Nonparametric regression for circular variables with different groups of observations

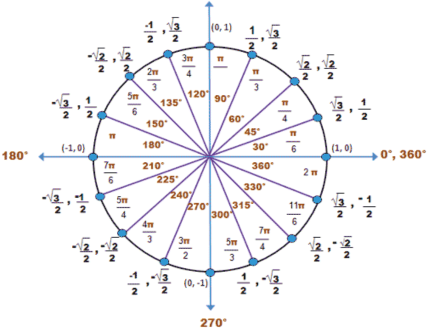
María Alonso Pena

Departmento de Estatística, Análise Matemática e Optimización
Universidade de Santiago de Compostela

Joint work with Rosa M. Crujeiras (USC)
and Jose Ameijeiras Alonso (KU Leuven)

What are circular data?

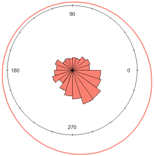
Observations which can be represented on the circumference of the unit circle and can be expressed as angles



Understanding circular data

Circular sample: $\Theta_1, ..., \Theta_n$

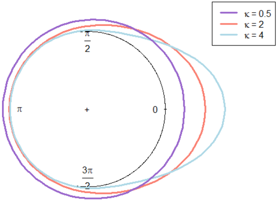
- Rose diagram (circular histogram)
- Circular densities



The von Mises density

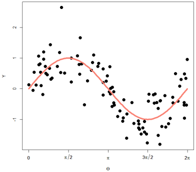
$$\Theta \sim f,\ \theta \in [0, 2\pi)$$

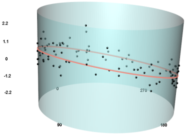
$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \mu))$$

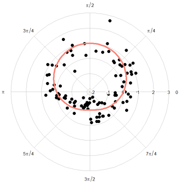


Regression with circular variables

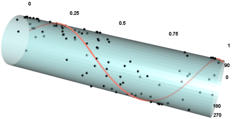
- Circular predictor - Linear response



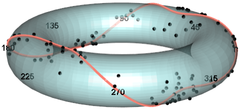




- Linear predictor - Circular response



- Circular predictor - Circular response



What have we done so far?

No-effect test

- Circular predictor - Linear response
- Linear predictor - Circular response
- Circular predictor - Circular response

ANCOVA models

- Circular predictor - Linear response
- Linear predictor - Circular response
- Circular predictor - Circular response

Alonso-Pena, M., Ameijeiras-Alonso, J. and Crujeiras, R.M.
Nonparametric tests for circular regression
Submitted

What have we done so far?

No-effect test

- Circular predictor - Linear response
- Linear predictor - Circular response
- Circular predictor - Circular response

ANCOVA models

- Circular predictor - Linear response
- Linear predictor - Circular response
- Circular predictor - Circular response

Modal regression (work in progress)

- Circular predictor - Linear response
- Linear predictor - Circular response
- Circular predictor - Circular response

What have we done so far?

No-effect test

- Circular predictor - Linear response
- Linear predictor - Circular response
- Circular predictor - Circular response

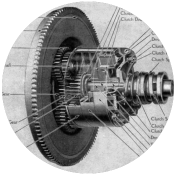
ANCOVA models

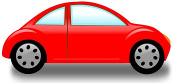
- **Circular predictor - Linear response**
- Linear predictor - Circular response
- Circular predictor - Circular response

Modal regression (work in progress)

- **Circular predictor - Linear response**
- Linear predictor - Circular response
- Circular predictor - Circular response

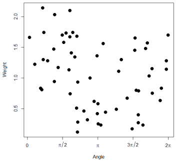
Motivating Circular-Linear regression



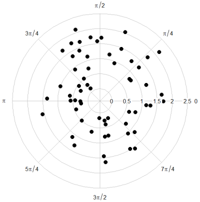


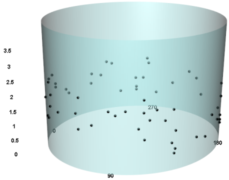
- Flywheels
- Present in cars' transmission systems
- Produce stability. Store rotational energy
- Balanced flywheels $\Rightarrow$ minimal vibration

Motivating Circular-Linear regression



- Θ: Angle of imbalance (circular)
- Y: balancing weight (real-valued)





Anderson-Cook, C.M (1999)
A tutorial on one-way analysis of circular-linear data
Journal of Quality Technology

