X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

An algorithm for estimating circular–linear densities

Eduardo García–Portugués[1]

[1]Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela.

ABSTRACT

Johnson and Wehrly (1978) introduced a parametric family for circular–linear densities with specified marginal distributions. In this work, an algorithm for estimating these circular–linear densities is presented. The performance of the proposed method is assessed by a simulation study, comparing the parametric approach with more flexible alternatives.

Keywords: Johnson and Wehrly family; Circular kernel estimation; Circular–linear data; Copula.

1. INTRODUCTION

Circular data appears in many diverse scientific fields where the measurements are angles or directions. Typical examples are the study of wind direction in meteorology and the angles of a molecule structure in biology. Circular–linear data arises, for example, when dealing with wind direction and pollutants concentration.

Johnson and Wehrly (1978) and Wehrly and Johnson (1980) present a method for obtaining joint circular–linear and circular–circular densities with specified marginals, respectively. Fernández–Durán (2007) adapts this proposal considering a new family of circular distributions based on nonnegative trigonometric sums (see Fernández–Durán, 2004). In this work, we present a procedure for modelling the relation between a circular and a linear variable through the estimation of a circular–linear density, also based on the ideas of Johnson and Wehrly (1978), but considering nonparametric kernel density estimators both in the circular and linear marginals and in the joining density function. With this approach, the lack of flexibility in the construction of circular–linear densities noticed by Fernández–Durán (2007) is overcome. Besides, the circular–linear density representation considered can be interpreted in terms of copula functions, which presents some computational advantages. The method has been applied for exploring the relation between wind direction and SO_2 concentration in monitoring stations close to a power plant located in Galicia (NW–Spain). The results can be seen in García–Portugués *et al.* (2011).

This paper is organized as follows. Some background on circular random variables and the Johnson and Wehrly family is introduced in Section 2. Section 3 is devoted to the estimation algorithm. Simulation results are shown in Section 4. Some final considerations are reported in Section 5.

2. CIRCULAR AND CIRCULAR–LINEAR DISTRIBUTIONS

Denote by Θ a circular random variable with support in the unit circle $\mathbb{S}^1$. A circular distribution $\Psi(\cdot)$ for Θ assigns a probability to each direction $(\cos(\theta), \sin(\theta))$ in the plane $\mathbb{R}^2$, characterized by the angle $\theta \in [0, 2\pi)$. A detailed review of descriptive and inferential methods for circular data can be found in Jammalamadaka and SenGupta (2001).

The von Mises distribution is the analogue of the normal distribution in circular random variables. This family of distributions, denoted by $vM(\mu, \kappa)$, is characterized by two parameters: $\mu \in [0, 2\pi)$, the circular mean and $\kappa \geq 0$, a circular concentration parameter around μ. For $\theta \in [0, 2\pi)$, the corresponding density function is given by

$$\varphi_{vM}(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \tag{1}$$

where $I_0(\cdot)$ is the modified Bessel function of first kind and order zero. The uniform circular distribution, $\varphi_U(\theta) = (2\pi)^{-1}$, is obtained as a particular case of the von Mises family, for $\kappa = 0$.

A circular–linear random variable (Θ, X) is supported in the cylinder $\mathbb{S}^1 \times \mathbb{R}$ and a circular–linear density, namely $p(\cdot, \cdot)$, must satisfy the periodicity condition in the circular argument, as well as the usual assumptions. Johnson and Wehrly (1978) propose a method for obtaining circular–linear densities with specified marginals. Denote by $\varphi(\cdot)$ and $f(\cdot)$ the circular and linear marginal densities, respectively, and by $\Psi(\cdot)$ and $F(\cdot)$ their corresponding marginal distributions. Let also $g(\cdot)$ be another circular density. Then:

$$p(\theta, x) = 2\pi g \left[2\pi \left(\Psi(\theta) + F(x)\right)\right] \varphi(\theta) f(x), \tag{2}$$

is a density for a circular–linear distribution for a random variable (Θ, X), with specified marginal densities $\varphi(\cdot)$ and $f(\cdot)$ (see Johnson and Wehrly (1978), Theorem 5).

