NONPARAMETRIC BAYESIAN DENSITY ESTIMATION ON A CIRCLE

MARK FISHER

Incomplete.

Abstract. This note describes nonparametric Bayesian density estimation for circular data using a Dirichlet Process Mixture model with the von Mises distribution as the kernel. A natural prior for the parameters produces a prior predictive distribution for the mean vector that is uniformly distributed on the unit disk.

Date: Original idea conceived December 2013. August 25, 2017 @ 19:06. *Filename*: "von Mises DPM".

1. Introduction

For an introduction to circular statistics see Pewsey et al. (2013). The authors introduce the notion of kernel density estimation (from a frequentist perspective) using the von Mises distribution as the kernel along with a variety of bandwidth selection criteria.

This note describes nonparametric Bayesian density estimation for circular data using a Dirichlet Process Mixture (DPM) model with the von Mises distribution as the kernel. A natural prior for the parameters produces a prior predictive distribution for the mean vector that is uniformly distributed on the unit disk. Other priors are entertained as well.

See Ghosh et al. (2003) for an early foray into this line of research. For more recent work, see Nuñez-Antonio et al. (2014) and the references therein. See also Damien and Walker (1999) who provide a Gibbs sampler (via the introduction of two auxiliary variables) for the parametric case with a conjugate prior. See also Brunner and Lo (1994).

Section 2 introduces the von Mises distribution on the unit circle. Section 3 describes parametric Bayesian inference using the von Mises distribution. This section covers material that is used in the section on the DPM. Before proceeding to the DPM, Section 4 provides a brief introduction to the Bayesian bootstrap. Section 5 presents the DPM model and provides a numerical example.

2. von Mises distribution

Circular (or directional) data are restricted to the interval $\theta \in [0, 2\,\pi)$. A simple parametric probability distribution for circular data is given by the von Mises distribution which has the following density:[1]

$$f(\theta|\phi) = f(\theta|\mu,\kappa) = \mathsf{vonMises}(\theta|\mu,\kappa) = 1_{[0,2\,\pi)}(\theta)\,\frac{e^{\kappa\,\cos(\theta-\mu)}}{2\,\pi\,I_0(\kappa)}, \tag{2.1}$$

where

$$\phi = (\mu,\kappa) \tag{2.2}$$

and $I_\nu(\,\cdot\,)$ is the modified Bessel function of the first kind (of order ν). Note that

$$\lim_{\theta\to 2\,\pi}\; f(\theta|\phi) = f(0|\phi). \tag{2.3}$$

Figure 1 shows plots of the von Mises distribution with various settings of the parameters μ and κ. Note that $\mu \in [0, 2\,\pi)$ is a measure of the location of the distribution and $\kappa \in [0,\infty)$ is measure of the precision.[2]

The density can be expressed in terms of rectangular coordinates as

$$\mathsf{vonMises}(x|m,\kappa) = \frac{e^{\kappa\, m^\top x}}{2\,\pi\,I_0(\kappa)}, \tag{2.4}$$

where $x = (\cos(\theta), \sin(\theta))$ and $m = (\cos(\mu), \sin(\mu))$, since $m^\top x = \cos(\theta-\mu)$.

---

[1] The indicator function included in (2.1) will be omitted henceforth.

[2] The precision parameter κ is often referred to as the "concentration parameter." In this paper we reserve that appellation for a different parameter (introduced below).

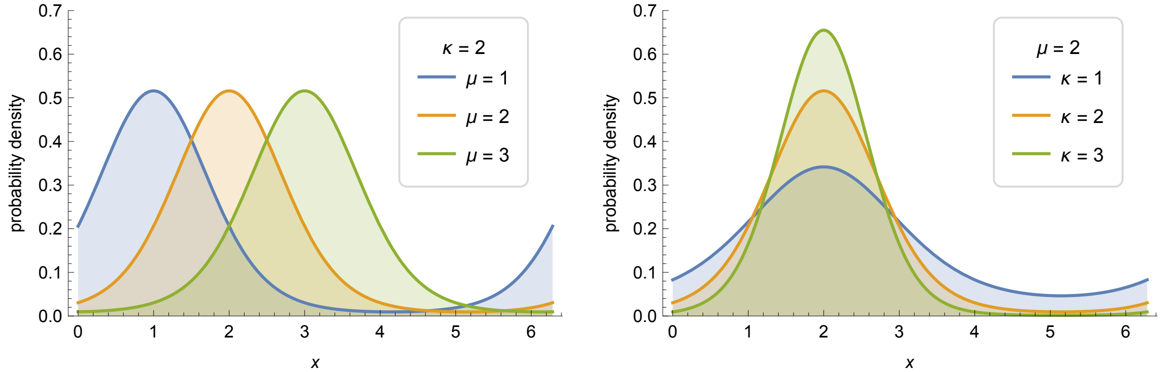

FIGURE 1. von Mises distribution shown with various settings of μ and κ.

A distribution on the unit circle implicitly defines a mean vector ξ characterized by direction and length. In particular, let

$$z = \int_0^{2\pi} e^{\boldsymbol{i}\,\theta}\, f(\theta|\phi)\, dx = e^{\boldsymbol{i}\,\mu}\, L(\kappa), \tag{2.5}$$

where $\boldsymbol{i}$ denotes the imaginary unit and

$$L(\kappa) := I_1(\kappa)/I_0(\kappa). \tag{2.6}$$

See Figure 2 for a plot of $\lambda = L(\kappa)$. Note $L'(\kappa) > 0$ and $\lambda \in [0, 1)$ for $\kappa \in [0, \infty)$. Define

$$\xi = (\xi_1, \xi_2) := \big(\Re(z), \Im(z)\big) = \big(\lambda\, \cos(\mu), \lambda\, \sin(\mu)\big). \tag{2.7}$$

The domain of ξ is the unit disk. Note $\lambda = \|\xi\|$ and $\mu = \text{atan2}(\xi)$, where

$$\text{atan2}\big(a\, \cos(\theta), a\, \sin(\theta)\big) := \theta \qquad \text{for } \theta \in [0, 2\,\pi), \tag{2.8}$$

for any $a > 0$.

3. Parametric Bayesian inference

In this section I describe parametric Bayesian inference as a stepping-stone on the way to nonparametric inference.

Let $\theta_{1:n} = (\theta_1, \ldots, \theta_n)$ denote the observations, which are independent draws from a von Mises distribution conditional a given value for the parameter θ. Note $\theta_i \in [0, 2\,\pi)$ and let

$$x_i = (\cos(\theta_i), \sin(\theta_i)). \tag{3.1}$$

Note $\|x_i\| = 1$.

