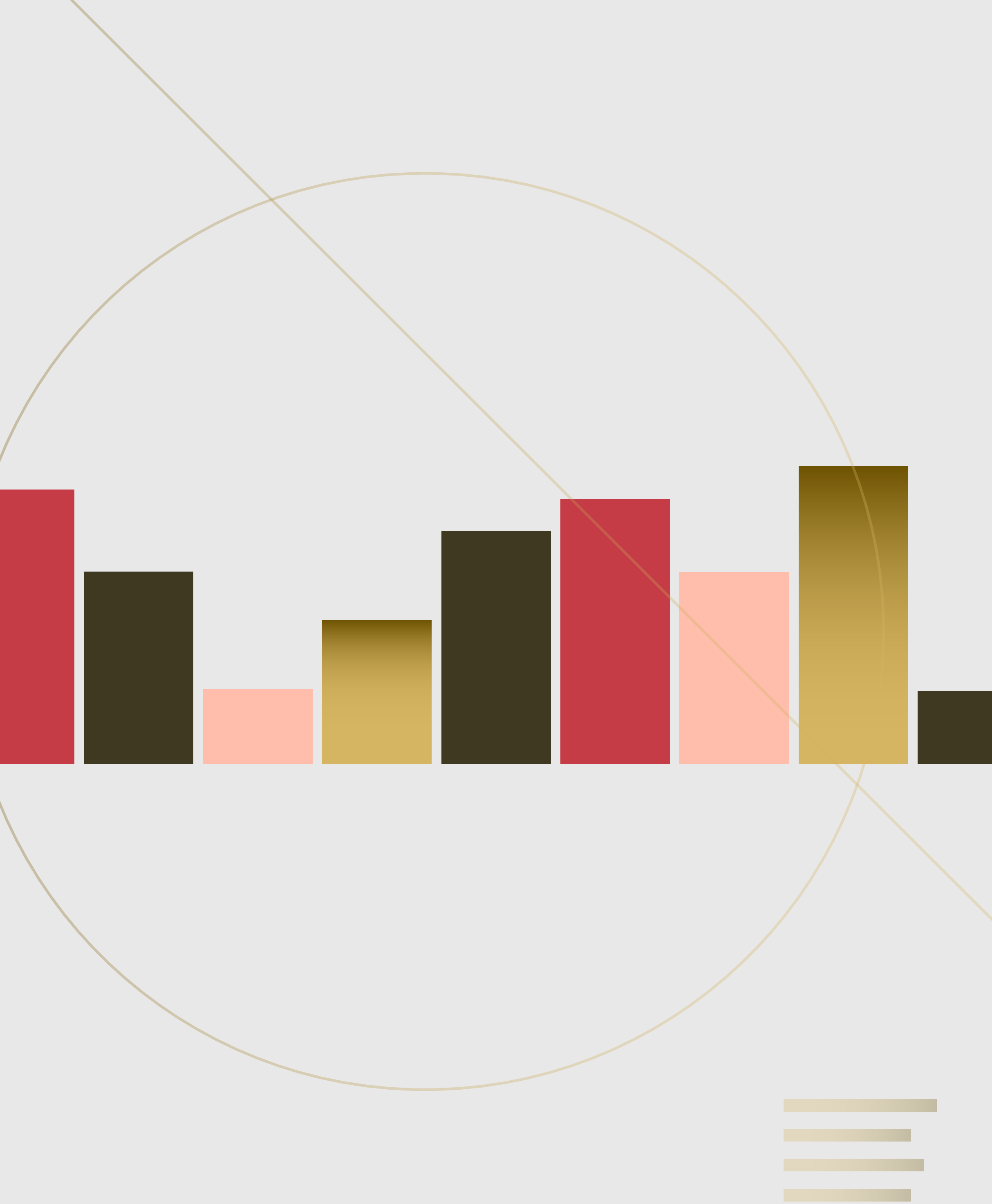

Marginal likelihood from the Gibbs output - Chib (1995)

Journal of American Statistical Association

Research Question - The problem of calculating marginal distribution (Normalizing constant) of the posterior density, which is an input in calculating Bayes factor, has proved extremely challenging as it involves integrating the likelihood function with respect to the prior density.

