CPSC 535
Trans-dimensional MCMC

AD

March 2007

- Bayesian Model Selection

- Bayesian Model Selection
- Metropolis-Hastings on a General State-Space

- Bayesian Model Selection
- Metropolis-Hastings on a General State-Space
- Trans-dimensional Markov chain Monte Carlo.

- Most Bayesian models discussed until now: prior $p(\theta)$ and likelihood $p(y|\theta)$. Using MCMC, we sample from

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{\int p(\theta)\, p(y|\theta)\, d\theta}.$$

- Most Bayesian models discussed until now: prior $p(\theta)$ and likelihood $p(y|\theta)$. Using MCMC, we sample from

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{\int p(\theta)\, p(y|\theta)\, d\theta}.$$

- We discuss several examples where the model under study is fully specified.

- Most Bayesian models discussed until now: prior $p(\theta)$ and likelihood $p(y|\theta)$. Using MCMC, we sample from

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{\int p(\theta)\, p(y|\theta)\, d\theta}.$$

- We discuss several examples where the model under study is fully specified.
- In practice, we might have a collection of candidate models. This class of problems include cases where “the number of unknowns is something you don’t know” (Green, 1995).

Bayesian Model Selection

- Assume we have a (countable) set $\mathcal{K}$ of candidate models then an associated Bayesian model is such that

Bayesian Model Selection

- Assume we have a (countable) set $\mathcal{K}$ of candidate models then an associated Bayesian model is such that
  - k denotes the model and has a prior probability $p(k)$.

Bayesian Model Selection

- Assume we have a (countable) set $\mathcal{K}$ of candidate models then an associated Bayesian model is such that
  - k denotes the model and has a prior probability $p(k)$.
  - $\theta_k \in \Theta_k$ is the unknown parameter associated to model k of prior $p(\theta_k | k)$.

Bayesian Model Selection

- Assume we have a (countable) set $\mathcal{K}$ of candidate models then an associated Bayesian model is such that
  - k denotes the model and has a prior probability $p(k)$.
  - $\theta_k \in \Theta_k$ is the unknown parameter associated to model k of prior $p(\theta_k | k)$.
  - The likelihood is $p(y | k, \theta_k)$.

Bayesian Model Selection

- Assume we have a (countable) set $\mathcal{K}$ of candidate models then an associated Bayesian model is such that
  - k denotes the model and has a prior probability $p(k)$.
  - $\theta_k \in \Theta_k$ is the unknown parameter associated to model k of prior $p(\theta_k | k)$.
  - The likelihood is $p(y | k, \theta_k)$.
- You can think of it as a "standard" Bayesian model of parameter $(k, \theta_k) \in \cup_{i \in \mathcal{K}} (\{i\} \times \Theta_i)$.

- The Bayes' rule gives the posterior

$$p\left(k,\theta_k \middle| y\right)=\frac{p\left(k\right)p\left(\theta_k \middle| k\right)p\left(y \middle| k,\theta_k\right)}{\sum_{i\in\mathcal{K}}\int_{\Theta_i}p\left(i\right)p\left(\theta_i \middle| i\right)p\left(y \middle| i,\theta_i\right)d\theta_i}$$

defined on $\cup_{i\in\mathcal{K}}\left(\{i\}\times\Theta_i\right)$.

- The Bayes' rule gives the posterior

$$p(k, \theta_k | y) = \frac{p(k)\, p(\theta_k | k)\, p(y | k, \theta_k)}{\sum_{i \in \mathcal{K}} \int_{\Theta_i} p(i)\, p(\theta_i | i)\, p(y | i, \theta_i)\, d\theta_i}$$

  defined on $\cup_{i \in \mathcal{K}} (\{i\} \times \Theta_i)$.
- From this posterior, we can compute

$$p(k | y) \text{ and } \frac{p(y | k)}{p(y | j)} = \frac{p(k | y)}{p(j | y)} \frac{p(j)}{p(k)}$$

  or performing Bayesian model averaging

$$p(y' | y) = \sum_{i \in \mathcal{K}} \int_{\Theta_i} p(y' | i, \theta_i)\, p(i, \theta_i | y)\, d\theta_i$$

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$$

and we have $\theta_k = \left(a_{1:k}, \sigma^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$.

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$$

and we have $\theta_k = \left(a_{1:k}, \sigma^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$.

- We need to defined a prior $p(k, \theta_k) = p(k)\, p(\theta_k | k)$, say

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } k \in \mathcal{K}, \\ p(\theta_k | k) &= \mathcal{N}\left(a_{1:k}; 0, \sigma^2 \delta^2 I_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$$

  and we have $\theta_k = \left(a_{1:k}, \sigma^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$.
- We need to defined a prior $p(k, \theta_k) = p(k)\, p(\theta_k | k)$, say

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } k \in \mathcal{K}, \\ p(\theta_k | k) &= \mathcal{N}\left(a_{1:k}; 0, \sigma^2 \delta^2 I_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- One should be careful, the parameters denoted similarly can have a different "meaning" so that computing say $p\left(\sigma^2 \middle| y\right)$ does not mean much.

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$$

  and we have $\theta_k = \left(a_{1:k}, \sigma^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$.
- We need to defined a prior $p(k, \theta_k) = p(k)\, p(\theta_k | k)$, say

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } k \in \mathcal{K}, \\ p(\theta_k | k) &= \mathcal{N}\left(a_{1:k}; 0, \sigma^2 \delta^2 I_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- One should be careful, the parameters denoted similarly can have a different "meaning" so that computing say $p\left(\sigma^2 \middle| y\right)$ does not mean much.
- Some authors favour a more precise notation $\theta_k = \left(a_{k,1:k}, \sigma_k^2\right)$ but this can be cumbersome.

Example: Finite Mixture of Gaussians

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by a mixture of k Gaussians

$$Y_n \sim \sum_{i=1}^{k} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right).$$

and we have $\theta_k = \left(\pi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2\right) \in S_k \times \mathbb{R}^k \times \left(\mathbb{R}^+\right)^k$.

