Light and Widely Applicable MCMC: Bayesian inference for large datasets

Florian Maire with Nial Friel & Pierre Alquier

Outline

Topics of interest / Keywords

- $\Rightarrow$ Bayesian inference for large datasets
- $\Rightarrow$ What is a posterior distribution?
- $\Rightarrow$ Markov chain Monte Carlo methods (MCMC)
- $\Rightarrow$ Our method: Light and Widely Applicable MCMC
- $\Rightarrow$ Applications: Shape recognition, Regression ...

Modeling the data

Let $Y_1, Y_2, \ldots, Y_N$ be N data ($\equiv$ observations)

Modeling the data

Let $Y_1, Y_2, \ldots, Y_N$ be N data ($\equiv$ observations)

- Y_i can represent any type of information:
  - a measurement of a physical experimentation,
  - a sensor response,
  - a survey,
  - a graph, an image...

Modeling the data

Let $Y_1, Y_2, \ldots, Y_N$ be N data ($\equiv$ observations)

- Y_i can represent any type of information:
  - a measurement of a physical experimentation,
  - a sensor response,
  - a survey,
  - a graph, an image...
- Y_i can be a real number/vector/matrix

Modeling the data

Let $Y_1, Y_2, \ldots, Y_N$ be N data ($\equiv$ observations)

- Y_i can represent any type of information:
  - a measurement of a physical experimentation,
  - a sensor response,
  - a survey,
  - a graph, an image...
- Y_i can be a real number/vector/matrix

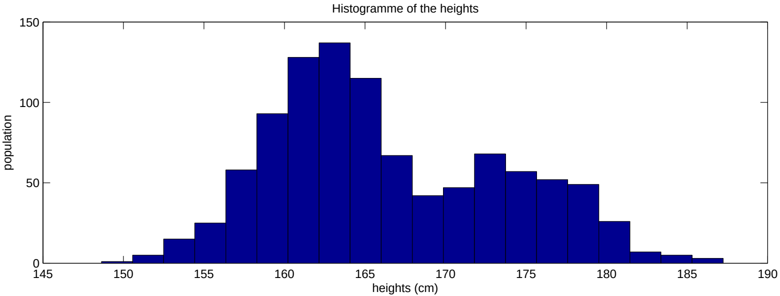
Interest: modeling Y_i is finding a probability distribution f such that

the distribution of $Y_1, Y_2, \ldots$ is roughly f

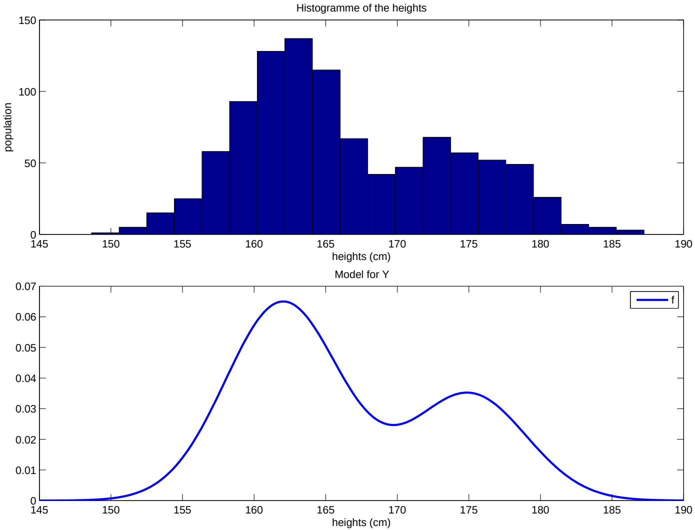
Height survey example

We asked to a group of $N = 1,000$ people their height

- Y_i is the i-th participant height...
- Y_i is a (positive!) real number (measure in cm)



Height survey example



Why modeling is important?

Here the model has been estimated with:

$$f(y) = \underbrace{.65}_{\text{prop. 1}} \times \mathcal{N}(\underbrace{163}_{\text{mean 1}}, \underbrace{4}_{\text{var 1}}, y) + .35 \times \mathcal{N}(172, 4.6, y)\,.$$

Why modeling is important?

Here the model has been estimated with:

$$f(y) = \underbrace{.65}_{\text{prop. 1}} \times \mathcal{N}(\underbrace{163}_{\text{mean 1}}, \underbrace{4}_{\text{var 1}}, y) + .35 \times \mathcal{N}(172, 4.6, y)\,.$$

Modeling Y_i will allow to:

- understand the uncertainty related to the phenomenon of interest
- predict new data $Y_{N+1}, Y_{N+2}, \ldots$
- simulate new data
- cluster the existing data
- estimate missing/latent data, etc.

Why modeling is important?

Here the model has been estimated with:

$$f(y) = \underbrace{.65}_{\text{prop. 1}} \times \mathcal{N}(\underbrace{163}_{\text{mean 1}}, \underbrace{4}_{\text{var 1}}, y) + .35 \times \mathcal{N}(172, 4.6, y)\,.$$

Modeling Y_i will allow to:

- understand the uncertainty related to the phenomenon of interest
- predict new data $Y_{N+1}, Y_{N+2}, \ldots$
- simulate new data
- cluster the existing data
- estimate missing/latent data, etc.

