Approximated Bayesian Inference and Applications to Large Data Sets

Florian Maire, UCD

joint work with : Nial Friel & Pierre Alquier

Working Group on Statistical Learning, 23th of September 2014

Outlines

Outlines

Bayesian inference at large

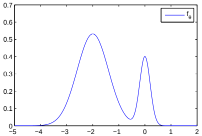
- Modelling & Data Analysis using Bayesian methods :

$\{Y_1, Y_2, \ldots, Y_N\}$ ⇝ **Black Box** ⇝

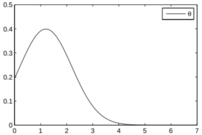
Data

- Physical measurements,
- Experimentation results,
- Surveys, etc...

Model selection:
Likelihood f_θ



Model fitting:
Post. Distribution θ



- Robustness and simplicity are attracting for a wide range of people / domains

Estimation of the parameter

- The data are random var. on $(\mathsf{Y}, \mathcal{Y})$ (typ. $\mathsf{Y} \subseteq \mathbb{R}^p$)
- The parameter θ is (regarded as) random var. on (Θ, ϑ) (typ. $\Theta \subseteq \mathbb{R}^d$)
- Given:
  - (i) a likelihood model $f_\theta \equiv f(\cdot \,|\, \theta)$ on $(\mathsf{Y}, \mathcal{Y})$,
  - (ii) a prior dist. p for θ on (Θ, ϑ)
- define the posterior distribution of θ given $Y_{1:N} = (Y_1, \ldots, Y_N) \in \mathsf{Y}^N$

$$\pi(\theta \,|\, Y_{1:N}) \propto f(Y_{1:N} \,|\, \theta) p(\theta)$$

- our primary objective is to gain knowledge of π, (we assume likelihood model and prior known and fixed...)

Markov chain Monte Carlo: the black box!!

- Seminal papers late 80's/early 90's [1] popularised the use of Markov chains targeting π to explore the state space Θ
- The Metropolis-Hastings (M–H) sampler being the most straightforward *black box*
- Start from some initial state $\theta_0 \in \Theta$. At step k:
  - (i) Propose a move $\tilde{\theta} \sim Q(\theta_k, \cdot)$
  - (ii) Set θ_{k+1} as the next state of the chain if event E_k is realized:

$$E_k = \left\{ U \leq \frac{f(\tilde{\theta} \mid Y_{1:N}) p(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)}{f(\theta_k \mid Y_{1:N}) p(\theta_k) Q(\theta_k, \tilde{\theta})}, \quad U \sim \text{Uni}(0,1) \right\}$$

What if N becomes larger and larger?? (*e.g* $N > 10^6$)

[1] Tanner & Wong (1987), Gelfand & Smith (1990), Tierney (1994),...

The N case

- A likelihood function evaluation has a complexity in $\mathcal{O}(N)$
- M–H (or other MCMC's) & optimization methods are severely hampered by a large N
- When comparing MCMC algorithms
  (i) Autocorrelation
  (ii) Asymptotic Variance
  (iii) Time of transition
- From this perspective, one can expect M–H to be badly ranked!
- Example: for a likelihood function

$$f(\,\cdot\,|\,\theta) = 0.8\,\mathcal{N}(0.3, 0.8) + 0.2\,\mathcal{N}(4, 1)$$

and *i.i.d.* data

N	1.000	10.000	100.000	10^6	10^7
M–H trans. CPU Time	0.016	0.151	1.53	15.40	151.87

Main Problematic

How to rescue the traditional Bayesian analysis methods

(i) from being overwhelmed by N,

(ii) while still preserving the *black box* thing?

Outlines

Profusion of Research on this topic over the last years

- *Exact* Methods:
  - Using unbiased estimate of $f(\theta \mid Y_{1:N})$ for all $\theta \in \Theta$

    (Pseudo-Marginal literature, Andrieu & Vihola 2012, Doucet et al 2012)
  - 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
- *Approximated* Methods with error control
  - A proxy of the M–H kernel with complexity $\leq \mathcal{O}(N)$

    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

And also at UCD!

