Reversible Jump MCMC

Nick Whiteley

02.11.2008

6.1 Bayesian multi-model inference

Multi-model inference: Idea

- So far we have assumed: We know what the true model is. $\rightsquigarrow$ all inference was done conditionally on this model being true.
- In reality there is often more than one plausible model.
- Examples:
    - Linear / quadratic regression
    - Choice of covariates in a regression model
    - Number of components in a mixture model

Bayesian Multi-model inference: Setup (1)

- Model $\mathcal{M}_k := \{f_k(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_k\}$.
- Hierarchical Bayesian model:
  - Prior distribution on the space of models (at least enumerable)
  $$\mathbb{P}(\mathcal{M}_k) = p_k \qquad \text{with } \sum_k p_k = 1.$$
  - Prior on the parameters in each model
  $$\boldsymbol{\theta}|\mathcal{M}_k \sim f_k^{\text{prior}}(\boldsymbol{\theta}).$$
  - Likelihood of observed data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is
  $$l_k(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\theta}) := \prod_{i=1}^{n} f_k(\mathbf{y}_i|\boldsymbol{\theta}).$$
- Posterior density of $\boldsymbol{\theta}$ in model $\mathcal{M}_k$ is
$$f_k^{\text{post}}(\boldsymbol{\theta}) = \frac{f_k^{\text{prior}}(\boldsymbol{\theta}) l_k(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\theta})}{\int_{\Theta_k} l_k^{\text{prior}}(\boldsymbol{\vartheta}) l_k(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}}$$

Bayesian Multi-model inference: Setup (2)

- Posterior probability that the data was generated by model $\mathcal{M}_k$

$$\mathbb{P}(\mathcal{M}_k|\mathbf{y}_1,\dots,\mathbf{y}_n) = \frac{p_k \int_{\Theta_k} f_k^{\text{prior}}(\boldsymbol{\theta}) l_k(\mathbf{y}_1,\dots,\mathbf{y}_n|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\sum_\kappa p_\kappa \int_{\Theta_\kappa} f_\kappa^{\text{prior}}(\boldsymbol{\theta}) l_\kappa(\mathbf{y}_1,\dots,\mathbf{y}_n|\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

- Comparison between two models $\mathcal{M}_{k_1}$ and $\mathcal{M}_{k_2}$ by the posterior odds

$$\frac{\mathbb{P}(\mathcal{M}_{k_1}|\mathbf{y}_1,\dots,\mathbf{y}_n)}{\mathbb{P}(\mathcal{M}_{k_2}|\mathbf{y}_1,\dots,\mathbf{y}_n)} = \frac{p_{k_1}}{p_{k_2}} \cdot \underbrace{\frac{\int_{\Theta_{k_1}} f_{k_1}^{\text{prior}}(\boldsymbol{\theta}) l_{k_1}(\mathbf{y}_1,\dots,\mathbf{y}_n|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{\Theta_{k_2}} f_{k_2}^{\text{prior}}(\boldsymbol{\theta}) l_{k_2}(\mathbf{y}_1,\dots,\mathbf{y}_n|\boldsymbol{\theta})\, d\boldsymbol{\theta}}}_{\text{“Bayes factor”}}.$$

- We can either ...
  - only consider the model with the highest posterior probability $\mathbb{P}(\mathcal{M}_k|\mathbf{y}_1,\dots,\mathbf{y}_n)$, or
  - perform *model averaging* using $\mathbb{P}(\mathcal{M}_k|\mathbf{y}_1,\dots,\mathbf{y}_n)$ as weights.

Bayesian Multi-model inference: Transdimensional simulation

Within model simulation Run a separate MCMC algorithm for each model.(not always possible due to the large (possibly infinite) number of candidate models / often not efficient)

Transdimensional simulation Sample directly from the joint posterior

$$f^{\text{post}}(k,\boldsymbol{\theta}) = \frac{p_k f_k^{\text{prior}}(\boldsymbol{\theta}) l_k(\mathbf{y}_1,\ldots,\mathbf{y}_n|\boldsymbol{\theta})}{\sum_\kappa p_\kappa \int_{\Theta_\kappa} l_\kappa^{\text{prior}}(\boldsymbol{\vartheta}) f_\kappa(\mathbf{y}_1,\ldots,\mathbf{y}_n|\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta}}$$

defined on *disjoint* union

$$\Theta := \bigcup_k \left(\{k\} \times \Theta_k\right).$$

(e.g. $\Theta_k := \mathbb{R}^k$ requires variable dimension simulation)

6.2 Another look at the M-H algorithm

Setup

Random walk Metropolis algorithm

Propose $\mathbf{X} := \mathbf{X}^{(t-1)} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim g$ (g symmetric distribution)

More generally, we will..

obtain $\mathbf{X}$ by some deterministic transformation of $(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)})$ with $\mathbf{U}^{(t-1)} \sim g_{1\to 2}$.

