SUBSAMPLING MCMC

Mattias Villani

Division of Statistics and Machine Learning
Department of Computer and Information Science
Linköping University

CLASS OVERVIEW

- Data subsampling in MCMC
- Gaussian processes and Optimization
- Distributed MCMC
- Topic models for Text

Lecture Overview

- Very brief intro to MCMC
- Pseudo-marginal Metropolis-Hastings (PMMH)
- A PMMH approach to data subsampling
- Alternative subsampling approaches

WHY DATA SUBSAMPLING?

- **Big data**. Data sets are getting bigger and bigger.
- **Bayesian inference** is the way to go.
- Bayesian inference is usually implemented using MCMC.
- **MCMC** can be very **slow** on large data sets. Evaluate the data density for each observation.
- The **likelihood can be costly** to evaluate (also on small data).

MCMC - THE BASIC IDEA

- Explore complicated joint posterior distributions $p(\theta|\mathbf{y})$ by **simulation**.
- Set up **Markov chain** $\theta^{(i)}|\theta^{(i-1)}$ for θ with $p(\theta|\mathbf{y})$ as **stationary distribution**.
- Draw are **autocorrelated ...**
- ...but sample averages ($\bar{\theta} = N^{-1} \sum_{i=1}^{N} \theta^{(i)}$) still converge to posterior expectations ($E(\theta|\mathbf{y})$).
- High autocorrelation means fewer **effective draws**

$$Var(\bar{\theta}) = \frac{\sigma^2}{N} \left(1 + 2 \sum_{k=1}^{\infty} \rho_k \right)$$

The Metropolis-Hastings algorithm

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

  1. Sample $\theta_p \sim q\left(\cdot|\theta^{(i-1)}\right)$ (the **proposal distribution**)
  2. Compute the **acceptance probability**

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

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

MCMC WITH AN UNBIASED LIKELIHOOD ESTIMATOR

- The **full likelihood** $p(\mathbf{y}|\theta)$ is **intractable** or very **costly to evaluate**.
- **Unbiased estimator** $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$ of the likelihood is available

$$\int \hat{p}(\mathbf{y}|\theta, \mathbf{u}) p(\mathbf{u}) d\mathbf{u} = p(\mathbf{y}|\theta)$$

- $u \sim p(u)$ are auxilliary variables used to compute $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$.
- **Importance sampling/particle filters** for latent variable $(\mathbf{x})$ models

$$p(\mathbf{y}|\theta) = \int p(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{x} = \int \frac{p(\mathbf{y}, \mathbf{x}|\theta)}{q_\theta(\mathbf{x})} q_\theta(\mathbf{x}) d\mathbf{x}$$

$$\hat{p}(\mathbf{y}|\theta, \mathbf{u}) = \frac{1}{m} \sum_{k=1}^{m} \frac{p(\mathbf{y}, \mathbf{x}^{(k)}|\theta)}{q_\theta(\mathbf{x}^{(k)})} \text{ where } \mathbf{x}^{(k)} \overset{iid}{\sim} q_\theta(\cdot)$$

  and the u's are the random numbers used to simulate from q_θ.
- **Subsampling**: $\mathbf{u}$ are indicators for selected observations.
- Let m be the number of u's (particles/subsample size).

MCMC WITH A UNBIASED LIKELIHOOD ESTIMATOR

- But is it OK to use a noisy estimate $\hat{p}(\mathbf{y}|\theta,\mathbf{u})$ of the likelihood in MH?
- The joint density

$$\tilde{p}(\theta,\mathbf{u}|\mathbf{y}) = \frac{\hat{p}(\mathbf{y}|\theta,\mathbf{u})p(\theta)p(\mathbf{u})}{p(\mathbf{y})}$$

has the correct marginal density $p(\theta|\mathbf{y})$ if $\hat{p}(\mathbf{y}|\theta,\mathbf{u})$ is **unbiased**

$$p(\mathbf{y}|\theta) = \int \hat{p}(\mathbf{y}|\theta,\mathbf{u})p(\mathbf{u})d\mathbf{u}$$

- This is easily seen from

$$\int \tilde{p}(\theta,\mathbf{u}|\mathbf{y})d\mathbf{u} = \frac{p(\theta)}{p(y)}\int \hat{p}(\mathbf{y}|\theta,\mathbf{u})p(\mathbf{u})d\mathbf{u} = \frac{p(\theta)p(\mathbf{y}|\theta)}{p(\mathbf{y})} = p(\theta|\mathbf{y})$$