- Connected with other Research activities at UCD
- Bayesian inference in large networks
  - Aidan Boland: Noisy M–H / Application to the Ising model
  - Lampros Bouranis: Composite Likelihood Inference / Application to Exponential Random Graph model
  - and probably others!

Korattikara et al. / Bardenet et al.

Roughly share the same idea:

- Rewrite the acceptance step of M–H as the realization of the event

$$E_k = \left\{ \frac{1}{N} \sum_{k=1}^{N} \log \frac{f(Y_k \mid \tilde{\theta})}{f(Y_k \mid \theta_k)} \geq \frac{1}{N} \log U \frac{p(\tilde{\theta})Q(\tilde{\theta}, \theta_k)}{p(\theta_k)Q(\theta_k, \tilde{\theta})}, \quad U \sim \mathrm{Uni}(0,1) \right\}$$

- Draw wo replacement, sub batch of data from the data set (successively) up until the event E_ℓ is realized:

$$\tilde{E}_{k,\ell} = \left\{ \left| \frac{1}{n_\ell} \sum_{k=1}^{n_\ell} \log \frac{f(Y_{u_k} \mid \tilde{\theta})}{f(Y_{u_k} \mid \theta_k)} - \psi(\theta_k, \tilde{\theta}, U) \right| > \eta_\ell \right\}$$

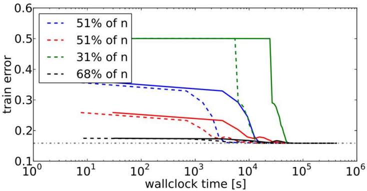
- The threshold η_ℓ is defined so that

$$\mathbb{P}[E_k = \tilde{E}_{k,\ell}] \geq \epsilon$$

So why should we keep on asking questions?!

Three main reasons:

- As the Markov chain gets closer to equilibrium $n_\ell \to N$ *i.e* all data are used
- Computational gains are highly model specific:



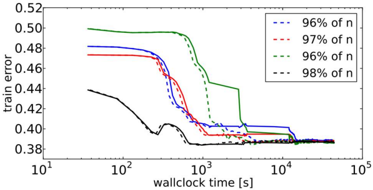


Figure: Two different classification tasks (*Covtype* Dataset (l) and a Synthetic 2D binary decision (r) in Bardenet et al.)

- only applicable for independent data

Our Motivations

- Design a new MCMC approach so that, by construction, each transition's complexity is deterministic in $\mathcal{O}(n)$, $n \ll N$
- Do not restrict to *i.i.d.* data
  - Markov models,
  - Time series,
  - Networks...
- While all the mentioned approaches have stand by the standard posterior distribution $\pi(\,\cdot \mid Y_{1:N})$, we rather investigate the feasibility / efficiency to learn from a changing subset data of size $n \ll N$
- We don't consider a pre-processing data reduction step (ACP, clustering,...) as we want a method as simple as it can gets, (*black-box*)

Outlines

Learning from a proxy of π

- We fix $n \in \mathbb{N}$, $n \ll N$
- Let $\mathcal{U}_n$ be the set of all possible integer combinations such that:

$$\mathcal{U}_n = \{U = (U_1, \ldots, U_n) \in [1, N]^n, \quad \forall (i,j) \in [1, n],\ U_i \neq U_j\}$$

- The question we address is twofold:

(i) Does it exist a subset $\mathcal{U}_n^\star \subseteq \mathcal{U}_n$ s.t.

for $U \in \mathcal{U}_n^\star$, $\quad \pi(\theta \mid Y_k, k \in U) := \pi(\theta \mid Y_U) \approx \pi(\theta \mid Y_{1:N})$

(ii) How can we find such $\mathcal{U}_n^\star$?

