Spectral Subsampling MCMC for Large-Scale Time Series

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

Overview

- Subsampling MCMC/HMC
- Fourier analysis from a statistical viewpoint
- The Whittle likelihood
- Spectral subsampling for stationary time series

- Slides: http://mattiasvillani.com/news

Collaborators down under - in chronological order

- Robert Kohn, UNSW Sydney
- Matias Quiroz, UTS Sydney
- Minh-Ngoc Tran, University of Sydney
- Khue-Dung Dang, UTS Sydney
- Robert Salomone, UNSW Sydney

The Metropolis-Hastings (MH) algorithm

- **Bayesian inference**

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

- Initialize $\theta^{(0)}$ and iterate for $k = 1, 2, ..., N$
  1. Sample $\theta_p \sim q\left(\cdot|\theta^{(k-1)}\right)$ (the **proposal distribution**)
  2. Accept θ_p with **acceptance probability**

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

- **Costly** to evaluate $L(\theta_p)$ when n is large. **Big data**.

Naive Subsampling MH

- Independent data - log-likelihood is a sum

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

- Estimate log-likelihood $\ell(\theta)$ from subsample of size $m \ll n$

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

- Unbiased: $\mathbb{E}_{\mathrm{u}}[\hat{\ell}(\theta, \mathrm{u})] = \ell(\theta)$.
- Run Pseudo-marginal MH with $\hat{L}(\theta, \mathrm{u}) = \exp(\hat{\ell}(\theta, \mathrm{u}))$.

- Initialize $(\theta^{(0)}, \mathrm{u}^{(0)})$ and iterate for $k = 1, 2, ..., N$
  1. Sample $\theta_p \sim q(\cdot|\theta^{(k-1)})$ and subsample $\mathrm{u}_p \sim p(\mathrm{u})$
  2. Accept (θ_p, u_p) with acceptance probability

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

Fixing Naive Subsampling MH

- Pseudo-marginal MH samples from $p(\theta|\boldsymbol{y})$ if $\hat{L}$ **is unbiased** [1]
  - **Approximate bias correction of** $\exp\left(\hat{\ell}(\theta, \mathrm{u})\right)$ [2]
    Theorem: Error in posterior approximation is $O(m^{-2}n^{-1})$. [3]
  - **Unbiased Block-Poisson estimator** + **Signed PMMH**. [4]

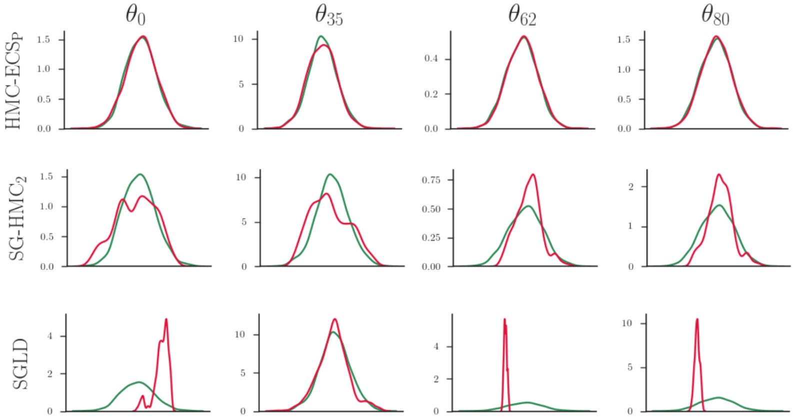
- **Low** $\mathbb{V}\left(\hat{L}\left(\theta, \mathrm{u}\right)\right)$ crucial for **efficient sampling**. Stuck.
  - **Difference estimator** and **control variates** [3, 5, 6]
  - **Optimal tuning** of m [4]
  - **Blocking**: only refresh part of the subsample [7, 8]

- **High-dim**: **Energy Conserving Subsampling HMC**.
  Estimate likelihood and Hamiltonian dynamics from **same** subsample. [9]

Logistic spline regression, 81 parameters

- Firm bankruptcy data. $n = 4\,748\,089$ firm-year obs.
- Subsample size: $m = 1000$.
- **Computational Time** (**CT**):
  - Computing time to obtain the equivalent of an iid draw.
  - Balances **computational cost** and **MCMC inefficiency**.
  - **Relative CT** (**RCT**)
- RCT vs HMC without subsampling: 7692.
- RCT vs state-of-the-art subsampling algorithms: 100-230.

