Bayesian Analysis of Circular Data Using Wrapped Distributions

Palanikumar Ravindran†

Roche Palo Alto, Palo Alto, USA.

Sujit K. Ghosh

North Carolina State University, Raleigh, USA.

Institute of Statistics Mimeo Series# 2564

Summary. Circular data arise in a number of different areas such as geological, meteorological, biological and industrial sciences. Standard statistical techniques can not be used to model circular data due to the circular geometry of the sample space. One of the common methods to analyze circular data is known as the *wrapping approach*. This approach is based on a simple fact that a probability distribution on a circle can be obtained by wrapping a probability distribution defined on the real line. A large class of probability distributions that are flexible to account for different features of circular data can be obtained by the aforementioned approach. However, the likelihood-based inference for wrapped distributions can be very complicated and computationally intensive. The EM algorithm to compute the MLE is feasible, but is computationally unsatisfactory. A data augmentation method using slice sampling is proposed to overcome such computational difficulties. The proposed method turns out to be flexible and computationally efficient to fit a wide class of wrapped distributions. In addition, a new model selection criteria for circular data is developed. Results from an extensive simulation study are presented to validate the performance of the proposed estimation method and the model selection criteria. Application to a real data set is also presented and parameter estimates are compared to those that are available in the literature.

Keywords: Bayesian inference; MCMC; Wrapped Cauchy; Wrapped Double Exponential; Wrapped Normal

†*Address for correspondence:* Palani Ravindran, Roche Palo Alto LLC, 3431 Hillview Ave, Palo Alto, CA 94304, USA.

E-mail: Palanikumar.Ravindran@roche.com

1. Introduction

Circular data arise from a variety of sources in our daily lives, where we consider the circle to be the sample space. Some common examples are the migration paths of birds and animals, wind directions, ocean current directions and patients' arrival times in an emergency ward of a hospital. Many examples of circular data can be found in various scientific fields such as earth sciences, meteorology, biology, physics, psychology and medicine. To motivate the use of circular data, we present a brief description of couple of examples from these fields.

In the field of Physics, before the discovery of isotopes, **?** proposed testing the hypothesis that atomic weights are integers subject to error. He converted the fractional parts of the atomic weights to angles. He regarded these angles as a random sample from a circular distribution with mean zero and tested for uniformity. This study resulted into the famous Von Mises distribution on the circle.

A very good source of circular data is in the field of biology. Migration path of birds and animals has been the subject of many studies. The objective of these studies is to ascertain whether the direction of migration is uniform. An example of the migration of turtles is given in Figure **??**. In this figure, we present the circular histogram plot of the data collected by Dr. E. Gould from John Hopkins University School of Hygiene and first cited by **?**. The data represents the directions taken by the sea turtles after laying their eggs. The predominant direction is $64°$, which is the direction the turtles took to return to the sea.

Standard statistical techniques cannot be used to analyze circular data. This is due to the circular geometry of the sample space. For example, the sample mean of a data set on the circle is not the usual sample mean. Let $y_1, y_2, \ldots, y_n$ be independent observations on the unit circle, such that $0 \leq y_j < 2\pi$, $j = 1, 2, \ldots, n$. The mean direction $\bar{y}$ is not given by the usual definition, $\frac{1}{n}\sum_{j=1}^{n} y_j$. This is illustrated in Figure **??**, where the dashed arrow is the direction represented by $\frac{1}{n}\sum_{j=1}^{n} y_j$ and the solid arrow represents the mean obtained by vector addition. To find the circular mean, we use vector addition techniques.

In general, the p^{th} theoretical moment is defined as $E(e^{ipY}) = \alpha_p + i\beta_p$, $i = \sqrt{-1}$, for $p = 1, 2, \ldots$. The mean direction is given by $\mu = \arctan(\beta_1/\alpha_1)$ and the mean resultant is defined as $\rho = \sqrt{\alpha_1^2 + \beta_1^2}$, so that $E(e^{iY}) = \rho e^{i\mu}$. In most applications it is of interest to estimate the location parameter μ and the scale parameter ρ. The sample p^{th} moment

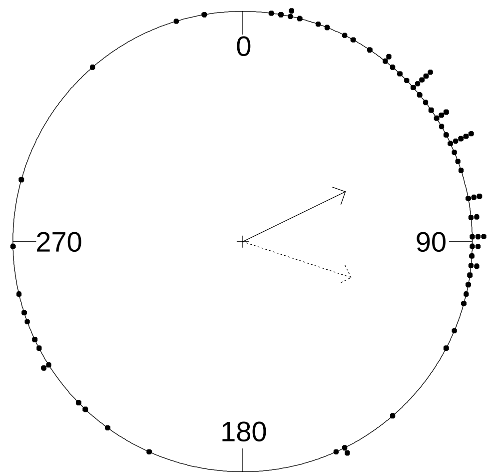

Fig. 1. Circular histogram plot of the turtle data. Solid line indicates circular mean and dashed line indicates the linear mean.

generally provides a consistent estimate of the theoretical p^{th} moment (See p.21 **?**). However such a nonparametric method cannot be easily generalized to regression problems using link function via μ. Typically parametric circular distribution are used to model the error part of the regression models (See **?**, ch.11). Also no finite sample standard errors of the sample moments are available. As a first step towards regression models, we investigate the performance of the estimation procedures for parametric models. Thus, in this article we focus only on estimating μ and ρ for a wide class of wrapped distributions.

