Block-Wise Pseudo-Marginal Metropolis-Hastings

M.-N. Tran* R. Kohn † M. Quiroz ‡§ M. Villani‡

March 25, 2016

Abstract

The pseudo-marginal Metropolis-Hastings approach is increasingly used for Bayesian inference in statistical models where the likelihood is analytically intractable but can be estimated unbiasedly, such as random effects models and state-space models, or for data subsampling in big data settings. In a seminal paper, Deligiannidis et al. (2015) show how the pseudo-marginal Metropolis-Hastings (PMMH) approach can be made much more efficient by correlating the underlying random numbers used to form the estimate of the likelihood at the current and proposed values of the unknown parameters. Their proposed approach greatly speeds up the standard PMMH algorithm, as it requires a much smaller number of particles to form the optimal likelihood estimate. We present a closely related alternative PMMH approach that divides the underlying random numbers mentioned above into blocks so that the likelihood estimates for the proposed and current values of the likelihood only differ by the random numbers in one block. Our approach is less general than that of Deligiannidis et al. (2015), but has the following advantages. First, it provides a more direct way to control the correlation between the logarithms of the estimates of the likelihood at the current and proposed values of the parameters. Second, the mathematical properties of the method are simplified and made more transparent compared to the treatment in Deligiannidis et al. (2015). Third, blocking is shown to be a natural way to carry out PMMH in, for example, panel data models and subsampling problems. We obtain theory and guidelines for selecting the optimal number of particles, and document large speed-ups in a panel data example and a subsampling problem.

Keywords. Intractable likelihood; Unbiasedness; Panel data; Data subsampling.

*Discipline of Business Analytics, University of Sydney
†School of Economics, UNSW School of Business
‡Department of Computer and Information Science, Linköping University
§Research Division, Sveriges Riksbank

1 Introduction

In many statistical applications the likelihood is analytically or computationally intractable, making it difficult to carry out Bayesian inference. An example of models where the likelihood is often intractable are generalized linear mixed models (GLMM) for longitudinal data, where random effects are used to account for the dependence between the observations measured on the same individual (Fitzmaurice et al., 2011; Bartolucci et al., 2012). The likelihood is intractable because it is an integral over the random effects, but it can be easily estimated unbiasedly using importance sampling. The second example that uses a variant of the unbiasedness idea, is that of unbiasedly estimating the log likelihood by subsampling, as in Quiroz et al. (2016). Subsampling is useful when the log likelihood is a sum of terms, with each term in the log likelihood expensive to evaluate, or when there are a very large number of such terms. Quiroz et al. (2016) estimate the log-likelihood unbiasedly in this way and then bias correct the resulting likelihood estimator to use within an pseudo marginal Metropolis-Hastings algorithm. They show analytically that the resulting posterior distribution of the parameters is a perturbation of the true target distribution, with the perturbation error being very small. State space models are a third class of models where the likelihood is often intractable but can be unbiasedly estimated using an importance sampling estimator (Shephard and Pitt, 1997; Durbin and Koopman, 1997) or by a particle filter estimator (Del Moral, 2004; Andrieu et al., 2010).

It is now well known in the literature that a direct way to overcome the problem of working with an intractable likelihood is to estimate the likelihood unbiasedly and use this estimate within an Markov chain Monte Carlo (MCMC) simulation on an expanded space that includes the random numbers used to construct the likelihood estimator. This was first considered by Lin et al. (2000) in the Physics literature and Beaumont (2003) in the Statistics literature; it was formally studied in Andrieu and Roberts (2009), who called the method pseudo-marginal Metropolis-Hastings (PMMH) and gave conditions for the chain to converge. Andrieu et al. (2010) use MCMC for doing inference in state space models where the likelihood is estimated unbiasedly by the particle filter. Flury and Shephard (2011) give an excellent discussion with illustrative examples of PMMH. Pitt et al. (2012) and Doucet et al. (2015) analyse the effect of estimating the likelihood and show that the variance of the log-likelihood estimator should be around 1 to obtain an optimal tradeoff between the efficiency of the Markov chain and the computational cost. See also Sherlock et al. (2015), who consider random walk proposals for the parameters, and show that the optimal variance of the log of the likelihood estimator can be somewhat higher in this case.

A key issue in estimating models by standard PMMH is that the variance of the log of the estimated likelihood grows linearly with the number of observations T. Hence, to keep the variance of the log of the estimated likelihood small and around 1 it is necessary for the number of particles N, used in constructing the likelihood estimator, to increase in proportion to T, which means that PMMH requires $O(T^2)$ operations at every MCMC iteration. In a seminal paper, Deligiannidis et al. (2015) propose correlating the random numbers used in constructing the estimators of the likelihood at the current and proposed values of the parameters. They show that by inducing a high correlation between these ensembles of random numbers it is only necessary to increase the number of particles N in proportion to $T^{\frac{1}{2}}$, reducing the PMMH algorithm to $O(T^{3/2})$ per iteration. This is a tremendous breakthrough in the ability of PMMH to be competitive with more traditional MCMC methods. However, at this stage

their derivations are essentially limited to panel data models.

Our article proposes an alternative approach to Deligiannidis et al. (2015) called the block-wise PMMH that also requires a much smaller number of particles for optimal performance. The block-wise PMMH divides the set of pseudo-random numbers into blocks and updates one block at a time, thus reducing the variation in the Metropolis-Hastings acceptance probability. This helps the chain to mix well even if highly variable estimates of the likelihood are used. As a result, only a small number of particles is needed at every iteration. We obtain theory and guidelines for selecting an optimal number of particles in the block-wise PMMH. The block-wise PMMH is not as general as the correlated PMMH method in Deligiannidis et al. (2015), but has a number of advantages. First, the correlation between the proposed and current values of the log likelihood estimators is controlled directly rather than indirectly and nonlinearly through the correlated ensembles of random numbers. Second, the theoretical justification and optimality properties of the block-wise PMMH are more transparent and easier to prove than the derivations in Deligiannidis et al. (2015). Third, blocking is a natural way to carry out a dependent PMMH in many problems such as panel data models and subsampling problems. Fourth, we show that the optimal number of particles required at each iteration of the block-wise PMMH is also $O(T^{3/2})$.

2 Block-wise pseudo-marginal Metropolis-Hastings algorithm

2.1 The PMMH algorithm

Let y be a set of observations with density $p(y|\theta)$, where $\theta \in \Theta$ is the vector of unknown parameters. We are interested in sampling from the posterior $\pi(\theta) \propto p(\theta)p(y|\theta)$ in models where the likelihood $p(y|\theta)$ is analytically or computationally intractable. We assume that $p(y|\theta)$ can be unbiasedly estimated by $\widehat{p}_N(y|\theta) = \widehat{p}_N(y|\theta,u)$, with u the set of independent pseudo-random uniform or standard normal variables used to compute $\widehat{p}_N(y|\theta)$; N is the number of importance samples or total number of particles used and the dimension of u is proportional N. Denote the density function of u by $p_N(u)$ and define a joint density of θ and u as

$$\pi_N(\theta,u) := p(\theta)\widehat{p}_N(y|\theta,u)p_N(u)/p(y). \tag{1}$$

Then, $\pi_N(\theta,u)$ admits $\pi(\theta)$ as its marginal density because $\int \widehat{p}_N(y|\theta,u)p(u)du = p(y|\theta)$ by unbiasedness. Therefore, we obtain samples from the posterior $\pi(\theta)$ by sampling from $\pi_N(\theta,u)$.