Example: Finite Mixture of Gaussians

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by a mixture of k Gaussians

$$Y_n \sim \sum_{i=1}^{k} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right).$$

  and we have $\theta_k = \left(\pi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2\right) \in S_k \times \mathbb{R}^k \times \left(\mathbb{R}^+\right)^k$.
- We need to defined a prior $p\left(k, \theta_k\right) = p\left(k\right) p\left(\theta_k \mid k\right)$, say

$$\begin{aligned} p\left(k\right) &= \frac{1}{k_{\max}}, \\ p\left(\theta_k \mid k\right) &= \mathcal{D}\left(\pi_{1:k}; 1, ..., 1\right) \prod_{i=1}^{k} \mathcal{N}\left(\mu_i; \alpha, \beta\right) \mathcal{IG}\left(\sigma_i^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

Example: Finite Mixture of Gaussians

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by a mixture of k Gaussians

$$Y_n \sim \sum_{i=1}^{k} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right).$$

  and we have $\theta_k = \left(\pi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2\right) \in S_k \times \mathbb{R}^k \times \left(\mathbb{R}^+\right)^k$.
- We need to defined a prior $p\left(k, \theta_k\right) = p\left(k\right) p\left(\theta_k \mid k\right)$, say

$$\begin{aligned} p\left(k\right) &= \frac{1}{k_{\max}}, \\ p\left(\theta_k \mid k\right) &= \mathcal{D}\left(\pi_{1:k}; 1, ..., 1\right) \prod_{i=1}^{k} \mathcal{N}\left(\mu_i; \alpha, \beta\right) \mathcal{IG}\left(\sigma_i^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- Some authors favour a more precise notation $\theta_k = \left(\pi_{k,1:k}, \mu_{k,1:k}, \sigma_{k,1:k}^2\right)$.

Problem Statement

- As stated before, Bayesian model selection problems corresponds to the case where the parameter space is simply $\cup_{k\in\mathcal{K}}\left(\{k\}\times\Theta_k\right)$.

Problem Statement

- As stated before, Bayesian model selection problems corresponds to the case where the parameter space is simply $\cup_{k\in\mathcal{K}}\left(\{k\}\times\Theta_k\right)$.
- Can we define MCMC algorithms - i.e. Markov chain kernels with fixed invariant distribution $\pi\left(k,\theta_k\right)=\pi\left(k\right)\pi_k\left(\theta_k\right)$- ?

Problem Statement

- As stated before, Bayesian model selection problems corresponds to the case where the parameter space is simply $\cup_{k \in \mathcal{K}} \left(\{k\} \times \Theta_k \right)$.
- Can we define MCMC algorithms - i.e. Markov chain kernels with fixed invariant distribution $\pi\left(k, \theta_k\right) = \pi\left(k\right) \pi_k\left(\theta_k\right)$- ?
- We are going to present a generalization of MH after revisiting first the MH algorithm.

- We say that a measure $\gamma(dx)$ admits a density with respect to a measure $\lambda(dx)$ if for any (measurable) set $A \in B(\mathcal{X})$

$$\lambda(A) = 0 \Rightarrow \gamma(A) = 0$$

and we call

$$\frac{\gamma(dx)}{\lambda(dx)} = f(x)$$

the density of $\gamma(dx)$ with respect to $\lambda(dx)$.

- We say that a measure $\gamma(dx)$ admits a density with respect to a measure $\lambda(dx)$ if for any (measurable) set $A \in B(\mathcal{X})$

$$\lambda(A) = 0 \Rightarrow \gamma(A) = 0$$

and we call

$$\frac{\gamma(dx)}{\lambda(dx)} = f(x)$$

the density of $\gamma(dx)$ with respect to $\lambda(dx)$.

- In 99% of the applications in statistics $\lambda(dx)$ is the Lebesgue measure dx and we write

$$\frac{\gamma(dx)}{\lambda(dx)} = \frac{\gamma(dx)}{dx} = \gamma(x).$$

- The standard MH algorithm where $\mathcal{X} \subset \mathbb{R}^d$ corresponds to

$$K\left(x, dx'\right) = \alpha\left(x, x'\right) q\left(x, dx'\right) + \left(1 - \int \alpha\left(x, z\right) q\left(x, dz\right)\right) \delta_x\left(dx'\right)$$

where

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

- The standard MH algorithm where $\mathcal{X} \subset \mathbb{R}^d$ corresponds to

$$K\left(x, dx'\right) = \alpha\left(x, x'\right) q\left(x, dx'\right) + \left(1 - \int \alpha\left(x, z\right) q\left(x, dz\right)\right) \delta_x\left(dx'\right)$$

where

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

- You should think of

$$\frac{\pi\left(x'\right) q\left(x', x\right)}{\pi\left(x\right) q\left(x, x'\right)}$$

not as just a "number"!

- The acceptance ratio corresponds to a ratio of probability measures -importance weight- defined on the same spaces

$$\frac{\pi\left(dx'\right)q\left(x',dx\right)}{\pi\left(dx\right)q\left(x,dx'\right)}=\frac{\pi\left(x'\right)dx'q\left(x',x\right)dx}{\pi\left(x\right)dxq\left(x,x'\right)dx'}=\frac{\pi\left(x'\right)q\left(x',x\right)}{\pi\left(x\right)q\left(x,x'\right)}.$$

- The acceptance ratio corresponds to a ratio of probability measures -importance weight- defined on the same spaces

$$\frac{\pi\left(dx'\right)q\left(x',dx\right)}{\pi\left(dx\right)q\left(x,dx'\right)}=\frac{\pi\left(x'\right)dx'q\left(x',x\right)dx}{\pi\left(x\right)dxq\left(x,x'\right)dx'}=\frac{\pi\left(x'\right)q\left(x',x\right)}{\pi\left(x\right)q\left(x,x'\right)}.$$

- You can only compared points defined on the same joint space. If you have $x=\left(x_1,x_2\right)$ and $\pi_1\left(dx_1\right)=\pi_1\left(x_1\right)dx_1$, $\pi_2\left(dx_1,dx_2\right)=\pi_2\left(x_1,x_2\right)dx_1dx_2$, you can compute numerically

$$\frac{\pi_2\left(x_1,x_2\right)}{\pi_1\left(x_1\right)}$$

but it means *nothing* as the measures π_1 and π_2 are not defined on the same space. You CANNOT compare a surface to a volume!

- In the general case where $\mathcal{X}$ is a union of subspaces of different dimensions, you might want to move from $x \in \mathbb{R}^d$ to $x' \in \mathbb{R}^{d'}$.

- In the general case where $\mathcal{X}$ is a union of subspaces of different dimensions, you might want to move from $x \in \mathbb{R}^d$ to $x' \in \mathbb{R}^{d'}$.
- To construct this move, you can use $u \in \mathbb{R}^r$ and $u' \in \mathbb{R}^{r'}$ and a *one-to-one differentiable* mapping h:$\mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}^{d'} \times \mathbb{R}^{r'}$

$$\left(x', u'\right) = h\left(x, u\right) \text{ where } u \sim g$$

and

$$\left(x, u\right) = h^{-1}\left(x', u'\right) \text{ where } u' \sim g'.$$

- In the general case where $\mathcal{X}$ is a union of subspaces of different dimensions, you might want to move from $x \in \mathbb{R}^d$ to $x' \in \mathbb{R}^{d'}$.
- To construct this move, you can use $u \in \mathbb{R}^r$ and $u' \in \mathbb{R}^{r'}$ and a *one-to-one differentiable* mapping $h: \mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}^{d'} \times \mathbb{R}^{r'}$

$$\left(x', u'\right) = h\left(x, u\right) \text{ where } u \sim g$$

  and

$$\left(x, u\right) = h^{-1}\left(x', u'\right) \text{ where } u' \sim g'.$$

- We need $d + r = d' + r'$ and typically, if $d < d'$, then $r' = 0$ and $r = d' - d$, that is in most case the variable u' is not introduced.

