Speeding Up MCMC by Efficient Data Subsampling

Matias Quiroz[1], Mattias Villani[2] and Robert Kohn[3]

[1]Sveriges Riksbank and Department of Statistics, Stockholm University

[2]Department of Computer and Information Science, Linköping University

[3]Australian Business School, University of New South Wales

August 28, 2014

Problem statement and idea

- **Problem:** Markov Chain Monte Carlo algorithms (MCMC) are very costly for complex models and/or Big Data. Can we do something about it?
- **Objective:** **Generic MCMC** algorithm being able to handle **large** data sets.
- **Achieved so far:** **Speeding up MCMC** for complex models. Good insight of the challenges with Big Data for "non-complex" models.
- **Big Data:** *Tall data*. Many observations, not necessary many variables.
  Example: Microeconomic data.

Problem statement and idea

- **Problem:** Markov Chain Monte Carlo algorithms (MCMC) are very costly for complex models and/or Big Data. Can we do something about it?
- **Objective:** **Generic MCMC** algorithm being able to handle **large** data sets.
- **Achieved so far:** **Speeding up MCMC** for complex models. Good insight of the challenges with Big Data for "non-complex" models.
- **Big Data:** *Tall data*. Many observations, not necessary many variables. **Example:** Microeconomic data.

- **The main idea:** Combine *MCMC* and *Survey sampling*.

MCMC

- **Notation:**
  - **Parameters** $\theta = (\theta_1, \ldots \theta_p)^T$
  - **Data** $y = (y_1, \ldots, y_n)^T$.
  - **Data distribution** $p(y_k|\theta)$
  - **Likelihood** $p(y|\theta) = \left(\prod_{k=1}^{n} p(y_k|\theta)\right)$
  - **posterior** $p(\theta|y) \propto \left(\prod_{k=1}^{n} p(y_k|\theta)\right) p(\theta)$

MCMC

- **Notation:**
  - **Parameters** $\theta = (\theta_1, \ldots \theta_p)^T$
  - **Data** $y = (y_1, \ldots, y_n)^T$.
  - **Data distribution** $p(y_k|\theta)$
  - **Likelihood** $p(y|\theta) = \left(\prod_{k=1}^{n} p(y_k|\theta)\right)$
  - **posterior** $p(\theta|y) \propto \left(\prod_{k=1}^{n} p(y_k|\theta)\right) p(\theta)$
- **MCMC:**
  - **In general:** MCMC gives N draws $\{x_j\}_{j=1}^{N}$ from *any* $p(x)$.
  - **For Bayesians:** $p(x) = p(\theta|y)$.
  - **Idea:** Construct a Markov Chain $\{\theta_j\}_{j=1}^{N}$ which admits $p(\theta|y)$ as **invariant distribution**.

MCMC, cont

- **Metropolis Hastings (M-H) algorithm:**
  set θ_c = guess
  let $\theta_1 = \theta_c$
  for j = 2:N
  $\theta_p \sim q(\theta_p|\theta_c)$ (proposal distribution)
  $\alpha = \min\left(1, \frac{p(\theta_p|y)/q(\theta_p|\theta_c)}{p(\theta_c|y)/q(\theta_c|\theta_p)}\right)$
  accept $\theta_j = \theta_p$ with probability α. If rejected set $\theta_j = \theta_c$
  set $\theta_c = \theta_j$
  endfor
  Output: $\{\theta_j\}_{j=1}^N$ draws from $p(\theta|y)$ (after discarding burn-in period)

- **Why is MCMC expensive?:** Need to evaluate $p(\theta_p|y) \propto \left(\prod_{k=1}^n p(y_k|\theta_p)\right) p(\theta_p)$. Massive product for large datasets. Complex $p(y_k|\theta_p)$.

Survey sampling and MCMC

- **Survey sampling:** Area of statistics which deals with **estimation** when **the population is finite.**
  Problem: *What is the total sales of all Swedish firms?*
  Key: *Which firms to include in the sample to answer this accurately?*
- Total sales = (finite) **population total**.

Survey sampling and MCMC

- **Survey sampling:** Area of statistics which deals with **estimation** when **the population is finite.**
  Problem: *What is the total sales of all Swedish firms?*
  Key: *Which firms to include in the sample to answer this accurately?*
- Total sales = (finite) **population total**.
- **Analogy:** In any given MCMC iteration **the full data log-likelihood** is a **population total**

$$l(\theta) = \log p(y|\theta) = \sum_{k=1}^{n} \log p(y_k|\theta).$$

- **In MCMC: Subsample data** and **estimate** $l(\theta)$ using **Survey sampling**. Plug in **the estimated likelihood** in the acceptance probability.
- **The estimated likelihood is noisy** - standard MCMC theory **does not apply**.