Let $q(\theta|\theta')$ be a proposal density for θ, conditional on θ', with θ' the current state. Let u' be the corresponding current set of pseudo-random numbers used to compute $\widehat{p}_N(y|\theta',u')$. The standard PMMH algorithm generates samples from $\pi(\theta)$ by generating a Markov chain with invariant density based on $\pi_N(\theta,u)$ using the Metropolis-Hastings algorithm with proposal density $q(\theta,u|\theta',u') = q(\theta|\theta')p_N(u)$. The proposal (θ,u) is accepted with probability

$$\alpha(\theta',u';\theta,u) := \min\left(1, \frac{\pi_N(\theta,u)}{\pi_N(\theta',u')}\frac{q(\theta',u'|\theta,u)}{q(\theta,u|\theta',u')}\right) = \min\left(1, \frac{p(\theta)\widehat{p}_N(y|\theta,u)}{p(\theta')\widehat{p}_N(y|\theta',u')}\frac{q(\theta'|\theta)}{q(\theta|\theta')}\right), \tag{2}$$

which is computable and it is usually unnecessary to store u and u'. In the standard PMMH scheme, a new independent set of pseudo-random numbers u is generated each time the likelihood estimate is computed.

For the standard PMMH algorithm, Pitt et al. (2012) and Doucet et al. (2015) show that the variance of $\log \widehat{p}_N(y|\theta,u)$ should be around 1 in order to obtain an optimal tradeoff between the computational cost and efficiency of the Markov chain in θ and u. However, in some problems it may be prohibitively expensive to take an N large enough to ensure that $\mathbb{V}(\log \widehat{p}_N(y|\theta,u)) \approx 1$.

2.2 The block-wise PMMH algorithm

In our block-wise PMMH algorithm, instead of generating a completely new set u when estimating the likelihood as previously done in the literature, we update u in blocks. Let us divide the set of variables u into G sets $u_{(1)},...,u_{(G)}$. The extended target (1) can be re-written as

$$\pi_N(\theta, u_{(1)}, ..., u_{(G)}) = p(\theta)\widehat{p}_N(y|\theta, u_{(1)}, ..., u_{(G)})p_N(u_{(1)}, ..., u_{(G)})/p(y). \tag{3}$$

Instead of updating the full set of (θ,u) as in the standard PMMH, we propose to update θ and a block $u_{(K)}$ at a time. The block index K is randomly selected from $1,...,G$ with $P(K=k)>0$ for every $k=1,...,G$. Typically, $P(K=k)=1/G$. This is similar to component-wise MCMC whose convergence is well established in the literature; see, e.g., Johnson et al. (2013). The acceptance probability (2) becomes

$$\min\left(1, \frac{p(\theta)\widehat{p}_N(y|\theta, u'_{(1)}, ..., u_{(k)}, ..., u'_{(G)})}{p(\theta')\widehat{p}_N(y|\theta', u'_{(1)}, ..., u'_{(k)}, ..., u'_{(G)})} \frac{q(\theta'|\theta)}{q(\theta|\theta')}\right). \tag{4}$$

Intuitively, by fixing all $u_{(j)}$'s except $u'_{(k)}$, the variation in the ratio of the likelihood estimates is reduced, which helps the chain to mix well.

3 Analysis of the block-wise PMMH

We now suppose that the likelihood can be written as a product of G independent terms,

$$p(y|\theta) = \prod_{k=1}^{G} p(y_{(k)}|\theta). \tag{5}$$

Example: panel data models. Consider a panel data model with T panels, which we divide into G groups $y_{(1)},...,y_{(G)}$ with approximately T/G panels in each.

Example: big data. Consider a big data set with T independent observations, which we divide into G groups $y_{(1)},...,y_{(G)}$ with T/G observations in each.

We assume that the k^{th} likelihood term $p(y_{(k)}|\theta)$ is estimated unbiasedly by $\widehat{p}_{N_{(k)}}(y_{(k)}|\theta,u_{(k)})$, where the $u_{(k)}$ are independent with $u_{(k)} \sim p_{N_{(k)}}(\cdot)$, and $N_{(k)}$ is the number of particles or importance samples or the size of $u_{(k)}$. Let $N := N_{(1)} + \cdots N_{(G)}$. An unbiased estimator of the likelihood is

$$\widehat{p}_N(y|\theta) := \prod_{k=1}^{G} \widehat{p}_{N_{(k)}}(y_{(k)}|\theta,u_{(k)}).$$

We now follow Pitt et al. (2012) and define $z_{(k)} = z_{(k)}(\theta,u_{(k)}) := \log \widehat{p}_{N_{(k)}}(y_{(k)}|\theta,u_{(k)}) - \log p(y_{(k)}|\theta)$ as the error in the log likelihood estimate of the kth block. Then, $z = z(\theta,u) := \log \widehat{p}_N(y|\theta,u) - \log p(y|\theta) = \sum_{k=1}^G z_{(k)}$ is the error in the log-likelihood estimate. The following result holds, whose proof is straightforward and therefore omitted.

Lemma 1. *Suppose that $u_{(1)},...,u_{(G)}$ are independent and generated from $p_{N_k}(u_{(k)})$ and suppose that $g_{N_k}(z_{(k)};\theta)$ is the distribution of the corresponding $z_{(k)}$. Then, the distribution of $\theta,z_{(1)},...,z_{(G)}$ based on (1) is*

$$\pi_N(\theta,z_{(1)},...,z_{(G)}) = \pi(\theta)\prod_{k=1}^G \exp(z_{(k)}) g_{N_{(k)}}(z_{(k)}|\theta), \tag{6}$$

which means that $\pi_{N_{(k)}}(z_{(k)}|\theta) = \exp(z_{(k)}) g_{N_k}(z_{(k)}|\theta)$ and the $z_{(k)}$ are independent conditional on θ, i.e., in $\pi_N(\cdot|\theta)$

We note that it is straightforward to show that the acceptance probability (2) can be rewritten as

$$\min\left\{1,\exp(z-z')\frac{\pi(\theta)}{\pi(\theta')}\frac{q(\theta'|\theta)}{q(\theta|\theta')}\right\} \tag{7}$$

where $z' = z(\theta',u')$. This is also the acceptance probability if we consider a Metropolis-Hastings scheme that samples from

$$\pi_N(\theta,z) = \pi(\theta)e^z g_N(z|\theta), \tag{8}$$

with $g_N(z|\theta)$ the density of z. Thus, as noted in Pitt et al. (2012), the properties of the Metropolis-Hastings scheme based on $\pi_N(\theta,u)$ are the same as those based on $\pi_N(\theta,z)$. As we show below, it will be easier to work with $\pi_N(\theta,z)$ than with $\pi_N(\theta,u)$ because z is a scalar.

Instead of updating θ and u as in the standard PMMH, the block-wise PMMH updates θ and a single block, $u_{(k)}$. The terms z and z' in the acceptance probability (7) are $z = \sum_{j=1,j\neq k}^G z_{(j)}(\theta,u_{(j)} = u'_{(j)}) + z_{(k)}(\theta,u_{(k)})$, $z' = \sum_{j=1}^G z_{(j)}(\theta',u'_{(j)})$. We use the following notation: $w \sim \mathfrak{N}(a,b^2)$ means that w has a normal distribution with mean a and variance b^2, and denote the density of w as $\mathfrak{n}(w;a,b^2)$.

We make the following assumptions.

Assumption 1. *Suppose that $u_{(1)},...,u_{(G)}$ are independent and generated from $p_{N_{(k)}}(u_{(k)})$ and,*

(i) For each group k, there is a $\gamma^2_{(k)}(\theta) > 0$ and an N_k such that

$$z_{(k)}(\theta,u_{(k)}) \sim \mathfrak{N}\left(-\frac{\gamma^2_{(k)}(\theta)}{2N_k}, \frac{\gamma^2_{(k)}(\theta)}{N_k}\right).$$

where N_k is not necessarily the same as $N_{(k)}$.

