The University of Sydney Business School
The University of Sydney

BUSINESS ANALYTICS WORKING PAPER SERIES

Speeding up MCMC by Efficient Data Subsampling

Matias Quiroz, Mattias Villani, Robert Kohn and Minh-Ngoc Tran

Abstract

We propose Subsampling MCMC, a Markov Chain Monte Carlo (MCMC) framework where the likelihood function for *n* observations is estimated from a random subset of *m* observations. We introduce a general and highly efficient unbiased estimator of the log-likelihood based on control variates obtained from clustering the data. The cost of computing the log-likelihood estimator is much smaller than that of the full log-likelihood used by standard MCMC. The likelihood estimate is bias-corrected and used in two correlated pseudo-marginal algorithms to sample from a perturbed posterior, for which we derive the asymptotic error with respect to n and m, respectively. A practical estimator of the error is proposed and we show that the error is negligible even for a very small *m* in our applications. We demonstrate that Subsampling MCMC is substantially more efficient than standard MCMC in terms of sampling efficiency for a given computational budget, and that it outperforms other subsampling methods for MCMC proposed in the literature.

Keywords: Bayesian inference, Estimated likelihood, Correlated pseudo-marginal, Block pseudo-marginal, Big Data, Survey sampling.

January 2017

BA Working Paper No: BAWP-2017-01
http://sydney.edu.au/business/business_analytics/research/working_papers

SPEEDING UP MCMC BY EFFICIENT DATA SUBSAMPLING

MATIAS QUIROZ, MATTIAS VILLANI, ROBERT KOHN AND MINH-NGOC TRAN

ABSTRACT. We propose Subsampling MCMC, a Markov Chain Monte Carlo (MCMC) framework where the likelihood function for n observations is estimated from a random subset of m observations. We introduce a general and highly efficient unbiased estimator of the log-likelihood based on control variates obtained from clustering the data. The cost of computing the log-likelihood estimator is much smaller than that of the full log-likelihood used by standard MCMC. The likelihood estimate is bias-corrected and used in two correlated pseudo-marginal algorithms to sample from a perturbed posterior, for which we derive the asymptotic error with respect to n and m, respectively. A practical estimator of the error is proposed and we show that the error is negligible even for a very small m in our applications. We demonstrate that Subsampling MCMC is substantially more efficient than standard MCMC in terms of sampling efficiency for a given computational budget, and that it outperforms other subsampling methods for MCMC proposed in the literature.

KEYWORDS: Bayesian inference, Estimated likelihood, Correlated pseudo-marginal, Block pseudo-marginal, Big Data, Survey sampling.

1. INTRODUCTION

The popularity of Bayesian methods increased significantly in the early 90's due to advances in computer technology and the introduction of powerful simulation algorithms such as Markov Chain Monte Carlo (MCMC) (Gelfand and Smith, 1990). However, posterior sampling with MCMC is time-consuming and there is an increasing awareness that new scalable algorithms are necessary for MCMC to remain an attractive choice for inference in data sets with a large number of observations.

Quiroz: *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University and Research Division, Sveriges Riksbank. E-mail: quiroz.matias@gmail.com.* Villani: *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University.* Kohn: *Australian School of Business, University of New South Wales.* Tran: *Discipline of Business Analytics, University of Sydney.* The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

Current research on scalable MCMC algorithms belongs to two major groups. The first group employs parallelism through the typical MapReduce scheme (Dean and Ghemawat, 2008) by partitioning the data and computing posteriors in a parallel and distributed manner. The resulting draws are subsequently combined into a single posterior distribution. The main difference within this group is how weighting is performed and whether the partitions communicate at runtime, see for example Scott et al., 2013; Neiswanger et al., 2013; Wang and Dunson, 2013; Minsker et al., 2014; Nemeth and Sherlock, 2016. Our approach belongs to the second group of methods that use a subsample of the data in each MCMC iteration to speed up the algorithm, which we refer to as subsampling MCMC, see Korattikara et al. (2014); Bardenet et al. (2014); Maclaurin and Adams (2014); Bardenet et al. (2015); Liu et al. (2015). Section 4.3 briefly outlines these approaches and Section 4.4 compares them against our methods. For a more extensive introduction to these methods and a broad overview of the problem in general, see the excellent review in Bardenet et al. (2015).

Our article presents a Metropolis-Hastings (MH) framework where the likelihood is estimated from a random subset of the data using highly efficient control variates for variance reduction. For models with an intractable likelihood function, Beaumont (2003) proposes to estimate the likelihood unbiasedly and run the MH algorithm on an extended space, which also includes the random variates underlying the likelihood estimate. Andrieu and Roberts (2009) develop theory for such Pseudo-Marginal MH (PM) algorithms, and prove that PM algorithms target the true posterior if the likelihood estimator is unbiased and almost surely positive. Obtaining unbiased likelihood estimators with low variability from subsampling is a major challenge, and previous attempts have failed to produce an MCMC sampler that does not get stuck (Korattikara et al., 2014; Bardenet et al., 2015). Moreover, ensuring that the unbiased likelihood estimator is also positive was shown by Jacob and Thiery (2015) to only be possible under assumptions that can only be satisfied by sampling the full data set (Bardenet et al., 2015). Quiroz et al. (2016) use the insights and techniques proposed here (control variates and correlated PM for subsampling) to produce an estimator with low variability and, in addition, target the absolute value of the estimate following Lyne et al.

(2015) that allow the estimator to occasionally be negative. Draws from the algorithm are corrected with an importance sampling step to obtain unbiased estimates of expectations of posterior functions. Let m and n be the subsample and population sizes. Our article provides an alternative approach that instead simulates by PM, but from a slightly perturbed target (because the likelihood estimator is slightly biased), where (i) the error can be estimated and made arbitrarily small at the rate $O(m^{-2})$ with n fixed and (ii) a smaller variance of the logarithm of the estimator of the likelihood can be achieved thus requiring a smaller m. We also study the error with respect to the number of observations n, when $m = m(n)$ and under certain assumptions of the control variates.

The variance of the estimator of the log-likelihood is crucial for the performance of PM algorithms: a large variance can easily produce extreme over-estimates of the likelihood and cause the Markov chain to get stuck for long periods. On the other hand, a too precise likelihood estimator might be unnecessarily costly. Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) analyze the variance of the log-likelihood estimator $\sigma^2_{LL,m,n}$ that maximizes the number of effective draws per unit of computing time. They conclude that the optimal number of particles m should be such that $\sigma^2_{LL,m,n}$ is around 1. Moreover, $m = O(n)$ is required to obtain the optimal value of the variance. Recent advances in PM algorithms correlate the particles at the current and proposed parameter value in the MH ratio (Deligiannidis et al., 2016; Dahlin et al., 2015) or use blocking (Tran et al., 2016a). This makes it possible to target $\sigma^2_{LL,m,n} \gg 1$ and the optimal variance can be obtained with $m = O(n^{1/2})$ (Deligiannidis et al., 2016; Tran et al., 2016a). Our article proposes a correlated pseudo-marginal approach for data subsampling that uses a copula transformation of the random variates in Deligiannidis et al. (2016). Tran et al. (2016a) derive an explicit expression of the correlation of the log-likelihood estimator at the current and proposed draw, which we use to compute the optimal subsample size $m = m(n)$ for our algorithm.

The paper is organized as follows. Section 2 introduces the general estimator and derives some important properties. Section 3 outlines the subsampling MCMC algorithm and its theoretical framework, including results on the accuracy of the perturbed posterior, and how

to select the rate of m to achieve the optimal $\sigma^2_{LL,m,n}$. Section 4 reviews other subsampling approaches and evaluates the performance of the proposed methodology on two examples. The second of these examples benchmarks our methods against the other approaches. Implementation details and proofs are placed in the online Appendices A and B.

2. Sampling-based Log-likelihood Estimators

2.1. **A log-likelihood estimator based on simple random sampling with efficient control variates.** Let $\{y_i, x_i\}_{i=1}^n$ denote the data, where y is a response vector and x is a vector of covariates. Let $\theta \in \Theta$ be the vector of parameters. Given conditionally independent observations we have the usual decomposition of the log-likelihood

$$\ell_{(n)}(\theta) \coloneqq \sum_{i=1}^{n} \ell_i(\theta), \quad \text{where } \ell_i(\theta) \coloneqq \log p(y_i|\theta, x_i) \tag{2.1}$$

is the *log-likelihood contribution* of the ith observation. For any given θ, (2.1) is a sum of a finite number of elements and estimating it is equivalent to the classical survey sampling problem of estimating a population total. See Särndal et al. (2003) for an introduction. We assume in (2.1) that the log-likelihood decomposes as a sum of terms where each term depends on a unique piece of data information. This applies to longitudinal problems where $\ell_i(\theta)$ is the log joint density of all measurements on the ith subject, and we sample subjects rather than individual observations. It also applies to certain time-series problems such as AR(p) processes, where the sample elements become $(y_t, \ldots, y_{t-p})$, for $t = p+1, \ldots, n$. Our examples in Section 4 use independent identically distributed (iid) and time series data.

Estimating (2.1) based on Simple Random Sampling (SRS), where any $\ell_i(\theta)$ is included with the same probability generally results in a dramatically large variance. Intuitively, since some $\ell_i(\theta)$ contribute significantly more to the sum in (2.1) they should be included in the sample with a larger probability, using so called Probability Proportional-to-Size (PPS) sampling. However, this requires each of the n sampling probabilities to be proportional to a measure of their size. Evaluating n size measures is likely to defeat the purpose of subsampling, except in cases when there is a computationally cheaper proxy than $\ell_i(\theta)$ that

can be utilized instead. Alternatively, one can make the $\{\ell_i(\theta)\}_{i=1}^n$ more homogeneous by using control variates so that the population elements are roughly of the same size and SRS is then expected to be efficient. Our article focuses on this case and proposes efficient control variates $q_{i,n}(\theta)$ such that the computational cost of the estimator is substantially less than $O(n)$. The dependence on n is due to $q_{i,n}(\theta)$ being an approximation of $l_i(\theta)$, which typically improves as more data is available.

Define the differences $d_{i,n}(\theta) \coloneqq \ell_i(\theta) - q_{i,n}(\theta)$ and let

$$\mu_{d,n}(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} d_{i,n}(\theta) \quad \text{and } \sigma^2_{d,n}(\theta) \coloneqq \frac{\sum_{i=1}^n \left(d_{i,n}(\theta) - \mu_{d,n}(\theta)\right)^2}{n}$$

be the mean and variance of the finite population $\{d_{i,n}(\theta)\}_{i=1}^n$. Let $u_1, \ldots, u_m$ be iid random variables such that $\Pr(u = k) = 1/n$ for $k = 1, \ldots, n$. The Difference Estimator (DE, Särndal et al., 2003) of $\ell_{(n)}(\theta)$ in (2.1) is

$$\widehat{\ell}_{(m,n)}(\theta) \coloneqq q_{(n)}(\theta) + n\widehat{\mu}_{d,n}(\theta), \quad \widehat{\mu}_{d,n}(\theta) \coloneqq \frac{1}{m}\sum_{i=1}^{m} d_{u_i,n}(\theta) \tag{2.2}$$

with $q_{(n)}(\theta) \coloneqq \sum_{i=1}^n q_{i,n}(\theta)$. It is straightforward to use unequal sampling probabilities with the DE, but the sampling probabilities need to be evaluated for every observation, which can be costly. The following Lemma gives some basic properties of the DE estimator.

Lemma 1. *Suppose that $\widehat{\ell}_{(m,n)}$ is the estimator of $\ell_{(n)}(\theta) = \ell(\theta)$ given by (2.2). Then*

i. $\mathrm{E}[\widehat{\mu}_{d,n}(\theta)] = \mu_{d,n}(\theta)$.

ii.
$$\mathrm{E}\left[\widehat{\ell}_{(m,n)}(\theta)\right] = l_{(n)}(\theta) \quad \textit{and} \quad \sigma^2_{LL,m,n} = \mathrm{V}\left[\widehat{\ell}_{(m,n)}(\theta)\right] = \frac{n^2\sigma^2_{d,n}}{m}.$$

iii. *$\widehat{\ell}_{(m,n)}(\theta)$ is asymptotically normal when $m \to \infty$ for fixed n and $\sigma^2_{d,n} < \infty$, or when both $m, n \to \infty$ with $m = Bn^\gamma$ for constants $B > 0$ and $\gamma > 0$ and $\sigma^3_{d,n} < \infty$.*

Proof. The proofs of parts i) and ii) are straightforward and omitted. The proof of iii) is in Appendix B. □

The assumptions of finite $\sigma^2_{d,n}$ and $\sigma^3_{d,n}$ in Lemma 1 part (iii) are highly non-restrictive because the random variables are discrete with a finite sample space: they are satisfied for any control variates that are finite. We use the following estimate of $\sigma^2_{d,n}$

$$\widehat{\sigma}^2_{d,n}(\theta) \coloneqq \frac{\sum_{i=1}^m \left(d_{u_i,n}(\theta) - \widehat{\mu}_{d,n}(\theta)\right)^2}{m}.$$

2.2. **Control variates for variance reduction.** To see the crucial role of variance reduction using control variates, we first note that the variance of the log-likelihood estimator should be around one for the standard PM (see e.g. Pitt et al., 2012; Doucet et al., 2015 and Section 1). Now, define

$$a_n \coloneqq \sup_{\theta\in\Theta} \sup_{i\in\{1,\ldots,n\}} \left|d_{i,n}(\theta) - \mu_{d,n}(\theta)\right|. \tag{2.3}$$

Throughout our article we assume that $a_n < \infty$ for a given n and also that $\limsup a_n < \infty$. This assumption is easily enforced if Θ is a compact space as long as the control variates are finite. The following lemma is straightforward to prove.

Lemma 2. *Suppose that* $\limsup a_n < \infty$. *Then* $\sigma^2_{d,n} = O(a_n^2)$ *and* $\sigma^2_{LL,m,n} = \frac{n^2 a_n^2 O(1)}{m}$.

According to Lemma 2, keeping the variance bounded as a function of n requires that $\frac{n^2 O(a_n^2)}{m} = O(1)$. This highlights the importance of the variance reduction: SRS without control variates scales poorly because $O(a_n^2) = O(1)$ and $m = O(n^2)$. On the other hand, with control variates that improve as, say $d_{i,n} = O(n^{-\alpha})$ with $\alpha > 0$, we have $O(a_n^2) = O(n^{-2\alpha})$ and $m = O(n^{2(1-\alpha)})$.

2.3. **Computational complexity.** The difference estimator in (2.2) needs to compute $q_{(n)}(\theta) = \sum_{i=1}^n q_{i,n}(\theta)$ in every MCMC iteration: this requires evaluating the control variates $q_{i,n}(\theta)$ for all data points. We now explore specific choices of $q_{i,n}$ that allow us to compute $\sum_{i=1}^n q_{i,n}(\theta)$ using substantially less evaluations than n. Denote the Computational Cost (CC) for the standard MH without subsampling which evaluates $\ell_{(n)} \coloneqq \sum_{i=1}^n \ell_i$ by

$$\mathrm{CC}[\ell_{(n)}(\theta)] \coloneqq n \cdot c_\ell,$$

where c_ℓ is the cost of evaluating a single log-likelihood contribution (assuming the cost is the same for all i). For the difference estimator in (2.2), we have

$$\mathrm{CC}\left[\widehat{\ell}_{(m,n)}(\theta)\right] \coloneqq n \cdot c_q + m \cdot c_\ell,$$

where c_q is the cost of computing a control variate. We now briefly describe two particular control variates that reduce the first term $n \cdot c_q$. See Appendix A for details regarding their implementation.

