Remarks on Propriety of the Posterior Distribution for GLM’s

1. **Theorem:**
   If $\pi(\beta) \propto 1$ and the MLE of $\beta$ exists for the GLM, then $p(\beta \mid y)$ is proper.

2. With Jeffreys’s prior, i.e., $\pi(\beta) \propto |X'\Delta V \Delta X|^{1/2}$, it can be shown that $p(\beta \mid y)$ is proper under some very general conditions. Note that Jeffreys’s prior is improper for the Poisson, gamma, and normal models. See Ibrahim and Laud (1991, *JASA*) for more details.

3. If $\pi(\beta)$ is proper, then of course $p(\beta \mid y)$ is always proper for any GLM.

Not all posterior densities are log-concave. In those situations, the adaptive rejection algorithm cannot be used. Therefore Gilks, Best, Tan (1995, *Applied Statistics*) developed another algorithm to handle the non log-concave case. The algorithm involves introducing a Metropolis step within the Gibbs sampler. They call the algorithm **Adaptive Rejection Metropolis Sampling** (ARMS). These algorithms have become popular because computer code has been provided by the authors for implementing these algorithms. The computer code is quite general and can be used for many types of problems.

The C code and documentation for ARMS is available from Gilks at wally.gilks@mrc-bsu.cam.ac.uk The idea of ARMS is that it uses the Adaptive Rejection Sampling
(ARS) part of the algorithm to construct a proposal density for the Metropolis algorithm. Thus ARS is used to create a good proposal density. Then ARS is appended to a single Metropolis-Hastings step, thus creating an ARMS within Gibbs chain. However, unlike ARS, ARMS will not produce independent samples from the desired density.

Let $f(x)$ denote the density we wish to sample from. Let $(x, y)$ denote the complete set of variables being sampled by the Gibbs sampler. Here, $x$ is the current variable to be sampled from its full conditional density, which is proportional to $f(x)$, where we notionally suppress the conditioning on $y$. Let $X_{cur}$ denote the current value of $x$ at a given iteration of the Gibbs sampler. The aim then is to replace $x_{cur}$ with a new value $x_M$ from $f$.

For ARMS, we construct a function $h_n(x)$ which is more complicated than ARS. Recall that for ARS, the adaptive envelope is given by the following steps.
Thus $g_n(x)$ is the normalized $h_n(x)$, i.e.,
\[ \int g_n(x) \, dx = 1. \]

**Model Selection and Model Comparison for GLM’s**

The computation of Bayes factors, HPD intervals, or posterior model probabilities will require MCMC techniques since the posterior distributions are not available in a closed form. It turns out that some novel MCMC algorithms can be developed for computing posterior model probabilities, in cases in which noninformative priors or informative priors are used. We now discuss some of these methods.

A popular method for computing posterior model probabilities using noninformative (but proper) priors was developed by George and McCulloch (1993, *JASA*), and George, McCulloch, and Tsay (1996). We now discuss these methods. The method discussed by George and McCulloch (1993) was written for the linear model but was extended to GLM’s by George, McCulloch, and Tsay (1996).

Consider the model
\[ Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I). \]

They consider a prior for each $\beta_i, \beta = (\beta_1, \ldots, \beta_p)'$ to be
a mixture of two normal densities, and thus

\[ \beta_i \mid \gamma_i \sim (1 - \gamma_i) \ N(0, \tau_i^2) + \gamma_i \ N(0, c_i^2 \tau_i^2) \]

where \( \gamma_i \) is a binary random variable with

\[ p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = p_i . \]

Note that when \( \gamma_i = 0 \), \( \beta_i \sim N(0, \tau_i^2) \) and when \( \gamma_i = 1 \), \( \beta_i \sim N(0, c_i^2 \tau_i^2) \). The interpretation of this is as follows. Set \( \tau_i \) \((\tau_i > 0)\) small so that if \( \gamma_i = 0 \), then \( \beta_i \) would probably be so small that it could “safely” be estimated by 0. Second, if \( c_i \) \((c_i > 1 \text{ always})\) is set large so that if \( \gamma_i = 1 \), then a non-zero estimate of \( \beta_i \) be probably included in the final model. Thus, the user must specify \((\tau_i, c_i), \ i = 1, \ldots, p\). Note here, that a priori, the \( \beta_i \)'s are not necessarily independent. Based on this interpretation, \( p_i \) may be thought of as the prior probability that \( \beta_i \) is not zero, or equivalently that \( X_i \) should be included in the model, where \( X_i \) denotes the \( i^{th} \) covariate.

