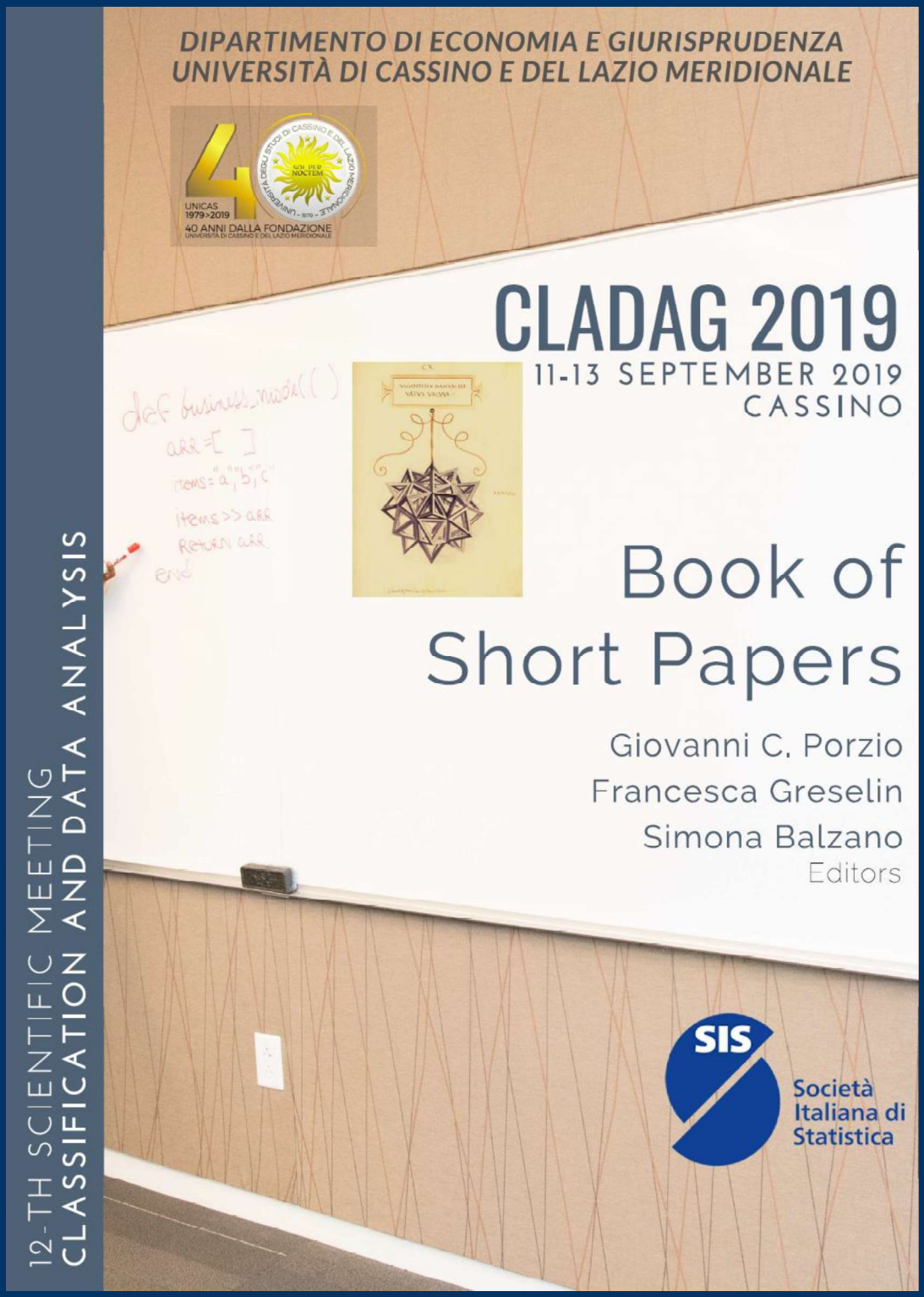
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA



EDIZIONI UNIVERSITA' DI CASSINO

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6

CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

LOCAL FITTING OF ANGULAR VARIABLES OBSERVED WITH ERROR

Marco Di Marzio[1], Stefania Fensore[1], Agnese Panzera[2]
and Charles C. Taylor[3]

[1] DSFPEQ, University of Chieti-Pescara, (e-mail: marco.dimarzio@unich.it, stefania.fensore@unich.it)

[2] DISIA, University of Florence, (e-mail: a.panzera@disia.unifi.it)

[3] Department of Statistics, University of Leeds,
(e-mail: charles@maths.leeds. ac.uk)

ABSTRACT: The problem of estimating a circular regression when the predictor is contaminated by errors is studied. Other than some estimators, we also present a novel smoothing degree selection rule.