(ii) For a given $\sigma^2 > 0$, let N_k be a function of θ, σ^2 and G such that $\mathbb{V}(z_{(k)}(\theta,u_{(k)})) = \sigma^2/G$, i.e. $N_k = N_k(\theta,\sigma^2,G) = G\gamma^2_{(k)}(\theta)/\sigma^2$. This means that $z_{(k)}(\theta,u_{(k)}) \sim \mathfrak{N}(-\sigma^2/(2G),\sigma^2/G)$ for each k.

We note that $N_{(k)}$ is the total number of particles used for the kth group, and will usually be different than N_k. In panel data models, $N_{(k)} = (T/G)N_k$. Assumption 1 is a stylized set of assumptions that will hold approximately when the N_k are sufficiently large. Part (i) is similar to the assumption made in Pitt et al. (2012) and ensures that if $z_{(k)}$ is normally distributed then the expected value $\mathbb{E}(\exp(z_{(k)})) = 1$ for each k; this is consistent with the estimator $\widehat{p}(y_{(k)}|\theta)$ being unbiased. For panel data models, part (i) is justified by Lemma 6. Assumption 1(ii) can be enforced for most panel data models and subsampling approaches because it is straightforward to estimate the variance of $z_{(k)}$ accurately for each k and θ.

We can now obtain the following lemma whose proof is straightforward and omitted.

Lemma 2. *Suppose that parts (i) and (ii) of Assumption 1 hold and* $\theta', u'_{(j)}, j = 1,...,G$ *come from* $\pi_N(\theta', u'_{(1)},...,u'_{(G)})$. *Define* $z' := \sum_{j=1}^{G} z_{(j)}(\theta', u'_{(j)})$ *and* $z := \sum_{j=1, j\neq k}^{G} z_{(j)}(\theta, u_{(j)} = u'_{(j)}) + z_{(k)}(\theta, u_{(k)})$, *where* $u_{(k)}$ *is generated from* $p_{N_k}(u_{(k)})$. *Let* $\rho = 1 - 1/G$. *Then,*

(i)

$$z' \sim \mathfrak{N}\left(\frac{\sigma^2}{2}, \sigma^2\right) \quad \textit{and} \quad z \sim \mathfrak{N}\left(\frac{\sigma^2(2\rho - 1)}{2}, \sigma^2\right).$$

(ii) $\mathrm{Corr}(z, z') = \rho$.

(iii)

$$z|z' \sim \mathfrak{N}\left(-\frac{\sigma^2}{2}(1-\rho) + \rho z', \sigma^2(1-\rho^2)\right).$$

We follow Pitt et al. (2012) and also assume that the proposal for θ is perfect. That is,

Assumption 2. $q(\theta|\theta') = \pi(\theta)$.

This assumption helps identify the effect of estimating the likelihood, as we assume a perfect proposal, and helps to simplify the derivation of the derivation of the guidelines for the optimal number of particles. It follows from Pitt et al. (2012) and Doucet et al. (2015) that this assumption results in a choice of σ that may be a little too low and so the number of particles required may be set a little too high, than optimal. However, we have found that this conservative strategy works well in practice because it makes the empirical applications more robust to the assumptions. Under Assumption 2, the acceptance probability (7) of the Metropolis-Hastings scheme becomes

$$\alpha(z', z; \rho, \sigma) = \min\left(1, e^{z - z'}\right), \tag{9}$$

The next lemma gives the conditional and unconditional acceptance probabilities of the Metropolis-Hastings scheme for z and θ and is proved in Appendix A

Lemma 3. *Suppose Assumptions 1 and 2 hold and* $\rho = 1 - 1/G$.

(i) The acceptance probability of the Metropolis-Hastings scheme conditional on $z'=z(\theta',u')$ is

$$\mathrm{P}(accept|z',\rho,\sigma)=\exp(-x+\tau^2/2)\Phi\Big(\frac{x}{\tau}-\tau\Big)+\Phi\Big(\frac{-x}{\tau}\Big)$$

with $x:=\left(z'+\frac{\sigma^2}{2}\right)(1-\rho)$ *and* $\tau=\sigma\sqrt{1-\rho^2}$.

(ii) The unconditional acceptance probability of the Metropolis-Hastings scheme is

$$\mathrm{P}(accept|\rho,\sigma)=2\left(1-\Phi\Big(\frac{\sigma\sqrt{1-\rho}}{\sqrt{2}}\Big)\right). \tag{10}$$

Suppose that we are interested in estimating $\pi(\varphi)=\int\varphi(\theta)\pi(\theta)d\theta$ for some scalar function $\varphi(\theta)$ of θ. Let $\{\theta^{[j]},z^{[j]},j=1,...,M\}$ be the draws obtained from the PMMH sampler after it has converged, and let the estimator of $\pi(\varphi)$ be $\widehat{\pi}(\varphi):=\frac{1}{M}\sum\varphi(\theta^{[j]})$. Then, we define the inefficiency of the estimator $\widehat{\pi}(\varphi)$ relative to an estimator based on an i.i.d. sample from $\pi(\theta)$ (as in Assumption 2) as

$$\mathrm{IF}(\varphi,\sigma,\rho):=\lim_{M\to\infty}M\mathbb{V}_{\mathrm{PMMH}}(\widehat{\pi}(\varphi))/\mathbb{V}(\varphi|y), \tag{11}$$

where $\mathbb{V}_{\mathrm{PMMH}}(\widehat{\pi}(\varphi))$ is the posterior variance of the estimator $\widehat{\pi}(\varphi)$ and $\mathbb{V}(\varphi|y):=\mathbb{E}_\pi(\varphi(\theta)^2)-[\mathbb{E}_\pi(\varphi(\theta))]^2$ is the posterior variance of φ so that $\mathbb{V}(\varphi|y)/M$ is the variance of the ideal estimator when $\theta^{[j]}\overset{iid}{\sim}\pi(\theta)$. We obtain the following result which shows that under our assumptions the inefficiency $\mathrm{IF}(\varphi,\sigma,\rho)$ is independent of φ and is a function only of σ and $\rho=1-1/G$. The proof is Appendix A. We call $\mathrm{IF}(\sigma,\rho)$ the inefficiency of the PMMH algorithm, for given ρ and σ, because under our assumptions it does not depend φ. From (8) and Lemma 2, we note that the posterior density of z conditional on θ is $\pi_N(z|\theta)=e^z g_N(z|\theta)=\mathfrak{n}(z;\sigma^2/2,\sigma^2)$, which does not depend on θ and is denoted $\pi(z|\sigma)$.

Lemma 4. *The inefficiency is given by*

$$\mathrm{IF}(\sigma,\rho)=1+2\mathbb{E}_{z'\sim\pi(z'|\sigma)}\left(\frac{1-k(z'|\sigma,\rho)}{k(z'|\sigma,\rho)}\right), \tag{12}$$

where $k(z'|\rho,\sigma)=\Pr(accept|z',\rho,\sigma)$ *is the acceptance probability of the MCMC scheme conditional on the previous iterate* z' *and is given by part (i) of Lemma 3.*

Similarly to Pitt et al. (2012), we obtain in Appendix B the computing time of the sampler as

$$\mathrm{CT}(\sigma,\rho):=\frac{\mathrm{IF}(\sigma,\rho)}{\sigma^2} \tag{13}$$

which takes into account the computing time and the mixing rate of the PMMH chain.

To simplify the notation in this section we often do not show dependence on ρ as it is assumed constant. In Section 5 we show that if we take $G=O(T^{\frac{1}{2}})$, then $\rho=1-O(T^{-\frac{1}{2}})$ and $N_k=O(T^{\frac{1}{2}})$ are optimal.