The likelihood is given by

$$p(\theta_{1:n}|\phi) = \prod_{i=1}^{n} f(\theta_i|\phi). \tag{3.2}$$

The posterior distribution for the parameters is

$$p(\phi|\theta_{1:n}) \propto p(\theta_{1:n}|\phi)\, p(\phi), \tag{3.3}$$

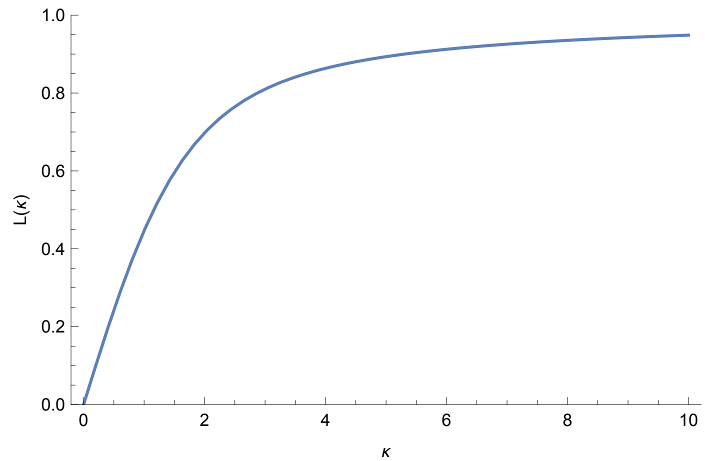

FIGURE 2. The length function, $\lambda = L(\kappa)$.

where $p(\phi)$ denotes the prior distribution for ϕ. The posterior predictive distribution for the next observation is

$$p(\theta_{n+1}|\theta_{1:n}) = \int f(\theta_{n+1}|\phi)\, p(\phi|\theta_{1:n})\, d\phi. \tag{3.4}$$

Sufficient statistics. Let

$$\widehat{\zeta} = (\widehat{\zeta}_1, \widehat{\zeta}_2) = \sum_{i=1}^{n} x_i. \tag{3.5}$$

Define

$$s = \|\widehat{\zeta}\| \qquad \text{and} \qquad \widehat{\mu} = \operatorname{atan2}(\widehat{\zeta}). \tag{3.6}$$

Sufficient statistics are $(\widehat{\zeta}, n)$ or equivalently $(\widehat{\mu}, s, n)$.

It is convenient to express the likelihood as[3]

$$p(\theta_{1:n}|\mu,\kappa) = \frac{\exp\big(s\,\kappa\,\cos(\widehat{\mu}-\mu)\big)}{I_0(\kappa)^n} = 2\,\pi\,\mathsf{vonMises}(\mu|\widehat{\mu}, s\,\kappa)\,\frac{I_0(s\,\kappa)}{I_0(\kappa)^n}. \tag{3.7}$$

omitting the factor $1/(2\,\pi)^n$. Note $p(\theta_{1:n}|\mu,\kappa=0)=1$.

Prior distribution. Essentially any prior distribution for $\phi = (\mu,\kappa)$ can be accommodated. I consider only isotopic priors for which μ is uniformly distributed over $[0, 2\,\pi)$:

$$p(\mu,\kappa) = \frac{1}{2\,\pi}\,p(\kappa). \tag{3.8}$$

The associated posterior distribution is given by

$$p(\mu,\kappa|\theta_{1:n}) = \mathsf{vonMises}(\mu|\widehat{\mu}, s\,\kappa)\,p(\kappa|\theta_{1:n}) \tag{3.9}$$

where

$$p(\kappa|\theta_{1:n}) = \frac{I_0(s\,\kappa)\,I_0(\kappa)^{-n}\,p(\kappa)}{\int I_0(s\,\kappa)\,I_0(\kappa)^{-n}\,p(\kappa)\,d\kappa}. \tag{3.10}$$

[3]One may confirm that $\widehat{\mu}$ and s/n are the maximum likelihood estimates of μ and λ.

The conditional distribution for μ is von Mises and does not depend on the prior for κ. The predictive distribution is

$$p(\theta_{n+1}|\theta_{1:n}) = \int p(\theta_{n+1}|\theta_{n+1},\kappa)\, p(\kappa|\theta_{1:n})\, d\kappa, \tag{3.11}$$

where

$$p(\theta_{n+1}|\theta_{1:n},\kappa) = \int_0^{2\pi} \mathsf{vonMises}(\theta_{n+1}|\mu,\kappa)\, \mathsf{vonMises}(\mu|\widehat{\mu}, s\,\kappa)\, d\mu = \frac{I_0\big(v(\theta_{n+1},\theta_{1:n})\,\kappa\big)}{2\,\pi\, I_0(\kappa)\, I_0(s\,\kappa)}, \tag{3.12}$$

and where

$$v(\theta_{n+1},\theta_{1:n}) := \sqrt{1 + s^2 + 2\,s\,\cos(\theta_{n+1} - \widehat{\mu})}. \tag{3.13}$$

The conjugate isotropic prior is given by[4]

$$p(\kappa) = \mathsf{Bessel}(\kappa|0,\underline{n}) = \frac{1}{C(0,\underline{n})\, I_0(\kappa)^{\underline{n}}}, \tag{3.14}$$

where

$$C(a,b) = \int_0^\infty \frac{I_0(a\,\kappa)}{I_0(\kappa)^b}\, d\kappa \tag{3.15}$$

and $\underline{n} > 0$ can be interpreted as the prior number of observations in favor of isotropy. Figure 3 shows plots of the prior for three values of $\underline{n}$. As $\underline{n} \to \infty$, the prior for κ becomes concentrated around $\kappa = 0$. As $\underline{n} \to 0$, the prior approaches zero everywhere pointwise. The limiting prior is improper.

Given the conjugate isotopic prior distribution, the posterior distribution for κ is

$$p(\kappa|\theta_{1:n}) = \mathsf{Bessel}(\kappa|s, n+\underline{n}) = \frac{I_0(s\,\kappa)}{C(s,\underline{n}+n)\, I_0(\kappa)^{n+\underline{n}}}. \tag{3.16}$$

The joint posterior is

$$p(\mu,\kappa|\theta_{1:n}) = \frac{1}{2\,\pi\, C(s,\underline{n}+n)}\, \frac{e^{s\,\kappa\,\cos(\mu-\widehat{\mu})}}{I_0(\kappa)^{\underline{n}+n}}, \tag{3.17}$$

and the marginal posterior for μ is

$$p(\mu|\theta_{1:n}) = \frac{1}{2\,\pi\, C(s,\underline{n}+n)} \int_0^\infty \frac{e^{s\,\kappa\,\cos(\mu-\widehat{\mu})}}{I_0(\kappa)^{\underline{n}+n}}\, d\kappa. \tag{3.18}$$

The predictive distribution becomes

$$p(\theta_{n+1}|\theta_{1:n}) = \frac{C\big(v(\theta_{n+1},\theta_{1:n}),\, \underline{n}+n+1\big)}{2\,\pi\, C(s,\underline{n}+n)}. \tag{3.19}$$

[4]The conjugate prior for (μ,κ) is presented in Appendix A.