First, consider the control variates in Bardenet et al. (2015) who propose to use a second order Taylor expansion of each $\ell_i(\theta)$ around some reference value $\theta^\star$, e.g. the maximum likelihood estimate. This reduces the complexity from n evaluations to a single one (similar to sufficient statistics for a normal model because $q_{i,n}(\theta)$ is quadratic in θ). As noted by Bardenet et al. (2015), this control variate is a poor approximation of $\ell_i(\theta)$ whenever the algorithm proposes a θ that is not near to $\theta^\star$, and will therefore work well only when the posterior is tightly concentrated around $\theta^\star$. As a remedy in the case of a less concentrated posterior, they suggest to occasionally recompute the control variates, expanding around the current θ in the MCMC (using all n observations).

We now propose a control variate that works well regardless of the posterior concentration. This control variate is based on clustering the data $\{z_i = (y_i, x_i)\}_{i=1}^n$ into K clusters that are kept fixed throughout the MCMC. At every MCMC iteration, we compute the exact log-likelihood contributions at all K centroids and use a second order Taylor expansion with respect to z_i at the centroid as a local approximation of ℓ_i around each centroid. This allows us to compute $\sum_{i=1}^n q_{i,n}(\theta)$ by simply scaling up quantities computed at the K centroids. The resulting estimator therefore has cost

$$\mathrm{CC}\left[\widehat{\ell}_{(m,n)}(\theta)\right] = K \cdot c_q + m \cdot c_\ell, \tag{2.4}$$

where typically $K \ll n$.

3. Subsampling MCMC Methodology

3.1. **MCMC with likelihood estimators from data subsampling.** We propose an efficient unbiased estimator $\widehat{\ell}_{(m,n)}(\theta)$ of the log-likelihood and then approximately bias-correct following Ceperley and Dewing (1999) (see also Nicholls et al., 2012) to obtain the 'bias-corrected' likelihood estimator

$$\widehat{L}_{(m,n)}(\theta, u) \coloneqq \exp\left(\widehat{\ell}_{(m,n)}(\theta) - \frac{n^2}{2m}\widehat{\sigma}^2_{d,n}(\theta)\right), \tag{3.1}$$

where $\widehat{\ell}_{(m,n)}(\theta)$ and $\widehat{\sigma}^2_{d,n}(\theta)$ are the estimators presented in Section 2.1. The form of (3.1) is motivated by the case when $\widehat{\ell}_{(m,n)} \sim \mathcal{N}(\ell_{(n)}(\theta), \sigma^2_{LL,m,n}(\theta))$ and $\sigma^2_{LL,m,n}$ is known, in which case all bias is removed. Normality holds asymptotically in both m and n by part (iii) of Lemma 1. However, the assumption of known variance is unrealistic because the computation requires the entire data set. The estimator in (3.1) is therefore expected to only be nearly unbiased. A main difference of our use of this estimator compared to Ceperley and Dewing (1999) and Nicholls et al. (2012) is that our approach is a pseudo-marginal, where the space explored by the Markov chain also includes the random variates used for estimating the likelihood. Other differences are that we use control variates, that we analyze the dependence on n in the analysis of the error, and that our convergence rate of the error (Theorem 1 below) is $O(m^{-2})$ as opposed to $O(m^{-1})$ in Nicholls et al. (2012).

We now outline how to carry out a pseudo-marginal MH with the approximately unbiased estimator in (3.1) and derive the asymptotic error in the stationary distribution. Denote the likelihood by $L_{(n)}(\theta) \coloneqq p(y|\theta)$, let $p_\Theta(\theta)$ be the prior and define the marginal likelihood $\overline{L}_{(n)} \coloneqq \int L_{(n)}(\theta)p_\Theta(\theta)d\theta$. Then the posterior is $\pi_{(n)}(\theta) = L_{(n)}(\theta)p_\Theta(\theta)/\overline{L}_{(n)}$. Let $p_U(u)$ be the distribution of the vector u of auxiliary variables corresponding to the subset of observations to include when estimating $L_{(n)}(\theta)$. Let $\widehat{L}_{(m,n)}(\theta, u)$, for fixed m and n, be a possibly biased estimator of $L_{(n)}(\theta)$ with expectation

$$L_{(m,n)}(\theta) = \int \widehat{L}_{(m,n)}(\theta, u)p_U(u)du.$$

Define

$$\overline{\pi}_{(m,n)}(\theta,u) \coloneqq \widehat{L}_{(m,n)}(\theta,u)p_U(u)p_\Theta(\theta)/\overline{L}_{(m,n)}, \text{ with } \overline{L}_{(m,n)} \coloneqq \int L_{(m,n)}(\theta)p_\Theta(\theta)d\theta, \tag{3.2}$$

on the augmented space (θ, u). It is straightforward to show that $\overline{\pi}_{(m,n)}(\theta,u)$ is a proper density with marginal

$$\overline{\pi}_{(m,n)}(\theta) \;=\; \int \overline{\pi}_{(m,n)}(\theta,u)du = L_{(m,n)}(\theta)p_\Theta(\theta)/\overline{L}_{(m,n)}.$$

The MCMC that targets (3.2) uses a joint proposal for θ and u given by

$$q_{\Theta,U}(\theta,u|\theta_c,u_c) = p_U(u)q_\Theta(\theta|\theta_c)$$

where c denotes the current state of the Markov chain. The PM acceptance probability becomes

$$\alpha = \min\left(1, \frac{\widehat{L}_{(m,n)}(\theta_p,u_p)p_\Theta(\theta_p)/q_\Theta(\theta_p|\theta_c)}{\widehat{L}_{(m,n)}(\theta_c,u_c)p_\Theta(\theta_c)/q_\Theta(\theta_c|\theta_p)}\right). \tag{3.3}$$

This expression is similar to the MH acceptance probability, but with the true likelihood replaced by its estimate. By Andrieu and Roberts (2009), the draws of θ obtained by this MH algorithm have $\overline{\pi}_{(m,n)}(\theta)$ as invariant distribution. If $\widehat{L}_{(m,n)}(\theta,u)$ is an unbiased estimator of $L_{(n)}(\theta)$, then the marginal of the augmented MCMC scheme above has $\overline{\pi}_{(m,n)}(\theta) = \pi_{(n)}(\theta)$ (the true posterior) as invariant distribution. However, if $\widehat{L}_{(m,n)}(\theta,u)$ is biased, the sampler is still valid but has a perturbed marginal $\overline{\pi}_{(m,n)}(\theta)$.

3.2. **Perturbation analysis - asymptotics.** Our next result gives the rate at which the perturbed target $\pi_{(m,n)}(\theta)$ approaches the true target posterior $\pi_{(n)}(\theta)$.

Theorem 1. *Suppose that a PM algorithm is implemented with the estimator in* $\widehat{L}_{(m,n)}(\theta,u)$ *in* (3.1) *and assume that* $n^3a_n^3/m^2 = o(1)$. *The following results hold for any* $\theta \in \Theta$, *where* Θ *is a compact space,*

i.
$$\frac{\left|L_{(m,n)}(\theta) - L_{(n)}(\theta)\right|}{L_{(n)}(\theta)} \leq O\left(\frac{n^2 a_n^2}{m^2}\right).$$

ii.
$$\frac{\left|\overline{\pi}_{(m,n)}(\theta) - \pi_{(n)}(\theta)\right|}{\pi_{(n)}(\theta)} \leq O\left(\frac{n^2 a_n^2}{m^2}\right).$$

iii. *Suppose that* $h(\theta)$ *is a function such that* $\mathrm{E}_{\pi_{(n)}}[|h(\theta)|] < \infty$. *Then*

$$\left|\frac{\mathrm{E}_{\overline{\pi}_{(m,n)}}[h(\theta)] - \mathrm{E}_{\pi_{(n)}}[h(\theta)]}{\mathrm{E}_{\pi_{(n)}}[h(\theta)]}\right| \leq O\left(\frac{n^2 a_n^2}{m^2}\right).$$

Note that we need to have $m = n^2 O(a_n^2)$ to target $\sigma^2_{LL,m,n} = \frac{n^2 O(a_n^2)}{m}$ at the optimal value around 1. By Theorem 1, this m gives a fractional error of the posterior which is within $O(n^{-2}a_n^{-2})$ of the true posterior. Note also that the perturbation error is $O\left(m^{-2}\right)$ for a fixed n.

3.3. **Approximating the perturbation error.** Theorem 1 is derived under essentially no assumptions on the estimator, and clearly displays the fast convergence of our perturbed posterior, but it does not provide a practically useful way to quantify the discrepancy between $\overline{\pi}_{(m,n)}(\theta)$ and $\pi_{(n)}(\theta)$. We now propose a way to estimate the point-wise fractional error in the perturbed posterior distribution

$$\text{error}(\theta) = \frac{\pi_{(m,n)}(\theta) - \pi_{(n)}(\theta)}{\pi_{(n)}(\theta)} = \left(\frac{L_{(m,n)}(\theta)}{L_{(n)}(\theta)}\right) \Big/ \left(\frac{\overline{L}_{(m,n)}}{\overline{L}_{(n)}}\right) - 1. \tag{3.4}$$

The following lemma is an application of the bivariate Central Limit Theorem (CLT).

Lemma 3.

$$\sqrt{m}\left(\begin{bmatrix}\widehat{\mu}_{d,n}\\ \widehat{\sigma}^2_{d,n}\end{bmatrix} - \begin{bmatrix}\mu_{d,n}\\ \sigma^2_{d,n}\end{bmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}\sigma^2_{d,n} & \varphi^{(3)}_{d,n}\\ \varphi^{(3)}_{d,n} & \sigma^4_{d,n} - \varphi^{(4)}_{d,n}\end{bmatrix}\right) \quad \text{as } m \to \infty,$$

with $\varphi^{(b)}_{d,n} = E[(d_{u_i,n} - \mu_{d,n})^b] = \sum_{i=1}^n (d_{i,n} - \mu_{d,n})^b / n$ *for* $b \geq 1$.

Lemma 4. *Suppose that*

$$\begin{bmatrix}\widehat{\mu}_{d,n}\\ \widehat{\sigma}^2_{d,n}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{d,n}\\ \sigma^2_{d,n}\end{bmatrix}, \bar{\Sigma} = \frac{1}{m}\begin{bmatrix}\sigma^2_{d,n} & \varphi^{(3)}_{d,n}\\ \varphi^{(3)}_{d,n} & \sigma^4_{d,n}-\varphi^{(4)}_{d,n}\end{bmatrix}\right). \tag{3.5}$$

Then, with $\Psi^{(b)}_{d,n} \coloneqq \varphi^{(b)}_{d,n}/\sigma^b_{d,n}$ *for* $b \geq 1$,

$$L_{(m,n)}(\theta) = \exp\left(\ell_{(n)} + \frac{\sigma^4_{LL,m,n}(\theta)}{8m}\left(1-\Psi^{(4)}_{d,n}(\theta)\right) - \frac{\sigma^3_{LL,m,n}(\theta)}{2\sqrt{m}}\Psi^{(3)}_{d,n}(\theta)\right). \tag{3.6}$$

Proof. Since $L_{(m,n)}(\theta) = \exp\left(q_{(n)}(\theta)\right)\mathrm{E}\left[\exp\left(n\widehat{\mu}_{d,n} - \frac{n^2}{2m}\widehat{\sigma}^2_{d,n}\right)\right]$ the result follows from the moment generating function (mgf) of the bivariate normal distribution in (3.5). □

It is easy to show that $\gamma^{(b)}_{d,n} = O(1)$ for any $b \geq 1$. From Lemma 4 it follows that the perturbation error in the likelihood $L_{(m,n)}(\theta)/L_{(n)}(\theta)$ depends on $\sigma_{LL,m,n}$, and will increase with it for sufficiently large $\sigma_{LL,m,n}$. It is important to note, however, that any constant factor c in $L_{(m,n)}(\theta)/L_{(n)}(\theta)$ that does not depend on θ will cancel out in the fractional posterior error in (3.4) since the same factor c will also appear in $\overline{L}_{(m,n)}/\overline{L}_{(n)}$. This observation leads to the following theoretically interesting remark.

Remark 1. *Suppose we run a PM algorithm and that we can, for any proposed* $\theta \in \Theta$*, where* Θ *is a compact space, choose* $m(\theta)$ *such that*

$$\frac{\sigma^4_{LL,m(\theta),n}(\theta)}{8m(\theta)}\left(1-\Psi^{(4)}_{d,n}(\theta)\right) - \frac{\sigma^3_{LL,m(\theta),n}(\theta)}{2\sqrt{m(\theta)}}\Psi^{(3)}_{d,n}(\theta) = c. \tag{3.7}$$

Then $\mathrm{error}(\theta) = 0$.

The constant c is an arbitrary choice: any c generates a specific $m_c(\theta)$ which ensures unbiasedness for any $\theta \in \Theta$. A natural choice of c would be to solve (3.7) based on the $m(\theta^\star)$ that targets the optimal $\sigma^2_{LL,m(\theta^\star),n}$, where $\theta^\star$ is the mode. Of course, this strategy uses all data and is thus not applicable, but illustrates an important property of our method.

In practice we can instead use the result in Lemma 4 to check $\mathrm{error}(\theta)$ in any given application as follows. The quantities $\sigma_{LL,m,n}(\theta)$ and $\gamma^{(b)}_{d,n}$ can be easily evaluated for any θ at the cost of evaluating $\ell_i(\theta)$ for all $i = 1, ..., n$, or estimated from a subsample. It

is also necessary to evaluate $\overline{L}_{(n)}$ and $\overline{L}_{(m,n)}$, which can be done with the usual Laplace approximation. Approximating $\overline{L}_{(n)}$ requires the Hessian of $\log L_{(n)}(\theta)$ evaluated at the mode, which can be obtained analytically from tedious differentiation or numerically by finite differences. A similar procedure applies for $\overline{L}_{(m,n)}$, but with $\log L_{(m,n)}$ from (3.6) where clearly the Hessian becomes analytically intractable, but finite differences are straightforward.

3.4. **Correlated proposals of u for subsampling.** Deligiannidis et al. (2016) and Dahlin et al. (2015) both propose a general method that correlates the current and proposed values of u_i. The advantage of this correlation is that it makes the variance of the difference in the logarithms of the estimated likelihoods appearing in (3.3) much smaller than that of each of the terms themselves. This leads in our context to requiring much smaller values of m, or equivalently, that we can target much higher values of $\sigma^2_{LL,m,n}$ than unity, provided we also check that error(θ) remains at an acceptable level.

For a correlated PM approach to subsampling, we let u be a vector of length n with binary elements u_i that determine if observation i is included ($u_i = 1$) when estimating the log-likelihood. Note that this is different from the above, where u contained the observation indices and was of length m. Moreover, here the sample size is random and we let $m^\star$ be the expected sample size. The sampling probabilities become $\Pr(u_i = 1) = m^\star/n$ for $i = 1, \ldots, n$. We use the auxiliary variable (particle) v in Deligiannidis et al. (2016) to induce dependence at the current u_i^c and proposed u_i^p sampling indicator through a Gaussian copula as we now explain. The correlated pseudo-marginal method uses a Gaussian auto-regressive kernel $\mathcal{K}(v_c, v_p)$ with a transition defined by $v_p = \phi v_c + \sqrt{1-\phi^2}\varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$. We also have $v_c \sim p(v) = \mathcal{N}(v|0,1)$ and $\mathcal{K}(v_c, v_p)$ is reversible with respect to $p(v)$. We sample the u_i's by first generating v_c and v_p and set $u_i^c = \mathcal{I}\left[\Phi(v_c) \leq \frac{m^\star}{n}\right]$ and $u_i^p = \mathcal{I}\left[\Phi(v_p) \leq \frac{m^\star}{n}\right]$, where Φ denotes the standard normal cdf. An equivalent approach is to generate u_i^p from a Markov chain with marginal $p(u_i^c = 1) = m^\star/n$, with transition probabilities $\Pr(u_i^p = 1|u_i^c = 1) = \kappa$ and $\Pr(u_i^p = 0|u_i^c = 0) = 1 - (1-\kappa)\frac{m^\star/n}{1-m^\star/n}$. The persistence parameter κ in the Markov chain is related to the AR persistence ϕ by the relation $\kappa = \frac{n}{m^\star}\Phi_2\left(\Phi^{-1}(m^\star/n), \Phi^{-1}(m^\star/n)|\phi\right)$, where $\Phi_2(\cdot,\cdot|\phi)$ is the cdf of bivariate standard normal variables with correlation ϕ.

