

Exact Bayesian Inference for the Bingham Distribution

Christopher J. Fallaize & Theodore Kypraios*

School of Mathematical Sciences, University of Nottingham, Nottingham, NG7 2RD, UK

June 27, 2018

Abstract

This paper is concerned with making Bayesian inference from data that are assumed to be drawn from a Bingham distribution. A barrier to the Bayesian approach is the parameter-dependent normalising constant of the Bingham distribution, which, even when it can be evaluated or accurately approximated, would have to be calculated at each iteration of an MCMC scheme, thereby greatly increasing the computational burden. We propose a method which enables exact (in Monte Carlo sense) Bayesian inference for the unknown parameters of the Bingham distribution by completely avoiding the need to evaluate this constant. We apply the method to simulated and real data, and illustrate that it is simpler to implement, faster, and performs better than an alternative algorithm that has recently been proposed in the literature.

1 Introduction

Observations that inherit a direction occur in many scientific disciplines (see, for example, Mardia and Jupp; 2000). For example, directional data arise naturally in the biomedical field for protein structure (Boomsma et al.; 2008), cell-cycle (Rueda et al.; 2009) and circadian clock experiments (Levine et al.; 2002); see also the references in Ehler and Galanis (2011). A distribution that has proved useful as a model for spherical data which arise as unsigned directions is the Bingham distribution (Bingham; 1974; Mardia and Jupp; 2000).

The Bingham distribution can be constructed by conditioning a zero-mean multivariate Normal (MVN) distribution to lie on the sphere \mathcal{S}^{q-1} of unit radius in \mathbb{R}^q . In particular, for a given matrix A of dimension $q \times q$, the density with respect to the uniform measure on \mathcal{S}^{q-1} is given by

$$f(\mathbf{x}; A) = \frac{\exp(-\mathbf{x}^T A \mathbf{x})}{c(A)}, \quad \mathbf{x}^T \mathbf{x} = 1 \text{ and } \mathbf{x} \in \mathbb{R}^q, \quad (1)$$

where $c(A)$ is the corresponding normalising constant.

Having observed some directional data, interest then lies in inference for the matrix A in (1). The likelihood of the observed data given the parameters can easily be written down and at first glance it appears that maximum likelihood inference for A is straightforward. However, inferring the matrix A is rather challenging. That is due to the fact that the likelihood of the observed

*Author for correspondence: theodore.kypraios@nottingham.ac.uk

data given the matrix A involves the parameter-dependent normalising constant $c(A)$ which, in the general case, is not available in closed form. Therefore this poses significant challenges to undertake statistical inference involving the Bingham distribution either in a frequentist or Bayesian setting.

Although a maximum likelihood estimator for A can be derived by iterative techniques which are based on being able to approximate $c(A)$ (see, for example, Kent; 1987; Kume and Wood; 2005, 2007; Sei and Kume; 2013), very little attention has been drawn in the literature concerning estimation of A within a Bayesian framework. Walker (2013) considered Bayesian inference for the Bingham distribution which removes the need to compute the normalising constant, using a (more general) method that was developed earlier (Walker; 2011) and cleverly gets around the intractable nature of the normalising constant. However, it requires the introduction of several latent variables and a Reversible-Jump Markov Chain Monte Carlo (RJMCMC) sampling scheme.

The main contribution of this paper is to show how one can draw Bayesian inference for the matrix A , by exploiting the recent developments in Bayesian computation for distributions with doubly intractable normalising constants (Møller et al.; 2006; Murray et al.; 2006). The main advantage of our approach is that it does not require any numerical approximation to $c(A)$ and hence, enables *exact* (in the Monte Carlo sense) Bayesian inference for A . Our method relies on being able to simulate *exact* samples from the Bingham distribution which can be done by employing an efficient rejection sampling algorithm proposed by Kent and Ganeiber (2012).

The rest of the paper is structured as follows. In Section 2 we introduce the family of Angular Central Gaussian distributions and illustrate how such distributions serve as efficient proposal densities to sample from the Bingham distribution. In Section 3 we describe our proposed algorithm while in Section 4 we illustrate our method both using simulated and real directional data from earthquakes in New Zealand. In Section 5 we discuss the computational aspects of our method as well as directions for future research.

2 Rejection Sampling

2.1 Preliminaries

Rejection sampling (Ripley; 1987) is a method for drawing independent samples from a distribution with probability density function $f(x) = f^*(x)/Z_f$ assuming that we can evaluate $f^*(x)$ for any value x , but may not necessarily know Z_f . Suppose that there exists another distribution, with probability density function $g(x) = g^*(x)/Z_g$, often termed an *envelope density*, from which we can easily draw independent samples and can evaluate $g^*(x)$ at any value x . We further assume that there exists a constant M^* for which $M^*g^*(x) \geq f^*(x) \forall x$. We can then draw samples from $f(x)$ as follows:

1. Draw a candidate value y from $g(x)$ and u from $U(0,1)$;
2. if $u \leq \frac{f^*(y)}{M^*g^*(y)}$ accept y ; otherwise reject y and go to step 1.

The set of accepted points provides a sample from the target density $f(x)$. It can be shown that the number of trials until a candidate is accepted has a geometric distribution with mean M , where

$$M = \sup_{x \in \mathcal{R}} \left\{ \frac{f(x)}{g(x)} \right\} < \infty. \quad (2)$$

Therefore, the algorithm will work efficiently provided that M is small or, in other words, the probability of acceptance ($1/M$) is large. Moreover, it is important to note that it is not necessary to know the normalising constants Z_f and Z_g to implement the algorithm; the only requirement is being able to draw from the envelope density $g(x)$ and knowledge of M^* (rather than M).

