Stat 535 C - Statistical Computing & Monte Carlo Methods

Lecture 14 - 28 February 2006

Arnaud Doucet

Email: arnaud@cs.ubc.ca

1.1– Outline

- More about the Metropolis-Hastings algorithm.
- Mixture and composition of kernels.
- “Hybrid” algorithms.
- Examples

2.1– Metropolis-Hastings algorithm

- Initialization:
  - Select deterministically or randomly $\theta^{(0)}$.
- Iteration i; $i \geq 1$:
  - Sample $\theta^* \sim q\left(\theta^{(i-1)}, \cdot\right)$ and compute

$$\alpha\left(\theta^{(i-1)}, \theta^*\right) = \min\left(1, \frac{\pi\left(\theta^*\right) q\left(\theta^*, \theta^{(i-1)}\right)}{\pi\left(\theta^{(i-1)}\right) q\left(\theta^{(i-1)}, \theta^*\right)}\right).$$

  - With probability $\alpha\left(\theta^{(i-1)}, \theta^*\right)$, set $\theta^{(i)} = \theta^*$; otherwise set $\theta^{(i)} = \theta^{(i-1)}$.

2.1– Metropolis-Hastings algorithm

- The transition kernel associated to the MH algorithm can be rewritten as

$$K(\theta, \theta') = \alpha(\theta, \theta') \, q(\theta, \theta') + \left(1 - \int \alpha(\theta, u) \, q(\theta, u) \, du\right) \delta_\theta(\theta') \,.$$

- The MH kernel is π-reversible hence π-invariant

$$\pi(\theta) \, K(\theta, \theta') = \pi(\theta') \, K(\theta', \theta) \Rightarrow \int \pi(\theta) \, K(\theta, \theta') \, d\theta = \pi(\theta')$$

- It is irreducible and aperiodic under very weak assumptions.

- **Independent proposal** $q(\theta, \theta') = q(\theta')$ then

$$\alpha(\theta, \theta') = \min\left(1, \left(\frac{\pi(\theta')}{q(\theta')}\right) / \left(\frac{\pi(\theta)}{q(\theta)}\right)\right)$$

- If we use an independent proposal, one should ensure

$$\frac{\pi(\theta)}{q(\theta)} \leq C \text{ for all } \theta.$$

• **Random walk** $q(\theta, \theta') = f(\theta' - \theta) = f(\theta - \theta')$ then

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right).$$

• If we use a random walk, one should ensure that the tails of distribution of the random walk increments are thick enough.

$\Rightarrow$ In all cases, the selection of $q(\theta, \theta')$ is tricky and is getting more difficult as the dimension of the parameter space is increasing.

2.2– Using gradient information to build the proposal

- We usually wants to sample candidates in regions of high probability masses.
- We can use

$$\theta' = \theta + \frac{\sigma^2}{2} \nabla \log \pi (\theta) + \sigma V \text{ where } V \sim \mathcal{N} (0, 1)$$

where σ^2 is selected such that the acceptance ratio is approximately 0.57.

- The motivation is that, we know that in continuous-time

$$d\theta_t = \frac{1}{2} \nabla \log \pi (\theta) + \sigma dW_t$$

admits π has an invariant distribution.

2.3– Local optimization

• To build $q(\theta, \theta')$, you can use complex deterministic strategies. Assume you are in θ and you want to propose

$$\theta' \sim \mathcal{N}\left(\varphi(\theta), \sigma^2\right).$$

• You do not need to have an explicit form for the mapping φ! As long as φ is a *deterministic* mapping, then it is fine. For example $\varphi(\theta)$ could be the local maximum of π closest to θ that has been determined using a gradient algorithm.

• To compute the acceptance probability of the candidate θ', you will need to compute $\varphi(\theta')$ and then you can compute the MH acceptance ratio.

3.1– Mixture of proposals

• In practice, random walk proposals can be used to explore locally the space whereas independent walk proposals can be used to jump into the space.

• So a good strategy can be to use a proposal distribution of the form

$$q\left(\theta,\theta'\right)=\lambda q_1\left(\theta'\right)+\left(1-\lambda\right)q_2\left(\theta,\theta'\right)$$

where $0<\lambda<1$.