The Pseudo-Marginal MH (PMMH) algorithm

- Initialize $\left(\theta^{(0)}, u^{(0)}\right)$ and iterate for $i = 1, 2, \ldots$

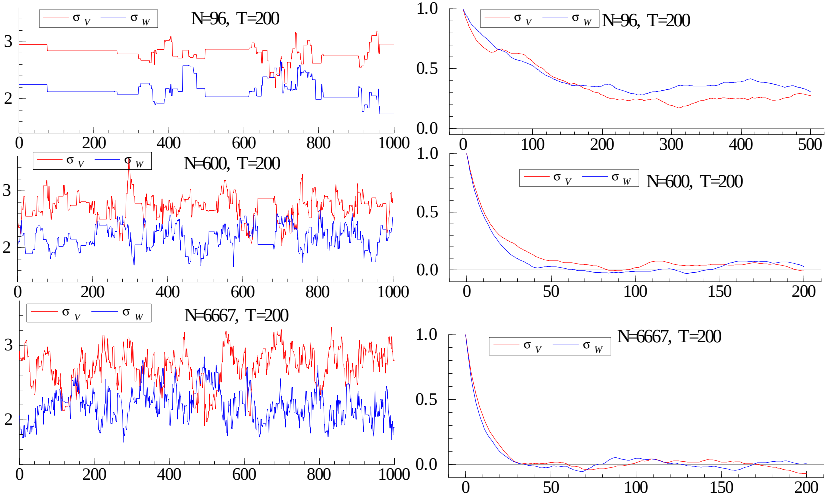
  1. Sample $\theta_p \sim q\left(\cdot|\theta^{(i-1)}\right)$ and $u_p \sim p_\theta(u)$ to obtain $\hat{p}(y|\theta_p, u)$
  2. Compute the **acceptance probability**

  $$\alpha = \min\left(1, \frac{\hat{p}\left(\mathbf{y}|\theta_p, u_p\right) p(\theta_p)}{\hat{p}\left(\mathbf{y}|\theta^{(i-1)}, u^{(i-1)}\right) p(\theta^{(i-1)})} \frac{q\left(\theta^{(i-1)}|\theta_p\right)}{q\left(\theta_p|\theta^{(i-1)}\right)}\right)$$

  3. With probability α set $\left(\theta^{(i)}, u^{(i)}\right) = (\theta_p, u_p)$ and $\left(\theta^{(i)}, u^{(i)}\right) = \left(\theta^{(i-1)}, u^{(i-1)}\right)$ otherwise.

- This MH has $\tilde{p}(\theta, \mathbf{u}|\mathbf{y})$ as stationary distribution with marginal $p(\theta|\mathbf{y})$.
- This result holds **irrespective of the variance** of $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$.
- **It's OK to replace the likelihood with an unbiased estimate! [1]**

THE NUMBER OF PARTICLES IN A STATE-SPACE MODEL (FROM MIKE PITT)



Optimal m - keep the variance around 1

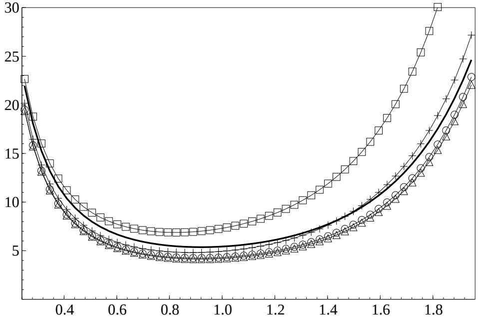
- **Large** $m \Rightarrow$ **costly** $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$, but **efficient MCMC**.
- **Small** $m \Rightarrow$ **inexpensive** $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$, but **inefficient MCMC**.
- Define the estimation error

$$z = \ln \hat{p}(\mathbf{y}|\theta, \mathbf{u}) - \ln p(\mathbf{y}|\theta)$$

  and $\sigma_z^2 = Var(z)$.
- Assumptions:
  - z is independent of θ
  - z is Gaussian
- **Optimal** m to maximize effective sample size per computational unit:
  - For good proposals for θ, set m so that $\sigma_z \approx 1$.
  - For bad proposals for θ, set m so that $\sigma_z \approx 1.7$.
  - Targeting $\sigma_z \approx 1.7$ when $\sigma_z \approx 1$ is optimal is much worse than targeting $\sigma_z \approx 1$ when $\sigma_z \approx 1.7$ is optimal.
  - Conservative is good: $\sigma_z \approx 1$. [2, 3]

