Statistics and Computing manuscript No.
(will be inserted by the editor)

Exact Bayesian Inference for the Bingham Distribution

Christopher J. Fallaize · Theodore Kypraios

Received: date / Accepted: date

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

Keywords

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

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

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 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),

Christopher J. Fallaize · Theodore Kypraios
School of Mathematical Sciences, University Park,University of Nottingham
Nottingham, NG7 2RD, United Kingdom
E-mail: theodore.kypraios@nottingham.ac.uk

S. Author
second address

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 et al. (2013).

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

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

Therefore, the algorithm will work efficiently provided that M is small or, in other words, the probability of acceptance $(1/M)$ is large. We should 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 which depends on the normalising constant of the likelihood function and cannot be computed).

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 $\boldsymbol{y}$ has a multivariate Normal distribution in $\mathbb{R}^q$ with mean $\mathbf{0}$ and variance covariance matrix Ψ, then the vector $\boldsymbol{x} = \boldsymbol{y}/||\boldsymbol{y}||$ 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

$$\begin{aligned} g(\boldsymbol{x}; \Psi) &= w_q^{-1} |\Psi|^{-1/2} \left(\boldsymbol{x}^T \Psi^{-1} \boldsymbol{x}\right)^{-q/2} \\ &= c_{\mathrm{ACG}}(\Psi) g^*(\boldsymbol{x}; \Psi) \end{aligned}$$

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

2.3 Rejection Sampling for the Bingham Distribution

Kent et al. (2013) 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, they proposed the following algorithm to simulate a value from the Bingham distribution with parameter matrix A:

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

Here, $f^*(\boldsymbol{y};A) = \exp(-\boldsymbol{y}^T A \boldsymbol{y})$ and $g^*(\boldsymbol{y};\Psi) = (\boldsymbol{y}^T \Psi^{-1} \boldsymbol{y})^{-\frac{q}{2}}$, 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 $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_q)$, then it can be shown that if $\boldsymbol{x}$ is drawn from a distribution with probability density function $f(\boldsymbol{x}; A)$, the corresponding random vector $\boldsymbol{y} = X^T V$ is drawn from a distribution with density $f(\boldsymbol{x}; \Lambda)$ (see, for example, Kume and Walker 2006; Kume and Wood 2007). Therefore, without loss of generality, we assume that $A = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_q)$. Moreover, to ensure identifiability, we assume that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_q = 0$ (Kent 1987). We discuss in Section 3.3 how one can draw inference for an arbitrary symmetric matrix A which may not necessarily be diagonal. In the case where $A = \Lambda$ the probability density function becomes

$$f(\boldsymbol{x}; \Lambda) = \frac{\exp\left\{ -\sum_{i=1}^{q-1} \lambda_i x_i^2 \right\}}{c(\Lambda)} \tag{3}$$

with respect to a uniform measure on the sphere and

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

Suppose $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is a sample of unit vectors in $\mathcal{S}^{q-1}$ from the Bingham distribution with density (3). Then the likelihood function is given by

$$\begin{aligned} L(\Lambda) &= \frac{1}{c(\Lambda)^n} \exp\left\{ -\sum_{i=1}^{q-1} \lambda_i \sum_{j=1}^{n} \left(x_j^i\right)^2 \right\} \\ &= \frac{1}{c(\Lambda)^n} \exp\left\{ -n \sum_{i=1}^{q-1} \lambda_i \tau_i \right\}, \end{aligned} \tag{4}$$

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

3.2 Bayesian Inference

We are interested in drawing Bayesian inference for the matrix Λ, or equivalently, for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{q-1})$. The likelihood function in (4) 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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{q-1}$ then the density of the posterior distribution of Λ up to proportionality given the data is as follows:

$$\begin{aligned} \pi(\boldsymbol{\lambda} | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto L(\Lambda) &\prod_{i=1}^{q-1} \exp\{-\lambda_i \mu_i\} \\ &\times \mathbf{1}\left(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{q-1}\right) \\ &= \frac{1}{c(\Lambda)^n} \exp\left\{ -\sum_{i=1}^{q-1} \lambda_i (n\tau_i + \mu_i) \right\} \\ &\times \mathbf{1}\left(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{q-1}\right). \end{aligned} \tag{5}$$

