Chapter 10

Multi-model inference

10.1 Introduction

If we consider several competing statistical models, any of which could serve as an explanation for our data, and would like to select the best of them, then we face a model selection (or a model choice, or a model comparison) problem. Instead of choosing a single best model, it might be more meaningful to combine somehow inferences obtained from all of the models, and then we may speak of model averaging. Such activities may also be called multi-model inference.

For example, in the binary regression setting with the explanatory variable x we might posit the model

$$[Y_i \mid \theta] \overset{\text{ind}}{\sim} B(F(\alpha + \beta x_i)), \qquad i = 1, \ldots, n,$$

where $B(p)$ is the Bernoulli distribution with success probability p, but we might want to consider several different link function F such as the logit, the probit and, say, the cdf of t distribution $\nu = 4$ degrees of freedom.

In a continuous regression problem with explanatory variable x, we might want to consider polynomials of degrees zero, one and two as the mean response,

$$\begin{array}{lrl}
\text{model 0:} & [Y_i \mid \alpha, \sigma^2] \overset{\text{ind}}{\sim} N(\alpha, \sigma^2), & i = 1, \ldots, n \\
\text{model 1:} & [Y_i \mid \alpha, \beta_1, \sigma^2] \overset{\text{ind}}{\sim} N(\alpha + \beta_1 x_i, \sigma^2), & i = 1, \ldots, n \\
\text{model 2:} & [Y_i \mid \alpha, \beta_1, \beta_2, \sigma^2] \overset{\text{ind}}{\sim} N(\alpha + \beta_1 x_i + \beta_2 x_i^2, \sigma^2), & i = 1, \ldots, n.
\end{array}$$

One commonly occurring situation is the variable selection problem. For instance, we might want to select which of the candidate variables to use as explanatory variables in a multiple regression problem.

The usual frequentist solution to model selection in the case of nested models is to perform a series of hypothesis tests. One statistical model is said to be nested within another model, if it is a special case of the other model. In the polynomial regression example, model 0 is a special case of model 1, and model 1 is a special case of model 2. In this example a frequentist statistician would probably select among these models by using F-tests. However, one may be bothered by the fact that we actually need to make multiple tests. How should we take this into account when selecting the size of the test?

Outside the linear model framework, a frequentist statistician would compare nested models by using the asymptotic χ^2 distribution of the likelihood ratio test (LRT) statistic, but the asymptotics is valid only when the simpler model does not correspond to a parameter value at the boundary of the parameter space of the more complex model. There are important statistical models (such as the linear mixed effects model) where a natural null hypothesis corresponds to a point at the boundary of the parameter space, and then the usual χ^2 asymptotics do not apply.

In contrast to the polynomial regression example, in the binary regression example there is no natural way to nest the models, and comparing the models by hypothesis tests would be problematic.

Besides hypothesis testing, a frequentist statistician might compare models using some information criterion, such as the Akaike information criterion, AIC. This approach does not suffer from the problems we identified in the hypothesis testing approach.

In the rest of this chapter we will discuss Bayesian techniques for model selection, or more generally, to multi-model inference. The basic idea is to introduce a single encompassing model which is a union of all the alternative models. Then we use Bayes rule to derive the posterior distribution. This requires that we have successfully specified the entire collection of candidate models we want to consider. This the $\mathcal{M}$-closed case instead of the more general $\mathcal{M}$-open case, where the ultimate model collection is not known ahead of time, see [1, Ch. 6] for a deep discussion on this and other assumptions and approaches a Bayesian statistician can use in multi-model inference.

The concepts we need are borrowed from the Bayesian approach to hypothesis testing. There is no requirement that the models should be nested with respect to one another, and no problem arises if one model corresponds to a parameter value at the boundary of the parameter space of another model.

To unify the discussion we make the following conventions. The alternative models are numbered $1, \ldots, K$. The parameter vector θ_m of model m belongs to the parameter space $S_m \subset \mathbb{R}^{d_m}$. The parameter vectors $\theta_m, m = 1, \ldots, K$ of the models are considered separate: no two models share any parameters.

For example, in the binary regression example the α and β parameters for the logit link and for the probit link and for the t link are considered separate, and we could label them, e.g., as

$$\theta_1 = (\alpha_1, \beta_1), \quad \theta_2 = (\alpha_2, \beta_2), \quad \theta_3 = (\alpha_3, \beta_3).$$

Here $S_1 = S_2 = S_3 = \mathbb{R}^2$, and $d_1 = d_2 = d_3 = 2$.

In the polynomial regression example the error variance parameters are considered separate parameters in all of the three models, the intercepts and slopes are considered separate parameters, and so on. We could label them, e.g., as

$$\theta_1 = (\alpha_0, \sigma_0^2), \quad \theta_2 = (\alpha_1, \beta_1, \sigma_1^2), \quad \theta_3 = (\alpha_2, \beta_{21}, \beta_{22}, \sigma_2^2).$$

Here $d_1 = 2$, $d_2 = 3$, $d_3 = 4$, and

$$S_1 = \mathbb{R} \times \mathbb{R}_+, \quad S_2 = \mathbb{R}^2 \times \mathbb{R}_+, \quad S_3 = \mathbb{R}^3 \times \mathbb{R}_+,$$

At first sight it may seem unnatural to separate the parameters which usually are denoted by the same symbol, such as α and σ^2 in the zeroth and the first degree polynomial regression models. To make it more acceptable, think of them in the following way.

- In the zeroth degree model α_0 is the "grand mean" and σ_0^2 is the error variance when there no explanatory variable is present in the model.
- In the first degree regression model α_1 is the intercept and σ_1^2 is the error variance when there is intercept and slope present in the model, and so on.

