Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **461,** 627–640 (2016)
Advance Access publication 2016 June 8

doi:10.1093/mnras/stw1335

SOPIE: an R package for the non-parametric estimation of the off-pulse interval of a pulsar light curve

Willem D. Schutte★ and Jan W. H. Swanepoel

School of Computer, Statistical and Mathematical Sciences, North-West University, 11 Hoffman Street, Potchefstroom 2531, South Africa

Accepted 2016 June 1. Received 2016 May 25; in original form 2015 October 29

ABSTRACT

An automated tool to derive the off-pulse interval of a light curve originating from a pulsar is needed. First, we derive a powerful and accurate non-parametric sequential estimation technique to estimate the off-pulse interval of a pulsar light curve in an objective manner. This is in contrast to the subjective 'eye-ball' (visual) technique, and complementary to the Bayesian Block method which is currently used in the literature. The second aim involves the development of a statistical package, necessary for the implementation of our new estimation technique. We develop a statistical procedure to estimate the off-pulse interval in the presence of noise. It is based on a sequential application of p-values obtained from goodness-of-fit tests for uniformity. The Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling and Rayleigh test statistics are applied. The details of the newly developed statistical package SOPIE (Sequential Off-Pulse Interval Estimation) are discussed. The developed estimation procedure is applied to simulated and real pulsar data. Finally, the SOPIE estimated off-pulse intervals of two pulsars are compared to the estimates obtained with the Bayesian Block method and yield very satisfactory results. We provide the code to implement the SOPIE package, which is publicly available at http://CRAN.R-project.org/package=SOPIE (Schutte).

Key words: methods: data analysis – methods: statistical – pulsars: general.

1 INTRODUCTION AND MOTIVATION

Since the launch of the *Fermi* Large Area Telescope (LAT) in 2008 (Atwood et al. 2009), the number of detected pulsars in the γ-ray domain has dramatically increased, currently[1] standing at in excess of 200. The opportunity to study a large number of these high-energy objects suddenly arose, and several research papers have been published on γ-ray pulsars and their associated pulsar wind nebulae (e.g. Abdo et al. 2009, 2010e,b,d,a,c; Ackermann et al. 2011; Grondin et al. 2013; Leung et al. 2014).

Over the years, statistical methods have played a major role in pulsar science. The H-test (De Jager, Raubenheimer & Swanepoel 1989; De Jager & Büsching 2010) has been widely used to characterize the significance of pulsations from newly detected pulsars and as a means for claiming a new pulsar discovery. Kerr (2011) has recently updated this technique by weighting each photon by its probability to have originated from the candidate pulsar using a spectral model, thereby improving background rejection and increasing its sensitivity for making new pulsar detections. Another example is afforded by the detection of a pulsar inside the CTA 1 supernova remnant within the first days following the launch of *Fermi* LAT (Abdo et al. 2008). This rapid detection was enabled by a novel semi-coherent 'photon arrival-time differencing' technique developed for sparse photon data (Atwood et al. 2006). Yet another new technique, originally developed to search for gravitational waves from pulsars (e.g. Rosado, Sesana & Gair 2015), enabled the first-ever discovery of a millisecond pulsar using only data from the γ-ray waveband (Pletsch et al. 2012a,b).

Although blind periodic searches have been successfully carried out using the LAT γ-ray data, detecting about a quarter of the current pulsar population in this way (e.g. Saz Parkinson et al. 2010), the usual mode of detection involves the use of pulsar parameters found from radio analyses. In order to gather and construct a pulse profile that is discernible above the background noise of the receiver, most pulsars require the coherent addition of many hundreds of pulses together through a process known as *(phase) folding* (Lorimer & Kramer 2005). In the context of γ-ray pulsars, when searching for new detections, one has to bear in mind that there are multiple sources of background photons. These are usually modelled by including a diffuse Galactic plus isotropic extragalactic diffuse background flux component in the spectral analysis (e.g. Johnson et al. 2013). Contamination by nearby sources may also play a role. Lastly, some young pulsars are surrounded by their associated nebulae, providing an additional background source. Therefore, a typical data set consists not only of pulsed radiation from the pulsar

★E-mail: wd.schutte@nwu.ac.za

[1] https://confluence.slac.stanford.edu/display/GLAMCOG/Public+List+of+LAT-Detected+Gamma-Ray+Pulsars

(sometimes referred to as the magnetospheric component or pulsed signal), but also of background. Such a data set would therefore contain photon arrival times t_i, with each arrival time representing either 'background' or pulsed radiation. Usually, the data set is pre-analysed so that the arrival times t_i are folded modulo 1 with respect to the pulsar's period P, which can be accurately determined from the *times of arrival* (TOAs). The unknown periodic density function (or light curve) $f(\theta)$ of the folded (modulo 1) arrival times can be represented as

$$f(\theta) = 1 - p + pf_s(\theta), \tag{1}$$

where $0 \leq p \leq 1$ represents the unknown signal strength of the periodic (or pulsed) signal and $f_s(\theta)$ is the unknown source function that characterizes the radiation pattern of the source. It is assumed that the background is uniformly distributed. The case $p = 0$ then corresponds to no signal (pure background), whereas $p = 1$ corresponds to no background (pure pulsed signal). The observed light curve $f(\theta)$ is thus represented as a mixture of the background and signal distribution.

The question of discriminating between the actual pulsed signal and the background has been a vital one, for at least two reasons. The first is that the so-called duty cycle of the pulsar light curve (the fraction of the rotational phase interval for which the pulsed signal is 'on') is a key prediction of emission models (as well as an indicator of pulsar geometry, e.g. smaller magnetic inclination angles should lead to larger duty cycles of the radio pulses). For example, Venter, Harding & Guillemot (2009) noted that standard two-pole caustic geometric pulsar models generally predict large duty cycles while outer gap models prefer sharper peaks and lower levels of emission outside the peaks. Accurate knowledge of the off-peak region's extent is therefore crucial for model discrimination. Secondly, one may wish to study the 'background source' surrounding a particular pulsar or population of pulsars. To do so requires the separation of the pulsed emission by the pulsar(s) from the steady background emission. The latter may be a pulsar wind nebula in the case of young pulsars, or a globular cluster in the case of a population of old pulsars. To study the 'background' signal, one has to accumulate data in phase intervals where the pulsed signal is switched 'off'. Such a 'gating' procedure has been applied during the analysis of γ-ray data to detect GeV pulsar wind nebulae (Ackermann et al. 2011). Another example is afforded by the detection of the millisecond pulsar J1823−3021A in the globular cluster NGC 6624: the flux upper limit derived for the cluster emission in the off-pulse phase of this pulsar implies that the cluster contains less than 32 γ-ray millisecond pulsars. It is important to constrain the size of the millisecond pulsar population in globular clusters, as this is a key ingredient in models of steady broad-band flux (e.g. Kopp et al. 2013; Zajczyk, Bednarek & Rudak 2013) or cosmic rays (e.g. Venter et al. 2015) initiated by such ensembles of pulsars. The region where the pulsed signal is 'off' therefore represents an important observable that is of both practical and theoretical use. Formally, we define this off-pulse window (also called the off-peak or off-pulse interval) as the period of time when the pulsar is 'off', implying the time period where the pulsar is not recognized as the source of (or 'responsible' for) any γ-rays that are received by the detection instruments.

Many researchers utilize the histogram-type estimated light curves to perform subsequent analyses on these curves in order to understand pulsar magnetospheres and their associated nebulae better (Saz Parkinson et al. 2010; Ray et al. 2011). Some analyses on these histogram-type estimated light curves are done with the 'eye-ball' technique, or visual inspection of the histogram in order to identify the off-peak phase interval (e.g. Saz Parkinson et al. 2010; Ng et al. 2014). Another less arbitrary method is to perform likelihood analysis in small phase bins, gradually increasing the width of consecutive bins (Abdo et al. 2012). For all of these width-varying bins only the left and right limits are changing while the centre of the bin is fixed at what one believes to be the centre of the off-pulse interval. The identified off-peak window is then used to estimate the unpulsed emission level, which in turn is used to study the properties of magnetospheric pulsar or pulsar wind nebula emission.

Related work in this regard is that developed by Scargle (1998), who introduced the Bayesian Blocks procedure for detecting localized structures (bursts), revealing pulse shapes, and generally characterizing intensity variations in the time series data and other forms of sequential data. Scargle et al. (2013) present a simple non-parametric modelling technique and an algorithm implementing it (an improved and generalized version of the Bayesian Blocks) that finds the optimal segmentation of the time series data in the observation interval. This algorithm (with modifications) is used in Abdo et al. (2013) to define the off-peak interval for a pulsars light curve. To avoid potential contamination from the trailing and leading edges of the peak, the authors reduce the extent of the block by 10 per cent on either side. It is also reported that the Bayesian Block algorithm fails to model weak peaks for a few pulsars. For these pulsars, the authors conservatively shrink the off-peak region.

In contrast to the *subjective* 'eye-ball' approach to identify the off-pulse interval visually, and *complementary* to the Bayesian Blocks approach, the SOPIE package is developed to implement an *objective* technique for estimating the off-pulse interval non-parametrically. The focus of our technique is the identification of the points (time interval) where the unknown source function $f_s(\theta)$ is discernible from the background noise in equation (1). Essentially, the goal of the algorithm is to estimate a time interval over which the total measured signal $f(\theta)$ is uniform. Therefore, two unknown points a and b need to be estimated from the data. Although the pulsar's unknown source function $f_s(\theta)$ is almost certainly never exactly 0, it is assumed that the signal decays to unmeasurable values in the off-pulse interval. Furthermore, the use of the term 'non-parametric' refers to the fact that we do not assume any known functional form of the source function $f_s(\theta)$ in equation (1).

In this paper, we describe the development of a non-parametric sequential estimation technique to objectively estimate the off-pulse interval of a pulsar light curve. The SOPIE package is therefore developed to estimate the unknown values a and b without assuming any known functional form of the source function. The primary source of data that will be used in the application of the SOPIE package is high-energy pulsar radiation. This package is an implementation of our proposed statistical technique to estimate the off-pulse region of a pulsar light curve, available for R (R Core Team 2015) from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/package=SOPIE (Schutte 2015). The remainder of this paper is structured as follows. A brief background on circular statistics is given in Section 2, followed by Section 3 that focuses on the off-pulse estimation algorithm. The complete package is then explored in Section 4 using simulated and practical examples. Some closing remarks are given in Section 5, while the appendices contain the details of the functions available in the SOPIE package.

