3 Model Choice

3.1 Introduction

What is Model choice? Model choice, in vague terms, is model comparison and, perhaps, model selection. In a "typical" or "standard" inferential problem we are interested in some null hypothesis such as $\theta = 0$. In a "typical" estimation problem, at least in the Bayesian setting, we wish to discover the "true" mean or distribution of a function of a parameter given the data, $\mathbb{E}(h(\theta) \mid \mathbf{Y})$ or $\pi(h(\theta) \mid \mathbf{Y})$.

Model choice can be distilled to the choice about several models in a, possibility infinite, set of models $\{\mathcal{M}_i\}_i$. This, in turn, can be viewed as an estimation problem on the index $i \in \{1, \ldots, p\}$, in the finite case, or $i \in \mathbb{N}$, in the infinite case.

In a typical Bayesian analysis we are given data $\mathbf{Y}$, a sampling distribution, or data model, $\pi(\mathbf{Y} \mid \theta)$, and a prior distribution $\pi(\theta)$. Our goal in this case is to draw inferences about θ given the data, via the posterior distribution $\pi(\theta \mid \mathbf{Y})$. When speaking of model choice, we are speaking of choosing the most appropriate data model, for the data at hand. This makes it difficult to condition on the data especially when we are entertaining questions such as: *Does π belong to the family $\{\pi_\theta : \theta \in \Theta\}$.*

Therefore, we will restrict our attention to the case where several parametric models are to be considered:

$$\mathcal{M}_i : \mathbf{Y} \sim \pi_i(\mathbf{Y} \mid \theta_i), \quad \theta_i \in \Theta_i, \quad i \in \mathcal{I}, \tag{21}$$

where the index set $\mathcal{I}$ may be infinite. This reduced perspective is less puzzling from a Bayesian point of view because we can construct a prior distribution for the parameters in each model and, in theory, construct a prior for the model index parameter i. The problem then reduces to a standard Bayesian posterior inference on i.

Let's look at two different examples: In the first example we are faced with choosing between a finite number of models. In the second example, we are faced with choosing from a (countably) infinite number of examples.

Example 1 *(Gelfand, 1996) For 5 orange trees, the growth of tree i is measured through the* *circumferences y_{it} at seven times T_t. The models under consideration are*

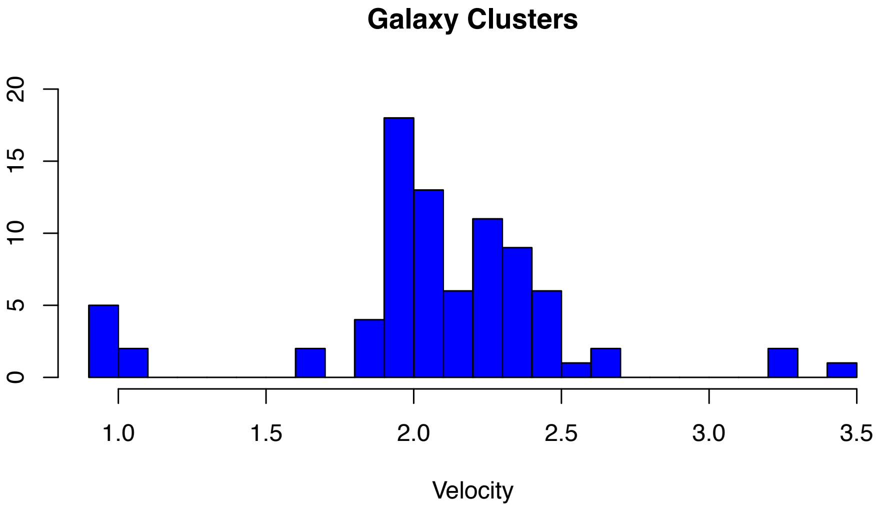
$$\begin{aligned}
\mathcal{M}_1 &: Y_{it} \sim N(\beta_{10} + b_{1i}, \sigma_1^2) \\
\mathcal{M}_2 &: Y_{it} \sim N(\beta_{20} + \beta_{21} T_t + b_{2i}, \sigma_2^2) \\
\mathcal{M}_3 &: Y_{it} \sim N\left(\frac{\beta_{30}}{(1 + \beta_{31}\exp(\beta_{32} T_t)}, \sigma_3^2\right) \\
\mathcal{M}_4 &: Y_{it} \sim N\left(\frac{\beta_{40} + b_{4i}}{(1 + \beta_{41}\exp(\beta_{42} T_t)}, \sigma_4^2\right),
\end{aligned}$$

where $b_{ji} \sim N(0, \tau^2)$ are random effects. The first model is a simple individual random effects model with no time effects. The second model includes a linear time effect. The third model is a non-linear model with no random effects and the fourth model extends the third by including random effects. Note that these models are not nested.

