Spectral Subsampling MCMC for Stationary Time Series

Robert Salomone[1] **Matias Quiroz**[2] **Robert Kohn**[1] **Mattias Villani**[3 4] **Minh-Ngoc Tran**[5]

Abstract

Bayesian inference using Markov Chain Monte Carlo (MCMC) on large datasets has developed rapidly in recent years. However, the underlying methods are generally limited to relatively simple settings where the data have specific forms of independence. We propose a novel technique for speeding up MCMC for time series data by efficient data subsampling in the frequency domain. For several challenging time series models, we demonstrate a speedup of up to two orders of magnitude while incurring negligible bias compared to MCMC on the full dataset. We also propose alternative control variates for variance reduction based on data grouping and coreset constructions.

1. INTRODUCTION

Bayesian inference has gained widespread use in Statistics and Machine Learning largely due to convenient and quite generally applicable Markov Chain Monte Carlo (MCMC) and Hamiltonian Monte Carlo (HMC) algorithms that simulate from the posterior distribution of the model parameters.

However, it is now increasingly common for datasets to contain millions or even billions of observations. This is particularly true for temporal data recorded by sensors at increasingly faster sampling rates. MCMC is often too slow for such big data problems and practitioners are replacing MCMC with more scalable approximate methods such as Variational Inference (Blei et al., 2017), Approximate Bayesian Computation (Marin et al., 2012) and Integrated Nested Laplace Approximation (Rue et al., 2009).

A recent strand of the literature instead proposes methods that speed up MCMC and HMC by data subsampling, where the costly likelihood evaluation in each MCMC iteration is replaced by an estimate from a subsample of data observations (Quiroz et al., 2019a; Dang et al., 2019) or by a weighted coreset of data points found by optimization (Campbell & Broderick, 2018; Campbell & Beronov, 2019; Campbell & Broderick, 2019). Data subsampling methods require that the log-likelihood is a sum, where each term depends on a unique piece of data — a condition satisfied for independent observations or for independent subjects in longitudinal data (with potentially dependent data within each subject) — but does not hold for general time series problems.

Our paper extends the applicability of previously proposed subsampling methods to stationary time series. The method is based on using the Fast Fourier Transform (FFT) to evaluate the likelihood function in the frequency domain for the periodogram data. The advantage of working in the frequency domain is that under quite general conditions the periodogram observations are known to be asymptotically independent and exponentially distributed with scale equal to the spectral density. The logarithm of this so called *Whittle likelihood* approximation of the likelihood is therefore a sum even when the data are dependent in the time domain. The asymptotic nature of the Whittle likelihood makes it especially suitable here since subsampling tends to be used for large-scale problems where the Whittle likelihood is expected to be accurate. Moreover, our algorithm can also be used with the recently proposed Debiased Whittle Likelihood (Sykulski et al., 2019) which gives better likelihood approximations for smaller datasets.

It is by now well established that efficient subsampling MCMC methods require likelihood estimators with low variance (Quiroz et al., 2019a;b). Variance reduction is typically achieved by using control variates that approximate the individual log-likelihood terms, often by assuming that these terms are approximately quadratic around a reference value. Our second contribution proposes a grouping strategy that makes the grouped log-likelihood terms more quadratic compared to that of the individual log-likelihood terms. In addition, when the number of observation in each group is large, we propose to use the coreset construction of Campbell & Broderick (2018) to approximate the grouped log-likelihood. This is advantageous when the grouped log-likelihood terms are not approximately quadratic. The new control variate approach thus confers additional robustness to the method, and is of independent interest beyond the time series setting.

[1]UNSW Sydney [2]University of Technology Sydney [3]Stockholm University [4]Linköping University [5]University of Sydney. Correspondence to: Robert Salomone <r.salomone@unsw.edu.au>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

The structure of our paper is as follows. Section 2 introduces the necessary frequency domain concepts and defines the Whittle likelihood. Section 3 gives an overview of the Subsampling MCMC approach of Quiroz et al. (2019a). Section 4 introduces our novel control variate schemes. Section 5 summarizes the results of experiments on examples of models that have previously not been feasible with large data methods, such as long memory stochastic volatility models.

2. DATA SUBSAMPLING USING THE WHITTLE LIKELIHOOD

2.1. Discrete Fourier Transformed Data

Let $\{X_t\}_{t=1}^n$ be a covariance stationary zero-mean time series with covariance function $\gamma(\tau) := \mathbb{E}X_t X_{t-\tau}$ for $\tau \in \mathbb{Z}$. The *spectral density* is the Fourier transform of $\gamma(\tau)$ (Lindgren, 2012)

$$f(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma(\tau) \exp(-\mathrm{i}\omega\tau), \tag{1}$$

where $\omega \in (-\pi, \pi]$ is called the *angular frequency*. The *discrete Fourier Transform* (DFT) of $\{X_t\}_{t=1}^n$ is the complex valued series

$$J(\omega_k) := \frac{1}{\sqrt{2\pi}} \sum_{t=1}^{n} X_t \exp(-\mathrm{i}\omega_k t), \tag{2}$$

for $\omega_1, \ldots, \omega_n$ in the set of Fourier frequencies

$$\Omega = \{2\pi k/n, \text{ for } k = -\lceil n/2 \rceil + 1, \ldots, \lfloor n/2 \rfloor\}.$$

The DFT is efficiently computed by the Fast Fourier Transform (FFT). The *periodogram* $\mathcal{I}(\omega_k) := n^{-1}|J(\omega_k)|^2$ is an asymptotically unbiased estimate of $f(\omega_k)$.

2.2. Frequency Domain Asymptotics

The DFT in (2) is a linear transformation that acts like a weighted average of time-domain data. A central limit theorem can therefore be used to prove that $J(\omega_k)$ are asymptotically independent complex Gaussian under quite general conditions (Shao et al., 2007, Corollary 2.1); see also Peligrad et al. (2010, Theorem 2.1) who establish the result for $\omega \in (0, 2\pi)$ almost-everywhere under even weaker conditions.

