VON MISES-BASED SOLUTIONS FOR PRACTICAL PROBLEMS

BAYESIAN CIRCULAR STATISTICS

KEES MULDER

BAYESIAN CIRCULAR STATISTICS VON MISES-BASED SOLUTIONS FOR PRACTICAL PROBLEMS KEES MULDER

Bayesian Circular Statistics

von Mises-based solutions for practical problems

Bayesiaanse Circulaire Statistiek

Oplossingen voor praktische problemen gebaseerd op de von Mises-verdeling

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 21 juni 2019 des middags te 4.15 uur

door

Kees Tim Mulder

geboren op 3 mei 1990
te Oegstgeest

Promotoren:
Prof.dr. I.G. Klugkist
Prof.dr. H.J.A. Hoijtink

The studies in this thesis were funded by the Netherlands Organization for Scientific Research (project 452-12-010).

Beoordelingscommissie:
A. Postma
D. Borsboom
E. Hamaker
C. Ley
Y. Rosseel

Bayesian Circular Statistics: von Mises-based solutions for practical problems
Proefschrift Universiteit Utrecht, Utrecht. - Met samenvatting in het Nederlands.
Cover illustration by Esther Kuik (estherkuik.com).

Contents

Introduction

1 Bayesian Estimation and Hypothesis Tests for a Circular GLM **1**
- 1.1 Bayesian circular GLM . . . 4
  - 1.1.1 Likelihood . . . 6
  - 1.1.2 Priors . . . 6
- 1.2 MCMC sampling . . . 9
  - 1.2.1 Sampling β_0 . . . 9
  - 1.2.2 Sampling κ . . . 9
  - 1.2.3 Sampling $\boldsymbol{\beta}$ . . . 9
  - 1.2.4 Sampling $\boldsymbol{\delta}$ . . . 11
- 1.3 Hypothesis tests . . . 11
  - 1.3.1 Equality constrained hypotheses . . . 12
  - 1.3.2 Inequality constrained hypotheses . . . 13
- 1.4 Simulation study . . . 14
  - 1.4.1 Simple regression . . . 14
  - 1.4.2 Factorial ANOVA . . . 15
  - 1.4.3 ANCOVA with four covariates . . . 16
  - 1.4.4 Bayes factors and posterior model probabilities . . . 17
- 1.5 Example . . . 18
  - 1.5.1 ANOVA model . . . 19
  - 1.5.2 ANCOVA model . . . 21
  - 1.5.3 Inequality constrained hypothesis . . . 22
- 1.6 Discussion . . . 23
- 1.7 Acknowledgements . . . 25
- 1.A Conditional distribution of β_0 . . . 26

2 Bayesian Tests for Circular Uniformity **27**
- 2.1 Frequentist tests of circular uniformity . . . 29
- 2.2 Tests with a von Mises alternative . . . 30

2.2.1 Choosing priors . . . 31
2.2.2 Priors based on the conjugate prior . . . 32
2.2.3 Jeffreys prior . . . 34
2.2.4 Simulation . . . 35
2.3 Tests with a Kernel Density alternative . . . 37
2.3.1 Simulation . . . 38
2.4 Examples . . . 39
2.4.1 Homing pigeon example 1 . . . 40
2.4.2 Homing pigeon example 2 . . . 41
2.5 Discussion . . . 42

3 Mixtures of Peaked Power Batschelet Distributions for Circular Data With Application to Saccade Directions 45
3.1 Family of Batschelet Distributions . . . 48
3.1.1 Inverse Batschelet distribution . . . 48
3.1.2 Power Batschelet distribution . . . 52
3.1.3 Measures of circular dispersion . . . 54
3.2 Inference for Batschelet mixtures . . . 55
3.2.1 EM Algorithm . . . 55
3.2.2 Bayesian inference . . . 56
3.2.3 Model identifiability . . . 59
3.2.4 Model selection and hypothesis testing . . . 60
3.3 Illustration . . . 62
3.3.1 Synthetic data . . . 62
3.3.2 Free-viewing data . . . 62
3.4 Discussion . . . 67
3.A Properties of the Power Batschelet Distribution . . . 70

4 Bayesian inference for mixtures of von Mises distributions using the Reversible Jump MCMC sampler 73
4.1 Von Mises Mixture Model . . . 76
4.1.1 Von Mises mixture density . . . 76
4.1.2 Likelihood . . . 77
4.1.3 Prior distributions . . . 78
4.2 Reversible jump MCMC for von Mises Mixtures . . . 79
4.2.1 Updating the weights w . . . 79
4.2.2 Updating component parameters μ and κ . . . 80
4.2.3 Updating the allocation z . . . 80
4.2.4 Dimensionality changing moves . . . 80
4.2.5 Label switching . . . 85
4.3 Simulation study . . . 86

4.3.1 Simulation scenarios . . . 86
4.3.2 Starting values . . . 86
4.3.3 Convergence . . . 87
4.3.4 Results . . . 87
4.4 Illustration . . . 90
4.4.1 Dataset . . . 90
4.4.2 Results . . . 91
4.5 Discussion . . . 92

5 Inference for Interval-Censored Circular Data With Application to Crime Times 95
5.1 Aoristic data . . . 98
5.2 Aoristic Fraction method . . . 100
5.3 Statistical models for aoristic data . . . 103
5.3.1 Circular data models . . . 103
5.3.2 Aoristic Likelihood . . . 104
5.3.3 Data augmentation . . . 107
5.4 Nonparametric models . . . 107
5.4.1 Dirichlet Process Mixture models . . . 108
5.5 Applications . . . 110
5.5.1 Ashby & Bowers data . . . 111
5.5.2 Montgomery Crime data . . . 112
5.5.3 Montly crime time trends . . . 113
5.6 Discussion . . . 115
5.7 Acknowledgements . . . 117
5.A Proof of variance overestimation using the aoristic fraction method . . . 118
5.A.1 Data on the real line . . . 118
5.A.2 Aoristic data . . . 120
5.B Von Mises Dirichlet Process Mixture model . . . 123
5.C Rejection sampling aoristic data . . . 125
5.D Prior independent of μ_0 . . . 126

6 circbayes: An R package for Bayesian circular statistics 129
6.1 Related work . . . 131
6.2 General package structure . . . 133
6.3 Circular statistics . . . 135
6.3.1 Summary statistics . . . 135
6.3.2 Probability distributions . . . 135
6.4 Bayesian Circular regression . . . 140
6.4.1 Circular regression with von Mises residuals . . . 141

6.4.2 Circular regression with Projected Normal residuals . . 142
6.4.3 Comparison between the two regression models 144
6.4.4 Hierarchical Projected Normal regression 144
6.5 Density estimation problems 145
6.5.1 Mixture models 146
6.5.2 Bayesian Nonparametrics for Circular data 147
6.6 Hypothesis tests and model comparison 149
6.6.1 Information criteria 151
6.6.2 Nested comparison Bayes Factors 151
6.6.3 Marginal Likelihood computation 152
6.6.4 Hypothesis testing examples 152
6.7 Discussion 155

Nederlandse samenvatting **157**

Dankwoord **161**

Curriculum Vitae **163**

Introduction

In a small café in the historical city center, three long-time friends meet for dinner. After some chit-chat, one of the friends, in a philosophical mood, asks: "*Hey, I wonder what our average birthday is?*"

Pondering over this somewhat odd question, the friends first conclude that the question is clearly more reasonable than a nonsensical question such as "*What is the average of three cows and a goat?*". Then again, a question such as "*What is the average of 5, 13, and 7?*" is obviously much easier to answer, as the object (or, *data*) to be averaged is simply a real number. As often happens with seemingly hard-to-answer puzzles, the question stuck in the mind, and the friends end up discussing the issue for two hours[1].

Thankfully, the friends are all born in the month February, on February 3, 7, and 26 to be precise. Therefore, they now decide to celebrate their triple birthday on their average birthday of February 12th, a simple average of the three days. However, just as the three friends rejoice the imminent financial benefits of throwing not three birthday parties but one, more of their friends start arriving.

The first new arrival has her birthday on January 7. Not a problem, conclude the friends. They come up with the scheme of labeling all days in the year consecutively, and averaging these. So, they average the new arrival's score of 7 with their birthdays, which are now labeled 34, 38, and 57. The new average is 34, which translates back to an average birthday of February 3rd. Content, the friends continue their dinner, although they are left with the sneaking suspicion that they have cheated, somehow.

The real trouble begins when four more friends arrive. Due to a mild statistical anomaly[2], they all happen to be born in December, on the 5th, the 6th, 16th, and 21st. Now, the strategy of consecutively labeling all the days in the year not only has become excessively laborious[3], it also leads to

[1] This true story was relayed to me by my friend and colleague Erik-Jan van Kesteren.

[2] The probability of this is described in a famous statistical problem, which also happens to be about birthdays (see https://en.wikipedia.org/wiki/Birthday_problem).

[3] Not to mention inevitable issues with leap years, although we will ignore those here.

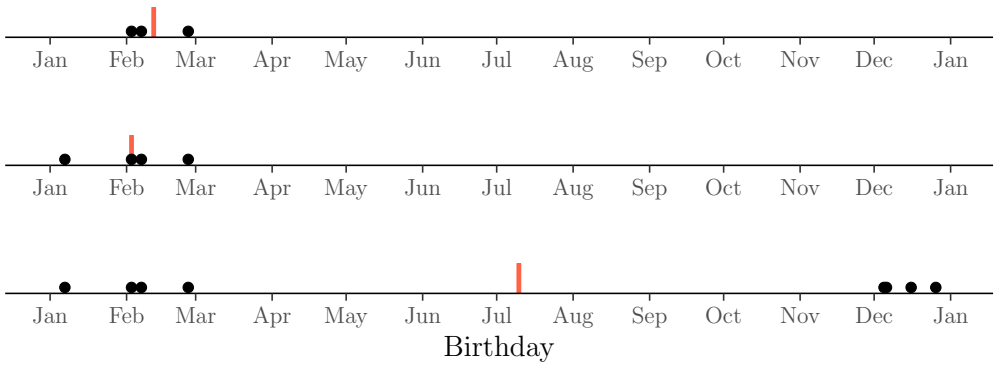


Figure 1: Birthdays of our group of friends (black dots), and the average birthday (red line) progression for our group of friends, as friends keep arriving.

some odd conclusions. The average of 7, 34, 38, 57, 339, 339, 350, and 355 is 190. This 190 translates to back to the day of July 9, which baffles the group of friends for two reasons.

First, the average birthday of July 9th is a summer day, and is nowhere near the birthday of any of the people in the group, who were all born in winter. Second, if the friends had started counting the days of the year at March 1st, just as good a day to start as the arbitrary January 1st, the resulting average birthday would have been completely different. Clearly, the strategy of consecutive labeling fails to accurately capture what we mean by 'average'.

So what happened here? Within the month of February, the days nicely follow each other, and the average seems to makes sense. But when we expand our scope to the full year, a problem arises. When the year progresses to its final day, the day of the year rolls around to the first day of the new year. The moment the previous year ends and the moment the new year begins are one and the same, so the beginning and end of the year are tied together. The day of the year is, in this sense, periodic (or *cyclical*), and it is exactly this periodicity which makes the friend group's attempts at creating a sensible average fail.

So, the friends go home unsatiated, mulling over possible solutions every once in a while in the weeks to follow. However, we will see that the solution to this problem is much simpler if we think of the days of the year not on the real line, but on the circle (see Figure 2). Because the ends of the scale (December 31 and January 1) are now connected correctly, there is no longer ambiguity about the average.

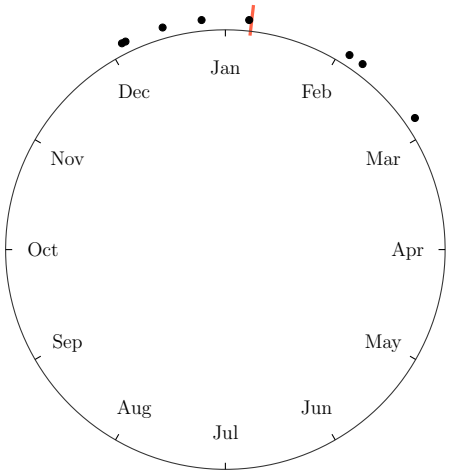


Figure 2: On the circle, the circular mean (red line) of the sample of birthdays (black dots) is a natural central tendency.

Circular data

As it turns out, this 'average birthday' problem is not as frivolous as it seems at first glance. A birthday is simply an example of an observed day of the year. Throughout statistics, examples can be found where the interest is in statistical modeling and prediction when in a year certain events occur. One example is modeling the distribution of extreme weather events (such as hurricanes) over the year. In contrast, the question of whether extreme weather events have increased in number over the last 30 years is quite different, because it is interested in the linear aspects of time, that is, temporal trends over multiple years. If we wish to truly understand extreme weather events, both parts must be understood, although the linear temporal aspects have traditionally received more attention.

Instead of yearly periodic data, one can also consider hours of the day, days of the week, days of the month, and so forth. In fact, many occurences of events are governed in some way by a periodical process, which must be taken into account when modeling if we wish to understand and control this process. For example, the number of patients arriving at the emergency room of a hospital differs strongly over the hours of the day.

In fact, temporal data is just one example of measurements on a periodic scale. Another major source are directly observed directions or angles, for example measured by the compass in degrees. For example, one might consider movement directions of animals, or orientations of rock formations. These

more directly provide us with data that is periodic. That is, two directions of 1° and 359° are very close, but this is not immediately apparent from the numerical values.

Besides directly observing circular data, in some cases it has turned out to be beneficial to transform a linear problem into a circular one. A prime example is the use of circumplex scales in psychology (Gurtman, 2015). Such scales extract a circular observation from a questionnaire, by means of a latent factor analysis method which imposes a circular structure on the correlation matrix (Browne, 1992). While usually latent factors are represented by linear variables, the circumplex structure can in some cases provide a simpler and more interpretable description of the underlying latent factors.

Regardless of the origin of the data, any data whose sample space is periodic requires us to take into account the periodicity during analysis. This periodic type of data is what is generally referred to as *circular data* (or rarely *angular data*), and this dissertation consists of methods to treat circular data problems.

Applications

Interestingly, circular data tend to be encountered in a niche of almost every scientific field. Therefore, circular statistics are used by a large number of somewhat disconnected research communities. These communities partly use their own specialized circular data methods, but also build on the general framework of circular statistics. To build some intuition for the variety of research for which circular models are used, we will recap a few examples of circular data applications here.

In meteorology, one is often concerned with modeling wind directions (Bowers et al., 2000). Wind directions can be directly measured, for example using a wind vane. A question of interest could be whether a certain condition, such as the SO2 concentration, is related to the observed wind direction (García-Portugués et al., 2013). Another question is to extract underlying groups of wind direction, using mixture models (Masseran et al., 2013).

In biology, circular data can arise while investigating vanishing directions of birds. For example, in one experiment, the question of interest how homing pigeons use magnetic information to navigate (Gagliardo et al., 2008). To this end, groups of homing pigeons were subjected to sections of their nerves in different ways. Their navigational ability was compared to a control group by measuring in what direction the pigeons would fly when released. A different circular distribution, here, tells us that the bisected nerve is useful in navigation.

In political science, quantifying how average crime times differ between

cities can influence policy (J. Gill & Hangartner, 2010). Here, the problem requires us to model the crime times over the hours of the day, and infer in what way cities differ in their crime time patterns.

Other examples can be found in fields such as aerospace (Kurz & Hanebeck, 2017), machine learning (Gopal & Yang, 2014), signal processing (Traa & Smaragdis, 2013), behavioural biology (Nuñez-Antonio et al., 2018), life sciences (Mardia, 2011), behavioural biology (Bulbert et al., 2015) and environmental sciences (Lagona, 2016; Lagona et al., 2015; Arnold & SenGupta, 2006).

While these problem vary wildly in flavor, all require a toolkit that can deal with circular data. In this dissertation, most motivating examples come from the field of social and behavioural sciences, where circular data is obtained from eye movements (Tatler & Vincent, 2009), experimental designs in motor behaviour research (Mechsner et al., 2001a; Kaas & Van Mier, 2006; Mechsner et al., 2007; Postma et al., 2008) and the aforementioned circumplex models (Leary, 1957; Gurtman & Pincus, 2003; Gurtman, 2009a), to name a few.

Units

A circular observation, which we will usually denote by θ, can be represented mathematically in a variety of ways. Among the most common are degrees, hours, unit vectors, complex numbers and radians. Degrees ($\theta \in [0°, 360°)$) are generally obtained from the compass, while hours are usually measured on a 24-hour clock, so $\theta \in$ [0:00, 24:00).

In some cases, thinking of angles as unit vectors is also beneficial. In this representation, we have a bivariate vector $\boldsymbol{x} = \{x_1, x_2\}^T = \{\cos\theta, \sin\theta\}^T$, and the restriction that the Euclidean norm $||\boldsymbol{x}|| = \sqrt{x_1^2 + x_2^2} = 1$.

An angle can also be represented as a complex number $\cos\theta + i\sin\theta$, where i is the imaginary unit. This can be elegant, particularly because by Euler's formula we have $\cos\theta + i\sin\theta = e^{i\theta}$, which provides a natural mapping between bivariate and univariate representation of angles.

Representing circular observations as radians, where $\theta \in [-\pi, \pi)$, is both ubiquitous and generally advantageous, for both mathematical and computational purposes. Therefore, throughout this dissertation circular observations will usually be represented as radians.

Circular statistics

Techniques for both descriptive and inferential analysis of circular data can be found in a branch of statistics called *circular statistics*. Circular statistics can also be seen as a subfield of *directional statistics*, where one considers analysis of data found on (hyper-)spheres. Because the circle is a special case of the hypersphere, methods from directional statistics are often also applicable to circular statistics.

In this section, we will provide a short introduction to circular statistics by first recapping some important literature, then describing three major approaches to dealing with circular data, and finally discussing circular descriptive and summary statistics.

Literature

A limited number of books have been written on the topic of (frequentist) circular and directional statistics. The seminal technical work is Mardia & Jupp (2000). A more practical handbooks are given by Fisher (1995) and a more recent book by Pewsey et al. (2013) focusing on running circular data models in R. Very recently, two more books have been published, one focusing recent statistical developments for directional data (Ley & Verdebout, 2017), another on recent applications of directional statistics (Ley & Verdebout, 2018). Other books exist with a more specialized focus, such as Batschelet (1981) and Jammalamadaka & Sengupta (2001).

Approaches

Three major distinct approaches for analysis of circular data are available in the literature.

The first is the *intrinsic* approach, where distributions are defined directly on the circle. The most popular of these is the von Mises distribution Von Mises (1918), given by

$$\mathcal{M}(\theta \mid \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu)\}, \tag{1}$$

where μ is the mean direction parameter, κ is a concentration parameter, and $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero. Note that in this probability density function, the cosine guarantees the probability being periodic over the circle.

The second is the *embedding* (or *projecting*) approach (Presnell et al., 1998; Nunez-Antonio & Gutierrez-Pena, 2005), where the circular observa-

tions are assumed to be projections of unobserved bivariate observations to the circle. That is, we assume there is some latent random variable r for each circular observation θ, and that $(r\cos\theta, r\sin\theta)^T \in \mathbb{R}^2$ has a bivariate distribution on the Euclidean plane, usually the bivariate Normal distribution. The main advantage is that this approach allows us to use bivariate models with circular observations. A disadvantage is that we must deal with the latent variable r, which makes computation slightly more difficult.

Thirdly, there is the *wrapping* approach, where distributions on the real line are wrapped around the circle (Ferrari, 2009). In this framework, we represent the circular observation as $\theta = x \mod 2\pi$, where $x \in \mathbb{R}$. This means that univariate distributions can be adapted for use with circular observation, leading for example to the Wrapped Normal distribution. However, for inference this means introducing a latent variable, as in the embedding approach. For the wrapping approach, this latent variable can be seen as some integer $k \in \mathbb{Z}$, and $x = \theta + 2\pi k$. Similar to the projected approach, this means that we pay for the flexibility of using regular univariate models by the burden of modeling a latent variable.

Arguably, a fourth approach exists and is quite common, which is to simply ignore the circular structure of the data and apply regular linear methods, perhaps after some helpful rotations. If the data is concentrated on the circle, such an approach may not have such dire consequences, and can serve as a good approximation to a method that correctly takes the circular nature of the data into account. However, the more the data is spread out over the circle, the more such approaches distort our view of reality.

In this dissertation, focus will be almost entirely on the intrinsic approach. More specifically, we will employ the von Mises distribution and distributions derived from the von Mises distribution throughout, for several reasons. First, the von Mises distribution is of fairly simple mathematical form. Second, the von Mises distribution has directly interpretable parameters. Third, Bayesian inference for the von Mises distribution has recently become more computationally feasible, mostly due to work by Forbes & Mardia (2015). Lastly, Bayesian inference is also facilitated by the fact that conjugate priors are available (Mardia & El-Atoum, 1976).

Descriptive and summary statistics

In the 'average birthday' problem at the beginning of this introduction, it has become clear that circular data require a new definition of the mean. The variance also requires a new definition. These are called the mean direction (often written $\bar{\theta}$) and circular variance. Only the main descriptive and summary statistics will be provided here. For a full overview, see Mardia &

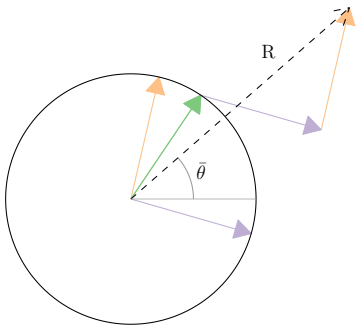


Figure 3: Example of obtaining the mean direction $\bar{\theta}$ and the mean resultant length R. The arrows inside the circle are unit vectors representing the circular observations. Two arrows are repeated outside the circle in the same color to show the summation of the unit vectors. The resultant vector (dashed line) has direction $\bar{\theta}$ and length R.

Jupp (2000).

To compute the mean direction, each angle in a circular dataset can be viewed as a vector of length 1 in direction θ_i. As illustrated in Figure 3, the summation of these vectors results in a vector in direction $\bar{\theta}$ which has length R. The mean direction $\bar{\theta}$ is an unbiased estimator of the mean direction of the von Mises distribution μ.

The resultant length R may be directly obtained from

$$R = \sqrt{\left(\sum_{i=1}^{n} \cos\theta_i\right)^2 + \left(\sum_{i=1}^{n} \sin\theta_i\right)^2},$$

which increases with concentration and sample size. For the von Mises distribution, R is a sufficient statistic for κ.

The mean resultant length can be computed as $\bar{R} = R/n$, which is a metric of concentration independent of the sample size. The mean resultant length is used to compute the circular variance, which is defined as $1 - \bar{R}$.

Bayesian statistics

The methods in this dissertation are almost all Bayesian methods. A discussion of the merits and downfalls of the frequentist or Bayesian approaches to statistics is beyond the scope of this work. However, we discuss how this work can be placed within current developments in the field of Bayesian circular

statistics, and highlight how Bayesian circular statistics may be advantageous above frequentist circular statistics.

Among the earliest developments in Bayesian circular statistics was the development of conjugate priors for the von Mises distribution (Mardia & El-Atoum, 1976). Since then, a variety of Bayesian circular methods have been developed, such as Markov chain Monte Carlo (MCMC) methods for the von Mises distribution (Damien & Walker, 1999; Forbes & Mardia, 2015), regression with von Mises residuals (J. Gill & Hangartner, 2010), wrapped models (Ravindran & Ghosh, 2011), semiparametric circular regression (George & Ghosh, 2006; McVinish & Mengersen, 2008; Bhattacharya & Sengupta, 2009), small area estimation (Hernandez-Stumpfhauser et al., 2016) and hierarchical regression with Projected Normal residuals (Nuñez-Antonio et al., 2011; Nuñez-Antonio & Gutiérrez-Peña, 2014). As of this writing, no books have been published that specifically focus on Bayesian circular statistics. However, Bayesian methods offer several appealing properties specific to circular statistics.

First, Bayesian statistics allows us to include prior knowledge on the parameters of the model. While prior knowledge on the current research question may or may not be available, sometimes prior knowledge is available at the level of the model, for example because the model produces nonsensical data for certain parameter values (as seen in Chapter 1, for example).

Second, Bayesian algorithms such as MCMC allow us to trivially obtain inference for functions of parameters. This can be exploited to fit a model using parameters which are easy to work with computationally and mathematically, while providing results for functions of parameters that are interpretable (for example, this is used for the Batschelet-type models of Chapter 3).

Third, Bayesian hypothesis testing provides a very flexible toolbox providing model comparison. This approach can be seen as symmetric in the sense that in Bayesian Hypothesis Testing (BHT), all hypotheses under consideration can be selected as the most likely hypothesis. In contrast, in traditional frequentist Null Hypothesis Significance Testing (NHST), the null hypothesis is special in that it can never be chosen as more likely than the alternative hypothesis, as it can only be rejected or non-rejected. Bayesian Hypothesis testing provides a unified framework for selecting either hypothesis under consideration. This framework is exploited in Chapters 1, 2, 3, and 6.

Finally, in statistical inference, we are generally concerned with not just parameter estimation, but also quantification of the uncertainty around the parameter estimates. However, some circular models are of such mathematical form that it is hard to derive frequentist standard errors. The Bayesian approach naturally takes into account the uncertainty around the parameters

of a model.

Aims & Outline

The goal of this dissertation is to develop Bayesian methods for circular statistics which solve practical research problems. To this end, it has been a goal to develop collaborations with researchers and organizations working in domains which encounter circular data. Two subgoals can be identified.

First, a solid foundation must be built for the use of Bayesian circular statistics. That is, the properties of these methods must be well-studied and understood. In addition, to allow applied researchers to employ these methods, they must also be available in user-friendly software packages. To this end, the vast majority of methods in this dissertation is available in easy-to-use R packages.

Second, specialized models must be developed which are useful for specific problems. While for some problems one can build on general methods for basic questions, researchers often require statistical methods that are geared toward the specific structure of their data or the research questions. Bayesian solutions for a selection of such problems are presented throughout this work.

These goals will be prevalent throughout the following six chapters of this dissertation. The chapters will be shortly described here.

In **Chapter 1**, regression with a circular outcome is addressed. A Bayesian version of the circular regression model by Fisher & Lee (1992) is developed. This model is essentially a Generalized Linear Model (GLM) where a linear prediction is mapped onto a circular outcome space, while the residuals have the von Mises distribution. It is shown that if we add categorical variables as dummy variables into the model, that the shape of the regression function is dependent on the arbitrary labeling of the groups. Therefore, an alternative version of this model is proposed, which circumvents this issue. Bayesian inference and hypothesis testing for this model is developed and investigated by means of a simulation study. An example is provided from the field of cognitive psychology, where an experimental design for measurement of haptic ability leads to circular observations (van Dijk et al., 2013).

Bayesian hypothesis tests for circular uniformity are developed in **Chapter 2**. Testing circular uniformity is a clear example of a hypothesis testing setting where one might be interested in finding evidence in favor of the null hypothesis. This chapter investigates a selection of proper priors and alternative hypotheses for Bayesian hypothesis tests of circular uniformity. Their performance is investigated by means of a simulation study.

The following three chapters, **Chapters 3, 4, and 5**, all represent re-

search projects which have emerged from collaborations with researchers which provided motivating data examples for which no satisfactory methods were available previously.

Chapter 3 is motivated by an example from eye movement research, previously published on in Van Renswoude et al. (2016). In eye movement research, a saccade is the jump one makes from one fixation point to another. The direction of a saccade is one of the properties of interest, and our goal is to adequately model a distribution of saccade directions. In the motivating example, we are interested in quantifying the differences between infants and adults in their respective saccade direction distribution. To do this, a Batschelet Mixture model is developed that captures the multimodal and peaked nature of saccade distributions. The version of the Inverse Batschelet distribution in Jones & Pewsey (2012) fits the peaked nature of the data, but is difficult to compute due because of its mathematical form. We propose an alternative Batschelet distribution which has a similar shape, but is much simpler to compute. Using bridge sampling (Meng & Wong, 1996; Gronau, Sarafoglou, et al., 2017), Bayes factors can be computed for these models. The model is then applied to saccade direction data, and it is shown that the original research question can now be directly evaluated.

A different approach to extending circular mixture models is taken in **Chapter 4**. While for saccade direction data the number of components of the mixture can usually be treated as known, this is not always the case. Here, we are motivated by a dataset containing music listening behaviour of a large sample of users of a streaming website. The goal here is to predict for each hour of the day the probability that a certain genre will be listened to. To this end, the temporal distribution of music listening for each genre must be inferred. A mixture model is appropriate, but the number of components can not easily be prespecified a priori. Therefore, the reversible jump MCMC sampler (Richardson & Green, 1997a) is applied, as this method allows the MCMC sampler to jump between mixture models of different dimension (that is, number of components). As a result, we can infer the number of components. We develop the required steps to apply this sampler for circular data, and apply this new sampler to the music listening data.

The motivating example in **Chapter 5** comes from the field of criminology. In crime modeling, a large amount of attention is devoted to spatial modeling of crime, while temporal differences receive comparatively less attention. However, understanding at which hour of the day crimes occur is essential in both understanding and predicting crime. However, for many types of crime, such as burglary and theft, crime times are usually not directly observed (as the offender is actively trying to not be observed). Therefore, crime times are usually interval-censored circular observations, which are called *aoristic*

data. Some descriptive analysis methods for aoristic data have been developed (Ashby & Bowers, 2013). We extend these methods by showing their limitations and developing methods for full statistical inference from aoristic data. Circular statistical models can be fit using the aoristic likelihood approach. Further, we develop nonparametric methods for density estimation of aoristic data, based on Bayesian Dirichlet Process Mixture models, with the von Mises as the base distribution.

The final chapter, **Chapter 6**, is an overview of the package **circbayes**. This package is a collection of models for Bayesian circular statistics, some of which have been developed in previous chapters. Any researcher interested in applying the methods in this work would do well to start with this chapter, as it provides a short introduction and comparison of different models for Bayesian circular statistics. We show how to perform Bayesian inference for several models, including von Mises, Batschelet, Projected Normal, circular regression, mixture models, and Dirichlet Process Mixture models. The package also includes methods to perform Bayesian hypothesis testing between any of the models included. As a result, it is able to perform Bayesian versions of most hypothesis tests common in frequentist circular statistics. With this toolbox, we hope to have provided an accessible entrypoint for the use of Bayesian circular statistics.

Chapter 1

Bayesian Estimation and Hypothesis Tests for a Circular GLM

Abstract

Motivated by a study from cognitive psychology, we develop a Generalized Linear Model for circular data within the Bayesian framework, using the von Mises distribution. Although circular data arise in a wide variety of scientific fields, the number of methods for their analysis is limited. Our model allows inclusion of both continuous and categorical covariates. In a frequentist setting, this type of model is plagued by the likelihood surface of its regression coefficients, which is not logarithmically concave. In a Bayesian context, a weakly informative prior solves this issue, while for other paramaters noninformative priors are available. In addition to an MCMC sampling algorithm, we develop Bayesian hypothesis tests based on the Bayes factor for both equality and inequality constrained hypotheses. In a simulation study, it can be seen that our method performs well. The analyses are available in the package circglmbayes. Finally, we apply this model to a dataset from experimental psychology, and show that it provides valuable insight for applied researchers. Extensions to dependent observations are within reach by means of the multivariate von Mises distribution.

This chapter is published as Mulder, K., & Klugkist, I. (2017). Bayesian estimation and hypothesis tests for a circular Generalized Linear Model. *Journal of Mathematical Psychology, 80*, 4–14. https://doi.org/10.1016/j.jmp.2017.07.001.

Author contributions: KM and IK designed the study. KM developed the statistical methods, simulation study and software package. KM wrote the paper. IK provided feedback on written work.

Circular data are measured in angles or directions, and are frequently encountered in scientific fields as diverse as life sciences (Mardia, 2011), behavioural biology (Bulbert et al., 2015), cognitive psychology (Kaas & Van Mier, 2006), bioinformatics (Mardia et al., 2008), political science (J. Gill & Hangartner, 2010) and environmental sciences (Lagona, 2016; Lagona et al., 2015; Arnold & SenGupta, 2006). In psychology, circular data occur often in motor behaviour research (Mechsner et al., 2001a, 2007; Postma et al., 2008; Baayen et al., 2012), as well as in the application of circumplex models (Leary, 1957; Gurtman & Pincus, 2003; Gurtman, 2009a). Circular data differ from linear data in the sense that circular data are measured in a periodical sample space. For example, an angle of 1° is quite close to an angle 359°, although linear intuition suggests otherwise.

Therefore, linear models may not properly describe the process that has generated the circular data of interest. Circular data analysis has been developed to deal with this, although attention to this type of analysis has been limited. Only slightly more than a handful of in-depth books on circular data analysis have been published (Fisher, 1995; Mardia & Jupp, 1999; Pewsey et al., 2013; Jammalamadaka & Sengupta, 2001), and in general, statistical methods for circular data are somewhat limited.

Here, attention is turned to analysis of datasets with a circular outcome, predicted by covariates that can be continuous (linear) or categorical. This leads to a structure similar to the Generalized Linear Model (GLM), which has both multiple regression and ANCOVA as special cases.

Three main approaches to circular data analysis might be distinguished. First, the intrinsic approach employs distributions directly defined on the circle (Fisher & Lee, 1992; Artes, 2008). Second, the wrapping approach 'wraps' a univariate distribution around the circle by taking the modulus of data on the real line (Ferrari, 2009; Coles, 1998). Third, the embedding approach projects points from a bivariate distribution to the circle (Nuñez-Antonio et al., 2011; Nuñez-Antonio & Gutiérrez-Peña, 2014; Wang & Gelfand, 2014; Hernandez-Stumpfhauser et al., 2017a; Maruotti, 2016). While the wrapping and embedding approach provide promising avenues of study in their own right, here attention is restricted to the intrinsic approach, as it might provide the most natural analysis of circular data.

Within the intrinsic approach, the circular analogue to the Normal distribution is the von Mises distribution (Von Mises, 1918). This symmetric unimodal distribution is given by

$$\mathcal{M}(\theta \mid \mu, \kappa) = [2\pi I_0(\kappa)]^{-1} \exp\left(\kappa \cos\left[\theta - \mu\right]\right), \tag{1.1}$$

where $\theta \in (-\pi, \pi)$ represents an angular measurement, $\mu \in (-\pi, \pi)$ represents the mean direction, $\kappa \in \mathbb{R}^+$ is a concentration parameter, and $I_0(\cdot)$

represents the modified Bessel function of the first kind and order zero. Some examples of frequentist methods that employ the von Mises distribution are a circular ANOVA (Watson & Williams, 1956), circular ANCOVA (Artes, 2008) and circular regression (Fisher & Lee, 1992). Here, a Bayesian analysis of such models will be developed.

Early approaches to Markov chain Monte Carlo (MCMC) sampling for the von Mises distribution provide a method for sampling μ when κ is known (Mardia & El-Atoum, 1976) and sampling both parameters for a single group of data (Damien & Walker, 1999). Guttorp & Lockhart (1988) present a conjugate prior for the von Mises model. Recent theoretical work has much improved the efficiency of the sampling of the concentration parameter of the von Mises distribution (Forbes & Mardia, 2015).

Some development has also taken place in the field of semiparametric inference for circular data models, often using Dirichlet process priors (Bhattacharya & Sengupta, 2009; Ghosh et al., 2003a; George & Ghosh, 2006; McVinish & Mengersen, 2008). In particular, Ghosh et al. (2003a) provide Bayes factors for the simple hypothesis test of equality of two means. However, these methods are generally complex, which makes it hard to extend these models, for example to include covariates. Therefore, we will focus on parametric models, with residuals following the von Mises distribution.

A Bayesian circular regression analysis has been developed by J. Gill & Hangartner (2010), using starting values from a frequentist iterative reweighted least squares (IRLS) algorithm, which is similar to that used by Fisher & Lee (1992). J. Gill & Hangartner (2010) note that the likelihood function of the regression coefficients from their model is not globally logarithmically concave, which might cause the algorithm to converge to a local maximum. To combat this, J. Gill & Hangartner (2010) advise careful inspection of the likelihood surface of the regression coefficients. Drawbacks of the approach taken by J. Gill & Hangartner (2010) are that a prior is not specified, the algorithm is slow, categorical predictors are not treated separately and for larger models it may be unclear whether the regression coefficients have converged to the global maximum.

Recent work has provided a multivariate extension of the von Mises distribution (Mardia et al., 2008; Mardia & Voss, 2014), which offers a promising new way of thinking about circular covariate models. The multivariate von Mises was applied in this context by Lagona (2016) within a Generalized Linear Model (GLM) setting, applying MCMC likelihood approximation as in Geyer & Thompson (1992) to compute maximum likelihood estimates. This approach is not Bayesian, but it is a promising approach because of its flexibility, allowing both the mean and concentration to be dependent on an arbitrary set of covariates, as well as allowing observations to be dependent.

There are three main drawbacks of the circular GLM approach to circular data analysis currently. First, the GLM approach is not free from the lack of concavity as described in J. Gill & Hangartner (2010), although this has not yet been investigated in detail. Second, the current approach does not have separate parameters for differences in group mean direction, which precludes the popular ANCOVA model to some extent. Third, Bayesian hypothesis tests for this model are not available, which limits its applicability.

The structure of this paper is as follows. The circular data GLM model is developed in a fully Bayesian setting in Section 1.1. The lack of concavity in the likelihood function will be examined, and suggestions will be formulated on how to deal with this issue. Details on the MCMC sampler are provided in Section 1.2. Section 1.3 outlines Bayesian hypothesis tests for this model, both for equality and inequality constrained hypotheses. Then, a simulation study for the method is provided in Section 1.4. Section 1.5 provides an application of our method to empirical data from cognitive psychology. Finally, Section 1.6 provides a short discussion.

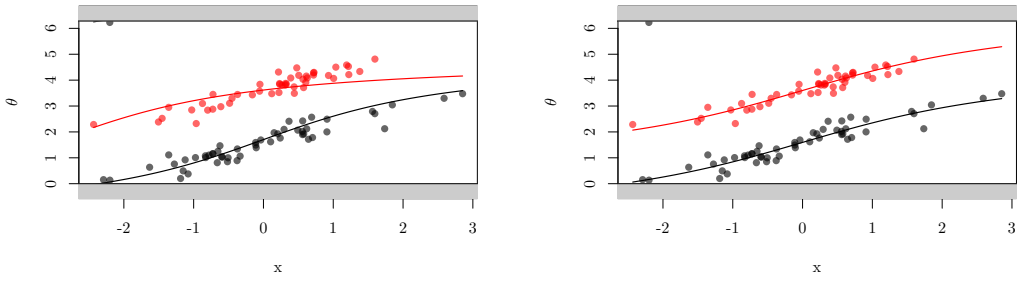
1.1 Bayesian circular GLM

Consider a dataset $\{\theta_i, \boldsymbol{x}_i, \boldsymbol{d}_i\}, (i = 1, \ldots, n)$, where θ_i is a circular outcome variable, $\boldsymbol{x}_i \in \mathbb{R}^K$ is a column vector of continuous linear covariates which are assumed to be standardized, and $\boldsymbol{d}_i \in \{0, 1\}^J$ is a column vector of dichotomous variables indicating group membership. Assume that each observed angle θ_i is generated independently from a von Mises distribution $\mathcal{M}(\theta_i \mid \mu_i, \kappa)$. Then, μ_i is chosen to be

$$\mu_i = \beta_0 + \boldsymbol{\delta}^T \boldsymbol{d}_i + g(\boldsymbol{\beta}^T \boldsymbol{x}_i), \tag{1.2}$$

where $\beta_0 \in [-\pi, \pi)$ is an offset parameter which serves as a circular intercept, $\boldsymbol{\delta} \in [-\pi, \pi)^J$ is a column vector of circular group difference parameters, $g(\cdot) : \mathbb{R} \to (-\pi, \pi)$ is a twice differentiable link function, and $\boldsymbol{\beta} \in \mathbb{R}^K$ is a column vector of regression coefficients. Jammalamadaka & Sengupta (2001) and Fisher & Lee (1992) discuss the choice of the link function. A common and natural choice for the link function is $g(x) = 2\tan^{-1}(x)$, which we will focus on here.

This model specification differs from the usual approach to circular regression models, as these generally set $\mu_i = \beta_0 + g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ (Fisher & Lee, 1992; J. Gill & Hangartner, 2010; Lagona, 2016). However, we view this model as unsatisfactory when including dichotomous predictors in $\boldsymbol{x}$, which we will illustrate in Figure 1.1. Consider a single dichotomous predictor d added to a model with a single continuous predictor x. The dichotomous



(a) $\mu_i = \beta_0 + g(\beta x_i + \delta d_i)$ (b) $\mu_i = \beta_0 + \delta d_i + g(\beta x_i)$

Figure 1.1: Prediction lines from two different models, which were fitted to a dataset with $n = 100$, and true parameters $\delta = 2, \beta_0 = \pi/2, \beta = 0.4, \kappa = 20$. The two models have (a) dichotomous predictors placed in the link function, and (b) dichotomous predictors treated separately.

predictor might be added into the model as $\mu = \beta_0 + g(\beta x + \delta d)$. Adding δ in the link function shifts the location of the prediction line, but also its shape. Therefore, the shape for $d = 0$ is fixed, but for $d = 1$ the shape is dependent on a free parameter, δ. This makes the shape of the prediction line (and therefore the analysis) depend on the arbitrary choice of reference group, which can be seen in Figure 1.1a. To solve this, we advocate setting $\mu = \beta_0 + \delta d + g(\beta x)$, the resulting prediction lines of which are shown in Figure 1.1b.

A comparable approach is taken in Artes (2008), where a separate intercept is estimated for each group. However, having a separate intercept for each group means that a factorial design with main effects only can not be specified. In many applications, especially in psychology, this is problematic. The approach here is more flexible in that it allows a researcher to either fit a model with main effects only, to fit a model with specific interactions, or to compare these models. In addition, Artes (2008) also describes a non-parallel case where the regression parameters are estimated separately for each group. This model can be obtained as a special case of the model provided here by including appropriate interaction terms in the model.

1.1.1 Likelihood

Denote the set of parameters by $\boldsymbol{\phi} = \{\beta_0, \kappa, \boldsymbol{\delta}, \boldsymbol{\beta}\}$. The joint likelihood for the GLM-type model is then given by

$$f(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d} \mid \boldsymbol{\phi}) = \prod_{i=1}^{n} \mathcal{M}(\theta_i \mid \mu_i, \kappa) \tag{1.3}$$

$$= \{2\pi I_0(\kappa)\}^{-n} \exp\left\{\kappa \sum_{i=1}^{n} \cos\left[\theta_i - \left(\beta_0 + \boldsymbol{\delta}^T \boldsymbol{d}_i + g(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right)\right]\right\}. \tag{1.4}$$

If we let $\psi_i = \theta_i - \boldsymbol{\delta}^T \boldsymbol{d}_i - g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, $(i = 1, \ldots, n)$, then we can recognize the conditional likelihood $f(\beta_0, \kappa \mid \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d})$ as the likelihood of the parameters of a von Mises distribution with mean direction β_0 and concentration κ, as shown in Appendix 1.A.

The conditional distribution of β_0 is $\mathcal{M}(\bar{\psi}, R_\psi \kappa)$, where $\bar{\psi}$ and R_ψ are the mean direction and resultant length of the vector $\boldsymbol{\psi}$, given by

$$\bar{\psi} = \operatorname{atan2}\left(\sum_{i=1}^{n} \sin\psi_i, \sum_{i=1}^{n} \cos\psi_i\right), \quad R_\psi = \sqrt{\left(\sum_{i=1}^{n} \cos\psi_i\right)^2 + \left(\sum_{i=1}^{n} \sin\psi_i\right)^2}.$$

Conditionals for κ, $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are not of simple form and require special attention.

1.1.2 Priors

The next step in the model specification is setting prior distributions for the parameters $\boldsymbol{\phi}$. We will focus on uninformative, default priors where possible. The joint prior is factored as

$$p(\boldsymbol{\phi}) \propto p(\beta_0, \kappa \mid \boldsymbol{\delta}, \boldsymbol{\beta}) p(\boldsymbol{\delta}) p(\boldsymbol{\beta}) \tag{1.5}$$

so that $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are independent. Furthermore,

$$p(\boldsymbol{\delta}) \propto \prod_{j=1}^{J} p(\delta_j), \quad p(\boldsymbol{\beta}) \propto \prod_{k=1}^{K} p(\beta_k). \tag{1.6}$$

Next, the choice of each of these priors is discussed.

$p(\delta_j)$

For each δ_j, the circular uniform distribution is a natural and uninformative default prior, so that

$$p(\delta_j) = \frac{1}{2\pi}, \ \forall j = 1, \ldots, J. \tag{1.7}$$

This prior indicates that given a mean direction for some reference group, there is no knowledge on the mean direction of group j.

$p(\beta)$

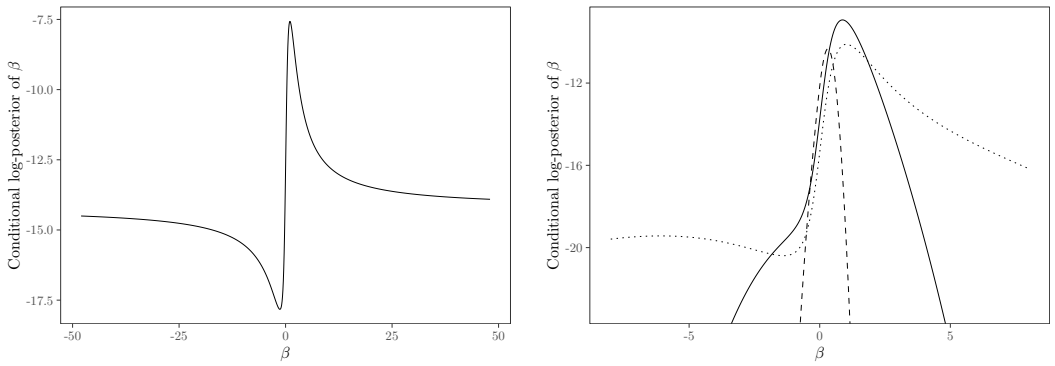
For each β_k there is no natural uninformative prior. A constant prior $p(\beta_k) \propto 1$ could be employed. However, as noted by Fisher (1995) and J. Gill & Hangartner (2010), this leads to a posterior of irregular form, including local maxima and non-zero asymptotes, as shown in Figure 1.2a.

However, because the linear predictors are standardized, the interpretation of the size of β is equal across studies. Therefore, we can determine a priori which values of β_k would be probable in practical research scenarios. In this case, if $\sum_{k=1}^{K} |\beta_k| > 1.5$, the majority of the probability mass of the data is on the semi-circle opposite of the group intercept ($\beta_0 + \boldsymbol{\delta}^T \boldsymbol{d}$), which is not likely in practice. This expectation can be translated to a weakly informative prior distribution. Here, this was done by setting the prior as

$$\beta_k \sim N(0, \sigma^2), \ \forall k = 1, \ldots, K, \tag{1.8}$$

where $N(\mu, \sigma^2)$ denotes the Normal distribution with mean μ and variance σ^2. For a Normal prior with any finite σ^2, there are no non-zero asymptotes in the conditional posterior of β for any values of $\beta_0, \kappa, \boldsymbol{\delta}$, because $\log f_N(x \mid 0, \sigma^2) \to -\infty$ as $|x| \to \infty$. As $\sigma^2 \to 0$, the prior becomes more informative and the posterior for β_k centers on 0. As $\sigma^2 \to \infty$, the prior becomes less informative, but the posterior becomes more irregular, with large plateaus and more local maxima. By default, we choose $\sigma^2 = 1$ so that the prior is the standard Normal distribution, which represents the weakly informative prior mentioned previously, for which values of $|\beta_k| > 1.5$ are a priori unlikely.

To illustrate this, the resulting posterior is compared to the posterior resulting from the constant prior in Figure 1.2. The posterior is based on a synthetic data set of 7 observations with a single predictor $\boldsymbol{x} = -3, \ldots, 3$, which is standardized and the outcome is then computed as $\theta_i = 2 \tan^{-1}(x_i) + \varepsilon_i$ where $\varepsilon_i \sim N(0, 1/10)$, so the true β is 1. In Figure 1.2, the conditional posterior of β is displayed given $\beta_0 = 0, \kappa = 1$. Figure 1.2a gives a zoomed-out view of the posterior resulting from the constant prior, where the asymptotes



(a) $p(\beta_k) \propto 1$

(b) $p(\beta_k) \propto N(0,\ 1/5)$ (dashed)
$p(\beta_k) \propto N(0,\ 1)$ (solid)
$p(\beta_k) \propto N(0,\ 5)$ (dotted)

Figure 1.2: Comparison of the conditional log-posterior of β_k when using (a) a constant prior, which means the log-posterior is equal to the log-likelihood, and (b) a Normal prior with three different values for σ^2.

are clearly visible. Figure 1.2b shows a zoomed-in view of the resulting conditional posterior from a prior with $\sigma^2 = 1/5$ (dashed), $\sigma^2 = 1$ (dashed), $\sigma^2 = 5$ (dotted). It can be seen that the $N(0, 5)$ prior takes a shape not unlike the one seen in Figure 1.2a, although the asymptote is avoided. The $N(0, 1/5)$ prior can be seen to have a strong influence on the posterior estimate for this data where the true $\beta = 1$. The $N(0, 1)$ prior represents a balance for which the asymptotes are solved, but the posterior estimates are very close to the maximum likelihood estimates. In practical settings generally $|\beta| << 1$, so that the influence of the prior will often be minimal. For these priors the posterior is not necessarily logarithmically concave, which might make optimization difficult, but which MCMC methods handle well.

$p(\beta_0, \kappa)$

For the von Mises part of the model we follow the conjugate prior provided by Guttorp & Lockhart (1988), given by

$$p(\beta_0, \kappa \mid \boldsymbol{\delta}, \boldsymbol{\beta}) \propto I_0(\kappa)^{-c} \exp\left[R_0 \kappa \cos(\beta_0 - \mu_0)\right]. \tag{1.9}$$

The prior hyperparameters $\{c, R_0, \mu_0\}$ can be interpreted as the prior sample size c, prior resultant length R_0 and prior mean direction μ_0 respectively of a hypothetical set of angles $\psi = \theta - \boldsymbol{\delta}^T \boldsymbol{d} - g\left(\boldsymbol{\beta}^T \boldsymbol{x}\right)$. Setting informative prior expectations for the parameters of the distribution of a random angle ψ might be difficult, because conditioning on $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ makes ψ hard to interpret.

However, an uninformative prior is easily obtained by setting $c = 0, R_0 = 0$, which is the approach taken here. Note that this does induce an improper (constant) prior on κ.

1.2 MCMC sampling

In this section, details will be discussed for the MCMC sampling procedure, given below as Algorithm 1. Usually, the algorithm converges fast and mixes rapidly, at least for smaller models. The following sections provide further details on sampling $\beta_0, \kappa, \boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

1.2.1 Sampling β_0

Using the likelihood discussed in Section 1.1.1 and the uninformative prior from Section 1.1.2, it can be seen that the conditional posterior distribution of β_0 is $\mathcal{M}(\bar{\psi}, R_\psi \kappa)$. To draw from this distribution, a new vector $\boldsymbol{\psi}$ is computed in each iteration, using the current values of $\{\boldsymbol{\beta}, \boldsymbol{\delta}\}$. Then, the corresponding values of $\bar{\psi}$, and R_ψ are computed. In this case, a Gibbs step can be applied, because it is straightforward to sample from the von Mises distribution, for example as in Best & Fisher (1979).

1.2.2 Sampling κ

Sampling κ is performed by employing a fast rejection sampler described by Forbes & Mardia (2015). This algorithm takes inputs $\{m, \zeta\}$ and returns a new value from the conditional distribution of κ. Here, with an uninformative prior on the von Mises model, $m = n$ and $\zeta = -R_\psi \cos(\beta_0 - \bar{\psi})/n$. For further details, see Forbes & Mardia (2015).

1.2.3 Sampling $\boldsymbol{\beta}$

Sampling $\boldsymbol{\beta}$ is performed by a Metropolis-Hastings step (Metropolis et al., 1953; Hastings, 1970a). However, because of the irregular shape of the posterior, a random walk on β_k may cause slow convergence. If the current value for β_k is further from zero, we might prefer to propose candidates that are further away from the current value.

