Spectral approaches to speed up Bayesian inference for large stationary time series data

Matias Quiroz[1]

[1]Department of Statistics, Stockholm University

December 2022

What this talk is about

- Spectral approaches to speed up computations and their application in subsampling algorithms.
- **Co-authors** (alphabetical order):
  - *Robert Kohn* (University of New South Wales).
  - *Robert Salomone* (Queensland University of Technology).
  - *Minh-Ngoc Tran* (University of Sydney).
  - *Mattias Villani* (Stockholm University).
- Applied to our previous work on **subsampling MCMC**.
  - Main paper: (Quiroz et al., 2019, JASA).
  - Textbook like review of our work (prior to the spectral approaches): (Quiroz et al., 2018b, Sankhya A).
- The main points of this talk:
  1. Spectral approaches for **univariate** time series and their implied independence that facilitate subsampling.
  2. Extend the approaches to **multivariate** time series.
- Slides on `www.matiasquiroz.com/news`.

Motivation for subsampling MCMC

- Markov chain Monte Carlo (**MCMC**) - Bayesian workhorse for 3 decades.
- This talk focuses on subsampling for **Metropolis-Hastings** (MH). See Dang et al. (2019) for Hamiltonian Monte Carlo, Gunawan et al. (2020) for sequential Monte Carlo and Quiroz et al. (2018a) for delayed acceptance MCMC.
- Metropolis-Hastings is often slow
  - Need to **evaluate the likelihood function** in each iteration.
  - **Many iterations** (sampling algorithm).
  - In time series models, the likelihood evaluation may be computationally expensive for **large time series data**.
- **Subsampling MCMC**: estimate the likelihood from a subsample in each MCMC iteration. Faster!

The Metropolis-Hastings algorithm

- Bayesians carry out inference based on $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- A general approach to generate posterior samples is to construct a Markov chain $\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{N}$ that has $p(\boldsymbol{\theta}|\boldsymbol{y})$ **invariant distribution** as $N \to \infty$.
- The **Metropolis-Hastings** algorithm achieves this as follows.
  1. Start at $\boldsymbol{\theta}_c = \boldsymbol{\theta}^{(0)}$ and set $j = 1$.

     Repeat step 2. to 3. N times.
  2. Propose $\boldsymbol{\theta}_p$ from a proposal (based on $\boldsymbol{\theta}_c$). Accept $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}_p$ with probability

     $$\alpha = \min\left(1, \frac{p\left(\boldsymbol{y}|\boldsymbol{\theta}_p\right)p(\boldsymbol{\theta}_p)}{p\left(\boldsymbol{y}|\boldsymbol{\theta}^{(j-1)}\right)p(\boldsymbol{\theta}^{(j-1)})}\right)$$ (symmetric proposal), else set $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}_c$.
  3. Set $j = j + 1$.
- Some challenges:
  - Likelihood $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(y_k|\boldsymbol{\theta})$ is expensive for large $\dim(\boldsymbol{y}) = n$.
  - For time series, $p(\boldsymbol{y}|\boldsymbol{\theta})$ is expensive even for moderately large n.

Key idea: The Pseudo-Marginal MH (PMMH) algorithm

- Can we speed up likelihood evaluations? (i.) **Data subsampling**. (ii.) **The Whittle likelihood** (time series).
- Data subsampling: Estimate $p(\boldsymbol{y}|\boldsymbol{\theta})$ with $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})$.
- $\boldsymbol{u}$ are **auxiliary variables**, serve the purpose of estimating $p(\boldsymbol{y}|\boldsymbol{\theta})$.
- Metropolis-Hastings with an estimated likelihood? Pseudo marginal!
- Samples from $\pi(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y}) \propto \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})p(\boldsymbol{\theta})p(\boldsymbol{u})$ $[p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})]$.
- Proposes $\boldsymbol{\theta}, \boldsymbol{u}$ and accepts/rejects them jointly. Like previous slide but with

$$\alpha = \min\left(1, \frac{\widehat{p}\left(\boldsymbol{y}|\boldsymbol{\theta}_p, \boldsymbol{u}_p\right)p(\boldsymbol{\theta}_p)}{\widehat{p}\left(\boldsymbol{y}|\boldsymbol{\theta}^{(j-1)}, \boldsymbol{u}^{(j-1)}\right)p(\boldsymbol{\theta}^{(j-1)})}\right).$$

- Targets $\pi(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y})$ with marginal $p(\boldsymbol{\theta}|\boldsymbol{y})$ if $\int \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u} = p(\boldsymbol{y}|\boldsymbol{\theta})$ (Beaumont, 2003; Andrieu and Roberts, 2009).
- True for any positive unbiased estimator, but large variance is inefficient.

