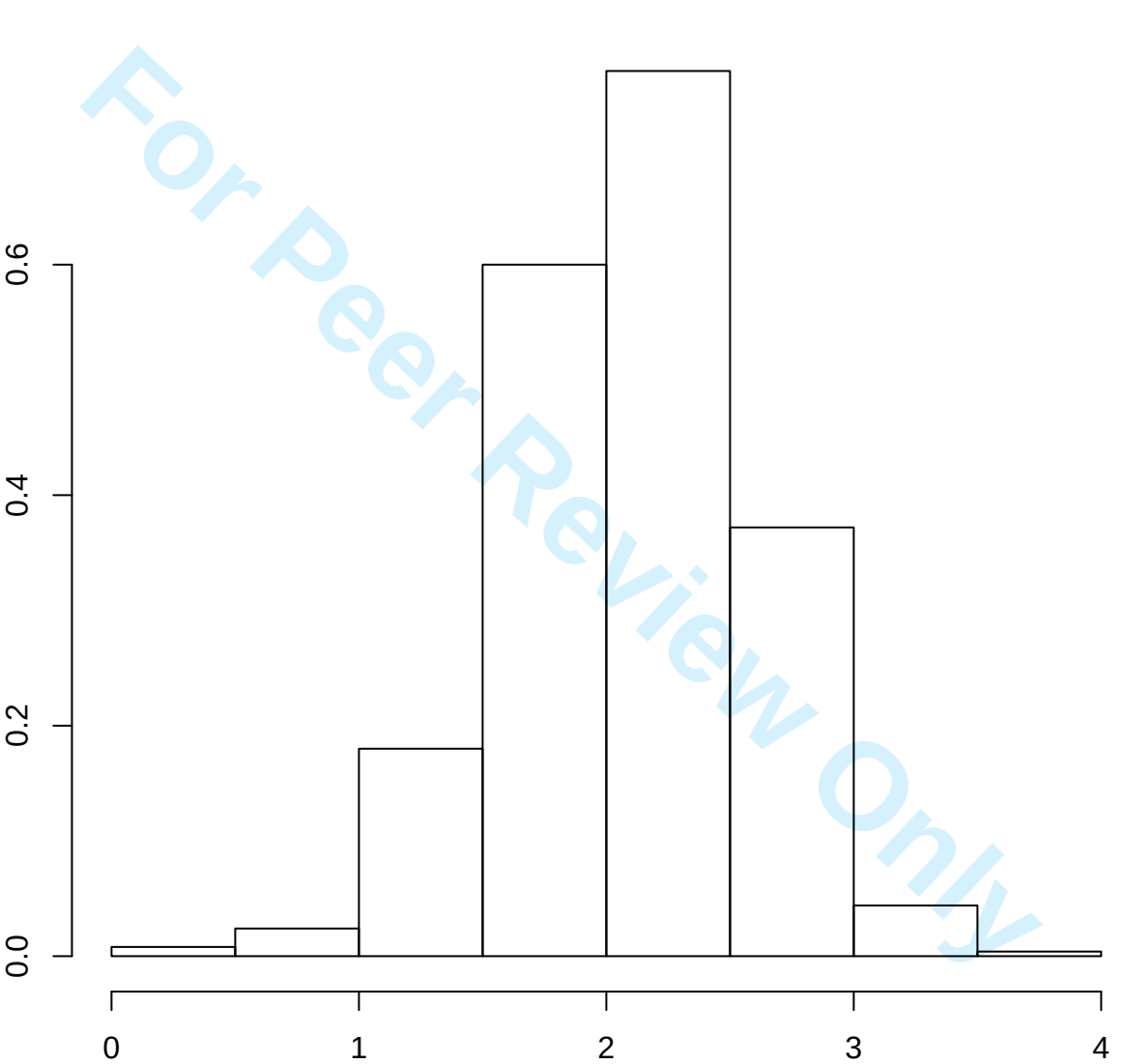
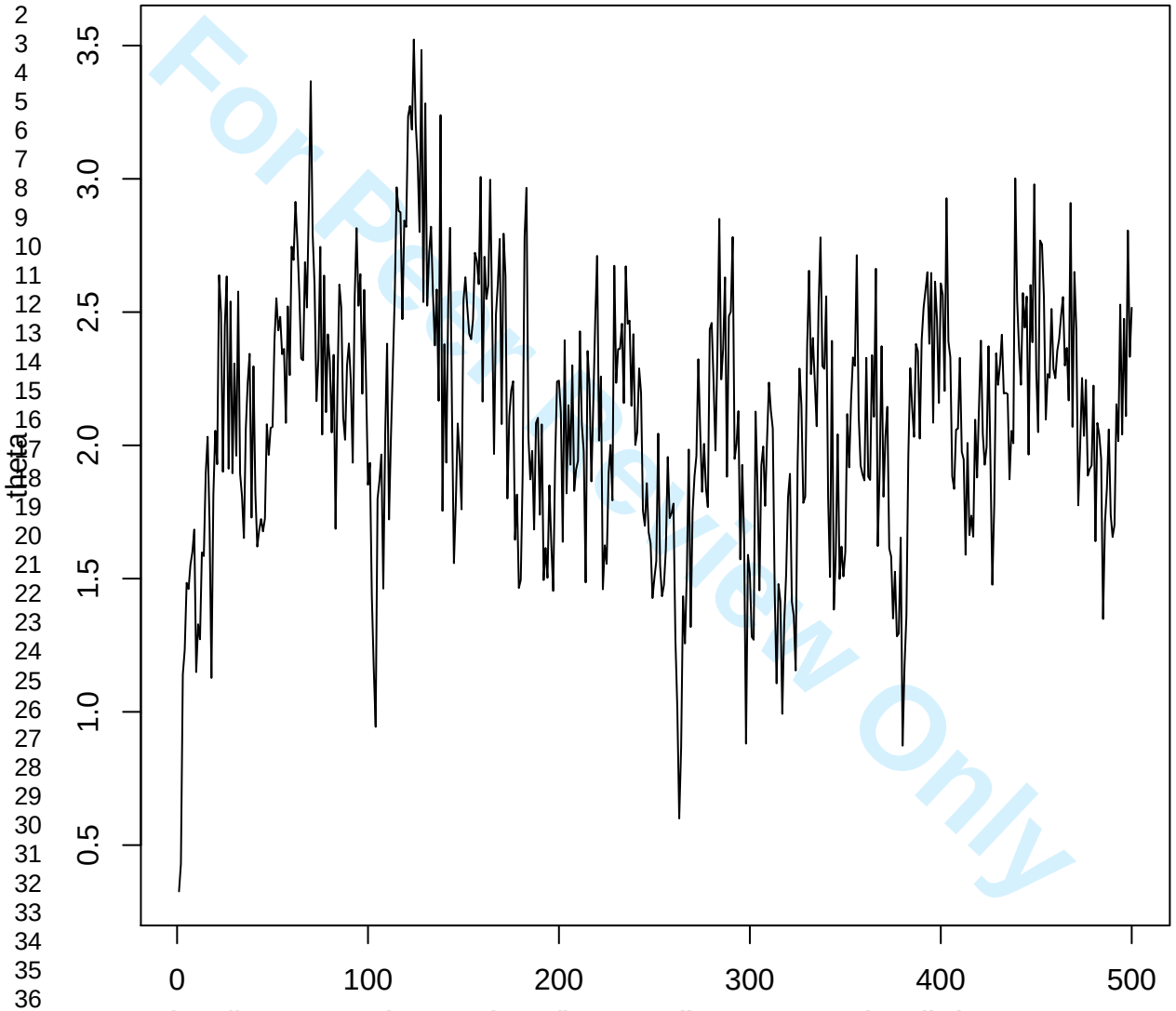
Posterior Sampling when the Normalizing Constant is Unknown