Furthermore, the real and imaginary parts are asymptotically independent. Denote a chi-squared random variable with r degrees of freedom as χ^2_r. The scaled periodogram ordinate $\mathcal{I}(\omega_k)/f(\omega_k)$ is asymptotically distributed as $\chi^2_2/2$ (i.e. standard exponential) for all $k \neq 0, n$; and as χ^2_1 for $k = 0$ and $k = n$. Hence, we have the following asymptotic distribution of the periodogram

$$\mathcal{I}(\omega_k) \sim \mathsf{Exp}(f(\omega_k)), \quad k = 1, \ldots, n-1 \tag{3}$$

independently as $n \to \infty$, with the exponential distribution in the scale parameterization, i.e., parameterized by its mean.

2.3. The Whittle Likelihood

The asymptotic distribution of the periodogram ordinates in (3) motivates the Whittle log-likelihood (Whittle, 1953) for a time series model with parameter vector $\boldsymbol{\theta}$:

$$\ell_W(\boldsymbol{\theta}) = -\sum_{k=1}^{\lfloor (n-1)/2 \rfloor} \Big(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{\mathcal{I}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} \Big), \tag{4}$$

where $f_{\boldsymbol{\theta}}(\omega)$ is the spectral density of the model. For real-valued data, both $f_{\boldsymbol{\theta}}(\omega_k)$ and $\mathcal{I}(\omega_k)$ are symmetric about the origin, so the Whittle log-likelihood is evaluated by summing only over the non-negative frequencies; for demeaned data, the term for $\omega_k = 0$ is removed from the likelihood as then $J(\omega_k) = 0$.

The Whittle log-likelihood has several desirable properties that enable scalable Bayesian inference:

- The periodogram does not depend on the parameter vector $\boldsymbol{\theta}$ and can therefore be computed before the MCMC at a cost of $\mathcal{O}(n \log n)$ via the Fast Fourier Transform algorithm. After this one-time cost, likelihood evaluations have the same $\mathcal{O}(n)$ cost as for independent data.
- The Whittle log-likelihood is a sum in the frequency domain and is therefore amenable to subsampling using the *same* algorithms developed for independent data in the time domain.
- As the Whittle log-likelihood relies on large sample properties of the periodogram, it is particularly suited to large datasets where subsampling MCMC and related methods are used.

Below, the term log-likelihood refers to the Whittle log-likelihood, and n to be the number of unique summands in (4) — i.e., we assume the FFT has been performed and we are working in the frequency domain.

3. Subsampling MCMC

The fundamental idea in the previous section can be used to extend *any* existing method for subsampling that requires conditionally independent data to the case of fitting a parametric stationary time series model with known spectral density. To provide a proof-of-concept in the form of numerical experiments, we focus on the Subsampling MCMC approach of Quiroz et al. (2019a), as it has been shown to give more accurate posterior inferences than other approaches such as Stochastic Gradient Langevin Dynamics

(SGLD) (Welling & Teh, 2011) and Stochastic Gradient Hamiltonian Monte Carlo (SG-HMC) (Chen et al., 2014)), see for example Dang et al. (2019).

We also propose novel control variates that we use with Subsampling MCMC which are also likely to be useful in further improving SG-HMC and SGLD.

3.1. MCMC with an estimated likelihood

Let $\pi(\boldsymbol{\theta}) \propto L_n(\boldsymbol{\theta})p(\boldsymbol{\theta})$ denote the posterior distribution from a sample of n observations with likelihood function $L_n(\boldsymbol{\theta})$. MCMC and HMC algorithms sample iteratively from $\pi(\boldsymbol{\theta})$ by proposing a parameter vector $\boldsymbol{\theta}^{(j)}$ at the jth iteration and accepting it with probability

$$\min\left\{1, \frac{L_n(\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)})}{L_n(\boldsymbol{\theta}^{(j-1)})p(\boldsymbol{\theta}^{(j-1)})}\cdot\frac{g(\boldsymbol{\theta}^{(j-1)}|\boldsymbol{\theta}^{(j)})}{g(\boldsymbol{\theta}^{(j)}|\boldsymbol{\theta}^{(j-1)})}\right\}, \quad (5)$$

where $g(\cdot|\cdot)$ is the proposal distribution. Repeated evaluations of the likelihood in the acceptance probability are costly when n is large. Quiroz et al. (2019a) propose speeding up MCMC for large n by replacing $L_n(\boldsymbol{\theta})$ with an estimate $\widehat{L}(\boldsymbol{\theta}, \mathbf{u})$ based on a small random subsample of $m \ll n$ observations, where $\mathbf{u} = (u_1, ..., u_m)$ indexes the selected observations.

Their algorithm samples $\boldsymbol{\theta}$ and $\mathbf{u}$ jointly from an extended target distribution $\tilde{\pi}(\boldsymbol{\theta}, \mathbf{u})$. Andrieu et al. (2009) prove that such *pseudo-marginal MCMC* algorithms sample from the full-data posterior $\pi(\boldsymbol{\theta})$ if the likelihood estimator is unbiased; i.e., $\mathbb{E}_{\mathbf{u}}\widehat{L}(\boldsymbol{\theta}, \mathbf{u}) = L(\boldsymbol{\theta})$.

Quiroz et al. (2019a) use an unbiased estimator of the log-likelihood $\widehat{\ell}(\boldsymbol{\theta}, \mathbf{u})$ and subsequently debias $\exp(\widehat{\ell}(\boldsymbol{\theta}, \mathbf{u}))$ to estimate the full-data likelihood. Although the debiasing approach in general cannot remove all bias, their pseudo-marginal sampler is still a valid MCMC algorithm, targeting a slightly perturbed posterior which is shown to be within $O(n^{-1}m^{-2})$ distance in total variation norm of the true posterior. See Quiroz et al. (2018b) for an alternative *completely* unbiased likelihood estimator and Dang et al. (2019) for an HMC extension.

3.2. Estimators based on control variates

Assume that the log-likelihood decomposes as a sum $\ell(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ell_k(\boldsymbol{\theta})$; either by assuming independent data or by using the Whittle likelihood in the frequency domain for temporally dependent data. A naive estimator of the log-likelihood is

$$\widehat{\ell}_{\text{naive}}(\boldsymbol{\theta}) := \frac{n}{m}\sum_{i=1}^{m} \ell_{u_i}(\boldsymbol{\theta}),$$

where $u_1, \dots, u_m \overset{\text{iid}}{\sim} \mathsf{Unif}(\{1, \dots, n\})$ and we suppress dependence on $\boldsymbol{u}$ in the notation for $\widehat{\ell}$ for notational clarity. This estimator typically has large variance and is prone to occasional gross overestimates of the likelihood causing the MCMC sampler to become stuck for extended periods and thus become very inefficient.