2.2 The Angular Central Gaussian Distribution

The family of the angular central Gaussian (ACG) distributions is an alternative to the family of the Bingham distributions for modelling antipodal symmetric directional data (Tyler; 1987). An angular central Gaussian distribution on the $(q - 1)$ -dimensional sphere \mathcal{S}^{q-1} can be obtained by projecting a multivariate Gaussian distribution in \mathbb{R}^q , $q \geq 2$, with mean zero onto \mathcal{S}^{q-1} with radius one. In other words, if the vector \mathbf{y} has a multivariate Normal distribution in \mathbb{R}^q with mean $\mathbf{0}$ and variance covariance matrix Ψ , then the vector $\mathbf{x} = \mathbf{y}/\|\mathbf{y}\|$ follows an ACG distribution on \mathcal{S}^{q-1} with $q \times q$ symmetric positive-definite parameter matrix Ψ (Mardia and Jupp; 2000). The probability density function of the ACG distribution with respect to the surface measure on \mathcal{S}^{q-1} is given by

$$g(\mathbf{x}; \Psi) = w_q^{-1} |\Psi|^{-1/2} (\mathbf{x}^T \Psi^{-1} \mathbf{x})^{-q/2} = c_{\text{ACG}}(\Psi) g^*(\mathbf{x}; \Psi) \quad (3)$$

where the constant $w_q = 2\pi^{q/2}/\Gamma(q/2)$ represents the surface area on \mathcal{S}^{q-1} . Denote by $c_{\text{ACG}}(\Psi) = w_q^{-1} |\Psi|^{-1/2}$ the normalising constant where Ψ is a $q \times q$ symmetric positive-definite matrix.

2.3 Rejection Sampling for the Bingham Distribution

Kent and Ganeiber (2012) have demonstrated that one can draw samples from the Bingham distribution using the ACG distribution as an envelope density within a rejection sampling framework. In particular, the following algorithm can be used to simulate a value from the Bingham distribution with parameter matrix A :

1. Set $\Psi^{-1} = I_q + \frac{2}{b}A$ and $M^* \geq \sup_{\mathbf{x}} \left\{ \frac{f^*(\mathbf{x})}{g^*(\mathbf{x})} \right\}$;
2. draw u from $U(0,1)$ and a candidate value \mathbf{y} from the ACG distribution on the sphere with parameter matrix Ψ ;
3. if $u < \frac{f^*(\mathbf{y}; A)}{M^* g^*(\mathbf{y}; \Psi)}$ accept \mathbf{y} ; otherwise reject \mathbf{y} and go to Step 1.

Here, $f^*(\mathbf{y}; A) = \exp(-\mathbf{y}^T A \mathbf{y})$, and $g^*(\mathbf{y}; \Psi) = (\mathbf{y}^T \Psi^{-1} \mathbf{y})^{-\frac{q}{2}}$, the unnormalized Bingham and ACG densities respectively, and $b < q$ is a tuning constant. We found that setting $b = 1$ as a default works well in many situations, but an optimal value can be found numerically by maximising the acceptance probability $1/M$ (see, for example, Ganeiber; 2012).

3 Bayesian Inference

3.1 Preliminaries

Consider the probability density function of the Bingham distribution as given in (1). If $A = V\Lambda V^T$ is the Singular Value Decomposition of A where V is orthogonal and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_q)$,

then it can be shown that if \mathbf{x} is drawn from a distribution with probability density function $f(\mathbf{x}; A)$, the corresponding random vector $\mathbf{y} = X^T V$ is drawn from a distribution with density $f(\mathbf{x}; \Lambda)$ (see, for example, Kume and Walker; 2006; Kume and Wood; 2007). Therefore, without loss of generality, we assume that $A = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_q)$. Moreover, to ensure identifiability, we assume that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q = 0$ (Kent; 1987). Therefore, the probability density function becomes

$$f(\mathbf{x}; \Lambda) = \frac{\exp\left\{-\sum_{i=1}^{q-1} \lambda_i x_i^2\right\}}{c(\Lambda)} \quad (4)$$

with respect to a uniform measure on the sphere and

$$c(\Lambda) = \int_{\mathbf{x} \in S^{q-1}} \exp\left\{-\sum_{i=1}^{q-1} \lambda_i x_i^2\right\} dS^{q-1}(\mathbf{x}).$$

Suppose $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ is a sample of unit vectors in S^{q-1} from the Bingham distribution with density (4). Then the likelihood function is given by

$$L(\Lambda) = \frac{1}{c(\Lambda)^n} \exp\left\{-\sum_{i=1}^{q-1} \lambda_i \sum_{j=1}^n (x_j^i)^2\right\} = \frac{1}{c(\Lambda)^n} \exp\left\{-n \sum_{i=1}^{q-1} \lambda_i \tau_i\right\}, \quad (5)$$

where $\tau_i = \frac{1}{n} \sum_{j=1}^n (x_j^i)^2$. The data can therefore be summarised by $(n, \tau_1, \dots, \tau_{q-1})$, and $(\tau_1, \dots, \tau_{q-1})$ are sufficient statistics for $(\lambda_1, \dots, \lambda_{q-1})$.

