Subsampling MCMC

Matias Quiroz[1,2], Mattias Villani[4,2], Minh-Ngoc Tran[3,2] and Robert Kohn[1,2]

[1]School of Economics, University of New South Wales Business School

[2]ARC Centre of Excellence for Mathematical & Statistical Frontiers

[3]Discipline of Business Analytics, University of Sydney Business School

[4]Division of Statistics and Machine Learning, Linköping University

Aug 2017

The Research Project

- **Scaling up** *Metropolis-Hastings* (MH) for **tall** (and somewhat wide) datasets.
- Scalability by **subsampling** the data.
- Talk based on four papers ...
  - Speeding Up MCMC by Efficient Data Subsampling
  - The Block Pseudo-Marginal Sampler
  - Exact Subsampling MCMC
  - Hamiltonian Monte Carlo with Energy Conserving Subsampling
    (main author Khue-Dung Dang)
- All papers are on **arXiv**.

The Metropolis-Hastings algorithm

- The **workhorse** for Bayesians for nearly **three decades**.
- Let θ and $y = (y_1, \ldots, y_n)$ denote the **parameter** and **data**.
- **Distribution** of interest

$$\pi(\theta) := p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- **Idea**: Simulate a Markov chain $\{\theta^{(j)}\}_{j=1}^{N}$ with invariant distribution $\pi(\theta)$.

The Metropolis-Hastings algorithm

- Initialize $\theta^{(0)}$ and iterate for $t = 1, 2, ...$

    1. Sample $\theta' \sim q\left(\cdot|\theta^{(t-1)}\right)$ (the **proposal distribution**)
    2. Compute the **acceptance probability**

    $$\alpha = \min\left(1, \frac{p(\mathbf{y}|\theta')p(\theta')}{p(\mathbf{y}|\theta^{(t-1)})p(\theta^{(t-1)})} \frac{q\left(\theta^{(t-1)}|\theta'\right)}{q\left(\theta'|\theta^{(t-1)}\right)}\right)$$

    3. With probability α set $\theta^{(i)} = \theta'$ and $\theta^{(i)} = \theta^{(i-1)}$ otherwise.

Subsampling the data to speed up computations

- **The likelihood** [data: $y = (y_1, \ldots, y_n)$]

$$p(y|\theta) = \exp\left(\ell_{(n)}(\theta)\right), \text{ where } \ell_{(n)}(\theta) = \sum_{k=1}^{n} \ell_k(\theta) \text{ with } \ell_k(\theta) = \log p(y_k|\theta),$$

- **MH computationally demanding** because **complete scan of the data** for each $\theta^{(j)}$.

- **Subsampling**: in each iteration
  1. Let $u = (u_1, \ldots, u_m), u_i \in \{1, \ldots, n\}$
  2. Sample m observations by Simple Random Sampling (SRS): $\ell_{u_1}(\theta), \ldots, \ell_{u_m}(\theta)$.
  3. Replace the log-likelihood $\ell_{(n)}(\theta)$ in α_{MH} by an **estimate** $\hat{\ell}_m(\theta)$.

- If $m \ll n$, a single iteration becomes much **faster**.

- **Computational Cost**: $\mathrm{CC}[\ell_{(n)}(\theta)] = O(n)$ and $\mathrm{CC}[\hat{\ell}_m(\theta)] = O(m)$.

OK to plug in an estimated likelihood in MH?

- **Pseudo-marginal MH** (Andrieu and Roberts, 2009).
- Targets the **augmented** posterior

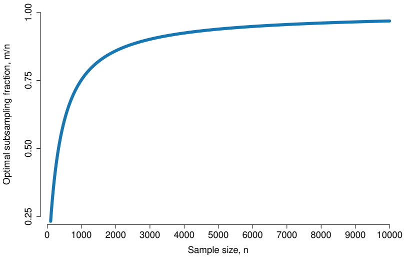
$$\overline{\pi}(\theta, u) = \frac{\hat{p}_m(y|\theta, u)p(u)p(\theta)}{p_m(y)}$$