Quiroz et al. (2019a) propose using a control variate to reduce the variance in the so-called *difference estimator*

$$\widehat{\ell}_{\text{diff}}(\boldsymbol{\theta}) := \sum_{k=1}^{n} q_k(\boldsymbol{\theta}) + \frac{n}{m}\sum_{i=1}^{m}\Big(\ell_{u_i}(\boldsymbol{\theta}) - q_{u_i}(\boldsymbol{\theta})\Big), \quad (6)$$

where $u_1, \dots, u_m \overset{\text{iid}}{\sim} \mathsf{Unif}(\{1, \dots, \mathsf{n}\})$. The $q_k(\boldsymbol{\theta})$ is the control variate for the kth observation. It is evident from (6) that the variance of $\widehat{\ell}_{\text{diff}}$ is small when the $q_k(\boldsymbol{\theta})$ approximates $\ell_k(\boldsymbol{\theta})$ well. Quiroz et al. (2019a) follow Bardenet et al. (2017) and use a second order Taylor expansion of $\ell_k(\boldsymbol{\theta})$ around some central value $\boldsymbol{\theta}^\star$ as the control variate:

$$\begin{aligned} q_k(\boldsymbol{\theta}) := \ell_k(\boldsymbol{\theta}^\star) &+ \nabla_{\boldsymbol{\theta}}\ell_k(\boldsymbol{\theta}^\star)^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\ &+ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \nabla^2_{\boldsymbol{\theta}}\ell_k(\boldsymbol{\theta}^\star)(\boldsymbol{\theta} - \boldsymbol{\theta}^\star). \end{aligned}$$

One advantage of this control variate is that the otherwise $O(n)$ term $\sum_{i=1}^{n} q_i(\boldsymbol{\theta})$ can be computed at $O(1)$ cost; see Bardenet et al. (2017).

3.3. Block Pseudo-Marginal Sampling

The acceptance probability in (5) reveals that it is actually the variability of the *ratio* of estimates at the proposed and current draw that matters for MCMC efficiency (Deligiannidis et al., 2018). Tran et al. (2016) propose a blocked pseudo-marginal scheme for subsampling that partitions the indicators in B blocks $\mathbf{u} = (\mathbf{u}_1, ..., \mathbf{u}_B)$ and only updates one of the blocks in each MCMC iteration. Under simplifying assumptions, Lemma 2 in Tran et al. (2016) shows that blocking induces a controllable correlation between subsequent estimates in the MCMC of the simple form

$$\text{Corr}\big(\widehat{\ell}(\boldsymbol{\theta}^{(j)}), \widehat{\ell}(\boldsymbol{\theta}^{(j-1)})\big) \approx 1 - 1/B.$$

4. Alternative Control Variates via Grouping

The control variates presented in Section 3.2 are only good approximations if the *individual* log-likelihood terms, $\ell_k(\boldsymbol{\theta})$, are approximately quadratic or $||\boldsymbol{\theta} - \boldsymbol{\theta}^\star||$ is sufficiently small. We propose two new control variates that may be preferable when this is not the case.

4.1. Grouped Quadratic Control Variates

Rather than sampling individual observations, we can sample observations in *groups*. The advantage of sampling groups is that the quadratic control variates are expected to be more accurate for the group as a whole compared to individual observations. The reason is that the Bernstein-von Mises theorem (asymptotic normality of the posterior) suggests an approximately quadratic log-likelihood for the group provided the number of observations in the group is large enough; see Tamaki (2008) for a Bernstein-von Mises theorem specifically for the Whittle likelihood.

Let $\mathcal{G}$ be a partition of the set of indices $\mathcal{U} = \{1, ..., n\}$ into $|\mathcal{G}|$ groups $G_1, \ldots, G_{|\mathcal{G}|}$; i.e. $\mathcal{U} = \cup_{k=1}^{|\mathcal{G}|} G_k$, where G_k is the set of data indices associated with the kth group. Similarly, write

$$\ell_{G_k}(\boldsymbol{\theta}) \coloneqq \sum_{i \in G_k} \ell_i(\boldsymbol{\theta})$$

for the sum of log-likelihood terms corresponding to the observations in the kth group, noting that $\ell = \ell_{\cup_k G_k} = \sum_k \ell_{G_k}$. Since $\ell_{G_k}(\boldsymbol{\theta})$ is based on $|G_k|$ observations we expect it to be closer to a quadratic function than the $\ell_i(\boldsymbol{\theta})$ belonging to the individual samples in the group. Now, define the control variate for group G_k as

$$\begin{aligned} q_{G_k}(\boldsymbol{\theta}) \coloneqq \ell_{G_k}(\boldsymbol{\theta}^\star) &+ \nabla_{\boldsymbol{\theta}} \ell_{G_k}(\boldsymbol{\theta}^\star)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^\star) \\ &+ \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^\star)^\top \nabla^2_{\boldsymbol{\theta}} \ell_{G_k}(\boldsymbol{\theta}^\star)(\boldsymbol{\theta} - \boldsymbol{\theta}^\star), \end{aligned}$$

where the same $\theta^\star$ is used for all groups. The *grouped difference estimator* for a sample of m groups is then

$$\widehat{\ell}_{\text{gr}}(\boldsymbol{\theta}) \coloneqq \sum_{k=1}^{|\mathcal{G}|} q_{G_k}(\boldsymbol{\theta}) + \frac{|\mathcal{G}|}{m} \sum_{i=1}^{m} \left(\ell_{G_{u_i}}(\boldsymbol{\theta}) - q_{G_{u_i}}(\boldsymbol{\theta}) \right), \tag{7}$$

where $u_1, \ldots, u_m \overset{\text{iid}}{\sim} \mathsf{Unif}(\{1, \ldots, |\mathcal{G}|\})$.