PMMH with dependent subsamples

- $\mathrm{V}\left(\log \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})\right) \approx 1$ is optimal (Pitt et al., 2012; Doucet et al., 2015).

- What really matters for PMMH is the variance of

$$\log \frac{\widehat{p}\left(\boldsymbol{y}|\boldsymbol{\theta}_p, \boldsymbol{u}_p\right)}{\widehat{p}\left(\boldsymbol{y}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{u}^{(i-1)}\right)} = \left[\log \widehat{p}_{\mathrm{prop}} - \log \widehat{p}_{\mathrm{curr}}\right].$$

- **Correlated pseudo marginal** (Deligiannidis et al., 2018): Correlate the $\boldsymbol{u}$s over Metropolis-Hastings iterations using an autoregressive proposal $\boldsymbol{u}^{(i)} = \phi \boldsymbol{u}^{(i-1)} + \epsilon$.

- **Block pseudo marginal** (Tran et al., 2016): Partition $\boldsymbol{u}$ into λ blocks and update only K of them jointly with $\boldsymbol{\theta}$ in each iteration.

- Can show that (under certain assumptions)

$$\mathrm{Corr}\left(\log \widehat{p}_{\mathrm{prop}}, \log \widehat{p}_{\mathrm{curr}}\right) \approx 1 - K/\lambda.$$

- Consequence: tolerates much larger variance (faster!) of $\log \widehat{p}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})$.

Bias-corrected log-likelihood based estimator (Quiroz et al., 2019)

- Data **subsampling estimator** of the log-likelihood for independent data

$$\widehat{\ell}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{u}) = \frac{n}{m}\sum_{i\in\boldsymbol{u}}\ell(y_i|\boldsymbol{\theta}), \mathrm{E}_{\boldsymbol{u}}\left(\widehat{\ell}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{u})\right) = \ell(\boldsymbol{\theta}) = \sum_{k=1}^{n}\ell(y_k|\boldsymbol{\theta}),$$

$\boldsymbol{u} = (u_1,\ldots,u_m)$, $\Pr(u_j = k) = 1/n$, $j = 1,\ldots.m$, and $k = 1,\ldots,n$.

- **Difference estimator** with **control variates** $q_k(\boldsymbol{\theta}) \approx \ell(y_k|\boldsymbol{\theta})$

$$\widehat{\ell}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{u}) = \sum_{k=1}^{n} q_k(\boldsymbol{\theta}) + \frac{n}{m}\sum_{i\in\boldsymbol{u}} d_i(\boldsymbol{\theta}),\ \ d_i(\boldsymbol{\theta}) = \ell(y_i|\boldsymbol{\theta}) - q_i(\boldsymbol{\theta}).$$

- Estimate $L(\boldsymbol{\theta}) = \exp\left(\ell(\boldsymbol{y}|\boldsymbol{\theta})\right)$ by **bias-correcting** $\exp\left(\widehat{\ell}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{u})\right)$.
- Bias-correction term estimated... $\widehat{p}(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{u})$ not unbiased anymore...
- ... still targets a **perturbed posterior** with TV-norm error of $\mathcal{O}(n^{-1}m^{-2})$.
- For example, if $m = \mathcal{O}(n^{1/2})$ then the error is $\mathcal{O}(n^{-2})$.

Beyond independent data via spectral methods

- Quiroz et al. (2019) use the quite restrictive assumption:

$$\ell(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ell(y_k|\boldsymbol{\theta}), \quad \left[\text{which comes from } p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(y_k|\boldsymbol{\theta})\right].$$

- Violated for many interesting models, including:
  - General **time series** problems.
  - **Spatially** dependent data.
  - Hyper-parameter estimation in **Gaussian processes**.
- This talk is on univariate and multivariate **stationary time series models**.
- We know that the data $\boldsymbol{y} = (Y_1, \ldots, Y_n)$ in the **time domain** are dependent.
- Transform the time domain data to the frequency domain using the **discrete Fourier transform** (DFT).
- Large sample properties of the DFT ensures (asymptotically) independent observations.

The discrete Fourier transform

- Transform the data from the **time domain** to the **frequency domain**.
- The **discrete Fourier transform** (DFT) of the time series $Y_t \in \mathbb{R}$,

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

  at **Fourier frequencies**

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

- Can be computed in $\mathcal{O}(n \log n)$ with the **fast Fourier transform** (FFT).
- The **periodogram**

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

  will be kth "data observation" in the frequency domain.

