TP 3: Hasting-Metropolis (and Gibbs) samplers

Exercise 1: Hasting-Metropolis within Gibbs – Stochastic Approximation EM

We observe a group of N (independent) individuals. For the i-th individual, we have k_i measurements $y_{i,j} \in \mathbb{R}$. In studies on the progression of diseases, measurements $y_{i,j}$ can be measures of weight, volume of brain structures, protein concentration, tumoral score, etc. over time. We assume that each measurement $y_{i,j}$ are independent and obtained at time $t_{i,j}$ with $t_{i,1} < \ldots < t_{i,k_i}$.

1.A – A population model for longitudinal data

We wish to model an average progression as well as individual-specific progressions of the measurements from the observations $(y_{i,j})_{i\in[\![1,N]\!],\, j\in[\![1,k_i]\!]}$. To do that, we consider a hierarchical model defined as follows.

i. We assume that the average trajectory is the straight line which goes through the point p_0 at time t_0 with velocity v_0

$$d(t) := p_0 + v_0(t - t_0)$$

where

$$p_0 \sim \mathcal{N}(\,\overline{p_0}, \sigma_{p_0}^2\,) \quad ; \quad t_0 \sim \mathcal{N}(\,\overline{t_0}, \sigma_{t_0}^2\,) \quad ; \quad v_0 \sim \mathcal{N}(\,\overline{v_0}, \sigma_{v_0}^2\,)$$

and $\sigma_{p_0}, \sigma_{t_0}, \sigma_{v_0}$ are fixed variance parameters. While we consider straight lines, we can also fix p_0.

ii. For the i-th individual, we assume a trajectory of progression of the form

$$d_i(t) := d(\,\alpha_i(t - t_0 - \tau_i) + t_0\,)\,.$$

The trajectory of the i-th individual corresponds to an affine reparametrization of the average trajectory. This affine reparametrization, given by $t \mapsto \alpha_i(t - t_0 - \tau_i) + t_0$, allows to characterize changes in speed and delay in the progression of the i-th individual with respect to the average trajectory. Moreover, we assume that for all i-th individual measurements

$$\begin{cases} y_{i,j} = d_i(t_{i,j}) + \varepsilon_{i,j} \quad \text{where} \quad \varepsilon_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \\ \alpha_i = \exp(\xi_i) \quad \text{where} \quad \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\xi^2) \\ \tau_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\tau^2) \end{cases} .$$

The parameters of the model are $\theta = (\,\overline{t_0}, \overline{v_0}, \sigma_\xi, \sigma_\tau, \sigma\,)$. For all $i \in [\![1, N]\!]$, $z_i = (\alpha_i, \tau_i)$ are random variables called *random effects* and $z_{pop} = (\,t_0, v_0\,)$ are called *fixed effects.* The fixed effects are used to model the group progression whereas random effects model individual progressions.
Likewise, we define $\theta_i = (\sigma_\xi, \sigma_\tau, \sigma)$ and $\theta_{pop} = (\,\overline{t_0}, \overline{v_0}\,)$.

We consider a bayesian framework and assume the following *a priori* on the parameters θ :

$$\overline{t_0} \sim \mathcal{N}(\overline{\overline{t_0}}, s_{t_0}^2) \quad ; \quad \overline{v_0} \sim \mathcal{N}(\overline{\overline{v_0}}, s_{v_0}^2)$$

$$\sigma_\xi^2 \sim \mathcal{W}^{-1}(v_\xi, m_\xi) \quad ; \quad \sigma_\tau^2 \sim \mathcal{W}^{-1}(v_\tau, m_\tau) \quad ; \quad \sigma^2 \sim \mathcal{W}^{-1}(v, m)\,.$$

where $\mathcal{W}^{-1}(v,m)$ ($v > 0$, $m \in \mathbb{N}^*$) is the inverse-Wishart distribution :

$$f_{\mathcal{W}^{-1}}(\sigma^2) \;=\; \frac{1}{\Gamma\left(\frac{m}{2}\right)} \frac{1}{\sigma^2} \left(\frac{v}{\sigma\sqrt{2}}\right)^m \exp\left(-\frac{v^2}{2\sigma^2}\right)\,.$$

1. Write the complete log-likelihood of the previous model $\log q(y, z, \theta)$ and show that the proposed model belongs to the curved exponential family.

