Informed Subsampling MCMC: Approximated Bayesian Inference for Large Datasets

Florian Maire, University College Dublin

joint work with : Nial Friel (UCD) & Pierre Alquier (ENSAE ParisTech)

Outline of talk

- Introduction/overview of literature on the Bayesian inference for tall data
- Generally, two types of approaches:
  - Divide-and-conquer: partition the data into subsets, process each batch separately and then combine the inferences.
  - Sub-sampling strategies: reduce the computational burden of Metropolis-Hastings.
- Our approach falls under the category of sub-sampling strategies.
- The main idea is to **fix** the subset size $n \ll N$ and to **focus on those sub-samples that are similar to the full data**, in terms of how close the summary statistics of the sub-sample is to summary statistics of the full data.
- It therefore shares some similarities with Approximate Bayesian Computation.

Outlines

Metropolis-Hastings sampler in Big Data problems

- Consider the posterior distribution

$$\pi(\theta \mid Y_1, \ldots, Y_N) \propto f(Y_1, \ldots, Y_N \mid \theta) p(\theta)$$

  where $f(\cdot \mid Y_1, \ldots, Y_N)$ is the likelihood model and p the prior distribution.

- Metropolis-Hastings simulates a Markov chain $\{\theta_k\}_k$ targeting $\pi(\cdot \mid Y_1, \ldots, Y_N)$, transition $\theta_k \to \theta_{k+1}$ follows:

  (1) draw $\tilde{\theta} \sim Q(\theta_k, \cdot)$
  (2) set $\theta_{k+1} = \tilde{\theta}$ with probability

$$A(\theta_k, \tilde{\theta}) = 1 \wedge \frac{f(\tilde{\theta} \mid Y_1, \ldots, Y_N) p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{f(\theta_k \mid Y_1, \ldots, Y_N) p(\theta_k) Q(\theta_k, \tilde{\theta})}$$

  and $\theta_{k+1} = \theta_k$ w.p. $1 - A(\theta_k, \tilde{\theta})$.

Another way of looking at the MH algorithm

Transition $\theta_k \to \theta_{k+1}$ follows:

(1) draw $\tilde{\theta} \sim Q(\theta_k, \cdot)$ and $W_k \sim \text{unif}(0, 1)$

(2-a) Let $E_k^{(N)}$ be the event

$$E_k^{(N)}(\theta_k, \tilde{\theta}, W_k) = \left\{ W_k \le 1 \wedge \frac{f(\tilde{\theta} \mid Y_1, \ldots, Y_N) p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{f(\theta_k \mid Y_1, \ldots, Y_N) p(\theta_k) Q(\theta_k, \tilde{\theta})} \right\}$$

(2-b) Set

$$\theta_{k+1} = \begin{cases} \tilde{\theta} & \text{if } E_k^{(N)}(\theta_k, \tilde{\theta}, W_k) \\ \theta_k & \text{otherwise} \end{cases}$$

$\Rightarrow$ A MH transition is thus a statistical hypothesis test: does $E_k^{(N)}$ occur or not?

Making the decision with sub-samples of data?

- Is it possible to make the <u>same decision as MH</u> (with a large probability), without computing $f(\tilde{\theta} \mid Y_1, \ldots, Y_N)$?
- Make the decision to accept/reject $\tilde{\theta}$ based on a subset of $n \ll N$ data:

$$E_k^{(n)}(\theta_k, \tilde{\theta}, W_k) = \left\{ W_k \leq 1 \wedge \frac{f(\tilde{\theta} \mid Y_1^*, \ldots, Y_n^*) p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{f(\theta_k \mid Y_1^*, \ldots, Y_n^*) p(\theta_k) Q(\theta_k, \tilde{\theta})} \right\}$$

  rather than on $E_k^{(N)}$.

Making the decision with sub-samples of data?

- Austerity in MCMC land: Cutting the M–H budget, Korattikara et al, 2013
- Towards scaling up MCMC: an adaptive subsampling approach, Bardenet et al, 2014
- On MCMC methods for tall data, Bardenet et al, 2015
- Random Projections in MCMC for tall data, Bardenet et al, 2016

However, these methods are no longer exact in that the chain $\{\theta_k\}_k$ does not admit $\pi(\cdot \mid Y_1, \ldots, Y_N)$ as stationary distribution.