• This algorithm is definitely valid as it is just a particular case of the MH algorithm.

• An alternative achieving the same purpose is to use a transition kernel

$$K(\theta,\theta') = \lambda K_1(\theta,\theta') + (1-\lambda) K_2(\theta,\theta')$$

where K_1 (resp. K_2) is an MH algorithm of proposal q_1 (resp. q_2).

• This algorithm is different from using $q(\theta,\theta') = \lambda q_1(\theta') + (1-\lambda) q_2(\theta,\theta')$. It is computationally cheaper and still valid as

$$\begin{aligned}\int \pi(\theta) K(\theta,\theta') \, d\theta &= \lambda \int \pi(\theta) K_1(\theta,\theta') \, d\theta + (1-\lambda) \int \pi(\theta) K_2(\theta,\theta') \, d\theta \\ &= \lambda \pi(\theta') + (1-\lambda) \pi(\theta') \\ &= \pi(\theta')\end{aligned}$$

• A sufficient condition to ensure that K is irreducible and aperiodic is to have either K_1 or K_2 irreducible and aperiodic.

• You do NOT need to have both kernels to be irreducible and aperiodic. In the limiting case, you could have $K_2(\theta, \theta') = \delta_\theta(\theta')$ and the total kernel K would still be irreducible and aperiodic if K_1 is irreducible and aperiodic.

• None of the kernels have to be irreducible and aperiodic to ensure that K is irreducible and aperiodic.

4.2– Composition of MH kernels

- Alternatively we can apply at each iteration of the algorithm first the kernel K_1 then the kernel K_2, i.e. in this case where have at iteration i

$$Z \sim K_1\left(\theta^{(i-1)}, \cdot\right) \text{ and } \theta^{(i)} \sim K_2\left(Z, \cdot\right).$$

- The composition of these kernels corresponds to

$$K\left(\theta, \theta'\right) = \int K_1\left(\theta, z\right) K_2\left(z, \theta'\right) dz.$$

- This algorithm admits the right invariant distribution as

$$\begin{aligned}
\int \pi\left(\theta\right) K\left(\theta, \theta'\right) d\theta &= \int \left(\int \pi\left(\theta\right) K_1\left(\theta, z\right) d\theta\right) K_2\left(z, \theta'\right) dz \\
&= \int \pi\left(z\right) K_2\left(z, \theta'\right) dz \\
&= \pi\left(\theta'\right)
\end{aligned}$$

4.2– Composition of MH kernels

- A sufficient condition to ensure that K is irreducible and aperiodic is to have either K_1 or K_2 irreducible and aperiodic.

- You do NOT need to have both kernels to be irreducible and aperiodic to have K irreducible and aperiodic, e.g. take K_1 irreducible and aperiodic and $K_2(\theta, \theta') = \delta_\theta(\theta')$.

- None of the kernels have to be irreducible and aperiodic to ensure that K is irreducible and aperiodic.

4.2– Composition of MH kernels

• The MH algorithm is a simple and very general algorithm to sample from a target distribution $\pi(\theta)$.

• In practice, the performance of the algorithm are choice of the proposal dis-

tribution is absolutely crucial
on the performance of the algorithm.

• In high dimensional problems, a simple MH algorithm will be useless.
It will be necessary to use a combination of MH kernels.
.... However for the time being you might not have realized the power of the mixture and composition of kernels.

4.3– Applications of Mixture and Composition of MH algorithms

- Consider the target distribution $\pi(\theta_1, \theta_2)$.

- We use two MH kernels to sample from this distribution,

  - the kernel K_1 updates θ_1 and keeps θ_2 fixed whereas
  - the kernel K_2 updates θ_2 and keeps θ_1 fixed.

- We then combine these kernels through mixture or composition.