The joining density $g(\cdot)$ can be interpreted as the copula density linked to $p(\cdot, \cdot)$ (see Nelsen (2006) for details on copulas). Hence, (2) can be written as

$$p(\theta, x) = c\left(\Psi(\theta), F(x)\right) \varphi(\theta) f(x),$$

where $c(\cdot, \cdot)$ is the copula density of (Θ, X). This representation in terms of copulas allows for efficient computations in the estimation algorithm (see García–Portugués *et al.*, 2011).

3. ESTIMATION ALGORITHM

Denote by $\{(\Theta_i, X_i)\}_{i=1}^n$ a data sample from a circular–linear random variable (Θ, X) and assume that the density $p(\cdot,\cdot)$ for (Θ, X) admits a representation such as the one in (2). In this joint circular–linear density model, three density functions must be estimated: the marginal densities $\varphi(\cdot)$ and $f(\cdot)$ (and also the corresponding distributions) and the joining circular density $g(\cdot)$. A new natural procedure for estimating $p(\cdot,\cdot)$ is given in the following algorithm:

Estimation algorithm

Step 1. Obtain estimators for the marginal densities $\hat{\varphi}(\cdot)$, $\hat{f}(\cdot)$ and the corresponding marginal distributions $\hat{\Psi}(\cdot)$, $\hat{F}(\cdot)$.

Step 2. Compute an artificial sample $\left\{2\pi(\hat{\Psi}(\theta_i) + \hat{F}(x_i))\right\}_{i=1}^n$ and estimate the joining circular density $\hat{g}(\cdot)$.

Step 3. Obtain the circular–linear density estimator as

$$\hat{p}\,(\theta, x) = 2\pi\hat{g}\left[2\pi(\hat{\Psi}(\theta) + \hat{F}(x))\right]\hat{\varphi}(\theta)\hat{f}(x).$$

Note that all the estimators involved in the algorithm are obtained in a strictly univariate way, which simplifies their computation. The estimation of the marginal densities in Step 1, as well as the circular joining density in Step 2, can be done by parametric methods (for example, Maximum Likelihood) or by nonparametric procedures (using kernel methods).

Hall *et al.* (1987) propose a nonparametric kernel density estimator for directional data in $\mathbb{S}^q$. For the circular case ($q = 1$), denoting by Θ a random variable with density $\varphi(\cdot)$, the circular kernel density estimation from a sample $\{\Theta_i\}_{i=1}^n$ is given by

$$\hat{\varphi}_\nu(\theta) = \frac{c_0(\nu)}{n}\sum_{i=1}^n L\left(\nu\cos(\theta - \Theta_i)\right), \tag{3}$$

where $L(\cdot)$ is the circular kernel, ν is the circular smoothing parameter and $c_0(\nu)$ is a constant such that $\hat{\varphi}_\nu(\cdot)$ is a density. Some differences should be noted with respect to the linear estimator. First, the kernel function $L(\cdot)$ must be a rapidly varying function, such as the exponential (which corresponds to the von Mises density introduced in (1)), quite different from the bell shaped linear kernels. Secondly, small values of ν oversmooth the density, whereas large values of ν cause undersmoothing, quite the contrary to the linear bandwidth behaviour. See Hall *et al.* (1987) for a detailed description of the estimator and its properties.

As in the linear case, smoothing parameter selection is also an issue in circular kernel density estimation and there are still some open questions. Taylor (2008) proposes some automatic bandwidth selection methods as a rule of thumb based on the von Mises distribution, and the log–likelihood cross–validatory bandwidth, jointly with some other robust bandwidth selectors.

4. SIMULATION STUDY

In order to check the performance of the estimation algorithm for circular–linear densities, different estimation alternatives have been implemented for the marginals and the joining density, specificaly: parametric estimation (by Maximum Likelihood, ML), nonparametric kernel smoothers and a mixed approach, considering parametric marginals and nonparametric joining density. The following examples (see Figure 1) from Johnson and Wehrly (1978) are considered for circular–linear data generation.