In the literature several statistical approaches are available to model circular data. In Section **??**, we use the wrapping method to generate a flexible class of circular distributions. In Section 3, we discuss classical and Bayesian methods to obtain estimates of the parameters of wrapped circular distributions. As the Bayesian methods have the advantage of obtaining a finite sample estimate of the variability (for example, the posterior standard deviation), we use slice sampling as proposed in **?** within a data augmentation method for parameter estimation. We propose to use two model selection criteria in Section 4. In Section 5, we present results from extensive simulation studies to validate the frequentist performance of the methods developed in Sections 3 and 4. In Section 6, we apply our method to a real data set involving the movement of ants. Finally we conclude with some discussions on future work in Section 7.

2. Statistical Approaches to model circular data

Many methods and statistical techniques have been developed to analyze and understand circular data. The popular approaches have been *embedding*, *wrapping* and *intrinsic* approaches. A good overview of the embedding and intrinsic approaches can be found in **?** and **?**. We describe only the wrapping approach in this article.

2.1. The Wrapping Approach

In the wrapping approach, given a known distribution on the real line, we wrap it around the circumference of a circle with unit radius. Technically this means that if U is a random variable on the real line, then the corresponding random variable Y on the circle is given by $Y = U (\text{mod } 2\pi)$. Equivalently, the wrapped version of U is obtained by defining

$$Y = U - 2\pi \left[\frac{U}{2\pi}\right], \tag{1}$$

where $[u]$ = largest integer $\leq u$. Let the distribution function of U on the real line be denoted by F. Then the distribution function of Y denoted by F_w can be obtained as,

$$F_w(y) = \Pr(Y \leq y) = \sum_{k=-\infty}^{\infty} \left[F(y + 2\pi k) - F(2\pi k)\right].$$

This implies that if the density f of U exists, then the wrapped density f_w is given by,

$$f_w(y) = \sum_{k=-\infty}^{\infty} f(y + 2\pi k), \quad 0 \leq y < 2\pi. \tag{2}$$

An excellent overview of the properties of the wrapped distributions can be found in **?**. In most cases the above series can not be written in closed form except in few cases such as Cauchy distribution. The density of the wrapped Cauchy (WC) distribution is given by

$$f_{WC}(y) = \frac{1}{\pi\sigma} \sum_{k=-\infty}^{\infty} \left[1 + \left(\frac{y+2\pi k-\mu}{\sigma}\right)^2\right]^{-2}, \quad 0 \leq y < 2\pi, \tag{3}$$

where μ is the location parameter and σ is the scale parameter. Using inversion theorem, (**??**) can be simplified to

$$f_{WC}(y) = \frac{1}{2\pi} \frac{1-\rho^2}{1+\rho^2-2\rho\cos(y-\mu)},$$

where $\rho = e^{-\sigma}$. We consider two other wrapped distributions, for which no closed form expressions are available for (**??**). For instance, a wrapped normal (WN) density is given by,

$$f_{WN}(y) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} e^{-(y+2\pi k-\mu)^2/(2\sigma^2)}, \quad 0 \leq y < 2\pi. \tag{4}$$

Another popular wrapped distribution, known as the wrapped double exponential (WDE) is described by it density,

$$f_{WDE}(y) = \frac{1}{2\sigma} \sum_{k=-\infty}^{\infty} e^{-|y+2\pi k-\mu|/\sigma}, \quad 0 \leq y < 2\pi. \tag{5}$$

In Section **??**, we describe a Monte Carlo method based on slice sampling to obtain estimate of μ and ρ under (**??**), (**??**) and (**??**).

It follows that, a rich class of distributions on the circle can be obtained using the wrapping technique, as we can wrap any known distribution on the real line to the circle. The main difficulty in working with this approach has been that in most cases, the form of the densities and distribution functions are large sums (e.g. (**??**) and (**??**)) and can not be simplified as closed forms except for few cases (e.g. (**??**)). Due to this complexity, maximum likelihood techniques for point estimation and hypothesis testing can not be easily implemented. The main contribution of this paper is to present a general approach to obtain parameter estimates for a wide class of wrapped distributions. This can be achieved as follows. For a given probability distribution on the circle, we make assumptions that the original circular distribution was distributed on a line and was wrapped to a circle. Therefore if we can *unwrap* the distribution on the circle and obtain a distribution on the real line, we can use all the standard statistical techniques for data on a real line. We propose to perform this using the data augmentation approach. We use Bayesian methods, so that we can easily obtain parameter uncertainty estimates based on the finite sample.

3. Parameter Estimation of Wrapped Distributions

It was shown by **?** that the maximum likelihood estimate for wrapped Cauchy exists and is unique for samples of size greater than two. They also gave a simple iterative algorithm which would always converge to the maximum likelihood estimate (MLE). Calculating the MLE for the wrapped Cauchy is possible because the density has a closed form representation. In general, wrapped distributions do not have closed form densities and consequently, computing the MLE is complicated.

A different approach for fitting wrapped distributions was given by **?** who used the Expectation Maximization (EM) algorithm techniques to obtain parameter estimates from the Wrapped Normal distribution. However, the E-step involves ratio of large infinite sums, which needs to be approximated at each step. This makes the algorithm computationally inefficient. In addition, the standard errors of the MLEs have to be evaluated based on large-sample theory. We propose an alternative method that is more computationally efficient and flexible to entertain a large class of wrapped distributions. Also, as a by-product, we acquire finite sample interval estimates of the parameters of the wrapped distributions. This is done using the data augmentation approach described in the following section.