- We can rewrite formally

$$\pi\left(dx\right) q\left(x, \left(dx', du'\right)\right) = \pi\left(x\right) g\left(u\right) dxdu$$

and

$$\pi\left(dx'\right) q\left(x', \left(dx, du\right)\right) = \pi\left(x'\right) g'\left(u'\right) dx'du'.$$

- We can rewrite formally

$$\pi(dx)\, q\left(x, \left(dx', du'\right)\right) = \pi(x)\, g(u)\, dxdu$$

and

$$\pi\left(dx'\right) q\left(x', (dx, du)\right) = \pi\left(x'\right) g'\left(u'\right) dx'du'.$$

- An acceptance ratio ensuring $\pi-$reversibility of this trans-dimensional move is given by

$$\frac{\pi\left(dx'\right) q\left(x', (dx, du)\right)}{\pi(dx)\, q\left(x, \left(dx', du'\right)\right)} = \frac{\pi\left(x'\right) g'\left(u'\right)}{\pi\left(x'\right) g(u)} \left| \frac{\partial\left(x', u'\right)}{\partial(x, u)} \right|.$$

- We can rewrite formally

$$\pi(dx)\, q\left(x, \left(dx', du'\right)\right) = \pi(x)\, g(u)\, dxdu$$

and

$$\pi\left(dx'\right) q\left(x', (dx, du)\right) = \pi\left(x'\right) g'\left(u'\right) dx' du'.$$

- An acceptance ratio ensuring $\pi-$reversibility of this trans-dimensional move is given by

$$\frac{\pi\left(dx'\right) q\left(x', (dx, du)\right)}{\pi(dx)\, q\left(x, (dx', du')\right)} = \frac{\pi\left(x'\right) g'\left(u'\right)}{\pi\left(x'\right) g(u)} \left|\frac{\partial\left(x', u'\right)}{\partial(x, u)}\right|.$$

- In this respect, the RJMCMC is an extension of standard MH as you need to introduce auxiliary variables u and u'.

- Assume we have a distribution defined on $\{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R} \times \mathbb{R}$. We want to propose some moves to go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$.

- Assume we have a distribution defined on $\{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R} \times \mathbb{R}$. We want to propose some moves to go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$.
- We can propose $u \sim g \in \mathbb{R}$ and set

$$(\theta_1, \theta_2) = h(\theta, u) = (\theta, u),$$

i.e. we do not need to introduce a variable u'. Its inverse is given by

$$(\theta, u) = h^{-1}(\theta_1, \theta_2) = (\theta_1, \theta_2).$$

- Assume we have a distribution defined on $\{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R} \times \mathbb{R}$. We want to propose some moves to go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$.
- We can propose $u \sim g \in \mathbb{R}$ and set

$$(\theta_1, \theta_2) = h(\theta, u) = (\theta, u),$$

i.e. we do not need to introduce a variable u'. Its inverse is given by

$$(\theta, u) = h^{-1}(\theta_1, \theta_2) = (\theta_1, \theta_2).$$

- The acceptance probability for this "birth" move is given by

$$\begin{aligned} & \min\left(1, \frac{\pi(2, \theta_1, \theta_2)}{\pi(1, \theta)} \frac{1}{g(u)} \left| \frac{\partial(\theta_1, \theta_2)}{\partial(\theta, u)} \right| \right) \\ = \; & \min\left(1, \frac{\pi(2, \theta_1, \theta_2)}{\pi(1, \theta_1) g(\theta_2)}\right). \end{aligned}$$

- The acceptance probability for the associated "death move" is

$$\min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta_1,\theta_2)} g(u) \left| \frac{\partial(\theta,u)}{\partial(\theta_1,\theta_2)} \right| \right) = \min\left(1, \frac{\pi(1,\theta) g(u)}{\pi(2,\theta,u)}\right)$$

- The acceptance probability for the associated “death move” is

$$\min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta_1,\theta_2)} g(u) \left| \frac{\partial(\theta,u)}{\partial(\theta_1,\theta_2)} \right| \right) = \min\left(1, \frac{\pi(1,\theta)\, g(u)}{\pi(2,\theta,u)}\right)$$

- Once the birth move is defined then the death move follows automatically. In the death move, we do not simulate from g but its expression still appear in the acceptance probability.

- To simplify notation -as in Green (1995) & Robert (2004)-, we don't emphasize that actually we can have the proposal g which is a function of the current point θ but it is possible!

- To simplify notation -as in Green (1995) & Robert (2004)-, we don't emphasize that actually we can have the proposal g which is a function of the current point θ but it is possible!
- We can propose $u \sim g(\cdot | \theta) \in \mathbb{R}$ and set

$$(\theta_1, \theta_2) = h(\theta, u) = (\theta, u).$$

Its inverse is given by

$$(\theta, u) = h^{-1}(\theta_1, \theta_2) = (\theta_1, \theta_2).$$

- To simplify notation -as in Green (1995) & Robert (2004)-, we don't emphasize that actually we can have the proposal g which is a function of the current point θ but it is possible!
- We can propose $u \sim g\left(\cdot|\,\theta\right) \in \mathbb{R}$ and set

$$(\theta_1, \theta_2) = h(\theta, u) = (\theta, u).$$

Its inverse is given by

$$(\theta, u) = h^{-1}(\theta_1, \theta_2) = (\theta_1, \theta_2).$$

- The acceptance probability for this "birth" move is given by

$$\begin{aligned} & \min\left(1, \frac{\pi(2, \theta_1, \theta_2)}{\pi(1, \theta)} \frac{1}{g(u|\,\theta)} \left|\frac{\partial(\theta_1, \theta_2)}{\partial(\theta, u)}\right|\right) \\ = & \min\left(1, \frac{\pi(2, \theta_1, \theta_2)}{\pi(1, \theta_1)\, g(\theta_2|\,\theta_1)}\right). \end{aligned}$$

- The acceptance probability for the associated "death move" is

$$\begin{aligned} & \min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta_1,\theta_2)} g(u|\theta) \left| \frac{\partial(\theta,u)}{\partial(\theta_1,\theta_2)} \right| \right) \\ = \; & \min\left(1, \frac{\pi(1,\theta)\, g(u|\theta)}{\pi(2,\theta,u)}\right) \end{aligned}$$

- The acceptance probability for the associated "death move" is

$$\begin{aligned} & \min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta_1,\theta_2)} g(u|\theta)\left|\frac{\partial(\theta,u)}{\partial(\theta_1,\theta_2)}\right|\right) \\ = & \min\left(1, \frac{\pi(1,\theta)\, g(u|\theta)}{\pi(2,\theta,u)}\right) \end{aligned}$$

- Once the birth move is defined then the death move follows automatically. In the death move, we do not simulate from g but its expression still appears in the acceptance probability.