2. Generate synthetic data from the model by taking some reasonable values for the parameters.

1.B – HM-SAEM – Hasting-Metropolis sampler

In order to estimate – by a *maximum a posteriori* for instance – the parameters of this statistical model, we will use the SAEM – Stochastic Approximation EM algorithm. However, this algorithm requires that we are able to sample from the *a posteriori* distribution, see algorithm 2.

We will use the Hasting-Metropolis algorithm to that end: actually a direct sampling is not possible in our context. Let $q(.|z)$ be the proposal distribution of the algorithm, *i.e.* the conditional probability of proposing a state z^* given the current state z, and π be a density defined on an open set $\mathcal{U}$ of $\mathbb{R}^n$. The Hasting-Metropolis algorithm targeting π writes:

Algorithm 1: Hasting-Metropolis Sampler

```
Given $z^{(0)}$
for $k = 0$ to maxIter do
    #Proposal: $z^* \sim q(.|z^{(k)})$
    #Acceptance-Rejection: $\alpha(z^{(k)}, z^*) = \min\left(1, \frac{q(z^{(k)}|z^*)\,\pi(z^*)}{q(z^*|z^{(k)})\,\pi(z^{(k)})}\right)$
    $z^{(k+1)} = \begin{cases} z^* \text{ with probability } \alpha(z^{(k)}, z^*) \\ z^{(k)} \text{ with probability } 1 - \alpha(z^{(k)}, z^*) \end{cases}$
end
```

3. Propose a Metropolis-Hastings sampler to sample from the *a posteriori* distribution of the latent variable $z = (z_{pop}, z_i)_{i \in [\![1,N]\!]} = (t_0, v_0, \xi_i, \tau_i)_{i \in [\![1,N]\!]} \in \mathbb{R}^{2N+2}$.

A natural choice for the proposal distribution is to consider a multivariate Gaussian distribution $\mathscr{N}(z, \sigma_{prop})$. Thus, the acceptance ratio simply writes $1 \wedge \frac{\pi(z^*)}{\pi(z^{(k)})}$. This algorithm is called Symmetric Random Walk Hasting-Metropolis algorithm.

We want to use the Expectation-Maximization algorithm to maximize the likelihood, especially as we have proved at the question 1 that the model belongs to the curved exponential family. Nevertheless, the expectation required by the EM algorithm $Q_k(\theta^{(k)}) = \mathbb{E}_y\left[q(y, z, \theta^{(k)})|y, \theta^{(k)}\right]$ cannot be calculated here, due to the latent variable z. So, we have to use a stochastic version of the EM algorithm, namely the SAEM algorithm: the Expectation step is split into two steps, the Simulation one and the Stochastic Approximation one, see Algorithm 2.

4. Compute the optimal parameters

$$\theta^{(k)} = \underset{\theta\in\Theta}{\operatorname{argmax}} \{-\Phi(\theta) + \langle\ S_k \mid \Psi(\theta)\ \rangle\}$$

and implement the HM-SAEM in order to find the MAP. In particular, we assume that the MAP exists. Use the question 2 to check your algorithm.

For step-sizes ε_k we can choose a parameter N_b – burn-in parameter – and define

$$\forall k \in \mathbb{N}, \qquad \begin{cases} 1 & \text{if } k \in [\![1, N_b]\!] \\ (k - N_b)^{-\alpha} & \text{otherwise} \end{cases}$$

where $\alpha \in [\frac{1}{2}, 1[$ is necessary to ensure the convergence of the MCMC-SAEM. See [AKT10, AK15].

> **Remark** : Contrary to Bayesian inference, where burn-in traditionally refers to a certain amount of samples which are discarded, here the term burn-in refers to memoryless approximation steps. In other words, during the burn-in phase, the information contained in $z^{(k)}$ is not used in the approximation of the sufficient statistics. In practice, the burn-in period is chosen to be half of the maximum number of iterations.