The next lemma shows the optimal σ under our assumptions as well as the corresponding acceptance rate. Its proof is in Appendix A.

Lemma 5. *Given that* $\rho = 1-1/G$ *is close to 1, the optimal* σ_{opt} *that minimizes* $\text{CT}(\sigma)$ *is approximately* $\sigma_{\text{opt}} \approx 2.16/\sqrt{1-\rho^2}$. *The unconditional acceptance rate* (10) *under this optimal choice of the tuning parameters is approximately 0.28.*

Our theory indicates that the efficiency of the Markov chain increases with G. Although one could set G to its maximum value T, using a G that is too large leads to the situation that some blocks $u_{(k)}$ might be not updated, and the conditions of Assumption 1 may not be satisfied because we require N_k reasonably large. The Markov chain then depends on the initial set of u and the PMMH may produce samples not from the correct target posterior. Let M be the length of the generated Markov chain. The average number of times that a block $u_{(k)}$ is updated is M/G. In general, G should be selected such that M/G is not too small. In the examples in this paper, if not otherwise stated, we set $G=100$, as we found that the efficiency is relatively insensitive to larger values of G.

A toy example. Suppose that we wish to sample from $\pi(\theta,z) = \pi(\theta)e^z g(z|\sigma)$ in which θ is the parameter of interest, with $\pi(\theta) = \mathfrak{N}(0,1)$ and $g(z|\sigma) = \mathfrak{N}(-\sigma^2/2,\sigma^2)$. Suppose further that z is divided into blocks $z = \sum_{k=1}^{G} z_{(k)}$ with $z_{(k)} \overset{iid}{\sim} \mathfrak{N}(-\sigma_G^2/2,\sigma_G^2)$, $\sigma_G^2 = \sigma^2/G$ and $G=100$.

First, we consider two sampling schemes, the standard PMMH and the block-wise PMMH, to sample from $\pi(\theta,z)$ with $\sigma_G^2 = 2.34$, i.e. $\sigma^2 = 234$. Suppose that (θ',z') is the current state. The proposal (θ,z) in the standard PMMH is generated by $\theta \sim \pi(\theta)$ and $z \sim g(z|\sigma)$. The proposal (θ,z) in the block-wise PMMH scheme is generated as follows. Let $z' = \sum_{k=1}^{G} z'_{(k)}$ be the current z-state and let k be an index uniformly generated from the set $\{1,...,G\}$. Sample $z_{(k)} \sim \mathfrak{N}(-\sigma_G^2/2,\sigma_G^2)$ and let $z = \sum_{j\neq k} z'_{(k)} + z_{(k)}$ be the proposal. Both schemes accept (θ,z) with probability $\min(1,e^{z-z'})$.

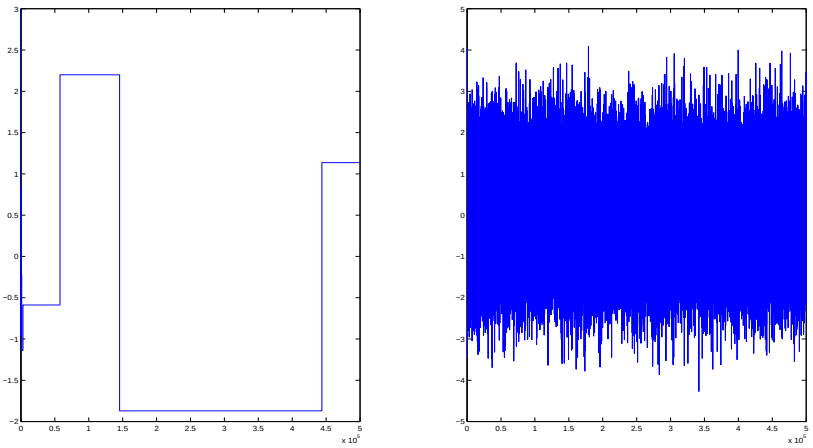

Figure 1: The samples of θ generated by the standard PMMH scheme (left) and the block-wise PMMH scheme (right). Both chains are initialized at 3 and run for 500,000 iterations.

Figure 1 plots the θ-samples generated by the standard PMMH scheme (left panel) and by the block-wise PMMH scheme (right panel). As expected, the standard PMMH chain is sticky because of the big variance of z, $\sigma^2 = 234$.

Now, in order to study the effect of σ^2 on the acceptance rate and computing time $\text{CT}(\sigma)$ of the sampler, Figure 2 shows the $\text{CT}(\sigma)$ and acceptance rates for various values of σ^2. The

figure shows that CT(σ) has a minimum value of 0.0263 at $\sigma^2 = 234$, where the acceptance rate is 0.279, which confirm the theory. Among all standard PMMH schemes with different σ^2, Pitt et al. (2012) show that the optimal scheme is the one with $\sigma^2 = 1$. We also run this optimal standard PMMH scheme and obtain a computing time value of CT($\sigma = 1$) = 5.32. Hence, the optimal block-wise PMMH is 5.32/0.0263 approximately 202 times more efficient than the optimal standard PMMH.

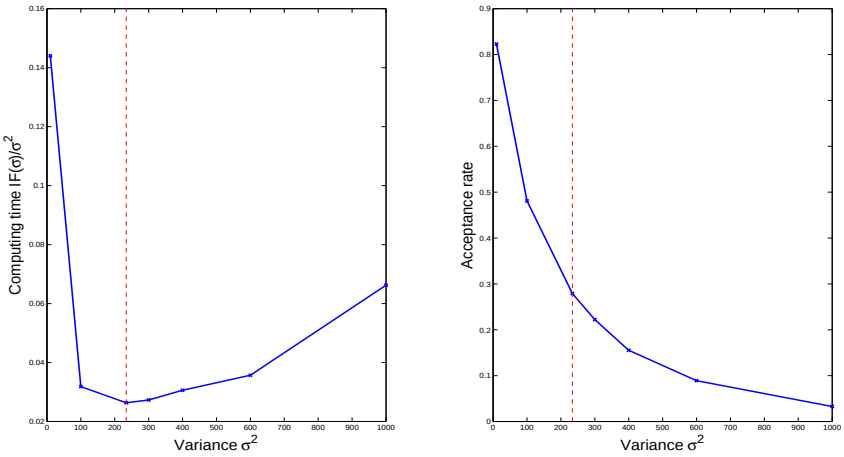

Figure 2: The left panel shows the computing time CT(σ) and the right panel shows the acceptance rate v.s. the variance σ^2. The dashed lines indicate the values w.r.t. the optimal variance $\sigma^2_{\text{opt}} = 234$.

3.1 Guidelines for selecting the optimal number of particles for panel data

We now consider the panel data case with the likelihood factorized as in (5). From Lemma 5, the optimal $\sigma^2_{\text{opt}} = 2.16^2/(1-\rho^2)$, and the optimal variance of the log-likelihood estimator based on each group is $\sigma^2_{\text{opt}}/G = 2.16^2/(1+\rho)$ is approximately 2.34 given that G approximately 100 is large. Hence, for each group k, we propose tuning the number of particles $N_{(k)} = N_{(k)}(\theta)$ such that $\mathbb{V}(z_{(k)}|\theta, N_{(k)})$ is approximately 2.34.

4 Applications

4.1 Panel data

A clinical trial is conducted to test the effectiveness of beta-carotene in preventing non-melanoma skin cancer (Greenberg et al., 1989). Patients were randomly assigned to a control or treatment group and biopsied once a year to ascertain the number of new skin cancers since the last examination. The response y_{ij} is a count of the number of new skin cancers in year j for the ith subject. Covariates include age, skin (1 if skin has burns and 0 otherwise), gender, exposure (a count of the number of previous skin cancers), year of follow-up and treatment (1

if the subject is in the treatment group and 0 otherwise). There are $m=1683$ subjects with complete covariate information.