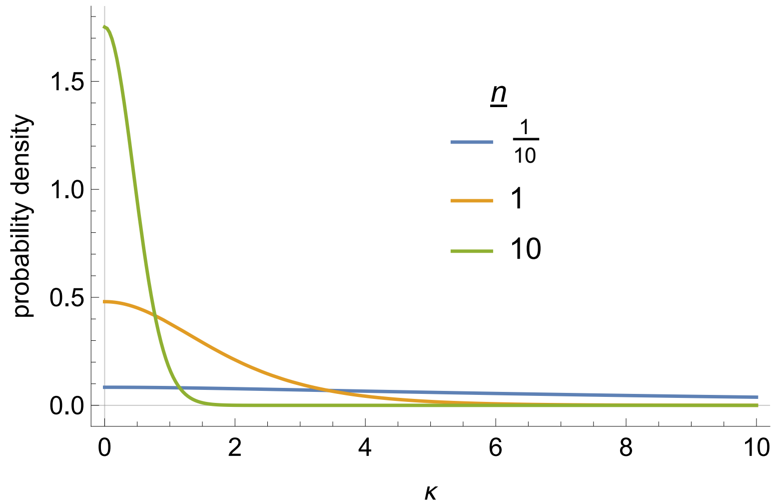

FIGURE 3. Conjugate prior for κ with $\underline{n} = 1/10, 1, 10$.

Bayes factor in favor of isotropy. Here I address the question as to the odds in favor of the proposition that the data come from an isotropic distribution. In this model, isotropy is characterized by $\kappa = 0$.

Let $\mathcal{B}$ denote the Bayes factor in favor of isotropy. Then

$$\mathcal{B} = \frac{p(\theta_{1:n}|\kappa = 0)}{p(\theta_{1:n})} = \frac{p(\kappa = 0|\theta_{1:n})}{p(\kappa = 0)}. \tag{3.20}$$

(The first equality follows from the definition of the Bayes factor; the second follows from Bayes rule.) First note

$$p(\theta_{1:n}|\kappa) = \int_0^{2\pi} p(\theta_{1:n}|\mu, \kappa)\, p(\mu)\, d\mu = \frac{I_0(s\,\kappa)}{I_0(\kappa)^n}. \tag{3.21}$$

Next note $p(\theta_{1:n}|\kappa = 0) = 1$ (since $I_0(0) = 1$). Therefore, $\mathcal{B} = 1/p(\theta_{1:n})$. The likelihood of the unrestricted model is the average likelihood according to the prior for κ:

$$p(\theta_{1:n}) = \int_0^\infty p(\theta_{1:n}|\kappa)\, p(\kappa)\, d\kappa. \tag{3.22}$$

For the conjugate prior we have

$$\mathcal{B}_{\underline{n}} = \frac{C(0, \underline{n})}{C(s, n + \underline{n})}, \tag{3.23}$$

where $C(a, b)$ is given in the Appendix.

Remark. A single observation provides no evidence regarding isotropy, regardless of the prior for κ, because $n = 1$, $\widehat{\lambda} = 1$, $p(\theta_1|\kappa) = 1$, and $p(\theta_1) = 1$. Nevertheless, the posterior predictive distribution for θ_2 will have a peak located at θ_1 and the precision of the peak will depend on the prior for κ.

Table 1. Roulette data (in degrees counterclockwise from east).

Direction								
43	45	52	61	75	88	88	279	357

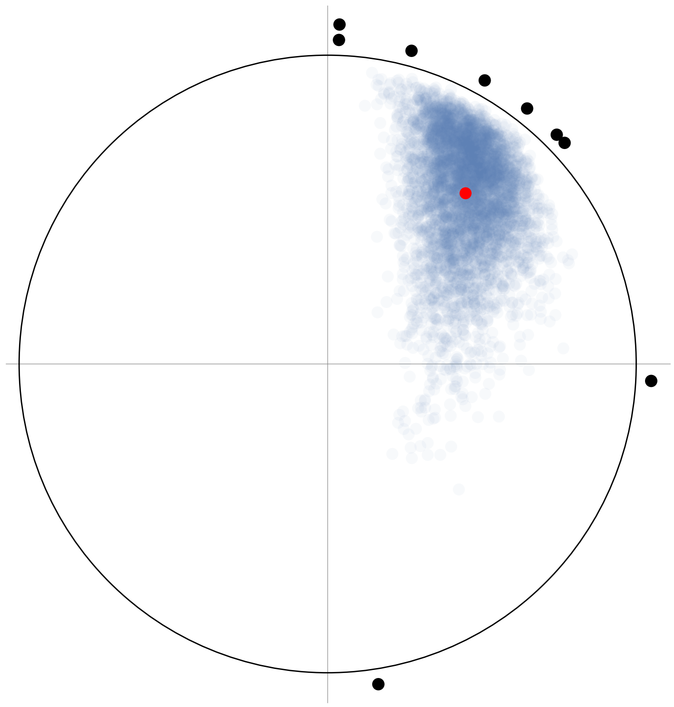

Figure 4. Circular plot of the roulette data (black dots) with sample mean vector $\widehat{\xi}$ (red dot). Also shown are bootstrap draws $\{\widehat{\xi}^{(b)}\}_{b=1}^{B}$ [see Section 4].

Example: Roulette data. As an example consider the roulette data presented in Table 1 and in Figure 4. There are nine observations of the final orientation of a roulette wheel after being spun.

Figure 7 shows the Bayes factor in favor of isotropy, $\mathcal{B}_{\underline{n}}$. The data favor isotropy only for small values of $\underline{n}$. Posterior predictive distributions are shown in Figure 6. Comparing Figures 7 and 6 reveals a peculiarity: The prior that produces evidence in favor of isotropy, $\underline{n} = 10^{-2}$, produces the predictive distribution that most deviates from isotropy. This is because this prior is the flattest (of those examined). This flatness produces the least shrinkage away from the likelihood. At the same time, the flatness puts the least prior density at $\kappa = 0$, which allows that posterior distribution to place more density there. Also, larger κ are associated with greater prior correlation between θ_1 and θ_2.

Once one entertains the idea that the isotropic model might be correct, it behooves one to continue to entertain both models using updated probabilities. That leads to Bayesian Model Averaging (BMA). Suppose the prior odds in favor of isotropy were even. Then the

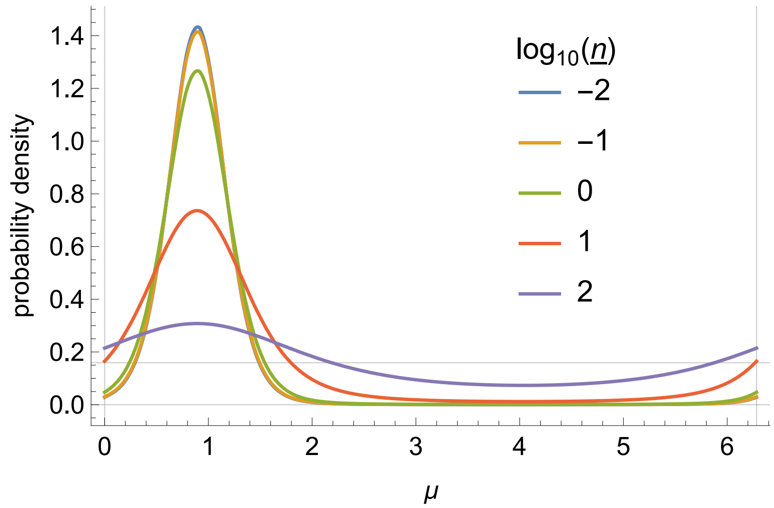

FIGURE 5. Posterior distributions for μ given $\log_{10}(\underline{n}) = -2, -1, 0, 1, 2$.

posterior odds would be given by $\mathcal{B}_{\underline{n}}$ and the posterior probability in favor of the isotropic model would be

$$\frac{\mathcal{B}_{\underline{n}}}{1+\mathcal{B}_{\underline{n}}} = \frac{C(0,\underline{n})}{C(0,\underline{n}) + C(s,\underline{n}+n)}. \tag{3.24}$$

The resulting predictive distribution is

$$p(\theta_{n+1}|\theta_{1:n}) = \frac{1}{2\,\pi}\,\frac{C(0,\underline{n}) + C\big(v(\theta_{n+1},\theta_{1:n}),\,\underline{n}+n+1\big)}{C(0,\underline{n}) + C(s,\underline{n}+n)}. \tag{3.25}$$

See Figure 8. As $\underline{n} \to 0$, the weight on the isotropic model attenuates the greater deviation from isotropy in the other component.