Optimal m - Var(z)≈ 1 (Pitt et al. 2012 [2])

- Effective sample size per minute as a function of σ_z



- Squares - $IF = 1$
- Crosses - $IF = 4$
- Circles - $IF = 20$
- Triangles - $IF = 80$
- Solid - Perfect proposal

Estimating the likelihood by subsampling

- **Log-likelihood** for independent observations:

$$\ell(\theta) = \ln p(y_1, ..., y_n|\theta) = \sum_{i=1}^{n} \ln p(y_i|\theta)$$

- **Log-likelihood contribution** of ith observation:

$$\ell_i(\theta) = \ln p(y_i|\theta)$$

- Applicable as long as we have **independent pieces of data**:
  - **Longitudinal data**. Subjects are independent, the observations for a given subject are not.
  - **Time series** with kth order Markov structure: $y_t|y_{t-1}, ..., y_{t-k}$.
  - **Textual data**. Documents are independent. Words within documents are not.
- Estimating the log-likelihood (a sum) is like estimating a population total. **Survey sampling**.

Simple random sampling does not work

- **Simple random sampling** (**SRS**) with replacement. At the jth draw:

$$\Pr(u_j = k) = \frac{1}{n}, \; k = 1, ..., n \text{ and } j = 1, ..., m$$

- Let $\mathbf{u} = (u_1, ..., u_m)$ record the sampled observations.
- **Unbiased estimator** of the **log-likelihood** (more on this later)

$$\hat{\ell}_{SRS}(\theta) = \frac{n}{m} \sum_{j=1}^{m} \ell_{u_j}(\theta)$$

- $\hat{\ell}_{SRS}(\theta)$ is **extremely variable**, even when m/n is large.
- **PMCMC stuck** when $\hat{\ell}_{SRS}(\theta)$ is sampled in the extreme right tail.
- Sampling without replacement does not help.

The difference estimator

- Idea: reduce the variance of $\hat{\ell}$ by control variates.
- Let $w_k(\theta)$ be an cheap approximation of $\ell_k(\theta)$.
- Trivial decomposition:

$$
\begin{aligned}
\ell(\theta) &= \sum_{k \in F} w_k(\theta) + \sum_{k \in F} [\ell_k(\theta) - w_k(\theta)] \\
&= \sum_{k \in F} w_k(\theta) + \sum_{k \in F} d_k(\theta)
\end{aligned}
$$

- $\sum_{k \in F} w_k(\theta)$ is known.
- $\sum_{k \in F} d_k(\theta)$ can be estimated by sampling like any population total.
- If $w_k(\theta)$ is a decent proxy for $\ell_k(\theta)$, the **differences** $d_k(\theta)$ should be **roughly equal in size** and small. SRS works!

Proxies by data clustering

- Idea: $\ell(\theta; y, \mathbf{x})$ and $\ell(\theta; y', \mathbf{x}')$ are likely to be similar when $(y, \mathbf{x})$ and $(y', \mathbf{x}')$ are close.

- Approximate $\ell(\theta; y, \mathbf{x}) \approx \ell(\theta; y_c, \mathbf{x}_c)$ where $(y_c, \mathbf{x}_c)$ is the **nearest cluster centroid**.

- Even better: use 2nd order **Taylor expansion** of $\ell(\theta; y, \mathbf{x})$ around $y_c, \mathbf{x}_c$ as proxy.

- Difference estimator can be computed using computations only at the C centroids. Data and parameters factorize in $\sum_{k \in F} w_k(\theta)$. Scalable!

- Curse of dimensionality can be dealt with by **dimension reduction** (clustering in PCA space) or other subspace clustering methods.

Bias-correction

- So far: **unbiased** estimators of the **log-likelihood**.
  We need unbiasedness for the likelihood.
- Let z denote the **error** in the log-likelihood estimate:

$$\hat{\ell}(\theta) = \ell(\theta) + z$$

  and $\sigma_z^2 = \mathrm{Var}(z)$.
- Assume $z \sim N(0, \sigma_z^2)$ and that σ_z^2 known. Then

$$\exp\left[\hat{\ell}(\theta) - \sigma_z^2/2\right]$$

  is unbiased for the likelihood.
- What if z is not Gaussian and σ_z^2 is estimated unbiasedly?