Following Donohue et al. (2011) we consider the mixed Poisson model with a random intercept

$$\begin{aligned} p(y_{ij}|\beta,\alpha_i) &= \text{Poisson}(\exp(\eta_{ij})), \\ \eta_{ij} &= \beta_0+\beta_1\text{Age}_i+\beta_2\text{Skin}_i+\beta_3\text{Gender}_i+\beta_4\text{Exposure}_{ij}+\alpha_i, \end{aligned}$$

where $\alpha_i \sim \mathfrak{N}(0,\sigma^2)$, $i=1,....,m=1683$, $j=1,...,n_i=5$. The likelihood is

$$p(y|\theta) = \prod_{i=1}^{m} p(y_i|\theta), \;\; p(y_i|\theta) = \int \left(\prod_{j=1}^{n_i} p(y_{ij}|\beta,\alpha_i) \right) p(\alpha_i|\sigma^2) d\alpha_i$$

with $\theta=(\beta,\sigma^2)$ the vector of model parameters to be estimated.

We run both the optimal standard PMMH and the optimal block-wise PMMH for 50,000 iterations with the first 10,000 discarded as burnins. For simplicity, each likelihood $p(y_i|\theta)$ is estimated by importance sampling with the natural importance sampler $p(\alpha_i|\sigma^2)$. For the standard PMMH, the number of particles is tuned so as the variance of the log-likelihood estimator is not bigger than 1. In the block-wise PMMH, we divide the data into $G=99$ groups, so that each group has 17 data points, and the variance of log-likelihood estimator in each group is tuned to not be bigger than 2.344. We note that the structure of the data in this example allows us to select different number of particles N_i for each individual y_i. Figure 3 plots the samples generated by the two algorithms.

As performance measures, we report the acceptance rate, the integrated autocorrelation time (IACT), the CPU times, and the computing time (CT). For a univariate Markov chain, the IACT is estimated by

$$\text{IACT} = 1 + 2\sum_{t=1}^{1000} \widehat{\rho}_t,$$

where $\widehat{\rho}_t$ are the sample autocorrelations. The IACT for a multivariate chain is averaged over the IACT values for the parameters. The computing time is the product of the IACT and the CPU time.

Table 1 summarizes the acceptance rates, the IACT ratio, the CPU ratio, and the CT ratio, with the blockwise PMMH as the baseline. As shown, the block-wise PMMH outperforms the standard PMMH. In particular, the block-wise PMMH is around 25 times more efficient than the standard PMMH in terms of computing time.

Methods	Acceptance	IACT ratio	CPU ratio	CT ratio
Standard PMMH	0.222	1.080	23.095	24.938
Block-wise PMMH	0.243	1	1	1

Table 1: Skin cancer example using the block-wise PMMH as the baseline

4.2 Subsampling MCMC

Quiroz et al. (2016) propose a data subsampling approach, based on the correlated PMMH algorithm in Deligiannidis et al. (2015), to speed up MCMC. They use a Gaussian copula to

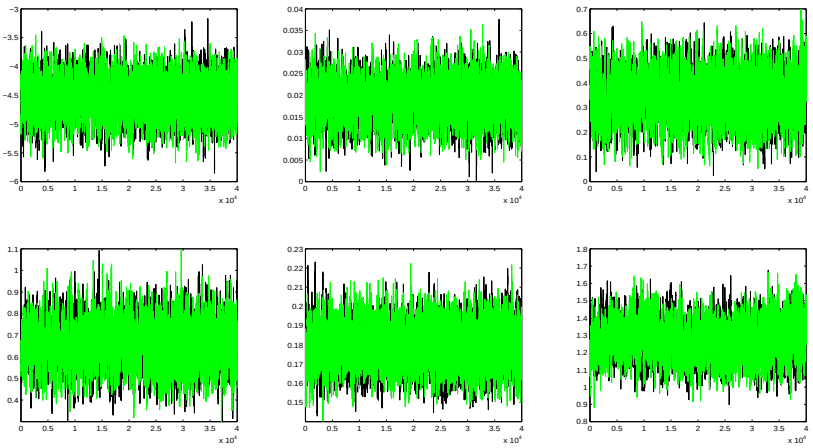

Figure 3: Plots of the standard PMMH chain (black) and block-wise PMMH chain (green) for the 6 parameters.

induce a high correlation between the binary vectors u' and u at the current and proposed draws, respectively. The ith element of u is a binary indicator determining if observation i is included in the subsample for estimating the likelihood. We now present an alternative subsampling approach based on block-wise PMMH.

Let

$$u=(u_1,...,u_N), \quad u_i \in \{1,...,T\} \text{ and independent for } i=1,...,N,$$

where N is the subsample size and u represents a vector of observation indices rather than selection indicators as in Quiroz et al. (2016); u is therefore of length N rather than T as in Quiroz et al. (2016). Block-wise PMMH creates G blocks, each with N/G indices. For large G, a high correlation in the MH log-ratio is naturally induced, since u and u' differ only at a small number of positions.

As in Quiroz et al. (2016), we consider the following two AR(1) processes with Student-t iid errors $\epsilon_t \sim t(\nu)$ with known degrees of freedom ν. The data generating processes are

$$y_t \quad = \begin{cases} \beta_0+\beta_1 y_{t-1}+\epsilon_t & ,[\mathrm{M}_1, \theta=(\beta_0=0.3,\beta_1=0.6)] \\ \mu+\rho(y_{t-1}-\mu)+\epsilon_t & ,[\mathrm{M}_2, \theta=(\mu=0.3,\rho=0.99)] \end{cases} \tag{14}$$

where $p(\epsilon_t) \propto (1+\epsilon_t^2/\nu)^{-(\nu+1)/2}$ with $\nu=5$ and the uniform priors

$$p(\beta_0,\beta_1) \overset{\text{ind.}}{=} \mathcal{U}(-5,5)\cdot\mathcal{U}(0,1) \quad \text{and} \quad p(\mu,\rho) \overset{\text{ind.}}{=} \mathcal{U}(-5,5)\cdot\mathcal{U}(0,1).$$

Define $\ell_t(\theta) := \log p(y_t|y_{t-1},\theta)$ and rewrite the log likelihood $\ell(\theta)$ as

$$\ell(\theta)=q(\theta)+d(\theta), \quad q(\theta)=\sum_{t=1}^{T} q_t(\theta), \quad d(\theta)=\sum_{t=1}^{T} d_t(\theta), \text{ with } d_t(\theta)=\ell_t(\theta)-q_t(\theta),$$

where $q_t(\theta) \approx \ell_t(\theta)$ is a control variate. We set $q_t(\theta)$ to a Taylor series approximation of $l_t(\theta)$ evaluated at the nearest centroid from a clustering in data space. This has the effect

of reducing the complexity of computing $q(\theta)$ from $O(T)$ to $O(C)$, where C is the number of centroids (Quiroz et al., 2016). Now, an unbiased estimate of $\ell(\theta)$ based on a simple random sample with replacement is

$$\widehat{\ell}(\theta)=\frac{T}{N}\sum_{i=1}^{N}d_{u_i}(\theta)+q(\theta),\quad \text{with } \Pr(u_i=t)=\frac{1}{T},\quad t=1,...,T. \tag{15}$$

The resulting likelihood estimate is $\widehat{p}_N(y|\theta,u)=\exp\Big(\widehat{\ell}(\theta)-\widehat{\sigma}^2(\theta)/2\Big)$, where $\widehat{\sigma}^2(\theta)$ is an unbiased estimate of the variance of (15). Quiroz et al. (2016) show that carrying out MCMC with this slightly biased likelihood estimator samples from a perturbed posterior that is very close to the correct posterior if N is sufficiently large in relation to $\sigma^2(\theta)$.