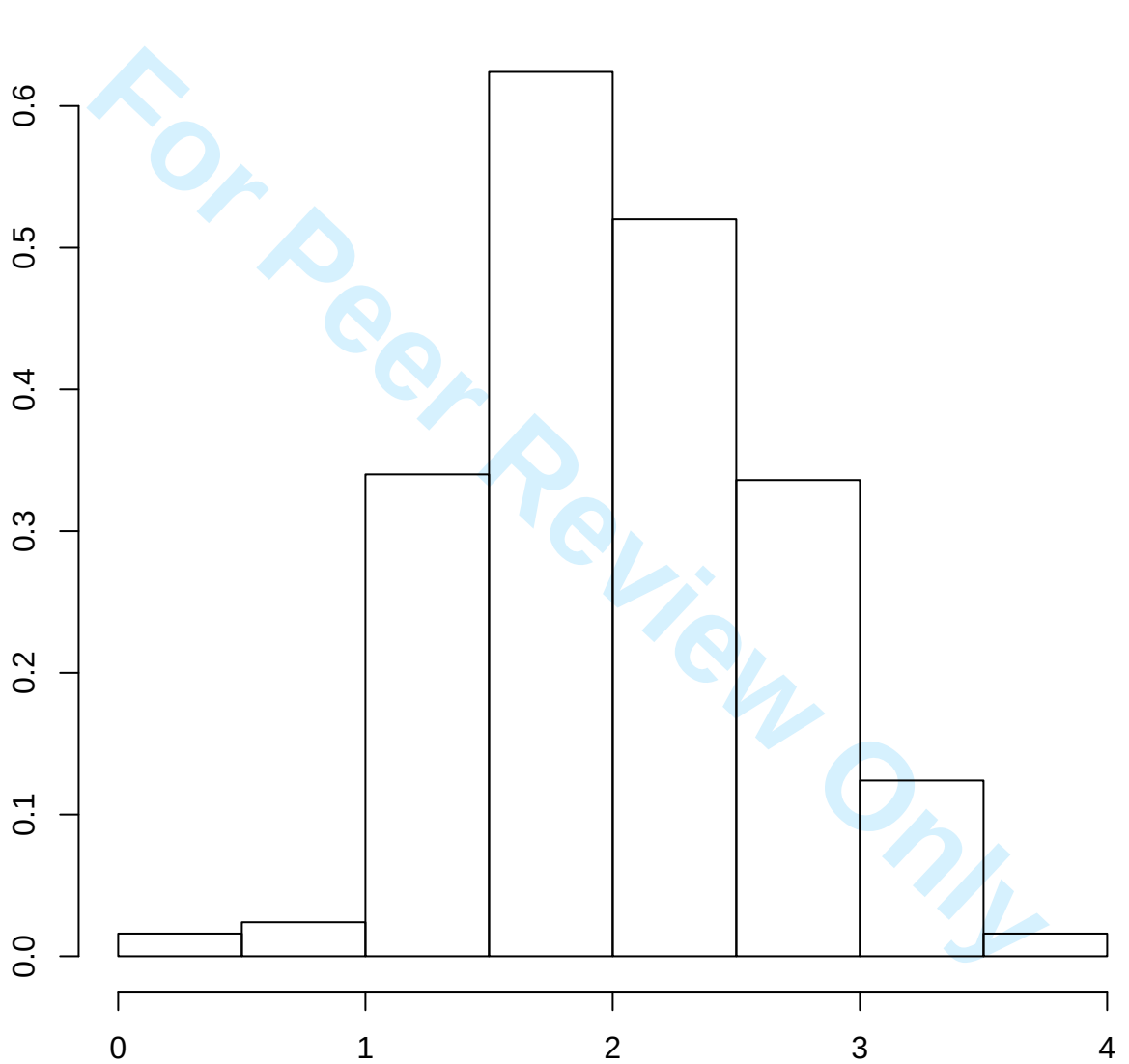
Stephen G. Walker

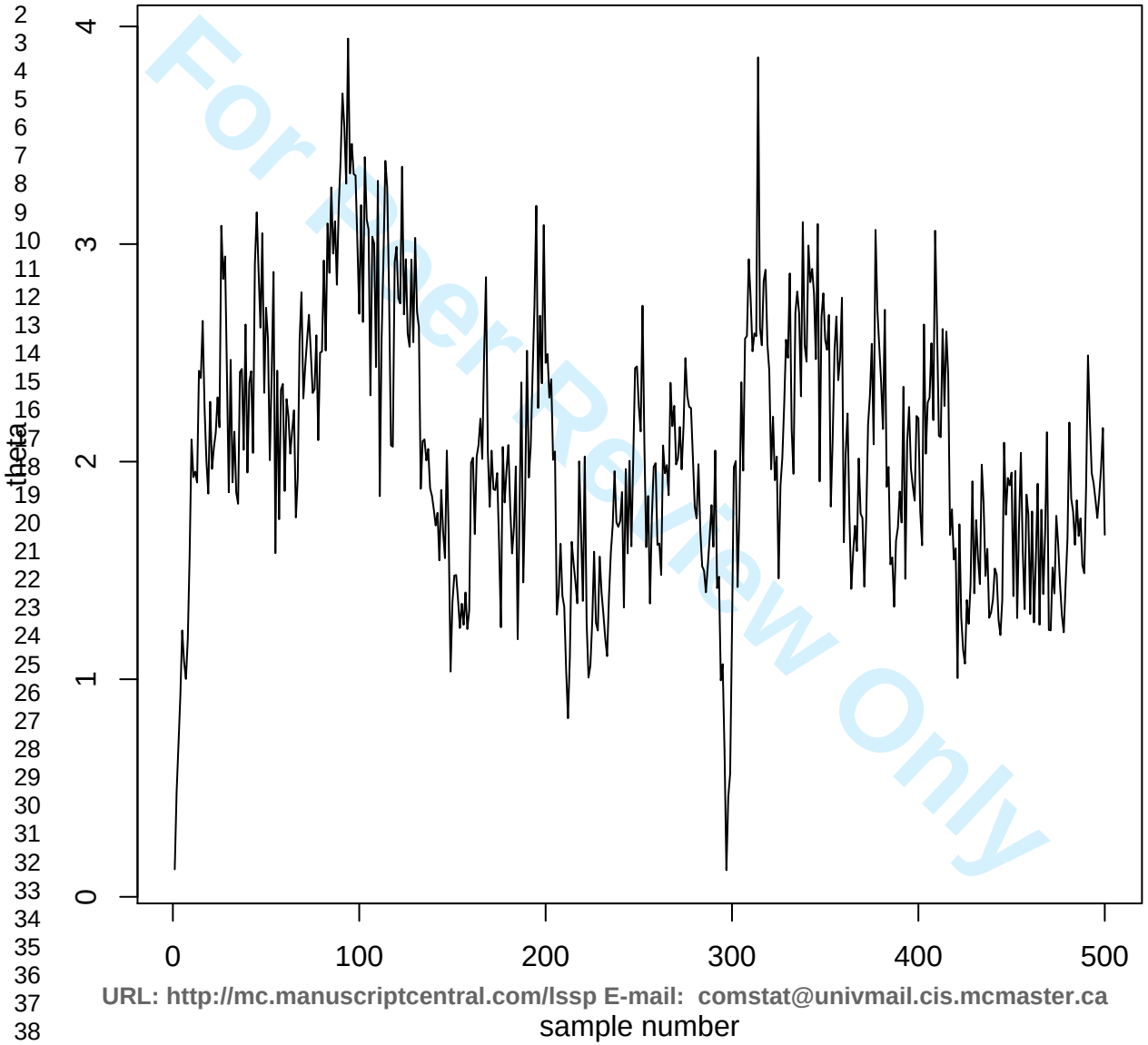
Posterior Sampling when the Normalizing Constant is Unknown