Representativeness of a subset of data

We introduce a *Summary Statistics* mapping, projecting a batch of data $\{Y_U,\ U \in \mathcal{U}_n\}$ onto a space of smaller dimension $\mathcal{S} \subseteq \mathbb{R}^m$

$$S_n : \mathcal{U}_n \to \mathcal{S}$$

Define the probability measure $\nu_{n,\epsilon}$ on the discrete state space $(\mathcal{U}_n, \mathscr{U}_n)$

$$\forall\, U \in \mathcal{U}_n, \quad \nu_{n,\epsilon}(U) = \frac{\Phi\left(\|S(U) - s\|/\epsilon\right)}{\sum_{V \in \mathcal{U}_n} \Phi\left(\|S(V) - s\|/\epsilon\right)}$$

where:

- $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is a kernel function
- $\epsilon > 0$ is a *bandwidth* attached to Φ
- $s = S_N(\{1, \ldots, N\})$ the summary statistic vector of the full data set

Heuristic

The intuition is that for all $(U, V) \in \mathcal{U}_n^2$

$$\nu_{n,\epsilon}(U) > \nu_{n,\epsilon}(V) \rightsquigarrow d(\pi; \tilde{\pi}(\,\cdot \mid Y_U)) \leq d(\pi; \tilde{\pi}(\,\cdot \mid Y_V)), \tag{1}$$

for some distance measure on the set of proba on (Θ, ϑ).

- (1) requires a *reasonable* choice of S_n to be meaningful
- Connection with ABC (Approximated Bayesian Computation)

ABC	Subset Inf.
$\theta \sim Q$	$U \sim R$
$\tilde{Y} \sim f(\cdot \mid \theta)$	$\theta \sim \pi(\,\cdot \mid Y_U)$
Accept θ with probability	
$\nu_{N,\epsilon}(\tilde{Y})$	$\nu_{n,\epsilon}(U)$

- Take advantage of ABC literature to design relevant S_n

The case of curved exponential family models

- Consider *i.i.d.* observations from some exponential model $Y_k \sim f(\,\cdot\,|\,\theta)$, where

$$f(y\,|\,\theta) = \exp\left\langle h(\theta), S(y)\right\rangle \Big/ \int_{\mathsf{Y}} \exp\left\langle h(\theta), S(y')\right\rangle \mathrm{d}y'$$

- Here, the choice of *Summary Statistics* in our approach is naturally provided by the *Sufficient Statistics* of the exponential model
- In this special case, we show that given $U \in \mathcal{U}_n$

$$\mathrm{KL}\left(\pi \,\|\, \tilde{\pi}(\,\cdot\,|\,Y_U)\right) \leq \Psi(n, N, Y_{1:N}, p) + B(U)\,, \tag{2}$$

where $B : \mathcal{U}_n \to \mathbb{R}^+$ such that for all $U \in \mathcal{U}_n$

$$B(U) = 0 \iff \frac{1}{N}\sum_{k=1}^{N} S(Y_k) = \frac{1}{n}\sum_{k\in U} S(Y_k)\,.$$

Regarding U as a missing parameter of the model

- These two arguments give credit to the intuition that "some subsets are better than others"
- Issues:
  - $\mathcal{U}_n^\star$ is unlikely to be restricted to a single element (esp. as $d \nearrow$)
  - and even in such a case, wouldn't it be more interesting to account for a collection of good subset
- A collection of good subsets may act somehow complementarily to track π
- Define the proxy of the target as

$$\tilde{\pi}_{n,\epsilon}(\theta \mid Y_{1:N}) = \sum_{U \in \mathcal{U}_n} \tilde{\pi}(\theta \mid Y_U)\nu_{n,\epsilon}(U)$$

yielding a mixture model with $\binom{n}{k}$ components...

First example: Probit model–1

Sample $(Y_1, \ldots, Y_N) \in (\{0\}, \{1\})^N$, independently from the model

$$(i)\ X_k \sim \mathcal{N}(\mu, 1), \qquad (ii)\ Y_k = \mathbb{1}_{\{X_k > 0\}}$$

Can we estimate $\mu \in \mathbb{R}$ from $\tilde{\pi}_{n,\epsilon}$ rather than from π?

