Bayesian inference using sub-posteriors.

A study of instable Markov chains using Metropolis-Hastings algorithm variations.

Lionel Riou-Durand

2nd year application internship report
2014-2015 - ENSAE

an internship performed at

Insight - Centre for Data Analytics
University College of Dublin - Ireland

Supervisor: Florian Maire
Referent teacher: Vincent Cottet
Internship dates: 8th of June - 15th of August, 2015.

Contents

Introduction

The internship I performed during this summer was a research internship in Statistics at the University College of Dublin. I was supervised by post-doctoral researcher Florian Maire, who first of all gave me a very interesting research topic which I present its original version in appendix, and provided me in the same time various bibliographical references which allowed me to understand slowly at first but surely the research problem he wanted to study. This was a precise problem but also a wide open and non-restricted one, and what is more, the freedom that was granted to me allowed me to find my pace and my way through the successive working questions I was faced. I particularly enjoyed this research internship for these reasons, it was also the first I did in theoretical Statistics.

The topic as I understood it is the following: given a parametrized statistical model, we are very often forced to use Bayesian inference to provide some statistical estimates of the parameters as soon as the model we are studying is too complicated. Indeed, if we cannot compute easily neither the usual frequentist estimators (such as the Maximum Likelihood for example), nor its variance or other dispersion information which would allow us to provide some confidence intervals, then a solution widely applied now is to perform some Bayesian inference using Monte Carlo Markov Chains algorithms.

The most famous one is the Metropolis-Hastings algorithm: given a prior distribution of the parameters, and a model of likelihood, also assumed to be the distribution of the data given the parameters, the Metropolis-Hastings algorithm is built as a Markov chain constructed in order to get the posterior distribution of the parameter as its stationary distribution. Using properties of Markov chains under some regularity assumptions, the constructed chain will therefore converge towards the targeted posterior distribution. And we will thus be able to provide some “Bayesian credible intervals”, which if the modelling and the prior distribution are not mistaking too much, can also be interpreted as we usually do with frequentist confidence intervals.

This algorithm is very used because it can be applied to a wide range of problems and it is also very easy to implement since we only need to know how to compute the prior distribution and the likelihood for arbitrary given parameters and data, and we do not require to compute normalizing constants which are often intractable in complicated models. However, these algorithms require the use of the whole dataset at each iteration, which depending on the computational storage capacity is not always doable. Indeed, as soon as the dataset is too big, either because the amount of data is huge, or because each unit of data is very heavy, the Metropolis-Hastings algorithm will only be doable on a relatively small subset, so that the storage capacity can handle it.

The problem we here want to study is: in a MCMC context, how can we perform Bayesian inference with a limited storage capacity? More precisely, we here assume we can get an infinite stream of data, with limited size, giving us new information, but the key problem is we cannot store this data and once we have changed our data subset the previous one is gone forever. Since we have an infinite stream of data, the use of only one data subset is very disappointing. So the research questions are the following: under these assumptions, how to perform Bayesian inference using MCMC algorithms? How and when can we do better than the classical Metropolis-Hastings algorithm performed on one unique subset? The precise aim of my work was therefore to study the stability and the asymptotic properties of the Markov chains defined by variations of the Metropolis-Hastings Algorithm involving more than one unique subset of data.

The nature of the work I did was thus theoretical, in the sense that there was no real dataset involved. The two first weeks were fully dedicated to the understanding of the problem and particularly to get used to notations. The learning of theoretical Markov chains properties took also a reasonable amount of time especially since the assumptions on finite or countable spaces which I studied during the year are a bit different on uncountable spaces.

All my work was not only abstract though, the following weeks I also performed some simulations on simple examples in order to see what happens when we use the classical M-H algorithm but involving more than a unique subset of data, ie changing the data while the algorithm still runs. The M-H algorithm was not designed to be used this way, so the purpose of these simulations was to compare the ”empirical” results of the stable Markov chain using one unique subset to other instable Markov chains, but using a lot more data. The problem was then foreseen as a kind of proper equilibrium to find between the loss of stability of the chain (directly linked with the data refresh speed), and the gain of information provided by the new data, for a given computational time.

In the light of these first results, showing on these examples at least, that the loss of stability was hugely compensated by the gain of information, even for high speed data refresh process, the decision was then made to study more rigorously the properties of such “instable” Markov chains, especially those in which the subset of data was replaced at each iteration. The simulations seemed to converge towards a kind of “average” posterior distribution, actually not so “disturbed” as we forethought, the challenges of the remaining weeks of the internship were thus focused on this new target distribution.

This “average” posterior distribution was indeed very interesting because of its properties. In particular, we showed that it would always provide better estimators than the classical M-H algorithm performed on one unique subset, and was particularly efficient for problems involving large enough subsets. The aim was then to find a way to sample from this new target distribution. So we tried either to prove that the stationary distribution of these Markov chains was close to this “average” posterior distribution, or to modify slightly the algorithm in order to get a stationary distribution perfectly coinciding with the “average” posterior wanted.

Very sadly, I cannot say that this research work leaded to astonishing discoveries! But still, this report presents some results and partial findings. The applicability of the algorithms founded depends on the assumptions made (tractable constants, large enough subsets, regularity assumptions). But above all, this experience was very rewarding for me, especially since it was the first time that my work was fully dedicated to mathematical research, with a supervisor almost always available to advise me when I needed it, and with the freedom to do my research through the way I wanted.

I chose to organize my report in three parts. The first chapter is an opening Gaussian example in one dimension, which despite being a single example helped me a lot to understand the problem and it brought me as well a good guidance. The second chapter presents the methodological approach we chose, in the light of these first results, to study variations of the Metropolis-Hastings algorithm. In this chapter I focus as well on the new target we chose: the “average” posterior distribution. The third part focuses then on the final results, the algorithms founded, solving more or less the problem depending on the assumptions made. This final part is constructed with partial findings, but also with ideas and calculations which did not lead to any results. This last chapter show as well which are the remaining open problems.

Chapter 1

An opening example

We focus here on a very simple example in one dimension: we want to perform inference on the mean parameter of a Gaussian distribution. We assume here and in the entire report that we can get independently and identically distributed random variables. Of course this simple example does not require MCMC algorithms and taking the empirical mean as an estimator, the problem involved by the limited storage capacity is not relevant since we could easily store the empirical mean for each subset of data and compute then the mean of the means to get an almost perfect estimate of the mean parameter. But we here do as if this was not doable. We try here to perform Bayesian inference of this mean parameter using the Metropolis-Hastings algorithm, just as we would normally do with some more complicated model. Why sticking to an unrealistic example then? Because its simplicity will allow us to analyze more easily the results from the simulations, but also to go deeper into the theoretical understanding.

1.1 General setting

- **True distribution :** $(Y_i)_i \;\; iid \;\; \sim \;\; \mathcal{N}(0,1) = \mathbb{P}_0$ assumed unknown.
- **Likelihood model :** $(Y_i|\mu)_{i=1,\dots,N} \;\; iid \;\; \sim \;\; \mathcal{N}(\mu,1)$

  Thus we know Y_i is normally distributed with variance equal to one but we don't know its mean.

  Now given a prior distribution of the parameter, we will be able to sample from the posterior distribution using the Metropolis-Hastings algorithm. What is more we want to study the stability properties of the Markov chains regardless of prior mistakes, we thus assume here we don't have any information on this parameter.

- **Improper prior :** $p(\mu) = 1$

  I chose to use Jeffrey's prior which is also an improper one, because it is considered by many as an uninformative one (at least for simple examples), but we could also take a very flat distribution, for example : $p(\mu) = \mathcal{N}(1, 100)$, the variations in the results would be negligible. After some calculations skipped in purpose, we can then get the posterior distribution:

- **Posterior distribution :** $\pi(\mu|Y_{1:N}) = \mathcal{N}(\bar{Y}_N, \frac{1}{N})$

With the previous flat prior taken as an example, we can see that the posterior distribution would be quite the same :

$$\pi(\mu|Y_{1:N}) = \mathcal{N}\left(\frac{(\sum_{i=1}^{N} Y_i) + 0.01}{N + 0.01}, \frac{1}{N + 0.01}\right)$$

In general cases we will not know as well the posterior distribution and will only be able to compute the prior and the likelihood for any given parameter. Once again, though we could here simulate directly this posterior Gaussian distribution, we will choose later to do as if we could not and we will use the Metropolis-Hastings algorithm. We here focus on estimating $\mathbb{P}_0(A)$ for an arbitrary A for any parameter θ of interest, by using the so called predictive posterior distribution. Given a new observation $Y_0 \sim \mathbb{P}_0$ independent from the data subset $Y_{1:N}$, we can define more generally the predictive posterior distribution noted $\mathbb{P}_*$, with the following density :

$$f_*(dy_0|y_1, ..., y_N) := \int_{\Theta} \mathcal{L}(dy_0|\theta)\pi(\theta|y_1, ..., y_N)d\theta$$

This predictive posterior distribution is a conditional one given the data. Obviously the more the modelling or the prior will be wrong, the more the estimates will be poor. And the smaller the data subset will be, the less accurate this estimator $\mathbb{P}_*$ will be.

Now for a given A, we have :

$$\begin{aligned}
\mathbb{P}_*(A) &= \int_A f_*(y_0|y_1, ..., y_N)dy_0 \\
&= \int_{\Theta} \left\{ \int_A \mathcal{L}(y_0|\theta)dy_0 \right\} \pi(\theta|y_1, ..., y_N)d\theta \\
&= \mathbb{E}\left[\left\{ \int_A \mathcal{L}(y_0|\theta)dy_0 \right\} \middle| y_1, ..., y_N \right]
\end{aligned}$$

So $\mathbb{P}_*$ can also be seen as a conditional expectation given the data subset in which the random variable θ is distributed with its posterior distribution: $\pi(\theta|y_1, ..., y_N)$. And thus assuming here to simplify that we are able to integrate the likelihood, we can then provide an estimate of $\mathbb{P}_*(A)$ for any A by sampling at first from the posterior distribution using the M-H algorithm and then by computing for each parameter sampled the integral in brackets. Considering the empirical mean of these integrals as an estimator of $\mathbb{P}_*(A)$, for a large number of M-H iterations, we will thus provide an almost perfect estimate of $\mathbb{P}_*(A)$.