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2010-0228.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	27-Dec-2010
Complete List of Authors:	Walker, Stephen G.; IMSAS
Keywords:	Bayesian inference, unknown normalizing constant, Gibbs sampling
Abstract:	This paper describes a means by which to undertake Bayesian posterior inference via sampling techniques when the normalizing constant is not computable and hence unavailable. The strategy relies on the introduction of latent variables which removes any integrals associated with the inaccessibility of the normalizing constant.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
revj.zip	









Posterior Sampling when the Normalizing Constant is Unknown

Stephen G. Walker *

Abstract

This paper describes a means by which to undertake Bayesian posterior inference via sampling techniques when the normalizing constant is not computable and hence unavailable. The strategy relies on the introduction of latent variables which removes any integrals associated with the inaccessibility of the normalizing constant.

Keywords: Bayesian inference, Gibbs sampling, Reversible Jump MCMC, Unknown normalizing constant.

1. Introduction. This paper considers the situation when a probability model is employed for which the normalizing constant is not computable. That is, for $y \in I$,

$$f(y|\theta) = \frac{g(y,\theta)}{\int_I g(s,\theta)\,\mathrm{d}s}$$

where $g(y,\theta)$ is known and computable, but

$$Z(\theta) = \int_I g(s,\theta)\,\mathrm{d}s$$

is uncomputable. Such a scenario arises naturally in a number of problems:

1. Censored data problems; for some density $g(y,\theta)$ and $A \subset I$ it is that
$$f(y|\theta) \propto g(y,\theta)\,\mathbf{1}(y \in A)$$
and the normalizing constant $\int_A g(y,\theta)\,\mathrm{d}y$ is not computable.

*Stephen G. Walker is Professor of Statistics, School of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, U. K. (email: S.G.Walker@kent.ac.uk)

2. Weighted sampling problems; for some density $g(y, \theta)$ and weight function $w(y)$, it is that

$$f(y|\theta) \propto w(y)\, g(y, \theta)$$

and $\int w(y)\, g(y, \theta)\, \mathrm{d}y$ is not computable.

We will provide a means by which to implement an exact Gibbs sampler based on latent variables; see Besag and Green (1993) and Damien et al. (1999). The aim here is to use latent variables to remove the integral from the denominator. Numerical methods work first by calculating the normalizing constant and then undertake posterior inference with the estimated normalizing constant. Path sampling (Gelman and Meng, 1998) is one popular approach and some applications are to be found in Pettitt et al. (2003).

Other methods rely on an original auxiliary variable scheme introduced by Moller et al. (2006). This ideas has been extended by Murray et al. (2006) and also by Adams et al. (2009). Specifically, this latter paper deals with case 2. above whereby it is also that w is modeled as a stochastic process in a Bayesian setting. These algorithms rely on proposal distributions associated with auxiliary variables which, within the framework of a Metropolis–Hastings sampler, yield an acceptance probability ratio which does not depend on the normalizing constant. However, difficulties emerge in obtaining good enough acceptance probabilities and also the need to sample the proposals which are forced in order to obtain the correct acceptance probability.

The method outlined in this paper uses auxiliary variables but only to remove the problem caused by the normalizing constant. The MCMC algorithm can then be constructed with no special considerations required to be taken into account. In Murray et al. (2006) it is stated that "No known method of defining auxiliary variables removes $Z(\theta)$ from the joint distribution". The joint distribution referred to is the one for (y, θ) once a prior $\pi(\theta)$ has been included; that is

$$f(y, \theta) = \frac{g(y, \theta)}{Z(\theta)}\, \pi(\theta).$$

The present paper indeed finds such auxiliary variables for removing $Z(\theta)$.

