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 an activity 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,

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

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. Should we take this into account in selecting the size of the test?

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.

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 exist 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 does not apply.

In addition to the hypothesis testing approach, a frequentist statistician might compare models using one of several information criteria, 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 is defined on 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,

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

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 of model m 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.,

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

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 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.

Another approach is to use the Laplace method to approximate the integral

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

This gives the marginal likelihood 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 $\tilde{\theta}$.

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

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^{N} p(y \mid \theta^{(i)}), \tag{10.8}$$

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 (10.8) 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

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

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,

$$\hat{p}(y) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{g(\theta^{(i)})}{p(y \mid \theta^{(i)})\, p(\theta^{(i)})} \right]^{-1}. \tag{10.10}$$

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 [8]. 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 (10.10) 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 [13]. 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 given in eq. (10.7). Taking logarithms and multiplying by minus two gives

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

where $\tilde{\theta}$ is the MAP estimate. To simplify this we note that in large samples log prior is negligible compared to the log likelihood, and the MAP estimate $\tilde{\theta}$ is roughly equal to the maximum likelihood estimate $\hat{\theta}$. Further, when the n data points are conditionally independent, then $Q \approx nI$, where I is the expected Fisher information for sample size one. Dropping terms which remain constant as the sample size n increases in the resulting approximation, we get the Bayes information criterion for model m, namely

$$\mathrm{BIC}_m = -2L_m + d_m \log n. \tag{10.11}$$

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.

Usually BIC_m is a poor approximation for to the marginal likelihood of model m, and 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 [12] have constructed a special prior, the unit information prior, under which $-\frac{1}{2}\,\mathrm{BIC}_m$ does give a good approximation to the logarithm of the model m marginal likelihood $\log p(y \mid m)$. If such a priors are adopted, then one may estimate the posterior model probabilities by

$$\hat{p}(m \mid y) = \frac{\exp(-\frac{1}{2}\,\mathrm{BIC}_m)}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\,\mathrm{BIC}_k)}. \tag{10.12}$$

However, this approximation is also used also under other kinds of priors.

BIC resembles the equally famous Akaike information criterion, AIC, which is given by

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

In addition, the alphabet soup of information criteria includes such acronyms as AIC_c or CAIC; QAIC and QAIC_c; TIC; HQ; WIC. See, e.g., Burnham and Anderson [2].

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, such as hierarchical models or random effects models, the number of parameters is not clearly defined. Spiegelhalter *et al.* [16] suggest that in such a situation one may use their deviance information criterion, DIC, defined by

$$\mathrm{DIC}_m = 2\overline{D(\theta_m, m)} - D(\bar{\theta}_m, m), \tag{10.13}$$

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 is the case with AIC and BIC, smaller DIC indicates a better model.

The authors interpret

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

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

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

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

$$S_{\text{sum}} = \cup_{m=1}^{K} \{m\} \times S_m \tag{10.14}$$

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

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

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 [10] 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 $\{1, \ldots, K\}$ and the Cartesian product of the parameter spaces of the models,

$$S_{\text{prod}} = S_1 \times S_2 \times \cdots \times S_K. \tag{10.15}$$

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

$$\theta = (\theta_1, \theta_2, \ldots, \theta_K), \tag{10.16}$$

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

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

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

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

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

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

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

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

of the joint posterior

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

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 $(m^{(i)}, \theta^{(i)}), i = 1, \ldots, N$ 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

$$p(m, \theta, y) = p(m)\, p(\theta, y \mid m), \tag{10.17}$$

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

$$p(\theta, y \mid m) = p(\theta_m \mid m)\, p(y \mid \theta_m, m) \prod_{k \neq m} g_k(\theta_k \mid y) \tag{10.18}$$

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 $g_k(\theta_k \mid y)$, $k = 1, \ldots, K$ 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 (10.17) and (10.18), 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 product space formulation in this case.

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 proposed to use the Gibbs sampler. For this we need to calculate the full conditionals. First,

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

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 we recognize that this full conditional is the posterior within model m. 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 [9] for a review.

10.7 Reversible jump MCMC

Green's reversible jump MCMC algorithm (RJMCMC) [10] 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

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

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.,

$$(\theta_k, u_k) = T_{m\to k}(\theta_m, u_m) \quad \Leftrightarrow \quad (\theta_m, u_m) = T_{k\to m}(\theta_k, u_k). \tag{10.19}$$

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,

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

Algorithm 20: One step of the RJMCMC algorithm.

Input: The current state of the chain is (m, θ_m).
Assumption: The correspondences (10.19) are diffeomorhpic.
Result: Proposed next value (k, θ_k) as well as the test ratio r.

Draw k from the pmf $\beta(k \mid m)$.
if $k = m$ **then**
    generate the proposal θ_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 θ_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

since otherwise such a diffeomorphism cannot exist.

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 [10] for more details. See the review articles [11] and [9] for more on Bayesian methods for model choice and model averaging.

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 [11] 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, one may consult the books [4, 14, 6, 7, 15]. These books also address measures for 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] Peter Congdon. *Bayesian Statistical Modelling*. Wiley, 2nd edition, 2006.

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] Anthony O'Hagan and Jonathan Forster. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, second edition, 2004.

[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.

Chapter 11

MCMC theory

In this chapter we will finally justify the usual MCMC algorithms theoretically using the machinery of general state space Markov chains. We will prove that the Markov chains corresponding to our MCMC algorithms have the correct invariant distributions, using the concept of reversibility of a Markov chain. Additionally, we will try to understand, what the concept of irreducibility of a Markov chain means and also touch on the topic of Markov chain central limit theorems.

