

Spectral Subsampling MCMC for Stationary Multivariate Time Series

Mattias Villani 🧑

**Department of Statistics
Stockholm University**

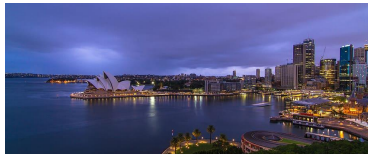


Overview

- Background: Subsampling MCMC/HMC
 - Spectral subsampling for multivariate time series
 - Application to vector ARTFIMA
-
- **Slides:** <http://mattiasvillani.com/news> (soon).

Joint work with some Aussies

- Subsampling MCMC for time series:
 - ▶ **Robert Kohn**, UNSW Sydney
 - ▶ **Matias Quiroz**, UTS Sydney
 - ▶ **Robert Salomone**, QUT Brisbane
- Subsampling MCMC/HMC (conditionally) independent data:
 - ▶ **Minh-Ngoc Tran**, Univ of Sydney
 - ▶ **Khue-Dung Dang**, Univ of Melbourne



Mattias Villani



Spectral Subsampling MCMC

Motivation

- **Long time series** are increasingly **common**:
 - ▶ high frequency financial transaction data
 - ▶ neuroimaging data with high temporal resolution
 - ▶ sensor data from robots
 - ▶ meteorological weather stations
 - ▶ GPS data used in urban traffic monitoring.
- Often **multivariate observations**.
- Automatic **decision making under uncertainty**.
- **Bayesian decisions**: maximize **posterior expected utility**.
- Posteriors by Markov Chain Monte Carlo (**MCMC**) simulation.
- **MCMC is slow** on large datasets. 🙄

The Metropolis-Hastings (MH) algorithm

■ Bayesian inference

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta})p(\boldsymbol{\theta})$$

■ Initialize $\boldsymbol{\theta}^{(0)}$ and iterate for $k = 1, 2, \dots, N$

1 Sample $\boldsymbol{\theta}_p \sim q(\cdot|\boldsymbol{\theta}^{(k-1)})$ (the **proposal distribution**)

2 Accept $\boldsymbol{\theta}_p$ with **acceptance probability**

$$\alpha = \min \left(1, \frac{L(\boldsymbol{\theta}_p)p(\boldsymbol{\theta}_p)}{L(\boldsymbol{\theta}^{(k-1)})p(\boldsymbol{\theta}^{(k-1)})} \frac{q(\boldsymbol{\theta}^{(k-1)}|\boldsymbol{\theta}_p)}{q(\boldsymbol{\theta}_p|\boldsymbol{\theta}^{(k-1)})} \right)$$

■ **Costly** to evaluate $L(\boldsymbol{\theta}_p)$ when n is large. **Big data**.

Naive Subsampling MH

- Independent data - **log-likelihood** is a **sum**

$$\ell(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log \ell_i(\boldsymbol{\theta}), \quad \ell_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} p(y_i|\boldsymbol{\theta})$$

- Unbiased estimate** of $\ell(\boldsymbol{\theta})$ from **subsample** of size $m \ll n$

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

- Run **Pseudo-marginal MH** with $\hat{L}(\boldsymbol{\theta}, \mathbf{u}) = \exp(\hat{\ell}(\boldsymbol{\theta}, \mathbf{u}))$.

- Initialize $(\boldsymbol{\theta}^{(0)}, \mathbf{u}^{(0)})$ and iterate for $k = 1, 2, \dots, N$
 - Sample $\boldsymbol{\theta}_p \sim q(\cdot | \boldsymbol{\theta}^{(k-1)})$ and subsample $\mathbf{u}_p \sim p(\mathbf{u})$
 - Accept $(\boldsymbol{\theta}_p, \mathbf{u}_p)$ with **acceptance probability**

$$\alpha = \min \left(1, \frac{\hat{L}(\boldsymbol{\theta}_p, \mathbf{u}_p) p(\boldsymbol{\theta}_p)}{\hat{L}(\boldsymbol{\theta}^{(k-1)}, \mathbf{u}^{(k-1)}) p(\boldsymbol{\theta}^{(k-1)})} \frac{q(\boldsymbol{\theta}^{(k-1)} | \boldsymbol{\theta}_p)}{q(\boldsymbol{\theta}_p | \boldsymbol{\theta}^{(k-1)})} \right)$$

