.
- CircSiZer considers nonparametric curve estimates for a wide range of smoothing parameters (τ).
- CircSiZer addresses the question of which features are really there.
- CircSiZer assesses the significance of such features by constructing confidence intervals for the derivative of the smoothed underlying curve at each location $\theta \in [0, 2\pi)$ and scale τ, $f'(\theta;\tau) \equiv \mathbb{E}(\hat{f}'(\theta;\tau))$.

Chaudhuri, P. and Marron, J. S. (1999)
SiZer for exploration of structures in curves.
Journal of the American Statistical Association, 94, 807–823.

Confidence interval

Given a significance level α and for a fixed value of $\tau > 0$ and with $\theta \in [0, 2\pi)$, confidence intervals are of the form

$$\left(\hat{f}'(\theta;\tau) - q^{(1-\alpha/2)} \cdot \widehat{\mathsf{sd}}(\hat{f}'(\theta;\tau)), \hat{f}'(\theta;\tau) - q^{(\alpha/2)} \cdot \widehat{\mathsf{sd}}(\hat{f}'(\theta;\tau))\right)$$

- $\hat{f}'(\theta;\tau)$ is the estimator of the derivative of the curve.
- $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are appropriate quantiles.
- $\widehat{\mathsf{sd}}(\hat{f}'(\theta;\tau))$ is an estimator of the std of $\hat{f}'(\theta;\tau)$.

Oliveira, M., RMC and Rodríguez–Casal, A. (2014)
CircSiZer: an exploratory tool for circular data.
Journal of Environmental and Ecological Statistics, 21, 143–159.

Quantiles: normal approximation

- Pointwise normal quantiles
  $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are the quantiles of order $(1-\alpha/2)$ and $\alpha/2$ of the standard normal distribution.

Quantiles: normal approximation

- Pointwise normal quantiles
  $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are the quantiles of order $(1-\alpha/2)$ and $\alpha/2$ of the standard normal distribution.

- Simultaneous normal quantiles
  $q^{(1-\alpha/2)} = -q^{(\alpha/2)} = \Phi^{-1}\left\{\frac{1+(1-\alpha)^{1/m(\tau)}}{2}\right\}$ where Φ^{-1} is the inverse of the standard normal distribution and

$$m(\tau) = \frac{n}{\text{avg}_{\theta \in \mathcal{D}_\tau} ESS(\theta;\tau)}$$

  where $ESS(\theta;\tau)$ is the Effective Sample Size for the pair (θ,τ) and $\mathcal{D}_\tau = \{\theta : ESS(\theta;\tau) \geq 5\}$.

Quantiles: bootstrap method

- Pointwise bootstrap quantiles
  $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are the sample quantiles of order $(1-\alpha/2)$ and $\alpha/2$ of $Z_1^*(\theta;\tau),\ldots,Z_B^*(\theta;\tau)$ where

$$Z_b^*(\theta;\tau)=\frac{\hat{f}'(\theta;\tau)^{*b}-\hat{f}'(\theta;\tau)}{\widehat{\mathsf{sd}}(\hat{f}'(\theta;\tau)^{*b})},\ b=1,\ldots,B$$

Quantiles: bootstrap method

- Pointwise bootstrap quantiles
  $q^{(1-\alpha/2)}$ and $q^{(\alpha/2)}$ are the sample quantiles of order $(1-\alpha/2)$ and $\alpha/2$ of $Z_1^*(\theta;\tau),\ldots,Z_B^*(\theta;\tau)$ where

$$Z_b^*(\theta;\tau) = \frac{\hat{f}'(\theta;\tau)^{*b} - \hat{f}'(\theta;\tau)}{\widehat{\mathsf{sd}}(\hat{f}'(\theta;\tau)^{*b})},\ b = 1,\ldots,B$$

- Simultaneous bootstrap quantiles
  $q^{(1-\alpha/2)}$ is the sample quantile of order $(1-\alpha/2)$ of $Z_{\mathsf{sup}}^{*1},\ldots,Z_{\mathsf{sup}}^{*B}$ and $q^{(\alpha/2)}$ is the sample quantile of order $\alpha/2$ of $Z_{\mathsf{inf}}^{*1},\ldots,Z_{\mathsf{inf}}^{*B}$ where

$$Z_{\mathsf{inf}}^{*b} = \inf_{\theta\in\mathcal{D}_\tau^*} Z_b^*(\theta;\tau) \text{ and } Z_{\mathsf{sup}}^{*b} = \sup_{\theta\in\mathcal{D}_\tau^*} Z_b^*(\theta;\tau), \quad b = 1,\ldots,B$$