Formulae :

$$m(y|M_k) = \overbrace{\int \underbrace{f(y|\theta_k, M_k)}_{\substack{\text{likelihood or} \\ \text{density function of} \\ y \text{ for } \theta_k \text{ and} \\ \text{model } k}} \underbrace{\pi(\theta_k|M_k)}_{\substack{\text{prior distribution} \\ \text{for } \theta_k \text{ for} \\ \text{model } k}} d\theta_k}^{\text{Integrating over } \theta_k \text{ to get marginal}}$$

$$f(y|\theta_k, M_k)\, \pi(\theta_k|M_k) = \underbrace{\pi(y, \theta_k|M_k)}_{\text{joint distribution}}$$

Previous studies solutions

1) Newton and Raftery (1994)

$$\hat{m}_{NR} = \left\{ \frac{1}{G} \sum_{i=1}^{G} \left(\frac{1}{f(y|\theta_k^{(g)}, M_k)} \right) \right\}^{-1}$$

Harmonic mean of the likelihood functions evaluated at $\theta_k^{(1)}, \theta_k^{(2)} \ldots \theta_k^{(g)}$

Limitations - not stable as inverse likelihood does not have finite variance.

2) Gelfand and Dey (1993)

$$\hat{m}_{GD} = \left\{ \frac{1}{G} \sum_{g=1}^{G} \left(\frac{\overbrace{p(\theta_k^{(g)})}^{\text{density with thin tails than the denominator. (tuning function)}}}{f(y|\theta_k^{(g)}, M_k)\, \pi(\theta_k^{(g)}|M_k)} \right) \right\}^{-1}$$

Limitation : Tuning function can be hard to find in high dimensional problems.

Derivation of the Approach

Let's set up the problem
θ = parameters of interest
y = data
z = latent variable (To allow for data augmentation).

Data augmentation — ??

It refers to the scheme augmenting the observed data. We do it when there are missing values or truncated data. Sometimes it is difficult or impossible to sample y directly, but there exist

a latent variable "y^*" s.t it is possible to conditionally sample $y|y^*$ and $y^*|y$.
Example: Probit regression explained later.

Basic marginal likelihood identity (BMI)

Bayes rule : $\pi(\theta|y) = \dfrac{f(y|\theta)\,\pi(\theta)}{m(y)}$

$$m(y) = \frac{f(y|\theta)\,\pi(\theta)}{\pi(\theta|y)} \quad - \text{BMI}$$

Now, in order to calculate LHS, we need to evaluate RHS. We can evaluate RHS at some fixed value of θ say θ^* as we already know likelihood, prior and posterior. For that we can exploit the already existing information that we have from Gibbs sampler - $\{\pi(\theta_r|y, \theta_s(s\neq r), z)\}_{r=1}^{B}$

On the logarithmic scale

$$\ln \hat{m}(y) = \ln f(y|\theta^*) + \ln \pi(\theta^*) - \ln \pi(\theta^*|y)$$

where $\hat{m}(y)$ is the estimator for $m(y)$.

The identity is valid for any θ^* but it is recommended to take θ^* from high density for efficiency.