Example 2 *(Roeder, 1992) This dataset has been used in many papers on mixture estimation. The dataset consists of 82 observations of galaxy velocities. For convenience, we will assume that this dataset can be represented as a mixture of normal distributions whose number of components k is unknown. Here, a component of the mixture can be interpreted as a cluster of galaxies. The models under consideration are*

$$\mathcal{M}_i : V_j \sim \sum_{\ell=1}^{i} p_{li} N(\mu_{li}, \sigma_{li}^2),$$

where $i \in \mathbb{N}$. Usually, we will pick some large upper bound N and restrict $i \in \{1, 2, \ldots, N\}$.



From both of these examples, it is obvious that there is often a high degree of arbitrariness involved in the selection of which models to choose. Furthermore, as in Example 2 the normality assumption is based on convenience—it may or may not be motivated by astrophysics.

3.2 Deviance information criterion

Spiegelhalter, Best and Carlin (2002) developed a Bayesian alternative to both AIC (Akaike's information criterion) and BIC (Bayesian, or Schwartz's, information criterion). Recall

$$\begin{aligned}\text{AIC} &= -2\ln\pi(\mathbf{Y} \mid \hat{\theta}) + 2p = D(\hat{\theta}) + 2p \\ \text{BIC} &= -2\ln\pi(\mathbf{Y} \mid \hat{\theta}) + p\ln(n) = D(\hat{\theta}) + p\ln(n).\end{aligned}$$

$D(\theta) = -2\ln\pi(\mathbf{Y} \mid \theta)$ is called the deviance. From a classical perspective, model comparison takes place by defining a measure of fit, typically a deviance statistic, and complexity, the number of free parameters in the model. A problem, from a Bayesian perspective, with both AIC and BIC is that they do not take into account the prior information. Another problem with AIC and BIC occurs in complex hierarchical models where the number of parameters is larger than the number of observations. AIC and BIC clearly cannot be applied in these settings.

Furthermore, in non-i.i.d. structures the definition of both the sample size, n, and the number of parameters, p, may be ambiguous (see Spiegelhalter, et al.). DIC is defined by

$$\begin{aligned}\text{DIC} &= \mathbb{E}(D(\theta) \mid \mathbf{Y}) + p_D \\ &= \mathbb{E}(D(\theta) \mid \mathbf{Y}) + \{\mathbb{E}(D(\theta) \mid \mathbf{Y}) - D(\mathbb{E}(\theta \mid \mathbf{Y}))\} \\ &= D(\mathbb{E}(\theta \mid \mathbf{Y})) + 2p_D.\end{aligned}$$

$\mathbb{E}(D(\theta) \mid Y)$ can be interpreted as a measure of *fit*, while p_D is a measure of *complexity*, also called the *effective number of parameters*. One advantage of DIC over the use of Bayes Factors is that it allows for improper priors since each model is considered separately. For Bayes factors, the prior model specification must be proper. Spiegelhalter et al. defined DIC in terms of the last equality above, in analogy with AIC and BIC.

Computationally, one usually has to rely on MCMC estimates to compute DIC. But this is trivial as $\mathbb{E}(D(\theta) \mid \mathbf{Y})$ is just the posterior expectation of an explicit function of θ and $D(\mathbb{E}(\theta \mid \mathbf{Y}))$ is just a function of the posterior mean of θ. That is, for an MCMC sample $\theta^{(1)}, \ldots, \theta^{(T)}$,

$$\mathbb{E}(D(\theta) \mid \mathbf{Y}) \approx T^{-1}\sum_{t=1}^{T} D(\theta^{(t)}) = -2T^{-1}\sum_{t=1}^{T} \ln\left(\pi\left(\mathbf{Y} \mid \theta^{(t)}\right)\right)$$

and

$$D(\mathbb{E}(\theta \mid \mathbf{Y})) \approx D\left(T^{-1}\sum_{t=1}^{T}\theta^{(t)}\right) = -2\ln\pi\left(\mathbf{Y} \mid \left(T^{-1}\sum_{t=1}^{T}\theta^{(t)}\right)\right).$$