Back to our example, let's assume now we have a new observation available, modeled as the others, ie $(Y_0|\mu) \sim \mathcal{N}(\mu, 1)$, and let's consider the posterior predictive distribution. Here we consider for instance $A =]-\infty\, ; \Phi^{-1}(0.05)]$, where Φ is the cumulative function of normal standard distribution. In this example we are indeed able to integrate the likelihood :

$$\begin{aligned}
\int_A \mathcal{L}(y_0|\mu)\, \mathrm{d}y_0 &= \mathbb{P}(Y_0 \leq \Phi^{-1}(0.05) \,|\, \mu) \\
&= \mathbb{P}(Y_0 - \mu \leq \Phi^{-1}(0.05) - \mu \,|\, \mu) \\
&= \Phi(\Phi^{-1}(0.05) - \mu)
\end{aligned}$$

Thus we have the following :

$$\mathbb{P}_*(A) = \mathbb{E}\Big[\Phi(\Phi^{-1}(0.05) - \mu)\Big| Y_{1:N}\Big]$$

What is more here, since the model is well specified, the predictive posterior distribution will converge towards the true distribution when the size of the data subset tends to infinity:

$$\mathbb{P}_*(A) \underset{N\to+\infty}{\longrightarrow} \mathbb{P}_0(A) = 0.05$$

However as soon as $N < \infty$, $\mathbb{P}_*(A)$ is now a random variable depending on the observations $Y_{1:N}$ which is also a biased estimator, in the sense that its expectation is not equal to $\mathbb{P}_0(A)$. Indeed, we have the following :

$$\begin{aligned}\mathbb{E}[\mathbb{P}_*(A)] &= \mathbb{E}\Big[\mathbb{E}\Big[\Phi(\Phi^{-1}(0.05) - \mu)\Big| Y_{1:N}\Big]\Big] \\ &= \mathbb{E}\Big[\Phi(\Phi^{-1}(0.05) - \mu)\Big]\end{aligned}$$

The random variable μ in this final expectation is now distributed with the "true distribution" of μ, which can also be seen as a kind of "average" posterior distribution since it is defined as follows:

$$\widetilde{\pi}_N(\mu) := \int_{\mathbb{R}^N} \pi(\mu|Y_{1:N})\mathbb{P}_0(Y_{1:N})dY_{1:N}$$

Considering here the posterior distribution and the true distribution of $Y_{1:N}$:

$$\pi(\mu|Y_{1:N}) = \mathcal{N}(\bar{Y}_N, \frac{1}{N}) \text{ and } \mathbb{P}_0(Y_{1:N}) = \mathcal{N}(0_{\mathbb{R}^N}, I_N)$$

And using properties of Gaussian vectors, we can prove the following :

$$\widetilde{\pi}_N(\mu) = \mathcal{N}(0, \frac{2}{N})$$

Since $\Phi(x)$ is convex when $x < 0$, and $\mathbb{P}(\Phi^{-1}(0.05) - \mu < 0) \approx 1$ for a large enough N. Thus using Jensen's inequality we can directly show why $\mathbb{P}_*(A)$ is actually biased in this example :

$$\mathbb{E}[\Phi(\Phi^{-1}(0.05) - \mu)] > \Phi(\Phi^{-1}(0.05) - \mathbb{E}[\mu]) = 0.05$$

$$\implies \boxed{\mathbb{E}[\mathbb{P}_*(A)] > \mathbb{P}_0(A)}$$

1.2 Batches of data implied by the limited storage capacity

Let's now consider we have a fixed storage capacity N. We assume though that we can have an infinite stream of data, the total length of data available is not taken as a constraint, which means we can consider B independent batches of data on the same design, with B very large, ie :

True distribution : $(Y^b_{i:N})^{b=1,\dots,B} \;\; iid \;\; \sim \;\; \mathcal{N}(0_{\mathbb{R}^N}, I_N)$

Likelihood model : $(Y^b_{i:N}|\mu)^{b=1,...,B} \ \ iid \ \ \sim \ \ \mathcal{N}(\mu_{\mathbb{R}^N}, I_N)$

If we perform the same Bayesian inference on each batch, using the improper prior as before and getting thus B different posterior distributions, we will have actually also B different posterior predictive distributions :

$$\mathbb{P}_{*,b}(A) := \mathbb{E}[\Phi(\Phi^{-1}(0.05) - \mu)|Y^b_{1:N}] \quad \text{for all } b = 1, ..., B$$

Thus if L is the total number of Metropolis-Hastings iterations and if we consider on each batch b the following estimator :

$$\hat{\mathbb{P}}^{(L)}_{*,b}(A) := \frac{1}{L}\sum_{l=1}^{L} \Phi(\Phi^{-1}(0.05) - \mu_{l,b})$$

where $(\mu_{l,b})_{l=1,...,L}$ are sampled from the posterior distribution of the batch b: $\pi(\mu|Y^b_{1:N})$, using the Metropolis-Hastings algorithm, we will have on each batch the following ergodic result :

$$\hat{\mathbb{P}}^{(L)}_{*,b}(A) \underset{L\to+\infty}{\longrightarrow} \mathbb{P}_{*,b}(A)$$

Now a thing to note is that even if we had unlimited computational time and if we could perform these simulations on a very large number of different data batches, using the empirical mean of all the different predictive posteriors would create a bias anyway, due to the limited size of each batch:

$$\begin{aligned}\frac{1}{B}\sum_{l=1}^{B} \hat{\mathbb{P}}^{(L)}_{*,b}(A) &\xrightarrow[L\to+\infty]{} \frac{1}{B}\sum_{l=1}^{B} \mathbb{P}_{*,b}(A) \\ &\xrightarrow[\substack{L\to+\infty \\ B\to+\infty}]{} \mathbb{E}[\mathbb{P}_*(A)] > \mathbb{P}_0(A)\end{aligned}$$

1.3 Limited computational time

Of course, if we allow in the same time L and B to be infinitely large, the "average" predictive posterior will be more accurate, considering the L^2 distance, simply because its variance will tends to 0. Indeed, breaking down the L^2 distance as the variance plus the square of the bias, we have the following :

$$\begin{aligned}\mathbb{E}\Big[(\mathbb{P}_*(A) - \mathbb{P}_0(A))^2\Big] &= \mathbb{V}\Big(\mathbb{P}_*(A)\Big) + \Big(\mathbb{E}[\mathbb{P}_*(A)] - \mathbb{P}_0(A)\Big)^2 \\ \text{with : } &\mathbb{V}\Big(\mathbb{P}_*(A)\Big) > 0 \ \text{ as soon as : } \ N < +\infty\end{aligned}$$

But a more realistic situation would be to consider of course a limited computational time. We have indeed to compare several scenarios on the same standard. We will never be able to sample a very large number of M-H iterations on each subset of data for a very large number of subset.

We could for example consider taking $B \times L$ as a constant, and compare several scenarios for different values of B for example. This way we would perform the M-H algorithm on B different subset and keep only L iterations of each simulation. As soon as we consider doing that, we have to take into account the fact that each M-H simulation will need a burn-in period to be removed to allow each distribution to converge towards its posterior. The relative computational time of these burn-in periods will not be negligible especially for scenarios in which we would consider a very large number of subsets B and a small number of M-H iterations L.

Another way of doing it then is to forget completely about burn-in periods, ie to consider a fixed number of Metropolis-Hastings iterations, but to change the data, and therefore the transition kernel, during the algorithm. Several scenarios will therefore be compared for different data refresh speeds.

The general MCMC setting will be changed and the Metropolis-Hastings algorithm will have a moving target distribution. We expect the Markov chain will not be stable anymore, its instability depending probably on the storage capacity, but also on the speed of the data-refresh process. What's more theoretical properties of such "disturbed" chains are more difficult to understand and formalize, and are therefore less well known. But allowing the algorithm to view a lot more of data may produce more accurate estimators, even if the chain is not stable.

We consider for this example that we can compute L=10 000 iterations of Metropolis-Hastings algorithm, and no more. But we assume for simplicity that we can easily switch the data subset between two iterations without any additional computational cost, although it might seem to be a strong assumption. Let's compare thus several scenarios: for a given storage capacity N, we let the number of data batches B used in total during the whole simulation vary from 1 to 10 000 by powers of ten.

Thus B is also a measure of the data refresh speed. At one extreme, if B=1 the whole algorithm is performed on a single batch of data, which means the Markov chain is stable, the transition kernel remains the same all along the process. At the other extreme, if B=10 000, each Metropolis-Hastings iteration is performed on a completely new batch, therefore on the one hand the amount of data used will be huge $(L \times N)$, on the other hand the transition kernel is constantly moving and since the M-H algorithm has not been designed in that purpose, we expect the algorithm will have no time to exploit all the information provided by each batch. Between these two extremes, we simply let B vary by powers of ten, so that the same number of iterations is performed on each batch of data.

This way if there is indeed a competing effect between stability and gain of information, we will be able to see it and assess it using the predictive posterior as before, replacing the posterior distribution by the distribution empirically simulated, by computing L^2 distances. Since the standard predictive posterior is biased, the comparison between the scenarios may not be easy. Thus the mean square errors will not only be computed between the estimates and the true distribution $\mathbb{P}_0(A)$ but also between the estimates and $\mathbb{E}[\mathbb{P}_*(A)]$.

Since computing the empirical L^2 distances requires the simulations to be repeated many times, I have only performed this analysis for 3 different storage capacity : N=10, N=100 and N=1 000, for each of them and for a given number of data subsets B we can get one unique estimate $\widehat{\mathbb{P}}(A)$ by performing the L=10 000 M-H iterations replacing the data as explained, by computing from the L=10 000 resulting parameters $(\mu_l)_{l=1,...,L}$, the following formula:

$$\widehat{\mathbb{P}}(A) := \frac{1}{L}\sum_{l=1}^{L}\Phi(\Phi^{-1}(0.05)-\mu_l)$$

Now for each scenario this estimation has to be repeated a sufficient number of times in order to get decent L^2 distances estimators. Due to computational time constraints, I was only able to repeat this estimation 200 times.