10.2 Marginal likelihood and Bayes factor

Handling multi-model inference in the Bayesian framework is easy, at least in principle. In the single encompassing model one needs, in addition to the parameter vectors of the different models $\theta_1, \theta_2, \ldots, \theta_K$, also a random variable M to indicate the model index. Then

$$P(M = m) \equiv p(m), \qquad m = 1, \ldots, K$$

are the prior model probabilities, which have to sum to one. Typically the prior model probabilities are chosen to be uniform. Further,

$$p(\theta_m \mid M = m) \equiv p(\theta_m \mid m),$$

is the prior on θ_m in model m,

$$p(y \mid \theta_m, M = m) \equiv p(y \mid \theta_m, m),$$

is the likelihood within model m, and

$$p(\theta_m \mid y, M = m) \equiv p(\theta_m \mid y, m)$$

is the posterior for θ_m within model m.

For model selection, the most interesting quantities are the posterior model probabilities,

$$P(M = m \mid y) \equiv p(m \mid y), \qquad m = 1, \ldots, K.$$

By Bayes rule,

$$p(m \mid y) = \frac{p(y \mid m)\, p(m)}{p(y)}, \quad \text{where} \quad p(y) = \sum_{m=1}^{K} p(y \mid m)\, p(m) \tag{10.1}$$

Here $p(y \mid m)$ is usually called the **marginal likelihood** of the data within model m, or simply the marginal likelihood of model m. Of course, this marginal likelihood is different from the marginal likelihood we discussed in connection with the EM algorithm. Other terms like *marginal density of the data*, *integrated likelihood*, *prior predictive (density)*, *predictive likelihood* or *evidence* are also all used in the literature. The marginal likelihood of model m is obtained by averaging the likelihood using the prior as the weight, both within model m, i.e.,

$$p(y \mid m) = \int p(y, \theta_m \mid m)\, \mathrm{d}\theta_m = \int p(\theta_m \mid m)\, p(y \mid \theta_m, m)\, \mathrm{d}\theta_m. \tag{10.2}$$

In other words, the marginal likelihood is the normalizing constant needed in order to make prior times likelihood within model m to integrate to one,

$$p(\theta_m \mid y, m) = \frac{p(\theta_m \mid m)\, p(y \mid \theta_m, m)}{p(y \mid m)}.$$

The **Bayes factor** BF_{kl} for comparing model k against model l is defined to be the *ratio of posterior to prior odds*, or in more detail, the posterior odds in favor of model k against model l divided by the prior odds in favor of model k against model l, i.e.,

$$\mathrm{BF}_{kl} = \frac{P(M = k \mid y)}{P(M = l \mid y)} \bigg/ \frac{P(M = k)}{P(M = l)} \tag{10.3}$$

By Bayes rule (10.1), the Bayes factor equals the ratio of the two marginal likelihoods,

$$\mathrm{BF}_{kl} = \frac{p(y \mid M = k)}{p(y \mid M = l)} \tag{10.4}$$

From this we see immediately that $\mathrm{BF}_{lk} = 1/\,\mathrm{BF}_{kl}$. There are tables available (due to Jeffreys and other people) for interpreting the value of the Bayes factor.

One can compute the posterior model probabilities $p(m \mid y)$, if one knows the prior model probabilities and either the marginal likelihoods for all the models, or the Bayes factors for all pairs of models. Having done this, we may restrict our attention to the best model which has the largest posterior probability. Alternatively we might want to consider all those models whose posterior probabilities are nearly equal to that of the best model.

If one needs to form predictions for future observations Y^* which are conditionally independent of the observations, then one might form the predictions by **model averaging**, i.e., by using the predictive distribution

$$\begin{aligned} p(y^* \mid y) &= \sum_{m=1}^{K} \int p(y^*, m, \theta_m \mid y) \,\mathrm{d}\theta_m \\ &= \sum_{m=1}^{K} \int p(y^* \mid m, \theta_m, y)\, p(m \mid y)\, p(\theta_m \mid m, y) \,\mathrm{d}\theta_m \\ &= \sum_{m=1}^{K} p(m \mid y) \int p(y^* \mid m, \theta_m)\, p(\theta_m \mid m, y) \,\mathrm{d}\theta_m, \end{aligned}$$

where on the last line we used the assumption that the data Y and the future observation Y^* are conditionally independent within each of the models m, conditionally on the parameter vector θ_m. The predictive distribution for future data is obtained by averaging the within-model predictive distributions using posterior model probabilities as weights.

Similarly, we could consider the posterior distribution of a function of the parameter vector, which is meaningful in all of the candidate models. In the binary regression example, such a parameter could be LD50 (lethal dose 50 %) which is defined as the value of the covariate x which gives success probability 50 %. Such a parameter could be estimated with model averaging.

In multi-model inference one should pay close attention to the formulation of the within-model prior distributions. While the within-model posterior distributions are usually robust against the specification of the within-model prior, the same is not true for the marginal likelihood. In particular, in a multi-model situation one cannot use improper priors for the following reason. If the prior for model m is improper, i.e.,

$$p(\theta_m \mid m) \propto h_m(\theta_m)$$

where the integral of h_m is infinite, then

$$c\, h_m(\theta_m), \qquad \text{with } c > 0 \text{ arbitrary},$$

is an equally valid expression for the within-model prior. Taking $h_m(\theta_m)$ as the prior within model m in eq. (10.2) leads to the result

$$p_1(y \mid m) = \int h_m(\theta_m)\, p(y \mid \theta_m, m)\, \mathrm{d}\theta_m$$

whereas the choice $c\, h_m(\theta_m)$ leads to the result

$$p_c(y \mid m) = c\, p_1(y \mid m).$$

Therefore, if the prior for model m is improper, then we cannot assign any meaning to the marginal likelihood for model m, and the same difficulty applies to the Bayes factor, as well.

