Spectral Subsampling MCMC for Stationary Time Series

Robert Salomone

UNSW Sydney

r.salomone@unsw.edu.au

joint work with Matias Quiroz, Robert Kohn,

Mattias Villani & Minh-Ngoc Tran.

ICML 2020

What this talk is about...

- **Extending Bayesian "big data" methods** beyond the simple models previously considered.
- Our goal is efficient and accurate (approximate) sampling from the **posterior**

$$p(\theta|\mathbf{y}) \propto \underbrace{p(\theta)}_{\text{prior}} \underbrace{\exp(\ell(\theta))}_{\text{likelihood}},$$

  when the log-likelihood $\ell(\theta)$ is not naturally amenable to subsampling.
- **Simple solution** to a tough problem, this work is a **proof-of-concept** that spectral methods are one promising direction forward.
- Spectral subsampling approach is general, but **Subsampling MCMC** is our method of choice - where we make additional contributions to improve robustness.

Background: Subsampling

- Subsampling methods **assume** the **log-likelihood is a sum**

$$\ell(\theta) = \sum_{i=1}^{n} \log p(y_i|\theta)$$

- Estimating $\ell(\theta)$ is like estimating a **population total**

$$\hat{\ell}(\theta, \mathbf{u}) = \frac{n}{m} \sum_{i \in \mathbf{u}} \log p(y_i|\theta)$$

- **Log-likelihood is a sum:**
  - for conditionally independent y_i
  - for longitudinal data when subjects are independent.
  - for very special time series, e.g. AR processes.
- General **time series** dependence? Spatial dependence?

An simple yet elusive Solution...

- Data $\mathbf{y} = (Y_1, \ldots, Y_n)$ in the **time domain** are dependent.
- Main Idea: **Transform data to independence.**
- Discrete Fourier Transform

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

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

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

($\mathcal{O}(n \log n)$ via FFT)

- Then, the **periodogram ordinates**

$$\mathcal{I}(\omega_k) = n^{-1} \left| J(\omega_k) \right|^2, \quad \omega_k \in \Omega,$$

are **asymptotically** (in n) conditionally (on $\boldsymbol{\theta}$) independent, thus

$$\ell_W(\mathcal{I}(\boldsymbol{\omega})|\boldsymbol{\theta}) = \sum_{w_k \in \Omega} \ell_W(\mathcal{I}(\omega_k)|\boldsymbol{\theta}).$$

- **It's a sum**!

Whittle likelihood

- Let $f(\cdot)$ denote the **spectral density** (Fourier transform of autocovariance function) of a covariance-stationary data-generating process.
- Then, **asympotically** as $n \to \infty$

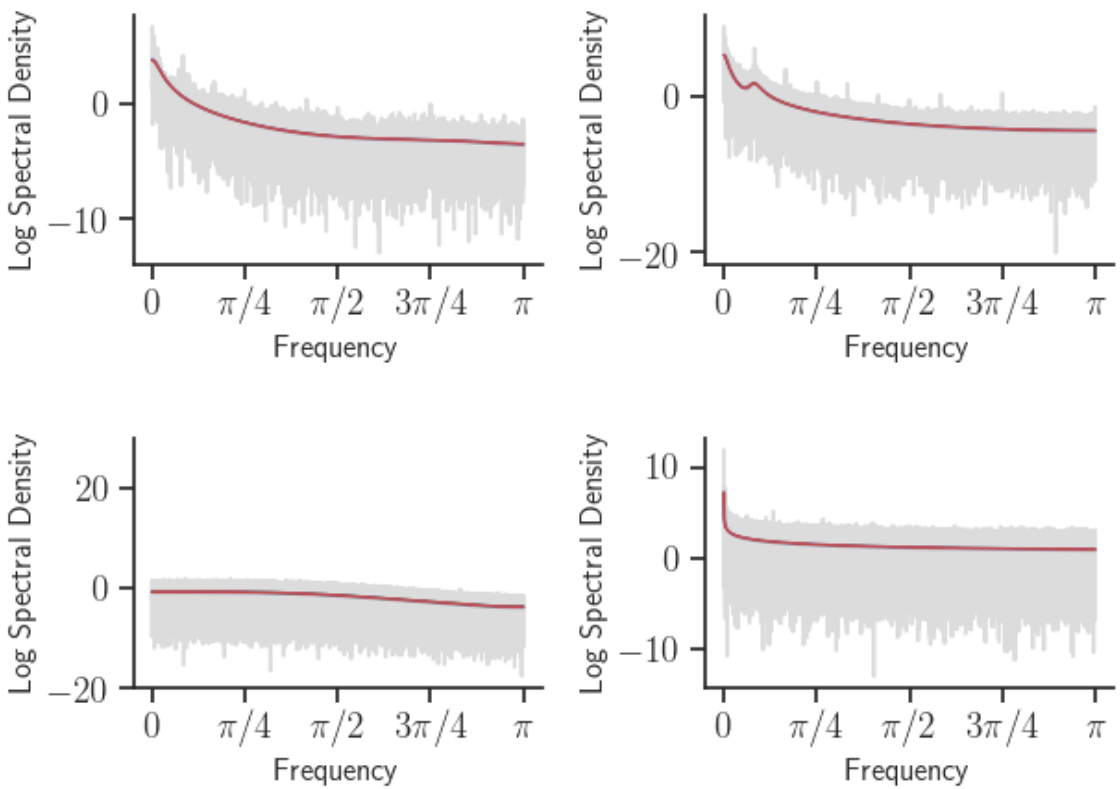
$$\mathcal{I}(\omega_k) \overset{\text{ind}}{\sim} \text{Exp}(f(\omega_k)^{-1}), \quad k = 1, \ldots, n-1.$$

