Stockholms universitet

Research Report

Department of Statistics

No. 2015:4

Speeding up MCMC by efficient data subsampling

Matias Quiroz
Mattias Villani
Robert Kohn

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

SPEEDING UP MCMC BY EFFICIENT DATA SUBSAMPLING

MATIAS QUIROZ, MATTIAS VILLANI AND ROBERT KOHN

ABSTRACT. The computing time for Markov Chain Monte Carlo (MCMC) algorithms can be prohibitively large for datasets with many observations, especially when the data density for each observation is costly to evaluate. We propose a framework where the likelihood function is estimated from a random subset of the data, resulting in substantially fewer density evaluations. The data subsets are selected using an efficient Probability Proportional-to-Size (PPS) sampling scheme, where the inclusion probability of an observation is proportional to an approximation of its contribution to the log-likelihood function. Three broad classes of approximations are presented. The proposed algorithm is shown to sample from a distribution that is within $O(m^{-\frac{1}{2}})$ of the true posterior, where m is the subsample size. Moreover, the constant in the $O(m^{-\frac{1}{2}})$ error bound of the likelihood is shown to be small and the approximation error is demonstrated to be negligible even for a small m in our applications. We propose a simple way to adaptively choose the sample size m during the MCMC to optimize sampling efficiency for a fixed computational budget. The method is applied to a bivariate probit model on a data set with half a million observations, and on a Weibull regression model with random effects for discrete-time survival data.

Supplementary material is available online.

KEYWORDS: Bayesian inference, Markov Chain Monte Carlo, Pseudo-marginal MCMC, Big Data, Probability Proportional-to-Size sampling, Numerical integration.

Quiroz: *Research Division, Sveriges Riksbank, SE-103 37 Stockholm, Sweden* and *Department of Statistics, Stockholm University. E-mail: quiroz.matias@gmail.com.* Villani: *Division of Statistics and Machine Learning, Department of Computer and Information Science, Linkoping University.* E-mail: mattias.villani@liu.se. Kohn: *Australian School of Business, University of New South Wales, UNSW, Sydney 2052.* 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. We would like to thank Eduardo Mendes for a helpful discussion on the proofs in the Appendix. 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.

1. Introduction

Markov Chain Monte Carlo (MCMC) methods have been the workhorse for sampling from the posterior distribution in Bayesian models since their introduction in statistics in the early 90's by Gelfand and Smith (1990). MCMC methods are desirable as they return the full posterior distribution without approximations. Furthermore, model diagnostics, model selection, and model regularization through variable selection, etc. may be obtained within a single run of the sampler. Models with costly likelihood evaluations pose a great challenge for MCMC algorithms, however, because of their demanding computing time. This is especially true for large datasets which are becoming increasingly more common in practical work. Because of this caveat, researchers tend to approach such problems using approximate methods such as Approximate Bayesian Computations (ABC) (Marin et al., 2012), Variational Bayes (VB) (Ormerod and Wand, 2010) or optimization techniques, e.g. stochastic approximation (Spall, 2005). However, the great drawback of these three approaches is that there are currently no theoretical results to assess the errors in the approximate posterior produced, nor even tell empirically if these errors are too large to be acceptable. If MCMC is the choice for inference in computationally demanding models, it is often via data augmentation through the Gibbs sampler, an approach which is well known to increase the inefficiency of the sampler, see e.g. Liu et al. (1994).

Our article presents a Metropolis-Hastings framework where the likelihood is estimated from a small random subset of the data. The data subsets are obtained by highly efficient probability proportional-to-size (PPS) sampling schemes. We show that the sampling design is absolutely crucial for the success of data subsampling in an MCMC context, which explains the poorly mixing MCMC chain based on the simple random sampling (SI) design found by Korattikara et al. (2013). PPS is many orders of magnitude more efficient than SI, leading to an MCMC chain with many more efficient draws for a given time budget compared to a regular MCMC on the full dataset, especially when the proposal is poor.

Beaumont (2003) introduces a Metropolis-Hastings algorithm based on an unbiased estimator of the likelihood. Andrieu and Roberts (2009) and Andrieu et al. (2010) prove

that such *Pseudo-marginal MCMC* (PMCMC) algorithms do indeed sample from the target posterior if the likelihood estimator is unbiased, regardless of the estimator variance. We note here that PMCMC is still valid even when the likelihood estimator is biased, but then samples from a perturbed target posterior. We exploit this and propose efficient estimators of the log-likelihood which are subsequently corrected to achieve near-unbiasedness for the likelihood. We prove that the distribution targeted by our sampling scheme is within $O(m^{-\frac{1}{2}})$ of the true posterior, where m is the subsample size. Moreover, it is also shown that the constant of proportionality in the $O(m^{-\frac{1}{2}})$ bound of the likelihood approximation is relatively small and that highly accurate results can therefore be obtained already for very small m. This accuracy is also verified in our two applications where the approximation errors are shown to be negligible. Focusing the estimation effort on the log-likelihood has several advantages. First, it makes it possible to adopt well-studied methods for estimating a finite population total (a sum) from the survey sampling field. Second, it allows us to use the simple rules in Pitt et al. (2012) and Doucet et al. (2015) to choose the optimal subsample size. Third, it enables us to choose m adaptively in each MCMC iteration which helps to control the variance of the log-likelihood estimator and therefore improve MCMC efficiency for a given computational budget.

The paper is organized as follows. Section 2 introduces our MCMC algorithm for sampling from the posterior distribution using random subsets of the data and proves some important properties of the method. Section 3 proposes a framework for estimating the likelihood by PPS sampling of data subsets with efficient sampling inclusion probabilities based on approximations of the data distribution. Section 4 evaluates the performance of the proposed methodology in two empirical applications. Section 5 concludes and discusses further research. Appendix A contains some implementation details. The proof of the main theorem is available in the supplementary material.

2. MCMC WITH SAMPLING-BASED LIKELIHOOD ESTIMATORS

2.1. **Theory.** Let $y = (y_1, \ldots, y_n)^T$ denote the vector of observed data and let θ be the vector of parameters. Potential dependence on covariates is suppressed in this section for notational clarity. Denote the likelihood by $p(y|\theta)$, and let $p(\theta)$ and $\pi(\theta) = p(\theta|y)$ denote the prior and the posterior for θ, respectively. Let u be a vector of auxiliary variables corresponding to the subset of observations to include when estimating $p(y|\theta)$. In our framework, the distribution of u is dependent on the data and the parameter, i.e. $p(u|\theta, y)$. Let $\hat{p}_m(y|\theta, u)$, for a fixed m, be a possibly biased estimator of $p(y|\theta)$ with expectation

$$p_m(y|\theta) = \int \hat{p}_m(y|\theta, u)p(u|\theta, y)du. \tag{2.1}$$

Define

$$\tilde{\pi}_m(\theta, u) = \hat{p}_m(y|\theta, u)p(u|\theta, y)p(\theta)/p_m(y), \text{ with } p_m(y) = \int p_m(y|\theta)p(\theta)d\theta, \tag{2.2}$$

on the augmented space (θ, u). The marginal density for θ given by

$$\pi_m(\theta) = \int \tilde{\pi}_m(\theta, u)du = \frac{p_m(y|\theta)p(\theta)}{p_m(y)}.$$

Informally, we note that for $m = \infty$,

$$p_m(y|\theta) = p(y|\theta), \quad p_m(y) = p(y) \quad \text{and } \pi_m(\theta) = \pi(\theta).$$

Draws from the joint posterior $\tilde{\pi}_m(\theta, u)$ are obtained by the M-H algorithm as follows. Move the Markov chain from the current state (θ_c, u_c) to (θ_p, u_p) by proposing $\theta_p \sim q(\theta|\theta_c)$ and $u_p \sim p(u|\theta_p, y)$ and accept with probability

$$\alpha = \min\left(1, \frac{\hat{p}_m(y|\theta_p, u_p)p(\theta_p)/q(\theta_p|\theta_c)}{\hat{p}_m(y|\theta_c, u_c)p(\theta_c)/q(\theta_c|\theta_p)}\right). \tag{2.3}$$

If the draw is rejected then $(\theta_p, u_p) = (\theta_c, u_c)$. We note that $u \sim p(u|\theta, y)$ is an auxiliary variable that arises when estimating $p(y|\theta)$. We can therefore regard u as being proposed from $p(u|\theta, y)$, i.e. $q(u|\theta, y) = p(u|\theta, y)$, so that it always cancel in the M-H ratio in Equation (2.3).

Moreover, this choice of q implies that $p(u|\theta, y)$ can be intractable as it is never evaluated, we only need to simulate from it.

By the results in Andrieu and Roberts (2009), the draws of θ obtained by this M-H algorithm has $\pi_m(\theta)$ as invariant distribution regardless of the variance of the likelihood estimator; the efficiency of the scheme depends crucially on this variance however, as discussed in Section 3.1. It is clear that if $\hat{p}_m(y|\theta, u)$ is an unbiased estimator of $p(y|\theta)$, then the marginal of the augmented MCMC scheme above has $\pi(\theta)$ as invariant distribution, but it is important to realize that it is still a valid sampler also when $\hat{p}_m(y|\theta, u)$ is biased, but then targets the distribution $\pi_m(\theta)$.