3.2 Bayesian Inference

We are interested in drawing Bayesian inference for the matrix Λ , or equivalently, for $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{q-1})$. The likelihood function in (5) reveals that the normalising constant $c(\Lambda)$ plays a crucial role. The fact that there does not exist a closed form expression for $c(\Lambda)$ makes Bayesian inference for Λ very challenging.

For example, if we assign independent Exponential prior distributions to the elements of $\boldsymbol{\lambda}$ with rate μ_i (i.e. mean $1/\mu_i$) subject to the constraint that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{q-1}$ then the density of the posterior distribution of Λ up to proportionality given the data is as follows:

$$\begin{aligned} \pi(\boldsymbol{\lambda} | \mathbf{x}_1, \dots, \mathbf{x}_n) &\propto L(\Lambda) \prod_{i=1}^{q-1} \exp\{-\lambda_i \mu_i\} \cdot \mathbf{1}(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{q-1}) \\ &= \frac{1}{c(\Lambda)^n} \exp\left\{-\sum_{i=1}^{q-1} \lambda_i (n\tau_i + \mu_i)\right\} \cdot \mathbf{1}(\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{q-1}). \end{aligned} \quad (6)$$

Consider the following Metropolis-Hastings algorithm which aims to draw samples from $\pi(\boldsymbol{\lambda} | \mathbf{x}_1, \dots, \mathbf{x}_n)$:

1. Suppose that the current state of the chain is $\boldsymbol{\lambda}^{\text{cur}}$;
2. Update $\boldsymbol{\lambda}$ using, for example, a random walk Metropolis step by proposing $\boldsymbol{\lambda}^{\text{can}} \sim N_{q-1}(\boldsymbol{\lambda}^{\text{cur}}, \Sigma)$;
3. Repeat steps 1-2.

Note that $N_{q-1}(\mathbf{m}, S)$ denotes the density of a multivariate Normal distribution with mean vector \mathbf{m} and variance-covariance matrix S . Step 2 of the above algorithm requires the evaluation of the ratio $\pi(\boldsymbol{\lambda}^{\text{can}}|\mathbf{x}_1, \dots, \mathbf{x}_n)/\pi(\boldsymbol{\lambda}^{\text{cur}}|\mathbf{x}_1, \dots, \mathbf{x}_n)$, which in turn involves evaluation of the ratio $c(\Lambda^{\text{can}})/c(\Lambda^{\text{cur}})$. Therefore, implementing the above algorithm requires an approximation of the normalising constant. In principle, one can employ one of the proposed methods in the literature which are based either on asymptotic expansions (Kent; 1987), saddlepoint approximations (Kume and Wood; 2005) or holonomic gradient methods (Sei and Kume; 2013). Although such an approach is feasible, in practice, it can be very computationally costly since the normalising constant would have to be approximated at every single MCMC iteration. Furthermore, despite how accurate these approximations may be, the stationary distribution of such an MCMC algorithm won't be the distribution of interest $\pi(\boldsymbol{\lambda}|\mathbf{x}_1, \dots, \mathbf{x}_n)$, but an approximation to it.

3.2.1 An Exchange Algorithm

The main contribution of this paper is to demonstrate that recent developments in Markov Chain Monte Carlo algorithms for the so-called doubly intractable distributions enable drawing exact Bayesian inference for the Bingham distribution without having to resort to any kind of approximations.

Møller et al. (2006) proposed an auxiliary variable MCMC algorithm to sample from doubly intractable distributions by introducing a cleverly chosen variable in to the Metropolis-Hastings (M-H) algorithm such that the normalising constants cancel in the M-H ratio. In order for their proposed algorithm to have good mixing and convergence properties, one should have access to some sort of typical value of the parameter of interest, for example a pseudo-likelihood estimator. A simpler version that avoids having to specify such an appropriate auxiliary variable was proposed in Murray et al. (2006). Although both approaches rely on being able to simulate realisations from the Bingham distribution (see Section 2.3), we choose to adapt to our context the approach presented in Murray et al. (2006) because it is simple and easy to implement, since a value of the parameter of interest does not need to be specified.

Consider augmenting the observed data with auxiliary data \mathbf{y} , so that the corresponding augmented posterior density becomes

$$\pi(\boldsymbol{\lambda}, \mathbf{y}, |\mathbf{x}) \propto \pi(\mathbf{x}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})\pi(\mathbf{y}|\boldsymbol{\lambda}), \quad (7)$$

where $\pi(\mathbf{y}|\boldsymbol{\lambda})$ is the same distribution as the original distribution on which the data \mathbf{x} is defined (i.e. in the present case, the Bingham distribution). Proposal values for updating the parameter $\boldsymbol{\lambda}$ are drawn from a proposal density $h(\cdot|\boldsymbol{\lambda})$, although in general this density does not have to depend on the variables $\boldsymbol{\lambda}$. For example, random walk proposals centred at $\boldsymbol{\lambda}$ or independence sampler proposals could be used.