3.1. The Data Augmentation Approach

The data augmentation approach was originally proposed by **?**. Extensions of this technique can be found in the works of **?**, **?** and **?** and references therein. In the context of circular data, **?** used it to study Von Mises distribution. **?** used it to study the Wrapped Bivariate Normal distribution and wrapped autoregressive process. We present a generic approach based on slice sampling (see Neal, 2003) that can be used for a broad class of wrapped distributions.

The main idea behind the data augmentation approach is to augment the original data with some "additional data" that would simplify the original likelihood to a form that is much easier to handle. In case of circular data, as $Y = U(\text{mod } 2\pi)$, the random variable U on the real line, can be represented as $U = Y + 2\pi K$, where Y is the observed data on the circle, and K is the number of times U was wrapped to obtain Y. Therefore, in this case if we were able to "add" the information on K, and thus *unwrap* Y, then we could observe U. However, given Y, as the value of K is not unique, we obtain the conditional probability distribution of K given Y. To illustrate the unwrapping method, let us consider a location scale family $\frac{1}{\sigma}f(\frac{y-\mu}{\sigma})$ on the real line. Then the corresponding wrapped density is obtained from (**??**) as,

$$f_w(y|\mu,\sigma) \quad = \quad \sum_{k=-\infty}^{\infty} \frac{1}{\sigma}f(\tfrac{y+2\pi k-\mu}{\sigma}) \tag{6}$$

In the above equation, we follow the convention that the location parameter $\mu = \mu_0(mod\ 2\pi)$, where μ_0 is the location parameter on the real line. In order to specify the full probability model we consider several prior distributions for (μ, σ). In most cases the mean resultant direction, ρ (as defined in Section 1) can be expressed as a function of σ. In general, we will write $\rho = h(\sigma)$, e.g. for wrapped normal family $\rho = e^{-\sigma^2/2}$, for wrapped Cauchy family $\rho = e^{-\sigma}$ and for wrapped double exponential family $\rho = \frac{1}{1+\sigma^2}$. Notice that by definition, $0 < \rho < 1$. A class of non-informative prior for (μ, ρ) can be specified by a joint density of (μ, ρ) as,

$$[\mu, \rho] \quad \propto \quad I_\mu(0, 2\pi)\rho^{a_\rho - 1}(1-\rho)^{a_\rho - 1},\ a_\rho > 0, \tag{7}$$

where $I_x(A)$ denotes the indicator function, i.e., $I_x(A) = 1$ if $x \in A$ and $I_x(A) = 0$, otherwise. Viewing the wrapped number K to be a random variable, from (**??**) we see that the conditional density of Y given $K = k$ and the parameters μ, ρ is given by

$$f_w(y|k, \mu, \sigma) \quad = \quad \frac{\frac{1}{\sigma} f(\frac{y+2\pi k-\mu}{\sigma}) I_y(0, 2\pi)}{F(\frac{2\pi(k+1)-\mu}{\sigma}) - F(\frac{2\pi k-\mu}{\sigma})}$$

which is a truncated density of a location-scale family. It also follows that the marginal density of K given the parameters (μ, σ) can be obtained as,

$$\Pr(K = k|\mu, \sigma) \quad = \quad F(\tfrac{2\pi(k+1)-\mu}{\sigma}) - F(\tfrac{2\pi k-\mu}{\sigma}),\ k = \ldots, -1, 0, 1, \ldots$$

Thus, we obtain a Bayesian hierarchical model by specifying the distribution of y given k, μ, σ, then the conditional distribution of k given μ, σ and finally the prior distribution for (μ, σ).

Given a random sample on the circle, $\boldsymbol{y}= \{y_1, y_2, y_3, \ldots, y_n\}, 0 \leq y_j < 2\pi, j = 1 \ldots n$, we unwrap the data by obtaining samples from the conditional distribution of $\boldsymbol{k}$ given μ, σ and the observed data $\boldsymbol{y}$, where $\boldsymbol{k} = \{k_1, k_2, k_3, \ldots, k_n\}$. This is will be referred to as the data augmentation step. Then conditional on the augmented data $\boldsymbol{y}, \boldsymbol{k}$, we obtain samples from the joint posterior distribution of (μ, σ) to complete the Gibbs cycle of the MCMC method. From these samples, we obtain the marginal posterior distribution of μ and σ given the observed data $\boldsymbol{y}$, using the Ergodic Theorem of Markov Chain.

We provide a generic method to implement the MCMC method for a general class of location scale family, whose density function is invertible. A function $y = f(x)$ is invertible if f can be analytically or numerically inverted, or if f can be factorized into functions, which can be analytically or numerically inverted. That is, we assume that $f(x) = \prod_{t=1}^{T} f_t(x)$ and $\{f_t^{-1}(y) : j = 1 \ldots T\}$ are explicitly known functions or functions that can be computed using numerical methods (see **?** for some examples on this kind of splits). For example,

$f(x) = e^{-x}e^{-e^{-x}}$ does not have an explicitly known inverse, but e^{-x} and $e^{-e^{-x}}$ have explicitly known inverses. Alternatively, it is also possible to compute the inverse of $e^{-x}e^{-e^{-x}}$ using numerical methods such as bisection method. But for our proposed algorithm we restrict ourselves to the former method. It is assumed that $\rho = h(\sigma)$ is a monotone decreasing function of σ and can be inverted. This is a reasonable assumption because in general, as $\sigma \downarrow 0$, $\rho \uparrow 1$ and as $\sigma \uparrow \infty$, $\rho \downarrow 0$. These assumptions are satisfied for most wrapped distributions including Wrapped Normal (WN), Wrapped Cauchy (WC) and Wrapped Double Exponential (WDE) distributions. The general method will work even if $\rho = h(\sigma)$ is not a monotone decreasing function of σ. These wrapped distributions are all symmetric and unimodal. However, this method is not restricted only to these distributions.