4.2. Grouped Coreset Control Variates

When the grouped log-likelihoods are far from quadratic, we propose an alternative method using *Bayesian Coresets* (Huggins et al., 2016) to construct control variates. An advantage of this approach is that it is unnecessary to select a central point $\boldsymbol{\theta}^\star$, or to rely on a quadratic expansion function which may be unsuitable.

Bayesian Coresets replace the true log-likelihood $\ell(\boldsymbol{\theta})$ with the approximation $\ell_C(\boldsymbol{\theta}) = \sum_{k=1}^n w_k \ell_k(\boldsymbol{\theta})$, where $\boldsymbol{w}$ is a *sparse* vector with a number of non-zero elements that is much less than n. Let $\widehat{\pi}$ denote some *weighting distribution* that has the same support as the posterior and can easily be sampled from — for example a Gaussian based on Laplace approximation, or the empirical distribution of samples from an MCMC on a smaller data set. The log-likelihood approximation ℓ_{C} is constructed using a greedy algorithm, which, after M steps, provides an approximate solution to

$$\text{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} \left\{ \mathbb{E}_{\widehat{\pi}} \left[\left(\ell(\boldsymbol{\theta}) - \ell_C(\boldsymbol{\theta}) \right)^2 \right] \right\}$$

subject to the constraints that $w_i \geq 0$ for $i = 1, \ldots, n$, and $\sum_{k=1}^n \mathbb{I}\{w_k > 0\} \leq M$, where M is a user-specified number of iterations in the coreset optimization procedure. We use the *Greedy Iterative Geodesic Ascent* (GIGA) method in Campbell & Broderick (2018) for tackling the optimization.

We propose approximating the group log-likelihoods $\ell_{G_k}(\boldsymbol{\theta}), k = 1, \ldots, |\mathcal{G}|$, by a separate coreset approximation for each group. We can use the grouped difference estimator in (7) with coreset approximations as control variates for each respective group. The coreset control variates are attractive as *by design* they approximate $\ell_{G_k}(\boldsymbol{\theta})$ well for each group using less density evaluations than the number of observations in the group. While the construction of coreset control variates requires $|\mathcal{G}|$ runs of the coreset procedure, each is only on a group of the dataset, so the overall effort is roughly that of the standard coreset approach, or less if run in parallel.

4.3. Perturbation of Subsampled Whittle Posterior

In this section, we present a result regarding the perturbation of the subsampled Whittle posterior to the exact Whittle posterior.

Let $\pi_n(\boldsymbol{\theta}) \propto L_n(\boldsymbol{\theta}) p(\boldsymbol{\theta})$ be the posterior based on the Whittle likelihood $L_n(\boldsymbol{\theta}) = \exp(\ell_n(\boldsymbol{\theta}))$ with n samples. Following Quiroz et al. (2019a) we define $\overline{\pi}_{n,m}(\boldsymbol{\theta}, \boldsymbol{u})$ as the target for $\boldsymbol{\theta}$ extended with the m (group) subsample indicators $\boldsymbol{u}$, and use MCMC to sample $\boldsymbol{\theta}$ and $\boldsymbol{u}$ jointly from the extended target. The MCMC algorithm produces valid draws from the marginal $\overline{\pi}_{n,m}(\boldsymbol{\theta}) = \int \overline{\pi}_{n,m}(\boldsymbol{\theta}, \boldsymbol{u}) d\boldsymbol{u}$. Note that $\overline{\pi}_{n,m}(\boldsymbol{\theta}) \neq \pi_n(\boldsymbol{\theta})$ as with the methodology in Quiroz et al. (2019a), it is not possible to eliminate all bias in $\exp(\ell_n(\boldsymbol{\theta}))$.

However, as a direct consequence of Quiroz et al. (2019a, Theorem 1), we have the following lemma, which shows that the perturbation error decreases rapidly.

Lemma 1 *Suppose that the regularity conditions discussed in the supplement are satisfied, and that the control variates in Section 4.1 are used, with the number of groups m depending on n in a manner such that $m(n) \to \infty$ as $n \to \infty$. Then,*

$$\int_{\boldsymbol{\theta}} \left| \overline{\pi}_{n,m(n)}(\boldsymbol{\theta}) - \pi_n(\boldsymbol{\theta}) \right| d\boldsymbol{\theta} = O\left(\frac{1}{nm(n)^2} \right).$$

Moreover, for any scalar-valued function h that satisfies $\limsup_{n \to \infty} \mathbb{E}_{\pi_n}[h^2(\boldsymbol{\theta})] < \infty$

$$\left| \mathbb{E}_{\overline{\pi}_{n,m(n)}} h(\boldsymbol{\theta}) - \mathbb{E}_{\pi_n} h(\boldsymbol{\theta}) \right| = O\left(\frac{1}{nm(n)^2} \right).$$

We highlight that the required regularity conditions are essentially those of Quiroz et al. (2019a) but where the posterior density has (4) as a (log)-likelihood function. Thus, these conditions are justified by results such as the asymptotic normality of the maximum Whittle-likelihood estimator (Fox & Taqqu, 1986) and the Bernstein-von Mises theorem for the Whittle measure (Tamaki, 2008).

5. EXPERIMENTS

5.1. Settings and Performance Measures

There are many ways to partition the dataset into groups for the control variates. We use the same number of samples for each group. The kth group is chosen by starting with the kth lowest frequency and then systematically sampling every $|\mathcal{G}|$ frequency after that. This way we ensure that each group contains periodogram ordinates across the entire frequency range. The homogeneity of the groups makes it possible to use the same $\boldsymbol{\theta}^\star$ for all groups.

We use the Laplace approximation of the posterior as weighting function in the coreset approximation, truncated to the region of admissible parameters. Each coreset is fitted using the GIGA algorithm for $M = 200$ iterations, using 500 random projections; see Campbell & Broderick (2018) for details. We note that the mode used in the Laplace approximation comes with no extra cost compared to full-data MCMC as the latter uses the mode as a starting value for the sampler and to build the covariance matrix of the random walk Metropolis proposal. Likewise, the Taylor control variates are constructed using the mode as $\boldsymbol{\theta}^\star$. Therefore, this part of the start-up cost is assumed to be the same for all algorithms. However, both control variates have additional start-up costs compared to full-data MCMC. The coreset control variate needs to perform the GIGA optimization, which makes M sweeps of the full dataset, using Mn density evaluations. As discussed above, this can in practice be done in parallel for each group, where each group uses $M|G_k|$ observations. Recall that $n = \sum_k |G_k|$, which explains the cost of Mn density evaluations. The Taylor control variate requires summing all the q_k once (first term in (6)), hence adding n to the total cost. For simplicity, assume all groups have the same number of observations $|G| = |G_k|$. During run time, full-data MCMC requires n density evaluations in each iteration, whereas the Taylor control variate uses $m|G|$ and the coreset control variate $|m|G + \sum_{k=1}^{|\mathcal{G}|} g_k$, where the second term is the cost of evaluating the summation of all $q_{G_k}(\boldsymbol{\theta})$, which is much faster than full-data MCMC if the coreset size g_k is small in relation to $|G_k|$.