The mixture prior for \( \beta_i \mid \gamma_i \) can be written in vector form as

\[ \beta \mid \gamma \sim N_p(0, D_\gamma RD_\gamma) , \]

where \( \gamma = (\gamma_1 \ldots, \gamma_p) \), \( R \) is the prior correlation matrix and

\[ D_\gamma = \text{diag}(a_1 \tau_i, \ldots, a_p \tau_p) , \]

where \( a_i = 1 \) if \( \gamma_i = 0 \) and \( a_i = c_i \) if \( \gamma_i = 1 \). Thus \( D_\gamma \) determines the scaling of the prior covariance matrix.
The final piece of the model is to specify a prior for $\sigma^2$. We take
\[
\sigma^2 | \gamma \sim IG \left( \frac{v_\gamma}{2}, \frac{v_\gamma \lambda_\gamma}{2} \right).
\]
The main reason for embedding the normal linear model in this hierarchical model is to obtain the marginal posterior of $\gamma$, i.e., $f(\gamma \mid y)$, where $f(\gamma \mid y) \propto f(y \mid \gamma) f(\gamma)$. The marginal posterior of $\gamma$ contains the relevant information about variable selection.

The prior for $\gamma$ can be taken as
\[
f(\gamma) = \prod_{i=1}^{p} p_i^{\gamma_i} (1 - p_i)^{1-\gamma_i}.
\]
Choices of $\tau_i$ and $c_i$ are motivated in Section 2.2 of the paper. To help guide the choice of $c_i$, it is useful to observe that the densities of $N(0, \tau_i^2)$ and $N(0, c_i^2 \tau_i^2)$ intersect at
\[
\xi(c_i) \tau_i \text{ when } \xi(c_i) = \sqrt{\frac{2 \log(c_i)c_i^2}{c_i^2-1}}.
\]
This implies that the density of $N(0, c_i^2 \tau_i^2)$ will be larger than the density of $N(0, \tau_i^2)$ if and only if $| \beta_i | > \xi(c_i) \tau_i$. It may also be useful to observe that $c_i$ is the ratio of the heights of $N(0, \tau_i^2)$ and $N(0, c_i^2 \tau_i^2)$ at 0. Thus $c_i$ can be interpreted as the prior odds that $X_i$ should be excluded when $\beta_i$ is very close to 0 (see Section 2.2). Choices for $R$ include $R = I$ or $R \propto (X'X)^{-1}$. See Sections (2.3 and 2.4).
Gibbs Sampling the Subsets

George and McCulloch (1993) propose a stochastic search variable selection (SSVS) to find the best possible subsets. The first part of SSVS entails specifying the hierarchical normal mixture model so that \( f(\gamma | y) \) puts most weight on the more “promising” subsets of predictors. The second part of SSVS entails extracting this information. Rather than calculate all \( 2^p \) posterior probabilities in \( f(\gamma | y) \), SSVS uses the Gibbs sampler to generate a sequence \( \gamma^1, \ldots, \gamma^n \), which in many cases converges rapidly in distribution to \( \gamma \sim f(\gamma | y) \). Such a sequence can be obtained quickly and efficiently, with far less effort than required to compute the entire posterior. The sequence \( \gamma^1, \ldots, \gamma^m \) will, with high probability in many cases, contain exactly the information relevant to variable selection. This is because those \( \gamma \) with highest probability will also appear most frequently and hence will be easiest to identify. These \( \gamma \) that appear infrequently or not at all are simply not of interest and an be disregarded. To see how SSVS works, let \( \beta^0, \sigma^0, \gamma^0, \beta^1, \sigma^1, \gamma^1, \ldots, \beta^j, \sigma^j, \gamma^j \) be a Gibbs sequence that is generated from the following scheme.

\[
\beta^j = f(\beta^j | y, \sigma^{j-1}, \gamma^{j-1}) \tag{1}
\]

\[
= N_p((\sigma^{j-1})^{-2} A_{\gamma_{j-1}} X' X \hat{\beta}_{LS}, A_{\gamma_{j-1}})
\]

where
\[
A_{\gamma_{j-1}} = \left((\sigma^{j-1})^{-2} X'X + D^{-1}_{\gamma_{j-1}} R^{-1} D^{-1}_{\gamma_{j-1}}\right)^{-1}.
\]

Next, the variance \(\sigma^j\) is obtained by sampling

\[
\sigma^j \sim f(\sigma^j \mid y, \beta^j, \gamma^{j-1})
= IG\left(\frac{n + v_{\gamma_{j-1}}}{2}, \frac{\|Y - X\beta^j\|^2 + v_{\gamma_{j-1}} \lambda_{\gamma_{j-1}}}{2}\right)
\]