- Settings: $N = 1000$, $n = 100$, $\epsilon = 1$, $S_n(U) = \frac{1}{n} \sum_{k \in U} Y_k$
- In this toy example, the likelihood evaluation is NOT in $\mathcal{O}(N)$ and the *exact* posterior writes:

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

- $\pi(\,\cdot \mid Y_{1:N})$ can be explored through standard M–H
- Similarly, given $U \in \mathcal{U}_n$, $\tilde{\pi}(\,\cdot \mid Y_U)$ can be estimated by standard M–H

First example: Probit model–2

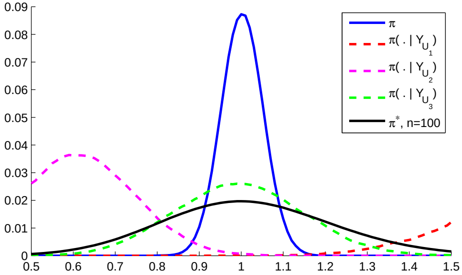
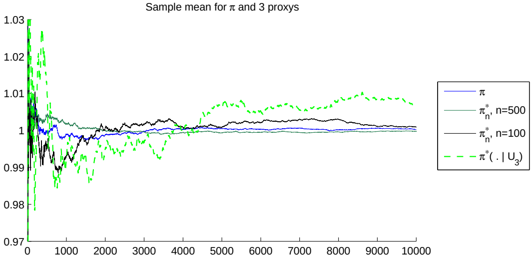


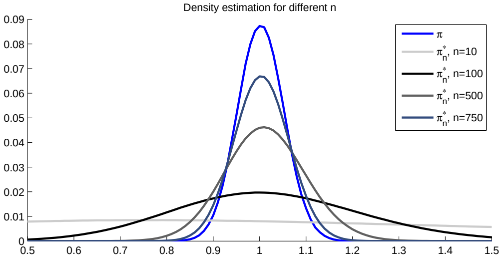
Figure: Density estimation – $S_n(U_1) = 0.71$, $S_n(U_2) = 0.77$, $S_n(U_3) = 0.84$, $S_N = 0.843$

- At first sight, $\tilde{\pi}_{n,\epsilon}$ remains far from π...
- However, our main interest is to approximate the expectation

$$\int_\Theta H(\theta)\pi(\mathrm{d}\theta \mid Y_{1:N}) \quad \text{by} \quad \int_\Theta H(\theta)\tilde{\pi}_{n,\epsilon}(\mathrm{d}\theta \mid Y_{1:N})$$

First example: Probit model–3





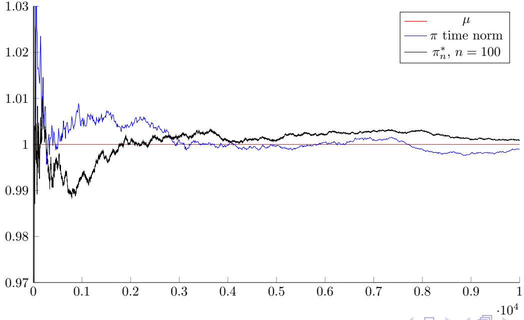
First example: Probit model–4

- Variance of the TCL estimate

$$\sigma_L^2 = \frac{1}{L}\mathrm{Var}\left(\sum_{k=1}^{L}\mu_k\right)$$

$\sigma_L^2 = 0.0105$ for π, $\sigma_L^2 = 0.0305$ for $\tilde{\pi}_{n,\epsilon}$ for $L = 10.000$

- but when we "time normalize":



A general approach

- In general, sampling from the mixture

$$\tilde{\pi}_{n,\epsilon}(\theta \mid Y_{1:N}) = \sum_{U \in \mathcal{U}_n} \tilde{\pi}(\theta \mid Y_U)\nu_{n,\epsilon}(U)$$

  is not feasible ($N \gg n$, model more complex than the Probit example...)