Details (1)

- Consider a transformation $\mathbf{T}_{1\to 2}$:

$$\mathbf{T}_{1\to 2} : (\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)}) \mapsto (\mathbf{X}, \mathbf{U}).$$

- Assume $\mathbf{T}_{1\to 2}$ is a diffeomorphism with inverse $\mathbf{T}_{2\to 1} = \mathbf{T}_{1\to 2}^{-1}$.
- Now consider an algorithm with the following proposal:
  - $1 \to 2$ with prob. $1/2$, draw $\mathbf{U}^{(t-1)} \sim g_1$ and set $(\mathbf{X}, \mathbf{U}) = \mathbf{T}_{1\to 2}(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)})$.
  - $2 \to 1$ with prob. $1/2$, draw $\mathbf{U}^{(t-1)} \sim g_2$ and set $(\mathbf{X}, \mathbf{U}) = \mathbf{T}_{2\to 1}(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)})$.
- What probability of acceptance gives detailed balance?

Details (2)

Detailed balance w.r.t. f can be achieved by setting the probability of acceptance for the move $1 \longrightarrow 2$ to

$$\alpha(\mathbf{x}|\mathbf{x}^{(t-1)}) = \min\left\{1, \frac{f(\mathbf{x})g_{2\rightarrow 1}(\mathbf{u})}{f(\mathbf{x}^{(t-1)})g_{1\rightarrow 2}(\mathbf{u}^{(t-1)})}\left|\frac{\partial \mathbf{T}_{1\rightarrow 2}(\mathbf{x}^{(t-1)}, \mathbf{u}^{(t-1)})}{\partial(\mathbf{x}^{(t-1)}, \mathbf{u}^{(t-1)})}\right|\right\}$$

Example 6.1 RW Metropolis in this framework

Consider the random walk Metropolis algorithm with

$$\mathbf{X} = \mathbf{X}^{(t-1)} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim g_{1\to 2},$$

In our new notation this can be seen as two (equivalent) moves

$$\begin{aligned}
(\mathbf{X}, \mathbf{U}) &= \mathbf{T}_{1\to 2}(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)}) = (\mathbf{X}^{(t-1)} + \mathbf{U}^{(t-1)}, \mathbf{U}^{(t-1)}), \\
&\qquad \mathbf{U}^{(t-1)} \sim g_{1\to 2} \\
(\mathbf{X}, \mathbf{U}) &= \mathbf{T}_{2\to 1}(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)}) = (\mathbf{X}^{(t-1)} - \mathbf{U}^{(t-1)}, \mathbf{U}^{(t-1)}), \\
&\qquad \mathbf{U}^{(t-1)} \sim g_{2\to 1} = g_{1\to 2}
\end{aligned}$$

We accept the newly proposed $\mathbf{X}$ with probability

$$\begin{aligned}
\alpha(\mathbf{X}|\mathbf{X}^{(t-1)}) &= \min\left\{1, \frac{f(\mathbf{X}) g_{2\to 1}(\mathbf{U})}{f(\mathbf{X}^{(t-1)}) g_{1\to 2}(\mathbf{U}^{(t-1)})} \left| \frac{\partial \mathbf{T}_{1\to 2}(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)})}{\partial(\mathbf{X}^{(t-1)}, \mathbf{U}^{(t-1)})} \right| \right\} \\
&= \min\left\{1, \frac{f(\mathbf{X})}{f(\mathbf{X}^{(t-1)})}\right\}.
\end{aligned}$$

6.3 The reversible jump algorithm

Idea of the reversible jump algorithm

- Algorithm from the previous slides does not require $\mathbf{X}^{(t-1)}$ and $\mathbf{X}^{(t)}$ to be of the same dimension.
- We only need $dim(\mathbf{X}^{(t-1)}) + dim(\mathbf{U}^{(t-1)}) = dim(\mathbf{X}) + dim(\mathbf{U})$.
- Idea: Use this framework to sample from the joint posterior

$$
\begin{aligned}
f^{\text{post}}(k, \boldsymbol{\theta}) &= \frac{p_k f_k^{\text{prior}}(\boldsymbol{\theta}) l_k(\mathbf{y}_1, \ldots, \mathbf{y}_n | \boldsymbol{\theta})}{\sum_\kappa p_\kappa \int_{\Theta_\kappa} f_\kappa^{\text{prior}}(\boldsymbol{\vartheta}) l_\kappa(\mathbf{y}_1, \ldots, \mathbf{y}_n | \boldsymbol{\vartheta}) \, d\boldsymbol{\vartheta}} \\
&\propto p_k f_k^{\text{prior}}(\boldsymbol{\theta}) l_k(\mathbf{y}_1, \ldots, \mathbf{y}_n | \boldsymbol{\theta})
\end{aligned}
$$

defined on

$$
\Theta := \bigcup_k \left(\{k\} \times \Theta_k \right)
$$

Green's reversible jump algorithm

Algorithm 6.1 (Reversible jump)