Fixing Naive Subsampling MH - Bias

- If \hat{L} unbiased then samples are from $p(\boldsymbol{\theta}|\mathbf{y})$ [1]
- Unbiased log-likelihood: $\mathbb{E}_{\mathbf{u}}[\hat{\ell}(\boldsymbol{\theta}, \mathbf{u})] = \ell(\boldsymbol{\theta})$
- But biased likelihood estimate: $\hat{L}(\boldsymbol{\theta}, \mathbf{u}) = \exp(\hat{\ell}(\boldsymbol{\theta}, \mathbf{u}))$
- Approximate bias correction of $\exp(\hat{\ell}(\boldsymbol{\theta}, \mathbf{u}))$ [2]

Theorem: Error in posterior approximation is $O\left(\frac{1}{m^2 n}\right)$ [3]

- Unbiased Block-Poisson estimator + Signed PMMH [4]

Fixing Naive Subsampling MH - Variance

- **Low** $\mathbb{V}(\hat{L}(\boldsymbol{\theta}, \mathbf{u}))$ is crucial for **efficient sampling**.
- **Difference estimator** with control variates [3]

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

- **Control variates** $q_{u_i}(\boldsymbol{\theta})$ by Taylor expansion around $\tilde{\boldsymbol{\theta}}$. [3, 5]
- **Optimal tuning** of subsample size m [6, 3, 4]
- **Blocking**: only refresh part of the subsample [7, 8]
- **Grouping observations** for improved control variates [9]
- **High-dim $\boldsymbol{\theta}$: Subsampling HMC**. [10]

Beyond independent data

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

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

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

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

- **Log-likelihood is a sum:**

- ▶ for conditionally independent y_i
- ▶ for longitudinal data when subjects are independent.
- ▶ for special time series, e.g. AR processes. Sample (x_t, x_{t-1}) .

- General **time series** dependence? Spatial dependence?

Spectral density of a stationary process

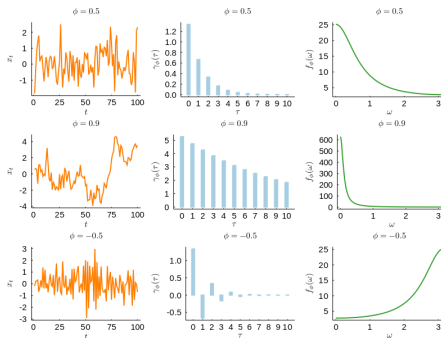
Autocovariance function

$$\gamma(\tau) = \mathbb{E}[(x_t - \mu)(x_{t-\tau} - \mu)], \quad \tau = 0, 1, \dots$$

Spectral density

$$f(\omega) \equiv \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma(\tau) \exp(-i\omega\tau) \quad \text{for } \omega \in (-\pi, \pi].$$

AR(1) process: $x_t = \phi x_{t-1} + \varepsilon_t$, $\varepsilon_t \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$



Multivariate Fourier analysis

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

$$\gamma_{\mathbf{x}}(\tau) = \text{Cov}(\mathbf{x}_t, \mathbf{x}_{t-\tau}) = [\gamma_{jk}(\tau)]_{j,k=1,\dots,r}$$

- **Spectral density matrix**

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

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

$$f_{jk}(\omega) = \sum_{\tau=-\infty}^{\infty} \gamma_{jk}(\tau) \exp(-i\omega\tau)$$

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

$$J(\omega_k) = \sum_{t=0}^{n-1} \mathbf{x}_t \exp(-i\omega_k t)$$

Subsampling MCMC for multivariate time series

- DFT asymptotically independent complex multiv normal [11]

$$n^{-1/2} J(\omega_k) \stackrel{\text{indep}}{\sim} \text{CN}(0, 2\pi f_{\mathbf{x}}(\omega_k)) \text{ as } n \rightarrow \infty.$$

- **Multivariate periodogram** is complex singular Wishart

$$I_T(\omega) = (2\pi n)^{-1} J(\omega) J_T(\omega)^H \sim \text{CW}(1, f_{\mathbf{x}}(\omega))$$

- **Multivariate Whittle** log-likelihood [12]

$$\ell_{\mathcal{W}}(\boldsymbol{\theta}) = - \sum_{\omega_k \in \Omega_n} (\log |f_{\mathbf{x}}(\omega_k)| + \text{tr} [f_{\mathbf{x}}(\omega_k)^{-1} I_T(\omega)])$$

- **Whittle log-likelihood is a sum. Subsample frequencies!**
- Whittle **biased for small n**
- ... but **subsampling** only relevant for **large n** .