1.4 Graphical results

I present in this report the following results of these simulations for a storage capacity N=100. Each histogram presents thus an empirical distribution of $\widehat{\mathbb{P}}(A)$ based on 200 simulations. Two vertical lines have been added, the red one corresponding to $\mathbb{P}_0(A) = 0.05$, the green one to $\mathbb{E}[\mathbb{P}_*(A)]$ (≈ 0.05171 in this example). Beware, the scale of the x-axis differs from one histogram to another.

Figure 1.1: B=1 data batch used

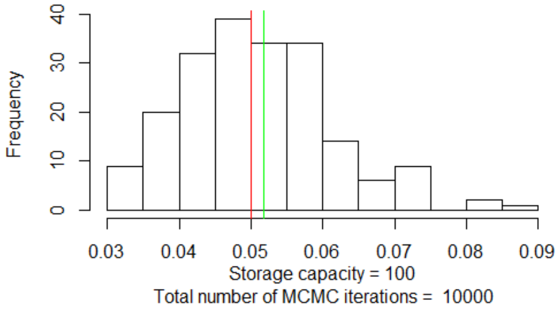


Figure 1.2: B=10 data batches used

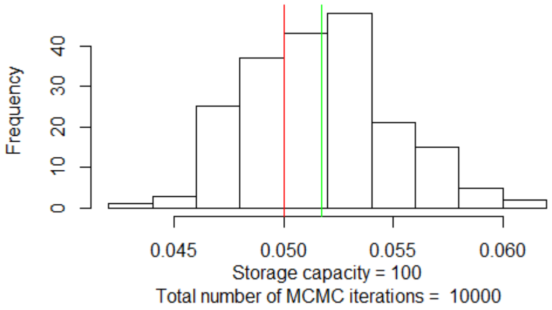


Figure 1.3: B=100 data batches used

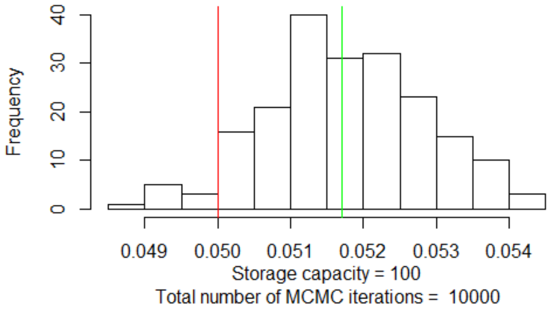


Figure 1.4: B=1000 data batch used

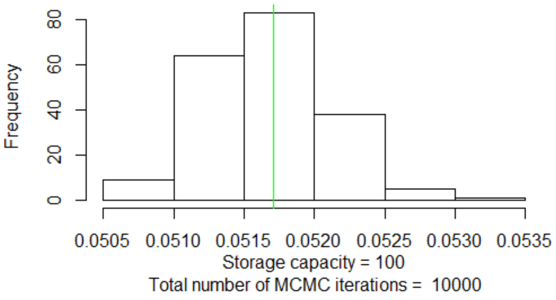
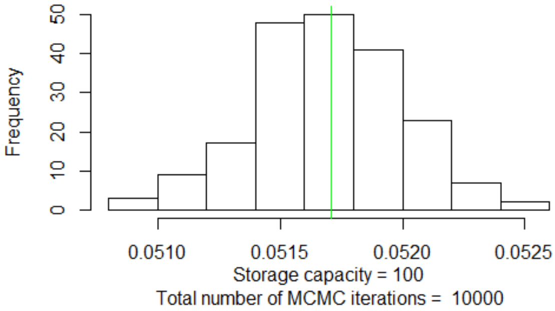


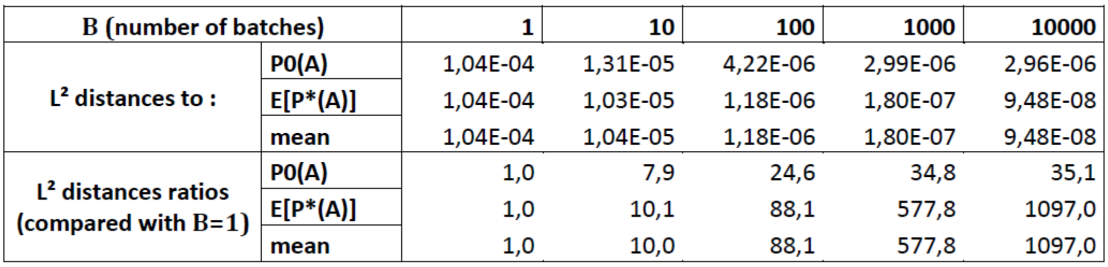
Figure 1.5: B=10000 data batches used



We can see that the variance drops down very fast as the number of data batches used increases. The bias induced by the limited storage capacity is here graphically highlighted, since the simulations are indeed converging towards $\mathbb{E}[\mathbb{P}_*(A)] \neq \mathbb{P}_0(A)$. Note that each empirical distribution is still centered on $\mathbb{E}[\mathbb{P}_*(A)]$ even for high speed of data refresh like B=10 000, which appeared to us as a surprisingly good result. Indeed we had no guaranty that such results would continue to hold for an "instable" Markov chain with a constantly moving kernel with no burn-in periods.

The empirical distances are presented in the first part of the following table. These are 3 different L^2 distances, the first one measured to the true probability $\mathbb{P}_0(A)$, the second one to the expectation of the predictive posterior $\mathbb{E}[\mathbb{P}_*(A)]$, and the final one to the empirical mean. The second part presents the ratios between these L^2 distances compared with the situation where one unique data subset is used. These ratios are also measuring the gain of variance due to the increase in the amount of data used during the simulation.

Figure 1.6: Summary of L^2 distances



B (number of batches)		1	10	100	1000	10000
L² distances to :	P0(A)	1,04E-04	1,31E-05	4,22E-06	2,99E-06	2,96E-06
	E[P*(A)]	1,04E-04	1,03E-05	1,18E-06	1,80E-07	9,48E-08
	mean	1,04E-04	1,04E-05	1,18E-06	1,80E-07	9,48E-08
L² distances ratios (compared with B=1)	P0(A)	1,0	7,9	24,6	34,8	35,1
	E[P*(A)]	1,0	10,1	88,1	577,8	1097,0
	mean	1,0	10,0	88,1	577,8	1097,0

As we noted when looking at the graphs, the simulations are centered on $\mathbb{E}[\mathbb{P}_*(A)]$, and so the L^2 distances to this expectation and the empirical variance are almost the same. All L^2 distances are dropping down when the number of data increases, even for high speed in the data refresh process. What is more, the gain of variance remains very high even for a very "instable" chain like B=10 000.

Indeed, if the simulations had been done on a i.i.d. context, without MCMC ie directly by simulating from the posterior which is doable in this simple example, then the variance would have decreased in a linear way: a ten times larger sample of data would have led to a ten times more accurate estimator.

This results seem indeed to hold for slow data refresh like B=10, just as if we had performed 10 different M-H algorithms each of them well separated by a sufficiently long burn-in period. But here these ratios are not linear, at some point, they decrease when the speed of the refresh is high, due to the increasing instability of the Markov chain. But from our point of view these ratios were still remaining very high considering that for 10 000 MCMC iterations in total, the use of one single iteration on each new data subset was still enough to reduce the variance by 1 000.

These surprisingly good results lead us to study more theoretically these "instable" Markov chains in the next chapter, and to focus particularly on this average predictive posterior towards which the simulations seemed to converge even for high data refresh speed. The next chapter is dedicated to the methodological approach we choose to follow during the rest of my internship.

Chapter 2

Methodological approach

In this chapter we present the theoretical approach we chose to answer the problem. The limited storage capacity is a problem in a MCMC context because we are unable to get an exact posterior from so-called sub-posteriors, ie several posterior distributions given different data subsets. Indeed we are unable to combine sub-posteriors to get an exact posterior, neither analytically nor by sampling, in the sense that we cannot for instance from two different posterior distributions $\pi(\theta|Y^1_{1:N})$ and $\pi(\theta|Y^2_{1:N})$, get easily the posterior distribution given the two subsets: $\pi(\theta|Y^1_{1:N}, Y^2_{1:N})$. Several works try to provide some algorithms aiming to approximate as best as possible these "global" posteriors, I studied some related research papers such as the following: Alexey Miroshnikov, Erin M. Conlon. Parallel Markov Chain Monte Carlo for Non-Gaussian Posterior Distributions. arXiv preprint arXiv:1506.03162, 2015. But here we are not trying to combine sub-posterior looking for a good "global" posterior, we have a different approach.