First approach: Austerity in MCMC land, Korattikara 2013

Rewriting $E_k^{(N)}$ in case of *i.i.d.* data

$$E_k^{(N)}(\theta_k, \tilde{\theta}, W_k) = \left\{ \underbrace{\frac{1}{N} \log \left(W_k \frac{p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{p(\theta_k) Q(\theta_k, \tilde{\theta})} \right)}_{\mu_0} \leq \underbrace{\frac{1}{N} \sum_{\ell=1}^{N} \log \frac{f(\tilde{\theta} \mid Y_\ell)}{f(\theta_k \mid Y_\ell)}}_{\mu} \right\}$$

- draw without replacement n data $Y_1^*, \ldots, Y_n^*$
- calculate $\tilde{\mu}^{(n)} = n^{-1} \sum_{\ell=1}^{n} \log \frac{f(\tilde{\theta} \mid Y_\ell^*)}{f(\theta_k \mid Y_\ell^*)}$
- test $H_0^{(n)} = \{\mu_0 = \tilde{\mu}^{(n)}\}$ vs $H_1^{(n)} = \{\mu_0 \neq \tilde{\mu}^{(n)}\}$
- subsample data until $\mathbb{P}(H_1^{(n)}) > 1 - \epsilon$
- make decision using $\tilde{\mu}^{(n)}$ instead of μ:

$$E_k^{(n)}(\theta_k, \tilde{\theta}, W_k) = \left\{ \mu_0 \leq \tilde{\mu}^{(n)} \right\}$$

Second approach: Confidence sampler (Bardenet et al, 2014, 2015)

Assume that a concentration inequality exists for the model, *i.e*

$$\forall\, n \leq N\,, \exists\, c_n > 0,\ \delta_n \in (0,1)\,, \qquad \mathbb{P}\left(\left|\mu - \tilde{\mu}^{(n)}\right| \leq c_n\right) \geq 1 - \delta_n\,.$$

For example, choose $\delta_n \in (0,1)$ and c_n is defined as

$$c_n(\delta_n) = \sigma_{n,\theta,\tilde{\theta}} \sqrt{\frac{2\log(3/\delta_n)}{n}} + \frac{6\, C_{\theta,\tilde{\theta}} \log(3/\delta_n)}{n}$$

where

- $\sigma_{n,\theta,\tilde{\theta}}$ is the sample standard deviation of $\{\log f(\tilde{\theta} \mid Y_i^*)/f(\theta \mid Y_i^*)\}_{i=1}^n$
- $C_{\theta,\tilde{\theta}} = \max_{i=1:n}(\log f(\tilde{\theta} \mid Y_i^*)/f(\theta \mid Y_i^*))$

$\Rightarrow$ draw data $Y_1^*, \ldots, Y_n^*$ such that as soon as:

$$\left|\tilde{\mu}^{(n)} - \mu_0\right| > c_n\,,$$

then decisions based on $E_k^{(n)}$ and $E_k^{(N)}$ are the same with probability $1 - \delta_n$.

Subsampling approaches

- Subsampling approaches share the same philosophy:
  Draw more data until a decision replicating MH can be made with a level of confidence.
- Bardenet et al.'s works offer more theoretical guarantees (*e.g* ergodicity, quantification of the error,...)
  But comes at the price of more intermediate calculations $\sigma_{n,\theta,\tilde{\theta}}$ and $C_{\theta,\tilde{\theta}}$.
- Critically, the adaptive subset size n tends to N as the chain is close to equilibrium.

Consensus Monte Carlo (Scott et al, 2013)

This approach exploits parallel computing in a very natural way.

- Split the dataset into S (independent) batches $Y_{1:N} = \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_S$ and note that

$$\pi(\theta \,|\, Y_1, \ldots, Y_N) \propto \Pi_{i=1}^{S} f(\boldsymbol{Y}_i) p(\theta)^{1/S}$$

- Generate S independent Markov chains (in parallel) targeting $\{\pi(\theta \,|\, \boldsymbol{Y}_i) \propto f(\boldsymbol{Y}_i) p(\theta)^{1/S}\}_{i=1}^{S}$
- Derive a weighted average of the S chains

$$\theta_k = \left\{ \sum_{i=1}^{S} W_i \right\}^{-1} \sum_{i=1}^{S} W_i \theta_k^{(i)}$$

- This is justifiable when π is Gaussian, but questions about the convergence of $\{\theta_k\}_k$ and the choice of $\{W_i\}_{i=1}^{S}$ remains open