Improved control variates by grouping

- Difference estimator:

$$\widehat{\ell}_{\text{diff}}(\boldsymbol{\theta}, \mathbf{u}) := \sum_{k=1}^n q_k(\boldsymbol{\theta}) + \frac{n}{m} \sum_{j=1}^m \left(\ell_{u_j}(\boldsymbol{\theta}) - q_{u_j}(\boldsymbol{\theta}) \right)$$

- Need $q_i(\boldsymbol{\theta}) \approx \ell_i(\boldsymbol{\theta})$ to have small $\text{Var}(\widehat{\ell}_{\text{diff}}(\boldsymbol{\theta}, \mathbf{u}))$.
- $q_i(\boldsymbol{\theta})$ by second order Taylor approximation of $\ell_i(\boldsymbol{\theta})$.
- **Group observations.** Treat partial sum as sampling unit: [9]
$$\ell(\boldsymbol{\theta}) = \underbrace{\ell_{\omega_1}(\boldsymbol{\theta}) + \ell_{\omega_2}(\boldsymbol{\theta}) + \dots \ell_{\omega_K}(\boldsymbol{\theta})}_{\text{group 1}} + \underbrace{\ell_{\omega_{K+1}}(\boldsymbol{\theta}) + \ell_{\omega_{K+2}}(\boldsymbol{\theta}) + \dots \ell_{\omega_{2K}}(\boldsymbol{\theta})}_{\text{group 2}} + \dots$$
- **Bernstein-von Mises:** partial sums quadratic as $K \rightarrow \infty$.
- Each group spans the whole spectral domain.

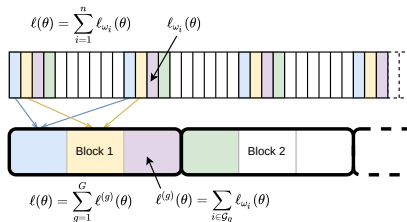
Blocking

- What really matters for MH is the variance of

$$\log \frac{\hat{p}(\mathbf{y}|\theta_p, \mathbf{u}_p)}{\hat{p}(\mathbf{y}|\theta^{(i-1)}, \mathbf{u}^{(i-1)})}$$

- Blocking:** [7]

- partition the groups in blocks
- update only a single block** at each iteration.



Univariate ARTFIMA

- **ARFIMA**(p, d, q) with fractional differencing d

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

- **Long memory**. $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| = \infty$. But stationary if $|d| < 1/2$.

$$(1-L)^d x_t \stackrel{\text{def}}{=} \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(1+d)}{\Gamma(1+d-j)j!} x_{t-j}$$

- **ARTFIMA** adds **tempering** parameter $\lambda \geq 0$ [13]

$$(1 - e^{-\lambda} L)^d x_t \stackrel{\text{def}}{=} \sum_{j=0}^{\infty} (-1)^j \frac{\Gamma(1+d)}{\Gamma(1+d-j)j!} e^{-\lambda j} x_{t-j}$$

- ▶ long range dependence in $\gamma(\tau)$ for small τ
- ▶ exponential decay for larger τ
- ▶ Stationary for all d and $\lambda > 0$.

Vector ARTFIMA(p, d, λ, q)

- Multivariate extension of ARTFIMA for r -dim \mathbf{x}_t [12]

$$\Phi_p(L)\Delta^{\mathbf{d},\boldsymbol{\lambda}}(\mathbf{x}_t - \boldsymbol{\mu}) = \Theta_q(L)\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \stackrel{\text{iid}}{\sim} N(0, \Sigma_\varepsilon)$$

where

$$\Delta^{\mathbf{d},\boldsymbol{\lambda}} \equiv \text{Diag}((1 - e^{-\lambda_1}L)^{d_1}, \dots, (1 - e^{-\lambda_r}L)^{d_r})$$

- VARTFIMA is **stationary** and causal for all \mathbf{d} and $\boldsymbol{\lambda} > \mathbf{0}$.
- **Spectral density matrix**

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

$$\mathbf{B} = \text{Diag}((1 - e^{-(\lambda_1+i\omega)})^{-d_1}, \dots, (1 - e^{-(\lambda_r+i\omega)})^{-d_r}).$$

- **Ansley-Kohn parametrization** of both Φ and Θ to ensure stationarity and invertibility.
- Aim: joint posterior

$$p(\Phi, \Theta, \mathbf{d}, \boldsymbol{\lambda} | \mathbf{x}_{1:n})$$

Three datasets for evaluation

■ Swedish temperatures

- ▶ Three locations: Arlanda, Bromma and Landvetter.
- ▶ Hourly data from February 1, 2008 until May 1, 2022.