- The proposal $\overline{q}_1(\theta,\theta')$ associated to $K_1(\theta,\theta')$ is given by

$$\overline{q}_1(\theta,\theta') = \overline{q}_1((\theta_1,\theta_2),(\theta_1',\theta_2')) = q_1((\theta_1,\theta_2),\theta_1')\,\delta_{\theta_2}(\theta_2').$$

- The acceptance probability is given by $\alpha_1(\theta,\theta') = \min(1, r_1(\theta,\theta'))$ where

$$\begin{aligned} r_1(\theta,\theta') &= \frac{\pi(\theta')\,\overline{q}_1(\theta',\theta)}{\pi(\theta)\,\overline{q}_1(\theta,\theta')} = \frac{\pi(\theta_1',\theta_2')\,q_1((\theta_1',\theta_2'),\theta_1)\,\delta_{\theta_2'}(\theta_2)}{\pi(\theta_1,\theta_2)\,q_1((\theta_1,\theta_2),\theta_1')\,\delta_{\theta_2}(\theta_2')} \\ &= \frac{\pi(\theta_1',\theta_2)\,q_1((\theta_1',\theta_2),\theta_1)}{\pi(\theta_1,\theta_2)\,q_1((\theta_1,\theta_2),\theta_1')} \\ &= \frac{\pi(\theta_1'|\,\theta_2)\,q_1((\theta_1',\theta_2),\theta_1)}{\pi(\theta_1|\,\theta_2)\,q_1((\theta_1,\theta_2),\theta_1')}. \end{aligned}$$

- This move is also equivalent to an MH step of invariant distribution $\pi(\theta_1|\,\theta_2)$.

4.4– Description of transition kernels

- The proposal $\overline{q}_2(\theta, \theta')$ associated to $K_2(\theta, \theta')$ is given by

$$\overline{q}_2(\theta, \theta') = \overline{q}_2((\theta_1, \theta_2), (\theta_1', \theta_2')) = \delta_{\theta_1}(\theta_1') \, q_2((\theta_1, \theta_2), \theta_2').$$

- The acceptance probability is given by $\alpha_2(\theta, \theta') = \min(1, r_2(\theta, \theta'))$ where

$$\begin{aligned} r(\theta, \theta') &= \frac{\pi(\theta') \overline{q}_2(\theta', \theta)}{\pi(\theta) \overline{q}_2(\theta, \theta')} = \frac{\pi(\theta_1', \theta_2') \, \delta_{\theta_1'}(\theta_1) \, q_2((\theta_1', \theta_2'), \theta_2)}{\pi(\theta_1, \theta_2) \, \delta_{\theta_1}(\theta_1') \, q_2((\theta_1, \theta_2), \theta_2')} \\ &= \frac{\pi(\theta_1, \theta_2') \, q_2((\theta_1, \theta_2'), \theta_2)}{\pi(\theta_1, \theta_2) \, q_2((\theta_1, \theta_2), \theta_2')} \\ &= \frac{\pi(\theta_2' | \theta_1) \, q_2((\theta_1, \theta_2'), \theta_2)}{\pi(\theta_2 | \theta_1) \, q_2((\theta_1, \theta_2), \theta_2')}. \end{aligned}$$

- This move is also equivalent to an MH step of invariant distribution $\pi(\theta_2 | \theta_1)$.

4.5– Composition of MH algorithms

Assume we use a composition of these kernels, then the resulting algorithm proceeds as follows at iteration i.

MH step to update component 1

- Sample $\theta_1^* \sim q_1\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \cdot\right)$ and compute

$$
{}_1\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \left(\theta_1^*, \theta_2^{(i-1)}\right)\right) = \min\left(1, \frac{\pi\left(\theta_1^* \middle| \theta_2^{(i-1)}\right) q_1\left(\left(\theta_1^*, \theta_2^{(i-1)}\right), \theta_1^{(i-1)}\right)}{\pi\left(\theta_1^{(i-1)} \middle| \theta_2^{(i-1)}\right) q_1\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \theta_1^*\right)}\right)
$$

- With probability $\alpha_1\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \left(\theta_1^*, \theta_2^{(i-1)}\right)\right)$, set $\theta_1^{(i)} = \theta_1^*$ and otherwise $\theta_1^{(i)} = \theta_1^{(i-1)}$.