- This yields a likelihood over **independent** but not identically distributed **exponentials** (each with expected value being the spectral density at its respective Fourier frequency).
- **Whittle's** log-likelihood:

$$\ell_W(\boldsymbol{\theta}) \equiv - \sum_{\omega_k \in \Omega} \left(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{\mathcal{I}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} \right)$$

- Whittle (1951) arrived at the above via different (!) arguments (Gaussian Processes, less general).

"Smoothing the Periodogram"



Subsampling for stationary time series

- Whittle is **biased for small** n, but **subsampling** is only relevant for **very large** n.

- **Subsampling** for stationary **time series** **?**
  1. **Compute periodogram** once before MCMC at cost $\mathcal{O}(n \log n)$.
  2. **Pick a model family** for which you **know the spectral density and stationary subset of parameters**.
  3. Estimate $\ell_W(\boldsymbol{\theta})$ by **subsampling frequencies** in your inference algorithm.
- We use **Subsampling MCMC**.

Subsampling Markov Chain Monte Carlo (MCMC): A Crash Course

- Subsampling MCMC[1] is a collection of techniques to **make standard MCMC work with subsampling**.
  1. Pseudo Marginal MCMC — Perform (principled) MCMC on **extended space** $\Theta \times \mathcal{U}$, where $\mathcal{U}$ is the data (frequency) subsamples.
  2. Use $\hat{L}(\theta) = \exp(\hat{\ell}(\theta))$ (biased estimator of $L(\theta)$) with a **bias correction mechanism**.
  3. **Control variates** for **reducing subsampling variance and further bias reduction** (essential for the method to work well!).
     - In our paper, we introduce **two new control variate** schemes based on grouping.

[1] Quiroz, Matias, et al., **Speeding up MCMC by efficient data subsampling**. *Journal of the American Statistical Association* 114.526 (2019): 831-843.

Numerical Experiments

A General Class of Model

- Autoregressive tempered fractionally integrated moving average (**ARTFIMA**) process.
- ARTFIMA(q, d, λ, p):

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

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is iid zero mean with variance σ^2,

$\phi_q(L) := 1-\phi_1 L-\cdots-\phi_q L^q$, and $\theta_p(L) := 1+\theta_1 L+\cdots+\theta_p L^p$,

with $L^k(Y_t) = Y_{t-k}$ (**lag-operator**), and

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

is the **tempered fractional differencing operator**.

- d - **fractional integration** parameter and $\lambda > 0$ - **tempering parameter**.
- Autoregressive parameters reparameterized in terms of **partial autocorrelations**. Stationarity via uniform priors on $(-1, 1)$.

General time series model of our examples, cont.