In order to specify the required conditional distributions, we use the notation $[\theta_1, \dots, \theta_n]$ to represent the joint density of $\theta_1, \theta_2, \dots, \theta_n$ and $[\theta_1|\theta_2, \dots, \theta_n]$ to represent the conditional density of θ_1 given $\theta_2, \dots, \theta_n$. By the term *full conditional density of* θ_1, we mean the conditional density of θ_1, given rest of the parameters.

The general method is as follows. The joint prior density of (μ, σ) is given by

$$[\mu, \sigma] \quad \propto \quad I_\mu(0, 2\pi) h(\sigma)^{a_\rho - 1}(1 - h(\sigma))^{a_\rho - 1}|h'(\sigma)|, a_\rho > 0,\ \rho = h(\sigma)$$

For each observed direction y_j, we augment a random wrapping number k_j. The joint density of $\boldsymbol{y}, \boldsymbol{k}, \mu$ and σ is given by,

$$\begin{aligned} [\boldsymbol{y}, \boldsymbol{k}, \mu, \sigma] &\propto \left[\boldsymbol{y}|\boldsymbol{k}, \mu, \sigma^2\right]\left[\boldsymbol{k}|\mu, \sigma^2\right]\left[\mu, \sigma^2\right] \\ &\propto \tfrac{1}{\sigma^{n+n_0}} \prod_{j=1}^{n} f(\tfrac{y_j - \mu + 2\pi k_j}{\sigma}) h(\sigma)^{a_\rho - 1 + n_1}(1 - h(\sigma))^{a_\rho - 1} h_1(\sigma) I_\mu(0, 2\pi), \end{aligned}$$

where $|h'(\sigma)|$ can be factorized as $\frac{1}{\sigma^{n_0}} h(\sigma)^{n_1} h_1(\sigma)$. The full conditional densities of $\boldsymbol{k}, \mu$ and σ are nonstandard densities. So we introduce auxiliary variables $\boldsymbol{z}$ and $\boldsymbol{v}$. Let $\boldsymbol{z} = \{z_0, z_1, z_2, z_3\}$ and $\boldsymbol{v} = \{v_1, v_2, v_3, \dots, v_n\}$, such that

$$[\boldsymbol{y}, \boldsymbol{k}, \mu, \sigma] \quad \propto \quad \int [\boldsymbol{y}, \boldsymbol{k}, \mu, \boldsymbol{z}, \boldsymbol{v}, \sigma]\, d\boldsymbol{z} d\boldsymbol{v}$$

The joint density of $\boldsymbol{y}, \boldsymbol{k}, \mu, \boldsymbol{z}, \boldsymbol{v}$ and σ is given by,

$$\begin{aligned} [\boldsymbol{y}, \boldsymbol{k}, \mu, \boldsymbol{z}, \boldsymbol{v}, \sigma] \propto\ & I_{z_0}\left(0, \tfrac{1}{\sigma^{n+n_0}}\right) \prod_{j=1}^{n} I_{v_j}\left(0, f(\tfrac{y_j - \mu + 2\pi k_j}{\sigma})\right) I_{z_1}\left(0, h(\sigma)^{a_\rho - 1 + n_1}\right) \\ & I_{z_3}(0, h_1(\sigma)) I_\mu(0, 2\pi) \left\{ I_{z_2}\left(0, (1 - h(\sigma))^{a_\rho - 1}\right) I(a_\rho \neq 1) + I(a_\rho = 1) \right\} \end{aligned}$$

It follows from the above equation that all the full conditional densities are standard distributions that can be easily sampled using subroutines available in SAS, Splus, R and other statistical softwares. In particular, as an illustration, we provide the standard full conditionals that are needed to fit a WDE model in the Appendix. The standard full conditional densities that are needed to fit other distributions can be found in the thesis of the first author.

4. Model selection

We now have a large class of parametric models that can be fitted to a given circular data. In order to select the best fitting model, we use Deviance Information Criteria proposed by **?** and a predictive loss approach developed by **?**.

Deviance Information Criteria (DIC) is defined as

$$DIC = 2\mathrm{E}(\mathrm{Dev}(\mu,\sigma)|\boldsymbol{y}) - \mathrm{Dev}(\mathrm{E}(\mu,\sigma|\boldsymbol{y}))$$

Deviance (**?**) is defined as twice the negative of the log-likelihood. For example, for any wrapped location-scale density, $f_w()$ the deviance (Dev) is given by,

$$\begin{aligned}\mathrm{Dev}(\mu,\sigma) &= -2log(\prod_{i=1}^{n} f_w(y_i)) \\ &= -2\sum_{i=1}^{n} log(\sum_{k=-\infty}^{\infty} \tfrac{1}{\sigma} f(\tfrac{y_i+2\pi k-\mu}{\sigma})) \\ &\approx -2\sum_{i=1}^{n} log(\sum_{k=-L}^{L} \tfrac{1}{\sigma} f(\tfrac{y_i+2\pi k-\mu}{\sigma})),\end{aligned}$$

where L is a very large positive number. Given a set of models, the model with the lowest DIC is chosen as the "best" model. For justifications on the use of DIC as a model selection criteria see **?**.