KEYWORDS: deconvolution, measurement error, Simex.

1 Introduction

Statistical regression models are generally based on the assumption that the independent variables have been measured exactly. However, sometimes the regressors are, for some reason, not directly observable or measured with errors. When this is the case specific models, known as *errors-in-variables* or *measurement error models*, have to be taken into account.

Formally, suppose that we are interested in estimating the regression of Y on X^*, denoted as m, and that our data are realizations from variables $X = X^* + \eta$ and Y, say $(x_1, y_1), \ldots, (x_n, y_n)$. A general model for this case could be

$$\begin{aligned} y_i &= m(x_i^*) + \zeta_i \\ x_i &= x_i^* + \eta_i \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$, where X^* and Y respectively refer to the predictor and response variable, ζ_is are observations of the random error term ζ, η_is are realizations of η. The unobserved variable X^* is always referred as latent or true variable. Usual assumptions include that ζ is independent from both X^* and η, the distribution of ζ is unknown but has mean 0 and constant variance, while the distribution of η is known.

Let f_X, f_{X^*} and f_η respectively denote the probability density function of X, X^* and η. Basic theoretical considerations suggest that f_X is the convolution between f_{X^*} and f_η:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X^*}(x-\nu) dF_\eta(\nu), \tag{2}$$

where F_η denotes the distribution function of η. As the consequence, the estimators of the free-error model are clearly not consistent. In such a context there are two approaches to obtain accurate estimates: deconvolution methods and explicit bias estimation and correction.

In this paper we address the measurement error case when data can be represented as points on a circumference. Specifically, we present a nonparametric deconvolution estimator along with a rule for smoothness selection.

2 Circular data

Angular or circular data are collected whenever observations are measured by means of a periodic scale. They are usually represented as points on the circumference of a circle with unit radius. Classical examples of such data are wind directions, animal movements, any phenomenon measured by the 24 h clock, etc. Once a zero direction and a sense of rotation have been arbitrarily chosen, these observations can be expressed as angles. Due to their periodic nature, circular data cannot be analysed by standard real-line methods, therefore in the last decades great attention has been devoted to circular statistics. For a comprehensive account, see the survey paper by Lee, 2010, and the references therein.

3 The estimator

Consider a pair of random angles (Θ, Δ), i.e. variables taking values on $[0, 2\pi)$. Given the random sample $(\Phi_1, \Delta_1), \ldots, (\Phi_n, \Delta_n)$, we can write model (1) as

$$\begin{aligned} \Delta_i &= (m(\Theta_i) + \varepsilon_i) \mathsf{mod}(2\pi), \\ \Phi_i &= \Theta_i + u_i, \end{aligned} \tag{3}$$

where Θ_is are independent copies of the circular latent variable Θ, the ε_is are i.i.d. random angles independent of the Θ_is, with zero mean direction and finite concentration, and the u_is are realizations of the random angle U independent of the Θ_is.

A local estimator for m at $\theta \in [0, 2\pi)$ can be defined as

$$\hat{m}(\theta;\kappa) = \mathsf{atan2}(\hat{m}_s(\theta;\kappa), \hat{m}_c(\theta;\kappa)), \tag{4}$$

with

$$\hat{m}_s(\theta;\kappa) = \sum_{i=1}^{n} \sin(\Delta_i) L_\kappa(\Theta_i - \theta),$$
$$\hat{m}_c(\theta;\kappa) = \sum_{i=1}^{n} \cos(\Delta_i) L_\kappa(\Theta_i - \theta),$$

where the function $\mathsf{atan2}(y,x)$ returns the angle between the x-axis and the vector from the origin to (x,y), and L_κ is a circular *deconvolution kernel* function depending on $\gamma_\ell(\kappa)$ and $\lambda_\ell(\kappa_U)$ which are, for $\ell \in \mathbb{Z}$, respectively, the ℓth Fourier coefficient of the periodic weight function K_κ and the error density f_U whose concentration parameter is κ_U:

$$L_\kappa(\theta) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{\ell=1}^{\infty} \frac{\gamma_\ell(\kappa)}{\lambda_\ell(\kappa_U)} \cos(\ell\theta) \right\}. \tag{5}$$