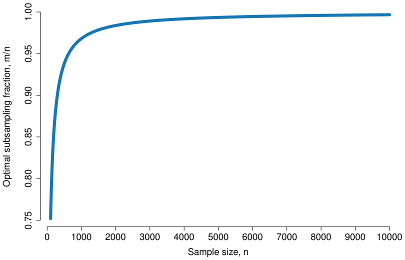
  by constructing **a Markov Chain** $\{\theta^{(j)}, u^{(j)}\}_{j=1}^{N}$ on an **augmented space**.
- θ-draws converges to $\pi(\theta)$ in distribution if the **likelihood estimator** is **unbiased**

$$\mathrm{E}_u[\hat{p}_m(y|\theta, u)] = p(y|\theta).$$

Estimator variance is crucial in pseudo-marginal MH

- **Too large** $\mathrm{V}[\hat{\ell}_m] \Rightarrow$ the chain gets **stuck**
- **Too small** $\mathrm{V}[\hat{\ell}_m] \Rightarrow$ unnecessarily **expensive** ($\mathrm{V}[\hat{\ell}_m] \propto 1/m$)
- Optimal: $\mathrm{V}[\hat{\ell}_m] \approx 1$ (Pitt et al., 2012). Gives a rule for tuning m.
- Targeting $\mathrm{V}[\hat{\ell}_m] \approx 1$ with Simple Random Sampling only obtainable by unreasonably large m.





Difference estimator and control variates

- Quiroz et al. (2016) achieve a small $\mathrm{V}[\hat{\ell}_m]$ using the difference estimator

$$\hat{\ell}_{DE}(\mathbf{y}|\theta,\mathbf{u}) \equiv \sum_{i=1}^{n} q_i(\theta) + \frac{n}{m}\sum_{k=1}^{m} d_{u_k}(\theta), \quad (1)$$

  where $d_i(\theta) = \ell_i(y_i|\theta) - q_i(\theta)$ and $q_i(\theta)$ is a **control variate** for $\ell_i(y_i|\theta)$.

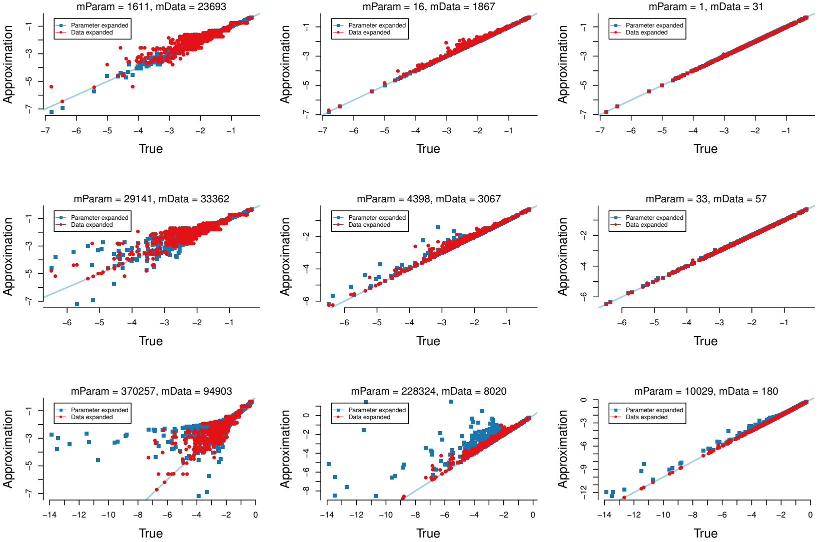
- **Data-expanded control variates** - clusters the data and Taylor expands around centroids

$$\ell_i(\mathbf{y}_i|\theta) \approx \ell_i(\mathbf{y}_{c_i}|\theta)+(\mathbf{y}_i-\mathbf{y}_{c_i})^T\nabla_{\mathbf{y}}\ell_i(\mathbf{y}|\theta)_{|\mathbf{y}=\mathbf{y}_{c_i}}+\frac{1}{2}(\mathbf{y}_i-\mathbf{y}_{c_i})^T\nabla^2_{\mathbf{y}\mathbf{y}^T}\ell_i(\mathbf{y}|\theta)_{|\mathbf{y}=\mathbf{y}_{c_i}}(\mathbf{y}_i-\quad (2)$$