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

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

Note that $N_{q-1}(\boldsymbol{m}, S)$ denotes the density of a multivariate Normal distribution with mean vector $\boldsymbol{m}$ and variance-covariance matrix S. Step 2 of the above algorithm requires the evaluation of the ratio $\pi(\boldsymbol{\lambda}^{\text{can}}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) / \pi(\boldsymbol{\lambda}^{\text{cur}}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, 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 $\pi(\boldsymbol{\lambda}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, 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 cleverly chosen variables in to the Metropolis-Hastings (M-H) algorithm such that the normalising constants cancel in the M-H ratio. Consider augmenting the observed data $\boldsymbol{x}$ with auxiliary data $\boldsymbol{y}$, which has the same state space as $\boldsymbol{x}$, leading to the joint distribution

$$\pi(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\lambda}) = \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda})\pi(\boldsymbol{x}, \boldsymbol{\lambda}),$$

where $\pi(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}|\boldsymbol{\lambda})\pi(\boldsymbol{\lambda}) \propto \pi(\boldsymbol{\lambda}|x)$, the target posterior distribution of interest. There is freedom of choice in the conditional density $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda})$. For example, this could be the same density as that of the data $\boldsymbol{x}$, evaluated at some fixed value of $\boldsymbol{\lambda}$, $\hat{\boldsymbol{\lambda}}$ say. This would require a good representative value of $\boldsymbol{\lambda}$, obtained for example from a pseudo-likelihood estimator, for good performance. Note also that it is necessary to store values of the auxiliary variables from one iteration to the next.

A simpler version that avoids having to specify the conditional density of the auxiliary variables 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. Proposals $\boldsymbol{\lambda}'$ are drawn from a density $h(\cdot|\boldsymbol{\lambda})$, although in general this density does not have to depend on the current state of $\boldsymbol{\lambda}$. For example, random walk proposals centred at $\boldsymbol{\lambda}$ or independence sampler proposals could be used. The algorithm proceeds as follows:

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

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

where $f^*(\boldsymbol{x}; A) = \exp(-\boldsymbol{x}^T A \boldsymbol{x})$ is the unnormalized Bingham density as previously and f is the normalized density. Under this scheme, the marginal distribution of $\boldsymbol{\lambda}$ is the target posterior distribution of interest, but crucially, 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. We are thus able to draw samples from our target posterior distribution, provided we can simulate exactly from the Bingham distribution. The exchange move can be interpreted as offering the observed data $\boldsymbol{x}$ to the proposed parameter $\boldsymbol{\lambda}'$ and similarly to offer the auxiliary data $\boldsymbol{y}$ the parameter $\boldsymbol{\lambda}$.

This algorithm, due to Murray et al. (2006), is a special case of a more general algorithm which draws two sets of new auxiliary variables $\boldsymbol{y}$ and $\boldsymbol{y}'$ at each iteration (Storvik 2011). It is the special case where only proposals such that $\boldsymbol{y} = \boldsymbol{y}'$ have non-zero probability, so that only one set of auxiliary variables need to be sampled. In addition to removing the need to store the auxiliary variables from the previous iteration, this choice also negates the need to specify $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda}')$, which is needed in both the algorithm of Møller et al. (2006) and the general algorithm just mentioned.

3.3 Bayesian Inference for an Arbitrary Symmetric Matrix A

Following Walker (2013) we have thus far assumed that the matrix A in Equation 1 is diagonal. However, when dealing with real datasets this may not necessarily the case. Therefore, we now describe how to draw Bayesian inference for a general symmetric matrix A, given data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ which are drawn from the corresponding Bingham distribution.

Denote by a_{ij} the elements of (the random) matrix A where $a_{ij} = a_{ji}$ for $i, j = 1, \ldots, q$. The likelihood function is given by $L(A) = \exp\left(-\sum_{i=1}^n \boldsymbol{x}_i^T A \boldsymbol{x}_i\right)/c(A)^n$. We assign independent Normal prior distributions to each a_{ij} for $i \geq j$ with mean zero and variance v and hence the posterior distribution has density proportional to

$$\frac{\exp\left\{-\sum_{i=1}^n \boldsymbol{x}_i^T A \boldsymbol{x}_i\right\} \exp\left\{-\frac{1}{2v}\boldsymbol{a}^T\boldsymbol{a}\right\}}{c(A)^n}, \tag{6}$$

where $\boldsymbol{a} = (a_1, \ldots, a_{n_a})$ is the vector of distinct elements of A, of which there are $n_a = q + \binom{q}{2}$.

To perform inference for A, based on a sample of n points assumed to be from a Bingham distribution, we can again apply the exchange algorithm. Suppose the current value of A is A^{cur}, with elements $\boldsymbol{a}^{\text{cur}}$. We first propose a move to a candidate value of A, A^{can} (with elements $\boldsymbol{a}^{\text{can}}$) , drawn from some density $h(\cdot|A^{\text{cur}})$. We then sample auxiliary data $\boldsymbol{y}$, by sampling n data points from the Bingham density (1) with parameter A^{can}. In this paper, we use random walk Metropolis updates for A, by drawing candidate values for the elements of A^{can} from a multivariate normal distribution with mean vector $\boldsymbol{a}^{\text{cur}}$ and covariance matrix $\sigma_a^2 I$. We find this satisfactory for our work here, but more elaborate proposals could be constructed, such as an independence sampler which draws candidate values from a good approximation to the posterior distribution of A; we do not consider such proposals further in this paper.