Hence, we can proceed with posterior inference for θ without first having to estimate the normalizing constant and without being forced to employ a special MCMc algorithm for which the normalizing constant only disappears in an acceptance probability ratio. The only condition under which we work is that g is bounded; so there exists some known constant $M < +\infty$ such that $g(y, \theta) \leq M$ for all θ and y. In all of the material which follows, and without loss of generality,

we will assume that $M = 1$. We will also assume that y belongs to some bounded interval which, again without loss of generality, we will assume to be the interval $[0, 1]$. This is to cover the general case

$$f(y|\theta) = \frac{g(y,\theta)}{\int_I g(s,\theta)\,\mathrm{d}s}.$$

However, in cases 1. and 2. described earlier no restrictions are required save A need be a bounded set and w needs to be bounded. We can also cover the case 2. with w modeled as a bounded stochastic process based on a transformed Gaussian process, as in Adams et al. (2009).

In Section 2 we will describe the latent model which provides the basis for a reversible jump MCMC algorithm for sampling the posterior distribution. The reversible jump algorithm is given in some detail in Section 3 and Section 4 contains some numerical illustrations.

2. The latent variables. The likelihood function based on a sample of size n from $f(y|\theta)$ is given by

$$f(y_1, \ldots, y_n|\theta) \propto \frac{\prod_{i=1}^n g(y_i,\theta)}{m(\theta)^n}$$

where

$$m(\theta) = \int_0^1 g(s,\theta)\,\mathrm{d}s.$$

When faced with a denominator of the type $m(\theta)^n$, a standard trick (see, e.g. Nieto at al., 2004) to remove some of the complexity is to use

$$f(v, y|\theta) \propto v^{n-1} \exp\{-v\, m(\theta)\} \prod_{i=1}^n g(y_i,\theta)$$

so that integrating out the v yields the likelihood function. But, when $m(\theta)$ is an uncomputable integral then little progress has been made here.

However, we can now introduce some further latent variables

$$(k, s_1, \ldots, s_k)$$

which removes the integral:

$$f(v, k, s^{(k)}, y|\theta) \propto \frac{e^{-v} v^{k+n-1}}{k!} \prod_{j=1}^k \{(1 - g(s_j,\theta))\mathbf{1}(0 < s_j < 1)\} \prod_{i=1}^n g(y_i,\theta),$$

where $s^{(k)} = (s_1, \ldots, s_k)$. Integrating out the $(s_j)_{j=1}^k$ and the summing over k returns the likelihood. One further idea which can ease

the sampling algorithm would be to introduce further latent variables $(u_j)_{j=1}^k$ which interact with the (s_j) via

$$\prod_{j=1}^{k} \mathbf{1}(u_j < 1 - g(s_j, \theta)).$$

This then provides us with a basis for the implementation of a MCMC (Smith and Roberts, 1993) for sampling the model. The only possible source of complication is the k variable which when changes the dimension of the model also changes. This can then be solved using ideas based on reversible jump MCMC (Green, 1995).

The variables $((u_j, s_j)_{j=1}^k, v, \theta)$, once a prior $\pi(\theta)$ has been specified, should not be difficult to sample from their full conditional density functions and so the next section is devoted to the sampling of the k, which needs some attention.

For completeness we provide the other full conditional densities here. Bur before doing this we note a simple procedure which is to remove the v latent variable. Integrating out v from $f(v, k, s^{(k)}, y|\theta)$ yields

$$f(k, s^{(k)}, y|\theta) \propto \left(\begin{array}{c} n+k-1 \\ k \end{array} \right) \prod_{j=1}^{k} \{(1 - g(s_j, \theta))\mathbf{1}(0 < s_j < 1)\} \prod_{i=1}^{n} g(y_i, \theta).$$

Now we can also confirm that integrating out the $s^{(k)}$ and summing over k returns the original likelihood due to the identity

$$\sum_{k=0}^{\infty} \left(\begin{array}{c} n+k-1 \\ k \end{array} \right) \xi^k = (1 - \xi)^{-n}$$

for any $0 < \xi < 1$ and $n \geq 1$.

The full conditional for u_j is uniform from the interval $(0, 1 - g(s_j, \theta))$ and similarly we take s_j uniformly from the interval $\{0 < s < 1 : g(s, \theta) < 1 - u_j\}$. Finally, we have the conditional for θ as

$$\pi(\theta|\cdots) \propto \left\{ \prod_{i=1}^{n} g(y_i, \theta) \right\} \pi(\theta)\mathbf{1}(s \in A),$$

where

$$A = \{\theta : g(s_j, \theta) < 1 - u_j \ \forall \ j\}.$$

We now turn to the sampling of k.

3. Sampling k. The form of reversible jump MCMC for k described here is based on the formulation presented in Godsill (2001). The idea

here is to complete the model with an infinite set of $(u_j, s_j)_{j=1}^{\infty}$ and construct a joint density $p(k, u, s)$ of the form

$$p(k,u,s) \propto \binom{n+k-1}{k} \prod_{j=1}^{k} \{\mathbf{1}(u_j < 1 - g(s_j,\theta))\mathbf{1}(0 < s_j < 1)\}$$
$$\times \prod_{j=k}^{\infty} p(u_{j+1}, s_{j+1}|u_j, s_j),$$

where $p(u_{j+1}, s_{j+1}|u_j, s_j)$ are to be specified density functions acting as proposals for the states moved to when k changes. In this case and formulation of a joint density, there is no dimension change when k changes and hence a standard Metropolis step can be implemented. The specific form involving the

$$\prod_{j=k}^{\infty} p(u_{j+1}, s_{j+1}|u_j, s_j)$$

term means that there is substantial canceling when the Metropolis acceptance probability is computed.

