On some computational methods for Bayesian model choice

Jean-Michel Marin

Université Montpellier 2

October 20, 2008

Outline

Model choice and model comparison

Choice between models

Several models available for the same observation

$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Bayesian resolution

Probabilise the entire model/parameter space

- allocate probabilities p_i to all models $\mathfrak{M}_i$
- define priors $\pi_i(\theta_i)$ for each parameter space Θ_i
- compute

$$\mathbb{P}(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

- take largest $\mathbb{P}(\mathfrak{M}_i|x)$ to determine “best” model, or use averaged predictive

$$\sum_j \mathbb{P}(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j, x)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Bayes factor

Definition (Bayes factors)

For models $\mathfrak{M}_1$ and $\mathfrak{M}_2$

$$B_{12} = \frac{\displaystyle\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\displaystyle\int_{\Theta_2} f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2}$$

[Jeffreys, 1939]

Outside decision-theoretic environment:

- Bayesian/marginal equivalent to the likelihood ratio
- Jeffreys' scale of evidence:
  - if $\log_{10}(B_{12})$ between 0 and 0.5, evidence against $\mathfrak{M}_2$ *weak*,
  - if $\log_{10}(B_{12})$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(B_{12})$ 1 and 2, evidence *strong* and
  - if $\log_{10}(B_{12})$ above 2, evidence *decisive*
- Requires the computation of the marginal/evidence under both hypotheses/models

Evidence

All these problems end up with a similar quantity, the *evidence*

$$\mathfrak{Z}_k = \int_{\Theta_k} \pi_k(\theta_k) f_k(x|\theta_k) \, \mathrm{d}\theta_k \,,$$

the marginal likelihood

Approximating $\mathfrak{Z}_k$ from posterior samples
Bridge sampling

If

$$\begin{aligned} \pi_1(\theta_1|x) &\propto \tilde{\pi}_1(\theta_1|x) \\ \pi_2(\theta_2|x) &\propto \tilde{\pi}_2(\theta_2|x) \end{aligned}$$

on same space $\Theta_1 = \Theta_2$, then

$$B_{12} \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i|x)}{\tilde{\pi}_2(\theta_i|x)} \qquad \theta_i \sim \pi_2(\cdot|x)$$

[Gelman & Meng, 1998; Chen, Shao & Ibrahim, 2000]

In addition

$$B_{12} = \frac{\int \tilde{\pi}_1(\theta|x)\alpha(\theta)\pi_2(\theta|x)\mathsf{d}\theta}{\int \tilde{\pi}_2(\theta|x)\alpha(\theta)\pi_1(\theta|x)\mathsf{d}\theta} \qquad \forall\, \alpha(\cdot)$$

$$\approx \frac{\frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i}|x)\alpha(\theta_{2i})}{\frac{1}{n_1}\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i}|x)\alpha(\theta_{1i})} \qquad \theta_{ji} \sim \pi_j(\cdot|x)$$

Approximating $\mathfrak{Z}_k$ from posterior samples
Harmonic means

Use of the identity

$$\mathbb{E}\left[\frac{\varphi(\theta_k)}{\pi_k(\theta_k)f_k(x|\theta_k)}\,|x\right] = \int \frac{\varphi(\theta_k)}{\pi_k(\theta_k)f_k(x|\theta_k)}\,\frac{\pi_k(\theta_k)f_k(x|\theta_k)}{\mathfrak{Z}_k}\,\mathrm{d}\theta_k = \frac{1}{\mathfrak{Z}_k}$$

no matter what the proposal $\varphi(\theta_k)$ is

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

Harmonic mean type: Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi(\theta)L(\theta)$ for the approximation

$$\widehat{\mathfrak{Z}_k} = 1 \Bigg/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi\left(\theta_k^{(t)}\right)}{\pi_k\left(\theta_k^{(t)}\right) f_k\left(x|\theta_k^{(t)}\right)}$$