2 CIRCULAR STATISTICS

It is important to note that data defined on a circle require different techniques to those defined on the line. Therefore, linear techniques

cannot always be applied directly to circular data. Specific statistical methods and techniques have been developed to handle circular data and can be found in published books such as Mardia (1972), Fisher (1993), Mardia (1992) and Mardia & Jupp (2000). The development in the field of density estimation pertaining to circular- and spherical data also received ample attention in the early 1990s, continuing into the new millennium (Hall, Watson & Cabrera 1987; Bai, Rao & Zhao 1988; Fisher 1989; Klemelä 2000; Agostinelli 2007; Taylor 2008). Density estimation, and specifically Kernel Density Estimation (KDE) on the straight line is, however, no new topic, and was first introduced by Rosenblatt (1956) and Parzen (1962). Other excellent references such as Scott (1992), Silverman (1986) and Wand & Jones (1995) also exist.

Although a full review of circular statistics will not be given, it is essential to have some understanding of the topic. Therefore, certain fundamental concepts relating to circular statistics will be given and explained, especially those concepts that relate to circular KDE. The motivation for emphasizing circular statistics and specifically the circular KDE is based on the fact that our off-pulse estimation algorithm uses the circular KDE as an input. The next subsection discusses some fundamental circular statistical ideas, followed by an overview of circular KDE techniques.

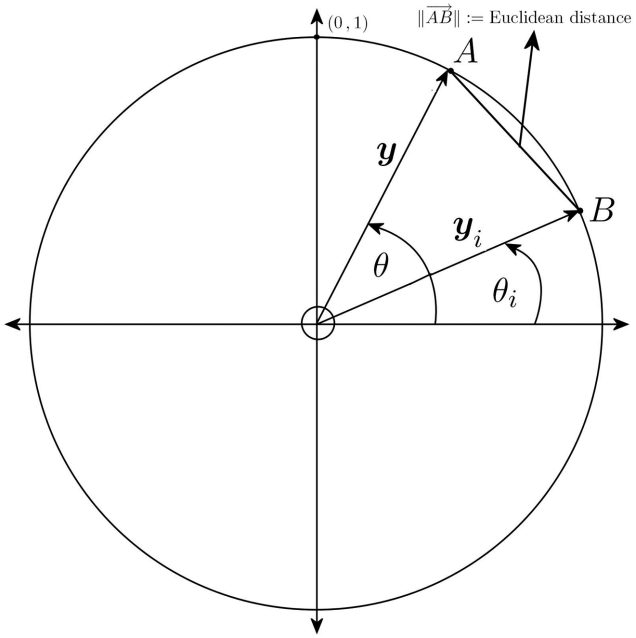

Figure 1. Representation of data on a circle classified as angles and Euclidean distance.

2.1 Circular ground work and notation

Observations on the (unit) circle can be regarded as unit vectors $\boldsymbol{y}$. Each point $\boldsymbol{y}$ on the circle can be represented as an angle θ or differently stated by $\boldsymbol{y} = \binom{\cos\theta}{\sin\theta}$.

It is important to state that all angles are measured in radians, and for all calculations, two points on the circle, namely θ and $\theta + 2\pi$ will be exactly the same point. Without loss of generality, it is also possible to convert sample observations on the interval $[0, 2\pi)$ to observations on the interval $[0, 1)$ by dividing with 2π.

A useful starting point for any analysis regarding circular data, is the definition of some measures of location, such as the mean direction, the median direction and the resultant length. The mean direction is obtained by treating the data as unit vectors and using the direction of their resultant vector. For a sample of unit vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, with corresponding angles $\theta_1, \ldots, \theta_n$, the resultant vector $\boldsymbol{E}$ is obtained by adding up the unit vectors componentwise:

$$\boldsymbol{E} = \sum_{i=1}^{n} \boldsymbol{y}_i. \tag{2}$$

This is a vector with length between 0 and n, and pointing in the mean direction $\bar{\theta}$ of the sample. The sample mean resultant length (standardized length) is given by:

$$\bar{R} = \frac{\|\boldsymbol{E}\|}{n}, \tag{3}$$

with $\bar{R} \in [0, 1]$, and $\|\cdot\|$ the Euclidean norm. If the data are closely clustered around the mean, then $\bar{R}$ is close to 1. However, if the data are evenly spread around the circle, $\bar{R}$ will be near zero. Hence, $\bar{R}$ is a natural measure of dispersion. Furthermore, the resultant vector $\boldsymbol{E}$ can be decomposed into two components, namely:

(i) the mean direction $\bar{\theta}$, and
(ii) the mean resultant length $\bar{R}$.

Fisher (1993) defines the sample median direction $\tilde{\theta}$ as the value of θ that maximises the function

$$d(\theta) = \sum_{i=1}^{n} |\pi - |\theta_i - \theta||. \tag{4}$$

For a random sample of unit vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ with corresponding angles $\theta_1, \ldots, \theta_n \in [0, 2\pi)$, it is important to define some distance measures between the sample observations on the circle. In order to thoroughly explain the idea of the distance measure between observations on a circle, the following definitions are required (the reader is also referred to Fig. 1).

Let

$$\boldsymbol{y} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \text{ and } \boldsymbol{y}_i = \begin{bmatrix} \cos\theta_i \\ \sin\theta_i \end{bmatrix}.$$

Equivalently one can denote point $A = (\cos\theta, \sin\theta)$ and point $B = (\cos\theta_i, \sin\theta_i)$. Also define $\|\boldsymbol{AB}\| :=$ Euclidean distance (straight line distance) between points A and B, where $\boldsymbol{AB}$ denotes the vector from A to B. Applying the law of cosines on the triangle OAB in Fig. 1, one obtains

$$\|\boldsymbol{AB}\|^2 = 2(1 - \cos(2\pi - |\theta - \theta_i|)). \tag{5}$$

The following definition is now introduced to define the distance between two angles θ and θ_i (Taylor 2008):

$$d_i(\theta) := \min(|\theta - \theta_i|, 2\pi - |\theta - \theta_i|), \tag{6}$$

where $|\theta - \theta_i|$ denotes the usual absolute value. This definition follows from the fact that the smallest angle between observations must be used as distance measure, even if the angles seem to be distant from one another on the real line.

The formulas in equations (5) and (6) will be used to define the circular KDE in the following section.

2.2 Circular kernel density estimation

Considerable literature exists that investigates the non-parametric estimation of a probability density function of a random variable through the use of kernel functions. Frequently, references such as Silverman (1986) and Wand & Jones (1995) apply KDE to data on the real line. There are some references that apply KDE to circular data, such as Hall et al. (1987), Bai et al. (1988), Fisher (1989), Klemelä (2000), Taylor (2008), Oliveira, Crujeiras & Rodriquez-Casal (2012) and Garcia-Portugues, Crujeiras & Gonzalez-Manteiga (2013). In order to develop standard notation that will be used throughout the text, some detail about KDE will be given.

Let X_1, ..., X_n be independent and identically distributed (IID) from some unknown density f on the real line. The well-known *kernel estimator* of the density f is given for some kernel function k by:

$$\hat{f}_{n,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right), \tag{7}$$

where h is the so-called bandwidth or smoothing parameter (Silverman 1986, p. 15).

Taylor (2008) proposes the use of angular distances when the data are on the circle, rather than using the Euclidean distance $x - X_i$ in (7). Applying (5) and (6), the kernel density estimator on the circle is defined for some given kernel function k by:

$$\hat{f}_{n,h}(\theta) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{1 - \cos(d_i(\theta))}{h}\right) \Big/ c_f(h), \tag{8}$$

where $\mathrm{d}_i(\theta)$ is given in equation (6), and

$$c_f(h) = \int_0^{2\pi} \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{1 - \cos(d_i(\theta))}{h}\right) \mathrm{d}\theta\,, \tag{9}$$

is known as the normalization constant to ensure that $\hat{f}_{n,h}(\theta)$ integrates to unity. In order to apply the circular KDE to data, some kernel function k must be chosen, together with a value for the smoothing parameter h. The next subsection will first expand on the choice of a kernel function k, followed by a subsection that will investigate some data-based choices of the smoothing parameter h for circular data.

2.2.1 *Choice of kernel function k*

It is important to note that various kernel functions can be chosen for the kernel estimator in equation (8). Popular choices of kernel functions include the Gaussian-like kernels in circular data such as the von Mises, and wrapped normal distribution with unbounded support. Other popular kernels with compact support are the so-called 'polynomial' kernels of the form:

$$k(x) = \begin{cases} \kappa_{rs}(1 - |x|^r)^s, & \text{if } -1 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where

$$\kappa_{rs} = \frac{r}{2B\left(s+1, \frac{1}{r}\right)}, \quad r > 0,\ s \ge 0,$$

with $B(\cdot, \cdot)$ denoting the beta-function. The rectangular kernel is obtained if $s = 0\,(\kappa_{r0} = \frac{1}{2})$; the Triangular kernel if $r = 1$, $s = 1\,(\kappa_{11} = 1)$; the Epanechnikov kernel if $r = 2, s = 1\,(\kappa_{21} = \frac{3}{4})$; the bi-weight kernel if $r = 2, s = 2\,(\kappa_{22} = \frac{15}{16})$; and the tri-weight kernel if $r = 2$, $s = 3\,(\kappa_{23} = \frac{35}{32})$. The simplest form of the Gaussian kernel can be obtained if $r = 2$, $s = \infty$ after a suitable rescaling (Loots 1995).

It is a well-known fact that, based on the definition of efficiency, the choice of kernel function is not the most important component of kernel density estimation, as there is little to choose between the various kernels on the basis of the mean integrated squared error (Silverman 1986; Wand & Jones 1995). Nevertheless, some kernel function must be chosen. Therefore, in the SOPIE package, only the Epanechnikov kernel is used. The reason for choosing the Epanechnikov kernel is because it is known to be an optimal kernel that minimises the asymptotic mean integrated squared error (Silverman 1986, p. 40). An extensive simulation study was conducted to verify the influence of the kernel function k on the estimated off-pulse interval. It was found that the choice of kernel function k does not significantly influence the estimated off-pulse interval (Schutte 2014, p. 58). The reason for this behaviour is due to the fact that the proposed algorithm (see Section 3.1) only utilizes the circular KDE as a first step to obtain a starting point for the algorithm to commence. The remainder of the algorithm is completely independent of a kernel estimator and utilizes the raw photon arrival times.