As noted above, in contrast to Section 2.1, u is a binary vector. We can instead use the Horvitz-Thompson (Horvitz and Thompson, 1952) which (under SRS) is

$$\widehat{d}_{(m^\star,n)} = \sum_{i=1}^{n} \frac{d_{i,n}}{m^\star/n} u_i,$$

and is unbiased for $d_{(n)}$. Note that we can write

$$\widehat{d}_{(m^\star,n)} = \frac{1}{m^\star} \sum_{i=1}^{n} n d_{i,n} u_i, \quad \text{with } \sigma^2_{LL,m^\star,n} = \frac{\sigma^2_{\xi,m^\star,n}}{m^\star}, \quad \text{where } \sigma^2_{\xi,m^\star,n} = n\left(1 - \frac{m^\star}{n}\right) \sum_{i=1}^{n} d^2_{i,n}$$

can be unbiasedly estimated by

$$\widehat{\sigma}^2_{\xi,m^\star,n} = n^2 \left(1 - \frac{m^\star}{n}\right) \frac{1}{m^\star} \sum_{i=1}^{n} d^2_{i,n} u_i.$$

3.5. **Block proposals of u for subsampling.** Tran et al. (2016a) propose the block PM algorithm and show that it is a natural way to correlate the estimation errors in panel data and also in subsampling problems such as ours. The method divides the vector of observation indices $u = (u_1, \ldots, u_m)$ into G blocks and then update one block at a time jointly with θ. By setting a large G, a high correlation ρ between the estimates at the proposed and current parameter values is induced, reducing the variability of the ratio of estimates. More precisely, they show that under certain assumptions $\rho_G = 1 - 1/G$.

3.6. **Optimal variance of the estimator.** Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) obtain the value of $\sigma^2_{LL,m,n}$ that optimizes the trade off between MCMC sampling efficiency and computational cost in standard PM. The consensus is that $\sigma^2_{LL,m,n}$ should be between $[1, 3.283]$ where, in general, the less efficient the proposal in the exact likelihood setting, the higher the optimal value of $\sigma^2_{LL,m,n}$. The optimal value is derived assuming that the cost of computing one MCMC sample is inversely proportional to $\sigma^2_{LL,m,n}$, so that the so called Computational Time (CT) to produce one equivalent to an iid draw is

$$\text{CT}(\sigma^2_{LL,m,n}) \;\propto\; \text{IF}(\sigma^2_{LL,m,n}) \times \frac{1}{\sigma^2_{LL,m,n}}, \text{ with } \text{IF}(\sigma^2_{LL,m,n}) = 1 + 2\sum_{l=1}^{\infty} \rho_l, \tag{3.8}$$

where IF is the Inefficiency Factor and ρ_l is the l-lag auto-correlation of the chain. In our approach we have to select both m and K, the number of clusters. The computational cost of a new cluster comes from evaluating ℓ_i at the centroid, but also from evaluating the gradient and Hessian of ℓ_i. An approximate upper bound for the cost of a new cluster is therefore $3c_\ell$, where c_ℓ is the cost of a single ℓ_i-evaluation. In many models one can however re-use computations when computing the gradient and Hessian, so the true cost is probably much closer to $1c_\ell$. Assuming that the cost of a new cluster is ωc_ℓ, for some $\omega > 0$, a reasonable measure of computational time is

$$\mathrm{CT}_{(m,K)}(\sigma^2_{LL,m,n}(K)) = \mathrm{IF}(\sigma^2_{LL,m,n}(K)) \times (\omega K + m). \tag{3.9}$$

This expression is similar to Tran et al. (2016b) who also take into account an overhead cost in their CT. We find m and K by standard numerical optimization using an expression for the IF (e.g. the ones derived in Pitt et al., 2012 for PM and Tran et al., 2016a for block PM). It should be noted that the optimal value $\sigma^2_{LL,m,n} \approx 1$ is obtained if m is much larger than K (and ω is not too large) because then (3.8) and (3.9) are approximately equal (up to a proportionality constant).

Tran et al. (2016a) show that the conditional variance of the log-likelihood estimator (conditional on only updating one block of u, keeping the others fixed) is $\tau^2_{m,n,G} = \sigma^2_{LL,m,n}(1-\rho^2_G)$. Let $G = G(m) = O(m^\beta)$, then it follows that (using Lemma 2 and $\rho_G(m) = 1-1/G(m)$) $\tau^2_{m,n,G} = O(1)$ is achieved if we take $m = O(n^\gamma)$ with

$$\gamma \;=\; \frac{2(1-\alpha)}{1+\beta}, \quad \text{and } \alpha \text{ in } a_n = O(n^{-\alpha}) \quad \text{as in (2.3).}$$

Note that if $\beta = 0$, i.e. G is constant as a function of m so that $\rho \to 0$ as $m \to \infty$, then $\gamma = 2(1-\alpha)$, which corresponds to the uncorrelated algorithm. We emphasize that it is the interaction of the control variates and the correlated mechanism that makes the method scale well. For example, using $G = O(\sqrt{m})$, the optimal m is sublinear in n if $\alpha > 1/4$. However, note that reducing γ lower the rates of the asymptotic errors in Theorem 1. Tran et al. (2016a) also derive the optimal value of $\sigma^2_{LL,m,n}$ to target under the assumption of a CT as

in (3.8), but conditional on a value of ρ. This value of $\sigma^2_{LL,m,n}$ will typically be much larger than 1.

4. Applications

4.1. **Settings for PM algorithms.** Our algorithms use the control variates obtained by data clustering followed by a Taylor series expansion in data space. The tuning parameters m and K in the PM algorithms are determined by optimizing the computational time CT in (3.9)

$$\sigma^2_{LL,m,n}(K) = \frac{n^2\sigma^2_{d,n}(K)}{m} \quad \text{and IF}(\sigma^2_{LL,m,n}(K)),$$

with respect to m and K. We estimate the relation $\sigma^2_{d,n}(K) = C_0K^\nu$ by, for each example, running our cluster algorithm on a grid of K and for each value of the grid we compute $\sigma^2_{d,n}$ at the likelihood mode $\theta^\star$. Given C_0 and ν, it is straightforward to use the expression for the IF in Pitt et al. (2012) (PM) and Tran et al. (2016a) (block PM) to minimize $\text{CT}_{(m,K)}$ in (3.9) and obtain m_{opt} and K_{opt} and the corresponding $\sigma^2_{\text{opt}} = \sigma^2_{LL,m_{\text{opt}},n}(K_{\text{opt}})$. The correlated PM uses $m^\star_{\text{opt}} = m_{\text{opt}}$ and the same value of K_{opt} as block PM. Table 1 shows a summary of the settings for the applications. Finally, we let $G = 100$ ($\rho = 0.99$) and $\phi = 0.9999$ ($\kappa = 0.9863$).

4.2. **Logistic regression.** Our first example uses a logistic regression model for modeling bankruptcy conditional on a set of firm-specific covariates and macroeconomic variables, see Giordani et al. (2014) for details on the covariates. The data set has $n = 4,748,089$ and 8 covariates. The model is

$$p(y_i|x_i,\beta) \;=\; \left(\frac{1}{1+\exp(x_i^T\beta)}\right)^{y_i}\left(\frac{1}{1+\exp(-x_i^T\beta)}\right)^{1-y_i}, \text{ with } \quad p(\beta) = \mathcal{N}(\beta|0,10I).$$

Since the bankruptcy observations ($y_i = 1$) are sparse (41,566 defaults), we only subsample the observations with $y_k = 0$ observations, i.e. the first term in

$$\ell(\theta) \;=\; \sum_{\{i;y_i=1\}}\ell_i(\theta) + \sum_{\{i;y_i=0\}}\ell_i(\theta),$$

TABLE 1. *Experimental setup in the applications.* n is the number of observation. The proposals are the Random Walk Metropolis (RWM) $q(\theta|\theta_c) = \mathcal{N}(\theta|\theta_c, \Sigma_{\theta^\star})$ and the Independent MH (IMH) $q(\theta) = t_{10}(\theta|\theta^\star, \Sigma_{\theta^\star})$, where the location parameter is $\theta^\star$ is the posterior mode and $\Sigma_{\theta^\star}$ is the negative inverse Hessian of the log-posterior evaluated at $\theta^\star$, both obtained from an initial numerical optimization. We denote the optimal sample size and number of clusters by m_{opt} and K_{opt}, and $\sigma^2_{LL,\text{opt}}$ is the corresponding optimal variance of the log-likelihood estimate. We use $N = 50,000$ iterates after discarding 5,000 iterates as burn-in.

Example	n	**Proposal**	$100m_{\text{opt}}/n$	$100K_{\text{opt}}/n$	$\sigma^2_{LL,\text{opt}}$
Logistic	$4{,}7\times10^6$	RWM/IMH			
Uncorr			8.615	4.967	0.27
Block / Corr			1.286	0.485	56.89
AR(1): M_1	10^5	RWM			
Uncorr			1.896	2.464	0.11
Block / Corr			0.757	0.993	12.41
AR(1): M_2	10^5	RWM			
Uncorr			4.561	8.192	0.11
Block / Corr			2.151	3.176	12.40

is always evaluated (and included in the CC). Figure 1 shows the sampling efficiency of the PM algorithms relative to that of the MH algorithm as measured by the Relative Computational Time (RCT) defined, for any sampler $\mathcal{A}$, as $\text{CT}_{\text{MH}}/\text{CT}_{\mathcal{A}}$. The figure also shows the Relative IF (RIF) , which is defined as $\text{IF}_{\mathcal{A}}/\text{IF}_{\text{MH}}$, where each IF is estimated using the coda package in `R` (Plummer et al., 2006). The figure shows that both correlated and block PM significantly outperform standard PM and also MH with respect to RCT. Figure 2 plots the Kernel Density Estimates (KDE) on the output from the three pseudo-marginal schemes and the exact MH approach. The figure shows that targeting a large $\sigma^2_{LL,m,n}$ (≈ 56) for the block and correlated PM samplers result in a very small bias in this application, with the approximation error in (3.4) being -0.01 for both the block and correlated PM and -0.0001 for the standard PM. Figure 2 suggests that this small perturbation is, for the correlated and block PM estimators, mostly due to β_3.

Finally, we also tried the exact subsampling approach (Quiroz et al., 2016) by setting $\text{E}[G] = 100$ (G is random in their approach) and the batch-size to 602, so that their prior

expected sample size corresponds to that for our correlated algorithms. For the RWM we obtained an RCT of ≈ 17 ($\omega = 3$) and ≈ 20 ($\omega = 1$) (average over parameters), with ω in (3.9).

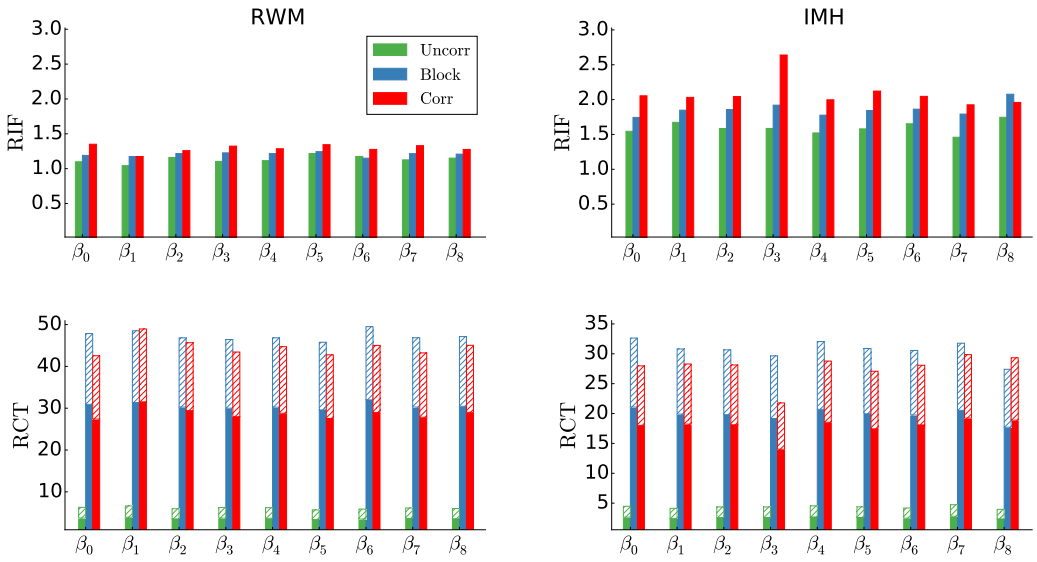

FIGURE 1. *Logistic regression example.* For algorithm $\mathcal{A}$ (uncorrelated (Uncorr), block (Block) and correlated (Corr) PM) the figure shows the Relative Inefficiency Factors (RIF) and Relative Computational Time for RWM proposal (left panel) and IMH (right panel). For RCT, the filled (dashed) bar correspond to $\omega = 3$ ($\omega = 1$) in (3.9).

4.3. **Other subsampling approaches.** We briefly discuss some other subsampling approaches and then benchmark them against our approach.

Korattikara et al. (2014) argue that using all data to take the simple decision to accept (or reject) a single parameter draw is a computationally inefficient strategy. Instead, they develop a sequence of t-tests, where each test is based on an increasing sample size and has a user specified error probability ϵ. The sequence is stopped when a decision of accepting (or rejecting) a single sample can be taken with a sufficiently small total error probability. They prove that the discrepancy between their approximate posterior and the true posterior can be made arbitrarily small by decreasing the total error probability of the test. However, it was empirically demonstrated in Bardenet et al. (2015) (see also Section 4.4) that the ϵ

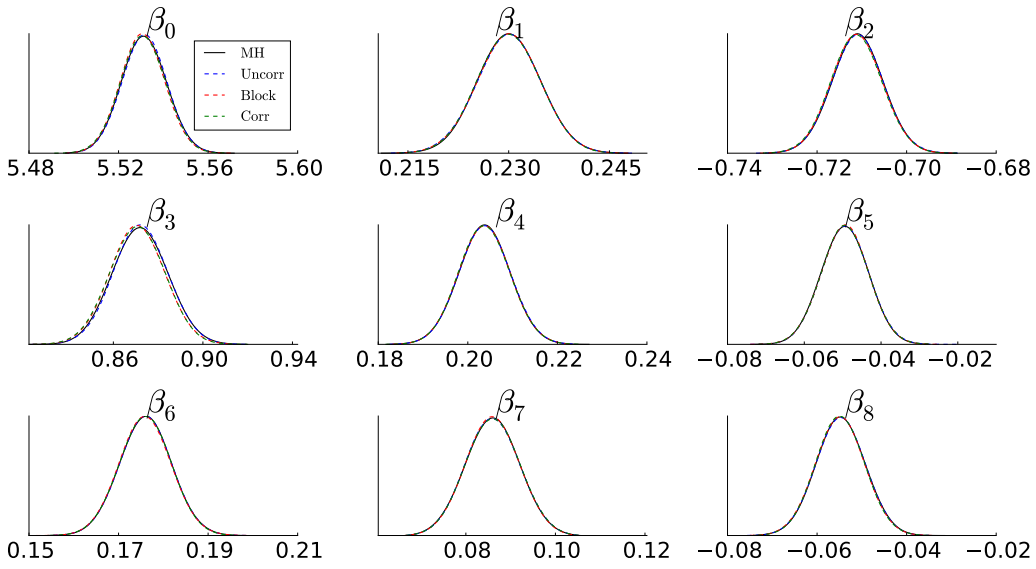

FIGURE 2. *Logistic regression example.* Kernel density estimates of marginal posteriors obtained by the IMH proposal. The figure shows the marginal posteriors obtained using the uncorrelated (Uncorr), block (Block) and correlated (Corr) PM (dashed blue, red and green, respectively) and MH (solid black line).

parameter must be set so low that virtually the full data set is required to obtain a reasonable approximation error.