Another approach. Here is another approach to exploring the odds for or against isotropy. Consider the *mixture model* (which is not an either/or model):

$$p(\theta_i|\mu,\kappa,\omega) = \omega\,\mathsf{Uniform}(\theta_i|0,2\,\pi) + (1-\omega)\,f(\theta_i|\mu,\kappa), \tag{3.26}$$

where $\omega \sim \mathsf{Uniform}(0,1)$. Marginal posterior distributions for ω are shown in Figure 9, each depending on a different prior for κ. Consider two specializations, ω_1 and ω_2. For each of the disributions, the Bayes factor in favor of the ω_2 model relative to the ω_1 model is

$$\frac{p(\omega_2|\theta_{1:n})}{p(\omega_1|\theta_{1:n})}. \tag{3.27}$$

In particular, $p(\omega = 1|\theta_{1:n})/p(\omega = 0|\theta_{1:n}) = \mathcal{B}_{\underline{n}}$, the Bayes factor in favor of isotropy discussed above. The plots in Figure 9 suggest that for $\underline{n} \le 1$ an intermediate value for ω can produce a model that is more likely than either of the two components.

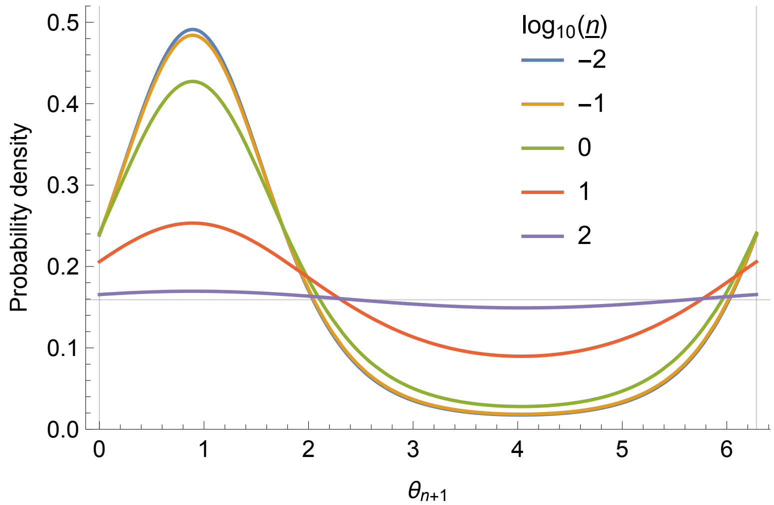

FIGURE 6. Predictive distributions for $\log_{10}(\underline{n}) = -2, -1, 0, 1, 2$.

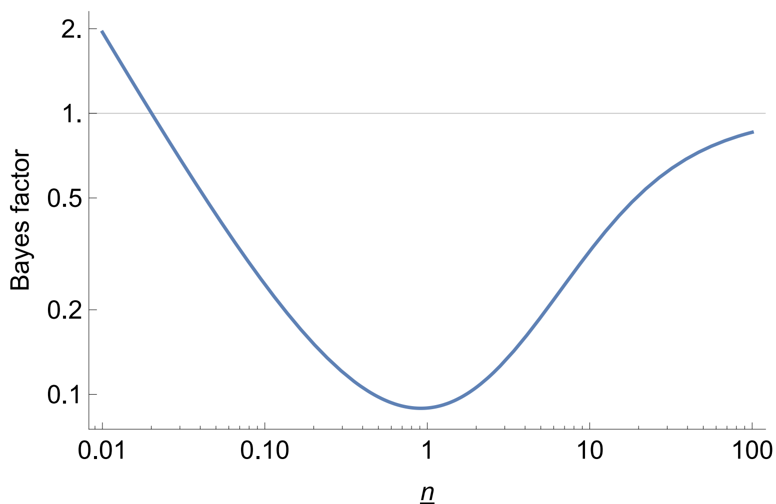

FIGURE 7. Bayes factor in favor of isotropy, $\mathcal{B}_{\underline{n}}$, as a function of $\underline{n}$, which parameterizes the conjugate prior for κ. See (3.23).

4. Bayesian bootstrap

Before proceeding to the fully nonparametric approach, I present the Bayesian bootstrap. For $b = 1, \ldots, B$, draw $u^{(b)} \in \Delta^{n-1}$ using a flat Dirichlet distribution. Then compute

$$\widehat{\xi}^{(b)} = \sum_{i=1}^{n} u_i^{(b)} x_i \qquad \text{and} \qquad \widehat{\mu}^{(b)} = \text{atan2}\big(\widehat{\xi}^{(b)}\big). \tag{4.1}$$

The draws $\{\widehat{\mu}^{(b)}\}_{b=1}^{B}$ can be used to approximate the posterior distribution for μ.

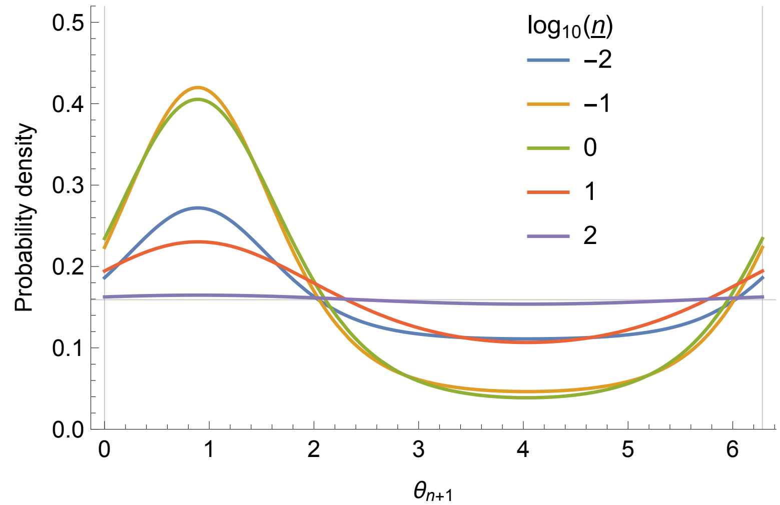

FIGURE 8. BMA predictive distributions for $\log_{10}(\underline{n}) = -2, -1, 0, 1, 2$.