2.1 Another target distribution

Indeed sampling easily from the posterior distribution given more than one storage capacity of data is not doable. But still, if we had unlimited computational time, one thing we could do would be to get a random subset of data b, then to sample one parameter from the posterior distribution of the batch b: $\pi(\theta|Y^b_{1:N})$ using the Metropolis-Hastings algorithm. This way the subset of data and the parameter sampled would be distributed from the following joint distribution: $\mathbb{P}_0(Y_{1:N})\pi(\theta|Y_{1:N})$. Repeating this sampling an infinite number of times then putting the different sampling together we could get this way samples independently and identically distributed from this previous joint distribution. Then, considering only the parameters we will get samples from the marginal distribution :

$$\widetilde{\pi}_N(\theta) := \int_{Y^N} \mathbb{P}_0(Y_{1:N})\pi(\theta|Y_{1:N})dY_{1:N}$$

This is an unrealistic assumption for sure, because each sample would require a different M-H algorithm with a sufficiently long burn-in period. But still, this kind of "average posterior distribution" is very interesting because when using it in the predictive posterior, estimator of the true distribution of the data $\mathbb{P}_0$ originally defined as :

$$\widehat{\mathbb{P}}_0(dy_0) := \int_{\Theta} \mathcal{L}(dy_0|\theta)\pi(\theta|Y_{1:N})d\theta$$

Replacing $\pi(\theta|Y_{1:N})$ by $\widetilde{\pi}_N(\theta)$ we could then get another "estimator" :

$$\widetilde{\mathbb{P}}_0(dy_0) := \int_\Theta \mathcal{L}(dy_0|\theta)\bigg\{\int_{Y^N} \mathbb{P}_0(Y_{1:N})\pi(\theta|Y_{1:N})dY_{1:N}\bigg\}d\theta$$

We can note that this "estimator" is no more a random variable, it is a constant indeed. What is more, by switching the two integrals we can see that this "estimator" is actually just the expectation of the predictive posterior :

$$\begin{aligned}\widetilde{\mathbb{P}}_0(dy_0) &= \int_{Y^N} \mathbb{P}_0(Y_{1:N})\bigg\{\int_\Theta \mathcal{L}(dy_0|\theta)\pi(\theta|Y_{1:N})d\theta\bigg\}dY_{1:N} \\ &= \int_{Y^N} \mathbb{P}_0(Y_{1:N})\Big\{\widehat{\mathbb{P}}_0(dy_0)\Big\}dY_{1:N} \\ &= \mathbb{E}_{\mathbb{P}_0}\Big[\widehat{\mathbb{P}}_0(dy_0)\Big]\end{aligned}$$

Unfortunately, as soon as $N < +\infty$, as we have shown it on a simple example in the first chapter, there is no reason for $\widehat{\mathbb{P}}_0(dy_0)$ to be unbiased. So $\widetilde{\mathbb{P}}_0(dy_0) \neq \mathbb{P}_0(dy_0)$ in general. But since we focus in our work on comparing new estimators from "instable" Markov chains to the stable one from the classical M-H algorithm, it becomes interesting to study this constant "estimator" and to compare it to the predictive posterior given one unique subset of data.

2.2 Some interesting properties

Once more, as we had foreseen it previously on a simple Gaussian example, if we could sample from $\widetilde{\pi}_N(\theta)$, we can show that considering the constant $\widetilde{\mathbb{P}}_0(dy_0)$ as an estimator of $\mathbb{P}_0(dy_0)$ will always be better than $\widehat{\mathbb{P}}_0(dy_0)$. Indeed, this is true in the sense of the L^2 distance, since we can decompose the mean square error of $\widehat{\mathbb{P}}_0(dy_0)$ as its variance plus the square of its bias:

$$\mathbb{E}_{\mathbb{P}_0}\Big[\Big(\widehat{\mathbb{P}}_0(dy_0) - \mathbb{P}_0(dy_0)\Big)^2\Big] = \mathbb{V}_{\mathbb{P}_0}\Big(\widehat{\mathbb{P}}_0(dy_0)\Big) + \Big(\widetilde{\mathbb{P}}_0(dy_0) - \mathbb{P}_0(dy_0)\Big)^2$$

What is more, if we assume the parametrized model to be well specified, and if there is indeed a true parameter, ie if there is $\theta_0 \in \Theta$ so that $\mathbb{P}_{\theta_0} = \mathbb{P}_0$, then when the size of the subsets, N is large enough, and when regularity assumptions holds, we can have then some classical asymptotic results. Noting $\widehat{\theta}_{ML}$ as the Maximum Likelihood estimator, $I_N(\theta_0)$ as the Fisher's Information Matrix, and $\|.\|_{TV}$ the total variation distance between two distributions, we have thus the standard following results:

- Bernstein Von Mises Theorem :

$$\mathbb{E}_{\mathbb{P}_0}\Big[\|\pi(\theta|Y_{1:N}) - \mathcal{N}\Big(\widehat{\theta}_{ML}, I_N^{-1}(\theta_0)\Big)(\theta)\|_{TV}\Big] \xrightarrow[N\to+\infty]{} 0$$

- Asymptotic Normality of the Maximum Likelihood :

$$I_N^{1/2}(\theta_0)(\widehat{\theta}_{ML} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Before this internship, among these results only the Asymptotic Normality of the Maximum Likelihood was well known for me. This is why I focus here a while on the Bernstein Von Mises Theorem. Considering the definition of the total variation distance, between any given two probability measures ν_1 and ν_2:

$$\|\nu_1(.) - \nu_2(.)\|_{TV} := \underset{A}{sup}|\nu_1(A) - \nu_2(A)|$$

Then we see that the Berstein Von Mises Theorem implies that the distance between the posterior distribution ; and a Gaussian distribution centered on the Maximum Likelihood estimator tends to zero when N tends to infinity. Thus for any subset of data of size N, as soon as N is large enough, we can approximate the posterior distribution directly by this Gaussian distribution. This distribution is a conditional one given the Maximum Likelihood estimator, and it does not depend on the prior distribution chosen. Thus, we can use it to replace the posterior, which was a conditional distribution given the whole subset. Moreover, note that we are not forced to be able to compute analytically neither the Maximum Likelihood nor the Fisher's Information so that this asymptotic result holds.

This is particularly useful here, because using the Bernstein Von Mises Theorem to approximate firstly $\pi(\theta|Y_{1:N})$ by $\pi(\theta|\widehat{\theta}_{ML}) := \mathcal{N}\Big(\widehat{\theta}_{ML}, I_N^{-1}(\theta_0)\Big)(\theta)$, and then again approximating the distribution of the Maximum Likelihood estimator by its Gaussian limit, we can then approximate $\widetilde{\pi}_N(\theta)$ as well:

$$\begin{aligned}\widetilde{\pi}_N(\theta) &= \int_{Y^N} \mathbb{P}_0(Y_{1:N})\pi(\theta|Y_{1:N})dY_{1:N} \\ &\approx \int_{\Theta} \mathcal{N}\Big(\theta_0, I_N^{-1}(\theta_0)\Big)(\widehat{\theta}_{ML}) \times \mathcal{N}\Big(\widehat{\theta}_{ML}, I_N^{-1}(\theta_0)\Big)(\theta)d\widehat{\theta}_{ML}\end{aligned}$$

Now using Gaussian vectors properties, we can show that this new marginal distribution is necessarily a Gaussian one, the mean and the variance parameters can then easily be founded using classical decomposition via conditional expectations:

$$\begin{aligned}\mathbb{E}\big[\theta\big] &= \mathbb{E}\big[\mathbb{E}[\theta|\widehat{\theta}_{ML}]\big] = \mathbb{E}\big[\widehat{\theta}_{ML}\big] = \theta_0 \\ \mathbb{V}\big(\theta\big) &= \mathbb{E}\big[\mathbb{V}(\theta|\widehat{\theta}_{ML})\big] + \mathbb{V}\big(\mathbb{E}[\theta|\widehat{\theta}_{ML}]\big) = \mathbb{E}\big[I_N^{-1}(\theta_0)\big] + \mathbb{V}\big(\widehat{\theta}_{ML}\big) = 2I_N^{-1}(\theta_0) \\ &\Longrightarrow \boxed{\widetilde{\pi}_N(\theta) \approx \mathcal{N}\Big(\theta_0, 2I_N^{-1}(\theta_0)\Big)(\theta)}\end{aligned}$$

And since this final distribution is centered on the true parameter θ_0, we could thus estimate it almost perfectly using sampling from $\widetilde{\pi}_N(\theta)$, simply by computing the empirical mean on a large size sample. What is more, as we just noticed, this remains doable even if we cannot compute analytically either the Maximum Likelihood or the Fisher's Information, which is very often the case for complicated models.

2.3 Dealing with the limited computational time

So it is worth trying to aim this "average" posterior distribution as best as possible, taking now into account the limited computational time, though $\widetilde{\pi}_N(\theta)$ may not be a perfect estimator, especially for small limited storage capacity. Of course, performing a large number of simulations, each using the M-H algorithm on a different subset is completely unrealistic. The purpose we have is thus to show that generally in a M-H algorithm using at each iteration a new subset of data, the resulting distribution of the parameter is at least a "good" approximation of the average posterior distribution. Indeed, the surprisingly good results of the simulations on a simple Gaussian example makes us question if this approximation can be generalized. Another similar purpose would be to modify the algorithm in a way that would make the stationary distribution of the Markov chain becomes directly the targeted average posterior.

The path we chose to follow required therefore to go through the theoretical side of the Metropolis-Hastings algorithm. Once more, since I was not familiar with Markov chains on uncountable spaces before this internship, I focus a while on general results related with the M-H algorithm. During my internship, I took this opportunity to study these properties more deeply than during my school year, with reference to the following paper: Gareth O. Roberts, Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. Probability Surveys Vol. 1, pages 20–71, 2004.

The M-H algorithm is a very handy and famous algorithm for the following reasons. From a complicated statistic model, which does not allow us to compute the classical frequentist estimators such as the Maximum Likelihood for example, sampling i.i.d random variables from the posterior distribution is not doable either in general. The M-H algorithm was designed in order to avoid this problem. Indeed, without requiring many assumptions, the algorithm creates a specific Markov chain with the posterior distribution targeted as its stationary one. This way, using Markov chains properties under some regularity assumptions, the chain created during the run of the algorithm will converge towards the posterior distribution. All of this has a price: the random variables sampled will not be independent anymore. But there are some ways to overcome the problem: either using ergodic results or keeping only widely spaced observations in order to get very low correlation between the variables.

2.3.1 Markov chains on uncountable spaces

Unlike the Markov chains on finite or countable spaces, to find the stationary distribution of a Markov chain defined on an uncountable space is not that easy. Indeed we cannot find it anymore by solving an equation system because we have now to face transition kernels which are probability densities with respect to Lebesgue's measure. A transition kernel is thus no more a matrix filled with transition probabilities to go from one state to another state, but a family of conditional densities $\Big(\mathbb{K}(dy|x)\Big)_{x\in\mathcal{X}}$ where $\mathcal{X}$ is an uncountable set such as $\mathbb{R}$ for example.