Bardenet et al. (2014) also rely on the idea of replacing the computation of the MH ratio with a hypothesis test. However, they use an exact (not relying on a CLT) confidence interval obtained through a concentration inequality, which allows the choice of a subsample large enough so that a decision can be taken with a user specified error probability. They prove that the posterior targeted by their algorithm can become arbitrarily close to the true posterior as the error probability decreases. To successfully implement the method, the range of the differences in log-likelihood contributions at the proposed and current samples must be cheap to compute. Moreover, the performance of the method is highly dependent on the variance of these differences.

Bardenet et al. (2015) improve on the sampler in Bardenet et al. (2014) by introducing control variates to obtain variance reduction of the differences. Furthermore, the method does not rely on a cheaply computed range, however, it does rely on a bound for the difference

between the log-likelihood contributions at the proposed and current sample, and that of the control variates. When the control variates are obtained via Taylor series approximations, they suggest using the Taylor-Lagrange inequality to obtain the bound. We show in Section 4.4 that the Taylor-Lagrange bound can sometimes be too crude, and it is then necessary to compensate with a very large subsample. Bardenet et al. (2015) show that their method dramatically outperforms, among others, Korattikara et al. (2014), Bardenet et al. (2014) and Firefly Monte Carlo (Maclaurin and Adams, 2014, see below).

Firefly Monte Carlo in (Maclaurin and Adams, 2014, and more recently, Liu et al., 2015) introduces an auxiliary variable for each observation which determines if it should be included in the evaluation of the posterior. The distribution of these variables are such that when they are integrated out, the marginal posterior is the true posterior of θ. Moreover, a lower bound for each likelihood term is introduced, which basically plays the role of replacing the observations that are not included in the evaluation of the posterior. The authors suggest using the Gibbs sampler, generating the parameters conditional on the auxiliary variables and vice versa. The method has been documented to be very inefficient, see e.g. Bardenet et al. (2015), see also Section 4.4.

4.4. **AR processes.** The running examples in Bardenet et al. (2015) use the normal model which is too simplistic for our method because the log-likelihood is quadratic in the data and therefore our control variates are perfect. We instead compare our method to alternative approaches using the following illustrative models. We consider the following two AR(1) models with Student-t iid errors $\epsilon_t \sim t(5)$ with 5 degrees of freedom

$$\mathrm{M}_1 : y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t \qquad [\theta = (\beta_0 = 0.3, \beta_1 = 0.6)]$$

$$\mathrm{M}_2 : y_t = \mu + \varrho(y_{t-1} - \mu) + \epsilon_t \quad [\theta = (\mu = 0.3, \varrho = 0.99)]$$

with priors

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

where $\mathcal{U}(\cdot|a,b)$ is the uniform density on the interval $[a,b]$. Model M_2, the so called steady state AR, is particularly interesting as ϱ close to 1 gives a weakly identified μ, with a posterior that concentrates very slowly as n increases.

We compare our method to the Austerity MH (Korattikara et al., 2014), Firefly Monte Carlo (Maclaurin and Adams, 2014), the confidence sampler (Bardenet et al., 2014) and the confidence sampler with proxies (Bardenet et al., 2015). We set the tuning parameters of the competing algorithms following Bardenet et al. (2015) with the following exceptions. First, we adapt during the burn-in phase to reach an acceptance probability of $\alpha = 0.35$ (instead of $\alpha = 0.50$), which is optimal for RWM with two parameters (Gelman et al., 1996). For the pseudo-marginals we instead use $\alpha = 0.15$ as the five parameter example in Sherlock et al. (2015). Second, the p-value of the t-test in the Austerity MH algorithm is set to $\epsilon = 0.01$ (instead of $\epsilon = 0.05$) to put the approximation error of the method on par with the other methods. Setting $\epsilon = 0.05$ gives an unusably poor approximation (and also produces a much lower RCT than our methods). Additionally, the confidence sampler with proxies (from a Taylor series approximation with respect to θ) requires that the third derivative can be bounded uniformly for every observation and any θ. This bound is achieved, without any extra computational cost, by computing on a θ-grid where the posterior mass is located.

TABLE 2. *AR-process example.* Mean of sampling fraction $f = m/n$ over MCMC iterations for models M_1 and M_2 with MH, uncorrelated PM (Uncorr), block PM (Block) and correlated PM (Corr), confidence sampler (Conf), confidence sampler with proxies (ConfProxy), Austerity MH (AustMH), and Firefly Monte Carlo (Firefly).

	MH	Uncorr	Block	Corr	Conf	ConfProxy	AustMH	Firefly
M_1	1.000	0.093	0.037	0.037	1.493	0.160	1.037	0.100
M_2	1.000	0.291	0.117	0.116	1.490	1.500	1.019	0.137

Table 2 shows the mean of the sampling fraction over MCMC iterations. We note that both confidence samplers and the Austerity MH estimate the numerator and denominator in each iteration, and therefore require twice as many evaluations in a given iteration as MCMC (in some cases evaluations from the previous iteration can be reused). It is clear

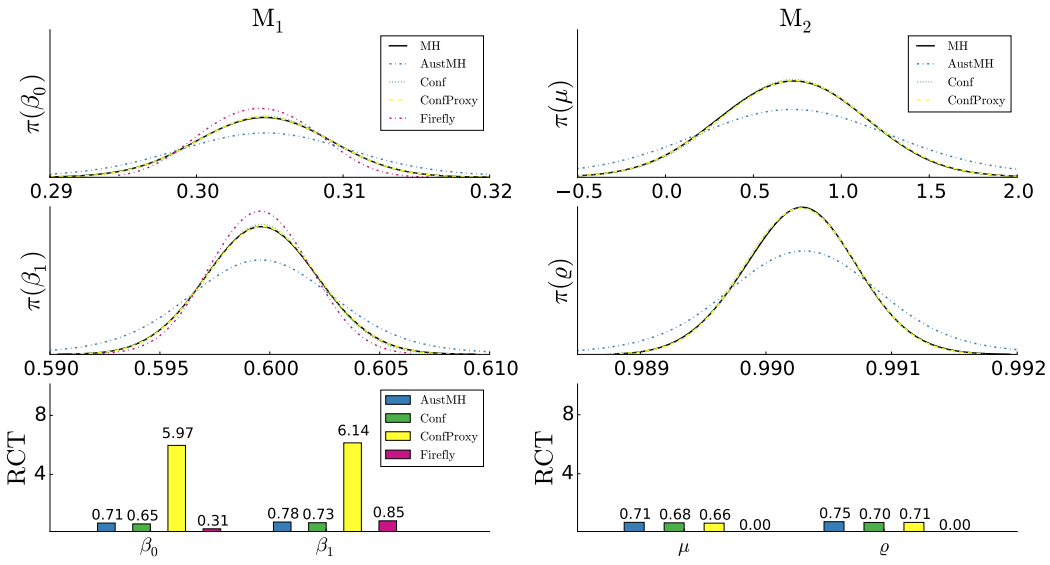

FIGURE 3. *AR-process example: Results for other subsampling algorithms.* The left and right panel, respectively, show the results for model M_1 and M_2. Each column shows the kernel density estimates of marginal posteriors (top two) and for algorithm $\mathcal{A}$ (confidence sampler (Conf), confidence sampler with proxies (ConfProxy), Austerity MH (AustMH), and Firefly Monte Carlo (Firefly)) the Relative Computational Time (RCT) (bottom).

that our algorithms makes very efficient use of a small subsample, especially the block and correlated PM samplers.

Figure 3 and 4 show the marginal posteriors obtained by, respectively, alternative sampling approaches and the several PM approaches. Moreover, the figures show the sampling efficiency of the different subsampling MCMC algorithms relative to that of the MH algorithm as measured by the Relative Computational Time. Figure 3 shows the striking result that many of these approaches are not more efficient than MH, except the confidence sampler with proxies for M_1. Regarding the approximation, it is evident that the Austerity MH still has a tempering effect (larger spread on the posterior) although ϵ is set so low that the full data set is sampled (see Table 2). The PM algorithms (and also the confidence samplers) provide excellent approximations: indeed, $\text{error}(\theta) \leq 10^{-6}$ in (3.4) for all our methods. Firefly Monte Carlo, although being an exact algorithm, is highly inefficient in this example, as also documented in Bardenet et al. (2015). In fact, for M_2, we were not able to obtain a

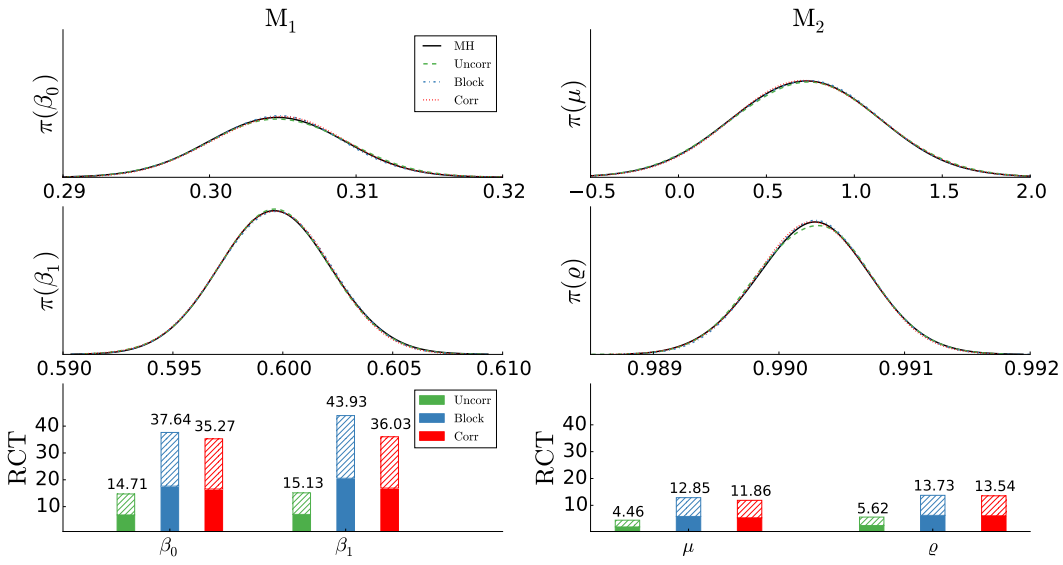

FIGURE 4. *AR-process example: Results for subsampling PM algorithms.* The left and right panel, respectively, show the results for model M_1 and M_2. Each column shows the kernel density estimates of marginal posteriors (top two) and for algorithm $\mathcal{A}$ (uncorrelated (Uncorr), block (Block) and correlated (Corr) PM) the Relative Computational Time (RCT) (bottom). For RCT, the filled (dashed) bar correspond to $\omega = 3$ ($\omega = 1$) in (3.9).

single effective sample out of $55,000$ iterations, and hence it was impossible to construct a kernel density estimate in this case. We also tried the exact subsampling in Quiroz et al. (2016), setting the tuning parameters to match the sample size used here as described in Section 4.2. For M_1 the sampler got stuck because the variance of the log of the estimator was too large (≈ 2070, compared to ≈ 12 for the estimator used here). For M_2 the exact subsampling produced an RCT of ≈ 4 ($\omega = 3$) and ≈ 7 ($\omega = 1$) (average over parameters), with ω in (3.9).

We conclude that the only viable subsampling MCMC approaches are the confidence sampler with proxies (Bardenet et al., 2015) and the PM approaches we propose. Moreover, a significant speed up is only obtained with the correlated PMs (both correlated and block). We acknowledge that we have put the confidence sampler with proxies in an unfavorable situation in M_2: the bound of its concentration inequality requires a bound of the remainder term in the Taylor series via the Taylor-Lagrange inequality, which is very hard for M_2.

We remark that the Taylor proxies with respect to θ work well in these two-dimensional examples, however, problems are encountered in the logistic example in Section 4.2; the expansion with respect to the data as proposed in Section 2.1 works much better. Moreover, an additional feature of expanding with respect to data is that an IMH proposal can be implemented straightforwardly, as the control variates are accurate for any θ.

5. Conclusions and Future Research

We propose a framework for speeding up MCMC for data sets with many observations by data subsampling. The following features are key to our approach. First, we introduce a highly efficient log-likelihood estimator, which incorporates information about each observation's contribution to the log-likelihood function, while only operating on a sparse set of the data. This results in a substantially lower computational cost. Second, we use the resulting likelihood estimate within a pseudo-marginal framework and sample from a perturbed posterior which, for fixed n, we prove to be within $O(m^{-2})$ of the true posterior. We also consider the asymptotic behavior with respect to n. Moreover, we provide a useful heuristic to approximate the error, given that m is large. Third, we propose a correlated pseudo-marginal approach to subsampling, which allows highly variable estimates of the likelihood without adversely affecting the mixing of the algorithm. The resulting algorithm is a highly efficient algorithm when taking into account the statistical efficiency and computational cost. Fourth, we use the correlated and block PM samplers to show that we can let the sample size m grow much more slowly as a function of n than the standard PM sampler to achieve the same inefficiency, and hence a much lower computational time. Finally, we document large speed ups relative to MH and, more importantly, we show that our method outperforms other recent subsampling approaches in the literature.

Future research concerns designing efficient proposals based on data subsampling, e.g. in hybrid Monte Carlo algorithms. It is also of interest to develop improved clustering methods to obtain control variates, especially in the presence of a large number of covariates.

6. Acknowledgments

Matias Quiroz was partially supported by VINNOVA grant 2010-02635. Robert Kohn's research was partially supported by Australian Research Council Center of Excellence grant CE140100049. Minh-Ngoc Tran was partially supported by a Business School Pilot Research grant. We thank the authors in Bardenet et al. (2015) for making their code publicly available, which facilitated the comparison against other subsampling approaches.

References

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

Bardenet, R., Doucet, A., and Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of The 31st International Conference on Machine Learning*, pages 405–413.

Bardenet, R., Doucet, A., and Holmes, C. (2015). On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Ceperley, D. and Dewing, M. (1999). The penalty method for random walks with uncertain energies. *The Journal of chemical physics*, 110(20):9812–9820.

Dahlin, J., Lindsten, F., Kronander, J., and Schön, T. B. (2015). Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. *arXiv preprint arXiv:1511.05483*.

Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2016). The correlated pseudo-marginal method. *arXiv preprint arXiv:1511.04992v3*.

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

asu075.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2014). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, 49(4):1071–1099.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Jacob, P. E. and Thiery, A. H. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784.

Korattikara, A., Chen, Y., and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 181–189.

Liu, S., Mingas, G., and Bouganis, C.-S. (2015). An exact MCMC accelerator under custom precision regimes. In *Field Programmable Technology (FPT), 2015 International Conference on*, pages 120–127. IEEE.

Lyne, A.-M., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467.

Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664.

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Nemeth, C. and Sherlock, C. (2016). Merging MCMC subposteriors through Gaussian-process approximations. *arXiv preprint arXiv:1605.08576*.

Nicholls, G. K., Fox, C., and Watt, A. M. (2012). Coupled MCMC with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*.