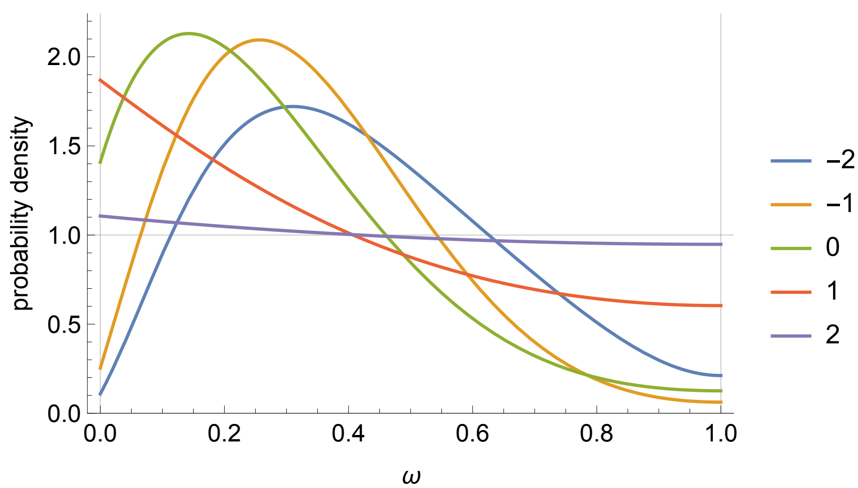

FIGURE 9. Posterior distributions for ω, depending on the prior for $\underline{n}$.

In addition, we can compute

$$\widehat{\lambda}^{(b)} = \|\widehat{\xi}^{(b)}\|. \tag{4.2}$$

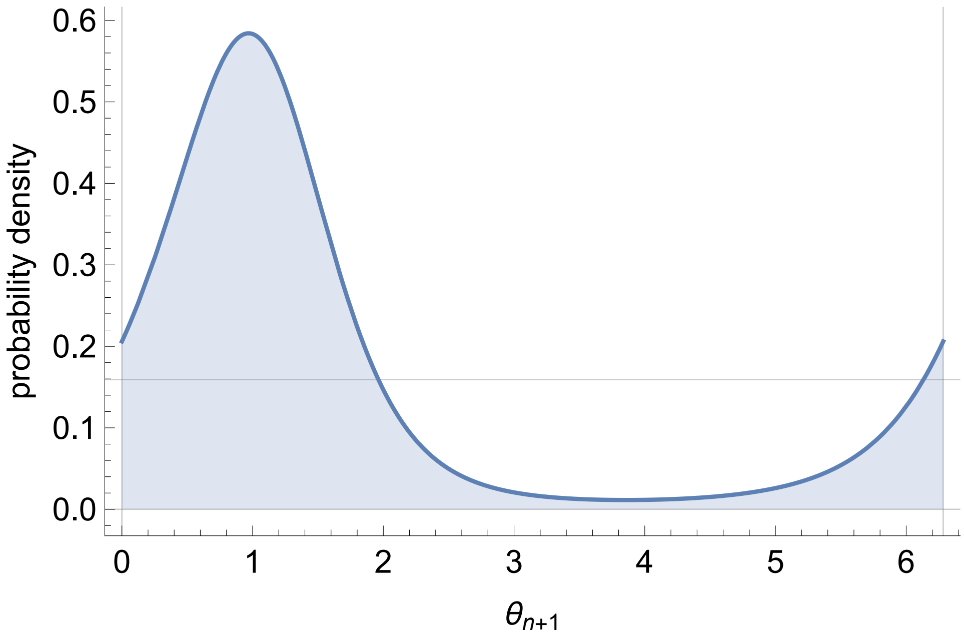

FIGURE 10. Predictive distribution using the Bayesian bootstrap.

Let $\widehat{\kappa}^{(b)} = K(\widehat{\lambda}^{(b)})$, where $K(\lambda)$ is the inverse relation between λ and κ.[5] Then the posterior predictive distribution is approximated by

$$p(\theta_{n+1}|\theta_{1:n}) \approx \frac{1}{B}\sum_{b=1}^{B} \mathsf{vonMises}(\theta_{n+1}|\widehat{\mu}^{(b)}, \widehat{\kappa}^{(b)}). \tag{4.3}$$

See Figure 10.

5. Nonparameteric Bayesian inference

The Dirichlet process mixture (DPM) model is a Bayesian nonparametric model that generalizes the parametric model presented in Section 3.

For our purposes, the DPM can be expressed in terms of an infinite-order mixture of von Mises distributions:

$$p(\theta_i|\psi) = \sum_{c=1}^{\infty} w_c\, f(\theta_i|\phi_c), \tag{5.1}$$

where $\psi = (\boldsymbol{w}, \boldsymbol{\phi})$ and $\boldsymbol{w} = (w_1, w_2, \ldots)$ denotes an infinite collection of nonnegative mixture weights that sum to one and $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots)$ denotes a corresponding collection of mixture-component parameters where $\phi_c = (\mu_c, \kappa_c)$ or $\phi_c = (\mu_c, \lambda_c)$ depending on which

[5]The inverse function $K = L^{-1}$ is not available in closed-form. Nevertheless, we can numerically construct K by evaluating the pair $(L(\kappa), \kappa)$ on a grid and forming an interpolating function from the result. The following grid works well: $\log_{10}(\kappa)$ varies from -10 to 10 in steps of 10^{-3}.

parametrization is more convenient. Assuming the data can be treated as conditionally independent, the likelihood is given by

$$p(\theta_{1:n}|\psi) = \prod_{i=1}^{n} p(\theta_i|\psi). \tag{5.2}$$

The prior for ψ can be expressed as

$$p(\psi) = p(\boldsymbol{w})\, p(\boldsymbol{\phi}) = p(\boldsymbol{w}) \prod_{c=1}^{\infty} p(\phi_c). \tag{5.3}$$

Let the prior for each ϕ_c be the same as for ϕ in the parametric section above. [See (3.8) and (3.14).] With this prior for $\boldsymbol{\phi}$, the prior predictive distribution is

$$p(\theta_i) = \int p(\theta_i|\psi)\, p(\psi)\, d\psi = \int f(\theta_i|\phi_c)\, p(\phi_c)\, d\phi_c = \mathsf{Uniform}(\theta_i|0, 2\,\pi), \tag{5.4}$$

which follows from the independence of $\boldsymbol{w}$ from $\boldsymbol{\phi}$ and the independence among the components of ψ and the isotropic prior.

It remains to specify the prior for the mixture weights. Let

$$\boldsymbol{w} \sim \mathsf{Stick}(\alpha), \tag{5.5}$$

where $\mathsf{Stick}(\alpha)$ denotes the stick-breaking distribution given by[6]

$$w_c = v_c \prod_{\ell=1}^{c-1} (1 - v_\ell) \qquad \text{where } v_c \sim \mathsf{Beta}(1, \alpha). \tag{5.6}$$

The parameter α controls the rate at which the weights decline on average. In particular, the weights decline geometrically in expectation:

$$E[w_c|\alpha] = \alpha^{c-1}\,(1+\alpha)^{-c}. \tag{5.7}$$

Note $E[w_1|\alpha] = 1/(1+\alpha)$ and $E\big[\sum_{c=m+1}^{\infty} w_c|\alpha\big] = \big(\alpha/(1+\alpha)\big)^m$.