Finally, the vector \(\gamma^j\) is obtained componentwise by sampling consecutively from the conditional distribution

\[
\gamma_i^j \sim f\left(\gamma_i^j \mid y, \beta^j, \sigma^j, \gamma_{(i)}^j\right)
= f\left(\gamma_i^j \mid \beta^j, \sigma^j, \gamma_{(i)}^j\right)
\]

where \(\gamma_{(i)}^j = (\gamma_1^j, \ldots, \gamma_{i-1}^j, \gamma_{i+1}^j \ldots, \gamma_{p}^j)^{-1})\).

Notice that the distribution of \(\gamma_i^j\) does \textbf{not} depend on \(y\).

This simplification reduces computational requirements and allows for fast convergence of the sequence \(\gamma^1, \ldots, \gamma^m\). Each \(\gamma_i^j\) is Bernoulli with

\[
p\left(\gamma_i^j = 1 \mid \beta^j, \sigma^j, \gamma_{(i)}^j\right) = \frac{a}{a + b},
\]

where

\[
a = f(\beta^j \mid \gamma_{(i)}^j, \gamma_i^j = 1) f(\sigma^j \mid \gamma_{(i)}^j, \gamma_i^j = 1) f(\gamma_{(i)}^j, \gamma_i^j = 1)
\]
and
\[ b = f(\beta^j | \gamma^j_{(i)}, \gamma^j_{i} = 0) \cdot f(\sigma^j | \gamma^j_{(i)}, \gamma^j_{i} = 0) \cdot f(\gamma^j_{(i)}, \gamma^j_{i} = 0). \]

By repeated successive sampling from (1), (2), (3) the Gibbs sequence
\[ \{ \beta^j, \sigma^j, \gamma^j, \quad j = 0, 1, \ldots \} \]
is obtained. As the length of this sequence is increased, the empirical distribution of the realized values of \( \gamma \) will converge to the actual posterior of \( f(\gamma | y) \).

The generalization of this method to GLM’s is straightforward. The same priors are used, except we change \( p(y | x, \beta) \) to a GLM. That is
\[
p(y | x, \beta) \propto \exp \left[ \sum_{i=1}^{n} \{ y_i \theta(x_i' \beta) - b(\theta(x_i' \beta)) \} \right].
\]
The advantages of SSVS are that
- it can handle large problems
- it is fast and efficient

Disadvantages are that it does not use informative priors, nor allows incorporation of useful prior information.

Mitchell and Beauchamp (1988, \textit{JASA}) also propose a method for doing Bayesian variable selection for the linear regression model using noninformative proper priors called
**spike and slab** priors. Their method requires computation of all $2^p$ posterior model probabilities, and hence is not as popular as the method of George and McCulloch (1993).


The basic problem is described as follows. Let $f_i(\theta)$, $i = 1, 2$ be two densities, each of which is known up to a normalizing constant. That is,

$$f_i(\theta) = \frac{f_i^*(\theta)}{c_i}, \quad \theta \in \Theta_i$$

where $\Theta_i$ is the parameter space for $f_i$, $i = 1, 2$.

The problem of estimating the ratio of two normalizing constants arises in the computation of Bayes factors and posterior model probabilities.

$$BF = \frac{\int p(y \mid \theta_1) \pi(\theta_1) \, d\theta_1}{\int p(y \mid \theta_2) \pi(\theta_2) \, d\theta_2}$$
\[ p(m \mid y) = \frac{c_1}{c_2}, \]

where \( c_1 \) is the normalizing constant of the posterior distribution of \([\theta_1 \mid y]\) and \( c_2 \) for \([\theta_2 \mid y]\). Here \( \theta_1 \) and \( \theta_2 \) denote parameter vectors from two different models. Also, in the computation of posterior model probabilities, we have

\[
P(m \mid y) = \frac{p(y \mid m) \ p(m)}{\sum_{m \in M} p(y \mid m) \ p(m)} \]

\[
= \frac{p(y \mid m)}{p(y \mid m^*)} \frac{p(m)}{p(y \mid m^*)} \]

\[
= \frac{c_m}{c_{m^*}} \ p(m) \]

\[
\sum_{m \in M} \left( \frac{c_m}{c_{m^*}} \right) \ p(m) .
\]

Let \( r = \) ratio of the two normalizing constants \( c_1 \) and \( c_2 \). That is,

\[ r = \frac{c_1}{c_2} . \]