Many researchers regard the sensitivity of the marginal likelihood to the within model prior specifications a very serious drawback. This difficulty has led to many proposals for model comparison which do not depend on marginal likelihoods and Bayes factors. However, we will continue to use them for the rest of this chapter. Therefore we suppose that

- we have specified the entire collection of candidate models (this the $\mathcal{M}$-closed assumption);
- we have successfully formulated proper and informative priors for each of the candidate models.

10.3 Approximating marginal likelihoods

If we use a conjugate prior in model m, then we can calculate its marginal likelihood analytically, e.g., by using Bayes rule in the form

$$p(y \mid m) = \frac{p(\theta_m \mid m)\, p(y \mid \theta_m, m)}{p(\theta_m \mid y, m)}, \tag{10.5}$$

where θ_m is any point in the parameter space of model m, and all the terms on the right-hand side (prior density, likelihood, and posterior density, each of them within model m, respectively) are available in a conjugate situation. This form of the Bayes rule is also known by the name *candidate's formula.* In order to simplify the notation, we will drop the conditioning on the model m from the notation for the rest of this section, since we will discuss estimating the marginal likelihood for a single model at a time. For example, in the rest of this section we will write candidate's formula (10.5) in the form

$$p(y) = \frac{p(\theta)\, p(y \mid \theta)}{p(\theta \mid y)}. \tag{10.6}$$

Hopefully, leaving the model under discussion implicit in the notation does not cause too much confusion to the reader. If it does, add conditioning on m to each of the subsequent formulas and add the subscript m to each occurrence of θ and modify the text accordingly.

When the marginal likelihood is not available analytically, we may try to estimate it. One idea is based on estimating the posterior ordinate $p(\theta \mid y)$ in candidate's formula (10.6) at some point θ_h having high posterior density (such as the posterior mean estimated by MCMC). The result can be called the *candidate's estimator* for the marginal likelihood. Suppose that the parameter can be divided into two blocks $\theta = (\theta_1, \theta_2)$ such that the full conditional distributions $p(\theta_1 \mid \theta_2, y)$ and $p(\theta_2 \mid \theta_1, y)$ are both available analytically. By the multiplication rule

$$p(\theta_1, \theta_2 \mid y) = p(\theta_1 \mid y)\, p(\theta_2 \mid \theta_1, y).$$

We might estimate the marginal posterior ordinate of θ_1 at $\theta_{h,1}$ by the Rao-Blackwellized estimate

$$\hat{p}(\theta_{h,1} \mid y) = \frac{1}{N} \sum_{i=1}^{N} p(\theta_{h,1} \mid \theta_2^{(i)}, y),$$

where $(\theta_1^{(i)}, \theta_2^{(i)}), i = 1, \ldots, N$ is a sample from the posterior, e.g., produced by MCMC. Then the joint posterior at $\theta_h = (\theta_{h,1}, \theta_{h,2})$ can be estimated by

$$\hat{p}(\theta_{h,1}, \theta_{h,2} \mid y) = \hat{p}(\theta_{h,1} \mid y)\, p(\theta_{h,2} \mid \theta_{h,1}, y).$$

This approach was proposed in Chib [5] where one can also find extensions to more than two blocks.

Approximating the marginal likelihood is an ideal application for Laplace's method. Recall that the basic idea of Laplace's method is to approximate a d-dimensional integral of the form

$$I = \int g(\theta)\, \exp(L(\theta))\, \mathrm{d}\theta$$

by replacing $L(\theta)$ by its quadratic approximation centered on the mode $\tilde{\theta}$ of $L(\theta)$ and by replacing $g(\theta)$ with $g(\tilde{\theta})$. The result was

$$I \approx \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, g(\tilde{\theta})\, \mathrm{e}^{L(\tilde{\theta})},$$

where Q is the negative Hessian of $L(\theta)$ evaluated at the mode $\tilde{\theta}$.

If we start from the representation

$$p(y) = \int p(\theta)\, p(y \mid \theta)\, \mathrm{d}\theta = \int \exp\left[\log\left(p(\theta)\, p(y \mid \theta)\right)\right]\, \mathrm{d}\theta,$$

and then apply Laplace's method, we get the approximation

$$\hat{p}_{\mathrm{Lap}}(y) = p(\tilde{\theta})\, p(y \mid \tilde{\theta})\, \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}, \tag{10.7}$$

where $\tilde{\theta}$ is the posterior mode (i.e. the maximum a posterior estimate, or MAP estimate), and and Q is the negative Hessian of the logarithm of the unnormalized posterior density

$$\theta \mapsto \log\left(p(\theta)\, p(y \mid \theta)\right)$$

evaluated at the mode $\tilde{\theta}$.

Another possibility is to start from the representation

$$p(y) = \int p(\theta)\, \exp\left[\log p(y \mid \theta)\right]\, \mathrm{d}\theta$$

and then integrate the quadratic approximation for the log-likelihood centered at its mode, the maximum likelihood estimate (MLE). This gives the result

$$\hat{p}_{\mathrm{Lap}}(y) = p(\hat{\theta})\, p(y \mid \hat{\theta})\, \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}, \tag{10.8}$$

where $\hat{\theta}$ is the MLE, and Q is now the *observed information matrix* (evaluated at the MLE), which is simply the negative Hessian of the log-likelihood evaluated at the MLE.

