BAYESIAN MODEL CRITICISM

Hedibert Freitas Lopes

The University of Chicago Booth School of Business
5807 South Woodlawn Avenue, Chicago, IL 60637
http://faculty.chicagobooth.edu/hedibert.lopes

hlopes@ChicagoBooth.edu

Based Gamerman and Lopes' (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman&Hall/CRC.

Outline

Bayesian model criticism

Suppose that the competing models can be enumerated and are represented by the set $M = \{M_1, M_2, \ldots\}$, and that the *true model* is in M (Bernardo and Smith, 1994).

The posterior model probability of model M_j is given by

$$Pr(M_j|y) \propto f(y|M_j)Pr(M_j)$$

where

$$f(y|M_j) = \int f(y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j$$

is the prior predictive density of model M_j and $Pr(M_j)$ is the prior model probability of model M_j.

Posterior odds

The posterior odds of model M_j relative to M_k is given by

$$\underbrace{\frac{Pr(M_j|y)}{Pr(M_k|y)}}_{\text{posterior odds}} = \underbrace{\frac{Pr(M_j)}{Pr(M_k)}}_{\text{prior odds}} \times \underbrace{\frac{f(y|M_j)}{f(y|M_k)}}_{\text{Bayes factor}} .$$

The Bayes factor can be viewed as the weighted likelihood ratio of M_j to M_k.

The main difficulty is the computation of the marginal likelihood or normalizing constant $f(y|M_j)$.

Bayes factor

Jeffreys (1961) recommends the use of the following rule of thumb to decide between models j and k:

$$\begin{aligned} B_{jk} > 100 &: \quad \text{decisive evidence against } k \\ 10 < B_{jk} \leq 100 &: \quad \text{strong evidence against } k \\ 3 < B_{jk} \leq 10 &: \quad \text{substantial evidence against } k \end{aligned}$$

Therefore, the posterior model probability for model j can be obtained from

$$\frac{1}{Pr(M_j|y)} = \sum_{M_k \in M} B_{kj} \frac{Pr(M_k)}{Pr(M_j)}.$$

Marginal likelihood

A basic ingredient for model assessment is given by the predictive density

$$f(y|M) = \int f(y|\theta, M)p(\theta|M)d\theta \ ,$$

which is the normalizing constant of the posterior distribution.

The predictive density can now be viewed as the likelihood of model M.

It is sometimes referred to as predictive likelihood, because it is obtained after marginalization of model parameters.

The predictive density can be written as the expectation of the likelihood with respect to the prior:

$$f(y) = E_p[f(y|\theta)].$$

Laplace-Metropolis estimator

Let m and V be the posterior mode and the asymptotic approximation for the posterior covariance matrix.

A normal approximation to $f(y)$ is given by

$$\widehat{f}_0(y) = (2\pi)^{d/2}|\widehat{V}|^{1/2}p(\widehat{m})f(y|\widehat{m})$$

Sampling-based approximations for m and V can be constructed from posterior draws $\theta^{(1)}, \ldots, \theta^{(N)}$:

- $\widehat{m} = \arg\max_{\theta^{(j)}} \pi(\theta^{(j)})$
- $\widehat{V} = \frac{1}{N}\sum_{j=1}^{N}(\theta^{(j)} - \bar{\theta})(\theta^{(j)} - \bar{\theta})'$, where $\bar{\theta} = \frac{1}{N}\sum_{j=1}^{N}\theta^{(j)}$.

Simple Monte Carlo

A direct Monte Carlo estimate is

$$\hat{f}_1(y) = \frac{1}{N}\sum_{j=1}^{N} f(y|\theta^{(j)})$$

where $\theta^{(1)}, \ldots, \theta^{(N)}$ is a sample from the prior distribution $p(\theta)$.

This estimator does not work well in cases of disagreement between prior and likelihood (Raftery, 1996, and McCulloch and Rossi, 1991).

Even for large values of n, this estimate will be influenced by a few sampled values, making it very unstable.

Monte Carlo via IS

An alternative is to perform importance sampling with the aim of boosting sampled values in regions where the integrand is large.