We generate $T=100{,}000$ observations from the models in (14) and run both the correlated PMMH and the block-wise PMMH for 55,000 iterations from which we discard the first 5,000 draws as burn-in. Using the same $\sigma^2(\theta)$ as in Quiroz et al. (2016) results in sample sizes N approximately 1300 and N approximately 2600 for models M_1 and M_2, respectively. For the block-wise PMMH we use $G=100$ so that each block has ≈ 13 observations for M_1 and and ≈ 26 observations for M_2. Also, following Quiroz et al. (2016), the persistence parameter in the correlated PMMH is set to $\phi=0.9999$, and we use a random walk proposal which is adapted during the burn-in phase to target $\alpha\approx 0.15$ (Sherlock et al., 2015).

Table 2 summarizes the performance measures introduced in Section 4.1. It is evident that the block-wise PMMH dramatically outperforms the correlated PMMH on the speed measures. This is because, as discussed above, the correlated PMMH requires T operations for generating the vector u. The block-wise PMMH moves only one block at a time, so that the update of the vector u requires T/G operations. Finally, Figure 4 plots the kernel density estimates of the block-wise PMMH and correlated PMMH vs the true posterior (obtained by standard MH) and shows that the estimates from both PMMH schemes are close to those from the MCMC.

Methods	Acceptance		IACT ratio		CPU ratio		CT ratio	
	M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2
Correlated PMMH	0.149	0.140	1.110	1.124	62.893	38.610	69.444	43.478
Block-wise PMMH	0.160	0.151	1	1	1	1	1	1

Table 2: Data subsampling example using block-wise PMMH as a baselin e.

5 Large sample in T analysis for panel data

In this section we discuss properties of the block-wise PMMH for large T for the panel data model discussed in Sections 3 and 4.1 and show that the total number of particles required per MCMC iteration is $O(T^{3/2})$ rather than $O(T^2)$ as in the standard PMMH. This result parallels that of Deligiannidis et al. (2015). We also justify part (i) of Assumption 1

Consider now the panel data model, with the panels in the kth group denoted by $\mathcal{G}_k$, and suppose that we use N_k particles for all panels $i\in\mathcal{G}_k$. Let $p(y_i|\theta)$ be the likelihood of the ith panel, and let $\widehat{p}_{N_i}(y_i|\theta,u_i)$ be the unbiased estimate of $p(y_i|\theta)$. We make the following assumption which will hold for most importance sampling estimates of the likelihood.

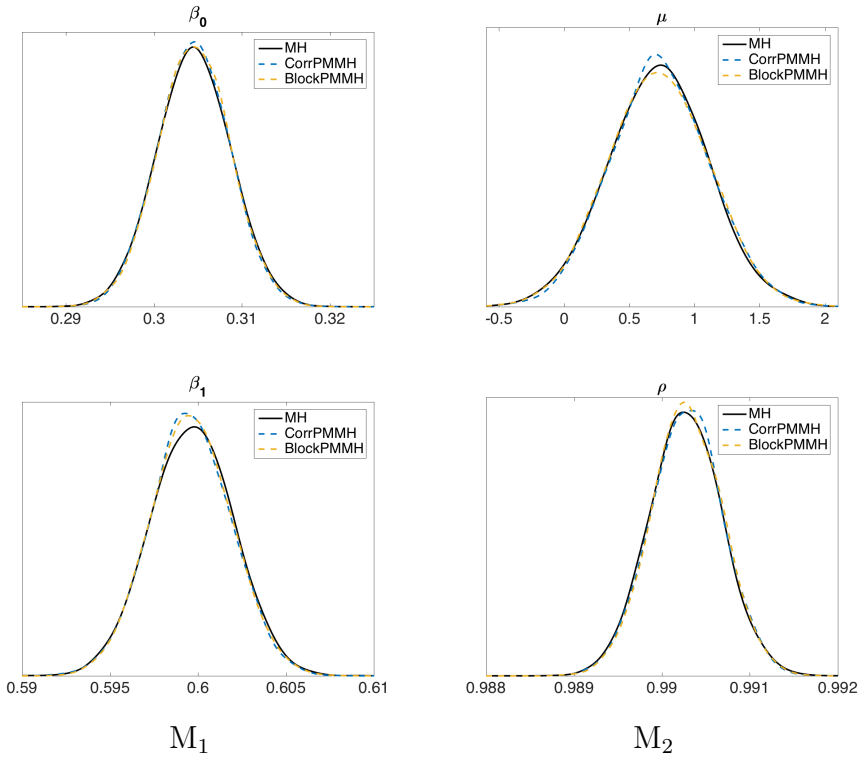

Figure 4: Kernel density estimates for correlated PMMH and blocking PMMH. The standard MH represents the true posterior

Assumption 3. *For each $i \in \mathcal{G}_k$ and parameter value θ, there exists a $A_i(\theta)^2$ such that for $N_k \to \infty$*

$$N_k^{\frac{1}{2}} \left(\widehat{p}_{N_i}(y_i;\theta,u_i) - p(y_i|\theta) \right) \overset{d}{\Rightarrow} \mathfrak{N}(0, A_i(\theta)^2) \tag{16}$$

Let $\gamma_i(\theta)^2 = A_i(\theta)^2 / p(y_i|\theta)^2$ and define $z_i(\theta,u_i) := \log\left(\widehat{p}_{N_i}(y_i;\theta,u_i)/p(y_i|\theta) \right)$. The following result holds. Its proof is straightforward and is omitted.

Lemma 6. *Suppose that Assumption 3 holds. Then, for $i \in \mathcal{G}_k$, as $N_k \to \infty$,*

$$N_k^{\frac{1}{2}} \left(z_i(\theta,u_i) + \frac{\gamma_i^2(\theta)}{2N_k} \right) \overset{d}{\Rightarrow} \mathfrak{N}(0, \gamma_i^2(\theta))$$

or, more informally,

$$z_i(\theta,u_i) \approx \mathfrak{N}\left(-\frac{\gamma_i^2(\theta)}{2N_k}, \frac{\gamma_i^2(\theta)}{N_k} \right)$$

The next corollary formalizes the results of this section. Its proof follows from the discussion immediately above and Lemma 6.

Corollary 1. *Define $\gamma_{(k)}^2(\theta) := \sum_{i \in \mathcal{G}_k} \gamma_i^2(\theta)$. Suppose that we take $G_T = O(T^{\frac{1}{2}})$, where the subscript T, here and below, indicates dependence on T. Then,*

(i) The number of elements in each group is $|\mathcal{G}_k| = O(T/G_T) = O(T^{\frac{1}{2}})$, $\gamma^2_{(k)}(\theta) = O(T/G_T) = O(T^{\frac{1}{2}})$, *and*

$$\rho_T = 1 - O(T^{-\frac{1}{2}}),\ \sigma^2_T = O(T^{\frac{1}{2}}) \quad \textit{and} \quad N_{k,T}(\theta, \tau^2) = O(T^{\frac{1}{2}}).$$

(ii) The total number of particles used to estimate the likelihood is $\sum_{k=1}^{G} N_{k,T}(\theta) \times (T/G) = O(T^{3/2})$.

(iii) Part (i) of Assumption 1 holds.

6 Conclusion

We have presented an efficient block-wise PMMH algorithm for Bayesian inference for models where the likelihood is a product as in panel data models and can be estimated unbiasedly or for big data problems where the log likelihood is a sum with many terms which can be estimated unbiasedly. The proposed algorithm divides the set of random numbers used to estimate the likelihood into blocks, and then updates the parameters of interest and each block at a time. The block-wise PMMH approach requires a much smaller number of particles in the likelihood estimation than the standard PMMH. Applications to panel data and subsampling MCMC strongly confirm the usefulness of the proposed methodology.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 72:1–33.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37:697–725.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, New York.