Therefore, motivated by the circular nature of the parameter space, candidates are generated by

$$\beta_k^* = \frac{\beta_k^{(cur)} + \tan(u\pi/2)}{1 - \beta_k^{(cur)} \tan(u\pi/2)}, \tag{1.10}$$

Algorithm 1 MCMC algorithm for circular GLM

Set $\boldsymbol{\phi}^{(1)} \leftarrow \{\beta_0^{(1)}, \kappa^{(1)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\delta}^{(1)}\}$, which are the given starting values.
for q $= 2, \ldots,$ Q **do**
  $\psi_i \leftarrow \theta_i - \boldsymbol{\delta}^T \boldsymbol{d}_i - g(\boldsymbol{\beta}^T \boldsymbol{x}_i),\ \forall\ i = 1, \ldots, n.$
  $R_\psi \leftarrow \sqrt{\left(\sum_{i=1}^n \cos\psi_i\right)^2 + \left(\sum_{i=1}^n \sin\psi_i\right)^2}.$
  $\bar{\psi} \leftarrow \text{atan2}\left[\sum_{i=1}^n \sin\psi_i, \sum_{i=1}^n \cos\psi_i\right].$
  Sample $\beta_0 \sim \mathcal{M}\left(\bar{\psi}, R_\psi \kappa\right).$
  $\zeta \leftarrow -n^{-1} R_\psi \cos\left(\beta_0 - \bar{\psi}\right).$
  Sample κ with `sampleKappa`(n, ζ) as in Forbes & Mardia (2015).
  for $j = 1, \ldots, J$ **do**
    Sample a candidate $\delta_j^* \sim \mathcal{M}\left(\delta_j, R_\psi \kappa\right).$
    $\alpha_{\delta_j} \leftarrow \log p\left(\delta_j^*, \boldsymbol{\phi}_{(-\delta_j)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}\right) - \log p\left(\delta_j, \boldsymbol{\phi}_{(-\delta_j)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}\right).$
    Sample $u_1 \sim U[0,1].$
    if $\alpha_{\delta_j} > \log u_1$ **then**
      $\delta_j \leftarrow \delta_j^*$
    end if
  end for
  for $k = 1, \ldots K$ **do**
    Sample $u_2 \sim U[-w, w]$ and $u_3 \sim U[0,1].$
    $\beta_k^* \leftarrow \left(\beta_k + \tan(u_2\pi/2)\right) / \left(1 - \beta_k \tan(u_2\pi/2)\right).$
    $\alpha_{\beta_k} \leftarrow \log p\left(\beta_k^*, \boldsymbol{\phi}_{(-\beta_k)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}\right) + \log f_{tc}\left(\beta_k^* \mid w\right) -$
      $\log p\left(\beta_k, \boldsymbol{\phi}_{(-\beta_k)} \,\middle|\, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}\right) - \log f_{tc}\left(\beta_k \mid w\right),$
    where $f_{tc}\left(x \mid w\right) = 1 / \left(w\pi\left[1 + x^2\right]\right).$
    if $\alpha_{\beta_k} > \log u_3$ **then**
      $\beta_k \leftarrow \beta_k^*$
    end if
  end for
  $\boldsymbol{\phi}^{(q)} \leftarrow \boldsymbol{\phi}$
end for

where u is a random variate from the uniform distribution $U(-w, w)$, with w a tuning parameter. Here, we choose $w = .05$. This procedure can be shown to be equivalent to drawing a proposal from the truncated Cauchy distribution $f\left(\beta_k^* \mid \beta_k^{(cur)}, w\right) = 1/\left(w\pi\left[1 + \beta_k^{*2}\right]\right)$ with bounds

$$\left[\frac{\beta_k^{(cur)} + \tan(-\pi w/2)}{1 - \beta_k^{(cur)}\tan(-\pi w/2)}, \frac{\beta_k^{(cur)} + \tan(\pi w/2)}{1 - \beta_k^{(cur)}\tan(\pi w/2)}\right].$$

Note that this proposal is not symmetric, although the Metropolis-Hastings ratio corrects for this lack of symmetry in the usual way.

1.2.4 Sampling $\boldsymbol{\delta}$

It can be seen that the conditional posterior of each δ_j is a convolution of two von Mises distributions, which itself is not von Mises. Mardia & Jupp (1999, p. 44) provide an approximation for such a convolution. Here, a slightly simpler approach is taking employing another Metropolis-Hastings step with von Mises proposals such that

$$\delta_j^* \sim \mathcal{M}\left(\delta_j^{(cur)}, R_\psi^{(cur)}\kappa^{(cur)}\right). \tag{1.11}$$

1.3 Hypothesis tests

In order to make decisions on a researcher's hypotheses, it is useful to consider hypothesis testing. A Bayesian approach to testing two discrete hypotheses against each other is by updating the prior odds of the hypotheses by multiplying them by the Bayes factor (Kass & Raftery, 1995; Jeffreys, 1961), in order to produce the posterior odds of the two hypotheses. From the posterior odds, we can obtain the posterior model probability of H_1 compared to H_0, $p(H_1 \mid D)$, where D represents the data, in this case $\{\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}\}$. This is an intuitive probability of interest, because it represents our current belief that H_1 is true rather than H_0, and this probability might further be used to make decisions.

Two types of hypothesis tests are considered here. First, we consider traditional equality constrained hypotheses, comparing a null hypothesis to an alternative. Second, we consider inequality constrained hypotheses (Hoijtink et al., 2008; Hoijtink, 2011), where we test whether a parameter is larger than some other parameter, which can be a function of the parameters in the model or a fixed value.

1.3.1 Equality constrained hypotheses

Consider two hypotheses about some model parameter γ,

$$H_0 : \gamma = \gamma_0, \quad H_1 : \gamma \in \Omega_\gamma, \tag{1.12}$$

where Ω_γ is the sample space of γ. The associated Bayes factor is given by

$$BF_{01} = \frac{p(D \mid H_0)}{p(D \mid H_1)}. \tag{1.13}$$

To obtain the Bayes factor, we must compute the quantity

$$p(D \mid H_s) = \int p(D, \boldsymbol{\phi}_s \mid H_s) d\boldsymbol{\phi}_s, \tag{1.14}$$

where $\boldsymbol{\phi}_s$ denotes the set of parameters in the model for hypothesis H_s. In general, this integral is not easy to compute, although special cases admit closed-form solutions. Here we will apply the Savage-Dickey method (Dickey et al., 1970; O'Hagan & Forster, 2004), following Wagenmakers et al. (2010). This method is based upon the result that, under some assumptions,

$$\frac{p(D \mid H_0)}{p(D \mid H_1)} = \frac{p(\gamma = \gamma_0 \mid D, H_1)}{p(\gamma = \gamma_0 \mid H_1)}, \tag{1.15}$$

which is a ratio of the posterior and prior probability of γ_0 under model H_1. In practice, this means that the probability of γ_0 under H_1 must be evaluated, both in the prior and the posterior. Although trivial in some conjugate situations, for the circular GLM this needs to be addressed separately for each parameter to which we wish to compute the Bayes factor.

In this study, this type of hypothesis test will be applied to both δ and β. For some β_k, the hypotheses under evaluation are

$$H_0 : \beta_k = 0, \quad H_1 : \beta_k \in \mathbb{R}. \tag{1.16}$$

Taking a Normal prior on β_k as before, $p(\beta_k = 0 \mid H_1) = p_N(0 \mid \mu = 0, \sigma^2 = 1) \approx .399$.

For some δ_j, the hypothesis under evaluation is

$$H_0 : \delta_j = 0, \quad H_1 : \delta_j \in [-\pi, \pi). \tag{1.17}$$

With a uniform prior on δ_j, $p(\delta_j = 0 \mid H_1) = 1/2\pi \approx .159$.

For $p(\beta_k = 0 \mid D, H_1)$ and $p(\delta_j = 0 \mid D, H_1)$, a simple estimate is obtained by computing the height of the histogram bar that would contain 10% of the observations and which would have γ_0 as its midpoint. A more sophisticated

approach could employ log-spline distributions on the real line (Stone et al., 1997) and on the circle (Ferreira, Juárez, Steel, et al., 2008).

We note that if γ_0 is far from the posterior samples of γ, this estimate will not be stable. However, Wagenmakers et al. (2010) note that evidence for H_1 is overwhelming by that point, which means that accuracy is much less important.

Another remark to be made is that this method is only valid if the nuisance parameters between the two hypotheses serve the same purpose. For a discussion, see Consonni et al. (2008).

1.3.2 Inequality constrained hypotheses

In practice, researchers often have directed (one-sided) hypotheses, which may be specified by using inequality constraints. Bayesian analysis of inequality constrained hypotheses has been studied by Klugkist et al. (2005) and Wetzels et al. (2010).

For some model parameter γ, a simple hypothesis to evaluate could be

$$H_0 : \gamma > \gamma_0, \quad H_1 : \gamma < \gamma_0.$$

In order to quantify our belief in these hypotheses, we employ an encompassing hypothesis $H_{unc} : \gamma \in \Omega_\gamma$, from which an MCMC sample $\boldsymbol{\gamma} = \{\gamma^{(1)}, \ldots, \gamma^{(Q)}\}$ is obtained (Klugkist et al., 2005). Then, assuming the encompassing prior does not favor either hypothesis, it can be shown that the Bayes factor for H_0 versus H_1 is given by

$$BF_{01} = \frac{p(D \mid H_0)}{p(D \mid H_1)} = \frac{p(D \mid H_0)/p(D \mid H_{unc})}{p(D \mid H_1)/p(D \mid H_{unc})} = \frac{\sum_{s=1}^{Q} I\left(\gamma^{(s)} \in \Omega_{\gamma|H_0}\right)}{\sum_{s=1}^{Q} I\left(\gamma^{(s)} \in \Omega_{\gamma|H_1}\right)}, \tag{1.18}$$

where $I(\cdot)$ is an indicator function, $\gamma^{(s)}$ is a sample from the unconstrained model H_{unc}, and $\Omega_{\gamma|H_s}$ is the admitted sample space for γ under hypothesis H_s.

This is a flexible approach, because it allows evaluation of any combination of inequality constrained hypotheses against each other. For example, consider a model with three groups, where we denote the mean directions by $\{\mu_1, \mu_2, \mu_3\}$. Then, a major advantage of the inequality constrained hypothesis approach is that it becomes easy to assess the model

$$\mu_1 > \mu_2 > \mu_3 \tag{1.19}$$

1.4 Simulation study

A simulation study was performed to assess the effectiveness of the proposed method. The sampler was implemented in Rcpp (Eddelbuettel & François, 2011), and analyzed in R (R Core Team, 2016). Generally, the method converges fast and mixes well for each cell in the simulation, so that a burn-in of 1000 and a number of iterations of 20000 was deemed sufficient, with no thinning. Three different models are considered. First, a circular regression scenario with a single linear predictor. Second, a 2×2 factorial ANOVA model with main effects only. Third, an ANCOVA model with a single grouping variable and four linear covariates. For all models, the artificial data featured (total) sample sizes $n = \{20, 100\}$, concentrations $\kappa = \{2, 10\}$ and circular intercept $\beta_0 = \pi/2$. Additional simulations were performed with $n = 50$, $\kappa = 5$ and $\beta, \delta = 0.2$, which are not shown here for brevity's sake, as they provided similar results to the other scenarios.

For each scenario, 5000 datasets were generated and subsequently analyzed. Point estimates obtained from the MCMC sampler are $\hat{\beta}_0$, the posterior mean direction of β_0, $\hat{\kappa}$, the posterior mode of κ, $\hat{\beta}$, the posterior mean of β, and $\hat{\delta}$, the posterior mean direction of δ. In addition, credible intervals are obtained from the posterior samples as well, by taking the circular quantiles of β_0, the Highest Posterior Density (HPD) interval of κ, the regular quantiles of $\boldsymbol{\beta}$, and the circular quantiles of $\boldsymbol{\delta}$. Circular quantiles of a set of angles $\boldsymbol{\theta}$ are obtained by computing the set of angles $(\boldsymbol{\theta} + \bar{\theta} - \pi)$, obtaining the linear quantiles, and finally subtracting $(\bar{\theta} - \pi)$ from the computed lower and upper bound.

In order to assess bias, point estimates were averaged over the datasets. For β_0, and $\boldsymbol{\delta}$, this means computation of the mean direction, for other values this refers to the regular linear mean. In addition, a coverage was obtained for each parameter by computing the proportion of the appropriate credible intervals that contained the true value. Finally, acceptance probabilities refer to the proportion of proposals for this parameter that were accepted by the Metropolis-Hastings step, which is applicable only for δ and β.

The three different scenarios for the predictors will be discussed separately in the following sections. The section will be concluded with a discussion of the behavior of the Bayes factors for these three scenarios.

1.4.1 Simple regression

For the simple regression data were simulated by first generating a vector $\boldsymbol{x}_r$ independently from the standard normal distribution $N(0, 1)$. This vector is then standardized by $\boldsymbol{x} = (\boldsymbol{x}_r - \bar{\boldsymbol{x}}_r)/\text{var}(\boldsymbol{x}_r)$. Then, the circular outcome is

Table 1.1: Results of the simulation study for the simple regression scenario. 'Cov.' denotes the 95% coverage for a specific parameter, while 'Acc.' denotes the acceptance probability. MCT denotes the mean computation time in seconds.

True			β_0		κ		β_1			
β	κ	n	Bias	Cov.	$\hat{\kappa}$	Cov.	$\hat{\beta}_1$	Cov.	Acc.	MCT
0.05	2	20	-0.00	0.95	2.09	0.97	0.05	0.94	0.84	0.58
		100	-0.00	0.95	1.99	0.98	0.05	0.95	0.66	2.26
	20	20	-0.00	0.94	22.70	0.95	0.05	0.94	0.47	0.59
		100	-0.00	0.95	20.36	0.95	0.05	0.94	0.23	2.37
0.80	2	20	-0.00	0.94	2.08	0.97	0.81	0.95	0.86	0.58
		100	0.00	0.96	2.00	0.98	0.81	0.95	0.72	2.33
	20	20	-0.00	0.94	22.37	0.95	0.80	0.95	0.55	0.60
		100	0.00	0.95	20.37	0.95	0.80	0.94	0.29	2.44

computed by

$$\theta_i = \pi/2 + g\left(\beta_1 x_i\right) + \varepsilon_i, \tag{1.20}$$

where β takes values $\{0.05, 0.8\}$, and $\varepsilon_i \sim \mathcal{M}(0, \kappa)$.

Table 1.1 shows the results of this simulation study. Performance is quite good, providing unbiased estimates and great coverage, which shows that the weakly informative prior on β_1 does not strongly harm the frequency properties of our Bayesian estimation procedure. Acceptance rates decline slightly for more concentrated data, which could be ameliorated in practice by tuning the proposals. The sampler runs quite fast, as the longest time a single analysis took was 2.44 seconds.

1.4.2 Factorial ANOVA

For the factorial ANOVA scenario, data were simulated by first generating two vectors, $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$, by randomly drawing either 0 or 1 with probability .5. This means that we generally create unbalanced designs, and the number of subjects in each group differs between the simulated datasets. This reflects a realistically broad range of possible research designs. The outcome is computed by

$$\theta_i = \pi/2 + \delta_1 d_{1i} + \delta_2 d_{2i} + \varepsilon_i, \tag{1.21}$$

where $\delta_1 = \delta_2$ takes values $\{0.05, 0.8\}$, and $\varepsilon_i \sim \mathcal{M}(0, \kappa)$.

Table 1.2: Results of the simulation study for the factorial ANOVA scenario. 'Cov.' denotes the 95% coverage for a specific parameter, while 'Acc.' denotes the acceptance probability. MCT denotes the mean computation time in seconds.

True			β_0		κ		δ_1			
δ	κ	n	Bias	Cov.	$\hat{\kappa}$	Cov.	$\hat{\delta}_1$	Cov.	Acc.	MCT
0.05	2	20	-0.00	0.95	2.10	0.97	0.06	0.95	0.78	0.70
		100	-0.00	0.95	1.99	0.98	0.05	0.95	0.78	2.63
	20	20	-0.00	0.93	22.75	0.95	0.05	0.93	0.79	0.71
		100	0.00	0.95	20.35	0.95	0.05	0.94	0.78	2.71
0.80	2	20	-0.00	0.95	2.10	0.97	0.80	0.96	0.79	0.69
		100	0.00	0.95	2.00	0.98	0.80	0.95	0.78	2.62
	20	20	-0.00	0.94	22.65	0.95	0.80	0.93	0.78	0.71
		100	-0.00	0.95	20.35	0.95	0.80	0.95	0.78	2.71

Table 1.2 shows the results of the simulation study. Once again, the sampler performs well in all situations. In contrast with β in the regression model, the acceptance rate of δ does not depend on n, κ, or true δ, because its proposal adapts itself to these parameters.

1.4.3 ANCOVA with four covariates

Here, a single grouping variable was generated as in Section 1.4.2, while the linear covariates were generated as in Section 1.4.1. Then, the outcome is computed by

$$\theta_i = \pi/2 + \delta_1 d_{1i} + g(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}) + \varepsilon_i, \tag{1.22}$$

where we have chosen the true values $\delta_1 = \beta_1 = \beta_2 = \beta_3 = \beta_4$ in all simulations, taking values $\{0.05, 0.8\}$, and $\varepsilon_i \sim \mathcal{M}(0, \kappa)$.

Table 1.3 shows the results of the simulation study. Only the results for β_1 and δ_1 are shown, because results the other regression parameters are almost identical. For most scenarios, the results are once again adequate. It can be seen that the scenario with strong effects ($\beta = \delta = 0.8$), estimates are often unsatisfactory. It should be noted that in these scenarios, the data show high variance. In general, if $\sum_{k=1}^{K} |\beta_k| > 1.5$, the data are spread all over the circle, such that many estimates of $\boldsymbol{\beta}$ are somewhat plausible, which results in an irregular posterior. Therefore, the sampler performs quite badly. Note that

Table 1.3: Results of the simulation study for the ANCOVA scenario. 'Cov.' denotes the 95% coverage for a specific parameter, while 'Acc.' denotes the acceptance probability. MCT denotes the mean computation time in seconds.

True			β_0		κ		δ_1			β_1			
β, δ	κ	n	Bias	Cov.	$\hat{\kappa}$	Cov.	$\hat{\delta}_1$	Cov.	Acc.	$\hat{\beta}_1$	Cov.	Acc.	MCT
0.05	2	20	0.01	0.95	1.96	0.94	0.04	0.96	0.79	0.01	0.96	0.85	1.24
		100	-0.00	0.95	1.99	0.98	0.05	0.95	0.78	0.05	0.95	0.67	4.11
	20	20	0.00	0.93	22.88	0.96	0.05	0.94	0.79	0.05	0.94	0.47	1.21
		100	0.00	0.95	20.36	0.95	0.05	0.95	0.78	0.05	0.95	0.23	4.11
0.80	2	20	-0.01	0.95	1.95	0.94	0.73	0.96	0.79	0.58	0.95	0.86	1.24
		100	0.00	0.87	1.78	0.88	0.79	0.82	0.78	0.60	0.88	0.73	4.32
	20	20	0.00	0.93	17.43	0.92	0.80	0.74	0.79	0.69	0.93	0.56	1.25
		100	-0.00	0.70	13.18	0.69	0.80	0.82	0.78	0.40	0.70	0.46	4.31

this is a property of the GLM approach to circular regression, rather than this specific model or implementation. In practice, however, we expect that this kind of dataset will almost never occur, although it might be advisable to monitor obtained estimates for this situation.

1.4.4 Bayes factors and posterior model probabilities

For hypothesis testing, Bayes factors and posterior model probabilities were obtained as detailed in Section 1.3. Figure 1.3 depicts boxplots of the posterior probability of the correct model for the inequality and equality hypotheses, for the three different models, where in each case all true $\beta = \delta = 0.05$. This means that in all cases H_1 is true, so values close to 1 indicate that in a given scenario greatly prefers the correct model. The hypotheses given are listed below.

- Regression
  - Test for equality: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$,
  - Test for inequality: $H_0 : \beta_1 < 0$ vs. $H_1 : \beta_1 > 0$,
- ANOVA and ANCOVA
  - Test for equality: $H_0 : \delta_1 = 0$ vs. $H_1 : \delta_1 \neq 0$,
  - Test for inequality: $H_0 : \delta_1 < 0$ vs. $H_1 : \delta_1 > 0$,

Generally, the correct hypothesis becomes favored as the sample size increases, as expected. In addition, there is less simulation variability when

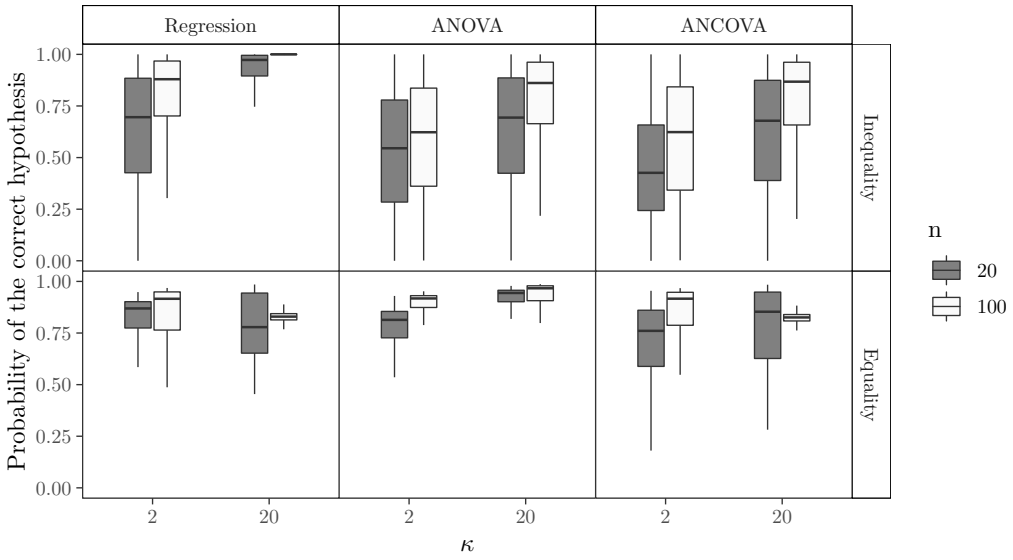


Figure 1.3: Boxplot of the posterior model probability assigned to the correct model in 5000 simulations with $\beta = \delta = 0.05$ for the inequality (top) and equality (bottom) hypotheses. For the regression scenario, inequality tests $\beta_1 > 0$ vs. $\beta_1 < 0$ and equality tests $\beta_1 \neq 0$ vs. $\beta_1 = 0$. For the ANOVA and ANCOVA models inequality tests $\delta_1 > 0$ vs. $\delta_1 < 0$ and equality tests $\delta_1 \neq 0$ vs. $\delta_1 = 0$.

n increases, shown by a smaller range in the boxplot. Compared to the inequality hypothesis, the equality hypothesis is more prone to pick up group differences in δ (ANOVA and ANCOVA model), as well as in the regression model when $\kappa = 2$.

1.5 Example

In this section, our method will be applied to data from van Dijk et al. (2013). In this study, an experiment was conducted to assess whether deafness enhances haptic perception. Haptic perception is assessed here by means of a haptic parallel setting task, where subjects are required to set two bars parallel, so that errors are measured in an angular differece to the reference direction. In this task, errors generally fall in the counterclockwise direction, which produces a positive score on the deviation from the target. Therefore, groups that are less apt at this task are expected to have stronger positive deviations, counterclockwise from the reference direction.

Three groups are distinguished: deaf subjects, sign language interpreters

Table 1.4: Summary statistics of the mean age (years), mean education (years) and mean direction of deviation (degrees).

	Age		Education		
	Mean	SD	Mean	SD	Deviation
Deaf	41.20	13.48	16.60	1.55	14.73
Interpreter	38.44	8.60	16.88	1.50	25.22
Control	44.75	9.73	17.06	1.44	28.82

and a control group. Table 1.4 shows some summary statistics of background variables for the three groups, as well as the main outcome, deviation. Note that the original study examines two different conditions in a repeated measures design, an "immediate" and a "delayed" condition, of which we only show the "immediate" outcome for illustration purposes. Therefore, the data under consideration are made independent.

The analysis will proceed as follows. First, a basic ANOVA model will be fitted to this dataset. Then, an ANCOVA model is examined. Lastly, informative inequality constrained hypotheses based on theory are evaluated.

1.5.1 ANOVA model

The goal in the ANOVA model is to assess whether the three groups differ. For this and all following models, the control group will be used as the reference group. The outcome in the ANOVA model is given by $\theta_i \sim \mathcal{M}(\mu_i, \kappa)$, where $\mu_i = \beta_0 + \delta_{df} d_{df} + \delta_{in} d_{in}$, where d_{df} and d_{in} are dummy variables for the deaf and interpreter group, respectively, and each δ is labeled appropriately. The main hypothesis is that the deaf group performs better than the control group, which could manifest itself as $\delta_{df} < 0$. Second, interpreters might also outperform the control group, which would mean $\delta_{in} < 0$.

A burn-in of 1000 was used, and the MCMC sampler was run for 100000 iterations. Figure 1.4 shows convergence plots for the four model parameters, where it can be seen that the sampler converges well.

Results are shown in Table 1.5. Estimates are given by the posterior mean direction of $\beta_0, \delta_{df}, \delta_{in}$ and the posterior mode of κ. The credible interval of δ_{df}, the difference between the deaf group mean direction and the control group mean direction, is given by (-0.42, -0.08), which can be seen as evidence for a non-zero group difference, as zero is not in this interval. The credible interval for the interpreter group mean direction is (-0.24, 0.11), which can be seen as evidence against a non-zero group difference.

Table 1.5: Results for the ANOVA model. LB and UB respectively represent lower and upper bound of the 95% credible interval of the given parameter.

	Estimate	LB	UB
β_0	0.51	0.38	0.63
κ	17.77	11.26	26.16
δ_{df}	-0.25	-0.42	-0.08
δ_{in}	-0.07	-0.24	0.11

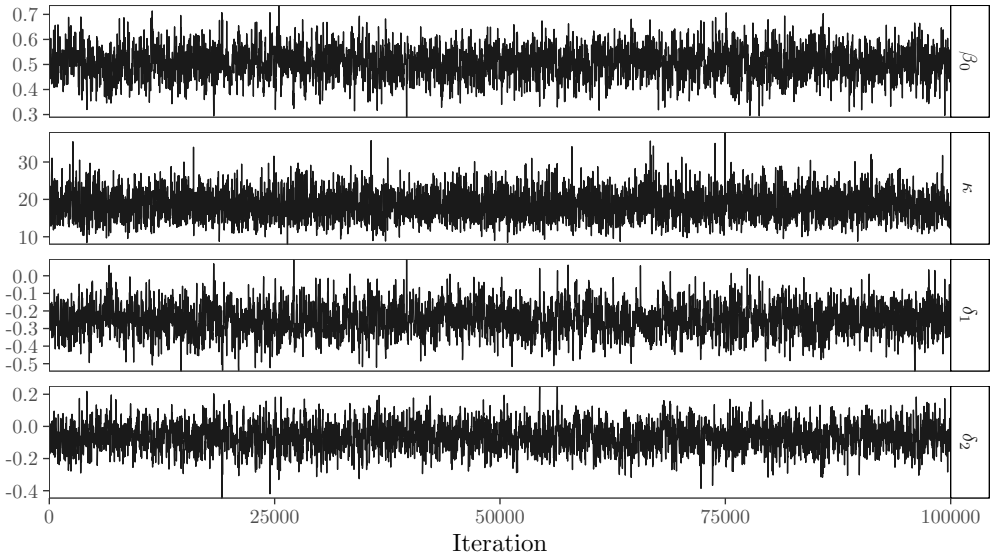


Figure 1.4: Convergence plots for the ANOVA model.

Table 1.6: Results for the ANCOVA model.

	Estimate	LB	UB
β_0	0.50	0.38	0.62
κ	18.30	11.54	27.19
δ_{df}	-0.21	-0.39	-0.04
δ_{in}	-0.08	-0.24	0.09
β_{age}	-0.02	-0.06	0.01
β_{hand}	0.03	-0.01	0.06

A more sophisticated approach is to employ the hypothesis tests that were developed. Employing the equality constrained hypotheses, a mild amount of support was found in favor of the hypothesis that deaf participants differ from the controls ($BF_{\mu_{cn}\neq\mu_{df}:\mu_{cn}=\mu_{df}} = 2.69$), while a mild amount of support was found against the hypothesis that deaf participants differ from sign language interpreters ($BF_{\mu_{in}\neq\mu_{df}:\mu_{in}=\mu_{df}} = 0.34$).

This highlights that this method is conservative in supporting alternative hypotheses such as $H_1 : \mu_{in} \neq \mu_{df}$. This is a result of the circular uniform prior on δ, which suggests that more subjective approach could be less conservative and more likely to pick up on group differences. For example, a von Mises prior on δ with $\kappa > 0$ could be used. This subjective prior represents the knowledge that mean directions of different groups are usually somewhat close together. In this case, selection of the κ to be used becomes a core issue, which represents a trade-off between the amount additional power for the hypothesis test and the amount of (potentially unwanted) prior information included in the analysis.

The inequality hypothesis tests show a large amount of support for the hypothesis that deaf participants perform better than the controls (where we have $BF_{\mu_{cn}>\mu_{df}:\mu_{cn}<\mu_{df}} = 267.82$), and for the hypothesis that deaf participants perform better than sign language interpreters ($BF_{\mu_{in}>\mu_{df}:\mu_{in}<\mu_{df}} = 52.28$).

1.5.2 ANCOVA model

In order to properly assess the effects found, it is useful to take into account theoretically relevant covariates. Here, we include age and handedness (that is, strength of hand preference) as covariates, and investigate the effect this has on previous conclusions. Table 1.6 shows the output of the ANCOVA model, with age and handedness as linear covariates.

From the results, it seems that the covariates do not have an effect on

performance. Under equal prior odds, the Bayes factors for the predictors indeed indicate 29.35 times more support for the hypothesis $\beta_{age} = 0$ compared to the alternative $\beta_{age} \neq 0$, and 19.08 times more support for the hypothesis $\beta_{hand} = 0$ than the alternative $\beta_{hand} \neq 0$. While controlling for covariates, the evidence for a negative δ_{df} is still substantive, indicating superior performance of deaf individuals over controls ($BF_{\mu_{df}<\mu_{cn} \,:\, \mu_{df}>\mu_{cn}} = 96.75$). As there is little support for the inclusion of covariates, they are omitted in subsequent analyses.

1.5.3 Inequality constrained hypothesis

Although van Dijk et al. (2013) do not evaluate inequality constrained hypotheses directly, the theories stated by the authors can be interpreted as such. First, they state:

> "On the basis of a greater proneness in visuospatial processing, we could expect a better developed haptic orientation processing ability in deaf individuals."

Then, with regards to sign language interpreters, they state:

> "The relative positions of the signer's hands are used to map spatial relations in the real world. We may speculate here that experienced signers can also do the reverse more easily: interpret the hand positions forced by inspecting the bars in the parallel setting task in absolute world reference frames. If so we would expect both deaf and hearing signers to outperform non-signing hearing controls but not to differ from each other."

These expectations can be mapped to hypotheses about the mean directions of the three groups, which are given by

$$\text{(Deaf) } \mu_{df} = \beta_0, \quad \text{(Interpreter) } \mu_{in} = \beta_0 + \delta_1, \quad \text{(Control) } \mu_{cn} = \beta_0 + \delta_2.$$

Following for example Rueda et al. (2009) and Baayen & Klugkist (2014), it is important to specify inequality constraints on circular data as either isotropic, or non-isotropic. Isotropic orderings are defined on the circle, and denote in which order the parameters are encountered as we move around the circle, relative to one another. Non-isotropic orderings are orderings relative to a fixed point on the circle. In this case, a type of non-isotropic orderings are considered where the hypothesis states that one parameter lies

in the semi-circle counterclockwise of another parameter. In this case, we have chosen to translate the expectations to the following hypotheses:

$$H_1 : \mu_{df} < (\mu_{in}, \mu_{cn}) < \mu_{df} + \pi, \quad H_2 : \mu_{cn} - \pi < (\mu_{df}, \mu_{in}) < \mu_{cn}. \quad (1.23)$$

Using the inequality constrained framework as described in Section 1.3.2, the support for either hypothesis can be quantified.

Following this method, we find that H_1 is true in 97.9% of the MCMC iterations, while H_2 occurs in 78% of the iterations, so both hypotheses are likely. From this, we find no conclusive evidence for either H_1 or H_2 ($BF_{H_1:H_2} = 1.25$). This means that although the study provides useful insight in the performance of deaf subjects, there is not enough evidence yet to decide on these two competing hypotheses.

1.6 Discussion

We developed a Bayesian circular GLM, with appropriate priors, proper treatment of dichotomous variables, and Bayesian hypothesis tests. Our method forms a middle ground between two veins of research into Bayesian analysis of circular data. On one hand, analysis of complex data shapes (Ghosh et al., 2003a; Ferreira, Juárez, Steel, et al., 2008; Fernández-Durán & Mercedes Gregorio-Domínguez, 2016a) provides modeling for a broad class of datasets but few possibilities for prediction and covariate models. On the other hand, circular regression models (Fisher & Lee, 1992; J. Gill & Hangartner, 2010; Lagona, 2016) provide prediction and covariates, but encounter problems with likelihood shapes and a lack of available hypothesis tests. The GLM aproach is promising because of its flexibility to allow for many different kinds of models, while also allowing straightforward extensions. Our method brings three main contributions to the literature.

First, we have shown that the irregular log-likelihood surface of the regression parameters $\boldsymbol{\beta}$ in a circular GLM can be dealt with naturally by employing a weakly informative prior that encapsulates our actual belief that extreme values for $\boldsymbol{\beta}$ are unlikely in applied research, while we still let the data overpower the prior. This is analogous to the widely accepted idea that very large effect sizes, say, Cohen's $d > 1$, are improbable in most scientific disciplines where empirical research is necessary, in particular the social sciences. If the method would indicate support for such large values of $|\beta|$, a researcher would not believe that the model is correct, and reassess it. Our prior simply represents the lack of belief in large values of $|\beta|$.

Second, we have separated the group difference parameters from linear covariates, in order to allow modeling of a large array of ANOVA designs,

including factorial and ANCOVA designs. This provides researchers with a straightforward way to map their hypotheses to a design. In addition, model comparison can easily be made possible through the DIC or WAIC (Gelman et al., 2003, Ch. 7).

Third, we have developed Bayesian hypothesis tests based on the Bayes factor for the circular data case. The tests employed here are pbased on the Savage-Dickey method advocated by Wagenmakers et al. (2010) and the inequality constrained aproach of Hoijtink (2011). Many Bayesian approaches to circular data analysis lack any form of hypothesis testing, which we view as limiting their ability to be applied in practice. In order to create statistical methods that are employed in practice, we must accomodate the desire for hypothesis testing, and compute posterior model probablities. Therefore, we have taken a step in this direction as well, showing how Bayesian hypothesis tests can be developed easily in the circular data context by using MCMC output.

Although the computational methods employed here are stable and allow for useful inferences, further consideration of useful hypotheses and their associated Bayes factors will be important for the applicability of the Bayesian paradigm to circular data analysis, in particular in behavioural research. Here, we have not provided Bayes factors based on estimation of the marginal likelihood as in Chib (1995) and subsequent work in this field, although this approach might be more flexible than the methods applied here. Another approach might be to attempt to develop priors that allow for closed-form Bayes factors in a similar vein as the g-prior in the linear case (Zellner, 1986; Liang et al., 2012). The computational simplicity of such Bayes factors is useful in many scenarios, although the complexity of the designs for such Bayes factors are usually limited.

In the broader scope of circular data analysis, our method can be seen as a Bayesian extension of the approaches of Artes (2008) and Lagona (2016). Further extensions of this model might ease the assumption of i.i.d. observations taken here by applying properties of the multivariate von Mises distribution (Mardia et al., 2008; Mardia & Voss, 2014), as in Lagona (2016).

In sum, the Bayesian approach provides a promising way to draw inference from circular data. Usual approaches are based on large sample or high concentration approximations (Artes, 2008) or bootstrap approaches for simple models (Baayen et al., 2012; Baayen & Klugkist, 2014). Our approach does not need such approximations, and provides a new direction for circular data analysis of GLM-type models.

1.7 Acknowledgements

This work was supported by a Vidi grant awarded to I. Klugkist from NWO, the Dutch Organization for Scientific Research (NWO 452-12-010).

The authors are grateful to two reviewers for helpful comments.

The authors are grateful to A. Postma for providing the illustrative data.

User-friendly code for the main analyses of the paper can be found in a GitHub package at https://github.com/keesmulder/circglmbayes.

All code for both the statistical tools, the simulation study and the paper is available online at
https://github.com/keesmulder/BayesMultCircCovariates.

1.A Conditional distribution of β_0

Here, it will be shown that β_0 conditionally has the von Mises distribution, that is

$$L(\beta_0 \mid \kappa, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}) \propto \mathcal{VM}(\beta_0 \mid \bar{\psi}, R_\psi \kappa). \tag{1.24}$$

The proof for the conditional distribution of β_0 in the GLM closely follows the derivation for the distribution of the mean direction μ of the von Mises distribution, which shows that $L(\mu \mid \kappa, \boldsymbol{\theta}) \propto \mathcal{VM}(\mu \mid \bar{\theta}, R_\theta \kappa)$.

The conditional likelihood of β_0 is given by

$$\begin{aligned} L(\beta_0 \mid \kappa, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}) &\propto \exp\left\{ \kappa \sum_{i=1}^{n} \cos\left[\theta_i - (\beta_0 + \boldsymbol{\delta}^T \boldsymbol{d}_i + g(\boldsymbol{\beta}^T \boldsymbol{x}_i))\right] \right\} \\ &= \exp\left\{ \kappa \sum_{i=1}^{n} \cos\left[\beta_0 - (\theta_i - \boldsymbol{\delta}^T \boldsymbol{d}_i - g(\boldsymbol{\beta}^T \boldsymbol{x}_i))\right] \right\}. \end{aligned}$$

We know that for any angle $\psi_i, i = 1, \ldots, n$,

$$C_\psi = \sum_{i=1}^{n} \cos(\psi_i), \quad S_\psi = \sum_{i=1}^{n} \sin(\psi_i), \quad R_\psi = \sqrt{C_\psi^2 + S_\psi^2},$$

$$\text{and} \quad \frac{C_\psi}{R_\psi} = \cos\bar{\psi}, \quad \frac{S_\psi}{R_\psi} = \sin\bar{\psi}, \quad \text{where} \quad \bar{\psi} = \operatorname{atan2}(S_\psi, C_\psi).$$

Thus, setting angle $\psi_i = \theta_i - \boldsymbol{\delta}^T \boldsymbol{d}_i - g(\boldsymbol{\beta}^T \boldsymbol{x}_i)$,

$$\begin{aligned} L(\beta_0 \mid \kappa, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{d}) &\propto \exp\left\{ \kappa \sum_{i=1}^{n} \cos(\beta_0 - \psi_i) \right\} \\ &= \exp\left\{ \kappa \left[\cos\beta_0 \sum_{i=1}^{n} \cos\psi_i + \sin\beta_0 \sum_{i=1}^{n} \sin\psi_i \right] \right\} \\ &= \exp\left\{ R_\psi \kappa \left[\cos\beta_0 \frac{C_\psi}{R_\psi} + \sin\beta_0 \frac{S_\psi}{R_\psi} \right] \right\} \\ &= \exp\left\{ R_\psi \kappa \left[\cos\beta_0 \cos\bar{\psi} + \sin\beta_0 \sin\bar{\psi} \right] \right\} \\ &= \exp\left\{ R_\psi \kappa \cos\left(\beta_0 - \bar{\psi}\right) \right\} \\ &\propto \mathcal{VM}(\beta_0 \mid \bar{\psi}, R_\psi \kappa). \end{aligned}$$

Chapter 2

Bayesian Tests for Circular Uniformity

Abstract

Circular data are data measured in angles or directions, which occur in a wide variety of scientific fields. An often investigated hypothesis is that of circular uniformity, or isotropy. Frequentist methods for assessing the circular uniformity null hypothesis exist, but do not allow the user faced with an insignificant result to distinguish lack of power and support for the null hypothesis. Bayesian hypothesis tests, which solve this issue and several others, are developed here. They are easy to compute and perform well, which is shown in a simulation. Two alternative hypotheses are considered. One is based on the von Mises distribution and performs well against unimodal alternatives. Another is based on a kernel density, which acts as an omnibus test against all other scenarios. Assessing the performance of the tests using different priors, it is shown that they are powerful and allow more elaborate conclusions than classical tests of circular uniformity.

This chapter is submitted to Journal of Statistical Planning and Inference as Mulder, K., & Klugkist, I. (tbd) Bayesian Tests for Circular Uniformity.

Author contributions: KM and IK designed the study. KM developed the statistical methods, simulation study and software package. KM wrote the paper. IK provided feedback on written work.

Circular data are measured in angles or directions. They are frequently encountered in scientific fields as diverse as life sciences (Mardia, 2011), behavioural biology (Bulbert et al., 2015), cognitive psychology (Kaas & Van Mier, 2006), bioinformatics (Mardia et al., 2008), political sciences (J. Gill & Hangartner, 2010) and environmental sciences (Arnold & SenGupta, 2006). In psychology, circular data occur often in motor behaviour research (Mechsner et al., 2001a, 2007; Postma et al., 2008; Baayen et al., 2012), as well as in the application of circumplex models (Gurtman & Pincus, 2003; Gurtman, 2009a; Leary, 1957). Circular data differ from linear data in the sense that circular data are measured in a periodical sample space. For example, an angle of 1° is quite close to an angle 359°, although linear intuition suggests otherwise.

A fundamental hypothesis of interest is that of circular uniformity. A test for circular uniformity can be used to assess a hypothesis of theoretical interest by itself, but can also be used as a preliminary assessment, because most tests performed in circular statistics are only valid if the data is non-uniform. Several methods for assessing circular uniformity exist in the frequentist framework. These will be reviewed in Section 2.1.

In the rest of this paper, Bayesian hypothesis tests will be added to this arsenal. In order to create a Bayesian test of circular uniformity, the Bayes factor will be employed, which is often hailed as the standard way of performing a Bayesian hypothesis test (Kass & Raftery, 1995; Jeffreys, 1961). A major advantage of this Bayesian method is that through specifying the alternative hypothesis and the associated prior, we can precisely quantify support for either the null hypothesis or the alternative hypothesis. Methods based on null hypothesis significance testing only signify whether or not the null hypothesis can be rejected, but never provide support in favor of the null hypothesis. In practice, failure to reject the null hypothesis in a frequentist test is often taken as evidence for the null. However, a failure to reject the null might just as well be caused by a lack of power, so that the evidence in the data is indifferent to circular uniformity. In contrast, the Bayesian hypothesis test developed here is able to provide support for the null hypothesis. This alleviates some of the well-described issues with null hypothesis significance testing, and is particularly useful for tests that are used as a preliminary assessment, such as circular uniformity tests.

To compute the Bayes factor, the so-called marginal likelihood must be obtained for each hypothesis. The marginal likelihood is the normalizing constant of the posterior, and the key ingredient of the Bayes factor. To obtain the marginal likelihood one must specify the prior distribution of the parameters in each hypothesis. The null hypothesis (circular uniformity) has no parameters, so no prior is needed for it. The alternative hypothesis,

however, requires specification of both a model for the data, would they be non-uniform, and second, a prior for the parameters in this model. This paper will investigate two alternative hypotheses, one based on the von Mises distribution and one based on a kernel density alternative. In addition, several priors will be investigated that can be used with each model.

The rest of the paper is structured as follows. A short review of frequentist tests of circular uniformity is provided in Section 2.1. The Bayesian circular uniformity test for a von Mises alternative is discussed in Section 2.2. The Bayesian circular uniformity test for a kernel density alternative, which functions as an omnibus test, is discussed in Section 2.3. The methods are applied to example datasets in Section 2.4. Section 2.5 provides a discussion.

2.1 Frequentist tests of circular uniformity

Here, we will shortly review frequentist tests of circular uniformity. Four commonly used tests are Kuiper's test (Kuiper, 1960), Rayleigh's test (Mardia & Jupp, 2000; Brazier, 1994), Rao's test of equal spacing (Rao, 1976) and Ajne's test (Ajne, 1968). Perhaps the most common of these is the Rayleigh test. Let $\boldsymbol{\theta} = (\theta_i = 1, \ldots, \theta_n)$ denote a set of data consisting of angles, and let the mean resultant length $\bar{R} = n^{-1}\sqrt{(\sum_{i=1}^{n} \cos\theta_i)^2 + (\sum_{i=1}^{n} \sin\theta_i)^2}$. Then the Rayleigh test statistic can be computed simply as $2n\bar{R}^2$, which has approximately a χ^2_2 distribution. It can be shown that the Rayleigh test is the most powerful test against von Mises alternatives, as well as Projected Normal (PN) alternatives (Bhattacharyya & Johnson, 1969). Although the Rayleigh test is consistent against unimodal alternatives, it is not consistent against alternatives that have resultant length $\rho = 0$, in particular distributions with antipodal symmetry (Mardia & Jupp, 2000).

Another test is Kuiper's test (Kuiper, 1960), which is based on the maximum difference between the theoretical and empirical distribution function. It is consistent against all alternatives to uniformity (Mardia & Jupp, 2000). A similar test uses Watson's U^2 statistic (Watson, 1961), which is instead based on the *mean* difference between the theoretical and empirical distribution function.

Several other tests for circular uniformity exist, among which Rao's equal spacing test (Rao, 1976), the range test (Laubscher & Rudolph, 1968), the Hodges-Ajne test (Hodges, 1955; Ajne, 1968), Ajne's A_n test (Ajne, 1968), and the Hermans-Rasson test (Hermans & Rasson, 1985). Somewhat more recently, a smooth test for circular uniformity was developed by Bogdan et al. (2002). A test specifically targeting multimodal alternatives was developed by Pycke (2010).

2.2 Tests with a von Mises alternative

In this section, a Bayesian hypothesis test for circular uniformity against a von Mises alternative will be developed. The von Mises distribution is a natural distribution on the circle, given by

$$\mathcal{M}(\theta \mid \mu, \kappa) = [2\pi I_0(\kappa)]^{-1} \exp\{\kappa \cos(\theta - \mu)\}, \tag{2.1}$$

where $\theta \in [0, 2\pi)$ is an angle, $\mu \in [0, 2\pi)$ is the mean direction, $\kappa \in \mathbb{R}^+$ is a concentration parameter and $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero.

The test will be based on the Bayes factor, which is the ratio of two marginal likelihoods, given by

$$BF_{10} = \frac{m_1(\boldsymbol{\theta})}{m_0(\boldsymbol{\theta})} = \frac{\int_{\boldsymbol{\phi}} p(\boldsymbol{\phi}, \boldsymbol{\theta} \mid H_1) d\boldsymbol{\phi}}{\int_{\boldsymbol{\phi}} p(\boldsymbol{\phi}, \boldsymbol{\theta} \mid H_0) d\boldsymbol{\phi}}, \tag{2.2}$$

where $\boldsymbol{\theta} = \theta_1, \ldots, \theta_n$ is a data set consisting of angles, and $\boldsymbol{\phi}$ is a vector of parameters belonging to the chosen model. For the von Mises distribution $\boldsymbol{\phi} = (\mu, \kappa)^T$. Because the null hypothesis does not feature parameters and assigns equal probability to each data point, $m_0(\boldsymbol{\theta})$ depends only on the sample size. The circular uniform distribution has $p(\theta) = (2\pi)^{-1} \forall\, \theta \in [0, 2\pi)$, so the marginal likelihood for H_0 is obtained by

$$m_0(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\theta_i) = (2\pi)^{-n}. \tag{2.3}$$

The marginal likelihood of H_1 is given by

$$m_1(\boldsymbol{\theta}) = \int_{\boldsymbol{\phi}} f(\boldsymbol{\phi}, \boldsymbol{\theta} \mid H_1) d\boldsymbol{\phi} = \int_0^{\infty} \int_0^{2\pi} f(\mu, \kappa, \boldsymbol{\theta} \mid H_1) d\mu d\kappa, \tag{2.4}$$

where

$$f(\mu, \kappa, \boldsymbol{\theta} \mid H_1) \propto p(\mu, \kappa \mid H_1) f(\boldsymbol{\theta} \mid \mu, \kappa, H_1) \tag{2.5}$$

is the kernel of the posterior, where the prior $p(\mu, \kappa \mid H_1)$ must still be chosen, and $f(\mu, \kappa \mid \boldsymbol{\theta}, H_1)$ is the likelihood. The likelihood of the von Mises distribution is given by

$$f(\boldsymbol{\theta} \mid \mu, \kappa, H_1) = \prod_{i=1}^{n} \mathcal{M}(\theta_i \mid \mu, \kappa) = [2\pi I_0(\kappa)]^{-n} \exp\left\{R\kappa \cos(\bar{\theta} - \mu)\right\}, \tag{2.6}$$

where R is the resultant length and $\bar{\theta}$ is the mean direction.

For any prior $p(\mu, \kappa)$ that does not depend on μ, the Bayes factor simplifies to

$$BF_{10} = (2\pi)^n \int_0^\infty p(\mu, \kappa) \int_0^{2\pi} [2\pi I_0(\kappa)]^{-n} \exp\left\{R\kappa \cos(\bar{\theta} - \mu)\right\} d\mu d\kappa \quad (2.7)$$

$$= \int_0^\infty I_0(\kappa)^{-n} p(\mu, \kappa) \int_0^{2\pi} \exp\left\{R\kappa \cos(\bar{\theta} - \mu)\right\} d\mu d\kappa \quad (2.8)$$

$$= 2\pi \int_0^\infty p(\mu, \kappa) \frac{I_0(R\kappa)}{I_0(\kappa)^n} d\kappa, \quad (2.9)$$

where the last step uses the fact that $I_0(x) = [2\pi]^{-1} \int_0^{2\pi} \exp\{x \cos\theta\} d\theta$. Thus, computation of the Bayes factor requires only univariate integration.

2.2.1 Choosing priors

Choosing the prior for this hypothesis test is not trivial. In principle, the prior for $\{\mu, \kappa\}$ should capture our actual belief about the possible values of the parameters, given that the alternative hypothesis is true. Although researchers are free to determine their own prior for this test, we propose some general guidelines for the set of possible priors to be considered here.

First, it should be noted that choosing improper priors generally do not result in useful Bayesian hypothesis tests. Therefore, proper priors will be used.

Second, if a test for circular uniformity is considered, the researcher will generally not already have an idea about the mean direction of the data if H_1 is true, because they are investigating whether there even is a preferred (mean) direction. Therefore, we suggest taking a circular uniform prior on μ. This is done by taking $p(\mu) = [2\pi]^{-1}$ and independent of κ, so that $p(\mu, \kappa) = p(\mu)p(\kappa) = p(\kappa)/(2\pi)$ and we can concern ourselves only with choosing the prior for κ.

Finally, a researcher that considers circular uniformity to be a reasonable hypothesis rarely expects strongly concentrated distributions, even if the alternative hypothesis were true. Therefore, we suggest setting a prior for κ that gives most of its probability to fairly low values of κ. Should the data follow a concentrated distribution anyway, the test will be powerful regardless.

In practice, whether these expectations are reasonable should be assessed by the researcher themselves. However, taking this approach allows us to build default methods that work well in most research scenarios in which the test would be applied. In the following sections, different choices for priors are considered, and for each the resulting test is assessed.

2.2.2 Priors based on the conjugate prior

A conjugate prior for the von Mises distribution was suggested by Guttorp & Lockhart (1988), and is given by

$$p(\mu, \kappa) \propto I_0(\kappa)^{-c} \exp\left\{R_0 \kappa \cos(\mu - \mu_0)\right\}, \tag{2.10}$$

where μ_0, R_0, and c are the prior mean, prior resultant length, and prior 'sample size', respectively. As discussed previously, we would like to remove the necessity to choose a prior mean μ_0. This can be done by putting a circular uniform prior on μ_0 and integrating it out so that

$$p(\mu, \kappa) \propto \int_0^{2\pi} [2\pi]^{-1} I_0(\kappa)^{-c} \exp\left\{R_0 \kappa \cos(\mu - \mu_0)\right\} d\mu_0 = \frac{I_0(R_0\kappa)}{I_0(\kappa)^c}, \tag{2.11}$$

which only depends on κ. Then, all that remains is choosing values for R_0 and c. It can easily be seen seen that imagining a single datapoint on the circle results in $R_0 = 1$ and $c = 1$, producing the constant prior on κ. Because the constant prior is improper and therefore invalid for hypothesis testing, we examine two valid options instead.