- The acceptance probability for the associated "death move" is

$$
\begin{aligned}
& \min\left(1, \frac{\pi\left(1, \theta\right)}{\pi\left(2, \theta_1, \theta_2\right)} g\left(u | \theta\right) \left| \frac{\partial\left(\theta, u\right)}{\partial\left(\theta_1, \theta_2\right)} \right| \right) \\
= & \min\left(1, \frac{\pi\left(1, \theta\right) g\left(u | \theta\right)}{\pi\left(2, \theta, u\right)}\right)
\end{aligned}
$$

- Once the birth move is defined then the death move follows automatically. In the death move, we do not simulate from g but its expression still appears in the acceptance probability.
- Clearly if we have $g\left(\theta_2 | \theta_1\right) = \pi\left(\theta_2 | 2, \theta_1\right)$ then the expressions simplify

$$
\begin{aligned}
\min\left(1, \frac{\pi\left(2, \theta_1, \theta_2\right)}{\pi\left(1, \theta_1\right) g\left(\theta_2 | \theta_1\right)}\right) &= \min\left(1, \frac{\pi\left(2, \theta_1\right)}{\pi\left(1, \theta_1\right)}\right), \\
\min\left(1, \frac{\pi\left(1, \theta\right) g\left(u | \theta\right)}{\pi\left(2, \theta, u\right)}\right) &= \min\left(1, \frac{\pi\left(1, \theta\right)}{\pi\left(2, \theta\right)}\right).
\end{aligned}
$$

- Assume we have a distribution defined on $\{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R} \times \mathbb{R}$. We want to propose some moves to go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$.

- Assume we have a distribution defined on $\{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R} \times \mathbb{R}$. We want to propose some moves to go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$.
- We can propose $u \sim g \in \mathbb{R}$ and set

$$(\theta_1, \theta_2) = h(\theta, u) = (\theta - u, \theta + u).$$

Its inverse is given by

$$(\theta, u) = h^{-1}(\theta_1, \theta_2) = \left(\frac{\theta_1 + \theta_2}{2}, \frac{\theta_2 - \theta_1}{2}\right).$$

- Assume we have a distribution defined on $\{1\} \times \mathbb{R} \cup \{2\} \times \mathbb{R} \times \mathbb{R}$. We want to propose some moves to go from $(1, \theta)$ to $(2, \theta_1, \theta_2)$.
- We can propose $u \sim g \in \mathbb{R}$ and set

$$(\theta_1, \theta_2) = h(\theta, u) = (\theta - u, \theta + u).$$

Its inverse is given by

$$(\theta, u) = h^{-1}(\theta_1, \theta_2) = \left(\frac{\theta_1 + \theta_2}{2}, \frac{\theta_2 - \theta_1}{2}\right).$$

- The acceptance probability for this "split" move is given by

$$\begin{aligned} & \min\left(1, \frac{\pi(2, \theta_1, \theta_2)}{\pi(1, \theta)} \frac{1}{g(u)} \left|\frac{\partial(\theta_1, \theta_2)}{\partial(\theta, u)}\right|\right) \\ = & \min\left(1, \frac{\pi(2, \theta_1, \theta_2)}{\pi\left(1, \frac{\theta_1+\theta_2}{2}\right)} \frac{2}{g\left(\frac{\theta_2-\theta_1}{2}\right)}\right). \end{aligned}$$

- The acceptance probability for the associated “merge move” is

$$\begin{aligned} & \min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta_1,\theta_2)} g(u) \left| \frac{\partial(\theta,u)}{\partial(\theta_1,\theta_2)} \right| \right) \\ = & \min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta-u,\theta+u)} \frac{g(u)}{2} \right) \end{aligned}$$

- The acceptance probability for the associated "merge move" is

$$
\begin{aligned}
& \min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta_1,\theta_2)} g(u) \left| \frac{\partial(\theta,u)}{\partial(\theta_1,\theta_2)} \right| \right) \\
= \; & \min\left(1, \frac{\pi(1,\theta)}{\pi(2,\theta-u,\theta+u)} \frac{g(u)}{2} \right)
\end{aligned}
$$

- Once the split move is defined then the merge move follows automatically. In the merge move, we do not simulate from g but its expression still appear in the acceptance probability.

- In practice, the algorithm is based on a combination of moves to move from $x = (k, \theta_k)$ to $x' = (k', \theta_{k'})$ indexed by $i \in \mathcal{M}$ and in this case we just need to have

$$\begin{aligned} & \int_{(x,x') \in A \times B} \pi(dx)\, \alpha_i(x, x')\, q_i(x, dx') \\ = & \int_{(x,x') \in A \times B} \pi(dx')\, \alpha_i(x', x)\, q_i(x', dx) \end{aligned}$$

to ensure that the kernel $P(x, B)$ defined for $x \notin B$

$$P(x, B) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \int_B \alpha_i(x, x')\, q_i(x, dx')$$

is π-reversible.

- In practice, the algorithm is based on a combination of moves to move from $x = (k, \theta_k)$ to $x' = (k', \theta_{k'})$ indexed by $i \in \mathcal{M}$ and in this case we just need to have

$$\begin{aligned} & \int_{(x,x') \in A \times B} \pi(dx) \, \alpha_i(x, x') \, q_i(x, dx') \\ = & \int_{(x,x') \in A \times B} \pi(dx') \, \alpha_i(x', x) \, q_i(x', dx) \end{aligned}$$

to ensure that the kernel $P(x, B)$ defined for $x \notin B$

$$P(x, B) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \int_B \alpha_i(x, x') \, q_i(x, dx')$$

is π-reversible.

- In practice, we would like to have

$$P(x, B) = \sum_{i \in \mathcal{M}} \int_B j_i(x) \, \alpha_i(x, x') \, q_i(x, dx')$$

where $j_i(x)$ is the probability of selecting the move i once we are in x and $\sum_{i \in \mathcal{M}} j_i(x) = 1$.