2.2.2 *Choice of smoothing parameter h*

One of the difficulties in non-parametric density estimation is to make a good data-based choice of the smoothing parameter h. Several excellent references on this topic exist, such as Silverman (1986), Hall et al. (1991), Sheather & Jones (1991), Jones, Marron & Sheather (1996) and Oliveira et al. (2012), to name but a few. When the data are observed in Euclidean space, there are many approaches to the problem, i.e. when the kernel function is taken as the Gaussian density in Euclidean space, a well-known choice for h is given by (Silverman 1986):

$$\hat{h} = 1.06\hat{\sigma}n^{-1/5}, \tag{11}$$

where $\hat{\sigma}$ is some estimate of the measure of dispersion, such as the linear standard deviation of the observations. Taylor (2008) proposes a class of smoothing parameter estimates for circular data which uses a robust measure of dispersion, such as the circular interquartile range, i.e.

$$\hat{h} = 1.06\frac{\mathrm{IQR}_\circ}{1.349}n^{-1/5}, \tag{12}$$

with $\mathrm{IQR}_\circ$ the estimated circular inter-quartile range defined in equation (B4).

It is evident that one can derive several different data-driven choices $\hat{h}$ based on different measures of dispersion. Appendix B contains a list of the estimated measures of dispersion that are used in the definition of the estimated smoothing parameters, available in the SOPIE package. In summary, nine different estimated smoothing parameters are available in SOPIE when the circular kernel density estimator is calculated for the data. In the discussion that follows, reference will only be made to the choice of smoothing parameter $\hat{h}_i$, $i = 1 \ldots 9$, according to Table 1.

An extensive simulation study was conducted to examine the influence of the estimated smoothing parameter $\hat{h}$ on the estimated off-pulse interval. The results show that most of the $\hat{h}$-choices can be used, since it only has a marginal influence on the Monte Carlo estimates of the bias and Mean Squared Error (MSE). In only a selected number of target populations it is found that $\hat{h}_4$, $\hat{h}_5$, $\hat{h}_7$, and $\hat{h}_8$ perform marginally worse than other choices of $\hat{h}$ based on bias and MSE. For small data sets, $\hat{h}_5$ performs poor, and would normally

Table 1. Smoothing parameter references and equations.

Smoothing parameter estimates
$\hat{h}_1 = 1.06sn^{-1/5}$
$\hat{h}_2 = 1.06s_\circ n^{-1/5}$
$\hat{h}_3 = 1.06\bar{D}_\circ n^{-1/5}$
$\hat{h}_4 = 1.06\lvert D_\circ\rvert n^{-1/5}$
$\hat{h}_5 = 1.06\mathrm{IQR}_\circ n^{-1/5}$
$\hat{h}_6 = \frac{1.06}{1.349}\mathrm{IQR}_\circ n^{-1/5}$
$\hat{h}_7 = 0.9s_\circ n^{-1/5}$
$\hat{h}_8 = \frac{0.9}{1.349}\mathrm{IQR}_\circ n^{-1/5}$
$\hat{h}_9 = \frac{1}{8}\sum_{i=1}^{8} h_i$

be closely related to $\hat{h}_6$.Therefore, it is recommended to use any one of $\hat{h}_1$, $\hat{h}_2$ or $\hat{h}_3$. Again it must be emphasized that the choice of the estimated smoothing parameter $\hat{h}$ does not significantly influence the estimated off-pulse interval, since the proposed algorithm only utilizes the circular KDE as a first step to obtain a starting point for the algorithm to commence. By default, SOPIE chooses $\hat{h}_1$ for the estimated smoothing parameter.

3 ESTIMATION OF THE OFF-PULSE INTERVAL

As mentioned throughout, the main aim is to estimate the endpoints of the interval on which the unknown light curve $f(\theta)$ is uniform. Following from the discussion in Section 1, suppose $p = 1$ in equation (1), implying that no background component is present in the light curve $f(\theta)$. If this is the case, then the off-pulse interval $[a, b]$ (as described in Section 1) can easily be estimated as follows:

Let $\theta_1, \theta_2, \ldots, \theta_n$ be a continuous random sample and denote the order statistics by

$$0 \leq \theta_{(1)} \leq \theta_{(2)} \leq \ldots \leq \theta_{(n)} \leq 2\pi.$$

The arc lengths between adjacent observations are defined by Mardia & Jupp (2000) as:

$$T_i = \theta_{(i+1)} - \theta_{(i)}, \quad 1 \leq i \leq n-1; \quad T_n = 2\pi - (\theta_{(n)} - \theta_{(1)}). \tag{13}$$

The index of the maximal arc length between two adjacent observations is then defined by

$$N = \arg\max_{1 \leq i \leq n} T_i. \tag{14}$$

It is now possible to obtain a non-parametric estimator for the interval $[a, b]$ in the absence of any background, which is given by

$$\hat{a} = \theta_{(N)} \quad \text{and} \quad \hat{b} = \theta_{(N+1)} \quad \text{if } N \leq n-1, \tag{15}$$

or

$$\hat{a} = \theta_{(N)} \quad \text{and} \quad \hat{b} = \theta_{(1)} \quad \text{if} \quad N = n. \tag{16}$$

Note that equation (16) implies that the on-pulse interval is estimate by $[\theta_{(1)}, \theta_{(N)}]$.

However, for typical astrophysical data, background is always present. We therefore propose a sequential estimation technique in the next section that will take this into account.

3.1 A new sequential method to estimate the off-pulse interval

The remainder of this paper is concerned with estimating the off-pulse interval $[a, b]$ in the presence of some background. Broadly speaking, this technique is based on a sequential application of p-values obtained from goodness-of-fit tests for the uniform distribution (see, e.g. D'Agostino & Stephens (1986) to name one reference). To be more specific, the well-known Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling and Rayleigh test statistics are applied sequentially on subintervals of [0, 1).

First, a result is stated that provides the theoretical justification for the proposed procedure. The proof of the result can be found in (Schutte 2014, p. 35). For this, some further notation needs to be introduced:

Let $\theta_1, \theta_2, \ldots, \theta_n$ be a sample of IID uniform random variables on the interval [0, 1] with corresponding order statistics

$$\theta_{(1)} \leq \theta_{(2)} \leq \ldots \leq \theta_{(n)}. \tag{17}$$

Suppose that r and s are integers such that $1 \leq r \leq s-1 \leq n-2$. Denote by $f(\theta^*_{r+1}, \ldots, \theta^*_s | \theta^*_r, \theta^*_{s+1})$ the joint conditional density function of $\theta_{(r+1)}, \ldots, \theta_{(s)}$ given that $\theta_{(r)} = \theta^*_r$ and $\theta_{(s+1)} = \theta^*_{s+1}$.

Proposition

$$f\left(\theta^*_{r+1}, \ldots, \theta^*_s | \theta^*_r, \theta^*_{s+1}\right)$$

$$= \begin{cases} \frac{(s-r)!}{(\theta^*_{s+1} - \theta^*_r)^{s-r}}, & \text{for } \theta^*_r \leq \theta^*_{r+1} \leq \ldots \leq \theta^*_s \leq \theta^*_{s+1}, \\ 0, & \text{elsewhere.} \end{cases} \tag{18}$$

Remark

The result stated in the Proposition can be interpreted as follows: The joint conditional density of $\theta_{(r+1)}, \ldots, \theta_{(s)}$, given that $\theta_{(r)} = \theta^*_r$ and $\theta_{(s+1)} = \theta^*_{s+1}$, is the same as the joint unconditional density of $s - r$ order statistics corresponding to an IID sample of size $s - r$ from the uniform distribution on the interval $[\theta^*_r, \theta^*_{s+1}]$. The proposition states that even though we are working on sub-intervals of [0, 1), the joint density of the order statistics under the null hypothesis of uniformity, remains uniform over the sub-interval under consideration.

The following algorithm is now proposed to estimate the off-pulse interval of a source function originating from a pulsar. The algorithm is implemented in the SOPIE package (Sequential Off-Pulse Interval Estimation).

Algorithm

(i) Calculate the point where the kernel density estimator $\hat{f}_{n,h}(\theta)$ attains its *global* minimum value:

$$x_1 := \arg\min_{\theta} \hat{f}_{n,h}(\theta). \tag{19}$$

Also, determine the next m *local* minimum points, i.e. for $i = 2, 3, \ldots, m$, let

$$x_i := \arg\min_{\theta \notin \{x_1, \ldots, x_{i-1}\}} \hat{f}_{n,h}(\theta). \tag{20}$$

(ii) For each of the selected minima points $x_1, x_2, \ldots, x_m$, find the nearest ordered observation to this point:
Let

$$k_i := \arg\min_{1 \leq j \leq n} |\theta_{(j)} - x_i|, \qquad i = 1, 2, \ldots, m, \tag{21}$$

then the nearest observation to x_i will be $\theta_{(k_i)}$, $i = 1, 2, \ldots, m$.

(iii) Duplicate the initial ordered observations between 0 and 1 to the left of 0 and to the right of 1. That is, define

$$\theta_{(-n+i)} = \theta_{(i)} - 1 \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{22}$$

$$\theta_{(n+i)} = \theta_{(i)} + 1 \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{23}$$

The result is the ordered observations $\theta_{(-n+1)}, \theta_{(-n+2)}, \ldots, \theta_{(0)}, \theta_{(1)}, \ldots, \theta_{(n)}, \theta_{(n+1)}, \ldots, \theta_{(2n)} \in [-1; 2]$.

(iv) For some specified integer g, define

$$n_g := \left\lfloor \frac{n-1}{g} \right\rfloor \quad \text{and} \quad \rho_g := \frac{n-1}{g}. \tag{24}$$

Evaluate ρ_g to determine whether it is an integer or not.