Next, we use the decision theoretic framework of **?** to define a model selection criteria based on minimizing a predictive loss. For circular data the usual loss functions on the real line are not well defined. We propose a Circular Predictive Discrepancy (CPD) measure based on posterior predictive distribution. Define $y^{obs} = (y_1, ..., y_n)$ as the observed data and $y^{pred} = (y_1^{pred}, ..., y_n^{pred})$ as the predictive data obtained from the following posterior predictive distribution,

$$p(y^{pred}|y^{obs}) = \int\int p(y^{pred}|\mu,\rho)p(\mu,\rho|y^{obs})d\mu\, d\rho,$$

where $p(y^{pred}|\mu, \rho)$ denotes the sampling distribution of the data, which is the wrapped density evaluated at y^{pred} given μ and ρ. $p(\mu, \rho|y^{obs})$ denotes the posterior distribution of the parameters (μ, ρ) given the observed data y^{obs}. In order to measure the discrepancy between observed and predicted quantities we use the following loss function and call it an Absolute Predicted Errors (APE) defined as,

$$APE \;=\; \sum_{i=1}^{n} \min\left(\left|y_i^{pred} - y_i\right|, 2\pi - \left|y_i^{pred} - y_i\right|\right).$$

Based on the above loss function we define the CPD as,

$$CPD \;=\; E[APE|y^{obs}]$$

Note that as in the case of linear loss functions, we can not break up the CPD into two terms involving a goodness-of-fit term and a penalty term. It is apparent that given a set of models, we will prefer a model with the lowest CPD.

We use both of the above model selection methods to choose from a given set of wrapped distributions. In order to see the performance of these model selection methods, we conduct a simulation study and monitor the percentage of success. Details of such a simulation study is presented in the next section.

5. Simulation studies

In this Section, we present couple of simulation studies to illustrate (i) the performance of the parameter estimation method presented in Section 3 and (ii) the performance of model selection criteria presented in Section 4. It is to be noted that we present only some selected results of a much broader simulation study which can be found in the doctoral thesis of the first author. In terms of parameter estimation, we found that when data are generated from the true model that is being fitted, the parameter estimates obtained by the data augmentation method (presented in Section 3) have very good frequentist properties in terms of low bias and nominal standard errors. Also based on 95% posterior intervals, we found that the coverage of such interval estimates are very close to the nominal level. These results have been excluded in this article. However, we present results from the study when the true model is misspecified. In order to study the sensitivity of the likelihood we generated data from Von Mises (VM) on the circle (0, 2π). To generate the data from

Von Mises distribution, we use the algorithm given by Best and Fisher (1978). For our simulations we fixed $\mu = \pi/2 \simeq 1.5708$ and $\rho = 0.5$. We generated $n = 50$ observations from the $VM(\pi/2, 0.5)$ and then fitted wrapped normal, wrapped Cauchy and wrapped double exponential distribution to each of the data generated from the above mentioned Von Mises distribution. We repeated the method 500 times to see the frequentist performance of the Bayes method. In addition, to observe model selection issues, we computed model selection criteria DIC and CPD to select the best model. In Table **??**, the percentages each of the fitted model chosen by the model selection criteria CPD and DIC are given. For model fitting, we used $a_\rho = 0.5$ for all priors. In our study (not shown here) we did not find the posterior summary to be very sensitive to the choice of a_ρ, provided $a_\rho \in (0, 2]$.

We computed the posterior mean, standard deviation (s.d), 2.5 percentile, median and 97.5 percentile for μ and ρ for each simulation. The percentiles are computed with 0 radians as the reference point on the circle. Usually, the reference point on the circle is chosen diametrically opposite to the sample mean, or where the samples are sparsely distributed. Note that, changing the reference point does not affect the circular mean. The simulation standard errors for each of these summary values are also computed. We also compute the coverage probability (c.p) for the 95% posterior interval given by the 2.5th and 97.5th percentile of the posterior distribution. After some preliminary studies, for all simulations, we choose the burn-in period to be 2000 samples (i.e. throw away first 2000 samples from the MCMC chain) and then keep 5000 samples after burn in, to obtain posterior summary values. All of these summary values are based on these final 5000 samples. In Table **??**, we present the empirical mean of the posterior summary values along with the standard error (s.e.) of the posterior summary values. Notice that the empirical s.e.'s match well with the estimated one given by the standard deviation of the posterior distribution. For instance, the average posterior standard deviation of μ under a WN model is 0.22, which matches well with the Monte Carlo standard error of 0.21. Also the empirical coverage probability for the estimation of μ using a WN model is exactly same as the nominal level. However, the results may not be that close for other models. For example, under a WDE model, the empirical coverage probability is about 92% for μ whereas it is about 94% for ρ. In general, we see that the location parameter μ is well estimated compared to the scale parameter ρ even under model misspecification. One of the benefits of using a Bayesian method which enables us to obtain the finite sample s.e. is that no large-sample theory is required and the

Table 1. Fitting WN, WC and WDE to VM(1.57, 0.5) distribution with $\rho \sim Beta(0.5, 0.5)$

WN	mean	std dev	2.5 %	50 %	97.5 %	cov prob
μ	1.57	0.22	1.15	1.57	2.01	0.95
s.e.	0.21	0.11	0.23	0.20	0.39	
ρ	0.47	0.08	0.31	0.47	0.60	0.91
s.e.	0.08	0.01	0.10	0.08	0.07	
CPD	0.33					
DIC	0.47					
WC	mean	std dev	2.5 %	50 %	97.5 %	cov prob
μ	1.57	0.20	1.18	1.57	1.97	0.93
s.e.	0.20	0.08	0.23	0.20	0.32	
ρ	0.44	0.07	0.30	0.45	0.58	0.87
s.e.	0.07	0.01	0.09	0.08	0.07	
CPD	0.18					
DIC	0.21					
WDE	mean	std dev	2.5 %	50 %	97.5 %	cov prob
μ	1.57	0.19	1.19	1.57	1.96	0.92
s.e.	0.20	0.07	0.23	0.20	0.28	
ρ	0.48	0.08	0.31	0.48	0.63	0.94
s.e.	0.08	0.01	0.10	0.09	0.07	
CPD	0.49					
DIC	0.32					