Let $\{y_k, x_k\}_{k=1}^n$ denote the data, where y is a potentially multivariate response vector and x is a vector of covariates. Given conditionally independent observations we have the usual decomposition of the log-likelihood

$$l(\theta) = \sum_{k=1}^{n} l_k(\theta), \tag{2.4}$$

where $l_k(\theta) = \log p(y_k|\theta, x_k)$ is the log-likelihood contribution of the kth observation. The log-likelihood is a sum and estimation of the log-likelihood is therefore equivalent to the classical survey sampling problem of estimating a finite population total from a subsample of the data observations (see Särndal et al. (2003) for an introduction). Note that the same is true for any problem where the log-likelihood decomposes as a sum of terms where each term depends on a unique piece of data information. The most obvious example are longitudinal problems where $l_k(\theta)$ is the log joint density of all measurements on the kth subject, and we sample subjects rather than individual observations. Similarly, it also applies to certain time-series problems with Markov dependence such as autoregressive processes. We consider estimators of the likelihood from a data subsample of size m of the following form

$$\hat{p}_m(y|\theta, u) = \exp\left(\hat{l}_m(\theta) - \hat{\sigma}_z^2(\theta)/2\right) \tag{2.5}$$

where $\hat{l}_m(\theta)$ is the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) of the log-likelihood

$$\hat{l}_m = \frac{1}{m}\sum_{i=1}^{m}\zeta_i, \text{ where } \zeta_i = \frac{l_{u_i}}{p_{u_i}}, \tag{2.6}$$

u_i are sampling indicator variables such that $u_i = k$ means that observation k was selected at the ith draw with replacement, with probability

$$p_k(\theta) = \Pr(u_i = k) = \Pr(\zeta_i = l_k(\theta)/p_k(\theta)).$$

It is easy to see that $\hat{l}_m(\theta)$ is unbiased for $l(\theta)$. Also, $z = \hat{l}_m(\theta) - l(\theta)$ denotes the estimation error, $\sigma_z^2 = Var(z)$ and

$$\hat{\sigma}_z^2 = \frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\frac{l_{u_i}(\theta)}{p_{u_i}} - \hat{l}_m(\theta)\right)^2 \tag{2.7}$$

is an unbiased estimator of σ_z^2. The selection probabilities $p_k(\theta)$ are determined by the sampling design and are discussed later. To motivate the form of $\hat{p}_m(y|\theta, u)$ in Equation (2.5), consider the case when $z \sim N(0, \sigma_z^2)$ and σ_z^2 is known. It is then easy to see that $\exp\left(\hat{l}_m(\theta) - \sigma_z^2(\theta)/2\right)$ is an unbiased estimator of the likelihood and the MCMC scheme above will sample from the target posterior $\pi(\theta)$. When z is not Gaussian and σ_z^2 is unknown, Theorem 1 below proves that the distribution targeted by the MCMC algorithm is within $O(m^{-\frac{1}{2}})$ of the true posterior, as are the posterior moments.

Theorem 1 is based on Assumption 1. Define

$$\gamma_\alpha(\theta) \;=\; \frac{E\left[(\zeta_1(\theta) - l)^\alpha\right]}{\left(\sigma_\zeta^2\right)^{\alpha/2}}, \quad \alpha \in \mathbb{N}$$

where $\sigma_\zeta^2 = V(\zeta_i) = \sum_{k=1}^{n}\left(\frac{l_k}{p_k} - l\right)^2 p_k$ so that $\sigma_z^2 = \sigma_\zeta^2/m$. We often omit dependence on θ to simplify notation.

Assumption 1. *We assume that the following results hold uniformly for θ where λ is a constant.*

i. *There exists a $K > 0$ such that $\gamma_4(\theta) \leq K$.*

ii.
$$E\left[\exp\left(\lambda\frac{\zeta_1 - l}{m}\right)\right] = 1 + \frac{\lambda^2}{2m}\sigma_z^2 + \frac{\lambda^3}{6m^{3/2}}\gamma_3\left(\sigma_z^2\right)^{3/2} + o(m^{-3/2}).$$

iii.
$$E\left[\exp\left(\lambda\frac{(\hat{l}_m - l)^2 - \sigma_z^2}{m-1}\right)\right] = 1 + \frac{\lambda^2\left(\sigma_z^2\right)^2}{2m(m-1)^2}(\gamma_4 + 2m - 3) + o(m^{-2}).$$

iv.
$$E\left[\exp\left(\lambda\frac{(\zeta_1 - l)^2 - \sigma_\zeta^2}{m(m-1)}\right)\right] = 1 + \frac{\lambda^2\left(\sigma_z^2\right)^2}{2(m-1)^2}(\gamma_4 - 1) + o(m^{-2}).$$

v. *The prior density $p(\theta)$ is proper.*

Remark. Part (ii) and (iv) in Assumption 1 can be easily verified in a given example by directly computing

$$E\left[g(\zeta)\right] = \sum_{k=1}^{n} g(l_k/p_k)p_k.$$

Part (iii) can be verified by simulation.

We now motivate part (ii)-(iv) of Assumption 1.

Part (ii): By a third order Taylor series expansion with remainder,

$$E\left[\exp\left(\lambda\frac{\zeta_1 - l}{m}\right)\right] = 1 + \frac{\lambda^2}{2m}\sigma_z^2 + \frac{\lambda^3}{6m^{3/2}}\gamma_3\left(\sigma_z^2\right)^{3/2} + \frac{\lambda^4}{24m^2}E\left[\left(\frac{\zeta_1 - l}{\sigma_\zeta}\right)^4\exp\left(\lambda\frac{\zeta_1'}{m}\right)\right]$$

where $|\lambda\zeta_1'| \leq |\lambda(\zeta_1 - l)|$. The assumption means that

$$\frac{1}{\sqrt{m}}E\left[\left(\frac{\zeta_1 - l}{\sigma_\zeta}\right)^4\exp\left(\lambda\frac{\zeta_1'}{m}\right)\right] \quad \text{is } o(1).$$

Part (iii): Define $\nu_m = (\hat{l}_m - l)^2 - \sigma_z^2$. We can show that

$$E[\nu_m] = 0 \quad \text{and} \quad V[\nu_m] = \frac{\left(\sigma_z^2\right)^2}{m}(\gamma_4 + 2m - 3).$$

Hence, by a second order Taylor series with remainder,

$$E\left[\exp\left(\lambda\frac{\nu_m}{m-1}\right)\right] = 1 + \frac{\lambda^2 V[\nu_m]}{2(m-1)^2} + \frac{\lambda^3}{6(m-1)^3}E\left[\nu_m^3\exp\left(\lambda\frac{v_m'}{m-1}\right)\right]$$

where $|\lambda v'_m| \leq |\lambda v_m|$. The assumption means that

$$\frac{1}{m-1} E\left[\nu_m^3 \exp\left(\lambda \frac{v'_m}{m-1}\right)\right] \quad \text{is } o(1).$$

Part (iv): Define $\psi_1 = \left((\zeta_1 - l)^2 - \sigma_\zeta^2\right)/m$. We can check that

$$E[\psi_1] = 0 \quad \text{and} \quad V[\psi_1] = \left(\sigma_z^2\right)^2 (\gamma_4 - 1).$$

Therefore, a second order Taylor series with remainder,

$$E\left[\exp\left(\lambda \frac{\psi_1}{m-1}\right)\right] = 1 + \frac{\lambda^2 V[\psi_1]}{2(m-1)^2} + \frac{\lambda^3}{6(m-1)^3} E\left[\psi_1^3 \exp\left(\lambda \frac{\psi'_1}{m-1}\right)\right]$$

where $|\lambda \psi'_1| \leq |\lambda \psi_1|$. The assumption states that

$$\frac{1}{m-1} E\left[\psi_1^3 \exp\left(\lambda \frac{\psi'_1}{m-1}\right)\right] \quad \text{is } o(1).$$

Theorem 1. *The following results hold subject to Assumption 1:*

i. *For any θ,*

$$\frac{|p_m(y|\theta) - p(y|\theta)|}{p(y|\theta)} \leq \frac{1}{\sqrt{m}} C_m(\theta)$$

where $C_m > 0$ is bounded.

ii.

$$\frac{|p_m(y) - p(y)|}{p(y)} \leq \frac{1}{\sqrt{m}} \sup_\theta C_m(\theta)$$

with C_m as above.

iii. *For any θ,*

$$\frac{|\pi_m(\theta) - \pi(\theta)|}{\pi(\theta)} \leq \frac{1}{\sqrt{m}} D_m(\theta)$$

where D_m is bounded and

$$D_m(\theta) = \tfrac{p(y)}{p_m(y)} \left(C_m(\theta) + \sup_\theta C_m(\theta)\right) \quad ,$$

with C_m as above.

iv. *Suppose that* $h(\theta)$ *is a function such that* $\int |h(\theta)|p(\theta)d\theta < \infty$. *Then*

$$\left| \int h(\theta)\pi_m(\theta)d\theta - \int h(\theta)\pi(\theta)d\theta \right| \leq \frac{1}{\sqrt{m}}\tilde{D} \int |h(\theta)|p(\theta)d\theta$$

where $\tilde{D} = \sup_\theta D_m(\theta)$ *with* D_m *as above. In particular*

$$\int |\pi_m(\theta) - \pi(\theta)|\, d\theta \leq \frac{1}{\sqrt{m}}\tilde{D}.$$

The proof of Theorem 1 and also the expression for C_m is given in the supplementary material. The constant C_m is illustrated to be relatively small (in comparison to $\sqrt{m}$) in our applications in Section 4. This results in highly accurate posterior approximations, even for relatively small values of m.

2.2. **The trade-off between efficiency and computing time.** Our PMCMC algorithm is faster but less efficient per iteration than a regular MCMC chain using all data observations because it uses a noisy likelihood based on a subsample. Increasing the noise results in reduced computing time per iteration, but reduces the efficiency of the Markov Chain, and vice versa. A measure that balances computing time and efficiency is *Efficient Draws Per Minute* (EDPM)

$$EDPM \quad = \quad \frac{N}{IF \times t}$$

where N = number of iterations, t = execution time and IF is the inefficiency factor, estimated by

$$IF \quad = \quad 1 + 2\sum_{l=1}^{L} \hat{\rho}_l,$$

where $\hat{\rho}_l$ is the sample autocorrelation at the lth lag of the chain and L is an upper limit such that $\rho_l \approx 0$ when $l > L$. The inefficiency factor, also called the integrated autocorrelation time, measures the number of MCMC draws that are equivalent to a single draw obtained using a sampler that produces independent draws. IF-values near 1 therefore suggest a very efficient algorithm.

In Section 4 we measure the performance of PMCMC vs MCMC by the Relative Efficient Draws Per Minute (REDPM) which is defined as

$$REDPM \quad = \quad \frac{EDPM^{PMCMC}}{EDPM^{MCMC}}.$$

2.3. **Choosing the optimal sampling fraction.** Pitt et al. (2012) and Doucet et al. (2015) define the computational time of the PMCMC as

$$CT(\sigma_z^2) \quad = \frac{IF^{PMCMC}(\sigma_z^2)}{\sigma_z^2} \quad , \tag{2.8}$$

i.e. the computing time is assumed to be inversely proportional to the variance of the estimator. To find an optimal value of σ_z^2, Pitt et al. (2012) assume that the proposal for θ is perfect, i.e. it is the target posterior distribution. They find that σ_z^2 around 1 is the optimal value, with a fairly benign region around 1. Doucet et al. (2015) allow for general proposals and obtain bounds for the optimal σ_z^2. Assuming that z is normal and that σ_z^2 is independent of θ, they show that the optimal value of σ_z^2 lies in the interval $[1.0, 3.2]$. In general, the less efficient the proposal in the exact likelihood setting, the higher the optimal value of σ_z^2. Sherlock et al. (2015) show (under different assumptions) that $V[\hat{l}(\theta)] = 3.283$ is optimal. Our view is that it is prudent to take $\sigma_z^2 \approx 1$, which is the conservative choice, and avoid the risk of obtaining a value of σ_z^2 that is catastrophically high.