If ρ_g is integer-valued, first consider k_1 (corresponding to x_1) and define for each $\ell = 1, 2, \ldots, n_g$ the following set of observations:

$$\chi_\ell := \{\theta_{(k_1)}, \theta_{(k_1+1)}, \ldots, \theta_{(k_1+\ell g)}, \theta_{(k_1+\ell g+1)}\}. \tag{25}$$

If ρ_g is not an integer, consider k_1 (corresponding to x_1) and define

$$\chi_\ell^0 := \begin{cases} \chi_\ell, & \ell = 1, 2, \ldots, n_g, \\ \{\theta_{(k_1)}, \theta_{(k_1+1)}, \theta_{(k_1+2)}, \ldots, \theta_{(k_1+n)}\}, & \ell = n_g + 1\,. \end{cases} \tag{26}$$

(v) Let $T_n(\chi_\ell)$ (or $T_n(\chi_\ell^0)$) be a given test statistic to test *uniformity* based on the observations in the set χ_ℓ (or χ_ℓ^0), when ρ_g is an integer (or not). Denote by P_ℓ the corresponding p-value. Four different tests for uniformity are applied, namely, the Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling and Rayleigh test (see Section 3.3 for details about the different goodness-of-fit tests). For each of the four goodness-of-fit tests, calculate P_ℓ sequentially for $\ell = 1, \ldots, N$, where N is a stopping time defined by

$N :=$ the smallest integer j, with $1 \le j \le n_g - r + 1$ for which

$\max_{j \le \ell \le j+r-1} P_\ell \le \alpha$, if ρ_g is an integer, or

$N :=$ the smallest integer j, with $1 \le j \le n_g - r + 2$ for which

$\max_{j \le \ell \le j+r-1} P_\ell \le \alpha$, if ρ_g is not an integer.

Note that it is possible that N does not exist. Here α is the so-called significance level which is usually taken as $\alpha = 0.01$, 0.05 or 0.1 in the statistical literature when hypothesis testing is conducted. Furthermore, r is an integer representing an applicable tuning parameter.

(vi) The right endpoint (boundary) of the unknown off-pulse interval $I = [a, b]$ is then estimated (when using x_1) by

$$\hat{b}_1 := \begin{cases} \theta_{(k_1+Ng+1)}, & \text{if } N \text{ exists and } k_1 + Ng + 1 \le n, \\ \theta_{(k_1+Ng-n+1)}, & \text{if } N \text{exists and } k_1 + Ng + 1 > n, \\ \theta_{(k_1)}, & \text{if } N \text{ does not exist.} \end{cases} \tag{27}$$

(vii) Repeat steps (iv)–(vi) for x_2, x_3, …, x_m to obtain $\hat{b}_2, \hat{b}_3, \ldots, \hat{b}_m$.

(viii) The estimate of the right endpoint (boundary) of the unknown off-pulse interval $I = [a, b]$ for a given goodness-of-fit test, is

$$\hat{b} := \frac{1}{m} \sum_{j=1}^{m} \hat{b}_j\,. \tag{28}$$

Once the estimated right endpoint of the unknown off-pulse interval is obtained for each of the four goodness-of-fit tests, the recommended estimated right endpoint is the median value of the four estimates.

Next, consider the estimation of a of the unknown off-pulse interval $I = [a, b]$. The procedure is similar to the procedure to estimate b, but for completeness, the details are provided where it differs from the previous. Therefore, follow a similar procedure as the above, but replace steps (iv)–(viii) with the following:

(iv) Define n_g and ρ_g as before and evaluate ρ_g to determine whether it is an integer or not.

If ρ_g is integer-valued, we first consider k_1 (corresponding to x_1) and define for each $\ell = 1, \ldots, n_g$ the following set of observations:

$$\chi_\ell := \{\theta_{(k_1-\ell g-1)}, \theta_{(k_1-\ell g)}, \ldots, \theta_{(k_1-1)}, \theta_{(k_1)}\}. \tag{29}$$

If ρ_g is not an integer, consider k_1 (corresponding to x_1) and define

$$\chi_\ell^0 := \begin{cases} \chi_\ell, & \ell = 1, \ldots, n_g, \\ \{\theta_{(k_1-n)}, \theta_{(k_1-n+1)}, \theta_{(k_1-n+2)}, \ldots, \theta_{(k_1)}\}, & \ell = n_g + 1. \end{cases} \tag{30}$$

(v) Define $T_n(\chi_\ell)$ (or $T_n(\chi_\ell^0)$), P_ℓ and N as previously.

(vi) The left endpoint (boundary) of the unknown off-pulse interval $I = [a, b]$ is then estimated (when using x_1) by

$$\hat{a}_1 := \begin{cases} \theta_{(k_1-Ng-1)}, & \text{if } N \text{ exists and } k_1 - Ng - 1 \ge 1, \\ \theta_{(k_1-Ng+n-1)}, & \text{if } N \text{ exists and } k_1 - Ng - 1 < 1\,, \\ \theta_{(k_1)}, & \text{if N does not exist.} \end{cases} \tag{31}$$

(vii) Repeat steps (iv)–(vi) for x_2, x_3, …, x_m to obtain $\hat{a}_2, \hat{a}_3, \ldots, \hat{a}_m$.

(viii) The estimate of the left endpoint (boundary) of the unknown off-pulse interval $I = [a, b]$ for a given goodness-of-fit test, is

$$\hat{a} := \frac{1}{m} \sum_{j=1}^{m} \hat{a}_j\,. \tag{32}$$

The recommended estimated left endpoint of the unknown off-pulse interval is again the median value of the four estimates obtained from the goodness-of-fit tests.

In order to facilitate a better understanding of the proposed algorithm, a flow chart of the algorithm is supplied in Fig. 2. The flow chart illustrates the broad details of the algorithm with references to the exact steps. Furthermore, it is evident from the algorithm that several tuning parameters must be specified. The next section will provide some details on the tuning parameters and recommended choices for each parameter.

3.2 Choice of tuning parameters in SOPIE

In the previous section, a new sequential method to estimate the off-pulse interval is proposed. Several tuning parameters are used in the algorithm and the aim of this section is to provide more detail about each of the tuning parameters, together with recommended choices for the values of the parameters.

3.2.1 The number of minimum points, m

The first tuning parameter is the choice of m, the number of local minimum points that are used. It is evident that the computation time of the algorithm is dependent on the choice of m, since the complete procedure is reiterated for every selected minimum point. Based on the results from the simulation study, it is clear that larger values of m influence the computation time of the procedure, although this is not considered as a serious constraint. What

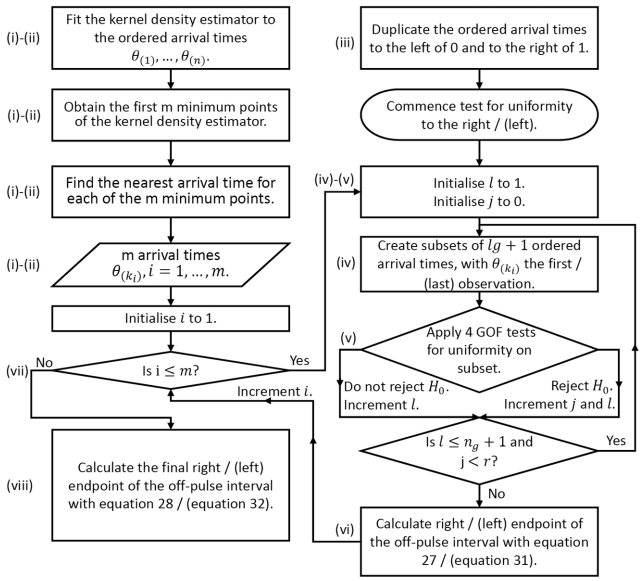

Figure 2. Flowchart showing the steps in the algorithm for the estimation of the off-pulse interval. The corresponding steps of the algorithm are indicated to the left of each item. Only broad details are given, while exact calculations are provided in the algorithm. Note that step (iii) in the top right-hand side of the flow chart has no predecessor and can thus be executed parallel to steps (i) and (ii).

is important, though, is the impact of m on the estimated off-pulse intervals. For all of the simulated target populations, none of the off-pulse estimations dramatically changes when $m = 1$ is compared to $m > 1$. Furthermore, for $m > 1$, averaging is used over all of the chosen minimum points m to obtain the estimated off-pulse interval. It is recommended to choose $m = 1$, but one could also choose any value up to $m = 5$. By default, SOPIE chooses $m = 1$.

3.2.2 *The incremental growth of each interval being tested for uniformity, g*

The tuning parameter g represents the value of the incremental growth of each subsequent interval over which uniformity is tested. Since uniformity is sequentially tested, with the interval used in the test growing by g observations in every iteration, the selection of g not only influences the computation time of the procedure, but also has an effect on the point where rejection of the hypothesis may take place. Based on the results of the simulation study, recommendations for the value of g will be given at the end of this section, since it was found that the significance level α is also related to the choice of r and g.

3.2.3 *The number of intervals of rejection, r*

The tuning parameter r represents the number of subsequent intervals that must result in the rejection of uniformity before SOPIE will stop. The choice of r must therefore be linked to the choice of g. For large values of g, the user takes the risk that uniformity is rejected for a certain (larger) interval, while it should have been rejected earlier (for a smaller interval). Small values of g may also result in the early rejection of uniformity, e.g. in the situation where a few observations may cause the rejection of uniformity, while uniformity is again confirmed when several more observations are included in the interval. Thus, for smaller values of g, it would be safer to select larger values of r, and vice versa. One should also note that, for a large value of r, there will be no influence on the value of $\hat{b}_1$ in equation (27) or $\hat{a}_1$ in equation (31) if rejection takes place for each interval after a certain point. Recommendations for choices of g and r are given in the next subsection.

3.2.4 *The significance level, α*

For any situation where a hypothesis is tested, it is required to choose the level of significance α. Since the algorithm sequentially applies goodness-of-fit tests, the value of α must be selected. In the simulation study it is found that the significance level α is also related to the choice of r and g. If only α is considered, use $\alpha \leq 0.05$ for large sample sizes, and $0.05 \leq \alpha \leq 0.10$ for small to moderate sample sizes. The simulation results also indicate that the value of α has a limited effect on the Cramér–von Mises goodness-of-fit test, but the Anderson–Darling and Kolmogorov–Smirnov goodness-of-fit tests are more sensitive to different choices of α. In general, it is recommended to use small α-value such as $\alpha = 0.01$ or $\alpha = 0.05$, with $1 \leq g \leq 40$ and $1 \leq r \leq 10$. By default, SOPIE chooses $\alpha = 0.05$, $g = 20$ and $r = 10$.

3.3 Goodness-of-fit tests for uniformity of circular data

The algorithm specified in Section 3.1 for estimating a and b is based on tests for uniformity on a circle. Throughout this paper, the Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling and Rayleigh goodness-of-fit tests are applied (D'Agostino & Stephens 1986; Mardia & Jupp 2000). Each of these test statistics is briefly discussed in the next subsections. The notation defined in Section 3.1 is used throughout.