One possibility for the $p(u_{j+1}, s_{j+1}|u_j, s_j)$ is based on an independent proposal density and given by

$$p(u_j, s_j) = \frac{\mathbf{1}(u_j < 1 - g(s_j,\theta))}{1 - g(s_j,\theta)} \mathbf{1}(0 < s_j < 1).$$

In fact there is little reason here to have a dependent proposal and it makes things simpler to work with.

Now suppose the chain is at state k and a proposal is made, with probability $q(k+1|k)$, to state $k+1$. Then we would need to sample (u_{k+1}, s_{k+1}) from $p(u_{k+1}, s_{k+1})$ and the move is accepted with probability

$$\min\left\{1, \frac{(n+k)\,(1 - g(s_{k+1},\theta))q(k|k+1)}{(k+1)\,q(k+1|k)}\right\}.$$

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

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

Note that it is a must for $q(1|0) = 1$, whereas for all other moves it seems reasonable for $q(k'|k) = \frac{1}{2}$ for all $|k - k'| = 1$.

4. Numerical illustrations. We start with a simple example whereby we take

$$g(y,\theta) = e^{-\theta y^2}$$

for $0 < y < 1$ and $\theta > 0$. Then, specifically, s_j is uniform from the interval $(\sqrt{-\theta^{-1}\log(1-u_j)}, 1)$ and θ has density

$$\pi(\theta)\exp\left\{-\theta\sum_{i=1}^{n} y_i^2\right\}\mathbf{1}(\theta > a)$$

where

$$a = \max_{1\le j\le k}\{-s_j^{-2}\log(1-u_j)\}.$$

We took the prior as $\pi(\theta) = e^{-\theta}$ so the full conditional for θ is easy to sample.

For the example, we took the true value of θ as 2; 100 observations were generated, which can be done by sampling truncated normal random variables with mean 0 and variance $1/(2\theta)$. The chain was run for 50,000 iterations and every 100th sample was used to construct the posterior distribution of θ presented in Figure 1. The posterior mean was 2.10. In Figure 2 we provide the trace of the plot of samples from the chain, exhibiting adequate mixing.

In the second illustration we take

$$g(y,\theta) = (1+y^2)^{-\theta}$$

for $0 < y < 1$ and $\theta > 0$. Again, we sample 100 observations with a true value of $\theta = 2$. The conditional distribution of s_j is uniform from the interval $(\sqrt{(1-u_j)^{-1/\theta}-1}, 1)$, and the density for θ is given by

$$\pi(\theta)\exp\left\{-\theta\sum_{i=1}^{n}\log(1+y_i^2)\right\}\mathbf{1}(\theta > a),$$

where

$$a = \max_{1\le j\le k}\left\{\frac{-\log(1-u_j)}{\log(1+s_j^2)}\right\}.$$

We again took a standard exponential prior for θ.

The posterior distribution of θ is presented in Figure 3. This is based on taking every 100th sample from a chain of length 50,000. The posterior mean is 2.04. The trace of samples, demonstrating adequate mixing, is presented in Figure 4.

In both cases the proposal probabilities for moving k was taken to be $\frac{1}{2}$; to $k+1$ and $k-1$. Except when $k = 0$ and then the proposal is probability 1 of moving to $k = 1$. In summary, the conditional densities are as follows:

- For $j = 1, \ldots, k$, $u_j \sim \text{Un}(0, 1 - g(s_j, \theta))$.

- For $j = 1, \ldots, k$, $s_j \sim \text{Un}\{0 < s < 1 : g(s, \theta) > 1 - u_j\}$.
- $\pi(\theta|\cdots) \propto \pi(\theta) \prod_{i=1}^{n} g(y_i, \theta) \mathbf{1}(\theta \in A)$, where

$$A = \{\theta : g(s_j, \theta) > 1 - u_j \ \forall \ j\}.$$

- Moving k as described in Section 3.

If at a particular iteration it is that $k = 0$, then the sampling of the (u, s) is not needed; and θ is sampled without the constraint of any $(\theta \in A)$.

5. Discussion. This paper has presented a means by which to undertake posterior sampling directly even when the normalizing constant is not available. There is no need to estimate it first using numerical methods nor ensure its removal only on the construction of a special MCMC algorithm. We have introduced auxiliary variables so that the normalising constant is removed prior to the introduction of any algorithm. Any MCMC can be used based on any proposal distribution. The only requirement is that in the most general case we require the $g(y, \theta)$ function to be bounded.

Acknowledgements. Thanks to a referee for constructive comments.

References.

Adams, R.P., Murray, I. and MacKay, D.J.C. (2009). Nonparametric Bayesian density modeling with Gaussian processes. In arXiv:0912.4896v1.