One can also use various Monte Carlo approaches to approximate the marginal likelihood. Since

$$p(y) = \int p(y \mid \theta)\, p(\theta)\, \mathrm{d}\theta,$$

naive Monte Carlo integration gives the estimate

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} p(y \mid \theta^{(i)}), \tag{10.9}$$

where we average the likelihood values using a sample $\theta^{(i)}, i = 1, \ldots, N$ from the prior $p(\theta)$. If the posterior corresponds to a large data set $y_1, \ldots, y_n$, then typically the model m likelihood is very peaked compared to the prior. In this situation the estimate (10.9) has typically huge variance, since very few of the sample points hit the region with high likelihood values, and these few values dominate the sum.

A better approach would be to write the marginal likelihood as

$$p(y) = \int \frac{p(y \mid \theta)\, p(\theta)}{g(\theta)}\, g(\theta)\, \mathrm{d}\theta,$$

where $g(\theta)$ is an importance sampling density for the model under consideration. This yields the importance sampling estimate

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} \frac{p(y \mid \theta^{(i)})\, p(\theta^{(i)})}{g(\theta^{(i)})}, \tag{10.10}$$

where $\theta^{(i)}, i = 1, \ldots, N$ is a sample drawn from the importance sampling density g. In order to obtain low variance, g should be an approximation to the posterior density, and g should have heavier tails than the true posterior. For example, g could be a multivariate t distribution centered on the posterior mode, the shape of which is chosen using an estimate of the posterior covariance matrix.

The marginal likelihood can also be estimated using an MCMC sample drawn from the posterior distribution $p(\theta \mid y)$. Let g be a probability density defined on the parameter space. Integrating the identity

$$g(\theta) = g(\theta) \frac{p(y)\, p(\theta \mid y)}{p(y \mid \theta)\, p(\theta)}$$

over the parameter space gives

$$\frac{1}{p(y)} = \int \frac{g(\theta)}{p(y \mid \theta)\, p(\theta)}\, p(\theta \mid y) \, \mathrm{d}\theta$$

If $\theta^{(i)}, i = 1, \ldots, N$ is a MCMC sample from the posterior, then we can estimate the marginal likelihood as follows,

$$\hat{p}(y) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta^{(i)})}{p(y \mid \theta^{(i)})\, p(\theta^{(i)})} \right]^{-1}. \tag{10.11}$$

Here we calculate the harmonic mean of prior times likelihood divided by the density g ordinates evaluated at the sample points, $p(y \mid \theta^{(i)})\, p(\theta^{(i)}) / g(\theta^{(i)})$. This is the *generalized harmonic mean estimator* suggested by Gelfand and Dey [9]. The function g should be chosen so that it has approximately the same shape as the posterior density $p(\theta \mid y)$ but in this case the tails of g should be thin compared to the tails of the posterior.

If one selects g to be the prior $p(\theta)$ then formula (10.11) suggests that one could estimate the marginal likelihood by calculating the harmonic mean of the likelihood values $p(y \mid \theta^{(i)})$. This is the (in)famous harmonic mean estimator first discussed by Newton and Raferty [14]. The harmonic mean estimator has typically infinite variance and is numerically unstable, and therefore should not be used at all.

Besides these, many other sampling-based approaches have been proposed in the literature (e.g., bridge sampling).

After all the marginal likelihoods $p(y \mid M = j)$ have been estimated one way or another, then one can estimate the posterior model probabilities based on eq. (10.1), i.e., by using

$$\hat{p}(m \mid y) = \frac{p(m)\, \hat{p}(y \mid m)}{\sum_{j=1}^{K} p(M = j)\, \hat{p}(y \mid M = j)}, \qquad m = 1, \ldots, K.$$

The denominator is just the sum of the numerators when m takes the values from 1 to K.

An obvious way to estimate the Bayes factor BF_{kl} is to calculate the ratio of two marginal likelihood estimators,

$$\widehat{\mathrm{BF}}_{kl} = \frac{\hat{p}(y \mid M = k)}{\hat{p}(y \mid M = l)}.$$

However, there are also more direct ways of estimating the Bayes factor, such as path sampling.

10.4 BIC and other information criteria

Information criteria consist of two parts: a measure of fit of the model to the data, and a penalty for the complexity of the model. The two most famous such criteria are AIC and BIC.

Our starting point for Schwarz's Bayes(ian) Information Criterion, BIC (other acronyms: SBIC, SBC, SIC), is the Laplace approximation to the marginal posterior based on the MLE (10.8). Taking logarithms and multiplying by minus two gives

$$-2\log p(y) \approx -2\log p(\hat{\theta}) - 2\log p(y \mid \hat{\theta}) - d\log(2\pi) + \log\det(Q).$$

where $\hat{\theta}$ is the MLE and Q is the observed information matrix (at the MLE). We concentrate on the case where we have n observations y_i which are conditionally independent, i.e.,

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta),$$

from which

$$\begin{aligned} \log p(y \mid \theta) &= \sum_{i=1}^{n} \log p(y_i \mid \theta) \\ Q &= n \left[\frac{1}{n} \sum_{i=1}^{n} (-1) \frac{\partial^2}{\partial\theta\,\partial\theta^T} \log p(y_i \mid \theta) \right]_{|\theta=\hat{\theta}} \end{aligned}$$