3.3.1 *Kolmogorov–Smirnov test for uniformity*

The test statistic for the Kolmogorov–Smirnov goodness-of-fit test is given for $i = 1, 2, \ldots, m$ by

$$T_{n,i}^{KS}(\chi_\ell) := \max\left(T_{n,i}^{D^+}(\chi_\ell), T_{n,i}^{D^-}(\chi_\ell)\right), \tag{33}$$

where

$$T_{n,i}^{D^+}(\chi_\ell) := \max_{1\leq j\leq \ell g}\left(\frac{j}{\ell g} - \frac{\theta_{(k_i+j)} - \theta_{(k_i)}}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}\right), \text{ and} \tag{34}$$

$$T_{n,i}^{D^-}(\chi_\ell) := \max_{1\leq j\leq \ell g}\left(\frac{\theta_{(k_i+j)} - \theta_{(k_i)}}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}} - \frac{j-1}{\ell g}\right), \tag{35}$$

for $\ell = 1, 2, \ldots, n_g$ (Stephens 1970). $T_{n,i}^{KS}(\chi_\ell^0)$ is defined similarly for $\ell = 1, 2, \ldots, n_g, n_g + 1$.

For calculating the p-values, the method of Marsaglia, Tsang & Wang (2003) is used. This method provides an accuracy of up to 15 digits for sample sizes ranging from 2 to at least 16 000.

3.3.2 *Cramér–von Mises test for uniformity*

The Cramér–von Mises goodness-of-fit test is a well-known and widely used non-parametric test. The test was introduced by Cramér (1928) and Von Mises (1931). The test statistic is given, for $i = 1, 2, \ldots, m$, by (Stephens 1970)

$$T_{n,i}^{CvM}(\chi_\ell) := \sum_{j=1}^{\ell g}\left(\frac{\theta_{(k_i+j)} - \theta_{(k_i)}}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}} - \frac{j-\frac{1}{2}}{\ell g}\right)^2 + \frac{1}{12\ell g}, \tag{36}$$

for $\ell = 1, 2, \ldots, n_g$. $T_{n,i}^{CvM}(\chi_\ell^0)$ is defined similarly for $\ell = 1, 2, \ldots, n_g, n_g + 1$.

For calculating the p-values, the modified version of the Cramér–von Mises statistic is used, which is given by

$$T_{n,i}^{CvM^*}(\chi_\ell) = \left(T_{n,i}^{CvM} - \frac{0.4}{\ell g} + \frac{0.6}{(\ell g)^2}\right)\left(1 + \frac{1}{\ell g}\right). \tag{37}$$

For the calculation of the p-values of this test, the program of Xiao, Gordon & Yakovlev (2007) was investigated. Due to the iterative nature of the proposed procedure and large sample sizes, too much computation time was used to calculate the exact p-values. Discrete p-values are calculated instead for the upper and lower tail according to Stephens (1970), with linear interpolation between discrete p-value levels.

3.3.3 *Anderson–Darling test for uniformity*

The Anderson–Darling test statistic is given, for $i = 1, 2, \ldots, m$, by (Stephens 1970)

$$T_{n,i}^{AD}(\chi_\ell) := -\ell g - \frac{1}{\ell g}\sum_{j=1}^{\ell g}(2j-1)\ln\left(\frac{\theta_{(k_i+j)} - \theta_{(k_i)}}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}\left(1 - \frac{\theta_{(k_i+\ell g-j+1)} - \theta_{(k_i)}}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}\right)\right), \tag{38}$$

for $\ell = 1, 2, \ldots, n_g$. $T_{n,i}^{AD}(\chi_\ell^0)$ is defined similarly for $\ell = 1, 2, \ldots, n_g, n_g + 1$. For the p-values, the two-term recursion method of Marsaglia & Marsaglia (2004) is used to obtain accuracy up to the fourth digit.

3.3.4 *Rayleigh test for uniformity*

The Rayleigh goodness-of-fit test is a test for uniformity on the circle proposed in 1894 by Lord Rayleigh (also known by the name of J. Strutt). The test is based on the sample mean resultant length $\bar{R}$ defined in equation (3). The relevant test statistic is $n\bar{R}^2$ (Mardia & Jupp 2000) and the p-values are calculated with the approximation proposed by Greenwood & Durand (1955). As mentioned in the first paragraph of Section 3.1, SOPIE is based on the sequential application of goodness-of-fit tests on *subintervals* of [0, 1]. By contrast, the Rayleigh goodness-of-fit test was developed to test uniformity of sample observations *on the entire circle*. It is, therefore, essential to *scale* the observations when SOPIE is applied to each subinterval, before calculating the resultant length $\bar{R}$.

Define the sample observations in the subinterval $[\theta_{(k_i)}, \theta_{(k_i+\ell g+1)}]$, for $i = 1, 2, \ldots, m$, by

$$\chi_{\ell,i} := \{\theta_{(k_i)}, \theta_{(k_i+1)}, \ldots, \theta_{(k_i+\ell g)}, \theta_{(k_i+\ell g+1)}\}, \tag{39}$$

for $\ell = 1, 2, \ldots, n_g$. $\chi_{\ell,i}^0$ is defined similarly for $\ell = 1, 2, \ldots, n_g, n_g + 1$. The *scaled* observations used in the calculation of $\bar{R}$, are given by

$$\chi_{\ell,i}^* := \left\{\frac{2\pi(\theta_{(k_i)} - \theta_{(k_i)})}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}, \frac{2\pi(\theta_{(k_i+1)} - \theta_{(k_i)})}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}, \ldots, \frac{2\pi(\theta_{(k_i+\ell g)} - \theta_{(k_i)})}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}, \frac{2\pi(\theta_{(k_i+\ell g+1)} - \theta_{(k_i)})}{\theta_{(k_i+\ell g+1)} - \theta_{(k_i)}}\right\}. \tag{40}$$

$\chi_{\ell,i}^{0*}$ is defined similarly for $\ell = 1, 2, \ldots, n_g, n_g + 1$.

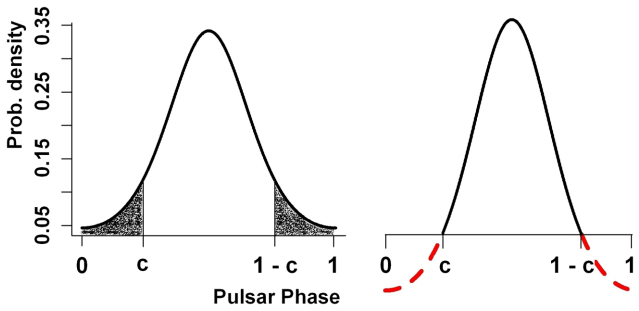

Figure 3. Scaling of the von Mises density function with $\mu = 0.5$ and $\kappa = 1$.

Remark

Note that, in the notation of the Proposition in Section 3.1 which provides the theoretical justification for SOPIE, $\theta_{(k_i)}$ and $\theta_{(k_i+\ell g+1)}$ play the role of x_r and x_{s+1}, respectively. Furthermore, since all four of the discussed goodness-of-fit tests are applied in SOPIE, the recommended estimated off-pulse interval is the median value obtained from the four estimates (as discussed in the algorithm). This recommended estimated off-pulse interval is depicted on the graphs produced as output from SOPIE.

4 APPLICATION OF SOPIE TO SIMULATED AND REAL DATA SETS

4.1 Application to simulated data sets

The von Mises distribution is a class of probability density functions that can be considered when using Monte Carlo simulation to evaluate the performance of SOPIE. Other classes of probability density functions can also be used, such as the Triangular distribution. Since SOPIE is developed to estimate the off-pulse interval, it is essential that random variates are obtained from a distribution where $f(\theta)$ in equation (1) is uniform for θ belonging to some finite subinterval of [0, 1). The process to create such a distribution is to remove some weight in the tails of the distribution in the intervals $[0, c]$ and $[1 - c, 1]$ for some value c. Furthermore, the distribution must be standardized to ensure that it is a density function. Fig. 3 is a depiction of this transformation for the von Mises density (for $\mu = 0.5$). The tinted part of the density on the left graph must be removed in order to end up with a density similar to the graph on the right. The dashed lines reflect the part of the original von Mises distribution that is no longer used. Mathematically, the density must be scaled from a support of [0, 1] to support on $[c, 1 - c]$. To generate random variates from such a scaled distribution, the accept–reject method is used (Robert & Casella 2010). The generated random variates from the rescaled distribution are thus a simulation of the source function $f_s(\theta)$ in equation (1). In order to satisfy the properties of $f(\theta)$, the random variates must be contaminated with uniform background proportional to $1 - p$. Thus, if n random variates are required, then one generates $\lfloor pn \rfloor$ values from the rescaled density and $n - \lfloor pn \rfloor$ values from a uniform density over the interval [0, 1), where $\lfloor pn \rfloor = \text{floor}(pn)$. Both these sets of random variates are then combined to produce a simulated set of data from $f(\theta)$. For a more detailed discussion of the von Mises density, the reader is referred to (Mardia & Jupp 2000, p. 36). The mathematical detail of the rescaling of densities can be found in Schutte (2014, p. 43, 47).

An extensive simulation study was conducted to evaluate the performance of SOPIE on different classes of distributions and on

Table 2. Monte Carlo estimates of bias for different goodness-of-fit tests when $\hat{h}_3$ is used with $\alpha = 0.05, m = 1$ and $r = 6$ for different values of g, based on a specific scaled von Mises distribution.

	Anderson–Darling		Cramér–von-Mises		Kolmogorov–Smirnov		Rayleigh	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
g=6	0.0213	−0.0223	0.0336	−0.0339	0.0242	−0.0299	0.0209	−0.0176
g=7	0.0224	−0.0198	0.0338	−0.0340	0.0242	−0.0259	0.0208	−0.0203
g=8	0.0233	−0.0196	0.0339	−0.0341	0.0239	−0.0240	0.0228	−0.0223
g=9	0.0205	−0.0202	0.0341	−0.0343	0.0237	−0.0233	0.0217	−0.0225
g=10	0.0198	−0.0183	0.0342	−0.0344	0.0241	−0.0221	0.0229	−0.0206
g=20	0.0207	−0.0223	0.0354	−0.0357	0.0227	−0.0237	0.0262	−0.0278
g=25	0.0216	−0.0230	0.0361	−0.0363	0.0234	−0.0245	0.0284	−0.0290

Table 3. Monte Carlo estimates of MSE for different goodness-of-fit tests when $\hat{h}_3$ is used with $\alpha = 0.05, m = 1$ and $r = 6$ for different values of g, based on a specific scaled von Mises distribution.