One known issue with DIC is that the effective number of parameters may be negative in some situations (although, by Jensen's inequality, it should always be nonnegative). In some random effects models, p_D does turn out to be negative. In these situations, it appears that DIC is not a viable solution to model choice as more complex models are favored as opposed to penalized. (Incidentally, their paper on DIC first appeared as a technical report in 1998. It did not get published until 2002.)

3.3 Bayes factors

Bayes factors are the most commonly accepted way to compare models. Let

$$\boldsymbol{\Theta} = \bigcup_{i\in I}\{i\} \times \Theta_i$$

denote the parameter space associated with the set of models (21), with the model indicator $i \in I$ now part of the parameters. Assume we can assign probabilities $p_i = \pi(\mathcal{M}_i)$ to the model indicators (or models $\mathcal{M}_i$) and also assign priors $\pi_i(\theta_i)$ on the parameter subspaces Θ_i. Then, by Bayes' theorem we can compute the posterior probability of model i:

$$\pi(\mathcal{M}_i \mid \mathbf{Y}) = \frac{p_i \int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\pi_i(\theta_i)d\theta_i}{\sum_{j\in I} p_j \int_{\Theta_j} \pi_j(\mathbf{Y} \mid \theta_j)\pi_j(\theta_j)d\theta_j}.$$

The Bayes factor, B_{ij} comparing model i to model j is defined as the ratio of the posterior probabilities to the prior probabilities:

$$\begin{aligned} B_{ij} &= \frac{\pi(\mathcal{M}_i \mid \mathbf{Y})}{\pi(\mathcal{M}_j \mid \mathbf{Y})} \Big/ \frac{\pi(\mathcal{M}_i)}{\pi(\mathcal{M}_j)} \\ &= \frac{\int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\pi_i(\theta_i)d\theta_i}{\int_{\Theta_j} \pi_j(\mathbf{Y} \mid \theta_j)\pi_j(\theta_j)d\theta_j} \\ &= \frac{\pi_i(\mathbf{Y})}{\pi_j(\mathbf{Y})}. \end{aligned}$$

Notice that the last two equations imply that the Bayes factor is the ratio of the marginal densities of the data under the two models. One may also think of this as the ratio of two normalizing constants. It is obvious from the above that the priors must be proper probability distributions. In the following two sections we will develop methods to estimate Bayes factors.

3.4 Importance sampling

We will begin with the computation of a single normalizing constant:

$$\pi_i(\mathbf{Y}) = \int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\pi_i(\theta_i)d\theta_i$$

and then the ratios of such integrals:

$$\frac{\int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\pi_i(\theta_i)d\theta_i}{\int_{\Theta_j} \pi_j(\mathbf{Y} \mid \theta_j)\pi_j(\theta_j)d\theta_j}.$$

Suppose we have an *importance* distribution with density g. Then

$$\begin{aligned} \pi_i(\mathbf{Y}) &= \int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\pi_i(\theta_i)d\theta_i \\ &= \int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\frac{\pi_i(\theta_i)}{g(\theta_i)}g(\theta_i)d\theta_i. \end{aligned}$$

Thus, a Monte Carlo estimate of the marginal density of $\mathbf{Y}$ is

$$\pi_i(\mathbf{Y}) \approx T^{-1}\sum_{t=1}^{T} \frac{\pi_i\left(\mathbf{Y} \mid \theta_i^{(t)}\right)\pi_i\left(\theta_i^{(t)}\right)}{g\left(\theta_i^{(t)}\right)},$$

where $\{\theta_i^{(t)}\}_t$ is a Monte Carlo or MCMC sample from g. One issue with the above formulation is that the variance of the estimator is finite only when

$$\mathbb{E}_g\left[\pi_i^2(\mathbf{Y} \mid \theta_i)\pi_i^2(\theta_i)/g^2(\theta_i)\right] < \infty.$$