Starting with $k^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ iterate for $t = 1, 2, \ldots$

1. Select new model $\mathcal{M}_k$ with probability $\rho_{k^{(t-1)} \to k}$.
   (With probability $\rho_{k^{(t-1)} \to k^{(t-1)}}$ update the parameters of $\mathcal{M}_{k^{(t-1)}}$ and skip the remaining steps.)
2. Generate $\mathbf{u}^{(t-1)} \sim g_{k^{(t-1)} \to k}$
3. Set $(\boldsymbol{\theta}, \mathbf{u}) := T_{k^{(t-1)} \to k}(\boldsymbol{\theta}^{(t-1)}, \mathbf{u}^{(t-1)})$.
4. Compute

$$\alpha = \min\left\{1, \frac{f^{\text{post}}(k, \boldsymbol{\theta}) \rho_{k \to k^{(t-1)}} g_{k \to k^{(t-1)}}(\mathbf{u})}{f^{\text{post}}(k^{(t-1)}, \boldsymbol{\theta}^{(t-1)}) \rho_{k^{(t-1)} \to k} g_{k^{(t-1)} \to k}(\mathbf{u}^{(t-1)})} \left| \frac{\partial \mathbf{T}_{k^{(t-1)} \to k}(\boldsymbol{\theta}^{(t-1)}, \mathbf{u}^{(t-1)})}{\partial(\boldsymbol{\theta}^{(t-1)}, \mathbf{u}^{(t-1)})} \right| \right\}$$

5. With probability α set $k^{(t)} = k$ and $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$, otherwise keep $k^{(t)} = k^{(t-1)}$ and $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Properties of the reversible jump algorithm

Remark 6.1

The probability of acceptance of the reversible jump algorithm does not depend on the normalisation constant of the joint posterior $f^{post}(k, \boldsymbol{\theta})$.

Proposition 6.1

The joint posterior $f^{\mathrm{post}}(k, \boldsymbol{\theta})$ is the invariant distribution of MC generated by the reversible jump algorithm, if all $\mathbf{T}_{k \to l}$ are diffeomorphisms with $\mathbf{T}_{l \to k} = \mathbf{T}_{k \to l}^{(-1)}$ and $\rho_{k \to l} \neq 0$ unless $\rho_{l \to k} = 0$.

Example 6.2: Model

- Two possible models $\mathcal{M}_1$ and $\mathcal{M}_2$.
- $\mathcal{M}_1$ has a single parameter $\theta \in [0,1]$.
- $\mathcal{M}_2$ has two parameters $\theta_1, \theta_2 \in D$ with triangular domain $D = \{(\theta_1, \theta_2) : 0 \le \theta_2 \le \theta_1 \le 1\}$.
- The joint posterior of $(k, \boldsymbol{\theta})$ is

$$f^{\text{post}}(k, \boldsymbol{\theta}) \propto p_k f_k^{\text{prior}}(\boldsymbol{\theta}) l_k(\mathbf{y}_1, \ldots, \mathbf{y}_n | \boldsymbol{\theta})$$

- We need to propose two moves $\mathbf{T}_{1\to2}$ and $\mathbf{T}_{2\to1}$ such that $\mathbf{T}_{1\to2} = \mathbf{T}_{2\to1}^{-1}$.

Example 6.2: Moves

- Move from $\mathcal{M}_2$ to $\mathcal{M}_1$: drop θ_2, i.e.

$$(\theta_1, U) := \mathbf{T}_{2\to 1}(\theta_1, \theta_2) = (\theta_1, \star)$$

- Move from $\mathcal{M}_1$ to $\mathcal{M}_2$: draw $U \sim \mathsf{U}[0,1]$ and

$$(\theta_1, \theta_2) = \mathbf{T}_{1\to 2}(\theta_1, U) = (\theta_1, U\theta_1)$$

- For the Jacobian we have that

$$\left| \frac{\partial \mathbf{T}_{1\to 2}(\theta, u)}{\partial(\theta, u)} \right| = \begin{vmatrix} 1 & 0 \\ u & \theta \end{vmatrix} = |\theta| = \theta.$$

- Additionally fixed dimensional move in $\mathcal{M}_1$ (with probability 0.5 update θ_1, otherwise attempt jump to $\mathcal{M}_2$).