Bias - Logistic spline regression, 81 parameters



Beyond independent data

- 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

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

- 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?

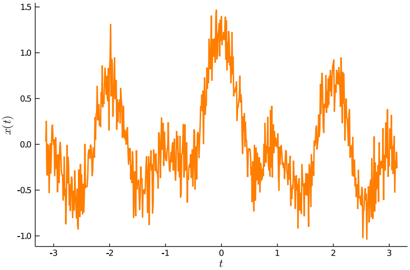
Discrete Fourier Transform - a statistical view

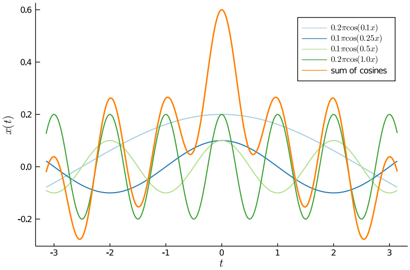
- **Fitting time trends** with **polynomial basis functions**

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon$$

- **Fitting time trends** with **periodic basis functions**

$$x_t = \beta_0 + \beta_1 \cos(0.1t) + \beta_2 \cos(t) + \beta_3 \cos(2t) + \varepsilon$$





Discrete Fourier Transform - a statistical view

- **Regress on trigonometric bases** $\cos(\omega_k t)$ and $\sin(\omega_k t)$ for **all Fourier frequencies**

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

- $\cos(\omega_k t)$ and $\sin(\omega_k t)$ are **orthogonal** functions/vectors.
- Regress on each basis separately. Each regression costs $O(n)$:

$$\hat{\beta}_k = \sum_{t=1}^{n} \cos(\omega_k t) x_t$$

- Total cost is $O(n^2)$. ☹

- **Fast Fourier Transform**: divide-and-conquer: $O(n \log n)$. ☺
- **FFT** is a **linear transformation** from $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ to $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)^T$:

$$\underset{(n\times 1)}{\hat{\boldsymbol{\beta}}} = \underset{(n\times n)}{\boldsymbol{T}} \cdot \underset{(n\times 1)}{\boldsymbol{x}}$$

Covariance and spectral density

- **(Auto)Covariance function**

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

- **Spectral density**

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

- **Discrete Fourier Transform** (DFT) of the time series

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

- The **periodogram**

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

- $\mathcal{I}(\omega_k)$ is asymptotically **unbiased for $f_\theta(\omega_k)$**. Not consistent.

AR(1) example

- $\mathrm{AR}(1)$ **process**

$$x_t = \phi x_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

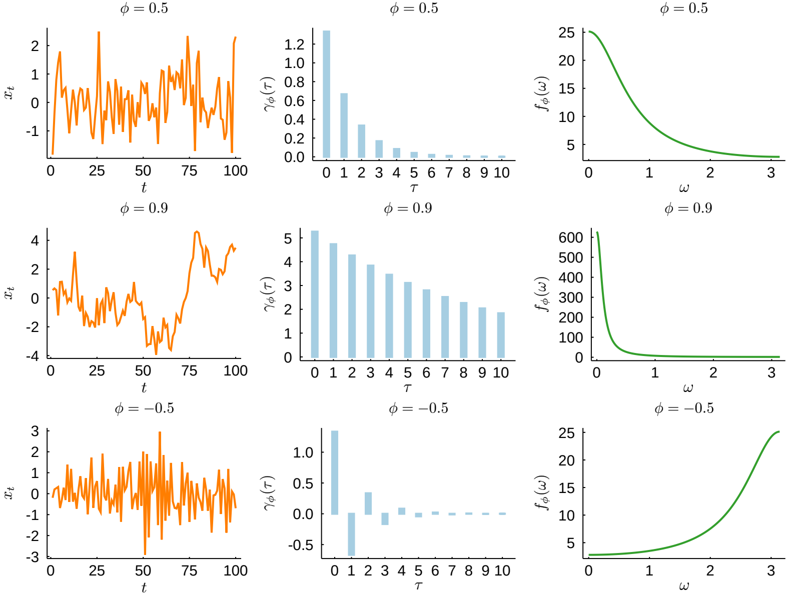
- **Covariance function**

$$\gamma_\phi(\tau) = \underbrace{\left(\frac{\sigma^2}{1-\phi^2}\right)}_{\mathbb{V}(x_t)} \underbrace{\phi^{|\tau|}}_{\text{ACF}}$$