In the general case, we can write $A = V\Lambda V^T$, where Λ is the diagonal matrix of parameter values considered previously, and V is an orthogonal matrix. Thus, inference for Λ can be performed by decomposing each sampled value of A in this manner, but we now also obtain posterior samples for the orthogonal component V. (Previously, when A was assumed to be diagonal, we simply had $V = I$.) Again, for identifiability, we apply the constraint $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q = 0$.

3.3.1 Example

To illustrate inference for general A, we apply our method to the data of Bingham (1974). The data consist of $n = 150$ measurements of a certain axis of interest relating to calcite grains from the Taconic Mountains of New York state, and yield a moment of inertia (or sum of squares and products) matrix

$$T = \sum_{i=1}^{150} \boldsymbol{x}_i \boldsymbol{x}_i^T = \begin{pmatrix} 76.5575 & 18.2147 & 12.2406 \\ 18.2147 & 46.7740 & 6.8589 \\ 12.2406 & 6.8589 & 26.6670 \end{pmatrix}.$$

We used the settings $v = 100$ and $\sigma_a^2 = 0.04$, and base inference on a sample of $N = 100000$ values from the posterior distribution (6). This value of σ_a^2 was found by experimentation to result in good mixing properties for the chain in this example, and altering σ_a^2 can affect the mixing and increase autocorrelation in the chain. As mentioned previously, more elaborate proposals could be constructed, but this is not pursued further here. For the diagonal component Λ, we obtain posterior median values $\lambda_1 = 3.631$ and $\lambda_2 = 1.963$, compared with maximum likelihood estimates of $\hat{\lambda}_1 = 3.518$ and $\hat{\lambda}_2 = 1.956$. For the orthogonal component of A, we have samples $V_1, \ldots, V_N$, obtained from the spectral decomposition of each sampled matrix A_i, $i = 1, \ldots, N$. To obtain a summary value, we first form the element-wise sample mean of V, $\bar{V} = \frac{1}{N}\sum_{i=1}^N V_i$. We then decompose this into $\bar{V} = \hat{V}K$, where $K = (\bar{V}^T\bar{V})^{1/2}$, the positive definite square root of $\bar{V}^T\bar{V}$, and $\hat{V}$ is an orthogonal matrix known as the polar part of $\bar{V}$ (Mardia and Jupp 2000); $\hat{V}$ is then our estimate of the orthogonal component V. We obtain

$$\hat{V} = \begin{pmatrix} 0.1795 & -0.4404 & 0.8797 \\ 0.1394 & 0.8966 & 0.4204 \\ -0.9738 & 0.0472 & 0.2223 \end{pmatrix}.$$

The corresponding maximum likelihood estimate is

$$\begin{pmatrix} -0.1723 & -0.4439 & 0.8794 \\ -0.1516 & 0.8940 & 0.4216 \\ 0.9733 & 0.0606 & 0.2213 \end{pmatrix}$$

(Bingham 1974), the columns of which are the eigenvectors of T. Again, our estimates agree closely (up to sign, since we have taken V to have determinant $+1$).

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 $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{100})$ 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) appear to be much better than those shown in 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 Dataset 1. A scatter plot of a sample from the joint posterior distribution is shown in Figure 1 (right panel), showing that our approach gives results which are consistent with the MLEs. Marginal posterior densities for λ_1 and λ_2 are shown in the bottom row of Figure 2, and ACF plots are shown in the bottom row of Figure 3. This example shows that our algorithm performs well when $\lambda_1 >> \lambda_2$ as well as when the difference between λ_1 and λ_2 is much smaller, as was the case in Dataset 1.

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 (B) axis is obtained via $B = P \times T$. (Arnold and Jupp (2013) label this the A axis, but we have used A for the parameter of the Bingham distribution.) 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 B axes only. In general, an orthogonal axial r-frame in $\mathbb{R}^p$, $r \leq p$, is an ordered set of r axes, $\{\pm u_1, \ldots, \pm u_r\}$, where $u_1, \ldots, u_r$ are orthonormal vectors in $\mathbb{R}^p$ (Arnold and Jupp 2013). The 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 B 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