The nonparametric model as expressed in (5.1) specializes to the parametric model described in Section 3 as a limiting case:

$$\lim_{\alpha \to 0} \sum_{c=1}^{\infty} p(\theta_i|\psi) = f(\theta_i|\phi_1). \tag{5.8}$$

Let the prior distribution for the concentration parameter be given by

$$p(\alpha) = \mathsf{Log\text{-}Logistic}(\alpha|1, 1) = \frac{1}{(1+\alpha)^2}. \tag{5.9}$$

With this prior, the Bayes factor in favor of the parametric model is given by $p(\alpha = 0|x_{1:n}, I)$.

[6] Start with a stick of length one. Break off the fraction v_1 leaving a stick of length $1 - v_1$. Then break off the fraction v_2 of the remaining stick leaving a stick of length $(1 - v_1)\,(1 - v_2)$. Continue in this manner.

TABLE 2. Turtle direction data. Orientation of 76 turtles after laying eggs.

Direction (in degrees) clockwise from north									
8	9	13	13	14	18	22	27	30	34
38	38	40	44	45	47	48	48	48	48
50	53	56	57	58	58	61	63	64	64
64	65	65	68	70	73	78	78	78	83
83	88	88	88	90	92	92	93	95	96
98	100	103	106	113	118	138	153	153	155
204	215	223	226	237	238	243	244	250	251
257	268	285	319	343	350				

Posterior predictive distribution. The goal of estimation is to compute the predictive distribution for the next observation x_{n+1} conditioned on the already-observed data $x_{1:n}$. This posterior predictive distribution can be expressed as

$$p(\theta_{n+1}|\theta_{1:n}) = \int p(\theta_{n+1}|\psi)\, p(\psi|\theta_{1:n})\, d\psi. \tag{5.10}$$

Given draws $\{\psi^{(r)}\}_{r=1}^R = \{(\boldsymbol{w}^{(r)}, \boldsymbol{\theta}^{(r)})\}_{r=1}^R$ from the posterior distribution, the posterior predictive distribution can be approximated (via Rao–Blackwellization) as

$$p(\theta_{n+1}|\theta_{1:n}) \approx \frac{1}{R}\sum_{r=1}^R p(\theta_{n+1}|\psi^{(r)}) = \frac{1}{R}\sum_{r=1}^R \sum_{c=1}^m w_c^{(r)}\, f(\theta_{n+1}|\phi_c^{(r)}), \tag{5.11}$$

where m is an upper bound chosen to provide an adequate approximation.[7]

Sampler. One may adopt the blocked Gibbs sampler described in Gelman et al. (2014, pp. 552–553). Details of the sampler can be found in Ishwaran and James (2001). This sampler relies on approximating $p(\theta_i|\psi)$ with a finite sum: Choose m large enough to make $\big(\alpha/(1+\alpha)\big)^m$ close enough to zero and set $v_m = 1$.

This sampler uses the classification variables $z_{1:n} = (z_1, \ldots, z_n)$, where $z_i = c$ signifies x_i is assigned to cluster c. The Gibbs sampling scheme involves cycling through the following full conditional posterior distributions:

$$p(z_{1:n}|\theta_{1:n}, \boldsymbol{w}, \boldsymbol{\phi}, \alpha) = \prod_{i=1}^n p(z_i|\theta_i, \boldsymbol{w}, \boldsymbol{\phi}) \tag{5.12a}$$

$$p(\boldsymbol{w}|\theta_{1:n}, z_{1:n}, \boldsymbol{\phi}, \alpha) = p(\boldsymbol{w}|z_{1:n}, \alpha) \tag{5.12b}$$

$$p(\boldsymbol{\phi}|\theta_{1:n}, z_{1:n}, \boldsymbol{w}, \alpha) = \prod_{c=1}^m p(\theta_c|\theta^c) \tag{5.12c}$$

$$p(\alpha|\theta_{1:n}, z_{1:n}, \boldsymbol{w}, \boldsymbol{\phi}) = p(\alpha|z_{1:n}), \tag{5.12d}$$

where θ^c is the collection of observations for which $z_i = c$.

[7] Alternatively, Walker's slice sampler could be used to avoid truncation.

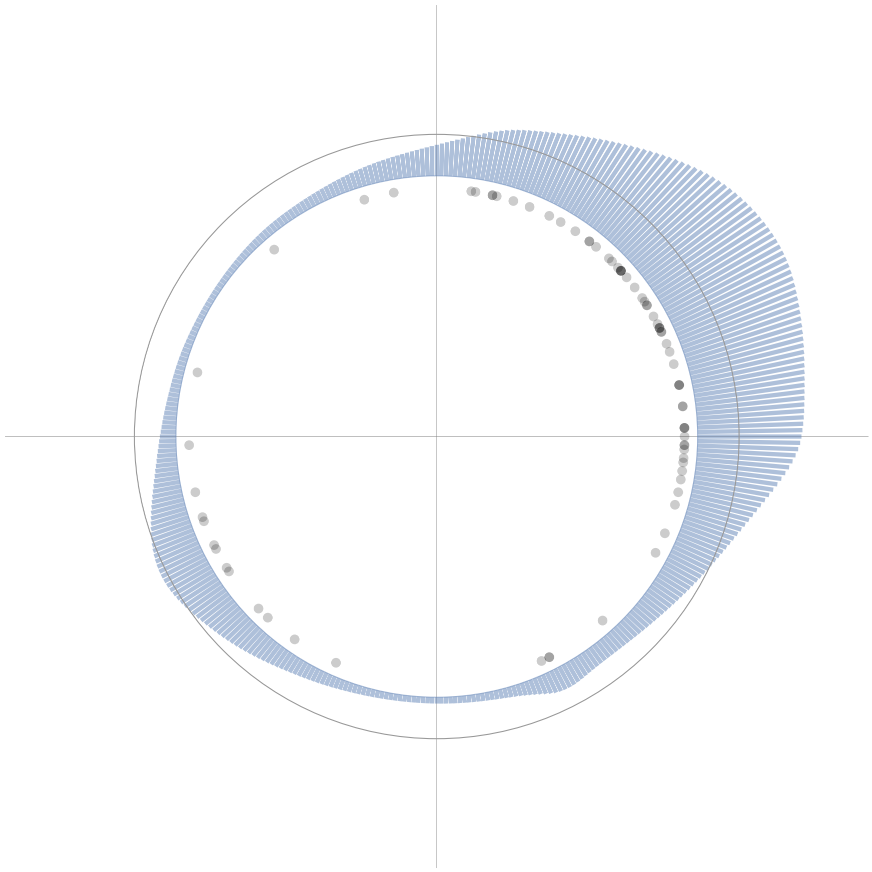

FIGURE 11. Turtle direction data and nonparametric density estimate.

