Bayesian Statistics

Bayes Factor

Nan Lin
Department of Mathematics
Washington University in St. Louis

Comparing two models

- Suppose we have two competing models/hypotheses: M_1 and M_2
- Assuming that either M_1 or M_2 is true, we would like to know what the data y tell us about the probabilities of either model being true, i.e. $p(M_1|y)$ and $p(M_2|y)$.
  - Look at posterior odds in favor of M_1

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{\frac{p(\mathbf{y}|M_1)p(M_1)}{p(\mathbf{y})}}{\frac{p(\mathbf{y}|M_2)p(M_2)}{p(\mathbf{y})}}$$

Bayes Factor

$$= \boxed{\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}} \frac{p(M_1)}{p(M_2)}$$

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

Bayes factor

- If both models are equally likely a priori ($p(M_1) = p(M_2)$), then the posterior odds equal the Bayes factor.

$$\frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}$$

  - Bayes factor is a measure of how much the data supports one model relative to the other.
  - Requires computing the marginal likelihood $p(y|M_1)$ and $p(y|M_2)$

$$\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} = \frac{\int_{\Theta_1} p(\mathbf{y}|\boldsymbol{\theta}_1, M_1)p(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_1}{\int_{\Theta_2} p(\mathbf{y}|\boldsymbol{\theta}_2, M_2)p(\boldsymbol{\theta}_2|M_2)d\boldsymbol{\theta}_2}$$

Interpretation of Bayes factor

Bayes Factor	Strength of Evidence for M_1
< 1	Negative (supports M_2)
1 to 3	Barely Worth Mentioning
3 to 10	Substantial
10 to 30	Strong
30 to 100	Very Strong
> 100	Decisive

Marginal likelihood

- The marginal likelihood of model M_k is

$$p(\mathbf{y}|M_k) = \int_{\Theta_k} p(\mathbf{y}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k$$

  - where θ_k are model parameters of model M_k
  - $p(\theta_k|M_k)$: prior for θ_k under model M_k

  - the expected value of the likelihood function taken over the prior density
  - the normalizing constant of the posterior $p(\theta_k|y)$ under model M_k → in general, it is difficult to obtain, and one has to consider approximating it

Model Selection

- Data: y
- $\mathcal{M}$: finite set of competing models
- $\boldsymbol{\theta}_j$: a distinct unknown parameter vector of dimension n_j corresponding to the jth model
- Prior: $\pi_j \equiv P(M = j)$ $\sum_{j \in \mathcal{M}} \pi_j = 1$
- Θ_j all possible values for $\boldsymbol{\theta}_j$

$$\boldsymbol{\theta}_j \in \Theta_j \subset \mathbb{R}^{n_j}$$

- $\boldsymbol{\theta}$: collection of all model specific $\boldsymbol{\theta}_j$

Model Selection

- Posterior probability

$$P(M = j|\mathbf{y}),$$

  - A single "best" model
  - Model averaging

- Bayes factor: Choice between two models

$$B_{21} = \frac{P(M = 2 \mid \mathbf{y})/P(M = 1 \mid \mathbf{y})}{P(M = 2)/P(M = 1)} = \frac{p(\mathbf{y} \mid M = 2)}{p(\mathbf{y} \mid M = 1)},$$

Estimating Marginal Likelihood

- Marginal likelihood

$$p(\mathbf{y}|M = j)$$

- Estimation
  - Ordinary Monte Carlo sampling
    - Difficult to implement for high-dimensional models
    - MCMC does not provide estimate of marginal likelihood directly
  - Include model indicator as a parameter in sampling
    - Product space search by Gibbs sampling
    - Metropolis-Hastings
    - Reversible Jump MCMC

Product space search

- Carlin and Chib (1995, JRSSB)
- Data likelihood of Model j $f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)$
- Prior of model j $p(\boldsymbol{\theta}_j|M = j).$
- Assumption:
  - M is merely an indicator of which $\boldsymbol{\theta}_j$ elevant to y
  - Y is independent of $\{\boldsymbol{\theta}_{j' \neq j}\}$ en the model indicator M
  - Proper priors are required
  - Prior independence among $\boldsymbol{\theta}_j$'s given M