Now consider the following algorithm:

1. Draw $\boldsymbol{\lambda}' \sim h(\cdot|\boldsymbol{\lambda})$;
2. Draw $\mathbf{y} \sim \pi(\cdot|\boldsymbol{\lambda}')$;
3. Propose the exchange move from $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}'$ with probability

$$\min \left(1, \frac{f^*(\mathbf{x}|\boldsymbol{\lambda}')\pi(\boldsymbol{\lambda}')h(\boldsymbol{\lambda}|\boldsymbol{\lambda}')f^*(\mathbf{y}|\boldsymbol{\lambda})}{f^*(\mathbf{x}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda})h(\boldsymbol{\lambda}'|\boldsymbol{\lambda})f^*(\mathbf{y}|\boldsymbol{\lambda}')} \times \frac{c(\Lambda)c(\Lambda')}{c(\Lambda)c(\Lambda')} \right),$$

where $f^*(\mathbf{x}; A) = \exp(-\mathbf{x}^T A \mathbf{x})$ is the unnormalized Bingham density as previously. This scheme targets the posterior distribution of interest (the marginal distribution of $\boldsymbol{\lambda}$ in (7)), but most importantly, note that all intractable normalising constants cancel above and below the fraction. Hence, the acceptance probability can be evaluated, unlike in the case of a standard Metropolis-Hastings scheme. In practice, the exchange move proposes to offer the observed data \mathbf{x} to the auxiliary parameter $\boldsymbol{\lambda}'$ and similarly to offer the auxiliary data \mathbf{y} the parameter $\boldsymbol{\lambda}$.

4 Applications

4.1 Artificial Data

We illustrate the proposed algorithm to sample from the posterior distribution of $\boldsymbol{\lambda}$ using artificial data.

Dataset 1

Consider a sample of $n = 100$ unit vectors $(\mathbf{x}_1, \dots, \mathbf{x}_{100})$ which result in the pair of sufficient statistics $(\tau_1, \tau_2) = (0.30, 0.32)$. We assign independent Exponential prior distributions with rate 0.01 (i.e. mean 100) to the parameters of interest λ_1 and λ_2 subject to the constraint that $\lambda_1 \geq \lambda_2$; note that we also implicitly assume that $\lambda_1 \geq \lambda_2 \geq \lambda_3 = 0$. We implemented the algorithm which was described in Section 3.2.1. The parameters were updated in blocks by proposing a candidate vector from a bivariate Normal distribution with mean the current values of the parameters and variance-covariance matrix σI , where I is the identity matrix and the samples were thinned, keeping every 10th value. Convergence was assessed by visual inspection of the Markov chains and we found that by using $\sigma = 1$ the mixing was good and achieved an acceptance rate between 25% and 30%. Figure 1 shows a scatter plot of the sample from the joint posterior distribution (left panel) whilst the marginal posterior densities for λ_1 and λ_2 are shown in the top row of Figure 2. The autocorrelation function (ACF) plots, shown in the top row of Figure 3 reveal good mixing properties of the MCMC algorithm and by (visual inspection) appears to be much better than the algorithm proposed by Walker (2013, Figure 1). Mardia and Zemroch (1977) report maximum likelihood estimates of $\hat{\lambda}_1 = 0.588$, $\hat{\lambda}_2 = 0.421$, with which our results broadly agree. Although in principle one can derive (approximate) confidence intervals based on some regularity conditions upon which it can be proved that the MLEs are (asymptotically) Normally distributed, an advantage of our (Bayesian) approach is that it allows quantification of the uncertainty of the parameters of interest in a probabilistic manner.

Dataset 2

We now consider an artificial dataset of 100 vectors which result in the pair of sufficient statistics $(\tau_1, \tau_2) = (0.02, 0.40)$ for which the maximum likelihood estimates are $\hat{\lambda}_1 = 25.31$, $\hat{\lambda}_2 = 0.762$ as reported by Mardia and Zemroch (1977). We implement the proposed algorithm by assigning the same prior distributions to λ_1 and λ_2 as for the Dataset 1. A scatter plot of a sample from the joint posterior distribution is shown in Figure 1, showing that our approach gives results which are consistent with the MLEs. This examples shows that our algorithm performs well even when $\lambda_1 \gg \lambda_2$.

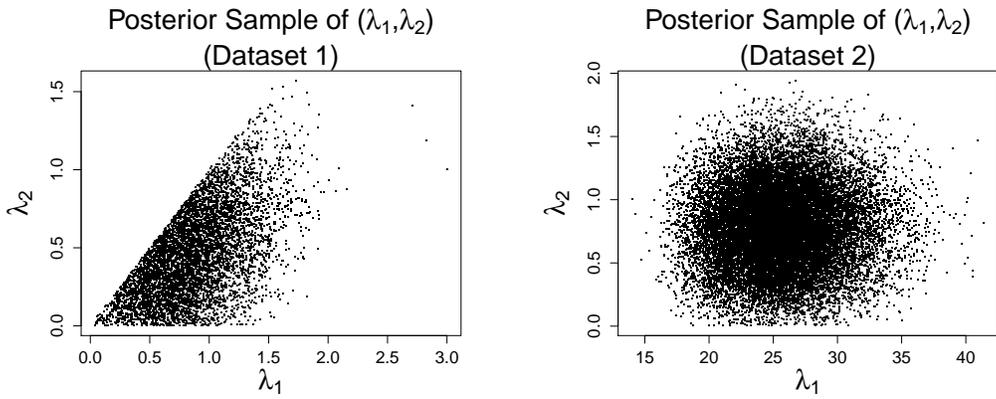


Figure 1: Sample from the joint posterior distribution of λ_1 and λ_2 for Dataset 1 (left) and Dataset 2 (right) as described in Section 4.