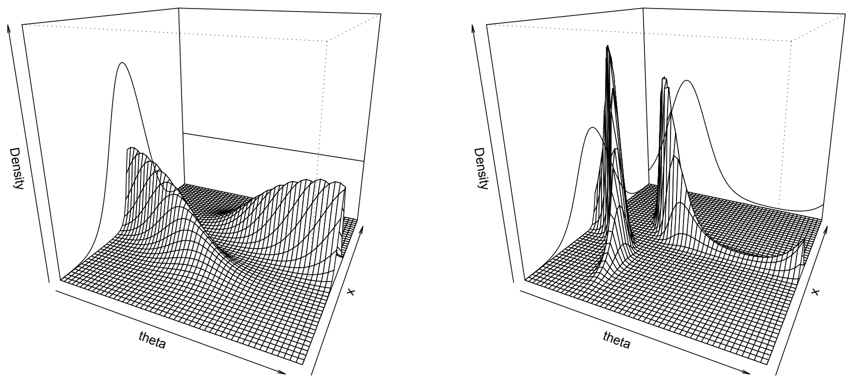

Figure 1: Left: Example 1 density surface with parameters $\mu = \pi$ and $\kappa = 2$. Right: Example 2 density surface with parameters $\mu = \pi$, $\kappa = 5$, $\mu_1 = \frac{\pi}{2}$ and $\kappa_1 = 2$.

Example 1 (Circular uniform and Normal marginal distributions). *Take $\varphi_U(\cdot)$ (circular uniform) and $f(\cdot) = \phi(\cdot)$ (standard Normal; $\Phi(\cdot)$ distribution) as marginal densities. Consider $g(\cdot)$, the joining density, as the von Mises density in (1) with parameters (μ, κ). The circular–linear density with margins $\varphi(\cdot)$ and $f(\cdot)$ is given by*

$$p_1(\theta, x) = \frac{1}{2\pi I_0(\kappa)} \exp\left[\kappa \cos(\theta - 2\pi\Phi(x) - \mu)\right] \phi(x). \tag{4}$$

Example 2 (von Mises and Normal marginal distributions). *Take the circular marginal as $\varphi_{vM}(\cdot)$ (von Mises density; Ψ_{vM} distribution) with parameters (μ_1, κ_1). With the same $f(\cdot)$ and $g(\cdot)$ as in the previous example, the joint circular–linear density is given by*

$$p_2(\theta, x) = \frac{1}{I_0(\kappa)} \exp\left[\kappa \cos\left(2\pi(\Psi_{vM}(\theta) - \Phi(x)) - \mu\right)\right] \varphi_{vM}(\theta)\phi(x). \tag{5}$$

As it has been commented in Section 2, the formulation of the joint circular–linear density in terms of copulas simplifies the simulation of random samples. This is achieved by simulating first the quantiles of (Θ, X) by the conditional copula, and then use the inversion method (see García–Portugués *et al.* (2011) for the simulation algorithm).

ML parametric density estimators, specifying the von Mises family for the circular distributions and the Normal family for the linear marginal, have been used for the purely parametric approach. Nonparametric estimation has been carried out using kernel methods: Gaussian kernel and cross–validatory bandwidth have been considered for obtaining $\hat{f}(\cdot)$ whereas for deriving $\hat{\varphi}(\cdot)$ and $\hat{g}(\cdot)$, the circular kernel density (3) has been implemented, with exponential kernel and cross–validatory bandwidth. In the mixed approach, ML has been used for obtaining $\hat{\varphi}(\cdot)$ and $\hat{f}(\cdot)$, whereas the circular kernel estimator has been considered for $\hat{g}(\cdot)$.

		MISE			Rel. efficiency	
	n	Parametric	Mixed	Nonparametric	Mixed	Nonparametric
Ex. 1	50	0.00538	0.00949	0.01680	0.56744	0.32077
	200	0.00128	0.00268	0.00518	0.48023	0.24831
	500	0.00051	0.00121	0.00249	0.42671	0.20778
	1000	0.00025	0.00066	0.00140	0.38969	0.18395
Ex. 2	50	0.04022	0.04828	0.09773	0.83305	0.41154
	200	0.01039	0.01368	0.03630	0.75950	0.28622
	500	0.00425	0.00595	0.01851	0.71404	0.22960
	1000	0.00213	0.00315	0.01066	0.67829	0.20056

Table 1: MISE for estimating the circular–linear density in Examples 1 and 2. Relative efficiencies for mixed and nonparametric estimations with respect to parametric estimation.