The conditional distribution for z_i is characterized by

$$p(z_i = c|\theta_{1:n}, \boldsymbol{w}, \boldsymbol{\phi}) \propto w_c\, f(\theta_i|\phi_c), \tag{5.13}$$

for $c = 1, \ldots, m$. Let n_c denote the multiplicity of c in $z_{1:n}$ (i.e., the number of times c occurs in $z_{1:n}$). Note $\sum_{c=1}^m n_c = n$. The weights w can be updated via the stick-breaking weights v:

$$v_c|z_{1:n} \sim \mathsf{Beta}\big(1 + n_c, \alpha + \textstyle\sum_{c'=c+1}^m n_{c'}\big) \tag{5.14}$$

for $c = 1, \ldots, m-1$. Finally, $\phi_c|\theta^c$ is updated as in a finite mixture model, with the parameters for the unoccupied clusters (for which $n_c = 0$) sampled directly from the prior $p(\phi_c)$. See Appendix B for additional details.

Finally, note that

$$p(\alpha|z_{1:n}) \propto p(z_{1:n}|\alpha)\, p(\alpha) \propto \frac{\alpha^d\, \Gamma(\alpha)}{\Gamma(n+\alpha)}\, p(\alpha), \tag{5.15}$$

where d is the number of occupied clusters (i.e., clusters for which $n_c > 0$).

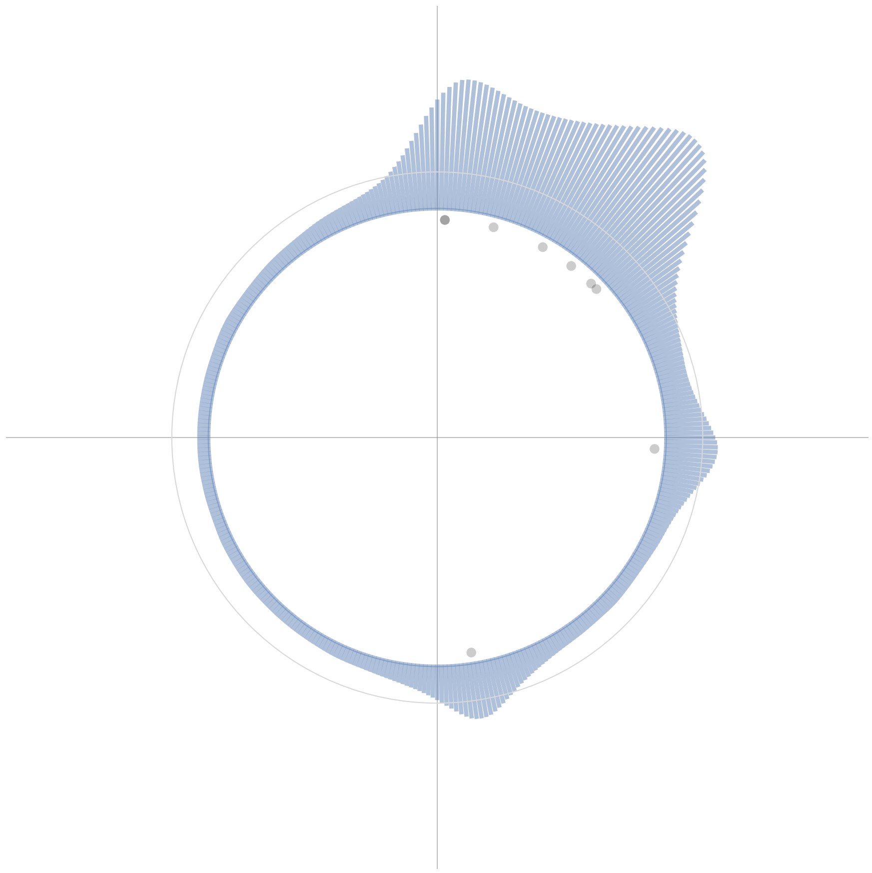

FIGURE 12. Roulette data and nonparametric density estimate.

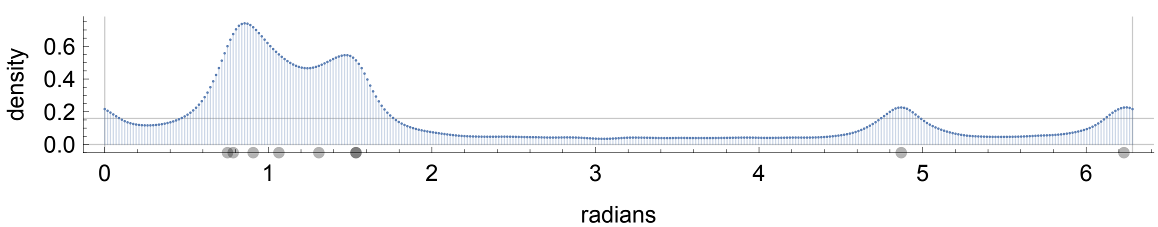

FIGURE 13. Roulette data and nonparametric density estimate.

Example. See Table 2 for the turtle direction data. [Gould's data cited by Stephens (1969). See Table 3, Sample 7 therein.]

See Figure 11 for the estimated density using the turtle direction data.

Appendix A. Conjugate prior

Guttorp and Lockhart (1988) present the conjugate prior for (μ, κ). As a preliminary, define the density for $x \geq 0$[8]

$$\mathsf{Bessel}(x|a,b) = \frac{I_0(a\,x)/I_0(x)^b}{C(a,b)}, \tag{A.1}$$

where $0 \leq a < b$ and

$$C(a,b) = \int_0^\infty I_0(a\,\kappa)/I_0(\kappa)^b \, d\kappa. \tag{A.2}$$

Note $\mathsf{Bessel}(1,b) = \mathsf{Bessel}(0,b-1)$. Also note $\mathsf{Bessel}(0|a,b) = 1/C(a,b)$.

The conjugate prior is characterized by $(\underline{\mu}, \underline{s}, \underline{n})$, subject to

$$0 \leq \underline{\mu} < 2\,\pi \qquad \text{and} \qquad 0 \leq \underline{s} < \underline{n}. \tag{A.3}$$

Define $\underline{\zeta} := \big(\underline{s}\,\cos(\underline{\mu}), \underline{s}\,\sin(\underline{\mu})\big)$ and note that $\underline{s} = \|\underline{\zeta}\|$ and $\underline{\mu} = \text{atan2}(\underline{\zeta})$. The conjugate prior can be expressed as

$$p(\mu,\kappa) = f(\mu|\underline{\mu}, \underline{s}\,\kappa)\,\mathsf{Bessel}(\kappa|\underline{s}, \underline{n}). \tag{A.4}$$

The isotropic conjugate prior is obtained by setting $\underline{s} = 0$:

$$p(\mu,\kappa) = \frac{1}{2\,\pi}\,\mathsf{Bessel}(\kappa|0,\underline{n}). \tag{A.5}$$

Let

$$\overline{n} = \underline{n} + n \tag{A.6a}$$

$$\overline{\zeta} = \underline{\zeta} + \widehat{\zeta} \tag{A.6b}$$

$$\overline{s} = \|\overline{\zeta}\| \tag{A.6c}$$

$$\overline{\mu} = \text{atan2}\big(\overline{\zeta}\big), \tag{A.6d}$$

where $(\widehat{\zeta}, n)$ are the sufficient statistics for $\theta_{1:n}$. Then the posterior distribution is given by[9]

$$p(\mu,\kappa|\theta_{1:n}) = f(\mu|\overline{\mu}, \overline{s}\,\kappa)\,\mathsf{Bessel}(\kappa|\overline{s}, \overline{n}). \tag{A.7}$$

With the isotropic prior, $\overline{\zeta} = \widehat{\zeta}$, $\overline{\mu} = \widehat{\mu}$, $\overline{s} = s$, and

$$p(\mu,\kappa|\theta_{1:n}) = f(\mu|\widehat{\mu}, s\,\kappa)\,\mathsf{Bessel}(\kappa|s, \underline{n} + n). \tag{A.8}$$

[8]The Bessel distribution is not standard and I do not know of any references for it.