The Whittle log-likelihood

- The **periodogram data** in the **frequency domain**,

$$(\mathcal{I}(\omega_1), \mathcal{I}(\omega_2), \ldots, \mathcal{I}(\omega_n)), \quad \text{where } \mathcal{I}(\omega_k) = n^{-1}|J(\omega_k)|^2.$$

- Asymptotically, as $n \to \infty$,

$$\mathcal{I}(\omega_k) \overset{\text{ind}}{\sim} \text{Exp}\left(f_{\boldsymbol{\theta}}(\omega_k)\right),$$

  where $f_{\boldsymbol{\theta}}$ is the **spectral density**,

$$f_{\boldsymbol{\theta}}(\omega) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} \gamma(\tau; \boldsymbol{\theta}) \exp(-\mathrm{i}\omega\tau),$$

  with **covariance function** $\gamma(\tau; \boldsymbol{\theta}) = \mathrm{E}[Y_t Y_{t-\tau}]$ of a covariance stationary zero-mean time series $\{Y_t\}_{t=1}^n$.

- Sketch of proof steps:
  - DFT is asymptotically **complex Gaussian** (by the CLT).
  - $|J(\omega_k)|^2$ is a sum of two squared **independent Gaussians**.
  - $\chi_2^2 = \text{Exp}(1/2)$.

Subsampling MCMC for univariate stationary time series

- **Whittle's** asymptotic approximation of the log-likelihood (Whittle, 1953)

$$\ell_{\mathcal{W}}(\boldsymbol{\theta}) = -\sum_{\omega_k \in \Omega} \left(\log f_{\boldsymbol{\theta}}(\omega_k) + \frac{\mathcal{I}(\omega_k)}{f_{\boldsymbol{\theta}}(\omega_k)} \right).$$

- May be biased for small n, but we consider large n (when subsampling is relevant).
- Unlike the time domain log-likelihood, the Whittle log-likelihood is a sum.
- **Spectral subsampling** for stationary **univariate time series** (Salomone et al., 2020)
  - **Compute periodogram** before MCMC at cost $\mathcal{O}(n \log n)$.
  - Estimate $\ell_{\mathcal{W}}(\boldsymbol{\theta})$ unbiasedly by **subsampling of frequencies**.
  - Use within a **pseudo-marginal** MH algorithm (Quiroz et al., 2019).

The (univariate) ARTFIMA model

- **ARIMA**(p, d, q), integer differences $d = 0, 1, 2, \ldots$ $(L^d Y_t = Y_{t-d})$

$$\phi_p(L)(1-L)^d Y_t = \theta_q(L)\varepsilon_t.$$

- **ARTFIMA**(p, d, λ, q) (Sabzikar et al., 2019)

$$\phi_p(L)(1-e^{-\lambda}L)^d Y_t = \theta_q(L)\varepsilon_t.$$

- Role of **fractional differencing** d. Can model longer memory since

$$(1-e^{-\lambda}L)^d Y_t = \sum_{j=0}^{\infty}(-1)^j \frac{\Gamma(1+d)}{\Gamma(1+d-j)j!} e^{-\lambda j} Y_{t-j}.$$

- The **ARTFIMA** model nests:
  - **ARIMA** ($\lambda = 0$ and d integer).
  - **ARFIMA** (Granger and Joyeux, 1980). $\lambda = 0, d \in \mathbb{Q}$. Stationary if $|d| < 0.5$. ARFIMA has so-called **long memory**: $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau; \boldsymbol{\theta})| = \infty$.

The ARTFIMA model, cont.

- Recall **ARTFIMA**(p, d, λ, q)

$$\phi_p(L)(1 - e^{-\lambda}L)^d x_t = \theta_q(L)\varepsilon_t.$$

- Role of the **tempering** parameter $\lambda \geq 0$.
  - long range dependence $\gamma(\tau; \boldsymbol{\theta})$ for small τ.
  - Exponential decay for larger τ: $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau; \boldsymbol{\theta})| < \infty$.
  - Stationary if $\lambda > 0$ for all $d \notin \mathbb{Z}$ (if AR and MA fulfill the usual conditions).
- The **spectral density** for **ARTFIMA**(p, d, λ, q)

$$f_{\boldsymbol{\theta}}(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\left|\theta(e^{-i\omega})\right|^2}{\left|\phi(e^{-i\omega})\right|^2} \left|1 - e^{-(\lambda + i\omega)}\right|^{-2d}.$$