One can argue (based on a multivariate version of the SLLN) that the average inside the square brackets is approximately equal to the corresponding expected value $J_1(\hat{\theta})$, the expected (or Fisher) information matrix due to a single observation, evaluated at the MLE, where

$$J_1(\theta) = -\int p(y \mid \theta) \frac{\partial^2}{\partial\theta\,\partial\theta^T} \log p(y \mid \theta)\,\mathrm{d}\theta.$$

Hence we approximate

$$Q \approx nJ_1(\hat{\theta}) \quad \Rightarrow \quad \det(Q) \approx n^d \det(J_1(\hat{\theta}))$$

This gives

$$-2\log p(y) \approx -2\log p(y \mid \hat{\theta}) + d\log n - 2\log p(\hat{\theta}) - d\log(2\pi) + \log\det(J_1(\hat{\theta})).$$

The final step is to drop all the terms which remain constant as the sample size n increases, and this gives the approximation

$$-2\log p(y) \approx -2\log p(y \mid \hat{\theta}) + d\log n.$$

We have now derived the Bayesian information criterion for model m, namely

$$\mathrm{BIC}_m = -2L_m + d_m \log n. \tag{10.12}$$

Here

$$L_m = \log p(y \mid \hat{\theta}_m, m)$$

is the maximized log-likelihood for model m, d_m is the dimensionality of the model m parameter space, and n is the sample size. (Warning: in the literature you will find several different definitions for BIC.) This criterion can be used for rough comparison of competing models: smaller values of BIC correspond to better models. Most of the time, more complex models lead automatically to higher values of the maximized likelihood, but the term $d_m \log n$ penalizes for increased model complexity.

The approximations involved in the derivation of BIC are rather crude, and therefore usually $\exp(-\frac{1}{2}\,\mathrm{BIC}_m)$ is a rather poor approximation to the marginal likelihood of model m. One should pay attention only to the differences

$$\Delta\,\mathrm{BIC}_{kl} = \mathrm{BIC}_k - \mathrm{BIC}_l = -2\log\frac{L_k}{L_l} + (d_k - d_l)\log n.$$

However, Kass and Wasserman [13] have constructed a special prior, the unit information prior, under which $\exp(-\frac{1}{2}\,\mathrm{BIC}_m)$ does give a good approximation to the model m marginal likelihood. Nevertheless, if we approximate $p(y \mid m)$ by $\exp(-\frac{1}{2}\,\mathrm{BIC}_m)$, and assume that the prior model probabilities are equal, then we may estimate the posterior model probabilities by

$$\hat{p}(m \mid y) = \frac{\exp(-\frac{1}{2}\,\mathrm{BIC}_m)}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\,\mathrm{BIC}_k)}. \tag{10.13}$$

BIC resembles the equally famous Akaike information criterion, AIC,

$$\mathrm{AIC}_m = -2L_m + 2d_m.$$

In addition, the alphabet soup of information criteria includes such acronyms as AIC_c (corrected AIC), cAIC (conditional AIC), mAIC; AFIC; BFIC; DIC; FIC; HQ; NIC; QAIC and QAIC_c; RIC; TIC; WIC. Furthermore, there are several other famous model selection criteria available, such as Mallows' C_p (for regression problems with normal errors), or Akaike's FPE (final prediction error). Also Rissanen's MDL (minimum description length) principle can be used. See, e.g., Burnham and Anderson [2] and Claeskens and Hjort [6].

In some statistical models it is not always clear what one should use as the sample size n in these information criteria. What is more, in complex models the number of parameters is not necessarily clearly defined. Spiegelhalter *et al.* [16] suggest that in such a situation one may use their deviance information criterion, DIC, defined by

$$\mathrm{DIC}_m = 2\overline{D(\theta_m, m)} - D(\bar{\theta}_m, m), \tag{10.14}$$

where $D(\theta_m, m)$ is the deviance, or minus twice the log-likelihood of model m,

$$D(\theta_m, m) = -2\log p(y \mid \theta_m, m),$$

$\bar{\theta}_m$ is the posterior mean of θ_m, and $\overline{D(\theta_m, m)}$ is the posterior mean of $D(\theta_m, m)$ within model m. These quantities are estimated using separate MCMC runs for each of the models. WinBUGS and OpenBUGS have automatic facilities for calculating DIC, and therefore it has become the widely used among Bayesian statisticians. As with AIC and BIC, smaller DIC indicates a better model.

The authors interpret

$$d_m^{\text{eff}} = \overline{D(\theta_m, m)} - D(\bar{\theta}_m, m)$$

as the number of effective parameters for model m, and therefore DIC_m can written in the form

$$\text{DIC}_m = D(\bar{\theta}_m, m) + 2d_m^{\text{eff}},$$

which shows its connection to AIC. The authors show that d_m^{eff} gives a reasonable definition for the effective number of parameters in many cases. If there is strong conflict between the prior and the data, then the effective number of parameters may turn out have a negative value, which does not make sense.

In order to use DIC, one must decide which expression to use as the likelihood. In complex statistical models, e.g., hierarchical models or random effects models, even this choice is not clear cut. Consider the hierarchical model, which has a prior on the hyperparameters ψ and which factorizes as follows

$$p(y, \theta, \psi) = p(y \mid \theta)\, p(\theta \mid \psi)\, p(\psi).$$

If one focuses the attention to the parameter vector θ, then the likelihood expression is $p(y \mid \theta)$. However, it would be equally valid to consider the vector ψ to be the true parameter vector. If one focuses on ψ, then one should select

$$p(y \mid \psi) = \int p(y, \theta \mid \psi)\, \mathrm{d}\psi = \int p(y \mid \theta)\, p(\theta \mid \psi)\, \mathrm{d}\psi$$

as the likelihood. In some models $p(y \mid \psi)$ is available in closed form. Otherwise, evaluating this likelihood may be problematic. Generally, the DIC values for $p(y \mid \theta)$ and $p(y \mid \psi)$ are different. Spiegelhalter *et al.* suggest that one should formulate clearly the focus of the analysis, and calculate DIC using the corresponding likelihood expression. They also point out that DIC_m changes, if one reparametrizes model m.