Table 2. Performance of CPD and DIC in selecting the best model

DGP		FIT		
		WN	WC	WDE
WN	CPD	0.59	0.07	0.34
	DIC	0.77	0.07	0.16
WC	CPD	0.08	0.44	0.48
	DIC	0.09	0.55	0.35
WDE	CPD	0.17	0.29	0.55
	DIC	0.24	0.38	0.38

performance of the estimates are quite good even from frequentist perspective. However, it is noted that the proposed Bayesian method is computationally more intensive than the comparable frequentist procedures. In SAS, on a Sparc 20 machine, on an average it took about 50 minutes to perform the entire simulation for a given wrapped distribution.

In Table **??**, we see that WN and WDE are performing well in estimating μ and ρ when they come from a Von Mises distribution. It is expected that WN will perform well because WN closely approximates the Von Mises distribution. WC model does not perform well in estimating the mean resultant length. The lower than nominal coverage probability of ρ indicates that WC is not robust in estimating the mean resultant length of the distribution when the distribution is misspecified. However, the location parameter is generally estimated very well. Using CPD, we find that the WDE model gets picked up most of the time (49%) as the best fit to VM data. On the other hand, DIC picks WN more often (47%) as the best fit to VM data.

In order to study the performance of the model selection criteria presented in Section 4, we generated data from Normal, Cauchy and Double Exponential distributions on the real line and wrapped them onto the circle (0, 2π). For our simulations, we fixed $\mu = \pi/2$ and $\rho = 0.5$. We generated $n = 50$ observations from the $WN(\pi/2, 0.5)$, $WC(\pi/2, 0.5)$ and $WDE(\pi/2, 0.5)$. We then fitted Wrapped Normal, Wrapped Cauchy and Wrapped Double Exponential distributions to each of the three datasets. We repeated the method 500 times to see the frequentist performance of model selection methods. For model fitting we used $a_\rho = 0.5$ for all priors. In Table **??**, the DGP refers to the data generating process from

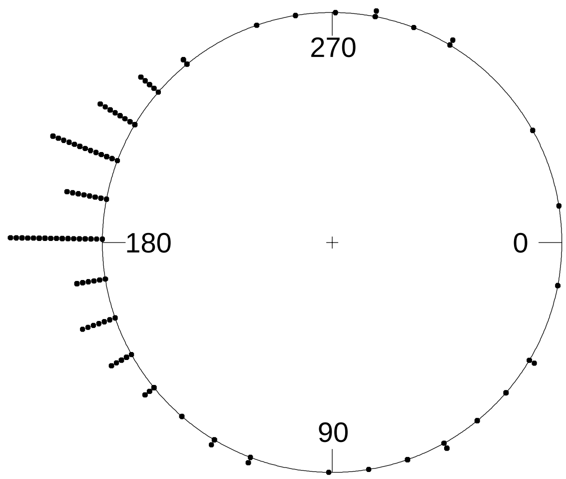

Fig. 2. Circular plot of the Jander's Ant data

WN, WC and WDE. For each of the three DGP, we fit WN, WC and WDE and report the percentages, each of the fitted model chosen by the model selection criteria DIC and CPD. Notice that each row sums to unity. For instance, when we generate data from a WN distribution and fit all three distributions, CPD picks up WN as the best model 59% of the time as compared to picking up 7% and remaining 34% by WC and WDE as the best model, respectively. On the other hand DIC in this case performs much better in the sense that it picks up WN as the best model 77% of the time. But the performance of DIC is not uniformly better than CPD, for instance when data are generated from WDE, CPD picks up the true model more often than the DIC. It appears that both CPD and DIC are performing reasonably well in picking the correct distribution, though more separation of the models are required for better performance. Notice that we used the same μ and ρ parameter values for each of the three distributions. Aside for the performance of the model selection criteria, another feature to observe is that, in general WDE model fits well even when the data is generated from other distributions like WN and WC. In this sense, we conclude that the estimation of μ and ρ based on a WDE model will be robust against model misspecification.

Table 3. Posterior summary values of (μ, ρ) based on fitting WN, WC and WDE models to the Ant data

WN	mean	s.d	2.5%	50%	97.5%
μ	3.13	0.13	2.88	3.13	3.40
ρ	0.53	0.05	0.42	0.53	0.62
CPD	1.16				
DIC	302.35				
WC	mean	s.d	2.5 %	50 %	97.5 %
μ	3.24	0.06	3.12	3.24	3.36
ρ	0.65	0.04	0.56	0.65	0.72
CPD	1.05				
DIC	266.97				
WDE	mean	s.d	2.5 %	50 %	97.5 %
μ	3.20	0.08	3.07	3.19	3.37
ρ	0.62	0.04	0.54	0.62	0.69
CPD	1.08				
DIC	275.40				

6. Application to real data sets

We analyze a data set that presents the orientation of the ants towards a black target when released in a round arena. This experiment was originally conducted by **?** and later mentioned in **?**. The data consists of 100 observations. For this data set, circular sample mean and resultant direction are 3.20 radians (183°) and 0.61 respectively.