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

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, Series B* **61**, 331–344.

Gelman, A. and Meng, X. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

Godsill, S.J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *J. Comp. Graph. Stats.*, **10**, 230–248.

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

Moller, J., Pettit, A.N., Reeves, R. and Bethelsen, K.K. (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451–458.

Murray, I., Ghahramani, Z., and MacKay, D.J.C. (2006). MCMC for doubly–intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 359–366.

Nieto–Barajas, L.E., Prünster, I. and Walker, S.G. (2004). Normalised random measures driven by increasing additive processes. *Annals of Statistics* **32**, 2343–2360.

Pettitt, A.N., Friel, N. and Reeves, R. (2003). Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice. *Journal of the Royal Statistical Society, Series B* **65**, 235–246.

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, Series B* **55**, 3–23.

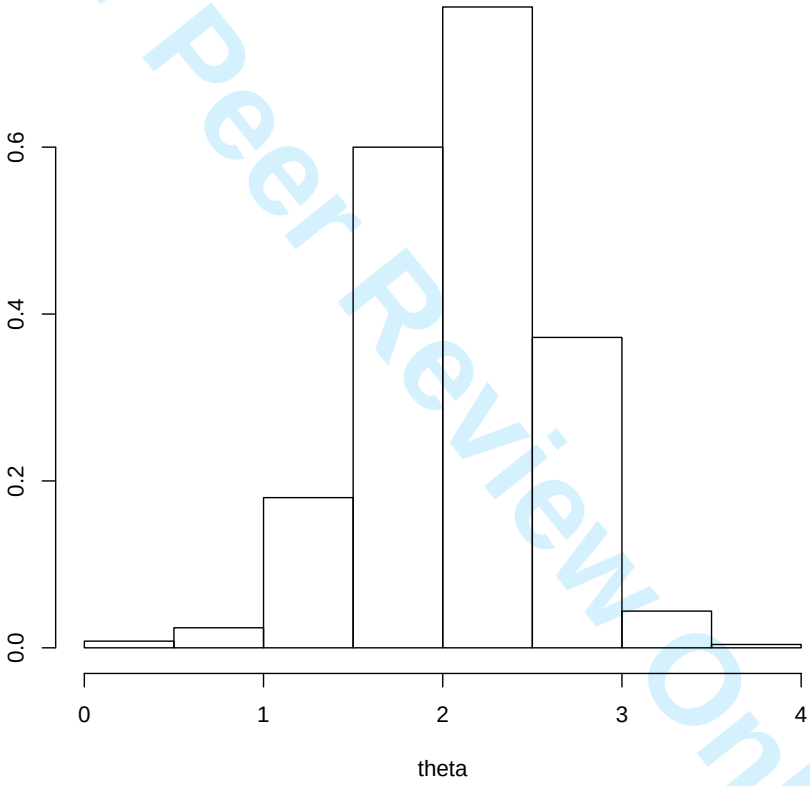


Figure 1: Density estimate for posterior distribution of θ for first illustration