4 Smoothing degree selection

In the context of measurement error the standard cross-validation criterion for the selection of the smoothing degree κ is not suitable. Indeed, if we knew the values $\Theta_1, \ldots, \Theta_n$ in addition to $(\Phi_1, \Delta_1), \ldots, (\Phi_n, \Delta_n)$ then we could compute the conventional cross-validation smoothing degree $\hat{\kappa}_0 = argmin CV_0(\kappa)$, with

$$CV_0(\kappa) = \frac{1}{n} \sum_{i=1}^{n} (1 - \cos(\Delta_i - \hat{m}_{-i}(\Theta_i))), \tag{6}$$

where $\hat{m}_{-i}$ denotes the version of $\hat{m}$ computed by omitting the ith pair of the sample. However, since Θ_is are unknown above criterion is not attainable.

However, a cross-validation idea could still be employed through a SIMEX (simulation-extrapolation) approach proposed by Delaigle and Hall, 2008 by following the steps listed below:

1. Generate two i.i.d. samples from U denoted as $u_1^*, \ldots, u_n^*$ and $u_1^{**}, \ldots, u_n^{**}$. Then, for $i = 1, \ldots, n$, define $\Phi_i^* = \Phi_i + u_i^*$ and $\Phi_i^{**} = \Phi_i + u_i^* + u_i^{**}$ and consider the problem of estimating two regression functions, m_1 and m_2, respectively from the contaminated data (Φ_i^*, Δ_i) and (Φ_i^{**}, Δ_i).

2. Define the objective functions $CV^*(\kappa)$ and $CV^{**}(\kappa)$

$$CV^*(\kappa) = \frac{1}{n}\sum_{i=1}^{n}(1-\cos(\Delta_i - \hat{m}_{1,-i}(\Phi_i)))$$
$$CV^{**}(\kappa) = \frac{1}{n}\sum_{i=1}^{n}(1-\cos(\Delta_i - \hat{m}_{2,-i}(\Phi_i^*)))$$

in order to obtain $\hat{\kappa}_1^* = argmin CV^*(\kappa)$ and $\hat{\kappa}_2^{**} = argmin CV^{**}(\kappa)$.

3. The dependence of $\hat{\kappa}_1^*$ on Φ_i^* and $\hat{\kappa}_2^{**}$ on Φ_i^{**} can be removed by averaging over a large number, say B, of CV^* and CV^{**} for different simulated sequences of $u_1^*,\ldots,u_n^*$ and $u_1^{**},\ldots,u_n^{**}$:

$$CV_1 = \frac{1}{B}\sum_{b=1}^{B} CV_b^*$$
$$CV_2 = \frac{1}{B}\sum_{b=1}^{B} CV_b^{**}$$

4. Then, we define, for $j = 0,1,2$,

$$\hat{\kappa}_j = argmin CV_j(\kappa). \tag{7}$$

Now, Φ^{**} approximates Φ^* in the same way that Φ^* approximates Φ and Φ approximates Θ. Therefore we expect that the relationship between $\hat{\kappa}_0$ and $\hat{\kappa}_1$ is similar to that between $\hat{\kappa}_1$ and $\hat{\kappa}_2$. As the final result, we get

$$\hat{\kappa}_0 = \hat{\kappa}_1^2/\hat{\kappa}_2. \tag{8}$$

References

LEE, A. 2010. Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 477–486.

DELAIGLE, A., HALL, P. 2008. Using SIMEX for Smoothing-Parameter Choice in Errors-in-Variables Problems. *Journal of the American Statistical Association*, **103**, 280–287.