We follow Quiroz et al. (2019a) and use the *computational time* (CT) as our measure of performance. This measure balances the cost (number of density evaluations as discussed above) and the efficiency of the Markov chain. It is defined as

$$\text{CT} := \text{IF} \times \text{number of density evaluations},$$

where the inefficiency factor (IF) is proportional to the asymptotic variance when estimating a univariate posterior mean based on MCMC output. The IF is interpreted as the number of (correlated) samples needed to obtain the equivalent of a single independent sample. It is convenient to measure the cost using density evaluations since it makes the comparisons implementation independent. We use the CODA package (Plummer et al., 2006) in R to estimate IF. Our measure of interest is the *relative* CT (RCT) which we define as the ratio between the CT of full-data MCMC and that of the subsampling algorithm of interest. Hence, values larger than one mean that the subsampling algorithm is more efficient when balancing computing cost (density evaluations) and statistical efficiency (variance of the posterior mean estimator).

5.2. Experiments

We consider several time series models for large data in our experiments, including the recently proposed class of autoregressive tempered fractionally integrated moving average (ARTFIMA) models and several of its widely-used special cases such as ARMA and ARFIMA — though we highlight that any model class for which the spectral density is known can be used. We also consider a stochastic volatility model with an underlying ARTFIMA process.

Sabzikar et al. (2019) defines Y_t as an $\text{ARTFIMA}(p, d, \lambda, q)$ process if

$$\phi_q(L)\Delta^{d,\lambda}(Y_t - \mu) = \theta_p(L)\varepsilon_t,$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an iid sequence of zero mean random variables with variance σ^2, $\phi_p(L) := 1 - \phi_1 L - \cdots - \phi_p L^p$, and $\theta_q(L) := 1 + \theta_1 L + \cdots + \theta_q L^q$, are the autoregressive and moving average lag polynomials, where L is the lag operator, i.e. $L^k(X_t) = X_{t-k}$. The *tempered fractional differencing operator* is defined by

$$\begin{aligned}\Delta^{d,\lambda}Y_t &:= (1 - e^{-\lambda}L)^d Y_t \\ &= \sum_{j=0}^{\infty}(-1)^j \frac{\Gamma(1+d)}{\Gamma(1+d-j)j!} e^{-\lambda j} Y_{t-j},\end{aligned}$$

where d is called the *fractional integration parameter* and λ is called the *tempering parameter*. To explain the role of the parameters d and λ, note that for $\lambda = 0$ and d a non-negative integer, $\Delta^{d,\lambda}Y_t$ reduces to simple differencing of order d and we obtain the AutoRegressive Integrated Moving Average (ARIMA) processes. Autoregressive Fractionally Integrated Moving Average (ARFIMA) (Granger & Joyeux, 1980) extends this class by allowing fractional differences, i.e. d need not be an integer. For $-0.5 < d < 0.5$, ARFIMA is stationary, and is of particular interest as it has long-range or long-memory dependence with an autocovariance function that dies off so slowly it is not absolutely summable, $\sum_{k=-\infty}^{\infty} |\gamma(k)| = \infty$. The tempering parameter $\lambda > 0$ in ARTFIMA allows for semi-long range dependence, i.e. ARFIMA-like long-range dependence for a number of lags beyond which the autocovariance decays exponentially fast.

Provided that $\lambda > 0$, $d \notin \mathbb{Z}$, and the roots of $\phi_p(z)$ lie outside of the unit circle in the complex plane, the ARTFIMA process is stationary (Sabzikar et al., 2019, Theorem 2.2) with spectral density

$$f(\omega) = \frac{\sigma^2}{2\pi} \left|1 - e^{-(\lambda+i\omega)}\right|^{-2d} \left|\frac{\theta_q(e^{-i\omega})}{\phi_p(e^{-\lambda i\omega})}\right|^2 . \quad (8)$$

We follow Barndorff-Nielsen & Schou (1973) and reparameterize the autoregressive parameters, ϕ_k for $k = 1, \ldots, p$, in terms of the *partial autocorrelations* $\widetilde{\boldsymbol{\phi}}_p = (\widetilde{\phi}_1, \ldots, \widetilde{\phi}_q)$. Stationarity can now be enforced by the conditions $|\widetilde{\phi}_k| < 1$, for $k = 1, \ldots, p$. We perform the same reparameterization to $\boldsymbol{\theta}_q$ to obtain $\widetilde{\boldsymbol{\theta}}_q$, which ensures the underlying process is invertible provided that the constraint $|\widetilde{\theta}_k| < 1$ for $k = 1, \ldots, q$ is satisfied. We use the priors $\widetilde{\boldsymbol{\phi}}_p \sim \mathsf{Unif}\big((-1,1)^p\big)$ and $\tilde{\boldsymbol{\theta}}_q \sim \mathsf{Unif}\big((-1,1)^q\big)$

A log-transformation is used for both σ^2 and λ, with both priors $\log(\sigma^2), \log(\lambda) \sim \mathcal{N}(0,1)$. For ARFIMA models, the fractional integration parameter d is parametrized by a scaled Fisher transformation $\widetilde{d} := \operatorname{arctanh}(2d)$ with prior $\widetilde{d} \sim \mathcal{N}(0,1)$, which is a weakly informative on $d = 0.5\tanh(\widetilde{d})$ in the region $\left(-\frac{1}{2}, \frac{1}{2}\right)$. For ARTFIMA models ($d$ not restricted to $(-0.5, 0.5)$) we set $d \sim \mathcal{N}(0,1)$.