Assuming that m is large enough our estimator is approximately unbiased. Furthermore, for each θ we can tune the sample size m to achieve a constant variance (independently of θ, see Section 2.4). We can therefore conveniently tune the subsample size following the simple rule of choosing m such that $\sigma_z^2 \approx 1$.

2.4. **Adaptive sampling fraction.** It is possible to adapt the sampling fraction $f = m/n$ in a given MCMC iteration if the variance is too large. The adaptation adds observations to reduce the variance of the log-likelihood estimator in Equation (2.6) at a given iteration. We now propose a fast and simple simulation of u such that $\hat{V}[\hat{l}(\theta)] < v_{max}$ where v_{max} is the user specified maximum variance tolerated in the log-likelihood estimate. In any given iteration, while $\hat{V}[\hat{l}(\theta)] > v_{max}$, simply increase m. Equation (2.7) can be iterated to compute

an accurate guess of the new sample size needed to bring the variance of the estimator down to v_{max}:

$$m^* = \frac{1}{v_{max}(m-1)} \sum_{i=1}^{m} \left(\frac{l_{u_i}(\theta)}{p_{u_i}} - \hat{l}(\theta) \right)^2. \tag{2.9}$$

3. Probability proportional-to-size sampling (PPS)

This section discusses the importance of efficient subsampling of data subsets in PMCMC and proposes a particularly attractive sampling scheme based on probability proportional-to-size sampling of the data.

3.1. Sampling variability and PMCMC efficiency.

PMCMC using the estimator in Equation (2.5) is theoretically guaranteed to converge to the posterior $\pi_m(\theta)$, but the efficiency of the sampling scheme depends crucially on the variance of the estimator. A large estimator variance can easily produce extreme over-estimates of the likelihood and cause the PMCMC chain to get stuck for long spells.

It is therefore crucial to use a combination of a sampling design and estimator that keeps the variance of the log-likelihood estimator around one (Pitt et al., 2012; Doucet et al., 2015) also for small subsample sizes m. As an example of design-estimator pairs that produces too much variability in the estimated likelihood, we consider simple random sampling without replacement (SI) and the usual unbiased estimator of the log-likelihood

$$\hat{l}_{SI}(\theta) = \frac{n}{m} \sum_{k \in S} l_k(\theta),$$

where S is the set of sampled observations. The SI scheme assigns equal inclusion probability to every observation. To illustrate the variability in $\hat{l}_{SI}(\theta)$ we consider the simple model $y_k \sim N(\theta, 0.1^2)$. Figure 1 shows $\hat{\sigma}_z = \left\{ \hat{V}[\hat{l}_{SI}(\theta)] \right\}^{1/2}$ for different choices of n and m. A large sampling fraction $f = m/n$ is needed to reduce the variance and this fraction increases as a function of the size of the data set. It seems impossible to achieve the optimal $V[\hat{l}_{SI}(\theta)] \approx 1$ with this sampling scheme, which explains why PMCMC based on the SI design does not work.

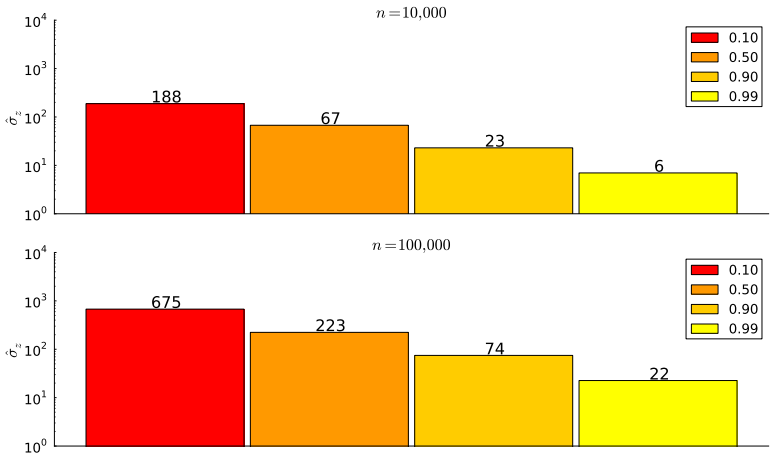

FIGURE 1. Estimated standard deviation of $\hat{l}_{SI}(\theta)$ on log-scale for the simple model $y_k \sim N(\theta, 0.1^2)$, for $n = 10,000$ (upper panel) and $n = 100,000$ (lower panel). The bars represent different sampling fractions, $f = \frac{m}{n}$.

3.2. **PPS.** The reason why equal-probability designs, such as SI, perform extremely poorly is because they treat all log-likelihood contributions $l_k(\theta)$ symmetrically. In practice, the log-likelihood contributions of some observations are much larger than others, and the sample should include those observations with higher probability. This is what the probability proportional-to-size (PPS) design achieves by sampling the kth unit with replacement with an inclusion probability p_k proportional to some measure of the unit's size.

We will use PPS sampling together with the Hansen-Hurwitz estimator in Equation (2.6). The choice of p_k is crucial for the variance of the estimator. Suppose we can choose sampling probabilities $p_k \propto l_k(\theta)$, so that $l_k(\theta)/p_k = c$ for all $k \in F$ where F is the index set of all observation and c is a constant. Then, in Equation (2.7), $\hat{l}(\theta) = c$ is constant and consequently $V[\hat{l}(\theta)] = 0$. This ideal estimator requires knowledge of $l_k(\theta)$ for all $k \in F$, in which case $l(\theta)$ is known so there is no point in subsampling. However, if we can construct sampling weights $w_k > 0$ so that $l_k(\theta)/w_k \approx c$ for all $k \in F$ and set $p_k \propto w_k$, then the ratio $l_k(\theta)/p_k$ will be approximately constant with small $V[\hat{l}(\theta)]$. It is evident that w_k must resemble the log-likelihood contribution $l_k(\theta)$; see below for some ways to achieve this.

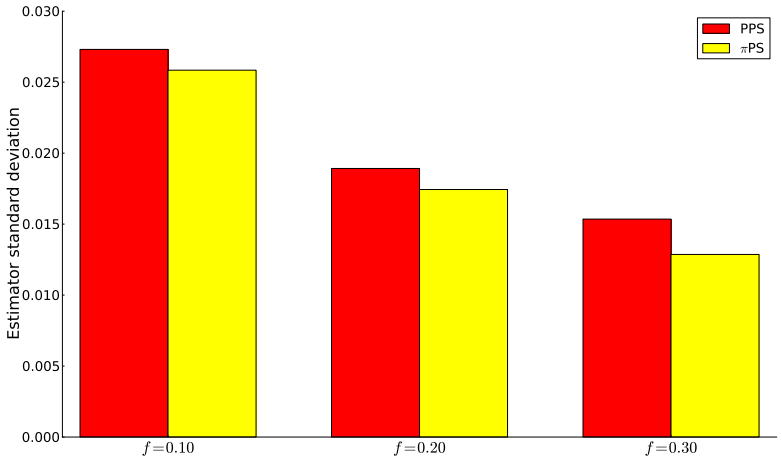

FIGURE 2. Comparing estimator standard deviation for πPS and PPS for different sampling fractions.

We will now compare the efficiency when probability proportional-to-size is instead implemented without replacement, which is called πPS sampling. Obtaining a single sample of size $m = 1,000$ from a population with $n = 10,000$, and subsequently estimating the variance takes 0.0009 sec for PPS compared to 65.1007 sec for πPS. Figure 2 shows the standard deviation of PPS vs πPS for the same simple model as in Figure 1. πPS is only marginally more efficient than PPS. Furthermore, by comparing Figure 1 and Figure 2 we see that both PPS and πPS are many orders of magnitude more efficient than SI.

We note that in our examples and for a number of other models, $l_k(\theta)$ is of the same sign for all k for a given θ, in which case we choose the sampling probabilities as above. More generally, $l_k(\theta)$ may take both positive and negative values so that it is impossible to construct $p_k > 0$ such that $l_k(\theta)/p_k \approx c$ for all k. However, for many such models we can write $l_k(\theta) = l_k^*(\theta) + d(\theta)$ where $l_k^*(\theta)$ has the same sign for all k for a given θ, and $d(\theta)$ does not depend on k. In this case we construct the sampling weights p_k based on $l_k^*(\theta)$. For example, consider the simple normal model with $y_k \sim N(\mu, \sigma_z^2)$ and $\theta = (\mu, \sigma^2)$. Then

$$l_k(\theta) = -\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2), \quad \text{with } l_k^*(\theta) = -\frac{1}{2\sigma^2}(y-\mu)^2 \text{ and } d(\theta) = -\frac{1}{2}\log(2\pi\sigma^2).$$

3.3. **Efficient PPS weights.** A crucial part of our PPS approach is the construction of the size weights, w_k. Since $w_k > 0$, let $w_k = |\tilde{l}_k(\theta)|$ where $\tilde{l}_k(\theta)$ is a proxy for the log-likelihood contribution $\log p(y_k|\theta, x_k)$. Note also that the goal is to reduce computing time so the construction of w_k needs to be fast. We now discuss three general ways to obtain the sampling weights.

A common approach for inference in models with computationally costly data density evaluations is to replace the model with an approximate *surrogate* model which is cheaper to evaluate. The choice of surrogate model depends on features of the original model, but may, for example, come from a low order Taylor series approximation of some aspect of the model or some other shortcut that makes the model easier to solve. The analysis is then performed as if the surrogate model is the true model. Our subsampling MCMC approach can instead use such a surrogate model in forming the weights w_k while still sampling from the correct posterior of the parameters in the true model. We refer to this as the *surrogate method* to obtain sampling weights.

Many models require time-consuming evaluations of $l_k(\theta)$ because some aspect of the model needs to be solved numerically. For example, an intractable integral may be approximated by Gaussian quadrature, a differential equation can be solved by the Runge-Kutta method, an optimum is found by Newton's method. Any numerical method depends on tuning parameters which control the accuracy of the solution. The sampling weights in our methodology can be computed from tuning parameters that give cruder, but much faster, evaluations of $l_k(\theta)$ (a coarse grid in numerical integration and in solving differential equations, a small number of Newton steps for optimization). The log-likelihood contributions for the sampled subset of observations are computed based on tuning parameters that give very accurate evaluations. We refer to these approaches as *numerical methods* for obtaining sampling weights.