to have a finite variance

E.g., use finite support kernels (like the Epanechnikov kernel) for φ

Standard importance sampling

Compare $\widehat{\mathfrak{Z}_k}$ with standard importance sampling approximation

$$\widetilde{\mathfrak{Z}_k} = \frac{1}{T} \sum_{t=1}^{T} \frac{\pi\left(\theta_k^{(t)}\right) f_k\left(x|\theta_k^{(t)}\right)}{\varphi(\theta_k^{(t)})}$$

where the $\theta_k^{(t)}$'s are generated from the density $\varphi(\cdot)$ (with fatter tails this time)

Approximating $\mathfrak{Z}_k$ using a mixture representation

Design a specific mixture for simulation purposes, with density

$$\tilde{\varphi}(\theta_k) \propto \omega_1 \pi_k(\theta_k) f_k(x|\theta_k) + \varphi(\theta_k) \,,$$

where $\varphi(\theta_k)$ is arbitrary (but normalised)
Note: ω_1 is not a probability weight

Corresponding MCMC (=Gibbs) sampler

At iteration t

1. Take $\delta^{(t)} = 1$ with probability

$$\omega_1 \pi_k(\theta_k^{(t-1)}) f_k(x|\theta_k^{(t-1)}) \Big/ \left(\omega_1 \pi_k(\theta_k^{(t-1)}) f_k(x|\theta_k^{(t-1)}) + \varphi(\theta_k^{(t-1)})\right)$$

   and $\delta^{(t)} = 2$ otherwise;
2. If $\delta^{(t)} = 1$, generate $\theta_k^{(t)} \sim \text{MCMC}(\theta_k^{(t-1)}, \cdot)$ where $\text{MCMC}(\theta, \theta')$ denotes an arbitrary MCMC kernel associated with the posterior $\pi_k(\theta|x) \propto \pi_k(\theta) f_k(x|\theta)$;
3. If $\delta^{(t)} = 2$, generate $\theta_k^{(t)} \sim \varphi(\cdot)$ independently

Rao-Blackwellised estimate

$$\hat{\xi} = \frac{1}{T} \sum_{t=1}^{T} \omega_1 \pi_k(\theta_k^{(t)}) f_k(x|\theta_k^{(t)}) \Big/ \omega_1 \pi_k(\theta_k^{(t)}) f_k(x|\theta_k^{(t)}) + \varphi(\theta_k^{(t)}) \,,$$

converges to $\omega_1 \mathfrak{Z}_k / \{\omega_1 \mathfrak{Z}_k + 1\}$
Deduce $\widetilde{\mathfrak{Z}_k}$ from

$$\omega_1 \widetilde{\mathfrak{Z}_k} / \{\omega_1 \widetilde{\mathfrak{Z}_k} + 1\} = \hat{\xi}$$

Chib's representation

Direct application of Bayes' theorem: given $x \sim f_k(x|\theta_k)$ and $\theta_k \sim \pi_k(\theta_k)$,

$$\mathfrak{Z}_k = \frac{f_k(x|\theta_k)\,\pi_k(\theta_k)}{\pi_k(\theta_k|x)}\,,$$

Use of an approximation

$$\widehat{\mathfrak{Z}_k} = \frac{f_k(x|\theta_k^*)\,\pi_k(\theta_k^*)}{\widehat{\pi_k}(\theta_k^*|\mathbf{x})}\,.$$

For missing variable $\mathbf{z}$ as in mixture models,

$$\widehat{\pi_k}(\theta_k^*|\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} \pi_k(\theta_k^*|x, z_k^{(t)})\,,$$

where the $z_k^{(t)}$'s are the latent variables simulated by a Gibbs sampler.
Difficulty caused by [non-]label switching overcome by imposing symmetry: since

$$\pi_k(\theta_k|x) = \pi_k(\sigma(\theta_k)|x) = \frac{1}{k!} \sum_{\sigma\in\mathfrak{S}} \pi_k(\sigma(\theta_k)|x)$$

for all σ's in $\mathfrak{S}_k$, set of all permutations of $\{1,\ldots,k\}$, use of

$$\widetilde{\pi_k}(\theta_k^*|x) = \frac{1}{T\,k!} \sum_{\sigma\in\mathfrak{S}_k} \sum_{t=1}^{T} \pi_k(\sigma(\theta_k^*)|x, z_k^{(t)})\,.$$