10.5 Sum space versus product space

In this section we discuss an embedding of the multi-model inference problem in the product-space formulation of the problem. We revert to the explicit notation of Section 10.2. Let

$$S_m \subset \mathbb{R}^{d_m}, \qquad m = 1, \ldots, K$$

be the parameter space of model m. We call the set

$$S_{\text{sum}} = \cup_{m=1}^{K} \{m\} \times S_m \tag{10.15}$$

the sum of the parameter spaces. (In topology, this would be called the topological sum, direct sum, disjoint union or coproduct of the spaces S_m.) Any point $x \in S_{\text{sum}}$ is of the form

$$x = (m, \theta_m), \qquad \text{where } m \in \{1, \ldots, K\} \text{ and } \theta_m \in S_m.$$

The quantities of inferential interest discussed in Section 10.2 can be defined based on the joint posterior

$$p(m, \theta_m \mid y), \qquad m \in \{1, \ldots, K\}, \quad \theta_M \in S_m,$$

which itself is defined on the sum space through the joint distribution specification

$$p(m, \theta_m, y) = p(m)\, p(\theta_m \mid m)\, p(y \mid \theta_m, m), \qquad m \in \{1, \dots, K\}, \quad \theta_M \in S_m.$$

Designing a MCMC algorithm which uses the sum space as its state space is challenging. For instance, the dimensionality of the parameter vector may change each time the model indicator changes. Specifying the sum-space formulation directly in BUGS seems to be impossible, since in the sum-space formulation parameter θ_m exits only when the model indicator has the value m. Green [11] was first to propose a trans-dimensional MCMC algorithm which works directly in the sum space, and called it the reversible jump MCMC (RJMCMC) algorithm.

Most of the other multi-model MCMC algorithms are conceptually based on the product-space formulation, where the state space is the Cartesian product of the model space $\{1, \dots, K\}$ and the Cartesian product of the parameter spaces of the models,

$$S_{\text{prod}} = S_1 \times S_2 \times \cdots \times S_K. \tag{10.16}$$

For the rest of the section, θ without a subscript will denote a point point $\theta \in S_{\text{prod}}$. It is of the form

$$\theta = (\theta_1, \theta_2, \dots, \theta_K), \tag{10.17}$$

where each of the $\theta_m \in S_m$. The product space is larger than the sum space, and the product-space formulation requires that we set up the joint distribution

$$p(m, \theta, y), \qquad m \in \{1, \dots, K\}, \quad \theta \in S_{\text{prod}}.$$

In contrast, in the sum-space formulation the parameters $\{\theta_k, k \neq m\}$ do not exist on the event $M = m$, and so we cannot speak of

$$p(m, \theta, y) = p(m, \theta_1, \dots, \theta_K, y)$$

within the sum-space formulation. We are obliged to set up the product-space formulation in such a way that the marginals

$$p(m, \theta_m, y), \qquad m \in \{1, \dots, K\}$$

remain the same as in the original sum-space formulation. For this reason we will not make a notational difference between the sum-space and the product-space formulation of the multi-model inference problem.

The preceding means that we embed the multi-model inference problem in the product-space formulation. While specifying the sum-space model is not possible in WinBUGS/OpenBUGS, it is straightforward to specify the product-space version of the same problem.

When we do posterior inference in the product-space formulation, only the marginals

$$p(m, \theta_m \mid y), \qquad m \in \{1, \dots, K\}$$

of the joint posterior

$$p(m, \theta \mid y) = p(m, \theta_1, \dots, \theta_K \mid y)$$

are of inferential relevance. The other aspects of the joint distribution are only devices, which allow us to work with the easiear product-space formulation.

If $(m^{(i)}, \theta^{(i)}), i = 1, \dots, N$ is a sample from the posterior $p(m, \theta \mid y)$, then for inference we use only the component $\theta^{(i)}_{m^{(i)}}$ of $\theta^{(i)}$, which is the parameter vector of that model $m^{(i)}$ which was visited during the i'th iteration. In particular, the posterior model probabilities $p(M = j \mid y)$ can be estimated by tabulating the relative frequencies of each of the possibilities $m^{(i)} = j$.

10.6 Carlin and Chib method

Carlin and Chib [3] use the product-space formulation, where

$$p(m, \theta, y) = p(m)\, p(\theta, y \mid m), \tag{10.18}$$

and $p(m)$ is the familiar model m prior probability. The conditional density $p(\theta, y \mid m)$ is selected to be

$$p(\theta, y \mid m) = p(\theta_m \mid m)\, p(y \mid \theta_m, m) \prod_{k \neq m} g_k(\theta_k \mid y) \tag{10.19}$$

Here $p(\theta_m \mid m)$ and $p(y \mid \theta_m, m)$ are the prior and the likelihood within model m, respectively. In addition, we need K densities $g_k(\theta_k \mid y)$, $k = 1, \dots, K$ which can be called *pseudo priors* or *linking densities*. The linking density $g_k(\theta_k \mid y)$ is an arbitrary density on the parameter space of model k. It can be shown that this is a valid formulation of the product-space joint density. No circularity results from allowing the linking densities to depend on the data. Further, this specification leads to the marginals $p(m, \theta_m, y)$ of the sum-space formulation irrespective of how one specifies the linking densities.