Therefore instrumental densities with lighter tails than π_i (i.e. such that the ratio π_i/g is unbounded) are inappropriate. Even if the ratio is bounded, the above importance sampling estimator may have quite large variance. To reduce this variance, note that

$$\int_{\Theta_i} \frac{\pi_i(\theta_i)}{g(\theta_i)}g(\theta_i)d\theta_i = 1.$$

Therefore, we have

$$\pi_i(\mathbf{Y}) = \frac{\int_{\Theta_i} \pi_i(\mathbf{Y} \mid \theta_i)\frac{\pi_i(\theta_i)}{g(\theta_i)}g(\theta_i)d\theta_i}{\int_{\Theta_i} \frac{\pi_i(\theta_i)}{g(\theta_i)}g(\theta_i)d\theta_i},$$

which can be estimated by

$$\pi_i(\mathbf{Y}) \approx \frac{\sum_{t=1}^{T} \frac{\pi_i\left(\mathbf{Y}|\theta_i^{(t)}\right)\pi_i\left(\theta_i^{(t)}\right)}{g\left(\theta_i^{(t)}\right)}}{\sum_{t=1}^{T} \frac{\pi_i\left(\theta_i^{(t)}\right)}{g\left(\theta_i^{(t)}\right)}} \equiv \widehat{\pi_i(\mathbf{Y})}. \tag{22}$$

Note that the above estimator is biased. The bias turns out to be small, however, and the improvement in the variance estimator is worth the tradeoff. Also note that g could be an unnormalized density as the normalizing constant cancels in the above estimator.

A compelling incentive for using importance sampling is that the same sample $\{\theta^{(t)}\}_t$ can be used for several models $\mathcal{M}_i$, if they all involve the same type of parameters.

How do we choose g, especially in high dimensional settings? As a first, and obvious choice, set $g(\theta) = \pi(\theta)$ in (22), thus we draw our samples from the prior. This leads to

$$\pi(\mathbf{Y}) \approx \frac{1}{T}\sum_{t=1}^{T} \pi(\mathbf{Y} \mid \theta^{(t)}).$$

However, it is often inefficient if the data is informative as most of the simulated values $\theta^{(t)}$ fall outside the modal region of the likelihood.

The next obvious choice is $g(\theta) = \pi(\mathbf{Y} \mid \theta)\pi(\theta)$. Here we use posterior draws from a MCMC simulation as g is proportional to the posterior). Then (22) becomes

$$\pi(\mathbf{Y}) \approx \frac{1}{T^{-1}\sum_{t=1}^{T}\left[\pi(\mathbf{Y} \mid \theta^{(t)})\right]^{-1}},$$

which is the harmonic mean of the likelihood. This is attractive since we can use the posterior draws from the MCMC simulation to estimate g and it allows for improper priors as long as the posterior is proper. The down side is that the variance is often infinite. A solution to this problem is *defensive importance sampling* by taking a mixture of $\pi(\theta \mid \mathbf{Y})$ and a distribution with fat tails $\varpi(\theta)$:

$$(1-\rho)\pi(\theta \mid \mathbf{Y}) + \rho\varpi(\theta), \quad \rho \ll 1.$$

For instance, $\varpi(\theta) = \pi(\theta)$. A second solution is to generate a sample from the posterior and approximate the marginal by

$$\pi(\mathbf{Y}) \approx \frac{1}{T^{-1}\sum_{t=1}^{T} h(\theta^{(t)})[\pi(\mathbf{Y} \mid \theta^{(t)})\pi(\theta^{(t)})]^{-1}}$$

where h is an arbitrary density. This estimator will have finite variance if

$$\int_{\Theta} \frac{h^2(\theta)}{\pi(\mathbf{Y} \mid \theta)\pi(\theta)} d\theta < \infty.$$

Therefore, in principle, h can be chosen to satisfy this constraint.