This approach is based on sampling from the importance density $g(\theta)$, since the predictive density can be rewritten as

$$f(y) = E_g\left[\frac{f(y|\theta)p(\theta)}{g(\theta)}\right].$$

This form motivates a new estimate

$$\hat{f}_2(y) = \frac{1}{N}\sum_{j=1}^{N}\frac{f(y|\theta^{(j)})p(\theta^{(j)})}{g(\theta^{(j)})}$$

where $\theta^{(1)}, \ldots, \theta^{(N)}$ is a sample from $g(\theta)$.

Sometimes g is only known up to a normalizing constant, i.e. $g(\theta) = kg^*(\theta)$, and the value of k must be estimated.

Noting that

$$k = \int kp(\theta)d\theta = \int \frac{p(\theta)}{g^*(\theta)} g(\theta)d\theta$$

leads to the estimator of k given by

$$\widehat{k} = \frac{1}{N}\sum_{j=1}^{N} \frac{p(\theta^{(j)})}{g^*(\theta^{(j)})}$$

where, again, the $\theta^{(j)}$ are sampled from g.

Replacing this estimate in $\widehat{f}_2(y)$ gives

$$\widehat{f}_3(y) = \frac{\sum_{j=1}^{N} f(y|\theta^{(j)})p(\theta^{(j)})/g^*(\theta^{(j)})}{\sum_{j=1}^{N} p(\theta^{(j)})/g^*(\theta^{(j)})} .$$

Harmonic mean

The harmonic mean (HM) estimator is obtained when $g(\theta)$ is the posterior $\pi(\theta)$:

$$\widehat{f}_4(y) = \left[\frac{1}{N} \sum_{j=1}^{N} \frac{1}{f(y|\theta^{(j)})} \right]^{-1}$$

for $\theta^{(1)}, \ldots, \theta^{(N)}$ from $\pi(\theta)$.

This is a very appealing estimator for its simplicity.

However, it is strongly affected by small likelihood values!

Newton-Raftery estimator

A compromise between $\widehat{f}_1$ and $\widehat{f}_4$ would lead to

$$g(\theta) = \delta p(\theta) + (1-\delta)\pi(\theta)$$

Problem: $f(y)$ needs to be known!

Solution via an iterative scheme:

$$\widehat{f}_5^{(i)}(y) = \frac{\sum_{j=1}^N \frac{p(y|\theta^{(j)})}{\delta \widehat{f}_5^{(i-1)}(y) + (1-\delta)p(y|\theta^{(j)})}}{\sum_{j=1}^N \frac{1}{\delta \widehat{f}_5^{(i-1)}(y) + (1-\delta)p(y|\theta^{(j)})}}$$

for $i = 1, 2, \ldots$ and, say, $\widehat{f}_5^{(0)} = \widehat{f}_4$.

Generalized HM

For any given density $g(\theta)$ it is easy to see that

$$\int g(\theta) \frac{f(y)\pi(\theta)}{f(y|\theta)p(\theta)} d\theta = 1$$

so,

$$f(y) = \left[\int \frac{g(\theta)}{f(y|\theta)p(\theta)} \pi(\theta) d\theta \right]^{-1} .$$

Therefore, sampling $\theta^{(1)}, \ldots, \theta^{(1)}$ from π leads to the estimate

$$\widehat{f}_6(y) = \left[\frac{1}{N} \sum_{j=1}^{N} \frac{g(\theta^{(j)})}{f(y|\theta^{(j)})p(\theta^{(j)})} \right]^{-1} .$$

Bridge sampler

Meng and Wong (1996) introduced the bridge sampling to estimate ratios of normalizing constants by noticing that

$$f(y) = \frac{E_g\{\alpha(\theta)p(\theta)p(y|\theta)\}}{E_\pi\{\alpha(\theta)g(\theta)\}}$$

for any *bridge* function $\alpha(\theta)$ with support encompassing both supports of the posterior density π and the proposal density g.

If $\alpha(\theta) = 1/g(\theta)$ then the bridge estimator reduces to the simple Monte Carlo estimator $\widehat{f}_1$.

Similarly, if $\alpha(\theta) = \{p(\theta)p(y|\theta)g(\theta)\}^{-1}$ then the bridge estimator is a variation of the harmonic mean estimator.