Standard deviation (density)

$$\widehat{\text{var}}\left(\hat{f}'(\theta;\nu)\right) = \widehat{\text{var}}\left(\tfrac{1}{n}\sum_{i=1}^{n} K'_\nu(\theta-\Theta_i)\right)$$

$$= \frac{1}{n} s^2\left(K'_\nu(\theta-\Theta_1),\ldots,K'_\nu(\theta-\Theta_n)\right)$$

where s^2 is the usual sample variance of n data, which in this context is formed by the derivative of the kernel centered at each sample value Θ_i, with $i = 1, \ldots, n$.

Standard deviation (regression)

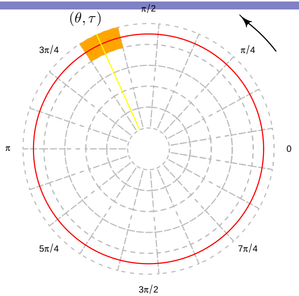
The estimator of the derivative of the regression function evaluated in a grid of angles in the interval $[0, 2\pi)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^t$, can be written as

$$\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}} = H\boldsymbol{Y}$$

where H is an $(N \times n)$ matrix and $\boldsymbol{Y}$ is the response vector.

- For fixed design:
$$\mathrm{var}(\hat{\boldsymbol{f}}'_{\boldsymbol{\theta}}) = H\Sigma H^t$$
where $\Sigma = \mathrm{diag}\left\{\sigma^2(\Theta_1), \ldots, \sigma^2(\Theta_n)\right\}$.
- For random design, the standard deviation is estimated by bootstrap.

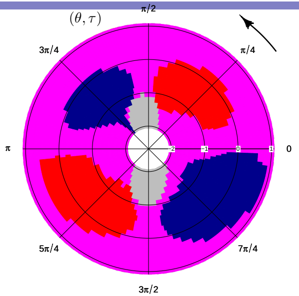
Construction of CircSiZer map



For each pair (θ, τ), with θ varying in $[0, 2\pi)$ and $\tau > 0$:

- Compute the confidence interval for $f'(\theta; \tau)$.
- If the interval is
  - above zero $\rightarrow$ the smoothed curve is significantly increasing $\rightarrow$ the location corresponding to the pair (θ, τ) is colored blue.
  - below zero $\rightarrow$ the smoothed curve is significantly decreasing $\rightarrow$ the location corresponding to the pair (θ, τ) is colored red.
  - contains zero $\rightarrow$ the derivative is not sig. dif. from zero $\rightarrow$ the location corresponding to the pair (θ, τ) is colored purple.
  - Location $(\theta, -\log_{10}(\nu))$ is coloured gray if there is not enough data.

Construction of CircSiZer map



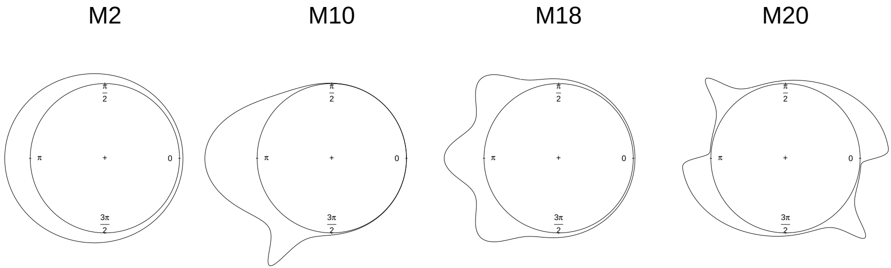
For each pair (θ, τ), with θ varying in $[0, 2\pi)$ and $\tau > 0$:

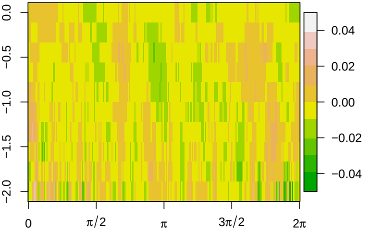
- Compute the confidence interval for $f'(\theta; \tau)$.
- If the interval is
  - above zero $\rightarrow$ the smoothed curve is significantly increasing $\rightarrow$ the location corresponding to the pair (θ, τ) is colored blue.
  - below zero $\rightarrow$ the smoothed curve is significantly decreasing $\rightarrow$ the location corresponding to the pair (θ, τ) is colored red.
  - contains zero $\rightarrow$ the derivative is not sig. dif. from zero $\rightarrow$ the location corresponding to the pair (θ, τ) is colored purple.
  - Location $(\theta, -\log_{10}(\nu))$ is coloured gray if there is not enough data.