We wish to estimate \( r \), where \( c_i \) is the normalizing constant of density \( f_i(\theta) \), i.e.,

\[ c_i = \int_{\Theta_i} f_i^*(\theta) \ d\theta . \]

We will discuss several methods for estimating \( r \).
Importance Sampling

This is the standard method, in which

\[ \hat{r} = \frac{1}{n_1} \sum_{j=1}^{n_1} \frac{f_1^*(\theta_{1j})}{g_1(\theta_{1j})} / \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{f_2^*(\theta_{2j})}{g_2(\theta_{2j})}, \]

where \( g_1 \) and \( g_2 \) are two importance sampling densities.

Alternatively we can view \( r \) as an expectation

\[ r = \frac{c_1}{c_2} = E_2 \left[ \frac{f_1^*(\theta)}{f_2^*(\theta)} \right] \]

\[ = \int \frac{f_1^*(\theta)}{f_2^*(\theta)} f_2(\theta) \, d\theta. \quad (4) \]

If we can sample from \( f_2(\theta) \), then \( r \) can be estimated by

\[ \hat{r} = \frac{1}{n} \sum_{i=1}^{n} \frac{f_1^*(\theta_{2i})}{f_2^*(\theta_{2i})}, \]

where \( \theta_{21}, \theta_{22}, \ldots, \theta_{2n} \) is a sample from \( f_2(\theta) \). We note that if the two densities \( f_1(\theta) \) and \( f_2(\theta) \) have little overlap, the importance sampling procedure will work poorly.
Bridge Sampling

Bridge sampling uses a generalization of (4) and writes

\[ r = \frac{c_1}{c_2} = \frac{E_2[f_1^*(\theta)\alpha(\theta)]}{E_1[f_2^*(\theta)\alpha(\theta)]}, \]

where \( \alpha \) is an arbitrary function defined on \( \Theta_1 \cap \Theta_2 \) such that

\[ 0 < \left| \int_{\Theta_1 \cap \Theta_2} \alpha(\theta) f_1^*(\theta) f_2^*(\theta) \, d\theta \right| < \infty. \]

Then, letting \( \theta_{i1}, \theta_{i2}, \ldots \theta_{in_i} \) be a random sample from \( f_i(\theta), \ i = 1, 2 \), we have

\[ \hat{r}_\alpha = \frac{n_2^{-1} \sum_{i=1}^{n_2} f_1^*(\theta_{2i}) \alpha(\theta_{2i})}{n_1^{-1} \sum_{i=1}^{n_1} f_2^*(\theta_{1i}) \alpha(\theta_{1i})}. \]

Let \( n = n_1 + n_2 \) and \( s_i = n_i/n, \ i = 1, 2 \), and assume that \( \lim_{n \to \infty} (s_i) > 0, \ i = 1, 2 \). Meng and Wong (1996) showed that the optimal choice of \( \alpha \) is given by

\[ \alpha_{opt} = \frac{c}{s_1 f_1(\theta) + s_2 f_2(\theta)}, \ \theta \in \Theta_1 \cap \Theta_2, \]

where \( f_i(\theta) = f_i^*(\theta)/c_i, \ i = 1, 2 \). \( \alpha_{opt} \) minimizes the relative mean square error

\[ \text{RE}^2(\hat{r}_\alpha) = \frac{E(\hat{r}_\alpha - r)^2}{r^2}. \]

Since \( c_1, c_2 \) are unknown, \( \alpha_{opt}(\theta) \) is not directly usable. As an alternative, Meng and Wong (1996) constructed the
following iterative estimator,

\[
\hat{r}^{(t+1)}_{opt} = \frac{\frac{1}{n_2} \sum_{i=1}^{n_2} f_1^*(\theta_{2i})/(s_1 f_1^*(\theta_{2i}) + s_2 \hat{r}_{opt}^{(t)} f_2^*(\theta_{2i}))}{\frac{1}{n_1} \sum_{i=1}^{n_1} f_2^*(\theta_{1i})/(s_1 f_1^*(\theta_{1i}) + s_2 \hat{r}_{opt}^{(t)} f_2^*(\theta_{1i}))}.
\]

They showed that for each \( t \geq 0 \), \( \hat{r}^{(t+1)}_{opt} \) provides a consistent estimator of \( r \), and that the unique limit, \( \hat{r}_{opt} \), achieves the asymptotic minimal relative mean-squared error. Meng and Wong (1996) also considered several non-iterative choices for \( \alpha(\theta) \), such as \( \alpha(\theta) = 1 \), 

\[
\alpha(\theta) = (f_1^*(\theta) f_2^*(\theta))^{-1/2}, \text{ and } \alpha(\theta) = (f_1^*(\theta) + f_2^*(\theta))^{-1}.
\]

The bridge sampling estimator will be unstable when \( f_1^*(\theta) \) and \( f_2^*(\theta) \) have little overlap. For such cases, the path sampling method of Gelman and Meng (1994) will substantially improve the simulation efficiency.