Estimating $\pi(\theta^* | y)$ using Gibbs Sampler

Two vector Block

Full conditionals : $\pi(\theta | y, z)$, $p(z | y, \theta)$

Gibbs sampler output : $\{\theta^{(g)}, z^{(g)}\}_{g=1}^{G}$

Posterior density which we are interested in (does not include z)

$$\pi(\theta | y) = \int \underbrace{\pi(\theta, z | y)}_{\text{target density in Gibbs samples}} dz$$

$$= \int \underbrace{\pi(\theta | y, z)}_{\text{we know the full conditional}} \pi(z | y) \, dz$$

Monte Carlo Estimate of $\pi(\theta | y)$ at θ^*

$$\hat{\pi}(\theta^* | y) = G^{-1} \sum_{i=1}^{G} \pi(\theta^* | y, z^{(g)})$$

$z^{(g)}$ is distributed from $\pi(z | y)$ as we are drawing θ and z from target distribution which is $\pi(\theta, z | y)$ which can be written as $\pi(\theta | y, z) \, \pi(z | y)$. Therefore, θ is coming from Left value & z is coming from right value. We could have written it as $\pi(z | y, \theta) \, \pi(\theta | y)$ as well.

Under regularity conditions. [Similar to gauss Markov assumption for OLS)

$$\hat{\pi}(\theta^*|y) \xrightarrow{a.s} \pi(\theta^*|y)$$

Recap of law of large numbers

Convergence in probability : $Z_n \xrightarrow{P} Z$

$$\lim_{n \to \infty} P(|Z_n - Z| \leq \delta) = 1 \quad \forall \delta$$

Convergence almost surely : $Z_n \xrightarrow{a.s} Z$

$$P(\lim_{n \to \infty} |Z_n - Z| \leq \delta) = 1$$

Regularity conditions

1) y_i is i.i.d with density $f(y;\theta)$
2) parameter space is compact (closed and bounded)
3) $\theta_0 = \max_{\theta \in \Theta} E_{\theta_0} \log f(Y_i;\theta)$ where θ_0 = true parameter
4) Likelihood function is continuous
5) $E_{\theta_0} \log f(Y_i;\theta)$ exist
6) $\sup_{\theta \in \Theta} \left| \frac{1}{n} L(y;\theta) - E_{\theta_0} \log f(Y_i;\theta) \right| < \delta$ almost surely $\forall \delta$.

$\therefore$ estimation of marginal likelihood will look like.

$$\ln \hat{m}(y) = \ln f(y|\theta^*) + \ln \pi(\theta^*) - \ln\left\{ G^{-1} \sum_{g=1}^{G} \pi(\theta^* | y, z^{(g)}) \right\}$$

<u>Three vector Block</u>

Full conditionals : $\pi(\theta_1 | y, \theta_2, z)$, $\pi(\theta_2 | y, \theta_1, z)$, $p(z | y, \theta)$
$\theta = \{\theta_1, \theta_2\}$

Again, the objective is to find $\pi(\theta^* | y)$ which will go in the derivation of marginal likelihood.

Posterior density which we are interested in (does not include z)

$$\pi(\theta_1^*, \theta_2^* | y) = \underset{1}{\pi(\theta_1^* | y)}\ \underset{2}{\pi(\theta_2^* | \theta_1^*, y)}$$

$$\text{Now, } \underset{1}{\pi(\theta_1^* | y)} = \iint \underbrace{\pi(\theta_1^*, \theta_2, z | y)}_{\text{target density in Gibbs sampler}} \, dz \, d\theta_2$$

$$= \iint \underbrace{\pi(\theta_1^* | \theta_2, z, y)}_{\text{we know the full conditionals}} \underbrace{\pi(z, \theta_2 | y)} \, dz \, d\theta_2$$

$$\therefore \hat{\pi}(\theta_1^* | y) = G^{-1} \sum_{i=1}^{G} \pi(\theta_1^* | \theta_2^{(g)}, z^{(g)}, y)$$

evaluated at θ^* and θ_2 & z are sampled from $\pi(z, \theta_2 | y)$

Now Considers ②

$$\pi(\theta_2^* \mid \theta_1^*, y) = \int \pi(\theta_2^* \mid \theta_1^*, y, z)\, \pi(z \mid \theta_1^*, y)\, dz$$

Draws of z from Gibbs samples is from $z|y$ and not $z|\theta_1, y$

$$\pi(\theta_1^*, \theta_2^*, z \mid y) = \pi(\theta_1^* \mid \theta_2^*, z, y)\, \pi(\theta_2^* \mid y, z)\, \pi(z \mid y)$$

In Gibbs samples we are sampling θ_1, θ_2 and z from $\pi(\theta_1^*, \theta_2^*, z|y)$ or in other words sampling θ_1^*, θ_2^* and z from RHS components respectively. RHS can be written in other sequences of θ_1, θ_2 & z as well but our problem set up demands us to make the RHS in this form only.