Now, more importantly, is that we wish to compute a ratio of normalizing constants. Given the above, then, it is obvious that we can estimate the two marginal densities, $\pi_1(\mathbf{Y})$ and

$\pi_2(\mathbf{Y})$, say, and take their ratio. Let g_1 be an importance function for π_1 with sample $\{\theta_1^{(t)}\}_{t=1}^{T_1}$ and let g_2 be an importance function for π_2 with sample $\{\theta_1^{(t)}\}_{t=1}^{T_2}$. Then

$$B_{12} \approx \frac{\widehat{\pi_1(\mathbf{Y})}}{\widehat{\pi_2(\mathbf{Y})}},$$

where the estimates of the marginal densities are given by (22) with the obvious modifications.

Let Θ_i be the support of $\pi_i(\theta \mid \mathbf{Y})$. Another importance sampling estimator of the ratio of two normalizing constants when $\Theta_1 \subset \Theta_2$ is given by the identity

$$\frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} = \mathbb{E}_{\pi_2(\theta|\mathbf{Y})}\left[\frac{\pi_1(\mathbf{Y},\theta)}{\pi_2(\mathbf{Y},\theta)}\right], \tag{23}$$

This equality is derived as follows:

$$\begin{aligned}
\mathbb{E}_{\pi_2(\theta|\mathbf{Y})}\left[\frac{\pi_1(\mathbf{Y},\theta)}{\pi_2(\mathbf{Y},\theta)}\right] &= \int_{\Theta_2} \frac{\pi_1(\mathbf{Y},\theta)}{\pi_2(\mathbf{Y},\theta)} \pi_2(\theta \mid \mathbf{Y}) d\theta \\
&= \int_{\Theta_2} \frac{\pi_1(\theta \mid \mathbf{Y})\pi_1(\mathbf{Y})}{\pi_2(\theta \mid \mathbf{Y})\pi_2(\mathbf{Y})} \pi_2(\theta \mid \mathbf{Y}) d\theta \\
&= \frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} \int_{\Theta_2} \pi_1(\theta \mid \mathbf{Y}) d\theta \\
&= \frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} \int_{\Theta_1} \pi_1(\theta \mid \mathbf{Y}) d\theta \\
&= \frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})}.
\end{aligned}$$

Given an MCMC sample $\{\theta^{(t)}\}_{t=1}^{T}$ from $\pi_2(\theta \mid \mathbf{Y})$ we have

$$B_{12} = \frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} \approx \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_1\left(\mathbf{Y},\theta^{(t)}\right)}{\pi_2\left(\mathbf{Y},\theta^{(t)}\right)} = \frac{1}{T}\sum_{t=1}^{T} \frac{\pi_1\left(\mathbf{Y} \mid \theta^{(t)}\right)\pi_1\left(\theta^{(t)}\right)}{\pi_2\left(\mathbf{Y} \mid \theta^{(t)}\right)\pi_2\left(\theta^{(t)}\right)}.$$

3.5 Bridge Sampling

In 1996, Meng and Wong introduced a generalization of importance sampling for the computation of ratios of normalizing constants, called *Bridge Sampling.* The idea is a generalization of importance sampling.

Let $\pi_1(\theta \mid \mathbf{Y}) = \pi_1(\mathbf{Y}, \theta)/\pi_1(\mathbf{Y})$ and $\pi_2(\theta \mid \mathbf{Y}) = \pi_2(\mathbf{Y}, \theta)/\pi_2(\mathbf{Y})$. Suppose Θ_1 is the support for π_1 and Θ_2 is the support for π_2. For any *bridge function* $h(\theta)$ defined on $\Theta_1 \cap \Theta_2$ such that

$$0 < \left| \int_{\Theta_1 \cap \Theta_2} h(\theta)\pi_1(\theta \mid \mathbf{Y})\pi_2(\theta \mid \mathbf{Y}) d\theta \right| < \infty,$$

we have

$$\frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} = \frac{\mathbb{E}_{\pi_2(\theta|\mathbf{Y})}\left[h(\theta)\pi_1(\mathbf{Y}, \theta)\right]}{\mathbb{E}_{\pi_1(\theta|\mathbf{Y})}\left[h(\theta)\pi_2(\mathbf{Y}, \theta)\right]}. \tag{24}$$

