CIRCULAR REGRESSION WITH MEASUREMENT ERRORS

Marco Di Marzio [1], Chiara Passamonti[1] and Charles Taylor[2]

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

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

ABSTRACT: We propose techniques for estimating a regression function when the predictor is circular. A case study on Carbon monoxide pollution is presented.

KEYWORDS: Characteristic function, deconvolution kernels, measurement errors.

1 Introduction

We propose a nonparametric regression estimator that is consistent in the presence of measurement error when predictor data are circular. Following the approach of Carroll & Hall, 1988 and Carroll *et al.*, 1995, we introduce a deconvolution-type estimator.

Some facts on the characteristic functions are worth to be recalled. The characteristic function of a circular random variable Θ, denoted as $\varphi_\Theta(\ell) = \alpha_\ell + i\beta_\ell$ satisfies $\varphi_\Theta(\ell) = \varphi_{\Theta+2\pi}(\ell), \ell \in \mathbb{Z}$, being zero elsewhere. Moreover, $\alpha_\ell = E[\cos(\ell\Theta)]$ and $\beta_\ell = E[\sin(\ell\Theta)]$, both are the coefficients in the Fourier series representation of f_Θ, and correspond to the ℓth *trigonometric moment* of Θ. Finally, $\beta_\ell = 0$ when f_Θ is symmetric. If f_Θ is square integrable on $[0, 2\pi)$, one can represent $f_\Theta(\theta)$, $\theta \in [0, 2\pi)$, as

$$\frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\varphi_\Theta(\ell)\exp(-i\ell\theta) = \frac{1}{2\pi}\left\{1 + 2\sum_{\ell=1}^{\infty}(\alpha_\ell\cos(\ell\theta) + \beta_\ell\sin(\ell\theta))\right\}. \quad (1)$$

Our estimator is described in Section 2. In Section 3, we model the carbon monoxide propagation due to wind direction in a region near Huston (Texas).

2 Model and estimator

We consider the case of a circular predictor and linear response. Given the random sample $(\Psi_1, Y_1), \ldots, (\Psi_n, Y_n)$, assume the regression model $Y_i = m(\Psi_i) +$

$\sigma(\Psi_i)e_i$, but it is available the sample $(\Phi_1, Y_1), \ldots, (\Phi_n, Y_n)$, modelled according $\Phi = (\Psi + \varepsilon) \mathsf{mod}(2\pi)$. Here we have that

- the e_is are i.i.d. real-random variables with zero mean and unit variance, and $\sigma^2(\cdot)$ is the conditional variance of Y;
- the Ψ_is are independent copies of the circular latent variable Ψ with density function f_Ψ;
- the ε_is are i.i.d. circular random variables independent of the (Ψ_i, e_i)'s, with a known density function f_ε which is symmetric around zero.

We assume that f_ε, f_Ψ and f_Φ are square integrable, and f_ε is a circular density allowing an absolutely convergent Fourier series representation.

A local estimator for m at $\psi \in [0, 2\pi)$, denoted by $\tilde{m}(\psi; \kappa)$, can be obtained by employing a *circular* deconvolution kernel. Using the inversion formula (1), and considering that for a symmetric function $\beta_\ell = 0$ for any ℓ, we have

$$\tilde{K}_\kappa(\phi) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{\ell=1}^{\infty} \frac{\gamma_\ell(\kappa)}{\lambda_\ell(\kappa_\varepsilon)} \cos(\ell\phi) \right\}, \tag{2}$$

with smoothing parameter $\kappa > 0$, where $\gamma_\ell(\kappa)$ and $\lambda_\ell(\kappa_\varepsilon)$, for $\ell \in \mathbb{Z}$, respectively are the ℓth Fourier coefficient of the periodic weight function K_κ and the error density f_ε whose concentration is κ_ε. The estimator is well defined when the error density has nonvanishing Fourier coefficients, $\gamma_\ell(\kappa)$ is not identically zero and $\sum_{\ell=1}^{\infty} \mid \gamma_\ell(\kappa)/\lambda_\ell(\kappa_\varepsilon) \mid < \infty$ for all $(\kappa, \kappa_\varepsilon) \in \mathbb{R}_+^2$, which, in turn, imply that both K_κ and $\tilde{K}_\kappa$ are square integrable functions.

The local constant estimator for m is defined by

$$\tilde{m}(\psi; \kappa) = \frac{\sum_{i=1}^{n} \tilde{K}_\kappa(\Phi_i - \psi) Y_i}{\sum_{i=1}^{n} \tilde{K}_\kappa(\Phi_i - \psi)}, \tag{3}$$

where $\tilde{K}_\kappa$ is a circular deconvolution kernel.