MH step to update component 2

- Sample $\theta_2^* \sim q_2\left(\left(\theta_1^{(i)}, \theta_2^{(i-1)}\right), \cdot\right)$ and compute

$$\alpha_2\left(\left(\theta_1^{(i)}, \theta_2^{(i-1)}\right), \left(\theta_1^{(i)}, \theta_2^{*}\right)\right) = \min\left(1, \frac{\pi\left(\theta_2^* \middle| \theta_1^{(i)}\right) q_2\left(\left(\theta_1^{(i)}, \theta_2^{*}\right), \theta_2^{(i-1)}\right)}{\pi\left(\theta_2^{(i-1)} \middle| \theta_1^{(i)}\right) q_2\left(\left(\theta_1^{(i)}, \theta_2^{(i-1)}\right), \theta_2^{*}\right)}\right)$$

- With probability $\alpha_2\left(\left(\theta_1^{(i)}, \theta_2^{(i-1)}\right), \left(\theta_1^{(i)}, \theta_1^{*}\right)\right)$, set $\theta_2^{(i)} = \theta_2^*$ otherwise $\theta_2^{(i)} = \theta_2^{(i-1)}$.

4.6– Mixture of MH algorithms

Assume we use a even mixture of these kernels, then the resulting algorithm proceeds as follows at iteration i.

- Sample the index of the component to update $J \sim U\{1,2\}$.

- Set $\theta_{-J}^{(i)} = \theta_{-J}^{(i-1)}$.

- Sample $\theta_J^* \sim q_J\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \cdot\right)$ and compute

$$\alpha_J\left(\left(\theta_1^{(i-1)}, \theta_2^{(i-1)}\right), \left(\theta_J^*, \theta_{-J}^{(i)}\right)\right) = \min\left(1, \frac{\pi\left(\theta_J^* \middle| \theta_{-J}^{(i)}\right) q_J\left(\left(\theta_J^*, \theta_{-J}^{(i)}\right), \theta_J^{(i-1)}\right)}{\pi\left(\theta_J^{(i-1)} \middle| \theta_{-J}^{(i)}\right) q_K\left(\left(\theta_J^{(i-1)}, \theta_{-J}^{(i)}\right), \theta_J^*\right)}\right).$$

- With probability $\alpha_J\left(\left(\theta_J^{(i-1)}, \theta_J^{(i-1)}\right), \left(\theta_J^*, \theta_{-J}^{(i)}\right)\right)$, set $\theta_J^{(i)} = \theta_J^*$ otherwise

$\theta_J^{(i)} = \theta_J^{(i-1)}$.

• It is clear that in such cases both K_1 and K_2 are NOT irreducible and aperiodic.

$\Rightarrow$ Each of them only update one component!!!!

• However, the composition and mixture of these kernels can be irreducible and aperiodic because then all the components are updated.

4.8– Back to the Gibbs sampler

- Consider now the case where

$$q_1\left(\left(\theta_1,\theta_2\right),\theta_1'\right)=\pi\left(\theta_1'\middle|\,\theta_2\right).$$

then

$$r_1\left(\theta,\theta'\right)=\frac{\pi\left(\theta_1'\middle|\,\theta_2\right)q_1\left(\left(\theta_1',\theta_2\right),\theta_1\right)}{\pi\left(\theta_1\middle|\,\theta_2\right)q_1\left(\left(\theta_1,\theta_2\right),\theta_1'\right)}=\frac{\pi\left(\theta_1'\middle|\,\theta_2\right)\pi\left(\theta_1\middle|\,\theta_2\right)}{\pi\left(\theta_1\middle|\,\theta_2\right)\pi\left(\theta_1'\middle|\,\theta_2\right)}=1$$

- Similarly if $q_2\left(\left(\theta_1,\theta_2\right),\theta_2'\right)=\pi\left(\theta_2'\middle|\,\theta_1\right)$ then $r_2\left(\theta,\theta'\right)=1$.

- If you take for proposal distributions in the MH kernels the full conditional distributions then you have the Gibbs sampler!