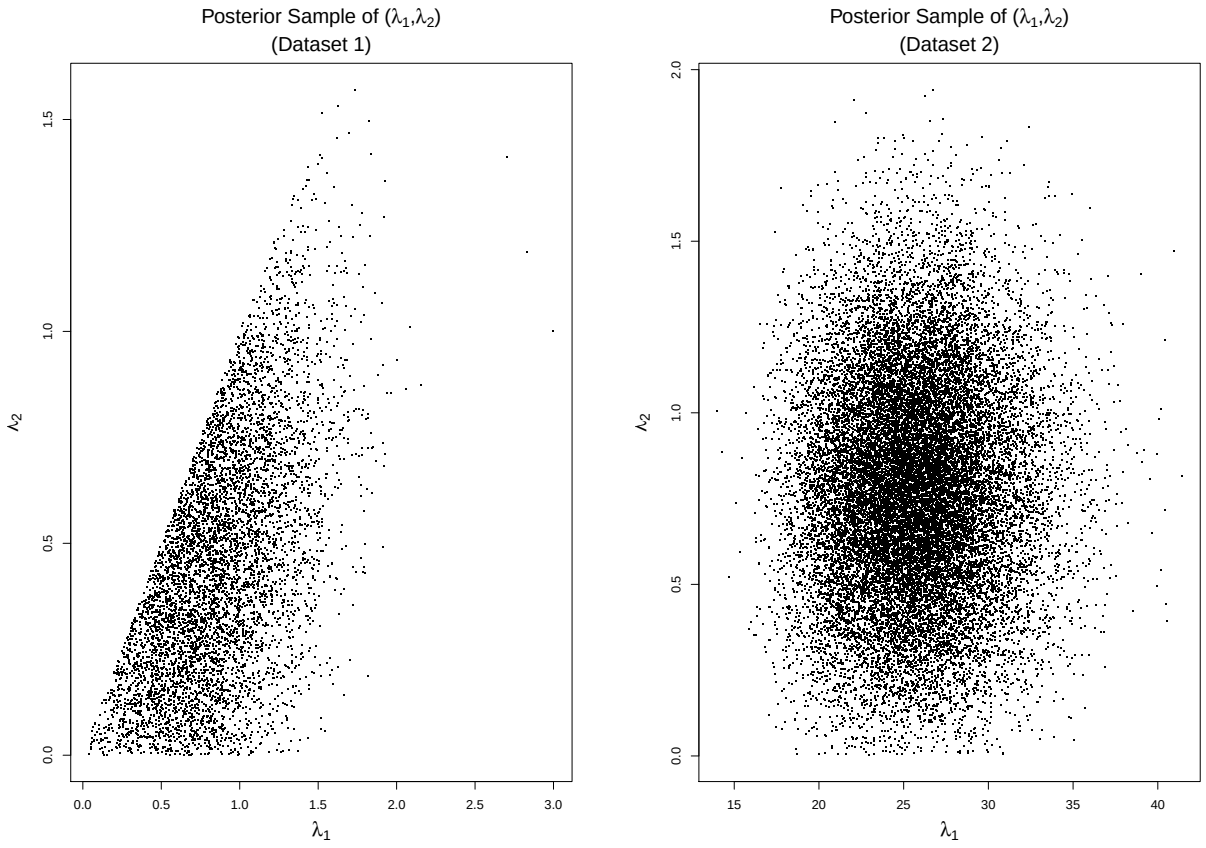

Fig. 1 Sample from the joint posterior distribution of λ_1 and λ_2 for Dataset 1 (left) and Dataset 2 (right) as described in Section 4.

our method by fitting Bingham models to the B axes from each of the individual clusters and considering the posterior distributions of the parameters of the diagonal component of the Bingham parameter matrix. We will denote these 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.

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 B 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 popu-