In order to check the performance of the procedure for estimating circular–linear densities, we consider the Mean Integrated Square Error (MISE) in the estimation of $p(\cdot,\cdot)$:

$$\text{MISE} = \iint \mathbb{E}\left[\hat{p}(\theta, x) - p(\theta, x)\right]^2 d\theta dx.$$

The MISE is approximated by Monte Carlo simulations, taking $M = 1000$ replicates. We compare the performance of the method using parametric estimation, nonparametric estimation and a mixed approach. Four sample sizes have been used: $n = 50$, $n = 200$, $n = 500$ and $n = 1000$. In the first example, the set–up parameters are $\mu = \pi$ and $\kappa = 2$. For the second example, we take $\mu = \pi$, $\kappa = 5$, $\mu_1 = \frac{\pi}{2}$ and $\kappa_1 = 2$.

Table 1 shows the simulation results for Examples 1 and 2. In all the cases, the MISE is reduced when increasing the sample size. Example 2 presents higher values for the MISE, and it is due to the estimation of a more complex structure in the circular marginal density (circular uniform in Example 1 and von Mises in Example 2). Obviously, the parametric method presents the lowest MISE values for all sample sizes in both examples, so it will be taken as a benchmark for computing the relative efficiencies of the nonparametric and mixed approaches. Relative efficiencies are obtained as the ratio between the MISE of the parametric method and the MISE of the mixed and nonparametric procedures. The relative efficiencies (see Table 1) are higher, in both examples, for the mixed approach, with better results for Example 2.

5. FINAL COMMENTS

The circular–linear density estimation algorithm based on Johnson and Wehrly (1978) proposal allows for the introduction of flexible estimators in the marginals. Although we have considered kernel density estimators for the marginals, other flexible density estimators could be used. In addition, the interpretation of the circular–linear density model in terms of copulas, enables the performance of simulation studies.

A natural question that may arise is the adequacy of model (2) for a certain circular–linear bivariate variable. In the simulation study presented in this paper, both examples satisfy this condition. To the best of our knowledge, there are no suitable tests for assessing if a circular–linear density can be expressed in this way.

6. ACKNOWLEDGEMENTS

The author acknowledge the support of Project MTM2008–03010, from the Spanish Ministry of Science and Project 10MDS207015PR from Dirección Xeral de I+D, Xunta de Galicia.

REFERENCES

Fernández–Durán, J.J. (2004) Circular distributions based on nonnegative trigonometric sums. *Biometrics*, **60**, 499-503.

Fernández–Durán, J.J. (2007) Models for circular–linear and circular–circular data constructed from circular distributions based on nonnegative trigonometric sums. *Biometrics*, **63**, 579-585.

García–Portugués, E., Crujeiras, R.M. and González–Manteiga, W. (2011) Exploring wind direction and SO_2 concentration by circular–linear density estimation. *Reports in Statistics and Operations Research.*

Hall, P., Watson, G.S. and Cabrera, J. (1987) Kernel density estimation with spherical data. *Biometrika*, **74**, 751-762.

Jammalamadaka, S.R. and SenGupta, A. (2001). *Topics in Circular Statistics.* World Scientific Press. Singapore.

Johnson, R. A. and Wehrly, T. (1978). Some angular–linear distributions and related regression models. *Journal of the American Statistical Association*, **73**, 602-606.

Nelsen, R.B. (2006) *An Introduction to Copulas.* Springer Series in Statistics. Springer.

Taylor, C.C. (2008) Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493-3500.

Wehrly, T. and Johnson, R.A. (1980) Bivariate models for dependence of angular observations and a related Markov process. *Biometrika*, **67**, 255-256.