First, the prior used by McVinish & Mengersen (2008) has $R_0 = 0, c = 1$, so that we obtain

$$p(\mu, \kappa) \propto I_0(\kappa)^{-1}. \tag{2.12}$$

This prior will be referred to as prior 2.12, and is displayed in Figure 2.1, in red.

Second, the prior could be taken to be proportional to the likelihood of an imagined dataset $\{a, a + \pi/2\}$, with a any angle. This imagined dataset has two angles at 90° from one another. This results in $R_0 = \sqrt{2}, c = 2$, so we obtain

$$p(\mu, \kappa) \propto I_0\left(\sqrt{2}\kappa\right) I_0(\kappa)^{-2}. \tag{2.13}$$

This prior will be referred to as prior 2.13, and is displayed in Figure 2.1, in blue. It can be seen that this prior has more mass at higher values of κ.

Denoting the normalizing constant of either prior by

$$g = 2\pi \int_0^{\infty} I_0(R_0\kappa) I_0(\kappa)^{-c} d\kappa,$$

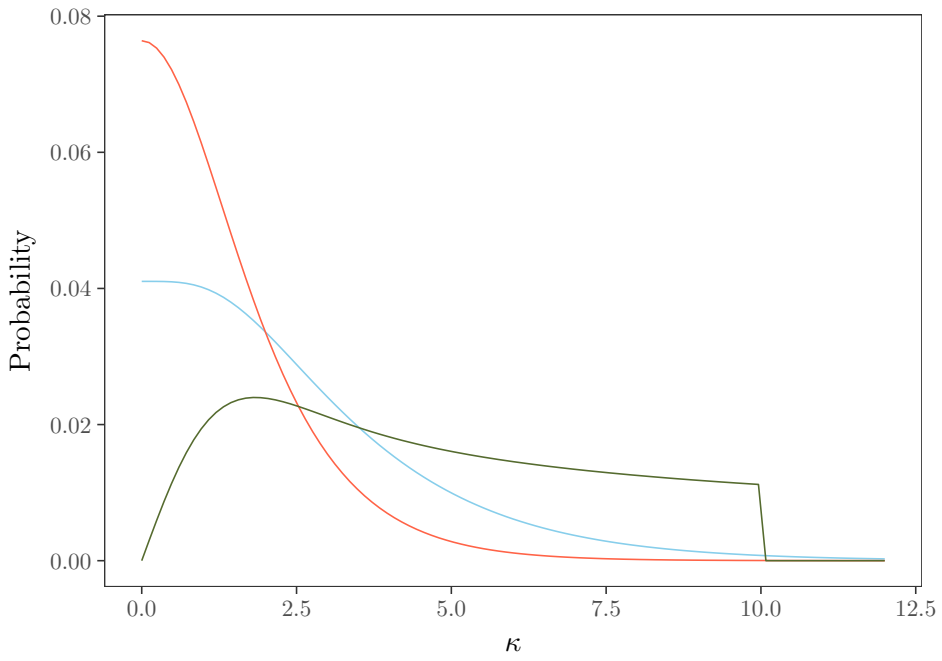


Figure 2.1: Graphs of three different choices of priors for κ: Prior 2.12 (red) has $R_0 = 0, c = 1$, prior 2.13 (blue) has $R_0 = \sqrt{2}, c = 2$, and the Jeffreys prior (green) has $\kappa_u = 10$.

the marginal likelihood for H_1 for these priors is

$$m_1(\boldsymbol{\theta}) = \int_0^\infty \int_0^{2\pi} p(\mu,\kappa) f(\boldsymbol{\theta} \mid \mu,\kappa) d\mu d\kappa \tag{2.14}$$

$$= g\,[2\pi]^{-n} \int_0^\infty \frac{I_0(R_0\kappa)}{I_0(\kappa)^c} I_0(\kappa)^{-n} \int_0^{2\pi} \exp\left\{R\kappa\cos\left(\bar{\theta}-\mu\right)\right\} d\mu d\kappa \tag{2.15}$$

$$= g\,[2\pi]^{-(n+1)} \int_0^\infty I_0(R_0\kappa) I_0(R\kappa) I_0(\kappa)^{-(n+c)} d\kappa. \tag{2.16}$$

The Bayes factor in favor of the alternative is then

$$BF_{10} = \frac{m_1(\boldsymbol{\theta})}{m_0(\boldsymbol{\theta})} = [2\pi]^n m_1(\boldsymbol{\theta}) = g\,[2\pi]^{-1} \int_0^\infty I_0(R_0\kappa) I_0(R\kappa) I_0(\kappa)^{-(n+c)} d\kappa. \tag{2.17}$$

This can be computed by univariate numerical integration.

2.2.3 Jeffreys prior

The Jeffreys prior is a common choice for non-informative priors, especially in low-dimensional parameter spaces as is the case here. The Jeffreys prior is proportional to the square root of the determinant of the Fisher Information Matrix $\mathcal{I}(\boldsymbol{\phi})$ for a single observation, so that for the von Mises distribution it is given by

$$p(\boldsymbol{\phi}) \propto \sqrt{\det\left[\mathcal{I}(\boldsymbol{\phi})\right]} = \sqrt{\kappa A(\kappa) A'(\kappa)}, \tag{2.18}$$

where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$ and $A'(\kappa) = \frac{d}{d\kappa} A(\kappa)$.

An attractive property of this prior is that it has $p(\kappa = 0) = 0$. However, this prior is improper, which means it can not be used directly in hypothesis testing. Therefore, we suggest to take a truncation of this prior from above at some value κ_u. A proper prior based on the Jeffreys prior is then given by

$$p(\mu,\kappa \mid \kappa_u) = \frac{I(\kappa < \kappa_u)\sqrt{\kappa A(\kappa) A'(\kappa)}}{2\pi \int_0^{\kappa_u} \sqrt{\kappa A(\kappa) A'(\kappa)} d\kappa}, \tag{2.19}$$

where $I(\cdot)$ is an indicator function. This prior with $\kappa_u = 10$ is shown in Figure 2.1, in green.

To choose κ_u, it might be thought of as an upper bound for the values of κ for which we will be able to find support. If the data favors a value of κ higher than κ_u, the marginal likelihood of the alternative hypothesis H_1 will be underestimated, although H_1 will still be preferred. Conversely, it should be noted that even if the likelihood strongly suggests $\kappa < \kappa_u$, the resulting

Bayes Factor will still depend on κ_u through the integral in the normalizing constant. The concern that a somewhat arbitrary choice must be made can be alleviated somewhat by performing a sensitivity analysis. In Section 2.2.4, it will be shown that the hypothesis test using this prior performs well even for some fixed values of κ_u.

The Bayes Factor is given by

$$BF_{10} = (2\pi)^n \int_0^\infty \int_0^{2\pi} p(\mu, \kappa) f(\boldsymbol{\theta} \mid \mu, \kappa) d\mu d\kappa \tag{2.20}$$

$$= \int_0^\infty p(\mu, \kappa) I_0(\kappa)^{-n} \int_0^{2\pi} \exp\left\{R\kappa \cos(\bar{\theta} - \mu)\right\} d\mu d\kappa \tag{2.21}$$

$$= 2\pi \left[\int_0^{\kappa_u} \sqrt{\kappa A(\kappa) A'(\kappa)} d\kappa\right]^{-1} \int_0^{\kappa_u} \sqrt{\kappa A(\kappa) A'(\kappa)} I_0(R\kappa) I_0(\kappa)^{-n} d\kappa. \tag{2.22}$$

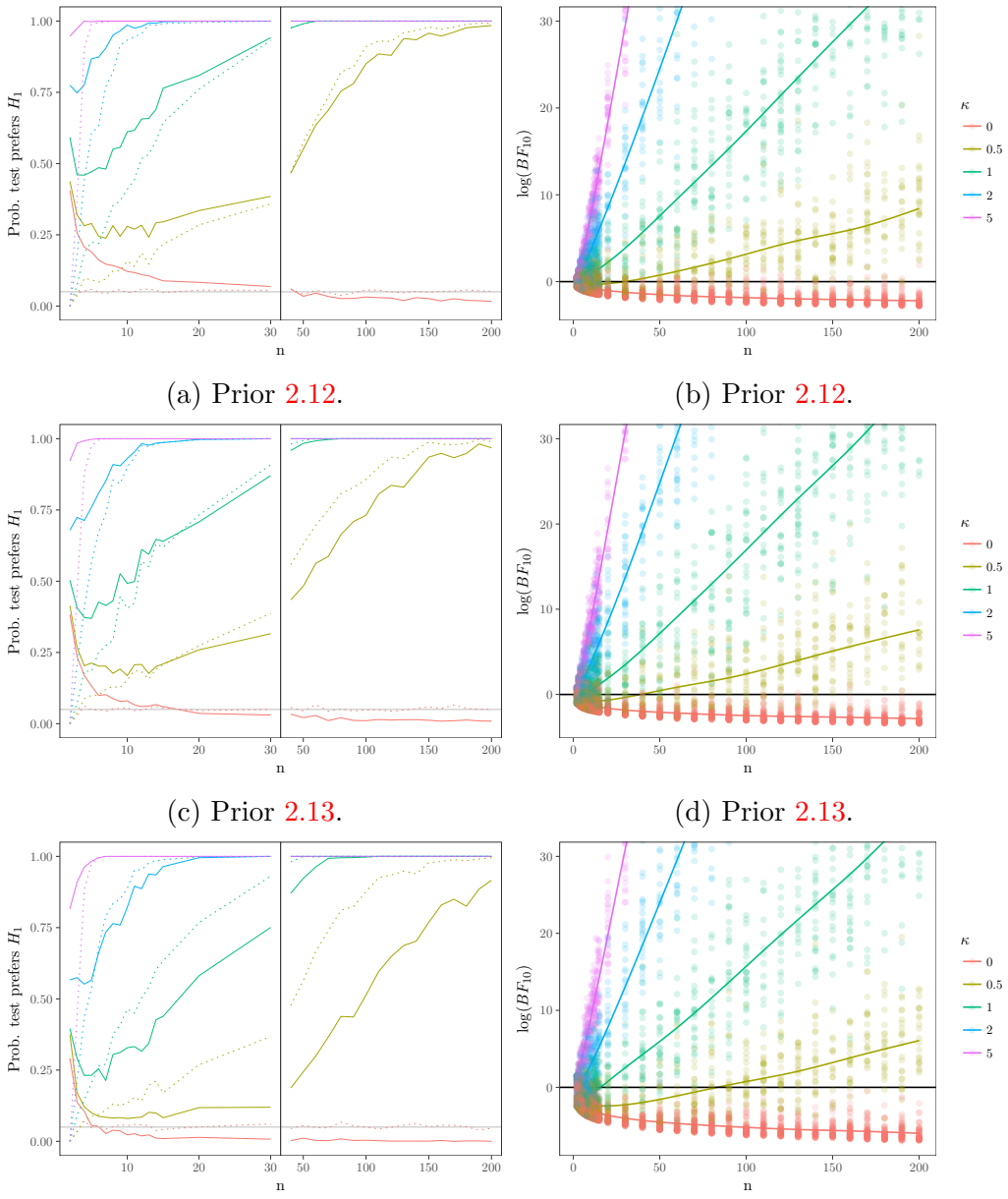
2.2.4 Simulation

In order to assess the performance of the Bayesian hypothesis tests with a von Mises alternative and the three priors discussed previously, a simulation study was performed. One million datasets were sampled from the von Mises distribution with κ set to $\{0, 0.5, 1, 2, 5\}$, where $\kappa = 0$ was used three times more often as it represents H_0. Samples sizes were randomly selected from $\{2, \ldots, 15, 20, 30, \ldots, 190, 200\}$.

Figure 2.2 shows the performance of $BF_{10} > 1$ as a decision criterion for all priors, as well as a plot of the obtained log Bayes factors. In general, all three tests perform well, and are particularly good at correctly classifying data generated under the null hypothesis. Prior 2.12 and prior 2.13 show very similar performance, although prior 2.13 is more prone to select H_0. The Jeffreys prior with $\kappa_u = 20$ is even more prone to select H_0. When data is almost uniform with $\kappa = 0.5$, the tests need a large sample size to select H_1 more than half of the time (around $n > 50$ for prior 2.12 and prior 2.13, and $n > 100$ for the Jeffreys prior with $\kappa_u = 20$).

Compared to the error rates of the Rayleigh test, the current test has better power in all situations but those with $\kappa = .5, n > 30$ and $\kappa = 0, n < 30$. For those cases, it can be seen in the plots on the right of Figure 2.2 that the Bayes factors that are produced are somewhat indecisive, so they may not be taken as evidence in favor of either hypothesis at all. Also, it can be seen that if H_0 is true, $p(H_0 \mid \boldsymbol{\theta}) \to 1$ as $n \to \infty$, which is not the case for the Rayleigh test.

It can be seen that in some cases, such as in 2.2a with $\kappa = 0.5$, increasing the sample size from 1 to 10 actually decreases the probability of selecting



(a) Prior 2.12. (b) Prior 2.12.

(c) Prior 2.13. (d) Prior 2.13.

(e) Jeffreys prior with $\kappa_u = 20$. (f) Jeffreys prior with $\kappa_u = 20$.

Figure 2.2: Results of the simulation study for prior 2.12 (top), prior 2.13 (middle) and the Jeffreys prior (bottom) with $\kappa_u = 20$. The left plots show the proportion of simulations which obtained a Bayes factor in favor of the alternative hypothesis ($BF_{10} > 1$). Error rates for the Rayleigh test with $\alpha = .05$ are provided as dotted lines, with the nominal significance displayed as a gray line at .05. The right plots show a subsample of the log-Bayes factors obtained for different sample sizes n and κ, as well as a solid trendline computed from the full simulation showing the average log Bayes factor for each sample size.

H_1, even though it is the true hypothesis. This is a known property of some Bayesian hypothesis tests. It should be noted that in these cases, the Bayes factor is shows indecision.

2.3 Tests with a Kernel Density alternative

If the von Mises alternative is insufficient, the correct alternative distribution to test against is often unknown. A pure Bayesian approach could be to formulate a set of possible models, and choose between this set of alternatives. However, this requires attempting to fit an infinite set of models which might be hard to do in practice.

Instead, it may be useful to fit a very flexible model as the alternative, which can mimic the true distribution well, so as to provide an omnibus test against many possible models. A kernel density fulfills this role, being able to approximate any density given enough data. Recent developments of kernel density methods for circular data have focused on kernel density bandwith selection and kernel regression (Di Marzio et al., 2009; Oliveira et al., 2012; Di Marzio et al., 2013; Oliveira et al., 2014).

Here, we will build a test for circular uniformity which uses a von Mises kernel density as the alternative. The pdf of the kernel density based on a dataset $\boldsymbol{\Theta} = \Theta_1, \ldots, \Theta_n$ is given by

$$f(\theta \mid \boldsymbol{\Theta}, \kappa) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{M}(\theta \mid \Theta_i, \kappa). \tag{2.23}$$

Our interest is to obtain a posterior for the bandwith κ, which is the only free parameter. However, if the likelihood is specified as

$$f(\boldsymbol{\Theta} \mid \kappa) = \prod_{j=1}^{n} f(\theta_j \mid \boldsymbol{\Theta}, \kappa) = \prod_{j=1}^{n} \sum_{i=1}^{n} \mathcal{M}(\theta_j \mid \Theta_i, \kappa). \tag{2.24}$$

then $\mathcal{M}(\theta_j \mid \Theta_i, \kappa) \to \infty$ if $i = j, \kappa \to \infty$. Therefore, following Hall et al. (1987), we specify the likelihood in a Leave-one-out cross-validation sense, by setting

$$f(\boldsymbol{\Theta} \mid \kappa) = \prod_{j=1}^{n} \sum_{i \neq j} \mathcal{M}(\theta_j \mid \Theta_i, \kappa). \tag{2.25}$$

This leads to the posterior

$$p(\kappa \mid \boldsymbol{\Theta}) \propto f(\kappa \mid \boldsymbol{\Theta}) p(\kappa) \tag{2.26}$$

where the prior for $p(\kappa)$ must still be set. Note that for this model, κ has a different interpretation than in the von Mises, so a different prior is in order. Specifically, in the von Mises model κ refers to the concentration of the full dataset, while in the kernel density model κ refers to the concentration around each separate data point. Therefore, higher values of κ should be considered likely a priori.

The Bayes factor is given by

$$BF_{10} = [2\pi]^n \int_0^\infty p(\kappa) \prod_{j=1}^{n} \sum_{i \neq j} [2\pi I_0(\kappa)]^{-1} \exp\left\{\kappa \cos(\theta_j - \Theta_i)\right\} d\kappa \tag{2.27}$$

$$= \int_0^\infty \frac{p(\kappa)}{I_0(\kappa)^n} \prod_{j=1}^{n} \sum_{i \neq j} \exp\left\{\kappa \cos(\theta_j - \Theta_i)\right\} d\kappa, \tag{2.28}$$

which is once again computed by univariate numerical integration.

For priors of the type discussed in Section 2.2.2, the Bayes factor for some R_0 and c can be written as

$$BF_{10} = \left[\int_0^\infty I_0(R_0\kappa) I_0(\kappa)^{-c} d\kappa\right]^{-1} \times \tag{2.29}$$

$$\int_0^\infty I_0(R_0\kappa) I_0(\kappa)^{-(n+c)} \prod_{j=1}^{n} \sum_{i \neq j} \exp\left\{\kappa \cos(\theta_j - \Theta_i)\right\} d\kappa. \tag{2.30}$$

Another good option for the kernel density model specifically is the Jeffreys prior discussed in 2.2.3, as it allows tuning κ_u to accomodate reasonably high values for the concentration. For this prior, the Bayes factor can be written as

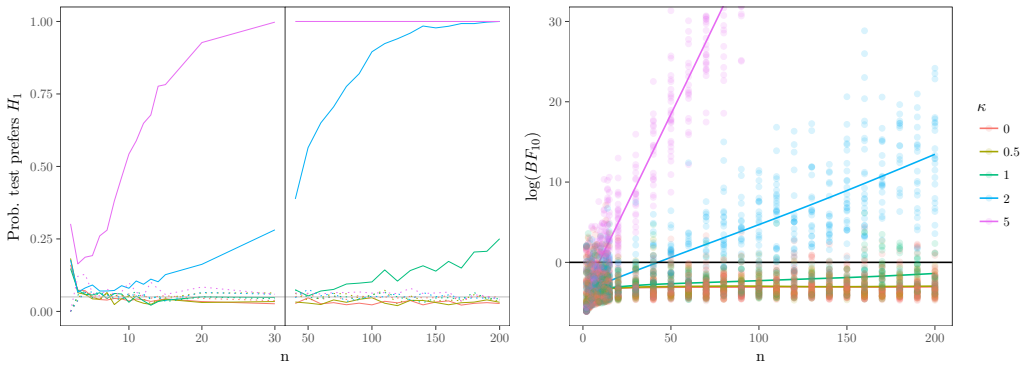
$$BF_{10} = \left[2\pi \int_0^{\kappa_u} \sqrt{\kappa A(\kappa) A'(\kappa)} d\kappa\right]^{-1} \times \tag{2.31}$$

$$\int_0^{\kappa_u} \frac{\sqrt{\kappa A(\kappa) A'(\kappa)}}{I_0(\kappa)^{n+1}} \prod_{j=1}^{n} \sum_{i \neq j} \exp\left\{\kappa \cos(\theta_j - \Theta_i)\right\} d\kappa. \tag{2.32}$$

2.3.1 Simulation

We assess the performance of the kernel based circular uniformity test for antipodal von Mises. The antipodal von Mises is an antipodally symmetric mixture of two von Mises distributions, where data was obtained by drawing from the pdf

$$f(\theta \mid \mu, \kappa) = \frac{1}{2}\mathcal{M}(\theta \mid \mu, \kappa) + \frac{1}{2}\mathcal{M}(\theta \mid \mu + \pi, \kappa). \tag{2.33}$$



(a) (b)

Figure 2.3: Performance for data from the antipodal von Mises distribution with various true values for κ. The left plot shows the proportion of simulations which obtained a Bayes factor in favor of the alternative hypothesis ($BF_{10} > 1$). Error rates for the Rayleigh test with $\alpha = .05$ are provided as dotted lines, with that nominal significance displayed as a gray line at .05. The right plot shows a subsample of the log Bayes factors obtained for different sample sizes n and κ, as well as a solid trendline computed from the full simulation showing the average log Bayes factor for each sample size.

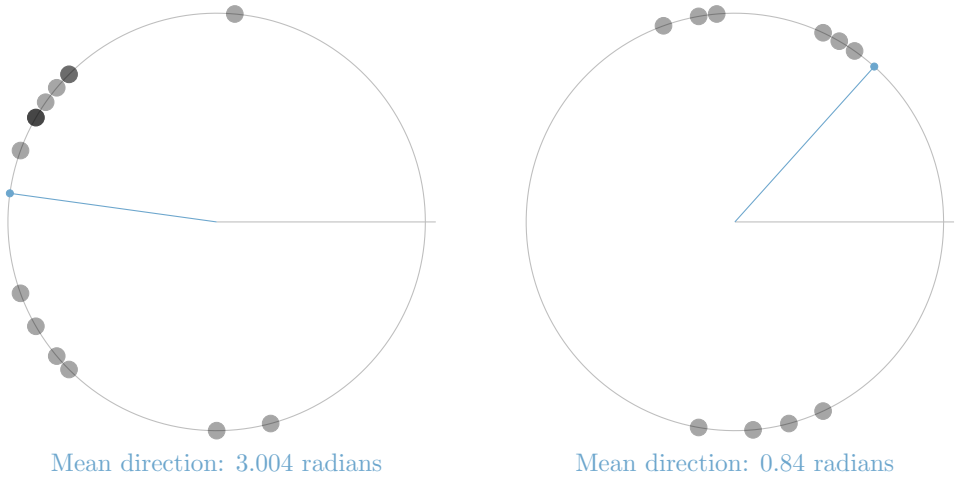
This alternative hypothesis is chosen to be especially hard for the von Mises based tests developed in Section 2.2. The setup in terms of sample sizes and chosen true values for κ is the same as in Section 2.2.4.

Results for data generated from the antipodal von Mises distribution are displayed in Figure 2.3. It can be seen that the Rayleigh test performs abysmally, which is expected, because it is based on rejection of H_0 for large values of the resultant length, which for the antipodal von Mises is zero on average. Our method picks up the difference with reasonable power when data was generated with $\kappa > 2$. In order to detect nonuniformity for antipodal von Mises data with $\kappa = 1$, a very large sample is needed, but it must be noted that antipodal data with small κ is almost uniform. Evidence in favor of H_0 is collected slowly, but with larger sample sizes, H_0 is selected more and more.

2.4 Examples

In this section, the method will be applied to two examples.

In a classic experiment on pigeon homing (Schmidt-Koenig, 1963), the vanishing angles of homing pigeons were measured, with the initial question



(a) Example data 1.

(b) Example data 2.

Figure 2.4: The two example datasets. For each subfigure, the blue line between the center and the circle depicts the mean direction, while the gray line depicts 0°.

of whether the vanishing direction is circular uniform or follows some other circular distribution. Two datasets from this experiment are depicted in Figure 2.4. In one experiment, also provided in Fisher (1995), fifteen homing pigeons were measured to have vanishing directions given by

$$\{85°,135°, 135°, 140°, 145°, 150°, 150°, 150°,\\ 160°, 285°, 200°, 210°, 220°, 225°, 270°\},$$

shown in Figure 2.4a). In another dataset, provided in Mardia & Jupp (2000), ten pigeons were measured to have vanishing directions

$$\{55°, 60°, 65°, 95°, 100°, 110°, 260°, 275°, 285°, 295°\},$$

shown in Figure 2.4b).

2.4.1 Homing pigeon example 1

The results for the first dataset are given in Table 2.1. For this dataset, reasonable hypotheses are that the data are either circular uniform (which we call H_0), or that the data follow a symmetric unimodal distribution, where we pick the von Mises distribution (which we call $H_{\mathcal{M}}$ here). These two hypotheses are evaluated as discussed in Section 2.2.2, using prior 2.12 given

by $p(\kappa) \propto I_0(\kappa)^{-1}$ because a low concentration is expected. Table 2.1 denotes the results of our hypothesis test, as well as the Rayleigh test for comparison. The log marginal likelihood of H_0 is -27.57, while the log marginal likelihood of $H_{\mathcal{M}}$ is -23.92, so $H_{\mathcal{M}}$ is most supported by the data. In fact, the Bayes factor in favor of $H_{\mathcal{M}}$ is 38.54, so that the posterior probability of $H_{\mathcal{M}}$ is 0.975, which constitutes strong support for this hypothesis.

2.4.2 Homing pigeon example 2

For the second dataset, the hypothesis that the data is bimodal is also reasonable, although we might not want to assume antipodal symmetry. To demonstrate the flexibility of the Bayesian approach, we evaluate three hypotheses jointly. The hypotheses are circular uniformity (H_0), the von Mises distribution ($H_{\mathcal{M}}$), and the kernel density alternative (H_k) described in Section 2.3. For the von Mises distribution, the same conjugate prior is used as before, $p(\kappa) \propto I_0(\kappa)^{-1}$. For the kernel density alternative, higher concentrations are more plausible than for the von Mises hypothesis, because dispersion in the final kernel density model is not exclusively determined by κ, but also by the spread of the data. Therefore, we pick the Jeffreys prior here, truncated above at 40. In a small sensitivity analysis for the truncation value (not reported further), the Bayes factor was robust to truncation values above 20, although setting the value extremely high will influence the marginal likelihood, and the inference as a result.

In order to compare the relative probability of each hypothesis, posterior model probabilities were computed. When choosing between a set of p models, we can compute the posterior model probability of model i, assuming equal prior model probabilities, as

$$p(H_i \mid \boldsymbol{\theta}) = \frac{m_i(\boldsymbol{\theta})}{\sum_{j=1}^{p} m_j(\boldsymbol{\theta})}, \tag{2.34}$$

where $m_a(\boldsymbol{\theta})$ denotes the marginal likelihood of model H_a. This will provide the relative probabilities of the models that are assessed.

Results are displayed in Table 2.2. The Rayleigh test is not significant ($p = 0.62$), suggesting no departure from uniformity. In contrast, our comparison of hypotheses shows a preference for the kernel density alternative,

Table 2.1: Results of example 1.

Bayes Factor	$p(H_0 \mid \boldsymbol{\theta})$	$p(H_{\mathcal{M}} \mid \boldsymbol{\theta})$	Rayleigh Statistic	Rayleigh p-value
38.542	0.025	0.975	0.637	0.001

Table 2.2: Results of example 2.

$p(H_0 \mid \boldsymbol{\theta})$	$p(H_{\mathcal{M}} \mid \boldsymbol{\theta})$	$p(H_k \mid \boldsymbol{\theta})$	Rayleigh Statistic	Rayleigh p-value
0.034	0.012	0.954	0.223	0.620

giving it a posterior model probability of 0.954. This can be seen as evidence that the data generation distribution is likely neither the circular uniform distribution nor the von Mises distribution. Rather, the correct model was likely not included in the set of models that were assessed, which should motivate the researcher to further investigate possible models. This result is easy to interpret and understand, and provides a more complete picture than the usual frequentist test.

2.5 Discussion

Bayesian hypothesis tests for assessing circular uniformity were developed in this paper. The Bayesian approach provides three major advantages for this type of hypothesis. First, the hypothesis of circular uniformity is precisely the type of hypothesis which might be true in reality, so that we would want to choose H_0 if the data supports it. The available frequentist tests do not support this, as an insignificant p-value does not allow us to draw conclusion on whether H_0 is true. Second, the Bayesian hypothesis test allows us to quantify the strength of the evidence, either in an odds ratio in the Bayes factor, or in an intuitive probability in the posterior model probability, which is more informative than the simple dichotomous decisions provided by null hypothesis tests. Third, the Bayesian framework allows us to add additional hypotheses to the comparison quite easily. In example 2 in Section 2.4.2, this is used by having the kernel density alternative effectively act as a "none of the above" category, motivating the researcher to search for a model that fits the data better.

Among the most central critiques of the Bayesian method (and Bayesian testing in particular) lies the difficulty in choosing priors, as this seemingly requires us to know in advance what distribution the data may have should the alternative hypothesis be true. Moreover, the conclusions drawn in Bayesian hypothesis tests are often highly dependent on seemingly arbitrary quantities, most notably the parameters of the prior distribution. However, when choosing a frequentist test for circular uniformity, one is faced with a plethora of tests (see Section 2.1) which are each most powerful against different alternatives. This choice closely mirrors the choice of the prior in the alternative hypothesis of a Bayesian hypothesis test. For example, this can be seen in

Landler et al. (2018), where different tests are recommended for different expected alternative distributions. In either case, we must use our expectations of the distribution of the data, should the alternative hypothesis be true. Furthermore, in Section 2.2.1 it was shown how the selection of priors can be dealt with to circumvent the concerns about their influence on the results.

Beyond circular uniformity, previously Bayesian analyses of circular models have been investigated from several viewpoints. Bayesian model assessment has been investigated for wrapped models (Ravindran & Ghosh, 2011), projected normal models (Nuñez-Antonio et al., 2015) and semiparametric intrinsic models (Bhattacharya & SenGupta, 2009; George & Ghosh, 2006). However, the only previously discussed Bayesian test for circular uniformity the authors are aware of is in McVinish & Mengersen (2008), where the alternative hypothesis is a Dirichlet process mixture of triangular distributions. Compared to that work, our focus is on adding parametric alternatives, simplifying computation, assessing performance of the Bayes factor, developing accessible computational tools and comparison of this method to frequentist methods, both conceptually and in a simulation study. Computation involved in evaluating the marginal likelihood of our models has been reduced to simple univariate numerical integration, which makes running these tests more straightforward and markedly faster. Also, the tools used in this paper are easily available from R through the package `BayesCircIsotropy`, available on GitHub.

Assessing the performance of the method, it was shown that the test is often powerful in selecting the correct model, both for the data from the null hypothesis as well as the alternative. The main difficulty in practice, as is often the case in Bayesian analyses, is selecting a prior. In general, choosing a prior with larger variance will allow us to find support for a larger set of true models, but required sample size to find this support will increase. As is shown in the simulation, some default options perform quite well in common research settings. In practice, it is often advisable to perform a prior sensitivity analysis.

Although the philosophical underpinnings of Bayesian hypothesis testing are not the focus of this paper, we will shortly connect the current work with the ongoing discussion. The Bayesian framework is sometimes touted as inductive, which would suggest Bayesian model comparison is sufficient to draw scientific conclusions from data. Recently, Gelman & Shalizi (2013) refute this claim outright and advocate model checking, as models are usually wrong. We generally follow the view of Morey et al. (2013) and note that tools developed here are useful to give preference between models, but do not necessarily provide inductive evidence in favor of the model assessed, such

as the von Mises model. The kernel density alternative presented in Section 2.3 functions as a form of model checking, circumventing the step of deciding on test statistics to be used in a posterior predictive check, or deciding on a specific alternative hypothesis to test against.

Finally, the approach of this paper is to apply Bayesian hypothesis testing to basic circular data analyses. Future work might attempt to obtain easily computable marginal likelihoods for more complex models. In circular data analysis, model selection is an important avenue that requires more attention.

Chapter 3

Mixtures of Peaked Power Batschelet Distributions for Circular Data With Application to Saccade Directions

Abstract

Circular data are encountered throughout a variety of scientific disciplines, such as in eye movement research as the direction of saccades. Motivated by such applications, mixtures of peaked circular distributions are developed. The peaked distributions are a novel family of Batschelet-type distributions, where the shape of the distribution is warped by means of a transformation function. Because the Inverse Batschelet distribution features an implicit inverse that is not computationally feasible for large or complex data, an alternative called the Power Batschelet distribution is introduced. This distribution is easy to compute and mimics the behaviour of the Inverse Batschelet distribution. Inference is performed in both the frequentist framework, through Expectation-Maximization (EM) and the bootstrap, and the Bayesian framework, through MCMC. All parameters can be fixed, which may be done by assumption to reduce the number of parameters. Model comparison can be performed through information criteria or through bridge sampling in the Bayesian framework, which allows performing a wealth of hypothesis tests through the Bayes factor. An `R` package, `flexcircmix`, is available to perform these analyses.

This chapter is submitted to Journal of Mathematical Psychology as Mulder, K., Klugkist, I., van Renswoude, D. & Visser, I. (tbd) Mixtures of Peaked Power Batschelet Distributions for Circular Data With Application to Saccade Directions.

Author contributions: KM, DVR, IV and IK developed the goals for the study. KM developed the statistical methods, simulation study and software package. DVR and IV provided the data for the motivating example. KM wrote the paper. IK, IV and DVR provided feedback on written work.

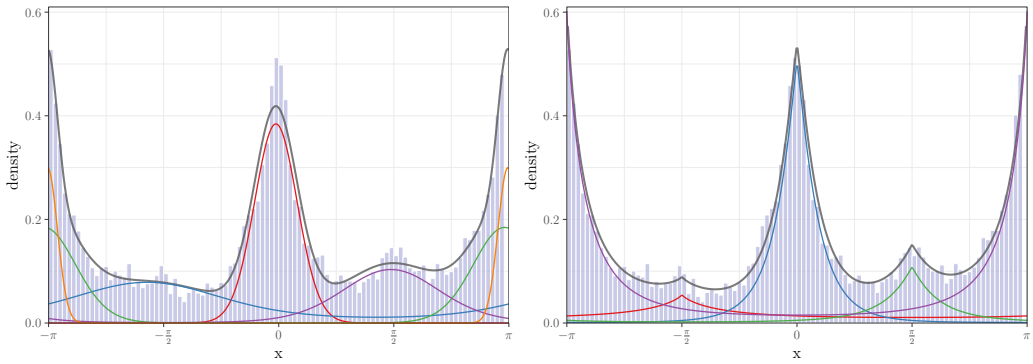
Eye movements are commonly used to study aspects of cognition and its development (Itti & Koch, 2001; Henderson, 2003). Eye movements consist of point fixations and movements between fixations called saccades. In particular, eye movements are of paramount importance in studying the top-down division of attention. For a review, see Rayner (2009).

In eye movement research, a major quantity of interest is the *saccade direction*, the angle between two consecutive fixations. For example, one topic of interest using saccade directions investigates the existence of general directional biases (Tatler & Vincent, 2009), such as a preference for saccades along the horizontal axis (Foulsham et al., 2008) or a preference for leftward saccades (Foulsham et al., 2013). Another topic of interest is eye movement behaviour when reading (Rayner, 2009). Furthermore, distributions of eye movement directions are used to assess the closeness of algorithms to human performance on a variety of eye movement tasks, such as visual search (Najemnik & Geisler, 2008) and saccadic decision making (Tatler et al., 2017; Engbert et al., 2015; Le Meur & Coutrot, 2016).

Previously, it has been difficult to directly analyze a sample of saccade directions. One pragmatic solution to the difficulty of analyzing saccade directions is categorizing the angles in a number of general directions, such as in Foulsham et al. (2008). However, such analyses have reduced power, provide less precise interpretation, and require arbitrary selection of a method of categorization.

A natural model of saccade directions can be obtained by viewing saccade directions as circular data, that is, data measured in angles. Circular data differ from linear data in the sense that circular data are measured in a periodical sample space. For example, an angle of 1° is quite close to an angle 359°, although linear intuition suggests otherwise. Circular data are frequently encountered in scientific fields as diverse as life sciences (Mardia, 2011), behavioural biology (Bulbert et al., 2015), cognitive psychology (Kaas & Van Mier, 2006), bioinformatics (Mardia et al., 2008), political sciences (J. Gill & Hangartner, 2010) and environmental sciences (Arnold & SenGupta, 2006). In this study, a model will be developed that leans on the field of circular statistics (Fisher, 1995; Mardia & Jupp, 2000; Pewsey et al., 2013) to provide satisfying inference for saccade direction data.

Previously, Van Renswoude et al. (2016) used mixtures of von Mises distributions to model saccade direction data, using the R package `movMF` (Hornik & Grün, 2014). The `movMF` package was developed in the more general case for von Mises-Fisher mixtures of distributions on $p-$dimensional hyperspheres, with circular mixtures resulting as a special case. The van Renswoude saccade direction data and the von Mises mixture fit are displayed in Figure 3.1a. It can be seen that the peakedness of the data is not captured well.



(a) Von Mises mixtures through `movMF`. (b) Proposed method.

Figure 3.1: A comparison of two approaches of analyzing saccade direction. Each colored distribution represents a single component of the mixture model.

The peaked mixture model in Figure 3.1b is the final model to be introduced in this work. It naturally incorporates peakedness and requires fewer parameters. It can clearly be seen that using the von Mises mixture approach for saccade direction data is not a natural fit, and as such has major drawbacks.

In this paper, four major drawbacks of analyzing saccade directions using the von Mises mixture method of Hornik & Grün (2014) will be addressed. First, the `movMF` approach provides estimates for the parameters of the model by using the Expectation-Maximization (EM) algorithm, but no measure of their uncertainty, such as confidence intervals or standard errors. Second, it can be seen that saccade direction distributions are very often sharp-peaked or flat-topped distributions, which are not directly modeled by this approach. Instead, the mixture model will deal with peaked data by fitting multiple components on a single mode, which precludes interpretation of the component parameters. Third, because the mixture model deals with peakedness by fitting multiple components for a single mode, it is impossible to compare variances of components, which is something of interest in many saccade direction studies, such as in Van Renswoude et al. (2016). Fourth, there is often a desire to fix component means (or other parameters) to pre-specified values, in order to improve power, which is not possible currently.

The model that will be developed in this paper for saccade direction data has two main characteristics. First, it will be a mixture of circular distributions. Second, it will employ flexible distributions in order to naturally model sharp-peaked and flat-topped components. Inference will be developed in a frequentist framework through an EM-algorithm and the bootstrap, and in a Bayesian framework through MCMC sampling. In order to speed up the

required computations, a new distribution will be introduced that mimics the behaviour of the symmetric density introduced in Jones & Pewsey (2012).

As a motivating example, this study will rely on saccade direction data which was previously published on in Van Renswoude et al. (2016). The main interest in this work is in describing behavioural differences in free-viewing between adults and infants (see also Aslin (2007)). The data consists of 12367 saccades from adults and 4832 saccades from infants. For details on data collection, see Van Renswoude et al. (2016). This data is plotted in Figure 3.2. Because the hypotheses of interest inform the development of the model, they will be revisited here. First, the researchers are interested in reaffirming a horizontal bias, that is, there are more saccade directions along the horizontal axis than the vertical axis. Second, infants are expected to have larger variance in their saccade directions. Third, the researchers are interested in the difference in the horizontal bias of infants and adults. For all of these hypotheses, currently one would be limited to descriptive analyses. The methods developed in this paper will allow full statistical inference. The methods are available in the `R` package `flexcircmix`, freely available on GitHub.

The structure of this paper will proceed as follows. In Section 3.1, the base distribution of the mixture model will be discussed, and the Power Batschelet distribution will be introduced. Inference for the resulting mixture model will be discussed in Section 3.2. The method will be illustrated on both synthetic data and the van Renswoude data in Section 3.3. Finally, some concluding remarks will be given in Section 3.4.

3.1 Family of Batschelet Distributions

In this section, we will introduce Batschelet-type distributions. First, the Inverse Batschelet distribution of Jones & Pewsey (2012) will be recapped. Then, this approach will be adapted into the Power Batschelet distribution. Lastly, a note on computing the circular variance for such distributions will be given.

3.1.1 Inverse Batschelet distribution

The Inverse Batschelet distribution is a peaked or flat-topped circular distribution. It is constructed by modifying a base distribution, for which the von Mises distribution will be used. The von Mises distribution can be seen as a circular analogue to the normal distribution. In the following, it will be introduced shortly.

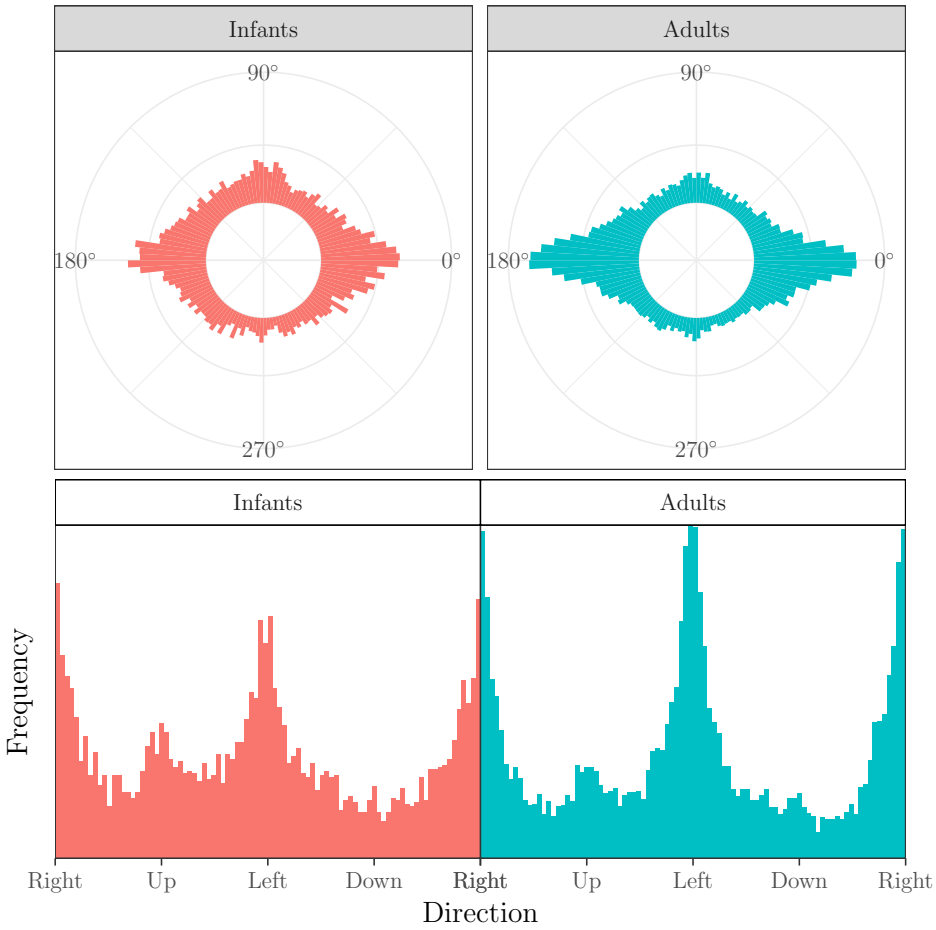


Figure 3.2: Plots of the example data. The top plot provides the data in polar coordinates. The bottom plot shows the data on the real line, where the left and right sides of the plot represent the same point on the circle.

Denote the unit circle by $\mathbb{S}^1$, and the set of observed angles (in radians) by $\boldsymbol{\theta} = \theta_1, \ldots, \theta_n$, with $\theta_i \in \mathbb{S}^1$. For notational simplicity, we assume $\theta \in [-\pi, \pi)$. The von Mises distribution is given by

$$\mathcal{M}(\theta \mid \mu, \kappa) = [2\pi I_0(\kappa)]^{-1} \exp\{\kappa \cos(\theta - \mu)\}, \tag{3.1}$$

where θ is the observed angle, $\mu \in [-\pi, \pi)$ is the mean direction, $\kappa \in \mathbb{R}^+$ is a concentration parameter and $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero. Note that this density is periodic, so $\mathcal{M}(\theta \mid \mu, \kappa) = \mathcal{M}(\theta + 2k\pi \mid \mu, \kappa)$, $\forall\ k \in \mathbb{Z}$. Various von Mises densities are displayed in Figure 3.1a as the separate components of the mixture.

Clearly, saccade directions tend to follow more peaked densities than the von Mises density. Two approaches to incorporate peakedness in the model are the Jones-Pewsey distribution (Jones & Pewsey, 2005) and the Inverse Batschelet distribution (Jones & Pewsey, 2012). Both options have the von Mises distribution as a special case. However, the latter is of somewhat simpler form and allows for more peaked distributions, so it will be employed here.

The core idea of the peaked densities developed in Batschelet (1981) is that given a circular density $f(\theta)$, a new distribution emerges if we take $f(\tau(\theta))$ for some bijective function τ which maps the circle onto itself. We will refer to all distributions obtained by this construction as **Batschelet distributions**, possibly with a prefix relating to the specific bijective function τ used. Attention will be limited to using the von Mises density as the base distribution f, such that we will work in practice with $f_\kappa(\tau(\theta - \mu))$, with κ the concentration parameter of the von Mises distribution.

The function originally used by Batschelet (1981) is given by

$$\tau(\theta) = \theta + \lambda \sin\theta, \tag{3.2}$$

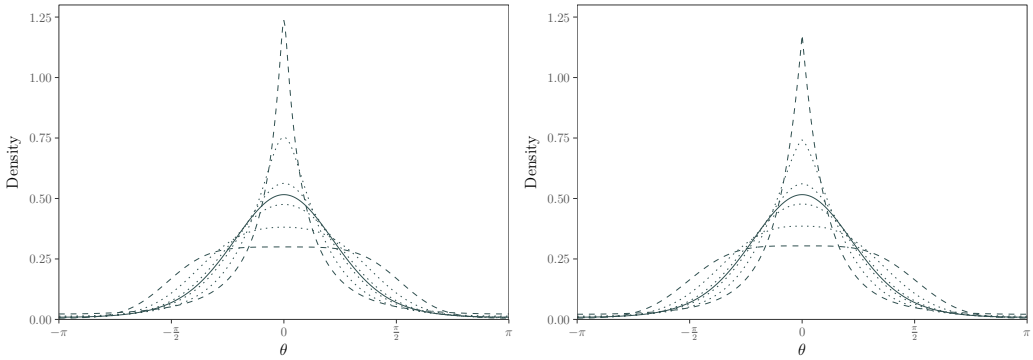
where the peakedness parameter $\lambda \in [-1, 1]$ can be used to obtain a family of flat-topped densities. Using the inverse of $\tau(\theta)$ instead results in a family of peaked densities (Abe et al., 2010; Pewsey et al., 2011). A family of densities incorporating both flat-topped and peaked members was developed in Jones & Pewsey (2012) and will be employed here.

The von Mises based symmetric Inverse Batschelet density is given by

$$f(\theta \mid \mu, \kappa, \lambda) = [2\pi I_0(\kappa) K_{\kappa,\lambda}]^{-1} \exp\{\kappa \cos t_\lambda(\theta - \mu)\} \tag{3.3}$$

where

$$t_\lambda(\theta) = \frac{1-\lambda}{1+\lambda}\theta + \frac{2\lambda}{1+\lambda} s_\lambda^{-1}(\theta) \tag{3.4}$$



(a) Inverse Batschelet distribution. (b) Power Batschelet distribution.

Figure 3.3: Two types of Batschelet distributions, based on the von Mises distribution with $\mu = 0, \kappa = 2$. In order of increasing height at $\theta = 0$, the peakedness parameter $\lambda = \{-.8, -.4, -.1, 0, .1, .4, .8\}$. In each figure, $\lambda = \{-.8, .8\}$ are dashed, while $\lambda = 0$ is a solid black line, with all others dotted. It can be seen that the densities are extraordinarily similar.

with $s_\lambda^{-1}(\theta)$ being the inverse of $s_\lambda(\theta) = \theta - \frac{1}{2}(1+\lambda)\sin(\theta)$, and

$$K_{\kappa,\lambda} = \frac{1+\lambda}{1-\lambda} - \frac{2\lambda}{1-\lambda}\int_{-\pi}^{\pi}[2\pi I_0(\kappa)]^{-1}\exp\left\{\kappa\cos\left(\theta-(1-\lambda)\sin(\theta)/2\right)\right\}d\theta. \tag{3.5}$$

Note that $t_\lambda(\theta)$ is not available analytically because $s_\lambda^{-1}(\theta)$ is not. Therefore, evaluation of the density requires both numerical integration and numerical inversion. The density is plotted with various values of λ in Figure 3.3a. It can be seen that the peaked distribution observed for the saccade data can be obtained from this distribution when $0 < \lambda \leq 1$.

Although for many applications the computational burden of numerically inverting a function for each density evaluation is acceptable, such computations quickly become burdensome upon incorporation of the density into a larger model, such as a mixture model. This same can occur when using certain methods for uncertainty quantification, such as MCMC or the bootstrap. Therefore, in order to be able to employ Batschelet distributions in a broader context of models, an alternative to $t_\lambda(\theta)$ will be introduced in the following section. The major advantage will be that the alternative will not require numerical inversion.

3.1.2 Power Batschelet distribution

In order to improve computational efficiency in more complex models, $t_\lambda(\theta)$ can be replaced by a function of similar shape, but more appealing computational properties. We propose

$$t^*_\lambda(\theta) = \text{sign}(\theta)\pi\left(\frac{|\theta|}{\pi}\right)^{\gamma(\lambda)}, \tag{3.6}$$

with $\gamma(\lambda) \in \mathbb{R}^+$, which has the basic properties required of it, namely to be a mapping of the circle onto itself, so long as $-\pi \leq \theta \leq \pi$ is assumed. In practice, this property is generally forced upon the original data θ_o by taking $\theta = [(\theta_o + \pi) \mod 2\pi] - \pi$. Note this does not change the angle, merely its numerical representation.

Next, the function $\gamma(\lambda)$ should be chosen such that changing λ mimics the behaviour of parameter λ of the Inverse Batschelet distribution in Equation 3.3. First, in order to keep the parametrization where $-1 \leq \lambda \leq 1$ with negative values corresponding to flat-topped densities, take

$$\gamma(\lambda) = \frac{1 - c\lambda}{1 + c\lambda}, \tag{3.7}$$

where c is some fixed constant, chosen such that t^*_λ closely approximates t_λ. In order to choose c, the difference between t^*_λ and t_λ was numerically minimized over values of c,[1] resulting in $c = 0.4052284$. The two functions t^*_λ and t_λ, are plotted together in Figure 3.4. It is clear that the functions, although they do not exactly coincide, are strongly comparable. In practical use, the resulting density of the Power Batschelet distribution was found to be evaluated more than several hundred times faster than the Inverse Batschelet distribution. The resulting density is shown in Figure 3.3b, where again, we conclude that the densities are strongly similar.

The new continuous function $t^*_\lambda(\theta)$ is trivial to compute and has several attractive properties. For example, note that the we simply have $t^{*-1}_\lambda(\theta) = t^*_{-\lambda}(\theta)$.

Using this function, the Power Batschelet distribution is then defined as

$$f_{PB}(\theta \mid \mu, \kappa, \lambda) = [K^*_{\kappa,\lambda}]^{-1} \exp\{\kappa \cos t^*_\lambda(\theta - \mu)\}, \tag{3.8}$$

[1] To be precise, c was chosen such that for a specific λ, the mean absolute difference between $t_\lambda(\theta)$ and $t^*_\lambda(\theta)$ evaluated at 100 points evenly spread on the circle was minimized. The final $c = 0.4052284$ is the average between the optimal c for $\lambda = 1$ and $\lambda = -1$. The two functions do not have coincide exactly, as being able to directly compare values for λ is only used for interpretability.

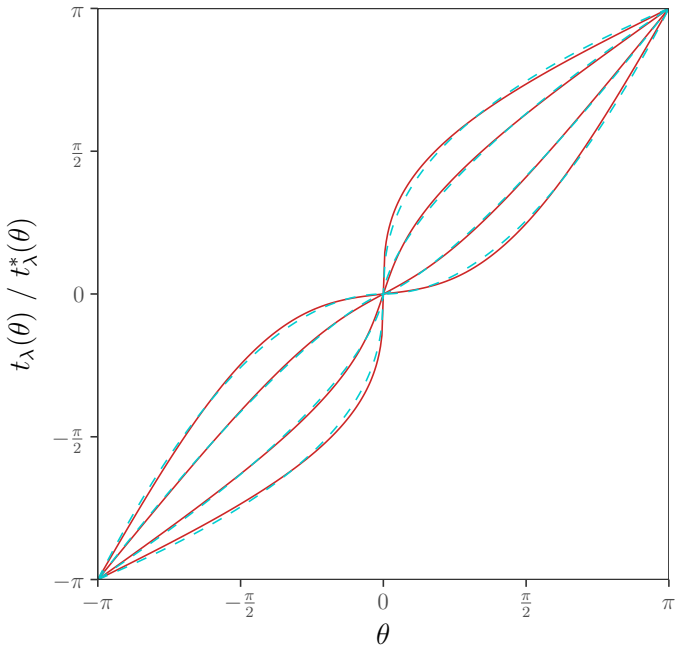


Figure 3.4: Comparison of t^*_λ (blue, dashed) used for the Power Batschelet distribution and t_λ (red, solid) used for the Inverse Batschelet distribution, with, in order of increasing height at $\pi/2$, peakedness parameter $\lambda = \{-.8, -.3, .5, 1\}$.

where

$$t^*_\lambda(\theta) = \text{sign}(\theta)\pi \left(\frac{|\theta|}{\pi} \right)^{\frac{1-0.4052284\lambda}{1+0.4052284\lambda}}, \tag{3.9}$$

and the inverse of the normalizing constant is

$$K^*_{\kappa,\lambda} = \int_{-\pi}^{\pi} \exp\{\kappa \cos t^*_\lambda(\theta - \mu)\} d\theta, \tag{3.10}$$

which must still be numerically integrated. The Power Batschelet distribution generally shares the properties of the Inverse Batschelet distribution, in that it is symmetric around μ and unimodal. Several further properties of this distribution are discussed in Appendix 3.A.