"All models are wrong, some are usefull", G. Box

Likelihood function

Most of the time, f is assumed to belong to a parametric distribution

$$Y_i \sim f \equiv f(\cdot \,|\, \theta), \qquad \theta \in \Theta\,.$$

- θ is called the parameter of the model (mean, variance, correlations...)
- θ can be a real number/vector/matrix
- Θ is the set of all possible parameters

The function $y \to f(y \,|\, \theta)$ is called the likelihood function

Likelihood function

Most of the time, f is assumed to belong to a parametric distribution

$$Y_i \sim f \equiv f(\cdot \mid \theta), \qquad \theta \in \Theta\,.$$

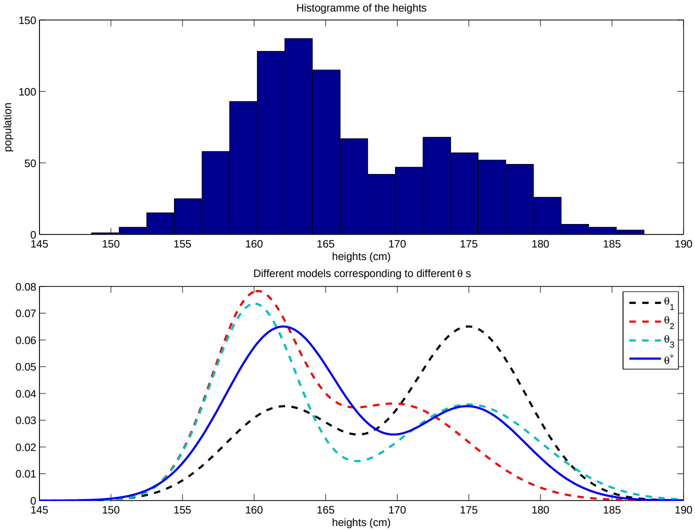
- θ is called the parameter of the model (mean, variance, correlations...)
- θ can be a real number/vector/matrix
- Θ is the set of all possible parameters

The function $y \to f(y \mid \theta)$ is called the likelihood function

The Question: how to estimate a "good" θ, say θ^*, such that

$f(\cdot \mid \theta^*)$ is a "good" model ??

Height survey example



The Bayesian approach

In Bayesian statistics, θ is regarded as a random variable

We are not looking for an unique good θ^* but for a good range of those!

The Bayesian approach

In Bayesian statistics, θ is regarded as a random variable

We are not looking for an unique good θ^* but for a good range of those!

An initial guess on θ is conveyed through its prior distribution p
$\Rightarrow$ so, in absence of data, the distribution of θ is simply

$$\Pr(\theta) = p(\theta)\,.$$

The Bayesian approach

In Bayesian statistics, θ is regarded as a random variable

We are not looking for an unique good θ^* but for a good range of those!

An initial guess on θ is conveyed through its prior distribution p
$\Rightarrow$ so, in absence of data, the distribution of θ is simply

$$\Pr(\theta) = p(\theta)\,.$$

Now, as soon as some data are observed, the distribution of θ is updated:

$$\Pr(\theta \mid Y_1, \dots, Y_N) \propto f(Y_1, \dots, Y_N \mid \theta) p(\theta)\,.$$

$\Rightarrow$ $\Pr(\,\cdot \mid Y_1, \dots, Y_N)$ is called the posterior distribution of θ given $Y_1, \dots, Y_N$.

The Bayesian question

Remember, our objective is to derive the location of the good θs to model $Y_1, \ldots, Y_N$.

The Bayesian question

Remember, our objective is to derive the location of the good θs to model $Y_1, \ldots, Y_N$.

In the Bayesian approach, we want to have a good knowledge of the posterior distribution

The Bayesian question

Remember, our objective is to derive the location of the good θs to model $Y_1, \ldots, Y_N$.

In the Bayesian approach, we want to have a good knowledge of the posterior distribution

The Bayesian question: Having observed $Y_1, \ldots, Y_N$, what is the probability that θ belongs to an interval/region $\mathcal{I}$?

$$\Rightarrow \Pr(\theta \in \mathcal{I} \mid Y_1, \ldots, Y_N) = \frac{\int_{\mathcal{I}} f(Y_1, \ldots, Y_N \mid \theta) p(\mathrm{d}\theta)}{\int_{\Theta} f(Y_1, \ldots, Y_N \mid \theta) p(\mathrm{d}\theta)}$$

The Bayesian question

Remember, our objective is to derive the location of the good θs to model $Y_1, \dots, Y_N$.

In the Bayesian approach, we want to have a good knowledge of the posterior distribution

The Bayesian question: Having observed $Y_1, \dots, Y_N$, what is the probability that θ belongs to an interval/region $\mathcal{I}$?

$$\Rightarrow \Pr(\theta \in \mathcal{I} \mid Y_1, \dots, Y_N) = \frac{\int_{\mathcal{I}} f(Y_1, \dots, Y_N \mid \theta) p(\mathrm{d}\theta)}{\int_{\Theta} f(Y_1, \dots, Y_N \mid \theta) p(\mathrm{d}\theta)}$$

Main issue: for realistic models, there is no hope to get this!

A computational solution

It is possible to approximate the quantity $\Pr(\theta \in \mathcal{I} \mid Y_1, \ldots, Y_N)$ with an arbitrary precision

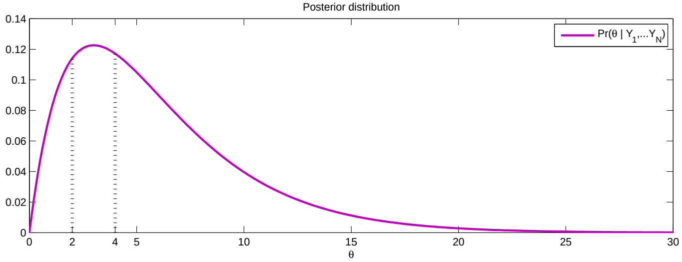
Provided that we can simulate samples from the posterior

$$\theta_1, \theta_2, \ldots \sim \Pr(\,\cdot \mid Y_1, \ldots, Y_N)$$

$\Rightarrow$ Example!