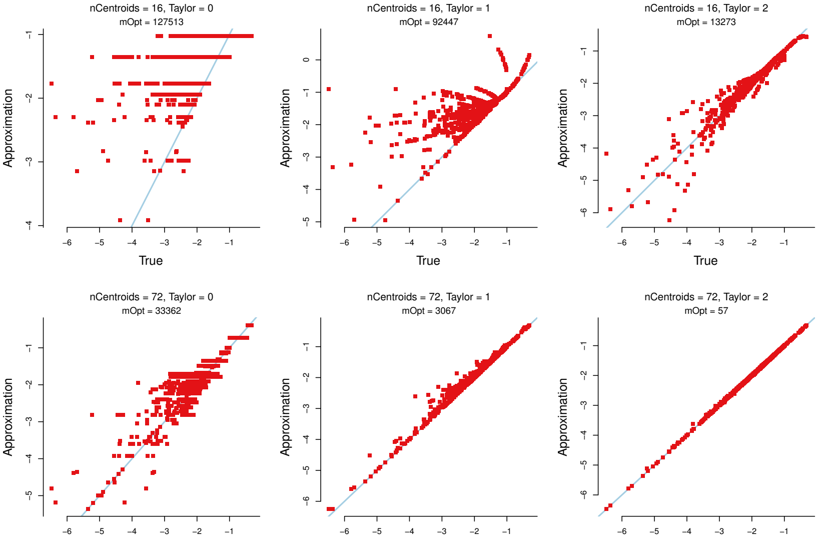
- **Parameter-expanded control variates** - Taylor expands around $\theta^\star$ (Bardenet et al. 2017)

$$\ell_i(\mathbf{y}_i|\theta) \approx \ell_i(\mathbf{y}_i|\theta^\star)+(\theta^\star-\theta)^T\nabla_{\theta}\ell_i(\mathbf{y}_i|\theta)_{|\theta=\theta^\star}+\frac{1}{2}(\theta^\star-\theta)^T\nabla^2_{\theta\theta^T}\ell_i(\mathbf{y}_i|\theta)_{|\theta=\theta^\star}(\theta^\star- \quad (3)$$

Parameter-expanded control variates sensitive to $\theta^{\star}$



Data-expanded control variates sensitive to number of clusters



Take it easy - deependent subsampling

- So far: completely new subsample in each MCMC iteration. Easy to get stuck.
- Deligiannidis et al. (2016) propose to **correlate the u** over iterations using an autoregressive proposal:

$$\mathbf{u}' = \phi \mathbf{u}^{(i-1)} + \sqrt{1-\phi^2}\varepsilon, \qquad \varepsilon \sim N(0,1). \tag{4}$$

- Quiroz et al. (2016) extend this to a subsampling context using a copula to correlate the binary u_i sampling selection indicators.
- Tran et al. (2017) propose an alternative approach to generate dependent subsampling. Partition $\mathbf{u}$ in blocks: $\mathbf{u} = (\mathbf{u}^{(1)}, ..., \mathbf{u}^{(G)})$. **Update a single block** jointly with θ in each iteration.
- By correlating the u over iterations we can tolerate a much larger $\mathrm{V}[\hat{\ell}_m]$ (Deligiannidis et al., 2016; Tran et al., 2017)
- Optimal with correlation (blocking): $\mathrm{V}[\hat{\ell}_m] \approx 234$ (Tran et al., 2017).

Approximate MCMC: near bias-correction

- The **Difference Estimator** (DE) is **unbiased for the log-likelihood**

$$\mathrm{E}[\hat{\ell}_{DE}(\mathbf{y}|\theta)] = \ell_{(n)}(\theta)$$

... but biased for the likelihood.

- Quiroz et al. (2016) propose an **approximate bias-correction**:

$$\exp\left(\hat{\ell}_{DE}(\mathbf{y}|\theta) - \sigma^2_{\hat{\ell}_{DE}}(\theta)/2\right), \tag{5}$$

where $\sigma^2_{\hat{\ell}_{DE}} = \mathrm{Var}(\hat{\ell}_{DE}(\mathbf{y}|\theta))$, which is estimated unbiasedly.
- Gives samples from a **perturbed posterior** $\pi_m(\theta) \neq \pi(\theta)$.
- Quiroz et al. (2016) show that

$$\int_\Theta |\pi_m(\theta) - \pi(\theta)|\, d\theta \leq O\left(\frac{p^3}{nm^2}\right),$$

where d is the dimension of θ.