	Anderson–Darling		Cramér–von-Mises		Kolmogorov–Smirnov		Rayleigh	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
g=6	0.0125	0.0217	0.0012	0.0012	0.0140	0.0220	0.0099	0.0102
g=7	0.0117	0.0179	0.0012	0.0012	0.0129	0.0178	0.0079	0.0082
g=8	0.0102	0.0157	0.0012	0.0012	0.0109	0.0154	0.0076	0.0071
g=9	0.0071	0.0145	0.0012	0.0012	0.0088	0.0131	0.0055	0.0057
g=10	0.0064	0.0111	0.0012	0.0012	0.0077	0.0119	0.0045	0.0044
g=20	0.0029	0.0051	0.0013	0.0013	0.0022	0.0049	0.0014	0.0024
g=25	0.0024	0.0038	0.0013	0.0013	0.0017	0.0031	0.0017	0.0019

Table 4. Monte Carlo estimates of bias for different goodness-of-fit tests when $\hat{h}_1$ is used with $\alpha = 0.01, m = 1$ and $r = 6$ for different values of g, based on a specific scaled Triangular distribution.

	Anderson–Darling		Cramér–von-Mises		Kolmogorov–Smirnov		Rayleigh	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
g=6	-0.0072	0.0070	−0.0387	0.0384	−0.0140	0.0117	−0.0214	0.0199
g=7	−0.0095	0.0090	−0.0387	0.0385	−0.0164	0.0130	−0.0228	0.0200
g=8	−0.0097	0.0094	−0.0387	0.0385	−0.0168	0.0136	−0.0240	0.0231
g=9	−0.0106	0.0117	−0.0387	0.0385	−0.0184	0.0141	−0.0250	0.0239
g=10	−0.0126	0.0130	−0.0388	0.0385	−0.0184	0.0152	−0.0253	0.0244
g=20	−0.0186	0.0166	−0.0390	0.0388	−0.0217	0.0212	−0.0307	0.0285
g=25	−0.0202	0.0172	−0.0392	0.0389	−0.0235	0.0213	−0.0317	0.0289

a wide array of data set parameters. The Monte Carlo simulation consisted of 1000 independent trials for each of the selected classes and parameter sets. The complete simulation study was done for two classes of densities, namely, the von Mises density and the Triangular density. The different parameters considered in the simulation study were the background level ($0.1 \leq 1 - p \leq 0.4$), the concentration parameter ($1 \leq \kappa \leq 4$), the number of observations in the simulated data set ($500 \leq n \leq 25000$) and the width of the on-pulse interval ($0.13 \leq a \leq 0.45$, $0.55 \leq b \leq 0.8$). The standard errors of all the averages of the Monte Carlo estimates of the bias and MSE were found to be negligibly small and are therefore omitted from the tables. In order to establish the accuracy and consistency of the estimated off-pulse interval, two measures are used, namely the bias and the MSE. The bias is used as measure of how well SOPIE estimates the theoretical (true) off-pulse interval. Since the theoretical off-pulse interval is known for each of the simulated data sets, the bias serves as a measure of the accuracy of the estimation technique. The closer the bias is to zero, the smaller the mean difference between the estimator and the theoretical value. The MSE criterion takes both the bias and the variance of an estimator into account. Suppose, for example, an unknown parameter ψ is estimated by $\hat{\psi}$ (some function of the data), then the MSE is defined by

$$\text{MSE} = E\left((\hat{\psi} - \psi)^2\right) \tag{41}$$

$$= (\text{Variance of } \hat{\psi}) + (\text{Bias of } \hat{\psi})^2. \tag{42}$$

This is an important measure, since an estimator with good properties should ideally control both the bias and the variance. In this paper, the results of only two Monte Carlo configurations (consisting of 1000 independent trials) are reported. The first configuration is based on a scaled Von Mises distribution with background level $1 - p = 0.2$, concentration parameter $\kappa = 2$, with $n = 5\,000$ observations and on-pulse interval $[a, b] = [0.3, 0.7]$. Tables 2 and 3 present some of the simulation results obtained from this specific configuration, for different choices of the tuning parameters of SOPIE. The second configuration is based on the scaled Triangular distribution with background level $1 - p = 0.4$, $n = 25\,000$ observations and off-pulse interval $[a, b] = [0.3, 0.7]$. Tables 4 and 5 present some of the simulation results obtained from this specific configuration, for different choices of the tuning parameters of SOPIE. Several other

Table 5. Monte Carlo estimates of MSE for different goodness-of-fit tests when $\hat{h}_1$ is used with $\alpha = 0.01, m = 1$ and $r = 6$ for different values of g, based on a specific scaled Triangular distribution.

	Anderson–Darling		Cramér–von-Mises		Kolmogorov–Smirnov		Rayleigh	
	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$	$\hat{a}$	$\hat{b}$
g=6	0.0039	0.0037	0.0015	0.0015	0.0036	0.0038	0.0039	0.0039
g=7	0.0035	0.0033	0.0015	0.0015	0.0030	0.0035	0.0036	0.0040
g=8	0.0034	0.0033	0.0015	0.0015	0.0030	0.0034	0.0035	0.0034
g=9	0.0033	0.0029	0.0015	0.0015	0.0027	0.0033	0.0033	0.0032
g=10	0.0029	0.0026	0.0015	0.0015	0.0028	0.0032	0.0032	0.0031
g=20	0.0018	0.0022	0.0016	0.0015	0.0022	0.0021	0.0023	0.0025
g=25	0.0016	0.0020	0.0016	0.0016	0.0020	0.0022	0.0021	0.0025

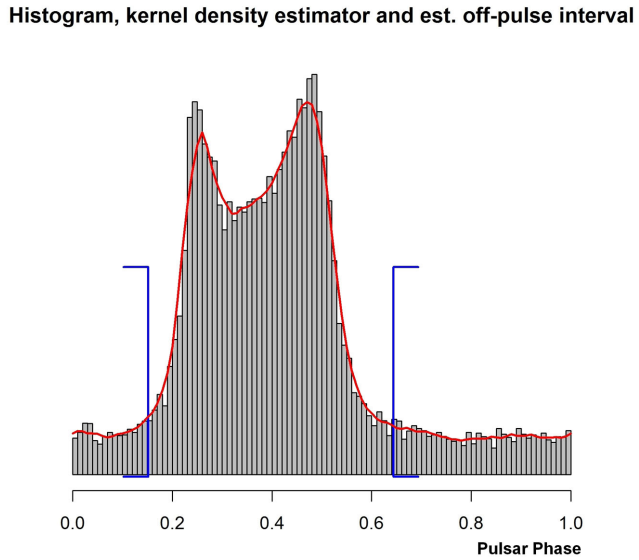

Figure 4. Histogram estimator with 100 classes (grey bars) of the γ-ray light curve of PSR J1709-4429 (energy >0.1 GeV) with circular kernel density estimator (red line) fitted to the data.The blue bracket represents the direction and boundaries of the estimated off-pulse interval obtained with SOPIE.

comparisons were made to inspect the influence of different choices of the arguments on SOPIE. For a wide array of different values of the arguments of SOPIE, accurate and consistent estimation followed (Schutte 2014).

4.2 Application to real data sets and comparison to Bayesian block method

The discussion in the previous section dealt with some results from a simulation study that was performed to establish the accuracy and consistency of the algorithm to estimate the off-pulse interval of an unknown source function. However, this research originated from an *astrophysical context*. Therefore, the proposed technique should be applied to pulsar data in order to assess the performance thereof. This section will provide these results and compare the end-point values of the off-pulse interval (estimated with SOPIE) to the values obtained with the 'eye-ball' technique and Bayesian Block method frequently used in the literature (as discussed in Section 1).

4.2.1 Example 1: PSR J1709-4429

Fig. 4 presents the graphical output generated with the SOPIE package. This figure is a histogram representation (with 100 classes) of a single cycle of the estimated γ-ray light curve of the PSR J1709-4429 pulsar (energy >0.1 GeV, $n = 21\,153$). The circular kernel density estimator is also fitted to the data and overlaid on the graph. Table 6 provides the estimated off-pulse interval.

Table 6. Off-pulse estimation for PSR J1709-4429 produced as output from SOPIE. Abbreviations used: CvM = Cramér–von Mises, KS= Kolmogorov–Smirnov AD = Anderson–Darling, Ray = Rayleigh.

	CvM	KS	AD	Ray.	MEDIAN
$\hat{a}$	0.615	0.652	0.692	0.635	0.644
$\hat{b}$	0.163	0.141	0.133	0.164	0.152

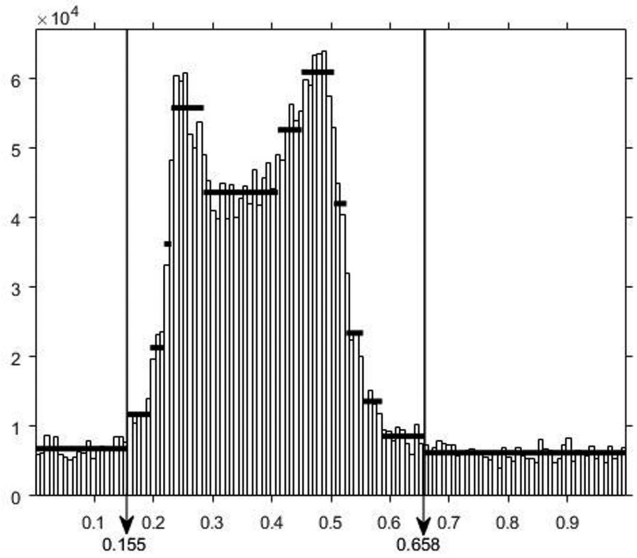

Figure 5. Bayesian Block analysis performed on PSR J1709-4429.