Product space search

- The sampler operates over the product space

$$\mathcal{M} \times \prod_{j \in \mathcal{M}} \Theta_j .$$

- Marginal likelihood

$$p(\mathbf{y}|M=j) = \int f(\mathbf{y}|\boldsymbol{\theta}, M=j) p(\boldsymbol{\theta}|M=j) d\boldsymbol{\theta}$$

$$= \int f(\mathbf{y}|\boldsymbol{\theta}_j, M=j) p(\boldsymbol{\theta}_j|M=j) d\boldsymbol{\theta}_j ,$$

- Remark: $p(\boldsymbol{\theta}_j|M \neq j)$ become irrelevant in this calculation, we may choose these pseudopriors in any way we like.

Search by Gibbs sampler

$$p(\boldsymbol{\theta}_j|\boldsymbol{\theta}_{j'\neq j}, M, \mathbf{y}) \propto \begin{cases} f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j) & \text{if } M = j \\ p(\boldsymbol{\theta}_j|M \neq j) & \text{if } M \neq j \end{cases},$$

and

$$P(M = j|\boldsymbol{\theta}, \mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)[\prod_{j'\in\mathcal{M}} p(\boldsymbol{\theta}_{j'}|M = j)]\pi_j}{\sum_{k\in\mathcal{M}}\{f(\mathbf{y}|\boldsymbol{\theta}_k, M = k)[\prod_{j'\in\mathcal{M}} p(\boldsymbol{\theta}_{j'}|M = k)]\pi_k\}},$$

Bayes Factor

- Provided the sampling chain for the model indicator mixes well, the posterior probability of model j can be estimated by

$$\widehat{P}(M = j|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} I(M^{(g)} = j),$$

- Bayes factor is estimated by

$$\widehat{B}_{jj'} = \frac{\widehat{P}(M = j|\mathbf{y})/\widehat{P}(M = j'|\mathbf{y})}{P(M = j)/P(M = j')}.$$

Choice of prior probability

- In general, $\pi_j = P(M = j)$ can be chosen arbitrarily
  - Its effect is divided out in the estimate of Bayes factor
- Often, they are chosen so that the algorithm visits each model in roughly equal proportion
  - Allows more accurate estimate of Bayes factor
  - Preliminary runs are needed to select computationally efficient values

More remarks

- Performance of this method is optimized when the pseudo-priors $p(\boldsymbol{\theta}_j|M \neq j)$ match the corresponding model specific priors as nearly as possible
- Draw back of the method
  - Draw must be made from each pseudo prior at each iteration to produce acceptably accurate results
  - If a large number of models are considered, the method becomes impractical

Metropolized product space search

- Dellaportas P., Forster J.J., Ntzoufras I. (2002). On Bayesian Model and Variable Selection Using MCMC. *Statistics and Computing, 12, 27-36.*
- A hybrid Gibbs-Metropolis strategy
- Model selection step is based on a proposal moving between models

Metropolized Carlin and Chib (MCC)

1. Let the current state be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.
2. Propose a new model j' with probability $h(j, j')$.
3. Generate $\boldsymbol{\theta}_{j'}$ from a pseudoprior $p(\boldsymbol{\theta}_{j'}|M \neq j')$ as in Carlin and Chib's method.
4. Accept the proposed move (from j to j') with probability

$$\alpha_{j \to j'} = \min\left\{1, \frac{f(\mathbf{y}|\boldsymbol{\theta}_{j'}, M = j')p(\boldsymbol{\theta}_{j'}|M = j')p(\boldsymbol{\theta}_j|M = j')\pi_{j'}h(j', j)}{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j)p(\boldsymbol{\theta}_{j'}|M = j)\pi_j h(j, j')}\right\}.$$

<u>Advantage</u>: Only needs to sample from the pseudo prior for the proposed model

Reversible jump MCMC

- Green (1995, Biometrika)
- This method operates on the union space

$$\mathcal{M} \times \bigcup_{j \in \mathcal{M}} \Theta_j.$$

- It generates a Markov chain that can jump between models with parameter spaces of different dimensions