MCMC with analytically intractable $p(y|\theta)$

- Forget **data subsampling**. Consider situations when $p(y|\theta)$ is analytically intractable.
- MCMC with **estimation of the likelihood**: Use *particles* u to construct an estimator $\hat{p}(y|\theta, u)$ of $p(y|\theta)$. **Pseudo-marginal MCMC** (PMCMC).
- PMCMC samples from $p(\theta, u|y)$ by constructing a Markov chain

$$\{\theta_j, u_j\}_{j=1}^{N}$$

and accepting with

$$\alpha = \min\left(1, \frac{\hat{p}(y|\theta_p, u_p)p(\theta_p)/q(\theta_p|\theta_c)}{\hat{p}(y|\theta_c, u_c)p(\theta_c)/q(\theta_c|\theta_p)}\right).$$

- **Note:** We have replaced the true likelihood with an **estimate**.
- **Andrieu and Roberts (2009):** The marginal distribution of θ admits $p(\theta|y)$ as invariant distribution, regardless of the variance!
- **Requirement:** *unbiased* likelihood estimator

$$p(y|\theta) = \int \hat{p}(y|\theta, u)p(u)du.$$

MCMC with analytically intractable $p(y|\theta)$, cont

- **In practice:** Efficiency and computing time depends on the variance.
- **Low variance:** Gives **efficient** draws but **expensive** to compute the estimator (more particles required)
- **High variance:** **Less efficient draws** but **faster to compute** (less particles required)
- **Trade-off** between **computing time** and **efficiency**. Doucet et al (2012) finds that an estimator with *standard deviation around 1* is optimal.
  Main message: **Choose the number of particles** so that this is fulfilled.

MCMC with data subsampling

- ... Back to **data subsampling**.
- Constructing an **unbiased estimator of the likelihood** using **subsampling of data** fits the framework in **PMCMC**.
- The particles u become the **selection indicators** for which observations to include for estimating the likelihood.
- **Key point:** We can obtain the exact same result by only using **a small fraction of the data** instead of the full data. **Speeds up our computations.**
- This was also noted by Korattikara et al (2013) but quickly dismissed. Why?

MCMC with data subsampling

- ... Back to **data subsampling**.
- Constructing an **unbiased estimator of the likelihood** using **subsampling of data** fits the framework in **PMCMC**.
- The particles u become the **selection indicators** for which observations to include for estimating the likelihood.
- **Key point:** We can obtain the exact same result by only using **a small fraction of the data** instead of the full data. **Speeds up our computations.**
- This was also noted by Korattikara et al (2013) but quickly dismissed. Why?
- **The variance of the estimator becomes too large** for PMCMC to be useful (the chain gets stuck)...

MCMC with data subsampling

- ... Back to **data subsampling**.
- Constructing an **unbiased estimator of the likelihood** using **subsampling of data** fits the framework in **PMCMC**.
- The particles u become the **selection indicators** for which observations to include for estimating the likelihood.
- **Key point:** We can obtain the exact same result by only using **a small fraction of the data** instead of the full data. **Speeds up our computations.**
- This was also noted by Korattikara et al (2013) but quickly dismissed. Why?
- **The variance of the estimator becomes too large** for PMCMC to be useful (the chain gets stuck)...
- ... but these conclusion are based on a **Simple random sampling design**.

MCMC with data subsampling

- ... Back to **data subsampling**.
- Constructing an **unbiased estimator of the likelihood** using **subsampling of data** fits the framework in **PMCMC**.
- The particles u become the **selection indicators** for which observations to include for estimating the likelihood.
- **Key point:** We can obtain the exact same result by only using **a small fraction of the data** instead of the full data. **Speeds up our computations.**
- This was also noted by Korattikara et al (2013) but quickly dismissed. Why?
- **The variance of the estimator becomes too large** for PMCMC to be useful (the chain gets stuck)...
- ... but these conclusion are based on a **Simple random sampling design**.
- **Our main contribution:** Design **efficient sampling schemes** to make PMCMC useful.