Example 1: Vancouver Temperatures (ARMA)

The first example considers the ARMA model for hourly temperature data for the city of Vancouver during the years 2012 to 2017 sourced from `openweathermap.org`. Using the relatively simple ARMA model makes it possible to compare with the posterior obtained by MCMC using the exact time domain likelihood from the Kalman filter. We use the stl function in R to remove the trend and yearly seasonal from the original series. We then confirm that the series passes the Augmented Dickey-Fuller and Phillips-Perron unit-root tests for stationarity. The final time series is of length $n = 44001$, yielding a likelihood with $n = 2.2 \times 10^4$ frequency terms. The auto.arima function in the Forecast (Hyndman & Khandakar, 2008) package in R was used for model selection, yielding an $\mathrm{ARMA}(2,3)$ model.

Example 2: Stockholm Temperatures (ARTFIMA)

The second example fits an ARTFIMA(2,2) model to hourly temperature data for the city of Stockholm during the years 1967 to 2018, obtained from the Swedish Meteorological and Hydrological Institute (`www.smhi.se/en`). The data preparation procedure is the same as in the last example. The series is of length $4.5 \times 10^5 + 1$ yielding 2.25×10^5 terms in the Whittle likelihood.

Example 3: Simulated ARMA Series (ARFIMA)

To assess the performance of our methods on even larger datasets, we simulate an ARMA(2,1) series of length $n = 5 \times 10^6 + 1$, and fit an ARFIMA(2,1) model. The parameters of the data generating process are $\boldsymbol{\phi} = (0.22, -0.1)$, $\boldsymbol{\theta} = (0.5)$, and $\sigma^2 = 1$.

Example 4: Bitcoin Prices Stochastic Volatility (ARTFIMA-SV)

Finally, we test our methods on a challenging ARTFIMA model extended with stochastic volatility. A general class of Stochastic Volatility (SV) models is

$$y_t = \exp(v_t/2)\xi_t, \quad (9)$$

where $\{\xi_t\}$ is an independent and identically distributed sequence having mean zero and unit variance, and v_t is a stationary process with parameter vector $\boldsymbol{\psi}$ and spectral density $f_v(\omega; \boldsymbol{\psi})$. Breidt et al. (1998) observe that the SV model in (9) can be estimated by noting that

$$\log y_t^2 = \mu + v_t + \varepsilon_t, \quad (10)$$

where $\mu \in \mathbb{R}$, and $\{\varepsilon_t\}$ is independent and identically distributed white noise process with zero mean and variance σ_ε^2. Thus, as the spectral density of ε_t is $f_\varepsilon(\omega) = \sigma_\varepsilon^2/2\pi$, fitting a parametric spectral density of the form

$$f(\omega; \boldsymbol{\psi}, \sigma_\epsilon^2) = f_v(\omega; \boldsymbol{\psi}) + \frac{\sigma_\epsilon^2}{2\pi},$$

to the log-squared series is equivalent to fitting the model in (9) to the original series. We set $\log\sigma_\varepsilon^2 \sim \mathcal{N}(0, 0.01)$ a priori.

We let v_t be an ARTFIMA(1,1) process, which generalizes the Long Memory Stochastic Volatility model of Breidt et al. (1998) to allow for tempered fractional differencing. This is a challenging model to estimate even without tempering since the computational cost of filtering to obtain the Gaussian ARFIMA likelihood scales poorly with the length of the time series (Chan & Palma, 1998). Tempering introduces additional difficulty as the ARTFIMA covariance function involves infinite sums involving the Gaussian hypergeometric function (Sabzikar et al., 2019, Equation (2.9)). We show that spectral subsampling MCMC is a computationally efficient way to obtain the posterior of the ARTFIMA-SV model for large datasets. We fit the model to a dataset of one-minute Bitcoin returns (prices from the exchange on `coinbase.com`) of length $n = 10^6 + 1$ (resulting in $n = 5 \times 10^5$ terms in the Whittle likelihood).

	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
n $(\times 10^4)$	2.2	22.5	250	50
$\lvert\mathcal{G}\rvert$ $(\times 10^3)$	1	1	1	1
$\lvert G\rvert$	22	225	2500	500
m $(\%\lvert\mathcal{G}\rvert)$	2	1	1	1
B	10	10	10	10
M	-	200	200	200
RP	-	500	500	500
$\bar{g}_k$	-	29	34	39

Table 1. Settings for models ARMA(2,3) ($\mathcal{M}_1$), ARTFIMA(2,2) ($\mathcal{M}_2$), ARFIMA(2,1) ($\mathcal{M}_3$) and ARTFIMA-SV(1,1) ($\mathcal{M}_4$). The table shows number of frequency observations (n), number of groups ($|\mathcal{G}|$), number of observations per group ($|G|$, the same for all groups), percentage of subsampled groups (m) and number of blocks in the block pseudo-marginal algorithm (B). The coreset settings are number of iterations of GIGA algorithm (M), number of random projections (RP) (see Campbell & Broderick (2018) for details) and the average size of the coreset ($\bar{g}_k$).

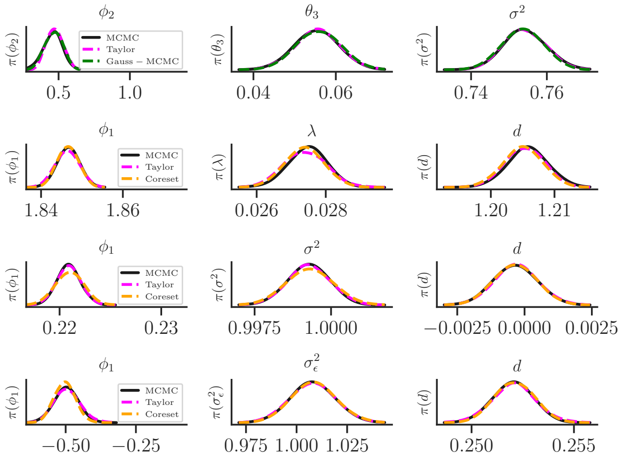

Figure 1. Kernel Density Estimates of some marginal distributions (see the supplementary for all parameters) for all examples. Each row corresponds to a single model (from the top: ARMA(2,3), ARTFIMA(2,2), ARFIMA(2,1) and ARTFIMA-SV(1,1)). MCMC is the full-data MCMC on the Whittle likelihood. Taylor and coreset are the Subsampling MCMC methods using the corresponding control variates. Gauss-MCMC is the full-data MCMC with the time-domain Gaussian likelihood.