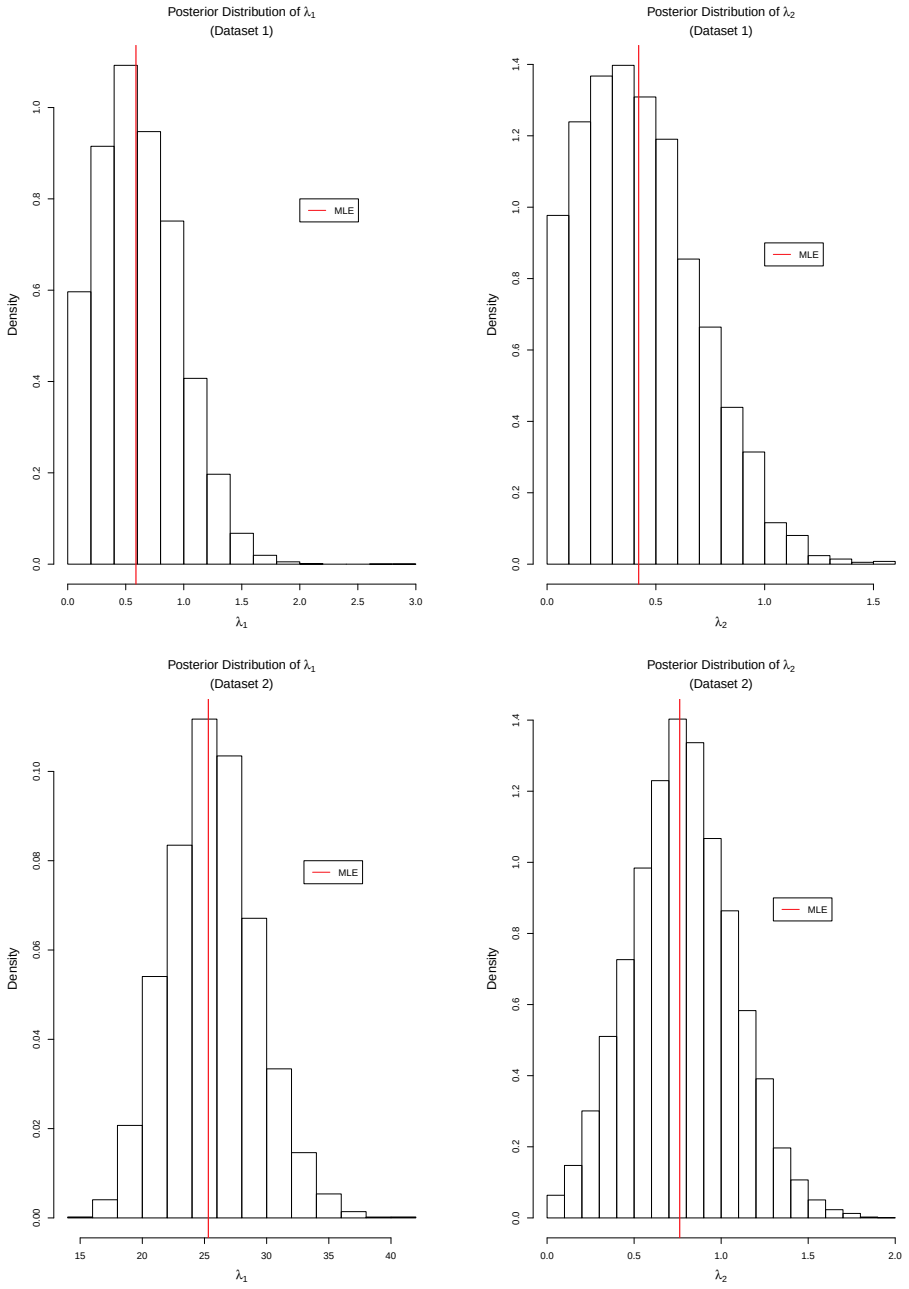

Fig. 2 Marginal posterior densities for λ_1 and λ_2 for Dataset 1 (top) and Dataset 2 (bottom) in Section 4.

lations, so again our analysis on the A axes agrees with this.

4.2.1 Inference for full A

As well as testing for differences between the samples, it is of interest to determine whether the B axes are vertical, since the observed P and T axes lie approximately in the horizontal plane. For the CCA cluster, our analysis yields an estimate of the dominant eigenvector of A of $(-0.0010, -0.0530, 0.9985)$, suggesting the B axes are close to vertical, or equivalently that the P and T axes are confined to the horizontal plane. For the CCB cluster, the estimate of the dominant eigenvector of A is $(-0.0442, -0.0119, 0.9990)$, which is again close to the vertical.