Our final proposed approach is based on approximating the map (surface) $(x, y) \to l_k(\theta; y, x)$ at each iteration in the MCMC and then use this approximation to predict the log-likelihood contribution of all the other observations. The approximation is constructed

from a small number of observations for which the exact log-likelihood contribution has been computed. Since the predictions need to be performed at every MCMC iteration, great care is needed however to control the computational complexity of the surface fit. We discuss Gaussian processes (GP) and thin-plate splines in detail in Appendix A, but other surface fitting methods may also be used. A key feature of our approach is that hyperparameters such as covariance kernel parameters in the GP are determined before the MCMC. This allows us to compute the log-likelihood predictions in each MCMC iteration by fast matrix-vector multiplication. Appendix A also proposes a surface fitting approach for high-dimensional data.

4. Applications

4.1. **Bivariate probit model.** Our first application illustrates how to use the surface estimation method to construct the sampling weights.

Data. Our data set contains annual observations for Swedish firms in the time period 1991-2008. We have in total $500,000$ firm-year observations from $56,257$ firms. This data set is analyzed in a different setting by Jacobson et al. (2013), Giordani et al. (2013) and Quiroz and Villani (2013), using bankruptcy as the dependent variable. The present paper uses a bivariate model to analyze the endogenous treatment effect of holding cash on bankruptcy while controlling for other variables. The cash ratio variable has many severe outliers and we therefore choose to use the binary variable excess cash, which is set to one at a given time period if the firm has more cash than the median firm, and zero otherwise.

Model. We use a bivariate probit model to model the data. Multivariate probit models can be analyzed using data augmentation via the Gibbs sampler as in Chib and Greenberg (1998), but for illustration purposes we will demonstrate our approach by evaluating the likelihood directly with bivariate normal integrals. The bivariate probit model is

$$\begin{aligned} y_1 = I(y_1^* > 0) \text{ with } y_1^* &= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \alpha y_2 + \epsilon_1 \\ y_2 = I(y_2^* > 0) \text{ with } y_2^* &= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_3 + \beta_{23}x_4 + \epsilon_2 \end{aligned} \tag{4.1}$$

where $y_1 =$ Bankrupt, $y_2 =$ Excess Cash and $x_1 =$ Earnings, $x_2 =$ Leverage, $x_3 =$ Tangible, $x_4 =$ Size (log Sales). All variables except size are scaled with respect to total assets. The errors ϵ_1 and ϵ_2 are standard Gaussian with correlation ρ.

Estimation. We estimate the model with M-H using a single block of parameters. We sample $11,000$ draws from the posterior and discard the first 1000 draws as burn-in. We use the Fisher transformation $\tilde{\rho} = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$ for ρ. Let

$$\theta \;=\; (\beta_{10}, \beta_{11}, \beta_{12}, \alpha, \beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}, \tilde{\rho})^T$$

and set the prior to $p(\theta) = N(0, 10 \cdot I)$ for simplicity. Two different samplers are considered for both MCMC and PMCMC: the Random walk Metropolis (RWM) and the Independent M-H (IMH). The RWM uses the Hessian $H(\theta^*)$ of $p(\theta|y)$ evaluated at the posterior mode θ^* obtained from numerical optimization and sets $q(\theta_p|\theta_c) = N(\theta_c, c_\lambda H^{-1}(\theta^*))$. The IMH uses $q(\theta_p) = t_\nu(\theta^*, H^{-1}(\theta^*))$, where t_ν is the multivariate Student-t distribution with $\nu = 10$ degrees of freedom.

For PMCMC, the approximate log-likelihood contributions are obtained using the thin-plate spline approach in Appendix A, with a probit link for each of the four response outcomes separately. We use $|V| = 25,000$ and put more knots in responses with more observations. We note that $l_k(\theta) < 0$ for all k and θ in both this example and the next one.

Results. Table 1 shows the number of Efficient Draws Per Minute (EDPM) and the relative EDPM (REDPM) for PMCMC vs MCMC for the two different proposals, IMH and RWM. The targeted variance is 1 for both IMH and RWM. PMCMC gives more efficient draws per minute than regular MCMC, especially for the RWM algorithm. Both cases use around 8% of the data to estimate the likelihood (plus another 5% for training the thin-plate spline) so we should expect PMCMC to be around 7.69 times faster (given no overhead costs). In this application the overhead costs are expensive because the data set is very large. Besides the cost associated with the thin-plate spline at each iteration, there is a one time cost for initializing PMCMC (determine V, estimate λ, etc). This needs to be done only once

for a given data set and the results can be stored for future use. The fraction of this cost decreases as more posterior samples are desired. The overhead cost from the PPS sampling at each iteration is negligible in comparison to obtaining the approximate log-likelihood contributions. The mean Relative Inefficiency Factors (RIF) for RWM and IMH are 1.20 and 2.60, respectively. This confirms the result in Doucet et al. (2015) that RIF decreases as the proposal deteriorates.

TABLE 1. *Bivariate probit example.* Efficient draws per minute for PMCMC vs MCMC and relative efficient draws per minute (REDPM). The table also shows the acceptance probability $\Pr(Acc)$ and the scaling c_λ used for the RWM proposal. v_{max} is the maximum variance allowed in the estimator. $\bar{V}[\hat{l}(\theta_p)]$ is the mean of the estimated variance ($< v_{max}$) at the proposed θ values. $\bar{f}$ is the mean sampling fraction.

	IMH			RWM		
	MCMC	PMCMC	REDPM	MCMC	PMCMC	REDPM
β_{11}	5.780	8.892	1.538	0.310	1.064	3.432
β_{12}	6.523	8.487	1.301	0.382	1.133	2.966
β_{13}	6.300	8.807	1.398	0.371	1.149	3.097
α	6.177	9.159	1.483	0.281	0.816	2.904
β_{21}	5.633	8.733	1.550	0.372	1.186	3.188
β_{22}	6.687	8.248	1.233	0.435	1.400	3.218
β_{23}	7.073	7.863	1.112	0.468	1.169	2.498
β_{24}	5.076	9.730	1.917	0.267	1.203	4.506
ρ	6.008	9.040	1.505	0.283	0.871	3.078
$\Pr(Acc)$	0.752	0.478		0.253	0.195	
c_λ	n/a	n/a		0.680	0.680	
v_{max}	n/a	1		n/a	1	
$\bar{V}[\hat{l}(\theta_p)]$	n/a	0.995		n/a	0.995	
$\bar{f}$	1	0.080		1	0.081	

For both proposals, we experimented with increasing the targeted variance for the log-likelihood estimator. The case $\hat{V}[\hat{l}(\theta)] \approx 3.283$ is of particular interest for the RWM, but also $\Pr(Acc) \approx 0.07$ and $c_\lambda = 2.562/\sqrt{d} = 0.854$ (Sherlock et al., 2015). Table 2 presents the results for the RWM proposal and shows that $\hat{V}[\hat{l}(\theta)] \approx 1$ (Doucet et al., 2015) is optimal for our application. The same conclusion is reached for the IMH proposal (not reported), i.e. increasing the targeted variance results in fewer efficient draws per minute.

TABLE 2. *Bivariate probit example.* Inefficiency factors (IF) and relative efficient draws per minute (REDPM) for three different PMCMC with RWM for different values of v_{max} and c_λ as indicated in the table. v_{max} is the maximum variance allowed in the estimator and c_λ is the scaling used for the RWM proposal. The table also shows the acceptance probability $\Pr(Acc)$, the mean of the estimated variance at the proposed θ values $\bar{V}[\hat{l}(\theta_p)]$ ($< v_{max}$). $\bar{f}$ is the mean sampling fraction.

	PMCMC1		PMCMC2		PMCMC3	
	IF	REPDM	IF	REPDM	IF	REPDM
β_{11}	36.605	3.432	60.296	2.777	43.013	3.890
β_{12}	34.359	2.966	45.612	2.979	53.249	2.550
β_{13}	33.879	3.097	45.114	3.102	60.412	2.315
α	47.702	2.904	58.642	3.149	49.495	3.730
β_{21}	32.827	3.188	64.909	2.151	55.411	2.516
β_{22}	27.82	3.218	63.283	1.885	55.031	2.168
β_{23}	33.326	2.498	50.276	2.207	62.273	1.780
β_{24}	32.376	4.506	43.499	4.472	58.335	3.330
ρ	44.695	3.078	50.725	3.618	48.682	3.763
$\Pr(Acc)$	0.195		0.125		0.085	
c_λ	0.680		0.854		1	
v_{max}	1		3.283		3.283	
$\bar{V}[\hat{l}(\theta_p)]$	0.995		3.248		3.247	
$\bar{f}$	0.081		0.025		0.024	

The marginal posteriors from the MCMC and PMCMC algorithms are nearly indistinguishable (not shown here). Figure 3 explores the upper bound of the fractional error in the likelihood approximation in part (i) of Theorem 1. The figure is created using 1000 values of θ, and for each θ the subsample m is adapted (if needed) such that $\sigma_z^2 < 1$.

4.2. **Weibull survival model with random effects.** Our second application models the bankruptcy response in the previous section as discrete-time survival data. In particular, we extend the discrete-time Weibull survival model in Quiroz and Villani (2013) with random effects. This example illustrates how the numerical method for constructing the sampling weights is used within our PMCMC framework.