**Path Sampling**

Instead of trying to estimate \( r = \frac{c_1}{c_2} \) directly, Gelman and Meng (1994) proposed the path sampling method to estimate the logarithm of \( r \), that is,

\[
\xi = -\log(r) = -\log \left( \frac{c_1}{c_2} \right).
\]

Suppose we write

\[
f_1(\theta) = \frac{f_1^*(\theta|\lambda_1)}{c_1(\lambda_1)},
\]
\[ f_2(\theta) = \frac{f_2^*(\theta|\lambda_2)}{c_2(\lambda_2)} , \]

where \( \lambda_1 \) and \( \lambda_2 \) are hyperparameters that index \( f_1(\theta) \) and \( f_2(\theta) \). Suppose there exists a \( \lambda \), such that when \( \lambda_1 = \lambda_2 = \lambda \), we have

\[ f_1(\theta) = \frac{f_1^*(\theta|\lambda_1)}{c_1(\lambda_1)} = \frac{f_2^*(\theta|\lambda_2)}{c_2(\lambda_2)} = f_2(\theta) . \]

Denote this common density by \( p(\theta|\lambda) \), and let \( p^*(\theta|\lambda) \) denote the unnormalized \( p(\theta|\lambda) \). That is,

\[ p(\theta|\lambda) = \frac{p^*(\theta|\lambda)}{c(\lambda)} . \]

Let

\[ U(\theta, \lambda) = \frac{\partial}{\partial \lambda} \left( \log[p^*(\theta|\lambda)] \right) , \]

and let \( \pi(\lambda) \) be a prior for \( \lambda \) on \([0, 1]\), and thus \( 0 \leq \lambda \leq 1 \). Gelman and Meng (1994) showed that

\[ \xi = -\log \left( \frac{c_1}{c_2} \right) = E \left[ \frac{U(\theta, \lambda)}{\pi(\lambda)} \right] , \]

where the expectation is with respect to the joint density

\[ f(\theta, \lambda) = f(\theta|\lambda)\pi(\lambda) . \]

Now let \( (\theta_i, \lambda_i), i = 1, \ldots, n \) be a random sample from \( f(\theta, \lambda) \). Then

\[ \hat{\xi} = \frac{1}{n} \sum_{i=1}^{n} \frac{U(\theta_i, \lambda_i)}{\pi(\lambda_i)} . \]
Here, $f(\theta, \lambda)$ serves as a **continuous path** to link $f_1(\theta|\lambda_1)$ and $f_2(\theta|\lambda_2)$. The Monte Carlo variance of $\hat{\xi}$ is

$$\text{Var}(\hat{\xi}) = \frac{1}{n} \left( \int_{0}^{1} \frac{E[U^2(\theta, \lambda)]}{\pi(\lambda)} \, d\lambda - \xi^2 \right),$$

where the expectation is taken with respect to $f(\theta|\lambda)$.

**Marginal Likelihood**

Let

$$m(x) = \int_{\Theta} L(\theta) \, \pi(\theta) \, d\theta$$

denote the normalizing constant of the posterior distribution of $\theta$. Chib (1995, JASA) considered the following identity:

$$m(x) = \frac{L(\theta)\pi(\theta)}{p(\theta|x)}.$$

Let $\theta^*$ be the posterior mean (or mode) of $\theta$, and let $\hat{p}(\theta^*|x)$ be an estimator for the posterior density evaluated at $\theta^*$. Chib (1995) obtained the following estimator for $m(x)$:

$$\hat{m}(x) = \frac{L(\theta^*)\pi(\theta^*)}{\hat{p}(\theta^*|x)}.$$

Then Chib (1995) developed a **data augmentation** scheme (see Tanner and Wong, 1987, JASA) to estimate $\hat{p}(\theta^*|x)$ by introducing latent variables. Chib’s method is particularly useful for multivariate problems when the full conditional densities are completely known.
\[ \hat{p}(\theta^*|x) \] can also be estimated by using the **Importance-Weighted Marginal Density Estimation** (IWMDE) method of Chen (1994, JASA). The IWMDE does not require completely known full conditional densities. Further, the IWMDE method can be used to estimate \( m(x) \) directly.

