Brazilian Journal of Probability and Statistics
2014, Vol. 28, No. 1, 61–72
DOI: 10.1214/12-BJPS193

A Bayesian analysis of the Bingham distribution

Stephen G. Walker
University of Kent

Abstract. This paper provides a means, using latent variables, to undertake a Bayesian analysis for the Bingham distribution. To date, this has been problematic due to a nonclosed form for the normalizing constant. Previous approaches have relied on approximating the constant; something which is unnecessary in the method adopted here.

1 Introduction

The Bingham distribution (Bingham, 1993; Mardia and Jupp, 2000) is obtained as a multivariate normal vector $x = (x_1, \ldots, x_p)$ constrained to lie on the unit sphere $S_p = \{x : x_1^2 + \cdots + x_p^2 = 1\}$. The density function is hence given, for some symmetric matrix $A = A'$, by

$$f(x|A) = \frac{\exp(-x'Ax)\mathbf{1}(x \in S_p)}{c(A)},$$

where $c(A)$ is the normalizing constant and given by

$$c(A) = \int_{S_p} \exp(-x'Ax)m(\mathrm{d}x),$$

and $m(\mathrm{d}x)$ represents surface area on the unit sphere.

It is the normalizing constant which presents problems when endeavoring to undertake statistical inference involving the Bingham distribution. As is well known, see, for example, Kent (1987), it is sufficient to consider A as a diagonal matrix (following rotation to principal axes) and therefore we will take $A = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Now, for identifiability, we can arbitrarily take λ_p as the smallest value and indeed can set it to 0. See Kent (1987), for example. Hence, this results in each $\lambda_l \geq 0$. Throughout the remainder of the paper, we will now use Λ to denote the diagonal matrix with elements $\lambda = (\lambda_l)$. Thus, we have

$$f(x|\Lambda) \propto \exp\left(-\sum_{l=1}^{p-1} \lambda_l x_l^2\right),$$

Key words and phrases. Latent variables, Markov chain Monte Carlo, normalizing constant, posterior distribution.
Received October 2011; accepted April 2012.

a density with respect to the normalized measure

$$\omega(\mathrm{d}x) = c_p\,\mathrm{d}x_1\cdots\mathrm{d}x_{p-1}\big(1-x_1^2-\cdots-x_{p-1}^2\big)^{-1/2}$$

on $\tilde{S}_p = \{(x_1,\ldots,x_{p-1}) : x_1^2+\cdots+x_{p-1}^2 \le 1\}$. Hence,

$$c_p^{-1} = \int_{\tilde{S}_p} \mathrm{d}x_1\cdots\mathrm{d}x_{p-1}\big(1-x_1^2-\cdots-x_{p-1}^2\big)^{-1/2}.$$

Now writing the density in full, complete with normalizing constant, we have

$$f(x|\Lambda) = \frac{\exp(-\sum_{l=1}^{p-1}\lambda_l x_l^2)}{\int_{\tilde{S}_p}\exp(-\sum_{l=1}^{p-1}\lambda_l x_l^2)\omega(x)\,\mathrm{d}x_1\cdots\mathrm{d}x_{p-1}},$$

where we now write $\omega(x) = c_p(1-x_1^2-\cdots-x_{p-1}^2)^{-1/2}$.

The likelihood function, based on a sample

$$\big(x_i = (x_{i,1},\ldots,x_{i,p-1})\big)_{i=1}^n,$$

is given by

$$L_n(\lambda) = \frac{\exp(-n\sum_{l=1}^{p-1}\lambda_l\tau_l)}{c(\Lambda)^n},$$

where

$$\tau_l = n^{-1}\sum_{i=1}^n x_{i,l}^2.$$

Hence, the data can be represented as $(n,\tau_1,\ldots,\tau_{p-1})$. From this, it is clear that the maximum likelihood estimator can be achieved by maximizing

$$L(\lambda) = \frac{\exp(-\sum_{l=1}^{p-1}\lambda_l\tau_l)}{c(\Lambda)}$$

and so the normalizing constant $c(\Lambda)$ plays a crucial role. In fact, the maximum likelihood estimator can be derived by iterative techniques once one has good approximations to the $c(\Lambda)$ and its derivatives. See Kent (1987), Kume and Wood (2005, 2007). Note that for the case $p=3$, Mardia and Zemroch (1977) provide maximum likelihood estimators for (λ_1,λ_2) based on various choices of (τ_1,τ_2). On the other hand, Dryden (2005) describes maximum likelihood when both p and n are large and based on asymptotic approximations.