- In this case reversibility is ensured if

$$\begin{aligned}&\int_{(x,x')\in A\times B}\pi\left(dx\right)j_i\left(x\right)\alpha_i\left(x,x'\right)q_i\left(x,dx'\right)\\&=\int_{(x,x')\in A\times B}\pi\left(dx'\right)j_i\left(x'\right)\alpha_i\left(x',x\right)q_i\left(x',dx\right)\end{aligned}$$

which is satisfied if

$$\alpha_i\left(x,x'\right)=\min\left(1,\frac{\pi\left(x'\right)j_i\left(x'\right)g_i'\left(u'\right)}{\pi\left(x\right)j_i\left(x\right)g_i\left(u\right)}\left|\frac{\partial\left(x',u'\right)}{\partial\left(x,u\right)}\right|\right).$$

- In this case reversibility is ensured if

$$\int_{(x,x')\in A\times B} \pi(dx)\, j_i(x)\, \alpha_i(x,x')\, q_i(x,dx') \\ = \int_{(x,x')\in A\times B} \pi(dx')\, j_i(x')\, \alpha_i(x',x)\, q_i(x',dx)$$

which is satisfied if

$$\alpha_i(x,x') = \min\left(1, \frac{\pi(x')\, j_i(x')\, g_i'(u')}{\pi(x)\, j_i(x)\, g_i(u)} \left|\frac{\partial(x',u')}{\partial(x,u)}\right|\right).$$

- In practice, we will only have a limited number of moves possible from each point x.

Reversible Jump MCMC Algorithm

- For each point $x = (k, \theta_k)$, we define a collection of potential moves selected randomly with probability $j_i(x)$ where $i \in \mathcal{M}$

Reversible Jump MCMC Algorithm

- For each point $x = (k, \theta_k)$, we define a collection of potential moves selected randomly with probability $j_i(x)$ where $i \in \mathcal{M}$
- To move from $x = (k, \theta_k)$ to $x' = (k', \theta_{k'})$, we build one (or several) deterministic differentiable and inversible mapping(s)

$$(\theta_{k'}, u_{k',k}) = T_{k,k'}(\theta_k, u_{k,k'})$$

where $u_{k,k'} \sim g_{k,k'}$ and $u_{k',k} \sim g_{k',k}$ and we accept the move with proba

$$\min\left(1, \frac{\pi(k', \theta_{k'})\, j_i(k', \theta_{k'})\, g_{k',k}(u_{k',k})}{\pi(k, \theta_k)\, j_i(k, \theta_k)\, g_{k,k'}(u_{k,k'})} \left|\frac{\partial T_{k,k'}(\theta_k, u_{k,k'})}{\partial(\theta_k, u_{k,k'})}\right|\right).$$

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \sim \mathcal{N}(0,1)$$

and we have $\theta_k = \left(a_{k,1:k}, \sigma_k^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$ where

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } k \in \mathcal{K}, \\ p(\theta_k | k) &= \mathcal{N}\left(a_{k,1:k}; 0, \sigma_k^2 \delta^2 I_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \sim \mathcal{N}(0,1)$$

and we have $\theta_k = \left(a_{k,1:k}, \sigma_k^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$ where

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } k \in \mathcal{K}, \\ p(\theta_k | k) &= \mathcal{N}\left(a_{k,1:k}; 0, \sigma_k^2 \delta^2 I_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- For sake of simplicity, we assume here that the initial conditions $y_{1-k_{\max}:0} = (0, ..., 0)$ are known and we want to sample from

$$p(\theta_k, k | y_{1:T}).$$

Example: Autoregression

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by an AR of order k

$$Y_n = \sum_{i=1}^{k} a_i Y_{n-i} + \sigma V_n \text{ where } V_n \sim \mathcal{N}(0,1)$$

and we have $\theta_k = \left(a_{k,1:k}, \sigma_k^2\right) \in \mathbb{R}^k \times \mathbb{R}^+$ where

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } k \in \mathcal{K}, \\ p(\theta_k | k) &= \mathcal{N}\left(a_{k,1:k}; 0, \sigma_k^2 \delta^2 I_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- For sake of simplicity, we assume here that the initial conditions $y_{1-k_{\max}:0} = (0, ..., 0)$ are known and we want to sample from

$$p(\theta_k, k | y_{1:T}).$$

- Note that this is not very clever as $p(k | y_{1:T})$ is known up to a normalizing constant!

- We propose the following moves. If we have $\left(k, a_{1:k}, \sigma_k^2\right)$ then with probability b_k we propose a birth move if $k \leq k_{\max}$, with proba u_k we propose an update move and with proba $d_k = 1 - b_k - u_k$ we propose a death move.

- We propose the following moves. If we have $\left(k, a_{1:k}, \sigma_k^2\right)$ then with probability b_k we propose a birth move if $k \leq k_{\max}$, with proba u_k we propose an update move and with proba $d_k = 1 - b_k - u_k$ we propose a death move.
- We have $d_1 = 0$ and $b_{k\max} = 0$.

- We propose the following moves. If we have $(k, a_{1:k}, \sigma_k^2)$ then with probability b_k we propose a birth move if $k \leq k_{\max}$, with proba u_k we propose an update move and with proba $d_k = 1 - b_k - u_k$ we propose a death move.
- We have $d_1 = 0$ and $b_{k\max} = 0$.
- The *update move* can simply done in a Gibbs step as

$$p\left(\theta_k \middle| y_{1:T}, k\right) = \mathcal{N}\left(a_{k,1:k}; m_k, \sigma^2 \Sigma_k\right) \mathcal{IG}\left(\sigma^2; \frac{\nu_k}{2}, \frac{\gamma_k}{2}\right)$$

- *Birth move*: We propose to move from k to $k+1$

$$\left(a_{k+1,1:k}, a_{k+1,k+1}, \sigma^2_{k+1}\right) = \left(a_{k,1:k}, u, \sigma^2_k\right) \text{ where } u \sim g_{k,k+1}$$

and the acceptance probability is

$$\min\left(1, \frac{p\left(a_{k,1:k}, u, \sigma^2_k, k+1 \middle| y_{1:T}\right) d_{k+1}}{p\left(a_{k,1:k}, \sigma^2_k, k \middle| y_{1:T}\right) b_k g_{k,k+1}(u)}\right).$$

- *Birth move*: We propose to move from k to $k+1$

$$\left(a_{k+1,1:k}, a_{k+1,k+1}, \sigma_{k+1}^2\right) = \left(a_{k,1:k}, u, \sigma_k^2\right) \text{ where } u \sim g_{k,k+1}$$

and the acceptance probability is

$$\min\left(1, \frac{p\left(a_{k,1:k}, u, \sigma_k^2, k+1 \middle| y_{1:T}\right) d_{k+1}}{p\left(a_{k,1:k}, \sigma_k^2, k \middle| y_{1:T}\right) b_k g_{k,k+1}(u)}\right).$$

- *Death move*: We propose to move from k to $k-1$

$$\left(a_{k-1,1:k-1}, u, \sigma_{k-1}^2\right) = \left(a_{k,1:k-1}, a_{k,k}, \sigma_k^2\right)$$

and the acceptance probability is

$$\min\left(1, \frac{p\left(a_{k,1:k-1}, \sigma_k^2, k-1 \middle| y_{1:T}\right) b_{k-1} g_{k-1,k}\left(a_{k,k}\right)}{p\left(a_{k,1:k}, \sigma_k^2, k \middle| y_{1:T}\right) d_k}\right).$$