Abdo et al. (2010a) reports an off-pulse interval for this pulsar as [0.66, 0.14] by using the subjective 'eye-ball' technique. The estimated off-pulse interval (using SOPIE) is $[\hat{a}, \hat{b}] = [0.644, 0.152]$ using parameter values of $\hat{h} = 1$, $\alpha = 0.05$, $g = 22$, and $r = 10$. When comparing the Bayesian Block method to our algorithm, it should be emphasized that only two specific 'change points' identified by the Bayesian Block method, are comparable to the estimated off-pulse interval. These 'change points'are the starting and ending x-axis values of the block with the lowest density, or differently stated, the block with the lowest y-axis value. When applying the Bayesian Block method to the same data, the start– and end points of the interval of lowest density for the PSR J1709-4429 pulsar, is equal to 0.658 and 0.155, respectively (see Fig. 5). When

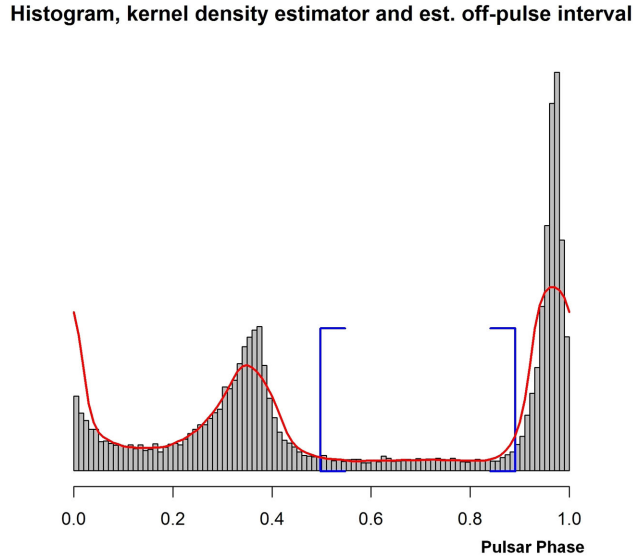

Figure 6. Histogram estimator with 100 classes (grey bars) of the γ-ray light curve of PSR J0534+2200 (Crab pulsar, energy >0.1 GeV) with circular kernel density estimator (red line) fitted to the data. The blue bracket represents the direction and boundaries of the estimated off-pulse interval obtained with SOPIE.

Table 7. Off-pulse estimation for PSR J0534+2200 produced as output from SOPIE. Abbreviations used: CvM = Cramér–von Mises, KS= Kolmogorov–Smirnov AD = Anderson–Darling, Ray = Rayleigh.

	CvM	KS	AD	Ray.	MEDIAN
$\hat{a}$	0.456	0.501	0.494	0.501	0.498
$\hat{b}$	0.895	0.886	0.881	0.897	0.891

comparing this interval to the SOPIE estimated interval of [0.644, 0.152], it is evident that these two intervals are very similar to each other, despite the fact that fundamentally different approaches are used.

4.2.2 *Example 2: PSR J0534+2200*

The second example of real life pulsar data used in the demonstration of SOPIE, is the well-known Crab pulsar. The data contains $n = 21\,145$ times of arrival (energy >0.1 GeV). Abdo et al. (2010b) reports an off-pulse interval of $[\hat{a}, \hat{b}] = [0.52, 0.87]$ for this pulsar, obtained with the 'eye-ball' technique. The estimated off-pulse interval (using SOPIE) is $[\hat{a}, \hat{b}] = [0.50, 0.89]$ using parameter values of $\hat{h} = 1$, $\alpha = 0.05$, $g = 22$, and $r = 10$. Fig. 6 is the graphical representation of the output obtained from SOPIE, and Table 7 presents the estimated off-pulse interval. When applying the Bayesian Block method to PSR J0534+2200, the start- and end points of the interval of lowest density is equal to 0.52 and 0.88, respectively (see Fig. 7). Again note the comparability of the intervals.

5 CONCLUSIONS AND FUTURE WORK

Our main goal was to derive a powerful and accurate non-parametric sequential estimation technique to estimate the off-pulse interval of a pulsar light curve in an objective manner. The second aim involved the development of an user friendly statistical package SOPIE that implements the new estimation technique in such a way that it is straightforward to apply the procedure to data consisting

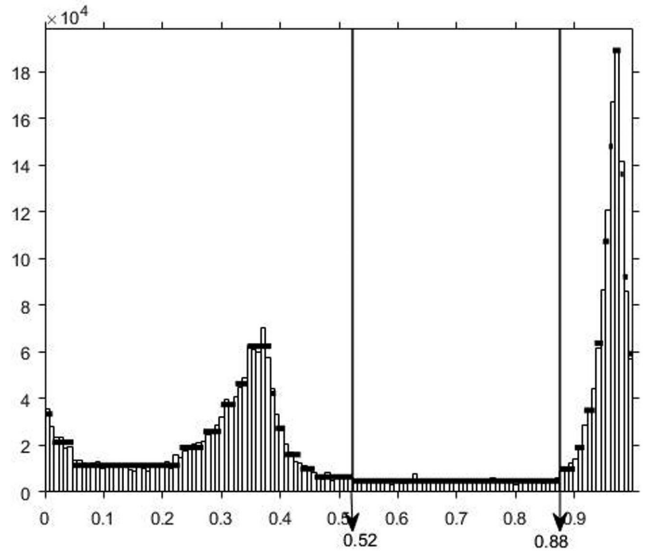

Figure 7. Bayesian Block analysis performed on PSR J0534+2200.

of photon arrival times. The major advantage of this new package is that it provides a procedure to objectively estimate the off-pulse interval of a pulsar light curve in the form of an easy-to-use software package. Our procedure is complementary to the subjective 'eye-ball' technique and Bayesian Block approach used in the literature.

In this paper, we provided background pertaining to circular statistics and circular density estimation to facilitate the understanding of our non-parametric sequential estimation technique. Furthermore, the performance of the technique was evaluated by means of a simulation study and the application to real-life pulsar data. Several aspects relating to the performance of the procedure are discussed in Schutte (2014), where the good properties of the newly developed procedure are illustrated on several more simulated and real-life pulsar data sets. Also, other aspects such as the choice of optimal parameter configurations are examined.

The SOPIE package consists of four main functions. The details of each of the functions are given in Appendix A. The `findh` function provides an estimated smoothing parameter to the `circ.kernel` function, which performs circular kernel density estimation on the data. The minima of the circular kernel density estimator are then utilized in the `a.estimate` and `b.estimate` functions, which provides the estimators for the off-pulse interval of the data. The SOPIE wrapper function allows the user to execute all of the mentioned functions using a single program call. The output is structured in a table containing the different estimators, together with a graph consisting of the histogram estimator, circular kernel density estimator and the recommended estimated off-pulse interval obtained from the different goodness-of-fit test statistics.

We plan to constantly update the SOPIE package to improve the estimator and to include some more goodness-of-fit tests. Some future directions includes the expansion of the number of circular kernel functions used in the kernel density estimation, and the investigation into the addition of a parameter that corresponds to a detection threshold. Another future direction is the identification of a second off-pulse region or interval. The current method can be adapted to accommodate a situation where more than one off-pulse interval exists, but some changes will have to be incorporated into the package. The SOPIE package is available on http://CRAN.R-project.org/package=SOPIE (Schutte 2015), which also contains a detailed guide on the functionality of the package.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation towards Christo Venter for the constructive remarks, careful reading and feedback given on several versions of the manuscript. The authors are grateful to two anonymous referees for critical reviews that have definitely improved the quality of the paper.

The second author thanks the National Research Foundation of South Africa for financial support.

REFERENCES

Abdo A. A. et al., 2008, Science, 322, 1218
Abdo A. et al., 2009, ApJ, 696, 1084
Abdo A. et al., 2010a, ApJS, 187, 460
Abdo A. et al., 2010b, ApJ, 708, 1254
Abdo A. et al., 2010c, AJ, 711, 64
Abdo A. et al., 2010d, ApJ, 713, 146
Abdo A. et al., 2010e, AJ, 714, 927
Abdo A. A. et al., 2012, ApJ, 744, 146
Abdo A. A. et al., 2013, ApJS, 208, 17
Abuzaid A., Mohamed I., Hussin A., 2011, Comput. Stat., 27, 381
Ackermann M. et al., 2011, AJ, 726, 35
Agostinelli C., 2007, Comput. Stat. Data Anal., 51, 5867
Atwood W. B., Ziegler M., Johnson R. P., Baughman B. M., 2006, ApJ, 652, L49
Atwood W. et al., 2009, ApJ, 697, 1071
Bai Z., Rao C. R., Zhao L., 1988, J. Multivariate Anal., 27, 24
Cramér H., 1928, Skandinavisk Aktuarietidskrift, 11, 141
D'Agostino R., Stephens M., eds, 1986, Goodness-of-Fit Techniques. Marcel Dekker, Inc, New York
De Jager O. C., Büsching I., 2010, A&A, 517, L9
De Jager O. C., Raubenheimer B. C., Swanepoel J. W. H., 1989, A&A, 221, 180
Fisher N., 1989, J. Struc. Geol., 11, 775
Fisher N., 1993, Statistical Analysis of Circular Data. Cambridge Univ. Press, Cambridge
Garcia-Portugues E., Crujeiras R. M., Gonzalez-Manteiga W., 2013, J. Multivariate Anal., 121, 152
Greenwood J., Durand D., 1955, Ann. Math. Stat., 26, 233
Grondin M., Romani R., Lemoine-Goumard M., Harding L. G. G., Reposeur T., 2013, ApJ, 774, 110
Hall P., Watson G., Cabrera J., 1987, Biometrika, 74, 751
Hall P., Sheather S., Jones M., Marron J., 1991, Biometrika, 78, 263
Johnson T. J. et al., 2013, ApJ, 778, 106
Jones M., Marron J., Sheather S., 1996, J. Am. Stat. Assoc., 91, 401
Kerr M., 2011, ApJ, 732, 38
Klemelä J., 2000, J. Multivariate Anal., 73, 18
Kopp A., Venter C., Büsching I., de Jager O. C., 2013, ApJ, 779, 126
Leung G., Takata J., Ng C., Kong A., Tam P., Hui C., Cheng K., 2014, ApJ, L13
Loots H., 1995, PhD thesis, Potchefstroom Univ. of Christian Higher Education
Lorimer D., Kramer M., 2005, Handbook of Pulsar Astronomy. Cambridge Univ. Press, Cambridge
Mardia K., 1972, Statistics of Directional Data. Academic Press, London
Mardia K., 1992, The Art of Statistical Science: A Tribute to G.S. Watson. John Wiley & Sons, New York
Mardia K., Jupp P., 2000, Directional Statistics. John Wiley & Sons, New York
Marsaglia G., Marsaglia J., 2004, J. Stat. Softw., 9, 1
Marsaglia G., Tsang W. W., Wang J., 2003, J. Stat. Softw., 8, 1
Ng C. et al., 2014, MNRAS, 439, 1865
Oliveira M., Crujeiras R., Rodriquez-Casal A., 2012, Comput. Stat. Data Anal., 56, 3898
Parzen E., 1962, Ann. Math. Stat., 33, 1065
Pletsch H. J. et al., 2012a, Science, 338, 1314
Pletsch H. J. et al., 2012b, ApJ, 744, 105
R Core Team 2015, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, available at https://www.R-project.org
Ray P. et al., 2011, ApJS, 194, 17
Robert C. P., Casella G., 2010, Introducing Monte Carlo Methods with R. Springer-Verlag, New York
Rosado P. A., Sesana A., Gair J., 2015, MNRAS, 451, 2417
Rosenblatt M., 1956, Ann. Math. Stat., 27, 832
Saz Parkinson P. M. et al., 2010, ApJ, 725, 571
Scargle J., 1998, ApJ, 504, 405
Scargle J., Norris J., Jackson B., Chiang J., 2013, ApJ, 764, 167
Schutte W. D., 2014, PhD thesis, North-West Univ., available at http://hdl.handle.net/10394/12199
Schutte W. D., 2015, SOPIE: Non-Parametric Estimation of the Off-Pulse Interval of a Pulsar. Available at http://CRAN.R-project.org/package=SOPIE
Scott D., 1992, Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons, New York
Sheather S., Jones M., 1991, J. R. Stat. Soc. Ser. B Stat. Methodol., 53, 683
Silverman B., 1986, Density Estimation for Statistics and Data Analysis. Chapman & Hall, New York
Stephens M., 1970, J. R. Stat. Soc. Ser. B Stat. Methodol., 32, 115
Taylor C., 2008, Comput. Stat. Data Anal., 52, 3493
Venter C., Harding A. K., Guillemot L., 2009, ApJ, 707, 800
Venter C., Kopp A., Harding A. K., Gonthier P. L., Büsching I., 2015, ApJ, 807, 130
Von Mises R., 1931, Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und Theoretischen Physik. Deuticke, Leipzig
Wand M., Jones M., 1995, Kernel Smoothing. Chapman & Hall, New York
Xiao Y., Gordon A., Yakovlev A., 2007, J. Stat. Softw., 17, 1
Zajczyk A., Bednarek W., Rudak B., 2013, MNRAS, 432, 3462