Meng and Wong (1996) showed that the optimal mean square error α function is

$$\alpha(\theta) = \{g(\theta) + (N_2/N_1)\pi(\theta)\}^{-1},$$

which depends on $f(y)$ itself.

By letting

$$\begin{array}{rcll} \omega_j & = & p(y|\theta^{(j)})p(\theta^{(j)})/g(\theta^{(j)}) & \quad \theta^{(1)},\ldots,\theta^{(N_1)} \sim \pi(\theta) \\ \tilde{\omega}_j & = & p(y|\tilde{\theta}^{(j)})p(\tilde{\theta}^{(j)})/g(\tilde{\theta}^{(j)}) & \quad \tilde{\theta}^{(j)},\ldots,\tilde{\theta}^{(N_2)} \sim g(\theta) \end{array}$$

they devised an iterative scheme:

$$\widehat{f}_7^{(i)}(y) = \frac{\frac{1}{N_2}\sum_{j=1}^{N_2} \frac{\tilde{\omega}_j}{s_1\tilde{\omega}_j + s_2\widehat{f}_7^{(i-1)}(y)}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{1}{s_1\omega_j + s_2\widehat{f}_7^{(i-1)}(y)}},$$

for $i = 1, 2, \ldots$, $s_1 = N_1/(N_1 + N_2)$, $s_2 = N_2/(N_1 + N_2)$ and, say, $\widehat{f}_7^{(0)} = \widehat{f}_4$.

Chib's estimator

Chib (1995) introduced an estimate of $\pi(\theta)$ when conditionals are available in closed form:

$$\widehat{\pi}(\theta) = \widehat{\pi}(\theta_1) \prod_{i=2}^{d} \widehat{\pi}(\theta_i|\theta_1, \ldots, \theta_{i-1})$$

with

$$\widehat{\pi}(\theta_i|\theta_1, \ldots, \theta_{i-1}) = \frac{1}{N} \sum_{j=1}^{N} \pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}^{(j)}, \ldots, \theta_d^{(j)})$$

and $(\theta_1^{(j)}, \ldots, \theta_d^{(j)})$, $j = 1, \ldots, n$, draws from $\pi(\theta)$. Thus

$$\widehat{f}_8(y) = \frac{f(y|\theta)p(\theta)}{\widehat{\pi}(\theta)} \qquad \forall\theta$$

Simple choices of θ are the mode and the mean but any value in that region should be adequate.

List of estimators of $f(y)$

Estimate	Proposal density/method
$\widehat{f}_0$	normal approximation
$\widehat{f}_1$	$p(\theta)$
$\widehat{f}_2$	unnormalized $g^*(\theta)$
$\widehat{f}_3$	unnormalized $g(\theta)$
$\widehat{f}_4$	$\pi(\theta)$
$\widehat{f}_5$	$\delta p(\theta) + (1-\delta)\pi(\theta)$
$\widehat{f}_6$	generalized harmonic mean
$\widehat{f}_7$	optimal bridge sampling
$\widehat{f}_8$	candidate's estimator: Gibbs

DiCiccio, Kass, Raftery and Wasserman (1997), Han and Carlin (2001) and Lopes and West (2004), among others, compared several of these estimators.

Example: Cauchy-normal

Model: $y_1, \ldots, y_n \sim N(\theta, \sigma^2)$, σ^2 known.
Cauchy prior: $p(\theta) = \pi^{-1}(1+\theta^2)^{-1}$.
Data: $\bar{y} = 7$ and $\sigma^2/n = 4.5$.
Posterior density for θ:

$$\pi(\theta) \propto (1+\theta^2)^{-1} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \bar{y})^2\right\} .$$

Error $= 100|\hat{f}(y) - f(y)|/f(y)$. For $\hat{f}_5$, $\delta = 0.1$.