A possible problem is that if $0 < \lambda \leq 1$, we have $0 < \gamma(\lambda) < 1$, and thus $\frac{dt^*_\lambda(\theta)}{d\theta}\Big|_{\theta=0} = \infty$, so the function is not twice differentiable for that range of λ, nor smooth. That is, the probability density is continuous, and so is $t^*_\lambda(\cdot)$, but its derivative is not, nor is $\frac{df_{PB}(\theta|\mu,\kappa,\lambda)}{d\theta}$. As a result, not all regularity conditions for maximum likelihood estimation are not met, in very similar fashion to the commonly used Laplace (double exponential) distribution. In addition, due to the role of this distribution as a close approximation to the Inverse Batschelet, results for the Power Batschelet distribution can be seen as an approximation to the results of the Inverse Batschelet distribution.

If one is concerned about the regularity conditions for the Power Batschelet distribution, the Inverse Batschelet distribution is an alternative which is also implemented in the package `flexcircmix`. However, any analysis with that method may several orders of magnitude longer. In practice, we have not run into any issues related to this, so we prefer the computational efficiency of the Power Batschelet distribution.

3.1.3 Measures of circular dispersion

While for the von Mises distribution the circular variance is known to decrease monotonically with increasing κ regardless of the other parameters, this does not hold true for Batschelet distributions, because the peakedness parameter λ also exerts strong influence on the circular variance. However, it is desirable to compare the circular variance across components in the mixture model discussed in the following sections. Therefore, we compute the circular variance v, given by $v = 1 - \rho$, where ρ is the population resultant length associated with the circular density $f(\theta \mid \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ denotes a vector of parameters. In the general case, it is given by

$$\rho = E[\cos \Theta] = \int_{-\pi}^{\pi} \cos\theta \; p(\theta \mid \boldsymbol{\phi}) d\theta. \tag{3.11}$$

If the data has the von Mises distribution, it is known that $\rho = \frac{I_1(\kappa)}{I_0(\kappa)}$ (Mardia & Jupp, 2000), but in general, computing ρ will require numerical integration. Denoting the normalizing constant by $C(\kappa, \lambda) = \left[\int_{-\pi}^{\pi} \exp\{\kappa \cos t_\lambda(\theta)\} d\theta\right]^{-1}$, we have

$$\rho(\kappa, \lambda) = E[\cos \Theta] = \int_{-\pi}^{\pi} \cos\theta \; p(\theta \mid \mu, \kappa, \lambda) d\theta \tag{3.12}$$

$$= C(\kappa, \lambda) \int_{-\pi}^{\pi} \cos\theta \;\; \exp\{\kappa \cos t_\lambda(\theta)\} d\theta. \tag{3.13}$$

This means we should need at most two numerical integrations. Lastly, the circular standard deviation can be computed by $\sigma_c = \sqrt{-2 \log \rho}$ (Fisher, 1995).

3.2 Inference for Batschelet mixtures

The mixture of Batschelet distributions is given by

$$f(\theta \mid \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{j=1}^{J} \alpha_j f_B(\theta \mid \mu_j, \kappa_j, \lambda_j), \tag{3.14}$$

where j indexes the J components in the mixture, α_j are component weights, and $f_B(\cdot)$ is the chosen density, either Inverse Batschelet or Power Batschelet.

First, in Section 3.2.1, an EM algorithm will be presented. Second, in Section 3.2.2, a method for inference through MCMC is presented. A note on identifiability is given in 3.2.3. Note that the number of components will be assumed to be known initially. In Section 3.2.4, model selection and hypothesis testing will be discussed, which can be used to select the number of components as well evaluate many types of hypotheses. For a discussion of direct inference on mixtures with an unknown number of components, see Richardson & Green (1997a).

3.2.1 EM Algorithm

Directly maximizing the observed data log-likelihood of a mixture model is generally difficult. Therefore, the EM-algorithm will be employed, which exploits the fact that the complete data maximum likelihood, that is, with observed labels, is easier to maximize.

The EM-algorithm consists of the following steps:

(Initialize) Define an $n \times J$ matrix $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J\}^T$, where $\boldsymbol{w}_j$ are n-vectors. Initialize the parameters $\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\lambda}, \boldsymbol{\alpha}$ at some user-specified values.

(E-step) Compute, for all i, j, the elements of $\boldsymbol{W}$ as

$$w_{i,j} = \frac{\alpha_j f_B(\theta_i \mid \mu_j, \kappa_j, \lambda_j)}{\sum_{s=1}^{J} \alpha_s f_B(\theta_i \mid \mu_s, \kappa_s, \lambda_s)}. \tag{3.15}$$

(M-step) For each component k, maximize

$$\ell(\mu_j, \kappa_j, \lambda_j, w_j \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} w_{i,j} \log f_B(\theta_i \mid \mu_j, \kappa_j, \lambda_j). \tag{3.16}$$

The maximization of this log-likelihood of the parameters of a Batschelet distribution may proceed through the Nelder-Mead simplex (Nelder & Mead, 1965), as was done in Jones & Pewsey (2012).

Note that the EM algorithm is not guaranteed to find a global maximum. Therefore, reasonable (or multiple) starting values should be used, in order to assess the validity of the final results. If avoiding convergence to local maxima is of particular importance, one could consider implementing a stochastic version of the EM algorithm (Diebolt & Ip, 1996; Nielsen et al., 2000). However, the Bayesian MCMC approach described in Section 3.2.2 shares the advantages of such methods.

In addition to obtaining estimates of the mixture model, it is essential to infer the uncertainty around these estimates. For mixture models, asymptotic standard errors obtained from inverting the Fisher Information generally require very large datasets in order to have desirable properties (McLachlan & Peel, 2004). Therefore, parameter uncertainty will need to be assessed either through bootstrapping, or through MCMC.

Bootstrapping (Efron & Tibshirani, 1994) was implemented through a nonparametric bootstrap. In order to reduce computational burden, the EM algorithm of each bootstrap sample was given the full data estimates as starting values.

3.2.2 Bayesian inference

A Bayesian analysis of the finite mixture of von Mises-based Batschelet distributions is available through MCMC sampling (Chib & Greenberg, 1995; Gilks et al., 1995). For an introduction focused on mixture models, see Frühwirth-Schnatter (2006). Besides providing uncertainty quantification naturally by

performing inference on the posterior distribution rather than a set of estimates, the Bayesian paradigm also provides computational advantages in this case. In particular, the MCMC algorithm is less likely to converge to local maxima.

As is common for Bayesian sampling for mixture models , the parameter space is augmented by a vector of latent variables $\boldsymbol{z} \in \{1, \ldots, J\}^n$ which contains a group label for each observation. By randomly assigning each observation to a group during every iteration, the problem simplifies to MCMC sampling for each component separately. First, in Section 3.2.2, priors for this model will be discussed. Then, the MCMC algorithm will be provided in Section 3.2.2.

Priors

Although subjective priors can be chosen in practical use, attention here will be restricted to (somewhat) non-informative priors. Priors are required for $\alpha_j, \mu_j, \kappa_j$ and λ_j, either jointly or separately. In principle, priors could even be set for the group assignments.

The component weights α_j are given the conjugate Dirichlet prior distribution, with vector prior parameter $\boldsymbol{n}_0 \in [\mathbb{R}^+]^J$. If $\boldsymbol{n}_0 = \mathbf{1}_J$, this prior is uninformative.

The mean directions $\mu_j \in [-\pi, \pi)$ are given a circular uniform prior, $p(\mu_j) = [2\pi]^{-1}$ which is proper.

The concentration parameters κ_j are given a constant prior $p(\kappa_j) \propto 1$, which is improper. In principle, the Jeffreys prior for the von Mises distribution could also be used. The Jeffreys prior is proportional to the square root of the determinant of the Fisher Information Matrix $\mathcal{I}(\boldsymbol{\phi})$, so that for the von Mises distribution it is given by

$$p(\boldsymbol{\phi}) \propto \sqrt{\det\left[\mathcal{I}(\boldsymbol{\phi})\right]} = \sqrt{\kappa A(\kappa) A'(\kappa)}, \tag{3.17}$$

where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$ and $A'(\kappa) = \frac{d}{d\kappa} A(\kappa)$. However, note that this is *not* the Jeffreys prior for the Inverse Batschelet distribution (as the given $\mathcal{I}(\boldsymbol{\phi})$ is the Fisher information of the von Mises), nor for a mixture of any circular distributions, nor proper. However, it can be used as a relatively diffuse default prior for cases in which very large values of κ are deemed unlikely. For the case of the von Mises distribution, Hornik & Grün (2013) show that the resulting posterior is almost surely proper if $n \geq 2$. A final alternative is to use a relatively diffuse non-conjugate proper prior, such as one from the gamma family of distributions.

The peakedness parameter λ can be given a proper uniform prior $p(\lambda_j) = 1/2, \lambda_j \in [-1, 1)$. However, Jones & Pewsey (2012) note that in maximum

likelihood estimation of the Inverse Batschelet model, estimates often fall on the boundary of the parameter space. Boundary avoiding priors can be used here to prevent this behaviour of the estimates. In particular, one may posit that large values of $|\lambda_j|$ are a priori unlikely. This belief can be captured in a rescaled Beta(a, b) prior, so that $p(\lambda_j) \propto f_{Beta}\left(\frac{\lambda+1}{2} \mid a, b\right)$. If $a = 1, b = 1$, this results in the uniform prior on $[-1, 1]$, while $1 < a, b \leq 2$ gives a range of priors which favor smaller values for $|\lambda_j|$, and thus less peaked and less flat-topped densities.

MCMC algorithm

In the application of MCMC sampling, only the group assignments in latent variable $\boldsymbol{z}$ and the mixture weights $\boldsymbol{\alpha}$ have known full conditional distributions. All other parameters are updated using the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970a). After selecting starting values, the algorithm will be performed for $m = 1, \ldots, M$ iterations, which will constitute a sample from the posterior distribution. One iteration m of the algorithm proceeds as follows:

(z_i) For each observation θ_i, sample $z_i \in 1, \ldots, J$ with group probabilities

$$P(z_i = j) = \frac{\alpha_j f_B(\theta_i \mid \mu_j, \kappa_j, \lambda_j)}{\sum_{s=1}^{J} \alpha_s f_B(\theta_i \mid \mu_s, \kappa_s, \lambda_s)}.$$

This represents assigning this observation to one of the possible mixture components.

($\boldsymbol{\alpha}$) Sample the vector of mixture weights

$$\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_J \sim \text{Dirichlet}(\boldsymbol{n} + \boldsymbol{n}_0),$$

where $\boldsymbol{n} = \{\sum_{i=1}^{n} I(z_i = 1), \ldots, \sum_{i=1}^{n} I(z_i = J)\}^T$, with $I(\cdot)$ the indicator function, and $\boldsymbol{n}_0$ the vector prior parameter for the Dirichlet, which is set to $\boldsymbol{n}_0 = \mathbf{1}_n$ by default for an uninformative prior. Note that the definition of $\boldsymbol{n}$ means that all we need to do to sample the mixture weights is counting the number of observations assigned to each group.

($\mu_j, \kappa_j, \lambda_j$) For each mixture component $j \in 1, \ldots, J$, sample parameters $\mu_j, \kappa_j, \lambda_j$. Note that none of these parameters have known distributions, so we resort to Metropolis-Hastings throughout.

(a) Sample μ_j using an MH-step. As for the proposal μ_j^*, we make use of the fact that the distribution reduces to the von Mises distribution if $\lambda = 0$. Therefore, we can draw from the known distribution of the mean of the von Mises distribution, because it will be somewhat close to the desired distribution. We can take either the previously sampled μ, by sampling from $\mu_j^* \sim \mathcal{M}\left(\mu_j^{(m-1)}, R_j\kappa\right)$, or use the current sample mean direction $\bar{\theta}$, by sampling from $\mu_j^* \sim \mathcal{M}\left(\bar{\theta}_j, R_j\kappa\right)$, where $\bar{\theta}_j$ and R_j are computed from the sample assigned to component j.

(b) Sample κ_j using a MH-step, using a gamma distribution with mean $\kappa^{(m-1)}$ as the proposal distribution. The variance of this gamma distribution is a tuning parameter, which can be changed to improve computational efficiency of the algorithm. If the variance is $\kappa^{(m-1)}$, the proposal is the χ^2-distribution with $\kappa^{(m-1)}$ degrees of freedom. In practice, setting the variance to $.05\kappa^{(m-1)}$ seems to work well.

(c) Sample λ_j using a MH-step, using a uniform proposal

$$U\left[\max(-1, \lambda_j^{(m-1)} - \varepsilon),\ \min(1, \lambda_j^{(m-1)} + \varepsilon)\right]. \tag{3.18}$$

Note that although the proposal distribution seems symmetric, this is not the case if the current value is less than ε from the boundary. Because the proposal is not symmetric, the proposal distribution must be included in the MH ratio. Again, ε is a tuning parameter, and we will set $\varepsilon = 0.01$.

It is well known that if parameters are strongly correlated, MCMC sampling can benefit from joint proposals for the correlated parameters. The parameters κ and λ are correlated, although not in an extreme fashion. Therefore, κ and λ may be sampled jointly, although this did not always prove beneficial in practice.

3.2.3 Model identifiability

In general, mixture models may not be identifiable (Teicher, 1963), a property which manifests itself most often through label switching, where component k represents a different unobserved subpopulation in different bootstrap or MCMC samples from the model. However, forcing an ordering on the means may be sufficient for identification (Everitt, 2004).

For the case of Batschelet mixtures on the circle, means are sometimes fixed by design, because this can be a reasonable assumption for saccade data.

If the means are fixed, the model is identifiable as long as $\{\kappa_j = 0, \lambda = 0\}$ in no more than one component. If this assumption is violated, any convex combination of mixture weights α_j of components where $\{\kappa_j = 0, \lambda = 0\}$ gives the same probability density, so the model is not identified. This is unlikely in practice and can simply be checked in the output.

If the means are not fixed but estimated, label switching may occur. In practical inference, if label switching has occurred, this would be evident in bootstrap or MCMC samples. If label switching has occurred, a post-processing step may be used to solve this issue (M. Stephens, 2000; Jasra et al., 2005a). However, care must be taken in ensuring a circular ordering rather than a linear ordering.

3.2.4 Model selection and hypothesis testing

It is often relevant to compare several models and select the best among them, for example to select the required number of mixture components. The most common approach to model selection is through information criteria such as AIC (Akaike, 1987) and BIC (Schwarz et al., 1978) in frequentist settings, and DIC (Spiegelhalter et al., 2002) and WAIC (Watanabe, 2010) in Bayesian settings (for an overview, see Wagenmakers & Waldorp (2006)). Such tools are provided in the `R` package `flexcircmix` accompanying this paper and provide an approximate comparison of the fit of various models.

However, the Bayesian approach also allows us to perform more sophisticated model comparisons naturally by comparing the models on their posterior model probability. Consider a set of Q models $\mathcal{M}_1, \ldots, \mathcal{M}_Q$ each indexed by a set of free parameters $\boldsymbol{\phi}$, that are to be compared on their probability after observing data. This set of models can also be hypotheses to be compared. For example, one could compare a 3-component model with a 4-component model, a model with mean directions fixed at the cardinal directions versus a model where mean directions can vary freely, or a model that allows peaked distributions (ie. $\lambda \in (-1, 1)$) versus a von Mises mixture model (ie. $\lambda = 0$).

In order to obtain the posterior model probability, the prior probability of the models under consideration must first be assessed. Here, and throughout the rest of this work, the models will be assumed to have equal prior probability, so $p(\mathcal{M}_s) = 1/Q$. As a result, the prior probability drops out of the rest of the formulae.

Then, regardless of the set of models under consideration, we can compute the posterior model probability

$$\text{pmp}(\mathcal{M}_s) = \frac{p(\mathcal{M}_s \mid \boldsymbol{x})}{\sum_{q=1}^{Q} p(\mathcal{M}_q \mid \boldsymbol{x})} \tag{3.19}$$

where $p(\mathcal{M}_s \mid \boldsymbol{x})$ is the *marginal likelihood*, given by

$$p(\mathcal{M}_s \mid \boldsymbol{x}) = \int_{\Omega_{\boldsymbol{\phi}}} p(\boldsymbol{x} \mid \boldsymbol{\phi}) d\boldsymbol{\phi}, \tag{3.20}$$

where $\Omega_{\boldsymbol{\phi}}$ is the sample space of the parameter vector for the model $\mathcal{M}_s$. This integral is in general not easy to compute and has sparked a wealth of methods for computing it (for an overview, see Ardia et al. (2012) and Friel & Wyse (2012)). Perhaps the most promising and stable sampling-based solution is found in bridge sampling (Meng & Wong, 1996), which was recently made more easily applicable as a post-processing step on MCMC output through the `R` package `bridgesampling` (Gronau, Sarafoglou, et al., 2017). Broadly speaking, bridge sampling produces an estimate of the marginal likelihood by evaluating additional samples from a known density that approximates the posterior. For details, see Gronau, Sarafoglou, et al. (2017) and Meng & Wong (1996).

Two issues arise for this specific application. First, the sample of mean direction parameters $\boldsymbol{\mu}_j$ lie on a circular parameter space. Bridge sampling will find a known density that approximates the posterior by using the linear mean and the covariance matrix of the MCMC samples, for example by using the multivariate normal density with the same mean vector and covariance matrix. The approximation need only be roughly correct, which is not necessarily the case for our model. For example, if we have a circular parameter with a mean direction near zero, some sampled values will lie in both intervals $[0, .1]$ and $[2\pi - .1, 2\pi)$. The linear mean will then incorrectly lie near π, and the linear variance will be far too large. To solve this, we will change the numerical representation of the mean direction sample of $\boldsymbol{\mu}_j$ such that it lends itself better to the linear approximation. To do this, first the posterior mean direction $\bar{\mu}_j$ is computed from the sample of mean directions $\boldsymbol{\mu}_j$. Then, by taking $\boldsymbol{\mu}_j^* = [(\boldsymbol{\mu}_j - \bar{\mu}_j + \pi) \bmod 2\pi] - \pi + \bar{\mu}_j$, a numerical representation is obtained that does not have any 'gaps' on the real line, but corresponds to the same set of angles $\boldsymbol{\mu}_j$.

The second issue is that the sample of component weight parameters $\alpha_j \in [0, 1]$ lie on a simplex, that is, they are constrained to sum to one. The bridge sampling usually lies on the real line, such as the aforementioned multivariate normal distribution, which means the constrained parameter space is ignored, so that almost surely invalid proposals are sampled. Therefore, these parameters are given a stick-breaking representation and are then logit-transformed, in a similar manner as in Stan (Carpenter et al., 2017). For details on the transformation, its inverse and associated Jacobian, see the Stan reference manual (Stan Development Team, 2017). The solutions for

both circular and simplex parameters were contributed to the latest version of the `bridgesampling` package.

3.3 Illustration

In order to illustrate the methods presented in this work, they will be applied to two examples.

First, the method is applied to a synthetic dataset in Section 3.3.1, where it is shown that the true parameters of a data generating process can be recovered. Then, in Section 3.3.2, the method is shown to provide new insights in the saccade direction data from Van Renswoude et al. (2016).

3.3.1 Synthetic data

Here, the methods developed in this paper will be applied to a synthetic data set for which parameter values are known. In order to sample from the Inverse Batschelet distribution, the sampling algorithm from Jones & Pewsey (2012) was applied. A data set consisting of 1000 angles was sampled with parameters $\boldsymbol{\mu} = \{-1, 1, 2\}$, $\boldsymbol{\kappa} = \{20, 4, 15\}$, $\boldsymbol{\lambda} = \{-.7, 0, .7\}$, and $\boldsymbol{\alpha} = \{.25, .25, .5\}$.

The results are shown in Table 3.1 and Figure 3.5. First, it is clear that the method is able to recover mean direction. Also, it can be seen that both the bootstrapped confidence intervals and the credible intervals generally include the true value, and cases in which this is not true can be attributed to sampling error.

As mentioned previously, the joint likelihood of $\{\kappa, \lambda\}$ is correlated. Because of this, it can be seen in Table 3.1 that neither κ nor λ can be estimated precisely, with both having confidence intervals that are quite wide. A remarkable property of this method is that while neither of the variance-related parameters is estimated very precisely, the circular standard deviation, computed as in Section 3.1.3, has tighter confidence intervals and is estimated more precisely, so inference on it will be more powerful than inference directly on κ or λ.

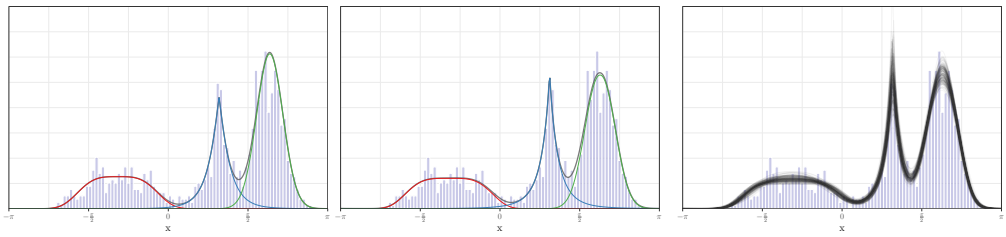
The components weights are estimated adequately in all cases.

3.3.2 Free-viewing data

Here, the method will be applied to a real world example, the free-viewing dataset that was originally published in Van Renswoude et al. (2016), shown again in Figure 3.6. As discussed in Section , there are several hypotheses one

Table 3.1: Synthetic data fits using the Power Batschelet distribution.

		EM	Boot. CI		Bayes (MCMC)		
	Truth	Est.	2.5%	97.5%	Median	2.5%	97.5%
μ_1	-1.00	-1.00	-1.09	-0.91	-1.00	-1.04	-0.96
κ_1	20.00	24.92	8.67	63.37	27.67	8.91	59.99
λ_1	-0.70	-0.81	-1.00	-0.47	-0.85	-0.99	-0.51
α_1	0.25	0.26	0.23	0.28	0.26	0.23	0.29
σ_{c1}	0.52	0.55	0.51	0.60	0.56	0.51	0.60
μ_2	1.00	0.98	0.95	1.01	1.00	0.97	1.03
κ_2	4.00	3.26	3.17	3.94	3.52	2.80	5.17
λ_2	0.70	0.89	0.66	1.00	0.81	0.54	0.98
α_2	0.25	0.26	0.23	0.29	0.27	0.23	0.31
σ_{c2}	0.34	0.42	0.33	0.44	0.38	0.27	0.53
μ_3	2.00	1.97	1.95	2.00	1.98	1.96	2.00
κ_3	15.00	16.29	10.87	44.57	25.08	11.37	67.42
λ_3	0.00	-0.09	-0.36	0.08	-0.22	-0.47	0.06
α_3	0.50	0.48	0.45	0.51	0.48	0.44	0.51
σ_{c3}	0.26	0.28	0.26	0.30	0.28	0.25	0.30



(a) True density. (b) Fit from EM algorithm. (c) MCMC fit.

Figure 3.5: Synthetic data plots using the power Batschelet distibution. The sample of synthetic data, plotted along with a sample of 200 densities of which the parameters were sampled in the MCMC. That is, the spread of probability density functions provides a rough uncertainty bound for the true probability at each point.

might wish to learn about using this dataset. One hypothesis of interest is whether infants have a larger circular variance for each of their mixture components. Another is whether the horizontal bias, that is, the preference for left-right movements, is weaker for infants than for adults. These hypotheses will be assessed here.

Because the mixture model has a fairly large number of parameters, it can be fruitful to consider fixing parameters about which we do not need to learn. For free-viewing data, modes are always observed oriented exactly in the cardinal directions, which will simplify our modeling problem somewhat. The mean directions can be chosen to be fixed at $\mu_1 = -\pi/2$ (upward), $\mu_2 = 0$ (rightward), $\mu_3 = \pi/2$ (downward), $\mu_4 = \pi$ (leftward). Because fewer components are needed in the Batschelet mixture and because the means are fixed, the model actually has fewer parameters than the von Mises mixture model. It is also possible to loosen this assumption slightly by placing a strong prior on the mean directions centered on the aforementioned cardinal directions.

For Bayesian inference, the priors were chosen according to the consideration in Section 3.2.2. To be specific, the prior for μ_j was circular uniform, for κ_j the Jeffreys prior of the von Mises distribution, for λ_j the prior was the rescaled beta distribution proportional to $f_{Beta}\left(\frac{\lambda+1}{2} \mid \sqrt{2}, \sqrt{2}\right)$, and finally the prior for α_j was Dirichlet($\boldsymbol{\alpha} = \{\sqrt{2}, \ldots, \sqrt{2}\}^T$). The MCMC algorithm was run for 46000 iterations, split into 46 parallel chains each having a burn in of 1000.

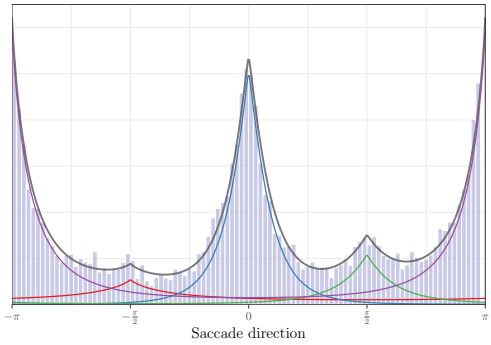
A bootstrap was run with 10000 bootstrap replications. For both adults and infants, the results from the EM algorithm with bootstrapped standard errors will be displayed together with the Bayesian approach in Table 3.2.

Adults

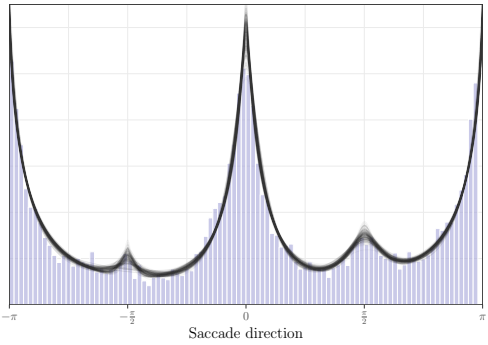
Results for the adult sample are displayed in Figures 3.6a and 3.6b. Visually, the model fit seems excellent, and it can be seen that observed distributions are generally quite peaked. For adults, it can be seen that all four components contribute to the overall shape of the overall shape of the model.

Infants

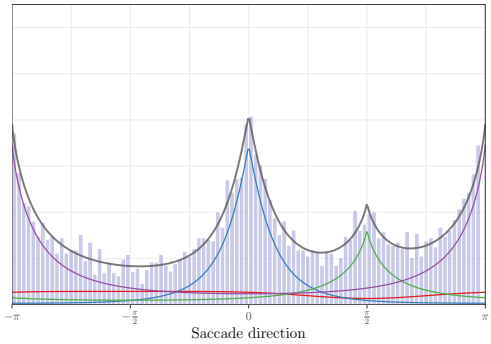
For infants, judging from the convergence plots in Figure 3.7, there might be grounds to assume that fewer than four components may suffice. Specifically, we can see that component 1 (upward) has component weight α that tends to zero. This can also be seen in Figure 3.6c, where the red (upward) component 1 is taken as almost completely flat.



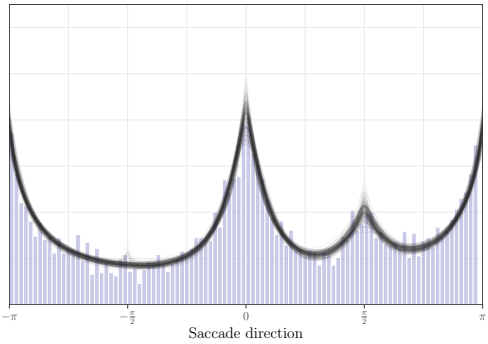
(a) EM fit for adults.



(b) MCMC fit for adults.



(c) EM fit for infants.



(d) MCMC fit for infants.

Figure 3.6: Free-viewing data fit for adults (top) and infants (bottom) using the power Batschelet distibution. The x-axis is given in radians, where 0 corresponds to the rightward direction, while $\pi/2$ corresponds to the downward direction. In the left plot, it can be seen that a third component is estimated for adults, while this is not the case for infants. In the right plot, the densities are plotted that result from sampled parameter sets from the MCMC.

Horizontal Bias comparison

The main question of whether the horizontal bias of adults and infants differ can be addressed by comparing the circular standard deviation in Table 3.2, as well as compare the component weights $\boldsymbol{\alpha}$.

The horizontal components are component 2 and 4. For the rightward component 2, with ($\mu_2 = 0$), the estimates are generally somewhat similar between adults and infants, as the confidence and credible intervals overlap.

For the leftward component 4 ($\mu_4 = \pi$), the confidence and credible intervals of adults and infants do not overlap, which can also be observed in Figure 3.6 by noting that this component has a different shape between Figures 3.6c and 3.6a. For the component weight α_4, adults have EM-estimate and 95% bootstrap confidence interval $\hat{\alpha}^{(EM)} = 0.465$ (0.443, 0.502) and posterior median and credible interval $\hat{\alpha}^{(MCMC)} = 0.542$ (0.431, 0.58), compared to infants which have $\hat{\alpha}^{(EM)} = 0.405$ (0.37, 0.455) and $\hat{\alpha}^{(MCMC)} = 0.472$ (0.28, 0.614).

For the circular standard deviation σ_{c4}, adults have $\hat{\sigma}_c^{(EM)} = 0.886$ (0.812, 0.964) and $\hat{\sigma}_c^{(MCMC)} = 0.995$ (0.786, 1.07), compared to infants which have $\hat{\sigma}_c^{(EM)} = 1.16$ (1.036, 1.258) and $\hat{\sigma}_c^{(MCMC)} = 1.161$ (0.749, 1.391). Therefore, infants seem to have a larger variance on this component than adults. From this, it can be concluded that the horizontal bias exists, and differs between adults and infants.

Hypothesis testing

Using the model comparison methods discussed in Section 3.2.4, several models of interest can be compared using Bayesian hypothesis tests.

First, it is worthwile to investigate whether in general infants and adults differ. This can be done by comparing the model that is discussed in Section 3.3.2, which allows separate parameters for infants and adults, to a model where both are given the same parameters (for which parameter estimates are not shown). For this model, the log Bayes Factor in favor of the model that has separate parameters is 52.7, which is associated with a posterior model probability (assuming equal prior odds) in favor of separate parameters close to 100%. Therefore, we have separated the groups throughout.

Second, as can be seen in Figure 3.6c, one may be unsure about the number of required components for infants. Although four components were assumed so far, three may suffice. Again, we find almost certain evidence, log Bayes Factor 70.7, posterior probability $\approx$ 100%, that the model with four components fits better than the model with only three components.

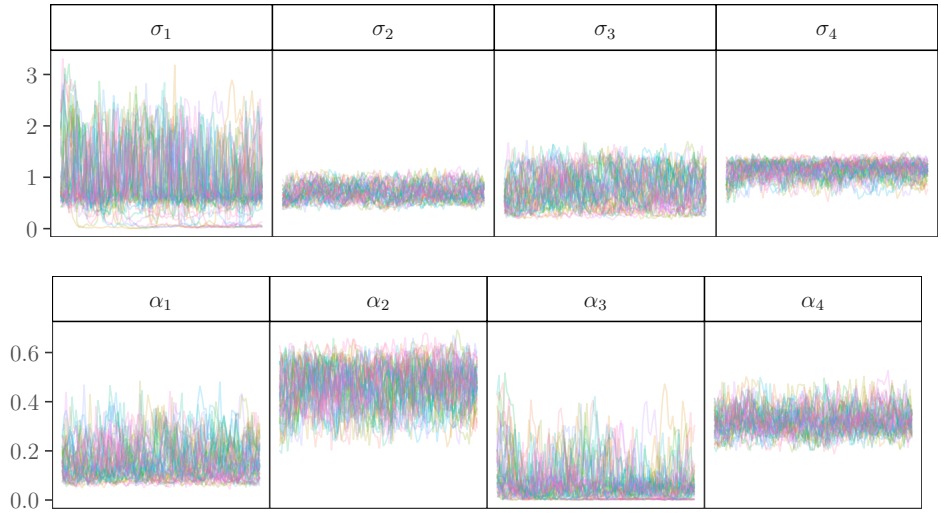


Figure 3.7: Convergence plot for the infant data. The plot shows 46 chains of 1000 MCMC iterations, with a thinning factor of 50.

Finally, the assumption that the means can be fixed to the cardinal directions can be checked. To test whether this was a valid assumption, a model with free means is run for both adults and infants (parameter estimates not shown). This model is then compared to the model with fixed means from before, which gives a log Bayes factor of 61 in favor of free means. This suggests that the data can be fit better by allowing the means to be freely estimated. This could be understood as a systematic bias of the means away from the cardinal directions contained in the stimuli used. However, for the sake of simplicity, the means were kept fixed throughout this paper.

3.4 Discussion

In this paper, a new mixture model for flexible circular distributions was developed. It can be used to distinguish clusters in samples of directions, for example those obtained from saccade directions. The main contribution is the development of mixture models for circular data that allow for peaked and flat-topped shapes. In order to do this, a new family of distributions was introduced, the Power Batschelet distributions, that mimic the distributions developed in Jones & Pewsey (2012), but that enjoy more appealing computational properties.

The method developed here can be used as a method to investigate

Table 3.2: Free-viewing fit using the Power Batschelet mixture model, for adults (left) and infants (right). Mean directions were fixed at $\mu_1 = -\pi/2$ (upward), $\mu_2 = 0$ (rightward), $\mu_3 = \pi/2$ (downward), $\mu_4 = \pi$ (leftward).

		Adults						Infants					
		EM	Boot. CI		Bayes (MCMC)			EM	Boot. CI		Bayes (MCMC)		
		Est.	2.5%	97.5%	Median	2.5%	97.5%	Est.	2.5%	97.5%	Median	2.5%	97.5%
Up	κ_1	0.81	0.69	1.03	5.94	0.85	666.66	0.38	0.25	0.91	6.76	0.14	635.15
	λ_1	1.00	1.00	1.00	0.41	-0.53	0.97	-1.00	-1.00	0.17	-0.56	-0.97	0.55
	α_1	0.11	0.09	0.12	0.02	0.01	0.12	0.15	0.11	0.18	0.05	0.00	0.28
	σ_{c1}	1.54	1.37	1.64	0.32	0.09	1.46	1.96	1.46	2.13	0.68	0.03	2.33
Right	κ_2	3.05	2.61	3.50	2.46	2.09	3.39	2.34	1.90	3.75	2.22	1.49	3.78
	λ_2	0.67	0.56	0.78	0.77	0.58	0.91	0.72	0.40	0.90	0.67	0.38	0.93
	α_2	0.32	0.31	0.35	0.36	0.31	0.40	0.27	0.23	0.32	0.32	0.23	0.43
	σ_{c2}	0.48	0.42	0.58	0.62	0.44	0.76	0.65	0.43	0.84	0.69	0.43	1.05
Down	κ_3	1.93	1.65	3.34	4.33	1.51	30.56	1.41	1.21	1.94	2.14	0.90	18.24
	λ_3	0.76	0.44	0.90	0.41	-0.20	0.85	0.99	0.61	1.00	0.65	-0.09	0.96
	α_3	0.10	0.07	0.12	0.06	0.05	0.14	0.18	0.15	0.20	0.13	0.06	0.34
	σ_{c3}	0.80	0.47	0.95	0.38	0.23	1.01	1.12	0.80	1.23	0.73	0.25	1.42
Left	κ_4	1.84	1.70	1.97	1.62	1.50	2.02	1.34	1.18	1.49	1.28	0.98	2.06
	λ_4	0.96	0.88	1.00	0.97	0.87	1.00	0.98	0.82	1.00	0.88	0.61	0.99
	α_4	0.47	0.44	0.50	0.54	0.43	0.58	0.41	0.37	0.46	0.47	0.28	0.61
	σ_{c4}	0.89	0.81	0.96	1.00	0.79	1.07	1.16	1.04	1.26	1.16	0.75	1.39

whether two sets of saccade directions differ from each other, as shown in Section 3.3.2. This allows eye-tracking researchers to answer more complex questions about their saccade data, and draw inference where this was previously not possible.

Although developed in the context of saccade directions, the method has potentially much broader applications. For example, wind directions are commonly modeled with circular distributions (Bowers et al., 2000; Holzmann et al., 2006; Bao et al., 2010), and sometimes feature peaked distributions. Also, observed arrival times can sometimes be governed by an event occuring at a single time point, causing strongly peaked distributions as well. Also, the method is a strong contender for any form of nonparametric fit on a set of univariate circular data. In that case, the current method functions as a flexible parametric alternative to a fully nonparametric analysis. The mixture of Batschelet distributions then allows much more extensive interpretation and inference than a nonparametric analysis typically would, at a minor cost of flexibility.

It should be noted that in general saccade data is time-series data where each observation is correlated with the previous observations. In the free-viewing paradigm, this autocorrelation is not of core interest, as the time series consist of only 3-5 saccades per viewed image. In general, fitting the saccade directions with a flexible model such as the one provided here while ignoring the time-series structure violates an assumption of the model. However, the time series structure of saccade directions usually does not follow a traditional autocorrelation structure, because, for example, reaching the end of a page results in negative autocorrelation. Therefore, this issue should

be addressed separately. The current approach does allow a powerful parametric comparison of different groups of saccade direction data, even when ignoring the autocorrelation. For some applications however, interest might specifically lie in the autocorrelation structure, and for those settings different models might be preferred.

In future studies, it may be fruitful to model the saccade direction jointly with saccade length, instead of the saccade direction separately. One promising model that has not been applied to the field of eye-tracking is the Abe-Ley model for cylindrical data (Abe & Ley, 2017). In this model, saccades with larger length naturally have a higher (circular) concentration, a property which is commonly observed in vision research. However, the complex form of such models means that extensions such as mixtures are not available, although a hidden Markov model has been developed (Lagona et al., 2015).

Finally, it should be noted that methods developed here are made accessible to eye-tracking researchers through the easy-to-use R package `flexcircmix`. This means that the methods can be readily applied to new data, without requiring extensive technical knowledge. Hopefully, the methods developed in this paper will provide a valuable new direction for eye-tracking researchers to perform more valid parametric inference on the angles of saccades throughout a broad range of applications.

3.A Properties of the Power Batschelet Distribution

The probability density function of the Power Batschelet distribution is defined as

$$f_{PB}(\theta \mid \mu, \kappa, \lambda) = [K^*_{\kappa,\lambda}]^{-1} \exp\{\kappa \cos t^*_\lambda(\theta - \mu)\}, \tag{3.21}$$

where

$$t^*_\lambda(\theta) = \text{sign}(\theta)\pi \left(\frac{|\theta|}{\pi}\right)^{\gamma(\lambda)}, \tag{3.22}$$

with $\gamma(\lambda) = \frac{1-c\lambda}{1+c\lambda}$, where $c = 0.4052284$ and the inverse of the normalizing constant being

$$K^*_{\kappa,\lambda} = \int_{-\pi}^{\pi} \exp\{\kappa \cos t^*_\lambda(\theta - \mu)\} d\theta, \tag{3.23}$$

which is usually numerically integrated.

Note that for any symmetric base density, such as the von Mises used in this case, the sign(θ) in Equation 3.22 is optional, because $f_{PB}((\theta - \mu) \mid \mu, \kappa, \lambda) = f_{PB}(-(\theta - \mu) \mid \mu, \kappa, \lambda)$.

The log-likelihood is

$$\ell(\mu, \kappa, \lambda \mid \boldsymbol{\theta}) = -\log[K^*_{\kappa,\lambda}] + \kappa \sum_{i=1}^{n} \cos t^*_\lambda(\theta_i - \mu), \tag{3.24}$$

so that, assuming $\theta_i \neq \mu \; \forall \; i \in 1, \ldots, n$, the score functions are

$$\frac{\partial \ell(\mu, \kappa, \lambda \mid \boldsymbol{\theta})}{\partial \mu} = \kappa \pi^{1-\gamma(\lambda)} \gamma(\lambda) \sum_{i=1}^{n} |\theta_i - \mu|^{\gamma(\lambda)-1} \sin t^*_\lambda(\theta_i - \mu) \tag{3.25}$$

$$\frac{\partial \ell(\mu, \kappa, \lambda \mid \boldsymbol{\theta})}{\partial \kappa} = -\left[K^*_{\kappa,\lambda}\right]^{-1} \int_{-\pi}^{\pi} \cos t^*_\lambda(\theta) e^{\kappa \cos t^*_\lambda(\theta)} d\theta + \sum_{i=1}^{n} \cos t^*_\lambda(\theta_i - \mu) \tag{3.26}$$

$$\frac{\partial \ell(\mu, \kappa, \lambda \mid \boldsymbol{\theta})}{\partial \lambda} = -\left[K^*_{\kappa,\lambda}\right]^{-1} \int_{-\pi}^{\pi} h(\theta, \mu, \kappa, \lambda) e^{\kappa \cos t^*_\lambda(\theta)} d\theta + \sum_{i=1}^{n} h(\theta_i, \mu, \kappa, \lambda) \tag{3.27}$$

where

$$h(\theta, \mu, \kappa, \lambda) = \frac{\partial \kappa \cos t^*_\lambda(\theta - \mu)}{\partial \lambda} \tag{3.28}$$

$$= \frac{\kappa \gamma'(\lambda) |\theta - \mu|^{\gamma(\lambda)}}{\pi^{\gamma(\lambda)-1}} \sin\left(\frac{|\theta - \mu|^{\gamma(\lambda)}}{\pi^{\gamma(\lambda)-1}}\right) \log\left(\frac{\pi}{|\theta - \mu|}\right), \tag{3.29}$$

with $\gamma'(\lambda) = \frac{2c}{(1+c\lambda)}$. Due to the form of the score functions, as well as the condition that $\theta_i \neq \mu \ \forall \ i \in 1, \ldots, n$, it is clear that the Hessian and Fisher Information will not be easy to work with. Therefore, it is preferred to work directly with the log-likelihood.

Chapter 4

Bayesian inference for mixtures of von Mises distributions using the Reversible Jump MCMC sampler

Abstract

Circular data are encountered in a variety of fields, but circular statistics tend to focus on unimodal distributions. A dataset on music listening behavior throughout the day motivates development of models for multi-modal circular data. As the number of required components is not known ahead of time, the reversible jump MCMC algorithm is adapted for circular data and presented. The performance of this sampler is investigated in a simulation study. At small samples ($n \leq 100$), the number of components is uncertain. At larger sample sizes ($n \geq 500$) the estimation of the number of components is accurate. Application to the music listening data shows interpretable results that correspond with intuition.

This chapter is submitted to Journal of Statistical Computation and Simulation as Mulder, K., Jongsma, P., Klugkist, I. (tbd) Bayesian inference for mixtures of von Mises distributions using the Reversible Jump MCMC sampler.

Author contributions: KM, PJ and IK designed the study. PJ and KM developed the statistical methods. PJ programmed the sampler and performed the simulation study. KM and PJ wrote the paper. KM and IK provided feedback on written work.

Circular data are data measured in angles or orientations in two-dimensional space. Examples include directions on a compass (0° - 360°), time of day (0 - 24 hours) or day of year (0 - 365 days). These data are encountered across behavioral research (Mechsner et al., 2001b; Gurtman, 2009b) and many other scientific disciplines.

Analysis of circular data requires special statistical methods due to the periodicity of the sample space. For example, the arithmetic mean of the two time points 00:30h and 23:30h would be 12:00h, while the circular mean is 00:00h, which is clearly a preferable central tendency in this case. Several books on circular statistics are available, in particular Pewsey et al. (2013), Mardia & Jupp (2009) and Fisher (1995).

This paper will focus on the modeling of multi-modal circular data with mixtures with an unknown number of components. Mixture models often assume the number of mixture components to be known, although this is rarely true in practice. As a solution, the number of mixture components is usually selected by comparing information criteria such as the AIC (Akaike, 1974) or BIC (Schwarz, 1978). Such an approach allows selection of a mixture model with the best-fitting number of components, but entirely ignores our uncertainty about the parameter determining the number of components.

A more natural and sophisticated approach is to treat the number of modes as unknown, and obtaining the uncertainty around the number of modes jointly with the rest of the analysis. The main contribution of this paper is to provide an algorithm to perform a fully Bayesian mixture model that correctly captures the uncertainty about the number of components, as well as showing its usefulness and interpretability with a real data example.

The motivating example for this paper is a data set on music listening behavior. The data was provided by music service `http://www.22tracks.com/`, and consists of the time of day (on the 24-hour clock) at which a user played a particular song. The songs that are being listened to are also categorized in genres such as 'Pop' or 'Deep House'. When a user visits the *22tracks* service, they are presented with a genre. The user can then choose to listen to this genre, select a different genre or stop using the service. Reducing the fraction of users that stop using the site is of direct interest to the music service. Currently, the initially presented genre is selected uniformly random over the genres. The aim of our analysis is to determine which genres are listened to most at certain times, so the most relevant genre at a given time can be presented. In addition, a model where parameters can be directly interpreted may allow us to understand what drives music listening behavior.

The base distribution for our circular mixture model will be the von Mises distribution, which is commonly used for analysis of circular data and can be considered the circular analogue of the Normal distribution. Von Mises mix-

ture models with a fixed number of modes have been developed previously in a frequentist setting (A Mooney et al., 2003), for example using the Expectation Maximization (EM) algorithm (McLachlan & Krishnan, 2007; Banerjee et al., 2005). Bayesian analysis for this type of model can be performed through Markov chain Monte Carlo (MCMC) sampling (Tierney, 1994; Besag et al., 1995). In particular, the high-dimensional variant of this mixture model has seen some popularity due to several appealing applications such as text mining, which has led to an R package for this model (Hornik & Grün, 2014). Such high-dimensional circular mixtures use the von Mises-Fisher distribution on hyperspheres, with the von Mises as a special case. We will focus on the circular mixture model only.

The core difficulty in employing MCMC samplers in such applications is that the parameter space is of variable dimension. That is, if there are more components in the mixture model, there are more parameters. Therefore, the usual MCMC approaches do not provide a way to explore the whole parameter space, and we must use a solution such as Reversible jump MCMC (Green, 1995; Richardson & Green, 1997b). In a reversible jump MCMC sampler, we allow moves between parameter spaces by use of a special case of the Metropolis-Hastings (MH) acceptance ratio (Hastings, 1970b). This paper provides a detailed account of the adaptation of a reversible jump sampler for von Mises mixtures.

Three major contributions are made to the field of mixture modeling for circular data. First, this paper presents the first application of the reversible jump sampler to this setting, which allows us to perform inference on the amount of components in the mixture model. Second, a novel split move, which makes use of the trigonometric properties of the von Mises distribution, allows the sampler to move across the parameter space efficiently. Lastly, a simulation study is performed to show that this method performs well in common research scenarios.

Several alternative approaches for Bayesian modeling of multi-modal circular data could be considered. Most are found in the field of Bayesian non-parametrics, such as Dirichlet process mixture models (Ghosh et al., 2003b), log-spline distributions (Ferreira, Juárez, & Steel, 2008) or a family of densities based on non-negative trigonometric sums (Fernández-Durán & Mercedes Gregorio-Domínguez, 2016b). Such approaches generally have the advantage of making fewer assumptions about the distribution of the data. However, none of these methods provide a way for direct inference on the number of subpopulations (ie. components) making up the mixture, and the parameters in the mixture model with unknown number of components are much more interpretable.

Therefore, the approach taken in this paper can be seen as a useful in-

between step between mixture models with a fixed number of components and non-parametric approaches. Compared to fixed-component mixture models, our approach is more realistic in the uncertainty about the amount of components, allows performing inference about the number of components, but also enables leaving the number of components to be uncertain. In particular, while information criteria based methods also allow selection of the most likely number of components, our approach provides a posterior probability distribution around the number of components, and as such acknowledges that the selected number of components can be wrong. Compared to non-parametric approaches, our approach feature much more interpretable parameters and inference, at the cost of taking more assumptions about the shape of the distribution. Concluding, if the number of components is known or not of interest, a fixed-component mixture model can be preferred for simplicity, while if density estimation is the goal a fully non-parametric approach may be the best choice. If the number of components is not known, of interest, and interpretation of the parameters of subpopulations is of interest, our method aligns the best with these goals.

The paper is organized as follows. Section 4.1 describes the model and chosen priors. Section 4.2 contains the description and implementation of each of the steps involved in sampling the model parameters. In Section 4.3 the performance of the sampler is investigated in a simulation study. The sampler is applied to the *22tracks* data in Section 4.4. Finally, the results are discussed in Section 4.5.

4.1 Von Mises Mixture Model

In this section, the von Mises-based mixture model will be developed. First, its general form will be given. Second, the likelihoods necessary for inference are discussed. Third, priors for this model are shortly discussed.

4.1.1 Von Mises mixture density

The von Mises distribution is a symmetric, unimodal distribution commonly used in the analysis of circular data. Its density is given by

$$f_{VM}(\theta \mid \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \mu)), \tag{4.1}$$

where $\theta \in [0, 2\pi)$ is an angle measured in radians, $\mu \in [0, 2\pi)$ is the mean direction, $\kappa \in [0, \infty)$ is a non-negative concentration parameter and $I_0(\cdot)$

is the modified Bessel function of the first kind and order zero. For an introduction into circular statistics, see (Mardia & Jupp, 2009).

When data consist of observations from multiple subpopulations for which the labels are not observed, the distribution of the pooled observations can be described by a mixture model. For example, times at which people listen to music are expected to coincide with daily events, such as dinner time or the daily commute, which are clustered around certain time points that show up as modes in the data set.

The density of the pooled observations can be expressed as a mixture

$$f(\theta \mid \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}) = \sum_{j=1}^{g} w_j f_{VM}(\theta \mid \mu_j, \kappa_j), \tag{4.2}$$

where $g \in \mathbb{N}^+$ is the number of components in the mixture, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_g\}$ and $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_g\}$ are vectors of distribution parameters of each von Mises component and weight vector $\boldsymbol{w} = \{w_1, \ldots, w_g\}$ contains the relative size of each component in the total sample. Weights, which are also sometimes called mixing probabilities, satisfy the usual constraints to lie on the simplex, that is $0 \leq w_j \leq 1$ and $\sum_{j=1}^{g} w_j = 1$.

4.1.2 Likelihood

For a dataset $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_N\}$ of observations from a mixture of von Mises components, the likelihood is

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, g \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{j=1}^{g} w_j f_{VM}(\theta_i \mid \mu_j, \kappa_j), \tag{4.3}$$

where it should be noted that the number of components g is treated as an unknown parameter instead of fixed, and that the lengths of $\boldsymbol{\mu}, \boldsymbol{\kappa}$ and $\boldsymbol{w}$ depend on g.