Deligiannidis, G., Doucet, A., and Pitt, M. (2015). The correlated pseudo-marginal method. Technical report.

Donohue, M. C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, 98:685–700.

Doucet, A., Pitt, M., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313.

Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84:669–684.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. John Wiley & Sons, Ltd, New Jersey, 2nd edition.

Flury, T. and Shephard, N. (2011). Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models. *Econometric Theory*, 1:1–24.

Greenberg, E. R., Baron, J. A., Stevens, M. M., Stukel, T. A., Mandel, J. S., Spencer, S. K., Elias, P. M., Lowe, N., Nierenberg, D. N., G., B., and Vance, J. C. (1989). The skin cancer prevention study: design of a clinical trial of beta-carotene among persons at high risk for nonmelanoma skin cancer. *Controlled Clinical Trials*, 10:153–166.

Johnson, A. A., Jones, G. L., and Neath, R. C. (2013). Component-wise Markov Chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Statistical Science*, 28(3):360–375.

Lin, L., Liu, K., and Sloan, J. (2000). A noisy Monte Carlo algorithm. *Physical Review D*, 61.

Pitt, M. K., Silva, R. S., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

Quiroz, M., Villani, M., and Kohn, R. (2016). Speeding up MCMC by efficient data subsampling. Technical report.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–667.

Sherlock, C., Thiery, A., Roberts, G., and Rosenthal, J. (2015). On the efficiency of the pseudo marginal random walk Metropolis algorithm. *The Annals of Statistics*, 43(1):238–275.

Appendix A Proofs

Proof of Lemma 3. To obtain the conditional acceptance probability, we use the following results,

$$\int_{-\infty}^{A} \exp(z)\mathfrak{n}(z;a,b^2)dz = \exp(a+b^2/2)\Phi\left(\frac{A-a-b^2}{b}\right) \tag{17}$$

$$\int_{A}^{\infty} \mathfrak{n}(z;a,b^2)dz = \Phi\left(\frac{a-A}{b}\right). \tag{18}$$

Now, as in Lemma 2, let $a(z') = \mathbb{E}(z|z') = -\sigma^2/2G + \rho z'$ and $\tau^2 = \mathbb{V}(z|z') = \sigma^2(1-\rho^2)$, so that the conditional density of z given z' is $\mathfrak{n}(z;a(z'),\tau^2)$, and using (17) and (18), the conditional

probability of acceptance is

$$\begin{aligned}\int \min(1,\exp(z-z'))\mathfrak{n}(z;a(z'),\tau^2)dz &= \int_{-\infty}^{z'} \exp(z-z')\mathfrak{n}(z;a(z'),\tau^2)dz + \int_{z'}^{\infty} \mathfrak{n}(z;a(z'),\tau^2)dz \\ &= \exp\left(a(z')-z'+\tau^2/2\right)\Phi\left(\frac{z'-a(z')-\tau^2}{\tau}\right)+\Phi\left(\frac{a(z')-z'}{\tau}\right) \\ &= \exp\left(-y+\tau^2/2\right)\Phi\left(\frac{y-\tau^2}{\tau}\right)+\Phi\left(\frac{-y}{\tau}\right),\end{aligned}$$

where $y := z'-a(z') = (1-\rho)(z'+\sigma^2/2)$. To obtain the unconditional acceptance probability, we need to take the expectation of the conditional probability with respect to z'. We note that $y \sim \mathfrak{N}\left(\frac{\sigma^2}{2G}, \frac{\sigma^2}{G^2}\right)$. Hence, the unconditional probability of acceptance is

$$\int \exp\left(-y+\tau^2/2\right)\Phi\left(\frac{y-\tau^2}{\tau}\right)\mathfrak{n}(y;\sigma^2/(2G),\sigma^2/G^2)dy + \int \Phi\left(\frac{-y}{\tau}\right)\mathfrak{n}(y;\sigma^2/(2G),\sigma^2/G^2)dy \tag{19}$$

To proceed further we use the following elementary results,

$$\exp(-y)\mathfrak{n}(y;a,b^2) = \exp(b^2/2-a)\mathfrak{n}(y;a-b^2,b^2) \tag{20}$$

$$\Phi\left(\frac{-a-c}{\sqrt{b^2+d^2}}\right) = \int \Phi\left(\frac{-y-a}{b}\right)\mathfrak{n}(y;c,d^2)dy \tag{21}$$

$$\Phi\left(\frac{-a+c}{\sqrt{b^2+d^2}}\right) = \int \Phi\left(\frac{y-a}{b}\right)\mathfrak{n}(y;c,d^2)dy. \tag{22}$$

Hence,

$$\begin{aligned}\exp\left(-y+\tau^2/2\right)\mathfrak{n}(y;\sigma^2/G,\sigma^2/G^2) &= \exp(\tau^2/2+\sigma^2/2G^2-\sigma^2/G)\mathfrak{n}(y;\sigma^2/G-\sigma^2/G^2,\sigma^2/G^2) \\ &= \mathfrak{n}(y;\rho\sigma^2/G,\sigma^2/G^2) \\ \int \Phi\left(\frac{y-\tau^2}{\tau}\right)\mathfrak{n}(y;\rho\sigma^2/G,\sigma^2/G^2)dy &= \Phi\left(\frac{\rho\sigma^2/G-\sigma^2(1-\rho^2)}{\sqrt{\sigma^2/G^2+\sigma^2(1-\rho^2)}}\right) \\ &= \Phi\left(-\frac{\sigma\sqrt{1-\rho}}{\sqrt{2}}\right) \\ \int \Phi\left(\frac{-y}{\tau}\right)\mathfrak{n}(y;\sigma^2/G,\sigma^2/G^2)dy &= \Phi\left(-\frac{\sigma\sqrt{1-\rho}}{\sqrt{2}}\right).\end{aligned}$$

□

Proof of Lemma 4. For notational simplicity, we write the proposal density $q(z|z';\rho,\sigma)$ as $q(z|z')$, the acceptance probability in (9) as $\alpha(z',z;\rho,\sigma)$ as $\alpha(z',z)$ and the acceptance probability $k(z'|\sigma,\rho)$, conditional on the previous iterate, as $k(z')$. Let $\{(\theta_j,z_j), j=1,...,M\}$ be iterates, after convergence, for the Markov chain produced by the PMMH sampling scheme. Then, the Markov transition distribution from (θ',z') to (θ,z) is

$$\begin{aligned}p(\theta',z';d\theta,dz) &= \alpha(z',z)\pi(\theta)q(z|z')d\theta dz + \left(1-\int \alpha(z',z^*)\pi(\theta^*)q(z^*|z')d\theta^* dz^*\right)\delta_{(\theta',z')}(\theta,z) \\ &= \alpha(z',z)\pi(\theta)q(z|z')d\theta dz + (1-k(z'|\sigma,\rho))\,\delta_{(\theta',z')}(d\theta,dz),\end{aligned}$$

$\delta_{\theta',z'}(d\theta,dz)$ is the probability measure concentrated at (θ',z').