Let us consider the case of two models ($K = 2$) in more detail. According to (10.18) and (10.19), the joint density $p(m, \theta, y)$ is

$$\begin{cases} p(M = 1)\, p(\theta_1 \mid M = 1)\, p(y \mid \theta_1, M = 1)\, g_2(\theta_2 \mid y) & \text{when } m = 1 \\ p(M = 2)\, p(\theta_2 \mid M = 2)\, p(y \mid \theta_2, M = 2)\, g_1(\theta_1 \mid y) & \text{when } m = 2. \end{cases}$$

We see easily that the marginal densities $p(m, \theta_m, y), m = 1, 2$ are the same as in the sum-space formulation: just integrate out

$$\begin{aligned} &\theta_2 \quad \text{from } p(m = 1, \theta_1, \theta_2, y) \\ &\theta_1 \quad \text{from } p(m = 2, \theta_1, \theta_2, y). \end{aligned}$$

Hence we have checked the validity of the specification.

While the specification of the linking densities $g_k(\theta_k \mid y)$ does not influence the validity of the product-space formulation, this matter does have a critical influence on the efficiency of the ensuing MCMC algorithm. A recommended choice is to select $g_k(\theta_k \mid y)$ to be a tractable approximation to the posterior distribution within model k, such as a multivariate normal approximation or a multivariate t approximation. Building such approximations usually requires pilot MCMC runs of all the models under consideration.

Carlin and Chib use the Gibbs sampler. For this we need the full conditionals. First,

$$p(m \mid \theta, y) \propto p(m, \theta, y), \qquad m = 1, \dots, K.$$

which is easy to simulate since it is a discrete distribution. Next,

$$p(\theta_m \mid M = m, \theta_{-m}, y) \propto p(\theta_m \mid M = m)\, p(y \mid \theta_m, M = m).$$

Hence this full conditional is the within model m posterior distribution. Finally, for $k \neq m$

$$p(\theta_k \mid M = m, \theta_{-k}, y) = g_k(\theta_k \mid y)$$

is the linking density for θ_k.

These full conditionals lead to a Gibbs sampler (or a Metropolis-within-Gibbs sampler), where one first selects a new value m^{cur} for the model indicator, drawing the new value from the full conditional $p(m \mid \theta, y)$. After this, one updates the parameter vectors of all the models. For m equal to m^{cur} (for the currently visited model), the new value for θ_m is drawn from the posterior of model m (and if this is not feasible, one may execute a M–H step for the same target $p(\theta_m \mid y, m)$, instead). For all other values of k, the new value of θ_k is drawn from the linking density $g_k(\theta_k \mid y)$.

Many other product-space algorithms have been developed as well, see [10] for a review.

10.7 Reversible jump MCMC

Green's reversible jump MCMC algorithm (RJMCMC) [11] uses a Markov chain whose state space is the sum space. We discuss a simplified version of RJMCMC, where there is only one type of move available for moving from model m to model k. We also assume that the distributions of the parameter vectors θ_m in all of the models are continuous.

The RJMCMC works like the Metropolis–Hastings algorithm. One first proposes a new state, and then accepts the proposed state as the new state of the Markov chain, if $v < r$, where r is the test ratio and v is a fresh uniform $\text{Uni}(0, 1)$ random variate. The difference lies in the details: how the proposed state is generated, and how the test ratio is calculated. The state space of the Markov chain is the sum space S_{sum}, and the target distribution π is the posterior distribution

$$\pi(m, \theta_m) = p(m, \theta_m \mid y), \qquad m \in \{1, \dots, K\}, \quad \theta_m \in S_m.$$

When the current state of the chain is (m, θ_m), then the proposal (k, θ_k) and the test ratio r are calculated as described in algorithm 20. The proposed model k is drawn from the pmf $\beta(\cdot \mid m)$. If $k = m$, then one executes an ordinary M–H step within model m. If $k \neq m$, then one proposes a new parameter vector θ_k in model k as follows. First one generates a noise vector u_m associated with θ_m from noise density $g(\cdot \mid \theta_m, m \to k)$ specific for the move $m \to k$. Then one calculates θ_k and u_k by applying the so called dimension-matching function $T_{m\to k}$. The dimension-matching functions are defined for all moves $m \neq k$, and they have to satisfy the following compatibility conditions, which are also called dimension-matching conditions.

We assume that for each move $m \to k$ where $m \neq k$ there exists a diffeomorphic correspondence

$$(\theta_k, u_k) = T_{m\to k}(\theta_m, u_m)$$

with inverse $T_{k\to m}$, i.e.,

$$(\theta_k, u_k) = T_{m\to k}(\theta_m, u_m) \quad \Leftrightarrow \quad (\theta_m, u_m) = T_{k\to m}(\theta_k, u_k). \tag{10.20}$$

Here u_m is the noise variable associated with θ_m and u_k is the noise variable associated with θ_k (for the move $m \to k$). Here the dimensions have to match,

$$\dim(\theta_m) + \dim(u_m) = \dim(\theta_k) + \dim(u_k),$$

since otherwise such a diffeomorphism cannot exist.