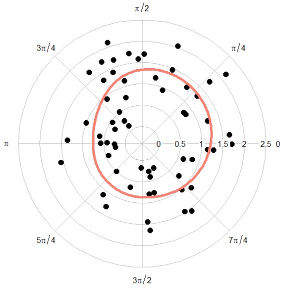
Circular-Linear regression

- The model:

$$Y_j = m(\Theta_j) + \varepsilon_j$$

- Local trigonometric fit

$$\beta_0 + \beta_1 \sin(\Theta_j - \theta)$$



- Estimation: $\hat{m}(\theta) = \hat{\beta}_0$, where
$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{j=1}^{n} K_\kappa(\theta - \Theta_j)[Y_j - (a + b\sin(\theta - \Theta_j))]^2$

Di Marzio M, Panzera A, and Taylor CC (2009)
Local polynomial regression for circular predictors
Statistics & Probability Letters

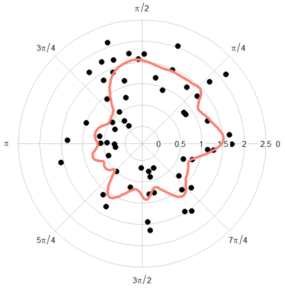
Circular-Linear regression

- The model:

$$Y_j = m(\Theta_j) + \varepsilon_j$$

- Local trigonometric fit

$$\beta_0 + \beta_1 \sin(\Theta_j - \theta)$$



- Estimation: $\hat{m}(\theta) = \hat{\beta}_0$, where
$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{j=1}^{n} K_\kappa(\theta - \Theta_j)[Y_j - (a + b\sin(\theta - \Theta_j))]^2$

Di Marzio M, Panzera A, and Taylor CC (2009)
Local polynomial regression for circular predictors
Statistics & Probability Letters

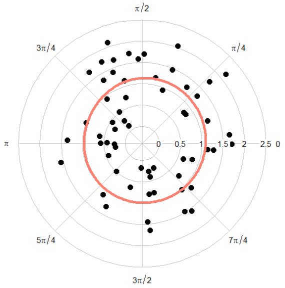
Circular-Linear regression

- The model:

$$Y_j = m(\Theta_j) + \varepsilon_j$$

- Local trigonometric fit

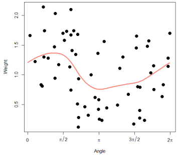
$$\beta_0 + \beta_1 \sin(\Theta_j - \theta)$$



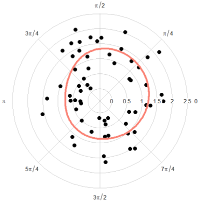
- Estimation: $\hat{m}(\theta) = \hat{\beta}_0$, where
$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(a,b)} \sum_{j=1}^{n} K_\kappa(\theta - \Theta_j)[Y_j - (a + b\sin(\theta - \Theta_j))]^2$

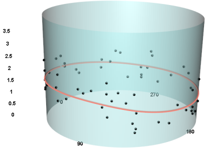
Di Marzio M, Panzera A, and Taylor CC (2009)
Local polynomial regression for circular predictors
Statistics & Probability Letters

Regression estimation of the flywheels data



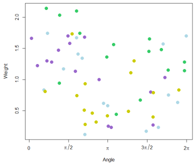
- Θ: Angle of imbalance (circular)
- Y: balancing weight (real-valued)

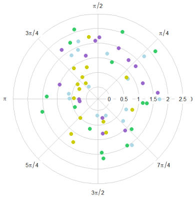




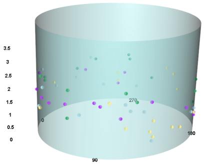
Anderson-Cook, C.M (1999)
A tutorial on one-way analysis of circular-linear data
Journal of Quality Technology

Motivating Circular-Linear ANCOVA



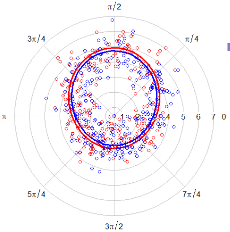


- Θ: Angle of imbalance (circular)
- Y: balancing weight (real-valued)
- Type of metal $i = 1, 2, 3, 4$



Anderson-Cook, C.M (1999)
A tutorial on one-way analysis of circular-linear data
Journal of Quality Technology

ANCOVA model



The model:

$$Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Equality test

$$H_0 : Y_{ij} = m(\Theta_{ij}) + \varepsilon_{ij}$$
$$H_1 : Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}$$

Test statistic

$$T_E = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [\hat{m}_i(\Theta_{ij}) - \hat{m}(\Theta_{ij})]^2$$