We plot a circular histogram of this data in Figure **??**. We fit WN, WC and WDE distributions to this data. We compute the posterior mean, standard deviation (s.d), 2.5 percentile, median and 97.5 percentile for μ and ρ. We have used a class of non-informative priors for (μ, ρ) which is given in equation (**??**), but present the results only for $a_\rho = 0.5$, in Table **??** .

While fitting the data, initially we rejected the first 2000 samples (burn-in period) of the MCMC chain and kept the next 5000 samples after burn-in to do posterior inference. However we observed some problem with convergence (using CODA). So we increased the

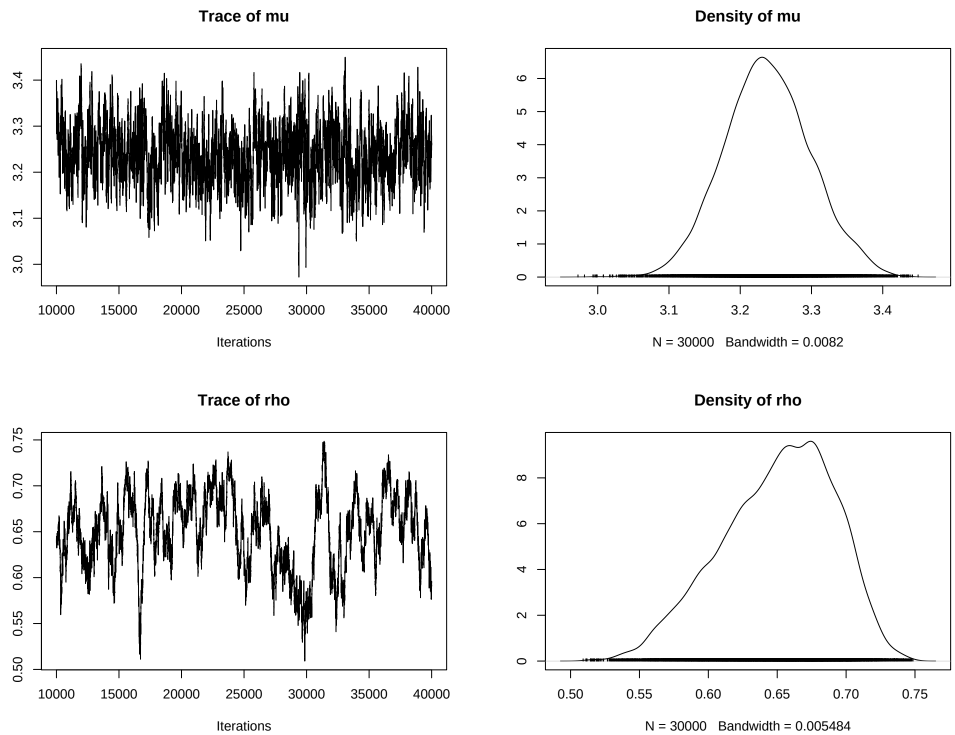

Fig. 3. MCMC output for parameters while fitting Wrapped Cauchy to ant data

burn-in time to 10000 samples and kept the next 30000 samples. We used multiple starting values for μ and ρ, viz. $\mu = 3.2$ and 0 and $\rho = 0.61$ and 0.99. We find that after 3000 samples, the two chains overlap each other and achieves good mixing. Carefully monitoring other diagnostics aspects of convergence, such as auto/cross correlations using CODA (not presented here), we assert that there was no apparent problem with MCMC convergence. Summary values of the parameters are given in Table **??**.

To compare the models we use CPD and DIC. A smaller value of the model selection criteria indicates a better fit. From Table **??**, we see that WC is performing best based on CPD and DIC. WDE estimates are close to WC estimates. But the estimates from WN are different from WC and WDE. Using CODA, we obtained the trace plots and the kernel density estimates of the parameters μ and ρ based on the wrapped Cauchy (WC) distribution, which are presented in Figure **??**.

This data has been previously analyzed by **?**. They find through hypothesis testing, that WN does not fit this data very well, which concurs with our finding as well. They recommend a family of symmetric wrapped stable distributions (SWS, see **?**, p. 52) or a mixture of SWS and circular uniform density for this data. Notice that WC belongs to the SWS family and hence this conclusion agrees with the result obtained from the CPD and DIC values. The test presented by **?** is based on a large-sample theory, whereas our findings are based finite sample method.

7. Discussion

In this article, we provided a generic method to fit a wide class of continuous wrapped densities on the circle using MCMC methodology. We showed that using this method, we can easily obtain samples from the joint posterior distribution, and hence, compute the posterior summary values such as the posterior mean, standard deviation, 2.5 percentile, median and 97.5 percentile for μ and ρ. From simulation-based studies, we found that parameter estimates based on WDE are robust to erroneous models. In addition, we found that the posterior distribution is not sensitive to the class of non-informative priors that we proposed. We also studied the performance of two model selection criteria viz., CPD and DIC, and using simulation studies we showed that both criteria perform well in selecting the correct model from a set of given models. As an extension to this work, we have developed regression models based on wrapped distributions and will be reported elsewhere. Another extension is to consider correlated circular observations such as time series model by wrapping autoregressive model on the line to a circle.

Acknowledgments

The authors would like to thank Dr. Kaushik Ghosh at the George Washington University, for providing the Splus code to plot the circular histograms.

References

Best, D. J. & Fisher, N. I. (1978) Efficient simulation of the Von Mises distribution. *Appl.Statist.*, **28**, 152-157.