MCMC WITH A BIASED LIKELIHOOD ESTIMATOR

- **Biased likelihood estimator**:

$$\hat{p}_{m,n}(\mathbf{y}|\theta,\mathbf{u}) = \exp\left[\hat{\ell}(\theta) - \hat{\sigma}_z^2/2\right]$$

- Define:
  - Perturbed likelihood: $p_{m,n}(\mathbf{y}|\theta) = \int \hat{p}_{m,n}(\mathbf{y}|\theta,\mathbf{u})p(\mathbf{u})d\mathbf{u}$
  - Perturbed marginal data density: $p_{m,n}(\mathbf{y}) = \int p_{m,n}(\mathbf{y}|\theta)p(\theta)d\theta$
  - **Perturbed posterior**: $p_{m,n}(\theta|\mathbf{y}) = p_{m,n}(\mathbf{y}|\theta)p(\theta)/p_{m,n}(\mathbf{y})$.

- A PMMH scheme targeting

$$\tilde{\pi}_{m,n}(\theta,\mathbf{u}|\mathbf{y}) = \frac{\hat{p}_{m,n}(\mathbf{y}|\theta,\mathbf{u})p(\theta)p(\mathbf{u})}{p_{m,n}(\mathbf{y})}$$

has $p_{m,n}(\theta|\mathbf{y})$ as invariant distribution.

MCMC WITH A BIASED LIKELIHOOD ESTIMATOR

- Let $m = O(n^{\gamma})$ [subsample size increases with n]
- Let $d_k = O(n^{-\alpha})$ [the quality of the proxies improve with n]

THEOREM

$$\frac{|p_{m,n}(\theta|\mathbf{y}) - p(\theta|\mathbf{y})|}{p(\theta|\mathbf{y})} \leq \begin{cases} O(m^{-1}) \\ O(n^{-a}) \end{cases}$$

where

$$a = \min \begin{cases} 3(\alpha - 1) + 2\gamma \\ 2(\alpha - 1) + \gamma \\ 4(\alpha - 1) + 3\gamma \end{cases}$$

Correlated PMMH

- Deligiannidis et al (2015) [4] and Dahlin et al. (2015) [5] propose to use auxillary variables $\mathbf{u}$ that are **correlated over the MCMC iterations**.
- **Autoregressive proposal**

$$\mathbf{u}_p = \rho \mathbf{u}_c + \sqrt{1-\rho^2}\varepsilon, \qquad \varepsilon \sim N(0, I_p)$$

- Correlated u's give a **lower variance** for the estimated likelihood ratio in the MH acc. prob.

$$\frac{\hat{p}(\mathbf{y}|\theta_p, \mathbf{u}_p)}{\hat{p}(\mathbf{y}|\theta_c, \mathbf{u}_c)}$$

- We can now tolerate a much larger variance of the likelihood estimator without getting stuck.

Correlated PMMH for Subsampling

- Quiroz et al. (2014, 3rd revision [6]) **correlate the binary selection indicators** over the MCMC iterations in data subsampling.
- **Gaussian copula** for binary variables. Latent variables in copula follow autoregressive proposal.
- Equivalent: **Markov Chain** $Pr(u_p = i | u_c = j) = \pi_{ij}$, where π_{00} and π_{11} are close to one, and the expected subsample size $E(u) = m^{\star}/n$ is set by the user.
- **Only small changes in the subsample** at a given iteration.
- We can tolerate a larger variance of $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$. **Smaller subsamples**!
- **Block-wise PMMH** [7] blocks the u's and updates a single block in every MCMC iteration. See my talk tomorrow.