The aim in the present paper is to demonstrate that Bayesian posterior inference can be implemented using latent variable techniques (Besag and Green, 1993; Damien et al., 1999) and which do not consequently require any numerical approximations. The normalizing constant can be replaced by a density function defined up to a constant of proportionality; the normalizing constant of which is the reciprocal of $c(\Lambda)$. Hence, if y denotes the latent variable, we consider a joint density

$p(x, y) = g(x)h(y)$ where $g(x) = \exp(-\sum_{l=1}^{p-1} \lambda_l x_l^2)$ and $\int h(y)\,\mathrm{d}y = 1/c(\Lambda)$. The idea then is that $h(y)$ is tractable and there is in particular easy access to Λ.

The approach described in the present paper can also be used to perform Bayesian analysis for the Fisher–Bingham distribution (Mardia, 1975). This has density function given up to proportionality by

$$f(x|A, \mu) \propto \exp(-x'Ax + \mu' x)\mathbf{1}(x'x = 1),$$

where $\mu = (\mu_1, \ldots, \mu_p)$. It is a more complicated latent model to describe for this family and hence we concentrate on the Bingham family of distributions.

There has been limited, if any, Bayesian analysis of data assumed to arise from the Bingham distribution. This would be due to the normalizing constant not having a closed from expression. The present paper represents an application of an algorithm, dealing with uncomputable normalizing constants, to be found in Walker (2011). Full details of the general algorithm are presented in Walker (2011), which includes a description of alternative techniques for dealing with incomputable normalizing constants. However, none of these alternatives explicitly make use of a latent model.

Describing the layout of the paper: In Section 2, we describe the function $h(\cdot)$ which when integrated yields $[c(\Lambda)]^{-1}$. Posterior inference via MCMC is presented in Section 3 and Section 4 contains a numerical illustration.

2 The latent model

For a sample of size n from the Bingham distribution, say $(x_1, \ldots, x_n)$, where each $x_i = (x_{i,1}, \ldots, x_{i,p-1})$, we wish to estimate Λ. For a Bayesian, the task is to construct a posterior distribution for Λ.

The latent variable introduced for the Bingham distribution to account for the normalizing constant is based on the ideas presented in Walker (2011). The basic idea is to take an intractable normalizing constant of the type

$$\left(\frac{1}{\int g(s)m(\mathrm{d}s)}\right)^n$$

for some bounded (by 1) function g and a probability density m. The key to dealing with this is to use the fact that

$$\sum_{k=0}^{\infty} \binom{n+k-1}{k} \left[\int m(\mathrm{d}s)(1 - g(s))\right]^k = \left(\frac{1}{\int g(s)m(\mathrm{d}s)}\right)^n.$$

We can then consider

$$\left[\int m(\mathrm{d}s)(1 - g(s))\right]^k$$

as the density of a latent model

$$\prod_{l=1}^{k}[1-g(s_l)],$$

where the s_l are independent and identically distributed from m. So the intractable normalizing constant is removed and taking expectations returns the required term. The specific details are now presented.

We define a joint density on $\{0,1,2,\ldots\}\otimes\tilde{S}_p^{\infty}\otimes[0,1]^{\infty}$ which will be the joint density of all the required latent variables. So if we integrate and sum this joint density over this space, the normalizing constant will be the reciprocal of what it is for the un-normalized Bingham distribution. Hence, the normalizing constant disappears when the latent model is placed next to the un-normalized Bingham distribution. The latent density is given as follows:

$$\begin{aligned} h_n(k,s_1,s_2,\ldots,u_1,u_2,\ldots) &\propto \binom{n+k-1}{k} \\ &\quad\times\prod_{j=1}^{k}\mathbf{1}\left\{u_j<1-\exp\left(-\sum_{l=1}^{p-1}\lambda_l s_{jl}^2\right)\right\}\omega^{\infty}(s)\times p_k(u|s), \end{aligned}$$

where $\omega^{\infty}(s)=\prod_{j=1}^{\infty}\omega(s_j)$, and each $\omega(s_j)$ is proportional to the function, with each s_j of the type $s_j=(s_{j1},\ldots,s_{jp-1})$,

$$\omega(s_j)\propto\left(1-s_{j1}^2-\cdots-s_{jp-1}^2\right)^{-1/2}$$

on $\tilde{S}_p$, and

$$p_k(u|s)=\prod_{j=k+1}^{\infty}\mathrm{Un}\left[u_j\middle|0,1-\exp\left(-\sum_{l=1}^{p-1}\lambda_l s_{jl}^2\right)\right].$$

Here $\mathrm{Un}[u_j|a,b]$ is the density of the uniform distribution on the interval $[a,b]$, evaluated at u_j.