$f(y)$	0.00963235	
Estimator	Estimate	Error (%)
$f(y)$	0.00963235	
$\hat{f}_0$	0.00932328	3.21
$\hat{f}_1$	0.00960189	0.32
$\hat{f}_4$	0.01055301	9.56
$\hat{f}_5$	0.00957345	0.61
$\hat{f}_6$	0.00962871	0.04
$\hat{f}_7$	0.01044794	8.47

Savage-Dickey Density Ratio

Suppose that M_2 is described by

$$p(y|\omega, \psi, M_2)$$

and M_1 is a restricted verison of M_2, ie.

$$p(y|\psi, M_1) \equiv p(y|\omega = \omega_0, \psi, M_2)$$

Suppose also that

$$\pi(\psi|\omega = \omega_0, M_2) = \pi(\psi|M_1)$$

Therefore, it can be proved that the Bayes factor is

$$\begin{aligned} B_{12} &= \frac{\pi(\omega = \omega_0|y, M_2)}{\pi(\omega = \omega_0|M_2)} \\ &\approx \frac{N^{-1}\sum_{n=1}^{N} \pi(\omega = \omega_0|\psi^{(n)}, y, M_2)}{\pi(\omega = \omega_0|M_2)} \end{aligned}$$

where $\{\psi^{(1)}, \ldots, \psi^{(N)}\} \sim \pi(\psi|y, M_2)$.

Example: Normality x Student-t

From Verdinelli-Wasserman (1995). Suppose that we have observations $x_1, \ldots, x_n$ and we would like to entertain two models:

$$\mathcal{M}_1 : x_i \sim N(\mu, \sigma^2) \quad \text{and} \quad \mathcal{M}_2 : x_i \sim t_\lambda(\mu, \sigma^2)$$

Letting $\omega = 1/\lambda$, $\mathcal{M}_1$ is a particular case of $\mathcal{M}_2$ when $\omega = \omega_0 = 0.0$, with $\psi = (\mu, \sigma^2)$.

Let us also assume that $\omega \sim U(0,1)$, with $\omega = 1$ corresponding to a Cauchy distribution, and that

$$\pi(\mu, \sigma^2 | \mathcal{M}_1) = \pi(\mu, \sigma^2, \omega | \mathcal{M}_2) \propto \sigma^{-2}$$

the Savage-Dickey formula holds and the Bayes factor is

$$B_{12} = \pi(\omega_0 | x, \mathcal{M}_2),$$

i.e., the marginal posterior of ω evaluated at 0.

Because $\pi(\mu, \sigma^2, \omega | x, \mathcal{M}_2)$ has no closed form solution, they use a Metropolis algorithm and sample (μ, σ^2, ω) from $q(\mu, \sigma^2, \omega)$,

$$q(\mu, \sigma^2, \omega) = \pi(\omega)\pi(\sigma^2 | x, \mathcal{M}_1)\pi(\mu | \sigma^2, x, \mathcal{M}_1)$$

ie.

$$\begin{aligned} \omega &\sim U(0,1) \\ \sigma^2 &\sim IG\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \\ \mu|\sigma^2 &\sim N(\bar{x}, \sigma^2/n) \end{aligned}$$

When $n = 100$ from $N(0,1)$, then $B_{12} = 3.79$ (standard error=0.145)

When $n = 100$ from $Cauchy(0,1)$, then $B_{12} = 0.000405$ (standard error=0.000240)

Reversible jump MCMC

Suppose that the competing models can be enumerable and are represented by the set $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$. Under model $\mathcal{M}_k$, the posterior distribution is

$$p(\theta_k|y, k) \propto p(y|\theta_k, k)p(\theta_k|k)$$

where $p(y|\theta_k, k)$ and $p(\theta_k|k)$ represent the probability model and the prior distribution of the parameters of model $\mathcal{M}_k$, respectively. Then,

$$p(\theta_k, k|y) \propto p(k)p(\theta_k|k, y)$$

The RJMCMC methods involve MH-type algorithms that move a simulation analysis between models defined by (k, θ_k) to $(k', \theta_{k'})$ with different defining dimensions k and k'.

The resulting Markov chain simulations jump between such distinct models and form samples from the joint distribution $p(\theta_k, k)$.

The algorithm are designed to be reversible so as to maintain detailed balance of a irreducible and aperiodic chain that converges to the correct target measure (See Green, 1995, and Gamerman and Lopes, 2006, chapter 7).

The RJMCMC algorithm

Step 0. Current state: (k, θ_k)

Step 1. Sample $\mathcal{M}_{k'}$ from $J(k \rightarrow k')$.

Step 2. Sample u from $q(u|\theta_k, k, k')$.