Example

Assume the following posterior distribution $\Pr(\theta \mid Y_1, \ldots, Y_N)$ is unknown



$$\Pr(\theta \in (2,4) \mid Y_1, \ldots, Y_N)?$$

$\Rightarrow$ in this case, we know it exactly: $\Pr(\theta \in (2,4) \mid Y_1, \ldots, Y_N) = .24$

(Surprisingly!) even if $\Pr(\theta \mid Y_1, \ldots, Y_n)$ is unknown,
it may be possible to get samples $\theta_1, \theta_2, \ldots$ from it

Example: 10 samples from $\Pr(\theta \mid Y_1, \ldots, Y_N)$

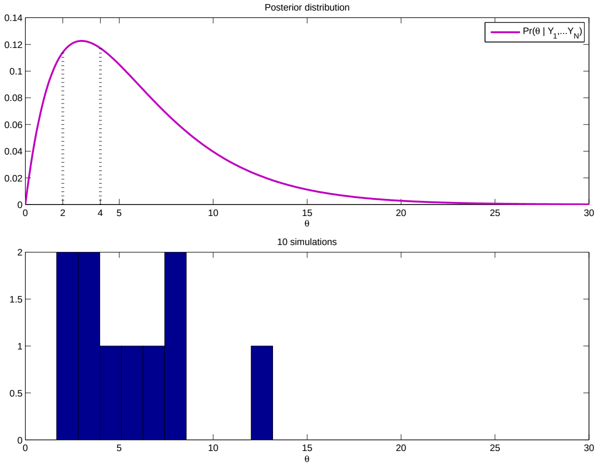


Figure: $\widehat{\Pr}(\theta \in (2,4) \mid Y_1, \ldots, Y_N) = .1$ (true=.24)

Example: 100 samples from $\Pr(\theta \mid Y_1, \dots, Y_N)$

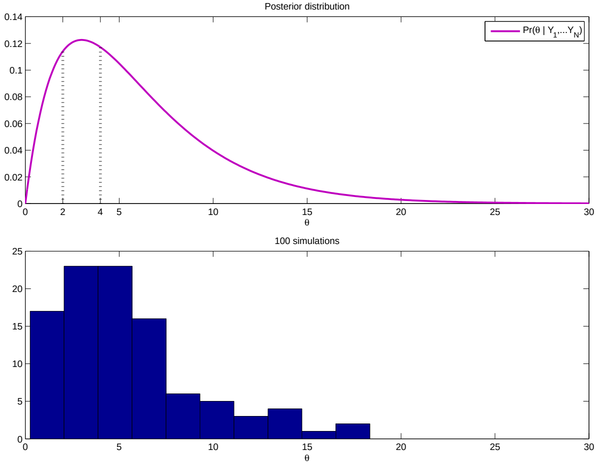


Figure: $\widehat{\Pr}(\theta \in (2, 4) \mid Y_1, \dots, Y_N) = .17$ (true=.24)

Example: 1000 samples from $\Pr(\theta \mid Y_1, \ldots, Y_N)$

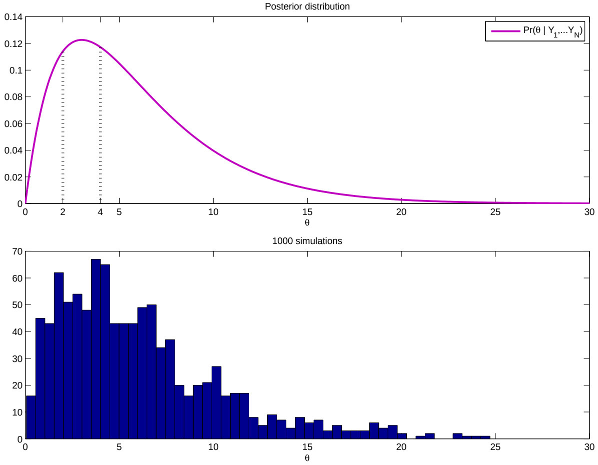


Figure: $\widehat{\Pr}(\theta \in (2,4) \mid Y_1, \ldots, Y_N) = .25$ (true=.24)

Example: 10^5 samples from $\Pr(\theta \mid Y_1, \ldots, Y_N)$

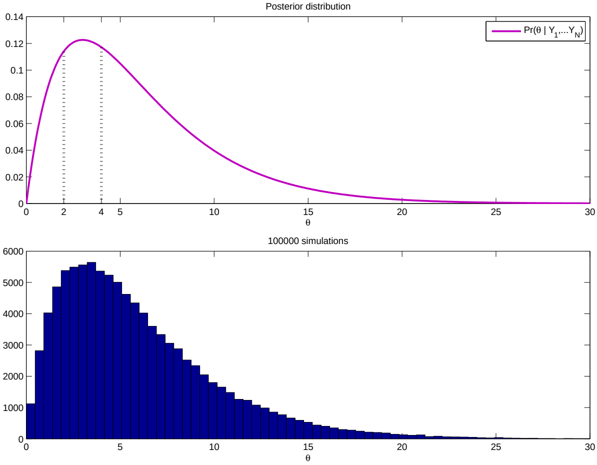


Figure: $\widehat{\Pr}(\theta \in (2,4) \mid Y_1, \ldots, Y_N) = .2408$ (true=.24)

Bridge to our actual problem

In a sense: the initial modeling problem reduces to a simulation problem

Bridge to our actual problem

In a sense: the initial modeling problem reduces to a simulation problem

Question: How do simulation methods cope when the number of data increases, *i.e* $N \to \infty$?

$\Rightarrow$ actually very badly!

We understand that $\Pr(\theta \mid Y_1, \ldots, Y_{100})$ might be less complicated that $\Pr(\theta \mid Y_1, \ldots, Y_{10^6})$

Outline

Topics of interest / Keywords

$\Rightarrow$ Bayesian inference ✓

$\Rightarrow$ Posterior distribution ✓

$\Rightarrow$ Markov chain Monte Carlo methods

$\Rightarrow$ our method: Light and Widely Applicable MCMC

$\Rightarrow$ Applications: Regression, Classification, Shape recognition...