RJMCMC

1. Let the current state of the Markov chain be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.
2. Propose a new model j' with probability $h(j, j')$.
3. Generate $\mathbf{u}$ from a proposal density $q(\mathbf{u}|\boldsymbol{\theta}_j, j, j')$.
4. Set $(\boldsymbol{\theta}'_{j'}, \mathbf{u}') = \mathbf{g}_{j,j'}(\boldsymbol{\theta}_j, \mathbf{u})$, where $\mathbf{g}_{j,j'}$ is a deterministic function that is 1-1 and onto. This is a "dimension-matching" function, specified so that $n_j + \dim(\mathbf{u}) = n_{j'} + \dim(\mathbf{u}')$.
5. Accept the proposed move (from j to j') with probability

$$\alpha_{j\to j'} = \min\left\{1, \frac{f(\mathbf{y}|\boldsymbol{\theta}'_{j'}, M = j')p(\boldsymbol{\theta}'_{j'}|M = j')\pi_{j'}h(j', j)q(\mathbf{u}'|\boldsymbol{\theta}_{j'}, j', j)}{f(\mathbf{y}|\boldsymbol{\theta}_j, M = j)p(\boldsymbol{\theta}_j|M = j)\pi_j h(j, j')q(\mathbf{u}|\boldsymbol{\theta}_j, j, j')} \times \left|\frac{\partial \mathbf{g}(\boldsymbol{\theta}_j, \mathbf{u})}{\partial(\boldsymbol{\theta}_j, \mathbf{u})}\right|\right\}. \quad (5)$$

Using Partial Analytic Structure (PAS)

- Godsill (2001, JCGS)
- Similar setup as in the CC method, but allows parameters to be shared between different models.

  there exists a subvector $(\boldsymbol{\theta}_{j'})_{\mathcal{U}}$ of the parameter vector $\boldsymbol{\theta}_{j'}$ for model j' such that $p((\boldsymbol{\theta}_{j'})_{\mathcal{U}} | (\boldsymbol{\theta}_{j'})_{-\mathcal{U}}, M = j', \mathbf{y})$ is available in closed form, and in the current model j, there exists an equivalent subvector $(\boldsymbol{\theta}_j)_{-\mathcal{U}}$ (the elements of $\boldsymbol{\theta}_j$ not in subvector $\mathcal{U}$) of the same dimension as $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$. Operationally,

- Avoids dimension matching

PAS

1. Let the current state be $(j, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is of dimension n_j.
2. Propose a new model j' with probability $h(j, j')$.
3. Set $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}} = (\boldsymbol{\theta}_j)_{-\mathcal{U}}$.
4. Accept the proposed move with probability

$$\alpha_{j \to j'} = \min\left\{1, \ \frac{p(j'|(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}, \mathbf{y})h(j', j)}{p(j|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y})h(j, j')}\right\}, \qquad (6)$$

   where $p(j|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y}) = \int p(j, (\boldsymbol{\theta}_j)_{\mathcal{U}}|(\boldsymbol{\theta}_j)_{-\mathcal{U}}, \mathbf{y})\, d(\boldsymbol{\theta}_j)_{\mathcal{U}}$.
5. If the model move is accepted, update the parameters of the new model $(\boldsymbol{\theta}_{j'})_{\mathcal{U}}$ and $(\boldsymbol{\theta}_{j'})_{-\mathcal{U}}$ using standard Gibbs or Metropolis–Hastings steps; otherwise, update the parameters of the old model $(\boldsymbol{\theta}_j)_{\mathcal{U}}$ and $(\boldsymbol{\theta}_j)_{-\mathcal{U}}$ using standard Gibbs or Metropolis–Hastings steps.

Marginal likelihood estimation (Chib,1995)

$$\forall \boldsymbol{\theta} \in \Theta, \; p(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) / p(\boldsymbol{\theta}|\mathbf{y}),$$

$$\log \hat{p}(\mathbf{y}) = \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}) + \log p(\tilde{\boldsymbol{\theta}}) - \log \hat{p}(\tilde{\boldsymbol{\theta}}|\mathbf{y}) \; ,$$

Marginal likelihood estimation (Chib 1995)

- Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ When all full conditional distributions for the parameters are in closed form

$$p(\tilde{\boldsymbol{\theta}}|\mathbf{y}) = p(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2|\mathbf{y}) = p(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1, \mathbf{y})p(\tilde{\boldsymbol{\theta}}_1|\mathbf{y}),$$

where $p(\tilde{\boldsymbol{\theta}}_1|\mathbf{y})$ can be estimated by

$$\hat{p}(\tilde{\boldsymbol{\theta}}_1|\mathbf{y}) = \frac{1}{G}\sum_{g=1}^{G} p(\tilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2^{(g)}, \mathbf{y})$$