Reversible jump

Idea: Set up a proper measure–theoretic framework for designing moves *between* models $\mathfrak{M}_k$

[Green, 1995]

Create a **reversible kernel** $\mathfrak{K}$ on $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ such that

$$\int_A \int_B \mathfrak{K}(x, dy)\pi(x)dx = \int_B \int_A \mathfrak{K}(y, dx)\pi(y)dy$$

for the invariant density π [x is of the form $(k, \theta^{(k)})$]

For a move between two models, $\mathfrak{M}_1$ and $\mathfrak{M}_2$, the Markov chain being in state $\theta_1 \in \mathfrak{M}_1$, denote by $\mathfrak{K}_{1\to 2}(\theta_1, d\theta)$ and $\mathfrak{K}_{2\to 1}(\theta_2, d\theta)$ the corresponding kernels, under the *detailed balance condition*

$$\pi(d\theta_1)\,\mathfrak{K}_{1\to 2}(\theta_1, d\theta) = \pi(d\theta_2)\,\mathfrak{K}_{2\to 1}(\theta_2, d\theta)\,,$$

and take, wlog, $\dim(\mathfrak{M}_2) > \dim(\mathfrak{M}_1)$.
Proposal expressed as

$$\theta_2 = \Psi_{1\to 2}(\theta_1, v_{1\to 2})$$

where $v_{1\to 2}$ is a random variable of dimension $\dim(\mathfrak{M}_2) - \dim(\mathfrak{M}_1)$, generated as

$$v_{1\to 2} \sim \varphi_{1\to 2}(v_{1\to 2})\,.$$

In this case, $\mathfrak{q}_{1\to 2}(\theta_1, d\theta_2)$ has density

$$\varphi_{1\to 2}(v_{1\to 2}) \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right|^{-1},$$

by the Jacobian rule.
If probability $\varpi_{1\to 2}$ of choosing move to $\mathfrak{M}_2$ while in $\mathfrak{M}_1$, acceptance probability reduces to

$$\alpha(\theta_1, v_{1\to 2}) = 1 \wedge \frac{\pi(\mathfrak{M}_2, \theta_2)\, \varpi_{2\to 1}}{\pi(\mathfrak{M}_1, \theta_1)\, \varpi_{1\to 2}\, \varphi_{1\to 2}(v_{1\to 2})} \left| \frac{\partial \Psi_{1\to 2}(\theta_1, v_{1\to 2})}{\partial(\theta_1, v_{1\to 2})} \right|.$$

ⓒDifficult calibration

Saturation schemes

Saturation of the parameter space $\mathfrak{H} = \bigcup_k \{k\} \times \Theta_k$ by creating

- a model index M
- pseudo-priors $\pi_j(\theta_j|M=k)$ for $j \neq k$

[Carlin & Chib, 1995]

Validation by

$$\mathbb{P}(M=k|x) = \int \mathbb{P}(M=k|x,\theta)\pi(\theta|x)\mathrm{d}\theta = \mathfrak{Z}_k$$

where the (marginal) posterior is

$$\begin{aligned}\pi(\theta|x) &= \sum_{k=1}^{D} \mathbb{P}(\theta, M=k|x) \\ &= \sum_{k=1}^{D} p_k\, \mathfrak{Z}_k\, \pi_k(\theta_k|x) \prod_{j\neq k} \pi_j(\theta_j|M=k)\,.\end{aligned}$$