Let \( \theta_i, i = 1, \ldots, n \) be a sample from \( p(\theta|x) \). Such a sample can be obtained by MCMC methods, such as the Gibbs sampler or by the Metropolis-Hastings algorithm. Then IWMDE yields a consistent estimator for \( m(x) \), given by

\[
\tilde{m}_{IWMDE}(x) = \left( \frac{1}{n} \sum_{i=1}^{n} \frac{w(\theta_i)}{L(\theta_i)\pi(\theta_i)} \right)^{-1},
\]

where \( w(\theta) \) is a weighted completely known density function with the same support as that of the posterior distribution. Chen (1994) discusses how to pick a good \( w \). Chen’s method for estimating \( m(x) \) does not require obtaining a \( \theta^* \). Chib’s estimator works well if \( \hat{p}(\theta^*|x) \) is a good approximation and \( \theta^* \) is a good estimator for the mean (or mode). Chen’s method works well if a good \( w(\theta) \) is chosen.

**Ratio Importance Sampling (RIS)**

Chen and Shao (1997, Annals of Statistics) propose a new method called **Ratio Importance Sampling** (RIS) to estimate \( r = \frac{c_1}{c_2} \). Let \( \Theta = \Theta_1 \cup \Theta_2 \) denote the entire
parameter space and let \( g(\theta) \) be an arbitrary density over \( \Theta \). Then

\[
    r = \frac{c_1}{c_2} = \frac{E_g[f^*_1(\theta)/g(\theta)]}{E_g[f^*_2(\theta)/g(\theta)]},
\]

where \( f^*_1(\theta) \) and \( f^*_2(\theta) \) are the unnormalized densities of interest, and the expectation is taken with respect to \( g(\theta) \).

Note that if \( g(\theta) = \frac{f^*_2(\theta)}{c_2} \), then Chen and Shao's identity leads to the identity

\[
    r = \frac{c_1}{c_2} = E_{f_2} \left[ \frac{f^*_1(\theta)}{f^*_2(\theta)} \right],
\]

where the expectation is with respect to \( f_2(\theta) \). Thus (6) can be estimated by

\[
    \hat{r} = \frac{1}{n} \sum_{i=1}^{n} \frac{f^*_1(\theta_{2i})}{f^*_2(\theta_{2i})},
\]

where \( \theta_{21}, \ldots, \theta_{2n} \) is a random sample from \( g(\theta) = f_2(\theta) = \frac{f^*_2(\theta)}{c_2} \).

Therefore, (5) is a generalization of (6). Given samples \( \theta_1, \ldots, \theta_n \) from \( g(\theta) \), the RIS estimator for \( r \) is

\[
    \hat{r}_g = \frac{\sum f^*_1(\theta_i)/g(\theta_i)}{\sum f^*_2(\theta_i)/g(\theta_i)}.
\]

It can be shown that \( \hat{r}_g \) is a consistent estimator of \( r \). Chen and Shao (1997) derive the optimal \( g(\theta) \), denoted \( g_{opt}(\theta) \), which maximizes the expected relative mean
squared error.

We now adapt the methods of Chen (1994) and Chen and Shao (1997) to variable subset selection. Let $\mathcal{M}$ denote the model space, $m$ is an arbitrary model in $\mathcal{M}$, and let $m^*$ denote the full model. Let $\beta^{(m)}$ denote the $k_m \times 1$ the regression coefficient vector for model $m$, and let $\beta \equiv \beta^{(m^*)}$ for ease of notation.

Write $\beta = (\beta^{(-m)}, \beta^{(m)})$, where $\beta^{(-m)}$ is the part of $\beta$ not in $\beta^{(m)}$. Using the result of Chen and Shao (1997), we have

$$\frac{c_m}{c_{m^*}} = \frac{p(y|m)}{p(y|m^*)} = E_{\beta|y} \left[ \frac{L(\beta^{(m)}) \pi(\beta^{(m)}) w(\beta^{(-m)}|\beta^{(m)})}{L(\beta) \pi(\beta)} \right],$$

where the expectation is taken with respect to the posterior density of $[\beta|y]$, i.e., the full model posterior density.