Data and likelihood. The data for each firm is recorded as a sequence $\{y_{ij}, x_{ij}\}_{j=1}^{n_i}$, where $y_{ij} = 1$ if firm i experiences the non-repeatable event bankruptcy at period j. The covariate

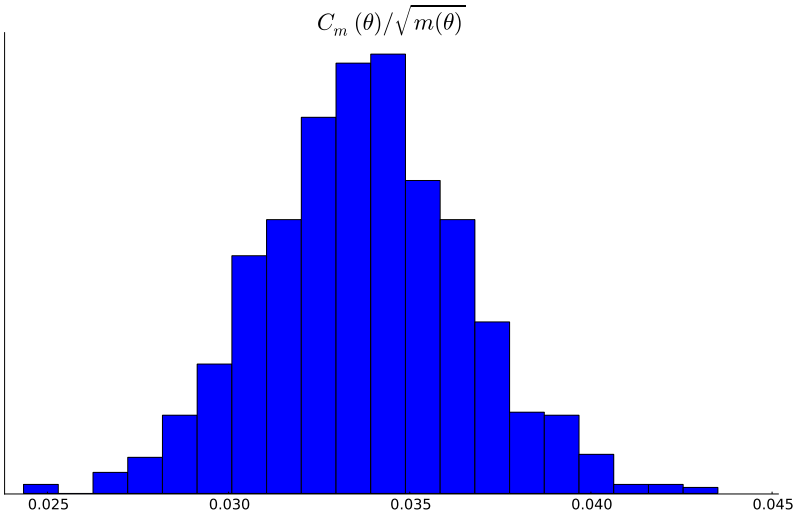

FIGURE 3. The bound for part (i) of Theorem 1 for the bivariate probit example. The figure shows the upper bound for the fractional error in the likelihood approximation computed over 1000 draws from the posterior. The subsample size $m(\theta)$ is chosen so that $\sigma_z^2 < 1$, on average $m \approx 40,000$ (8% of the full sample size) and $\sigma_z^2 = 0.95$ on average.

vector is $x_{ij}^T = (1, x_{ij}^1, \ldots, x_{ij}^5)^T$ where x^1 - earnings before interest and taxes over total assets (earnings ratio), x^2 - total liabilities over total assets (leverage ratio), x^3 - cash and liquid assets over total liabilities (cash ratio), x^4 - logarithm of (deflated) total sales (size) and x^5 - logarithm of firm age in years. We consider a data set with $2,000$ individual firms and $14,663$ firm-year observations.

The hazard probability, Pr(default at time t|non-default at time $t-1$), is modeled as

$$h_t(x_{ij}) \quad = \quad 1 - \exp\left(-\lambda\left(t_{ij}^{\rho} - t_{i(j-1)}^{\rho}\right)\right),$$

where

$$\log(\lambda) = \gamma_i + x_{ij}^T\beta_\lambda \text{ and } \log(\rho) = x_{ij}^T\beta_\rho, \text{ with } \gamma_i \overset{iid}{\sim} N(0,\tau^2). \tag{4.2}$$

The density of observations y_i for firm i, conditional on the random effect γ_i, is

$$p(y_i|\beta_\gamma, \beta_\rho, \gamma_i) = \prod_{j=1}^{n_i}(1 - h(x_{ij}))^{1-y_{ij}}h(x_{ij})^{y_{ij}},$$

with $y_i = (y_{i1}, \dots y_{in_i})$. The log-likelihood is obtained by integrating out γ_i for each firm so that

$$\log p(y|\beta_\gamma, \beta_\rho) = \sum_{i=1}^{n} \log \left(\int p(y_i|\beta_\gamma, \beta_\rho, \gamma_i) p(\gamma_i) d\gamma_i \right). \tag{4.3}$$

Estimation. We estimate the model with M-H using a single block of parameters. We sample $11,000$ draws from the posterior and discard the first 1000 draws as burn-in. For the variance parameter in the random effect we use $\tilde{\tau}^2 = \log(\tau^2)$. Let $\theta = (\beta_\lambda, \beta_\rho, \tilde{\tau}^2)^T$ and set the prior $p(\theta) = N(0, 10 \cdot I)$ for simplicity. We use the same RWM and IMH proposal as in Section 4.1 to sample the posterior (but with $\nu = 5$).

The integrals in Equation (4.3) are computed using the trapezoidal rule, where the tuning parameter is the step-size h. The value $h = 0.01$ is considered to give the "true value" of the integral and is determined by computing the log-likelihood evaluated at the posterior mode on a grid of h, and then choose the point where the difference in log-likelihood between two consecutive points is of the order 10^{-10}. This is a reasonable level of accuracy in the vast majority of numerical applications.

PMCMC uses a larger step-size to approximate the integral expression for $l_k(\theta)$ in the PPS-weights. It is important to understand the effect of increasing h. There are regions of the parameter space where a larger h still gives a good approximation of the integral. However, for other regions it can have a detrimental effect on the variance as Figure 4 illustrates for $h = 1.25$.

Results. Table 3 shows the relative efficient draws per minute for MCMC vs three different PMCMC (varying the step-size of the trapezoidal method in the approximate $l_k(\theta)$) using the RWM to propose θ. Many more efficient draws per minute are obtained with PMCMC and this is particularly true for the step-sizes $h = 0.50$ and $h = 1.25$. The decrease in REDPM with $h = 1.50$ is due to bad approximations over large regions of the parameter space so that the algorithm enters the adaptive phase too often, which results in increased computational time. It is difficult to obtain $\bar{V}[\hat{l}(\theta_p)] \approx 1$ and still be efficient with respect to computational time. This is because there is a sharp jump from a good approximation (low

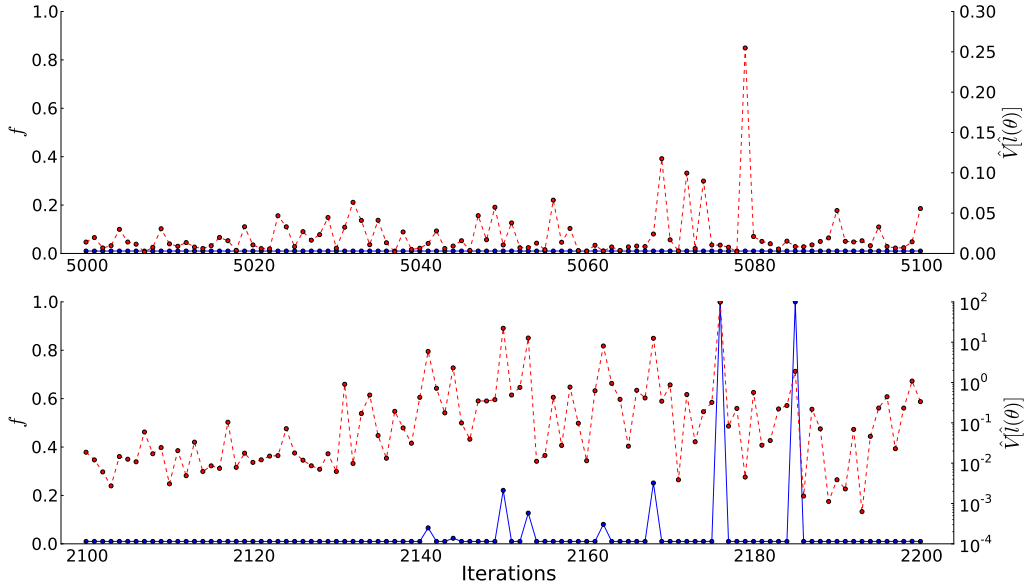

FIGURE 4. Sampling fraction and estimator variance at proposed point (before adaptation) against PMCMC iterations. The figure shows the adapted sampling fraction f, which is set to 0.01 at start (blue solid line and left y-axis), and the corresponding $\hat{V}[\hat{l}(\theta)]$ (red dashed line and right y-axis) for some iterations in the PMCMC. The upper panel shows the sampler with a RWM proposal exploring an area of the parameter space where the log-likelihood approximation is accurate. The lower panel (with a logarithmic right y-axis) shows the same example but when the approximation is poor.

variance) and a bad approximation (high variance) and therefore we have to set h so that $\hat{V}[\hat{l}(\theta_p)] > 1$ for any iteration and then bring it down to 1 by adapting the sampling fraction.

The PMCMC implementation with the IMH proposal was successful. This is in contrast to regular MCMC on the full data set that got stuck for very long periods and produced unacceptably large IFs.

Figure 5 compares the marginal posteriors of β_λ and τ^2 in Equation (4.2), obtained from MCMC ($h = 0.01$) and PMCMC ($h = 0.5$) with a RWM proposal. The PMCMC uses only 1% of the data (m =20 firms) and obtains very accurate posteriors. Furthermore, Figure 6 explores the bound for the fractional error in the likelihood approximation in Theorem 1. The bound is extremely small, which explains the accurate approximation of the marginal posteriors in Figure 5.

TABLE 3. *Survival Weibull example.* Relative efficient draws per minute for PMCMC ($h = 0.50, 1.25, 1.50$) with a RWM proposal. The MCMC is carried out with $h = 0.01$ and $\Pr(Acc) = 0.237$. The table also shows the acceptance probability $\Pr(Acc)$ and the scaling c_λ used for the RWM for PMCMC. v_{max} is the maximum variance allowed in the estimator. $\bar{V}[\hat{l}(\theta_p)]$ is the mean of the estimated variance at the proposed θ values. f_{start} is the initial sampling fraction set in the algorithm, which is then increased if $\hat{V}[\hat{l}(\theta_p)] > 1$. $\bar{f}$ is the mean sampling fraction.

	REDPM		
	$h = 0.50$	$h = 1.25$	$h = 1.50$
$\beta_{\lambda 1}$	3.980	4.848	2.746
$\beta_{\lambda 2}$	6.883	7.648	6.580
$\beta_{\lambda 3}$	3.208	3.651	2.521
$\beta_{\lambda 4}$	3.863	5.532	3.169
$\beta_{\lambda 5}$	3.138	2.557	1.953
$\beta_{\lambda 6}$	4.965	5.358	2.543
$\beta_{\rho 1}$	5.424	4.359	3.657
$\beta_{\rho 2}$	4.224	3.829	4.854
$\beta_{\rho 3}$	5.465	5.998	4.338
$\beta_{\rho 4}$	4.385	3.705	3.419
$\beta_{\rho 5}$	3.248	3.478	2.400
$\beta_{\rho 6}$	6.268	6.898	4.221
τ^2	5.068	5.996	4.241
$Pr(Acc)$	0.237	0.240	0.215
c_λ	0.416	0.416	0.404
v_{max}	1	1	1
$\bar{V}[\hat{l}(\theta_p)]$	0.188×10^{-3}	0.156	0.415
f_{start}	0.010	0.010	0.030
$\bar{f}$	0.010	0.011	0.078

It should be noted that while a cruder numerical integration works well for obtaining the PPS-weights, it cannot be used to speed up regular MCMC on the full data set. The reason is that there are regions of the parameter space where a crude numerical integration gives a very misleading likelihood value, see Figure 4 for an example, and the MCMC will then sample from the wrong posterior.