Notations

- Let n be **the size of the population** and let m be the **sample size.**
- **Notations:** Let y be the response and x the covariates

$$L_k(\theta) = p(y_k|\theta, x_k)$$

$$L(\theta) = \prod_{k=1}^{n} L_k(\theta)$$

$$l_k(\theta) = \log p(y_k|\theta, x_k)$$

$$l(\theta) = \sum_{k=1}^{n} l_k(\theta)$$

- **Goal: Sample** m **observations** and construct $\hat{l}(\theta)$ such that $E[\hat{l}(\theta)] = l(\theta)$ and std$[\hat{l}(\theta)] \approx 1$ (Doucet et al, 2012).

Estimating a population total using Simple random sampling

- **Survey sampling literature** (Särndal et al, 2003)
- **Unbiased estimation** using **Simple random sampling** (SI) *without replacement*:

$$\hat{l}(\theta) = \frac{n}{m} \sum_{k \in S(u)} l_k(\theta) = \frac{n}{m} \sum_{k=1}^{n} l_k(\theta) u_k$$

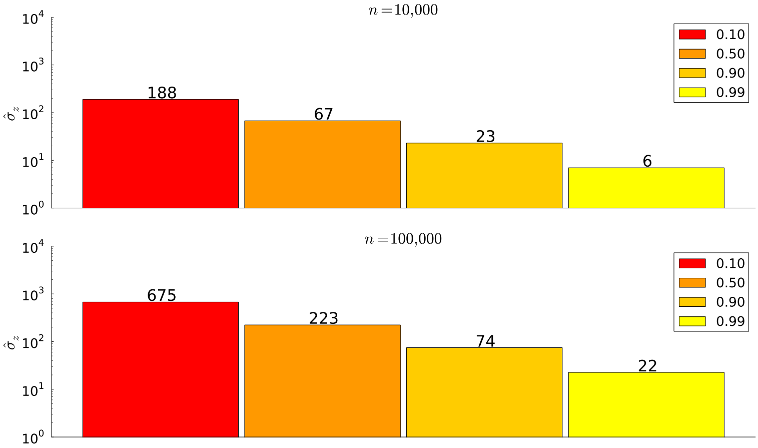
  $S(u)$ - the index-set of sampled observations. $|S(u)| = m$.
  $u = (u_1, \ldots, u_n)^T$ binary selection indicators.
  All observations **equally probable to be selected**: $\pi_k = P(u_k = 1) = m/n$.

- **Unbiased variance estimator**

$$\hat{V}[\hat{l}(\theta)] = n^2 \frac{(1-f)}{m} s_S^2$$

  where $f = \frac{m}{n}$ is the *sampling fraction* and $s_S^2 = \frac{1}{m-1} \sum_{k \in S} (l_k(\theta) - \bar{l}_S(\theta))^2$

Simple random sampling does not work



Estimating a population total using Probability proportional-to-size

- **SI does not work** because it treats all $\log p(y_k|\theta, x_k)$ symmetrically ($\pi_k = P(u_k = 1) = m/n$). **Proportional-to-size sampling** a better idea.

Estimating a population total using Probability proportional-to-size

- **SI does not work** because it treats all $\log p(y_k|\theta, x_k)$ symmetrically ($\pi_k = P(u_k = 1) = m/n$). **Proportional-to-size sampling** a better idea.
- **Unbiased estimation using general π_k: Horvitz-Thompson estimator** for the **population total**:

$$\hat{l}(\theta) = \sum_{k \in S(u)} \frac{l_k(\theta)}{\pi_k}$$

- **Unbiased variance estimator**

$$\hat{V}[\hat{l}(\theta)] = \sum_{k \in S} \sum_{l \in S} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{l_k(\theta)}{\pi_k} \frac{l_l(\theta)}{\pi_l}$$

$\pi_{kl} = P(u_k = 1, u_l = 1)$

- **How to choose π_k?**

Estimating a population total using Probability proportional-to-size, cont

- Assume we choose $\pi_k \propto l_k(\theta)$, i.e. $\frac{l_k(\theta)}{\pi_k} = c$
- Then
$$\hat{l}(\theta) = \sum_{k \in S(u)} \frac{l_k(\theta)}{\pi_k} = mc$$
is constant so $V[\hat{l}(\theta)] = 0$.
- **Ideal estimator.** Requires $l_k(\theta)$ for $k = 1, \ldots, n$. $l(\theta)$ is exactly known in this case. No point in subsampling.