Example 6.2: Resulting algorithm (1)

- If the current model is $\mathcal{M}_1$ (i.e. $k^{(t-1)} = 1$):
  - With probability $1/2$ perform an update of $\theta^{(t-1)}$ within model $\mathcal{M}_1$, i.e.
    1. Generate $\theta_1 \sim \mathsf{U}[0,1]$.
    2. Compute the probability of acceptance
$$\alpha = \min\left\{1, \frac{f_1^{\text{prior}}(\theta_1) l_1(\mathbf{y}_1, \dots, \mathbf{y}_n | \theta_1)}{f_1^{\text{prior}}(\theta_1^{(t-1)}) l_1(\mathbf{y}_1, \dots, \mathbf{y}_n | \theta_1^{(t-1)})}\right\}$$
    3. With probability α set $\theta^{(t)} = \theta$, otherwise keep $\theta^{(t)} = \theta^{(t-1)}$.
  - Otherwise attempt a jump to model $\mathcal{M}_2$, i.e.
    1. Generate $u^{(t-1)} \sim \mathsf{U}[0,1]$
    2. Set $(\theta_1, \theta_2) := T_{1\to 2}(\theta^{(t-1)}, u^{(t-1)}) = (\theta^{(t-1)}, u^{(t-1)}\theta^{(t-1)})$.
    3. Compute
$$\alpha = \min\left\{1, \frac{p_2 \cdot f_2^{\text{prior}}(\theta_1, \theta_2) l_2(\mathbf{y}_1, \dots, \mathbf{y}_n | \theta_1, \theta_2) \cdot 1}{p_1 \cdot f_1^{\text{prior}}(\theta_1^{(t-1)}) l_1(\mathbf{y}_1, \dots, \mathbf{y}_n | \theta_1^{(t-1)}) \cdot 1/2 \cdot 1} \cdot \theta_1^{(t-1)}\right\}$$
    4. With probability α set $k^{(t)} = 2$ and $\boldsymbol{\theta}^{(t)} = (\theta_1, \theta_2)$, otherwise keep $k = 1$ and $\theta^{(t)} = \theta^{(t-1)}$.

Example 6.2: Resulting algorithm (2)

- Otherwise, if the current model is $\mathcal{M}_2$ (i.e. $k^{(t-1)} = 2$) attempt a jump to $\mathcal{M}_1$:
  1. Set $(\theta, u) := T_{2\to 1}(\theta_1^{(t-1)}, \theta_2^{(t-1)}) = (\theta_1^{(t-1)}, \theta_2^{(t-1)}/\theta_1^{(t-1)})$.
  2. Compute

$$\alpha = \min\left\{1, \frac{p_1 \cdot f_1^{\text{prior}}(\theta_1) l_1(\mathbf{y}_1, \ldots, \mathbf{y}_n|\theta_1) \cdot 1/2 \cdot 1}{p_2 \cdot f_2^{\text{prior}}(\theta_1^{(t-1)}, \theta_2^{(t-1)}) l_2(\mathbf{y}_1, \ldots, \mathbf{y}_n|\theta_1^{(t-1)}, \theta_2^{(t-1)}) \cdot 1} \cdot \frac{1}{\theta_1^{(t-1)}}\right\}$$

  3. With probability α set $k^{(t)} = 1$ and $\theta^{(t)} = \theta$, otherwise keep $k = 2$ and $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t-1)}, \theta_2^{(t-1)})$.

Example 6.3: Mixtures of Gaussians: Model

- Recall the density of a mixture of Gaussians

$$f(y_i|\pi_1,\ldots,\pi_k,\mu_1,\ldots,\mu_k,\tau_1,\ldots,\tau_k) = \sum_{\kappa=1}^{k} \pi_\kappa \phi_{(\mu_\kappa,1/\tau_\kappa)}(y_i)$$

  with $\boldsymbol{\theta} = (\pi_1,\ldots,\pi_k,\mu_1,\ldots,\mu_k,\tau_1,\ldots,\tau_k)$.
- To ensure identifiability: $\mu_1 < \ldots < \mu_k$.
- Previously: Number of components k treated as known.
- Now: We want to estimate k as well as $\boldsymbol{\theta}$.
- Suitable priors:

  Number of components discrete $p_k = \mathbb{P}(\mathcal{M}_k)$
  Prior probabilities $(\pi_1,\ldots,\pi_k)|\mathcal{M}_k \sim \mathsf{Dirichlet}(\ldots)$
  Mean of population κ $\mu_\kappa|\mathcal{M}_k \sim \mathsf{N}(\ldots)$
  Precision of population κ $\tau_\kappa|\mathcal{M}_k \sim \mathsf{Gamma}(\ldots)$

Example 6.3: Mixtures of Gaussians: Transdimensional moves

- Transdimensional moves correspond to adding / removing components
- Several possibilities for such pairs of moves:
  - Add a new component ("birth") / drop a component ("death") without updating the other components.
  - Split an existing component into two / merge two components into one.