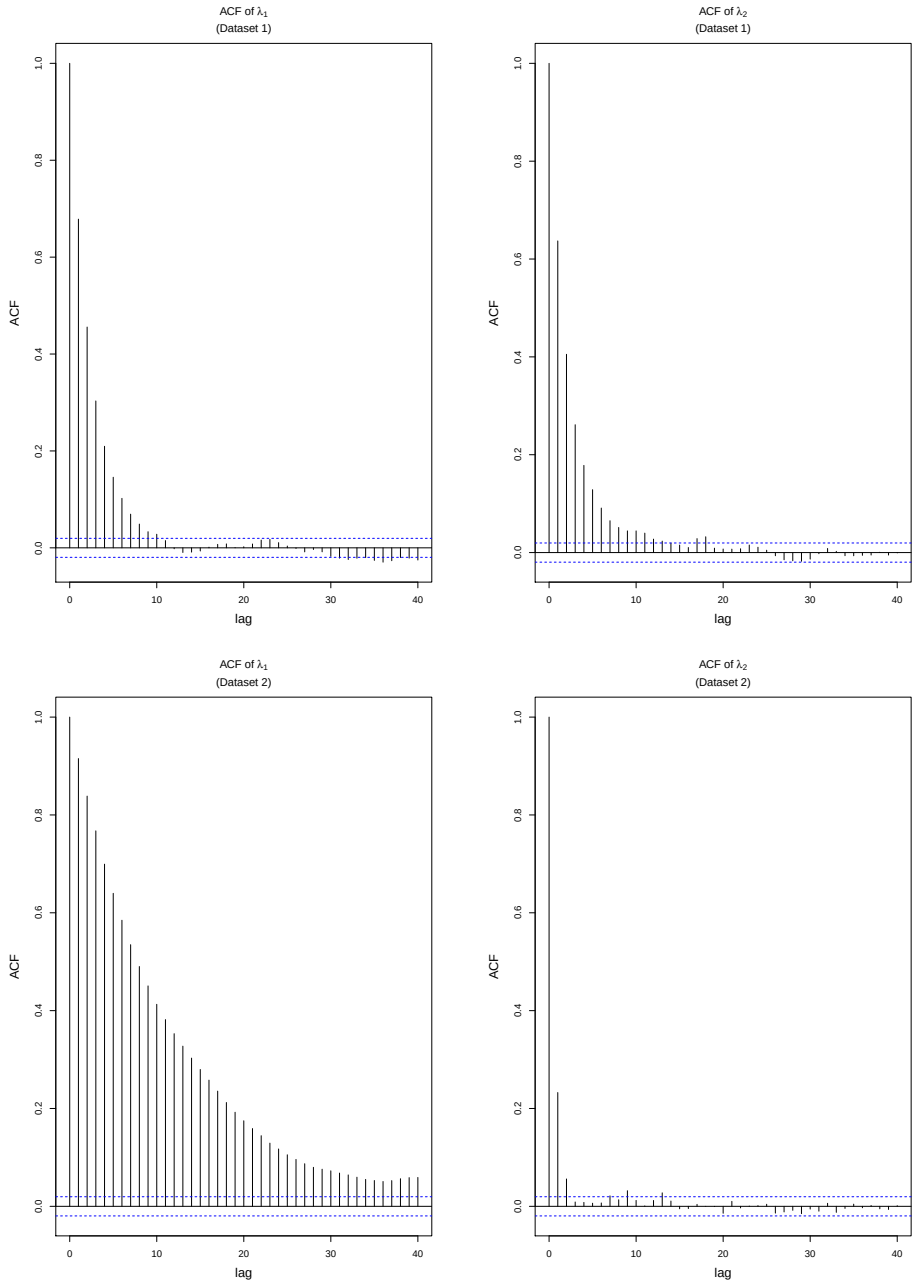

Fig. 3 ACFs for λ_1 and λ_2 for Dataset 1 (top) and Dataset 2 (bottom) in Section 4.

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 are not 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 con-

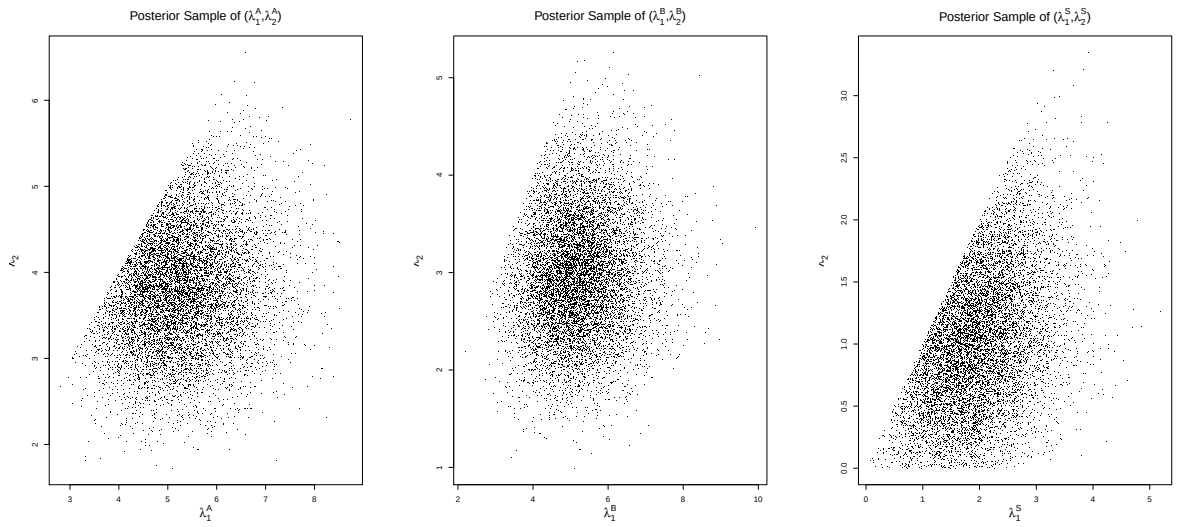

Fig. 4 Posterior samples for differences in λ_1 and λ_2 for the two sets of Christchurch data (left) and South Island and Christchurch data A (right). This shows a clear difference between the South Island and Christchurch data, but suggests no difference between the two sets of Christchurch data.