Exact methods

MCMC methods producing a chain that admits $\pi(\theta \mid Y_1, \ldots, Y_N)$ as invariant distribution:

- Using unbiased estimate of $f(\theta \mid Y_{1:N})$
  Pseudo-Marginal literature: Andrieu & Vihola 2012, Doucet et al 2012, Quiroz et al, 2016
- A sub-optimal M–H transition kernel
  Accelerating M–H algorithms: Delayed acceptance with prefetching, Banterle et al, 2014
- An auxiliary variable MCMC, under strong assumptions
  FireFly Monte Carlo: Exact MCMC with subsets of data, MacLaurin et al, 2014

An alternative approach

Definition

Let K be the M–H kernel targeting $\pi(\cdot \mid Y_1, \ldots, Y_N)$
Let $U \subset \{1, \ldots, N\}$ and K_U be the M–H kernel targeting $\pi(\cdot \mid Y_U)$

Assumption

$\tau(K) = \mathcal{O}(N)$ *and for* $U = \{1, \ldots, n\}$, $\tau(K_U) = \mathcal{O}(n)$

For a given CPU budget τ_0:

- the number of M-H iterations is fixed (potentially low if $N \gg 1$)
- can we derive an algorithm that achieves an arbitrary large number of iterations for a small subset size n?

Inference based on *subposteriors*

Definition

Let Y_U be a subset of $Y_{1:N}$ of size n and $\bar{\pi}_n$ be a scaled subposterior

$$\bar{\pi}_n(\theta \mid Y_U) \propto f(Y_U \mid \theta)^{N/n} p(\theta)$$

Outlines

Exponential family: an optimality result

Assume that f belongs to the curved exponential family

$$f(y \mid \theta) \propto \psi(\theta) \exp\{\phi(\theta)^{\mathsf{T}} S(y)\}\,.$$

Definition

For any subset $U \in \mathcal{U}_n$, define the vector of sufficient statistics between the whole dataset and the sub-sample Y_U as:

$$\Delta_n(U) = \sum_{k=1}^{N} S(y_k) - \frac{N}{n} \sum_{k \in U} S(y_k)$$

Exponential family: an optimality result

The KL divergence between two measure π and $\bar{\pi}$ is defined as $\mathrm{KL}(\pi, \bar{\pi}) = \mathbb{E}_\pi\{\log \pi(\theta)/\bar{\pi}(\theta)\}$.

Proposition

For any $U \in \mathcal{U}_n$, the following inequality holds:

$$\mathrm{KL}(\pi, \bar{\pi}_n(\,\cdot \mid Y_U)) \leq B(Y, U)\,,$$

where

$$B(Y, U) = \log \mathbb{E}_\pi \exp\left\{\|\mathbb{E}_\pi(\phi(\theta)) - \phi(\theta)\|\, \|\Delta_n(U)\|\right\}\ .$$

Corollary

1. *For any subset $U \in \mathcal{U}_n$ such that $(1/N)\sum_{k=1}^N S(Y_k) = (1/n)\sum_{k \in U} S(Y_k)$, then $\pi = \bar{\pi}(\,\cdot \mid Y_U)$ π-almost everywhere.*
2. *Let $(U_1, U_2) \in \mathcal{U}_n^2$. Assume $\|\Delta_n(U_1)\| \leq \|\Delta_n(U_2)\|$, then $B(Y, U_1) \leq B(Y, U_2)$.*

Optimal result in asymptotic regime (when $N \to \infty$)

If a Bernstein-von Mises theorem holds for π, *i.e* π can be approximated by a Normal distribution $\tilde{\pi} = N(\theta^{\mathrm{MLE}}, (1/N)\{I(\theta^{\mathrm{MLE}})\}^{-1})$.

Definition

Define $\widehat{\mathrm{KL}}_n(U)$ as the Kullback-Leibler divergence between the asymptotic approximation of π and $\bar{\pi}_n(\,\cdot\,|\,Y_U)$:

$$\widehat{\mathrm{KL}}_n(U) = \mathbb{E}_{\tilde{\pi}} \log \frac{\pi(\theta \mid Y)}{\bar{\pi}_n(\theta \mid Y_U)} \,.$$

Proposition

Let $(U_1, U_2) \in \mathcal{U}_n^2$.
Assume that for all $i \in \{1, \dots, d\}$, $|\Delta_n(U_1)^{(i)}| \leq |\Delta_n(U_2)^{(i)}|$, where $|\Delta_n(U_1)^{(i)}|$ refers to the i-th element of $\Delta_n(U_1)$. Then $\widehat{\mathrm{KL}}_n(U_1) \leq \widehat{\mathrm{KL}}_n(U_2)$.