Firefly Monte Carlo algorithm [8]

- **Augmenting** the data points with subset selection indicators.
- Assume a **lower bound** $b_k(\theta) \leq L_k(\theta)$ for likelihood contributions.
- **Augment** each y_k with a **binary indicator** z_k with distribution

$$p(z_k|y_k,\theta) = \left(\frac{L_k(\theta) - b_k(\theta)}{L_k(\theta)}\right)^{z_k} \left(\frac{b_k(\theta)}{L_k(\theta)}\right)^{1-z_k}$$

- Marginalizing out the z_k returns the posterior $p(\theta|\mathbf{y})$. [8]
- The likelihood contributions $L_k(\theta)$ only appears in terms where $z_k = 1$

$$L_k(\theta)p(z_k|y_k,\theta) = \begin{cases} L_k(\theta) - b_k(\theta) & \text{if } z_k = 1 \\ b_k(\theta) & \text{if } z_k = 0 \end{cases}$$

- **Gibbs sampling:** sample z_k from its full conditional. If **the bound is tight** most z_k will be zero, i.e. small subsample.
- Posterior on augmented space ($\prod_{k=1}^n b_k(\theta)$ often in $O(1)$ time)

$$p(\theta,\mathbf{z}|\mathbf{y}) = p(\theta)\prod_{k=1}^{n} b_k(\theta) \prod_{k:z_k=1}\left(\frac{L_k(\theta) - b_k(\theta)}{L_k(\theta)}\right)$$

Statistical tests with a certain confidence

- The following methods are of this nature
  1. Austerity Metropolis-Hastings (Korattikaria et al., 2014) [9].
  2. Confidence sampler (Bardenet et al., 2014) [10].
  3. Confidence sampler with proxies (Bardenet et al., 2015) [11]
- **Key idea**: The acceptance decision in **Metropolis-Hastings**

  $u \leq \alpha(\theta, \theta') = \exp\left[\ell(\theta') - \ell(\theta)\right]$ (symmetric proposal and flat prior)

  can be written

  $$\log(u) \quad \leq \quad \ell(\theta') - \ell(\theta) = n\left[\bar{\ell}(\theta') - \bar{\ell}(\theta)\right], \quad \left[\bar{\ell}(\theta) = l(\theta)/n\right].$$
- Let $\Lambda_n(\theta, \theta') = \bar{\ell}(\theta') - \bar{\ell}(\theta)$. We see that M-H **accepts a move if**

  $$\Lambda_n(\theta, \theta') \geq \frac{1}{n}\log(u) = \psi_0(\theta', \theta)$$

  and **rejects if the opposite**.
- **Base the acceptance decision** on a subset of data of size m, i.e. use $\Lambda_m^*(\theta, \theta')$ to determine if $\Lambda_n(\theta, \theta') > \psi_0(\theta, \theta')$

Statistical tests with a certain confidence

- **Korattikaria et al. (2014)** [9]: Statistical test:

$$
\begin{aligned}
H_0 &: \Lambda_n(\theta, \theta') = \psi_0(\theta, \theta') \\
H_1 &: \Lambda_n(\theta, \theta') \neq \psi_0(\theta, \theta')
\end{aligned}
$$

- **Normalized test statistic** is asymptotically Student-t by CLT.
- **Algorithm**: Start with a **small fraction of data**.
  1. Can the decision of **rejecting** H_0 be taken with a specified error probability?
  2. **If Yes**: accept the sample (if $\Lambda^*_m(\theta, \theta') > \psi_0$) and reject if the opposite
  3. **If No**: sample more data and ask **1.** again.
- **Drawbacks:** Relies on many CLTs. Approx may be poor when CLT is violated [11].

Statistical tests with a certain confidence

- Bardenet et. al. (2014) [10]: Use **concentration bounds** (no CLT):

$$\Pr(|\Lambda_m^*(\theta,\theta') - \Lambda_n(\theta,\theta')| \leq c_m) \geq 1-\delta,$$

  where c_m is the **concentration bound** and δ is the user specified error probability.
- **Keep sampling** data until we know that the event

$$\left\{|\Lambda_m^*(\theta,\theta') - \Lambda_n(\theta,\theta')| \leq c_m\right\}$$

  is **true** (with a certain "confidence").
- **Accept** the sample if $\Lambda_n^*(\theta,\theta') > \psi_0$, otherwise reject.
- **Important:** c_n is a function of **the variance** and **the range** of the "population"

$$\ell_k(\theta') - \ell_k(\theta).$$

- **Drawback: the range is typically** $O(n)$ in non-trivial models (Bardenet et. al., 2015) [11].