Example 6.3: Mixtures of Gaussians: Birth move

- Birth move $k \longrightarrow k+1$: add a new component by ...
  - drawing the mean μ_{k+1} and the precision τ_{k+1} from the corresponding prior distributions.
  - drawing $\pi_{k+1} \sim \text{Beta}(1, k)$ and rescaling the other prior probabilities to $\pi_\kappa = \pi_\kappa^{(t-1)}(1 - \pi_{k+1}^{(t)})$ $(\kappa = 1, \ldots, k)$.
- In our reversible jump notation:

$$u_1^{(t-1)} \sim g_1 = f_{\text{Beta}(1,k)}, \qquad u_2^{(t-1)} \sim g_2 = f_\mu^{\text{prior}}, \qquad u_3^{(t-1)} \sim g_3 = f_\tau^{\text{prior}}$$

$$\begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \\ \pi_{k+1} \\ \vdots \\ \mu_{k+1} \\ \vdots \\ \tau_{k+1} \end{pmatrix} = T_{k \to k+1} \begin{pmatrix} \pi_1^{(t-1)} \\ \vdots \\ \pi_k^{(t-1)} \\ \vdots \\ u_1^{(t-1)} \\ u_2^{(t-1)} \\ u_3^{(t-1)} \end{pmatrix} = \begin{pmatrix} \pi_1^{(t-1)}(1 - u_1^{(t-1)}) \\ \vdots \\ \pi_k^{(t-1)}(1 - u_1^{(t-1)}) \\ u_1^{(t-1)} \\ \vdots \\ u_2^{(t-1)} \\ \vdots \\ u_3^{(t-1)} \end{pmatrix}$$

Example 6.3: Mixtures of Gaussians: Death move

- Death move $k+1 \longrightarrow k$: remove the κ-th component
- For simplicity assume that we remove the $\kappa = k+1$-st component
- Note that we need to rescale the prior probabilities to $\pi_\iota = \pi_\iota/(1-\pi_{k+1}^{(t-1)})$ $(\iota = 1, \ldots, k)$.
- In our reversible jump notation

$$\begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \\ \vdots \\ u_1 \\ u_2 \\ u_3 \end{pmatrix} = T_{k+1\to k} \begin{pmatrix} \pi_1^{(t-1)} \\ \vdots \\ \pi_{k+1}^{(t-1)} \\ \vdots \\ \mu_{k+1}^{(t-1)} \\ \vdots \\ \tau_{k+1}^{(t-1)} \end{pmatrix} = \begin{pmatrix} \pi_1^{(t-1)}/(1-\pi_{k+1}^{(t-1)}) \\ \vdots \\ \pi_k^{(t-1)}/(1-\pi_{k+1}^{(t-1)}) \\ \vdots \\ \pi_{k+1}^{(t-1)} \\ \mu_{k+1}^{(t-1)} \\ \tau_{k+1}^{(t-1)} \end{pmatrix}$$

Example 6.3: Mixtures of Gaussians: Probabilities of acceptance

- We have $T_{k+1\to k} = T_{k\to k+1}^{-1}$
- The determinant of the Jacobian of the birth move is $T_{k\to k+1}$ is

$$(1-u_1^{(t-1)})^k$$

- Probability of acceptance of the birth move:

$$\min\left\{1, \frac{p_{k+1} f_{k+1}^{\text{prior}}(\boldsymbol{\theta}) l(y_1,\ldots,y_n|\boldsymbol{\theta})}{p_k f_k^{\text{prior}}(\boldsymbol{\theta}^{(t-1)}) l(y_1,\ldots,y_n|\boldsymbol{\theta}^{(t-1)})} \cdot \frac{(k+1)!}{k!} \right.$$
$$\left. \cdot \frac{\rho_{k+1\to k}/(k+1)}{\rho_{k\to k+1} g_1(u_1^{(t-1)}) g_2(u_2^{(t-1)}) g_3(u_3^{(t-1)})} \cdot (1-u_1^{(t-1)})^k \right\}$$

- Probability of acceptance of the death move is the reciprocal.