Note that while the REDPM values are already large in this example with a univariate random effect, it is clear that increasing the dimension of the random effects would lead to even larger relative efficiency gains. It should also be noted that the numerical integration

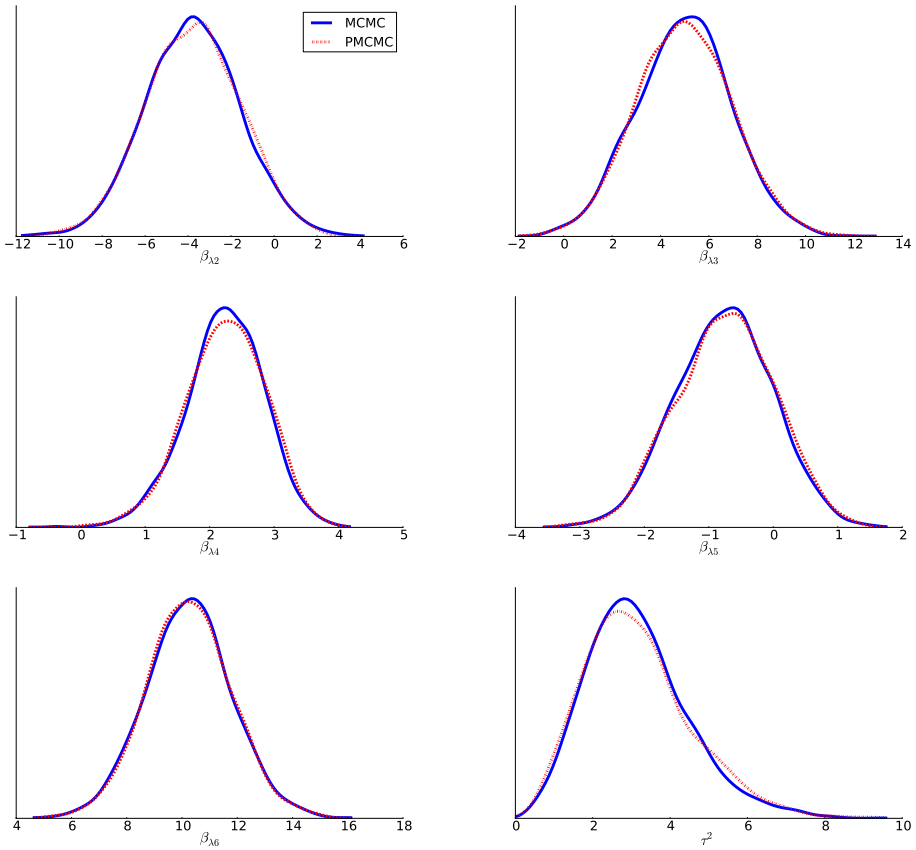

FIGURE 5. Marginal posterior distributions for MCMC (solid blue line) vs PMCMC (dashed red line) using a RWM proposal.

algorithm used here is highly vectorized and the computational cost decreases rather slowly with increased step sizes. Other numerical methods have a computational cost which is linear in the tuning parameter, and for such problems our subsampling method with PPS-weights based on a relaxed tuning parameter would be dramatically better than MCMC on the full sample.

5. Conclusions and Future Research

We propose a new framework for speeding up MCMC on models with time-consuming likelihood functions. The algorithms use efficient subsampling of the data and an estimator of the likelihood in a PMCMC scheme that is shown to sample from a posterior which is within $O(m^{-\frac{1}{2}})$ of the true posterior, where m is the subsample size. Moreover, the constant of proportionality in the error bound of the likelihood is shown to be small and our empirical

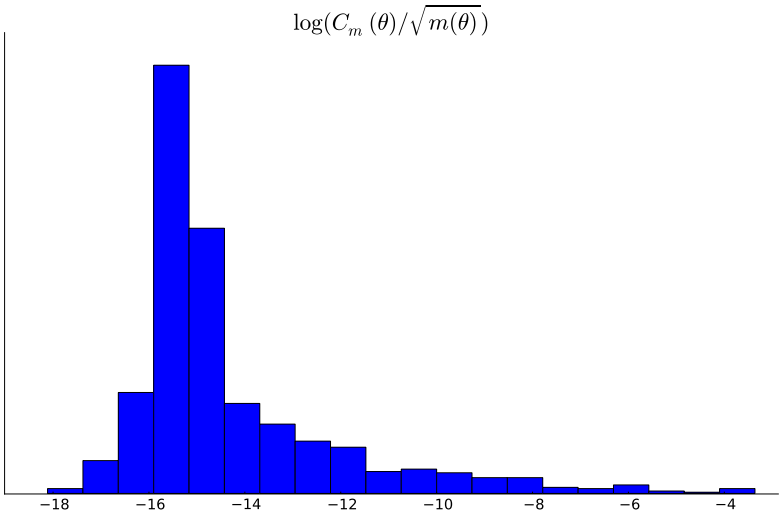

FIGURE 6. The bound for part (i) of Theorem 1 for the Weibull survival example. The figure shows the upper bound for the (log) fractional error in the likelihood approximation computed over 1000 draws from the posterior. The subsample size $m(\theta)$ is chosen so that $\sigma_z^2 < 1$, on average $m \approx 20$ (1% of the full sample size) and $\sigma_z^2 = 0.015$ on average.

applications clearly illustrate the accuracy of our approach. We demonstrate why simple schemes such as SI will not work, and how Probability Proportional-to-Size (PPS) sampling can reduce the variance of the likelihood estimator by many orders of magnitude by assigning higher sampling inclusion probabilities to observations that make a larger contribution to the likelihood. We propose three different strategies for obtaining efficient PPS sampling probabilities.

We argue that the assumptions in Doucet et al. (2015) are satisfied, and the sample size can therefore be conveniently tuned to target a variance of the log-likelihood estimator of around 1 for an optimal trade-off between efficiency and computing time. This choice is conservative and thus minimizes the risk of the PMCMC chain getting stuck, which is also evident in our applications. We propose an adaptive strategy for setting the sampling fraction so that the variance of the estimator is around 1 with a tight lower bound, hence avoiding including unnecessarily many observations for computing the likelihood.

The proposed algorithm is evaluated in two examples. The first example is a bivariate probit model with $500,000$ observations. The second example is a discrete time Weibull

model with random effects. In both examples we find that the proposed algorithm generates more efficient posterior sample draws per minute than the corresponding MCMC on the full data set. This is true in particular for weaker proposals, a result consistent with previous literature.

The use of efficient sampling schemes for subsampling data in PMCMC opens up many interesting directions for future research, including the development of alternative sampling designs and more efficient likelihood estimators. It is our hope that the proposed framework will motivate researchers in survey sampling to contribute to this important area of improving the speed and efficiency of MCMC methods for complex problems.

6. Supplementary Material

Online appendix: Proof of Theorem 1.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

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

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

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

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

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.

Giordani, P., Jacobson, T., Von Schedvin, E., and Villani, M. (2013). Taking the twists into account: Predicting firm bankruptcy risk with splines of financial ratios. *Journal of Financial and Quantitative Analysis*, forthcoming.

Hansen, M. H. and Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362.

Jacobson, T., Lindé, J., and Roszbach, K. (2013). Firm default and aggregate fluctuations. *Journal of European Economic Association*, 11:945–972.

Korattikara, A., Chen, Y., and Welling, M. (2013). Austerity in MCMC land: Cutting the Metropolis- Hastings budget. *arXiv preprint arXiv:1304.5299*.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

Ma, S., Racine, J. S., and Yang, L. (2011). Spline regression in the presence of categorical predictors. Technical report, Working paper, McMaster University and Michigan State University.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Ormerod, J. and Wand, M. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

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.

Quiroz, M. and Villani, M. (2013). Dynamic mixture-of-experts models for longitudinal and discrete-time survival data. Technical report, Sveriges Riksbank Working Paper Series.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The design and analysis of computer experiments*. Springer.

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

Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *To appear in Annals of Statistics*.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons.

Appendix A. pps weights from surface fitting

Let $d = (x^T, y^T)^T$ and assume that we have computed the log-likelihood contribution $l(\theta; d)$ for all d in a small fixed subset V of the full dataset for the θ in the current MCMC iteration. Let $l_V(\theta)$ collect these $|V|$ values, and let V^c be the complement of V, i.e. the remaining observations. Given that $l(\theta; d)$ is known for all $d \in V$, the population total can be decomposed as

$$l(\theta) \;=\; \sum_{d \in V} l(\theta; d) + \sum_{d \in V^c} l(\theta; d),$$

where the first term is known. Our sample is therefore only drawn from the observations in V^c. A natural way to approximate $l(\theta; d)$ at any $d \in V^c$ is by a noise-free Gaussian Process (GP); see Rasmussen and Williams (2006) for an introduction to GPs. That is, we use a GP prior $l(\theta; d) \sim GP\,[0, k(d, d')]$, where $k(d, d')$ is a positive definite covariance kernel, and update it to a GP posterior using the exact data density evaluations for the observations in V. The zero mean in the GP prior can be replaced by any crude surrogate model, if available. The predicted log-likelihood contributions for the observations $d \in V^c$ are given by

$$\hat{l}_{V^c}(\theta) = K(d_{V^c}, d_V) K(d_V, d_V)^{-1} l_V(\theta),$$

where d_A is the vector of data points in set A, $K(d_{V^c}, d_V)$ is the $|V^c| \times |V|$ covariance matrix between d_{V^c} and d_V based on the covariance kernel $k(d, d')$, and $K(d_V, d_V)$ is the $|V| \times |V|$ covariance matrix of d_V.

A commonly used kernel function is the squared exponential kernel

$$k(d, d') = \sigma_f^2 \exp\left(-\frac{1}{2\ell^2} \|d - d'\|^2\right).$$

The three-parameter Matern kernel or the ARD kernel are also attractive choices (Rasmussen and Williams, 2006). The hyperparameters σ_f and ℓ can be set before the MCMC by minimizing the prediction errors $\left\| l_{V^c}(\hat{\theta}) - \hat{l}_{V^c}(\hat{\theta}) \right\|$ at some estimate $\hat{\theta}$, e.g. the mode of the likelihood obtained by Newton's method. Since the subset V is fixed throughout the MCMC and the hyperparameters optimized before the MCMC, the matrix $K(d_V, d_V)^{-1}$ does not change over the MCMC iterations. Computing the predictions $\hat{l}_{V^c}(\theta)$ therefore involves only the matrix-vector multiplication $a = K(d_V, d_V)^{-1} l_V(\theta)$ followed by the matrix-vector multiplication $K(d_{V^c}, d_V)a$. This is typically fast compared to computing $l_{V^c}(\theta)$ which can be prohibitively expensive for complex density evaluations. The experimental designs in Sacks et al. (1989) and Santner et al. (2003) can be used to select the subset of observations in V optimally. The computations for the GP can nevertheless be costly on large datasets (but see the approximate GP methods in Rasmussen and Williams 2006, Ch. 8) and we now propose an alternative approximation based on spline regression.