$\Rightarrow$ This is a stronger result on partial ordering on subsets not on the KL bound but on the KL itself.

Example (Toy Example: probit model)

Simulate $N = 10,000$ *observations* $Y_1, \ldots, Y_N$

- $X_k \sim \mathcal{N}(\theta, 1)$
- $Y_k \mid X_k = \delta_{(X_k > 0)}(\,\cdot\,)$

$$\pi(\theta \mid Y_{1:N}) \propto p(\theta)(1-\alpha(\theta))^N \left(\alpha(\theta)/1-\alpha(\theta)\right)^{\sum_{k=1}^N Y_k}$$

$$\bar{\pi}_n(\theta \mid Y_U) \propto p(\theta)(1-\alpha(\theta))^N \left(\alpha(\theta)/1-\alpha(\theta)\right)^{\frac{N}{n}\sum_{k\in U} Y_k}$$

where $\alpha(\theta) = \mathbb{P}\{X_k > 0 \mid X_k \sim \mathcal{N}(\theta, 1)\}$.
Define

$$|\Delta_n(U)| = \left|\sum_{k=1}^N Y_k - \frac{N}{n}\sum_{k\in U} Y_k\right|$$

(note that $\sum_{k=1}^N Y_k$ *is a sufficient statistics.)*

Probit Example: $n = 100$

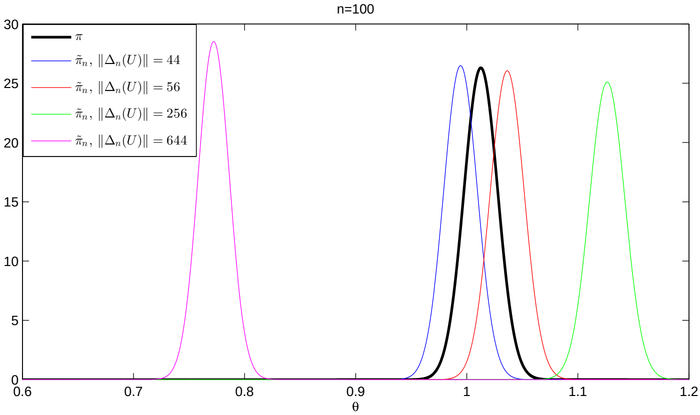


Figure: Sub-posteriors with different subsets U of size $n = 100$.

Probit Example: $n = 1,000$

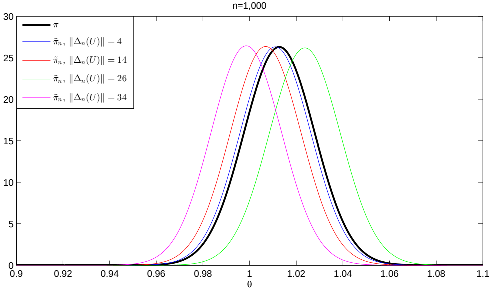


Figure: Sub-posteriors with different subsets U of size $n = 1,000$.

$\|\Delta_n(U)\|$	$\mathrm{KL}(\pi, \tilde{\pi}_n(\cdot \mid Y_U))$	$B(Y, U)$
4	0.004	0.04
14	0.11	0.18
26	0.19	0.29
34	0.41	0.54

Table: Comparison of the KL divergence between π and some sub-posterior distributions with different $\|\Delta_n(U)\|$.

Outlines

From summary statistics to sufficient statistics

$y \mapsto f(y \mid \theta)$ is now any likelihood model and the data are no longer assumed to be independent.