Pitt, M. K., Silva, R. d. 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.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Quiroz, M., Villani, M., and Kohn, R. (2016). Exact subsampling MCMC. *arXiv preprint arXiv:1603.08232*.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E., and McCulloch, R. (2013). Bayes and big data: the consensus Monte Carlo algorithm. In *EFaBBayes 250" conference*, volume 16.

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

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016a). Block-wise pseudo-marginal Metropolis-Hastings. *arXiv preprint arXiv:1603.02485v2*.

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2016b). Importance sampling squared for Bayesian inference in latent variable models. *arXiv preprint arXiv:1309.3339v4*.

Wang, X. and Dunson, D. B. (2013). Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

SPEEDING UP MCMC BY EFFICIENT DATA SUBSAMPLING

MATIAS QUIROZ, MATTIAS VILLANI, ROBERT KOHN AND MINH-NGOC TRAN

ABSTRACT. We propose Subsampling MCMC, a Markov Chain Monte Carlo (MCMC) framework where the likelihood function for n observations is estimated from a random subset of m observations. We introduce a general and highly efficient unbiased estimator of the log-likelihood based on control variates obtained from clustering the data. The cost of computing the log-likelihood estimator is much smaller than that of the full log-likelihood used by standard MCMC. The likelihood estimate is bias-corrected and used in two correlated pseudo-marginal algorithms to sample from a perturbed posterior, for which we derive the asymptotic error with respect to n and m, respectively. A practical estimator of the error is proposed and we show that the error is negligible even for a very small m in our applications. We demonstrate that Subsampling MCMC is substantially more efficient than standard MCMC in terms of sampling efficiency for a given computational budget, and that it outperforms other subsampling methods for MCMC proposed in the literature.

KEYWORDS: Bayesian inference, Estimated likelihood, Correlated pseudo-marginal, Block pseudo-marginal, Big Data, Survey sampling.

1. INTRODUCTION

The popularity of Bayesian methods increased significantly in the early 90's due to advances in computer technology and the introduction of powerful simulation algorithms such as Markov Chain Monte Carlo (MCMC) (Gelfand and Smith, 1990). However, posterior sampling with MCMC is time-consuming and there is an increasing awareness that new scalable algorithms are necessary for MCMC to remain an attractive choice for inference in data sets with a large number of observations.

Quiroz: *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University and Research Division, Sveriges Riksbank. E-mail: quiroz.matias@gmail.com.* Villani: *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University.* Kohn: *Australian School of Business, University of New South Wales.* Tran: *Discipline of Business Analytics, University of Sydney.* The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank.

Current research on scalable MCMC algorithms belongs to two major groups. The first group employs parallelism through the typical MapReduce scheme (Dean and Ghemawat, 2008) by partitioning the data and computing posteriors in a parallel and distributed manner. The resulting draws are subsequently combined into a single posterior distribution. The main difference within this group is how weighting is performed and whether the partitions communicate at runtime, see for example Scott et al., 2013; Neiswanger et al., 2013; Wang and Dunson, 2013; Minsker et al., 2014; Nemeth and Sherlock, 2016. Our approach belongs to the second group of methods that use a subsample of the data in each MCMC iteration to speed up the algorithm, which we refer to as subsampling MCMC, see Korattikara et al. (2014); Bardenet et al. (2014); Maclaurin and Adams (2014); Bardenet et al. (2015); Liu et al. (2015). Section 4.3 briefly outlines these approaches and Section 4.4 compares them against our methods. For a more extensive introduction to these methods and a broad overview of the problem in general, see the excellent review in Bardenet et al. (2015).

Our article presents a Metropolis-Hastings (MH) framework where the likelihood is estimated from a random subset of the data using highly efficient control variates for variance reduction. For models with an intractable likelihood function, Beaumont (2003) proposes to estimate the likelihood unbiasedly and run the MH algorithm on an extended space, which also includes the random variates underlying the likelihood estimate. Andrieu and Roberts (2009) develop theory for such Pseudo-Marginal MH (PM) algorithms, and prove that PM algorithms target the true posterior if the likelihood estimator is unbiased and almost surely positive. Obtaining unbiased likelihood estimators with low variability from subsampling is a major challenge, and previous attempts have failed to produce an MCMC sampler that does not get stuck (Korattikara et al., 2014; Bardenet et al., 2015). Moreover, ensuring that the unbiased likelihood estimator is also positive was shown by Jacob and Thiery (2015) to only be possible under assumptions that can only be satisfied by sampling the full data set (Bardenet et al., 2015). Quiroz et al. (2016) use the insights and techniques proposed here (control variates and correlated PM for subsampling) to produce an estimator with low variability and, in addition, target the absolute value of the estimate following Lyne et al.

(2015) that allow the estimator to occasionally be negative. Draws from the algorithm are corrected with an importance sampling step to obtain unbiased estimates of expectations of posterior functions. Let m and n be the subsample and population sizes. Our article provides an alternative approach that instead simulates by PM, but from a slightly perturbed target (because the likelihood estimator is slightly biased), where (i) the error can be estimated and made arbitrarily small at the rate $O(m^{-2})$ with n fixed and (ii) a smaller variance of the logarithm of the estimator of the likelihood can be achieved thus requiring a smaller m. We also study the error with respect to the number of observations n, when $m = m(n)$ and under certain assumptions of the control variates.

The variance of the estimator of the log-likelihood is crucial for the performance of PM algorithms: a large variance can easily produce extreme over-estimates of the likelihood and cause the Markov chain to get stuck for long periods. On the other hand, a too precise likelihood estimator might be unnecessarily costly. Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) analyze the variance of the log-likelihood estimator $\sigma^2_{LL,m,n}$ that maximizes the number of effective draws per unit of computing time. They conclude that the optimal number of particles m should be such that $\sigma^2_{LL,m,n}$ is around 1. Moreover, $m = O(n)$ is required to obtain the optimal value of the variance. Recent advances in PM algorithms correlate the particles at the current and proposed parameter value in the MH ratio (Deligiannidis et al., 2016; Dahlin et al., 2015) or use blocking (Tran et al., 2016a). This makes it possible to target $\sigma^2_{LL,m,n} \gg 1$ and the optimal variance can be obtained with $m = O(n^{1/2})$ (Deligiannidis et al., 2016; Tran et al., 2016a). Our article proposes a correlated pseudo-marginal approach for data subsampling that uses a copula transformation of the random variates in Deligiannidis et al. (2016). Tran et al. (2016a) derive an explicit expression of the correlation of the log-likelihood estimator at the current and proposed draw, which we use to compute the optimal subsample size $m = m(n)$ for our algorithm.

The paper is organized as follows. Section 2 introduces the general estimator and derives some important properties. Section 3 outlines the subsampling MCMC algorithm and its theoretical framework, including results on the accuracy of the perturbed posterior, and how

to select the rate of m to achieve the optimal $\sigma^2_{LL,m,n}$. Section 4 reviews other subsampling approaches and evaluates the performance of the proposed methodology on two examples. The second of these examples benchmarks our methods against the other approaches. Implementation details and proofs are placed in the online Appendices A and B.

2. Sampling-based Log-likelihood Estimators

2.1. **A log-likelihood estimator based on simple random sampling with efficient control variates.** Let $\{y_i, x_i\}_{i=1}^n$ denote the data, where y is a response vector and x is a vector of covariates. Let $\theta \in \Theta$ be the vector of parameters. Given conditionally independent observations we have the usual decomposition of the log-likelihood

$$\ell_{(n)}(\theta) \coloneqq \sum_{i=1}^{n} \ell_i(\theta), \quad \text{where } \ell_i(\theta) \coloneqq \log p(y_i|\theta, x_i) \tag{2.1}$$

is the *log-likelihood contribution* of the ith observation. For any given θ, (2.1) is a sum of a finite number of elements and estimating it is equivalent to the classical survey sampling problem of estimating a population total. See Särndal et al. (2003) for an introduction. We assume in (2.1) that the log-likelihood decomposes as a sum of terms where each term depends on a unique piece of data information. This applies to longitudinal problems where $\ell_i(\theta)$ is the log joint density of all measurements on the ith subject, and we sample subjects rather than individual observations. It also applies to certain time-series problems such as AR(p) processes, where the sample elements become $(y_t, \ldots, y_{t-p})$, for $t = p+1, \ldots, n$. Our examples in Section 4 use independent identically distributed (iid) and time series data.

Estimating (2.1) based on Simple Random Sampling (SRS), where any $\ell_i(\theta)$ is included with the same probability generally results in a dramatically large variance. Intuitively, since some $\ell_i(\theta)$ contribute significantly more to the sum in (2.1) they should be included in the sample with a larger probability, using so called Probability Proportional-to-Size (PPS) sampling. However, this requires each of the n sampling probabilities to be proportional to a measure of their size. Evaluating n size measures is likely to defeat the purpose of subsampling, except in cases when there is a computationally cheaper proxy than $\ell_i(\theta)$ that

can be utilized instead. Alternatively, one can make the $\{\ell_i(\theta)\}_{i=1}^n$ more homogeneous by using control variates so that the population elements are roughly of the same size and SRS is then expected to be efficient. Our article focuses on this case and proposes efficient control variates $q_{i,n}(\theta)$ such that the computational cost of the estimator is substantially less than $O(n)$. The dependence on n is due to $q_{i,n}(\theta)$ being an approximation of $l_i(\theta)$, which typically improves as more data is available.

Define the differences $d_{i,n}(\theta) \coloneqq \ell_i(\theta) - q_{i,n}(\theta)$ and let

$$\mu_{d,n}(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} d_{i,n}(\theta) \quad \text{and } \sigma^2_{d,n}(\theta) \coloneqq \frac{\sum_{i=1}^n \left(d_{i,n}(\theta) - \mu_{d,n}(\theta)\right)^2}{n}$$

be the mean and variance of the finite population $\{d_{i,n}(\theta)\}_{i=1}^n$. Let $u_1, \ldots, u_m$ be iid random variables such that $\Pr(u = k) = 1/n$ for $k = 1, \ldots, n$. The Difference Estimator (DE, Särndal et al., 2003) of $\ell_{(n)}(\theta)$ in (2.1) is

$$\widehat{\ell}_{(m,n)}(\theta) \coloneqq q_{(n)}(\theta) + n\widehat{\mu}_{d,n}(\theta), \quad \widehat{\mu}_{d,n}(\theta) \coloneqq \frac{1}{m}\sum_{i=1}^{m} d_{u_i,n}(\theta) \tag{2.2}$$

with $q_{(n)}(\theta) \coloneqq \sum_{i=1}^n q_{i,n}(\theta)$. It is straightforward to use unequal sampling probabilities with the DE, but the sampling probabilities need to be evaluated for every observation, which can be costly. The following Lemma gives some basic properties of the DE estimator.

Lemma 1. *Suppose that $\widehat{\ell}_{(m,n)}$ is the estimator of $\ell_{(n)}(\theta) = \ell(\theta)$ given by (2.2). Then*

i. $\mathrm{E}[\widehat{\mu}_{d,n}(\theta)] = \mu_{d,n}(\theta)$.

ii.
$$\mathrm{E}\left[\widehat{\ell}_{(m,n)}(\theta)\right] = l_{(n)}(\theta) \quad \textit{and} \quad \sigma^2_{LL,m,n} = \mathrm{V}\left[\widehat{\ell}_{(m,n)}(\theta)\right] = \frac{n^2\sigma^2_{d,n}}{m}.$$

iii. *$\widehat{\ell}_{(m,n)}(\theta)$ is asymptotically normal when $m \to \infty$ for fixed n and $\sigma^2_{d,n} < \infty$, or when both $m, n \to \infty$ with $m = Bn^\gamma$ for constants $B > 0$ and $\gamma > 0$ and $\sigma^3_{d,n} < \infty$.*

Proof. The proofs of parts i) and ii) are straightforward and omitted. The proof of iii) is in Appendix B. □

The assumptions of finite $\sigma^2_{d,n}$ and $\sigma^3_{d,n}$ in Lemma 1 part (iii) are highly non-restrictive because the random variables are discrete with a finite sample space: they are satisfied for any control variates that are finite. We use the following estimate of $\sigma^2_{d,n}$

$$\widehat{\sigma}^2_{d,n}(\theta) \coloneqq \frac{\sum_{i=1}^m \left(d_{u_i,n}(\theta) - \widehat{\mu}_{d,n}(\theta)\right)^2}{m}.$$

2.2. **Control variates for variance reduction.** To see the crucial role of variance reduction using control variates, we first note that the variance of the log-likelihood estimator should be around one for the standard PM (see e.g. Pitt et al., 2012; Doucet et al., 2015 and Section 1). Now, define

$$a_n \coloneqq \sup_{\theta\in\Theta} \sup_{i\in\{1,\ldots,n\}} \left|d_{i,n}(\theta) - \mu_{d,n}(\theta)\right|. \tag{2.3}$$

Throughout our article we assume that $a_n < \infty$ for a given n and also that $\limsup a_n < \infty$. This assumption is easily enforced if Θ is a compact space as long as the control variates are finite. The following lemma is straightforward to prove.

Lemma 2. *Suppose that* $\limsup a_n < \infty$. *Then* $\sigma^2_{d,n} = O(a_n^2)$ *and* $\sigma^2_{LL,m,n} = \frac{n^2 a_n^2 O(1)}{m}$.

According to Lemma 2, keeping the variance bounded as a function of n requires that $\frac{n^2 O(a_n^2)}{m} = O(1)$. This highlights the importance of the variance reduction: SRS without control variates scales poorly because $O(a_n^2) = O(1)$ and $m = O(n^2)$. On the other hand, with control variates that improve as, say $d_{i,n} = O(n^{-\alpha})$ with $\alpha > 0$, we have $O(a_n^2) = O(n^{-2\alpha})$ and $m = O(n^{2(1-\alpha)})$.

2.3. **Computational complexity.** The difference estimator in (2.2) needs to compute $q_{(n)}(\theta) = \sum_{i=1}^n q_{i,n}(\theta)$ in every MCMC iteration: this requires evaluating the control variates $q_{i,n}(\theta)$ for all data points. We now explore specific choices of $q_{i,n}$ that allow us to compute $\sum_{i=1}^n q_{i,n}(\theta)$ using substantially less evaluations than n. Denote the Computational Cost (CC) for the standard MH without subsampling which evaluates $\ell_{(n)} \coloneqq \sum_{i=1}^n \ell_i$ by

$$\mathrm{CC}[\ell_{(n)}(\theta)] \coloneqq n \cdot c_\ell,$$

where c_ℓ is the cost of evaluating a single log-likelihood contribution (assuming the cost is the same for all i). For the difference estimator in (2.2), we have

$$\mathrm{CC}\left[\widehat{\ell}_{(m,n)}(\theta)\right] \coloneqq n \cdot c_q + m \cdot c_\ell,$$

where c_q is the cost of computing a control variate. We now briefly describe two particular control variates that reduce the first term $n \cdot c_q$. See Appendix A for details regarding their implementation.

First, consider the control variates in Bardenet et al. (2015) who propose to use a second order Taylor expansion of each $\ell_i(\theta)$ around some reference value $\theta^\star$, e.g. the maximum likelihood estimate. This reduces the complexity from n evaluations to a single one (similar to sufficient statistics for a normal model because $q_{i,n}(\theta)$ is quadratic in θ). As noted by Bardenet et al. (2015), this control variate is a poor approximation of $\ell_i(\theta)$ whenever the algorithm proposes a θ that is not near to $\theta^\star$, and will therefore work well only when the posterior is tightly concentrated around $\theta^\star$. As a remedy in the case of a less concentrated posterior, they suggest to occasionally recompute the control variates, expanding around the current θ in the MCMC (using all n observations).