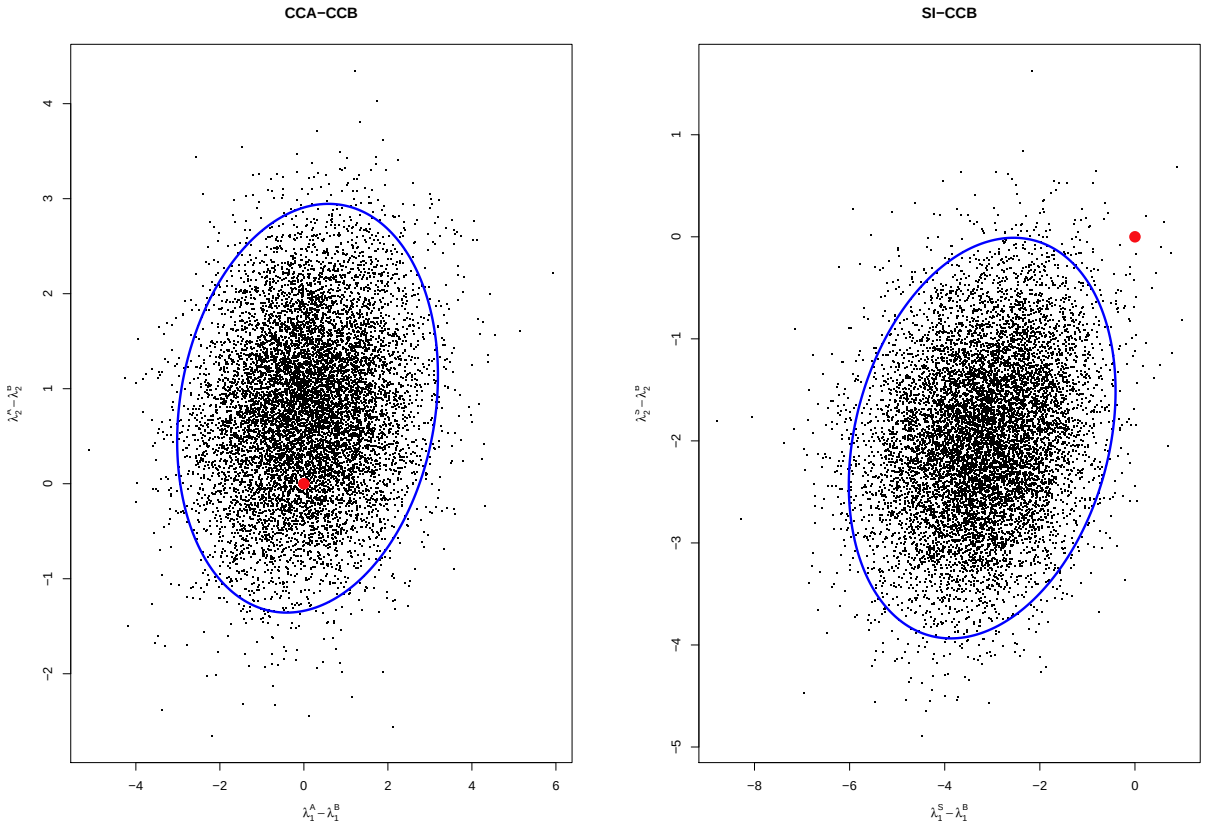

Fig. 5 Posterior samples for differences in λ_1 and λ_2 for the two sets of Christchurch data (left) and South Island and Christchurch data A (right). This shows a clear difference between the South Island and Christchurch data, but suggests no difference between the two sets of Christchurch data.

stant 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 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 processor[1]; 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. In particular, we found when we varied q from 3 to 7 that the probability of acceptance in the rejection sampling step was around 78%, 45%, 30%, 18% and 10% respectively. These numbers reveal why we found our algorithm to be very efficient for all our examples and efficient for at least a moderate number of dimensions. However, we anticipate that it will become slower (in CPU time) for $q \geq 7$.

In both the simulated and real datasets we chose the proposal distribution $h(\boldsymbol{\lambda}^{'}|\boldsymbol{\lambda})$ to be a multivariate Normal distribution with mean the current value of the chain and variance covariance matrix equal to σI. Any drawn values which did not satisfy the constraint $\lambda_1 > \ldots > \lambda_{q-1}$ were rejected straight away. The probability of the constraint being satisfied will decrease with q and in consequence such a proposal will become very inefficient for large values of q. Therefore, we have also implemented an alternative proposal distribution in which the constraint is implicitly taken account. Consider for illustration the case where $q = 3$; we draw $\lambda_2^{'}$ from a Normal distribution centered at the current value of λ_2 and then we draw $\lambda_1^{'}$ from a (truncated) Normal distribution with mean λ_1 subject to the constraint that $\lambda_1^{'} > \lambda_2^{'}$. It is easy to see how such a proposal can be generalised for $q > 3$. We have performed simulation studies (results not shown) and found that such a proposal distribution is more efficient than the standard random walk Metropolis when q and/or the difference between the successive values of the eigenvalue λ, $i = 1, \ldots, q-1$ is large.

With regards to the choice of prior distributions we assigned independent Exponential distributions with rate μ to each λ_i, $i = 1, \ldots q-1$ and independent Normal distributions with mean zero and variance v for the elements of A. We have used largely uninformative prior distributions in all applications and in particular we chose $\mu = 10^{-2}$ and $v = 10^2$. However, we performed some prior sensitivity analysis by choosing different values for both hyperparameters, e.g. 10^{-3} and 10^{-1} for μ and 10 and 10^3 for v and we found that there no material change in the inferred posterior distributions.

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 mo-

[1] Our code is available upon request.

ments of the Bingham distribution. Finally, we would like to thank Ian Dryden for commenting on an earlier draft of this manuscript.

References

R Arnold and P E Jupp. Statistics of orthogonal axial frames. *Biometrika*, 100(3):571–586, 2013.

Erlend Aune, Daniel P Simpson, and Jo Eidsvik. Parameter estimation in high dimensional gaussian distributions. *Statistics and Computing*, pages 1–17, 2012.