Parallelism test

$$H_0 : Y_{ij} = \gamma_i + m(\Theta_{ij}) + \varepsilon_{ij}$$
$$H_1 : Y_{ij} = m_i(\Theta_{ij}) + \varepsilon_{ij}$$

Test statistic

$$T_P = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} [\hat{\gamma}_i + \hat{m}(\Theta_{ij}) - \hat{m}_i(\Theta_{ij})]^2$$

Calibration. Distributions under H_0 approximated to a $a + c\chi^2_b$

Young, S. and Bowman, A.W. (1995)
Nonparametric analysis of covariance
Biometrics

Estimation of the parallelism parameter γ.

Model under H_0 in matrix notation:

$$\boldsymbol{Y} = \boldsymbol{D}\boldsymbol{\gamma} + \boldsymbol{m} + \boldsymbol{\varepsilon}$$

Given $\boldsymbol{\gamma}$, the regression function is estimated as

$$\widehat{\boldsymbol{m}} = \boldsymbol{S}(\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{\gamma})$$

and the estimator of the parallelism parameter is

$$\hat{\boldsymbol{\gamma}} = [\boldsymbol{D'}(\boldsymbol{I}_n - \boldsymbol{S}_{\mathbf{1}})'(\boldsymbol{I}_n - \boldsymbol{S}_{\mathbf{1}})\boldsymbol{D}]^{-1}\boldsymbol{D}'(\boldsymbol{I}_n - \boldsymbol{S}_{\mathbf{1}})'(\boldsymbol{I}_n - \boldsymbol{S}_{\mathbf{1}})\boldsymbol{Y}$$

Variance estimation

$\hat{\sigma}^2$ is obtained by using *periodic pseudoresiduals*:

$$\begin{aligned}\tilde{\varepsilon}_{i[j]} &= \frac{\Theta_{i[j+1]} - \Theta_{i[j]}}{\Theta_{i[j+1]} - \Theta_{i[j-1]}} Y_{i[j-1]} + \frac{\Theta_{i[j]} - \Theta_{i[j-1]}}{\Theta_{i[j+1]} - \Theta_{i[j-1]}} Y_{i[j+1]} - Y_{i[j]} \\ &= a_{i[j]} Y_{i[j-1]} + b_{i[j+1]} Y_{i[j+1]} - Y_{i[j]}\end{aligned}$$

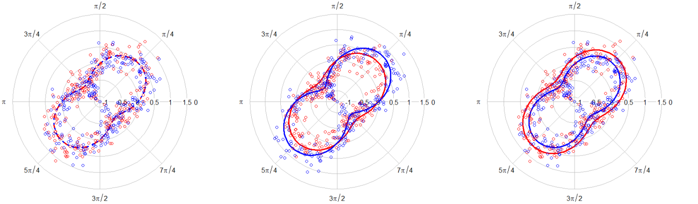
The variance in each group and the total variance are estimated as

$$\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{c_{i[j]}^2} \tilde{\varepsilon}_{i[j]}^2, \qquad \hat{\sigma}^2 = \frac{1}{n-I} \sum_{i=1}^{I} n_i \hat{\sigma}_i^2, \quad \text{with } c_{i[j]}^2 = a_{i[j]}^2 + b_{i[j]}^2 + 1$$

Gasser, T, Sroka, L and Jenne-Steinmetz C (1986)
Residual variance and residual pattern in nonlinear regression
Biometrika

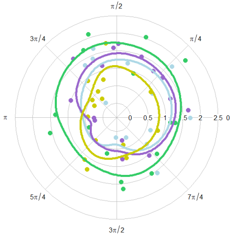
Simulation study

- Group 1: $Y = \cos\Theta \sin\Theta + \varepsilon$
- Group 2: $Y = \beta \cos\Theta \sin\Theta + \varepsilon, \quad \beta = 1, 1.5, 1.75, \; \gamma = 0.2$

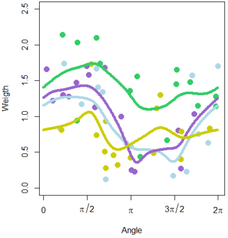


	Equality			Parallelism		
(n_1, n_2)	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$	$\beta = 1$	$\beta = 1.5$	$\beta = 1.75$
$(50, 50)$	.055	.519	.917	.048	.579	.915
$(50, 100)$	.042	.679	.987	.052	.730	.974
$(100, 100)$	.058	.915	1	.053	.932	1
$(100, 250)$	.043	.987	1	.054	.985	1
$(250, 250)$	.046	1	1	.064	1	1