[9]Damien and Walker (1999) provide a Gibbs sampler (via the introduction of two auxiliary variables) for drawing from (A.7). The advantage of a Gibbs sampler over a Metropolis-Hastings sampler is that no data-specific tuning is required. However, as noted above, the parametric model may be easily estimated without resorting to sampling. Nevertheless, as part of a DPM sampling may be required.

Appendix B. Sampler details

See Best and Fisher (1979) for drawing $\mu|\kappa$ and Forbes and Mardia (2015) for drawing $\kappa|\mu$. Forbes and Mardia (2015) characterize what they call the Bessel exponential distribution:[10]

$$\textsf{Bessel-Exp}(\kappa|\beta_0, \eta) \propto \frac{e^{-\beta_0\,\eta\,\kappa}}{I_0(\kappa)^\eta}, \tag{B.1}$$

where $\beta_0 > -1$ and $\eta > 0$. Note that $\textsf{Bessel-Exp}(0, b) = \textsf{Bessel}(0, b)$.

Draws from the prior can be had via

$$p(\mu) = \textsf{Uniform}(\mu|0, 2\,\pi) \tag{B.2}$$

$$p(\kappa|\underline{n}) = \textsf{Bessel-Exp}(\kappa|0, \underline{n}). \tag{B.3}$$

Draws from the joint posterior distribution for (μ, κ) can be had via a Gibbs sampler:

$$p(\mu|\theta_{1:n}, \kappa) = f(\mu|\widehat{\mu}, s\,\kappa) \tag{B.4}$$

$$p(\kappa|\theta_{1:n}, \mu, \underline{n}) = \textsf{Bessel-Exp}\left(\kappa \,\middle|\, \frac{-s\cos(\mu - \widehat{\mu})}{n + \underline{n}},\, n + \underline{n}\right). \tag{B.5}$$

Appendix C. Extensions

(1) Toroidal data. Joint observations on two circular random variables:

$$\theta_i = (\theta_{i1}, \theta_{i2}) \in [0, 2\,\pi)^2.$$

It should be straightforward to use orthogonal kernels.

(2) Axial data, for which θ and $\theta + \pi$ (in radians) cannot be distinguished.

(3) Spherical data. And with gaps.

(4) Estimate densities when there are measurement gaps.

(5) Markov switching to model changes in direction.

(6) Put a point mass on uniformity ($\kappa = 0$) in the kernel:

$$f(\theta|\phi) = \begin{cases} \omega\,\textsf{Uniform}(\theta|0, 2\,\pi) & \kappa = 0 \\ (1-\omega)\,\textsf{vonMises}(\theta|\mu, \kappa) & \kappa > 0 \end{cases} \tag{C.1}$$

Comparing $p(\omega = 1|\theta_{1:n})$ with $p(\omega = 0|\theta_{1:n})$ will tell us something about isotropy.

(7) Compare and contrast with fitting a functional form to observations taken at various locations on the unit circle.
   - Suppose we have a collection of microphones at a single location that listen in different directions.
   - The input is non-negative.
   - This takes us into simplex regression where the basis densities are von Mises distributions:

$$g(\theta|w) = B \sum_{j=1}^{k} w_j\,\textsf{vonMises}(\theta|\mu_j, k)$$

[10] Algorithm 1 in Forbes and Mardia (2015) fails when $\kappa_0 \le 0$ (see line 5). In a personal email, Forbes suggests fixing the algorithm by modifying line 4 as follows: $c_1 = \max[0, 1/2 + \{1 - 1/(2\,\eta)\}/(2\,\eta)]$.

where $B > 0$, $w = (w_1, \ldots, w_k) \in \Delta^{k-1}$, and $\mu_j = 2\,\pi\,(j-1)/k$.

(8) Latent variable density estimation
- Compute well-informed prior for μ_{n+1}
- Predictive distribution $p(\mu_{n+1}|\Theta_{1:n})$
- Likelihood $p(\Theta_{1:n}|\mu_{1:n}) = \prod_{i=1}^{n} p(\Theta_i|\mu_i)$

$$p(\Theta_i|\mu_i) = \int_0^\infty \frac{e^{s_i,\kappa_i \cos(\mu_i - \widehat{\mu}_i)}}{I_0(\kappa_i)^{n_i}}\, d\kappa_i \tag{C.2}$$

  - κ_i is nuisance parameter; integrate it out with flat prior (this requires $n_i \geq 2$)
- Open-minded prior for μ_i
  - $p(\mu_i|\psi) = \sum_{c=1}^{\infty} w_c\, \mathsf{vonMises}(\mu_i|a_c, b_c)$
  - $p(a_c, b_c) = \frac{1}{2\,\pi}\, \mathsf{Bessel}(b_c|\nu)$
  - Prior predictive: $p(\mu_i) = \frac{1}{2\,\pi}$

References

Best, D. J. and N. I. Fisher (1979). Efficient simulation of the von Mises distribution. *Journal of the Royal Statistical Soceity. Series C (Applied Statistics) 28*(2), 152–157.

Brunner, L. J. and A. Y. Lo (1994). Nonparametric Bayes methods for directional data. *The Canadian Journal of Statistics 22*(3), 401–412.

Damien, P. and S. Walker (1999). A full Bayesian analysis of circular data using the von Mises distribution. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 27*(2), 291–298.

Forbes, P. G. M. and K. V. Mardia (2015). A fast algorithm for sampling from the posterior of a von Mises distribution. *Journal of Statistical Computation and Simulation 85*(13), 2693–2701.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis* (Third ed.). CRC Press.

Ghosh, K., S. R. Jammalamadaka, and R. C. Tiwari (2003). Semiparametric Bayesian techniques for problems in circular data. *Journal of Applied Statistics 30*(2), 145–161.

Guttorp, P. and R. A. Lockhart (1988). Finding the location of a signal: A bayesian analysis. *Journal of the American Statistical Association 83*(402), 322–330.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

Nuñez-Antonio, G., M. C. Ausín, and M. P. Wiper (2014). 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.

Pewsey, A., M. Neuhäuser, and G. D. Ruxton (2013). *Circular Statistics in R.* Oxford University Press.

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

Federal Reserve Bank of Atlanta, Research Department, 1000 Peachtree Street N.E., Atlanta, GA 30309–4470
E-mail address: mark.fisher@atl.frb.org
URL: http://www.markfisher.net