How to draw z from $f(z \mid \theta_1^*, y)$?

Continue the original gibbs samples for G runs with full conditionals.

$$\pi(\theta_2 \mid y, \theta_1^*, z) \quad \& \quad p(z \mid y, \theta_1^*, \theta_2) \quad -(*)$$

Then, Sampling will be from $\pi(\theta_2, z \mid y, \theta_1^*)$

$$= \pi(\theta_2 \mid z, y, \theta_1^*)\, \pi(z \mid y, \theta_1^*)$$

$$\therefore \quad \hat{\pi}(\theta_2^* | y, \theta_1^*) = G^{-1} \sum_{i=1}^{G} \pi(\theta_2^* | y, \theta_1^*, z^{(j)})$$

where $z^{(j)}$ are sampled from $(*)$

$$\ln \hat{m}(y) = \ln f(y|\theta^*) + \ln \pi(\theta^*) - \ln \hat{\pi}(\theta_1^* | y) - \ln(\theta_2^* | y, \theta_1^*)$$

Multiple Blocks

$$\text{posterior density} = \pi(\theta^* | y) = \pi(\theta_1^*, \theta_2^* \cdots \theta_B^* | y)$$

$$= \underbrace{\pi(\theta_1^* | y)}_{\substack{\text{This can be estimated} \\ \text{using usual Gibbs sampler} \\ \text{draws from full conditionals}}} \pi(\theta_2^* | y, \theta_1^*) \cdots \pi(\theta_B^* | y, \theta_1^* \cdots \theta_{B-1}^*)$$

$$\pi(\theta_1^* | y) = \iint \cdots \int \underbrace{\pi(\theta_1^*, \theta_2 \cdots \theta_B, z | y)}_{\text{usual draw from gibbs sampler}} \, d\theta_2 \, d\theta_3 \cdots d\theta_B \, dz$$

$$\pi(\theta_1^* | y) = \iint \cdots \int \pi(\theta_1^* | \theta_2 \cdots \theta_B, z, y) \, \pi(\theta_2 | \theta_3 \cdots \theta_B, z, y) \cdots \pi(z|y) \, d\theta_2 \, d\theta_3 \cdots dz$$

$$\hat{\pi}(\theta_1^* | y) = G^{-1} \sum_{i=1}^{G} \pi(\theta_1^* | \theta_2^{(g)}, \theta_3^{(g)} \cdots, \theta_B^{(g)}, z^{(g)}, y)$$

Sampled from $\pi(\theta_B | z, y)$

Sampled from $\pi(z|y)$

Sampled from $\pi(\theta_2 | y, \theta_3 \cdots \theta_B, z, y)$

Sampled from $\pi(\theta_3 | y, \theta_4 \cdots \theta_B, z)$

Any other general term is written as

$$\pi(\theta_r^* \mid y, \theta_1^*, \theta_2^*, \ldots, \theta_{r-1}^*) = \iiint\cdots\int \pi(\theta_r^*, \theta_l(l>r), z \mid y, \theta_1^* \cdots \theta_{r-1}^*)$$

$$= \iiint \overbrace{\pi(\theta_r^* \mid \theta_l(l>r), z, y, \theta_1^* \cdots \theta_{r-1}^*)}^{\text{conditional density}}$$
$$\pi(\theta_l(l>r) \mid z, y, \cdots \theta_{r-1}^*) \cdots \pi(z \mid y, \theta_1^* \cdots \theta_{r-1}^*)\, d\theta_l(l>r)\, dz$$