Definition

Let

- $S : \mathsf{Y} \to \mathsf{S}$ *be a mapping of* *summary* *statistics*
- $\Delta_n(U) = S(Y_1, \ldots, Y_N) - \frac{N}{n} S(Y_U)$
- $\mathcal{U}_n$ *be the set of all possible subset of* $\{1, \ldots, N\}$ *of size* n

For each $U \in \mathcal{U}_n$, a *weight* $\nu_{n,\epsilon}(U)$ is assigned to the subset of data Y_U

$$\nu_{n,\epsilon}(U) \propto \exp\left\{-\epsilon \|\Delta_n(U)\|^2\right\} .$$

- $\epsilon \to \infty$: all the subsets have the same weight
- $\epsilon \to 0$: the mass is centered on the most representative subset

Informed Subsampling MCMC

Recall that Metropolis-Hastings produces a chain $\{\theta_k\}_k$

$$(i)\ \tilde{\theta} \sim Q(\theta,\,\cdot\,) \qquad (ii)\ A(\theta,\tilde{\theta}) = 1 \wedge \frac{f(\tilde{\theta}\,|\,Y_1,\ldots,Y_N)p(\tilde{\theta})Q(\tilde{\theta},\theta)}{f(\theta\,|\,Y_1,\ldots,Y_N)p(\theta)Q(\theta,\tilde{\theta})}$$

Our idea is to define a chain $\{\theta_k\}_k$ that evolves as follows:

$$(i)\ \tilde{\theta} \sim Q(\theta,\,\cdot\,) \qquad (ii)\ A(\theta,\tilde{\theta}) = 1 \wedge \frac{f(\tilde{\theta}\,|\,Y_U)p(\tilde{\theta})Q(\tilde{\theta},\theta)}{f(\theta\,|\,Y_U)p(\theta)Q(\theta,\tilde{\theta})}$$

This chain targets $\bar{\pi}_n(\theta\,|\,Y_U)$ which is of little interest, since likely to be far from π (see probit Example).

Informed Subsampling MCMC

Based on the analysis of exponential models, we consider the following algorithm. It produces a chain $\{\theta_k, U_k\}_k$ as follows.

1. Update the subset:
   1.1 propose $U' \sim R(U_k, \cdot)$
   1.2 set U_{k+1} with probability $1 \wedge \exp\{\epsilon\left(\|\Delta_n(U_k)\| - \|\Delta_n(U')\|\right)\}$
2. Update the parameter:
   2.1 propose $\theta' \sim Q(\theta_k, \cdot)$
   2.2 set θ_{k+1} with probability $\tilde{A}(\theta, \theta' | U_{k+1}) = 1 \wedge \tilde{\alpha}(\theta, \theta' | U_{k+1})$ where

$$\tilde{\alpha}(\theta, \theta' | U) = \frac{f(\theta' \mid Y_{U_{k+1}}) p(\theta') Q(\theta', \theta)}{f(\theta \mid Y_{U_{k+1}}) p(\theta) Q(\theta, \theta')} .$$

Note that the first step is independent of θ_k.
In fact, it is straightforward to show that $\{U_k\}_k$ is ν_n-reversible Markov chain.

Convergence of perturbed Markov chains

Consider a Metropolis-Hastings algorithm whose ratio $\alpha(\theta, \theta')$ is perturbed through some noisy auxiliary variables U: $\tilde{\alpha}(\theta, \theta' \mid U)$.

Proposition (Alquier 2016, Corollary 2.3)

If we can bound the expected error between α and $\tilde{\alpha}$ s.t.

$$\mathbb{E}\left\{|\alpha(\theta, \theta') - \tilde{\alpha}(\theta, \theta' \mid U)|\right\} \leq \delta(\theta, \theta'),$$

then:

$$\lim_{n \to \infty} \|\pi - \mu \tilde{P}^n\| \leq \kappa \sup_{\theta \in \Theta} \int_{\Theta} Q(\theta, \mathrm{d}\theta') \delta(\theta, \theta'), \tag{1}$$

where

- *$\tilde{P}$ is the transition kernel of the noisy algorithm*
- *κ is a constant depending on the efficiency of the non-noisy Metropolis Hastings chain.*

Convergence of $\{\theta_k\}_k$

We cast the analysis of Informed Subsampling chain $\{\theta_k\}_k$ in the noisy MCMC framework.

Proposition

Under regularity assumption on the function $\theta \mapsto f(Y_U \mid \theta)^{N/n} / f(Y \mid \theta)$, there is a constant λ such that

$$\mathbb{E}_\nu \left\{ \left| \alpha(\theta, \theta') - \tilde{\alpha}(\theta, \theta' \mid U) \right| \right\} \leq \alpha(\theta, \theta') \lambda \|\theta - \theta'\| \Phi(\theta) ,$$

where

$$\Phi(\theta) = \mathbb{E}_\nu \left\{ \frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}} \right\} \propto \sum_{U \in \mathcal{U}_n} \nu(U) \frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}} .$$

The RHS's expectation can be unstable if an unappropriate weight distribution is used.