CircSiZer performance (density)

Results based on 1000 samples of size $n = 250$:

- Pointwise vs. simultaneous.
- Normal vs. bootstrap.
- CircSiZer for detecting modes?





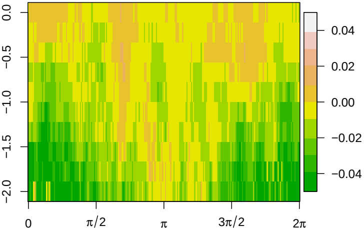


Figure: Differences between empirical and nominal coverage: normal (left) and bootstrap (right). Model M2 (von Mises).

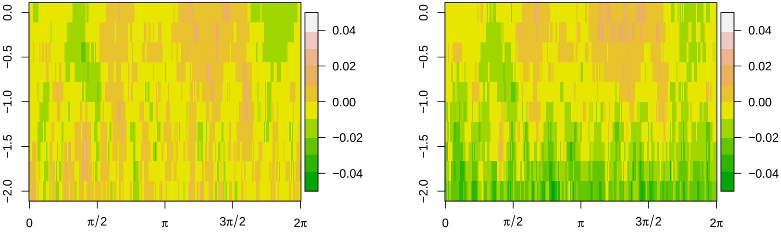


Figure: Differences between empirical and nominal coverage: normal (left) and bootstrap (right). Uniform model.

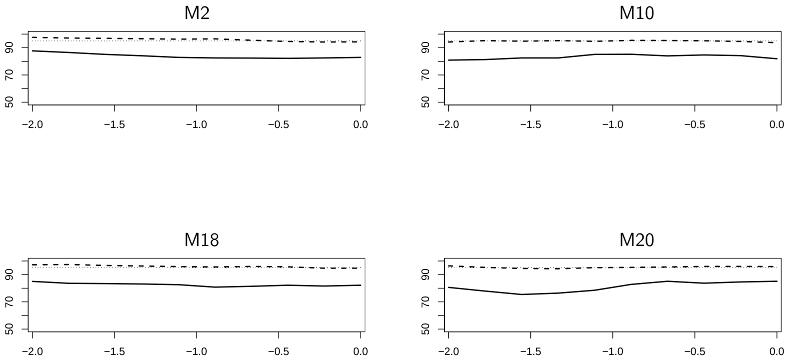
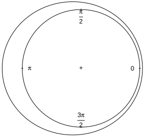


Figure: Global coverages for simultaneous normal (solid line) and simultaneous bootstrap (dashed line) for a range of smoothing values.

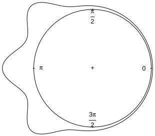
M2



	Modes	-2.00	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00
PN	0	220	159	104	22	4	0	0	0	0	0
	1	601	706	815	945	994	1000	1000	1000	1000	1000
	2	153	124	79	33	2	0	0	0	0	0
	3	25	11	2	0	0	0	0	0	0	0
	4	1	0	0	0	0	0	0	0	0	0
SB	0	997	995	986	915	620	114	2	0	0	0
	1	3	5	14	85	380	886	998	1000	1000	1000

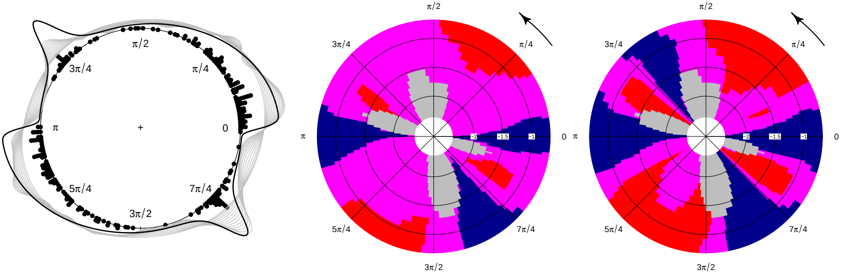
Table: Number of modes flagged by CircSiZer map with pointwise normal and simultaneous bootstrap confidence intervals for model M2.