Statistical tests with a certain confidence

- **Bardenet et. al. (2015) [11]** improves on this idea by introducing proxies $w_k(\theta, \theta') \approx \ell_k(\theta') - \ell_k(\theta)$.
- **Control-variates** to reduce the variance.
- The proxies are obtained by a **second order Taylor approximation** w.r.t the parameter.
- **Same procedure** as in Bardenet et. al. (2014), but now on

$$\ell_k(\theta') - \ell_k(\theta) - w_k(\theta, \theta')$$

- **The range** in the concentration bound is replaced by an estimate of **the remainder** of the Taylor series via the **Taylor-Lagrange inequality**.
- **Major drawback: Very difficult** to obtain a (tight) bound on the third order derivatives, even for reasonably **simplistic models.**

AR process example

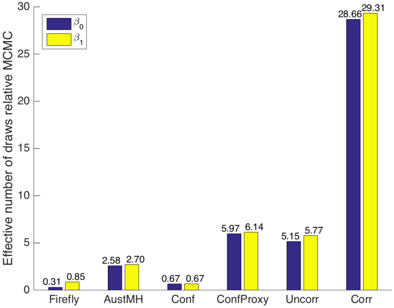
- **AR(1) process with student-t noise**

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

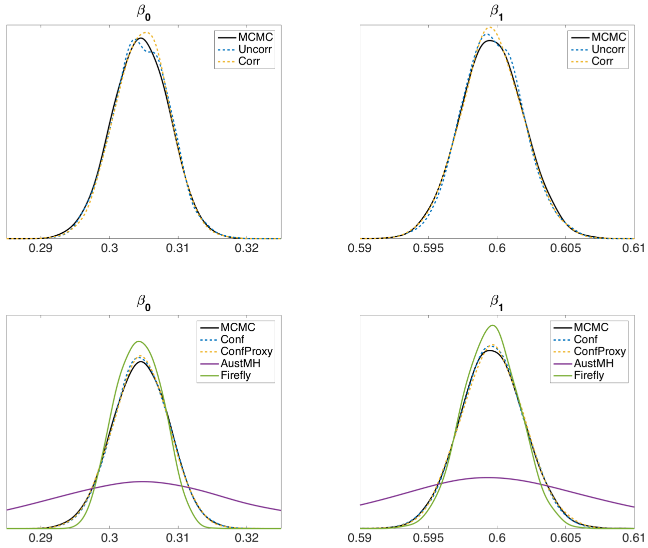
- **Aim:** posterior of β_0, β_1 with known $\nu = 5$ based on a sample with 100,000 observations.
- Posterior is more or less a spike. Confidence sampler should preform well.

Subsample fraction - AR

Uncorr	Corr	Conf	ConfProxy	AustMH	Firefly
0.055	0.023	1.493	0.161	0.197	0.100



AR process example



Steady state AR process example

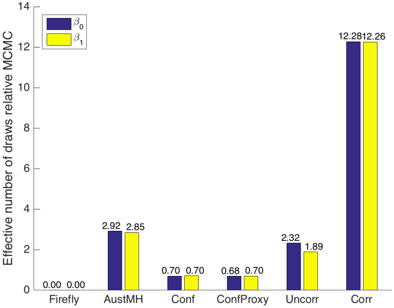
- **AR(1) process with student-t noise**

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

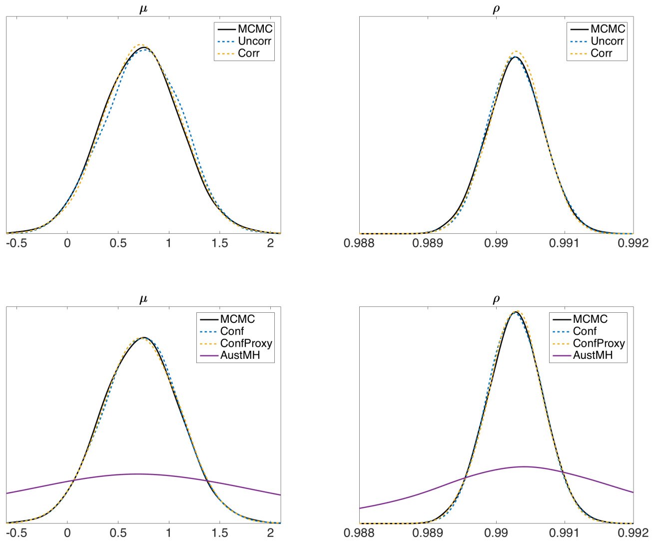
- **Aim:** posterior of μ, ρ with known $\nu = 5$ based on a sample with 100,000 observations.
- ρ is close to one in the data, so posterior of μ concentrates very slowly.