Flywheel data

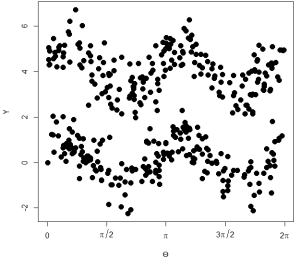


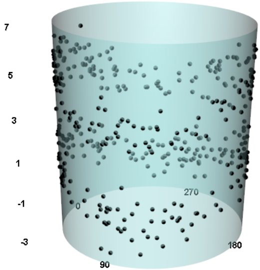
Balancing weight vs. angle of imbalance: are the regression curves equal for the 4 metals? Are they parallel?



- Equality: with cross-validation concentration, p-value is 0.0263.
- Parallelism: with cross-validation concentration, p-value is 0.4695.

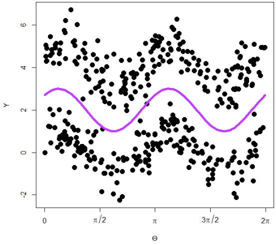
What if we don't know the groups?

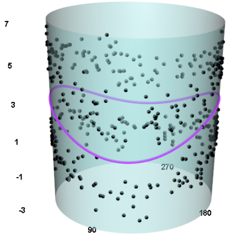




Regression to the mean is not adequate to explain the data

$$m(\theta) = \mathbb{E}(Y|\theta)$$

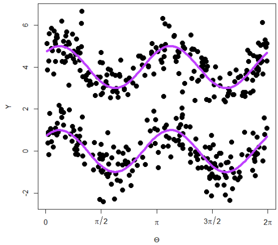


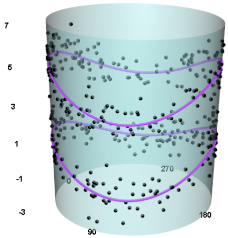


The modal regression multifunction

The conditional density is not unimodal!

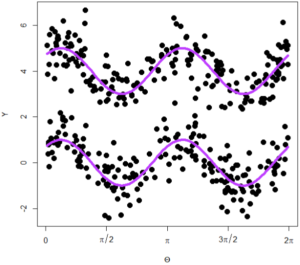
$$M(\theta) = \Big\{ \text{local maxima of } f(y|\theta) \Big\}$$





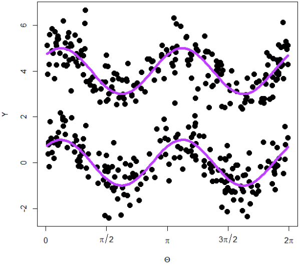
Einbeck, J. and Tutz, G. (2006)
Modelling beyond regression functions: an application of multimodal regression to speed-flow data
Applied Statistics

The modal regression multifunction



$$M(\theta) = \left\{ y : \frac{\partial}{\partial y} f(y|\theta) = 0, \frac{\partial^2}{\partial y^2} f(y|\theta) < 0 \right\}$$
$$= \left\{ y : \frac{\partial}{\partial y} f(\theta, y) = 0, \frac{\partial^2}{\partial y^2} f(\theta, y) < 0 \right\}$$

The modal regression multifunction

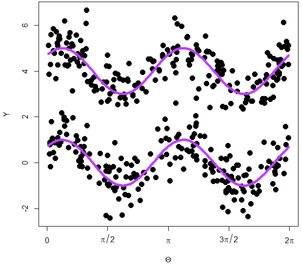


$$M(\theta) = \left\{ y : \frac{\partial}{\partial y} f(y|\theta) = 0, \frac{\partial^2}{\partial y^2} f(y|\theta) < 0 \right\}$$
$$= \left\{ y : \frac{\partial}{\partial y} f(\theta, y) = 0, \frac{\partial^2}{\partial y^2} f(\theta, y) < 0 \right\}$$

$$\hat{f}(\theta, y) = \frac{1}{n} \sum_{j=1}^{n} K_\kappa \left(\theta - \Theta_j \right) G_h \left(y - Y_j \right)$$

García-Portugués, E., Crujeiras, R. M. and González-Manteiga, W. (2013)
Kernel density estimation for directional-linear data
Journal of Multivariate Analysis