The chain still goes from one state $x \in \mathcal{X}$ to another $y \in \mathcal{X}$ and so on, but probabilities can now only be measured on wide enough sets such as intervals. Stationary distributions still exist but require the use of integrals in their definition, a distribution π is thus defined as stationary if:

$$\int_{x\in\mathcal{X}} \pi(dx)\mathbb{K}(dy|x) = \pi(dy)$$

On a finite space, this is easily solved since we can write down all the equations induced by this definition and deduce the stationary distribution from an equation system. Here this is not doable anymore. Actually we cannot solve this equation, the thing we do is that from a candidate distribution we prove that it is indeed a stationary one, using in general the notion of reversibility since the integral is most of the time intractable. A Markov chain is defined as reversible with respect to π if:

$$\pi(dx)\mathbb{K}(dy|x) = \pi(dy)\mathbb{K}(dx|y)$$

If a Markov chain is indeed reversible with respect to π, then we can easily show that π is a stationary distribution of the chain. We do not focus too much here on regularity assumptions, such as ϕ-irreducibility and aperiodicity, which are closely linked with the ones on countable spaces although a bit differently formalized, much more complicated to understand, but also not really useful to the understanding of the research problem we want to solve. But under these regularity assumptions, the Markov chain will converge towards the unique stationary distribution. The main asset of the M-H algorithm is to build a regular Markov chain which has as its stationary distribution the posterior distribution wanted.

2.3.2 The Metropolis-Hastings design

Indeed, the Metropolis-Hastings algorithm is designed as follows. We have first to choose a conditional distribution easy to sample from, also called the proposal distribution: $q(dy|x)$ defined on $\mathcal{X} = \Theta$. Then from a initial parameter $\theta_1 \in \Theta$, at each iteration $i \geq 1$, a new parameter is sampled from the proposal distribution: $q(d\theta_{i+1}|\theta_i)$. This new parameter θ_{i+1} is then accepted with the following probability:

$$\alpha(\theta_i, \theta_{i+1}) := min\Big[1, \frac{\pi(\theta_{i+1}|Y_{1:N})q(\theta_i|\theta_{i+1})}{\pi(\theta_i|Y_{1:N})q(\theta_{i+1}|\theta_i)}\Big]$$

This algorithm can actually also be formalized as a Markov chain with the following transition kernel, the rejection part of the algorithm requiring a mixture with a Dirac's measure:

$$\mathbb{K}(d\theta_{i+1}|\theta_i) := q(\theta_{i+1}|\theta_i) \times min\Big[1, \frac{\pi(\theta_{i+1}|Y_{1:N})q(\theta_i|\theta_{i+1})}{\pi(\theta_i|Y_{1:N})q(\theta_{i+1}|\theta_i)}\Big] d\theta_{i+1} \; + \; c_{\theta_i}\delta_{\theta_i}(d\theta_{i+1})$$
$$\text{where : } c_{\theta_i} \; := \; 1 \; - \int_{y \in \mathcal{X}} min\Big[1, \frac{\pi(y|Y_{1:N})q(\theta_i|y)}{\pi(\theta_i|Y_{1:N})q(y|\theta_i)}\Big] dy$$

First of all this Markov chain is indeed reversible with respect to the posterior distribution $\pi(.|Y_{1:N})$: indeed thanks to the symmetry of the minimum function, we can show that $\pi(dx|Y_{1:N})\mathbb{K}(dy|x) = \pi(dy|Y_{1:N})\mathbb{K}(dx|y)$. This is the guaranty that the posterior distribution is indeed the stationary distribution of the Markov chain, and so that the run of the M-H algorithm will indeed converge towards the posterior distribution wanted. But what is more, the M-H algorithm only requires to be able to compute the prior distribution and the Likelihood for a given data set and parameter, avoiding in the same time some usual intractability issues. Indeed we have in the acceptance ratio the following:

$$\frac{\pi(\theta_{i+1}|Y_{1:N})}{\pi(\theta_i|Y_{1:N})} = \frac{\mathcal{L}(Y_{1:N}|\theta_{i+1})p(\theta_{i+1})}{\mathcal{L}(Y_{1:N}|\theta_i)p(\theta_i)} \times \frac{\mathcal{Z}(Y_{1:N})}{\mathcal{Z}(Y_{1:N})}$$

And in this ratio the following distribution: $\mathcal{Z}(Y_{1:N}) := \int_\Theta \mathcal{L}(Y_{1:N}|\theta)p(\theta)d\theta$, generally intractable, simplifies.

2.3.3 Modelling the switch of data to build an algorithm

Now, coming back to our research question, we want to study the formalized kernel of the classical Metropolis-Hastings algorithm in which data is changed at each iteration. Several ways of formalizing this kernel are possible, we present these in the next chapter. From this formalization we intend to build some algorithms which deal with these constantly moving target distribution, keeping as a purpose to sample in a limited computational time from the average posterior distribution, which has some interesting properties as we just shown.

The aim of this work is to show either that the improper use of the classical M-H algorithm still allows us to sample from a good approximation of the targeted average posterior, or to find new variations of the M-H algorithm adjusted in a way that would correctly take into account the constantly changing data, and if possible to sample exactly from the average posterior. Indeed since our empirical results are good with a very simple and improper use of this classical Metropolis-Hastings algorithm, we think our objective is reachable, and that there might be other ways to sample directly from the average posterior we are aiming for.

We present therefore in the next chapter partial findings of this work, depending on the assumptions made, about the regularity of the model, the size of the data subsets, and the tractability of some constants. All these results use some variations of the Metropolis-Hastings algorithm, which share the purpose of making the average posterior becoming the stationary distribution of the Markov chain defined by the algorithm, therefore to sample from $\widetilde{\pi}_N(\theta)$ in a reasonable computational time, and hence to provide better algorithms than sampling from one unique posterior.

Chapter 3

Results

This chapter is dedicated to the presentation of all the results, the findings, but also all the wrong paths that we followed, in order to conceive an algorithm able to sample from the average posterior distribution: $\widetilde{\pi}_N(\theta)$. We were particularly influenced by the following research paper: Gareth O. Roberts, Jeffrey S. Rosenthal. Coupling and Ergodicity of Adaptive MCMC. Journal of Applied Probability, Volume 44, Number 2, pages 458-475, 2007. ; although their research question was a bit different in the sense that they prove convergence results for changing transition kernels which all share the same stationary distribution, whereas in our case the kernels all have a different target distribution. Note that these are only theoretical results, of course the algorithms have been quickly tested with simulations, but because of time constraints, no real dataset was ever involved to study more deeply the robustness of these algorithms.

3.1 A formalized kernel

Of course performing a M-H algorithm on each subset for a large number of subsets is not doable, thus we are focusing here on the formalized kernel corresponding to a M-H algorithm in which data is completely changed at each iteration. To simplify let's note $\Gamma_i := Y^i_{1:N}$ as the ith subset of data. We can then write this use of the M-H algorithm as a transition kernel in two different ways:

- By a chain defined only on the space of the parameters Θ, in which the switch of data is completely took into account into the transition kernel:

$$\mathbb{K}\Big(\theta_{i+1}|\theta_i\Big) := \int_{Y^N} \mathbb{P}_0(\Gamma) q(\theta_{i+1}|\theta_i) min\Bigg(1, \frac{\pi(\theta_{i+1}|\Gamma)}{\pi(\theta_i|\Gamma)} \frac{q(\theta_i|\theta_{i+1})}{q(\theta_{i+1}|\theta_i)}\Bigg) d\Gamma \; + \; c_{\theta_i}\delta_{\theta_i}(\theta_{i+1})$$

  This way, a new data set is sampled at each iteration, corresponding to $\mathbb{P}_0(\Gamma)$. The new parameter is drawn with the classical acceptance ratio of the M-H algorithm given the new data subset Γ just sampled. But we are focusing here only on the parameter, thus we consider only the marginal distribution by integrating over Y^N.

- By an extended chain defined at the same time on the space of the parameters and on the space of the data, $\Theta \times Y^N$, where the switch of data is considered as a random walk included in a wider

Markov chain:

$$\mathbb{K}\Big((\Gamma_{i+1},\theta_{i+1})|(\Gamma_i,\theta_i)\Big) := \mathbb{P}_0(\Gamma_{i+1})q(\theta_{i+1}|\theta_i)min\Bigg(1, \frac{\pi(\theta_{i+1}|\Gamma_{i+1})}{\pi(\theta_i|\Gamma_{i+1})}\frac{q(\theta_i|\theta_{i+1})}{q(\theta_{i+1}|\theta_i)}\Bigg) + c_{(\Gamma_i,\theta_i)}\delta_{(\Gamma_i,\theta_i)}(\Gamma_{i+1},\theta_{i+1})$$

This kernel corresponds to the exact same algorithm in practice, but this is conceptually a bit different: here we consider the change of data as part of the chain, and we focus more generally at a joint distribution over $\Theta \times Y^N$. Firstly a new batch of data is drawn: $\mathbb{P}_0(\Gamma_{i+1})$. And then conditionally to this new dataset, a classical M-H iteration is performed to draw a new parameter. The reject part formalized by a Dirac's measure is now concerning at the same time the parameter and the data. Of course extracting from this extended chain, only the chain of the parameters would directly lead to the first case.

The first attempts to study this use of the M-H algorithm were made considering the first of the two kernels. We can also see this kernel as a moving kernel, each kernel targeting a new posterior distribution. We thus tried to find out if the average posterior $\widetilde{\pi}_N(\theta)$ was close to the stationary distribution, which was not an easy task because the kernel is not easy to handle. We already knew that the stationary distribution was probably a noisy version of the average posterior. But since the empirical results were very close to this distribution, we wanted to find out if the average posterior and the stationary distribution might coincide asymptotically. Using the Bernstein Von Mises theorem and the asymptotic normality of the Maximum Likelihood, the kernel was indeed easier to handle, it became possible to compute its expectation. But the symmetry we were looking for, to guaranty the reversibility of the chain, and thus the convergence, was never reached, neither with the average posterior even with its asymptotic version, nor by any other distribution we tried which includes all the Gaussian ones.

At the end of this internship we still have no theoretical proof of how far the stationary distribution of this Markov chain will be from the better understood target $\widetilde{\pi}_N(\theta)$. But still, this algorithm seems to perform well on simple Gaussian examples, in the sense that the stationary distribution observed empirically is close to $\widetilde{\pi}_N(\theta)$. Since the results are good with Gaussian examples, we can ask ourselves if it is possible to prove that this algorithm performs well asymptotically in general, when we consider for example large enough N and holding regularity assumptions so that Bernstein Von Mises theorem and asymptotic normality of the Maximum Likelihood applies, and this way to be approximately back to a Gaussian case. But no way was found to generalize it properly.