Convergence of $\{\theta_k\}_k$

Proposition

For exponential models, we prove the following:

$$\frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}} = o(1/\nu(U))\,.$$

and thus $\Phi(\theta)$ *is bounded.*

For general models, this proposition serves as a way to validate summary statistics:

Rule

Let f be a general likelihood model and S a possible summary statistics vector. If there is a β such that

$$|\log f(Y \mid \theta) - (N/n) \log f(Y_U \mid \theta)| \leq \beta \|\Delta_n(U)\|\,,$$

then S is sensible choice of summary statistics.

Outlines

Example 1: estimation of template shapes

Data are of handwritten digits (MNIST database)

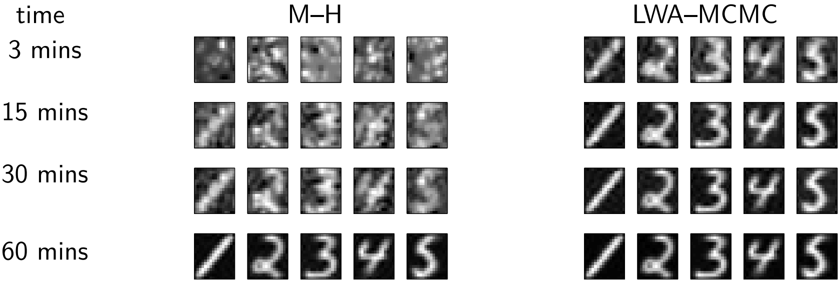
Figure: example of data

- The dataset contains $N = 10,000$ images of size 16x16
- Each image belongs to a class $I_k \in \{1, \ldots, 5\}$ assumed to be known
- The model can be written as:

$$I_k = i, \quad Y_k = \phi(\theta_i) + \sigma^2 \varepsilon_k \,, \qquad \varepsilon_k \sim \mathcal{N}(0,1)\,.$$

Example 1: estimation of template shapes

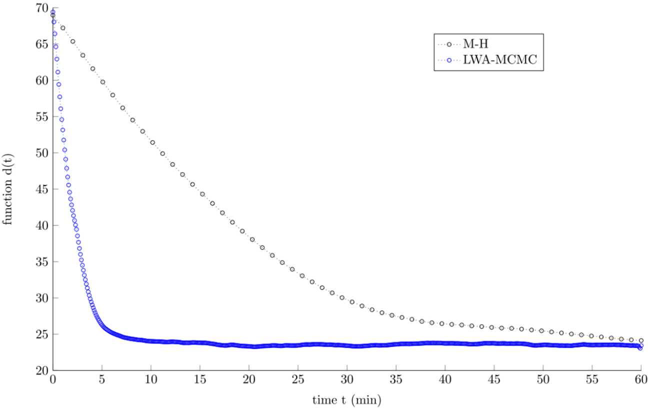
- Computational budget: $\tau_0 = 60$ mins.,
- We compare M–H and LWA-MCMC with subset of $n = 100$ digits, $\epsilon = 1$ and $S(U) = (S_1(U), \dots, S_5(U))$ with $S_i(U) = \sum_{k \in U} I_k$
- $\tau_{\text{MH}} = 41.2$ secs and $\tau_{\text{Informed Subsampling-MCMC}} = 0.7$ secs ($60 \times$ faster)



Example 1: estimation of template shapes

Consider the metric $d(t) = \sum_{i=1}^{5} \left\| \theta_i^* - \frac{1}{L(t)} \sum_{\ell=1}^{L(t)} \theta_{i,\ell} \right\|$, where:

- $L(t)$ is the number of iterations completed at time t
- θ_i^* is the map of model i (estimated from stochastic EM)



Example 1: Sampling at stationarity

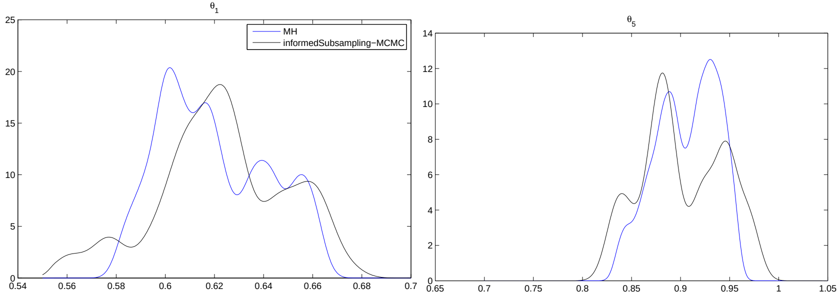


Figure: Comparing the true and approximate marginal distribution of one parameter of θ_1 (left) and one parameter of θ_5 (right)