Simulation of a posterior distribution

The most popular method to get samples from $\theta \sim \Pr(\cdot \mid Y_1, \ldots, Y_N)$ is called Metropolis–Hastings (M–H) (1956)

Simulation of a posterior distribution

The most popular method to get samples from $\theta \sim \Pr(\cdot \mid Y_1, \ldots, Y_N)$ is called Metropolis–Hastings (M–H) (1956)

$\Rightarrow$ implementable as long as $\theta \to f(Y_1, \ldots, Y_N \mid \theta)p(\theta)$ is available

Simulation of a posterior distribution

The most popular method to get samples from $\theta \sim \Pr(\cdot \mid Y_1, \ldots, Y_N)$ is called Metropolis–Hastings (M–H) (1956)

$\Rightarrow$ implementable as long as $\theta \to f(Y_1, \ldots, Y_N \mid \theta)p(\theta)$ is available

$\Rightarrow$ theoretically justified

Simulation of a posterior distribution

The most popular method to get samples from $\theta \sim \Pr(\,\cdot \mid Y_1, \ldots, Y_N)$ is called Metropolis–Hastings (M–H) (1956)

$\Rightarrow$ implementable as long as $\theta \to f(Y_1, \ldots, Y_N \mid \theta) p(\theta)$ is available

$\Rightarrow$ theoretically justified

$\Rightarrow$ $50,000+$ citations for the two main papers of the method (G Scholar)

Simulation of a posterior distribution

The most popular method to get samples from $\theta \sim \Pr(\cdot \mid Y_1, \ldots, Y_N)$ is called Metropolis–Hastings (M–H) (1956)

- $\Rightarrow$ implementable as long as $\theta \to f(Y_1, \ldots, Y_N \mid \theta)p(\theta)$ is available
- $\Rightarrow$ theoretically justified
- $\Rightarrow$ $50,000+$ citations for the two main papers of the method (G Scholar)

M–H is a particular instance of a general class of simulation methods called *Markov chain Monte Carlo* algorithms (MCMC)

Metropolis–Hastings algorithm

M–H simulates a non–independent sequence of parameters starting with a random $\theta_0 \in \Theta$

$$\theta_0 \to \theta_1 \to \theta_2 \to \ldots \to \theta_k \,,$$

with the following two steps:

Metropolis–Hastings algorithm

M–H simulates a non–independent sequence of parameters starting with a random $\theta_0 \in \Theta$

$$\theta_0 \to \theta_1 \to \theta_2 \to \ldots \to \theta_k \,,$$

with the following two steps:

(i) from θ_k, propose a candidate $\tilde{\theta} \sim Q(\theta_k, \cdot)$

Metropolis–Hastings algorithm

M–H simulates a non–independent sequence of parameters starting with a random $\theta_0 \in \Theta$

$$\theta_0 \to \theta_1 \to \theta_2 \to \ldots \to \theta_k \,,$$

with the following two steps:

(i) from θ_k, propose a candidate $\tilde{\theta} \sim Q(\theta_k, \cdot)$

(ii) set $\theta_{k+1} = \tilde{\theta}$ with proba.:

$$\alpha(\theta_k, \tilde{\theta}) = \min\left\{1, \frac{f(Y_1, \ldots, Y_N \,|\, \tilde{\theta}) p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{f(Y_1, \ldots, Y_N \,|\, \theta_k) p(\theta_k) Q(\theta_k, \tilde{\theta})}\right\}$$

and $\theta_{k+1} = \theta_k$ otherwise.

Metropolis–Hastings algorithm

M–H simulates a non–independent sequence of parameters starting with a random $\theta_0 \in \Theta$

$$\theta_0 \to \theta_1 \to \theta_2 \to \ldots \to \theta_k \,,$$

with the following two steps:

(i) from θ_k, propose a candidate $\tilde{\theta} \sim Q(\theta_k, \cdot)$

(ii) set $\theta_{k+1} = \tilde{\theta}$ with proba.:

$$\alpha(\theta_k, \tilde{\theta}) = \min\left\{1, \frac{f(Y_1, \ldots, Y_N \,|\, \tilde{\theta}) p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{f(Y_1, \ldots, Y_N \,|\, \theta_k) p(\theta_k) Q(\theta_k, \tilde{\theta})}\right\}$$

and $\theta_{k+1} = \theta_k$ otherwise.

Critical issue: If N is very large, computing $f(Y_1, \ldots, Y_N \,|\, \tilde{\theta})$ at each iteration is prohibitively expensive

Shifting the question

Remember, we want to estimate for any interval $\mathcal{I}$

$$\Pr(\theta \in \mathcal{I} \mid Y_1, \ldots, Y_N) \approx \frac{1}{L} \sum_{\ell=1}^{L} \#\{\theta_\ell \in \mathcal{I}\} \,.$$

Shifting the question

Remember, we want to estimate for any interval $\mathcal{I}$

$$\Pr(\theta \in \mathcal{I} \mid Y_1, \dots, Y_N) \approx \frac{1}{L} \sum_{\ell=1}^{L} \#\{\theta_\ell \in \mathcal{I}\}\,.$$

Classical: How big L should be to reach a given precision?

Shifting the question

Remember, we want to estimate for any interval $\mathcal{I}$

$$\Pr(\theta \in \mathcal{I} \mid Y_1, \ldots, Y_N) \approx \frac{1}{L} \sum_{\ell=1}^{L} \#\{\theta_\ell \in \mathcal{I}\}\,.$$

Classical: How big L should be to reach a given precision?

pb: this does not consider the simulation burden generated by each sample

Shifting the question

Remember, we want to estimate for any interval $\mathcal{I}$

$$\Pr(\theta \in \mathcal{I} \mid Y_1, \ldots, Y_N) \approx \frac{1}{L} \sum_{\ell=1}^{L} \#\{\theta_\ell \in \mathcal{I}\}\,.$$

Classical: How big L should be to reach a given precision?

pb: this does not consider the simulation burden generated by each sample

Topical: For a given CPU budget, how can we derive a tradeoff between precision and feasibility?