The weight function $w(\beta^{(-m)}|\beta^{(m)})$ is a completely known conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$. Chen (1994) shows that the best choice of $w(\beta^{(-m)}|\beta^{(m)})$ is

$$w(\beta^{(-m)}|\beta^{(m)}) = p(\beta^{(-m)}|\beta^{(m)}, y).$$

Since a closed form for $p(\beta^{(-m)}|\beta^{(m)}, y)$ is typically not available, we follow an empirical procedure provided by Chen (1994) to select $w(\beta^{(-m)}|\beta^{(m)})$.

Specifically, using the posterior sample

$$\{\beta(j), j = 1, \ldots, N\},$$

we construct the posterior mean and
covariance matrix, denoted \((\tilde{\beta}, \tilde{\Sigma})\), and then we choose 
\(w(\beta(-m) | \beta(m))\) to be the conditional density of the normal distribution \(N_p(\tilde{\beta}, \tilde{\Sigma})\) for \(\beta(-m) | \beta(m)\). Thus following the Monte Carlo method of Chen and Shao (1997) and using the posterior sample \(\{\beta(j), j = 1, \ldots, N\}\) from the full model, the posterior probability of model \(m\) can be estimated by 

\[
\hat{p}(m|y) = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{L(\beta(j)^{(m)}) \pi(\beta(j)^{(m)}) w(\beta(j)^{(-m)} | \beta(j)^{(m)})}{L(\beta(j)^{(j)}) \pi(\beta(j)^{(j)})} \right) p(m) \\
\sum_{k=1}^{2^p} \frac{1}{N} \sum_{j=1}^{N} \left( \frac{L(\beta(j)^{(k)}) \pi(\beta(j)^{(k)}) w(\beta(j)^{(-k)} | \beta(j)^{(k)})}{L(\beta(j)^{(j)}) \pi(\beta(j)^{(j)})} \right) p(k)
\]

for \(m = 1, \ldots, 2^p\), where \(\beta\) is \(p \times 1\), and 
\(\beta(j) = (\beta(j)^{(-m)}, \beta(j)^{(m)})\).

The weight function \(w(\beta(-m) | \beta(m))\) has the form 

\[
w(\beta(-m) | \beta(m)) = (2\pi)^{-\frac{(p-k_m)}{2}} |\Sigma_{11.2m}|^{-\frac{1}{2}} \\
x \exp \left\{ -\frac{1}{2} (\beta(-m) - \tilde{\mu}_{11.2m})' \Sigma_{11.2m}^{-1} (\beta(-m) - \tilde{\mu}_{11.2m}) \right\}
\]

where 

\[
\Sigma_{11.2m} = \tilde{\Sigma}_{11m} - \tilde{\Sigma}_{12m} \tilde{\Sigma}_{22m}^{-1} \tilde{\Sigma}_{12m}'
\]

\(\tilde{\Sigma}_{11m}\) is the covariance matrix from the marginal distribution of \(\beta(-m)\), \(\tilde{\Sigma}_{12m}\) consists of the covariances between \(\beta(-m)\) and \(\beta(m)\), and \(\tilde{\Sigma}_{22m}\) is the covariance matrix of the marginal distribution of \(\beta(m)\) with respect to the joint normal distribution \(N_p(\tilde{\beta}, \tilde{\Sigma})\) for the vector
\( \beta = \beta^{(m^*)} \). Also, we note that
\[
\hat{\mu}_{11.2m} = \hat{\mu}^{(-m)} + \tilde{\Sigma}_{12m} \Sigma_{22m}^{-1} (\beta^{(m)} - \hat{\mu}^{(m)}) ,
\]
where \( \hat{\mu}^{(-m)} \) is the mean of the marginal distribution of \( \beta^{(-m)} \) implied by the \( N_p(\tilde{\beta}, \tilde{\Sigma}) \) distribution.

We can see why Chen’s method is quite powerful for variable selection problems.

1. We only need samples from the full model posterior distribution

2. The weight function \( w \) is not hard to choose.

The advantage of Chen’s formula for \( \hat{p}(m|y) \) is that samples only from the full model are needed to compute \( \hat{p}(m|y) \) for all \( m \in \mathcal{M} \).

In a general model selection problem, we have
\[
\frac{c_m}{c_m^*} = \frac{\sum_{j=1}^{N} \frac{L(\beta_{(j)}|m)\pi(\beta_{(j)}|m)}{g(\beta_{(j)})}}{\sum_{j=1}^{N} \frac{L(\beta_{(j)}|m^*)\pi(\beta_{(j)}|m^*)}{g(\beta_{(j)})}} ,
\]
where \( g \) can be taken to be a normal density and \( \beta_{(j)} \), \( j = 1, \ldots, N \) are samples from \( g \).