- Compare to the **spectral density** for **ARMA**(p, q)

$$f_{\boldsymbol{\theta}}(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\left|\theta(e^{-i\omega})\right|^2}{\left|\phi(e^{-i\omega})\right|^2}.$$

Going multivariate (Villani et al., 2022)

- **Autocovariance matrix function** for time series $\boldsymbol{Y}_t \in \mathbb{R}^r$

$$\gamma_{\boldsymbol{Y}}(\tau) = \mathrm{Cov}(\boldsymbol{Y}_t, \boldsymbol{Y}_{t-\tau}), \text{ for } \tau \in \mathbb{Z}.$$

- **Spectral density matrix**

$$f_{\boldsymbol{Y}}(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_{\boldsymbol{Y}}(\tau) \exp(-i\omega\tau).$$

where off-diagonal elements are the **cross-spectral densities**

$$f_{jk}(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma_{jk}(\tau) \exp(-i\omega\tau), \text{ for } \omega \in (-\pi, \pi].$$

- **Multivariate** discrete Fourier transform (**DFT**)

$$J(\omega_k) = \frac{1}{\sqrt{2\pi}} \sum_{t=0}^{n-1} \boldsymbol{Y}_t \exp(-i\omega_k t).$$

Multivariate Fourier analysis, cont.

- DFT are **asymptotically** independent complex normal (Brillinger, 2001)

$$n^{-1/2}J(\omega_k) \overset{\text{ind}}{\sim} \mathrm{CN}(0, f_{\boldsymbol{Y}}(\omega_k)) \text{ as } n \to \infty.$$

- **Multivariate periodogram** is complex singular Wishart ($r > 1$)

$$I_T(\omega_k) = n^{-1}J(\omega_k)J(\omega_k)^H \sim \mathrm{CW}(1, f_{\boldsymbol{Y}}(\omega_k)),$$

  where $\boldsymbol{A}^H = \overline{\boldsymbol{A}}^\top$ is the conjugate transpose and $\overline{\boldsymbol{A}}$ is the matrix of complex conjugates of the elements of $\boldsymbol{A}$.

- **Multivariate Whittle** log-likelihood

$$\ell_{\mathcal{W}}(\boldsymbol{\theta}) = -\sum_{\omega_k \in \Omega} \left(\log |f_{\boldsymbol{Y}}(\omega_k)| + \mathrm{tr}\left[f_{\boldsymbol{Y}}(\omega_k)^{-1} I_T(\omega_k) \right] \right).$$

- It is a sum (still) — subsampling MCMC.

The multivariate ARTFIMA (VARTFIMA) model

- We propose a **multivariate extension** of the ARTFIMA (Sabzikar et al., 2019) process.
- The vector ARTFIMA (**VARTFIMA**)

$$\Phi_p(L)\Delta^{d,\lambda}\boldsymbol{Y}_t = \Theta_q(L)\boldsymbol{\varepsilon}_t \quad \boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_\varepsilon)$$

where $\Delta^{d,\lambda}$ is the **tempered fractional differencing operator** defined as

$$\Delta^{\mathbf{d},\lambda}\boldsymbol{Y}_t = \left((1-e^{-\lambda_1}L)^{d_1}Y_{1,t}, \ldots, (1-e^{-\lambda_r}L)^{d_r}Y_{r,t}\right)^\top.$$

- ARTFIMA nested ARMA, **VARTFIMA nests VARMA** (vector ARMA).
- **Spectral density matrix** (see (Villani et al., 2022, Theorem 1))

$$f_{\boldsymbol{Y}}(\omega) = \frac{1}{2\pi}\mathbf{B}\Phi_p^{-1}(e^{-i\omega})\Theta_q(e^{-i\omega})\Sigma_\varepsilon\Theta_q^H(e^{-i\omega})\Phi_p^{-H}(e^{-i\omega})\mathbf{B}^H,$$

where $\mathbf{B} = \mathrm{Diag}\left((1-e^{-(\lambda_1+i\omega)})^{-d_1}, \ldots, (1-e^{-(\lambda_r+i\omega)})^{-d_r}\right)$.