- The performance are obviously very dependent on the selection of the proposal distribution. We select whenever possible the full conditional distribution, i.e. we have
$u = a_{k+1,k+1} \sim p\left(a_{k+1,k+1} \middle| y_{1:T}, a_{k,1:k}, \sigma_k^2, k+1 \right)$ and

$$
\begin{aligned}
& \min\left(1, \frac{p\left(a_{k,1:k}, u, \sigma_k^2, k+1 \middle| y_{1:T} \right) d_{k+1}}{p\left(a_{k,1:k}, \sigma_k^2, k \middle| y_{1:T} \right) b_k p\left(u \middle| y_{1:T}, a_{k,1:k}, \sigma_k^2, k+1 \right)} \right) \\
= \; & \min\left(1, \frac{p\left(a_{k,1:k}, \sigma_k^2, k+1 \middle| y_{1:T} \right) d_{k+1}}{p\left(a_{k,1:k}, \sigma_k^2, k \middle| y_{1:T} \right) b_k} \right).
\end{aligned}
$$

- The performance are obviously very dependent on the selection of the proposal distribution. We select whenever possible the full conditional distribution, i.e. we have
$u = a_{k+1,k+1} \sim p\left(a_{k+1,k+1} \middle| y_{1:T}, a_{k,1:k}, \sigma_k^2, k+1 \right)$ and

$$\begin{aligned} & \min\left(1, \frac{p\left(a_{k,1:k}, u, \sigma_k^2, k+1 \middle| y_{1:T} \right) d_{k+1}}{p\left(a_{k,1:k}, \sigma_k^2, k \middle| y_{1:T} \right) b_k p\left(u \middle| y_{1:T}, a_{k,1:k}, \sigma_k^2, k+1 \right)}\right) \\ = & \min\left(1, \frac{p\left(a_{k,1:k}, \sigma_k^2, k+1 \middle| y_{1:T} \right) d_{k+1}}{p\left(a_{k,1:k}, \sigma_k^2, k \middle| y_{1:T} \right) b_k}\right). \end{aligned}$$

- In such cases, it is actually possible to reject a candidate before sampling it!

- We simulate 200 data with $k = 5$ and use 10,000 iterations of RJMCMC.

- We simulate 200 data with $k = 5$ and use 10,000 iterations of RJMCMC.
- The algorithm output is $\left(k^{(i)}, \theta_k^{(i)}\right) \sim p\left(\theta_k, k \middle| y\right)$ (asymptotically).

- We simulate 200 data with $k = 5$ and use 10,000 iterations of RJMCMC.
- The algorithm output is $\left(k^{(i)}, \theta_k^{(i)}\right) \sim p\left(\theta_k, k \mid y\right)$ (asymptotically).
- The histogram of $\left(k^{(i)}\right)$ yields an estimate of $p\left(k \mid y\right)$.

- We simulate 200 data with $k = 5$ and use 10,000 iterations of RJMCMC.
- The algorithm output is $\left(k^{(i)}, \theta_k^{(i)}\right) \sim p\left(\theta_k, k \mid y\right)$ (asymptotically).
- The histogram of $\left(k^{(i)}\right)$ yields an estimate of $p\left(k \mid y\right)$.
- Histograms of $\left(\theta_k^{(i)}\right)$ for which $k^{(i)} = k_0$ yields estimates of $p\left(\theta_{k_0} \mid y, k_0\right)$.

- We simulate 200 data with $k = 5$ and use 10,000 iterations of RJMCMC.
- The algorithm output is $\left(k^{(i)}, \theta_k^{(i)}\right) \sim p\left(\theta_k, k \mid y\right)$ (asymptotically).
- The histogram of $\left(k^{(i)}\right)$ yields an estimate of $p\left(k \mid y\right)$.
- Histograms of $\left(\theta_k^{(i)}\right)$ for which $k^{(i)} = k_0$ yields estimates of $p\left(\theta_{k_0} \mid y, k_0\right)$.
- The algorithm provides us with an estimate of $p\left(k \mid y\right)$ which matches analytical expressions.

Example: Finite Mixture of Gaussians

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by a mixture of k Gaussians

$$Y_n \sim \sum_{i=1}^{k} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right).$$

and we have $\theta_k = \left(\pi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2\right) \in S_k \times \mathbb{R}^k \times \left(\mathbb{R}^+\right)^k$.

Example: Finite Mixture of Gaussians

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by a mixture of k Gaussians

$$Y_n \sim \sum_{i=1}^{k} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right).$$

and we have $\theta_k = \left(\pi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2\right) \in S_k \times \mathbb{R}^k \times \left(\mathbb{R}^+\right)^k$.

- We need to defined a prior $p(k, \theta_k) = p(k)\, p(\theta_k | k)$, say

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } \in \mathcal{K} \\ p(\theta_k | k) &= \mathcal{D}(\pi_{k,1:k}; 1, ..., 1) \prod_{i=1}^{k} \mathcal{N}(\mu_{k,i}; \alpha, \beta)\, \mathcal{IG}\left(\sigma_{k,i}^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

Example: Finite Mixture of Gaussians

- The model $k \in \mathcal{K} = \{1, ..., k_{\max}\}$ is given by a mixture of k Gaussians

$$Y_n \sim \sum_{i=1}^{k} \pi_i \mathcal{N}\left(\mu_i, \sigma_i^2\right).$$

and we have $\theta_k = \left(\pi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2\right) \in S_k \times \mathbb{R}^k \times \left(\mathbb{R}^+\right)^k$.

- We need to defined a prior $p(k, \theta_k) = p(k)\, p(\theta_k | k)$, say

$$\begin{aligned} p(k) &= k_{\max}^{-1} \text{ for } \in \mathcal{K} \\ p(\theta_k | k) &= \mathcal{D}(\pi_{k,1:k}; 1, ..., 1) \prod_{i=1}^{k} \mathcal{N}(\mu_{k,i}; \alpha, \beta)\, \mathcal{IG}\left(\sigma_{k,i}^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2}\right). \end{aligned}$$

- Given T data, we are interested in $\pi(k, \theta_k | y_{1:T})$.

- When k is fixed, we will use Gibbs steps to sample from $\pi\left(\theta_k, z_{1:T}\middle| y_{1:T}, k\right)$ where $z_{1:T}$ are the discrete latent variables such that $\Pr\left(z_n = i\middle| k, \theta_k\right) = \pi_{k,i}$.

- When k is fixed, we will use Gibbs steps to sample from $\pi\left(\theta_k, z_{1:T} \middle| y_{1:T}, k\right)$ where $z_{1:T}$ are the discrete latent variables such that $\Pr\left(z_n = i \middle| k, \theta_k\right) = \pi_{k,i}$.
- To allow to move in the model space, we define a birth and death move.