- **General ARTFIMA likelihood is intractable**, but the spectral density of **ARTFIMA** is simple

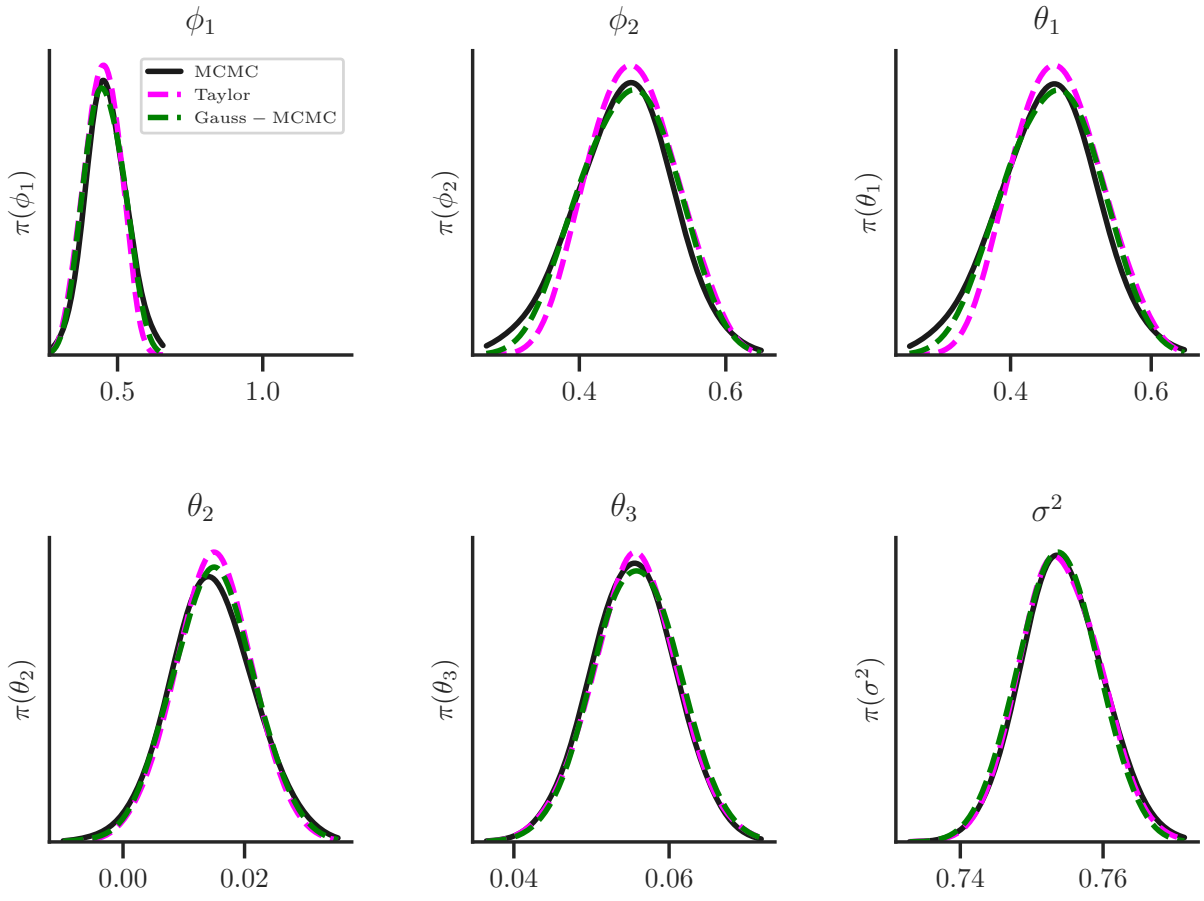
$$f_{\boldsymbol{\theta}}(\omega) = \frac{\sigma^2}{2\pi} \left|1 - e^{-(\lambda + \mathrm{i}\omega)}\right|^{-2d} \left|\frac{\theta_p(e^{-\mathrm{i}\omega})}{\phi_q(e^{-\lambda \mathrm{i}\omega})}\right|^2 .$$

- For ARMA ($d = 0$), **time-domain likelihood** efficiently computed via the **Kalman-filter**. ARMA makes it possible to **assess the accuracy** of the Whittle approximation.

- We fit an ARMA(2,3) to Vancouver temperature data, $44{,}000$ observations, after removing **trend** and **seasonal** effects.