If we now integrate out the u, we obtain

$$h(k,s_1,s_2,\ldots)\propto\binom{n+k-1}{k}\prod_{j=1}^{k}\left\{1-\exp\left(-\sum_{l=1}^{p-1}\lambda_l s_{jl}^2\right)\right\}\omega^{\infty}(s).$$

If we now integrate out the s, we obtain

$$h(k)\propto\binom{n+k-1}{k}\left\{1-\int_{\tilde{S}_p}\exp\left(-\sum_{l=1}^{p-1}\lambda_l s_l^2\right)\omega(s)\,\mathrm{d}s\right\}^{k}.$$

We now use the result that for any $0 < q < 1$ it is that

$$\sum_{k=0}^{\infty} \binom{n+k-1}{k} (1-q)^k = q^{-n}.$$

Therefore, summing over k, we obtain

$$\sum_{k=0}^{\infty} h(k) \propto \left\{ \int_{\tilde{S}_p} \exp\left(-\sum_{l=1}^{p-1} \lambda_l s_l^2 \right) \omega(s)\, \mathrm{d}s \right\}^{-n}.$$

Consequently, if we consider the joint density

$$f(x,k,s,u|\lambda) \propto \exp\left(-\sum_{l=1}^{p-1} \lambda_l \sum_{i=1}^{n} x_{i,l}^2 \right) \binom{n+k-1}{k} \times \prod_{j=1}^{k} \mathbf{1}\left\{ u_j < 1 - \exp\left(-\sum_{l=1}^{p-1} \lambda_l s_{jl}^2 \right) \right\} \omega^{\infty}(s) \times p_k(u|s)$$

we note two points. The first is that the normalizing constant for this joint density does not depend on the $\lambda = (\lambda_l)$ and, second, the marginal density of x is precisely as we want it.

We should mention that the infinite dimensional (s,u) is not strictly necessary; we actually only need $(s_1,\ldots,s_k,u_1,\ldots,u_k)$ to define an appropriate latent model. We do introduce the full infinite dimensional model here as it assists with the Markov Chain Monte Carlo (MCMC) algorithm described in the next section; this avoids a reversible jump algorithm as there is no dimension change when we move k. More details and references are provided in the next section, but we mention here that the formulation of the infinite latent model is described in Godsill (2001).

Also note a phenomenon slightly unusual with latent models which is that x and (k,s,u) are independent, given λ. Indeed, it is λ that connects the two parts; effectively the numerator and the denominator.

3 Sampling the posterior

We will use a MCMC (Smith and Roberts, 1993) approach to sampling the full joint density. This will involve setting up a Markov chain which samples values (k,s,u,λ). Before proceeding, we briefly discuss the infinite dimensions of s and u since this might suggest we are unable to sample the correct posterior distribution. In fact, we do not need to sample all the variables in order to construct a MCMC algorithm with the correct stationary density. All the variables we do not sample (i.e., $(s_{k+1},u_{k+1}),\ldots$) are taken from parts of the joint density which have no effect on the sampling of λ. Hence, we ensure the λ samples are coming from

the correct stationary density. To complete the Bayesian model, we assign a prior distribution $\pi(\lambda)$ for λ. This will be of the form $\pi(\lambda) = \prod_{l=1}^{p-1} \pi(\lambda_l)$.

Suppose at a particular iteration of the Markov chain we are currently at the state k and have $s_1, \ldots, s_k$ and λ. Then we sample u_j from the uniform distribution on

$$\left(0, 1 - \exp\left(-\sum_{l=1}^{p-1} \lambda_l s_{jl}^2\right)\right),$$

independently for $j = 1, \ldots, k$.