3.2 A new extended chain

After a lot of unsuccessful attempts, the study of the second kernel, which seemed redundant at first sight, actually led us to another perspective. Indeed, let's consider the extended Markov chain $(\Gamma_b, \theta_b)_{b\geq 0}$ where $\Gamma_b := Y^b_{1:N}$, defined with the following kernel :

$$\mathbb{K}\Big((\Gamma_{i+1},\theta_{i+1})|(\Gamma_i,\theta_i)\Big) := \mathbb{P}_0(\Gamma_{i+1})q(\theta_{i+1}|\theta_i)min\bigg(1, \frac{\pi(\theta_{i+1}|\Gamma_{i+1})}{\pi(\theta_i|\Gamma_i)}\frac{q(\theta_i|\theta_{i+1})}{q(\theta_{i+1}|\theta_i)}\bigg) \\ + c_{(\Gamma_i,\theta_i)}\delta_{(\Gamma_i,\theta_i)}(\Gamma_{i+1},\theta_{i+1})$$

This kernel is a bit different from the one we were studying, the only difference comes from the acceptance-reject ratio. This new Markov chain still corresponds to an algorithm in which at each iteration a new data subset Γ_{i+1} is sampled, but it is followed by a slightly modified version of the acceptance-reject ratio of the M-H algorithm, indeed the initial ratio of the posteriors $\frac{\pi(\theta_{i+1}|\Gamma_{i+1})}{\pi(\theta_i|\Gamma_{i+1})}$ is simply replaced by $\frac{\pi(\theta_{i+1}|\Gamma_{i+1})}{\pi(\theta_i|\Gamma_i)}$, and thus not only the new data subset but also the previous one plays now a role in the reject step.

Now we can easily show that this Markov chain is reversible with respect to the following joint distribution $\mathbb{P}_0(\Gamma_i)\pi(\theta_i|\Gamma_i)$:

$$\begin{aligned}&\mathbb{P}_0(\Gamma_i)\pi(\theta_i|\Gamma_i)\mathbb{K}\Big((\Gamma_{i+1},\theta_{i+1})|(\Gamma_i,\theta_i)\Big)\\ &\qquad= \mathbb{P}_0(\Gamma_i)\pi(\theta_i|\Gamma_i)\mathbb{P}_0(\Gamma_{i+1})q(\theta_{i+1}|\theta_i)min\bigg(1, \frac{\pi(\theta_{i+1}|\Gamma_{i+1})}{\pi(\theta_i|\Gamma_i)}\frac{q(\theta_i|\theta_{i+1})}{q(\theta_{i+1}|\theta_i)}\bigg)\\ &\qquad\quad + c_{(\Gamma_i,\theta_i)}\delta_{(\Gamma_i,\theta_i)}(\Gamma_{i+1},\theta_{i+1})\\ &\qquad= \mathbb{P}_0(\Gamma_i)\mathbb{P}_0(\Gamma_{i+1})min\bigg(\pi(\theta_i|\Gamma_i)q(\theta_{i+1}|\theta_i), \pi(\theta_{i+1}|\Gamma_{i+1})q(\theta_i|\theta_{i+1})\bigg)\\ &\qquad\quad + c_{(\Gamma_i,\theta_i)}\delta_{(\Gamma_i,\theta_i)}(\Gamma_{i+1},\theta_{i+1}) \ \text{ which is symmetric.}\end{aligned}$$

Therefore, it means that this joint distribution is a stationary one, and that under some regularity assumptions this extended Markov chain will converge towards it. Now looking only at the parameters, we finally get the average posterior $\widetilde{\pi}_N(\theta_i)$ as its marginal distribution. This new version of the algorithm is thus a guaranty for us to be able to sample from the average posterior distribution in a reasonable computational time, using a simple MCMC algorithm.

Unfortunately this algorithm is not widely applicable. The Markov chain has indeed the right stationary distribution but the switch in the acceptance ratio reveals some tractability issues, which were solved by the standard M-H algorithm but pose a problem here :

$$\frac{\pi(\theta_{i+1}|\Gamma_{i+1})}{\pi(\theta_i|\Gamma_i)} = \frac{\mathcal{L}(\Gamma_{i+1}|\theta_{i+1})p(\theta_{i+1})}{\mathcal{L}(\Gamma_i|\theta_i)p(\theta_i)} \times \frac{\mathcal{Z}(\Gamma_i)}{\mathcal{Z}(\Gamma_{i+1})}$$

Indeed, the ratio now have two constants, that does not simplify. This means that this new algorithm can actually only be used if the following distribution, also called the evidence :

$$\mathcal{Z}(\Gamma) := \int_\Theta p(\theta)\mathcal{L}(\Gamma|\theta)d\theta \quad \text{is tractable.}$$

But this is not the case for a very wide range of problems. Indeed most of the time for complicated models only the Likelihood and the prior can be computed. This is the fact that the M-H algorithm only requires these two conditions that makes it applicable in almost all the situations. Here the solution we found is not convincing enough because the problems in which the evidence is directly computable, are often simple enough problems to be solved without the need of MCMC, directly by frequentist inference for instance.

The problem posed by these two intractable constants seemed to be related to another research problem called "doubly intractable", which pushed me to study the following research paper: I. Murray, Z. Ghahramani, D. J. C. MacKay. MCMC for doubly intractable distributions. Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence, pages 359-366, 2006. These type of problems are actually not really similar to our case, because in the paper the constants that does not simplify are functions depending on the parameters whereas in our case these are constants depending on the data subsets. We have tried to find ways to adapt the solutions provided, as the exchange algorithm for instance, to our problem, but all the attempts were unsuccessful.

We did not find any way to approximate this ratio neither. Since the data is changing all the time, an estimation of the evidence during the algorithm seem complicated. What is more, the aim of this research internship was to find simple algorithms and to avoid too much computational burden, this is why we did not investigate this way further.

3.3 An asymptotic solution

We then found another solution using the classical asymptotic results already presented. Indeed, if the model is well specified, the size of the data subset N is large enough, and regularity assumptions hold, another variation of the M-H algorithm is conceivable, if we are able to provide a good estimate of the Fisher's Information Matrix. Note that since we can in that case approximate the posterior distribution $\pi(\theta|Y_{1:N})$ by a Gaussian one with the the Fisher's Information as its variance, the estimation of the Fisher's Information may be done in a first step using a standard Metropolis-Hastings algorithm on a single batch of data, but maybe this procedure can be improved in order to estimate the Fisher's Information directly during the algorithm.

To simplify, let's assume here that $dim(\Theta) = 1$ and that $q(\theta_{i+1}|\theta_i)$ is symmetric, the result can be generalized easily to an asymmetric distribution. We use in the following lines the following notations: θ_0 as the true parameter, $\lambda := \widehat{\theta}_{ML}$ as the estimator of the Maximum Likelihood, and $\Sigma := I_N^{-1}$ as the inverse of the Fisher's Information. Let's consider the following kernel :

$$\mathbb{K}\Big(\theta_{i+1}|\theta_i\Big) := \int_{Y^N} \mathbb{P}_0(\Gamma)q(\theta_{i+1}|\theta_i)min\Big(1, \frac{\pi(\theta_{i+1}|\Gamma)}{\pi(\theta_i|\Gamma)} exp\Big\{ - \frac{(\theta_{i+1}-\theta_i)^2}{2\Sigma}\Big\}\Big)d\Gamma \, + \, c_{\theta_i}\delta_{\theta_i}(\theta_{i+1})$$

We consider here a variation of the M-H algorithm in which the acceptance ratio has a corrective term: $exp\Big\{ - \frac{(\theta_{i+1}-\theta_i)^2}{2\Sigma}\Big\}$ which is lower than one, and decrease at an exponential speed with the distance between two consecutive parameters. As we said it is only computable if we know Σ i.e. if we are able to get a good approximation of the Fisher's Information. Now under these assumptions when N is large enough, we can approximate like we already did previously the posterior distribution by a conditional

Gaussian distribution given the estimator of the Maximum Likelihood, and we can approximate as well the distribution of the Maximum Likelihood by a Gaussian one, centered on the true parameter θ_0:

$$\mathbb{K}\Big(\theta_{i+1}|\theta_i\Big) \approx \int_{\mathbb{R}} \mathcal{N}(\theta_i,\Sigma)(\lambda)q(\theta_{i+1}|\theta_i)min\Big(1, \frac{\mathcal{N}(\lambda,\Sigma)(\theta_{i+1})}{\mathcal{N}(\lambda,\Sigma)(\theta_i)}exp\Big\{-\frac{(\theta_{i+1}-\theta_i)^2}{2\Sigma}\Big\}\Big)d\lambda + c_{\theta_i}\delta_{\theta_i}(\theta_{i+1})$$

From that approximation, multiplying the kernel by a Gaussian distribution, we get a very pleasant result after some developments combined with the variable change $\gamma := \lambda - \frac{(\theta_{i+1}-\theta_i)}{2}$. Indeed after these calculations we get the following expression:

$$\mathcal{N}(\theta_0,\Sigma)(\theta_i)\mathbb{K}\Big(\theta_{i+1}|\theta_i\Big) \approx \frac{q(\theta_{i+1}|\theta_i)}{2\pi\Sigma}\int_{\mathbb{R}} min\Big(exp\Big\{-\frac{1}{2\Sigma}\Big[(\gamma-\theta_0-\frac{(\theta_i-\theta_{i+1})}{2})^2+(\theta_i-\theta_0)^2\Big]\Big\}...$$

$$..., exp\Big\{-\frac{1}{2\Sigma}\Big[(\gamma-\theta_0-\frac{(\theta_{i+1}-\theta_i)}{2})^2+(\theta_{i+1}-\theta_0)^2\Big]\Big\}\Big)d\gamma$$

This expression is symmetric, which assures us that, asymptotically, the chain is reversible with respect to the following distribution: $\mathcal{N}(\theta_0,\Sigma)(\theta_i)$. This final distribution is centered on the true parameter, and its variance is twice lower than the Gaussian approximation of $\widetilde{\pi}_N(\theta)$ for large N. Thus this variance reduction will produce an even more accurate estimator than the average posterior, but it is actually not that important. The key thing is that from a large sample from this final distribution, we can estimate almost perfectly the true parameter θ_0 with a simple empirical mean.