In order to perform inference on ($\boldsymbol{w}$, $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$, g), it will be convenient to include a latent vector $\boldsymbol{z} = \{z_1, \ldots, z_N\}$ that encodes the component to which observation θ_i is attributed. Parameters $z_1, \ldots, z_N$ are realizations of categorical random variable $Z_1, \ldots, Z_N$, such that

$$P(Z_i = j \mid \boldsymbol{w}) = w_j, \qquad (i = 1, \ldots, N; j = 1, \ldots, g). \tag{4.4}$$

The reason for introducing this parameter vector is that conditional on $\boldsymbol{Z}$, $\theta_1, \ldots, \theta_N$ are independent observations from their respective component

$$p(\theta_i \mid Z_i = j, \boldsymbol{\mu}, \boldsymbol{\kappa}) = f_{VM}(\theta_i \mid \mu_j, \kappa_j), \tag{4.5}$$

where inference for μ_j and κ_j is markedly easier than in the mixture likelihood in Equation 4.3, because it can be done as if the model is simply a single von Mises component. The vector $\boldsymbol{z}$ is called the allocation vector and will be updated as part of the MCMC procedure. With this allocation vector, the expression for the likelihood of the parameters of each component is simply

$$\mathcal{L}(\mu_{z_i}, \kappa_{z_i} \mid \boldsymbol{\theta}, \boldsymbol{z}) = \prod_{i=1}^{N} f_{VM}(\theta_i \mid \mu_{z_i}, \kappa_{z_i}). \tag{4.6}$$

As per usual in the Bayesian framework, inference will be performed on the posterior distribution, which is given by

$$p(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{z} \mid \boldsymbol{\theta}, g) \propto p(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{z}, g)\mathcal{L}(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{z}, g \mid \boldsymbol{\theta}), \tag{4.7}$$

where the prior $p(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{z}, g)$ will be discussed next.

4.1.3 Prior distributions

Although informative priors could be used in practice, we will focus on providing uninformative priors for the parameters of the von Mises components and their weights. The joint prior $p(\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, g)$ is assumed to factor into several independent priors, which will be discussed in turn.

For the von Mises parameters μ_j and κ_j, a conjugate prior (Mardia & El-Atoum, 1976; Guttorp & Lockhart, 1988) is used, which is given uninformative prior hyperparameters. For μ_j, this is the circular uniform distribution, which we will write as $p(\mu_j) \sim \mathcal{U}(0, 2\pi)$, where $U(a, b)$ is the uniform distribution from a to b.

For κ_j this is a constant prior $p(\kappa_j) \propto 1$. Both priors represent a lack of knowledge about these parameters. A more informative prior for κ_j can also be set in the conjugate prior, for example if highly concentrated von Mises distributions are not expected to represent real subpopulations.

The prior for $\boldsymbol{w}$ is the Dirichlet distribution $p(\boldsymbol{w}) = \mathcal{D}(1, 1, \ldots, 1)$, which assigns equal probability to all combinations of weights.

The prior for the number of components g is chosen as $p(g) \propto \text{geom}(0.05)^N$ such that $p(g) \propto 0.05(1-0.05)^{gN}$. The geometric distribution is raised to the power N, the number of observations, as a method for penalizing complexity. While somewhat of a pragmatic choice, this prior performs well in practice. The prior prevents overfitting and can be interpreted as the belief that a parsimonious model is preferred, irrespective of the number of observations.

4.2 Reversible jump MCMC for von Mises Mixtures

Bayesian inference for the von Mises mixture model will proceed by sampling from the posterior in Equation 4.7 using MCMC sampling. As mentioned, standard MCMC will not be able to deal with the changing dimensionality in the parameter space after g changes, and therefore we will resort to reversible jump MCMC to solve this issue.

The reversible jump MCMC algorithm consists of five move types. These moves can be divided into fixed-dimension move types and dimension changing move types. The fixed-dimension moves are the standard moves for MCMC on mixture models. They do not change the component count g and thus do not alter the dimensionality of the parameter space. These moves are

1. updating the weights $\boldsymbol{w}$;
2. updating component parameters ($\boldsymbol{\mu}$, $\boldsymbol{\kappa}$);
3. updating the allocation $\boldsymbol{z}$.

When g is known for a mixture of von Mises components, a sampler consisting of just these three moves would be sufficient.

In many cases however, g is not known and should be estimated as part of the MCMC procedure. This can be achieved by including two more move types, which are the reversible jump move types. They are

4. splitting a component in two, or combining two components;
5. the birth or death of an empty component.

Both of these move types change g by 1 and update the other parameters ($\boldsymbol{w}$, $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$, $\boldsymbol{z}$) accordingly. In our implementation, the move types 1-5 are performed in order. One complete pass over each of these moves will be called an *iteration* and is the time step of the algorithm. The chosen implementations of these move types will be discussed in detail in the following sections.

4.2.1 Updating the weights w

Weights $\boldsymbol{w}$ can be drawn directly from their full conditional distribution $p(\boldsymbol{w} \mid \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{z}, g)$, which is Dirichlet and dependent only on the current allocation $\boldsymbol{z}$. It is given by

$$\boldsymbol{w} \mid \boldsymbol{z} \sim \mathcal{D}(n_1 + 1, ..., n_g + 1), \tag{4.8}$$

where n_j is the number of observations allocated to component j

$$n_j = \sum_{i=1}^{N} \mathbf{1}_{z_i=j}, \tag{4.9}$$

where $\mathbf{1}$ is an indicator function.

4.2.2 Updating component parameters μ and κ

The conditional posterior distribution of each μ_j is von Mises and given by

$$\mu_j \mid \kappa_j, \boldsymbol{\theta}_j \sim VM\left(\bar{\theta}_j, R_j\kappa_j\right), \tag{4.10}$$

where $\boldsymbol{\theta}_j$ is the vector of observations currently allocated to component j and $\bar{\theta}_j$ and R_j are the mean direction and the resultant length respectively, which can be computed as in (Mardia & Jupp, 2009, p. 15).

The conditional distribution of κ can be expressed as

$$f(\kappa_j \mid \mu_j, \boldsymbol{\theta}_j) \propto I_0(\kappa_j)^{-n_j} \exp\left\{R_j\kappa_j \cos\left(\mu_j - \bar{\theta}\right)\right\}. \tag{4.11}$$

It is not straightforward to sample from this distribution. The method proposed by Forbes & Mardia (2014) is applied, which uses a rejection sampler to produce a sample from the full conditional distribution.

4.2.3 Updating the allocation z

Allocation z_i for each observation is sampled based on the relative densities of the components. For observation i this is given by

$$P(Z_i = j \mid \theta_i, w_j, \mu_j, \kappa_j) = \frac{w_j f_{VM}(\theta_i \mid \mu_j, \kappa_j)}{\sum_{h=1}^{g} w_h f_{VM}(\theta_i \mid \mu_h, \kappa_h)}. \tag{4.12}$$

This is the categorical or 'multinouilli' distribution, and is simple to sample from.

4.2.4 Dimensionality changing moves

For the dimensionality changing moves we make use of reversible jump moves which are a special case of a Metropolis-Hastings step (Richardson & Green, 1997b). The goal is to allow the sampler to move from a current state, which we'll denote by $x = (\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{z})$, to another state x', which has a different number of dimensions than x.

This move is implemented by sampling a random vector $\boldsymbol{u}$ that is independent of x, the current state of the sampler. The proposal for a new state x' can then be expressed as an invertible function $x'(x, \boldsymbol{u})$, to be chosen later, which maps x and u jointly to a proposal x'. It is required that the move is designed as a pair, such that there also exists the reverse function $x(x', \boldsymbol{u})$, which is why the algorithm is called *reversible jump*. Essentially, we develop a bridge between two spaces of different dimension, and as a result are able to change the dimensionality.

Given a random vector u and the invertible function $x'(x, \boldsymbol{u})$, we can accept or reject the proposal x' using a Metropolis-Hastings (MH) acceptance ratio, which can be written as

$$\min\left\{1, \frac{p(x'|y)}{p(x|y)} \frac{r_m(x')}{r_m(x)q(\boldsymbol{u})} \left|\frac{\delta x'}{\delta(x, \boldsymbol{u})}\right|\right\}, \tag{4.13}$$

where $p(x' \mid y)/p(x \mid y)$ is the ordinary ratio of posterior probability of states x and x', $r_m(x)$ is the probabilty of choosing move type m from state x, and $q(\boldsymbol{u})$ is the density function of $\boldsymbol{u}$, and the final term $\left|\frac{\delta x'}{\delta(x,\boldsymbol{u})}\right|$ is the Jacobian that arises from the change of parameter space from $(x,\ \boldsymbol{u})$ to x'.

The reversible jump moves will dictate how to sample proposals for a new state x' given a vector $\boldsymbol{u}$, after which the proposal is accepted or rejected based on the MH ratio just described. The precise form of this MH ratio depends on the move type and the chosen function $x'(x, \boldsymbol{u})$. Developing the move types and their associated invertible functions represents a large chunk of the work involved in implementing the reversible jump algorithm for a specific model. Next, some sensible choices for the von Mises model will be discussed.

Split or combine move

The split or combine move is designed as a reversible pair, as is required in the reversible jump framework. That is, any proposed split move is associated with a combine move that would undo it. A split move takes one component and replaces it with two new components. Conversely, a combine move joins two existing components into a single component.

Constructing split/combine proposals for reversible jump MCMC samplers can be done using moment matching (Brooks et al., 2003), where the moments of a combined component are defined to be the sum of the moments of the split components. In the case of von Mises components, this is not straightforward, because the second (linear) moment of a von Mises distribution is mathematically intractable. Rather, trigonometric moments can be

used. The first trigonometric moments of a von Mises component with parameters μ and κ are given by $\alpha = E[\cos(\theta)] = \rho\cos(\mu)$ and $\beta = E[\sin(\theta)] = \rho\sin(\mu)$, where the mean resultant length is $\rho = A(\kappa) = I_1(\kappa)/I_0(\kappa)$ and $A(\kappa)$ can be approximated (Mardia & Jupp, 2009, p. 40).

To make sure that the trigonometric moments represent a valid von Mises distributions, the point described by (α, β) must lie on the unit disc, which means that $-1 \leq \alpha \leq 1$, $-1 \leq \beta \leq 1$, and most importantly

$$\sqrt{\alpha^2 + \beta^2} \leq 1. \tag{4.14}$$

This is important for the reversible jump algorithm, because whenever any dimensionality changing move occurs, any new component must also satisfy these constraints.

The constraint of mapping to valid von Mises components, along with the reversibility condition, limit the set of possible moves. However, any move that follows these limitations will be valid in the sense that it will correctly sample from the desired posterior. As long as the limitations are met, we are free to select move types based on computational efficiency, for example. Computational efficiency will be attained when the proposals are likely to be accepted, which in turn is more likely when the proposals are in some sense 'close' to the original components. This will lead the specific choices for the combine and split moves, which will be discussed next.

Combine move

In the combine move, two current components, say j_1 and j_2, are combined into a single new component j^*. The combine move can be obtained from a simple weighted sum of the trigonometric moments. That is, the new combined component has trigonometric moments that are a weighted average between the two components that it stems from.

The parameters of the new component are defined by their trigonometric moments and component weight, $(w_{j^*}, \alpha_{j^*}, \beta_{j^*})$, by computing

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2}, \\ w_{j^*}\alpha_{j^*} &= w_{j_1}\alpha_{j_1} + w_{j_2}\alpha_{j_2}, \\ w_{j^*}\beta_{j^*} &= w_{j_1}\beta_{j_1} + w_{j_2}\beta_{j_2}. \end{aligned} \tag{4.15}$$

It can be shown that the $(w_{j^*}, \alpha_{j^*}, \beta_{j^*})$ correspond to a valid von Mises distribution and weight, due to the convexity of the unit disc (for $\alpha_{j^*}, \beta_{j^*}$) and the convexity of the unit interval for the weight w_{j^*}.

Split move

In the split move, we start from joint component j^* and split it into two components, j^1 and j^2. The split move must also conform to (4.15) to fulfill the requirement of reversibility, but we must be more careful than in the combine move to prevent the trigonometric moments falling outside the allowed range. We will solve this by proposing the split components from the largest possible disc that is centered at the trigonometric moments of j^*, while being covered by the unit disc. This last property ensures that all proposals are valid.

Next we will discuss how exactly we draw the proposals from within this disc. We can do this by drawing vector $\boldsymbol{u}$ from

$$u_1 \sim \mathcal{U}(0, 0.5) \qquad u_2 \sim \mathcal{U}(0, 2\pi) \qquad u_3 \sim \text{Beta}(2, 1), \tag{4.16}$$

where $\text{Beta}(a, b)$ is the beta distribution. After drawing this vector, we can obtain our split components by computing

$$\begin{aligned} \rho_{max} &= (1 - \rho_{j^*})u_3 \\ w_{j_1} &= w_{j^*}u_1 & w_{j_2} &= w_{j^*}(1 - u_1) \\ \alpha_{j_1} &= \rho_{j^*} - \cos(u_2)\rho_{max} & \alpha_{j_2} &= \rho_{j^*} + \cos(u_2)\rho_{max}w_{j_1}/w_{j_2}, \\ \beta_{j_1} &= \sin(u_2)\rho_{max}, & \beta_{j_2} &= -\sin(u_2)\rho_{max}w_{j_1}/w_{j_2}. \end{aligned} \tag{4.17}$$

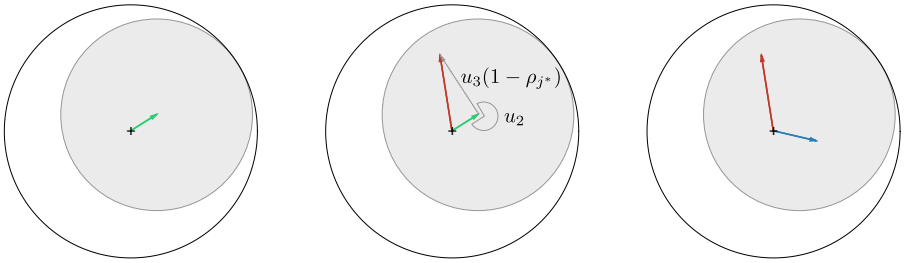
As discussed, different choice are possible, but these were found to perform well in practice. To aid understanding, this procedure is given a visual representation in Figure 4.1, which will be discussed step by step next.

In step 1 (4.1a), the von Mises component j^* is represented by its trigonometric moments α_{j^*} and β_{j^*} as an arrow. The two new components' trigonometric moments must fall inside the unit circle, as to satisfy constraint (4.14). To do this, a disc with radius $1 - \rho_{j^*}$ centered at $(\alpha_{j^*}, \beta_{j^*})$ is indicated in grey in the figure, from which the split components will be sampled.

In step 2 (4.1b), the first new von Mises component j_1 is placed relative to the original component. The random direction u_2 determines in what direction the new trigonometric moment of j_1 will lie. The trigonometric moments of the proposal $(\alpha_{j_1}, \beta_{j_1})$ are then chosen to lie in this direction, a distance of $u_3(1 - \rho_{j^*})$ away from $(\alpha_{j^*}, \beta_{j^*})$.

Step 3 (4.1c) places the second new von Mises component. Given the original component j^* and the first new component j_1, the moments for the second component j_2 are placed. They are found in the opposite direction from $(\alpha_{j^*}, \beta_{j^*})$, that is $u_2 + \pi$. The distance is determined depending on the ratio of the two weights. This can be computed as given in (4.15).

The probability of performing a split move as opposed to a combine move $r_m(x)$ is set to $\frac{1}{2}$, independent of the current state of the MCMC sampler. It



(a) Step 1 (b) Step 2 (c) Step 3

Figure 4.1: Construction of split proposal. Step 1 illustrates the mean resultant vector for the von Mises component to be split. In Step 2, the first split component is determined as a function of random vector $\boldsymbol{u}$. Step 3 shows the first and second split component, where the second split component follows from the combination of trigonometric moments.

then follows that the probability of the corresponding combine move $r_m(x') = 1 - r_m(x) = \frac{1}{2}$ and their ratio $r_m(x')/r_m(x) = 1$. This can result in an attempted combine move when $g = 1$, which is immediately rejected.

The Jacobian for the split move is straightforward to derive, but long in form and given by

$$\begin{aligned}\left|\frac{\delta x'}{\delta(x,u)}\right| =& \frac{(R_{j^*}-1)^2 R_{j^*} w_{j^*}(1-2u_1)^2 u_3}{u_1 - 1} \times \\ & \left(2(R_{j^*}-1)R_{j^*}\cos(u_2)u_3 + (1-2R_{j^*})u_3^2 + R_{j^*}^2(1+u_3^2)\right)^{-1/2} \times \\ & \left[(1-2R_{j^*})u_1^2 u_3^2 + R_{j^*}^2(1-2u_1+u_1^2(1+u_3^2)) - \right. \\ & \left. 2(R_{j^*}-1)R_{j^*}\cos(u_2)(u_1-1)u_1 u_3\right]^{-1/2}.\end{aligned}$$

The inverse of this Jacobian is used for the combine move.

Birth or death move

A birth move introduces a new component into the mixture, without assigning any observations to this component. Its inverse, a death move, removes a component that has no observations.

The proposal for a birth move consists of drawing parameters (w_{j^*}, μ_{j^*}, κ_{j^*}) for a new component. They are chosen from a proposal distribution as

$$v_{j^*} \sim \mathcal{U}(0,1) \qquad \mu_{j^*} \sim \mathcal{U}(0,2\pi) \qquad \kappa_{j^*} \sim \chi^2_{10}. \tag{4.18}$$

These parameters are then used to construct vector $\boldsymbol{u} = (v_{j^*}, \mu_{j^*}, \kappa_{j^*})$. The weights of the other components need to be rescaled such that the sum of

weights remains 1. The new weights are given by $w'_j = w_j(1 - v_{j^*})$, for $\{j \in 1, \ldots, g\}$.

Notably, as no observations are allocated to the newly created component, the likelihood of the data is unaltered by the move. Additionally, as with the split or combine move type, the probability of performing a birth move $r_m(x)$ is set equal to the probability of performing the corresponding death move $r_m(x')$, independent of the state of the MCMC sampler. Therefore, the acceptance probability (4.13) can be simplified to

$$\min\left\{1, \frac{p(x')}{p(x)} \frac{1}{q(u)} \left|\frac{\delta x'}{\delta(x,u)}\right|\right\}. \tag{4.19}$$

The Jacobian for a birth move is given by

$$\left|\frac{\delta x'}{\delta(x,u)}\right| = (1 - w_{j^*})^g. \tag{4.20}$$

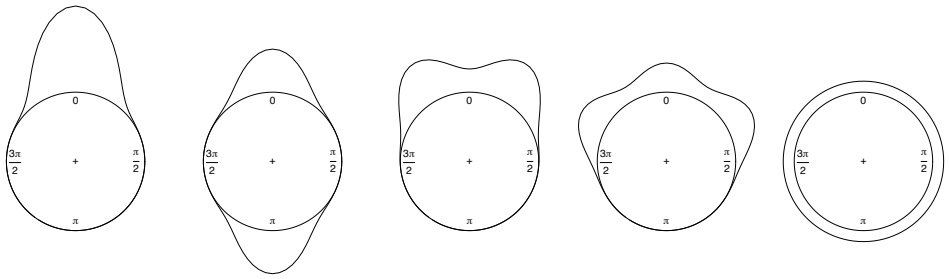
For a death move, vector $\boldsymbol{u}$ is given by the component parameters of the component that is removed j^*, $\boldsymbol{u} = (w_{j^*}, \mu_{j^*}, \kappa_{j^*})$. Its MH acceptance ratio is the inverse of the MH acceptance ratio of the corresponding birth move.

4.2.5 Label switching

When fitting a mixture model with a fixed number of components g, label switching (Jasra et al., 2005b) can occur when, for example, the means of two components are close and by random chance switch order. This can also be seen as an identifiability problem. When applying a sampler that can jump between parameter spaces and change the number of components as part of the MCMC chain, label switching is very likely to occur, regardless of the distance between means, because split moves do not define an order for the created components.

A simple method of dealing with label switching is by imposing an identifiability constraint, for example by requiring the means to be ordered. This is a crude method that does not work in the case of circular data, since two means will always be ordered the same way on a circle.

A solution is provided by M. Stephens (2000) in the form of a post-processing step to define the most likely allocation of sampled parameters to individual components and does not rely on constraints. The second post-processing algorithm featured in the paper is applied to the samples belonging to each specific g separately.



(a) (b) (c) (d) (e)

Figure 4.2: Visualization of simulation scenarios used for investigating sampler performance.

4.3 Simulation study

A simulation study was performed to investigate the relative performance of the MCMC sampler in different scenarios. The sampler was implemented in R (R Core Team, 2015) using the `circular` package (Agostinelli & Lund, 2013). The source code has been made available at `https://github.com/pieterjongsma/circular-rjmcmc`.

4.3.1 Simulation scenarios

In order to assess the performance in a variety of settings, five true data generating processes were selected to represent either common or particularly difficult mixture datasets to fit. These scenarios are visualised in Figure 4.2 and consist of (4.2a) a single von Mises component where $\mu_1 = 0$ and $\kappa_1 = 10$, (4.2b) two von Mises components where $\mu_1 = 0$, $\mu_2 = \pi$ and $\kappa_1 = \kappa_2 = 10$, (4.2c) two von Mises components where $\mu_1 = -\pi/6$, $\mu_2 = \pi/6$ and $\kappa_1 = \kappa_2 = 10$, (4.2d) three von Mises components where $\mu_1 = -\pi/3$, $\mu_2 = 0$, $\mu_3 = \pi/3$ and $\kappa_1 = \kappa_2 = \kappa_3 = 10$ and (4.2e) a uniform von Mises component ($\kappa = 0$). Each scenario is simulated with 50, 100, 250, 500, 1000, 2500 and 10,000 observations across 1000 replications.

4.3.2 Starting values

The MCMC sampler is initialized with a single component ($g = 1$) and parameters for the component drawn analogous to a birth move

$$w = 1 \qquad \mu \sim \mathcal{U}(0, 2\pi) \qquad \kappa \sim \chi^2_{10}. \tag{4.21}$$

The observations are all attributed to this single component by setting the allocation vector as $z_i = 1$ for $\{i \in 1, \ldots, N\}$.

4.3.3 Convergence

The convergence of a reversible jump MCMC algorithm is difficult to assess using conventional methods such as the inspection of the sampled values of a parameter. Due to the changing dimensionality, the posterior distributions of individual component parameters depend on the component count g. The chains for these components are expected to jump with every change of g and would therefore not provide a valid measure of convergence. Instead, the likelihood $p(x \mid \boldsymbol{\theta})$ is calculated for each iteration of the MCMC chain for which convergence of the posterior probability is assessed visually. All chains, regardless of starting value and variation in replicated data, converged within a burn-in of 10,000 iterations. After the burn-in, the next 5,000 iterations are retained and used to describe the posterior properties of the simulated dataset.

4.3.4 Results

For brevity, this section will be focused on the ability of the sampler to recover the number of components g correctly. In addition, performance of model parameters will be assessed for a subset of simulations.

Number of components g

The results for the number of components g of the simulation study are summarized in Table 4.1. It shows the fraction of replications in which the maximum a posteriori (MAP) estimate of g, g_{MAP}, which is the posterior mode, was equal to the simulated g, g_{TRUE}. Furthermore, the posterior distribution of g is given, averaged over all replications.

Results with few observations ($n = 50$) show high uncertainty about the number of components. For these replications, the mode of the posterior distribution for g was rarely equal to the simulated g. As expected, the estimation of g then improves with the number of observations. Most scenarios show a near 100% correct mode g_{MAP} at 1000 observations or more. One exception is scenario 4.2b for $n = 10000$. Here, g is overestimated and accuracy is seemingly worse than at a smaller sample size. It should be noted that an overestimation of g does not necessarily indicate a problem with the MCMC method. A model with a higher number of components may have a higher likelihood of the data. The chosen prior for g is intended to counter

Table 4.1: Simulation results for each of the scenarios in Figure 4.2 with sample sizes ranging from $n = 50$ to $n = 10000$. Each row represents 1000 replications. The fraction of replications where the estimated g was equal to the simulated g is given under $g_{MAP} = g_{TRUE}$. In addition, the posterior distribution $p(g \mid \theta)$ is given as the average over all replications.

			$p(g \mid \boldsymbol{\theta})$				
Scenario	n	$g_{MAP} = g_{TRUE}$	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g \geq 5$
4.2a	50	0.40	0.29	0.29	0.21	0.12	0.10
	100	0.73	0.54	0.32	0.11	0.03	0.00
	250	0.88	0.69	0.27	0.04	0.00	0.00
	500	0.94	0.72	0.25	0.03	0.00	0.00
	1000	0.96	0.76	0.22	0.02	0.00	0.00
	2500	0.99	0.82	0.17	0.01	0.00	0.00
	10000	1.00	0.85	0.14	0.01	0.00	0.00
4.2b	50	0.12	0.00	0.10	0.15	0.16	0.59
	100	0.66	0.00	0.50	0.31	0.12	0.06
	250	0.91	0.00	0.81	0.18	0.01	0.00
	500	0.88	0.00	0.77	0.21	0.01	0.00
	1000	0.81	0.00	0.71	0.27	0.02	0.00
	2500	0.71	0.00	0.63	0.34	0.03	0.00
	10000	0.58	0.00	0.54	0.44	0.02	0.00
4.2c	50	0.30	0.05	0.23	0.29	0.21	0.22
	100	0.66	0.08	0.51	0.30	0.09	0.03
	250	0.93	0.03	0.83	0.13	0.01	0.00
	500	0.97	0.01	0.85	0.13	0.01	0.00
	1000	0.98	0.00	0.86	0.13	0.01	0.00
	2500	0.98	0.00	0.89	0.11	0.00	0.00
	10000	1.00	0.00	0.94	0.06	0.00	0.00
4.2d	50	0.30	0.01	0.11	0.22	0.22	0.44
	100	0.60	0.01	0.25	0.45	0.21	0.08
	250	0.84	0.00	0.15	0.78	0.07	0.00
	500	0.96	0.00	0.04	0.91	0.06	0.00
	1000	0.97	0.00	0.02	0.94	0.04	0.00
	2500	0.97	0.00	0.02	0.95	0.03	0.00
	10000	0.98	0.00	0.01	0.98	0.01	0.00
4.2e	50		0.01	0.03	0.05	0.07	0.84
	100		0.15	0.25	0.25	0.17	0.18
	250		0.34	0.37	0.20	0.07	0.02
	500		0.36	0.38	0.19	0.06	0.01
	1000		0.37	0.37	0.19	0.05	0.01
	2500		0.37	0.38	0.19	0.05	0.01
	10000		0.37	0.37	0.21	0.04	0.01

Table 4.2: Parameter estimates for scenario 4.2d with parameters $\mu_1 = -1.05$, $\mu_2 = 0$, $\mu_3 = 1.05$ and $\kappa_1 = \kappa_2 = \kappa_3 = 10$.

n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_3$
50	-0.99	0.01	1.02	183.7	82.0	192.9
100	-1.04	0.00	1.07	161.8	62.6	133.9
250	-1.05	-0.01	1.05	31.6	34.7	17.8
500	-1.05	0.00	1.06	11.9	8.0	10.6
1000	-1.05	0.00	1.05	10.7	8.2	10.5
2500	-1.04	0.01	1.04	10.0	9.7	10.1
10000	-1.04	0.00	1.04	10.0	9.7	10.0

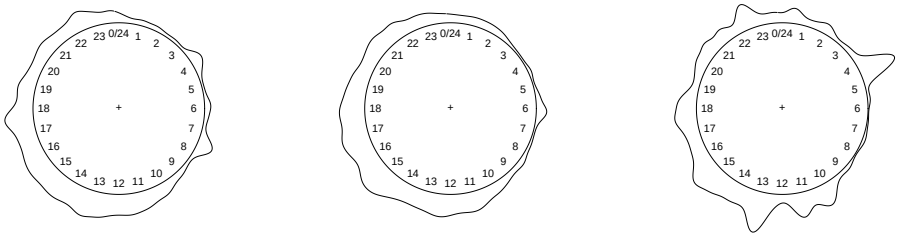
this effect, such that the simpler model is favored. The prior is seemingly not powerful enough for scenario 4.2b where $n = 1000$ or larger.

For the uniform scenario 4.2e, the column $g_{MAP} = g_{TRUE}$, showing the correspondence between the mode of the posterior distribution and the simulated g, has been omitted. Although this data was simulated as a single component with $\kappa = 0$, the interpretation of 'true' g of this distribution is ambiguous. For larger datasets, the method favors a small number of components, as expected.

Parameter estimates

Results for the recovery of the von Mises parameters $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$, are summarized for scenario 4.2d in Table 4.2. To obtain these estimates, only the MCMC states with three components are retained, so that $g = 3$ as in the data generating process of scenario 4.2d. For each simulated data set, MAP estimates for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\kappa}}$, are computed by estimating the posterior modes from the MCMC sample. Then, these MAP estimates are averaged over all simulated datasets and presented.

The estimates of component parameters for scenario 4.2d show that in general the method is able to recover the true parameter estimates without bias, even with small sample size ($n = 50$). The concentration parameter $\boldsymbol{\kappa}$ is overestimated with small samples, but estimates are reasonable for samples where $n \geq 500$. The concentration parameter of the central component, κ_2, is systematically underestimated in this scenario. Most likely, the central is assigned some observations that belong to its two neighbors, and as a result is estimated as less concentrated than the true data generating process.



(a) Indie Electronic ($n = 11,393$)

(b) Relax ($n = 10,657$)

(c) Deep House ($n = 4,757$)

Figure 4.3: Kernel density estimates of observations for analyzed *22tracks* genres, using a von Mises kernel with $\kappa = 200$. The period of the circle is 24 hours.

4.4 Illustration

As a motivating example, we will apply the reversible jump MCMC sampler to a dataset of listening behavior, made available by *22tracks*. This data will first be described and visualized in Section 4.4.1. Then, results from applying the sampler to the data are discussed in Section 4.4.2.

4.4.1 Dataset

The data provided by *22tracks* contain metrics of all users of the service over the span of one week (January 4-10, 2016). The data consist of the time of day (00:00h to 23:59h) at which a user played a particular song, categorized by the genre this song was in. In this paper, a subset of the data is used as an illustration. These consist of all observations categorized under one of three genres that were selected arbitrarily. The genres are Indie Electronic, Relax and Deep House. Figure 4.3 shows kernel density estimates of each genre. It can be seen that depending on the genre, the data features a different number of modes, although determining the precise number of modes is difficult without running the mixture model.

The first goal of this analysis is to estimate the genre that is most likely to be selected at any given time. Supposedly, users listen to the genres available on the *22tracks* service at different times of day. For example, Pop might be a genre that users listen to throughout the day while Deep House is preferred during the night. Quantifying this behavior is valuable for the music service, as it allows them to present the most appropriate genres to users when they

Table 4.3: Posterior probability of component counts $p(g)$ for selected *22tracks* genres.

	g							
Genre	1	2	3	4	5	6	7	8
Indie Electronic	0.01	0.52	0.36	0.10	0.01	0.00	0.00	0.00
Relax	0.00	0.25	0.55	0.18	0.02	0.00	0.00	0.00
Deep House	0.00	0.00	0.00	0.02	0.30	0.37	0.28	0.03

visit the site at a particular time.

The second goal is to understand music listening behavior through the parameters of our mixture components, as the times at which we listen to music are a reflection of life in our society. The mixture components are then interpreted as a subpopulation of observations that correspond to a certain category of music listening, such as listening while working, during transit, or while dancing.

4.4.2 Results

We apply the mixture model to each genre separately, with starting values as described in Section 4.3.2, using a burn-in of 10,000 iterations and retaining the next 100,000 iterations for inference. The posterior distributions for component count g are summarized in Table 4.3. The posteriors are quite different, with the Deep House genre showing a notably higher estimated component count, which is in accordance with the data as displayed in Figure 4.3.

To obtain estimates for the other parameters $\boldsymbol{w}$, $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$, only the samples for which $g = g_{MAP}$ are used. For Deep House, these are samples where $g = 6$, for Indie Electronic $g = 2$ and for Relax $g = 3$. The parameters have been summarized in Table 4.4, ordered by the weights $\boldsymbol{w}$.

The Indie Electronic genre shows two broad components spanning most of the day. One is centered at the middle of the day (14:15h) and one is centered in the evening (21:13h), most likely corresponding to listening while working and listening at home during the evening. The components have a small concentration, suggesting only slight preference for these times. Such broad components are necessary, because listening occurs throughout the day. Similarly, the Relax genre is given three components with small concentration.

For the Deep House genre, six components provided the best fit. The first three components are broad and similar to the components found for the other genres. The central times are later in the day, as one might expect

Table 4.4: Estimated component parameters for individual genres in *22tracks* data. LB and UB indicate the lower and upper bound of the 95% density of the data density of this von Mises component. Components have been ordered according to their respective weights.

					95% density	
Genre	j	$\hat{w}_j$	$\hat{\mu}_j$	$\hat{\kappa}_j$	LB	UB
Indie Electronic	1	0.72	3.73 (14:15h)	0.89	1.04 (03:58h)	0.15 (00:33h)
	2	0.28	5.55 (21:13h)	0.72	2.77 (10:36h)	2.05 (07:50h)
Relax	1	0.54	4.11 (15:43h)	1.20	1.63 (06:15h)	0.31 (01:11h)
	2	0.33	5.49 (20:59h)	0.85	2.78 (10:37h)	1.93 (07:22h)
	3	0.13	3.09 (11:49h)	1.75	1.10 (04:12h)	5.09 (19:27h)
Deep House	1	0.30	3.99 (15:15h)	2.07	2.25 (08:35h)	5.74 (21:55h)
	2	0.28	6.04 (23:04h)	1.64	3.95 (15:05h)	1.85 (07:03h)
	3	0.17	5.13 (19:36h)	1.59	2.99 (11:27h)	0.98 (03:45h)
	4	0.10	2.79 (10:41h)	33.81	2.45 (09:23h)	3.13 (11:59h)
	5	0.08	1.16 (04:27h)	635.81	1.09 (04:09h)	1.24 (04:45h)
	6	0.07	3.40 (12:58h)	648.16	3.32 (12:41h)	3.47 (13:16h)

for this type of music. The sampler was also able to detect and fit the strong concentrations of observations at 10:41h, 04:27h and 12:58h. It is unlikely that these patterns have been created by actual users. More likely, they indicate a special attribute of the data. For example, a computer bot instead of an actual person could have triggered a large amount of plays in a short time span. Although this does not tell us anything about the behavior of actual users, it is still an interesting property that the sampler detects quite well. In fact, such a component has direct financial implications for this business, as such plays can be rejected to save costs.

Compared to a kernel density model, this provides a much simpler and more interpretable summary of the data. The posterior distribution also provides uncertainty around all of these estimates, although these are not shown here for brevity.

4.5 Discussion

We have presented a method for Bayesian inference of von Mises mixture distribution. Previous work has assumed the number of components to be known, which is an assumption we have relaxed by employing the reversible jump MCMC algorithm. The main contributions included a novel set of dimensionality changing moves based on the trigonometric properties of the von Mises distribution. In addition, the performance of the method was investigated in a simulation study. Generally, the method performed well. An

illustration was provided on music listening behavior, showing the interpretation of this method.

Results of the simulation study showed that the estimation of the number of components g was accurate for the majority of the simulated sample sizes, so the proposed split and combine moves successfully move between parameter spaces. In one scenario (4.2b), g is overestimated at a very large sample count ($n = 10000$). In this case the proposed prior for g appears insufficient. A different choice of prior might be able to counter this effect and seems a topic for further investigation. It should be noted that although undesirable, an overfitted mixture is not necessarily problematic in application. The estimation of parameters $\boldsymbol{w}$, $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ of the individual components is not directly affected and these parameters remain interpretable. Furthermore, the component weights allow us to gauge the relative importance of each component.

Application to the *22tracks* data provides an example for the interpretation of reversible jump MCMC sampler output. It should be noted that the observation counts in the *22tracks* data were higher than what showed the most accurate estimation of the simulated component count in the simulation study and as such we did not infer much from the estimated g. Because the parametric von Mises model is easy to interpret, one can compare the results with intuition. The estimated component parameters $\boldsymbol{\mu}$ and $\boldsymbol{\kappa}$ in the provided example seem reasonable as they indicate listening to occur during daytime and in the evening.

In conclusion, the method presented in this paper provide a reversible jump MCMC sampler that is shown to perform well on simulated data as well as a real world example.

Chapter 5

Inference for Interval-Censored Circular Data With Application to Crime Times

Abstract

Temporal data are ubiquitous throughout crime analysis, but victims often do not know the exact crime time, instead only providing a time interval. However, methods currently available and for this type of data, called aoristic data, are shown here to be systematically biased upwards in their estimate of the spread of the data and are limited in their inferential ability. To solve these issues, we propose two other statistical approaches to recover the true crime time density. The first is based on the likelihood, and can be used to obtain maximum likelihood estimates for circular data models using aoristic data. The second is a non-parametric approach based on the Dirichlet Process that is more flexible in the types of true crime time densities it can recover. In several examples, these methods are shown work very well in practice, allowing better prediction and inference for the crime time density. The methods developed here are available in an easy-to-use R package called aoristicinference. Future work should incorporate these methods more fully into spatio-temporal models.

This chapter is in preparation for submission to Journal Quantitative Criminology as Mulder, K. and Ruiter, S. (tbd) Dealing With Partially Observed Crime Times.

Author contributions: KM and SR designed the study. KM developed the statistical methods, created the software package and performed the analyses. KM wrote the paper. SR provided the motivating data example and gave feedback on written work.

Crime analysis is concerned with understanding and predicting crime by evaluating core aspects of crimes such as their location, time, targets, offenders and the availability of guardians. When tasked with understanding or predicting whether a certain crime is likely to occur, one must know where the crime is to occur, but also at what time of the day. Burglaries, for example, are less likely to occur in the evening. Temporal variation, such as the hour of day, contributes more to the overall variation in crime likelihood than any other type of variation (Felson & Poulsen, 2003). However, in the crime analysis literature the temporal dimension is routinely ignored (Ratcliffe & McCullagh, 1998; Ratcliffe, 2000).

A central difficulty of the temporal dimension is that most crime times are not directly observed. That is, when a crime is recorded, the victim is often uncertain of when the crime occured exactly, but is only able to say that the crime happened after start time s and before some end time e. Therefore, one is left with a set of intervals of time (s, e) in which the crime occurred. Figure 5.1 shows three examples of such intervals. This type of crime time data where the exact time is not known is called *aoristic data*, with the interval (s, e) called an *aoristic interval*. Using such data severely complicates crime time analysis: if only the start times are used, the analysis is biased towards estimating crimes as happening too early, while using the end times gives estimates that are too late.

A more sophisticated approach to deal with aoristic data is what Ashby & Bowers (2013) call *aoristic analysis* (Ratcliffe & McCullagh, 1998; Ratcliffe, 2000). This method consists of giving each observation a total weight of 1, and spreading this weight out over the length of the interval. Thus, for a crime that took place in the interval 15:00 - 18:00, the hours in this interval each get a weight of $1/3$ from this observation. Therefore, the more certainty that a crime occured within a certain hour, the more probability is assigned to it. Then, the weights of all aoristic intervals are averaged within some chosen unit of time (e.g. an hour) to get an estimate of the proportion of crimes occuring during that unit of time. Because the weight is called the *aoristic fraction*, we will refer to this approach as the *aoristic fraction method* to distinguish it from other aoristic data analysis methods. Details for the aoristic fraction method are given in Section 5.2.

The goal in aoristic analyses is to obtain an estimated *crime time density*, that is, a mathematical function $\hat{p}(t)$ that takes in an interval of time and returns the proportion of crimes that will happen in that interval. If the estimate is good, we know well when crimes will happen. It should be noted that we can find estimated crime time densities at an aggregate level, or if additional predicting covariates are available, the estimated crime time density conditional on these covariates. This last approach is particularly

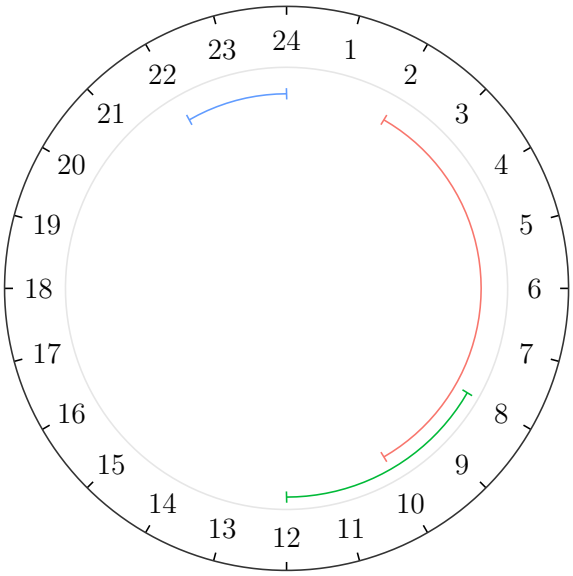


Figure 5.1: Example of three aoristic intervals representing possible crime times, which are 2 AM - 10 AM (red), 8 AM - 12 AM (green), and 10 PM - 12 PM (blue).

valuable for purposes of imputing the crime time, so that it can be used in further analysis.

From a statistical point of view, this crime time density is an estimate of a probability density function. In fact, the hope of any aoristic analysis is to approach the true probability density function of crime times, so we can evaluate any aoristic method by seeing whether it comes close to the true crime time density.

Currently available methods, such as the one described by Ashby & Bowers (2013), have several major drawbacks. The drawbacks we will address here are that the currently available methods (a) systematically overrepresent the variance of the data, (b) are not built on a solid statistical foundation so that they do not provide valid statistical inference about the true distribution of crime times, (c) do not take into account model uncertainty caused by small samples, (d) can not be embedded in a larger statistical model, (e) split time in categories which leads to arbitrary choices, and (f) do not allow mixing aoristic data with data where times are directly observed. These drawbacks limit both predictive performance as well as the use of these models for understanding patterns in crime. In this work, methods will be developed which are easier to use, built on a more solid statistical foundation, and which can be embedded in a larger model. This method is rooted in circular statistics,

which is the branch of statistics that deals with periodical sample spaces, such as the 24-hour clock in this case. It should be noted that all examples and discussion will be based on aoristic data on the 24-hour clock, but all methods are equally applicable to other time periods, such as weeks, months and years.

Throughout this paper, these drawbacks will be expanded on and addressed. The rest of this paper will be structured as follows. In Section 5.1, a general introduction to aoristic data will be given. The Aoristic Fraction method and its drawbacks will be addressed in Section 5.2. Section 5.3 will provide a way to estimate parametric statistical models using aoristic data, while Section 5.4 will add a non-parametric Bayesian method to these approaches, which is more realistic for multimodal crime time data. Section 5.5 will show how this method can be applied to several problems in crime time analysis. Finally, 5.6 has some concluding remarks.

5.1 Aoristic data

Aoristic data are temporal data that contain intervals (s, e) instead of having all observed time points. Such data arises in crime analysis when crimes happen while the victim is not present. For some types of crime close to all records are aoristic, such as for the bike thefts analyzed in Ashby & Bowers (2013) or property theft, such as burglaries. For other crimes, a combination of directly observed times (e.g. caught red-handed) and aoristic intervals is observed. This mixed data type can be a problem for some aoristic analysis methods, such as the aoristic fraction method described in the next section.

Importantly, aoristic data or temporal data refers here to the time of the day, the day of the week, month, or year. To be precise, this is the cyclic, or periodical, aspect of time. Time can also be viewed linearly, for example by making statements about crime increasing in a certain area over time, or using a time series model for the changing frequency of crime, or including year as a covariate in a model for crime. Such analyses can be combined with the aoristic methods under consideration, but are not the focus of this work.

The main issue of observing intervals of time instead of directly observing time points is that descriptive or inferential analysis methods are generally developed for directly observed data, so they do not allow intervals as inputs. Therefore, the aoristic data must be dealt with, either through a pragmatic solution or a specialized analysis method for it.

Simple ways to deal with aoristic data include ignoring the aoristic data, taking only the start points of the intervals, the end points, mid points, or a random point in the interval, which are compared in Ashby & Bowers (2013).

Removing all aoristic intervals from the data set is wasteful and often leaves us with little to no data. Using start times of the interval will clearly cause us to expect crime earlier than it really happens, while taking the end times of the interval will lead to estimates that are too late. Taking the midpoints is a better solution, but this means that we allege to know the true crime time, and as such overestimates our certainty in our analyses. Sampling a random point (uniformly) in the interval is the most valid amongst these pragmatic options, but can cause the results to depend on the random sampling. Therefore, the aforementioned aoristic fraction method is often recommended, which will be examined in the following section.

The aoristic data problem can also be seen as a missing data problem, where the crime times are not completely missing but rather partially observed. In this view, the random sampling approach essentially amounts to single imputation of missing data with the distribution of the data uniform in the interval. Single imputation is generally not recommended (Van Buuren, 2018). Commonly used missing data analysis methods use multiple imputations, and do not impute a data point uniformly from its possible places, but use the rest of the data to estimate the distribution of the missing data points, and draw imputations from this distribution. A multiple imputation approach could be a possible solution for aoristic data, but it difficult to define the crime time density within an aoristic interval based on the other aoristic intervals. Therefore, we will focus on other solutions.

Before we continue with an investigation of aoristic analysis methods, it is helpful to consider two ways in which aoristic data makes it more difficult to learn about the true crime time density.

First, the uncertainty about the location of the true crime time in the interval also means increased uncertainty in any statistical model that incorporates it. Therefore, making decisions using aoristic data requires a larger sample size than using known times. However, in this case it is still possible to create aoristic analysis methods that do not introduce any systematic bias in the estimated crime time density.

However, a systematic unobserved preference within the interval can cause bias which can not be solved, although this has not yet been investigated in previous work on aoristic analysis. To see this, note that the true crime time may have a tendency to be located near either the beginning or the end of the interval. For example, if offenders observe their victims, and strike when they leave their property, true crime times will have a tendency to occur in the beginning of the interval. If the unobserved true crime time is t_{actual}, then we can also think about the start time s as the true value minus some difference, or $s = t_{\text{actual}} - \delta^l$, and the end time e as the true time plus some difference, $e = t_{\text{actual}} + \delta^u$. Note that the differences δ^l and δ^u are also unobserved (unless

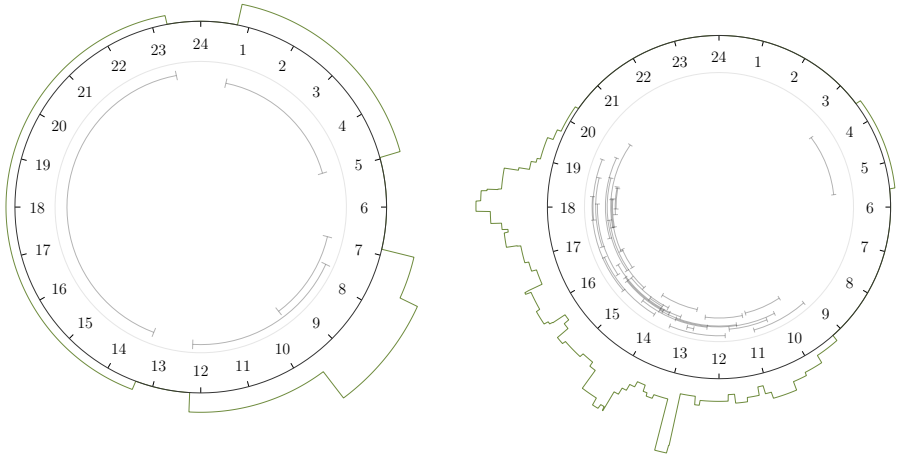
they are zero). If for some interval δ^l is larger than δ^u or vice versa, there is no problem because these will cancel out on average. However, if one is larger than another on average over the whole dataset, then any estimate of t_{actual} is systematically biased with no way for the crime analyst to detect this. From the missing data viewpoint, this is similar to the concept of missing not at random (MNAR), in that there is no information in our current dataset that can solve this issue. The only solution is to introduce outside information into the analysis, such as prior knowledge about the intervals or a sufficiently sized known-time dataset. Due to this difficulty, throughout this work we will follow other work on aoristic analysis in taking the assumption that δ^l and δ^u have the same average, that is, the true crime time is equally likely to be near the start or the end of the interval.

As a final unrelated note, temporal data (including aoristic data) is usually treated by splitting times into broad categories, such as hours ('15:00 - 16:00'), three or four groups (such as 'evening') (Pereira et al., 2016) or dichotomous splits ('5 AM - 5 PM') (Felson & Poulsen, 2003). While easy to perform, categorization requires the crime analyst to choose the amount of categories and where to place cutpoints, a choice which can influence the resulting conclusions. In addition, categorization treats adjacent categories, say '15:00 - 16:00' and '16:00 - 17:00', as being just as distant as, say, '15:00 - 16:00' and '02:00 - 03:00'. Certainly, we should be able to 'borrow information' from adjacent categories, which is especially relevant if there is limited data. In fact, from a statistical viewpoint, we can. Therefore, we will employ continuous methods throughout, because they do not pose signficant additional challenge in either statistical or computational methods.

5.2 Aoristic Fraction method

The aoristic fraction method is a descriptive method for aoristic data discussed in Ratcliffe & McCullagh (1998), which is similar various other methods that have been described over time (Gottlieb et al., 1994; Rayment, 1995; Brown, 1998). It can be described as a circular histogram that is created from aoristic intervals. While the random point method would sample a value from the uniform distribution within each interval (s, e), the aoristic fraction method keeps this uniform distribution and treats it as a building block for the histogram.

The aoristic fraction method uses the aoristic function, which is a function that captures the information in the data, as an approximation of the crime time density. The final analysis takes the form of a plot of this function, either directly on a 24-hour clock such as one displayed in Figure 5.2, which



(a) (b)

Figure 5.2: Two examples of observed aoristic intervals (grey, inside circle) and the output of the aoristic fraction method (green, outside circle) denoting the estimate of the crime time density. Left, a small data with four observations. Right, a more realistic dataset with 30 observed intervals.

shows both the observed aoristic intervals and the resulting aoristic function, or on a map such as generated by the `R` package `aoristic` (Kikuchi, 2015). To compute the aoristic function, for each observation i we take the interval (s_i, e_i) and compute the length of the interval $e_i - s_i$. Then we give all values within the observed interval weight $\frac{1}{e_i - s_i}$, so that shorter intervals have more weight. Finally, we sum these weights up for all observed intervals to obtain the estimated crime time density. Mathematically, we can write this function at time t as

$$\hat{p}_{AF}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(e_i < t < s_i)}{e_i - s_i} \tag{5.1}$$

where $I(e_i < t < s_i)$ an indicator function, which is 1 if t is in aoristic interval i, and 0 otherwise.

The aoristic function serves as estimate of the probability density at each time point of a crime occuring. That is, if the true crime times are distributed such that a crime at time point t has probability density $p(t)$, then the aoristic fraction method is an estimate of this, which we called $\hat{p}_{AF}(t)$. This estimate can then be used to make decisions, perhaps by calculating the expected percentage of crimes that will occur within a certain time frame. For example, the proportion of crimes expected to occur between 18:00 and 19:00 is $\hat{p}(18 <$

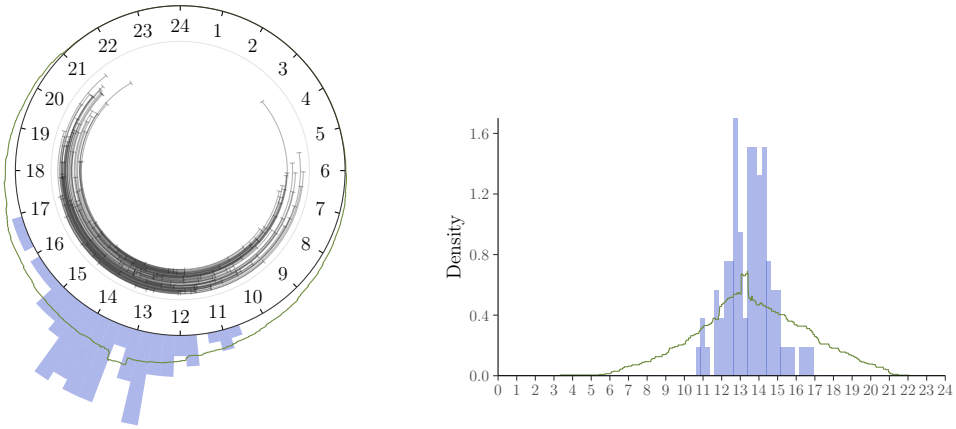
$t < 19) = \int_{18}^{19} \hat{p}_{AF}(t)dt = 0.069$, so 6.9%. This type of information is directly useful in crime prevention, such as in police resource allocation through the timing of police shifts.

While this method has several appealing properties, a major problem of the aoristic fraction method is that it systematically overestimates the variance of the crime time density. This means that any estimate derived from these methods is systematically biased towards crime times that are more spread out over the 24-hour clock. Here, this will be explained through an example, but this fact can also be proven mathematically, which is done in Appendix 5.A.