Note that the above bounds on the integral imples that $\Theta_1 \cap \Theta_2 \neq \emptyset$. To see (24),

$$\begin{aligned}
\frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} &= \frac{\pi_1(\mathbf{Y}, \theta)/\pi_1(\theta \mid \mathbf{Y})}{\pi_2(\mathbf{Y}, \theta)/\pi_2(\theta \mid \mathbf{Y})} && \Longrightarrow \\
h(\theta)\pi_1(\mathbf{Y})\pi_2(\mathbf{Y}, \theta)\pi_1(\theta \mid \mathbf{Y}) &= h(\theta)\pi_2(\mathbf{Y})\pi_1(\mathbf{Y}, \theta)\pi_2(\theta \mid \mathbf{Y}) && \Longrightarrow \\
\pi_1(\mathbf{Y}) \int_{\Theta_1} h(\theta)\pi_2(\mathbf{Y}, \theta)\pi_1(\theta \mid \mathbf{Y}) d\theta &= \pi_2(\mathbf{Y}) \int_{\Theta_2} h(\theta)\pi_1(\mathbf{Y}, \theta)\pi_2(\theta \mid \mathbf{Y}) d\theta && \Longrightarrow \\
\frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} &= \frac{\mathbb{E}_{\pi_2(\theta|\mathbf{Y})}\left[h(\theta)\pi_1(\mathbf{Y}, \theta)\right]}{\mathbb{E}_{\pi_1(\theta|\mathbf{Y})}\left[h(\theta)\pi_2(\mathbf{Y}, \theta)\right]}.
\end{aligned}$$

Let $\{\theta_i^{(t)}\}_{t=1}^{T_i}$ denote a random sample from $\pi_i(\theta \mid \mathbf{Y})$. Then

$$B_{12} = \frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} \approx \frac{T_2^{-1} \sum_{t=1}^{T^2} \pi_1\left(\mathbf{Y} \mid \theta_2^{(t)}\right) \pi_1\left(\theta_2^{(t)}\right) h\left(\theta_2^{(t)}\right)}{T_1^{-1} \sum_{t=1}^{T^1} \pi_2\left(\mathbf{Y} \mid \theta_1^{(t)}\right) \pi_2\left(\theta_1^{(t)}\right) h\left(\theta_1^{(t)}\right)}.$$

The expression (24) unifies many identities found in the literature on simulating Bayes factors and normalizing constants. For example, taking $h(\theta) = \pi_2^{-1}(\mathbf{Y}, \theta)$ when $\Theta_1 \subset \Theta_2$ leads to (23). When $\Theta_1 = \Theta_2$ we can take $h(\theta) = \pi_1^{-1}(\mathbf{Y}, \theta)\pi_2^{-1}(\mathbf{Y}, \theta)$ which gives

$$\frac{\pi_1(\mathbf{Y})}{\pi_2(\mathbf{Y})} = \frac{\mathbb{E}_{\pi_2(\theta|\mathbf{Y})}\left[\pi_2^{-1}(\mathbf{Y}, \theta)\right]}{\mathbb{E}_{\pi_1(\theta|\mathbf{Y})}\left[\pi_1^{-1}(\mathbf{Y}, \theta)\right]}.$$

Note that this is a generalization of the *harmonic rule* of Newton and Raftery (1994) and Gelfand and Dey (1994).

3.6 LPML

One more model selection criterion is sometimes called the *log pseudomarginal likelihood* (Geisser and Eddy, 1979). It uses predictive densities and can be used to compute pseudo

Bayes factors. LPML tends to be more stable than computing Bayes Factors, can be used with improper priors and does not suffer some of the same problems that DIC suffers.

Suppose, under model $\mathcal{M}$ the data are conditionally independent given a parameter θ. Thus

$$\pi(y \mid \theta, \mathcal{M}) = \prod_{i=1}^{n} \pi_i(y_i \mid \theta, \mathcal{M}).$$

The idea is to replace $f(y \mid \mathcal{M})$, which is need to compute Bayes factors, with a predictive version—called the pseudomarginal likelihood:

$$\hat{\pi}(y \mid \mathcal{M}) = \prod_{i=1}^{n} \pi_i(y_i \mid y_{-1},\, \mathcal{M})$$

where y_{-i} is the vector of all subject's observed data expect for the ith subject. The products on the right hand side are called the *conditional predictive ordinates* (CPO$_i$).