The modal regression multifunction



$$M(\theta) = \left\{ y : \frac{\partial}{\partial y} f(y|\theta) = 0, \frac{\partial^2}{\partial y^2} f(y|\theta) < 0 \right\}$$

$$= \left\{ y : \frac{\partial}{\partial y} f(\theta, y) = 0, \frac{\partial^2}{\partial y^2} f(\theta, y) < 0 \right\}$$

$$\hat{f}(\theta, y) = \frac{1}{n} \sum_{j=1}^{n} K_\kappa (\theta - \Theta_j) G_h (y - Y_j)$$

$$\hat{M}(\theta) = \left\{ y : \frac{\partial}{\partial y} \hat{f}(\theta, y) = 0, \frac{\partial^2}{\partial y^2} \hat{f}(\theta, y) < 0 \right\}$$

García-Portugués, E., Crujeiras, R. M. and González-Manteiga, W. (2013)
Kernel density estimation for directional-linear data
Journal of Multivariate Analysis

Estimation through an adaptation of the mean-shift algorithm

G_h is a radially symmetric kernel (e.g. the normal kernel)

$$\frac{\partial}{\partial y}\hat{f}(\theta, y) = 0 \iff y = \frac{\sum_{j=1}^{n} K_\kappa(\theta - \Theta_j) \exp\left\{\frac{-(y - Y_j)^2}{2h^2}\right\} Y_j}{\sum_{j=1}^{n} K_\kappa(\theta - \Theta_j) \exp\left\{\frac{-(y - Y_j)^2}{2h^2}\right\}}$$

Given a starting value y_0,

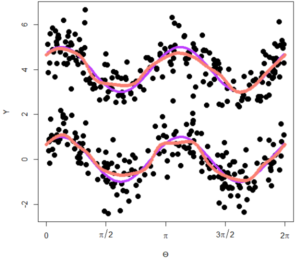
$$\boldsymbol{y_l} = \frac{\sum_{j=1}^{n} K_\kappa(\theta - \Theta_j) \exp\left\{\frac{-(\boldsymbol{y_{l-1}} - Y_j)^2}{2h^2}\right\} Y_j}{\sum_{j=1}^{n} K_\kappa(\theta - \Theta_j) \exp\left\{\frac{-(\boldsymbol{y_{l-1}} - Y_j)^2}{2h^2}\right\}}, \quad l = 1, 2, ...$$

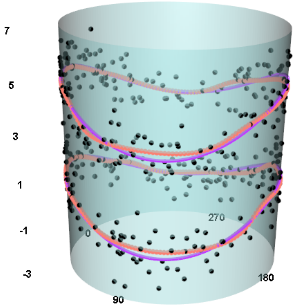
until convergence is reached.

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

Estimated modal regression multifunction

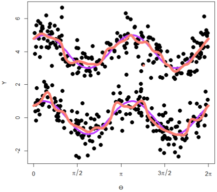
When using several starting values, we obtain the different local modes



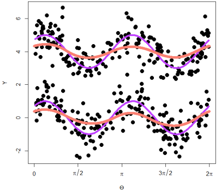


The smoothing parameters

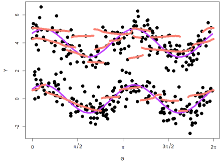
Large κ



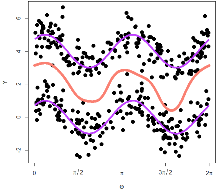
Small κ



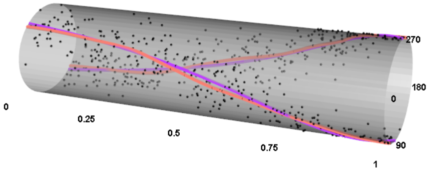
Small h

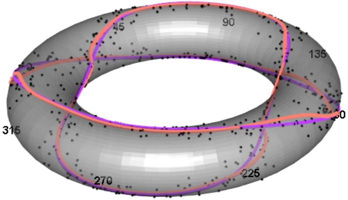


Large h



Modal regression for circular responses





To sum up

- ANCOVA model when a categorical variable is provided
- Modal regression when the conditional density is multimodal

Future (present) work

- Asymptotic properties of the modal regression estimator
- Bandwidth selection methods for modal regression
- Modal regression for circular responses

Nonparametric regression for circular variables with different groups of observations

María Alonso Pena

Departmento de Estatística, Análise Matemática e Optimización
Universidade de Santiago de Compostela

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Joint work with Rosa M. Crujeiras (USC)
and Jose Ameijeiras Alonso (KU Leuven)