Spline surfaces with thin-plate radial basis functions can be used to approximate the log-likelihood contributions. As before, define $d = (x^T, y^T)^T$ and denote the thin-plate spline approximation by $g(d; \gamma) = \sum_{m=1}^{M} \gamma_m g_m(d)$ where $g_m : \mathbb{R}^{\dim(d)} \to \mathbb{R}$ is $g_m(d) = ||d - \xi_m||^2 \log(||d - \xi_m||)$ and ξ_m is the mth knot. The knot locations are chosen with the k-means algorithm on the data space d before running the MCMC algorithm. The training set V must give good coverage of the data space in general, and the boundary in particular, as it is used to predict $l_{V^c}(\theta)$.

The predicted log-likelihood contributions are

$$\hat{l}_{V^c}(\theta) = B_{V^c}(B_V^T B_V + \lambda I)^{-1} B_V^T l_V(\theta),$$

where B_V denotes the basis-expanded matrix from M knots for the observations in V, and $\lambda > 0$ is a shrinkage factor. Analogously to the GP case, λ is set before the MCMC to the value that minimizes the prediction errors $\left\| l_{V^c}(\hat{\theta}) - \hat{l}_{V^c}(\hat{\theta}) \right\|$, and V, B_V, B_{V^c} and λ are fixed throughout the MCMC. When $M \ll |V|$ the computational complexity of the thin-plate spline is much lower than that of the GP, and is therefore likely to be the preferred method on large datasets.

The error in $\hat{l}_{V^c}(\theta)$ resulting from λ being evaluated at $\hat{\theta}$ rather than θ can be reduced as follows. We can adjust the predicted $\hat{l}(\theta; d)$ at any data point $d \in V^c$ by using the known prediction error $l(\theta; d_*) - \hat{l}(\theta; d_*)$ for the observation $d_* \in V$ which is closest to $d \in V^c$. This refinement will work well when the paired observation are close (V is a good coverage of the data space) and the log-likelihood contribution does not change drastically between these points.

For some models, it is convenient to transform $l_V(\theta)$ with a link function before regressing on data space. One example is dose-response models where, e.g., the logit-link is suitable for transforming $p(y_k|x_k, \theta) \in [0, 1]$. In models with categorical response we can run separate regressions for each of the categories (i.e. $d = x$). In the presence of many categories, spline regression with categorical predictors as in Ma et al. (2011) can be applied.

Local approximations such as the thin-plate spline are likely to give poorly estimated sampling weights for high dimensional data. In this case we suggest a dimension reduction of the data space as follows. Suppose the model has P features, i.e. model parameters linked to covariates. Each feature is a function of $x_p^T \theta_p$ where x_p and θ_p denote the covariates and the parameter for the pth feature, respectively. Let $\tilde{d}_p = x_p^T$ be the data corresponding to the pth feature. In some cases it is also connected to the response in the log-density and then $\tilde{d}_p = (x_p^T, y^T)^T$. Define the mapping $h_p : \mathbb{R}^{\dim(\tilde{d}_p)} \to \mathbb{R}$ with inputs $\tilde{d}_p$ and θ_p. The reduced data is now $\tilde{d} = (h_1, \ldots, h_P)^T$ with $P << d$ (in practice often $P \leq 2$), and

$l_k(\theta)$ is parametrized in terms of $\tilde{d}$. We proceed as before, but since the data changes the basis expanded matrix B is recomputed in each MCMC iteration. This approach works well when the computational cost of the complex model dominates the relatively cheap cost of computing the surface fit. All quantities that were previously fixed can be determined similarly by using $\tilde{d}_p$ evaluated at the posterior mode.

Supplementary Material for "Speeding Up MCMC By Efficient Data Subsampling"

Our proof of Theorem 1 makes use of the following lemmas.

Lemma 1. *Suppose that b_m and b'_m are bounded sequences. Then*

i.

$$\left(1 + \frac{x}{m} + \frac{b_m}{m^{3/2}} + \frac{o(1)}{m^{3/2}}\right)^m / \exp(x) \;=\; 1 + \frac{b_m}{m^{1/2}} + \frac{o(1)}{m^{1/2}}.$$

ii.

$$\left(1 + \frac{b'_m}{(m-1)^2} + o(m^{-2})\right)^m \;=\; 1 + \frac{b'_m}{m-1} + o(m^{-1}).$$

Proof. The proof of both (i) and (ii) is a straightforward application of a first order Taylor series with remainder. □

Lemma 2. *Suppose that X and Y are two random variables with $E[X] = 0$ and $E[Y] = 0$. Then,*

i.

$$0 \leq E\left[\exp(X+Y)\right] - 1 \leq \left(E\left[\exp(2X)\right] - 1\right)^{\frac{1}{2}} \left(E\left[\exp(2Y)\right] - 1\right)^{\frac{1}{2}} + \left(E\left[\exp(X)\right] - 1\right) + \left(E\left[\exp(Y)\right] - 1\right).$$

ii.

$$0 \leq E\left[\exp(X)\left(\exp(Y) - 1\right)\right] \;\leq\; \left(E\left[\exp(2X)\right]\right)^{\frac{1}{2}} \left(E\left[\exp(2Y)\right] - 1\right)^{\frac{1}{2}}.$$

Proof. To prove (i), note that for any r.v ξ, with $E[\xi] = 0$, we can show that $E[\exp(\xi)] \geq 1$. Now,

$$\exp(X+Y) - 1 \;=\; \left(\exp(X) - 1\right)\left(\exp(Y) - 1\right) + \exp(X) - 1 + \exp(Y) - 1,$$

and by Cauchy-Schwartz

$$0 \leq E\left[\left(\exp(X)-1\right)\left(\exp(Y)-1\right)\right] \;\leq\; \left(E\left[\left(\exp(X)-1\right)^2\right]\right)^{\frac{1}{2}}\left(E\left[\left(\exp(Y)-1\right)^2\right]\right)^{\frac{1}{2}}.$$

The proof is completed by noting that

$$\text{(S.1)}\quad E\left[\left(\exp(X)-1\right)^2\right] \;=\; E\left[\exp(2X)\right]-1-2\left(E\left[\exp(X)\right]-1\right) \leq E\left[\exp(2X)\right]-1,$$

and similarly for Y.

To prove (ii), by Cauchy-Schwartz

$$0 \leq E\left[\exp(X)\left(\exp(Y)-1\right)\right] \;\leq\; \left(E\left[\exp(2X)\right]\right)^{\frac{1}{2}}\left(E\left[\left(\exp(Y)-1\right)^2\right]\right)^{\frac{1}{2}},$$

and applying Equation (S.1) for Y completes the result. □

Lemma 3. *Suppose that λ is a constant. Then the following results hold*

i.

$$E\left[\exp\left(\lambda(\hat{l}_m - l)\right)\right] / \exp\left(\frac{\lambda^2}{2}\sigma_z^2\right) \;=\; 1+\lambda^3\frac{b_m}{m^{1/2}}+\frac{\Lambda_m^{(1)}(\lambda)}{m^{1/2}},$$

where

$$b_m = \left(\sigma_z^2\right)^{3/2}\gamma_3/6 \quad \textit{and} \quad \Lambda_m^{(1)}(\lambda) = o(1)$$

can be computed from Equation (S.4) *by direct evaluation for any θ.*

ii.

$$\text{(S.2)}\qquad E\left[\exp\left(\lambda\frac{(\hat{l}_m - l)^2-\sigma_z^2}{m-1}\right)\right]-1 \;=\; \lambda^2\frac{\tilde{b}_m}{(m-1)^2}+\frac{\Lambda_m^{(2)}(\lambda)}{m^2}$$

where

$$\tilde{b}_m = \frac{\left(\sigma_z^2\right)^2}{2m}\left(\gamma_4+2m-3\right) \quad \textit{and} \quad \Lambda_m^{(2)}(\lambda) = o(1)$$

can be computed from Equation (S.2) *by simulation for any θ.*

iii. *Let*

$$Y_m = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left((\zeta_i - l)^2 - \sigma_\zeta^2 \right), \tag{S.3}$$

Then,

$$E\left[\exp(\lambda Y_m)\right] - 1 = \lambda^2 \frac{b_m^*}{m-1} + \frac{\Lambda_m^{(3)}(\lambda)}{m}$$

where

$$b_m^* = \frac{\left(\sigma_z^2\right)^2}{2}(\gamma_4 - 1) \quad \textit{and} \quad \Lambda_m^{(3)}(\lambda) = o(1)$$

can be computed from Equation (S.7) *by direct evaluation for any* θ.

Proof. To prove (i) consider $E\left[\exp(\lambda(\hat{l}_m - l))\right]$ with $\lambda(\hat{l}_m - l) = \frac{1}{m}\sum_{i=1}^{m} \lambda(\zeta_i - l)$. By part (ii) of Assumption 1

$$E\left[\exp\left(\lambda \frac{\zeta_1 - l}{m}\right)\right] = 1 + \frac{\lambda^2}{2m}\sigma_z^2 + \frac{\lambda^3 b_m}{m^{3/2}} + \frac{c'_m}{m^{3/2}}$$

with

$$b_m = \gamma_3 \left(\sigma_z^2\right)^{3/2}/6 \quad \text{and} \quad c'_m = o(1).$$

Because the ζ_i's are iid, and by part (i) of Lemma 1 (with $x = \frac{\lambda^2}{2}\sigma_z^2$ and b_m bounded),

$$\begin{aligned} E\left[\exp(\lambda(\hat{l}_m - l))\right] &= \left(E\left[\exp\left(\lambda \frac{\zeta_1 - l}{m}\right)\right]\right)^m \\ &= \exp\left(\frac{\lambda^2}{2}\sigma_z^2\right)\left(1 + \frac{b_m}{m^{1/2}} + \frac{\Lambda_m^{(1)}(\lambda)}{m^{1/2}}\right) \end{aligned} \tag{S.4}$$

and $\Lambda_m^{(1)}(\lambda) = o(1)$, which concludes (i).