We now propose a control variate that works well regardless of the posterior concentration. This control variate is based on clustering the data $\{z_i = (y_i, x_i)\}_{i=1}^n$ into K clusters that are kept fixed throughout the MCMC. At every MCMC iteration, we compute the exact log-likelihood contributions at all K centroids and use a second order Taylor expansion with respect to z_i at the centroid as a local approximation of ℓ_i around each centroid. This allows us to compute $\sum_{i=1}^n q_{i,n}(\theta)$ by simply scaling up quantities computed at the K centroids. The resulting estimator therefore has cost

$$\mathrm{CC}\left[\widehat{\ell}_{(m,n)}(\theta)\right] = K \cdot c_q + m \cdot c_\ell, \tag{2.4}$$

where typically $K \ll n$.

3. Subsampling MCMC Methodology

3.1. **MCMC with likelihood estimators from data subsampling.** We propose an efficient unbiased estimator $\widehat{\ell}_{(m,n)}(\theta)$ of the log-likelihood and then approximately bias-correct following Ceperley and Dewing (1999) (see also Nicholls et al., 2012) to obtain the 'bias-corrected' likelihood estimator

$$\widehat{L}_{(m,n)}(\theta, u) \coloneqq \exp\left(\widehat{\ell}_{(m,n)}(\theta) - \frac{n^2}{2m}\widehat{\sigma}^2_{d,n}(\theta)\right), \tag{3.1}$$

where $\widehat{\ell}_{(m,n)}(\theta)$ and $\widehat{\sigma}^2_{d,n}(\theta)$ are the estimators presented in Section 2.1. The form of (3.1) is motivated by the case when $\widehat{\ell}_{(m,n)} \sim \mathcal{N}(\ell_{(n)}(\theta), \sigma^2_{LL,m,n}(\theta))$ and $\sigma^2_{LL,m,n}$ is known, in which case all bias is removed. Normality holds asymptotically in both m and n by part (iii) of Lemma 1. However, the assumption of known variance is unrealistic because the computation requires the entire data set. The estimator in (3.1) is therefore expected to only be nearly unbiased. A main difference of our use of this estimator compared to Ceperley and Dewing (1999) and Nicholls et al. (2012) is that our approach is a pseudo-marginal, where the space explored by the Markov chain also includes the random variates used for estimating the likelihood. Other differences are that we use control variates, that we analyze the dependence on n in the analysis of the error, and that our convergence rate of the error (Theorem 1 below) is $O(m^{-2})$ as opposed to $O(m^{-1})$ in Nicholls et al. (2012).

We now outline how to carry out a pseudo-marginal MH with the approximately unbiased estimator in (3.1) and derive the asymptotic error in the stationary distribution. Denote the likelihood by $L_{(n)}(\theta) \coloneqq p(y|\theta)$, let $p_\Theta(\theta)$ be the prior and define the marginal likelihood $\overline{L}_{(n)} \coloneqq \int L_{(n)}(\theta)p_\Theta(\theta)d\theta$. Then the posterior is $\pi_{(n)}(\theta) = L_{(n)}(\theta)p_\Theta(\theta)/\overline{L}_{(n)}$. Let $p_U(u)$ be the distribution of the vector u of auxiliary variables corresponding to the subset of observations to include when estimating $L_{(n)}(\theta)$. Let $\widehat{L}_{(m,n)}(\theta, u)$, for fixed m and n, be a possibly biased estimator of $L_{(n)}(\theta)$ with expectation

$$L_{(m,n)}(\theta) = \int \widehat{L}_{(m,n)}(\theta, u)p_U(u)du.$$

Define

$$\overline{\pi}_{(m,n)}(\theta,u) \coloneqq \widehat{L}_{(m,n)}(\theta,u)p_U(u)p_\Theta(\theta)/\overline{L}_{(m,n)}, \text{ with } \overline{L}_{(m,n)} \coloneqq \int L_{(m,n)}(\theta)p_\Theta(\theta)d\theta, \tag{3.2}$$

on the augmented space (θ, u). It is straightforward to show that $\overline{\pi}_{(m,n)}(\theta,u)$ is a proper density with marginal

$$\overline{\pi}_{(m,n)}(\theta) \;=\; \int \overline{\pi}_{(m,n)}(\theta,u)du = L_{(m,n)}(\theta)p_\Theta(\theta)/\overline{L}_{(m,n)}.$$

The MCMC that targets (3.2) uses a joint proposal for θ and u given by

$$q_{\Theta,U}(\theta,u|\theta_c,u_c) = p_U(u)q_\Theta(\theta|\theta_c)$$

where c denotes the current state of the Markov chain. The PM acceptance probability becomes

$$\alpha = \min\left(1, \frac{\widehat{L}_{(m,n)}(\theta_p,u_p)p_\Theta(\theta_p)/q_\Theta(\theta_p|\theta_c)}{\widehat{L}_{(m,n)}(\theta_c,u_c)p_\Theta(\theta_c)/q_\Theta(\theta_c|\theta_p)}\right). \tag{3.3}$$

This expression is similar to the MH acceptance probability, but with the true likelihood replaced by its estimate. By Andrieu and Roberts (2009), the draws of θ obtained by this MH algorithm have $\overline{\pi}_{(m,n)}(\theta)$ as invariant distribution. If $\widehat{L}_{(m,n)}(\theta,u)$ is an unbiased estimator of $L_{(n)}(\theta)$, then the marginal of the augmented MCMC scheme above has $\overline{\pi}_{(m,n)}(\theta) = \pi_{(n)}(\theta)$ (the true posterior) as invariant distribution. However, if $\widehat{L}_{(m,n)}(\theta,u)$ is biased, the sampler is still valid but has a perturbed marginal $\overline{\pi}_{(m,n)}(\theta)$.

3.2. **Perturbation analysis - asymptotics.** Our next result gives the rate at which the perturbed target $\pi_{(m,n)}(\theta)$ approaches the true target posterior $\pi_{(n)}(\theta)$.

Theorem 1. *Suppose that a PM algorithm is implemented with the estimator in* $\widehat{L}_{(m,n)}(\theta,u)$ *in* (3.1) *and assume that* $n^3a_n^3/m^2 = o(1)$. *The following results hold for any* $\theta \in \Theta$, *where* Θ *is a compact space,*

i.
$$\frac{\left|L_{(m,n)}(\theta) - L_{(n)}(\theta)\right|}{L_{(n)}(\theta)} \leq O\left(\frac{n^2 a_n^2}{m^2}\right).$$

ii.
$$\frac{\left|\overline{\pi}_{(m,n)}(\theta) - \pi_{(n)}(\theta)\right|}{\pi_{(n)}(\theta)} \leq O\left(\frac{n^2 a_n^2}{m^2}\right).$$

iii. *Suppose that* $h(\theta)$ *is a function such that* $\mathrm{E}_{\pi_{(n)}}[|h(\theta)|] < \infty$. *Then*

$$\left|\frac{\mathrm{E}_{\overline{\pi}_{(m,n)}}[h(\theta)] - \mathrm{E}_{\pi_{(n)}}[h(\theta)]}{\mathrm{E}_{\pi_{(n)}}[h(\theta)]}\right| \leq O\left(\frac{n^2 a_n^2}{m^2}\right).$$

Note that we need to have $m = n^2 O(a_n^2)$ to target $\sigma^2_{LL,m,n} = \frac{n^2 O(a_n^2)}{m}$ at the optimal value around 1. By Theorem 1, this m gives a fractional error of the posterior which is within $O(n^{-2}a_n^{-2})$ of the true posterior. Note also that the perturbation error is $O\left(m^{-2}\right)$ for a fixed n.

3.3. **Approximating the perturbation error.** Theorem 1 is derived under essentially no assumptions on the estimator, and clearly displays the fast convergence of our perturbed posterior, but it does not provide a practically useful way to quantify the discrepancy between $\overline{\pi}_{(m,n)}(\theta)$ and $\pi_{(n)}(\theta)$. We now propose a way to estimate the point-wise fractional error in the perturbed posterior distribution

$$\text{error}(\theta) = \frac{\pi_{(m,n)}(\theta) - \pi_{(n)}(\theta)}{\pi_{(n)}(\theta)} = \left(\frac{L_{(m,n)}(\theta)}{L_{(n)}(\theta)}\right) \Big/ \left(\frac{\overline{L}_{(m,n)}}{\overline{L}_{(n)}}\right) - 1. \tag{3.4}$$

The following lemma is an application of the bivariate Central Limit Theorem (CLT).

Lemma 3.

$$\sqrt{m}\left(\begin{bmatrix}\widehat{\mu}_{d,n}\\ \widehat{\sigma}^2_{d,n}\end{bmatrix} - \begin{bmatrix}\mu_{d,n}\\ \sigma^2_{d,n}\end{bmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}\sigma^2_{d,n} & \varphi^{(3)}_{d,n}\\ \varphi^{(3)}_{d,n} & \sigma^4_{d,n} - \varphi^{(4)}_{d,n}\end{bmatrix}\right) \quad \text{as } m \to \infty,$$

with $\varphi^{(b)}_{d,n} = E[(d_{u_i,n} - \mu_{d,n})^b] = \sum_{i=1}^n (d_{i,n} - \mu_{d,n})^b / n$ *for* $b \geq 1$.

Lemma 4. *Suppose that*

$$\begin{bmatrix}\widehat{\mu}_{d,n}\\ \widehat{\sigma}^2_{d,n}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_{d,n}\\ \sigma^2_{d,n}\end{bmatrix}, \bar{\Sigma} = \frac{1}{m}\begin{bmatrix}\sigma^2_{d,n} & \varphi^{(3)}_{d,n}\\ \varphi^{(3)}_{d,n} & \sigma^4_{d,n}-\varphi^{(4)}_{d,n}\end{bmatrix}\right). \tag{3.5}$$

Then, with $\Psi^{(b)}_{d,n} \coloneqq \varphi^{(b)}_{d,n}/\sigma^b_{d,n}$ *for* $b \geq 1$,

$$L_{(m,n)}(\theta) = \exp\left(\ell_{(n)} + \frac{\sigma^4_{LL,m,n}(\theta)}{8m}\left(1-\Psi^{(4)}_{d,n}(\theta)\right) - \frac{\sigma^3_{LL,m,n}(\theta)}{2\sqrt{m}}\Psi^{(3)}_{d,n}(\theta)\right). \tag{3.6}$$

Proof. Since $L_{(m,n)}(\theta) = \exp\left(q_{(n)}(\theta)\right)\mathrm{E}\left[\exp\left(n\widehat{\mu}_{d,n} - \frac{n^2}{2m}\widehat{\sigma}^2_{d,n}\right)\right]$ the result follows from the moment generating function (mgf) of the bivariate normal distribution in (3.5). □

It is easy to show that $\gamma^{(b)}_{d,n} = O(1)$ for any $b \geq 1$. From Lemma 4 it follows that the perturbation error in the likelihood $L_{(m,n)}(\theta)/L_{(n)}(\theta)$ depends on $\sigma_{LL,m,n}$, and will increase with it for sufficiently large $\sigma_{LL,m,n}$. It is important to note, however, that any constant factor c in $L_{(m,n)}(\theta)/L_{(n)}(\theta)$ that does not depend on θ will cancel out in the fractional posterior error in (3.4) since the same factor c will also appear in $\overline{L}_{(m,n)}/\overline{L}_{(n)}$. This observation leads to the following theoretically interesting remark.

Remark 1. *Suppose we run a PM algorithm and that we can, for any proposed* $\theta \in \Theta$*, where* Θ *is a compact space, choose* $m(\theta)$ *such that*

$$\frac{\sigma^4_{LL,m(\theta),n}(\theta)}{8m(\theta)}\left(1-\Psi^{(4)}_{d,n}(\theta)\right) - \frac{\sigma^3_{LL,m(\theta),n}(\theta)}{2\sqrt{m(\theta)}}\Psi^{(3)}_{d,n}(\theta) = c. \tag{3.7}$$

Then $\mathrm{error}(\theta) = 0$.

The constant c is an arbitrary choice: any c generates a specific $m_c(\theta)$ which ensures unbiasedness for any $\theta \in \Theta$. A natural choice of c would be to solve (3.7) based on the $m(\theta^\star)$ that targets the optimal $\sigma^2_{LL,m(\theta^\star),n}$, where $\theta^\star$ is the mode. Of course, this strategy uses all data and is thus not applicable, but illustrates an important property of our method.

In practice we can instead use the result in Lemma 4 to check $\mathrm{error}(\theta)$ in any given application as follows. The quantities $\sigma_{LL,m,n}(\theta)$ and $\gamma^{(b)}_{d,n}$ can be easily evaluated for any θ at the cost of evaluating $\ell_i(\theta)$ for all $i = 1, ..., n$, or estimated from a subsample. It

is also necessary to evaluate $\overline{L}_{(n)}$ and $\overline{L}_{(m,n)}$, which can be done with the usual Laplace approximation. Approximating $\overline{L}_{(n)}$ requires the Hessian of $\log L_{(n)}(\theta)$ evaluated at the mode, which can be obtained analytically from tedious differentiation or numerically by finite differences. A similar procedure applies for $\overline{L}_{(m,n)}$, but with $\log L_{(m,n)}$ from (3.6) where clearly the Hessian becomes analytically intractable, but finite differences are straightforward.

3.4. **Correlated proposals of u for subsampling.** Deligiannidis et al. (2016) and Dahlin et al. (2015) both propose a general method that correlates the current and proposed values of u_i. The advantage of this correlation is that it makes the variance of the difference in the logarithms of the estimated likelihoods appearing in (3.3) much smaller than that of each of the terms themselves. This leads in our context to requiring much smaller values of m, or equivalently, that we can target much higher values of $\sigma^2_{LL,m,n}$ than unity, provided we also check that error(θ) remains at an acceptable level.

For a correlated PM approach to subsampling, we let u be a vector of length n with binary elements u_i that determine if observation i is included ($u_i = 1$) when estimating the log-likelihood. Note that this is different from the above, where u contained the observation indices and was of length m. Moreover, here the sample size is random and we let $m^\star$ be the expected sample size. The sampling probabilities become $\Pr(u_i = 1) = m^\star/n$ for $i = 1, \ldots, n$. We use the auxiliary variable (particle) v in Deligiannidis et al. (2016) to induce dependence at the current u_i^c and proposed u_i^p sampling indicator through a Gaussian copula as we now explain. The correlated pseudo-marginal method uses a Gaussian auto-regressive kernel $\mathcal{K}(v_c, v_p)$ with a transition defined by $v_p = \phi v_c + \sqrt{1-\phi^2}\varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$. We also have $v_c \sim p(v) = \mathcal{N}(v|0,1)$ and $\mathcal{K}(v_c, v_p)$ is reversible with respect to $p(v)$. We sample the u_i's by first generating v_c and v_p and set $u_i^c = \mathcal{I}\left[\Phi(v_c) \leq \frac{m^\star}{n}\right]$ and $u_i^p = \mathcal{I}\left[\Phi(v_p) \leq \frac{m^\star}{n}\right]$, where Φ denotes the standard normal cdf. An equivalent approach is to generate u_i^p from a Markov chain with marginal $p(u_i^c = 1) = m^\star/n$, with transition probabilities $\Pr(u_i^p = 1|u_i^c = 1) = \kappa$ and $\Pr(u_i^p = 0|u_i^c = 0) = 1 - (1-\kappa)\frac{m^\star/n}{1-m^\star/n}$. The persistence parameter κ in the Markov chain is related to the AR persistence ϕ by the relation $\kappa = \frac{n}{m^\star}\Phi_2\left(\Phi^{-1}(m^\star/n), \Phi^{-1}(m^\star/n)|\phi\right)$, where $\Phi_2(\cdot,\cdot|\phi)$ is the cdf of bivariate standard normal variables with correlation ϕ.