Run a Markov chain $(M^{(t)}, \theta_1^{(t)}, \ldots, \theta_D^{(t)})$ with stationary distribution $\mathbb{P}(\theta, M = k|x)$ by

1. Pick $M^{(t)} = k$ with probability $\mathbb{P}(\theta^{(t-1)}, M = k|x)$
2. Generate $\theta_k^{(t-1)}$ from the posterior $\pi_k(\theta_k|x)$ [or MCMC step]
3. Generate $\theta_j^{(t-1)}$ $(j \neq k)$ from the pseudo-prior $\pi_j(\theta_j|M = k)$

Approximate $\mathbb{P}(M = k|x) = \mathfrak{Z}_k$ by

$$\check{\mathfrak{Z}}_k \propto p_k \sum_{t=1}^{T} f_k(x|\theta_k^{(t)})\, \pi_k(\theta_k^{(t)}) \prod_{j\neq k} \pi_j(\theta_j^{(t)}|M = k)$$

$$\Bigg/ \sum_{\ell=1}^{D} \varrho_\ell\, f_\ell(x|\theta_\ell^{(t)})\, \pi_\ell(\theta_\ell^{(t)}) \prod_{j\neq \ell} \pi_j(\theta_j^{(t)}|M = \ell)$$

Scott's (2002) mistake

Suggest estimating $\mathbb{P}(M = k|y)$ by

$$\widetilde{\mathfrak{Z}_k} \propto p_k \sum_{t=1}^{T} \left\{ f_k(y|\theta_k^{(t)}) \Big/ \sum_{j=1}^{D} \varrho_j \, f_j(y|\theta_j^{(t)}) \right\} ,$$

simultaneously and independently, D MCMC chains

$$(\theta_k^{(t)})_t \, , \qquad 1 \leq k \leq D \, ,$$

with stationary distributions $\pi_k(\theta_k|y)$
instead of above joint

Congdon's (2006) mistake

Using flat pseudo-priors [prohibited!], uses instead

$$\widehat{\mathfrak{Z}_k} \propto p_k \sum_{t=1}^{T} \left\{ f_k(y|\theta_k^{(t)})\pi_k(\theta_k^{(t)}) \Bigg/ \sum_{j=1}^{D} \varrho_j \, f_j(y|\theta_j^{(t)})\pi_j(\theta_j^{(t)}) \right\} ,$$

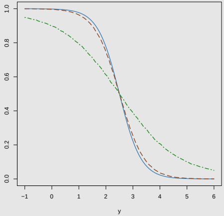
where again the $\theta_k^{(t)}$'s are MCMC chains with stationary distributions $\pi_k(\theta_k|y)$

Examples (1)

Example (Model choice (2))

Normal model $\mathfrak{M}_1 : y|\theta \sim \mathcal{N}(\theta, 1)$ with $\theta \sim \mathcal{N}(0, 1)$ vs. normal model $\mathfrak{M}_2 : y|\theta \sim \mathcal{N}(\theta, 1)$ with $\theta \sim \mathcal{N}(5, 1)$

Comparison of both approximations with $\mathbb{P}(M = 1|y)$: Scott's (2002) (green and mixed dashes) and Congdon's (2006) (brown and long dashes) ($N = 10^4$ simulations).



Examples (2)

Example (Model choice (3))

Model $\mathfrak{M}_1 : y|\omega \sim \mathcal{N}(0, 1/\omega)$ with $\omega \sim \mathcal{E}xp(a)$ vs. $\mathfrak{M}_2 : \exp(y)|\lambda \sim \mathcal{E}xp(\lambda)$ with $\lambda \sim \mathcal{E}xp(b)$.

Comparison of Congdon's (2006) (brown and dashed lines) with $\mathbb{P}(M = 1|y)$ when (a, b) is equal to $(.24, 8.9)$, $(.56, .7)$, $(4.1, .46)$ and $(.98, .081)$, resp. ($N = 10^4$ simulations).