Subsample fraction - Steady State AR

Uncorr	Corr	Conf	ConfProxy	AustMH	Firefly
0.159	0.059	1.489	1.497	0.189	0.134



Steady State AR



Logistic regression

- Bankruptcy of Swedish firms (**binary response**).
- Annual observations during 1991-2008.
- Nearly **5 million** firm-year observations.
- Eight **covariates**: earnings before interest and taxes, total liabilities, cash and liquid assets, tangible assets, log deflated total sales, log firm age, GDP growth rate and the repo rate.

Results logistic regression

- Target $\sigma_Z^2 = 7$. $C = 3\%$ of n and subsample size $m = 0.5\%$ of n.
- Average $RED = 3.51$ for **uncorrelated** PMMH.
- **Average** $RED = 10.51$ for **correlated** PMMH.

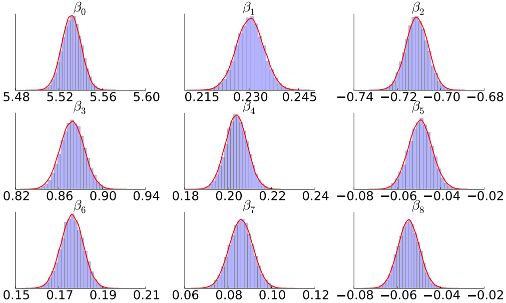


Figure: Marginal posteriors from MCMC on all data (kernel density estimates in red) and correlated PMMH (histograms).

Not only for tall data

- Discrete-time survival data.
- Firm bankruptcy.
- Dataset used has only 2000 firms.
- **Weibull regression** with covariates in both parameters.
- Random intercept for each firm. **Time-consuming evaluation of log-likelihood contributions**.
- Proxies obtained by crude and fast numerical integration of random effects.

NOT ONLY FOR TALL DATA

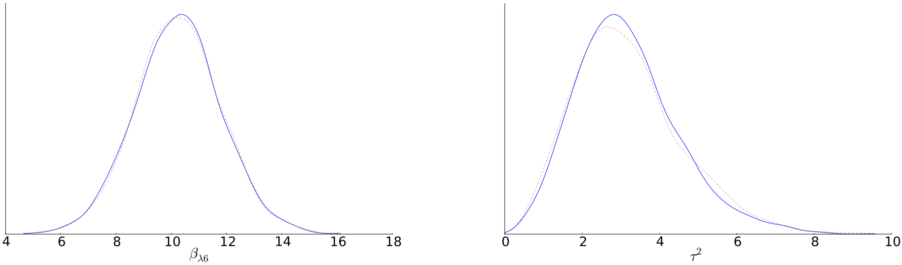


FIGURE 5. Marginal posterior distributions for MCMC (solid blue line) vs PMCMC (dashed red line) using a RWM proposal.

C. Andrieu and G. O. Roberts, "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, pp. 697–725, 2009.

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.

A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn, "Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator," *To appear in Biometrika*, 2015.

G. Deligiannidis, A. Doucet, and M. K. Pitt, "The correlated pseudo-marginal method," *arXiv preprint arXiv:1511.04992*, 2015.

J. Dahlin, F. Lindsten, J. Kronander, and T. B. Schön, "Accelerating pseudo-marginal metropolis-hastings by correlating auxiliary variables," *arXiv preprint arXiv:1511.05483*, 2015.

M. Quiroz, M. Villani, and R. Kohn, "Speeding up mcmc by efficient data subsampling," *arXiv preprint arXiv:1404.4178*, 2014.

M.-N. Tran, R. Kohn, M. Quiroz, and M. Villani, "Block-wise pseudo-marginal metropolis-hastings," *arXiv preprint arXiv:1603.02485*, 2016.

D. Maclaurin and R. P. Adams, "Firefly Monte Carlo: Exact MCMC with subsets of data," *arXiv preprint arXiv:1403.5693*, 2014.

A. Korattikara, Y. Chen, and M. Welling, "Austerity in MCMC land: Cutting the Metropolis-Hastings budget," *arXiv preprint arXiv:1304.5299*, 2013.

R. Bardenet, A. Doucet, and C. Holmes, "Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 405–413, 2014.

R. Bardenet, A. Doucet, and C. Holmes, "On Markov chain Monte Carlo methods for tall data," *arXiv preprint arXiv:1505.02827*, 2015.