- **Spectral density**

$$f_\phi(\omega) = \frac{1}{2\pi}\left(\frac{\sigma^2}{1+\phi^2-2\phi\cos(\omega)}\right)$$

AR(1) example



$\phi = 0.5$
$\phi = 0.5$
$\phi = 0.5$
x_t
t
$\gamma_\phi(\tau)$
τ
$f_\phi(\omega)$
ω
$\phi = 0.9$
$\phi = 0.9$
$\phi = 0.9$
x_t
t
$\gamma_\phi(\tau)$
τ
$f_\phi(\omega)$
ω
$\phi = -0.5$
$\phi = -0.5$
$\phi = -0.5$
x_t
t
$\gamma_\phi(\tau)$
τ
$f_\phi(\omega)$
ω

Whittle likelihood

- The periodogram

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

- Asympotically as $n \to \infty$

$$\mathcal{I}(\omega_k) \overset{indep}{\sim} \mathrm{Exponential}(f_{\boldsymbol{\theta}}(\omega_k)), \quad k = 1, \ldots, n$$

- Proof. Asymptotically:
  - DFT is complex Gaussian (CLT).
  - $|J(\omega_k)|^2$ is sum of two squared independent Gaussians.
  - $\chi^2_2 = \mathrm{Exp}(1/2)$.
- Same info in time series $\{x_t\}_{t=1}^n$ and periodogram $\{\mathcal{I}(\omega_k)\}_{k=1}^n$
- Whittle's asymptotic approximation of the log-likelihood:

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

Subsampling MCMC for stationary time series

- **Whittle log-likelihood is a sum**. **Subsampling!**

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

- Whittle may be **biased for small n**.
- But **subsampling** is only relevant for **large n**.
- **Subsampling** for stationary **time series** [6]
  - **Compute periodogram** before MCMC at cost $O(n \log n)$.
  - Estimate $\ell_W(\boldsymbol{\theta})$ by systematic **subsampling of frequencies**.

ARTFIMA models

- **ARIMA**(p, d, q) with integer differences $d = 0, 1, 2, \ldots$

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

- ARIMA: exponential decay of autocorrelations.
- **ARFIMA** allows for **fractional** d. **Long memory**.

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

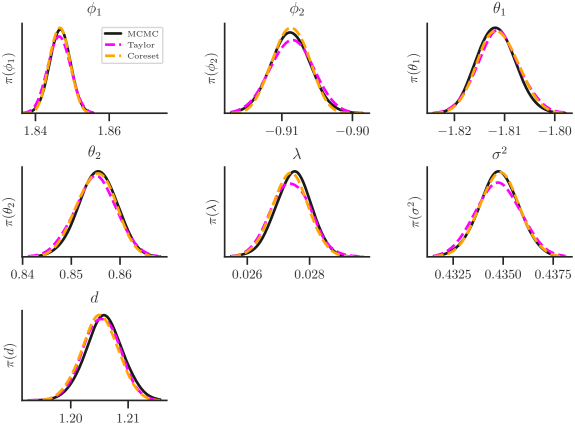
- ARFIMA: $\sum_{\tau=-\infty}^{\infty} |\gamma(\tau)| = \infty$. But stationary if $|d| < 1/2$.
- **ARTFIMA** adds **tempering** parameter $\lambda \geq 0$

$$\phi(L)(1-e^{-\lambda}L)^d y_t = \theta(L)\varepsilon_t$$

- ARTFIMA:
  - long range dependence $\gamma(\tau)$ for small τ
  - exponential decay for larger τ.
  - Stationary for all d and $\lambda > 0$.
- Autocovariances intractable. Spectral density in simple form.

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

- 450 000 hourly temperature readings during 1967-2018.
- Nearly 100 times more effective draws per minute than MCMC.



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/HMC.
- **Very large speed-ups** compared to regular MCMC/HMC.
- Future extensions:
  - **More theory** on approximation accuracy
  - Multidimensional FFT for **spatial data**
  - **Debiased 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.

- R. Salomone, M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran, "Spectral subsampling mcmc for stationary time series," *arXiv preprint arXiv:1910.13627*, 2019.
- 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.
- K.-D. Dang, M. Quiroz, R. Kohn, M.-N. Tran, and M. Villani, "Hamiltonian monte carlo with energy conserving subsampling," *arXiv preprint arXiv:1708.00955*, 2017.