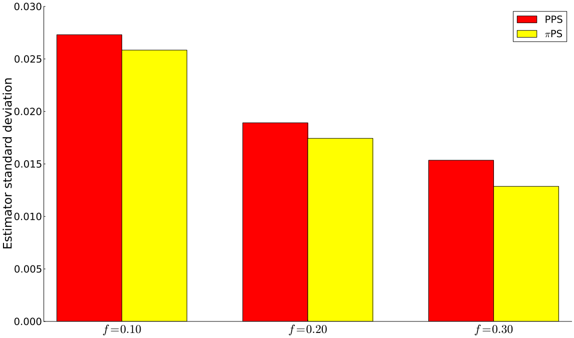
Estimating a population total using Probability proportional-to-size, cont

- Assume we choose $\pi_k \propto l_k(\theta)$, i.e. $\frac{l_k(\theta)}{\pi_k} = c$
- Then
$$\hat{l}(\theta) = \sum_{k \in S(u)} \frac{l_k(\theta)}{\pi_k} = mc$$
is constant so $V[\hat{l}(\theta)] = 0$.
- **Ideal estimator.** Requires $l_k(\theta)$ for $k = 1, \ldots, n$. $l(\theta)$ is exactly known in this case. No point in subsampling.
- **Assume** we can construct $w_k > 0$ such that $\frac{l_k(\theta)}{w_k} \approx c$ for all k.
- **Set**
$$\pi_k = \frac{w_k}{\sum_{k=1}^{n} w_k}$$
then $\frac{l_k(\theta)}{\pi_k}$ is approximately constant and $V[\hat{l}(\theta)]$ small.
- w_k needs to be a good **proxy** of $l_k(\theta)$. More on this later.

Estimating a population total using Probability proportional-to-size, cont

- This Probability proportional-to-size **without replacement** is known as πPS sampling. Without replacement makes πPS **computationally intractable for large** n.
- **PPS-sampling** is the equivalent when sampling is done **with replacement**.
- **PPS** has slightly higher variance **but is much faster. PPS** is our **final choice**.

Standard deviation of PPS and πPS



f = Sampling fraction

Important:
Note the gain in efficiency compared to Simple random sampling (SI).
For SI $\hat{\sigma} = 188$ for $f = 0.10$.

Bias-correction

- **Unbiasedness** for our Survey sampling estimators is on the **logaritmic scale.**
- **PMCMC** requires **unbiasedness** in the **ordinary scale**.
- Need to **bias-correct** $\hat{L}(\theta) = \exp\left(\hat{l}(\theta)\right)$.
- **Bias-correction** can be avoided using **Generalized Poisson Estimator** (Estimates $L(\theta)$ directly). Needs an extra Monte Carlo step + $\hat{L}(\theta) > 0$.
- **In the paper** a bias-correction based on **asymptotics** of $\hat{l}(\theta)$ is proposed. **Fast** and **effective** in practice.

Constructing efficient sampling weights

- **Recall:** Requirement $\frac{l_k(\theta)}{w_k} \approx c$
- Many models have **surrogate/approximate** models for inference - use this as w_k. **Exact inference** with a **minimum of density evaluations.**

Constructing efficient sampling weights

- **Recall:** Requirement $\frac{l_k(\theta)}{w_k} \approx c$
- Many models have **surrogate/approximate** models for inference - use this as w_k. **Exact inference** with a **minimum of density evaluations.**
- **Wanted: An approximation** of the **log-likelihood contribution** $l(\theta; d)$ for any data point $d = (y, x)$ and parameter vector θ. **Surface estimation.**

Constructing efficient sampling weights

- **Recall:** Requirement $\frac{l_k(\theta)}{w_k} \approx c$
- Many models have **surrogate/approximate** models for inference - use this as w_k. **Exact inference** with a **minimum of density evaluations.**
- **Wanted: An approximation** of the **log-likelihood contribution** $l(\theta; d)$ for any data point $d = (y, x)$ and parameter vector θ. **Surface estimation.**
- **"Predicting machine": Noise free Gaussian Process** (GP) or **Regularized thin-plate splines** (TPS).
- **Usage: Train** using a **small fixed set** of training points V. **In each iteration:** Compute $l_V(\theta)$. Predict $l_k(\theta)$ for the rest.
- **Fast.** Only matrix-vector multiplications.