- Example: $d = O(\sqrt{n})$ and $m = O(\sqrt{n})$ gives an $O(n^{-1/2})$ error.
- Example: d and n fixed gives an $O(m^{-2})$ error.
- Quiroz et al. (2016) also give a simple practical formula for the error.

Exact MCMC: unbiased estimate of likelihood

- **The Poisson Estimator (PE)** (e.g. Papaspiliopoulos, 2009):
  - Sample $G \sim \text{Poisson}(\lambda)$
  - Sample the selected observations within each batch: $u^{(g)}$, $g = 1, ..., G$.
  - Compute a log-likelihood estimator $\hat{\ell}^{(g)}$ for each batch.
  - Let a be a constant and compute

$$\hat{p}_m(y|\theta, u, G) = \exp(a + \lambda) \prod_{g=1}^{G} \left(\frac{\hat{\ell}^{(g)} - a}{\lambda} \right)$$

- Two sources of randomness in PE: i) **G** and ii) $u^{(g)}$, $g = 1, ..., G$.
- **PE is unbiased for the likelihood**, but usually $\mathrm{V}[\hat{\ell}_{PE}] > \mathrm{V}[\hat{\ell}_{DE}]$.
- We do **correlated Pseudo-marginal** on both G (copula) and $u^{(g)}$.
- The Rhee-Glynn estimator (Bardenet et al. 2017) has larger variance.

Pseudo-marginal MH with the Poisson Estimator

- Plug in the **Poisson Estimator** in the **pseudo-marginal MH algorithm**.
- Pseudo-marginal is just MH on an **extended target**

$$\overline{\pi}(\theta, u, G) \propto \hat{p}_m(y|\theta, u, G)p(u, G)p(\theta).$$

- **Jacob and Thiery (2015)**: PE is a.s. **nonnegative if only if** a is a lower bound for $\hat{\ell}^{(g)}$ for all $g = 1, ..., G$.
- **Two problems with obtaining a lower bound** a:
  1. We need to know ℓ_k, for $k = 1, \ldots, n$. **No point in subsampling**!
  2. Too conservative a. Gives **HUGE** variance.
- **Quiroz et al. (2017)**: Use soft lower bound $\tilde{a}$ such that $\Pr(\hat{p}_m(y|\theta, u) \geq 0)$ is close, but not equal, to unity.
- Soft lower bound $\Rightarrow$ Poisson estimator can be negative.

Pseudo-marginal MH with a non-positive estimator

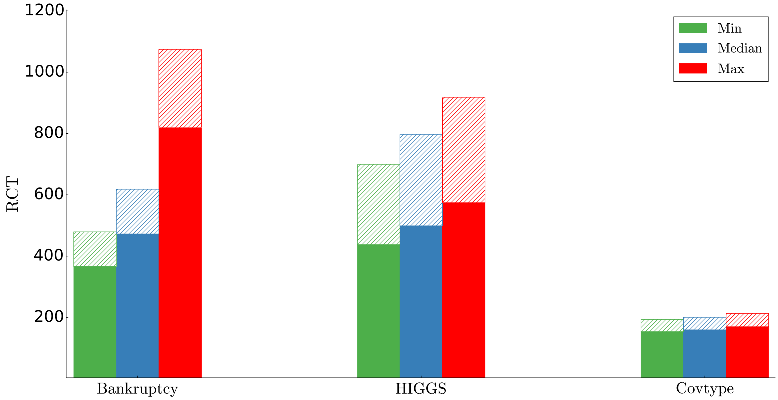
- **Lyne et al. (2015)**:
  - Run **Pseudo-marginal** on absolute measure $|\overline{\pi}(\theta, u, G)|$, but store the sign of $\overline{\pi}(\theta, u, G)$ in each iteration.
  - Use **importance sampling** on the iterates to correct for the sign, $s(\theta^{(j)})$, to estimate $\mathrm{I} = \mathrm{E}_{\theta}[\psi(\theta)]$ for any function $\psi(\theta)$

$$\hat{\mathrm{I}} = \frac{\sum_{j=1}^{N} \psi(\theta^{(j)}) s(\theta^{(j)})}{\sum_{j=1}^{N} s(\theta^{(j)})}.$$

- **Optimal** $\mathrm{V}[\hat{\mathrm{I}}]$: All signs positive (or negative). **Worst when 50-50**.
- Decreasing $\tilde{a} \Rightarrow$ increases probability of positive signs, but increases $\mathrm{V}[\hat{\ell}_{PE}]$.
- Increasing $\tilde{a} \Rightarrow$ decreasing $\mathrm{V}[\hat{\ell}_{PE}]$, but prob of positive signs closer to 0.5.