Example 2: Auto regressive model Example (AR(2))

An AR(2) model, parameterized by $\theta = (\theta_1\, \theta_2, \theta_3)$

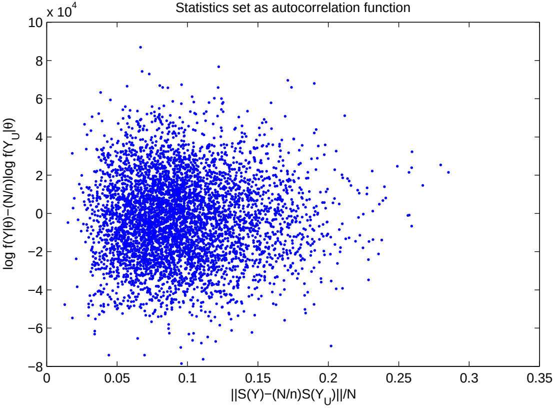
$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \zeta_t\,, \quad \zeta_t \sim \mathcal{N}(0, \theta_3^2)\,.$$

- Simulation of a TS of $N = 10^6$ observations,
- Approximate inference with $n \in \{10^2, 10^3\}$
- Different Summary statistics are tried:
  - $S(y') = \{\rho_1(y'), \ldots, \rho_5(y')\}$ where $\rho_i(y')$ is the i-th lag sample autocorrelation
  - $S(y') = \theta^{\mathrm{YW}}(y')$, the estimation of θ via Yule Walker method based on data y'
- Different ϵ were used.
- Our approach is tested versus MH implementation (prior, proposal, etc.) proposed in Chib, Understanding Metropolis

Samples from $\pi(\,\cdot \mid y)$ were obtained via MH on the whole dataset (a laborious work!).

Example 2: validation of summary statistics

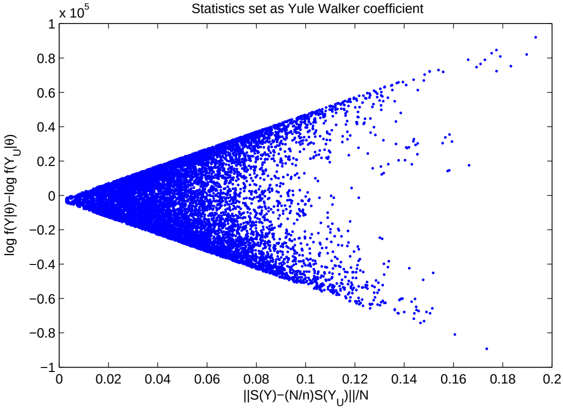
Try S defined as the estimated ACF (first 5 coefficients)



$\Rightarrow$ S rejected, $\phi(\theta)$ is unstable.

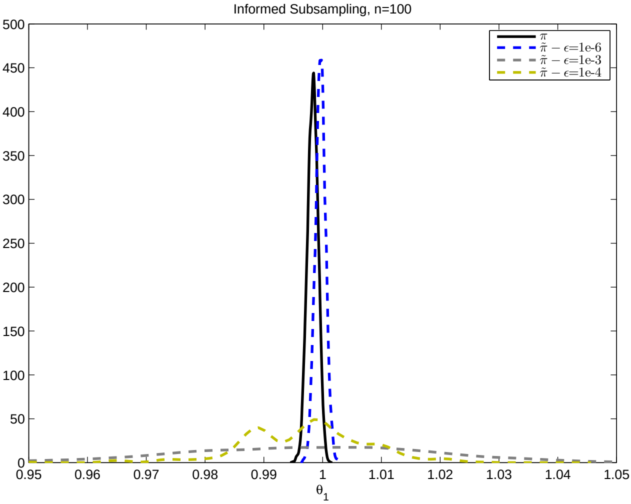
Example 2: validation of summary statistics