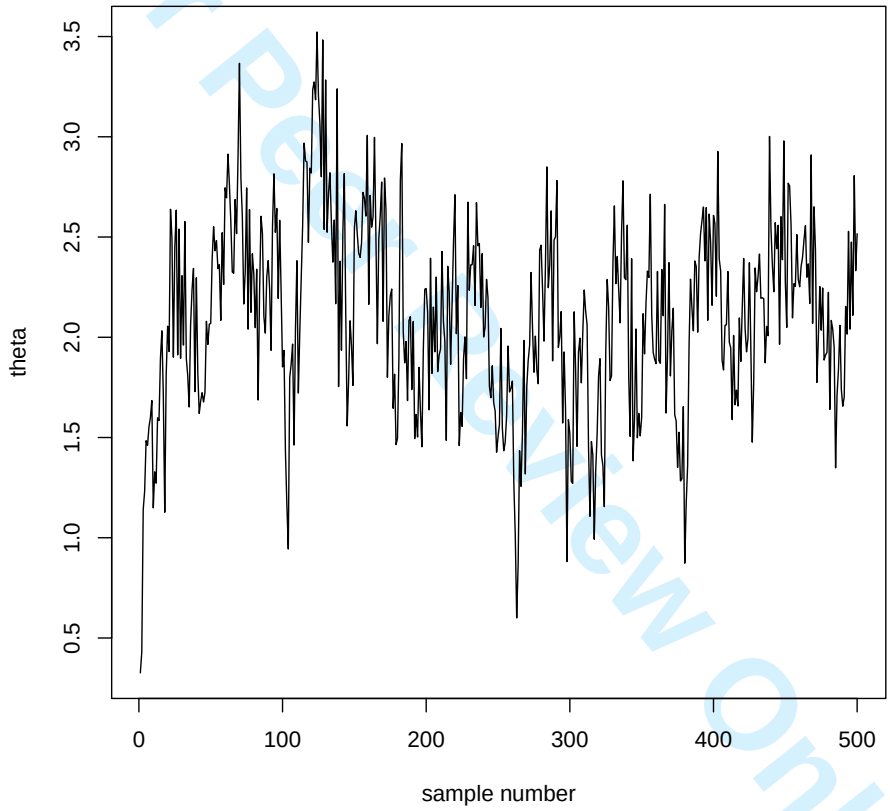


Figure 2: Trace of samples from chain for first illustration

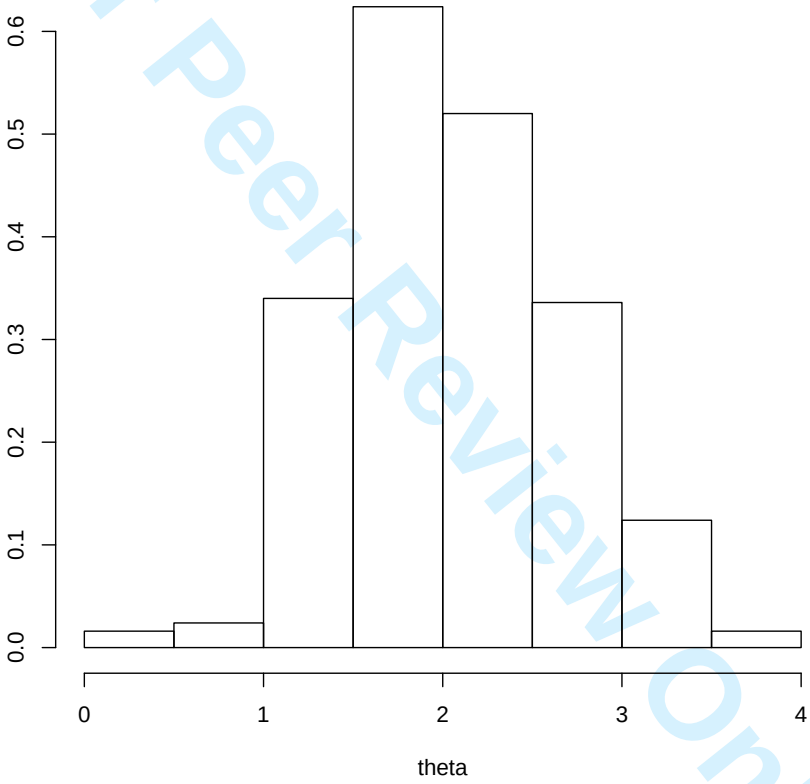


Figure 3: Density estimate for posterior distribution of θ for second illustration

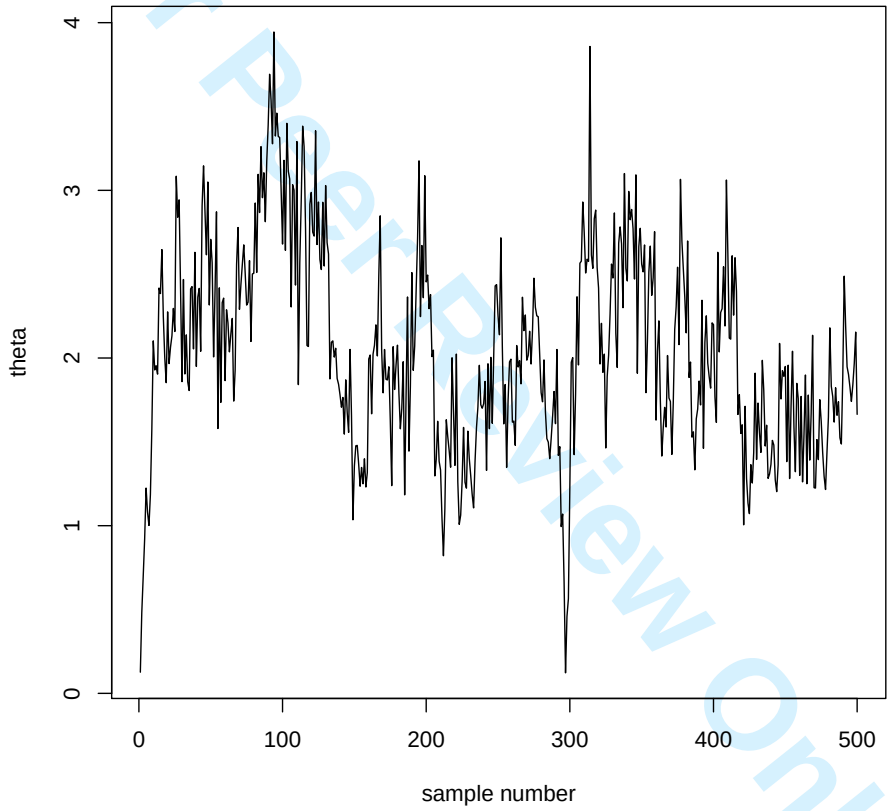


Figure 4: Trace of samples from chain for second illustration