Logistic Regression examples

- Logistic regression on three datasets:
  - Bankruptcy - $n = 4.7$ millons and $d = 9$
  - HIGGS - $n = 1.1$ millions and $d = 21$
  - CovType - $n = 550K$ and $d = 11$
- Combination of data- and parameter expanded control variates.
- Blocking u.



HMC with energy conserving subsampling

- **Hamiltonian Monte Carlo** (HMC) has proven to be successful in **high-dimensional spaces**.
- HMC augments the posterior $\pi(\theta)$ with fictitious **momentum variables** $\mathbf{m} \in \mathbb{R}^d$, and targets

$$\bar{\pi}(\theta, \mathbf{m}) \propto \exp(-\mathcal{H}(\theta, \mathbf{m})), \tag{6}$$

where $\mathcal{H}$ is the so called **Hamiltonian**

$$\mathcal{H}(\theta, \mathbf{m}) = \mathcal{U}(\theta) + \mathcal{K}(\mathbf{m}), \tag{7}$$

where in HMC

$$\mathcal{U}(\theta) = -\log[p(\mathbf{y}|\theta)p(\theta)] \text{ and } \mathcal{K}(\mathbf{m}) = \frac{1}{2}\mathbf{m}^T\mathbf{M}^{-1}\mathbf{m}, \tag{8}$$

and $\mathbf{M}$ is a $d \times d$ positive definite matrix.
- Initial momentum from $\mathbf{m} \sim N(0, \mathbf{M})$ is used to propagate both θ and $\mathbf{m}$ over time t along a trajectory mapped out by the Hamiltonian dynamics

$$\nabla_t \theta = \nabla_{\mathbf{m}} \mathcal{H}(\theta, \mathbf{m}) = \mathbf{M}^{-1}\mathbf{m} \tag{9}$$

$$\nabla_t \mathbf{m} = -\nabla_\theta \mathcal{H}(\theta, \mathbf{m}) = -\nabla_\theta \mathcal{U}(\theta). \tag{10}$$

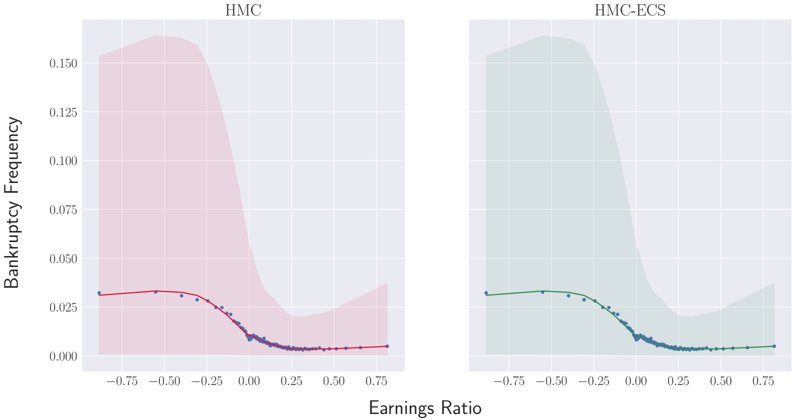
HMC with energy conserving subsampling

- HMC requires evaluating the gradient throughout the many steps of simulated dynamics. **Computationally costly**, especially for large data sets.
- Betancourt (2015) **estimates the gradient** in the simulated dynamics from a random **subsample**. MH acceptance probability is compute on the full dataset. Conclusion: α_{MH} decreases quickly with d.
- **Our approach**: Pseudo-marginal with α_{MH} evaluated using likelihood estimates on the **same data subset** as used in simulating the dynamics.
- HMC-within-Gibbs Algorithm
  - i $u|\theta, \mathbf{m}, y$ - MH-step for **subsample**
  - ii $\theta, \mathbf{m}|u, y$ - HMC-step for **parameters** θ and **momentum variables** p
- Our **HMC-ECS** algorithm **conserves the energy** (α_{MH} doesn't drop).
- **Approximate** (Quiroz et al., 2016) or **Exact** (Quiroz et al., 2017).