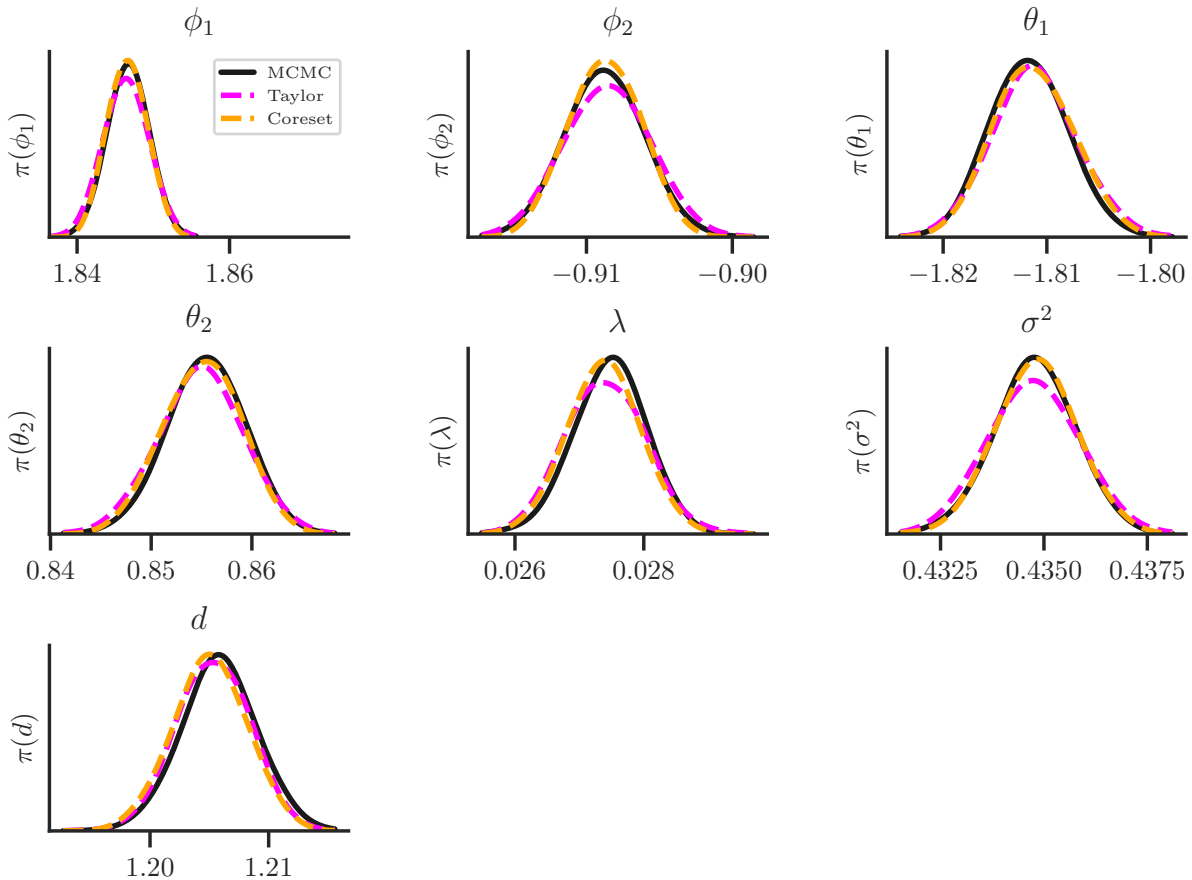
ARMA(p, q) for Vancouver temperature

- $44\,000$ hourly temperature readings.
- Nearly 100 times more effective draws per minute than MCMC **(Relative Computational Time)**.
- Even with this "small" dataset, Whittle is very accurate!



ARTFIMA(p, d, λ, q) for Stockholm temperature

- $450\,000$ hourly temperature readings during 1967-2018.
- Again, nearly 100 times more effective draws per minute than MCMC.



Bonus: SV Models

- **Stochastic Volatility** (SV) models

$$y_t = \exp(v_t/2)\xi_t, \tag{1}$$

  - $\{v_t\}$ is a stationary process with parameter vector $\boldsymbol{\psi}$ and spectral density $f_v(\omega;\boldsymbol{\psi})$ — AR(1) is standard.
  - $\{\xi_t\}$ is an iid sequence having mean zero and unit variance.
- We have

$$\log y_t^2 = \mu + v_t + \varepsilon_t, \tag{2}$$

  where $\{\varepsilon_t\}$ is a zero-mean white noise process w/ variance σ_ε^2.
- So, fitting

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

  to the data $\{\log y_t^2\}$ is **equivalent** to fitting (1)! No latents!
- We fit with v_t ARTFIMA$(1,1)$ to a **million data points** (Bitcoin prices) — see paper.

Conclusions

- **Whittle log-likelihood** is fast to compute and is a sum.
- **Whittle enables subsampling** for time series.
- **Systematic subsampling** of periodogram frequencies to speed up MCMC.
- **Significant speed-ups** compared to regular MCMC.
- Future extensions:
  - **Theory** on approximation accuracy
  - Multidimensional FFT for **spatial data**
  - **Debiased Approaches**