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:
(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \,|\, \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \,|\, \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data
$\Rightarrow$ assumptions hold for f

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data
$\Rightarrow$ assumptions hold for f
$\Rightarrow$ in practice, a large portion of the data are used (75-90%)

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data
$\Rightarrow$ assumptions hold for f
$\Rightarrow$ in practice, a large portion of the data are used (75-90%)

Our work aimed at finding a method:

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data
$\Rightarrow$ assumptions hold for f
$\Rightarrow$ in practice, a large portion of the data are used (75-90%)

Our work aimed at finding a method:

- Widely applicable (universal, like M–H)

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data
$\Rightarrow$ assumptions hold for f
$\Rightarrow$ in practice, a large portion of the data are used (75-90%)

Our work aimed at finding a method:

- Widely applicable (universal, like M–H)
- Light (with a controlled CPU cost)

Scaling up Metropolis–Hastings

From 2010 onwards, bypassing this severe limitation in the hot topic (tens of those papers in the best stat journals)

Most of the proposed strategies consist in:

(Sol-1) using an unbiased estimate of the likelihood $f(Y_1, \ldots, Y_N \mid \tilde{\theta})$
$\Rightarrow$ pb specific + estimate might be as costly as likelihood eval.

(Sol-2) taking *with a prescribed probability* the same decision as the original M–H but using a subset of the data
$\Rightarrow$ assumptions hold for f
$\Rightarrow$ in practice, a large portion of the data are used (75-90%)

Our work aimed at finding a method:

- Widely applicable (universal, like M–H)
- Light (with a controlled CPU cost)
- Simple (Black-box/few parameters)

Light and Widely Applicable (LWA) MCMC: motivation

Consider an exponential model *i.e* a likelihood function of the type:

$$f(y \mid \theta) = \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} / Z(\theta)\,, \qquad Z(\theta) = \int \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} \mathrm{d}y\,.$$

Light and Widely Applicable (LWA) MCMC: motivation

Consider an exponential model *i.e* a likelihood function of the type:

$$f(y \mid \theta) = \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} / Z(\theta)\,, \qquad Z(\theta) = \int \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} \mathrm{d}y\,.$$

For the likelihood of a set of independent data, we have:

$$f(Y_1, \ldots, Y_N \mid \theta)^{1/N} = \exp\left\{h(\theta)^{\mathsf{T}} \frac{1}{N} \sum_{\ell=1}^{N} S(Y_\ell)\right\} / Z(\theta)\,.$$

Light and Widely Applicable (LWA) MCMC: motivation

Consider an exponential model *i.e* a likelihood function of the type:

$$f(y \mid \theta) = \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} / Z(\theta)\,, \qquad Z(\theta) = \int \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} \mathrm{d}y\,.$$

For the likelihood of a set of independent data, we have:

$$f(Y_1, \ldots, Y_N \mid \theta)^{1/N} = \exp\left\{h(\theta)^{\mathsf{T}} \frac{1}{N} \sum_{\ell=1}^{N} S(Y_\ell)\right\} / Z(\theta)\,.$$

Now consider a subset of those data $\{Y_\ell,\, \ell \in \mathsf{U}\}$ of <u>size n</u>

$$f(Y_\ell,\, \ell \in \mathsf{U} \mid \theta)^{1/n} = \exp\left\{h(\theta)^{\mathsf{T}} \frac{1}{n} \sum_{\ell \in \mathsf{U}} S(Y_\ell)\right\} / Z(\theta)\,.$$

Light and Widely Applicable (LWA) MCMC: motivation

Consider an exponential model *i.e* a likelihood function of the type:

$$f(y \mid \theta) = \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} / Z(\theta)\,, \qquad Z(\theta) = \int \exp\left\{h(\theta)^{\mathsf{T}} S(y)\right\} \mathrm{d}y\,.$$

For the likelihood of a set of independent data, we have:

$$f(Y_1, \ldots, Y_N \mid \theta)^{1/N} = \exp\left\{h(\theta)^{\mathsf{T}} \frac{1}{N} \sum_{\ell=1}^{N} S(Y_\ell)\right\} / Z(\theta)\,.$$

Now consider a subset of those data $\{Y_\ell,\, \ell \in \mathrm{U}\}$ of <u>size *n*</u>

$$f(Y_\ell,\, \ell \in \mathrm{U} \mid \theta)^{1/n} = \exp\left\{h(\theta)^{\mathsf{T}} \frac{1}{n} \sum_{\ell \in \mathrm{U}} S(Y_\ell)\right\} / Z(\theta)\,.$$

Think about the case where: $\frac{1}{n} \sum_{\ell \in \mathrm{U}} S(Y_\ell) = \frac{1}{N} \sum_{\ell=1}^{N} S(Y_\ell)$

LWA–MCMC: the intuition

For this (very specific) setup: inference based on N data $Y_1, \ldots, Y_N$ is the same as using a subset of n data $\{Y_\ell,\, \ell \in \mathrm{U}\}$ provided that

$$\frac{1}{N}\sum_{\ell=1}^{N} S(Y_\ell) = \frac{1}{n}\sum_{\ell\in\mathrm{U}} S(Y_\ell)\,.$$

LWA–MCMC: the intuition

For this (very specific) setup: inference based on N data $Y_1, \ldots, Y_N$ is the same as using a subset of n data $\{Y_\ell,\, \ell \in \mathrm{U}\}$ provided that

$$\frac{1}{N}\sum_{\ell=1}^{N} S(Y_\ell) = \frac{1}{n}\sum_{\ell \in \mathrm{U}} S(Y_\ell)\,.$$

Our intuition: for any type of models, if it exists a subset $\{Y_\ell\,, \ell \in \mathrm{U}\}$ s.t.

$$\mathcal{L}(Y_1, \ldots, Y_N) \approx \mathcal{L}(Y_\ell, \ell \in \mathrm{U})\,,$$

then

$$\Pr(\,\cdot \mid Y_1, \ldots, Y_N) \approx \Pr(\,\cdot \mid Y_\ell,\, \ell \in \mathrm{U})\,.$$

LWA–MCMC: critical questions

- How to assess that $\mathcal{L}(Y_1, \ldots, Y_N) \approx \mathcal{L}(Y_\ell, \ell \in \mathsf{U})$?