$\theta_l(l>r)$ and z are not sample from the above distributions.

z is sampled from $z|y$ rather than $z|y, \theta_1^* \cdots \theta_{r-1}^*$

θ_{r+1} is sampled from $\theta_{r+1} \mid r+2, \cdots, r_B, z, y$ rather than $\theta_{r+1} \mid r+2, \cdots, r_B, z, y, \theta_1^* \cdots \theta_{r-1}^*$

So, to estimate the term, continue sampling with conditional densities of $\{\theta_r, \theta_{r+1} \cdots, \theta_B, z)$ and substitute $\theta_1^*, \theta_2^* \cdots \theta_{r-1}^*$ instead of $\theta_1, \theta_2 \cdots \theta_{r-1}$. (* *)

$$\hat{\pi}(\theta_r \mid y, \theta_s^*(s<r)) = G^{-1} \sum_{j=1}^{G} \pi(\theta_r^* \mid y, \theta_1^*, \theta_2^* \cdots \theta_{r-1}^*, \theta_l^{(j)}(l>r), z^{(j)})$$

where $\theta_l^{(j)}(l>r)$ and $z^{(j)}$ are the results from extra sampling done in (* *)

$$\ln \hat{m}(y) = \ln f(y \mid \theta^*) + \ln \pi(\theta^*) - \sum_{r=1}^{B} \ln \hat{\pi}(\theta_r^* \mid y, \theta_s^*(s<r))$$

Bayes factor

Bayes factor to compare model k and j

$$\hat{B}_{kj} = \exp\{\ln \hat{m}(y|M_k) - \ln \hat{m}(y|M_j)\}$$

Application : Binary Probit model

$$y = \begin{cases} 1 & \text{if cancer has spread} \\ 0 & \text{otherwise} \end{cases}$$

x_1 = age
x_2 = level of serum acid phosphate
x_3 = X ray result (Dummy)
x_4 = Size of tumor (Dummy)
x_5 = Grade of tumor (Dummy)

53 observations

Examined 9 different models

Probit model : $Pr(y=1|M_k) = \underbrace{\phi(x_{ik}' \beta_k)}_{\text{CDF of standard normal distribution}}$

as y is binary, the likelihood of y is multiplying iid bernoulli distribution for each i.

$$f(y \mid M_k, \beta_k) = \prod_{i=1}^{53} [\Phi(x_{ik}' \beta_k)]^{y_i} [1 - \Phi(x_{ik}' \beta_k)]^{1-y_i}$$

Setting up the gibbs sampler (for any model k)

Define a latent variable z s.t

$$z_i \sim N(x_i'\beta, 1) \; ; \quad y_i = \underbrace{I(z_i > 0)}_{\text{indicator function}}$$

$$\Pr(z > 0) = \Phi(x_i'\beta)$$

$\pi(\beta)$ - prior distribution

$\pi(\beta \mid y, z)$ - posterior distribution (calculated)

$$p(z_i \mid y, \beta) \propto \phi(z_i \mid x_i'\beta, 1)\, I[0, \infty] \text{ if } y_i = 1$$

$$\propto \phi(z_i \mid x_i'\beta, 1)\, I(-\infty, 0] \text{ if } y_i = 0$$

where $\phi(\cdot \mid \mu, 1)\, I[a, b]$ is the normal density truncated to the interval $[a, b]$.

Then log of marginal likelihood of model k is

$$\ln f(y \mid M_k) = \ln f(y \mid M_k, \beta_k^*) + \ln \pi(\beta_k) - \ln \left\{ \frac{1}{5000} \sum_{g=1}^{5000} \phi(\beta_k^* \mid y, z^{(g)}) \right\}$$

Bayes factor can be calculated to compare models