As noted above, in contrast to Section 2.1, u is a binary vector. We can instead use the Horvitz-Thompson (Horvitz and Thompson, 1952) which (under SRS) is

$$\widehat{d}_{(m^\star,n)} = \sum_{i=1}^{n} \frac{d_{i,n}}{m^\star/n} u_i,$$

and is unbiased for $d_{(n)}$. Note that we can write

$$\widehat{d}_{(m^\star,n)} = \frac{1}{m^\star} \sum_{i=1}^{n} n d_{i,n} u_i, \quad \text{with } \sigma^2_{LL,m^\star,n} = \frac{\sigma^2_{\xi,m^\star,n}}{m^\star}, \quad \text{where } \sigma^2_{\xi,m^\star,n} = n\left(1 - \frac{m^\star}{n}\right) \sum_{i=1}^{n} d^2_{i,n}$$

can be unbiasedly estimated by

$$\widehat{\sigma}^2_{\xi,m^\star,n} = n^2 \left(1 - \frac{m^\star}{n}\right) \frac{1}{m^\star} \sum_{i=1}^{n} d^2_{i,n} u_i.$$

3.5. **Block proposals of u for subsampling.** Tran et al. (2016a) propose the block PM algorithm and show that it is a natural way to correlate the estimation errors in panel data and also in subsampling problems such as ours. The method divides the vector of observation indices $u = (u_1, \ldots, u_m)$ into G blocks and then update one block at a time jointly with θ. By setting a large G, a high correlation ρ between the estimates at the proposed and current parameter values is induced, reducing the variability of the ratio of estimates. More precisely, they show that under certain assumptions $\rho_G = 1 - 1/G$.

3.6. **Optimal variance of the estimator.** Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) obtain the value of $\sigma^2_{LL,m,n}$ that optimizes the trade off between MCMC sampling efficiency and computational cost in standard PM. The consensus is that $\sigma^2_{LL,m,n}$ should be between $[1, 3.283]$ where, in general, the less efficient the proposal in the exact likelihood setting, the higher the optimal value of $\sigma^2_{LL,m,n}$. The optimal value is derived assuming that the cost of computing one MCMC sample is inversely proportional to $\sigma^2_{LL,m,n}$, so that the so called Computational Time (CT) to produce one equivalent to an iid draw is

$$\text{CT}(\sigma^2_{LL,m,n}) \;\propto\; \text{IF}(\sigma^2_{LL,m,n}) \times \frac{1}{\sigma^2_{LL,m,n}}, \text{ with } \text{IF}(\sigma^2_{LL,m,n}) = 1 + 2\sum_{l=1}^{\infty} \rho_l, \tag{3.8}$$

where IF is the Inefficiency Factor and ρ_l is the l-lag auto-correlation of the chain. In our approach we have to select both m and K, the number of clusters. The computational cost of a new cluster comes from evaluating ℓ_i at the centroid, but also from evaluating the gradient and Hessian of ℓ_i. An approximate upper bound for the cost of a new cluster is therefore $3c_\ell$, where c_ℓ is the cost of a single ℓ_i-evaluation. In many models one can however re-use computations when computing the gradient and Hessian, so the true cost is probably much closer to $1c_\ell$. Assuming that the cost of a new cluster is ωc_ℓ, for some $\omega > 0$, a reasonable measure of computational time is

$$\mathrm{CT}_{(m,K)}(\sigma^2_{LL,m,n}(K)) = \mathrm{IF}(\sigma^2_{LL,m,n}(K)) \times (\omega K + m). \tag{3.9}$$

This expression is similar to Tran et al. (2016b) who also take into account an overhead cost in their CT. We find m and K by standard numerical optimization using an expression for the IF (e.g. the ones derived in Pitt et al., 2012 for PM and Tran et al., 2016a for block PM). It should be noted that the optimal value $\sigma^2_{LL,m,n} \approx 1$ is obtained if m is much larger than K (and ω is not too large) because then (3.8) and (3.9) are approximately equal (up to a proportionality constant).

Tran et al. (2016a) show that the conditional variance of the log-likelihood estimator (conditional on only updating one block of u, keeping the others fixed) is $\tau^2_{m,n,G} = \sigma^2_{LL,m,n}(1-\rho^2_G)$. Let $G = G(m) = O(m^\beta)$, then it follows that (using Lemma 2 and $\rho_G(m) = 1-1/G(m)$) $\tau^2_{m,n,G} = O(1)$ is achieved if we take $m = O(n^\gamma)$ with

$$\gamma \;=\; \frac{2(1-\alpha)}{1+\beta}, \quad \text{and } \alpha \text{ in } a_n = O(n^{-\alpha}) \quad \text{as in (2.3).}$$

Note that if $\beta = 0$, i.e. G is constant as a function of m so that $\rho \to 0$ as $m \to \infty$, then $\gamma = 2(1-\alpha)$, which corresponds to the uncorrelated algorithm. We emphasize that it is the interaction of the control variates and the correlated mechanism that makes the method scale well. For example, using $G = O(\sqrt{m})$, the optimal m is sublinear in n if $\alpha > 1/4$. However, note that reducing γ lower the rates of the asymptotic errors in Theorem 1. Tran et al. (2016a) also derive the optimal value of $\sigma^2_{LL,m,n}$ to target under the assumption of a CT as

in (3.8), but conditional on a value of ρ. This value of $\sigma^2_{LL,m,n}$ will typically be much larger than 1.

4. Applications

4.1. **Settings for PM algorithms.** Our algorithms use the control variates obtained by data clustering followed by a Taylor series expansion in data space. The tuning parameters m and K in the PM algorithms are determined by optimizing the computational time CT in (3.9)

$$\sigma^2_{LL,m,n}(K) = \frac{n^2\sigma^2_{d,n}(K)}{m} \quad \text{and IF}(\sigma^2_{LL,m,n}(K)),$$

with respect to m and K. We estimate the relation $\sigma^2_{d,n}(K) = C_0K^\nu$ by, for each example, running our cluster algorithm on a grid of K and for each value of the grid we compute $\sigma^2_{d,n}$ at the likelihood mode $\theta^\star$. Given C_0 and ν, it is straightforward to use the expression for the IF in Pitt et al. (2012) (PM) and Tran et al. (2016a) (block PM) to minimize $\text{CT}_{(m,K)}$ in (3.9) and obtain m_{opt} and K_{opt} and the corresponding $\sigma^2_{\text{opt}} = \sigma^2_{LL,m_{\text{opt}},n}(K_{\text{opt}})$. The correlated PM uses $m^\star_{\text{opt}} = m_{\text{opt}}$ and the same value of K_{opt} as block PM. Table 1 shows a summary of the settings for the applications. Finally, we let $G = 100$ ($\rho = 0.99$) and $\phi = 0.9999$ ($\kappa = 0.9863$).

4.2. **Logistic regression.** Our first example uses a logistic regression model for modeling bankruptcy conditional on a set of firm-specific covariates and macroeconomic variables, see Giordani et al. (2014) for details on the covariates. The data set has $n = 4,748,089$ and 8 covariates. The model is

$$p(y_i|x_i,\beta) \;=\; \left(\frac{1}{1+\exp(x_i^T\beta)}\right)^{y_i}\left(\frac{1}{1+\exp(-x_i^T\beta)}\right)^{1-y_i}, \text{ with } \quad p(\beta) = \mathcal{N}(\beta|0,10I).$$

Since the bankruptcy observations ($y_i = 1$) are sparse (41,566 defaults), we only subsample the observations with $y_k = 0$ observations, i.e. the first term in

$$\ell(\theta) \;=\; \sum_{\{i;y_i=1\}}\ell_i(\theta) + \sum_{\{i;y_i=0\}}\ell_i(\theta),$$

TABLE 1. *Experimental setup in the applications.* n is the number of observation. The proposals are the Random Walk Metropolis (RWM) $q(\theta|\theta_c) = \mathcal{N}(\theta|\theta_c, \Sigma_{\theta^\star})$ and the Independent MH (IMH) $q(\theta) = t_{10}(\theta|\theta^\star, \Sigma_{\theta^\star})$, where the location parameter is $\theta^\star$ is the posterior mode and $\Sigma_{\theta^\star}$ is the negative inverse Hessian of the log-posterior evaluated at $\theta^\star$, both obtained from an initial numerical optimization. We denote the optimal sample size and number of clusters by m_{opt} and K_{opt}, and $\sigma^2_{LL,\text{opt}}$ is the corresponding optimal variance of the log-likelihood estimate. We use $N = 50,000$ iterates after discarding 5,000 iterates as burn-in.

Example	n	**Proposal**	$100m_{\text{opt}}/n$	$100K_{\text{opt}}/n$	$\sigma^2_{LL,\text{opt}}$
Logistic	$4{,}7\times10^6$	RWM/IMH			
Uncorr			8.615	4.967	0.27
Block / Corr			1.286	0.485	56.89
AR(1): M_1	10^5	RWM			
Uncorr			1.896	2.464	0.11
Block / Corr			0.757	0.993	12.41
AR(1): M_2	10^5	RWM			
Uncorr			4.561	8.192	0.11
Block / Corr			2.151	3.176	12.40

is always evaluated (and included in the CC). Figure 1 shows the sampling efficiency of the PM algorithms relative to that of the MH algorithm as measured by the Relative Computational Time (RCT) defined, for any sampler $\mathcal{A}$, as $\text{CT}_{\text{MH}}/\text{CT}_{\mathcal{A}}$. The figure also shows the Relative IF (RIF) , which is defined as $\text{IF}_{\mathcal{A}}/\text{IF}_{\text{MH}}$, where each IF is estimated using the coda package in `R` (Plummer et al., 2006). The figure shows that both correlated and block PM significantly outperform standard PM and also MH with respect to RCT. Figure 2 plots the Kernel Density Estimates (KDE) on the output from the three pseudo-marginal schemes and the exact MH approach. The figure shows that targeting a large $\sigma^2_{LL,m,n}$ (≈ 56) for the block and correlated PM samplers result in a very small bias in this application, with the approximation error in (3.4) being -0.01 for both the block and correlated PM and -0.0001 for the standard PM. Figure 2 suggests that this small perturbation is, for the correlated and block PM estimators, mostly due to β_3.

Finally, we also tried the exact subsampling approach (Quiroz et al., 2016) by setting $\text{E}[G] = 100$ (G is random in their approach) and the batch-size to 602, so that their prior

expected sample size corresponds to that for our correlated algorithms. For the RWM we obtained an RCT of ≈ 17 ($\omega = 3$) and ≈ 20 ($\omega = 1$) (average over parameters), with ω in (3.9).

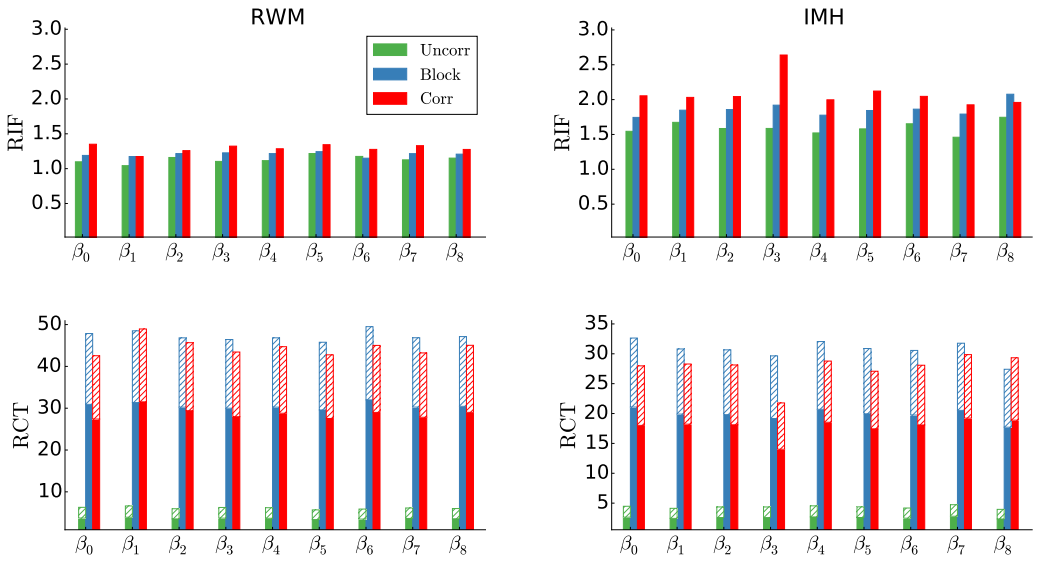

FIGURE 1. *Logistic regression example.* For algorithm $\mathcal{A}$ (uncorrelated (Uncorr), block (Block) and correlated (Corr) PM) the figure shows the Relative Inefficiency Factors (RIF) and Relative Computational Time for RWM proposal (left panel) and IMH (right panel). For RCT, the filled (dashed) bar correspond to $\omega = 3$ ($\omega = 1$) in (3.9).

4.3. **Other subsampling approaches.** We briefly discuss some other subsampling approaches and then benchmark them against our approach.

Korattikara et al. (2014) argue that using all data to take the simple decision to accept (or reject) a single parameter draw is a computationally inefficient strategy. Instead, they develop a sequence of t-tests, where each test is based on an increasing sample size and has a user specified error probability ϵ. The sequence is stopped when a decision of accepting (or rejecting) a single sample can be taken with a sufficiently small total error probability. They prove that the discrepancy between their approximate posterior and the true posterior can be made arbitrarily small by decreasing the total error probability of the test. However, it was empirically demonstrated in Bardenet et al. (2015) (see also Section 4.4) that the ϵ

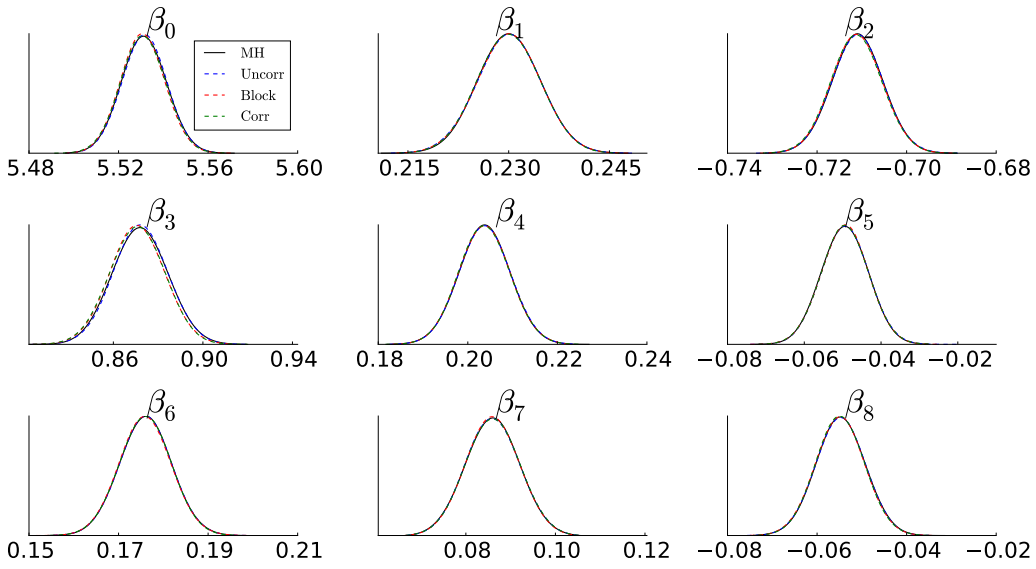

FIGURE 2. *Logistic regression example.* Kernel density estimates of marginal posteriors obtained by the IMH proposal. The figure shows the marginal posteriors obtained using the uncorrelated (Uncorr), block (Block) and correlated (Corr) PM (dashed blue, red and green, respectively) and MH (solid black line).

parameter must be set so low that virtually the full data set is required to obtain a reasonable approximation error.