Chib (1995)

$$\log \hat{p}(\mathbf{y}) = \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) + \log p(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$$
$$- \log p(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1, \mathbf{y}) - \log \hat{p}(\tilde{\boldsymbol{\theta}}_1|\mathbf{y}),$$

- The first three terms on the right side are available in close form
- The last term on the right side can be estimated from Gibbs steps

Chib and Jeliazkov (2001)

- Estimation of the last term requires knowing the normalizing constant
  - Not applicable to Metropolis-Hastings
- Let the acceptance probability be

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y}) = \min\{1, [p(\boldsymbol{\theta}'|\mathbf{y})q(\boldsymbol{\theta}', \boldsymbol{\theta}|\mathbf{y})]/[p(\boldsymbol{\theta}|\mathbf{y})q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})]\},$$

Chib and Jeliazkov (2001)

$$p(\tilde{\boldsymbol{\theta}}|\mathbf{y}) = \frac{E_1\{\alpha(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}|\mathbf{y})q(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}|\mathbf{y})\}}{E_2\{\alpha(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}|\mathbf{y})\}},$$

where E_1 is the expectation with respect to the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ and E_2 is the expectation with respect to the candidate density $q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}|\mathbf{y})$. The numerator is then estimated by averaging the product in braces with respect to draws from the posterior, and the denominator is estimated by averaging the acceptance probability with respect to draws from $q(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}|\mathbf{y})$,

Example: linear regression model

Table 1. Radiata Pine Compressive Strength Data

Case (i)	y_i	x_i	z_i	Case (i)	y_i	x_i	z_i
1	3040	29.2	25.4	22	3840	30.7	30.7
2	2470	24.7	22.2	23	3800	32.7	32.6
3	3610	32.3	32.2	24	4600	32.6	32.5
4	3480	31.3	31.0	25	1900	22.1	20.8
5	3810	31.5	30.9	26	2530	25.3	23.1
6	2330	24.5	23.9	27	2920	30.8	29.8
7	1800	19.9	19.2	28	4990	38.9	38.1
8	3110	27.3	27.2	29	1670	22.1	21.3
9	3160	27.1	26.3	30	3310	29.2	28.5
10	2310	24.0	23.9	31	3450	30.1	29.2
11	4360	33.8	33.2	32	3600	31.4	31.4
12	1880	21.5	21.0	33	2850	26.7	25.9
13	3670	32.2	29.0	34	1590	22.1	21.4
14	1740	22.5	22.0	35	3770	30.3	29.8
15	2250	27.5	23.8	36	3850	32.0	30.6
16	2650	25.6	25.3	37	2480	23.2	22.6
17	4970	34.5	34.2	38	3570	30.3	30.3
18	2620	26.2	25.7	39	2620	29.9	23.8
19	2900	26.7	26.4	40	1890	20.8	18.4
20	1670	21.1	20.0	41	3030	33.2	29.4
21	2540	24.1	23.9	42	3030	28.2	28.2

Two models

$$M = 1:\ y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,$$

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2), \qquad i = 1, \ldots, n,$$

$$M = 2:\ y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i,$$

$$\eta_i \overset{\text{iid}}{\sim} N(0, \tau^2), \qquad i = 1, \ldots, n.$$

$\mathcal{M} = \{1,\ 2\}$, $\boldsymbol{\theta}_1 = (\alpha, \beta, \sigma^2)^\top$, and $\boldsymbol{\theta}_2 = (\gamma, \delta, \tau^2)^\top$.

Priors

adopt $N((3000, 185)^\top, \mathrm{Diag}(10^6, 10^4))$ priors on $(\alpha, \beta)^\top$ and $(\gamma, \delta)^\top$, and $\mathrm{IG}(3, (2 \cdot 300^2)^{-1})$ priors on σ^2 and τ^2, where $\mathrm{IG}(a, b)$ represents the inverse gamma distribution with density function $f(v) = \exp(-\frac{1}{bv})/[\Gamma(a) b^a v^{a+1}]$, $v > 0$, $a, b > 0$.