Evaluating the PMCMC algorithm

- We evaluate the algorithm on a data set containing **half a million observations**.
- **Model: Bivariate probit** with **endogenous** treatment effect

$$\begin{aligned} y_1^* &= \beta_{10} + \beta_{11} \cdot x_1 + \beta_{12} \cdot x_2 + \alpha \cdot y_2 + \varepsilon_1 \\ y_2^* &= \beta_{20} + \beta_{21} \cdot x_1 + \beta_{22} \cdot x_3 + \beta_{23} \cdot x_4 + \varepsilon_2 \\ y_1 &= I(y_1^* > 0) \\ y_2 &= I(y_2^* > 0) \end{aligned}$$

  where ε_1 and ε_2 are standard Gaussian with correlation ρ.
- Variables:
  - y_1 = Bankrupt, y_2 = Excess cash
  - x_1 = Earnings, x_2 = Leverage, x_3 = Fixed assets, x_4 = Firm size.
- **Time-consuming likelihood** (bivariate normal integral).
- **PMCMC implemented** with TPS. 5% of the data to train TPS. 8% data on average to estimate likelihood.

Evaluating the PMCMC algorithm, cont

- **Measure efficiency** through Inefficiency Factor (IF)

$$IF = 1 + 2\sum_{l=1}^{\infty} \rho_l$$

where ρ_l is the correlation at the lth lag of the (P)MCMC chain

- Compare the **Efficient Draws Per Minute (EDPM)**

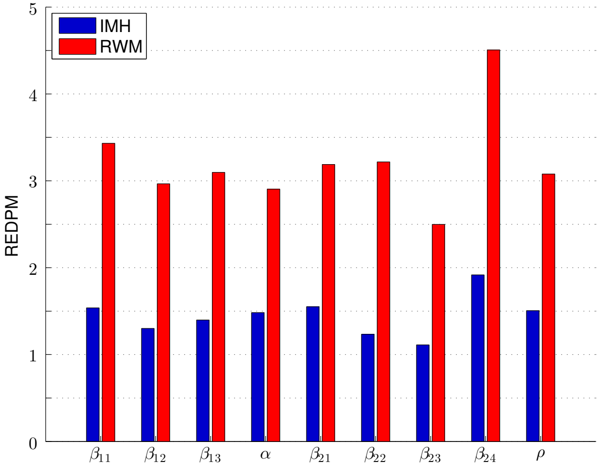
$$EDPM = \frac{N}{IF \times t}$$

- **Relative EDPM** (REDPM)

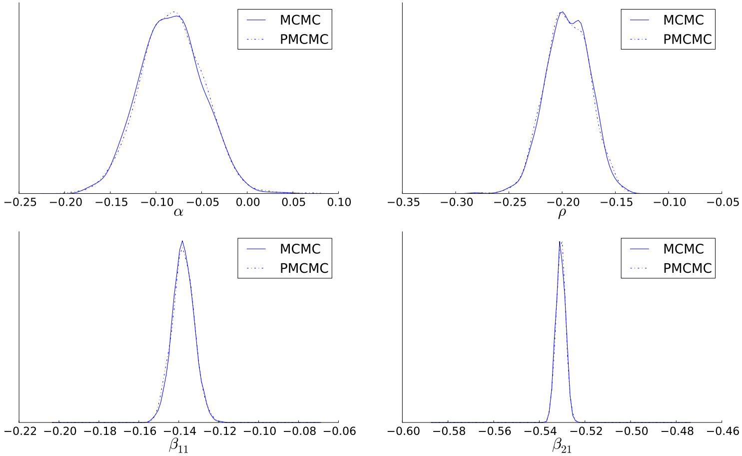
$$REDPM = \frac{EDPM^{PMCMC}}{EDPM^{MCMC}}$$

- Evaluate using two proposals: **Independent Metropolis Hastings** (IMH, efficient). **Random Walk Metropolis** (RWM, inefficient)

Comparing Relative Efficient Draws Per Minute for different proposals



Some marginal posteriors: PMCMC vs MCMC



Conclusions

- We have proposed a **general framework** for **Pseudo-marginal MCMC** based on **efficient data subsampling**.
- **Gaussian Process** or **Regularized thin-plate splines** to construct **efficient PPS-weights**.
- **More efficient draws per minute** in firm data application. **Biggest gain for weaker proposals** - consistent with theoretical results in Doucet et al. (2012).

The End

Thank you for listening!

References

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

Doucet, A., Pitt, M. and Kohn, R (2012). Efficient implementation of Markov Chain Monte Carlo when using an unbiased likelihood estimator. *arXiv preprint arXiv:1210.1871*.

Särndal C.-E., Swensson B., and Wretman, J. (2003). *Model assisted survey sampling*.