LWA–MCMC: critical questions

- How to assess that $\mathcal{L}(Y_1, \ldots, Y_N) \approx \mathcal{L}(Y_\ell, \ell \in \mathsf{U})$?

$\Rightarrow$ we use a set of (pb specific) summary statistics S (moments, quantiles,...) and assign to each subset U of size n a weight

$$\omega(\mathsf{U}) \propto \exp\left\{-\epsilon \left\| \frac{1}{N}\sum_{\ell=1}^{N} S(Y_\ell) - \frac{1}{n}\sum_{\ell \in \mathsf{U}} S(Y_\ell) \right\|^2 \right\}, \qquad \epsilon > 0\,.$$

LWA–MCMC: critical questions

• How to assess that $\mathcal{L}(Y_1, \dots, Y_N) \approx \mathcal{L}(Y_\ell, \ell \in \mathsf{U})$?
$\Rightarrow$ we use a set of (pb specific) summary statistics S (moments, quantiles,...) and assign to each subset U of size n a weight

$$\omega(\mathsf{U}) \propto \exp\left\{-\epsilon \left\| \frac{1}{N}\sum_{\ell=1}^{N} S(Y_\ell) - \frac{1}{n}\sum_{\ell \in \mathsf{U}} S(Y_\ell) \right\|^2 \right\}, \qquad \epsilon > 0\,.$$

• Which subset to choose from? (there are $\binom{N}{n}$ of them)

LWA–MCMC: critical questions

• How to assess that $\mathcal{L}(Y_1, \ldots, Y_N) \approx \mathcal{L}(Y_\ell, \ell \in \mathsf{U})$?
$\Rightarrow$ we use a set of (pb specific) summary statistics S (moments, quantiles,...) and assign to each subset U of size n a weight

$$\omega(\mathsf{U}) \propto \exp\left\{-\epsilon \left\| \frac{1}{N}\sum_{\ell=1}^{N} S(Y_\ell) - \frac{1}{n}\sum_{\ell \in \mathsf{U}} S(Y_\ell) \right\|^2 \right\}, \qquad \epsilon > 0\,.$$

• Which subset to choose from? (there are $\binom{N}{n}$ of them)
$\Rightarrow$ we refuse to choose!
$\Rightarrow$ each subset U should be involve in the process according to $\omega(\mathsf{U})$
$\Rightarrow$ the different subsets will act complimentarily

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \ldots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_\mathrm{k})\,,$$

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \ldots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_k)\,,$$

step (i) Refreshing the subset U_k

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \ldots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_k),$$

step (i) Refreshing the subset U_k
$\Rightarrow$ propose a new subset $\tilde{\mathrm{U}} \sim K(\mathrm{U}_\mathrm{k}, \cdot)$

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \ldots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_k)\,,$$

step (i) Refreshing the subset U_k

$\Rightarrow$ propose a new subset $\tilde{\mathrm{U}} \sim K(\mathrm{U}_\mathrm{k}, \cdot)$

$\Rightarrow$ set $\mathrm{U}_{\mathrm{k}+1} = \tilde{\mathrm{U}}$ with proba. $\min(1, \omega(\tilde{\mathrm{U}})/\omega(\mathrm{U}_\mathrm{k}))$ and $\mathrm{U}_{\mathrm{k}+1} = \mathrm{U}_\mathrm{k}$ otherwise

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \ldots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_k)\,,$$

step (i) Refreshing the subset U_k

$\Rightarrow$ propose a new subset $\tilde{\mathrm{U}} \sim K(\mathrm{U}_\mathrm{k}, \cdot)$

$\Rightarrow$ set $\mathrm{U}_{\mathrm{k}+1} = \tilde{\mathrm{U}}$ with proba. $\min(1, \omega(\tilde{\mathrm{U}})/\omega(\mathrm{U}_\mathrm{k}))$ and $\mathrm{U}_{\mathrm{k}+1} = \mathrm{U}_\mathrm{k}$ otherwise

step (ii) Simulation of θ_{k+1}

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \dots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_k)\,,$$

step (i) <u>Refreshing the subset U_k</u>

$\Rightarrow$ propose a new subset $\tilde{\mathrm{U}} \sim K(\mathrm{U}_\mathrm{k}, \cdot)$

$\Rightarrow$ set $\mathrm{U}_{\mathrm{k}+1} = \tilde{\mathrm{U}}$ with proba. $\min(1, \omega(\tilde{\mathrm{U}})/\omega(\mathrm{U}_\mathrm{k}))$ and $\mathrm{U}_{\mathrm{k}+1} = \mathrm{U}_\mathrm{k}$ otherwise

step (ii) <u>Simulation of θ_{k+1}</u>

$\Rightarrow$ from θ_k, propose a candidate $\tilde{\theta} \sim Q(\theta_k, \cdot)$

Light and Widely Applicable MCMC: the algorithm

Starting with a random $\theta_0 \in \Theta$, $\mathrm{U}_0 \subseteq \{1, \ldots, N\}$ and $|\mathrm{U}_0| = n$

$$(\theta_0, \mathrm{U}_0) \to (\theta_1, \mathrm{U}_1) \to (\theta_2, \mathrm{U}_2) \to \cdots \to (\theta_k, \mathrm{U}_k)\,,$$

step (i) Refreshing the subset U_k

$\Rightarrow$ propose a new subset $\tilde{\mathrm{U}} \sim K(\mathrm{U}_k, \cdot)$