Step 3. Set $(\theta_{k'}, u') = g_{k,k'}(\theta_k, u)$, where $g_{k,k'}(\cdot)$ is a bijection between (θ_k, u) and $(\theta_{k'}, u')$, where u and u' play the role of matching the dimensions of both vectors.

Step 4. The acceptance probability of the new model, $(\theta_{k'}, k')$ can be calculated as the minimum between one and

$$\underbrace{\frac{p(y|\theta_{k'}, k')p(\theta_{k'})p(k')}{p(y|\theta_k, k)p(\theta_k)p(k)}}_{\text{model ratio}} \underbrace{\frac{J(k' \rightarrow k)q(u'|\theta_{k'}, k', k)}{J(k \rightarrow k')q(u|\theta_k, k, k')} \left| \frac{\partial g_{k,k'}(\theta_k, u)}{\partial(\theta_k, u)} \right|}_{\text{proposal ratio}}$$

$\widehat{p}(k|y)$

Looping through steps 1-4 above L times produces a sample $\{k_l, l = 1, \ldots, L\}$ for the model indicators and $Pr(k|y)$ can be estimated by

$$\widehat{P}r(k|y) = \frac{1}{L}\sum_{l=1}^{L} 1_k(k_l)$$

where $1_k(k_l) = 1$ if $k = k_l$ and zero otherwise.

Choice of $q(u|k, \theta_k, k')$

The choice of the model proposal probabilities, $J(k \rightarrow k')$, and the proposal densities, $q(u|k, \theta_k, k')$, must be cautiously made, especially in highly parameterized problems.

Independent sampler: If all parameters of the proposed model are generated from the proposal distribution, then $(\theta_{k'}, u') = (u, \theta_k)$ and the Jacobian is one.

Standard Metropolis-Hastings: When the proposed model k' equals the current model k, the loop through steps 1-4 corresponds to the traditional Metropolis-Hastings algorithm.

Optimal choice

If $p(\theta_k|y,k)$ is available in close form for each model $\mathcal{M}_k$, then $q(u'|\theta_{k'},k',k) = p(\theta_k|y,k)$ and the acceptance probability reduces to the minimum between one and

$$\frac{p(k')p(y|k')}{p(k)p(y|k)}\frac{J(k' \to k)}{J(k \to k')}$$

since $p(y|\theta_k,k)p(\theta_k)p(k) = p(\theta_k,k|y)p(y|k)$.

The Jacobian equals one and $p(y|k)$ is available in close form.

If $J(k' \to k) = J(k \to k')$, then the acceptance probability is the posterior odds ratio from model $\mathcal{M}_{k'}$ to model $\mathcal{M}_k$.

In this case, the move is automatically accepted when model $\mathcal{M}_{k'}$ has higher posterior probability than model $\mathcal{M}_k$; otherwise the posterior odds ratio determines how likely is to move to a lower posterior probability model.

Metropolized Carlin-Chib

Let $\Theta = (\theta_k, \theta_{-k})$ be the vector containing the parameters of all competing models. Then the joint posterior of (Θ, k) is

$$p(\Theta, k|y) \propto p(k)p(y|\theta_k, k)p(\theta_k|k)p(\theta_{-k}|\theta_k, k)$$

where $p(\theta_{-k}|\theta_k, k)$ are *pseudo-prior* densities.

Carlin and Chib (1995) proposed a Gibbs sampler where the full posterior conditional distributions are

$$p(\theta_k|y, k, \theta_{-k}) \propto \begin{cases} p(y|\theta_k, k)p(\theta_k|k) & \text{if } k = k' \\ p(\theta_k|k') & \text{if } k = k' \end{cases}$$

and

$$p(k|\Theta, y) \propto p(y|\theta_k, k)p(k) \prod_{m \in \mathcal{M}} p(\theta_m|k)$$

Notice that the pseudo-prior densities and the RJMCMC's proposal densities have similar functions.

As a matter of fact, Carlin and Chib (1995) suggest using pseudo-prior distributions that are close to the posterior distributions within each competing model.

The main problem with Carlin and Chib's Gibbs sampler is the need of evaluating and drawing from the pseudo-prior distributions at each iteration of the MCMC scheme.