4.9– General hybrid algorithm

• Generally speaking, to sample from $\pi(\theta)$ where $\theta = (\theta_1, ..., \theta_p)$, we can use the following algorithm at iteration i.

• Iteration i; $i \geq 1$:

For $k = 1 : p$

- Sample $\theta_k^{(i)}$ using an MH step of proposal distribution

$q_k\left(\left(\theta_{-k}^{(i)}, \theta_k^{(i-1)}\right), \theta_k'\right)$ and target $\pi\left(\theta_k \middle| \theta_{-k}^{(i)}\right)$.

where $\theta_{-k}^{(i)} = \left(\theta_1^{(i)}, ..., \theta_{k-1}^{(i)}, \theta_{k+1}^{(i-1)}, ..., \theta_p^{(i-1)}\right)$.

4.9– General hybrid algorithm

- If we have $q_k\left(\theta_{1:p}, \theta'_k\right) = \pi\left(\theta'_k \middle| \theta_{-k}\right)$ then we are back to the Gibbs sampler.

- We can update some parameters according to $\pi\left(\theta'_k \middle| \theta_{-k}\right)$ (and the move is automatically accepted) and others according to different proposals.

- **Example**: Assume we have $\pi\left(\theta_1, \theta_2\right)$ where it is easy to sample from $\pi\left(\theta_1 \middle| \theta_2\right)$ and then use an MH step of invariant distribution $\pi\left(\theta_2 \middle| \theta_1\right)$.

At iteration i.

- Sample $\theta_1^{(i)} \sim \pi\left(\theta_1 \middle| \theta_2^{(i-1)}\right)$.

- Sample $\theta_2^{(i)}$ using one MH step of proposal distribution $q_2\left(\left(\theta_1^{(i)}, \theta_2^{(i-1)}\right), \theta_2\right)$ and target $\pi\left(\theta_2 \middle| \theta_1^{(i)}\right)$.

Remark: There is NO NEED to run the MH algorithm multiple steps to ensure that $\theta_2^{(i)} \sim \pi\left(\theta_2 \middle| \theta_2^{(i-1)}\right)$.

4.10– Alternative acceptance probabilities

- The standard MH algorithm uses the acceptance probability

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta') q(\theta', \theta)}{\pi(\theta) q(\theta, \theta')}\right).$$

- This is not necessary and one can also use any function

$$\alpha(\theta, \theta') = \frac{\delta(\theta, \theta')}{\pi(\theta) q(\theta, \theta')}$$

which is such that

$$\delta(\theta, \theta') = \delta(\theta', \theta) \text{ and } 0 \leq \alpha(\theta, \theta') \leq 1$$

- Example (Baker, 1965):

$$\alpha(\theta, \theta') = \frac{\pi(\theta') q(\theta', \theta)}{\pi(\theta') q(\theta', \theta) + \pi(\theta) q(\theta, \theta')}.$$

- Indeed one can check that
$$K(\theta,\theta') = \alpha(\theta,\theta')\,q(\theta,\theta') + \left(1 - \int \alpha(\theta,u)\,q(\theta,u)\,du\right)\delta_{\theta}(\theta')$$
is π-reversible.

- We have

$$\begin{aligned}
\pi(\theta)\,\alpha(\theta,\theta')\,q(\theta,\theta') &= \pi(\theta)\,\frac{\delta(\theta,\theta')}{\pi(\theta)\,q(\theta,\theta')}q(\theta,\theta') \\
&= \delta(\theta,\theta') \\
&= \delta(\theta',\theta) \\
&= \pi(\theta')\,\alpha(\theta',\theta)\,q(\theta',\theta)\,.
\end{aligned}$$

- The MH acceptance is favoured as it increases the acceptance probability.

4.11– Discussion

- In practice, we divide the parameter space $\theta = (\theta_1, ..., \theta_p)$.

- We update each parameter θ_k according to an MH step of propsal distribution $q_k(\theta_{1:p}, \theta'_k) = q_k((\theta_{-k}, \theta_k), \theta'_k)$ and invariant distribution $\pi(\theta_k | \theta_{-k})$.

- You are now equipped to fit advanced statistical models...