Suppose there is a neighbourhood where burglaries mostly take place in the afternoon, say, between 11:00 and 16:00. However, we are not able to observe the crime times as the victims are almost never present. Victims will provide us with the start time s they left their premises, for example leaving for work, and the end time e when they returned. For illustration, we have sampled true crime times and intervals, computed the aoristic function and have plotted these together in Figure 5.3. Due to the fact that most work schedules are similar, we will obtain a large amount of observations that have intervals roughly similar to (8:00, 19:00). However, this means that the aoristic fraction method will give a large amount of weight to the time frames 8:00 - 11:00 and 16:00 - 19:00, which there were very few true crime times in that region.

A final way to think of the aoristic function is that if we represent each aoristic interval by the uniform distribution on that interval, the aoristic function is the average of these uniform distributions at some time point t. This implies that we know nothing about where in the interval the data point is most likely to lie. That is, we do not think any point in the interval is more likely than another. However, the whole point of analyzing the dataset is to decide which crime times are more likely than others, and we are usually able to do this. Therefore, the uniform distribution on the aoristic interval is at odds with what we purport to do in our analysis.

Concluding, the aoristic fraction method is an appealing descriptive method, because its plot shows at a glance what the data looks like, but it does not allow us to infer the true crime time density. While descriptive statistics tells us something about the data, inferential statistics attempt to estimate the form of the true data generating process, which is generally the ultimate goal of crime analysis and prediction. The shortcomings of the descriptive aoristic fraction method motivate us to develop an inferential analysis method which treats the aoristic intervals as a missing data problem. That is, the true crime times are unknown, but we will try to use as much information as possible to understand where in the interval the true crime time is most likely to be.



(a) (b)

Figure 5.3: Example of how the aoristic fraction method overestimates the variance. The simulated true crime times are displayed as the blue histogram, while the approximation to this distribution is plotted as the dark green aoristic function.

5.3 Statistical models for aoristic data

In this section, inferential statistical methods will be developed for aoristic data. A crime analyst is using data to learn something about the world, and draws conclusions based on some statistical model. Initially, we are uncertain about the parameters in this statistical model, but as more data comes in we learn more about the parameters of the model and estimate them with more certainty.

In order to perform statistical inference on aoristic data, we need to set up some statistical model to learn, as well as some way to learn about them. In the next section we will shortly recap statistical models for temporal data, and in the section thereafter we will present a method to deal with the aoristic property of the data.

5.3.1 Circular data models

Statistical models for circular data, such as the temporal data under consideration here, have been developed in the field of circular statistics (Fisher, 1995; Mardia & Jupp, 2000; Pewsey et al., 2013). Circular data can consist of measurements in angles, directions, or times on the 24-hour clock, for ex-

ample. This type of data can be found throughout discliplines, such as life sciences (Mardia, 2011), behavioural biology (Bulbert et al., 2015), cognitive psychology (Kaas & Van Mier, 2006), bioinformatics (Mardia et al., 2008), political science (J. Gill & Hangartner, 2010) and environmental sciences (Lagona, 2016; Lagona et al., 2015; Arnold & SenGupta, 2006). Circular data differ from linear data in the sense that circular data are measured in a periodical sample space. For example, an angle of 1° is quite close to an angle 359°, although linear intuition suggests otherwise. Similarly, times on the 24-hour clock have 23:59 and 0:01 being close to one another, while the numerical representation suggests otherwise. As a result, models for circular data must be different from linear models, such as the Normal distribution.

A natural analogue of the Normal distribution for circular data is the von Mises distribution (Von Mises, 1918), which can be written as the probability density

$$p(t \mid \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left\{\kappa \cos(t - \mu)\right\}, \tag{5.2}$$

where t is a circular observation, such as a crime time, μ is the mean direction parameter, κ is a concentration parameter, and $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero. This is a unimodal, symmetric model for circular data which will be used as a building block for more complex models throughout this work.

In order to perform statistical inference, we use the data to estimate population values for μ and κ. In addition, we are always interested in quantifying the uncertainty in our estimates (caused by our limited sample size), which might be given as a standard error in frequentist inference or as the posterior distribution in Bayesian inference. Either way, in order to do this, the likelihood function is used, which will be discussed in the following section.

5.3.2 Aoristic Likelihood

The information contained in the data is usually entered into the statistical model through the likelihood function. The likelihood function of the von Mises distribution can be specified as

$$L(\boldsymbol{\phi} \mid \boldsymbol{t}) = \prod_{i=1}^{n} p(t_i \mid \mu, \kappa), \tag{5.3}$$

where μ and κ are the aforementioned parameters of the von Mises distribution, and $p(t_i \mid \mu, \kappa)$ is the von Mises density given in 5.2, which could also be replaced by any other density if we so desire. Usually, we would algebraically

or numerically optimize this likelihood function to obtain maximum likelihood estimates (MLEs), which can be plugged into the von Mises density to obtain an estimate of the crime time density. However, this formulation supposes all crime times are directly observed.

In order to adapt the likelihood for aoristic data, some ideas from the survival analysis literature can be used. In survival analysis, data are usually *right-censored*, where it is known that an event takes place after a certain time, but not exactly when. Aoristic data, then, is *interval-censored*, where an event has taken place after a certain time, as well as before some later time, but where in this is interval the event took place is not known. Interval-censored data analysis is considered in medical statistics (Klein et al., 2013), and two specialized books exist as well (Sun, 2007; Chen et al., 2012).

Based on theory for interval-censored data, we can define an aoristic version of the likelihood of the von Mises distribution, by writing

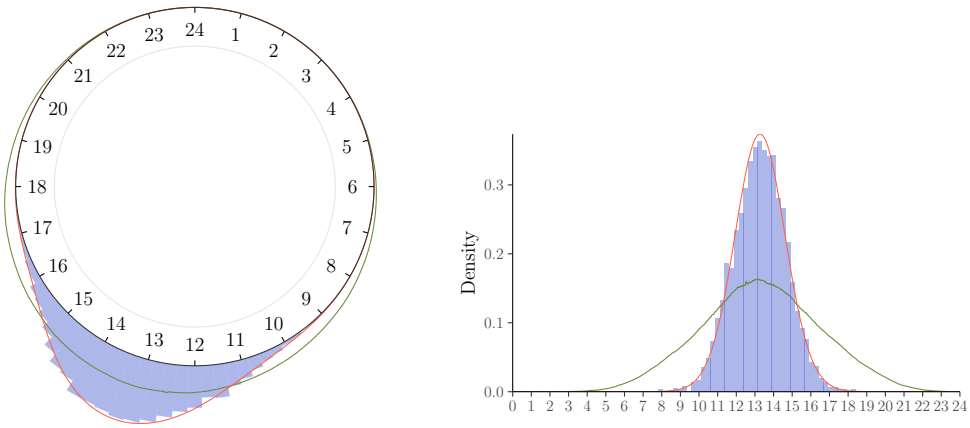
$$L(\mu, \kappa \mid \boldsymbol{t}_a) = \prod_{i=1}^{n} \int_{s_i}^{e_i} \frac{p(t \mid \mu, \kappa)}{e_i - s_i} dt \tag{5.4}$$

$$= \prod_{i=1}^{n} \frac{1}{e_i - s_i} \int_{s_i}^{e_i} p(t \mid \mu, \kappa) dt \tag{5.5}$$

$$= \prod_{i=1}^{n} \frac{F(e_i \mid \mu, \kappa) - F(s_i \mid \mu, \kappa)}{e_i - s_i}, \tag{5.6}$$

where $F(t \mid \mu, \kappa)$ is the cumulative distribution function (CDF) of the von Mises distribution. The computation of this aoristic likelihood is somewhat more complicated, but still feasible. To get estimates for μ and κ, this aoristic likelihood is optimized numerically to get maximum likelihood estimates. This approach can be simplified somewhat by filling in the unbiased estimator for μ which is given in 5.A.2, so that the aoristic likelihood only needs to be optimized for κ. It should be noted that if a different circular data model is desired, one can simply use the desired distribution in place of the cumulative distribution function of the von Mises distribution.

This aoristic likelihood correctly takes into account the uncertainty introduced by the aoristic intervals, but does not overestimate the spread of the crime time density as the aoristic fraction method does. This is shown in Figure 5.4, where it can be seen that estimated crime time density from the aoristic likelihood (red) can correctly recover the true distribution of crime times (blue histogram), while the aoristic fraction method (green) fails. Note that the blue histogram containing the true burglary times is never observed, as we only obtain aoristic intervals. Still, the aoristic likelihood recovers the shape of the true crime time distribution.



(a) (b)

Figure 5.4: Example of how the von Mises model estimated by the aoristic likelihood (red) correctly reconstructs the true distribution of the crime times (blue histogram), while the aoristic fraction method (green) does not.

Besides maximum likelihood estimates, one might be interested in obtaining an uncertainty around these estimates, perhaps in the form of a confidence interval or credible interval. Often, these can be obtained by deriving standard errors, but this is difficult due to the mathematical form of the aoristic likelihood. Therefore, one must resort to resampling methods, such as bootstrapping (Davison & Hinkley, 1997), or Bayesian approaches such as MCMC sampling to obtain the uncertainty around the maximum likelihood estimates.

We implemented bootstrapping using the R package `boot` (Canty & Ripley, 2017). The final results give an estimated mean of 13:16, with bootstrap confidence interval (CI) computed as (13:12, 13:21). Throughout the paper, we will write give bootstrap confidence intervals between parentheses after a maximum likelihood estimate, ie. MLE (Lower Bound, Upper Bound). The concentration parameter has an estimated value of 8.16 (7.64, 8.75). The true values under which the data was sampled are a true mean direction of 13:15 and a true concentration parameter of 8, so even though the data was made aoristic, the aoristic likelihood approach is able to precisely recover the true values using aoristic data.

5.3.3 Data augmentation

The aoristic likelihood approach is the most appealing method for most simple applications such as fitting a single unimodal distribution on a set of aoristic intervals. However, to compute our uncertainty in these estimates (such as through standard errors), or to incorporate this into more complex models, it can be beneficial to employ data augmentation strategies, such as for the nonparametric models discussed in the following section. Data augmentation is also the standard way to deal with missing data in Bayesian inference (Gelman et al., 2003).

Data augmentation strategies (Tanner & Wong, 1987; Gelfand & Smith, 1990; Van Dyk & Meng, 2001) make use of the fact that a statistical model implies a certain probability density $p(t \mid \boldsymbol{\phi})$ for any crime time t. Therefore, if we have our statistical model, we could sample the aoristic crime times from their intervals according to this probability density, which is the current estimated crime time density. For example, in the burglary example, the sampled values for each aoristic interval could come from a von Mises distribution truncated at the ends of the interval (for details of two ways to do this, see Appendix 5.C). The issue with this strategy is that the statistical model can only be estimated if the data are imputed, but the data are imputed using the statistical model, leading to circular reasoning. Therefore, such an approach usually involves some sort of iterative method, starting with the crime times uniformly sampled in the aoristic intervals, then estimating the statistical model, sampling the crime times again according to the new model, estimating the model again with the new data points, and so on until some sort of convergence.

The advantage of the data augmentation approach is that in each iteration, estimating the statistical model can be performed with off-the-shelf algorithms made for directly observed data. The disadvantage of this approach is that the resulting algorithm will require some sort of iterative updating, which can require more computational effort, as well as depending on random sampling. It is therefore recommended to use data augmentation only when necessary due to the computational complexity of a chosen model, such as in the following section.

5.4 Nonparametric models

While the aoristic likelihood approach works well for somewhat simple models, crime time data often displays more complex patterns which are not captured by such models. Crime is governed by a dynamic complex system,

society, where many unobserved and unknown factors contribute to why a crime occurs at exactly a certain time at exactly a certain place. As a result, crime times might not follow a simple unimodal distribution. For example, burglaries are known to occur either while victims are away during the day, or are asleep during the night, which would be a multimodal distribution. In addition, crime time densities are often skewed or violate distributional assumptions in some other way.

If the crime time density does not follow the distributional assumptions of the von Mises distribution, for example, these crime times might be a good fit for a nonparametric statistical model that does not make any assumption on the shape of the data distribution. Nonparametric models for interval censored data were first investigated in Turnbull (1974, 1976), but such models are not applicable to aoristic data. Therefore, we will apply Dirichlet process mixture models, which will be discussed next.

5.4.1 Dirichlet Process Mixture models

Among the most appealing nonparametric models are Dirichlet process mixture (DPM) models (Ferguson, 1973; Antoniak, 1974; Neal, 2000) found in the field of Bayesian nonparametrics (Hjort et al., 2010). For an introduction, see Gelman et al. (2003, ch. 23). DPM models are very flexible statistical models that are able to capture any underlying true distribution. Due to the increasing feasibility of computation for DPM models, they have seen an enormous increase in popularity over the last 20 years. Therefore, we believe that they represent a very promising approach for crime time modeling. A technical treatment of DPM models is beyond the scope of this paper, but we will recap some of its relevant properties here.

First, Dirichlet process models can fit any data distribution. This means that whether the true crime time density is multimodal, peaked, flat, or irregularly distributed in some other way, the DPM model will be able to learn this pattern after enough data, and thus give a good estimate of the true crime time density. Conceptually, this is true because the DPM model defines a prior over all possible probability distributions on the circle. Therefore, fear that our chosen statistical model does not fit the data is much less of a concern for such models. This is a property that this method shares with the aoristic fraction method, but without the attached issues discussed in Section 5.2.

Second, among the most appealing properties of the DPM model for crime time analysis is that we can not only compute an estimate of the probability of a crime happening in any desired time interval, but moreover we can compute the uncertainty around this estimate. For example, after running the DPM model, we might say the estimated probability of a crime happening

between 13:17 and 14:04 is 8.4%. This is already very useful, but the model also provides us a 95% credible interval, which might be, say, (0.4%, 19.4%). Because this interval is quite wide, we can conclude that more data are required to give a more precise estimate, but that the true probability of crime happening in this interval is unlikely to be larger than 19.4%. This is important, because it allows us to know when our predictions are unreliable because we have based them on too little data.

Third, because the DPM model is still a statistical model, it can be extended and connected to other statistical models. For example, DPM models can be embedded in hierarchical model (Teh et al., 2005), dynamic models (Ren et al., 2008), spatial models (Duan et al., 2007) or regression models (Chib & Greenberg, 2010). This would not be possible with the aoristic fraction method, for instance. Temporal and spatial analyses of crime are often said to be too separated in crime analysis (Grubesic & Mack, 2008), which we could address by developing a statistical model for the temporal aspects.

The DPM model takes the form of a mixture model with a varying number of components. As a result, we must choose some base distribution which serves as a building block for the DPM model. The base distribution is often chosen for computational convenience, because the resulting DPM is so flexible that it is not very sensitive to the choice of base distribution. More important is that the base distribution must be assigned a prior, as in any Bayesian analysis. This prior can indeed influence the resulting analysis if the sample size is small, but reasonable choices are available.

Dirichlet processes have been employed for circular data in a handful of papers, but none of them have treated aoristic data. Hernandez-Stumpfhauser et al. (2016) develops a Dirichlet process using the projected Normal distribution as the base distribution and applies this to small area estimation. McVinish & Mengersen (2008) develop a Dirichlet process model based on triangular distributions on the circle. DPM models have also been used to overcome problems in circular regression (Ghosh et al., 2003a; George & Ghosh, 2006). Several other papers in this field have approached this in varying ways (Nuñez-Antonio et al., 2015).

We will use the von Mises distribution as the base distribution, with an uninformative prior on the parameters. Details of the model are given in Appendix 5.B. The computation of DPM models tends to be relatively computationally intensive, and still requires us to choose a solution for the aoristic property of our data. Two approaches for computation of our DPM model with aoristic data will be discussed next.

Augmented sampler

The main problem to deal with when applying the DPM model is the fact that there are aoristic observations. If there were no aoristic observations, then a von Mises based Dirichlet Process could be performed. Therefore Dirichlet process models for interval-censored data were first investigated in the successive substitution sampling of Doss (1994), which represents an application of the data augmentation strategy as discussed in Section 5.3.3.

A technical issue arises here that is unique to aoristic data. Usually, data augmentation strategies for censored observations sample values by direct rejection. That is, a candidate is sampled from the full distribution, and if it falls in the required interval, it is accepted. If not, the process is repeated until acceptance. However, aoristic data might include both large and small intervals. If an aoristic interval is small, the probability of acceptance can also become very small, so a large number of candidates must be sampled before acceptance. In such cases, an alternative envelope rejection sampling method is used. Therefore, the sampling algorithm is chosen adaptively. That is, the rejection probability is estimated, and if it is too small (for example acceptance probability below 10%), we will use the envelope rejection sampler. For details, see Appendix 5.C.

Marginal sampler

An alternative to the augmented sampler is to use the aoristic likelihood discussed in Section 5.3.2 directly. It can simply be substituted in wherever the likelihood is used in the DPM computation. The downside here compared to the augmented sampler is each iteration of the algorithm takes a significantly larger amount of time. The upside is that the marginal sampler requires fewer random sampling steps and allows a larger set of priors than the augmented sampler. For most situations however, the augmented sampler will perform better and should be used.

5.5 Applications

This section will provide several applications of our method. Apart from results that directly stem from our approaches to estimating the crime time density, we will also show several further calculations that can be made which may be of interest. It should be noted that all of these analyses can simply be performed using the `R` package aoristicinference.

5.5.1 Ashby & Bowers data

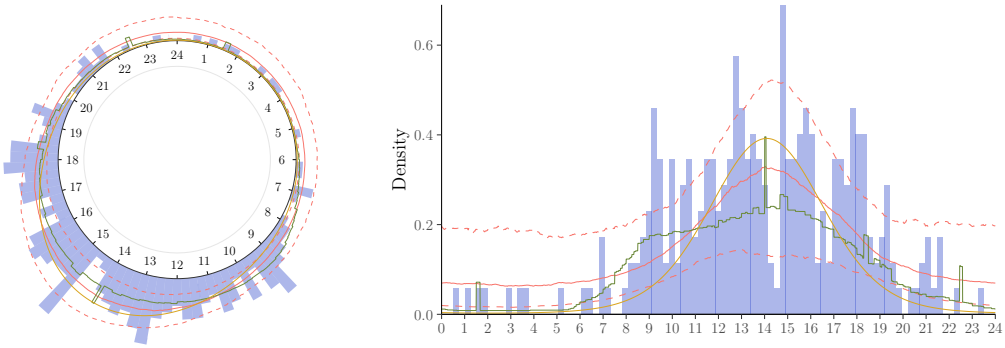
In Ashby & Bowers (2013), the authors exemplify the aoristic fraction method by aoristic data obtained on bike theft in London. This paper provides a fantastic example dataset, because the authors have painstakingly gone through CCTV footage of the subway station under consideration to determine the true crime times for each of the police reports. This means that this dataset has the rare property that it is both ecologically valid, as well as having the true crime time t_{actual} available. Therefore, this dataset is an excellent opportunity to compare different methods.

The data consists of 242 aoristic intervals, each with an attached t_{actual}. The data are broadly unimodal, with most bike thefts taking place in the afternoon. The aoristic intervals are quite large, in this case, with a mean of 8.93 hours. Some intervals were larger than 24 hours, which are removed for this illustration, along with instances where t_{actual} was outside the aoristic interval (which occurs most likely due to reporting errors by the victim).

It should be noted that the original dataset contains 263 observations, of which 21 crime times were known precisely due to being caught in the act, for instance. These were not analyzed previously, but both the aoristic likelihood method and the DPM model can use a mix of known times and aoristic data.

The results for the three methods considered in Sections 5.2, 5.3 and 5.4 are displayed in Figure 5.5. Because the data are mostly unimodal, the three methods can be seen to be broadly in agreement. All three options give a unimodal estimated crime time density, although the aoristic fraction method gives a more jagged estimate of the crime time density, as well as having a higher variance. The aoristic likelihood estimates the mean at 14:27 (13:48, 15:10), with concentration 2.15 (1.77, 2.67). In the fit of the DPM model it can be seen that an uncertainty around the crime time density is also provided, which would become smaller if more data would be obtained.

An advantage of the DPM model, as mentioned previously, is that it is possible to compute the proportion of crimes happen within a certain time frame along with its uncertainty. In particular, Ashby & Bowers (2013) focus on the proportion of crime times occuring during each of three police shifts: (7:00 - 15:00), (15:00 - 23:00), and (23:00 - 07:00). During the morning shift, the model suggests estimates 40.4% of the bike thefts to occur during it, with a credible interval (CI) of (19.5%, 58.8%). The evening shift expects to obtain 35.7% with CI (17.5%, 59.4%), with the night shift is estimated to have 16.5% with CI (2.7%, 46.2%). Although it is clear that most bike thefts at this location occur during the morning and evening shift, the main conclusion to be drawn from this analysis is the uncertainty in our conclusions



(a) (b)

Figure 5.5: Three different models run on the data of Ashby & Bowers (2013). The 242 aoristic intervals are not displayed, but the true theft times are displayed as the blue histogram. The estimated crime time densities, which attempt to fit this histogram, are the aoristic fraction method (green), the maximum likelihood estimate of a von Mises distribution fit with the Aoristic Likelihood (yellow), and the Dirichlet Process Mixture model and its credible interval (red solid line, with 80% credible interval red and dashed).

from the model is much larger than previous analyses have made it out to be. The cause for this can be found in the fact that the aoristic intervals provide less information than precise times, so the 'effective sample size' is much lower than the 242 observations we appeared to have.

5.5.2 Montgomery Crime data

The city of Montgomery is among several cities to publish open data of crimes in the city that includes a start and end date. This dataset was obtained from the Montgomery Open Crime Data website. The dataset contains 44299 observations of crimes in Montgomery observed in the years 2016 - 2018. Many crime types are provided, but we will focus on two types of property crime, theft from building and burglary, as these are the most interesting aoristic data points. Crimes missing start or end times are removed, while those with start and end times that exactly equal or up to 2 minutesn apart are treated as observed crime times. All others are treated as aoristic.

For this type of data, the Aoristic Fraction approach is a good way to obtain an initial descriptive analysis of the data. Figure 5.6 shows the results of all three methods for the two crime types in the Montgomery data. Note

that a true evaluation of which approach performs best here is not possible as the true crime times are not known.

'Theft from building', defined[1] as *'A theft from within a building which is either open to the general public or where the offender has legal access'*, is broadly unimodal, with a peak around 15:00. All three methods perform reasonably well. The aoristic likelihood method is able to fully capture the shape of this distribution, with an estimated mean direction of 15:30 (15:20, 15:40), with concentration 1.12 (1.07, 1.17). The DPM model fits the data slightly better, and is fairly certain about the distribution due to the size of the dataset.

For 'burglary', defined[2] as *'The unlawful entry into a building or other structure with the intent to commit a felony or a theft'*, the distribution is bimodal, with one peak in the middle of the night (around 3:00), and another around lunchtime at 13:00. These times correspond to times when properties are most likely left unattended. In this case, the aoristic likelihood fails to fit the data well because it attempts to fit a unimodal density to the bimodal dataset, and as a result gives a uniform result, with concentration 0 (0, 0.07). The DPM model still captures the distribution adequately.

5.5.3 Montly crime time trends

The DPM model can also provide inference for any function of the parameters of the model, which allows a wealth of further analyses. As an example, we show how to investigate seasonal changes in crime times. First, we run the DPM model separately for each month, giving a non-parametric fit along with the uncertainty around this fit. Second, we compute the mean direction of the resulting crime time density for 1000 MCMC samples from the DPM model, giving us an estimate of the mean direction for this month as well as the uncertainty around it. Then, this is plotted in Figure 5.7, where the black line provides the estimated mean crime time for each month, and the grey area provides the uncertainty around this, which was obtained as a credible interval from the mean direction in each of the MCMC samples. From it, we can conclude that there seems to be a general trend of slightly earlier burglary times during the summer, although the uncertainty around this trend is still fairly large.

[1] https://ucr.fbi.gov/nibrs/2011/resources/nibrs-offense-definitions
[2] https://ucr.fbi.gov/nibrs/2011/resources/nibrs-offense-definitions

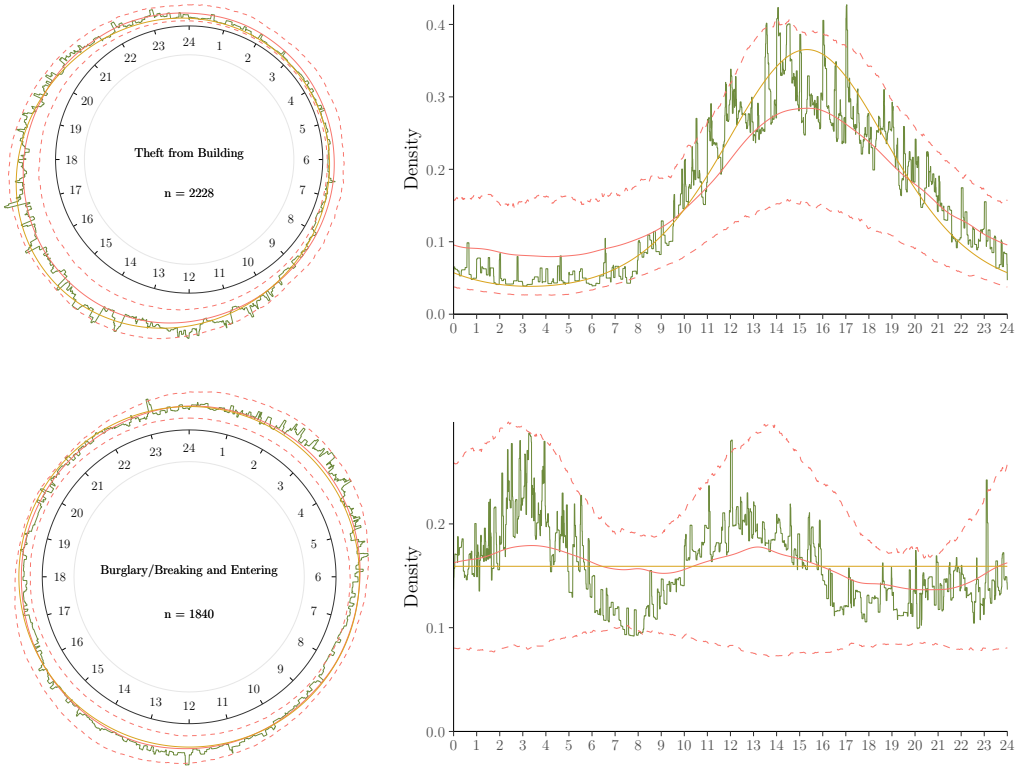


Figure 5.6: The three methods discussed in this paper applied to two crime types, theft from building and burglary. The methods displayed are the aoristic fraction method (green), the maximum likelihood estimate of a von Mises distribution fit with the Aoristic Likelihood (yellow), and the Dirichlet Process Mixture model and its credible interval (red solid line, with 80% credible interval red and dashed).

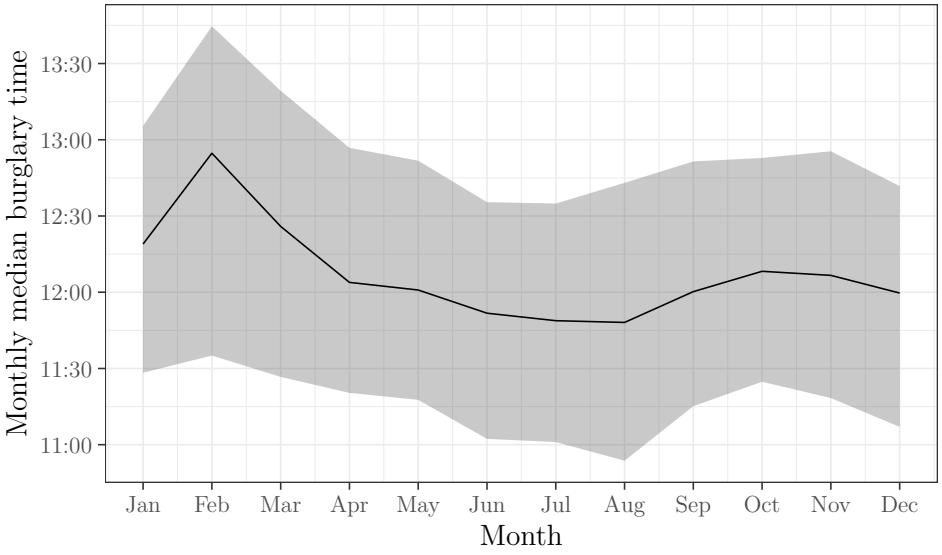


Figure 5.7: Median burglary time from Dirichlet Process Model plotted as it changes per month (black line), and its confidence interval (grey area).

5.6 Discussion

In this paper, we have developed novel methods for dealing with aoristic data in crime analysis and beyond. Three methods were presented. The aoristic fraction method is a descriptive method for which major downsides have been established, primarily a systematic overestimation of the variance of the crime time density. We have shown how the aoristic likelihood method provides a solid method for fitting unimodal crime times, using all information contained in the aoristic data.

The Dirichlet Process Mixture model provides a natural extension of these models, allowing us to fit any crime time distribution regardless of the number of modes or the shape of the distribution. This is an appealing property that this model shares with the aoristic fraction method, but as opposed to that method the DPM model is built on a solid statistical foundation. This provides the DPM three main advantages. First, it uses all information in the data. Second, it provides measures of uncertainty in our model and conclusions. Third, it can be extended and connected to other statistical models.

Choosing between aoristic likelihood methods and nonparametric approaches is a choice made dependent on the specific crime type under investigation. Some crime types are fit well with a symmetric unimodal model, and for these, the aoristic likelihood approach is simpler, faster, more inter-

pretable and easier. However, if the crime time density for this crime type is bimodal or otherwise differently distributed, the DPM model provides an extremely flexible extension that is able to adapt to any shape of the data distribution.

A major hurdle in employing aoristic analyses is the complexity of implementing the computational methods, so that crime analysts are usually not inclined to incorporate aoristic analyses in their workflow, as evidenced by several authors mentioning the importance of aoristic methods without implementing any. To address this, analyses in this work have been implemented in an easy to use `R` package, aoristicinference. It provides introductions and user guides, so that users not familiar with the details of the statistical methods in this work should also be able to use the methods.

A strongly related method to the aoristic fraction method is what is called *Interstitial Crime Analysis* (ICA) (A. Gill et al., 2014). Interstitial Crime Anaysis is a technique not for unknown time but unknown location. For example, Newton et al. (2014) use this technique to estimate the proportion of thefts that occurs between any two metro stations of the London Underground. From a statistical perspective, this method suffers from the same problems as the aoristic fraction method. In fact, Appendix 5.A.1 is a proof of why the variance will be systematically overestimated, with the minor difference that ICA is usually done on a categorical sample space. Therefore, a future investigation into an "Interstitial Likelihood" could prove fruitful.

A limitation of any aoristic analysis method is that if the true crime times are systematically more likely situated toward the start or end of the intervals, our analysis will be biased. This is strongly related to the assumption of independence of the intervals and the true crime times. That is, in reality the intervals can depend on the crime time, for example if an offender waits until they see their chosen victim leave their house. This could also happen if offenders and victims both follow a fixed pattern, for example if burglaries are usually in the afternoon, while the victims are always gone during working hours, and thus provide intervals close to working hours. It is impossible to infer such behaviours from a fully aoristic dataset, akin to the problem of Missing Not At Random (MNAR) missing data. However, one might attempt to test this assumption if a sufficiently sized known-time dataset is available. Then, it could be checked whether the known-time and aorisitic datasets produce differrent crime time densities. If so, we are indeed violating our assumptions. So long as this is not the case, our method extracts the most information from a set of (partly) aoristic data, given the constraints we have.

With these tools for addressing aoristic data, future work should focus on combining these methods with spatial analyses, in particular point process

models as in Wang & Gelfand (2014). A combined spatio-temporal model that uses aoristic data to their fullest extent would provide a powerful tool for understanding both when and where crimes occur, as well as for predictive policing.

5.7 Acknowledgements

The authors are grateful to Matthew Ashby for providing the data from Ashby & Bowers (2013).

5.A Proof of variance overestimation using the aoristic fraction method

The fact that the variance of the data in density estimation methods such as the aoristic fraction method can be proven mathematically. First, for familiarity, this will be done for the linear case, where the data lies on the real line. Second, we will show the same proof for circular interval censored data, such as aoristic data.

5.A.1 Data on the real line

Here, we will show that for data on the real line using a density estimate based on the interval-censored histogram (ICH) method, essentially the linear analogue of the aoristic fraction method, leads to an overestimate of the variance in the linear case.

Let $X \in \mathbb{R}$ be a random variable with some unknown distribution, but where $E[X] = \mu$ and $\text{Var}[X] = \sigma^2$. This X is not directly observed, but rather we observe an interval (a, b), knowing $a \leq x \leq b$. We can expand the bounds as $a = x - \delta_i^l$ and $b = x + \delta_i^u$, where δ_i^l is the distance between the start of the interval and the unobserved true x, and δ_i^u the distance further from x to the end of the interval. Note that δ_i^l and δ_i^u are random variables themselves, with unknown distribution. A core assumption that we will make is that δ_i^l and δ_i^u have the same distribution. Then, we also assume that the expected difference between x and an interval bound is $E[\delta_i^l] = E[\delta_i^l] > 0$. That is, the data are actually censored. As a result, $E[\delta_i^u - \delta_i^l] = 0$.

The procedure under investigation is to estimate the unknown density $p(x)$ by the interval-censored histogram, which is an average of uniform distributions for each interval, that is

$$\hat{p}_{ICH}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(a_i < x < b_i)}{b_i - a_i},$$

where $I(\cdot)$ is the indicator function. The question we will evaluate is whether the variance of this density, which we will call $\hat{\sigma}^2_{ICH}$, overestimates the true variance σ^2 of X. Thus, the question is whether $E[\hat{\sigma}^2_{ICH}] > \sigma^2$.

First, we will need the expectation to produce an estimate of the variance.

Note that our estimate of the expected value of X using the ICH density is

$$\hat{\mu} = \int_{\mathbb{R}} x \; \hat{p}_{ICH}(x) dx \tag{5.7}$$

$$= \int_{\mathbb{R}} x \frac{1}{n} \sum_{i=1}^{n} \frac{I(a_i < x < b_i)}{b_i - a_i} dx \tag{5.8}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{a_i}^{b_i} x dx = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{2} \tag{5.9}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{\delta_i^u - \delta_i^l}{2}, \tag{5.10}$$

where the last form is not observed, but helps us realise that if we take the expectation of the unknown interval lengths, $\frac{E[\delta_i^u] - E[\delta_i^l]}{2} = 0$. Then, $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. So, if δ_i^l and δ_i^u have the same expectation, the estimate $\hat{\mu}$ is an unbiased estimator of μ.

For simplicity, we will assume that the data are centered so that $\hat{\mu} = 0$. Then, we can create an unbiased estimator of σ^2 by

$$\hat{\sigma}^2_{ICH} = \left(1 - \frac{1}{n}\right) \text{Var}[X] = \left(1 - \frac{1}{n}\right) E[X^2] \tag{5.11}$$

$$= \int_{\mathbb{R}} x^2 \frac{1}{n-1} \sum_{i=1}^{n} \frac{I(a_i < x < b_i)}{b_i - a_i} dx \tag{5.12}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{3} \frac{b_i^3 - a_i^3}{b_i - a_i} = \frac{1}{3} \frac{1}{n-1} \sum_{i=1}^{n} a_i^2 + a_i b_i + b_i^2 \tag{5.13}$$

Then, expanding the bounds, and taking the expectation over δ_i^l and δ_i^u,

$$E_{\delta_i^l}[E_{\delta_i^u}[\hat{\sigma}^2_{ICH}]] = E_{\delta_i^l}\Big[E_{\delta_i^u}\Big[\frac{1}{3}\frac{1}{n-1}\sum_{i=1}^{n} 3x_i^2 - 2\delta_i^l x_i + \delta_i^{l\,2} + \delta_i^u x_i - \tag{5.14}$$

$$\delta_i^l \delta_i^u + 2\delta_i^u x_i + \delta_i^{u2}\Big]\Big] \tag{5.15}$$

$$= \frac{1}{3}\left\{\frac{1}{n-1}\sum_{i=1}^{n} 3x_i^2 + E_{\delta_i^l}\left[E_{\delta_i^u}\left[\delta_i^{l\,2} + \delta_i^{u2} - \delta_i^l \delta_i^u\right]\right]\right\} \tag{5.16}$$

$$= \left\{\frac{1}{n-1}\sum_{i=1}^{n} x_i^2\right\} \tag{5.17}$$

$$+ \frac{1}{3}\left\{\frac{1}{n-1}\sum_{i=1}^{n} E_{\delta_i^l}\left[E_{\delta_i^u}\left[\delta_i^{l\,2} + \delta_i^{u2} - \delta_i^l \delta_i^u\right]\right]\right\}. \tag{5.18}$$

We can recognize that $\frac{1}{n-1}\sum_{i=1}^{n} x_i^2$ is simply an unbiased estimator of the variance. Therefore, taking the expectation over x,

$$E_x[E_{\delta_i^l}[E_{\delta_i^u}[\hat{\sigma}^2_{ICH}]]] = E_x\left[\frac{1}{n-1}\sum_{i=1}^{n} x_i^2\right] + \tag{5.19}$$

$$\frac{1}{3}\left\{\frac{1}{n-1}\sum_{i=1}^{n} E_{\delta_i^l}\left[E_{\delta_i^u}\left[{\delta_i^l}^2 + {\delta_i^u}^2 - \delta_i^l\delta_i^u\right]\right]\right\} \tag{5.20}$$

$$= \sigma^2 + \frac{1}{3}\left\{\frac{1}{n-1}\sum_{i=1}^{n} E_{\delta_i^l}\left[E_{\delta_i^u}\left[{\delta_i^l}^2 + {\delta_i^u}^2 - \delta_i^l\delta_i^u\right]\right]\right\} \tag{5.21}$$

However, note that ${\delta_i^l}^2 + {\delta_i^u}^2 - \delta_i^l\delta_i^u \geq 0$, with equality if and only if $\delta_i^l = \delta_i^u = 0$, which would mean there was no interval-censoring. Because the second term is positive, we have that the variance is overestimated by $\hat{\sigma}^2_{ICH}$, that is, it is biased upwards. From this last equation, it is clear that the upwards bias depends on the distribution of δ_i^l and δ_i^u. Specifically, the bias is $E[\hat{\sigma}^2_{ICH} - \sigma] = \frac{1}{3}\left\{\frac{1}{n-1}\sum_{i=1}^{n} E_{\delta_i^l}\left[E_{\delta_i^u}\left[{\delta_i^l}^2 + {\delta_i^u}^2 - \delta_i^l\delta_i^u\right]\right]\right\}$.

5.A.2 Aoristic data

In the circular case, let $\theta \in [0, 2\pi)$ be a circular random variable, which again is not directly observed, but rather only up to an interval $a_i \leq \theta \leq b_i$, with again $a = \theta - \delta_i^l$ and $b = \theta + \delta_i^u$. This time, the intervals are assumed to be bounded on the semicircle, that is $\delta_i^l \in [0, \pi)$, and $\delta_i^u \in [0, \pi)$. Again, assume that δ_i^l and δ_i^u have some unknown distribution.

The distribution of θ is unknown, but let's assume it has a population mean direction μ. An unbiased estimator of μ is given by

$$\bar{\theta} = \text{atan2}(\sum_{i=1}^{n}\sin(\theta_i), \sum_{i=1}^{n}\cos(\theta_i))$$

(Mardia & Jupp, 2000). For aoristic data, an unbiased estimator of μ is

$$\tilde{\theta} = \text{atan2}\left(\sum_{i=1}^{n}\sin(a_i) + \sin(b_i), \sum_{i=1}^{n}\cos(a_i) + \cos(b_i)\right). \tag{5.22}$$

This is equal to taking the mean direction of the midpoints of the aoristic intervals. To show that this is unbiased, taking the expectation over the

distribution of δ_i^l and δ_i^u, we obtain

$$E_{\delta_i^l,\delta_i^u}\left[\sum_{i=1}^{n}\cos(a_i)+\cos(b_i)\right]=\sum_{i=1}^{n}\cos(\theta_i)E_{\delta_i^l,\delta_i^u}\left[\cos(\delta_i^l)+\cos(\delta_i^u)\right] \quad (5.23)$$

$$E_{\delta_i^l,\delta_i^u}\left[\sum_{i=1}^{n}\sin(a_i)+\sin(b_i)\right]=\sum_{i=1}^{n}\sin(\theta_i)E_{\delta_i^l,\delta_i^u}\left[\cos(\delta_i^l)+\cos(\delta_i^u)\right]. \quad (5.24)$$

Note that this last term $E_{\delta_i^l,\delta_i^u}\left[\cos(\delta_i^l)+\cos(\delta_i^u)\right]$ is the same for both components. Also, for any constant $q>0$ we have $\text{atan2}(qs,qc)=\text{atan2}(s,c)$. So, if we set $q=E_{\delta_i^l,\delta_i^u}\left[\cos(\delta_i^l)+\cos(\delta_i^u)\right]>0$, we can see

$$\text{atan2}\left(q\sum_{i=1}^{n}\sin(\theta_i),q\sum_{i=1}^{n}\cos(\theta_i)\right)=\text{atan2}\left(\sum_{i=1}^{n}\sin(\theta_i),\sum_{i=1}^{n}\cos(\theta_i)\right)=\bar{\theta}, \quad (5.25)$$

which is an unbiased estimator of μ.

The population variance is given by $1-\rho$, where $\rho=\int_0^{2\pi}\cos(\theta-\mu)p(\theta)d\theta$ is the population resultant length. If we center data by subtracting the mean direction estimate $\tilde{\theta}$ such that the mean direction is zero, then we simply have $\rho=\int_0^{2\pi}\cos\theta p(\theta)d\theta$, with an unbiased estimator being $\frac{1}{n}\sum_{i=1}^{n}\cos\theta_i$.

The aoristic fraction method leads us to estimate the resultant length as

$$\hat{\rho}_{AF}=\int_0^{2\pi}\cos\theta\ \hat{p}_{AF}(\theta)d\theta \quad (5.26)$$

$$=\int_0^{2\pi}\cos\theta\frac{1}{n}\sum_{i=1}^{n}\frac{I(a_i<x<b_i)}{b_i-a_i}d\theta \quad (5.27)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b_i-a_i}\int_{a_i}^{b_i}\cos\theta d\theta \quad (5.28)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\frac{\sin b_i-\sin a_i}{b_i-a_i} \quad (5.29)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\frac{\sin(\theta_i+\delta_i^u)-\sin(\theta_i-\delta_i^l)}{b_i-a_i} \quad (5.30)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\frac{\sin\theta_i\cos\delta_i^u+\cos\theta_i\sin\delta_i^u-\sin\theta_i\cos\delta_i^l+\cos\theta_i\sin\delta_i^l}{b_i-a_i}. \quad (5.31)$$

Then, taking the expectation over θ, δ_i^l and δ_i^u, recalling $E[\sin\theta]=0$, this

gives

$$E[\hat{\rho}_{AF}] = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{E[\delta_i^l] + E[\delta_i^u]} \Big[E[\sin\theta_i]E[\cos\delta_i^u] + E[\cos\theta_i]E[\sin\delta_i^u] \quad (5.32)$$

$$- E[\sin\theta_i]E[\cos\delta_i^l] + E[\cos\theta_i]E[\sin\delta_i^l] \Big] \quad (5.33)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{E[\cos\theta_i]E[\sin\delta_i^u] + E[\cos\theta_i]E[\sin\delta_i^l]}{E[\delta_i^l] + E[\delta_i^u]} \quad (5.34)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \rho \frac{E[\sin\delta_i^l] + E[\sin\delta_i^u]}{E[\delta_i^l] + E[\delta_i^u]}. \quad (5.35)$$

Then, set the distributions of δ_i^l and δ_i^u equal again to the distribution of some δ, to get

$$E[\hat{\rho}_{AF}] = \frac{1}{n}\sum_{i=1}^{n} \rho \frac{E[\sin\delta_i]}{E[\delta_i]}. \quad (5.36)$$

Finally, if $0 < \delta < \pi$, that is, there is interval censoring, we know that $\sin\delta < \delta$, so $\frac{E[\sin\delta_i]}{E[\delta_i]} < 1$, and the bias is

$$E[\hat{\rho}_{AF} - \rho] = \frac{1}{n}\sum_{i=1}^{n} \rho \frac{E[\sin\delta_i]}{E[\delta_i]} - \rho = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{E[\sin\delta_i]}{E[\delta_i]} - 1 \right) \rho < 0. \quad (5.37)$$

Therefore, the estimate of the resultant length given by the aoristic fraction method $\hat{\rho}_{AF}$ is biased downwards, so that the circular variance $1 - \hat{\rho}_{AF}$ is biased upwards.

5.B Von Mises Dirichlet Process Mixture model

We propose to use a von Mises based Dirichlet Process mixture (DPM) model. The von Mises DPM model on circular observation $\theta_i \in [0, 2\pi)$ can be written as

$$\theta_i \mid (\mu_i, \kappa_i) \sim \mathcal{M}(\mu_i, \kappa_i) \tag{5.38}$$
$$(\mu_i, \kappa_i) \mid G \sim G \tag{5.39}$$
$$G \sim \mathrm{DP}(P_0, \alpha). \tag{5.40}$$

The base distribution P_0 is the conjugate prior for the von Mises distribution, that is

$$p(\mu, \kappa \mid \mu_0, R_0, n_0) \propto [I_0(\kappa)]^{-n_0} \exp\left\{R_0 \kappa \cos(\mu - \mu_0)\right\}, \tag{5.41}$$

where μ_0 is the prior mean, R_0 is the prior resultant length, and n_0 is somewhat like a prior sample size.

Because this prior is conjugate, the posterior has the same form, which is

$$p(\mu, \kappa \mid \mu_n, R_n, m) \propto [I_0(\kappa)]^{-m} \exp\left\{R_n \kappa \cos(\mu - \mu_n)\right\}, \tag{5.42}$$

where, setting $S_n = \sum_{i=1}^{n} \sin(\theta_i) + R_0 \sin(\mu_0)$ and $C_n = \sum_{i=1}^{n} \cos(\theta_i) + R_0 \cos(\mu_0)$, the posterior mean direction is $\mu_n = \operatorname{atan2}(S_n, C_n)$, the posterior resultant length is given by $R_n = \sqrt{C_n^2 + S_n^2}$, and the posterior sample size is $m = n + n_0$.

For computation, we also require the prior predictive distribution of a data point $\tilde{y}$. It is given by

$$p(\tilde{y} \mid \mu_0, R_0, n_0) = \int_0^{\infty} \int_0^{2\pi} p(\mu, \kappa \mid \mu_0, R_0, n_0) \mathcal{M}(\tilde{y} \mid \mu, \kappa) d\mu d\kappa \tag{5.43}$$
$$= \frac{1}{C} \frac{1}{(2\pi)^{n_0}} \int_0^{\infty} \frac{I_0(R_p \kappa)}{I_0(\kappa)^{n_p}} d\kappa \tag{5.44}$$

where $p(\mu, \kappa \mid \mu_0, R_0, n_0)$ is the base distribution P_0, $\mathcal{M}(\tilde{y} \mid \mu, \kappa)$ is the probability density function of the von Mises distribution,

$$R_p = \sqrt{\left(\cos \tilde{y} + R_0 \cos \mu_0\right)^2 + \left(\sin \tilde{y} + R_0 \sin \mu_0\right)^2}, \tag{5.45}$$

$n_p = n_0 + 1$, and

$$C = (2\pi)^{1-n_0} \int_0^{\infty} \frac{I_0(R_0 \kappa)}{I_0(\kappa)^{n_0}} d\kappa \tag{5.46}$$

is the normalizing constant of the base distribution P_0.

To obtain a Dirichlet process that is uninformative with regards to the mean direction, we must take $R_0 = 0$ so that the terms with μ_0 disappear in the posterior computations. If we want a non-informative prior to limit the influence of the prior, this suggests setting $R_0 = 0$, $n_0 = 1$, and μ_0 any value, as it is irrelevant now. The downside of such an approach is that it puts most weight on components that have low concentration, an issue which is addressed in Appendix 5.D.

Computation was implemented using the R package dirichletprocess (J. Ross & Markwick, 2018), to which methods for circular data and aoristic data were contributed. This package uses the Gibbs sampling schemes from Neal (2000).

5.C Rejection sampling aoristic data

A computational issue arises in rejection sampling for aoristic data because it will tend to contain both very small and larger intervals. Usually, data augmentation will use direct rejection sampling, such as in Doss (1994). This direct rejection algorithm will attempt to sample a value t from interval (a, b) which has distribution $p(t \mid \boldsymbol{\phi}, a, b) \propto p(t \mid \boldsymbol{\phi}) I(a < t < b)$, where $\boldsymbol{\phi}$ are the parametrs of the distribution and $I(\cdot)$ is the indicator function. This algorithm can be summarized as follows.

1. Sample a value $t^* \sim p(t \mid \boldsymbol{\phi})$.
2. If $a < t < b$, set $t = t^*$. Otherwise, go back to step 1.

However, if we have a very small interval, say (17:18, 17:21), the algorithm will perform have very low acceptance probability and thus perform badly. This is because the acceptance probability of the direct rejection algorithm is $\int_a^b p(t \mid \boldsymbol{\phi}) dt$ which can be quite small for small intervals.

An alternative is to use envelope rejection sampling (Gilks & Wild, 1992), which can be summarized as follows.

1. Compute the maximum value of the distribution $p(t \mid \boldsymbol{\phi})$ within the interval, that is $m = \max_{t:a<t<b} p(t \mid \boldsymbol{\phi})$.
2. Sample a value $t^* \sim U[a, b]$, that is from the uniform distribution between a and b.
3. Sample $u \sim U[0, 1]$. If $um < p(t^* \mid \boldsymbol{\phi})$, set $t = t^*$. Otherwise, go back to step 2.

This algorithm has rejection probability $m(b-a) - \int_a^b p(t \mid \boldsymbol{\phi}) dt$. This should make it clear that the envelope rejection and direct rejection will often perform differently in terms of acceptance probability.

For this reason, we employ a adaptive strategy for the sampling. A final issue is that the acceptance probability is generally not available in closed form and costly to compute. Therefore, we approximate the distribution by a Normal distribution. For the von Mises distribution with mean direction μ and concentration κ, we can approximate it by a Normal distribution $N(\mu, \frac{1}{\kappa})$. If the approximated acceptance probability for direct rejection is smaller than some number, say 10%, we switch to the envelope rejection strategy.

5.D Prior independent of μ_0

In the Dirichlet Process, we have $\{\mu, \kappa\} \sim DP(\alpha P_0)$. The base measure P_0 is the conjugate prior for the von Mises distribution, that is

$$p(\mu, \kappa \mid \mu_0, R_0, n_0) \propto [I_0(\kappa)]^{-n_0} \exp\{R_0 \kappa \cos(\mu - \mu_0)\}. \tag{5.47}$$

As R_0 gets closer to n_0, more weight is given to higher κ, so more concentrated components, in the direction of μ_0. Note that the prior mean direction μ_0 is irrelevant if $R_0 = 0$, so a common prior undecided about μ_0 has $R_0 = 0, n_0 = 1$.

However, we would like to put prior mass on more concentrated components, while remaining undecided on μ_0. A prior with this property can be obtained by integrating out μ_0, which results in the prior

$$p(\mu, \kappa \mid R_0, n_0) \propto \int_0^{2\pi} [I_0(\kappa)]^{-n_0} \exp\{R_0 \kappa \cos(\mu - \mu_0)\} d\mu_0 = \frac{I_0(R_0\kappa)}{I_0(\kappa)^{n_0}}. \tag{5.48}$$

This prior, however, will lead to heavily increased computational burden, because it is no longer conjugate.

To regain conjugacy, the sampler can be augmented by taking $\mu_0 \sim U(0, 2\pi)$, so $p(\mu_0) = 1/(2\pi)$. Then, sampling μ_0 each time, the base measure, averaged over μ_0, will be

$$E_{\mu_0}[p(\mu, \kappa \mid \mu_0, R_0, n_0)] = \int_0^{2\pi} p(\mu, \kappa \mid \mu_0, R_0, n_0) p(\mu_0) d\mu_0 \tag{5.49}$$

$$= \int_0^{2\pi} [I_0(\kappa)]^{-n_0} \exp\{R_0 \kappa \cos(\mu - \mu_0)\} \frac{1}{2\pi} d\mu_0 \tag{5.50}$$

$$\propto p(\mu, \kappa \mid R_0, n_0). \tag{5.51}$$

This last one is exactly as required. Therefore, the μ_0 can be simply randomly sampled uniformly over the circle in each iteration, and filled into the conjugate model.