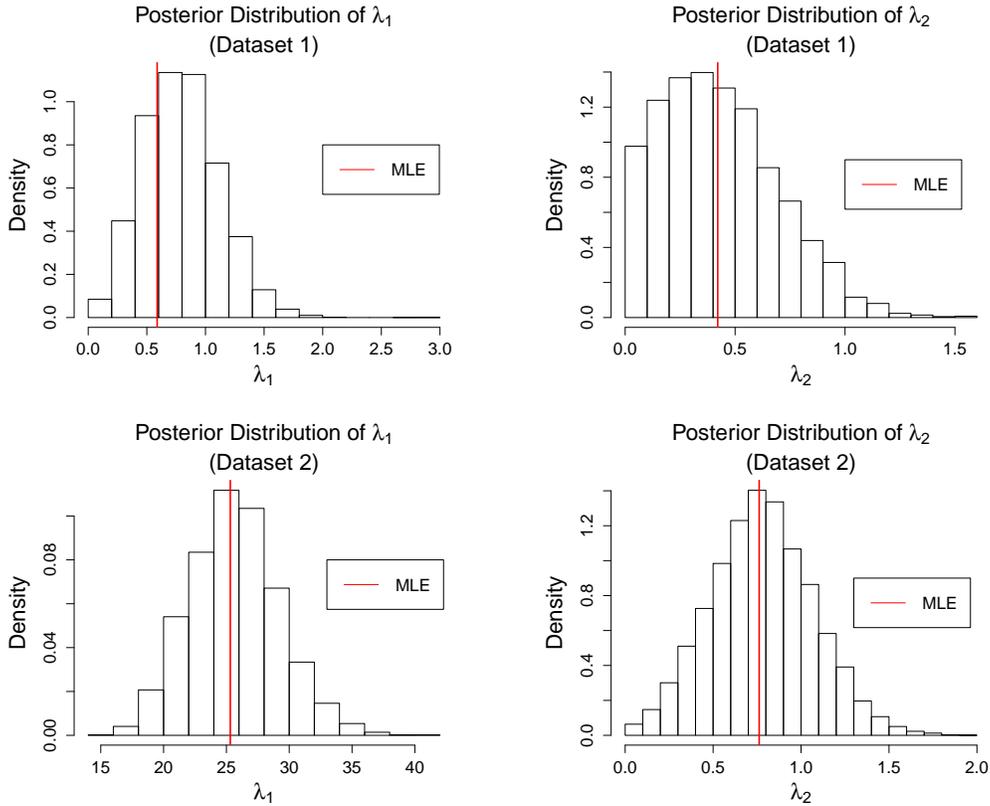


Figure 2: Marginal posterior densities and ACFs for λ_1 and λ_2 for Dataset 1 (top) and Dataset 2 (bottom) in Section 4.

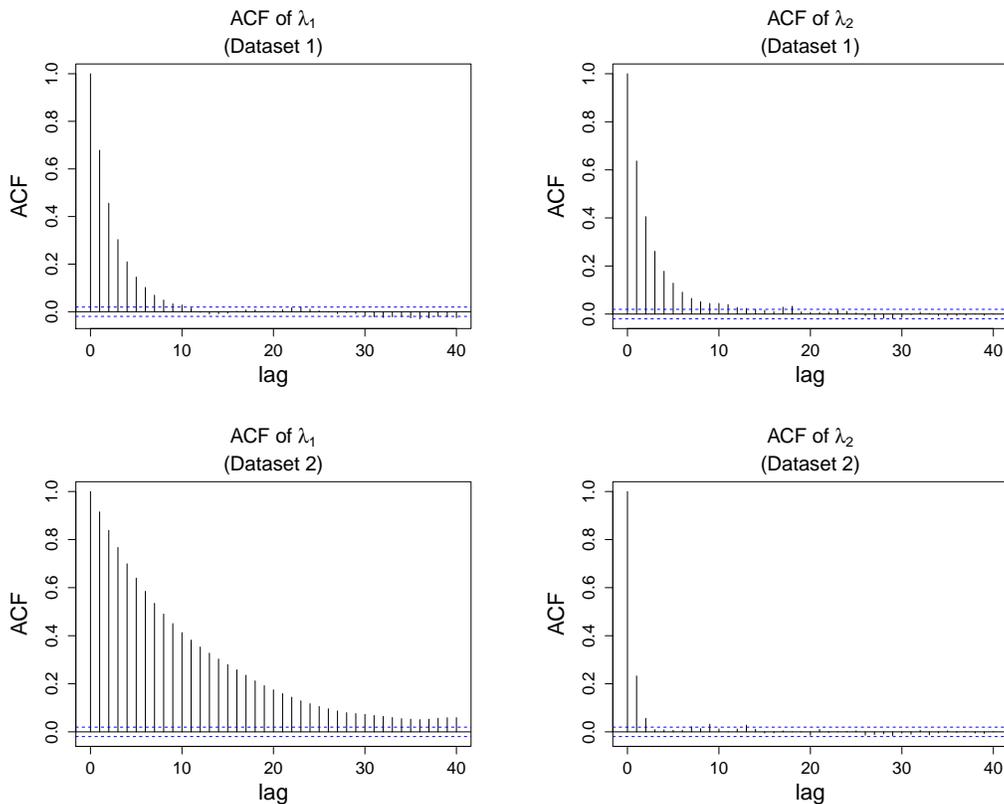


Figure 3: ACFs for λ_1 and λ_2 for Dataset 1 (top) and Dataset 2 (bottom) in Section 4.