- When k is fixed, we will use Gibbs steps to sample from $\pi\left(\theta_k, z_{1:T} \middle| y_{1:T}, k\right)$ where $z_{1:T}$ are the discrete latent variables such that $\Pr\left(z_n = i \middle| k, \theta_k\right) = \pi_{k,i}$.
- To allow to move in the model space, we define a birth and death move.
- The birth and death moves use as a target $\pi\left(\theta_k \middle| y_{1:T}, k\right)$ and not $\pi\left(\theta_k, z_{1:T} \middle| y_{1:T}, k\right) \Rightarrow$ Reduced dimensionality, easier to design moves.

- We propose a naive move to go from $k \longrightarrow k+1$ where $j \sim \mathcal{U}_{\{1,\ldots,k+1\}}$

$$\begin{aligned} \mu_{k+1,-j} &= \mu_{k,1:k},\ \sigma^2_{k+1,-j} = \sigma^2_{k,1:k}, \\ \pi_{k+1,-j} &= \left(1 - \pi_{k+1,j}\right) \pi_{k,-j}, \end{aligned}$$

where $\left(\pi_{k+1,j}, \mu_{k+1,j}, \sigma^2_{k+1,j}\right) \sim g_{k,k+1}$ (prior distribution in practice).

- We propose a naive move to go from $k \longrightarrow k+1$ where $j \sim \mathcal{U}_{\{1,\ldots,k+1\}}$

$$\begin{aligned} \mu_{k+1,-j} &= \mu_{k,1:k},\ \sigma^2_{k+1,-j} = \sigma^2_{k,1:k}, \\ \pi_{k+1,-j} &= (1-\pi_{k+1,j})\,\pi_{k,-j}, \end{aligned}$$

  where $\left(\pi_{k+1,j}, \mu_{k+1,j}, \sigma^2_{k+1,j}\right) \sim g_{k,k+1}$ (prior distribution in practice).
- The Jacobian of the transformation is $(1-\pi_{k+1,j})^{k-1}$ (only $k-1$ "true" variables for $\pi_{k,-j}$)

- We propose a naive move to go from $k \longrightarrow k+1$ where $j \sim \mathcal{U}_{\{1,\ldots,k+1\}}$

$$\begin{aligned} \mu_{k+1,-j} &= \mu_{k,1:k},\ \sigma^2_{k+1,-j} = \sigma^2_{k,1:k}, \\ \pi_{k+1,-j} &= (1-\pi_{k+1,j})\,\pi_{k,-j}, \end{aligned}$$

where $\left(\pi_{k+1,j}, \mu_{k+1,j}, \sigma^2_{k+1,j}\right) \sim g_{k,k+1}$ (prior distribution in practice).

- The Jacobian of the transformation is $(1-\pi_{k+1,j})^{k-1}$ (only $k-1$ "true" variables for $\pi_{k,-j}$)
- Now one has to be careful when considering the reverse death move. Assume the death move going from $k+1 \longrightarrow k$ by removing the component j.

- The acceptance probability of the birth move is given by $\min(1, A)$ where

$$A = \frac{\pi\left(k+1, \pi_{k+1,1:k+1}, \mu_{k+1,1:k+1}, \sigma^2_{k+1,1:k+1} \middle| y_{1:T}\right)}{\pi\left(k, \pi_{k,1:k}, \mu_{k,1:k}, \sigma^2_{k,1:k} \middle| y_{1:T}\right)} \times \frac{\left(d_{k+1,k}/(k+1)\right)\left(1-\pi_{k+1,j}\right)^{k-1}}{\left(b_{k,k+1}/(k+1)\right) g_{k,k+1}\left(\pi_{k+1,j}, \mu_{k+1,j}, \sigma_j^2\right)}.$$

- The acceptance probability of the birth move is given by $\min(1, A)$ where

$$A = \frac{\pi\left(k+1, \pi_{k+1,1:k+1}, \mu_{k+1,1:k+1}, \sigma^2_{k+1,1:k+1} \middle| y_{1:T}\right)}{\pi\left(k, \pi_{k,1:k}, \mu_{k,1:k}, \sigma^2_{k,1:k} \middle| y_{1:T}\right)} \times \frac{\left(d_{k+1,k} / (k+1)\right)\left(1-\pi_{k+1,j}\right)^{k-1}}{\left(b_{k,k+1} / (k+1)\right) g_{k,k+1}\left(\pi_{k+1,j}, \mu_{k+1,j}, \sigma_j^2\right)}.$$

- This move will work properly if the prior is not too diffuse. Otherwise the acceptance probability will be small.

- The acceptance probability of the birth move is given by $\min(1, A)$ where

$$A = \frac{\pi\left(k+1, \pi_{k+1,1:k+1}, \mu_{k+1,1:k+1}, \sigma^2_{k+1,1:k+1} \middle| y_{1:T}\right)}{\pi\left(k, \pi_{k,1:k}, \mu_{k,1:k}, \sigma^2_{k,1:k} \middle| y_{1:T}\right)} \times \frac{\left(d_{k+1,k}/\left(k+1\right)\right)\left(1-\pi_{k+1,j}\right)^{k-1}}{\left(b_{k,k+1}/\left(k+1\right)\right) g_{k,k+1}\left(\pi_{k+1,j}, \mu_{k+1,j}, \sigma^2_j\right)}.$$

- This move will work properly if the prior is not too diffuse. Otherwise the acceptance probability will be small.
- We have $(k+1)$ birth moves to move from $k \rightarrow k+1$ and $k+1$ associated death moves.

- To move from $k \rightarrow k+1$, one can also select a split move of the component $j \sim \mathcal{U}_{\{1,\ldots,k\}}$

$$
\begin{aligned}
\pi_{k+1,j} &= u_1 \pi_{k,j},\ \pi_{k+1,j+1} = (1-u_1)\,\pi_{k,j}, \\
\mu_{k+1,j} &= u_2 \mu_{k,j},\ \mu_{k+1,j+1} = \frac{\pi_{k,j} - \pi_{k+1,j} u_2}{\pi_{k,j} - \pi_{k+1,j}} \mu_{k,j}, \\
\sigma^2_{k+1,j} &= u_3 \sigma^2_{k,j},\ \sigma^2_{k+1,j+1} = \frac{\pi_{k,j} - \pi_{k+1,j} u_3}{\pi_{k,j} - \pi_{k+1,j}} \sigma^2_{k,j}
\end{aligned}
$$

with $u_1, u_2, u_3 \sim \mathcal{U}\,(0,1)$.