Algorithm 2: MCMC-SAEM (for curved exponential family)

```
Given data $y$ and intial guess $\theta^{(0)}$
#Initialization : $z^{(0)} = 0$, $S^{(0)} = 0$ and step-sizes $(\varepsilon_k)_{k\geqslant 0}$.
for $k = 0$ to maxIter do
    #Simulation $z^{(k+1)} \sim q(.|y, \theta^{(k)})$   (sampler intialized at $z^{(k)}$)
    #Stochastic Approximation : $S^{(k+1)} = S^{(k)} + \varepsilon_k\big(S(y, z^{(k+1)}) - S^{(k)}\big)$,
    #Maximization : $\theta^{(k+1)} = \underset{\theta\in\Theta}{\operatorname{argmax}}\big\{-\Phi(\theta^{(k)}) + \big\langle S^{(k+1)} \mid \Psi(\theta^{(k)})\big\rangle\big\}$
end
```

1.C – HMwG-SAEM – Hasting-Metropolis within Gibbs sampler

However, the dimension of the latent variable z may become high if we consider a large cohort and so the *a posteriori* distribution of the latent variable difficult to sample. In that case, we can use a Gibbs sampler which consists in generating an instance from the distribution of each (sub)-variable in turn, conditional on the current values of the other (sub)-variables. Gibbs sampling is more generally applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easy (or at least, easier) to sample from.

If we consider π, a density defined on an open set $\mathcal{U}$ of $\mathbb{R}^n$ $(n \geqslant 2)$ and if we denote, for $\ell \in [\![1, n]\!]$, π_ℓ the ℓ^{th} full conditional of π given by

$$\pi_\ell(z_\ell \mid z_{-\ell}) \propto \pi(z)$$

where $z_{-\ell} = \{z_1, \dots, z_{\ell-1}, z_{\ell+1}, \dots, z_n\}$, we recall that the classical Gibbs sampler writes as follows :

Algorithm 3: Gibbs Sampler

```
Given $z^{(k)} = (z_1^{(k)}, \dots, z_n^{(k)})$
for $\ell = 1$ to $n$ do
    $z_\ell^{(k+1)} \sim \pi_\ell(z_\ell \mid z_1^{(k+1)}, \dots, z_{\ell-1}^{(k+1)}, z_{\ell+1}^{(k)}, \dots, z_n^{(k)})$        $(\star)$
end
```

When direct sampling from the full conditionals is not possible, the step $(\star)$ is often replaced with a Metropolis-Hastings step. The resulting MCMC algorithm is called *hybrid Gibbs sampler* or *Metropolis-Hastings within Gibbs sampler.*

5. Propose a Metropolis-Hastings within Gibbs sampler to sample from the *a posteriori* distribution of $z_i = (\xi_i, \tau_i)$.

6. Likewise, propose a HMwG sampler for the *a posteriori* distribution of $z_{pop} = (t_0, v_0)$.

7. Implement the HMwG-SAEM in order to find the MAP.

We can improve the sampling step for big dataset by considering a Block HMwG sampler instead of a "one-at-a-time" as described above HMwG sampler. In the Block version, each Metropolis-Hastings step of the algorithm consists in a multivariate symmetric random walk. Then, the Block MHwG sampler updates simultaneously block (or sets) of latent variables given the others.

8. Explain what is the advantages of a Block Gibbs sampler over a "one-at-a-time" Gibbs sampler for our model.

9. Implement a Block HMwG sampler by choosing a block for the fixed effects and a block by individuals.

The model studied in this exercise is a very simplified version of the model proposed by Jean-Baptiste Schiratti in his Phd-Thesis. For more details, you can refer to [SACD15, Sch16, COA17].

Exercise 2: Multiplicative Hasting-Metropolis

Let f be a density function on $]-1, 1[$. We consider the *multiplicative Hasting-Metropolis algorithm* defined as follows.

> Let X be the current state of the Markov chain.
>
> (i) We sample ε from the probability density function f and a random variable $\mathcal{B}$ from the Bernoulli distribution with parameter $\frac{1}{2}$.
>
> (ii) If $\mathcal{B} = 1$, we set $Y = \varepsilon X$. Otherwise, we set $Y = \frac{X}{\varepsilon}$. Then, we accept the candidate Y with a probability given by $\alpha(X, Y)$, the usual Hasting-Metropolis acceptation ratio.

1. Determine the density of the jumping distribution $Y \sim q(X, Y)$.

2. Compute the acceptation ratio α so that the chain has a given distribution π as invariant distribution.

3. Implement this sampler for two different target distributions : the first one being a distribution from which we can sample using the inverse transform method and the second one is of your choice.