Coles, S. (1998) Inference for circular distributions and processes. *Statistics and Computing*, **8**, 105-113.

Damien, P., Wakefield, J. & Walker, S. (1999) Gibbs sampling for Bayesian non-conjugate hierarchical models by using auxiliary variables. *J. R. Statist. Soc. B*, **61**, 331-344.

Damien, P. & Walker, S. (1999) A full Bayesian analysis of circular data using von Mises distribution. *The Canadian Journal of Statistics*, **27**, 291-298.

Fisher, N. I. (1993) *Statistical analysis of circular data.*, Cambridge Univ. Press, Cambridge.

Fisher, N. I. & Lee, A.J. (1994) Time series analysis of circular data. *J. R. Statist. Soc. B*, **56**, 327-339.

Gelfand, A. E. & Ghosh, S. K. (1998) Model choice: A minimum posterior predictive loss function. *Biometrika*, **85**, 1-11.

Higdon, D. M. (1998) Auxiliary variable methods for Markov Chain Monte Carlo with applications. *J. Am. Statist. Assoc.*, **93**, 585-595.

Jander, R. (1957) Die optische Richtungsorientierung der roten Waldameise (*Formica rufa L.*). *Zeitschrift Fur Vergleichende Physiologie*, **40**, 162-238.

Kent, J. T. & Tyler, D. E. (1988) Maximum likelihood estimation for the wrapped Cauchy distribution. *Journal of Applied Statistics*, **15**, 247-254.

Mardia, K. V. (1972) *Statistics of directional data.*, Academic Press, London.

Mardia, K. V. & Jupp, P. E. (1999) *Directional statistics.*, Wiley, Chichester.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models, 2nd ed.*, Chapman and Hall, New York.

Neal, R.(2003) Slice sampling (with discussion). *Annals of Statistics*, **31**, 705-767.

Sengupta, A & Pal, C (2001) On optimal tests for isotropy against the symmetric wrapped stable-circular uniform mixture family. *Journal of Applied Statistics*, **28**, 129-143.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002) Bayesian measures of model complexity and fit, (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series B*, **64**, 583-639.

Stephens, M. A. (1969) *Techniques for directional data.* Technical Report 150, Department of Statistics, Stanford University.

Tanner, M. A. & Wong W. H. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.*, **82**, 528-540.

van Dyk, D. A. & Meng, X. L. (2001) The art of data augmentation. *Journal of Computational and Graphical Statistics*, **10**, 1-50.

Von Mises, R. (1918) Über die "Ganzzahligkeit" der Atomgewicht und verwandte Fragen. *Physikalische Zeitschrift*, **19**, 490-500.

Appendix: Full conditional densities for the WDE model

The full conditional densities of $\boldsymbol{k}$, $\boldsymbol{v}$, $\boldsymbol{z}$, μ and σ^2 are given by

$$\begin{aligned}
[k_j|\boldsymbol{y},\boldsymbol{k_{-j}},\mu,\boldsymbol{z},\boldsymbol{v},\sigma] &\propto I_{v_j}\left(0, e^{-\frac{|y_j-\mu+2\pi k_j|}{\sigma}}\right),\\
\Rightarrow k_j|\boldsymbol{y},\boldsymbol{k_{-j}},\mu,\boldsymbol{z},\boldsymbol{v},\sigma &\sim DU\left[\left\lceil \tfrac{1}{2\pi}\left(\mu - y_j + \sigma\log v_j\right)\right\rceil, \left\lfloor \tfrac{1}{2\pi}\left(\mu - y_j - \sigma\log v_j\right)\right\rfloor\right],\\
&\quad\text{where } DU \text{ stands for Discrete Uniform.}\\
v_j|\boldsymbol{y},\boldsymbol{k},\mu,\boldsymbol{z},\boldsymbol{v_{-j}},\sigma &\sim U\left[0, e^{-\frac{|y_j-\mu+2\pi k_j|}{\sigma}}\right],\\
z_0|\boldsymbol{y},\boldsymbol{k},\mu,\boldsymbol{z_{-0}},\boldsymbol{v},\sigma &\sim U\left[0, \tfrac{1}{\sigma^{n-2a_\rho+1}}\right],\\
z_1|\boldsymbol{y},\boldsymbol{k},\mu,\boldsymbol{z_{-1}},\boldsymbol{v},\sigma &\sim U\left[0, \tfrac{1}{(1+\sigma^2)^{2a_\rho}}\right],\\
\mu|\boldsymbol{y},\boldsymbol{k},\boldsymbol{z},\boldsymbol{v},\sigma &\sim U\left[m_\mu, M_\mu\right],\\
\text{where } m_\mu &= \max_{j=1}^{n}\left[y_j + 2\pi k_j + \sigma\log v_j\right]\bigvee 0 \text{ and}\\
M_\mu &= \min_{j=1}^{n}\left[y_j + 2\pi k_j - \sigma\log v_j\right]\bigwedge (2\pi).\\
\sigma|\boldsymbol{y},\boldsymbol{k},\mu,\boldsymbol{z},\boldsymbol{v} &\sim U\left[m_\sigma, M_\sigma\right],\\
\text{where } m_\sigma &= \max_{j=1}^{n}\left[\tfrac{|y_j-\mu+2\pi k_j|}{-\log v_j}\right] \text{ and}\\
M_\sigma &= \frac{1}{z_0^{\frac{1}{n-2a_\rho+1}}}\bigwedge\sqrt{\frac{1-z_1^{\frac{1}{2a_\rho}}}{z_1^{\frac{1}{2a_\rho}}}},
\end{aligned}$$

where $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.