- The associated merge move is

$$
\begin{aligned}
\pi_{k,j} &= \pi_{k+1,j} + \pi_{k+1,j+1}, \\
\pi_{k,j}\mu_{k,j} &= \pi_{k+1,j}\mu_{k+1,j} + \pi_{k+1,j+1}\mu_{k+1,j+1}, \\
\pi_{k,j}\sigma^2_{k,j} &= \pi_{k+1,j}\sigma^2_{k+1,j} + \pi_{k+1,j+1}\sigma^2_{k+1,j+1}.
\end{aligned}
$$

- The associated merge move is

$$\begin{aligned}
\pi_{k,j} &= \pi_{k+1,j} + \pi_{k+1,j+1}, \\
\pi_{k,j}\mu_{k,j} &= \pi_{k+1,j}\mu_{k+1,j} + \pi_{k+1,j+1}\mu_{k+1,j+1}, \\
\pi_{k,j}\sigma^2_{k,j} &= \pi_{k+1,j}\sigma^2_{k+1,j} + \pi_{k+1,j+1}\sigma^2_{k+1,j+1}.
\end{aligned}$$

- The Jacobian of the transformation of the split is given by

$$\left| \frac{\partial\left(\pi_{k+1,1:k+1}, \mu_{k+1,1:k+1}, \sigma^2_{k+1,1:k+1}\right)}{\partial\left(\pi_{k,1:k}, \mu_{k,1:k}, \sigma^2_{k,1:k}, u_1, u_2, u_3\right)} \right| = \frac{\pi^3_{k,j}}{\left(1-u_1\right)^2} \left|\mu_{k,j}\right| \sigma^2_{k,j}.$$

- It follows that the acceptance probability of the split move with $j \sim \mathcal{U}_{\{1,\ldots,k\}}$ is $\min(1, A)$ where

$$
\begin{aligned}
A &= \frac{\pi\left(k+1, \pi_{k+1,1:k+1}, \mu_{k+1,1:k+1}, \sigma^2_{k+1,1:k+1} \,\middle|\, y_{1:T}\right)}{\pi\left(k, \pi_{k,1:k}, \mu_{k,1:k}, \sigma^2_{k,1:k} \,\middle|\, y_{1:T}\right)} \\
&\quad \times \frac{(m_{k+1,k}/k)}{(s_{k,k+1}/k)} \times \frac{\pi^3_{k,j}}{(1-u_1)^2} \left|\mu_{k,j}\right| \sigma^2_{k,j}.
\end{aligned}
$$

- It follows that the acceptance probability of the split move with $j \sim \mathcal{U}_{\{1,\ldots,k\}}$ is $\min(1, A)$ where

$$A = \frac{\pi\left(k+1, \pi_{k+1,1:k+1}, \mu_{k+1,1:k+1}, \sigma^2_{k+1,1:k+1} \middle| y_{1:T}\right)}{\pi\left(k, \pi_{k,1:k}, \mu_{k,1:k}, \sigma^2_{k,1:k} \middle| y_{1:T}\right)} \times \frac{(m_{k+1,k}/k)}{(s_{k,k+1}/k)} \times \frac{\pi^3_{k,j}}{(1-u_1)^2} \left|\mu_{k,j}\right| \sigma^2_{k,j}.$$

- You should think of the split move as a mixture of k split moves and you have k associated merge moves.

- We set $k_{\max} = 20$ and we select (rather) informative priors following Green & Richardson (1999). In practice, it is worth using a hierarchical prior.

- We set $k_{\max} = 20$ and we select (rather) informative priors following Green & Richardson (1999). In practice, it is worth using a hierarchical prior.
- We run the algorithm for over 1,000,000 iterations.

- We set $k_{\max} = 20$ and we select (rather) informative priors following Green & Richardson (1999). In practice, it is worth using a hierarchical prior.
- We run the algorithm for over 1,000,000 iterations.
- We set additional constraints on the mean $\mu_{k,1} < \mu_{k,2} < .... < \mu_{k,k}$.

- We set $k_{\max} = 20$ and we select (rather) informative priors following Green & Richardson (1999). In practice, it is worth using a hierarchical prior.
- We run the algorithm for over 1,000,000 iterations.
- We set additional constraints on the mean $\mu_{k,1} < \mu_{k,2} < .... < \mu_{k,k}$.
- The cumulative averages stabilize very quickly.

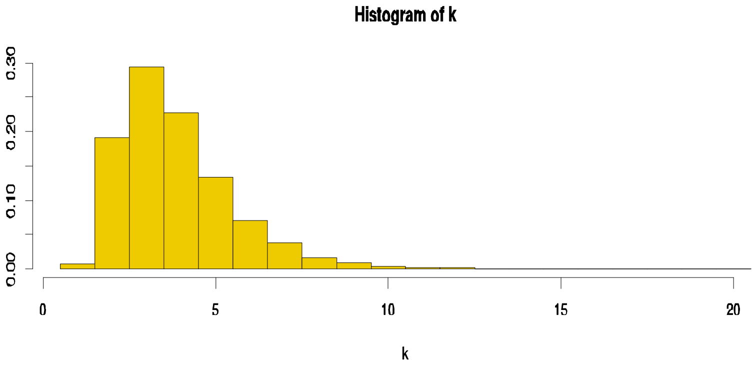

Figure: Estimation of the marginal posterior distribution $p(k|y_{1:T})$

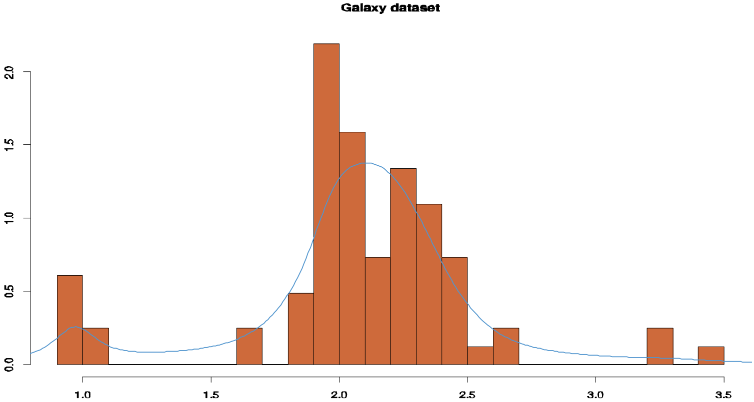

Figure: Estimation of $\mathbb{E}\left[f\left(y \middle| k, \theta_k \right) \middle| y_{1:T} \right]$

- Trans-dimensional MCMC allows us to implement numerically problems with Bayesian model uncertainty.

- Trans-dimensional MCMC allows us to implement numerically problems with Bayesian model uncertainty.
- Practical implementation is relatively easy, theory behind not so easy...

- Trans-dimensional MCMC allows us to implement numerically problems with Bayesian model uncertainty.
- Practical implementation is relatively easy, theory behind not so easy...
- Designing efficient trans-dimensional MCMC algorithms is still a research problem.