M18



	Modes	-2.00	-1.78	-1.56	-1.33	-1.11	-0.89	-0.67	-0.44	-0.22	0.00
PN	0	6	0	0	0	0	0	0	0	0	0
	1	219	187	234	553	966	998	1000	1000	1000	1000
	2	489	464	484	406	33	2	0	0	0	0
	3	282	343	279	40	1	0	0	0	0	0
	4	4	6	3	1	0	0	0	0	0	0
SB	0	700	341	73	2	0	0	0	0	0	0
	1	288	593	813	977	1000	1000	1000	1000	1000	1000
	2	12	62	112	21	0	0	0	0	0	0
	3	0	4	2	0	0	0	0	0	0	0

Table: Number of modes flagged by CircSiZer map with pointwise normal and simultaneous bootstrap confidence intervals for model M18.



KDEs for a sample with $n = 250$ data. Simultaneous CircSizer map (center) and pointwise CircSizer map (right).

CircSiZer performance (regression)

- Models:

$$f_1(\theta) = \sin(\theta)$$

$$f_2(\theta) = \sin(\theta - 1.2\pi) + 3\exp(-10(15(\theta - \pi)/(2\pi)^2))$$

- Performance with local linear and spline smoothers:
  - Fixed design.
  - Simultaneous bootstrap.

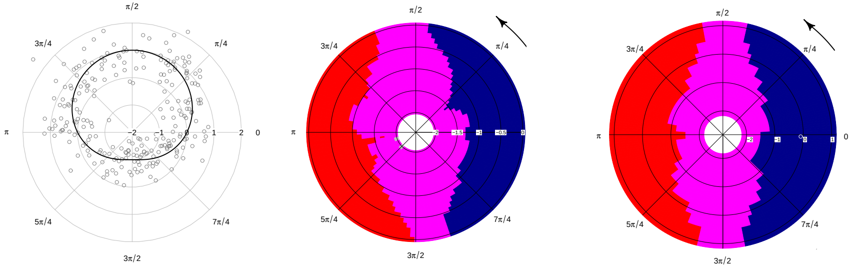


Figure: Regression estimation for a sample with $n = 250$ data. Center: based on local linear. Right: based on periodic spline. (Bootstrap global)

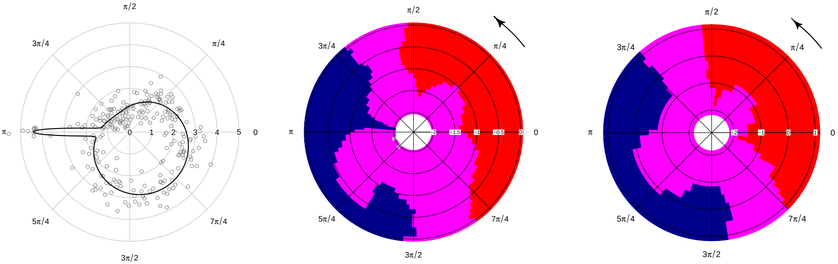


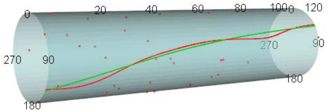
Figure: Regression estimation for a sample with $n = 250$ data. Center: based on local linear. Right: based on periodic spline. (Bootstrap global)

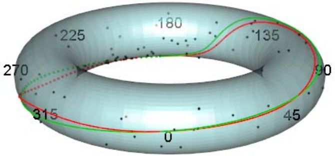
Take home message

- Pointwise normal (PN) is preferred to pointwise bootstrap.
- Simultaneous bootstrap (SB) is preferred to simultaneous normal.
- Both PN and SB provide useful information...

Extensions

- Circular–circular?
- Linear–circular?
- Higher dimensions?
- ... visualization...





Library NPCirc

Oliveira, M., Crujeiras, R. M. and Rodríguez–Casal, A. (2013)
NPCirc: An R package for nonparametric circular methods.
R package version 2.0.0. URL http://www.R-project.org/package=NPCirc.

Data set	Description
...	...
`circsizer.density`	CircSiZer map for density
`circsizer.map`	CircSiZer map
`circsizer.regression`	CircSiZer map for regression
...	...

Library NPCirc

Oliveira, M., Crujeiras, R. M. and Rodríguez–Casal, A. (2013)
NPCirc: An R package for nonparametric circular methods.
R package version 2.0.0. URL http://www.R-project.org/package=NPCirc.

Data set	Description
...	...
`circsizer.density`	CircSiZer map for density
`circsizer.map`	CircSiZer map
`circsizer.regression`	CircSiZer map for regression
...	...

John M. Chambers Statistical Software Award 2014

Thanks for your attention!