Then the LPLM is given by

$$\text{LPLM} = \sum_{i=1}^{n} \ln(\text{CPO}_i).$$

To find the pseudo Bayes factor we simply exponentiate the difference of the LPLMs between two models.

CPO$_i$ and LPLM are simple to compute (Gelfand and Dey, 1994). First, we have that

$$\text{CPO}_i^{-1} = \int_{\Theta} \frac{1}{\pi_i\left(y_i \mid \theta, \mathcal{M}\right)} \pi(\theta \mid y) d\theta = \mathbb{E}_{\pi(\theta|y)}\left[\frac{1}{\pi_i\left(y_i \mid \theta, \mathcal{M}\right)}\right]. \tag{25}$$

Notice that this integral does not depend on predictive densities. We can work directly with the sampling distribution under model $\mathcal{M}$. This vastly simplifies computation. We will drop $\mathcal{M}$ from the proof of (25):

$$\begin{aligned}
\text{CPO}_i = \pi_i(y_i \mid y_{-i}) &= \int_{\Theta} \pi_i(y_i \mid \theta)\pi(\theta \mid y_{-i}) d\theta \\
&= \int_{\Theta} \pi_i(y_i \mid \theta) \frac{\prod_{j \neq i} \pi_j(y_j \mid \theta)\pi(\theta)}{\pi(y_{-1})} d\theta \\
&= \frac{\int_{\Theta} \prod_{j=1}^{n} \pi_j(y_j \mid \theta)\pi(\theta) d\theta}{\int_{\Theta} \prod_{j \neq i} \pi_j(y_j \mid \theta)\pi(\theta) d\theta}.
\end{aligned}$$

So that

$$
\begin{aligned}
\mathrm{CPO}_i^{-1} = \frac{1}{\pi_i(y_i \mid y_{-1})} &= \frac{\int_\Theta \prod_{j \neq i} \pi_j(y_j \mid \theta)\pi(\theta)d\theta}{\int_\Theta \prod_{j=1}^n pi_j(y_j \mid \theta)\pi(\theta)d\theta} \\
&= \frac{\int_\Theta \prod_{j \neq i} \pi_j(y_j \mid \theta)\pi(\theta)d\theta}{\pi(y)} \\
&= \int_\Theta \frac{1}{\pi_i(y_i \mid \theta)} \prod_{j=1}^n \pi_j(y_j \mid \theta)\pi(\theta)/\pi(y)d\theta \\
&= \int_\Theta \frac{1}{\pi_i(y_i \mid \theta)} \pi(\theta \mid y)d\theta.
\end{aligned}
$$

Now since CPO_i^{-1} is the posterior expectation of a function, it is easy to approximate via an MCMC sample. Suppose we have an MCMC sample $\theta^{(1)}, \ldots, \theta^{(m)}$. Then

$$
\mathrm{CPO}_i^{-1} \approx \frac{1}{m} \sum_{j=1}^m \frac{1}{\pi_i\left(y_i \mid \theta^{(j)}, \mathcal{M}\right)} \equiv \widehat{\mathrm{CPO}_i^{-1}}.
$$

Then can estimate the LPML:

$$
\widehat{\mathrm{LPML}} = \sum_{i=1}^n -\ln\left(\widehat{\mathrm{CPO}_i^{-1}}\right).
$$

3.7 Model averaging

A natural approach, from the Bayesian perspective, is to account for model uncertainty by including all models under consideration for future decisions. This is a rather different approach than model selection. Model averaging accounts for the uncertainty inherent in choosing one model over all other models under consideration. By choosing a "best" model, one ignores the uncertainty in choosing that model and thereafter one ignores the uncertainty about that choice in all subsequent steps.

Model averaging gives more honest predictions of future observations when one is faced with "choosing from several different models". Given a sample $\mathbf{y}$ and a choice of models $\{\mathcal{M}_i\}$, with posterior probabilities $\Pr(\mathcal{M}_i \mid \mathbf{y})$, the model averaged predictive density of a new observation x is

$$
\pi(x \mid \mathbf{y}) = \sum_i \Pr(\mathcal{M}_i \mid \mathbf{y}) \int_{\Theta_i} \pi_i(x \mid \theta_i)\pi_i(\theta_i \mid \mathbf{y})d\theta_i.
$$