ML for Econometricians
Subsampling MCMC

Mattias Villani

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

Lecture Overview

- Data subsampling
- MCMC 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 $\hat{p}(\mathbf{y}|\theta, \mathbf{u})$.
- **Subsampling**: $\mathbf{u}$ are indicators for selected observations.
- Let m be the number of u's (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]**

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: target $\sigma_z^2 \in [1, 3.3]$, depending on how good the proposal for θ is. [2, 3]

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.

Subsampling MCMC

- Quiroz and co-authors [4, 5, 6, 7] develop much improved subsampling MCMC based on pseudo-marginal methods.
- Innovations:
  - **Control variates** to reduce the variance
  - **Depedent subsamples** over iterations. Allows much noisier estimators.
  - **Hamiltonian proposals** for high-dimensional problems
- More on this in my keynote talk ...

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** u_k with distribution

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

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

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

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

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

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_m 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**

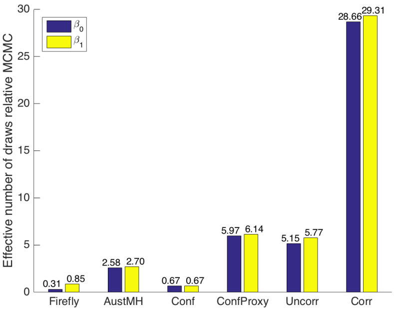
$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t, \quad \epsilon_t \sim t(\nu) \text{ iid}$$

- **Aim:** posterior of β_0, β_1 with known $\nu = 5$ based on a sample with 100,000 observations.

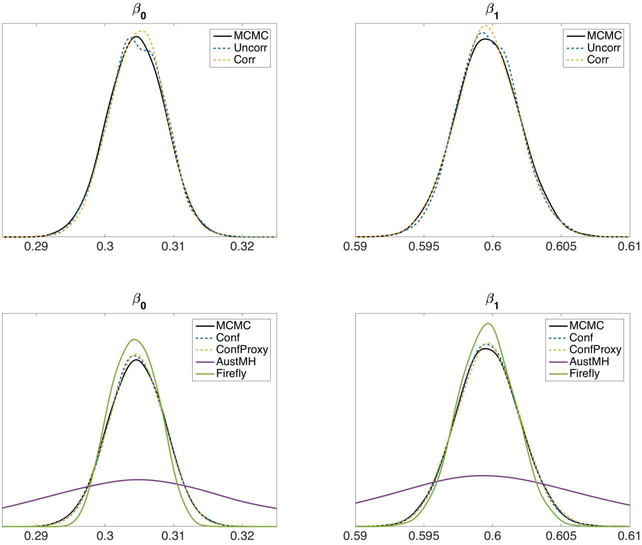
- Posterior is more or less a spike. Confidence sampler should preform well.

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

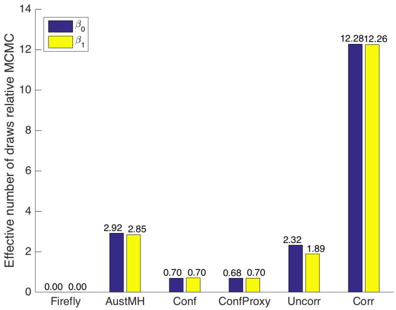
- **AR(1) process with student-t noise**

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

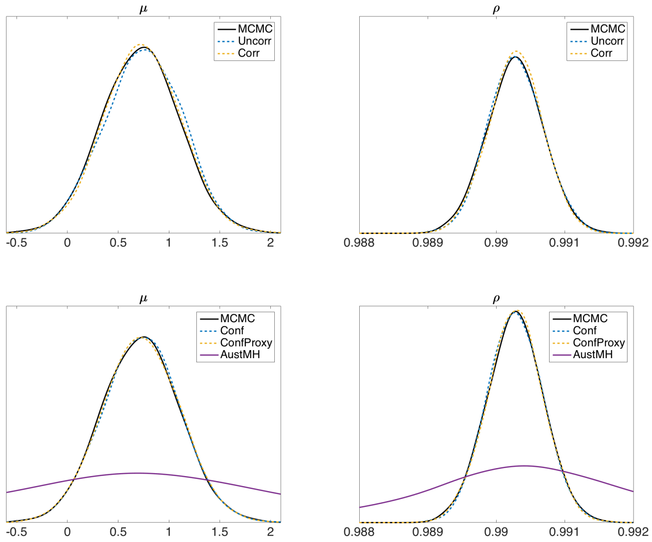
- **Aim:** posterior of μ, ρ with known $\nu = 5$ based on a sample with 100,000 observations.
- ρ is close to one in the data, so posterior of μ concentrates very slowly.

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



- 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.
- M. Quiroz, M. Villani, and R. Kohn, “Speeding up mcmc by efficient data subsampling,” *arXiv preprint arXiv:1404.4178*, 2014.
- M. Quiroz, M. Villani, and R. Kohn, “Exact subsampling mcmc,” *arXiv preprint arXiv:1603.08232*, 2016.

- M. Quiroz, M.-N. Tran, M. Villani, and R. Kohn, "Speeding up mcmc by delayed acceptance and data subsampling," *Journal of Computational and Graphical Statistics*, 2017.
- 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.
- 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.