$\Rightarrow$ set $\mathrm{U}_{k+1} = \tilde{\mathrm{U}}$ with proba. $\min(1, \omega(\tilde{\mathrm{U}})/\omega(\mathrm{U}_k))$ and $\mathrm{U}_{k+1} = \mathrm{U}_k$ otherwise

step (ii) Simulation of θ_{k+1}

$\Rightarrow$ from θ_k, propose a candidate $\tilde{\theta} \sim Q(\theta_k, \cdot)$

$\Rightarrow$ set $\theta_{k+1} = \tilde{\theta}$ with proba.:

$$\alpha(\theta_k, \tilde{\theta} \,|\, \mathrm{U}_{k+1}) = \min\left\{1, \frac{f(Y_\ell, \ell \in \mathrm{U}_{k+1} \,|\, \tilde{\theta})p(\tilde{\theta})Q(\tilde{\theta}, \theta_k)}{f(Y_\ell, \ell \in \mathrm{U}_{k+1} \,|\, \theta_k)p(\theta_k)Q(\theta_k, \tilde{\theta})}\right\}$$

and $\theta_{k+1} = \theta_k$ otherwise.

LWA–MCMC: summary

Given the CPU budget τ available

LWA–MCMC: summary

Given the CPU budget τ available

(i) select the subset size n

$\Rightarrow$ to get a reasonable nb of samples $\theta_1, \theta_2, \ldots$ using the ressource τ

LWA–MCMC: summary

Given the CPU budget τ available

(i) select the subset size n
$\Rightarrow$ to get a reasonable nb of samples $\theta_1, \theta_2, \ldots$ using the ressource τ

(ii) decide on a set of summary statistics S and on ϵ

LWA–MCMC: summary

Given the CPU budget τ available

(i) select the subset size n
$\Rightarrow$ to get a reasonable nb of samples $\theta_1, \theta_2, \ldots$ using the ressource τ

(ii) decide on a set of summary statistics S and on ϵ

(iii) run LWA–MCMC
$\Rightarrow$ retrieve samples $\theta_1, \theta_2, \ldots, \theta_L$

LWA–MCMC: summary

Given the CPU budget τ available

(i) select the subset size n
$\Rightarrow$ to get a reasonable nb of samples $\theta_1, \theta_2, \ldots$ using the ressource τ

(ii) decide on a set of summary statistics S and on ϵ

(iii) run LWA–MCMC
$\Rightarrow$ retrieve samples $\theta_1, \theta_2, \ldots, \theta_L$

The set of "good" parameters can now be found in an interval $\mathcal{I}^*$ such that

$$\Pr(\theta \in \mathcal{I}^* \mid Y_1, \ldots, Y_N) \approx \frac{1}{L} \sum_{\ell=1}^{L} \#\{\theta_\ell \in \mathcal{I}^*\}$$

is sufficiently high.

Outline

Topics of interest / Keywords

⇒ Bayesian inference ✓

⇒ Posterior distribution ✓

⇒ Markov chain Monte Carlo methods ✓

⇒ our method: Light and Widely Applicable MCMC ✓

⇒ Applications: Regression, Classification, Shape recognition...

Example: estimation of template shapes

Data are 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 writes:

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

Example: estimation of template shapes

Computational budget: 60 mins, we choose $n = 100$

S is the proportion of digits of each class

$\Rightarrow$ We maintain in each subset the correct proportion of 1,2,etc.

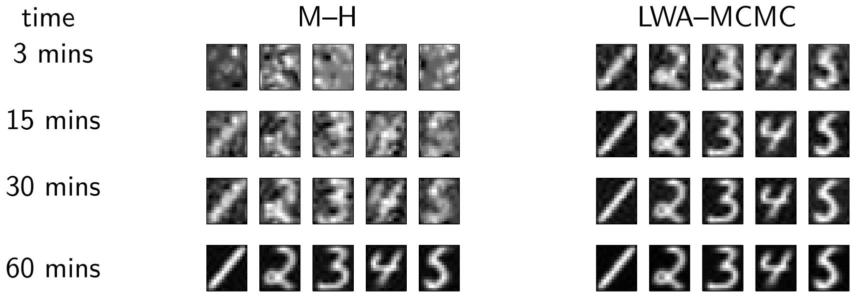
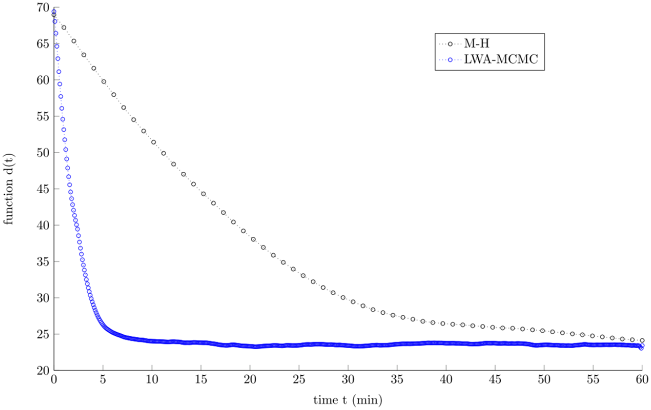


Figure: Efficiency of template estimation through M–H and LWA–MCMC.