- We propose an MCMC algorithm on the extended state space $(\Theta \times \mathcal{U}_n, \vartheta \otimes \mathscr{U}_n)$ with target distribution

$$\tilde{\pi}_{n,\epsilon}(\theta, U \mid Y_{1:N}) = \tilde{\pi}(\theta \mid Y_U)\nu_{n,\epsilon}(U)$$

- The Markov chain $\{(\theta_k, U_k),\ k \in \mathbb{N}\}$ will marginally target $\tilde{\pi}_{n,\epsilon}(\,\cdot \mid Y_{1:N})$

The ideal Markov chain

The desired scheme of the chain would be as follow:

$$
\begin{tikzcd}
U_0 \arrow[dr] & & U_k \arrow[r] \arrow[dd] \arrow[ddr] & U_{k+1} \arrow[r] \arrow[dd] \arrow[ddr] & U_{k+2} \arrow[r] \arrow[dd] \arrow[ddr] & \cdots \\
& (\ldots) \arrow[ur] \arrow[dr] & & & & \\
\theta_0 \arrow[ur] & & \theta_k \arrow[r] & \theta_{k+1} \arrow[r] & \theta_{k+2} \arrow[r] & \cdots
\end{tikzcd}
$$

Figure: Intertwined structure of the desired Markov chain

To avoid getting stuck on some *optimal* block of data:

(i) make two distinct decisions for a move on Θ and on $\mathcal{U}_n$

(ii) U_{k+1} should depend only on U_k for optimal mixing

mimicking independence sampler (if $\nu_{n,\epsilon}$ could be drawn from!)

The Markov chain we actually use...

$$
\begin{tikzcd}
U_0 \arrow[dr] & & U_k \arrow[r, "\text{(i) propose: } U' \sim r(U_k,\cdot)", "\text{(ii) accept wp } \alpha_1(U_k, U')"'] & U_{k+1} \arrow[r, "\text{copy}"] & U_{k+2} = U_{k+1} \arrow[r] & \cdots \\
& (\ldots) \arrow[ur] \arrow[dr] & & & & \\
\theta_0 \arrow[ur] & & \theta_k \arrow[r, "\text{copy}"] & \theta_{k+1} = \theta_k \arrow[r, "\text{(i) propose: } \theta' \sim q(\theta_{k+1}, U_{k+1};\cdot)", "\text{(ii) accept wp } \alpha_2(\theta_{k+1}, U_{k+1}; U')"'] & \theta_{k+2} \arrow[r] & \cdots
\end{tikzcd}
$$

Figure: A Markov chain with two independent decisions

We haven't been able yet to find a proper way to make the marginal chain $\{U_k,\ k \in \mathbb{N}\}$ independent of $\{\theta_k,\ k \in \mathbb{N}\}$

ARMA model

Observation $\{Y_t,\ t \in \mathbb{N}\}$

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

where

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

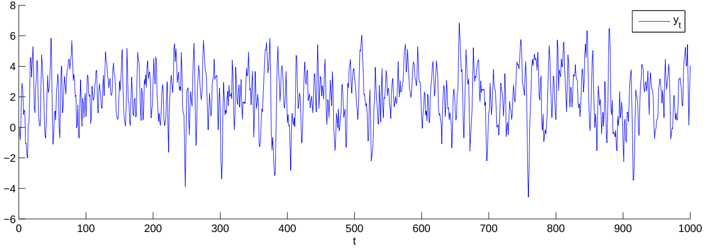


Figure: Realization of an ARMA of lenght $T = 10.000$

Influence of ϵ on α estimate

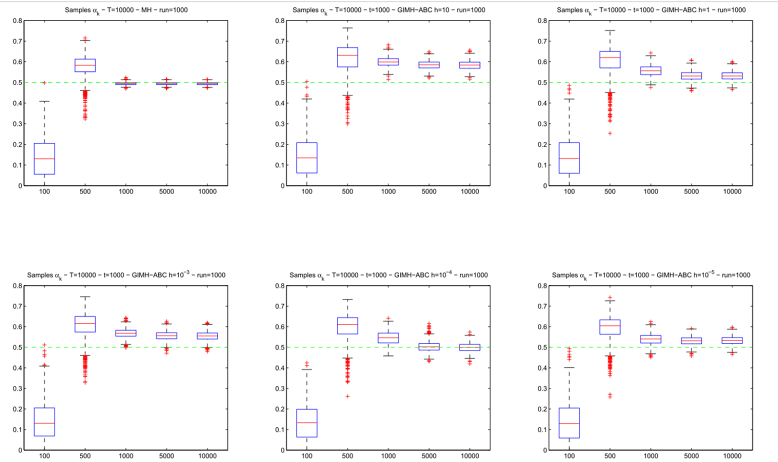


Figure: M–H top left & Approximated Bayesian Inference bottom (five different ϵ)