■ Water velocity

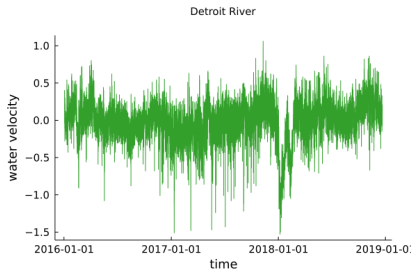
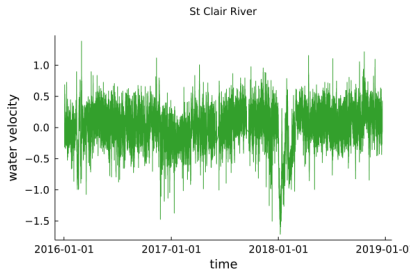
- ▶ Mean water velocity every 12th minute at two locations on opposite sides of Lake St Clair.
- ▶ 130,001 observations from Jan 3, 2016 until Dec 21, 2018.

■ Air pollution in Stockholm

- ▶ Nitrogen dioxide (NO₂) and particulate matter (PM₁₀) pollution at two streets in central Stockholm.
- ▶ Hourly data for the time period February 16, 2010 until October 31, 2015.

■ Subsample: 1% of sample, using control variates for groups.

Water velocity data



Model selection via BIC approximation

AR	MA	Water Velocity		Temperature		Pollution	
		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

Computational times

■ Computational Time (CT)

CT = Inefficiency factor \times Compute time for single draw

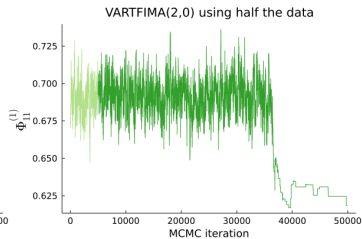
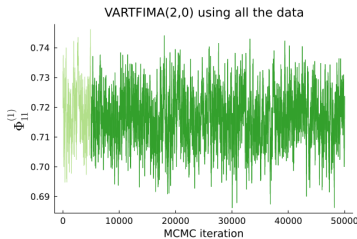
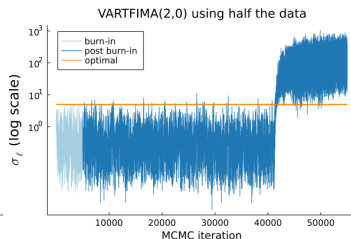
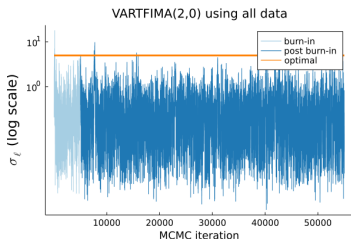
■ Relative Computational Time (RCT):

$$\text{RCT} = \frac{\text{CT MCMC full data sample}}{\text{CT Spectral subsampling MCMC}}$$

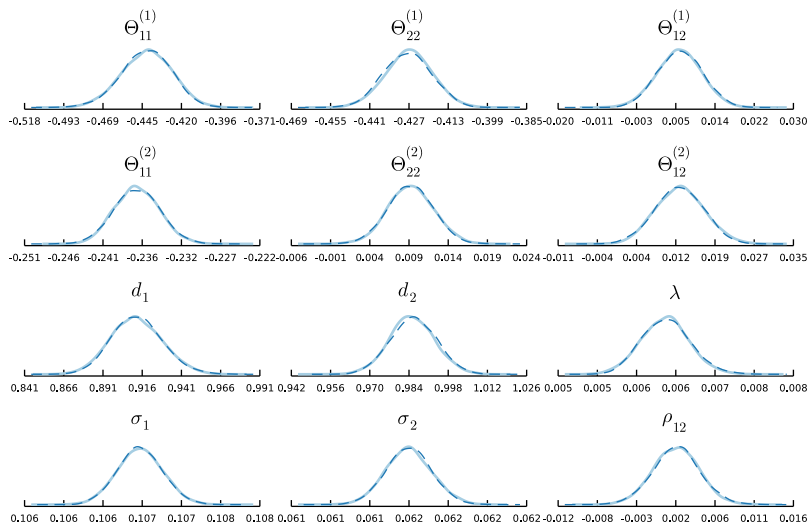
Dataset	Model	Min	Mean	Max
Water velocity	VARTFIMA(0,2)	87	98	125
Temperature	VARTFIMA(2,0)	68	89	114