An alternative is to marginalize the posterior summary statistics μ_n, R_n, m over the uniform distribution on μ_0. We obtain $E_{\mu_0}[\mu_n] = \bar{\theta}$, and $E_{\mu_0}[m] = n + n_0$, with R_n being the average distance of the points on a circle with center $(\sum_{i=1}^n \cos(\theta_i), \sum_{i=1}^n \sin(\theta_i))$ and radius R_0. This is given by

$$E_{\mu_0}[R_n] = \frac{1}{2\pi} \int_0^{2\pi} \sqrt{R_0^2 + R^2 + 2R_0 R \cos(t)} dt \tag{5.52}$$

$$= \frac{2[R_0 + R]}{\pi} E_2\left(\frac{2\sqrt{R_0 R}}{R_0 + R}\right) \tag{5.53}$$

where R is the resultant length of the data, and $E_2(\cdot)$ is the complete elliptic integral of the second kind. There is no closed form available for this last integral, but efficient algorithms are available.

A final option is introduce a data dependence in the prior by setting $\mu_0 = \bar{\theta}$ if there is data, and $\mu_0 \sim U[0, 2\pi)$ if there is not. This prior can be set in terms of posterior parameters μ_n, R_n, m because if $\mu_0 = \bar{\theta}$, we have $\mu_n = \bar{\theta}$, $R_n = R_0 + R$, and $m = n_0 + n$. The data dependence in this prior is sometimes seen as problematic in the Bayesian community (Darnieder, 2011), but the ease of use by setting this in the posterior parameters and conjugacy while allowing for more concentrated priors makes it an appealing alternative.

Chapter 6

circbayes: An R package for Bayesian circular statistics

Abstract

Circular data, consisting of observations measured in a periodic sample space, are encountered throughout a variety of disciplines. The field of circular statistics provides statistical models for the analysis of circular data. Within circular statistics, Bayesian methods are also becoming more prevalent, particularly due to the flexibility in developing sophisticated models. However, Bayesian methods for circular data are not yet widely available in easy-to-use packages. The package **circbayes** Bayesian inference for a wide variety of circular data models, so that it is easy to apply and compare multiple models on the same dataset. Circular regression models allow regressing circular outcomes on circular, linear or categorical predictors, and can be performed with von Mises or Projected Normal residuals. Hierarchical (mixed-effects) regression is available in the Projected Normal framework. Finite Mixture models and non-parametric Dirichlet Process Mixture models provide flexible methods to fit any circular density. Flexible Bayesian Hypothesis testing methods are provided which can perform tests between any of the previous models, which leads to Bayesian versions of many frequentist hypothesis tests of circular data. Together, the package **circbayes** provides a comprehensive selection of basic models for Bayesian analysis of circular data.

This chapter is in preparation for submission to Journal of Statistical Software as Mulder, K., Cremers, J. and Klugkist, I. (tbd) **circbayes**: An R package for Bayesian circular statistics.

Author contributions: KM developed the new software package, building on packages by KM and JC. KM wrote the paper. IK and JC gave feedback on written work.

Circular data consist of quantities defined on the unit circle, such that the sample space is periodic. This type of data can be generated by a number of natural mechanisms or measurement instruments. Wind directions, for example, can directly be observed as angles. Temporal data can be viewed as periodic if one is interested in the time of day, and similarly day of the week, month or year, which are repeating patterns over time. A polar transformation on bivariate data results in a linear and a directional component, which may represent the processes of interest better than the original bivariate data. Common between these examples is that the resulting data can be represented as angles in radians or degrees, and that they require specialized statistical methods, from a branch of statistics called circular statistics.

To see this, one might consider two angles 10° and 350°. Treating these as numbers on the real line results in a difference of 340° and a arithmetic mean of 180°, which do not correspond with an intuitive difference and mean. Instead, their difference on the circle (the arc length between them) is 20°, and their mean direction is 0°.

Circular data are encountered in a wide variety of scientific fields, including biology (Nuñez-Antonio et al., 2018), aerospace (Kurz & Hanebeck, 2017), political sciences (J. Gill & Hangartner, 2010), machine learning (Gopal & Yang, 2014), signal processing (Traa & Smaragdis, 2013), life sciences (Mardia, 2011), motor behaviour research (Mechsner et al., 2001a, 2007; Postma et al., 2008; Baayen et al., 2012), behavioural biology (Bulbert et al., 2015), psychology (Leary, 1957; Gurtman & Pincus, 2003; Kaas & Van Mier, 2006; Gurtman, 2009a), bioinformatics (Mardia et al., 2008) and environmental sciences (Lagona, 2016; Lagona et al., 2015; Arnold & SenGupta, 2006).

Statistical models for circular data have been developed in the field of circular statistics (Fisher, 1995; Jammalamadaka & Sengupta, 2001; Mardia & Jupp, 2000; Pewsey et al., 2013; Ley & Verdebout, 2017, 2018). Besides redefining the difference, mean, and other summary statistics, circular statistics relies on probability distributions defined on the circle, such as the von Mises distribution (Von Mises, 1918) or Projected Normal distribution (Kendall, 1974) (which is also known as the Displaced or Offset Normal distribution).

Several packages to deal with circular data exist (and will be recapped in Section 6.1), but no packages are available for Bayesian inference of circular data. However, there are several major advantages of the Bayesian approach, some of which are more general, while others are specific to the field of circular data. General advantages of the Bayesian approach are, among others, the ability to include prior information if desired, while objective approaches are also possible (Berger, 2006), the lack of asymptotic assumptions (such that analyses are also valid for small samples), flexibility in developing models, and an appealing hypothesis testing toolkit in the vein of Jeffreys (1961).

Some other properties of the Bayesian approach are particularly attractive when dealing with circular data. First, in the Bayesian approach it is trivially easy to obtain additional inference on functions of the parameters of a model. This is beneficial if the model form does not directly include interpretable parameters. This is exploited for the regression parameters of the Projected Normal regression model (Section 6.4.2) and the circular variance component of the Inverse Batschelet distribution (Section 6.5.1). Second, some models such as the Projected Normal regression models use data augmentation to facilitate estimation and inference, which is relatively simple in the Bayesian approach. Third, some likelihoods encountered in circular statistics are somewhat ill-behaved, such as those of the regression parameters of the von Mises regression model. This can be solved by introducing a specific prior on these parameters. Fourth, some circular data models are of such mathematical form that deriving standard errors is difficult, while the Bayesian approach allows us to compute our full uncertainty around the estimated parameters by investigating the posterior distribution.

To facilitate the use of Bayesian circular statistics, in this paper we introduce the R package **circbayes**, which is a diverse collection of methods for Bayesian inference on circular data. These methods include density estimation, regression models, hypothesis tests, mixture models and non-parametric models. Inference is generally performed through custom Markov chain Monte Carlo (MCMC) algorithms.

The rest of this paper will be structured as follows. Section 6.1 summarizes related packages for circular data analysis. Some general remarks on the structure of the package are given in Section 6.2. Basics of circular statistics are recapped in Section 6.3. Section 6.4 provides Bayesian inference for two types of circular regression models, one von Mises based and the other Projected Normal based. Density estimation problems are addressed in Section 6.5, using mixture models, both parametric in form as well as using a Dirichlet Process prior. Section 6.6 provides several Bayesian hypothesis tests, including examples for testing circular uniformity as well as tests for group differences (such as ANOVA). Finally, some concluding remarks are given in Section 6.7.

6.1 Related work

Three different groups of software packages related to circular statistics can be identified.

The first group are general packages for circular statistics including functions to obtain summary statistics, plots, hypothesis tests, and regression

models. General packages for circular statistics include the Stata toolbox **CIRCSTAT** (Cox, 1998), the MATLAB toolbox **CircStats** (Berens, 2009), and the R package **circular** (Lund et al., 2013). The package **circbayes** falls in this first category, and is unique in the fact that it provides Bayesian inference for a variety of problems in circular data.

Second, specific problems may require a single solution from the field of circular statistics, which are then packaged and often published in domain-specific journals. For example, in behavioural biology animal movements can be modeled using hidden Markov Models with circular regression components, which are implemented in **momentuHMM** (McClintock & Michelot, 2018). In criminology, spatial analysis with an interval-censored crime time can be performed using the package **aoristic** (Kikuchi, 2015). The package **NPCirc** (Oliveira et al., 2014) contains a variety of methods for non-parametric models on the circle. In cell biology, the cell division cycle calls for isotonic circular regression models, provided in **isocir** (Barragán et al., 2013). The R package **circE** (Grassi et al., 2010) implements circumplex models, which represent a type of latent factor analysis where the items can reasonably be constrained on a unit circle (Browne, 1992). Methods for analysis of circular data by wrapping distributions on the real line around the circle are provided in the R package **wrapped** (Nadarajah & Zhang, 2017). Directional distributions for use within dynamic Bayesian networks are provided by **Mocapy++** (Paluszewski & Hamelryck, 2010), which has useful applications in bioinformatics.

The third group are packages made for directional data, defined on p-dimensional hyperspheres. Because the circle is a 1-dimensional hypersphere, these packages often also provide functionality for circular data. An early package in MATLAB is **SPAK** (Leong & Carlile, 1998), providing spherical data analysis and visualization. Von Mises-Fisher Mixture models in any dimension can be fit in R using **movMF** (Hornik & Grün, 2014). The MATLAB package **libDirectional** (Kurz et al., 2017) implements methods for several sophisticated directional probability distributions as well as filtering. Methods specific to operations on objects oriented in three-dimensional space are available in the R packages **orientlib** (Murdoch et al., 2003) and **rotations** (Stanfill et al., 2014).

An alternative approach to perform Bayesian analyses of circular data is to use one of the probabilistic programming frameworks popular in Bayesian statistics, such as JAGS, BUGS or Stan. An issue is that these generally lack support for circular observations and circular parameters, although this may be circumvented by representing these as bivariate vectors constrained to lie on the unit circle. Some rudimentary support for circular probability distributions is available in some of these frameworks, usually for the von Mises

distribution only. In JAGS, this requires installing an extension (Stoneking et al., 2016).

The main problem in working with circular data in these probabilistic programming frameworks is that both the MCMC algorithm as well as methods for analysis of the MCMC output generally do not take into account the circular nature of the parameters. For example, to summarise a credible interval of a circular mean parameter, we must use circular credible intervals, which none of the probabilistic programming frameworks provides. An additional issue is that the MCMC algorithms implemented in **circbayes** often use custom steps in order to speed up the algorithm in some way, many of which are impossible or time-consuming to implement in these frameworks. Finally, plotting also requires attention for circular parameters, as standard plotting functions do not show the circular nature of the parameter space.

6.2 General package structure

The package **circbayes** is a collection of Bayesian inference algorithms for statistical models of circular observations, along with methods to analyze and compare the resulting fits. Each model is called through a fit function. Table 6.1 provides an overview of the names of the available methods as well as some of their properties, which will be described in more detail in the rest of this paper.

While programming languages usually work with angles described by radians, some users prefer to use other units such as degrees or hours. In this case, the user is recommended to work with their circular data as a `circular` S3 object from the package **circular**. Care must be taken that the `circular` object is given the correct unit type of the data, by setting the `units` argument accordingly.

```
R> th_numeric  <- c(100, 180, 340)
R> th_circular <- circular(th_numeric, units = "degrees")
```

Then, if these are entered into the fit functions of **circbayes**, they are automatically transformed to radians for internal use.

Because inference is generally done through posterior simulation (such as MCMC), several methods are always available. Results from the MCMC algorithms are returned as S3 objects (named by appending `_mod` to the function name), which provide easy access to these methods. In particular, methods such as `posterior_samples()`, `plot()`, `print()`, `coef()`, `marg_lik()` and `inf_crit()` are available for all models. Throughout all outputs, the circular parameters, such as mean direction parameters (ie. μ) are treated

Model	Fit function	MCMC	Marginal Likelihood
Circular uniform	-	-	$(2\pi)^{-n}$
Von Mises	`vm_posterior`	Yes	$\int_0^\infty I_0(R_n\kappa)/I_0(\kappa)^m d\kappa$
Projected Normal	`pn_posterior`	Yes	Bridgesampling
Batschelet	`bat_posterior`	Yes	Bridgesampling
Von Mises Regression	`vm_reg`	Yes	Bridgesampling
Projected Normal Regression	`pn_reg`	Yes	Bridgesampling
Hierarchical PN Regression	`pn_me_reg`	Yes	Bridgesampling
VM Mixture Models	`vm_mix`	Yes	Bridgesampling
Batschelet Mixture Models	`bat_mix`	Yes	Bridgesampling
Dirichlet Process Models	`vm_dpm`	Yes	-

Table 6.1: Overview of methods available for each model in **circbayes**. The columns represent the name of the model, the function to fit the model, whether the model is sampled with MCMC, how the marginal likelihood is calculated (see Section 6.6).

as such, so 'mean' or 'median' refers to the circular mean and median, for example. Lower and upper bounds given are generally highest posterior density (HPD) intervals. Modes are estimated as the midpoint of a short (ie. 10%) HPD interval (Venter, 1967).

The `plot()` methods for each model follow a similar pattern. Univariate models can be visualized either in polar coordinates (which is emphasizes the circular nature of the data) or on the cartesian plane (which sometimes facilitates comparisons along the y-axis), along with the best fit for the probability density function. Regression models can be plot as bivariate plots where a predictor can be selected to place on the x-axis. In addition, we exploit the fact that each iteration of an MCMC sample implies either a probability density function (given the current parameters) or a regression function (given the current parameters). As a result, we can plot the uncertainty as a number of samples from the probability density or regression functions from the posterior, using the option `n_samples = 100`, for example. Also, to reduce visual noise, the uncertainty can be plotted as credible intervals around the probability density or regression functions, using the option `add_ci = TRUE`.

Typical options for MCMC methods are available throughout. Setting `niter` allows the user to set the number of iterations to sample, while `burnin` allows the user to set the number of initial iterations to burn should the sampler take some time to converge, and finally `thin` allows the user to set a thinning factor, which will only keep a fraction of `1/thin` of the posterior samples, which can be useful to save on memory.

6.3 Circular statistics

Circular data requires descriptive statistics and probability distributions that are distinct from the usual linear statistics. Because these will be used throughout, they will shortly be recapped here. However, an extensive set of descriptive statistics is available in **circular**, so functions for these are not repeated in **circbayes**.

6.3.1 Summary statistics

Let $\boldsymbol{\theta} = \theta_1, \ldots, \theta_n$ be a set of circular observations, where $\theta_i \in [-\pi, \pi)$. The resultant length $R \in [0, n]$ is a measure of spread and depends on the sum of cosines and the sum of sines of the circular observations, so that

$$C = \sum_{i=1}^{n} \cos\theta_i, \qquad S = \sum_{i=1}^{n} \sin\theta_i, \qquad R = \sqrt{C^2 + S^2}. \tag{6.1}$$

One may also compute the mean resultant length $\bar{R} = R/n$, which is useful because $\bar{R} \in [0, 1]$, so it is comparable between datasets of different sizes.

The most important central tendency for circular data, the mean direction $\bar{\theta}$, can then be defined as

$$\cos\bar{\theta} = C/R, \qquad \sin\bar{\theta} = S/R, \tag{6.2}$$

which computationally can be obtained through $\bar{\theta} = \text{atan2}(S, C)$, where atan2 is the two-argument arctangent function, which computes the direction from the origin to the point (C, S). This function is directly implemented in R as `atan2`.

6.3.2 Probability distributions

Two approaches to derive circular distributions are used in **circbayes**. First, the intrinsic approach is described in Section 6.3.2. Second, the embedding approach is described in Section 6.3.2. Throughout, the models will be applied to the wind data (Agostinelli, 2007; Fisher, 1995), which is available in **circular** as `circular::wind`. The results are plotted in Figure 6.1.

Intrinsic approach

The intrinsic approach directly defines probability distributions on the circle. We provide methods for the von Mises distribution and an extension which allows for peaked and flat-topped densities, called Batschelet distributions.

Von Mises distribution The von Mises distribution is a symmetric and unimodal distribution on the circle, given by

$$\mathcal{M}(\theta \mid \mu, \kappa) = [2\pi I_0(\kappa)]^{-1} \exp\{\kappa \cos(\theta - \mu)\}, \tag{6.3}$$

where μ is the mean direction parameter, κ is a concentration parameter, and $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero. While of simple form, the Bessel function in the normalizing constant makes direct inference for $\{\mu, \kappa\}$ difficult.

To facilitate computation, we use conjugate priors where possible. A conjugate prior for (μ, κ) is given by Guttorp & Lockhart (1988) as

$$p(\mu, \kappa) \propto [I_0(\kappa)]^{-c} \exp\{R_0 \kappa \cos(\mu - \mu_0)\}, \tag{6.4}$$

where μ_0 is the prior mean direction, R_0 is the prior resultant length, and c can be interpreted as a prior 'sample size'. Then, the von Mises posterior is given by

$$p(\mu, \kappa \mid \boldsymbol{\theta}) \propto [I_0(\kappa)]^{-m} \exp\{R_n \kappa \cos(\mu - \mu_n)\}, \tag{6.5}$$

where the posterior summary statistics (μ_n, R_n, m) are computed by first obtaining $C_n = C + C_0$, $S_n = S + S_0$, then computing $R_n = \sqrt{C_n^2 + S_n^2}$, $\mu_n = \text{atan2}(S_n, C_n)$, and $m = n + c$.

Bayesian inference for μ and κ can proceed through Gibbs sampling (Chib & Greenberg, 1995). The full conditional distribution of μ is von Mises, and given by

$$p(\mu \mid \kappa, \boldsymbol{\theta}) = \mathcal{M}(\mu_n, R_n \kappa) \tag{6.6}$$

for which an efficient algorithm is available from Best & Fisher (1981). While that algorithm for random sampling from the von Mises distribution is available in **circular**, for efficiency a version was implemented in C++ through **Rcpp** (Eddelbuettel & François, 2011), and provided in **circbayes** as `rvm`.

```
R> th <- rvm(100, mu = 2, kp = 10)
```

The conditional distribution of κ is not of closed form. The most efficient algorithm to sample κ is given by Forbes & Mardia (2015), so we have implemented their rejection sampler in C++ as well.

An MCMC sample from the von Mises posterior with an uninformative prior can be obtained by setting the prior parameters to $\mu_0 = 0, R_0 = 0, c = 0$, or an informative prior can be specified, by setting $c > 0$ and μ_0 and R_0 accordingly.

```
R> # Prior order is mu_0, R_0, c.
R> vm_mod <- vm_posterior(circular::wind, prior = c(0, 0, 0))
R> vm_mod
```

```
       estimate    se  2.5% 97.5%
mu        0.293 0.054 0.186 0.400
kappa     1.728 0.149 1.466 2.053
```

Batschelet Distribution The family of densities introduced by Batschelet (1981) allows for peaked or flat-topped distributions on the circle. They are constructed by transforming the input of the probability density θ through a function $t_\lambda(\theta)$. This value is then passed into some base density f_0 of choice (Abe et al., 2010; Pewsey et al., 2011). Then, the Batschelet distribution is given by $f(\theta) \propto f_0(t_\lambda(\theta))$, where $\lambda \in (-1, 1)$ determines the shape of the distribution, usually with $\lambda = 0$ meaning no transformation. We implement two choices of $t_\lambda(\theta)$, but we always take the von Mises distribution as the base density.

The von Mises based symmetric Inverse Batschelet (IB) density as given in Jones & Pewsey (2012) can be written as

$$f_{IB}(\theta \mid \mu, \kappa, \lambda) = [2\pi I_0(\kappa) K_{\kappa,\lambda}]^{-1} \exp\{\kappa \cos t_\lambda(\theta - \mu)\} \tag{6.7}$$

where

$$t_\lambda(\theta) = \frac{1-\lambda}{1+\lambda}\theta + \frac{2\lambda}{1+\lambda} s_\lambda^{-1}(\theta), \tag{6.8}$$

with $s_\lambda^{-1}(\theta)$ is the inverse of $s_\lambda(\theta) = \theta - \frac{1}{2}(1+\lambda)\sin(\theta)$, and the normalizing constant given by

$$K_{\kappa,\lambda} = \frac{1+\lambda}{1-\lambda} - \frac{2\lambda}{1-\lambda} \int_{-\pi}^{\pi} [2\pi I_0(\kappa)]^{-1} \exp\left\{\kappa \cos\left(\theta - (1-\lambda)\sin(\theta)/2\right)\right\} d\theta. \tag{6.9}$$

Because of the implicit inverse in the transformation function, this probability density is relatively costly to compute, requiring a numerical inverse.

Therefore, an alternative transformation function is also available which results in the so-called Power Batschelet distribution. It is constructed to approximate the shape of the IB distribution, while being easier to compute. The transformation function used for the Power Batschelet distribution is given by

$$t_\lambda^*(\theta) = \text{sign}(\theta)\pi \left(\frac{|\theta|}{\pi}\right)^{\gamma(\lambda)}, \tag{6.10}$$

where $\gamma(\lambda) = \frac{1-c\lambda}{1+c\lambda}$, where c is some fixed constant, set to $c = 0.4052284$ to approximate the IB density closely. The normalizing constant is changed accordingly.

Regardless of the type of Batschelet distribution, there no longer is a conjugate prior. Therefore, one can select any non-conjugate prior to be used

for the parameters of this model. The priors are assumed to be independent, however, and can be provided as functions that return log-probability of the prior (up to an additive constant). For example, we can obtain a sample from the posterior of the inverse Batschelet distribution with a circular uniform prior on μ (the default), a Gamma(2, 0.1) prior on κ, and a $Beta(3, 3)$ prior on $(\lambda + 1)/2$ (where the shifting is done to place the prior on the range of λ), as follows.

```
R> kp_logp  <- function(kp) dgamma(kp, 2, 1/10,
R+                                 log = TRUE)
R> lam_logp <- function(lam) dbeta((lam + 1)/2, 3, 3,
R+                                 log = TRUE)
R>
R> pb_mod <- bat_posterior(circular::wind,
R+                         kp_logprior_fun  = kp_logp,
R+                         lam_logprior_fun = lam_logp)
R> pb_mod

               mean median    se  2.5% 97.5%
mu            0.172  0.154 0.035 0.135 0.286
kp            1.775  1.782 0.179 1.604 1.924
lam           0.790  0.834 0.190 0.017 0.942
circ_variance 0.328  0.325 0.031 0.285 0.386
circ_sd       0.891  0.887 0.052 0.820 0.988
```

This distribution does not have interpretable parameters beyond μ, because κ and λ both determine variance and shape. Therefore, results are also provided for two derived quantities, the circular variance and circular standard deviation, as given in (Fisher, 1995). The Power Batschelet can be selected by setting `bat_type = "power"`, while the Inverse Batschelet is obtained by `bat_type = "inverse"`.

Embedding approach

The embedding (or projected) approach treats the circular observations by seeing them as projected bivariate observations. The idea behind this projection is that we do not have to conduct inference on the circular variable $\Theta \in [-\pi, \pi)$ directly, but we can indirectly conduct inference on a bivariate variable $Y \in \mathbb{R}^2$. The bivariate variable Y is not observed but inferred from Θ. Their relation is given by

$$\boldsymbol{y} = \boldsymbol{u}r, \tag{6.11}$$

where $\boldsymbol{u} = (\cos\theta, \sin\theta)$ and $r = ||\boldsymbol{y}||$. A common example of a projected distribution is the Projected Normal (PN) distribution, given by

$$PN(\theta \mid \boldsymbol{\mu}, \boldsymbol{I}) = \frac{1}{2\pi} e^{-\frac{1}{2}||\boldsymbol{\mu}||^2} \left[1 + \frac{\boldsymbol{u}^t \boldsymbol{\mu} \Phi(\boldsymbol{u}^t \boldsymbol{\mu})}{\phi(\boldsymbol{u}^t \boldsymbol{\mu})}\right], \tag{6.12}$$

where $\theta \in [-\pi, \pi)$ is the circular observation, $\boldsymbol{\mu} = (\mu_1, \mu_2)^t \in \mathbb{R}^2$, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{u} = (\cos\theta, \sin\theta)^t$ and $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and probability density function of the standard normal distribution, respectively. The Projected Normal distribution is rotationally symmetric about its mean direction $\boldsymbol{\mu}/||\boldsymbol{\mu}||$ and its concentration is dependent on $||\boldsymbol{\mu}||^2$ (see Kendall (1974) for the exact form of this relation). Note that the covariance matrix is specified to be identity for identification purposes. An alternative parametrization of the Projected Normal distribution can be found in Wang & Gelfand (2013).

A conjugate prior for $\boldsymbol{\mu}$ is a bivariate Normal distribution $N_2(\boldsymbol{\mu}_0, \lambda_0 \boldsymbol{I})$, resulting in a bivariate Normal conditional posterior given by (Nunez-Antonio & Gutierrez-Pena, 2005)

$$p(\boldsymbol{\mu} \mid \boldsymbol{X}) = N_2(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n), \tag{6.13}$$

where $\boldsymbol{X}$ is a $n \times 2$ design matrix, $\boldsymbol{\mu}_n$ is a two dimensional vector where each dimension $j \in I, II$ is defined by $(\lambda_0 \mu_0^j + n\bar{\boldsymbol{X}}_j)/(\lambda_0 + n)$ and n equals the sample size. The 2×2 matrix $\boldsymbol{\Lambda}_n$ is diagonal with j$^t h$ element given by $\lambda_0 + n$. The conditional density for r is given by

$$p(r \mid \boldsymbol{\theta}, \boldsymbol{\mu}) \propto r \exp\{-0.5r^2 + br\}, \tag{6.14}$$

where $b = \boldsymbol{u}^t \boldsymbol{\mu}$, and $\boldsymbol{\theta} = \theta_1, \ldots \theta_n$.

We can sample from the posterior of the Projected Normal distribution using a Gibbs sampler (Chib & Greenberg, 1995) with a Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970a) step, as done by Nunez-Antonio & Gutierrez-Pena (2005), or by a slice sampling (Hernandez-Stumpfhauser et al., 2017b) step for r. In **circbayes** this second option was implemented in C++ through **Rcpp**.

To obtain an MCMC sample from the Projected Normal posterior we can use `pn_posterior`.

```
R> pn_mod <- pn_posterior(circular::wind)
R> pn_mod
```

```
     mean  mode    se  2.5% 97.5%
mu1 1.296 1.277 0.085 1.151 1.461
mu2 0.306 0.289 0.063 0.181 0.431
```

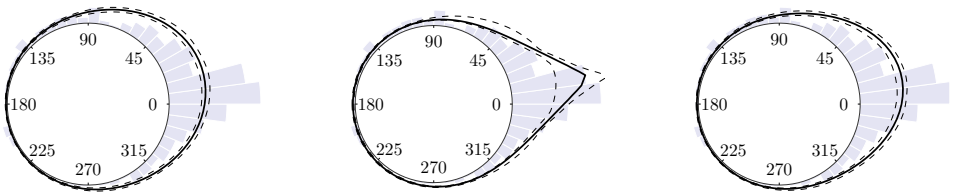


Figure 6.1: Three posterior fits of univariate symmetric circular probability densities, which are the von Mises distribution (left), the Power Batschelet distribution (middle) and the Projected Normal distribution (right). Dashed lines give the 95 percent credible interval of the probability density.

Alternatives

Two alternative approaches not used in **circbayes** are the wrapping approach and the stereographic projection approach. The wrapping approach augments the circular observations such that they can be represented as $\theta = x \mod 2\pi$, where $x \in \mathbb{R}$ is some unobserved data point which has a known distribution such as the Normal distribution. To reverse this, we can say that $x = \theta + k2\pi$, where $k \in \mathbb{Z}$ is a latent wrapping factor. Methods for wrapped models are available in the package **wrapped** (Nadarajah & Zhang, 2017). Another option is given by the stereographic projection approach, which uses projection from the real line to a circle below it (Abe et al., 2010), but few implementations or extensions exist.

6.4 Bayesian Circular regression

Circular regression refers to regression models where the outcome is a circular observation. If the circular observation θ is a predictor instead, often adding $\cos\theta$ and $\sin\theta$ to the predictors suffices (Fisher, 1995).

For circular regression, **circbayes** implements two approaches which we find the most promising. The first is a Generalized Linear Model (GLM) with an arctangent link function and von Mises residuals. We will call this the von Mises (VM) regression model, and it will be discussed in Section 6.4.1.

The second model is the Projected Normal (PN) regression model, where the residuals have the Projected Normal distribution. While the Projected Normal distribution is more difficult to work with, this leads to a regression function that is simpler to work with. This model will be discussed in Section 6.4.2.

Both models will exemplified using the `Motor` data (Puglisi et al., 2017) from **bpnreg**. The resulting models are plotted in Figure 6.2.

6.4.1 Circular regression with von Mises residuals

Consider the problem of regressing a circular outcome θ, on a column vector of continuous linear predictors $\boldsymbol{x}_i \in \mathbb{R}^K$, assumed to be standardized, and a column vector of dichotomous predictors $\boldsymbol{d}_i \in \{0, 1\}^J$. Assume that each observed angle θ_i is generated independently from a von Mises distribution $\mathcal{M}(\theta_i \mid \mu_i, \kappa)$. Then, μ_i is chosen to be

$$\mu_i = \beta_0 + \boldsymbol{\delta}^T \boldsymbol{d}_i + g(\boldsymbol{\beta}^T \boldsymbol{x}_i), \tag{6.15}$$

where $\beta_0 \in [-\pi, \pi)$ is an offset parameter which serves as a circular intercept, $\boldsymbol{\delta} \in [-\pi, \pi)^J$ is a column vector of circular group difference parameters, $g(\cdot) : \mathbb{R} \to (-\pi, \pi)$ is a twice differentiable link function, and $\boldsymbol{\beta} \in \mathbb{R}^K$ is a column vector of regression coefficients. The link function is often chosen to be $g(x) = 2 \arctan(x)$, where $\arctan(\cdot)$ is the arctangent, the inverse tangent function.

This model was first described by Fisher & Lee (1992), with an initial application of a Bayesian version of this method given by J. Gill & Hangartner (2010). The version in **circbayes** given above, where categorical predictors are treated separately, was developed in Mulder & Klugkist (2017).

The priors are assumed to be independent between $\{\beta_0, \kappa\}$, which has the conjugate prior for the parameters of the von Mises distribution, and the predictors $\{\boldsymbol{\beta}, \boldsymbol{\delta}\}$, which are given a Normal prior. As described by J. Gill & Hangartner (2010), the likelihood of the regression coefficients $\{\boldsymbol{\beta}, \boldsymbol{\delta}\}$ is ill-behaved in the sense that it can be multimodal and have non-zero asymptotes. If an uninformative prior is used, the posterior inherits this undesirable behaviour. This issue can be solved by giving these parameters a Normal$(0, 1)$ prior (see Mulder & Klugkist (2017)), which is only very weakly informative for this model if the predictors are standardized. Therefore, this is the default.

The VM regression model can be performed using the function `vm_reg`.

```
R> Motor$AvAmp <- scale(Motor$AvAmp)
```

```
R> fit_vmreg <- vm_reg(Phaserad ~ AvAmp, data = Motor,
R+                     prior_b0kp = c(0, 0, 0),
R+                     prior_btdt = c(0, 1))
R> coef(fit_vmreg)

           estimate         se        2.5%     97.5%
Intercept 0.6128075 0.15718544  0.29447560 0.8995695
Kappa     1.4656703 0.38317651  0.85921974 2.3125049
AvAmp     0.0599043 0.07670142 -0.09664822 0.2016793
```

The intercept β_0 and concentration parameter κ are sampled as described in Section 6.3.2. The regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are sampled using a Metropolis-Hastings step. For details, see (Mulder & Klugkist, 2017). This function will automatically select dichotomous predictors and treat them accordingly. Posterior summaries can be obtained through `coef(fit_vmreg)`.

6.4.2 Circular regression with Projected Normal residuals

In the Projected Normal (PN) regression model (Nuñez-Antonio et al., 2011), the bivariate mean vector μ is made dependent on a set of K predictors $\boldsymbol{X}$. This is done by defining the mean vector in (6.12) as

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_i^I \\ \mu_i^{II} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{\beta}^I)^t \boldsymbol{X}_i^I \\ (\boldsymbol{\beta}^{II})^t \boldsymbol{X}_i^{II} \end{pmatrix}, \tag{6.16}$$

where $\boldsymbol{\beta}^I \in \mathbb{R}^K$ and $\boldsymbol{\beta}^{II} \in \mathbb{R}^K$ are vectors with regression coefficients and $\boldsymbol{X}_i^I$ and $\boldsymbol{X}_i^{II}$ are design matrices, which are often equal but may also be different. As in the VM regression model, the mean direction is dependent on the predictors through a link function.

In **circbayes** PN regression is implemented through the function `pn_reg()`. We can fit the PN regression model as follows.

```
R> fit_pnreg <- pn_reg(Phaserad ~ AvAmp, data = Motor)
R> print(coef(fit_pnreg)$linear, digits = 3)

               mean    mode    se   2.5% 97.5%
(Intercept)  0.9529  0.9317 0.188  0.599 1.339
AvAmp       -0.0149 -0.0589 0.199 -0.385 0.396
(Intercept)  0.6640  0.6742 0.183  0.307 0.999
AvAmp        0.1073  0.1558 0.187 -0.226 0.500
```

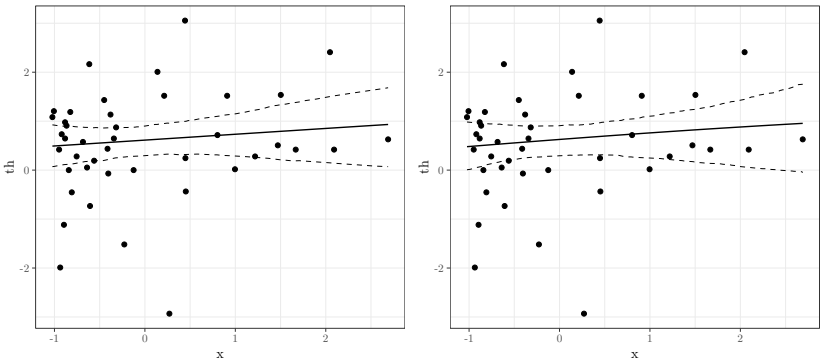


Figure 6.2: Univariate regression plots for the von Mises (left) and Projected Normal (right) regression models. The dashed lines provide the 95 percent credible interval of the regression line.

Bayesian inference is performed using a Gibbs sampler with slice sampling step for r (Hernandez-Stumpfhauser et al., 2017b) and implemented in C++ through **Rcpp**.

Regression Coefficients

The vectors $\boldsymbol{\beta}^I$ and $\boldsymbol{\beta}^{II}$ contain the linear regression coefficients. MCMC samples can be obtained using `posterior_samples()`.

Posterior summaries can be obtained through `coef_lin(fit_pnreg)`.

Note that due to the parametrization in bivariate space there are two regression coefficients for each predictor. In order to facilitate interpretation, three circular regression coefficients were introduced in Cremers, Mulder, & Klugkist (2018). These coefficients in circular space are obtained by a reparametrization of the marginal effect of one predictor x by

$$\begin{aligned}\hat{\theta} = \text{atan2}\left(\hat{y}^{II},\ \hat{y}^{I}\right) &= \text{atan2}\left(\beta_0^{II} + \beta_1^{II}x,\ \beta_0^{I} + \beta_1^{I}x\right) \\ &= a_c + \arctan\left[b_c\left(x - a_x\right)\right], \end{aligned} \tag{6.17}$$

where atan2 is the link function, $\hat{\theta}$ is the circular predicted value, $\hat{y}^{II}$ and $\hat{y}^{I}$ are the predicted values in bivariate space and b_c is the circular regression coefficient. Note that the two other regression coefficients SAM and AS are dependent on both b_c and x and return the slope at different parts of the circular regression line. Posterior summaries of the circular regression coefficients can be obtained through `coef_circ(fit_pnreg)`. For categorical predictors, the posterior descriptives of the predicted circular values of the different (combinations of) categories are returned.

6.4.3 Comparison between the two regression models

Although the structure of the two models is quite different, both the VM and PN regression models predict a point on the circle conditional on a set of predictors, and both have a fairly simple residual structure. In fact, both methods feature a regression function that takes the shape of an arctangent function. Therefore, one might wonder how to select a regression model for any practical problem. We will highlight three differences.

A first major distinction between the two models is that homogeneity of variance is enforced for the VM regression model, while the PN regression model has an inherent heterogeneity of variance, so that the circular residual variance changes as the predicted value changes. Whether heterogeneity of variance is required depends on whether the data exhibits this property. In order to test this, one may perform model comparison after running both models.

A second major distinction is that for the PN regression model, a single predictor is able to move the circular outcome across a semicircle, while in the VM regression model the reach of a single predictor can be selected, but is usually the full circle.

Thirdly, the PN regression model admits extensions in a somewhat more straightforward manner, because existing solutions for bivariate data can simply be plugged into the model. This has for instance led to a hierarchical version of the PN regression model, which will be presented next.

To make the selection between the two regression models, we recommend applying both methods to the dataset and applying the methods discussed in Section 6.6 to make a selection of the best-fitting method.

6.4.4 Hierarchical Projected Normal regression

The Hierarchical Projected Normal regression model is useful for modeling dependent circular data (Nuñez-Antonio & Gutiérrez-Peña, 2014). We define the mean vector in (6.12) as

$$\boldsymbol{\mu}_{ij} = \begin{pmatrix} \mu_{ij}^{I} \\ \mu_{ij}^{II} \end{pmatrix} = \begin{pmatrix} (\boldsymbol{\beta}^{I})^{t}\boldsymbol{X}_{ij}^{I} + (\boldsymbol{b}_{i}^{I})^{t}\boldsymbol{Z}_{ij}^{I} \\ (\boldsymbol{\beta}^{II})^{t}\boldsymbol{X}_{ij}^{II} + (\boldsymbol{b}_{i}^{II})^{t}\boldsymbol{Z}_{ij}^{II} \end{pmatrix}, \tag{6.18}$$

where $\boldsymbol{\beta}^{I}$ and $\boldsymbol{\beta}^{II}$ are vectors with fixed effect coefficients and intercept, $\boldsymbol{b}_{i}^{I}$ and $\boldsymbol{b}_{i}^{II}$ are vectors with random effects for each group i (or for individuals i when we have longitudinal observations) and $\boldsymbol{X}_{ij}^{I}$, $\boldsymbol{X}_{ij}^{II}$, $\boldsymbol{Z}_{ij}^{I}$ and $\boldsymbol{Z}_{ij}^{II}$ are design matrices.

In **circbayes** hierarchical PN regression is implemented through the function `pn_me_reg()`. Using the `Maps` data (Warren et al., 2017) we can fit a hierarchical PN regression model as follows:

```
R> pmer_mod <- pn_me_reg(formula = Error.rad ~ Maze +
R+                         Trial.type + L.c + (1 | Subject),
R+                       data = Maps)
R> print(coef(pmer_mod)$linear, digits = 3)
```

```
                mean     mode     se    2.5%   97.5%
(Intercept)  2.38992  2.41007 0.2600  1.9144  2.8698
Maze1       -0.91917 -1.04614 0.3162 -1.5074 -0.3499
Trial.type1  0.38481  0.40711 0.2550 -0.1355  0.8628
L.c          0.00546  0.01248 0.0406 -0.0810  0.0772
(Intercept) -0.57438 -0.58326 0.1598 -0.8888 -0.2594
Maze1        0.51194  0.43819 0.2121  0.0892  0.8939
Trial.type1  1.15050  1.16576 0.1852  0.8111  1.5337
L.c         -0.00778  0.00576 0.0310 -0.0709  0.0478
```

Bayesian estimation is performed using a Gibbs sampler with slice sampling step for r (Hernandez-Stumpfhauser et al., 2017b).

Regression Coefficients

MCMC samples for the fixed and random effects can be obtained by using the function `posterior_samples()`, posterior summaries can be obtained through `coef_lin()` and `coef_ran()`. To obtain posterior summaries for the circular coefficients (Cremers, Pennings, et al., 2018) we use `coef_circ()` and `coef_ran(pmer_mod, type = 'circular')`.

6.5 Density estimation problems

A different class of circular data problems deals with estimating densities on the circle, where the goal is to predict or understand where on the circle the circular observations are likely to occur. This is particularly common for multimodal circular observations, especially derived from temporal measurements spread across the hours of the day or the days of the year. For example, density estimation have been succesfully applied in predicting wildfires (Ameijeiras-Alonso et al., 2018; Ley & Verdebout, 2018) which exhibit a strong seasonal pattern over the year. Another example is the prediction

of crime times over the day (Ashby & Bowers, 2013). Outside the temporal domain, circular mixtures have been applied to eye movement directions (Van Renswoude et al., 2016).

The simplest approach is to obtain the posterior for a symmetric unimodal density, such as those described in Section 6.3.2. For this, methods are provided for von Mises, Inverse Batschelet and Projected Normal distributions. All three require MCMC sampling and were described in Section 6.3.2. However, often the data under consideration is multimodal or otherwise irregular. For such data, **circbayes** implements two models which can fit a wide variety of circular data distributions.

The first is a family of mixture distributions, containing both von Mises and Batschelet Mixtures. Mixture models are appropriate if we expect the circular observations to be generated by multiple distinct data generating processes, for which labels were not observed. In these cases, interest is in parameter estimation, posterior prediction of new observations, and possibly reassigning the labels of the unobserved classes. A disadvantage is that the number of underlying components must be chosen. These models are discussed in Section 6.5.1.

The second is the Dirichlet process mixture (DPM) model. This is a model from the field of Bayesian non-parametrics, where the number of components does not need to be set, while the model can still learn any data generating density. The DPM model is useful when one does not care about or know about the number of components required, but rather wishes to flexibly estimate possible probability densities on the circle. This model is discussed in Section 6.5.2.

These models will be exemplified on the turtles data set (M. A. Stephens, 1969; Fisher, 1995), provided in **circular** as `circular::fisherB3c`. They are plotted in Figure 6.3.

6.5.1 Mixture models

The mixture models available in **circbayes** are of the form

$$f(\theta \mid \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{j=1}^{J} \alpha_j f_B(\theta \mid \mu_j, \kappa_j, \lambda_j), \tag{6.19}$$

where j indexes the J components in the mixture, α_j are component weights, and $f_B(\cdot)$ is the chosen probability density for each component in the mixture. In this case, the probability density can be either the von Mises distribution (in which case we fix $\boldsymbol{\lambda} = \mathbf{0}_J$, a zero vector) or the Inverse Batschelet distribution.

The von Mises mixture can be performed by

```
R> th <- circular::fisherB3c
R> vm_mix_fit <- vm_mix(circular::fisherB3c, n_comp = 2)
R> vm_mix_fit

Mixture of power Batschelet distributions, using method 'bayes'.
              mu        kp lam      alph   circ_var   circ_sd
comp_1 -2.078875 10.334360   0 0.1565991 0.04968782 0.3192640
comp_2  1.108694  2.464985   0 0.8434009 0.23886648 0.7388457
```

Mixtures of Batschelet distributions, either Power or Inverse, can be performed using the `bat_mix()` command.

```
R> pb_mix_fit <- bat_mix(th = th, n_comp = 2,
R+                       bat_type = "power")
R> pb_mix_fit

               mean  median      se   2.5%    97.5%
mu_1         -2.107  -2.109   0.089 -2.247   -1.976
mu_2          1.100   1.097   0.064  0.977    1.223
kp_1        318.996 230.728 310.043  4.561 1012.603
kp_2          2.767   2.552   0.870  1.592    4.917
lam_1        -0.659  -0.745   0.262 -0.994   -0.083
lam_2        -0.039  -0.096   0.261 -0.420    0.458
alph_1        0.161   0.160   0.048  0.077    0.273
alph_2        0.839   0.840   0.048  0.727    0.923
circ_var_1    0.052   0.047   0.035  0.018    0.142
circ_var_2    0.248   0.241   0.052  0.155    0.355
circ_sd_1     0.316   0.310   0.093  0.189    0.554
circ_sd_2     0.752   0.743   0.092  0.580    0.936
```

For this type of mixture, any parameter can be fixed to specific values by providing an $4 \times J$ matrix of parameter values to the argument `fixed_pmat`, where any value to be freely estimated is set to `NA`.

6.5.2 Bayesian Nonparametrics for Circular data

The Dirichlet Process Mixture model is a popular method in Bayesian nonparametrics due to its ability to fit data from any data generating process. It essentially is an infinite mixture model, where the effective number of mixtures is regulated by a stick-breaking prior (Ishwaran & James, 2001).

Therefore, it can be seen as a mixture model for which we do not need to select a fixed number of components. Another way to view these methods is as a Bayesian alternative to kernel density methods such as those in **NPCirc** (Oliveira et al., 2014), although the DPM model is a true statistical model rather than a descriptive method. For a more in-depth introduction, see Gelman et al. (2003, ch. 23).

We provide methods for circular distributions that can be used with the **dirichletprocess** package (J. Ross & Markwick, 2018). For an in-depth introduction to the DPM model and the use of the **dirichletprocess** package, see its vignette. The Von Mises Dirichlet process model can be written as

$$\theta_i \mid (\mu_i, \kappa_i) \sim \mathcal{M}(\mu_i, \kappa_i) \tag{6.20}$$
$$(\mu_i, \kappa_i) \mid G \sim G \tag{6.21}$$
$$G \sim \mathrm{DP}(P_0, \alpha), \tag{6.22}$$

where P_0 is a base distribution (or base measure) on the parameters of the von Mises distribution, and α is a parameter of the Dirichlet Process which influences the number of mixture components used. We use the conjugate prior given in Equation 6.4 as the base distribution. In contrast to earlier applications of this prior, it must be proper here (that is, integrate to 1). It is generally sufficient to set $c \geq 1$, although note that this means that if we also set $R_0 > 0$, the prior is informative towards the mean direction μ_0. Therefore, the default sets $\{\mu_0 = 0, R_0 = 0, c = 1\}$.

The parameter α is given a Gamma(a, b) prior and sampled in the MCMC algorithm to avoid the need to tune this parameter. Setting the parameters of this prior $\{a, b\}$ must still be done with care. To evaluate this, one can run `DiagnosticPlots()` on the resulting object, which will provide an figure where one can see whether the sampled values of α differ strongly from the prior.

This Dirichlet Process Mixture model can be run by

```
R> dpm_model <- vm_dpm(th = th,
R+                     alphaPriors = c(2, 4),
R+                     g0Priors = c(0, 0, 1))
R> dpm_model

Dirichlet process object run for 1000 iterations.

  Mixing distribution       vonmises
  Base measure parameters    0, 0, 1
  Alpha Prior parameters        2, 4
```

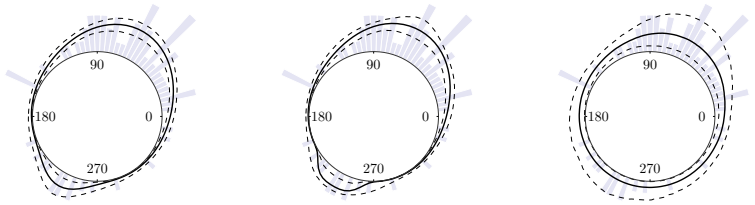


Figure 6.3: Density estimation plots, for the von Mises mixture (left), Batschelet mixture (middle), and Dirichlet Process Mixture (right) models. Dashed lines indicate the 95 percent credible interval of the probability density.

```
Conjugacy                 conjugate
Sample size                      76

Mean number of clusters        3.73
Median alpha                   0.52
```

In each iteration of the MCMC algorithm of the DPM model, the number of parameters changes. Therefore, the posterior samples are provided as a list.

6.6 Hypothesis tests and model comparison

Circular statistics packages such as **circular** often include a large variety of frequentist hypothesis tests. Bayesian counterparts of almost all of such tests are available in **circbayes**. A Bayesian approach to testing two discrete hypotheses against one another is to update the prior odds of two hypotheses by multiplying them by the Bayes factor (Kass & Raftery, 1995; Jeffreys, 1961), in order to produce the posterior odds of the two hypotheses. It should be noted that *hypotheses* here refers to distinct statistical models from which the data may have been generated, and we will therefore refer to them as model M_s with an associated set of parameters $\boldsymbol{\phi}_s$. In fact, the Bayes factor is computed as the ratio of the marginal likelihoods of the two models

$$BF_{12} = \frac{p(D \mid M_1)}{p(D \mid M_2)}, \tag{6.23}$$

where D represents the data, and the marginal likelihood is given by

$$p(D \mid M_s) = \int p(D, \boldsymbol{\phi}_s \mid M_s) d\boldsymbol{\phi}_s, \tag{6.24}$$

where the integral is over the parameter space of $\boldsymbol{\phi}_s$.

If prior probability $p(M_s)$ is available for model M_s, it can be introduced by multiplying the marginal likelihood with it in the Bayes factor, so that we obtain the posterior odds of each model given by

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{p(D \mid M_1)p(M_1)}{p(D \mid M_2)p(M_2)}. \tag{6.25}$$

If all models are a priori assumed equally likely, we have $p(M_s \mid D) \propto p(D \mid M_s)$, so the Bayes factor equals the posterior odds.

An advantage of the Bayesian hypothesis test is that it is possible to find evidence in favor of either hypothesis, so that all hypothesis under consideration could be chosen as most likely. This is in stark contrast with the frequentist Null Hypothesis Significance Testing (NHST) approach, where the null hypothesis can at most attain the status of 'not rejected'.

A disadvantage of the Bayesian hypothesis test is that care must be taken towards the influence of the prior. Improper priors, or priors chosen unwisely, can result in untrustworthy results. Therefore, we generally recommend to employ the Bayesian hypothesis test only with proper and at least weakly informative priors. In addition, a sensitivity analysis can provide useful insight in whether the prior is influential for a hypothesis test.

Besides the posterior odds between two hypotheses, we can also obtain the posterior model probability of each model within a set of models $M_1, \ldots, M_Q$. Assuming equal prior probabilities, we can compute

$$p(M_s \mid D) = \frac{p(D \mid M_s)}{\sum_{q=1}^{Q} p(D \mid M_q)}. \tag{6.26}$$

This is an intuitive probability of interest, because it represents our current belief that M_s is true rather than one of the other models under consideration. This probability might further be used to make decisions, or in Bayesian model averaging (Hoeting et al., 1999; Bao et al., 2010).

Due to the quite general form of the Bayesian hypothesis test, any two statistical models can be compared as long as their marginal likelihoods can be computed. Section 6.6.1 will first provide a simple alternative form of model comparison, which is to compare models on their (Bayesian) information criteria (Gelman et al., 2003, Ch. 7) instead of computing Bayes factors. However, Bayes factors are generally preferred for model comparison. If

nested models are compared, Bayes factors can be obtained somewhat easily, as discussed in Section 6.6.2. A more general solution using bridge sampling, which works for all models, is discussed in Section 6.6.3.

6.6.1 Information criteria

For all models in **circbayes**, information criteria can be obtained by running `inf_crit()` on the model.

```
R> inf_crit(fit_vmreg)
```

```
               value
AIC_Bayes 125.193987
DIC       125.633310
DIC_alt   126.486741
p_DIC       3.219662
p_DIC_alt   3.646377
p_WAIC1     3.005935
p_WAIC2     3.226626
WAIC1     125.419584
WAIC2     125.860965
```

For most models, we provide the DIC (Spiegelhalter et al., 2002) and the WAIC (Watanabe, 2010). For a discussion of the strengths and weaknesses of these, see Gelman et al. (2003, Ch. 7). Model comparison on the basis of these information criteria can be seen as an approximation to model comparison using the marginal likelihood, so they should only be preferred if computing the marginal likelihood is difficult or impossible.

6.6.2 Nested comparison Bayes Factors

Sometimes interest is in comparing a model M_s with a nested submodel M_s^*, which only differs in that some parameters are set to a fixed value. For example, one may compare a regression model with regression coefficient β with a regression model where that $\beta = 0$. For those cases, the Savage-Dickey method of obtaining Bayes factors is applicable (Dickey et al., 1970; O'Hagan & Forster, 2004; Wagenmakers et al., 2010) and markedly easier to compute than the marginal likelihood. Similarly, if one wishes to compare hypotheses related to inequalities, such as $\beta < 0$ versus $\beta > 0$, then the encompassing prior approach is applicable Klugkist et al. (2005); Wetzels et al. (2010). Bayes factors for both approaches are provided for the VM regression model.