4. Evaluate, in each case, the match of your samples with the true distribution.

Exercise 3: Data augmentation

Let $f\colon (x, y) \in \mathbb{R}^p \times \mathbb{R}^q \mapsto f(x, y) \in \mathbb{R}^+$ be a density with respect to the Lebesgue measure on $\mathbb{R}^{p+q}$. Let us define

$$f_X(x) := \int f(x, y)\,\mathrm{d}y \; ; \qquad f_Y(y) = \int f(x, y)\,\mathrm{d}x \; ;$$

and

$$\forall y \in \mathbf{Y} := \{\, y \in \mathbb{R}^q \mid f_Y(y) > 0 \,\}\,, \qquad f_{X|Y}(x, y) := \frac{f(x, y)}{f_Y(y)} \; ;$$

$$\forall x \in \mathbf{X} := \{\, x \in \mathbb{R}^p \mid f_X(x) > 0 \,\}\,, \qquad f_{X|Y}(x, y) := \frac{f(x, y)}{f_Y(y)} \, .$$

We define a bivariate chain $\{(X_n, Y_n), n \geqslant 0\}$ as in the following algorithm.

Algorithm 4: Data augmentation

```
Given $(X_0, Y_0) \in \mathbb{R}^p \times \mathbb{R}^q$ and $N \in \mathbb{N}$
for $n = 1$ to $N$ do
    $X_n \sim f_{X|Y}(\cdot, Y_{n-1})$
    $Y_n \sim f_{Y|X}(X_n, \cdot)$
end
return $\{(X_n, Y_n), 0 \leqslant n \leqslant N\}$
```

1. Show that the bivariate process $\{(X_n, Y_n), n \geqslant 0\}$ is a Markov chain. Give the expression of its transition kernel as a function of the quantities defined above.

2. Show that $\{Y_n, n \geqslant 0\}$ is a Markov chain : give the expression of its transition kernel and prove that $f_Y(y)\,\mathrm{d}y$ is invariant for this kernel.

Hereafter, we consider the case when

$$f(x,y) = \frac{4}{\sqrt{2\pi}} y^{\frac{3}{2}} \exp\left[-y\left(\frac{x^2}{2}+2\right)\right] \mathbb{1}_{\mathbb{R}^+}(y)$$

3. Describe a Gibbs algorithm to approximate the distribution on $\mathbb{R} \times \mathbb{R}$ with density f.

We can use a gamma distribution sampler : `numpy.random.gamma` in python or `gamrnd` from the Statistics and Machine Learning Toolbox in Matlab. We can also find a Matlab toolbox-free Gamma Generator in the Handbook of Monte Carlo Methods [KTB13] :
`https://people.smp.uq.edu.au/DirkKroese/montecarlohandbook/probdist/`.

4. Let H be a bounded function on $\mathbb{R}$. Explain how to approximate

$$\int_{\mathbb{R}} \frac{H(x)}{(4+x^2)^{\frac{5}{2}}}\,\mathrm{d}x$$

from the output $\{(X_n, Y_n), 0 \leqslant n \leqslant N\}$ of this Gibbs sampler.

References

[AK15] Stéphanie Allassonnière and Estelle Kuhn. Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. *Computational Statistics & Data Analysis*, 91:4 – 19, 2015.

[AKT10] Stéphanie Allassonnière, Estelle Kuhn, and Alain Trouvé. Construction of bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernoulli*, 16(3):641–678, 08 2010.

[COA17] Juliette Chevallier, Stéphane Oudard, and Stéphanie Allassonnière. Learning spatiotemporal piecewise-geodesic trajectories from longitudinal manifold-valued data. In *Advances in Neural Information Processing Systems 31*. 2017.

[Dut12] Somak Dutta. Multiplicative random walk metropolis-hastings on the real line. *Sankhya B*, 74(2):315–342, 2012.

[KTB13] D.P. Kroese, T. Taimre, and Z.I. Botev. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics. Wiley, 2013.

[SACD15] Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. Learning spatiotemporal trajectories from manifold-valued longitudinal data. In *Advances in Neural Information Processing Systems 28*. 2015.

[Sch16] Jean-Baptiste Schiratti. *Models and algorithms to learn spatiotemporal changes from longitudinal manifold-valued observations*. PhD thesis, École polytechnique, 2016.