This problem can be overwhelmingly exacerbated in large situations where the number of competing models is relatively large.

Dellaportas, Forster and Ntzoufras (1998) and Godsill (1998) proposed "Metropolizing" Carlin and Chib's Gibbs sampler:

Step 0. Current state: (θ_k, k)

Step 1. Sample $\mathcal{M}_{k'}$ from $J(k \rightarrow k')$;

Step 2. Sample $\theta_{k'}$ from $p(\theta_{k'}|k)$;

Step 3. The acceptance probability is $min(1, A)$

$$
\begin{aligned}
A &= \frac{p(y|\theta_{k'}, k')p(k')J(k' \rightarrow k)\prod_{m\in\mathcal{M}} p(\theta_m|k')}{p(y|\theta_k, k)p(k)J(k \rightarrow k')\prod_{m\in\mathcal{M}} p(\theta_m|k)} \\
&= \frac{p(y|\theta_{k'}, k')p(k')J(k' \rightarrow k)p(\theta_{k'}|k')p(\theta_k|k')}{p(y|\theta_k, k)p(k)J(k \rightarrow k')p(\theta_k|k)p(\theta_{k'}|k)}.
\end{aligned}
$$

Pseudo-priors and RJMCMC's proposals play similar roles and the closer their are to the competing models' posterior probabilities the better the sampler mixing.

Bayesian Model Averaging

See Hoeting, Madigan, Raftery and Volinsky (1999), *Statistical Science*, 14, 382-401.

Let $\mathcal{M}$ denote the set that indexes all entertained models.

Assume that Δ is an outcome of interest, such as the future value y_{t+k}, or an elasticity well defined across models, etc. The posterior distribution for Δ is

$$p(\Delta|y) = \sum_{m \in \mathcal{M}} p(\Delta|m, y) Pr(m|y)$$

for data y and posterior model probability

$$Pr(m|y) = \frac{p(y|m)Pr(m)}{p(y)}$$

where $Pr(m)$ is the prior probability model.

Posterior predictive criterion

Gelfand and Ghosh (1998) introduced a posterior predictive criterion that, under squared error loss, favors the model M_j which minimizes

$$D_j^G = P_j^G + G_j^G$$

where

$$P_j^G = \sum_{t=1}^{n} V(\tilde{y}_t | y, M_j)$$

$$G_j^G = \sum_{t=1}^{n} [y_t - E(\tilde{y}_t | y, M_j)]^2$$

and $(\tilde{y}_1, \ldots, \tilde{y}_n)$ are predictions/replicates of y.

The first term, P_j, is a penalty term for model complexity.

The second term, G_j, accounts for goodness of fit.

More general losses

Gelfand and Ghosh (1998) also derived the criteria for more general error loss functions.

Expectations $E(\tilde{y}_t|y, M_j)$ and variances $V(\tilde{y}_t|y, M_j)$ are computed under posterior predictive densities, ie.

$$E[h(\tilde{y}_t)|y, M_j] \quad = \quad \int\int h(\tilde{y}_t)f(\tilde{y}_t|y, \theta_j, M_j)\pi(\theta_j|M_j)d\theta_j d\tilde{y}_t$$

for $h(\tilde{y}_t) = \tilde{y}_t$ and $h(\tilde{y}_t) = \tilde{y}_t^2$.

The above integral can be approximated via Monte Carlo.

Deviance Information Criterion

See Spiegelhalter, Best, Carlin and van der Linde (2002), *JRSS-B*, 64, 583-616.

If $\theta^* = E(\theta|y)$ and $D(\theta) = -2\log p(y|\theta)$ is the deviance, then the DIC generalizes the AIC

$$
\begin{aligned}
DIC &= \bar{D} + p_D \\
&= \text{goodness of fit} + \text{model complexity}
\end{aligned}
$$

where $\bar{D} = E_{\theta|y}(D(\theta))$ and $p_D = \bar{D} - D(\theta^*)$.

The p_D is the *effective number of parameters*.

Small values of DIC suggests a better-fitting model.

DIC is computationally attractive criterion since its two terms can be easily computed during an MCMC run.

Let $\theta^{(1)}, \ldots, \theta^{(M)}$ be an MCMC sample from $p(\theta|y)$.