4.2 Earthquake data

As an illustration of an application to real data, we consider an analysis of earthquake data recently analysed by Arnold and Jupp (2013). An earthquake gives rise to three orthogonal axes, and geophysicists are interested in analysing such data in order to compare earthquakes at different locations and/or at different times. An earthquake gives rise to a pair of orthogonal axes, known as the compressional (P) and tensional (T) axes, from which a third axis, known as the null (A) axis is obtained via $A = P \times T$. Each of these quantities are determined only up to sign, and so models for axial data are appropriate. The data can be treated as orthogonal axial 3-frames in \mathbb{R}^3 and analysed accordingly, as in Arnold and Jupp (2013), but we will illustrate our method using the A axes only. In general, an orthogonal axial r -frame in \mathbb{R}^p , $r \leq p$, is an ordered set of r axes, $\{\pm u_1, \dots, \pm u_r\}$, where u_1, \dots, u_r are orthonormal vectors in \mathbb{R}^p (Arnold and Jupp; 2013). The more familiar case of data on the sphere \mathcal{S}^2 is the special case corresponding to $p = 3, r = 1$, which is the case we consider here.

The data consist of three clusters of observations relating to earthquakes in New Zealand. The first two clusters each consist of 50 observations near Christchurch which took place before and after a large earthquake on 22 February 2011, and we will label these two clusters CCA and CCB respectively. For these two clusters, the P and T axes are quite highly concentrated in the horizontal plane, and as a result the majority of the A axes are concentrated about the vertical axis. It is of interest to geophysicists to establish whether there is a difference between

the pattern of earthquakes before and after the large earthquake. The third cluster is a more diverse set of 32 observations obtained from earthquakes in the north of New Zealand’s South Island, and we will label this cluster SI. We will illustrate our method by fitting Bingham models to the A axes from each of the individual clusters and considering the posterior distributions of the Bingham parameters. We will denote the parameters from the CCA, CCB and SI models as λ_i^A , λ_i^B and λ_i^S respectively, $i = 1, 2$.

The observations for the two clusters of observations near Christchurch yield sample data of $(\tau_1^A, \tau_2^A) = (0.1152360, 0.1571938)$ for CCA and $(\tau_1^B, \tau_2^B) = (0.1127693, 0.1987671)$ for CCB. The data for the South Island observations are $(\tau_1^S, \tau_2^S) = (0.2288201, 0.3035098)$. We fit each dataset separately by implementing the proposed algorithm. Exponential prior distributions to all parameters of interest (mean 100) were assigned, subject to the constraint that $\lambda_1^j \geq \lambda_2^j$ for $j = A, B, S$. Scatter plots from the joint posterior distributions of the parameters from each individual analysis are shown in Figure 4. The plots for CCA and CCB look fairly similar, although λ_2 is a little lower for the CCB cluster. The plot for SI cluster suggests that these data are somewhat different.

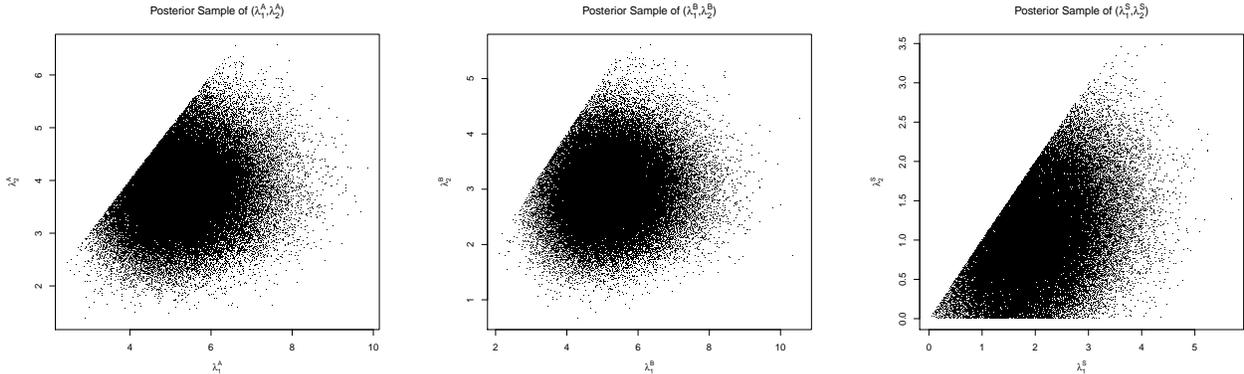


Figure 4: Posterior samples for differences in λ_1 and λ_2 for the two sets of Christchurch data (left) and South Island and Christchurch data A (right). This shows a clear difference between the South Island and Christchurch data, but suggests no difference between the two sets of Christchurch data.

To establish more formally if there is any evidence of a difference between the two Christchurch clusters, we consider the bivariate quantity $(\lambda_1^A - \lambda_1^B, \lambda_2^A - \lambda_2^B)$. If there is no difference between the two clusters, then this quantity should be $(0, 0)$. In Figure 5 (left panel), we show the posterior sample of this quantity, and a 95% probability region obtained by fitting a bivariate normal distribution with parameters estimated from this sample. The origin is contained comfortably within this region, suggesting there is no real evidence for a difference between the two clusters. Arnold and Jupp (2013) obtained a p -value of 0.890 from a test of equality for the two populations based on treating the data as full axial frames, and our analysis on the A axes alone agrees with this.

The right panel of Figure 5 shows a similar plot for the quantity $(\lambda_1^A - \lambda_1^S, \lambda_2^A - \lambda_2^S)$. Here, the origin lies outside the 95% probability region, suggesting a difference between the first

Christchurch cluster and the South Island cluster. Arnold and Jupp (2013) give a p -value of less than 0.001 for equality of the two populations, so again our analysis on the A axes agrees with this.