Questions of interest

- **Question 1 (Q1)**: Is the proposed **VARTFIMA** better than **VARMA**?
- **Question 2 (Q2)**: Is **subsampling MCMC** for the multivariate Whittle likelihood accurate? Compare vs MCMC on the multivariate Whittle likelihood **using all data**.
- **Question 3 (Q3)**: How accurate is the **multivariate Whittle approximation** to the true posterior (obtained via the time domain likelihood)?
- **Note**: The time domain likelihood (and thus the posterior) is intractable for VARTFIMA with large datasets. Can only test the above (Q3) for VARMA.
- **Question 4 (Q4)**: How much "faster" is **subsampling** compared to using the multivariate Whittle likelihood on **the full dataset**?

Applications: Hydrology, meteorology, environment

- Application 1: Measurements of mean **water velocity** in two locations in Detroit river, located on opposite sides of Lake St Clair. 130,000 observations.
- Application 2: Measurements of **temperature** in three airport locations in Sweden (Arlanda, Bromma and Landvetter). 124,000 observations.
- Application 3: Measurements of **two pollution types** (**nitrogen dioxide** (NO2) and **particulate matter** (PM10)) at two streets in central Stockholm. 50,000 observations.

Raw data water velocity

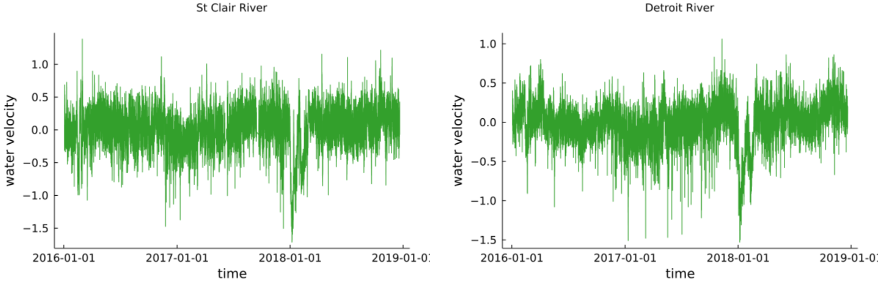


Figure 1: Water velocity data.

Raw data temperature

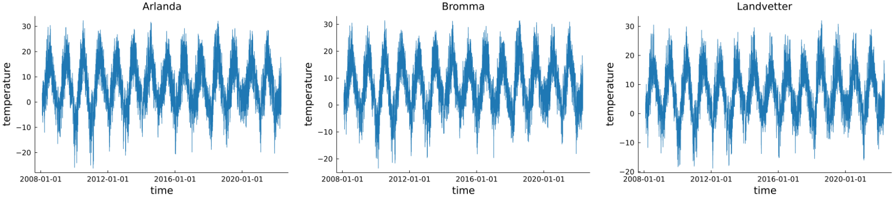


Figure 2: Swedish temperature data after interpolation, but before deseasoning.

Raw data pollution

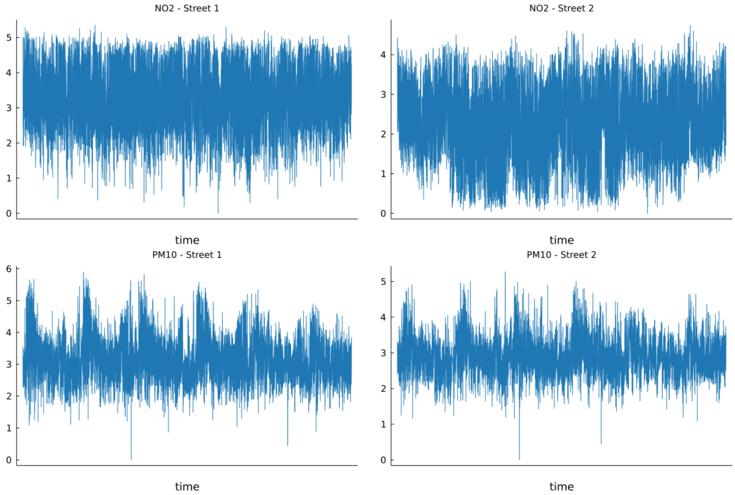


Figure 3: Stockholm air pollution data after interpolation and logarithmic transform, but before deseasoning.

Models and priors

- We estimate **VARMA**() and **VARTFIMA**() models for each dataset.
- Model selection using the **BIC approximation** of the log marginal likelihood

$$\log p_{\mathrm{BIC}}(\boldsymbol{Y}) = \log p(\boldsymbol{Y}|\widehat{\boldsymbol{\theta}}) - \frac{k \log n}{2},$$

  where k is the number of estimated parameters, n is the length of the time series and $\widehat{\boldsymbol{\theta}}$ is the **maximum likelihood estimate**.
- **Minnesota-style** prior for the AR and MA coefficients. Normal with diagonal covariance matrix:

$$v_{ij,l} = \begin{cases} (\lambda_0/l)^2, & \text{if } i = j \\ (\lambda_0 \theta_0 \sigma_i / l\sigma_j)^2 & \text{if } i \neq j. \end{cases}$$

  We set $\lambda_0 = 1$ and $\theta_0 = 0.2$. Normal priors for the rest (**transformed scale**).
- Ansley and Kohn (1986) parameterisation of AR and MA parts. **Ensures stationarity** — facilitates MCMC proposal (**unconstrained space**).