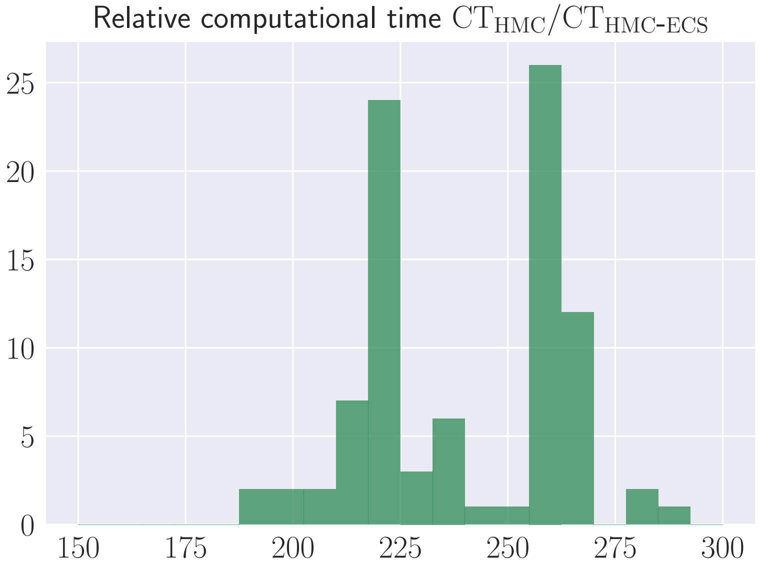
HMC-ECS on firm bankruptcy data

- Before:
  - $n = 4.7$ millons data points.
  - **logistic regression** with $d = 9$ parameters.
- Here:
  - $n = 4.7$ millons data points.
  - **additive splines logistic regression** with $d = 89$ parameters (10 knots + linear term for each covariate + intercept)
- HMC: $\alpha_{MH} = 81.8\%$
- HMC-ECS: $\alpha_{MH} = 79.3\%$

Approximate HMC-ECS on firm bankruptcy data



Approximate HMC-ECS on firm bankruptcy data



Summary

- Scalable frameworks for efficient data subsampling to speed up MCMC:
  - **Approximate** - Faster, approximate but controlled error
  - **Exact** - Slower than approximate, but guaranteed to be exact.
- **Variance reduction** by:
  - control variates
  - dependent subsamples
- **HMC with energy conserving subsampling** for high-dimensional parameter spaces.
- 1-3 orders of magnitude as many effective draws per computational time compared to MH on full dataset.

References

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

Dang, K-D., Quiroz, M., Kohn, R., Tran, M.-N. and M., Villani, M. (2017). Hamiltonian Monte Carlo with Energy Conserving Subsampling. *arXiv preprint*.

Deligiannidis, G., Doucet, A. and Pitt, M., K. (2016). The correlated pseudo-marginal method. *arXiv preprint arXiv:1511.04992v3*.

Jacob, P. E. and Thiery, A. H. (2015). On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769-784.

Lyne, A. M., Girolami, M., Atchade, Y., Strathmann, H., and Simpson, D. (2015). On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods.*Statistical Science*, 30(4):443-467.

Maclaurin, D. and Adams, R. P. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence* .

References, cont.

Papaspiliopoulos, O. (2009). A methodological framework for Monte Carlo probabilistic inference for diffusion processes. URL: `http://wrap.warwick.ac.uk/35220/1/WRAP_Papaspiliopoulos_09-31w.pdf`.

Pitt, M. K., Silva, R. d. S., 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., Villani, M., Kohn, R. and Tran, M.-N. (2016). Speeding up MCMC by efficient data subsampling. *arXiv preprint arXiv:1404.4178v4*.

Quiroz, M., Tran, M.-N., Villani and M., Kohn, R. (2017). Exact subsampling MCMC. *arXiv preprint arXiv:1603.08232v3*.

Tran, M.-N., Kohn, R., Quiroz, M. and Villani, M. (2017). The block pseudo-marginal sampler. *arXiv preprint arXiv:1603.02485v4*.