5.3. Results

Table 1 displays the settings for each example. Note that only for the ARMA(2,3) model is it computationally feasible to compare with the posterior based on the exact time domain likelihood. Furthermore, the number of observations per group in the Vancouver temperature data is too small (22) for the coreset control variates to be useful.

Figure 1 displays a selection of kernel density estimates for marginal posteriors across all examples. Note that the incurred bias from subsampling is neglible, especially for the Taylor series control variate. The coreset control variate results in a higher variance and thus a larger perturbation error (Quiroz et al., 2019a). Figure 1 also shows that the posterior based on the Whittle likelihood is close to the Gaussian time-domain posterior for the ARMA(2,3) example. We stress that we can not make this comparison for the other models because, as discussed in Section 5.2, the Gaussian time-domain likelihood is not computationally feasible for large n for the ARFIMA and ARTFIMA models. Further, no such comparison can be made in the ARTFIMA-SV case as the model is non-Gaussian (Breidt et al., 1998).

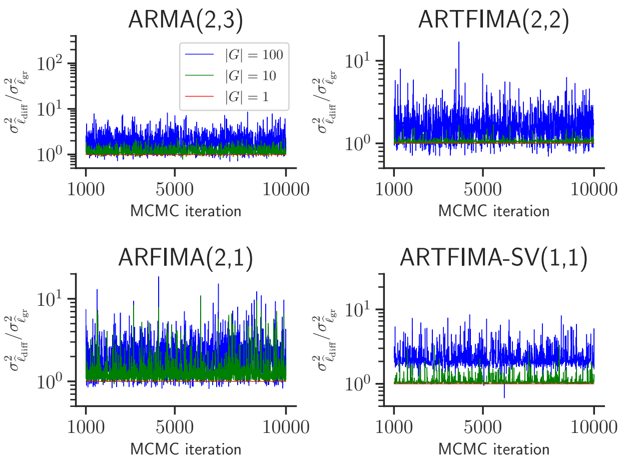

Figure 2. Effect of grouping the Taylor series control variate for all examples. The figure shows the relative variance reduction (in log-scale) with no grouping (i.e. one observation per group $G| = 1$) as baseline over the MCMC iterations post burn-in. A value larger than 1 means that the corresponding control variate is more efficient compared to the baseline.

Figure 2 shows that in general the use of grouping in the control variates reduces the variance, especially for the more complex models. This experiment is performed using the last 2001 time observations for each model ($n = 1000$ in the frequency domain) to prevent $||\boldsymbol{\theta} - \boldsymbol{\theta}^\star||$ becoming too small.

Figure 3 reports the relative computational time with MCMC on the full-data Whittle likelihood as baseline for all parameters — showing that our method introduces close to two orders of magnitude speedup on these examples.

Finally, Figure 4 plots the periodogram and posterior mean spectral density for all models using our Subsampling MCMC method and, for comparison, the full-data MCMC (on the Whittle likelihood) method. The figure confirms the accuracy of our method and, moreover, that we recover the true spectral density perfectly in a simulated data setting (ARFIMA(2,1)). Recall that due to (4), the fitting of a Stationary time series model is essentially a univariate regression problem for the spectral density. Thus, the figure also shows that the models we consider capture features of the real world datasets while avoiding overfitting.

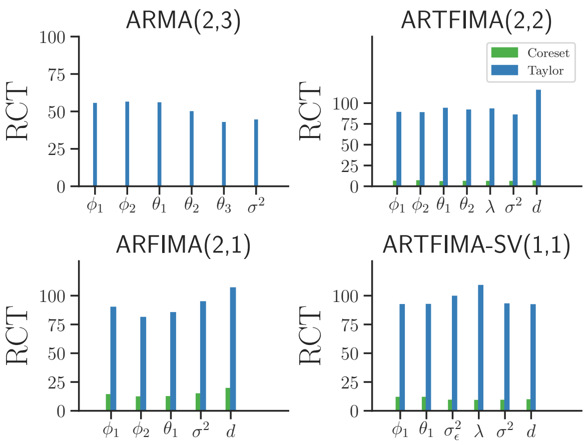

Figure 3. Relative computational time each parameter in all examples. Results are relative to full-data MCMC (larger is better for Subsampling MCMC).

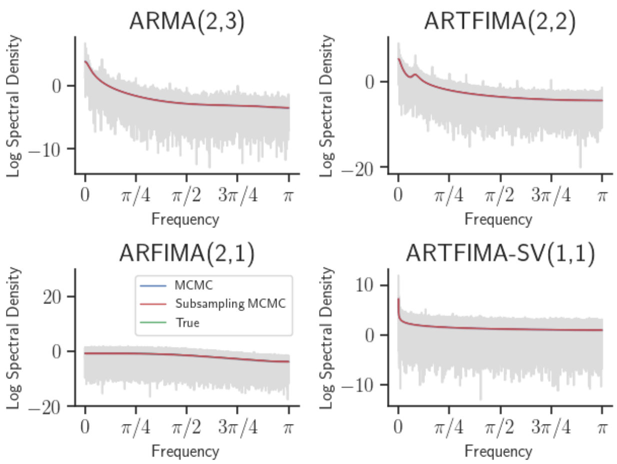

Figure 4. Periodogram data (grey) in log-scale with posterior means of the (log) spectral density estimate from MCMC on full-data Whittle likelihood and from Subsampling MCMC on the subsampled Whittle likelihood for all four models. The figure also shows the true spectral density, which is only available for the ARFIMA(2,1) example (simulated data). All posterior means are indistinguishable, showing the accuracy of our method and, moreover, the accuracy of the Whittle approximation for the ARFIMA(2,1) example.