Once this has been done, we can sample s_j, which comes from the density proportional to $\omega(s_j)$ on $\tilde{S}_p$, subject to the constraint

$$\sum_{l=1}^{p-1} \lambda_l s_{jl}^2 > -\log(1 - u_j),$$

again, independently for $j = 1, \ldots, k$. This is best done componentwise; so each s_{jl} is coming from the density given by

$$p(s_{jl}|\cdots) \propto \left(1 - s_{jl}^2 - \sum_{m \neq l} s_{jm}^2\right)^{-1/2} \times \mathbf{1}\left(s_{jl}^2 > \lambda_l^{-1}\left(-\log(1 - u_j) - \sum_{m \neq l} \lambda_m s_{jm}^2\right)\right).$$

This is easy to sample, as it is equivalent to sampling a truncated beta distribution.

Each λ_l, for $l = 1, \ldots, p-1$, would then be sampled from the density function

$$\pi(\lambda_l|\cdots) \propto \pi(\lambda_l) \exp\left(-\lambda_l \sum_{i=1}^{n} x_{i,l}^2\right) \mathbf{1}(\lambda_l \in A_l),$$

where

$$A_l = \left\{\lambda_l : \lambda_l > \max_{j=1,\ldots,k} \psi_{jl}\right\}$$

and

$$\psi_{jl} = \frac{-\log(1 - u_j) - \sum_{m \neq l} \lambda_m s_{jm}^2}{s_{jl}^2}.$$

This is not a difficult density to sample, particularly if the choice of $\pi(\lambda_l)$ is an exponential distribution, in which case the density becomes a truncated exponential.

Finally, we need to sample from the full conditional for k. This can be done with a Metropolis–Hastings step. At k, and assume for now that $k \neq 0$ or 1, then a proposal is made to either $k - 1$ or $k + 1$, with a probability of $1/2$ each. If the move to $k + 1$ is proposed then we would need to sample, while in state k, from

$p_k(u_{k+1}, s_{k+1})$, which would be sampling s_{k+1} from $\omega(s_{k+1})$ on $\tilde{S}_p$ and then u_{k+1} given s_{k+1} as uniform from the interval

$$\left(0, 1 - \exp\left(-\sum_{l=1}^{p-1} \lambda_l s_{k+1l}^2\right)\right).$$

The move from k to $k+1$ is accepted with probability

$$\min\left\{1, \frac{(n+k)(1-g(s_{k+1}, \lambda))}{k+1}\right\},$$

where

$$g(s, \lambda) = \exp\left(-\sum_{l=1}^{p-1} \lambda_l s_l^2\right).$$

On the other hand, if the proposal is made to go to state $k-1$, then this move is accepted with probability

$$\min\left\{1, \frac{k}{(n+k-1)(1-g(s_k, \lambda))}\right\}.$$

The alterations to these probabilities when $k = 1$ or $k = 0$ are obvious bearing in mind that a proposal from $k = 0$ to $k = 1$ must happen with probability 1. The k moves described here and the joint latent model are basically the Godsill (2001) formulation of reversible jump MCMC (Green, 1995).

An interesting scenario arises when we take $n = 1$. For here we have, after integrating out the (s, u),

$$h(k|\lambda) \propto \left\{1 - \int_{\tilde{S}_p} \exp\left(-\sum_{l=1}^{p-1} \lambda_l s_l^2\right) \omega(s)\, \mathrm{d}s\right\}^k.$$

Hence, $k+1$ given λ is a geometric distribution and as is well known then the mean of $k+1$ is given by

$$\frac{1}{\int_{\tilde{S}_p} \exp(-\sum_{l=1}^{p-1} \lambda_l s_l^2) \omega(s)\, \mathrm{d}s}$$

which is the normalizing constant. Thus, by running an MCMC algorithm with λ fixed and with $n = 1$, then an estimate of the normalizing constant is available. However, there seems no purpose in estimating this constant and the interesting aspect of the approach used in this paper is that the normalizing constant is not needed to be estimated at all.

4 Numerical illustrations

We start off by doing a comparison with the maximum likelihood estimators presented in Mardia and Zemroch (1977) based on $p = 3$. We choose the point $(\tau_1, \tau_2) = (0.30, 0.32)$ and in order to allow a good comparison we will need to take n large in the Bayesian approach so that the Bayes estimate, which is always taken to be the posterior mean, and maximum likelihood can coincide. From the tables in Mardia and Zemroch (1977), we obtain $(\widehat{\lambda}_1, \widehat{\lambda}_2) = (0.588, 0.421)$.