Then,

$$\begin{aligned}\bar{D} &\approx \frac{1}{M}\sum_{i=1}^{M} D\left(\theta^{(i)}\right)\\ &= -2M^{-1}\sum_{i=1}^{M}\log p\left(y|\theta^{(i)}\right)\end{aligned}$$

and

$$D(\theta^*) \approx D(\bar{\theta}) = -2\log p(y|\bar{\theta})$$

where $\bar{\theta} = M^{-1}\sum_{i=1}^{M}\theta^{(i)}$.

Example: cycles-to-failure times

Cycles-to-failure times for airplane yarns.

86	146	251	653	98	249	400	292	131
169	175	176	76	264	15	364	195	262
88	264	157	220	42	321	180	198	38
20	61	121	282	224	149	180	325	250
196	90	229	166	38	337	65	151	341
40	40	135	597	246	211	180	93	315
353	571	124	279	81	186	497	182	423
185	229	400	338	290	398	71	246	185
188	568	55	55	61	244	20	284	393
396	203	829	239	236	286	194	277	143
198	264	105	203	124	137	135	350	193
188								

Gamma, log-normal and Weibull models:

$$\begin{aligned} M_1 &: y_i \sim G(\alpha,\beta), & \alpha,\beta > 0 \\ M_2 &: y_i \sim LN(\mu,\sigma^2), & \mu \in R, \sigma^2 > 0 \\ M_3 &: y_i \sim Weibull(\gamma,\delta) & \gamma,\delta > 0, \end{aligned}$$

for $i = 1, \ldots, n$.

Under model M_2, μ and σ^2 are the mean and the variance of $\log\ y_i$, respectively.

Under model M_3, $p(y_i|\gamma,\delta) = \gamma y_i^{\gamma-1}\delta^{-\gamma}e^{-(y_i/\delta)^\gamma}$.

Flat priors were considered for

$$
\begin{aligned}
\theta_1 &= (\log\alpha, \log\beta) \\
\theta_2 &= (\mu, \log\sigma^2) \\
\theta_3 &= (\log\gamma, \log\delta)
\end{aligned}
$$

It is also easy to see that,

$$
\begin{aligned}
E(y|\theta_1, M_1) &= \alpha/\beta \\
V(y|\theta_1, M_1) &= \alpha/\beta^2 \\
E(y|\theta_2, M_2) &= \exp\{\mu + 0.5\sigma^2\} \\
V(y|\theta_2, M_2) &= \exp\{2\mu + \sigma^2\}(e^{\sigma^2} - 1) \\
E(y|\theta_3, M_3) &= \delta\Gamma(1/\gamma)/\gamma \\
V(y|\theta_3, M_3) &= \delta^2\left[2\Gamma(2/\gamma) - \Gamma(1/\gamma)^2/\gamma\right]/\gamma
\end{aligned}
$$

Weighted resampling schemes, with bivariate normal importance functions, were used to sample from the posterior distributions.

Proposals: $q_i(\theta_i) = f_N(\theta_i; \tilde{\theta}_i, V_i)$

$$
\begin{aligned}
\tilde{\theta}_1 &= (0.15, 0.2)' \\
\tilde{\theta}_2 &= (5.16, -0.26)' \\
\tilde{\theta}_3 &= (0.47, 5.51)' \\
V_1 &= \text{diag}\,(0.15, 0.2) \\
V_2 &= \text{diag}\,(0.087, 0.085) \\
V_3 &= \text{diag}\,(0.087, 0.101)
\end{aligned}
$$

Posterior means, standard deviations and 95% credibility intervals:

M1

- α: $2.24, 0.21, (1.84, 2.68)$
- β: $0.01, 0.001, (0.008, 0.012)$

M2

- α: $5.16, 0.06,\ (5.05,\ 5.27)$
- β: $0.77, 0.04, (0.69, 0.86)$

M3

- α: $1.60, 0.09, (1.42, 1.79)$
- β: $248.71, 13.88, (222.47, 276.62)$

DIC indicates that both the Gamma and the Weibull models are relatively similar with the Weibull model performing slightly better.

	Model	DIC
M_1	Gamma	1253.445
M_2	Log-normal	1265.842
M_3	Weibull	1253.051