Q1: BIC — Tempered fractional differencing is good

- **Question 1**: Is the proposed **VARTFIMA** better than **VARMA**?

		Water Velocity		Temperature		Pollution	
AR	MA	No TFI	TFI	No TFI	TFI	No TFI	TFI
1	0	737079	759123	327097	334122	363760	366022
0	1	588297	759457	61320	332888	306068	365658
2	0	749650	761200	335201	**335757**	365522	**366266**
0	2	621765	**761786**	93256	333948	325717	366142
1	1	758838	761305	333582	335647	365762	**366267**

Table 1: BIC approximation of the log marginal likelihood for different models for each of the three datasets. A higher value indicates a better model fit. **The highest value for each dataset is marked in bold font red.**

- **Answer**: Yes, the tempered fractional differencing is better than no tempered fractional differencing.

Q2: Best model for temperature data ($k = 24$)

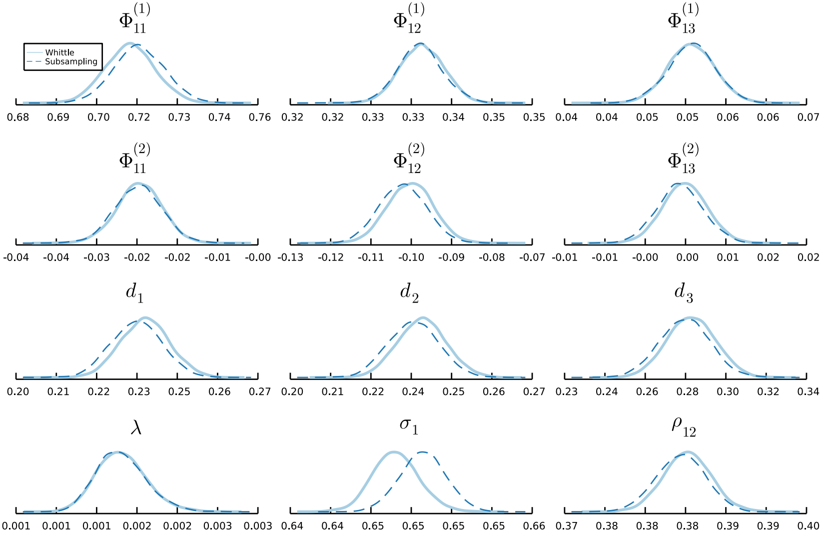


Figure 4: Kernel density estimates of a subset of the marginal posterior densities for the VARTFIMA(2,0) model fitted to the Swedish temperature data.

Q2: Best model for Water velocity data ($k = 11$)

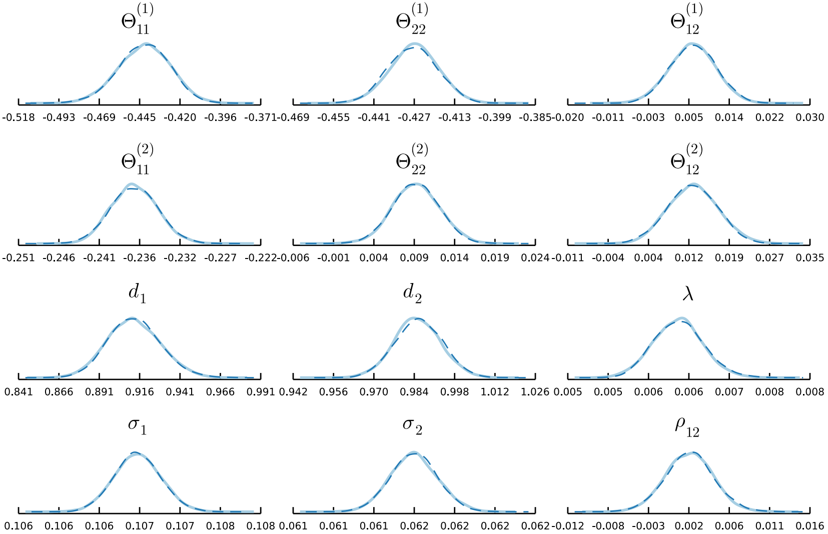


Figure 5: Kernel density estimates of a subset of the marginal posterior densities for the VARTFIMA(0,2) model fitted to the Water velocity data.