Try S defined as the Yule Walker coefficients

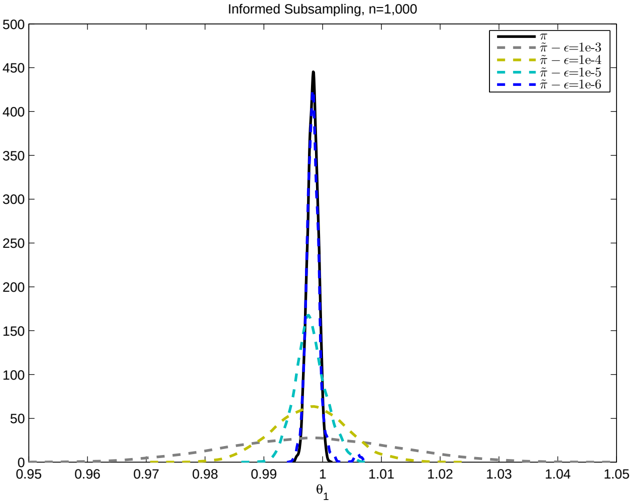


$\Rightarrow$ S accepted since the log ratio does not grow faster than linearly in $\|S(Y) - (N/n)S(Y_U)\|$.

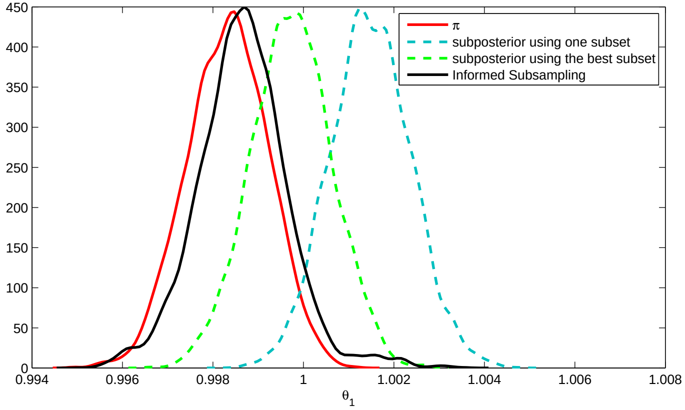
Example 2: marginal inference of θ_1 ($n = 100$)



Example 2: marginal inference θ_1 ($n = 1,000$)



Example 2: marginal inference θ_1 ($n = 1,000$)



Comparing with inference provided by a subposterior $\tilde{\pi}_n(\theta_1 \mid Y_U)$ given a fixed subset U: green is the best subset (as per measured by S) and gray is a subset picked at random.

Example 2: joint inference (θ_2, θ_3)

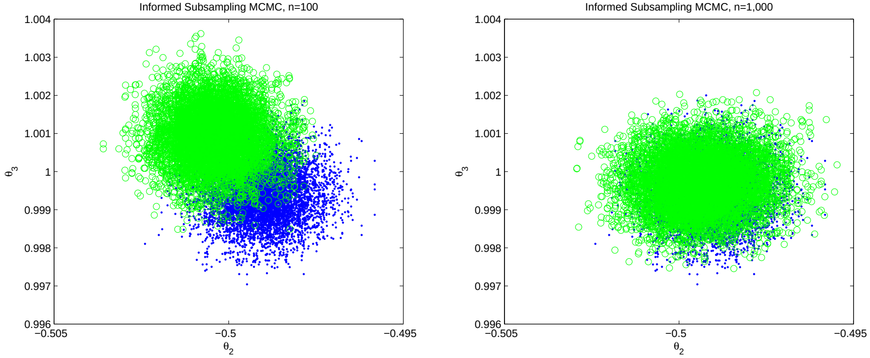


Figure: Samples from $\pi(\theta_2, \theta_3 \mid Y)$ obtained using Metropolis-Hastings (blue) and from $\tilde{\pi}_n(\theta_2, \theta_3 \mid Y)$ obtained using Informed Subsampling MCMC (green), with $n = 100$ (left) and $n = 1,000$ right.

Conclusions

"Uninformed" Subsampling MCMC

- Are designed so as to control locally the decision error wrt to the MH algorithm.
- Checking conditions in which this framework applies may be difficult in practice.
- The number of likelihood evaluation is not fixed, questioning the computational efficiency.
- Subsample the data uniformly at random

By contrast, our Informed Subsampling MCMC scheme

- Allows to control deterministically the MH transition complexity.
- Subsamples according to their fidelity to this full dataset, through summary statistics.
- Allows to control only asymptotically the distance between the chain distribution and the true posterior.