Theorem 1. *Given the* $[0, 2\pi) \times \mathbb{R}$*-valued random sample* $(\Psi_1, Y_1), \ldots, (\Psi_n, Y_n)$*, consider the local constant estimator. If*

i) K_κ *is a second sin-order kernel admitting a convergent Fourier series representation* $1/(2\pi)\{1 + 2\sum_{\ell=1}^{\infty} \gamma_\ell(\kappa)\cos(\ell\theta)\}$*, with* κ *increasing with n in such a way that, for* $\ell \in \mathbb{Z}^+$*,*
$\lim_{n\to\infty} \frac{1-\gamma_\ell(\kappa)}{1-\gamma_2(\kappa)} = \frac{\ell^2}{4}$,
$\lim_{n\to\infty} \gamma_\ell(\kappa) = 1$ *and* $\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{\infty} \gamma_\ell^2(\kappa) = 0$,

ii) the second derivative of the regression function m is continuous,
iii) the conditional variance σ^2 is continuous, and the density f_Ψ is continuously differentiable,

then

$$\mathsf{E}[\hat{m}(\psi;\kappa)] - m(\psi) = \frac{(1-\gamma_2(\kappa))}{4}\left\{m''(\psi) + \frac{2m'(\psi)f'_\Psi(\psi)}{f_\Psi(\psi)}\right\} + o(1-\gamma_2(\kappa)),$$

$$\mathsf{Var}[\hat{m}(\psi;\kappa)] = \frac{\left(1+2\sum_{\ell=1}^{\infty}\gamma_\ell^2(\kappa)\right)}{2\pi n f_\Psi(\psi)}\sigma^2(\psi) + o\left(\frac{\sum_{\ell=1}^{\infty}\gamma_\ell^2(\kappa)}{n}\right).$$

We notice that, as in the Euclidean setting, the measurement error has no effect on the asymptotic bias of the estimator, which, when the predictor observed with error is circular (linear respectively), depends only on the second moment of the classical kernel K_κ (K_h resp.). The asymptotic variance, similarly to the Euclidean setting, depends on the Fourier coefficients (characteristic function resp.) of the error density appearing in roughness of the deconvolution kernel $\tilde{K}_\kappa$ ($\tilde{K}_h$ resp.).

3 Pollution and surface wind data

Usually, air pollution in a region strongly depends on wind direction. We consider data from the Texas Commission on Environmental Quality, where the response variable is the amount of carbon monoxide (CO) while the explanatory variable is the wind direction. We have selected a site near Houston ("North Loop") in Harris County at Latitude: 29.81^o North and Longitude: -95.39^o West using data from 2018*. The data are collected hourly, but we have calculated the average daily wind direction (using the directional average), and the average daily CO (in parts per million). These daily averages were "thinned" to reduce serial correlation resulting in 183 observations from alternate days. We initially fit a parametric model in which CO (y) is related to wind direction (ϕ) using a sine-cosine model $Y_i = \beta_0 + \beta_1 \sin\Phi_i + \beta_2 \cos\Phi_i + e_i$. This gives fitted values $\hat{\beta}_0 = 0.568, \hat{\beta}_1 = -0.173, \hat{\beta}_2 = 0.074$. The CO pollution is highest when the wind is coming from the south (2.73 radians). Then, we fit a standard circular-linear nonparametric regression, in which the measurements are treated as error free. The smoothing parameter (chosen by leave-one-out

*https://www.tceq.texas.gov/

cross-validation) was selected as $\kappa = 7.77$ for a von Mises kernel. For this model, the maximum CO occurs at 2.11 radians.

Finally, in this circular-linear case, we use a error-in-variables model for the observed wind direction which can be approximated by a wrapped Normal error with zero mean and concentration equal to 0.9. The estimated CO is then given using equation (3), in which κ was found by leave-one-out cross-validation to be 3.35. The three curves, depicted in Figure 1, show that, in the last case, the curve appears to be somewhat less smooth than the error-free model estimate. The nonparametric errors-in-variables model has residual sum of squares equal to 1.91, whereas the parametric model is slightly larger (2.40) and the error-free model very similar (1.99). The maximum estimated CO occurs at $\phi = 2.17$ for the errors-in-variables model.

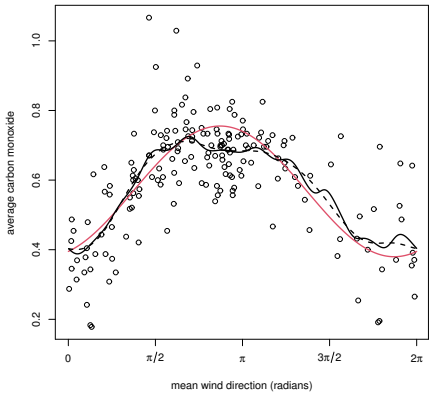

Figure 1. *Carbon monoxide vs wind direction at Houston North Loop monitoring station — alternate daily averages for 2018. Parametric sin/cos model (red), fitted nonparametric errors in variables model (black) and standard circular-linear (no error model) kernel regression (dashed).*

References

CARROLL, R.J., & HALL, P. 1988. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, **83**, 1184–1186.

CARROLL, R.J., RUPPERT, D., & STEFANSKI, L.A. 1995. *Measurement Error in Nonlinear Models*. New York: Chapman and Hall.