Variance of log-likelihood estimator is crucial

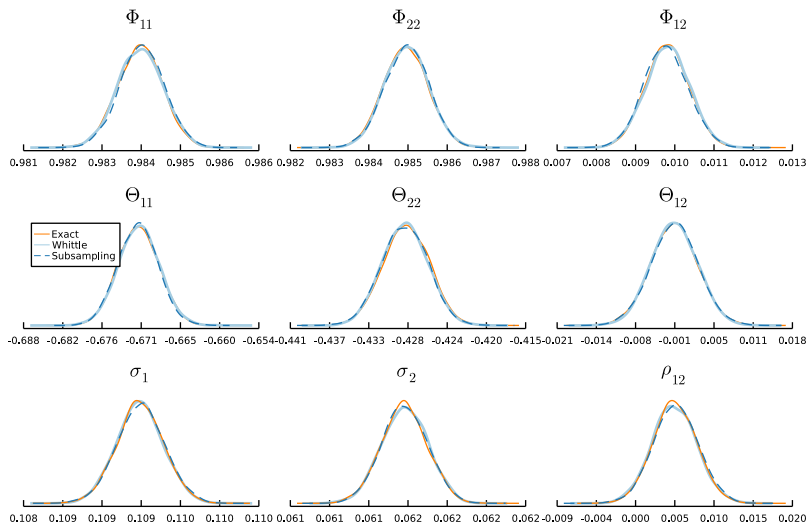
- Spectral subsampling can fail when $\text{Var}(\hat{\ell})$ is too large.
- VARTFIMA(2,0) for Swedish temperature data:



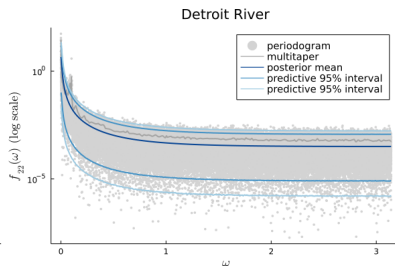
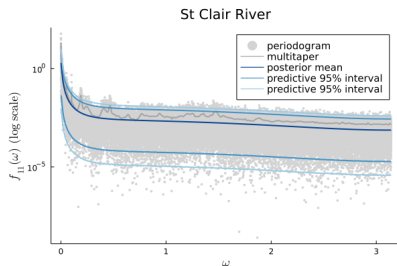
VARTFIMA(0,2) - Subsampling is accurate



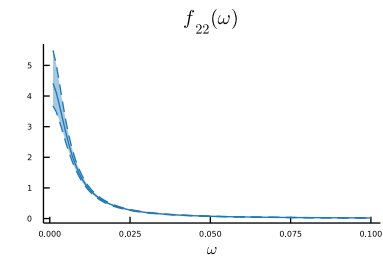
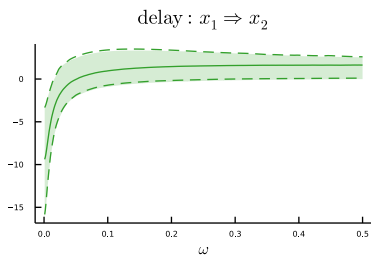
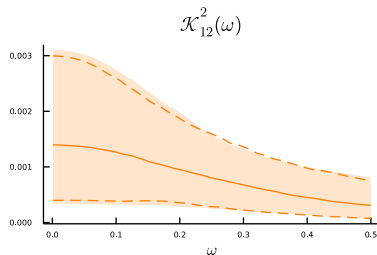
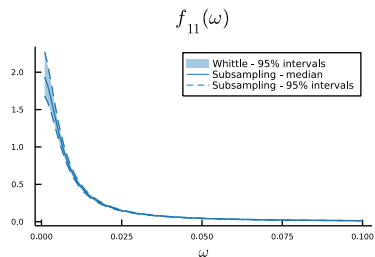
VARMA(1,1) - Whittle is accurate