To fix the pseudopriors, we suppose that $(\alpha, \beta)^\top | M = 2 \sim N((3000, 185)^\top, \mathrm{Diag}(52^2, 12^2))$, $(\gamma, \delta)^\top | M = 1 \sim N((3000, 185)^\top, \mathrm{Diag}(43^2, 9^2))$, $\sigma^2 | M = 2 \sim \mathrm{IG}(3, (2 \cdot 300^2)^{-1})$, and $\tau^2 | M = 1 \sim \mathrm{IG}(3, (2 \cdot 300^2)^{-1})$. Independence among the components of the pseudopriors is also assumed;

We choose π_1 and π_2 to be .9995 and .0005,

For MCC

- h(1,1)=h(1,2)=h(2,1)=h(2,2)=0.5

$$\alpha_{1\to 2} = \min\left\{1,\ \frac{f(\mathbf{y}|\gamma,\delta,\tau^2,M=2)p(\gamma,\delta,\tau^2|M=2)p(\alpha,\beta,\sigma^2|M=2)\pi_2}{f(\mathbf{y}|\alpha,\beta,\sigma^2,M=1)p(\alpha,\beta,\sigma^2|M=1)p(\gamma,\delta,\tau^2|M=1)\pi_1}\right\};$$

$$\alpha_{2\to 1} = \min\left\{1,\ \frac{f(\mathbf{y}|\alpha,\beta,\sigma^2,M=1)p(\alpha,\beta,\sigma^2|M=1)p(\gamma,\delta,\tau^2|M=1)\pi_1}{f(\mathbf{y}|\gamma,\delta,\tau^2,M=2)p(\gamma,\delta,\tau^2|M=2)p(\alpha,\beta,\sigma^2|M=2)\pi_2}\right\}.$$

For RJMCMC

- Dimension matching is automatically satisfied
- Due to similarities between two models

set $(\alpha, \beta, \lambda)^\top = (\gamma, \delta, \omega)^\top$ or $(\gamma, \delta, \omega)^\top = (\alpha, \beta, \lambda)^\top$. The acceptance probabilities are then given by

$$\alpha_{1\to 2} = \min\left\{1,\ \frac{f(\mathbf{y}|\gamma, \delta, \omega, M=2)\pi_2}{f(\mathbf{y}|\alpha, \beta, \lambda, M=1)\pi_1}\right\},$$

and

$$\alpha_{2\to 1} = \min\left\{1,\ \frac{f(\mathbf{y}|\alpha, \beta, \lambda, M=1)\pi_1}{f(\mathbf{y}|\gamma, \delta, \omega, M=2)\pi_2}\right\}.$$

Chib's method

- Two block gibbs sampler
  - (regression coefficients, variance)

Results

- By numerical integration, the true Bayes factor should be 4862 in favor of Model 2

Table 2. Comparison of Methods for Simple Linear Regression Example

Method	$\widehat{P}(M=2\|y)$	*SD*	*95% CI for $P(M=2\|y)$*	$\widehat{B}_{21}$	$\hat{\rho}(1)$	*Pr(move)*	*Time*
CC	.70806	.001721	(.70469, .71144)	4848.4	.567	.179	22.8″
MCC	.71195	.002061	(.70791, .71599)	4940.7	.673	.134	12.2″
RJ-M	.70861	.004058	(.70066, .71657)	4861.3	.589	.170	18.7″
RJ-G	.70906	.002394	(.70437, .71376)	4871.9	.593	.168	7.9″
RJ-R	.70750	.002004	(.70357, .71142)	4835.1	.660	.141	6.7″
PAS	.71035	.001800	(.70682, .71388)	4902.4	.591	.168	7.8″
Chib-1				4860.7			13.6″
Chib-2				4860.3			14.0″
Target	.70865			4862			

NOTE: RJ-M: reversible jump using Metropolis steps if the current model is proposed; RJ-G: reversible jump using Gibbs steps if the current model is proposed; PAS: Godsill's partial analytic structure method, where σ^2 and τ^2 are treated as the same parameter; RJ-R: reversible jump on the reduced model (i.e., with the regression coefficients integrated out); Chib-1: Chib's method evaluated at posterior means; Chib-2: Chib's method evaluated at frequentist LS solutions.