Example: estimation of template shapes

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



Example: regression in ARMA model

Data: a very long time series $\{Y_t,\ t \in \mathbb{N}\}$

$$Y_{t+1} = \alpha Y_t + \beta Z_t + Z_{t+1} + \gamma$$

where

- $Z_{t+1} \sim \mathcal{N}(0, \sigma^2)$
- we set $\alpha = 0.5$, $\beta = 0.7$, $\gamma = 1$, $\sigma = 1$
- Summary statistic S: autocorrelation time

We want to estimate $\theta = (\alpha, \beta, \gamma)$ from the observations $Y_1, Y_2, \ldots$

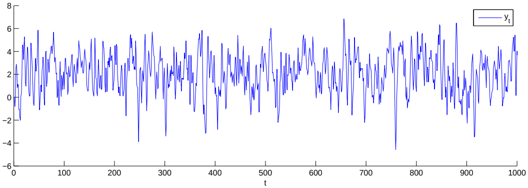


Figure: Realization of an ARMA of lenght $T = 10^7$

Example: regression in ARMA model

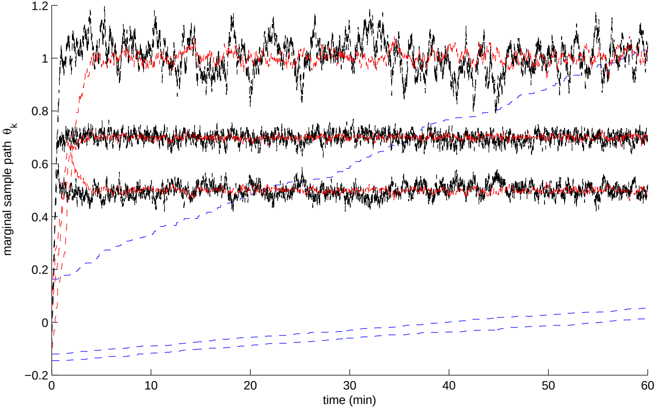
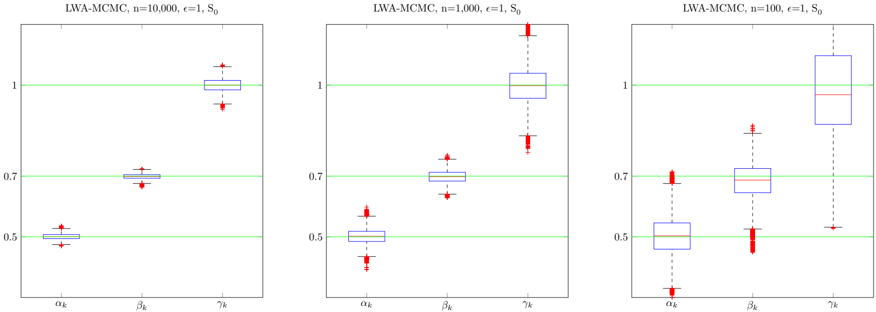


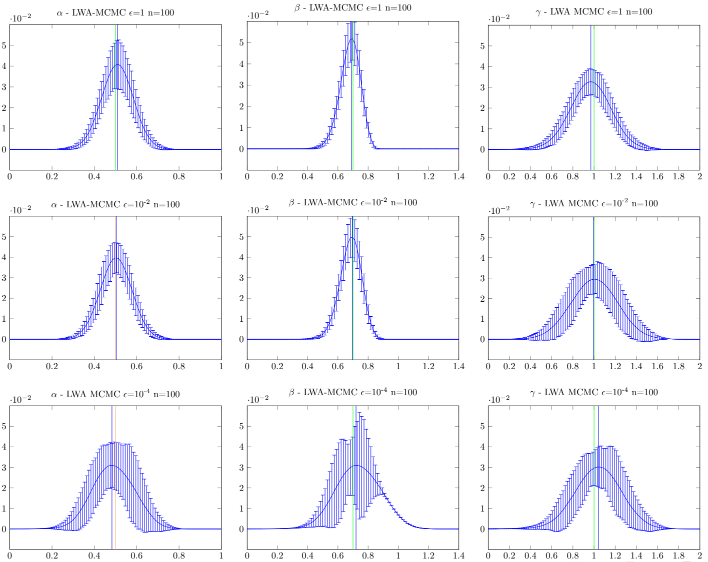
Figure: M–H (dashed, blue) and LWA–MCMC (dashed, black $n = 100$ and red $n = 1000$)

Example: regression in ARMA model



Evolution of the estimates of α, β and γ for different subset sizes $n = 10,000$, $n = 1000$ and $n = 100$

Example: regression in ARMA model



Conclusion

- LWA–MCMC approach works on subsets of data which are representative of the full data set

Conclusion

- LWA–MCMC approach works on subsets of data which are representative of the full data set
- for a given CPU budget, the number of samples L from a proxy of $\Pr(\,\cdot \mid Y_1, \ldots, Y_N)$ can be made arbitrarily large

Conclusion

- LWA–MCMC approach works on subsets of data which are representative of the full data set
- for a given CPU budget, the number of samples L from a proxy of $\Pr(\cdot \mid Y_1, \ldots, Y_N)$ can be made arbitrarily large
- obviously as $n/N \ll 1$, results deteriorate...

Conclusion

- LWA–MCMC approach works on subsets of data which are representative of the full data set
- for a given CPU budget, the number of samples L from a proxy of $\Pr(\cdot \mid Y_1, \ldots, Y_N)$ can be made arbitrarily large
- obviously as $n/N \ll 1$, results deteriorate...
- but inference can be corrected by being more "picky" with respect to the subsets

Conclusion

- LWA–MCMC approach works on subsets of data which are representative of the full data set
- for a given CPU budget, the number of samples L from a proxy of $\Pr(\cdot \mid Y_1, \ldots, Y_N)$ can be made arbitrarily large
- obviously as $n/N \ll 1$, results deteriorate...
- but inference can be corrected by being more "picky" with respect to the subsets

$\Rightarrow$ A ready-to-use method for efficient Bayesian inference in large data contexts

$\Rightarrow$ Work ahead investigates the theoretical implication of our approximation

Thank you for your interest!