The Markov chain described in Section 3 was run with 10,000 iterations, keeping every 10th sample for estimation purposes, so that the Bayes estimates are based on 1000 samples. No burn-in was used. The prior for each λ_l is taken to be exponential with parameter 0.01. In this case, the Bayes estimates are $(0.580, 0.366)$. Thinning chains is by no means widely used and for a mathematical treatment of this practice see MacEachern and Berliner (1994). However, I prefer working with reduced autocorrelated chains and perhaps running the chain longer as a consequence. There are many authors who also employ thinned chains; see, for example, Meyer et al. (2003).

Next, we take $(n, \tau_1, \tau_2) = (20, 0.20, 0.25)$. This time the Markov chain described in Section 3 was run for 100,000 iterations. Every 100th value was collected for estimating λ. Even though the autocorrelation of the un-thinned chain is high, taking every 100th value leads to much reduced autocorrelation. This is evident in Figure 1 where the satisfactory nature of the autocorrelation functions is evident.

Posterior estimation therefore was possible and the estimate of λ_1 was 2.04 and the estimate of λ_2 was 1.51. In Figure 2, we show the trace of the running averages of the output of the chain for the two series λ_1 and λ_2, which are seen converging to the parameter estimates.

We now look at the full posterior distributions. We take $(n, \tau_1, \tau_2) = (10, 0.20, 0.25)$. Figure 3 contains the following items: The marginal posterior distributions for λ_1 and λ_2 are presented. These are based on a chain run for 100,000 iterations with every 100th sample used for constructing the distributions. For this example, and using the same output, we plot the joint samples. The samples appear uncorrelated and indeed the estimated correlation is 0.1. The samples of k which are generated at each iteration over the entire 100,000 iterations are also plotted. The mean value is approximately 30 which means the average number of latent variables at each iteration is 60. This is not necessarily a particularly high number. Modern computing capabilities should be able to handle such an array of latent variables and thinning the chain is the natural idea. In fact, running the chain over 100,000 iterations takes a matter of minutes on a laptop, with the algorithm coded in Scilab (http://www.scilab.org/).

The length of time of run of the chain will depend on k, since a large k involves the simulation of more variables within each iteration. The value of k will be large

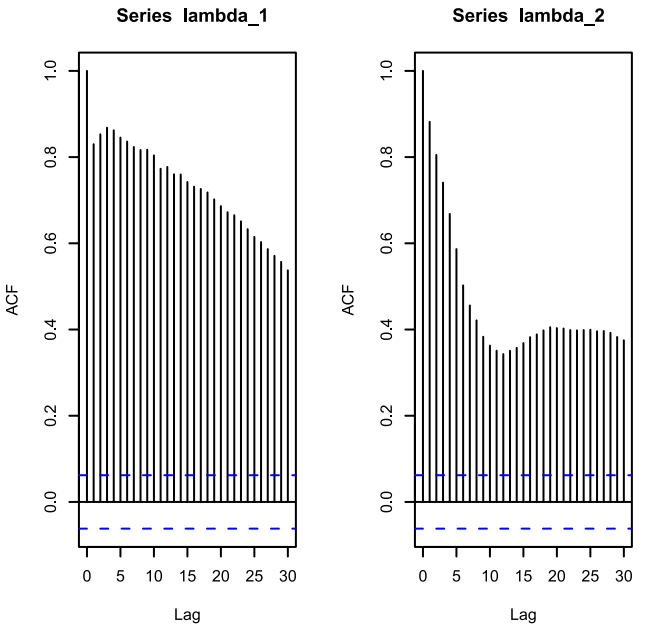

Figure 1 *Autocorrelation function for series* λ_1 *and* λ_2 *taking every* 100*th sample from the Markov chain.*

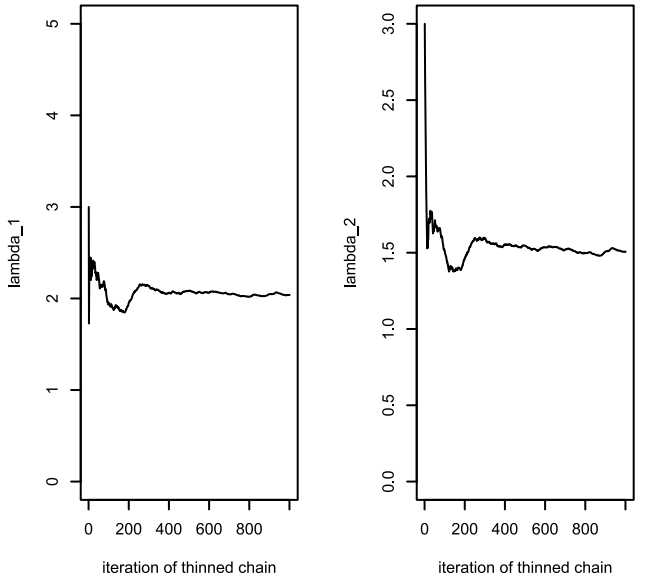

Figure 2 *Trace of running averages for* λ_1 *and* λ_2 *using every* 100*th sample from the Markov chain.*