Define $\nu_m = (\hat{l}_m - l)^2 - \sigma_z^2$. The proof follows from part (iii) of Assumption 1

$$E\left[\exp\left(\lambda \frac{\nu_m}{m-1}\right)\right] = 1 + \frac{\lambda^2}{(m-1)^2}\tilde{b}_m + \frac{\Lambda_m^{(2)}(\lambda)}{m^2}, \tag{S.5}$$

with

$$\tilde{b}_m = \frac{\left(\sigma_z^2\right)^2}{2m}\left(\gamma_4 + 2m - 3\right) \quad \text{and} \quad \Lambda_m^{(2)}(\lambda) = o(1).$$

Proof of part (iii). By part (iv) of Assumption 1,

$$E\left[\exp\left(\lambda\frac{(\zeta_1-l)^2-\sigma_\zeta^2}{m(m-1)}\right)\right]=1+\frac{b'_m}{(m-1)^2}+\frac{k_m}{(m-1)^2} \tag{S.6}$$

where

$$b'_m=\frac{\lambda^2\left(\sigma_z^2\right)^2}{2}(\gamma_4-1)\quad\text{and}\quad k_m=o(1).$$

The ζ_i's are iid, and by part (ii) of Lemma 1 (b'_m is bounded),

$$\begin{aligned}E\left[\exp\left(\lambda Y_m\right)\right] &= \left(E\left[\exp\left(\lambda\frac{(\zeta_1-l)^2-\sigma_\zeta^2}{m(m-1)}\right)\right]\right)^m\\ &= 1+\frac{b'_m}{m-1}+\frac{\Lambda_m^{(3)}(\lambda)}{m}\end{aligned} \tag{S.7}$$

and $\Lambda_m^{(3)}(\lambda)=o(1)$ which concludes the proof. □

Lemma 4. *Let $b_m,\tilde{b}_m,b_m^*$ and $\Lambda_m^{(1)},\Lambda_m^{(2)},\Lambda_m^{(3)}$ be defined as in Lemma 3. Under Assumption 1:*

i.

$$\left|E\left[\exp\left(\hat{l}_m-\frac{1}{2}\sigma_z^2\right)\right]-\exp(l)\right| = \exp(l)\frac{1}{\sqrt{m}}\left|A_m\right| \quad \textit{where } A_m=b_m+\Lambda_m^{(1)}(1)$$

and $A_m=O(1)$.

ii.

$$\left|E\left[\exp(\hat{l}_m-\frac{1}{2}\hat{\sigma}_z^2)\right]-E\left[\exp(\hat{l}_m-\frac{1}{2}\sigma_z^2)\right]\right| \leq \exp(l)\frac{1}{\sqrt{m}}\left|B_m\right|$$

where $B_m=\sqrt{m}U_1^{1/2}U_2^{1/2}=O(1)$,

$$U_1=\exp(\sigma_z^2)\left(1+8\frac{b_m}{m^{1/2}}+\frac{\Lambda_m^{(1)}(2)}{m^{1/2}}\right)$$

$$U_2 = \left(\frac{4\tilde{b}_m}{(m-1)^2} + \frac{\Lambda_m^{(2)}(2)}{m^2}\right)^{1/2} \left(\frac{4b_m^*}{m-1} + \frac{\Lambda_m^{(3)}(-2)}{m}\right)^{1/2} + \frac{\tilde{b}_m}{(m-1)^2} + \frac{\Lambda_m^{(2)}(1)}{m^2} + \frac{b_m^*}{m-1} + \frac{\Lambda_m^{(3)}(-1)}{m}$$

Proof. By part (i) of Lemma 3 (with $\lambda = 1$),

$$E\left[\exp\left(\hat{l}_m - \frac{1}{2}\sigma_z^2\right)\right] - \exp(l) \;=\; \exp(l)\left(\frac{b_m}{m^{1/2}} + \frac{\Lambda_m^{(1)}(1)}{m^{1/2}}\right)$$

and taking the absolute value proves (i).

To prove part (ii), note that

$$\left|E\left[\exp(\hat{l}_m - \frac{1}{2}\hat{\sigma}_z^2)\right] - E\left[\exp(\hat{l}_m - \frac{1}{2}\sigma_z^2)\right]\right| \;=\; \exp\left(l - \frac{1}{2}\sigma_z^2\right) \times \left(E\left[\exp(V_m)(\exp(W_m) - 1)\right]\right) \tag{S.8}$$

where

$$V_m = \hat{l}_m - l \quad \text{and} \quad W_m = \frac{1}{2}(\sigma_z^2 - \hat{\sigma}_z^2).$$

By part (ii) of Lemma 2

$$0 \le E\left[\exp(V_m)(\exp(W_m) - 1)\right] \;\le\; \left(E\left[\exp(2V_m)\right]\right)^{\frac{1}{2}} \left(E\left[\exp(2W_m)\right] - 1\right)^{\frac{1}{2}}.$$

From part (i) of Lemma 3 (with $\lambda = 2$)

$$E\left[\exp(2V_m)\right] \;=\; \exp(2\sigma_z^2)\left(1 + 8\frac{b_m}{m^{1/2}} + \frac{\Lambda_m^{(1)}(2)}{m^{1/2}}\right)$$

and $U_1 = \exp\left(-\sigma_z^2\right) E\left[\exp(2V_m)\right] = O(1)$.

We now derive the upper bound U_2 of $E\left[\exp(2W_m)\right] - 1$. We can show that

$$W_m \;=\; \frac{1}{2}(X_m - Y_m),$$

with $X_m = \left((\hat{l}_m - l)^2 - \sigma_z^2\right)/(m-1)$ and Y_m as in Equation (S.3). By part (i) of Lemma 2

$$E\left[\exp(2W_m)\right] - 1 = E\left[\exp(X_m - Y_m)\right] - 1 \leq \left(E\left[\exp(2X_m)\right] - 1\right)^{\frac{1}{2}} \left(E\left[\exp(-2Y_m)\right] - 1\right)^{\frac{1}{2}} + \left(E\left[\exp(X_m)\right] - 1\right) + \left(E\left[\exp(-Y_m)\right] - 1\right).$$

For X_m the expectations (with $\lambda = 1, 2$) follow from part (ii) of Lemma 3

$$E\left[\exp(2X_m)\right] - 1 \;=\; 4\frac{\tilde{b}_m}{(m-1)^2} + \frac{\Lambda_m^{(2)}(2)}{m^2}.$$

Similarly for Y_m (with $\lambda = -1, -2$), but using part (iii) of Lemma 3

$$E\left[\exp(-2Y_m)\right] - 1 \;=\; 4\frac{b_m^*}{m-1} + \frac{\Lambda_m^{(3)}(-2)}{m}.$$

Since $U_2 = O(m^{-1})$ we conclude that $B_m = O(1)$. □

After the following important remark we are now ready to prove Theorem 1.

Remark. The proof of Theorem 1 involves integrating over θ. To save space we have suppressed the dependence of θ everywhere and will continue to do so. However, we stress the fact that this notation indeed makes all assumptions to hold for any value of θ. When θ changes, the sampling weights change to fulfill the approximate proportionality assumption. In practice this is done by recomputing the weights using any of the three methods proposed in the paper.

Proof of Theorem 1. Proof of (i). Since

$$p_m(y|\theta) - p(y|\theta) \;=\; E\left[\exp(\hat{l}_m - \frac{1}{2}\hat{\sigma}_z^2)\right] - E\left[\exp(\hat{l}_m - \frac{1}{2}\sigma_z^2)\right] + E\left[\exp(\hat{l}_m - \frac{1}{2}\sigma_z^2)\right] - \exp(l)$$

the proof follows directly from Lemma 4. It also follows that $C_m(\theta) = |A_m(\theta)| + |B_m(\theta)|$ is bounded.

To prove part (ii), consider

$$|p_m(y) - p(y)| \leq \int |p_m(y|\theta) - p(y|\theta)|\, p(\theta) d\theta.$$

From part (i),

$$|p_m(y|\theta) - p(y|\theta)| \leq p(y|\theta)\frac{1}{\sqrt{m}} C_m(\theta) \leq p(y|\theta)\frac{1}{\sqrt{m}} \sup_\theta C_m(\theta) \quad \text{uniformly for } \theta,$$

and since $\int p(y|\theta)p(\theta)d\theta = p(y)$ the result follows.

To prove part (iii), we note that

$$\begin{aligned}
\pi_m(\theta) - \pi(\theta) &= p_m(y|\theta)p(\theta)/p_m(y) - p(y|\theta)p(\theta)/p(y) \\
&= (p_m(y|\theta) - p(y|\theta))\, \frac{p(\theta)}{p(y)} \frac{p(y)}{p_m(y)} \\
&\quad -p(y|\theta)p(\theta)\left(\frac{1}{p(y)} - \frac{1}{p_m(y)}\right).
\end{aligned}$$

By part (i),

$$\begin{aligned}
|\pi_m(\theta) - \pi(\theta)| &\leq |p_m(y|\theta) - p(y|\theta)|\, \frac{p(\theta)}{p_m(y)} + p(y|\theta)p(\theta)\left|\frac{1}{p(y)} - \frac{1}{p_m(y)}\right| \\
&\leq p(y|\theta)\frac{1}{\sqrt{m}} C_m(\theta)\frac{p(\theta)}{p_m(y)} + p(y|\theta)p(\theta)\left|\frac{1}{p(y)} - \frac{1}{p_m(y)}\right|.
\end{aligned}$$

From part (ii),

$$\begin{aligned}
\left|\frac{1}{p(y)} - \frac{1}{p_m(y)}\right| &= \frac{1}{p(y)p_m(y)}\, |p_m(y) - p(y)| \\
&\leq \frac{1}{p(y)p_m(y)} p(y) \frac{1}{\sqrt{m}} \sup_\theta C_m(\theta).
\end{aligned}$$

Hence

$$\begin{aligned}
|\pi_m(\theta) - \pi(\theta)| &\leq \frac{p(\theta)p(y|\theta)}{p(y)} \frac{1}{\sqrt{m}} C_m(\theta)\frac{p(y)}{p_m(y)} + p(\theta)p(y|\theta)\frac{p(y)}{p(y)p_m(y)}\frac{1}{\sqrt{m}} \sup_\theta C_m(\theta) \\
&= \pi(\theta)\frac{1}{\sqrt{m}}\left(\frac{p(y)}{p_m(y)} C_m(\theta) + \frac{p(y)}{p_m(y)} \sup_\theta C_m(\theta)\right)
\end{aligned}$$

which proves part (iii). The expression between brackets is D_m and is clearly bounded.

Proof of part (iv). From part (iii)

$$|\pi_m(\theta) - \pi_m(\theta)| \leq \pi(\theta)\frac{1}{\sqrt{m}}D_m(\theta) \leq \pi(\theta)\frac{1}{\sqrt{m}}\sup_\theta D_m(\theta) \quad \text{uniformly for } \theta.$$

Therefore

$$\left|\int h(\theta)(\pi_m(\theta) - \pi(\theta))d\theta\right| \leq \frac{1}{\sqrt{m}}\sup_\theta D_m(\theta)\int |h(\theta)|p(\theta)d\theta,$$

and the last statement follows by taking $h(\theta)$ as the identity function.

□