Consider now the space of functions

$$\mathfrak{F}=\Big\{\widetilde{\varphi}:\widetilde{\Theta}=\Theta\otimes\mathbb{R}\mapsto\mathbb{R},\widetilde{\varphi}=\varphi(\theta)\psi(z),\pi(\varphi):=\mathbb{E}_{\theta\sim\pi(\theta)}(\varphi)=0,\pi(\varphi^2):=\mathbb{E}_{\theta\sim\pi(\theta)}(\varphi^2)<\infty,$$
$$\pi(\psi^2):=E_{z\sim\pi(z)}(\psi)^2<\infty\Big\}.$$

We define the operator $P:\mathfrak{F}\mapsto\mathfrak{F}$ as

$$\begin{aligned}(P\widetilde{\varphi})(\theta,z):&=\int\widetilde{\varphi}(\theta^*,z^*)p(\theta,z;\theta^*,z^*)d\theta^*dz^*\\&=\pi(\varphi)\int\psi(z)\alpha(z,z^*)q(z^*|z)dz^*+\widetilde{\varphi}(\theta)(1-k(z))\\&=\varphi(\theta)\psi(z)(1-k(z)).\end{aligned}$$

as $\pi(\varphi)=0$ by assumption. It is straightforward to check that $(P^j\widetilde{\varphi})(\theta,z)=\varphi(\theta)\psi(z)(1-k(z))^j$ and that $(P\widetilde{\varphi})(\theta_{j-1},z_{j-1})=E(\widetilde{\varphi}(\theta_j,z_j)|\theta_{j-1},z_{j-1})$. Hence, $(P^j\varphi)(\theta_0,\theta_0)=\widetilde{\varphi}(\theta_0,z_0)(1-k(z_0))^j$.

We now consider $\widetilde{\varphi}(\theta,z)=\varphi(\theta)\psi(z)$ with $\psi(z)\equiv1$ so that $\widetilde{\varphi}\in\mathfrak{F}$; suppose also that $(\theta_0,z_0)\sim\widetilde{\pi}_N$. Define $c_j:=\mathrm{Cov}(\widetilde{\varphi}(\theta_j,z_j),\widetilde{\varphi}(\theta_0,z_0))=\mathrm{Cov}(\varphi(\theta_j),\varphi(\theta_0))$. Then,

$$\begin{aligned}c_j&=\mathbb{E}\big(\widetilde{\varphi}(\theta_j,z_j)\widetilde{\varphi}(\theta_0,z_0)\big)\\&=\mathbb{E}_{(\theta_0,z_0)\sim\widetilde{\pi}_N}\big(\mathbb{E}(\widetilde{\varphi}(\theta_j,z_j)|\theta_0,z_0)\widetilde{\varphi}(\theta_0,z_0)\big)\\&=\mathbb{E}_{(\theta_0,z_0)\sim\widetilde{\pi}_N}\big((1-k(z_0))^j\widetilde{\varphi}(\theta_0,z_0)^2\big)\\&=\mathbb{E}_{z_0\sim\widetilde{\pi}_N(z)}\big((1-k(z_0))^j\big)\mathbb{E}_{\theta_0\sim\pi}(\varphi(\theta_0)^2\big)\end{aligned}$$

because z_0 only depends on σ by construction

$$=\mathbb{E}_{z_0\sim\widetilde{\pi}_N(z)}\big((1-k(z_0))^j\big)c_0.$$

The inefficiency IF is defined as

$$\mathrm{IF}=(c_0+2\sum_{j=1}^{\infty}c_j)/c_0=1+2\sum_{j=1}^{\infty}\mathbb{E}_{z\sim\widetilde{\pi}_N(z)}\Big(\big(1-k(z)\big)^j\Big)=1+2\mathbb{E}_{z\sim\widetilde{\pi}_N(z)}\Big(\frac{1-k(z)}{k(z)}\Big)$$

as required.

Proof of Lemma 5. From Lemma 2, $\pi(z'|\sigma)=\mathfrak{n}(z';\sigma^2/2,\sigma^2)$. Let $\omega:=[(1-\rho)(z'+\sigma^2/2)-\tau^2]/\tau$. Then,

$$\omega\sim\mathfrak{N}\Big(-\frac{\rho\tau}{1+\rho},\frac{1-\rho}{1+\rho}\Big),$$

and we note that the variance of ω just depends on ρ. For ρ close to 1, the variance of ω is approximately $1/2G$, which is very small. Thus, ω will be concentrated close to its mean $\omega^*:=-\rho\tau/(1+\rho)$. Define $p^*(\omega|\tau):=1-k(z'|\rho,\sigma)=\Phi(\omega+\tau)+\exp(-\omega\tau-\tau^2/2)\Phi(\omega)$. Then,

$$\mathrm{IF}(\sigma,\rho)=\int\frac{1+p^*(\omega|\tau)}{1-p^*(\omega|\tau)}\mathfrak{n}(\omega;-\frac{\rho\tau}{1+\rho},\frac{1-\rho}{1+\rho}))d\omega;$$

it is convenient to write as $\mathrm{IF}(\sigma,\rho)$ as $\mathrm{IF}(\tau|\rho)$, which as we will optimize the computing time over τ keeping ρ fixed. Let,

$$f(\omega;\tau) := \frac{1+p^*(\omega|\tau)}{1-p^*(\omega|\tau)}$$

Using the 4th order Taylor series expansion of $f(w;\tau)$ at $\omega=\omega^*$, the inefficiency factor can be approximated by

$$\mathrm{IF}_{\mathrm{approx}}(\tau|\rho) = f(\omega^*|\tau) + \frac{1}{2}\frac{1-\rho}{1+\rho} f^{(2)}(\omega^*|\tau) + \frac{1}{8}\left(\frac{1-\rho}{1+\rho}\right)^2 f^{(4)}(\omega^*|\tau),$$

which is considered as a function of τ with ρ fixed. This approximation will be very good because, as noted, the variance of ω is very small for G large. So the computing time $\mathrm{CT}(\sigma,\rho) = \mathrm{IF}(\sigma,\rho)/\sigma^2$, is approximated by

$$\mathrm{CT}_{\mathrm{approx}}(\tau|\rho) = 1-\rho^2)\frac{\mathrm{IF}_{\mathrm{approx}}(\tau|\rho)}{\tau^2} \propto \frac{\mathrm{IF}_{\mathrm{approx}}(\tau|\rho)}{\tau^2}$$

Minimizing this term over τ^2, for ρ close to 1, we find that $\mathrm{CT}(\tau)$ is minimized at $\tau \approx 2.16$. So the optimal $\sigma_{\mathrm{opt}} \approx 2.16/\sqrt{1-\rho^2}$.

Then the unconditional acceptance rate (10) is

$$\begin{aligned} P(\mathrm{accept}|\rho,\sigma_{\mathrm{opt}}) &= 2\left(1-\Phi\Big(\frac{\sigma_{\mathrm{opt}}\sqrt{1-\rho}}{\sqrt{2}}\Big)\right) \\ &= 2\left(1-\Phi\Big(\frac{\sigma_{\mathrm{opt}}\sqrt{1-\rho^2}}{\sqrt{2(1+\rho)}}\Big)\right) \\ &\approx 2\left(1-\Phi\Big(\frac{2.16}{2}\Big)\right) \approx 0.28. \end{aligned}$$

□

Appendix B Derivation of the expression (13) for Computing Time

The average computing time required to give the same accuracy in terms of variance as M iterates $\theta_1,\ldots,\theta_M$ from an iid sampler from $\pi(\theta)$ is

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{G} N_k(\theta_i)\mathrm{IF}(\sigma,\rho) = \frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{G}\frac{G\gamma_{(k)}^2(\theta_i)}{\sigma^2}\mathrm{IF}(\sigma,\rho) \to \left(G\sum_{k=1}^{G}\overline{\gamma}_{(k)}^2\right)\frac{\mathrm{IF}(\sigma,\rho)}{\sigma^2}$$

as $M\to\infty$, where $\overline{\gamma}_{(k)}^2 = \mathbb{E}_{\theta\sim\pi}(\gamma_j(\theta))$. Hence the computing time is proportional to $\mathrm{CT} = \frac{\mathrm{IF}(\sigma,\rho)}{\sigma^2}$. □