6.6.3 Marginal Likelihood computation

To compute the marginal likelihood in (6.24) we must integrate once over each parameter of the model. Therefore, models with more than a few parameters tend to have a marginal likelihood that is difficult to compute, because numerical integration is no longer an option. For those models, there exists a literature of methods to approximate the marginal likelihood (for an overview, see Friel & Wyse (2012) and Ardia et al. (2012)).

We employ bridge sampling (Meng & Wong, 1996; Gronau, Sarafoglou, et al., 2017), which is implementend in the package **bridgesampling** (Gronau, Singmann, & Wagenmakers, 2017) to which methods for circular variables were contributed. An overview of the methods used for computing Bayesian hypothesis tests for each of the models in **circbayes** is given in Table 6.1. For almost all of the models, computing the marginal likelihood through bridge sampling is possible.

This most general form of the Bayesian hypothesis test thus allows us to compare any two models, which do not have to be nested. For example, this approach can be used to test whether the VM or PN regression model fits the data better, which set of predictors perform well in a regression, or whether a mixture model requires three or four components.

6.6.4 Hypothesis testing examples

This section will provide some examples of hypothesis tests that could be performed in the framework described previously.

Testing for circular uniformity

An important initial test is whether the data is uniform. Specific frequentists tests are most powerful against a specific alternative. Similarly, Bayesian hypothesis tests require selection of a specific alternative. The equivalent of the Rayleigh test (Mardia & Jupp, 2000; Brazier, 1994) in the Bayesian paradigm can be performed by comparing the von Mises marginal likelihood with the circular uniform marginal likelihood.

```
R> th <- circular::fisherB9c
R>
R> vm_mod <- vm_posterior(th, prior = c(0, 0, 1))
R>
R> cu_ml  <- marg_lik_circ_unif(th)
R> vm_ml  <- marg_lik(vm_mod)
R> bht_compare(uniform = cu_ml, von_mises = vm_ml)
```

```
Bayesian Hypothesis Test
    Comparing 2 models: uniform, von_mises

[Log Marginal Likelihood]
  uniform von_mises
-512.7677 -513.9991

[Posterior Model Probabilities]
  uniform von_mises
 0.774065  0.225935

[Pairwise log Bayes Factors]
               Versus:
Support for:      uniform von_mises
      uniform    0.000000  1.231408
      von_mises -1.231408  0.000000
```

For this dataset, the uniform distribution is slightly preferred, although the Bayes factor, which is 3.43, is still somewhat undecided.

Goodness-of-fit

Comparison of the goodness of fit of several models is common in circular statistics (for example, in Pewsey et al. (2013)), and allows us to select the correct probability density for a certain set of data. We apply this to the classic red ant dataset (Jander, 1957; Fisher, 1995), which is available in **circular** as `circular::fisherB7c`.

```
R> ants <- circular::fisherB7c
R>
R> vm_ants <- vm_posterior(ants)
R> pb_ants <- bat_posterior(ants, bat_type = "power")
R> pn_ants <- pn_posterior(ants)
R> bht_compare(von_mises = marg_lik(vm_ants),
R+             powerbat  = marg_lik(pb_ants),
R+             projnorm  = marg_lik(pn_ants))

Bayesian Hypothesis Test
    Comparing 3 models: von_mises, powerbat, projnorm

[Log Marginal Likelihood]
von_mises  powerbat  projnorm
```

```
-144.1331 -135.5867 -223.4015

[Posterior Model Probabilities]
   von_mises     powerbat     projnorm
1.942138e-04 9.998058e-01 7.285745e-39

[Pairwise log Bayes Factors]
               Versus:
Support for:     von_mises   powerbat projnorm
      von_mises   0.000000  -8.546356 79.26835
      powerbat    8.546356   0.000000 87.81470
      projnorm  -79.268348 -87.814705  0.00000
```

Here, we can conclude that due to the peaked shape of the data, the Batschelet distribution fits the best.

Multi-sample (ANOVA-type) test

Often, the question of interest is whether multiple samples have the same or different distributions. We show this on the `pigeons` data (Gagliardo et al., 2008; Fisher, 1995) from **circular**. The question is whether bearings of pigeons differ dependent on a treatment condition. One model under consideration has all three groups with equal distribution. Two others have the three groups with equal concentration, but different means (called homogeneous in variance), while a third has all three groups different in both mean and variance (called heterogeneous in variance). The homogeneous model must be run in a regression context, while the heterogeneous model just consists of separate runs of the von Mises posterior. Instead of choosing a specific alternative, we investigate all three models at the same time and compare.

```
R> bear <- circular(pigeons$bearing, units = "degrees")
R> trt  <- pigeons$treatment
R> pigeons$bearing <- bear
R> bearc           <- bear[trt == "c"]
R> bearv1          <- bear[trt == "v1"]
R> bearon          <- bear[trt == "on"]
R>
R> nodiff_pigeons <- vm_posterior(pigeons$bearing)
R> hom_pigeons    <- vm_reg(bearing ~ treatment,
R+                          data = pigeons)
R> het_pigeonsc   <- vm_posterior(bearc)
```

```
R> het_pigeonsv1  <- vm_posterior(bearv1)
R> het_pigeonson  <- vm_posterior(bearon)
R>
R>
R> bht_compare(nodiff    = marg_lik(nodiff_pigeons),
R+             homogen   = marg_lik(hom_pigeons),
R+             heterogen = sum(marg_lik(het_pigeonsc),
R+                             marg_lik(het_pigeonsv1),
R+                             marg_lik(het_pigeonson)))

Bayesian Hypothesis Test
    Comparing 3 models: nodiff, homogen, heterogen

[Log Marginal Likelihood]
   nodiff   homogen heterogen
-161.5862 -160.7505 -145.8102

[Posterior Model Probabilities]
      nodiff      homogen    heterogen
1.407858e-07 3.247305e-07 9.999995e-01

[Pairwise log Bayes Factors]
               Versus:
Support for:       nodiff   homogen heterogen
      nodiff     0.000000 -0.835756 -15.77603
      homogen    0.835756  0.000000 -14.94027
      heterogen 15.776026 14.940270   0.00000
```

From this test, we can conclude that the heterogeneous model is strongly preferred. Therefore, we can conclude that the treatment influences both the mean direction and the circular variance.

6.7 Discussion

The R package **circbayes** was introduced which performs Bayesian inference for a wide variety of circular data models. The methods introduced include regression models, mixture models, non-parametric models, and hypothesis testing between the models.

Because the package provides a single front-end for Bayesian versions of many of the most popular circular data methods, we hope users are able to easily try out several models and compare them, to find the one that fits the

data best. Particularly, the suite of Bayesian hypothesis tests allows allows a user to very flexibly compare models.

In addition, providing easy access to Bayesian methods for circular data might motivate users to try out Bayesian approaches in situations where they may provide modeling advantages. Such situations may range form employing informative priors, performing goodness-of-fit tests between several competing models, or performing hierarchical regression models.

Nederlandse samenvatting

Circulaire data bevinden zich op een periodieke schaal. Dat wil zeggen, een circulaire observatie is gemeten in richtingen of hoeken, waarbij de uiteinden van de schaal naast elkaar liggen, zoals bijvoorbeeld 359° en 0° dat doen. Zo kun je bijvoorbeeld denken aan het probleem van het bepalen van de gemiddelde verjaardag van een groep vrienden. Omdat het jaar na 31 december direct doorgaat naar 1 januari, is het moeilijk om een intuitief gemiddelde uit te rekenen. Simpelweg het aantal dagen tellen vanaf 1 januari en die aantallen middelen lijkt een redelijke aanpak, maar levert rare resultaten op als bijvoorbeeld twee mensen in januari en december zijn geboren: dan ligt hun gemiddelde verjaardag in de zomer. Voor circulaire data zijn dus aparte statistische methoden nodig.

We komen deze circulaire data tegen in een verscheidenheid aan vakgebieden. Zo kun je denken aan observaties van windrichtingen, het uur van de dag waarop een patienten op de eerste hulp binnen komen of de richting waarop een kikker wegduikt als hij door een vogel of een slang wordt bedreigd (voor wie benieuwd is: van een slang weg, maar naar een vogel toe in de hoop er onderdoor te glippen (Bulbert et al., 2015)). In sociale wetenschappen voegen we daar bijvoorbeeld ook nog de zogeheten *circumplexmodellen* aan toe, waarvan de roos van Leary (Leary, 1957) een van de bekendste voorbeelden is. Circumplexmodellen zijn latente factormodellen, waarbij een onobserveerbare eigenschap geëxtraheerd wordt uit een vragenlijst. Echter, waar normaal gesproken de latente factoren als reële getallen worden weergegeven, worden ze in een circumplex model weergegeven als een circulaire variabele, door in de correlatiematrix een circulaire ordening te vereisen (Browne, 1992). Dit levert voor sommige psychologische constructen een meetinstrument op met minder parameters waarvan de interpretatie makkelijker is.

Deze verschillende voorbeelden leveren obervaties θ op, die vervolgens gemeten zijn in bijvoorbeeld graden ($\theta \in [0° - 360°)$), uren van de dag (op de 24-uursklok, dus $\theta \in [0{:}00 - 24{:}00)$), eenheidsvectoren of radialen.

De tak van de statistiek zich bezig houdt met analyse van circulaire data wordt *circulaire statistiek* genoemd. De circulaire statistiek biedt een nieuwe

definitie van het gemiddelde en de variantie. Het circulaire gemiddelde, de *gemiddelde richting*, van de data $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ wordt berekend door te kijken naar de richting waar de resultante vector $(\sum_{i=1}^{n} \cos\theta_i, \sum_{i=1}^{n} \sin\theta_i)^T$ wijst (voor een voorbeeld, zie Figuur 3 voorin dit boek). De variantie is gerelateerd aan de lengte R van de resultante vector. Voorbij deze basis bestaat het veld van circulaire statistiek uit een grote variëteit van modellen voor verschillende onderzoeksvragen, zoals regressiemodellen, hypothesetoetsen en het schatten van multimodale kansverdelingen. Voor een overzicht, zie Fisher (1995), Mardia & Jupp (2000) of Pewsey et al. (2013).

Het doel van dit proefschrift is om Bayesiaanse circulaire statistiek te ontwikkelen waarmee praktische problemen opgelost kunnen worden. Daarom is contact gezocht met verscheidene onderzoekers die werken met circulaire data. Hierin zijn twee subdoelen te onderscheiden.

Eerst moeten de eigenschappen van basale Bayesiaanse circulaire methoden worden bestudeerd en begrepen. Daarnaast moeten deze methoden beschikbaar zijn in makkelijk te gebruiken software-pakketten, zodat uitgebreide kennis van wiskundige statistiek geen vereiste is om de modellen te kunnen draaien. Om dit te bewerkstelligen zijn voor het merendeel van de methoden in deze dissertatie gebruikersvriendelijke softwarepakketen gemaakt in de statistische programmeertaal R.

Ten tweede is het belangrijk om gespecialiseerde methoden te ontwikkelen voor specifieke problemen. In dit proefschrift worden een aantal problemen van toegepaste onderzoekers in sociale wetenschappen opgelost aan de hand van geavanceerde Bayesiaanse circulaire statistiek.

Deze doelen zullen aangepakt worden in een zestal hoofdstukken. Hierna volgt een kort overzicht van de belangrijkste bevindingen van elk.

In **Hoofdstuk 1** ontwikkelen we regressiemodellen met een circulaire uitkomst. Dit leidt tot een Bayesiaanse versie van het circulaire regressiemodel van Fisher & Lee (1992), in essentie een circulaire vorm van het gegeneraliseerde lineaire model (GLM). Omdat het oorspronkelijke model een andere vorm aanneemt als de arbitraire labeling van categorische predictoren wordt veranderd, passen we dit model aan door categorische predictoren apart te behandelen. Daarnaast wordt Bayesiaanse inferentie mogelijk gemaakt, zowel schatting als hypothesetoetsing, door middel van *Markov Chain Monte Carlo* (MCMC) methoden. Dit nieuwe model wordt toegepast op een voorbeeld uit de cognitieve psychologie.

Hoofdstuk 2 behandelt Bayesiaanse hypothesetoetsen voor circulaire uniformiteit. Bayesiaanse hypothesetoetsen hebben het voordeel dat ze niet alleen een verwerping of niet-verwerping van een nulhypothese kunnen opleveren, maar ook evidentie kunnen vinden voor het aannemen van de nulhypothese. Dit is een gewenste eigenschap als we testen of een circulaire dataset

uniform verdeeld is over de cirkel of niet. We ontwikkelen verschillende hypothesetoetsen, en testen deze in een simulatiestudie.

In het volgende hoofdstuk, **Hoofdstuk 3**, worden modellen ontwikkeld aan de hand van een voorbeeld uit oogbewegingenonderzoek. In dit onderzoek krijgen baby's en volwassenen foto's te zien waar ze kortstondig vrijelijk op mogen rondkijken. We zijn hier geinteresseerd in verchillen in oogbewegingengedrag tussen deze twee groepen, baby's en volwassenen. Een van de aspecten is hoe de richtingen van oogbewegingen, die *saccades* worden genoemd, verdeeld zijn. Hiervoor is een circulaire kansverdeling nodig die de multimodale natuur van deze data (men kijkt voornamelijk precies naar boven, links, rechts, en onder) goed kunnen vatten. Daarnaast zijn de toppen van de componenten sterk gepiekt. Om deze data te kunnen modelleren worden Batschelet-type mixture-modellen ontwikkeld, waarmee de multimodale en gepiekte eigenschappen van de data goed kunnen worden gevat. Omdat deze modellen computationeel erg intensief zijn, wordt een alternatief voorgesteld dat vele malen sneller te berekenen is. Hiermee kunnen de vragen die de onderzoekers hebben direct getoetst worden.

In **Hoofdstuk 4** worden ook mixture-modellen gebruikt, maar deze worden toegepast in een veld waar het niet makkelijk is om het aantal benodigde mixture-componenten van te voren vast te stellen. Het voorbeeld bestaat uit muziekluisterdata, waarbij de vraag is op welke tijd van de dag men luistert naar welk muziekgenre. We modelleren dit met een mixture-model, waarbij elk mixture-component een groep observaties modelleert dat zich rond een bepaald moment op de dag samenschoolt. Je kunt je bijvoorbeeld voorstellen dat een genre als 'Techno' door een bepaalde groep gebruikers tijdens het werken wordt beluisterd, terwijl een andere groep gebruikers dit genre slechts 's nachts luistert. Omdat deze componenten daarmee een aspect van dynamisch gedrag uit de maatschappij verklaren, en dit gedrag per genre verschilt, hebben we een methode nodig die ook kan leren over het benodigde aantal mixture-componenten waaruit de data bestaat. Om dit te doen gebruiken we het *reversible jump* ('omkeerbare sprong') algoritme (Richardson & Green, 1997a). Deze aanpak zorgt ervoor dat ons mixture-model directe inferentie kan uitvoeren op het aantal benodigde componenten. Dit levert een flexibele methode die nieuwe inzichten oplevert voor de muziekluisterdata.

Het voorbeeld in **Hoofdstuk 5** komt uit de criminologie. Sommige strafbare feiten komen veel vaker op bepaalde tijden van de dag voor dan op andere tijden. Fietsendiefstal op het station komt bijvoorbeeld 's nachts veel meer voor dan midden op de dag. Echter, als we de geobserveerde tijden willen modelleren lopen we tegen een probleem aan: er zijn meestal geen geobserveerde tijden, omdat een dief nu juist probeert om niet te worden geobserveerd. De data bestaat dus uit intervallen van tijd, bijvoorbeeld 08:20

- 17:50, waarin het slachtoffer niet bij het gestolen object aanwezig was. Dit komt voor als het slachtoffer om 08:20 de fiets parkeert en om 17:50 terugkomt om te zien dat de fiets is gestolen. Dit soort data heet *aoristische* data. Hiervoor zijn basale descriptieve tools ontwikkeld (Ashby & Bowers, 2013) in de criminologie. We laten zien dat zulke methoden nuttig zijn om de data te plotten, maar de variantie overschatten van de kansverdeling van de diefstaltijden. Daarom ontwikkelen wij methoden voor statistische inferentie van aoristische data. We laten zien hoe een circulaire kansverdeling kan worden geschat aan de hand van aoristische data. Daarnaast ontwikkelen we Bayesiaanse *Dirichlet Process Mixture*-modellen, die ook de multimodale vormen van realistische aoristische data goed kunnen vatten.

Tot slot wordt in **Hoofdstuk 6** het softwarepakket **circbayes** gepresenteerd, waarmee Bayesiaanse circulaire statistiek kan worden uitgevoerd. Hierin vinden we veel basale modellen, alsmede sommige modellen die in eerdere hoofdstukken zijn gepresenteerd. Zo zitten er methoden in voor von Mises, Batschelet en *Projected Normal* verdelingen, alsmede circulaire regressiemodellen, mixture-modellen en Dirichletprocesmodellen. Daarnaast kunnen deze modellen met elkaar worden vergeleken aan de hand van Bayesiaanse hypothesetoetsen, door gebruik te maken van het R pakket **bridgesampling** (Gronau, Singmann, & Wagenmakers, 2017). Hierdoor zijn van de meeste frequentistische circulaire hypothesetoetsen ook Bayesiaanse versies beschikbaar in dit pakket. Met deze gereedschapsdoos hopen we dat toegepaste onderzoekers een toegankelijke ingang hebben in het gebruik van Bayesiaanse circulaire statistiek.

Dankwoord

Een aantal mensen direct of indirect bijgedragen aan het afronden van dit proefschrift. Hier wil ik hen graag bedanken.

Ten eerste wil ik mijn dank uitspreken naar mijn promotoren. Irene, je bent écht een onwijs goede supervisor. Je wist wanneer je me moest bijsturen als ik teveel *off on a tangent* ging, maar je had ook een ongelofelijk vertrouwen in mij en de projecten die ik verzon. Bovenal kwam ik altijd enthousiast uit onze meetings, was er oneindige ruimte om mezelf te ontwikkelen op de vlakken die ik wilde, en was het ook nog gezellig. Bedankt. Herbert, bedankt voor je enthousiasme en je gastvrijheid in de statistiekwereld.

Mijn co-auteurs: Jolien, we zijn op veel vlakken antipodaal symmetrisch, maar ik heb enorm genoten van onze discussies, het zeldzaam hoge niveau waarop we elkaars werk konden begrijpen, en het verzinnen van artiestennamen voor je toekomste muziekcarriére. Pieter, ik heb het gevoel dat door met jou te werken, ik ook leerde hoe het is voor anderen om met mij te werken. Bedankt voor ons ontzettend plezierige proces, je humor en je constante stroom creatieve ideeën. Sven, ik heb met enorm veel plezier met je gewerkt in ons project over circulaire mediatie. Bedankt voor je energie.

Ingmar, Daan en Stijn, jullie hebben me het *waarom* van mijn proefschrift laten zien, door me mee te nemen in wat jullie zouden kunnen willen weten. Ontzettend bedankt daarvoor, en voor de fascinerende discussies over jullie domein en mijn domein.

Mijn paranimfen, Bart en Fayette: Jullie zijn allebei bijzonder goed in meedenken en luisteren, en hebben me hiermee ontzettend geholpen. Bart, bedankt voor al onze koffietjes bij de Village, ik heb er echt heel veel aan gehad. Fayette, wat was het ontzettend fijn om met jouw creativiteit en vrolijkheid de kamer te delen.

During this project I've had the pleasure to be surrounded by an exceedingly inspiring academic world. First, the department of Methods & Statistics at Utrecht University: thanks for all the lovely conversations, the atmosphere and the openness. Second, I've had the opportunity to meet many fantastic

people, whether at IOPS, ISBA, O'Bayes, ADISTA, BAYSM, CMStatistics, or any other conference or event: thanks so much for the engaging, stimulating and just plain fun conversations we've had. Also, thanks a lot to Arthur Pewsey for hosting my research visit to sunny Cáceres, Spain.

To Sanne, Anne, Erik-Jan, Duco, Oisín, Corine, Ayoub, Maria, Hidde, Sjoerd, Jeroen, Emmeke, Sander, Thomas, and anyone who ever joined our lunch conversations in the past: thanks for making the department not only an intellectually stimulating workplace, but also a very pleasant one. Also, I'm sorry for breaking my oath by handing in my dissertation before we scored 40 in lunch Jenga. I would love to make it up to you all one day.

The same goes for all the other colleagues at the Methods & Statistics department, who I've taught with, shared a room with, talked with, had dinner with, or worked with in any way. You are all really smart and cool people, and I'm really grateful for the memories I've made with you all along the way.

Iedereen van mijn masterjaar van M&S, met wie ik ooit dit statistiek-avontuur ingestapt ben: We hebben veel van elkaar geleerd, en ik ben blij dat ik jullie de afgelopen jaren nog veel gezien heb in het wereldje.

Esther, bedankt voor de rol die je bent gaan spelen in mijn leven, waardoor je pas bij het illustreren van de kaft er achter kwam waar mijn proefschrift eigenlijk over ging. Maar ook bedankt voor de mooie illustratie zelf, natuurlijk!

Mijn vrienden: Bruno, Jan, Bas, Michelle, Joeri, Ewout, Lisa, Joost, Romy, FW, Anne, Steffie, Casper, oud-huisgenoten, de donderdag-club, mensen van de Sunday Assembly, IOPS-borrelaars, YLL-maten, Young Statisticians, spelletjesmensen, feestjesmensen, festivalmensen, en alle anderen, bedankt voor de ontspanning, de steun, de drankjes, de gesprekken, de dansjes en de gezelligheid.

Loes, bedankt voor je aandacht naar mijn persoonlijke proces in het maken van dit proefschrift, en je vrolijke desinteresse in de inhoud ervan, waardoor ik op het eiland mijn hoofd echt hélemaal leeg kon maken. Johan, bedankt voor de fijne wandelingen door de de bossen van Radio Kootwijk. Maartje, bedankt voor de steun.

Gief, Ella, Dowie, en de rest van de schoonfamilie/kerstfamilie: bedankt voor de buitengewoon warme ontvangst en de fijne kerstweken, die elk jaar een mooi moment van reflectie en nieuwe start waren.

Tot slot: Lisa, van gekke dansjes tot het kopen van ons huis, je hebt heel de PhD lang kleur gegeven aan mijn leven. Ik kan me geen beter supportteam voorstellen. Bedankt voor de muziek, de warmte, het luisteren.

Curriculum Vitae

Kees Mulder was born on the 3^{rd} of May, 1990, in Oegstgeest, the Netherlands. In 2011, he finished his BSc. in Psychology with a minor in Methodology & Statistics. In 2012, he enrolled in the Master's programme *Methods and Statistics of Social and Behavioural Sciences* at Utrecht University. During this time he worked at CITO and CentERdata as a (research) assistant, took part in an interdisciplinary Honours programme (the *Young Leaders League*), took place on an education advisory committee, and was president of an interdisciplinary student society called the *Louis Bonaparte Society*. In 2014, he finished the Master's *cum laude*, with his Master's thesis *Extending Bayesian Analysis of Circular Data to Comparison of Multiple Groups* being selected as one of three final nominees for Utrecht University's *Best Master's Thesis award*.

In 2014, Kees started a PhD on Bayesian Circular Statistics at Utrecht University. During his PhD, he also taught classes on programming in R, Data Science, and Structural Equation Modeling.

In addition to research and teaching, Kees was active in Utrecht University's representational body for PhD's, *Prout*, where he served as chair from 2016 - 2017. In addition, he took place on the board of the *Young Statisticians*, the junior subdivision of the Dutch Society for Statistics, the *VVS-OR*.

As of April 2019, he works as a data science consultant at EY VODW.

Publications

Mulder, K., & Klugkist, I. (2017). Bayesian estimation and hypothesis tests for a circular Generalized Linear Model. *Journal of mathematical psychology*, 80, 4-14.

Zerr, P., Gayet, S., **Mulder, K.**, Pinto, Y., Sligte, I., & Van der Stigchel, S. (2017). Remapping high-capacity, pre-attentive, fragile sensory memory. *Scientific reports*, 7(1), 15940.

Klugkist, I., Cremers, J. & **Mulder, K.** (2018). Bayesian analysis of circular data in social and behavioral sciences. In Ley, C. & Verdebout, T. *Applied Directional Statistics: Modern Methods and Case Studies* (pp. 156 - 199). Chapman & Hall/CRC Press.

Cremers, J., **Mulder, K.**, & Klugkist, I. (2018). Circular interpretation of regression coefficients. *British journal of mathematical and statistical psychology*, 71(1), 75-95.

Veen, D., Stoel, D., Schalken, N., **Mulder, K.**, & van de Schoot, R. (2018). Using the Data Agreement Criterion to Rank Experts' Beliefs. *Entropy*, 20(8), 592.

Bibliography

Abe, T., & Ley, C. (2017). A tractable, parsimonious and flexible model for cylindrical data, with applications. *Econometrics and Statistics*, *4*, 91–104.

Abe, T., Shimizu, K., & Pewsey, A. (2010). Symmetric unimodal models for directional data motivated by inverse stereographic projection. *Journal of the Japan Statistical Society*, *40*(1), 45–61.

Agostinelli, C. (2007). Robust estimation for circular data. *Computational Statistics & Data Analysis*, *51*(12), 5867–5875.

Agostinelli, C., & Lund, U. (2013). R package `circular`: Circular statistics (version 0.4-7) [Computer software manual]. CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA. Retrieved from https://r-forge.r-project.org/projects/circular/

Ajne, B. (1968). A simple test for uniformity of a circular distribution. *Biometrika*, *55*(2), 343–354.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Akaike, H. (1987). Factor analysis and aic. In *Selected papers of hirotugu akaike* (pp. 371–386). Springer.

Ameijeiras-Alonso, J., Crujeiras, R. M., & Rodríguez Casal, A. (2018). Directional statistics for wildfires. *Applied Directional Statistics: Modern Methods and Case Studies*, 119.

A Mooney, J., Helms, P. J., & Jolliffe, I. T. (2003, January). Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, *41*(3-4), 505–513.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 1152–1174.

Ardia, D., Baştürk, N., Hoogerheide, L., & Van Dijk, H. K. (2012). A comparative study of monte carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics & Data Analysis*, *56*(11), 3398–3414.

Arnold, B. C., & SenGupta, A. (2006). Recent advances in the analyses of directional data in ecological and environmental sciences. *Environmental and Ecological Statistics*, *13*(3), 253–256.

Artes, R. (2008). Hypothesis tests for covariance analysis models for circular data. *Communications in Statistics - Theory and Methods*, *37*(10), 1632–1640.

Ashby, M. P. J., & Bowers, K. J. (2013). A comparison of methods for temporal analysis of aoristic crime. *Crime Science*, *2*(1), 1.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Baayen, C., & Klugkist, I. (2014). Evaluating order-constrained hypotheses for circular data from a between-within subjects design. *Psychological Methods*, *19*(3), 398.

Baayen, C., Klugkist, I., & Mechsner, F. (2012). A test of order-constrained hypotheses for circular data with applications to human movement science. *Journal of Motor Behavior*, *44*(5), 351–363.

Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005). Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of machine Learning research*, *6*(Sep), 1345–1382.

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P., & Raftery, A. E. (2010). Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, *138*(5), 1811–1821.

Barragán, S., Fernández, M. A., Rueda, C., & Peddada, S. D. (2013). isocir: An R package for constrained inference using isotonic regression for circular data, with an application to cell biology. *Journal of Statistical Software*, *54*(4).

Batschelet, E. (1981). Circular statistics in biology. *Academic Press, London*, *111*, 388.

Berens, P. (2009). Circstat: a matlab toolbox for circular statistics. *Journal of Statistical Software*, *31*(10), 1–21.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402.

Besag, J., Green, P., Higdon, D., & Mengersen, K. (1995, February). Bayesian computation and stochastic systems. *Statistical science*, *10*(1), 3–41.

Best, D. J., & Fisher, N. I. (1979). Efficient simulation of the von Mises distribution. *Applied Statistics*, *28*, 152–157.

Best, D. J., & Fisher, N. I. (1981). The bias of the maximum likelihood estimators of the von Mises-fisher concentration parameters. *Communications in Statistics - Simulation and Computation*, *10*(5), 493–502.

Bhattacharya, S., & Sengupta, A. (2009). Bayesian analysis of semiparametric linear-circular models. *Journal of Agricultural, Biological, and Environmental Statistics*, *14*(1), 33–65.

Bhattacharya, S., & SenGupta, A. (2009). Bayesian inference for circular distributions with unknown normalising constants. *Journal of Statistical Planning and Inference*, *139*(12), 4179–4192.

Bhattacharyya, G. K., & Johnson, R. A. (1969). On hodges's bivariate sign test and a test for uniformity of a circular distribution. *Biometrika*, *56*(2), 446–449.

Bogdan, M., Bogdan, K., & Futschik, A. (2002). A data driven smooth test for circular uniformity. *Annals of the Institute of Statistical Mathematics*, *54*(1), 29–44.

Bowers, J., Morton, I., & Mould, G. (2000). Directional statistics of the wind and waves. *Applied Ocean Research*, *22*(1), 13–30.

Brazier, K. T. S. (1994). Confidence intervals from the rayleigh test. *Monthly Notices of the Royal Astronomical Society*, *268*(3), 709–712.

Brooks, S. P., Giudici, P., & Roberts, G. O. (2003). Efficient construction of reversible-jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 3–55.

Brown, D. E. (1998). The regional crime analysis program (recap): a framework for mining data to catch criminals. In *Systems, man, and cybernetics, 1998. 1998 ieee international conference on* (Vol. 3, pp. 2848–2853).

Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, *57*(4), 469–497.

Bulbert, M. W., Page, R. A., & Bernal, X. E. (2015). Danger comes from all fronts: predator-dependent escape tactics of túngara frogs. *PloS one*, *10*(4), e0120546.

Canty, A., & Ripley, B. D. (2017). boot: Bootstrap r (s-plus) functions [Computer software manual]. (R package version 1.3-20)

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chen, D.-G., Sun, J., & Peace, K. E. (2012). *Interval-censored time-to-event data: Methods and applications*. Chapman and Hall/CRC.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, *49*(4), 327–335.

Chib, S., & Greenberg, E. (2010). Additive cubic spline regression with Dirichlet process mixture errors. *Journal of Econometrics*, *156*(2), 322–336.

Coles, S. (1998). Inference for circular distributions and processes. *Statistics and Computing*, *8*(2), 105–113.

Consonni, G., Veronese, P., et al. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, *23*(3), 332–353.

Cox, N. (1998). Circstat: Stata modules to calculate circular statistics.

Cremers, J., Mulder, K., & Klugkist, I. (2018). Circular interpretation of regression coefficients. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 75-95. doi: 10.1111/bmsp.12108

Cremers, J., Pennings, H. J., Mainhard, M. T., & Klugkist, I. (2018). Longitudinal circular modelling of circumplex measurements for interpersonal behavior. *Manuscript submitted for publication*.

Damien, P., & Walker, S. (1999). A full Bayesian analysis of circular data using the von Mises distribution. *Canadian Journal of Statistics*, *27*(2), 291–298.

Darnieder, W. F. (2011). *Bayesian methods for data-dependent priors* (Unpublished doctoral dissertation). The Ohio State University.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press. Retrieved from http://statwww.epfl.ch/davison/BMA/ (ISBN 0-521-57391-2)

Dickey, J. M., Lientz, B. P., et al. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: method and application. In *Markov chain Monte Carlo in practice* (pp. 259–273). Springer.

Di Marzio, M., Panzera, A., & Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters*, *79*(19), 2066–2075.

Di Marzio, M., Panzera, A., & Taylor, C. C. (2013). Non-parametric regression for circular responses. *Scandinavian Journal of Statistics*, *40*(2), 238–255.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics*, 1763–1786.

Duan, J. A., Guindani, M., & Gelfand, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, *94*(4), 809–825.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. Retrieved from http://www.jstatsoft.org/v40/i08/

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Engbert, R., Trukenbrod, H. A., Barthelmé, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, *15*(1), 14–14.

Everitt, B. S. (2004). Mixture distributions—i. *Encyclopedia of Statistical Sciences*, *7*.

Felson, M., & Poulsen, E. (2003). Simple indicators of crime by time of day. *International Journal of Forecasting*, *19*(4), 595–601.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 209–230.

Fernández-Durán, J. J., & Mercedes Gregorio-Domínguez, M. (2016a). Bayesian analysis of circular distributions based on non-negative trigonometric sums. *Journal of Statistical Computation and Simulation*, 1–13.

Fernández-Durán, J. J., & Mercedes Gregorio-Domínguez, M. (2016b, February). Bayesian analysis of circular distributions based on non-negative trigonometric sums. *Journal of Statistical Computation and Simulation*, *86*(16), 3175–3187.

Ferrari, C. (2009). *The wrapping approach for circular data Bayesian modeling* (Unpublished doctoral dissertation).

Ferreira, J. T. A. S., Juárez, M. A., & Steel, M. F. J. (2008, June). Directional log-spline distributions. *Bayesian Analysis*, *3*(2), 297–316.

Ferreira, J. T. A. S., Juárez, M. A., Steel, M. F. J., et al. (2008). Directional log-spline distributions. *Bayesian Analysis*, *3*(2), 297–316.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

Fisher, N. I., & Lee, A. J. (1992). Regression models for an angular response. *Biometrics*, 665–677.

Forbes, P. G. M., & Mardia, K. V. (2014, June). A fast algorithm for sampling from the posterior of a von Mises distribution. *Journal of Statistical Computation and Simulation*, *85*(13), 2693–2701.

Forbes, P. G. M., & Mardia, K. V. (2015). A fast algorithm for sampling from the posterior of a von Mises distribution. *Journal of Statistical Computation and Simulation*, *85*(13), 2693–2701.

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, *78*, 14–25.

Foulsham, T., Kingstone, A., & Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, *48*(17), 1777–1790.

Friel, N., & Wyse, J. (2012). Estimating the evidence–a review. *Statistica Neerlandica*, *66*(3), 288–308.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gagliardo, A., Ioalè, P., Savini, M., & Wild, M. (2008). Navigational abilities of homing pigeons deprived of olfactory or trigeminally mediated magnetic information when young. *Journal of Experimental Biology*, *211*(13), 2046–2051.

García-Portugués, E., Crujeiras, R. M., & González-Manteiga, W. (2013). Exploring wind direction and so2 concentration by circular–linear density estimation. *Stochastic Environmental Research and Risk Assessment*, *27*(5), 1055–1067.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. CRC press.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

George, B. J., & Ghosh, K. (2006). A semiparametric Bayesian model for circular-linear regression. *Communications in Statistics: Simulation and Computation*, *35*(4), 911–923.

Geyer, C. J., & Thompson, E. A. (1992). Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 657–699.

Ghosh, K., Jammalamadaka, R., & Tiwari, R. (2003a). Semiparametric Bayesian techniques for problems in circular data. *Journal of Applied Statistics*, *30*(2), 145–161.

Ghosh, K., Jammalamadaka, R., & Tiwari, R. (2003b, February). Semiparametric Bayesian Techniques for Problems in Circular Data. *Journal of Applied Statistics*, *30*(2), 145–161.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 337–348.

Gill, A., Partridge, H., & Newton, A. D. (2014). Interstitial crime analysis. *JDi Brief*.

Gill, J., & Hangartner, D. (2010). Circular data in political science and how to handle it. *Political Analysis*, *18*(3), 316–336.

Gopal, S., & Yang, Y. (2014). Von mises-fisher clustering models. In *International conference on machine learning* (pp. 154–162).

Gottlieb, S., Arenberg, S. I., Singh, R., et al. (1994). *Crime analysis: From first report to final arrest*. Alpha Publishing Montclair, CA.

Grassi, M., Luccio, R., & Di Blas, L. (2010). Circe: An r implementation of browne's circular stochastic process model. *Behavior Research Methods*, *42*(1), 55–73.

Green, P. J. (1995, December). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ... Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: an R package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Grubesic, T. H., & Mack, E. A. (2008). Spatio-temporal interaction of urban crime. *Journal of Quantitative Criminology*, *24*(3), 285–306.

Gurtman, M. B. (2009a). Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass*, *3*(4), 601–619.

Gurtman, M. B. (2009b, July). Exploring Personality with the Interpersonal Circumplex. *Social and Personality Psychology Compass*, *3*(4), 601–619.

Gurtman, M. B. (2015). Circumplex models. *The Encyclopedia of Clinical Psychology*.

Gurtman, M. B., & Pincus, A. L. (2003). The circumplex model: Methods and research applications. *Handbook of Psychology*.

Guttorp, P., & Lockhart, R. A. (1988). Finding the location of a signal: A Bayesian analysis. *Journal of the American Statistical Association*, *83*(402), 322–330.

Hall, P., Watson, G. S., & Cabrera, J. (1987). Kernel density estimation with spherical data. *Biometrika*, *74*(4), 751–762.

Hastings, W. K. (1970a). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hastings, W. K. (1970b, April). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498–504.

Hermans, M., & Rasson, J. P. (1985). A new sobolev test for uniformity on the circle. *Biometrika*, *72*(3), 698–702.

Hernandez-Stumpfhauser, D., Breidt, F. J., & Opsomer, J. D. (2016). Hierarchical Bayesian small area estimation for circular data. *Canadian Journal of Statistics*, *44*(4), 416–430.

Hernandez-Stumpfhauser, D., Breidt, F. J., van der Woerd, M. J., et al. (2017a). The general projected normal distribution of arbitrary dimension: Modeling and Bayesian inference. *Bayesian Analysis*, *12*(1), 113–133.

Hernandez-Stumpfhauser, D., Breidt, F. J., van der Woerd, M. J., et al. (2017b). The general projected normal distribution of arbitrary dimension: Modeling and Bayesian inference. *Bayesian Analysis*, *12*(1), 113–133.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics* (Vol. 28). Cambridge University Press.

Hodges, J. L. (1955). A bivariate sign test. *The Annals of Mathematical Statistics*, *26*(3), 523–527.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses*. Springer Science & Business Media.

Holzmann, H., Munk, A., Suster, M., & Zucchini, W. (2006). Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, *13*(3), 325–347.

Hornik, K., & Grün, B. (2013). On conjugate families and Jeffreys priors for von Mises–fisher distributions. *Journal of Statistical Planning and Inference*, *143*(5), 992–999.

Hornik, K., & Grün, B. (2014). movmf: An R package for fitting mixtures of von Mises-fisher distributions. *Journal of Statistical Software*, *58*(10), 1–31.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.

J. Ross, G., & Markwick, D. (2018). dirichletprocess: Build Dirichlet process objects for Bayesian modelling [Computer software manual]. Retrieved from https://github.com/dm13450/dirichletprocess (R package version 0.2.1.900)

Jammalamadaka, S. R., & Sengupta, A. (2001). *Topics in circular statistics* (Vol. 5). World Scientific.

Jander, R. (1957). Die optische richtungsorientierung der roten waldameise (formica ruea l.). *Zeitschrift für vergleichende Physiologie*, *40*(2), 162–238.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005a). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 50–67.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005b, February). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical science*, *20*(1), 50–67.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Clarendon Press.

Jones, M. C., & Pewsey, A. (2005). A family of symmetric distributions on the circle. *Journal of the American Statistical Association*, *100*(472), 1422–1428.

Jones, M. C., & Pewsey, A. (2012). Inverse Batschelet distributions for circular data. *Biometrics*, *68*(1), 183–193.

Kaas, A. L., & Van Mier, H. I. (2006). Haptic spatial matching in near peripersonal space. *Experimental Brain Research*, *170*(3), 403–413.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kendall, D. G. (1974). Pole-seeking brownian motion and bird navigation. *Journal of the Royal Statistical Society. Series B*, *37*, 97–133.

Kikuchi, G. (2015). Package 'aoristic'.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (2013). *Handbook of survival analysis*. Chapman and Hall/CRC.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods*, *10*(4), 477.

Kuiper, N. H. (1960). Tests concerning random points on a circle. In *Indagationes mathematicae (proceedings)* (Vol. 63, pp. 38–47).

Kurz, G., Gilitschenski, I., Pfaff, F., Drude, L., Hanebeck, U. D., Haeb-Umbach, R., & Siegwart, R. Y. (2017). Directional statistics and filtering using libdirectional. *arXiv preprint arXiv:1712.09718*.

Kurz, G., & Hanebeck, U. D. (2017). Deterministic sampling on the torus for bivariate circular estimation. *IEEE Transactions on Aerospace and Electronic Systems*, *53*(1), 530–534.

Lagona, F. (2016). Regression analysis of correlated circular data based on the multivariate von Mises distribution. *Environmental and Ecological Statistics*, *23*(1), 89–113.

Lagona, F., Picone, M., Maruotti, A., & Cosoli, S. (2015). A hidden Markov approach to the analysis of space–time environmental data with linear and circular components. *Stochastic Environmental Research and Risk Assessment*, *29*(2), 397–409.

Landler, L., Ruxton, G. D., & Malkemper, E. P. (2018). Circular data in biology: advice for effectively implementing statistical procedures. *Behavioral Ecology and Sociobiology*, *72*(8), 128.

Laubscher, N. F., & Rudolph, G. J. (1968). *A distribution arising from random points on the circumfence of a circle*. National Research Institute for Mathematical Sciences (South Africa).

Leary, T. (1957). *Interpersonal diagnosis of personality*. New York: Ronald Press.

Le Meur, O., & Coutrot, A. (2016). Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision Research*, *121*, 72–84.

Leong, P., & Carlile, S. (1998). Methods for spherical data analysis and visualization. *Journal of Neuroscience Methods*, *80*(2), 191–200.

Ley, C., & Verdebout, T. (2017). *Modern directional statistics*. Chapman and Hall/CRC.

Ley, C., & Verdebout, T. (2018). *Applied directional statistics: Modern methods and case studies*. CRC Press.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*.

Lund, U., Agostinelli, C., & Agostinelli, M. C. (2013). Package 'circular'. *Repository CRAN*.

Mardia, K. V. (2011). How new shape analysis and directional statistics are advancing modern life-sciences. In *Int. statistical inst.: Proc. 58th world statistical congress.*

Mardia, K. V., & El-Atoum, S. (1976). Bayesian inference for the von Mises-Fisher distribution. *Biometrika*, *63*(1), 203–206.

Mardia, K. V., Hughes, G., Taylor, C. C., & Singh, H. (2008). A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, *36*(1), 99–109.

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

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics* (Vol. 494). John Wiley & Sons.

Mardia, K. V., & Jupp, P. E. (2009). *Directional Statistics*. Wiley.

Mardia, K. V., & Voss, J. (2014). Some fundamental properties of a multivariate von Mises distribution. *Communications in Statistics-Theory and Methods*, *43*(6), 1132–1144.

Maruotti, A. (2016). Analyzing longitudinal circular data by Projected Normal models: a semi-parametric approach based on finite mixture models. *Environmental and Ecological Statistics*, 1–21.

Masseran, N., Razali, A. M., Ibrahim, K., & Latif, M. T. (2013). Fitting a mixture of von Mises distributions in order to model data on wind direction in Peninsular Malaysia. *Energy Conversion and Management*, *72*, 94–102.

McClintock, B. T., & Michelot, T. (2018). momentuhmm: R package for generalized hidden Markov models of animal movement. *Methods in Ecology and Evolution*, *9*(6), 1518–1530.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McVinish, R., & Mengersen, K. (2008). Semiparametric Bayesian circular statistics. *Computational Statistics & Data Analysis*, *52*(10), 4722–4730.

Mechsner, F., Kerzel, D., Knoblich, G., & Prinz, W. (2001a). Perceptual basis of bimanual coordination. *Nature*, *414*(6859), 69–73.

Mechsner, F., Kerzel, D., Knoblich, G., & Prinz, W. (2001b, November). Perceptual basis of bimanual coordination. *Nature*, *414*(6859), 69–73.

Mechsner, F., Stenneken, P., Cole, J., Aschersleben, G., & Prinz, W. (2007). Bimanual circling in deafferented patients: Evidence for a role of visual forward models. *Journal of neuropsychology*, *1*(2), 259–282.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831–860.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 68–75.

Mulder, K., & Klugkist, I. (2017). Bayesian estimation and hypothesis tests for a circular generalized linear model. *Journal of Mathematical Psychology*, *80*, 4–14. doi: https://doi.org/10.1016/j.jmp.2017.07.001

Murdoch, D., et al. (2003). orientlib: An R package for orientation data. *Journal of Statistical Software*, *8*(19), 1–11.

Nadarajah, S., & Zhang, Y. (2017). Wrapped: An R package for circular data. *PloS one*, *12*(12), e0188512.

Najemnik, J., & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. *Journal of Vision*, *8*(3), 4–4.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Newton, A. D., Partridge, H., & Gill, A. (2014). Above and below: measuring crime risk in and around underground mass transit systems. *Crime Science*, *3*(1), 1.

Nielsen, S. F., et al. (2000). The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, *6*(3), 457–489.

Nuñez-Antonio, G., Ausín, M. C., & Wiper, M. P. (2015). Bayesian nonparametric models of circular variables based on Dirichlet process mixtures of normal distributions. *Journal of Agricultural, Biological, and Environmental Statistics*, *20*(1), 47–64.

Nunez-Antonio, G., & Gutierrez-Pena, E. (2005). A Bayesian analysis of directional data using the Projected Normal distribution. *Journal of Applied Statistics*, *32*(10), 995–1001.

Nuñez-Antonio, G., & Gutiérrez-Peña, E. (2014). A Bayesian model for longitudinal circular data based on the Projected Normal distribution. *Computational Statistics & Data Analysis*, *71*, 506–519.

Nuñez-Antonio, G., Gutiérrez-Peña, E., & Escarela, G. (2011). A Bayesian regression model for circular data based on the Projected Normal distribution. *Statistical Modelling*, *11*(3), 185–201.

Nuñez-Antonio, G., Mendoza, M., Contreras-Cristán, A., Gutiérrez-Peña, E., & Mendoza, E. (2018). Bayesian nonparametric inference for the overlap of daily animal activity patterns. *Environmental and Ecological Statistics*, 1–24.

O'Hagan, A., & Forster, J. J. (2004). *Kendall's advanced theory of statistics, volume 2b: Bayesian inference* (Vol. 2). Arnold.

Oliveira, M., Crujeiras, R., & Rodríguez-Casal, A. (2014). Npcirc: An R package for nonparametric circular methods. *Journal of Statistical Software*, *61*(1), 1–26. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v061i09 doi: 10.18637/jss.v061.i09

Oliveira, M., Crujeiras, R. M., & Rodríguez-Casal, A. (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics & Data Analysis*, *56*(12), 3898–3908.

Paluszewski, M., & Hamelryck, T. (2010). Mocapy++-a toolkit for inference and learning in dynamic Bayesian networks. *BMC bioinformatics*, *11*(1), 126.

Pereira, D. V., Andresen, M. A., & Mota, C. M. (2016). A temporal and spatial analysis of homicides. *Journal of Environmental Psychology*, *46*, 116–124.

Pewsey, A., Neuhäuser, M., & Ruxton, G. D. (2013). *Circular Statistics in R*. OUP Oxford.

Pewsey, A., Shimizu, K., & de la Cruz, R. (2011). On an extension of the von Mises distribution due to Batschelet. *Journal of Applied Statistics*, *38*(5), 1073–1085.

Postma, A., Zuidhoek, S., Noordzij, M. L., & Kappers, A. M. L. (2008). Keep an eye on your hands: on the role of visual mechanisms in processing of haptic space. *Cognitive Processing*, *9*(1), 63–68.

Presnell, B., Morrison, S. P., & Littell, R. C. (1998). Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, *93*(443), 1068–1077.

Puglisi, G., Leonetti, A., Landau, A., Fornia, L., Cerri, G., & Borroni, P. (2017). The role of attention in human motor resonance. *PloS one*, *12*(5), e0177457. doi: 10.1371/journal.pone.0177457

Pycke, J.-R. (2010). Some tests for uniformity of circular distributions powerful against multimodal alternatives. *Canadian Journal of Statistics*, *38*(1), 80–96.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rao, J. S. (1976). Some tests based on arc-lengths for the circle. *Sankhyā: The Indian Journal of Statistics, Series B*, 329–338.

Ratcliffe, J. H. (2000). Aoristic analysis: the spatial interpretation of unspecific temporal events. *International Journal of Geographical Information Science*, *14*(7), 669–679.

Ratcliffe, J. H., & McCullagh, M. J. (1998). Aoristic crime analysis. *International Journal of Geographical Information Science*, *12*(7), 751–764.

Ravindran, P., & Ghosh, S. K. (2011). Bayesian analysis of circular data using wrapped distributions. *Journal of Statistical Theory and Practice*, *5*(4), 547–561.

Rayment, M. R. (1995). Spatial and temporal crime analysis techniques. *Vancouver Police Department Technical Memorandum*.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Ren, L., Dunson, D. B., & Carin, L. (2008). The dynamic hierarchical Dirichlet process. In *Proceedings of the 25th international conference on machine learning* (pp. 824–831).

Richardson, S., & Green, P. (1997b). On Bayesian Analysis of Mixtures with Unknown Number of Components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(4), 731–792.

Richardson, S., & Green, P. J. (1997a). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), 731–792.

Rueda, C., Fernández, M. A., & Peddada, S. D. (2009). Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell cycle genes. *Journal of the American Statistical Association*, *104*(485), 338–347.

Schmidt-Koenig, K. (1963). On the role of the loft, the distance and site of release in pigeon homing (the" cross-loft experiment"). *The Biological Bulletin*, *125*(1), 154–164.

Schwarz, G. (1978, March). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stan Development Team. (2017). *Stan modeling language users guide and reference manual.* Retrieved from http://mc-stan.org/ (Version 2.17.0)

Stanfill, B., Hofmann, H., & Genschel, U. (2014). rotations: An R package for so (3) data. *The R Journal*, *6*(1), 68–78.

Stephens, M. (2000, January). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809.

Stephens, M. A. (1969). *Techniques for directional data* (Tech. Rep.). STANFORD UNIV CA DEPT OF STATISTICS.

Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K., et al. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, *25*(4), 1371–1470.

Stoneking, C., Wabersich, D., & Oberauer, K. (2016). *jags-vonmises.* https://github.com/yeagle/jags-vonmises. GitHub.

Sun, J. (2007). *The statistical analysis of interval-censored failure time data*. Springer Science & Business Media.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528–540.

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. S. (2017). Latest: A model of saccadic decisions in space and time. *Psychological Review*, *124*(3), 267.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6-7), 1029–1054.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in neural information processing systems* (pp. 1385–1392).

Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 1265–1269.

Tierney, L. (1994, December). Markov chains for exploring posterior distributions. *The Annals of Statistics*, *22*(4), 1701–1728.

Traa, J., & Smaragdis, P. (2013). A wrapped kalman filter for azimuthal speaker tracking. *IEEE Signal Processing Letters*, *20*(12), 1257–1260.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, *69*(345), 169–173.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–295.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

van Dijk, R., Kappers, A. M. L., & Postma, A. (2013). Superior spatial touch: improved haptic orientation processing in deaf individuals. *Experimental Brain Research*, *230*(3), 283–289.

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

Van Renswoude, D. R., Johnson, S. P., Raijmakers, M. E. J., & Visser, I. (2016). Do infants have the horizontal bias? *Infant Behavior and Development*, *44*, 38–48.

Venter, J. H. (1967). On estimation of the mode. *The Annals of Mathematical Statistics*, *38*(5), 1446–1455.

Von Mises, R. (1918). Uber die ganzzahligkeit der atomgewichte und verwandte fragen. *Phys. Z*, *19*, 490–500.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., & Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychology*, *50*(2), 99 - 100. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022249606000022 (Special Issue on Model Selection: Theoretical Developments and Applications) doi: https://doi.org/10.1016/j.jmp.2005.01.005

Wang, F., & Gelfand, A. E. (2013). Directional data analysis under the general Projected Normal distribution. *Statistical Methodology*, *10*(1), 113–127.

Wang, F., & Gelfand, A. E. (2014). Modeling space and space-time directional data using projected gaussian processes. *Journal of the American Statistical Association*, *109*(508), 1565–1580.

Warren, W. H., Rothman, D. B., Schnapp, B. H., & Ericson, J. D. (2017). Wormholes in virtual space: From cognitive maps to cognitive graphs. *Cognition*, *166*, 152–163. doi: 10.1016/j.cognition.2017.05.020

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(Dec), 3571–3594.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, *48*(1/2), 109–114.

Watson, G. S., & Williams, E. J. (1956). On the construction of significance tests on the circle and the sphere. *Biometrika*, *43*(3/4), 344–352.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the savage–dickey density ratio. *Computational Statistics & Data Analysis*, *54*(9), 2094–2102.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association*, *81*(394), 446–451.