Christopher Bingham. An antipodally symmetric distribution on the sphere. *Ann. Statist.*, 2:1201–1225, 1974. ISSN 0090-5364.

W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci U S A*, 105(26):8932–7, 2008.

Alberto Caimo and Nial Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33(1): 41–55, 2011.

Martin Ehler and Jennifer Galanis. Frame theory in directional statistics. *Statistics and Probability Letters*, 81(8): 1046–1051, 2011.

Richard G Everitt. Bayesian parameter estimation for latent markov random fields and social networks. *Journal of Computational and Graphical Statistics*, 21(4):940–960, 2012.

Asaad M. Ganeiber. *Estimation and simulation in directional and statistical shape models*. PhD thesis, University of Leeds, 2012.

Mark Girolami, Anne-Marie Lyne, Heiko Strathmann, Daniel Simpson, and Yves Atchade. Playing russian roulette with intractable likelihoods. *ArXiv preprint; arXiv:1306.4032*, 2013.

John T. Kent. Asymptotic expansions for the Bingham distribution. *J. Roy. Statist. Soc. Ser. C*, 36(2):139–144, 1987. ISSN 0035-9254. doi: 10.2307/2347545. URL http://dx.doi.org/10.2307/2347545.

John T. Kent, Asaad M. Ganeiber, and Kanti V. Mardia. A new method to simulate the Bingham and related distributions in directional data analysis with applications. *ArXiv Preprint*, 2013. http://arxiv.org/abs/1310.8110.

A. Kume and Andrew T. A. Wood. Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants. *Biometrika*, 92(2):465–476, 2005. ISSN 0006-3444. doi: 10.1093/biomet/92.2.465. URL http://dx.doi.org/10.1093/biomet/92.2.465.

A. Kume and Andrew T. A. Wood. On the derivatives of the normalising constant of the Bingham distribution. *Statist. Probab. Lett.*, 77(8):832–837, 2007. ISSN 0167-7152. doi: 10.1016/j.spl.2006.12.003. URL http://dx.doi.org/10.1016/j.spl.2006.12.003.

Alfred Kume and Stephen G. Walker. Sampling from compositional and directional distributions. *Stat. Comput.*, 16(3):261–265, 2006. ISSN 0960-3174. doi: 10.1007/s11222-006-8077-9. URL http://dx.doi.org/10.1007/s11222-006-8077-9.

Joel D Levine, Pablo Funes, Harold B Dowse, and Jeffrey C Hall. Resetting the circadian clock by social experience in drosophila melanogaster. *Science Signaling*, 298(5600): 2010–2012, 2002.

K V Mardia and P J Zemroch. Table of maximum likelihood estimates for the bingham distribution. *Statist. Comput. Simul.*, 6:29–34, 1977.

Kanti V. Mardia and Peter E. Jupp. *Directional statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2000. ISBN 0-471-95333-4. Revised reprint of ıt Statistics of directional data by Mardia [MR0336854 (49 #1627)].

J. Møller, A. N. Pettitt, R. Reeves, and K. K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006. ISSN 0006-3444. doi: 10.1093/biomet/93.2.451. URL http://dx.doi.org/10.1093/biomet/93.2.451.

Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press, 2006.

Brian D. Ripley. *Stochastic simulation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987. ISBN 0-471-81884-4. doi: 10.1002/9780470316726. URL http://dx.doi.org/10.1002/9780470316726.

Cristina Rueda, Miguel A. Fernández, and Shyamal Das Peddada. 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, 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.0120. URL http://dx.doi.org/10.1198/jasa.2009.0120.

Tomonari Sei and Alfred Kume. Calculating the normalising constant of the bingham distribution on the sphere using the holonomic gradient method. *Statistics and Computing*, pages 1–12, 2013. ISSN 0960-3174. doi: 10.1007/s11222-013-9434-0. URL http://dx.doi.org/10.1007/s11222-013-9434-0.

Geir Storvik. On the flexibility of metropolishastings acceptance probabilities in auxiliary variable proposal generation. *Scandinavian Journal of Statistics*, 38(2):342–358, 2011. ISSN 1467-9469. doi: 10.1111/j.1467-9469.2010.00709.x. URL http://dx.doi.org/10.1111/j.1467-9469.2010.00709.x.

David E. Tyler. Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika*, 74(3):579–589, 1987. ISSN 0006-3444. doi: 10.1093/biomet/74.3.579. URL http://dx.doi.org/10.1093/biomet/74.3.579.

Stephen G. Walker. Posterior sampling when the normalizing constant is unknown. *Comm. Statist. Simulation Comput.*, 40(5):784–792, 2011. ISSN 0361-0918. doi: 10.1080/03610918.2011.555042. URL http://dx.doi.org/10.1080/03610918.2011.555042.

Stephen G. Walker. Bayesian estimation of the Bingham distribution. *BJPS*, 2013. To appear.