If such a \( g \) is hard to find and the models are not comparable, one could use Chen’s modification of Chib’s
method. That is, $c_m$ is estimated by

$$c_m = \left( \frac{1}{N} \sum_{j=1}^{N} \frac{w(\beta_{(j)}|m)}{L(\beta_{(j)}|m)\pi(\beta_{(j)}|m)} \right)^{-1},$$

where $\beta_{(j)}$, $j = 1, \ldots, N$ are samples from the posterior distribution of $[\beta|y,m]$.

We note that this method is quite general and can be used for other types of models as well.
Model Checking

The Bayesian model checking tools are not as plentiful as for the linear model, but there are some basic guidelines to checking a model (GLM or not) that we state here. Gelman (Chapter 4) has an excellent chapter on model checking. Here are some basic notions in model checking.

- Comparing your predictive distribution of future observations to data that have actually occurred. Construct your predictive distribution. Evaluate it at a few points and see if those look like your current data. Thus, we construct

\[ p(z \mid y, x_f) , \]

then evaluate \( p(z \mid y, x_f) \) at several \( (z, x_f) \) and observe to see if these are consistent with your data. Another check involves drawing simulated values from the predictive distribution of replicated data and compare these samples to the observed data.

General Diagnostic Statistics

We can compute general discrepancy measures such as

\[ T(y, \theta) = \Sigma (y_i - E(y_i \mid \theta))^2 / \text{Var}(y_i \mid \theta). \]

Then once \( T(y, \theta) \) is computed, a Bayesian p-value can be computed as

\[ \text{Bayes p-value} = P(T(y, \theta) \geq c \mid y) . \]
Or perhaps a better way to compute the p-value is to take random draws from the predictive distribution $z \mid y$ and compute

$$\text{Bayes p-value} = P(T(z, \theta) \geq T(y, \theta) \mid y),$$

where the probability is taken over the posterior of $\theta$. Yet another version of the Bayesian p-value would average over $z$, i.e.

$$\text{Bayes p-value} = \int \int I(T(z, \theta) \geq T(y, \theta)) p(\theta \mid y) p(z \mid \theta) \ d\theta \ dz.$$  

Bayesian p-values are tail areas under the posterior distribution, and thus are easy to interpret. See the Textbook pages 169-174 for more details.

- **Sensitivity Analysis**
  
  Any type of model checking must also include sensitivity analyses with regards to changing prior distributions, prior parameters, and even the likelihood. One should check sensitivity with respect to
  
  - posterior mean, mode, variance, quantiles
  - HPD regions
  - Bayes factors
  - posterior model probabilities

  A sensitivity analysis is a crucial aspect of checking a model.
In a sensitivity analysis, checking the likelihood itself is important. In some cases, the model is not good because of the likelihood. In these cases, we can consider model expansion (see Gelman et. al., p.177) by

a) adding more parameters

b) a more general distribution ($t$ instead of normal)

c) a more general model that contains other models as special cases (as in the variable selection problem)

d) consider multivariate response models instead of univariate response models
Random Effects Models

Many longitudinal studies are designed to investigate changes over time in a characteristic that is measured repeatedly for each study participant. For example, in medical studies measurements such as blood pressure, cholesterol level, or lung volume may be taken on each individual at different time points and possibly changing experimental conditions. The most common type of model for repeated measurements is the linear random effects model of Laird and Ware (1982, *Biometrics*).

The model of Laird and Ware (1982) is described as follows: For a given individual $i$ with $n_i$ repeated measurements, the random effects model for outcome vector $y_i$ is given by

$$y_i = X_i \beta + Z_i b_i + \epsilon_i, \quad i = 1, \ldots, n$$

where $y_i$ is $n_i \times 1$, $X_i$ is an $n_i \times p$ matrix of fixed covariates, $\beta$ is a $p \times 1$ vector of regression coefficients, commonly referred to as fixed effects, $Z_i$ is an $n_i \times q$ matrix of covariates for the $q \times 1$ vector of random effects $b_i$, and $\epsilon_i$ is an $n \times 1$ vector of random errors. It is standard in implementations of this model to assume $\epsilon_i$ and $b_i$ are independent, and both are normally distributed with

$$\epsilon_i \sim N_{n_i}(0, \sigma^2 I_{n_i})$$

and

$$b_i \sim N_q(0, D)$$
where $I_{n_i}$ denotes the $n_i \times n_i$ identity matrix. Under these assumptions

$$y_i \mid \beta, b_i \sim N_{n_i}(X_i\beta + Z_i b_i, \sigma^2 I_{n_i}) .$$