Bardenet et al. (2014) also rely on the idea of replacing the computation of the MH ratio with a hypothesis test. However, they use an exact (not relying on a CLT) confidence interval obtained through a concentration inequality, which allows the choice of a subsample large enough so that a decision can be taken with a user specified error probability. They prove that the posterior targeted by their algorithm can become arbitrarily close to the true posterior as the error probability decreases. To successfully implement the method, the range of the differences in log-likelihood contributions at the proposed and current samples must be cheap to compute. Moreover, the performance of the method is highly dependent on the variance of these differences.

Bardenet et al. (2015) improve on the sampler in Bardenet et al. (2014) by introducing control variates to obtain variance reduction of the differences. Furthermore, the method does not rely on a cheaply computed range, however, it does rely on a bound for the difference

between the log-likelihood contributions at the proposed and current sample, and that of the control variates. When the control variates are obtained via Taylor series approximations, they suggest using the Taylor-Lagrange inequality to obtain the bound. We show in Section 4.4 that the Taylor-Lagrange bound can sometimes be too crude, and it is then necessary to compensate with a very large subsample. Bardenet et al. (2015) show that their method dramatically outperforms, among others, Korattikara et al. (2014), Bardenet et al. (2014) and Firefly Monte Carlo (Maclaurin and Adams, 2014, see below).

Firefly Monte Carlo in (Maclaurin and Adams, 2014, and more recently, Liu et al., 2015) introduces an auxiliary variable for each observation which determines if it should be included in the evaluation of the posterior. The distribution of these variables are such that when they are integrated out, the marginal posterior is the true posterior of θ. Moreover, a lower bound for each likelihood term is introduced, which basically plays the role of replacing the observations that are not included in the evaluation of the posterior. The authors suggest using the Gibbs sampler, generating the parameters conditional on the auxiliary variables and vice versa. The method has been documented to be very inefficient, see e.g. Bardenet et al. (2015), see also Section 4.4.

4.4. **AR processes.** The running examples in Bardenet et al. (2015) use the normal model which is too simplistic for our method because the log-likelihood is quadratic in the data and therefore our control variates are perfect. We instead compare our method to alternative approaches using the following illustrative models. We consider the following two AR(1) models with Student-t iid errors $\epsilon_t \sim t(5)$ with 5 degrees of freedom

$$\mathrm{M}_1 : y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t \qquad [\theta = (\beta_0 = 0.3, \beta_1 = 0.6)]$$

$$\mathrm{M}_2 : y_t = \mu + \varrho(y_{t-1} - \mu) + \epsilon_t \quad [\theta = (\mu = 0.3, \varrho = 0.99)]$$

with priors

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

where $\mathcal{U}(\cdot|a,b)$ is the uniform density on the interval $[a,b]$. Model M_2, the so called steady state AR, is particularly interesting as ϱ close to 1 gives a weakly identified μ, with a posterior that concentrates very slowly as n increases.

We compare our method to the Austerity MH (Korattikara et al., 2014), Firefly Monte Carlo (Maclaurin and Adams, 2014), the confidence sampler (Bardenet et al., 2014) and the confidence sampler with proxies (Bardenet et al., 2015). We set the tuning parameters of the competing algorithms following Bardenet et al. (2015) with the following exceptions. First, we adapt during the burn-in phase to reach an acceptance probability of $\alpha = 0.35$ (instead of $\alpha = 0.50$), which is optimal for RWM with two parameters (Gelman et al., 1996). For the pseudo-marginals we instead use $\alpha = 0.15$ as the five parameter example in Sherlock et al. (2015). Second, the p-value of the t-test in the Austerity MH algorithm is set to $\epsilon = 0.01$ (instead of $\epsilon = 0.05$) to put the approximation error of the method on par with the other methods. Setting $\epsilon = 0.05$ gives an unusably poor approximation (and also produces a much lower RCT than our methods). Additionally, the confidence sampler with proxies (from a Taylor series approximation with respect to θ) requires that the third derivative can be bounded uniformly for every observation and any θ. This bound is achieved, without any extra computational cost, by computing on a θ-grid where the posterior mass is located.

TABLE 2. *AR-process example.* Mean of sampling fraction $f = m/n$ over MCMC iterations for models M_1 and M_2 with MH, uncorrelated PM (Uncorr), block PM (Block) and correlated PM (Corr), confidence sampler (Conf), confidence sampler with proxies (ConfProxy), Austerity MH (AustMH), and Firefly Monte Carlo (Firefly).

	MH	Uncorr	Block	Corr	Conf	ConfProxy	AustMH	Firefly
M_1	1.000	0.093	0.037	0.037	1.493	0.160	1.037	0.100
M_2	1.000	0.291	0.117	0.116	1.490	1.500	1.019	0.137

Table 2 shows the mean of the sampling fraction over MCMC iterations. We note that both confidence samplers and the Austerity MH estimate the numerator and denominator in each iteration, and therefore require twice as many evaluations in a given iteration as MCMC (in some cases evaluations from the previous iteration can be reused). It is clear

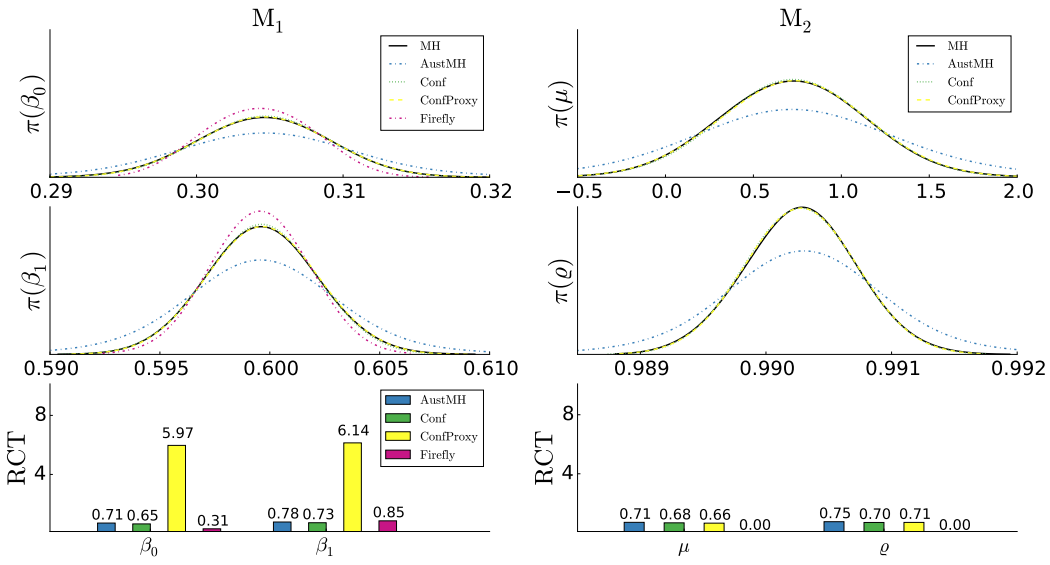

FIGURE 3. *AR-process example: Results for other subsampling algorithms.* The left and right panel, respectively, show the results for model M_1 and M_2. Each column shows the kernel density estimates of marginal posteriors (top two) and for algorithm $\mathcal{A}$ (confidence sampler (Conf), confidence sampler with proxies (ConfProxy), Austerity MH (AustMH), and Firefly Monte Carlo (Firefly)) the Relative Computational Time (RCT) (bottom).

that our algorithms makes very efficient use of a small subsample, especially the block and correlated PM samplers.

Figure 3 and 4 show the marginal posteriors obtained by, respectively, alternative sampling approaches and the several PM approaches. Moreover, the figures show the sampling efficiency of the different subsampling MCMC algorithms relative to that of the MH algorithm as measured by the Relative Computational Time. Figure 3 shows the striking result that many of these approaches are not more efficient than MH, except the confidence sampler with proxies for M_1. Regarding the approximation, it is evident that the Austerity MH still has a tempering effect (larger spread on the posterior) although ϵ is set so low that the full data set is sampled (see Table 2). The PM algorithms (and also the confidence samplers) provide excellent approximations: indeed, $\text{error}(\theta) \leq 10^{-6}$ in (3.4) for all our methods. Firefly Monte Carlo, although being an exact algorithm, is highly inefficient in this example, as also documented in Bardenet et al. (2015). In fact, for M_2, we were not able to obtain a

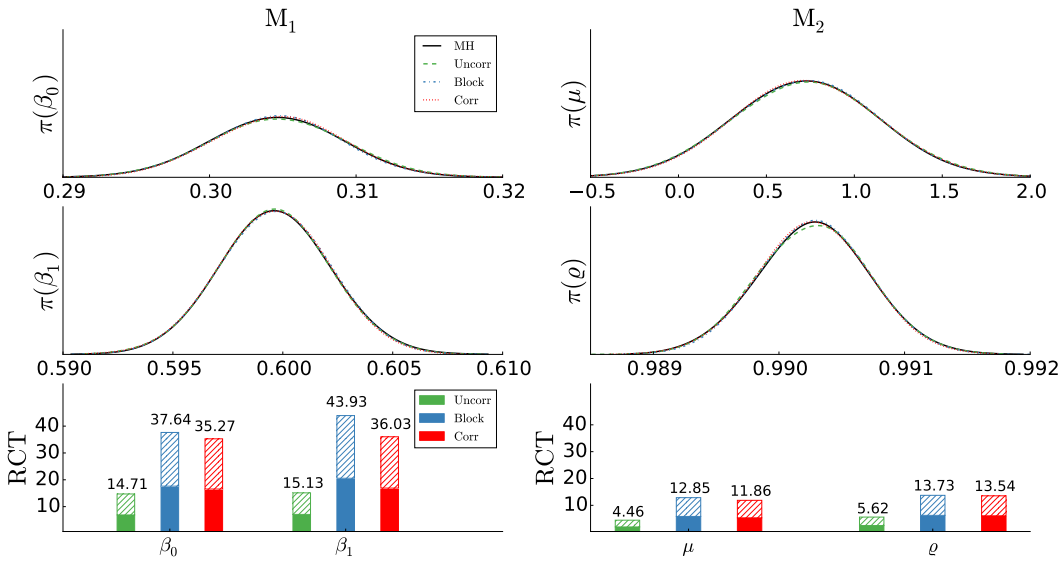

FIGURE 4. *AR-process example: Results for subsampling PM algorithms.* The left and right panel, respectively, show the results for model M_1 and M_2. Each column shows the kernel density estimates of marginal posteriors (top two) and for algorithm $\mathcal{A}$ (uncorrelated (Uncorr), block (Block) and correlated (Corr) PM) the Relative Computational Time (RCT) (bottom). For RCT, the filled (dashed) bar correspond to $\omega = 3$ ($\omega = 1$) in (3.9).

single effective sample out of $55,000$ iterations, and hence it was impossible to construct a kernel density estimate in this case. We also tried the exact subsampling in Quiroz et al. (2016), setting the tuning parameters to match the sample size used here as described in Section 4.2. For M_1 the sampler got stuck because the variance of the log of the estimator was too large (≈ 2070, compared to ≈ 12 for the estimator used here). For M_2 the exact subsampling produced an RCT of ≈ 4 ($\omega = 3$) and ≈ 7 ($\omega = 1$) (average over parameters), with ω in (3.9).

We conclude that the only viable subsampling MCMC approaches are the confidence sampler with proxies (Bardenet et al., 2015) and the PM approaches we propose. Moreover, a significant speed up is only obtained with the correlated PMs (both correlated and block). We acknowledge that we have put the confidence sampler with proxies in an unfavorable situation in M_2: the bound of its concentration inequality requires a bound of the remainder term in the Taylor series via the Taylor-Lagrange inequality, which is very hard for M_2.

We remark that the Taylor proxies with respect to θ work well in these two-dimensional examples, however, problems are encountered in the logistic example in Section 4.2; the expansion with respect to the data as proposed in Section 2.1 works much better. Moreover, an additional feature of expanding with respect to data is that an IMH proposal can be implemented straightforwardly, as the control variates are accurate for any θ.

5. Conclusions and Future Research

We propose a framework for speeding up MCMC for data sets with many observations by data subsampling. The following features are key to our approach. First, we introduce a highly efficient log-likelihood estimator, which incorporates information about each observation's contribution to the log-likelihood function, while only operating on a sparse set of the data. This results in a substantially lower computational cost. Second, we use the resulting likelihood estimate within a pseudo-marginal framework and sample from a perturbed posterior which, for fixed n, we prove to be within $O(m^{-2})$ of the true posterior. We also consider the asymptotic behavior with respect to n. Moreover, we provide a useful heuristic to approximate the error, given that m is large. Third, we propose a correlated pseudo-marginal approach to subsampling, which allows highly variable estimates of the likelihood without adversely affecting the mixing of the algorithm. The resulting algorithm is a highly efficient algorithm when taking into account the statistical efficiency and computational cost. Fourth, we use the correlated and block PM samplers to show that we can let the sample size m grow much more slowly as a function of n than the standard PM sampler to achieve the same inefficiency, and hence a much lower computational time. Finally, we document large speed ups relative to MH and, more importantly, we show that our method outperforms other recent subsampling approaches in the literature.

Future research concerns designing efficient proposals based on data subsampling, e.g. in hybrid Monte Carlo algorithms. It is also of interest to develop improved clustering methods to obtain control variates, especially in the presence of a large number of covariates.

6. Acknowledgments

Matias Quiroz was partially supported by VINNOVA grant 2010-02635. Robert Kohn's research was partially supported by Australian Research Council Center of Excellence grant CE140100049. Minh-Ngoc Tran was partially supported by a Business School Pilot Research grant. We thank the authors in Bardenet et al. (2015) for making their code publicly available, which facilitated the comparison against other subsampling approaches.

References

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

Bardenet, R., Doucet, A., and Holmes, C. (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *Proceedings of The 31st International Conference on Machine Learning*, pages 405–413.

Bardenet, R., Doucet, A., and Holmes, C. (2015). On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160.

Ceperley, D. and Dewing, M. (1999). The penalty method for random walks with uncertain energies. *The Journal of chemical physics*, 110(20):9812–9820.

Dahlin, J., Lindsten, F., Kronander, J., and Schön, T. B. (2015). Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. *arXiv preprint arXiv:1511.05483*.

Dean, J. and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2016). The correlated pseudo-marginal method. *arXiv preprint arXiv:1511.04992v3*.

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

asu075.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. *Bayesian statistics*, 5(599-608):42.

Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2014). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, 49(4):1071–1099.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Jacob, P. E. and Thiery, A. H. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784.

Korattikara, A., Chen, Y., and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 181–189.

Liu, S., Mingas, G., and Bouganis, C.-S. (2015). An exact MCMC accelerator under custom precision regimes. In *Field Programmable Technology (FPT), 2015 International Conference on*, pages 120–127. IEEE.

Lyne, A.-M., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467.

Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664.

Neiswanger, W., Wang, C., and Xing, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.

Nemeth, C. and Sherlock, C. (2016). Merging MCMC subposteriors through Gaussian-process approximations. *arXiv preprint arXiv:1605.08576*.

Nicholls, G. K., Fox, C., and Watt, A. M. (2012). Coupled MCMC with a randomized acceptance probability. *arXiv preprint arXiv:1205.6857*.

Pitt, M. K., Silva, R. d. 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.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Quiroz, M., Villani, M., and Kohn, R. (2016). Exact subsampling MCMC. *arXiv preprint arXiv:1603.08232*.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E., and McCulloch, R. (2013). Bayes and big data: the consensus Monte Carlo algorithm. In *EFaBBayes 250" conference*, volume 16.

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

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016a). Block-wise pseudo-marginal Metropolis-Hastings. *arXiv preprint arXiv:1603.02485v2*.

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2016b). Importance sampling squared for Bayesian inference in latent variable models. *arXiv preprint arXiv:1309.3339v4*.

Wang, X. and Dunson, D. B. (2013). Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.