3.4 Graphical results

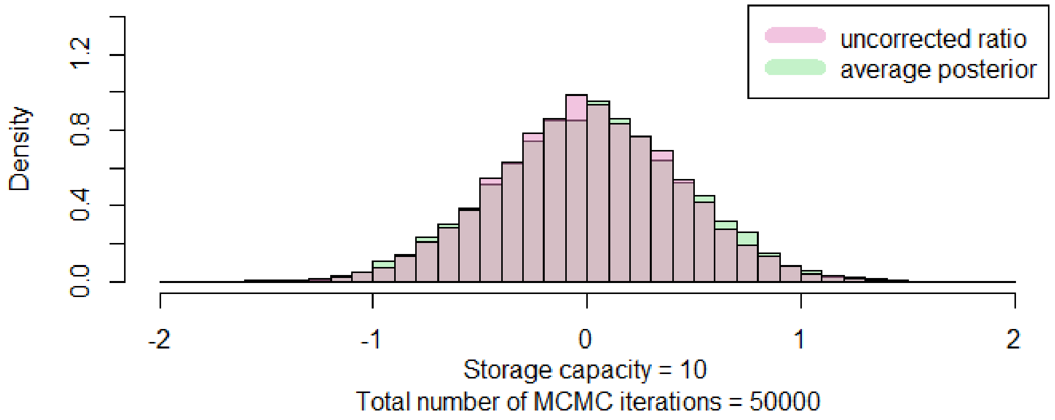
I close this chapter with some quick tests of the algorithms presented. Taking our simple Gaussian example back, we have compared three different variations of the Metropolis-Algorithm, each of them leading to a different stationary distribution. The purpose is to perform Bayesian inference on the mean parameter of a Gaussian distribution, which its variance is known equal to one, while the unknown mean parameter is zero. We want to do this inference as if the model was too complicated to allow us to perform any other type of inference than the use of MCMC. As we just presented, each algorithm is a M-H variation in which the data changes at each step.

The first algorithm is the standard uncorrected version of the M-H ratio. The second one is the algorithm based on an extended Markov chain, which if we assume being able to compute the evidence for a given subset of data, samples directly from the average posterior distribution. The third and final one needs asymptotic results, but with a good estimation of the Fisher's Information, it then allows to sample from a Gaussian distribution centered on the true parameter with a variance even twice lower than the variance of the average posterior.

For a limited storage capacity fixed at 10 observations per batch, we have performed 50 000 iterations of each algorithm. The final algorithm has only been tested directly on a Gaussian example which works thus even for small data batches, whereas we would have liked to test it on a more complicated model and assess its efficiency depending on the more or less large storage capacity chosen. Apart from time constraints, note that the study of a complicated model would often be compromised by the fact that a complex Likelihood would make the posterior become less well known, and so the average posterior would become an intractable distribution.

We present hereafter the empirical distributions sampled on the two following graphs:

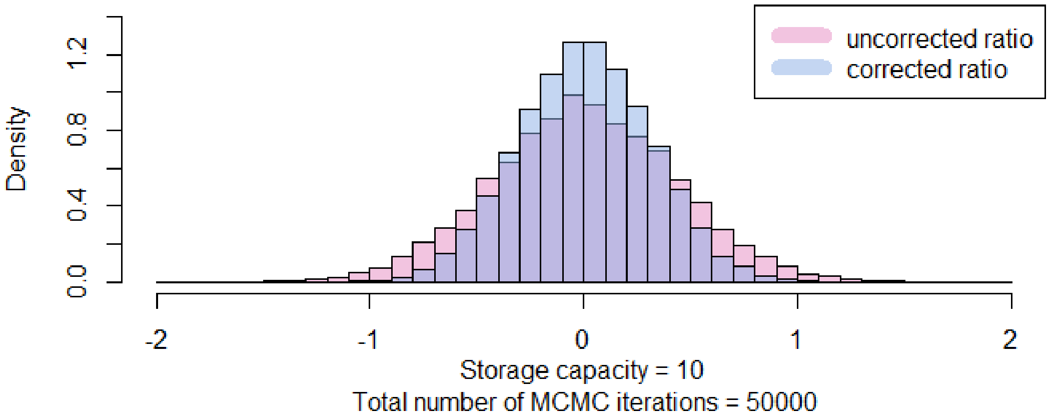
Figure 3.1: Stationary distributions, targeting the average posterior:



For the uncorrected ratio, the empirical mean after 50000 iterations was 0.0031 and the empirical variance was 0.1862. For the second algorithm which has the guaranty of targeting the average posterior, the empirical mean was 0.0082 and the empirical variance was 0.2006. On this first histogram, we can see that the two distributions are actually very close. The algorithm from the extended chain works well, indeed it has approximately the right empirical mean and variance, theoretically 0 and $\frac{2}{N} = 0.2$.

However, we can guess that stationary distribution from the uncorrected ratio is a bit noisy because it has here a smaller variance than expected, though in this example, this improper use of the M-H algorithm would works fine to estimate the mean parameter because it is centered on the true parameter. I have the intuition that this comes from the symmetry of the Gaussian posterior distributions, causing the badly known random walk to be symmetric as well and thus to be centered on the true parameter. But I think if we had tested it on a very asymmetric example with small data batches, this improper use of the Metropolis-Hastings algorithm would have induced a bias into the resulting stationary distribution.

Figure 3.2: Stationary distributions, correction of the acceptance ratio:



For the third and final algorithm, with a corrected acceptance ratio, the empirical mean after 50 000 iterations was 0.0046 and the empirical variance was 0.0983. Theoretically, the stationary distribution is a $\mathcal{N}\left(0, \frac{1}{N} = 0.1\right)$. The variance is indeed twice lower than with the average posterior. Thus the empirical results show that this algorithm works fine as soon as we are on a Gaussian case and we know the Fisher's Information.

But in real cases the Normality comes from approximations based on asymptotic results. We would have liked to assess the efficiency of this algorithm depending on the quality of these approximations. We would also have liked to see if the sampling was sensitive to mistakes made in the estimation of the Fisher's Information.

These results show that the M-H variations found works fine in practice, though I have only performed some simple tests. But the efficiency of these algorithms needs to be truly measured on more complex examples, involving multidimensional ones. Robustness to the statistical model complexity should be assessed in some way, especially for the final algorithm which is based on approximations. Finally, sensitivity to these approximations' mistakes would complete a proper study of these variations.

Conclusion

My research internship was about studying the stability of some classical MCMC algorithms in a context of limited storage capacity of the data in the computer memory. Beginning with one of the most famous ones, the Metropolis-Hastings algorithm, we showed on a simple Gaussian example that a direct and improper use of the M-H algorithm in which data was replaced at each step was actually very stable.

We then changed our approach and focused on the M-H algorithm during the whole internship. We showed that with some simple variations of the M-H acceptance ratio were able to provide new algorithms targeting some kinds of "average" posterior distributions. And we showed that these distributions would always provide better estimators than the classical use of one unique subset of data.

However, these findings remain partial because their applicability depends on some assumptions (tractable evidence, large enough storage capacity, regularity). From these results, many paths remain unexplored at the end of my internship:

Firstly, can the good results of the standard M-H algorithm with an uncorrected ratio be properly formalized? Do these good results directly come from the symmetry of the Gaussian distributions? In that case it could mean that for large data subsets this simple use of the M-H algorithm may directly provide symmetric stationary distributions centered on the true parameter.

Secondly, in a context where data is constantly replaced, is it possible to deal with intractable evidence? Is it possible for instance to estimate these unknown constants in the acceptance ratio while the algorithm still runs, in order to build a noisy version of this algorithm? Indeed we have no real solution yet for small data subsets, for which posterior distributions may be completely asymmetric, that is why we are very interested in solving this issue.

Finally, in the last algorithm the estimation of the Fisher's Information is necessary before performing the simulations. Maybe it is possible to adapt it in a way that would estimate the Fisher's Information directly during the algorithm. What is more, can the efficiency of this last algorithm hold empirically considering not so large data subsets?

Bibliography

References

(1) Alexey Miroshnikov, Erin M. Conlon. Parallel Markov Chain Monte Carlo for Non-Gaussian Posterior Distributions. arXiv preprint arXiv:1506.03162, 2015.

(2) Gareth O. Roberts, Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. Probability Surveys Vol. 1, pages 20–71, 2004.

(3) Gareth O. Roberts, Jeffrey S. Rosenthal. Coupling and Ergodicity of Adaptive MCMC. Journal of Applied Probability, Volume 44, Number 2, pages 458-475, 2007.

(4) I. Murray, Z. Ghahramani, D. J. C. MacKay. MCMC for doubly intractable distributions. Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence, pages 359-366, 2006.

Related papers

(a) Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy monte carlo: Convergence of markov chains with approximate transition kernels. Statistics and Computing, pages 1–19, 2014.

(b) Nicolas Chopin. A sequential particle filter method for static models. Biometrika, 89(3):539–552, 2002.

(c) Florian Maire, Nial Friel, and Pierre Alquier. Light and widely applicable mcmc: Approximate bayesian inference for large datasets. arXiv preprint arXiv:1503.04178, 2015.

(d) Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6):1087–1092, 1953.

(e) Natesh S Pillai and Aaron Smith. Ergodicity of approximate mcmc chains with applications to large data sets. arXiv preprint arXiv:1405.0182, 2014.

APPENDIX

As a unique appendix, I chose to present the original version of the research topic that Florian Maire provided me at the beginning of this internship. This way, one can see the evolution from the research question posed at first, to the solution we found to those problems.