Influence of ϵ on γ estimate

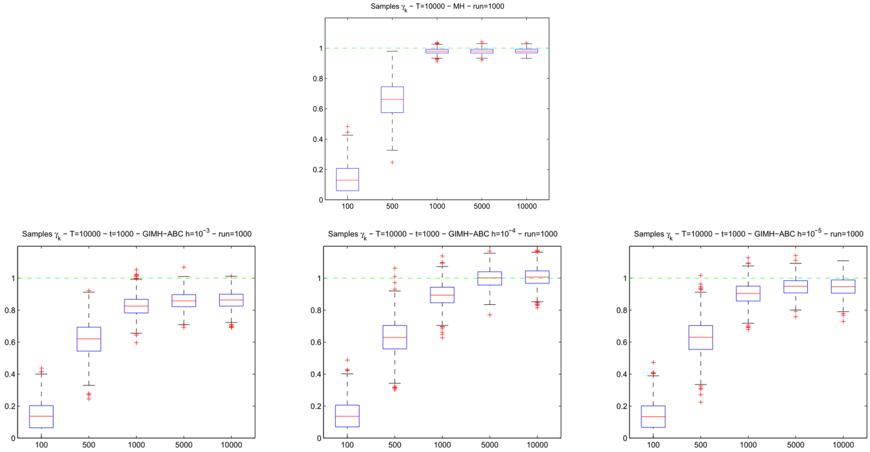


Figure: M–H top row Approximated Bayesian Inference bottom (three different ϵ)

Optimal behaviour

- Same kind of trend of β estimate
- For a fixed choice of summary statistics, there seems to exist an optimal ϵ

It is not that surprising, indeed

- $\epsilon \gg 1 \Rightarrow$ the choice of subset is not discriminant enough
- $\epsilon \ll 1 \Rightarrow$ in contrary we have

$$\tilde{\pi}_{n,\epsilon}(\theta \mid Y_{1:N}) \to \tilde{\pi}(\theta \mid U^{\star})$$

- so a proper mixture lies in-between...

Guidelines: if we trust S_n, then ϵ can be arbitrary low

A last example in high dimension

Reconstruction of template images from a handwritten digits data set. The parameter α we estimate has dimension $d = 256$, we have $N = 10.000$ observations each of size 15×15. Here $n = 100$ and S_n the mixture index.

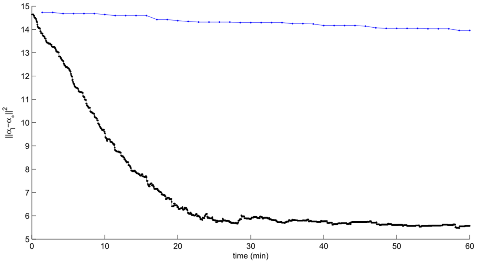


Figure: Distance from true templates: blue M–H and black Approximates Bayesian Inference

Perspectives

- Our approach targets a proxy of the true posterior which, provided a decent choice of summary statistics, achieves satisfactorily Bayesian inference at a fixed computational time
- Bardenet et al. & Korratikara don't know precisely the distribution they target...
- Theoretical analysis of the algorithm is difficult since

$$\tilde{\pi}_{n,\epsilon}(\theta \mid Y_{1:N}) = \sum_{U \in \mathcal{U}_n} \tilde{\pi}(\theta \mid Y_U) \nu_{n,\epsilon}(U)$$

  is intractable...

Further...

- Compare with Bardenet et. al simulations
- Search for the Intertwined Markov chain kernel...