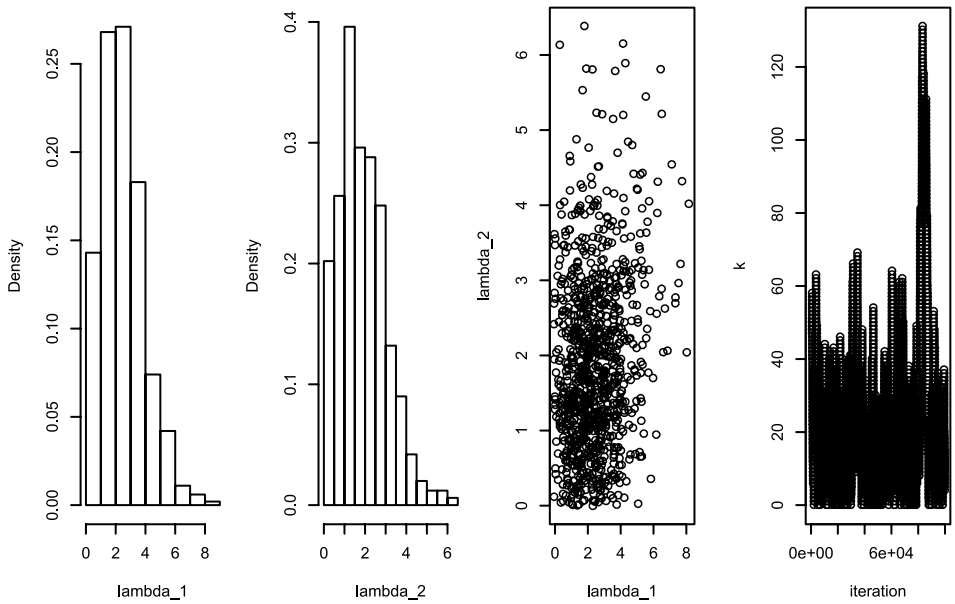

Figure 3 *For the data* $(n, \tau_1, \tau_2) = (20, 0.20, 0.25)$. *Histogram estimates of the marginal posterior densities for* λ_1 *and* λ_2 *with the joint posterior samples obtained from the MCMC output. Alongside are the output of the k samples.*

when the denominator is small; see Section 3 for an explanation of this. The denominator will be small when the λ are large and consequently another run of the chain was made taking $n = 20$ with $(\tau_1, \tau_2) = (0.02, 0.04)$. These yield large values of λ, the Bayes estimate of λ_1 is 23.7 and the Bayes estimate of λ_2 is 12.3. The chain now took just over $1/2$ hour to complete, that is, for the 100,000 iterations of the chain to finish. A plot of the posterior distributions of λ_1 and λ_2 are given in Figure 4, their marginal and joint. More interesting here, though not illustrated, is the value of k reached; the value increases steadily throughout the run of the chain and the value of k stabilizes at about 650 after about 50,000 iterations.

5 Discussion

The paper represents an application of a new approach for dealing with intractable normalizing constants. The aim is to use two sets of latent variables. First, to put the normalizing constant into a position whereby it can be written as an expectation, and second to remove the expectation so a full latent model materializes. The latent model avoids the problems of having to approximate the normalizing constant and also of being forced to find specific Metropolis samplers which remove the normalizing constant in the accept/reject ratios. Note for the method demonstrated here, the normalizing constant is dealt with before any sampler is