Algorithm 20: One step of the RJMCMC algorithm.

```
Input: The current state of the chain is $(m, \theta_m)$.
Assumption: The correspondences (10.20) are diffeomorphic.
Result: Proposed next value $(k, \theta_k)$ as well as the test ratio $r$.
Draw $k$ from the pmf $\beta(k \mid m)$.
if $k = m$ then
    generate the proposal $\theta_k$ with some M–H proposal mechanism within
    model $m$, and calculate $r$ with the ordinary formula for the M–H ratio.
else
    Draw the noise variable $u_m$ from density $g(u_m \mid \theta_m, m \to k)$. (This
    step is omitted, if the move $m \to k$ is deterministic.)
    Calculate $\theta_k$ and $u_k$ by the diffeomorphic correspondence specific for
    the move $m \to k$,
        $(\theta_k, u_k) \leftarrow T_{m\to k}(\theta_m, u_m)$.
    Calculate $r$ by
        $r \leftarrow \frac{\pi(k, \theta_k)}{\pi(m, \theta_m)} \frac{\beta(m \mid k)}{\beta(k \mid m)} \frac{g(u_k \mid \theta_k, k \to m)}{g(u_m \mid \theta_m, m \to k)} \left| \frac{\partial(\theta_k, u_k)}{\partial(\theta_m, u_m)} \right|$
end
```

Notice the following points concerning this method.

- When we calculate the test ratio r for the move $m \to k$, we have to use the quantities $\beta(m \mid k)$ and $g(u_k \mid \theta_k, k \to m)$ which correspond to the distributions from which we simulate, when the current state is (k, θ_k) and the move is selected to be $k \to m$.

- The Jacobian is the Jacobian of the transformation which maps (θ_m, u_m) to (θ_k, u_k), when the move is $m \to k$, i.e.,

  $$\frac{\partial(\theta_k, u_k)}{\partial(\theta_m, u_m)} = \frac{\partial T_{m\to k}(\theta_m, u_m)}{\partial(\theta_m, u_m)}.$$

  We will see in Sec. 11.8 that the Jacobian term arises from the change-of-variables formula for integrals, the reason being the fact that the proposal θ_k is calculated in an indirect way, by applying the deterministic function $T_{m\to k}$ to the pair (θ_m, u_m).

- One of the moves $m \to k$ or $k \to m$ can deterministic. If the move $m \to k$ is deterministic, then the associated noise variable, u_m is not defined nor simulated, the dimension-matching function is $(\theta_k, u_k) = T_{m \to k}(\theta_m)$, and the noise density value, $g(u_m \mid \theta_m, m \to k)$ gets replaced by the constant one. The same rules apply, when the move $k \to m$ is deterministic.

- The target density ratio is calculated by

$$\frac{\pi(k, \theta_k)}{\pi(m, \theta_m)} = \frac{P(M = k)}{P(M = m)} \frac{p(\theta_k \mid M = k)}{p(\theta_m \mid M = m)} \frac{p(y \mid M = k, \theta_k)}{p(y \mid M = m, \theta_m)}$$

- The test ratio r can be described verbally as

$$r = (\text{prior ratio}) \times (\text{likelihood ratio}) \times (\text{proposal ratio}) \times (\text{Jacobian})$$

It is possible to extend the method to the situation where we have discrete components in the state vectors θ_m of some of the models m. It is also possible to have more than one type of move between any given models. See the original paper by Green [11] for more details. The choice of the dimension-matching functions is critical to ensure good mixing of the Markov chain. In this respect, Green's automatic generic trans-dimensional sampler [12] seems to be very promising.

10.8 Discussion

In this chapter we have seen many different approaches for estimating the posterior model probabilities, which are central quantities both for model selection and model averaging. One approach is to estimate the marginal likelihoods for all of the models, and a distinct approach is to set up an MCMC algorithm which works over the model space and the parameter spaces of each of the models. Many variations are possible within each of the two approaches. What are the pros and cons of these approaches?

If the list of candidate models is short, then it is usually easy to estimate the marginal likelihoods for each of the models separately. However, if the list of candidate models is large and if it is suspected that only few of the models are supported by the data, then the best option might be to implement a multi-model MCMC sampler. However, getting the multi-model sampler to mix across the different models can be a challenging exercise and might require investigating pilot runs within each of the candidate models. Mixing within the parameter space of a single model is usually very much easier to achieve.

10.9 Literature

In addition to the original articles, see the books [4, 15, 7, 8],which also address model checking (model assessment, model criticism) which we have neglected in this chapter.

Bibliography

[1] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* John Wiley & Sons, 2000. First published in 1994.

[2] Kenneth B. Burnham and David R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer, 2nd edition, 2002.

[3] Bradley P. Carlin and Siddhartha Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57:473–484, 1995.

[4] Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis.* Chapman & Hall/CRC, 3rd edition, 2009.

[5] Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321, 1995.

[6] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging.* Cambridge University Press, 2008.

[7] Peter Congdon. *Bayesian Statistical Modelling.* Wiley, 2nd edition, 2006.

[8] Dani Gamerman and Hedibert F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall/CRC, second edition, 2006.

[9] A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, 56:501–514, 1994.

[10] Simon J. Godsill. On the relationship between Markov chain Monte Carlo mthods for model uncertainty. *Journal of Computational and Graphical Statistics*, 10:230–248, 2001.

[11] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[12] Peter J. Green. Trans-dimensional Markov chain Monte Carlo. In Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems.* Oxford University Pres, 2003.

[13] R. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Scwharz criterion. *Journal of the American Statistical Association*, 90:928–934, 1995.

[14] M. A. Newton and A. E. Raferty. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994.

[15] Ioannis Ntzoufras. *Bayesian Modeling Using WinBUGS.* Wiley, 2009.

[16] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64:583–639, 2002.