Q2: Best model for pollution data ($k = 47$)

- Subsampling MCMC does not work for this model for the given n and m.
- Chain gets stuck because $\widehat{\sigma}^2_{\widehat{\ell}} = \mathrm{V}\left(\widehat{\ell}(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{u})\right)$ is too large.
- $\widehat{\sigma}^2_{\widehat{\ell}} = \mathcal{O}(m^{-1}n^{-1})$ for the control variate we use (Quiroz et al., 2019).
- Example with $n = 62{,}000$ (instead of $n = 124{,}000$) for the Swedish temperature data on the next slide shows how the chain gets stuck.

Subsampling MCMC with smaller n fails for the Swedish temperature dataset

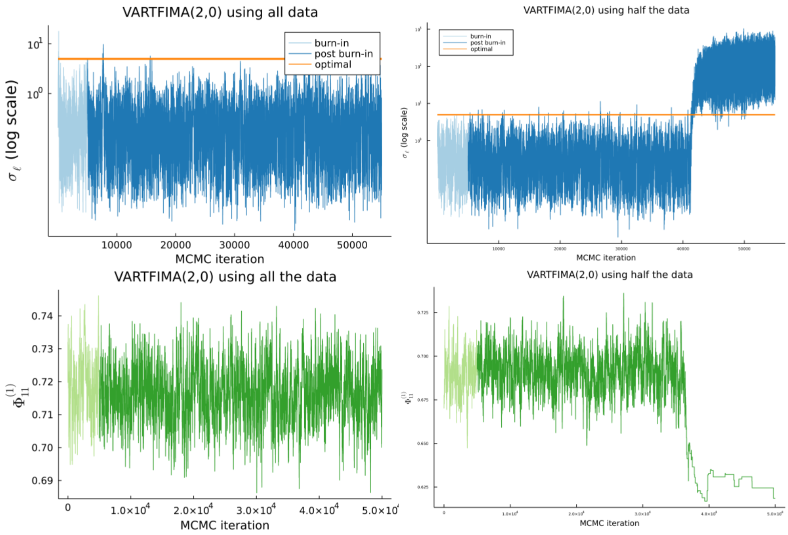
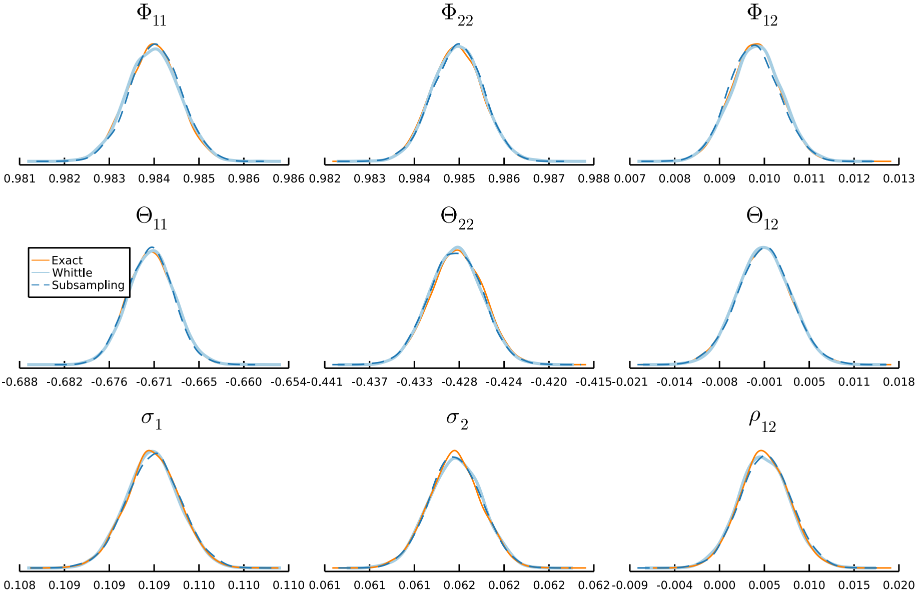
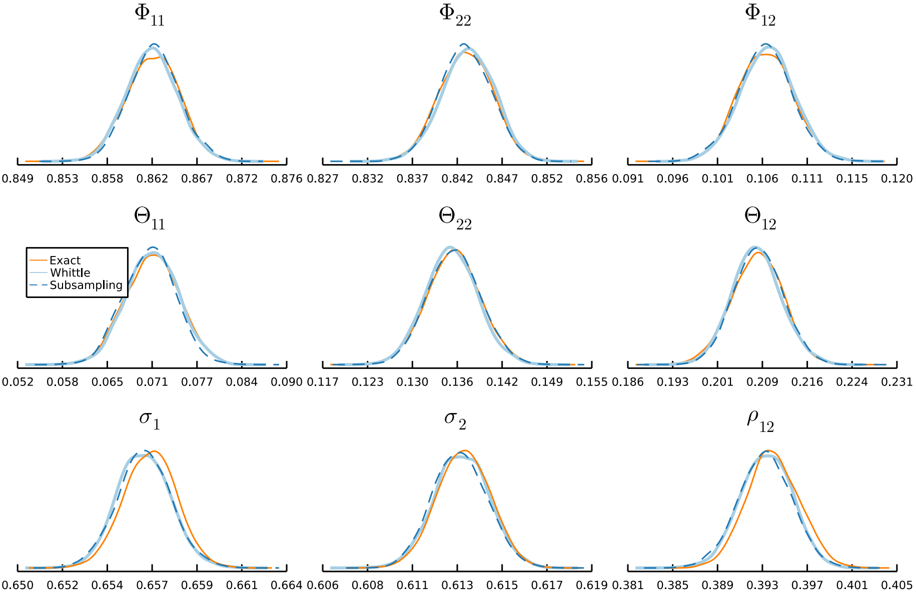