APPENDIX A: AVAILABLE FUNCTIONS IN THE SOPIE PACKAGE

The SOPIE package consists of four main functions. Each of these functions will be discussed in terms of its functioning, structure, arguments and output.

(i) `findh` is used to obtain the estimated smoothing parameter $\hat{h}$ that will be used in the circular kernel density estimator, defined in equation (8).

This function is structured as follows.

```
findh(data, h = 1, to = 1)
```

The arguments within this function are:

(a) `data` – A vector or data frame containing the data within which to find the estimated smoothing parameter $\hat{h}$ that will be used in the circular kernel density estimator.

(b) `h` – A scalar value from 1 to 9, specifying which smoothing parameter to calculate according to Table 1.

(c) `to` – A scalar value specifying the maximum domain of `data`. Values will usually either be 1 or 2π.

The output produced by the function is a single real value representing the rounded value (to 2 decimal places) of the estimated smoothing parameter.

(ii) `circ.kernel` is used to perform circular kernel density estimation on the sample data set in order to obtain the minimum points of the kernel density estimator. This is essentially step (i) of the suggested procedure in Section 3.1. The output

can also be used to draw a graph of the circular kernel density estimator.

This function is structured as follows.

```
circ.kernel(data, sp, to = 1, grid = 512, m
= 1)
```

The arguments within this function are the following.

(a) `data` – A vector or data frame containing the data on which the kernel density estimator will be calculated.
(b) `sp` – A real value ($0 < sp < 1$) for the smoothing parameter. This value can be obtained with the `findh` function.
(c) `to` – A scalar value specifying the maximum domain of `data`. Values will usually either be 1 or 2π.
(d) `grid` – A scalar value specifying the number of equally spaced grid points to be used in estimating the kernel density.
(e) `m` – A scalar value specifying the number of local minimum points included in the output. This value correspond to the value of m in step (i) of the algorithm given in Section 3.1.

The output produced by the function is a list with the following items.

(a) `x` is a vector that represents the equally spaced grid points used during the kernel density estimation.
(b) `y` is a vector that represents the density-values of the kernel density estimator in each of the grid points (`x`).
(c) `minimum` is a vector that represents the kernel grid point(s) of lowest density derived from the circular kernel density estimator. The number of points included, will depend on the choice of m above.

(iii) `a.estimate` and `b.estimate` is almost identical functions. `a.estimate` is the function used to obtain the estimated values of a, i.e. $\hat{a}$. This is essentially the implementation of steps (iii) to (viii) of the algorithm in Section 3.1. `b.estimate` is the function used to obtain the estimated values of b, i.e. $\hat{b}$.

These functions are structured as follows.

```
a.estimate(data, to = 1, min_points, alpha
= 0.05, g = 1, r = 1)
b.estimate(data, to = 1, min_points, alpha
= 0.05, g = 1, r = 1)
```

The same arguments are used in both functions.

(a) `data` – A vector or data frame containing the data which are used to estimate a (or b).
(b) `to` – A scalar value specifying the maximum domain of `data`. Values will usually either be 1 or 2π.
(c) `min.points` – A scalar or vector containing the value(s) of the minimum points calculated during the kernel density estimation. This argument does not represent the index value(s) of the observation(s) within `data`. This is the value of $\theta_{(k_i)}$ in step (ii) of the algorithm in Section 3.1. The value(s) can be obtained with the function `circ.kernel`.
(d) `alpha` – A real number (<1) specifying the level of significance (α) that will be used during the sequential application of the goodness-of-fit tests for uniformity in estimating the off-pulse interval.
(e) `g` – A scalar value specifying the value of the incremental growth of each subsequent interval over which uniformity is tested. In the suggested procedure, uniformity is sequentially tested, with the interval used in the test growing by `g` observations after every iteration. The selection of `g` not only influences the computation time of the procedure, but also has an effect on the point where rejection of the hypothesis takes place. For large values of `g`, the user takes the risk that uniformity is rejected for a certain (larger) interval, while it should have been rejected earlier (for a smaller interval). On the other hand, a very small choice of `g` results in long execution times. Small values of `g` may also result in the early rejection of uniformity, e.g. in the situation where a few observations may cause the rejection of uniformity, while uniformity is again confirmed when several more observations are included in the interval. If the user suspects that this situation may occur, the problem can be avoided by selecting a larger value of the integer `r`.
(f) `r` – A scalar value that represents the number of subsequent intervals that must result in the rejection of uniformity before the function will stop. The choice of `r` must therefore be linked to the choice of `g` as explained above (see steps (iv) and (iv) of the algorithm). For smaller values of `g`, it would be safer to select larger values of `r`, and vice versa. Since small values of `g` may result in a temporary rejection of uniformity for an interval, a larger value of `r` would prevent the method from immediately stopping at the first occurrence of rejection. It is very important to note that, for a large value of `r`, there will be no influence on the estimation of a (or b) if rejection takes place for each interval after a certain point.

The output produced by these functions is a list with the following items.

(a) `summary` is a vector that contains the estimated value of a (or b) for each of the four goodness-of-fit tests, namely the Anderson–Darling, Kolmogorov–Smirnov, Cramér–von-Mises and the Rayleigh goodness-of-fit test.
(b) `general` is a list containing the function call, the minimum value(s) used in the estimation, the level of significance (α), the value of `g` and the value of `r`.

(iv) `SOPIE` is a wrapper function in the sense that it utilises all of the above functions to produce the estimated off-pulse intervals in an easy readable matrix format, together with a graph consisting of the histogram estimator of the sample data, the kernel density estimator, and a visual representation of the estimated off-pulse intervals.

This function is structured as follows.

```
SOPIE(data = NULL, h = 1, to = 1, alpha =
0.05, g = 20, r = 10, m = 1, grid = 512)
```

The arguments within this function is similar to the arguments already described in the functions above.
The output produced by the function is as follows.

(a) `summary` is a matrix that contains the estimated values of a and b for each of the four goodness-of-fit tests, namely the Anderson–Darling, Kolmogorov–Smirnov, Cramér–von-Mises and the Rayleigh goodness-of-fit test. Based on the four estimated values of a and b, the median values of a and b are also calculated. This median off-pulse interval is the recommended interval and also the interval that is depicted on the graph.
(b) `general` is a list containing the function call, the minimum value(s) used in the estimation, the level of significance (α), the value of `g` and the value of `r`.
(c) A histogram estimate of the data is produced with the circular kernel density estimate overlaid, together with an indication of the estimated median off-pulse interval derived from the four goodness-of-fit tests (illustrated with two solid vertical lines).

APPENDIX B: LIST OF ESTIMATED MEASURES OF DISPERSION USED IN THE DEFINITION OF THE ESTIMATED SMOOTHING PARAMETERS

In Table 1, nine different estimated smoothing parameters are defined based on the following list of estimated measures of dispersion of the data:

(i) Linear standard deviation, denoted by s.
(ii) Square root of the circular variance given by

$$s_\circ = \sqrt{1 - \bar{R}}. \tag{B1}$$

(iii) Circular mean deviation, defined by

$$\bar{D}_\circ = \frac{1}{n}\sum_{i=1}^{n}\{\pi - |\pi - |\theta_i - \bar{\theta}||\}. \tag{B2}$$

(iv) Circular median absolute deviation, given by

$$|D_\circ| = \text{median}\left(|\theta_1 - \tilde{\theta}|, \ldots, |\theta_n - \tilde{\theta}|\right). \tag{B3}$$

(v) Circular interquartile range, defined by

$$\text{IQR}_\circ = 2\pi - (Q_3 - Q_1), \tag{B4}$$

where Q_1 and Q_3 are obtained by classifying the sample observations into two groups based on their location with respect to the sample median direction $\tilde{\theta}$. Q_1 can therefore be considered as the median of the first group and Q_3 as the median of the second group. If the value of Q_1 is larger than the value of Q_3, then the labels are simply interchanged (Abuzaid, Mohamed & Hussin 2011).

This paper has been typeset from a TEX/LATEX file prepared by the author.