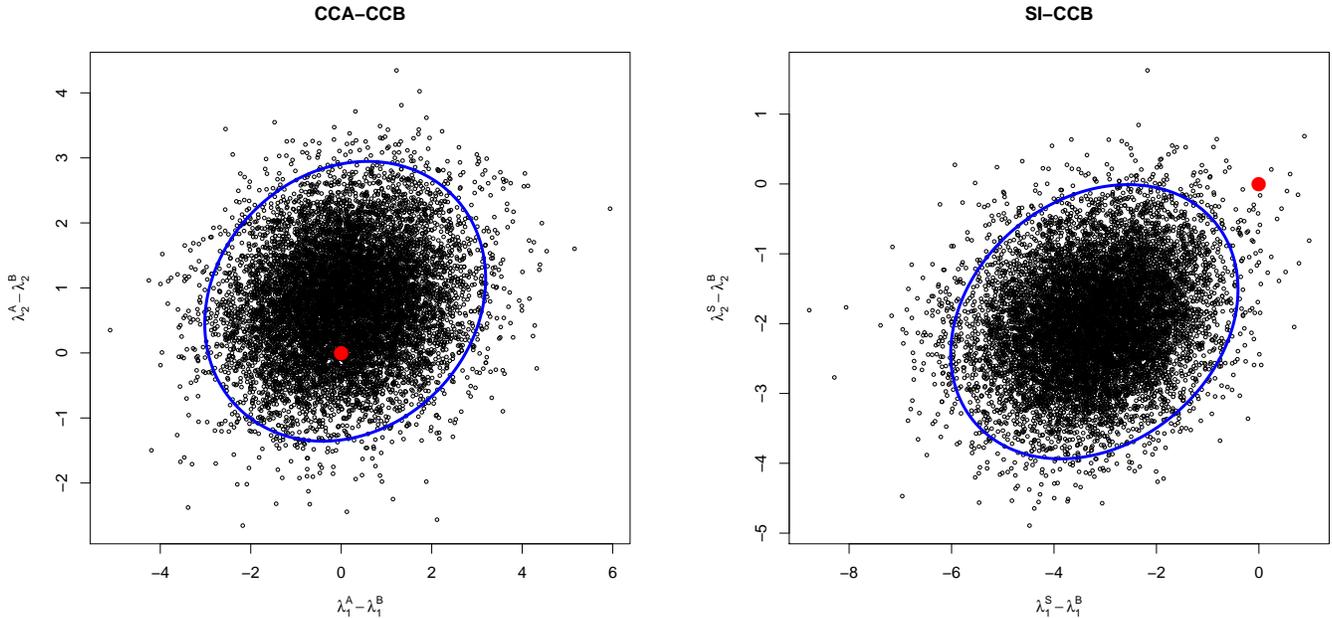


Figure 5: Posterior samples for differences in λ_1 and λ_2 for the two sets of Christchurch data (left) and South Island and Christchurch data A (right). This shows a clear difference between the South Island and Christchurch data, but suggests no difference between the two sets of Christchurch data.

5 Discussion

There is a growing area of applications that require inference over doubly intractable distributions including directional statistics, social networks (Caimo and Friel; 2011), latent Markov random fields (Everitt; 2012), and large-scale spatial statistics (Aune et al.; 2012) to name but a few. Most conventional inferential methods for such problems relied on approximating the normalising constant and embedded the latter into a standard MCMC algorithm (e.g. Metropolis-Hastings). Such approaches not only are only approximate in the sense that the target distribution is an approximation to the true posterior distribution of interest, but they can also suffer from being very computationally intensive. It is only until fairly recently that algorithms which avoid the need of approximating/evaluating the normalising constant became available; see Møller et al. (2006); Murray et al. (2006); Walker (2011); Girolami et al. (2013).

In this paper we were concerned with exact Bayesian inference for the Bingham distribution which has been a difficult task so far. We proposed an MCMC algorithm which allows us to draw samples from the posterior distribution of interest without having to approximate this constant. We have shown that the MCMC scheme is i) fairly straightforward to implement, ii)

mixes very well in a relatively short number of sweeps and iii) does not require the specification of good guesses of the unknown parameters. We have applied our method to both real and simulated data, and showed that the results agree with maximum likelihood estimates for the parameters. However, we believe that a fully Bayesian approach has the benefit of providing an honest assessment of the uncertainty of the parameter estimates and allows exploration of any non-linear correlations between the parameters of interest. In comparison to the approach recently proposed by Walker (2013) (which also avoids approximating the normalising constant) we argue that our algorithm is easier to implement, runs faster and the Markov chains appear to mix better.

In terms of the further computational aspects, our algorithm is not computationally intensive and this is particularly true for the number of dimensions that are commonly met in practice (e.g. $q = 3$). For all the results presented here, we ran our MCMC chains for 10^6 iterations for each of the simulated and real data examples, which we found to be sufficient for good mixing in all cases. Our method was implemented in C++ and each example took between 20 and 30 seconds on a desktop PC with 3.1GHz processor¹; note, that is considerably faster than the algorithm proposed by Walker (2013) in which “running 10^5 iterations takes a matter of minutes on a standard laptop”. In general the time taken for our proposed algorithm will depend on the number of auxiliary data points n that need to be simulated, as well as the efficiency of the underlying rejection algorithm for the particular parameter values at each iteration. In addition, the efficiency of the rejection algorithm is likely to deteriorate as the dimension q increases, but we found it to be very efficient for all our examples and it is reasonably efficient for at least a moderate number of dimensions according to simulations by Ganeiber (2012).