6. Discussion

We introduce novel methods allowing efficient Bayesian inference in stationary models for large time series datasets. The idea is simple and elegant: overcome the lack of independence in the data by transforming them to the frequency domain where they are independent. This work is, to our knowledge, the first to extend scalable MCMC algorithms to models with temporal dependence. The article is focused on Subsampling MCMC, but the ideas introduced here can be directly applied to many other scalable inference approaches, e.g.,those in Quiroz et al. (2018b;a) and Cornish et al. (2019). Subsampling of periodogram frequencies also extends beyond the MCMC setting, for example to *Doubly* Stochastic Variational Inference (Titsias & Lázaro-Gredilla, 2014). We also introduce novel control variate schemes based on grouping and coresets to improve the robustness of Subsampling MCMC.

Two immediate and interesting extensions of our methods are to multiple time series and locally stationary time series; Dahlhaus et al. (2000) define a suitable analogue to the univariate Whittle likelihood for these cases. A third extension is to semi-parametric and non-parametric spectral density estimation as in for example Carter & Kohn (1997), Choudhuri et al. (2004), and Edwards et al. (2019).

More generally, we remark that development of specialized subsampling methodogy for the case where the data-generating process is not believed to be have an absolutely summable covariance function would also be interesting, e.g., based on the methods of Chopin et al. (2013), who provide an alternate likelihood with superior properties to the Whittle Likelihood in such cases.

Finally, it has not escaped our notice that our approach can be directly extended to spatial and spatio-temporal data using the multidimensional DFT (see Peligrad & Zhang (2017) for results regarding the asymptotic distribution of the Fourier transform in the spatial setting).

Acknowledgements

Salomone, Quiroz, and Kohn were partially supported by the Australian Centre of Excellence for Mathematical and Statistical Frontiers, under Australian Research Council grant CE140100049. Villani was supported by the Swedish Foundation for Strategic Research (Smart Systems: RIT 15-0097).

References

Andrieu, C., Roberts, G. O., et al. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

Bardenet, R., Doucet, A., and Holmes, C. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Barndorff-Nielsen, O. and Schou, G. On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis*, 3(4):408–419, 1973.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Breidt, F., Crato, N., and de Lima, P. The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics*, 83(1):325 – 348, 1998.

Campbell, T. and Beronov, B. Sparse variational inference: Bayesian coresets from scratch. *arXiv preprint arXiv:1906.03329*, 2019.

Campbell, T. and Broderick, T. Bayesian coreset construction via greedy iterative geodesic ascent. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 698–706, 2018.

Campbell, T. and Broderick, T. Automated scalable Bayesian inference via Hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.

Carter, C. K. and Kohn, R. Semiparametric Bayesian inference for time series with mixed spectra. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):255–268, 1997.

Chan, N. H. and Palma, W. State space modeling of long-memory processes. *Annals of Statistics*, pp. 719–740, 1998.

Chen, T., Fox, E., and Guestrin, C. Stochastic gradient Hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691, 2014.

Chopin, N., Rousseau, J., and Liseo, B. Computational aspects of Bayesian spectral density estimation. *Journal of Computational and Graphical Statistics*, 22(3):533–557, 2013.

Choudhuri, N., Ghosal, S., and Roy, A. Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059, 2004.

Cornish, R., Vanetti, P., Bouchard-Côté, A., Deligiannidis, G., and Doucet, A. Scalable Metropolis-Hastings for exact Bayesian inference with large datasets. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pp. 1351–1360, 2019.

Dahlhaus, R. et al. A likelihood approximation for locally stationary processes. *The Annals of Statistics*, 28(6):1762–1794, 2000.

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. Hamiltonian Monte Carlo with energy conserving subsampling. *Journal of Machine Learning Research*, 20 (100):1–31, 2019.

Deligiannidis, G., Doucet, A., and Pitt, M. K. The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):839–870, 2018.

Edwards, M. C., Meyer, R., and Christensen, N. Bayesian nonparametric spectral density estimation using B-spline priors. *Statistics and Computing*, 29(1):67–78, Jan 2019.

Fox, R. and Taqqu, M. Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *The Annals of Statistics*, 14(2):517–532, 1986.

Granger, C. W. and Joyeux, R. An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis*, 1(1):15–29, 1980.

Huggins, J., Campbell, T., and Broderick, T. Coresets for scalable Bayesian logistic regression. In *Advances in Neural Information Processing Systems*, pp. 4080–4088, 2016.

Hyndman, R. and Khandakar, Y. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(3):1–22, 2008.

Lindgren, G. *Stationary stochastic processes: theory and applications*. Chapman and Hall/CRC, 2012.

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

Peligrad, M. and Zhang, N. Central limit theorem for Fourier transform and periodogram of random fields. *Bernoulli*, 25, 2017.

Peligrad, M., Wu, W. B., et al. Central limit theorem for Fourier transforms of stationary processes. *The Annals of Probability*, 38(5):2009–2022, 2010.

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

Quiroz, M., Tran, M.-N., Villani, M., and Kohn, R. Speeding up MCMC by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics*, 27 (1):12–22, 2018a.

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. The block-Poisson estimator for optimally tuned exact subsampling MCMC. *arXiv preprint arXiv:1603.08232v5*, 2018b.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, 114(526):831–843, 2019a.

Quiroz, M., Villani, M., Kohn, R., Tran, M.-N., and Dang, K.-D. Subsampling MCMC - An introduction for the survey statistician. *Sankhya A*, 80(1):33–69, 2019b.

Rue, H., Martino, S., and Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71 (2):319–392, 2009.

Sabzikar, F., McLeod, A. I., and Meerschaert, M. M. Parameter estimation for ARTFIMA time series. *Journal of Statistical Planning and Inference*, 200:129 – 145, 2019.

Shao, X., Wu, W. B., et al. Asymptotic spectral theory for nonlinear time series. *The Annals of Statistics*, 35(4): 1773–1801, 2007.

Sykulski, A. M., Guillaumin, A. P., Olhede, S. C., Early, J. J., and Lilly, J. M. The debiased Whittle likelihood. *Biometrika*, 106(2):251–266, 2019.

Tamaki, K. The Bernstein-von Mises theorem for stationary processes. *Journal of the Japan Statistical Society*, 38(2): 311–323, 2008.

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, pp. 1971–1979, 2014.

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. The block pseudo-marginal sampler. *arXiv preprint arXiv:1603.02485*, 2016.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Whittle, P. Estimation and information in stationary time series. *Arkiv för matematik*, 2(5):423–434, 1953.