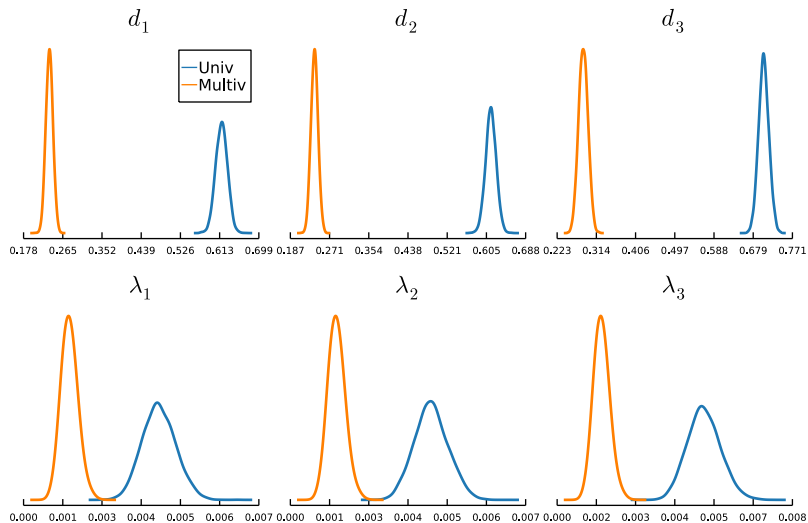
VARTFIMA(0,2) - good model fit



VARTFIMA(0,2) - coherence and delay (phase)








Swedish temperature data



Conclusions

- **Whittle log-likelihood** is fast to compute and is a sum.
- **Whittle enables subsampling** for time series.
- **Subsampling of matrix periodogram** data to speed up MCMC/HMC for multivariate time series.
- **Very large speed-ups** compared to regular MCMC/HMC.
- Future extensions:
 - ▶ Better **control variates for high-dim θ**
 - ▶ **Spatial data**
 - ▶ **Better Whittle**

References

-  C. Andrieu and G. O. Roberts, “The pseudo-marginal approach for efficient Monte Carlo computations,” *The Annals of Statistics*, pp. 697–725, 2009.
-  D. Ceperley and M. Dewing, “The penalty method for random walks with uncertain energies,” *The Journal of chemical physics*, vol. 110, no. 20, pp. 9812–9820, 1999.
-  M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran, “Speeding up mcmc by efficient data subsampling,” *Journal of the American Statistical Association*, no. forthcoming, pp. 1–35, 2018.
-  M. Quiroz, M.-N. Tran, M. Villani, R. Kohn, and K.-D. Dang, “The block-Poisson estimator for optimally tuned exact subsampling MCMC,” *arXiv preprint arXiv:1603.08232*, 2018.
-  R. Bardenet, A. Doucet, and C. Holmes, “On markov chain monte carlo methods for tall data,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1515–1557, 2017.

-  M. K. Pitt, R. d. S. Silva, P. Giordani, and R. Kohn, “On some properties of Markov chain Monte Carlo simulation methods based on the particle filter,” *Journal of Econometrics*, vol. 171, no. 2, pp. 134–151, 2012.
-  M.-N. Tran, R. Kohn, M. Quiroz, and M. Villani, “Block-wise pseudo-marginal metropolis-hastings,” *arXiv preprint arXiv:1603.02485*, 2016.
-  G. Deligiannidis, A. Doucet, and M. K. Pitt, “The correlated pseudo-marginal method,” *arXiv preprint arXiv:1511.04992*, 2015.
-  R. Salomone, M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran, “Spectral subsampling mcmc for stationary time series,” *ICML2020 and arXiv:1910.13627*, 2020.
-  K.-D. Dang, M. Quiroz, R. Kohn, M.-N. Tran, and M. Villani, “Hamiltonian monte carlo with energy conserving subsampling,” *Journal of Machine Learning Research*, 2019, vol. 20, pp. 1–31, 2019.



D. R. Brillinger, *Time series: data analysis and theory*.
SIAM, 2001.



M. Villani, M. Quiroz, R. Kohn, and R. Salomone, “Spectral subsampling mcmc for stationary multivariate time series with applications to vector artfima processes,” *Econometrics and Statistics*, 2023.



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