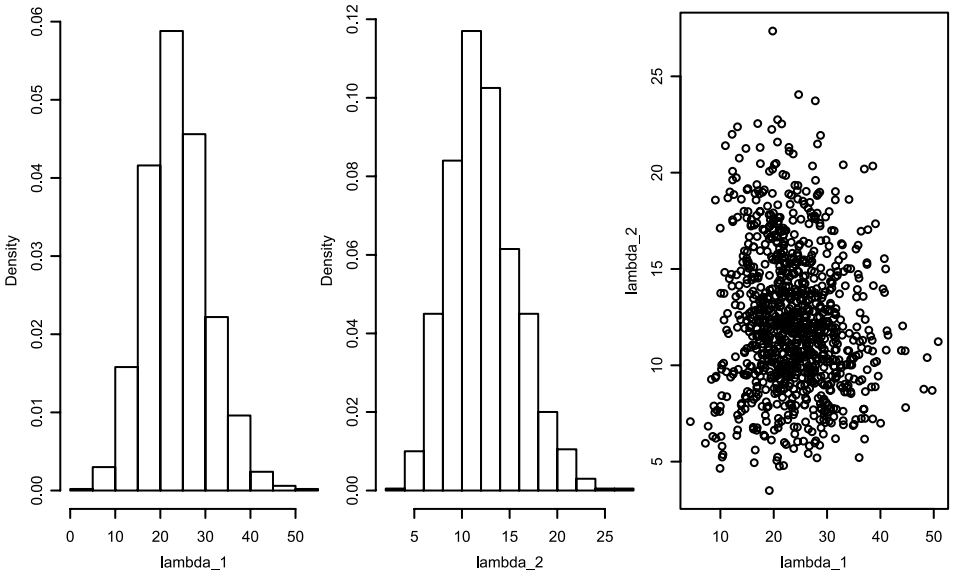

Figure 4 *For the data* $(n, \tau_1, \tau_2) = (20, 0.02, 0.04)$. *Histogram estimates of the marginal posterior densities for* λ_1 *and* λ_2 *with the joint posterior samples obtained from the MCMC output.*

even thought about or needed. In fact, an alternative Gibbs sampler could also be employed but would probably work out inferior to the Metropolis step of the type discussed by Godsill (2001).

Issues with the chain taking a long time to run, though the experience is that this is never excessive, amounted to a matter of at most 1 hour to complete 100,000 iterations. This case arises when the (τ_l) are small, leading to large λ_l, and hence leading to a large k. A large k means more latent variables have to be sampled at each iteration. For moderate τ_l the chain runs very fast, taking a matter of minutes for 100,000 iterations, which by MCMC standards is quite reasonable.

Acknowledgments

The author is grateful for the comments of an Associate Editor and referee which improved the presentation of the paper.

References

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Ser. B* **55**, 25–37. MR1210422

Bingham, C. (1974). An antipodally symmetric distribution on the sphere. *The Annals of Statistics* **2**, 1201–1205. MR0397988

Damien, P., Wakefield, J. C. and Walker, S. G. (1999). Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. *Journal of the Royal Statistical Society, Ser. B* **61**, 331–344. MR1680334

Dryden, I. L. (2005). Statistical analysis on high-dimensional spheres and shape spaces. *The Annals of Statistics* **33**, 1643–1665. MR2166558

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* **10**, 230–248. MR1939699

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732. MR1380810

Kent, J. T. (1987). Asymptotic expansions for the Bingham distribution. *Applied Statistics* **36**, 139–144. MR0897453

Kume, A. and Wood, A. T. A. (2005). Saddlepoint approximations for the Bingham and Fisher–Bingham normalizing constant. *Biometrika* **92**, 465–476. MR2201371

Kume, A. and Wood, A. T. A. (2007). On the derivatives of the normalizing constant of the Bingham distribution. *Statistics and Probability Letters* **77**, 832–837. MR2369690

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician* **48**, 188–190.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society, Ser. B* **37**, 349–393. MR0402998

Mardia, K. V. and Zemroch, P. J. (1977). Maximum likelihood estimators for the Bingham distribution. *Journal of Statistical Computation and Simulation* **6**, 29–34.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Chichester: Wiley. MR1828667

Meyer, R., Fournier, D. A. and Berg, A. (2003). Stochastic volatility: Bayesian computation using automatic differentiation and the extended Kalman filter. *Econometrics Journal* **6**, 407–419. MR2028243

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Ser. B* **55**, 3–23. MR1210421

Walker, S. G. (2011). Posterior sampling when the normalizing constant is unknown. *Communications in Statistics* **40**, 784–792. MR2783887

School of Mathematics, Statistics & Actuarial Science
University of Kent
Canterbury, Kent, CT2 7NZ
UK
E-mail: S.G.Walker@kent.ac.uk