Statistical inference, in general, is not limited to parameter estimation. Therefore, a possible direction for future research within this context is to develop methodology to enable calculation of the model evidence (marginal likelihood). This quantity is vital in Bayesian model choice and knowledge of it will allow a formal comparison between competing models for a given dataset such as the application presented in Section 4.2.

Acknowledgements

The authors are most grateful to Richard Arnold and Peter Jupp for providing the earthquake data and John Kent for providing a Fortran program to compute moments of the Bingham distribution. Finally, we would like to thank Ian Dryden for commenting on an earlier draft of this manuscript.

References

- Arnold, R. and Jupp, P. E. (2013). Statistics of orthogonal axial frames, *Biometrika* **100**(3): 571–586.
- Aune, E., Simpson, D. P. and Eidsvik, J. (2012). Parameter estimation in high dimensional gaussian distributions, *Statistics and Computing* pp. 1–17.

¹Our code is available upon request.

- Bingham, C. (1974). An antipodally symmetric distribution on the sphere, *Ann. Statist.* **2**: 1201–1225.
- Boomsma, W., Mardia, K., Taylor, C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure., *Proc Natl Acad Sci U S A* **105**(26): 8932–7.
- Caimo, A. and Friel, N. (2011). Bayesian inference for exponential random graph models, *Social Networks* **33**(1): 41–55.
- Ehler, M. and Galanis, J. (2011). Frame theory in directional statistics, *Statistics and Probability Letters* **81**(8): 1046–1051.
- Everitt, R. G. (2012). Bayesian parameter estimation for latent markov random fields and social networks, *Journal of Computational and Graphical Statistics* **21**(4): 940–960.
- Ganeiber, A. M. (2012). *Estimation and simulation in directional and statistical shape models*, PhD thesis, University of Leeds.
- Girolami, M., Lyne, A.-M., Strathmann, H., Simpson, D. and Atchade, Y. (2013). Playing russian roulette with intractable likelihoods, *ArXiv preprint; arXiv:1306.4032* .
- Kent, J. T. (1987). Asymptotic expansions for the Bingham distribution, *J. Roy. Statist. Soc. Ser. C* **36**(2): 139–144.
URL: <http://dx.doi.org/10.2307/2347545>
- Kent, J. T. and Ganeiber, A. M. (2012). Simulation of the Bingham distribution in directional statistics, in K. V. Mardia, A. Gusnato, A. D. Riley and J. Voss (eds), *New Statistics and Modern Natural Sciences*, Leeds University Press, Leeds, pp. 57–59.
- Kume, A. and Walker, S. G. (2006). Sampling from compositional and directional distributions, *Stat. Comput.* **16**(3): 261–265.
URL: <http://dx.doi.org/10.1007/s11222-006-8077-9>
- Kume, A. and Wood, A. T. A. (2005). Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants, *Biometrika* **92**(2): 465–476.
URL: <http://dx.doi.org/10.1093/biomet/92.2.465>
- Kume, A. and Wood, A. T. A. (2007). On the derivatives of the normalising constant of the Bingham distribution, *Statist. Probab. Lett.* **77**(8): 832–837.
URL: <http://dx.doi.org/10.1016/j.spl.2006.12.003>
- Levine, J. D., Funes, P., Dowse, H. B. and Hall, J. C. (2002). Resetting the circadian clock by social experience in drosophila melanogaster, *Science Signaling* **298**(5600): 2010–2012.
- Mardia, K. V. and Jupp, P. E. (2000). *Directional statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester. Revised reprint of it Statistics of directional data by Mardia [MR0336854 (49 #1627)].
- Mardia, K. V. and Zemroch, P. J. (1977). Table of maximum likelihood estimates for the bingham distribution, *Statist. Comput. Simul.* **6**: 29–34.

- Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants, *Biometrika* **93**(2): 451–458.
URL: <http://dx.doi.org/10.1093/biomet/93.2.451>
- Murray, I., Ghahramani, Z. and MacKay, D. J. C. (2006). MCMC for doubly-intractable distributions, *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, AUAI Press, pp. 359–366.
- Ripley, B. D. (1987). *Stochastic simulation*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York.
URL: <http://dx.doi.org/10.1002/9780470316726>
- Rueda, C., Fernández, M. A. and 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, *J. Amer. Statist. Assoc.* **104**(485): 338–347.
URL: <http://dx.doi.org/10.1198/jasa.2009.0120>
- Sei, T. and Kume, A. (2013). Calculating the normalising constant of the bingham distribution on the sphere using the holonomic gradient method, *Statistics and Computing* pp. 1–12.
URL: <http://dx.doi.org/10.1007/s11222-013-9434-0>
- Tyler, D. E. (1987). Statistical analysis for the angular central Gaussian distribution on the sphere, *Biometrika* **74**(3): 579–589.
URL: <http://dx.doi.org/10.1093/biomet/74.3.579>
- Walker, S. G. (2011). Posterior sampling when the normalizing constant is unknown, *Comm. Statist. Simulation Comput.* **40**(5): 784–792.
URL: <http://dx.doi.org/10.1080/03610918.2011.555042>
- Walker, S. G. (2013). Bayesian estimation of the Bingham distribution, *BJPS* . To appear.