AR-process toy example

- **Model** M_1:

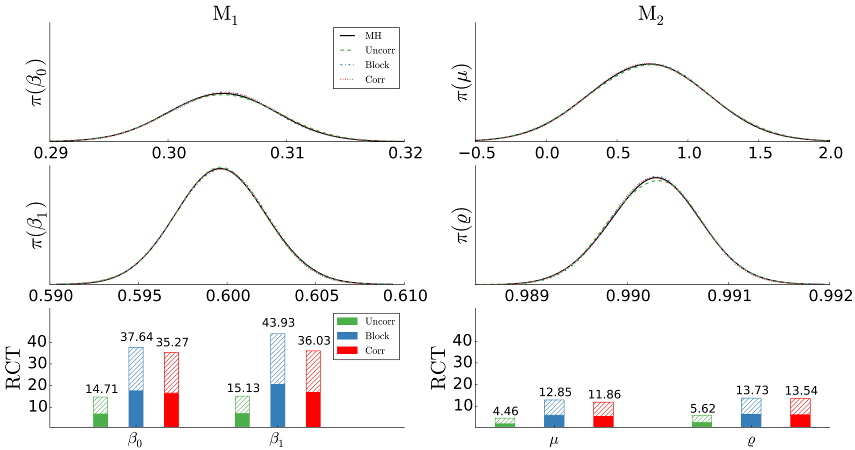
$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t, \quad \epsilon_t \sim t(\nu = 5) \text{ iid.}$$

- **Model** M_2:

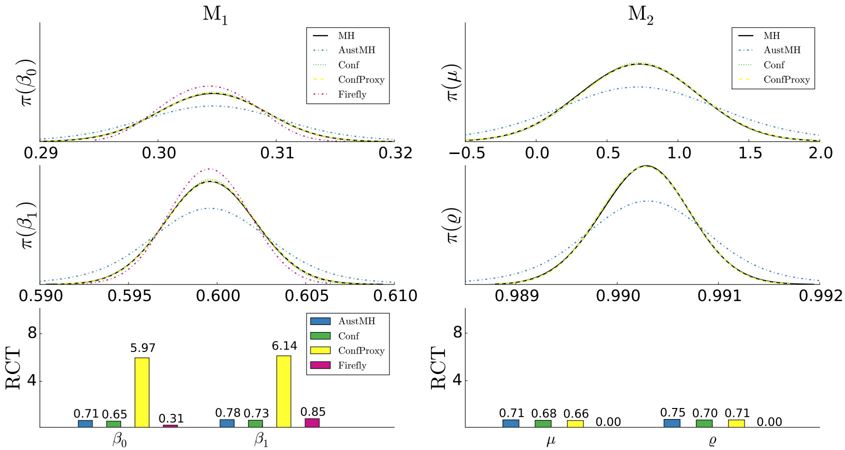
$$y_t = \mu + \rho(y_{t-1} - \mu) + \epsilon_t, \quad \epsilon_t \sim t(\nu = 5) \text{ iid.}$$

- **Generate** $100,000$ observations from the DGP.

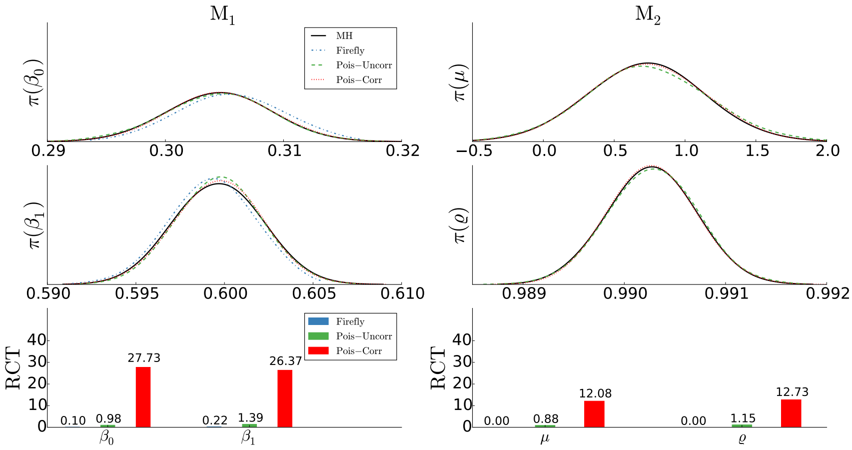
AR process example - approximate approaches



AR process example - alternative approaches



AR process example - exact approaches



AR process example - exact approaches