Stability of Markov chains Monte Carlo methods in the context of a moving target distribution

June 8, 2015

General setting

One of the core challenge in Bayesian statistics is to infer the distribution of a parameter $\theta \in \Theta$ ($\Theta \subseteq \mathbb{R}^d$) related to a random process of interest $Y \in \mathsf{Y}$ ($\mathsf{Y} \subseteq \mathbb{R}^m$), with unknown distribution $\mathbb{P}_0$. We observe n realizations $\{y_1, \ldots, y_n\}$ of this process and assume a prior distribution p for θ, defined on (Θ, ϑ). The so–called posterior distribution of θ given the observations $\{y_1, \ldots, y_n\}$ admits for density

$$\pi_*(\theta \,|\, y_1, \ldots, y_n) \propto f(y_1, \ldots, y_n \,|\, \theta) p(\theta)\,, \tag{1}$$

where for all $\theta \in \Theta$, $y \to f(y \,|\, \theta)$ is the likelihood of one (or a set of) observation(s).

Knowing the posterior distribution π_* can be desirable in order to model the data $Y \in \mathsf{Y}$ through the distribution $\mathbb{P}_*$ defined on $(\mathsf{Y}, \mathcal{Y})$, often referred to as the posterior predictive distribution, whose density is defined by:

$$f_*(y) = \int_\Theta f(y \,|\, \theta) \pi_*(\mathrm{d}\theta \,|\, y_1, \ldots, y_n)\,.$$

In realistic situations, f_* will often not be available analytically as integrating with respect to π_* might not be doable. Therefore, for any set $A \in \mathcal{Y}$, the probability $\mathbb{P}_*(A)$ will be intractable, hence preventing to predict or simulate new data.

However, we note that:

(Situation–1) if for any $\theta \in \Theta$, $\int_A f(y \,|\, \theta)\mathrm{d}y$ is tractable

$$\mathbb{P}_*(A) = \int_\Theta \left\{ \int_A f(y \,|\, \theta)\mathrm{d}y \right\} \pi_*(\mathrm{d}\theta \,|\, y_1, \ldots, y_n) = \mathbb{E}_* \left\{ \int_A f(y \,|\, \theta)\mathrm{d}y \right\},$$

where the expectation is taken under the posterior distribution π_*

(Situation–2) otherwise

$$\mathbb{P}_*(A) = \int_\Theta \int_\mathsf{Y} \mathbb{1}_A(y) f(\mathrm{d}y \,|\, \theta) \pi_*(\mathrm{d}\theta \,|\, y_1, \ldots, y_n) = \mathbb{E}\left\{\mathbb{1}_A(Y)\right\}\,,$$

where the expectation is taken under the distribution $f(y \,|\, \theta)\pi_*(\theta \,|\, y_1, \ldots, y_n)\mathrm{d}\theta\mathrm{d}y$.

Therefore, an approximation of $\mathbb{P}_*(A)$ can always be obtained through numerical integration, provided that samples either from π_* or $f(\,\cdot\,|\,\theta)\pi_*(\,\cdot\,|\, y_1, \ldots, y_n)$ are available. Indeed, defining

$$\hat{\mathbb{P}}_*^{(L)}(A) = \begin{cases} \frac{1}{L}\sum_{\ell=1}^L \int_A f(y \,|\, \theta_\ell)\mathrm{d}y\,, & \theta_\ell \sim \pi_*(\,\cdot\,|\, y_1, \ldots, y_n) \\ \frac{1}{L}\sum_{\ell=1}^L \mathbb{1}_A(Y_\ell)\,, & (Y_\ell, \theta_\ell) \sim f(\,\cdot\,|\,\theta)\pi_*(\,\cdot\,|\, y_1, \ldots, y_n) \end{cases} \tag{2}$$

we have $\hat{\mathbb{P}}_*^{(L)}(A) \to \mathbb{P}_*(A)$ as $L \to \infty$ almost surely.

Research question

For simplicity, assume (Situation–1), *i.e* $\theta \to \int_A f(y \mid \theta)\mathrm{d}\theta$ is analytically tractable (Considering Situation–2 would not change the approach). The estimate $\hat{\mathbb{P}}_*^{(L)}(A)$ can be computed, provided that *i.i.d.* samples $\{\theta_\ell\,, \ell \in \mathbb{N}\}$ from π_* are available. However, this is generally not the case and one typically resorts to approximative simulation techniques such as Markov chain Monte Carlo methods to get non *i.i.d.* samples from π_*.

The main question we want to investigate is related to the quality of the estimate $\hat{\mathbb{P}}_*^{(L)}(A)$ (2). We consider the following two approaches:

Approach–1 on the one hand, non *i.i.d.* samples from the posterior π_* can be routinely obtained using Markov chain Monte Carlo methods provided that a transition kernel $\theta_k \sim P_*(\theta_{k-1}, \ldots)$ leaving π_* invariant is available. This involves the simulation of a Markov chain $\{\theta_\ell\,, \ell \in \mathbb{N}\}$ with a fixed target distribution π_*:

$$\theta_0 \sim \nu \longrightarrow \cdots \longrightarrow \theta_k \sim P_*(\theta_{k-1}, \cdot) \longrightarrow \theta_{k+1} \sim P_*(\theta_k, \cdot) \longrightarrow \cdots\,, \tag{3}$$

which can be typically achieved using the Metropolis–Hastings sampler [4]. The theory states that, under mild assumptions, a law of large number and a central limit theorem will allow to get an estimate of $\hat{\mathbb{P}}_*^{(L)}(A)$, whose theoretical property are well understood.

Approach–2 on the other hand, assume that the observed measurements change over time (some new observations $y_{n+1}, y_{n+2}, \ldots$ might come in and others might become unavailable). This yields a sequence of distributions $\{\pi_t\}_{t>0}$ where π_t is the posterior distribution (1) defined conditioning on the data available at time t, say $y^{(t)} = (y_1^{(t)}, \ldots, y_{n_t}^{(t)})$:

$$\pi_t(\theta \mid y^{(t)}) \propto f(y_1^{(t)}, \ldots, y_{n_t}^{(t)} \mid \theta)p(\theta)\,. \tag{4}$$

For simplicity, assume that n_t is a constant, say $n_t = n$. Getting sequential samples from $\pi_1, \pi_2, \ldots$ is not as straightforward as in the situation of a fixed target distribution π_* (Approach–1). An option is to use a self-normalized Importance Sampling method *ala* Iterated Batch Importance Sampling [2]. The alternative we want to study is the use of the Markov chain

$$\hat{\theta}_0 \sim \nu \longrightarrow \cdots \longrightarrow \hat{\theta}_k \sim P_k(\hat{\theta}_{k-1}, \cdot) \longrightarrow \hat{\theta}_{k+1} \sim P_{k+1}(\hat{\theta}_k, \cdot) \longrightarrow \cdots\,, \tag{5}$$

where for all k, $P_k(\hat{\theta}_{k-1}, \cdot)$ is a transition kernel having π_k as stationary distribution. Compared to the Markov chain $\{\theta_\ell\,, \ell \in \mathbb{N}\}$ (3), most of the theoretical arguments justifying the stabilization are lost.

In the perspective of estimating $\mathbb{P}_*(A)$ with $\hat{\mathbb{P}}_*^{(L)}(A)$ (2), we want to study the difference between the *disturbed* Markov chain $\{\hat{\theta}_\ell\,, \ell \in \mathbb{N}\}$ (5) and the *plain* Markov chain $\{\theta_\ell\,, \ell \in \mathbb{N}\}$ (3). Indeed, there seems to be two competing effects:

- on the one hand, the Markov chain $\{\theta_\ell\,, \ell \in \mathbb{N}\}$ will be stationary and converge in distribution towards $\pi_*(\,\cdot \mid y_1, \ldots, y_n)$ which will allow to get a unbiased estimate in (2).

- on the other hand, even though $\{\hat{\theta}_\ell, \ell \in \mathbb{N}\}$ is likely to be unstable because the target distributions constantly changes, the fact that the chain $\{\hat{\theta}_\ell, \ell \in \mathbb{N}\}$ makes use of many more observations might be beneficial as well and might stabilize the estimate (2) in an alternative way.

For example, assuming first that the data $\{Y_k\}_k$ are *i.i.d.* realizations of some unknown distribution $\mathbb{P}_0$, is there any chance to observe a stabilizing effect from the fact that the *disturbed* chain allows to learn from a larger collection of data?

Related problems

- A significant number of recent contributions have allowed to understand the theoretical properties of a variety of *disturbed* Markov chains (eg [1], [5]), but not in the perspective of a estimating $\mathbb{P}_*(A)$ (2).

- In a recent work with Nial and Pierre [3], we have proposed a methodology (Light and Widely Applicable MCMC) allowing to reduce by an arbitrary low magnitude the CPU burden generated by large datasets that are being used in Bayesian inference. The main finding is that, if a prohibitively large number of N data are available, a Markov chain targeting a sequence of different posterior distributions $\{\pi_t\}_t$ (4) each involving a subset of n data, will be efficient provided that the subsets of data are refreshed according to their *representativeness* with respect to the full dataset.

- Adil's internship was about comparing different strategies to choose a subset of n observations from a prohibitively large dataset $\{y_1, \dots, y_N\}$. The main finding of the analysis is in accordance with [3]. More precisely, a posterior distribution involving a subset of n data will be closer to the posterior involving the full dataset, if the subset of n data resembles the full dataset.

- In our work, we do not assume that a large dataset is available but rather that a stream of data is available. The stream is arbitrary, ie the data are imposed to us. Its size can be taken as constant. A typical application would be estimating the posterior distribution of a parameter of interest given data that are acquired by a sensor with limited storage resource: data will change overtime. In this context, we want to investigate methods and theoretical arguments for estimation of probabilities $\mathbb{P}_*(A)$ for any $A \in \mathcal{Y}$, (3).

References

[1] Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy monte carlo: Convergence of markov chains with approximate transition kernels. Statistics and Computing, pages 1–19, 2014.

[2] Nicolas Chopin. A sequential particle filter method for static models. Biometrika, 89(3):539–552, 2002.

[3] Florian Maire, Nial Friel, and Pierre Alquier. Light and widely applicable mcmc: Approximate bayesian inference for large datasets. arXiv preprint arXiv:1503.04178, 2015.

[4] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. The journal of chemical physics, 21(6):1087–1092, 1953.

[5] Natesh S Pillai and Aaron Smith. Ergodicity of approximate mcmc chains with applications to large data sets. arXiv preprint arXiv:1405.0182, 2014.