Figure 6: Subsampling MCMC fails for the smaller dataset.

Q3: VARMA(1, 1) - Water velocity ($k = 11$)



Q3: VARMA(1, 1) - Temperatures ($k = 24$)



Q4: What about the speed-up of subsampling MCMC?

- **Relative computational time (RCT)**:

$$\text{RCT} = \frac{\text{CT MCMC full sample}}{\text{CT spectral subsampling MCMC}}.$$

Dataset	Model	Min	Mean	Max
Water velocity	VARTFIMA(0,2)	87	98	125
Temperature	VARTFIMA(2,0)	68	89	114

Table 2: Relative computational time (RCT) for comparing MCMC using the full dataset to spectral subsampling MCMC. The value 1 indicates that spectral subsampling MCMC and MCMC are equally efficient, and values larger than 1 indicate that spectral subsampling MCMC is the better algorithm.

- Recall that subsampling MCMC does not work for the pollution example when $n = 50{,}000$.

Conclusion and future work

- Presented a simple idea that extends **subsampling MCMC** beyond the conditionally independent observations setting.
- Useful for any subsampling approach (not just MCMC).
- Considered an application of subsampling MCMC for **multivariate** time series models.
- Presented the novel vector time series model **VARTFIMA**.
- Villani et al. (2022) show that VARTFIMA predicts each time series better than univariate ARTFIMA for the temperature data, especially for longer prediction horizons.
- Subsampling MCMC can break down if the control variates are inaccurate.
  - More efficient control variate constructions.
  - Other estimation algorithms that are less sensitive to the variance, e.g. variational Bayes.
- We are currently working on extending our approach to **spatial problems**.

Thank you for listening!

Questions?

References I

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

Ansley, C. F. and Kohn, R. (1986). A note on reparameterizing a vector autoregressive moving average model to enforce stationarity. *Journal of Statistical Computation and Simulation*, 24(2):99–106.

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

Brillinger, D. R. (2001). *Time series: Data Analysis and Theory*. SIAM.

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

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

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

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

References II

Gunawan, D., Dang, K.-D., Quiroz, M., Kohn, R., and Tran, M.-N. (2020). Subsampling sequential Monte Carlo for static Bayesian models. *Statistics and Computing*, 30:1741–1758.

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

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

Quiroz, M., Villani, M., Kohn, R., Tran, M.-N., and Dang, K.-D. (2018b). Subsampling MCMC-an introduction for the survey statistician. *Sankhya A*, 80(1):33–69.

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

Salomone, R., Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2020). Spectral subsampling MCMC for stationary time series. In *Proceedings of the Thirty-seventh International Conference on Machine Learning*, pages 10510–10519.

References III

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

Villani, M., Quiroz, M., Kohn, R., and Salomone, R. (2022). Spectral subsampling MCMC for stationary multivariate time series with applications to vector ARTFIMA processes. *Econometrics and Statistics*, (To appear).

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