11.1 Transition kernel

Let S be the state space of a homogeneous Markov chain

$$\Theta^{(0)}, \Theta^{(1)}, \Theta^{(2)}, \dots$$

This means that each of the RVs $\Theta^{(i)}$ takes values in the space S. S is usually some subset of the Euclidean space. When the chain corresponds to a MCMC algorithm, where the support of the target distribution is not the whole space under consideration, then we usually choose S equal to the support of the target distribution.

Let $K(\theta, A)$ be the transition (probability) kernel of the homogeneous Markov chain, i.e., we suppose that for all $A \subset S$ we have

$$K(\theta, A) = P(\Theta^{(t+1)} \in A \mid \Theta^{(t)} = \theta). \tag{11.1}$$

As a function of $A \subset S$, the transition kernel $K(\theta, A)$ is the conditional distribution of $\Theta^{(t+1)}$ given that $\Theta^{(t)} = \theta$. Of course,

$$K(\theta, S) = 1 \qquad \forall \theta.$$

If μ is the initial distribution of the chain, i.e.,

$$\mu(A) = P(\Theta^{(0)} \in A), \qquad A \subset S,$$

then the joint distribution of $\Theta^{(0)}$ and $\Theta^{(1)}$ is

$$P_\mu(\Theta^{(0)} \in A, \Theta^{(1)} \in B) = \int_A \mu(\mathrm{d}\theta_0) K(\theta_0, B).$$

Hence the distribution of the next state is

$$P_\mu(\Theta^{(1)} \in B) = \int \mu(\mathrm{d}\theta)\, K(\theta, B), \qquad B \subset S. \tag{11.2}$$

When the domain of integration is not indicated, as here, the integral is taken over the whole space S. Here the integral is the Lebesgue integral of the function $\theta \mapsto K(\theta, B)$ with respect to the measure μ. We write the initial distribution itself, or its density, as a subscript to the P-symbol, if need be.

Recall that we call $\pi(\theta)$ a density even if it represents a discrete distribution with respect to some components of θ and a continuous distribution for others. Then integrals involving the density $\pi(\theta)$ can actually be sums with respect to some components of θ and integrals with respect to the others. If the initial distribution has a density $\pi(\theta)$, then the initial distribution itself is given by

$$\mu(A) = \int_A \pi(\theta)\, \mathrm{d}\theta.$$

In that case, the distribution of the next state given in (11.2) can be written as

$$P_\mu(\Theta^{(1)} \in B) = \int \pi(\theta)\, K(\theta, B)\, \mathrm{d}\theta \qquad B \subset S. \tag{11.3}$$

However, this distribution may or may not admit a density; which case obtains depends on the nature of the transition kernel.

In some cases (but not always) the transition kernel can be obtained from a transition density $k(\theta_1 \mid \theta_0)$ by integration,

$$K(\theta_0, B) = \int_B k(\theta_1 \mid \theta_0)\, \mathrm{d}\theta_1.$$

In such a case $k(\theta_1 \mid \theta_0)$ is the conditional density of $\Theta^{(1)}$ conditionally on $\Theta^{(0)} = \theta_0$. If the initial distribution has the density π, then (11.3) can be written as

$$P_\pi(\Theta^{(1)} \in B) = \int_{\theta_1 \in B} \int \pi(\theta_0)\, k(\theta_1 \mid \theta_0)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_0.$$

That is, the density of $\Theta^{(1)}$ can be obtained from the joint density $\pi(\theta_0)\, k(\theta_1 \mid \theta_0)$ by marginalization.

The joint distribution of the states $\Theta^{(0)}, \Theta^{(1)}$ and $\Theta^{(2)}$ is determined by

$$P_\mu(\Theta^{(0)} \in A_0, \Theta^{(1)} \in A_1, \Theta^{(2)} \in A_2) \\ = \int_{\theta_0 \in A_0} \int_{\theta_1 \in A_1} \mu(\mathrm{d}\theta_0)\, K(\theta_0, \mathrm{d}\theta_1)\, K(\theta_1, A_2)$$

where μ is the initial distribution. If the initial distribution has density π, and the transition kernel can be obtained from transition density $k(\theta_1 \mid \theta_0)$, then the previous formula just states that the joint density of $\Theta^{(0)}$, $\Theta^{(1)}$ and $\Theta^{(2)}$ is

$$\pi(\theta_0)\, k(\theta_1 \mid \theta_0)\, k(\theta_2 \mid \theta_1).$$

Iterating, we see that the initial distribution μ and the transition kernel K together determine the distribution of the homogeneous Markov chain.

11.2 Invariant distribution and reversibility

The density $\pi(\theta)$ is an **invariant density** (or stationary density or equilibrium density) of the chain (or of its transition kernel), if the Markov chain preserves it in the following sense. When the initial state has the invariant distribution corresponding to the invariant density, then all the consecutive states have to have the same invariant distribution. In particular, when the initial distribution has the invariant density π, then the the distribution of $\Theta^{(1)}$ also has to have the density π. That is,

$$P_\pi(\Theta^{(0)} \in B) = P_\pi(\Theta^{(1)} \in B), \qquad \forall B \subset S. \tag{11.4}$$

If this holds, then by induction also all the consecutive states have the same invariant distribution, so this requirement is equivalent with the requirement that π is the invariant density of the Markov chain.

By (11.3), the requirement (11.4) can also be written in terms of the transition kernel,

$$\int_B \pi(\theta)\, \mathrm{d}\theta = \int \pi(\theta)\, K(\theta, B)\, \mathrm{d}\theta, \qquad \forall B \subset S. \tag{11.5}$$

A given transition kernel may have more than one invariant densities. E.g., the kernel

$$K(\theta, A) = 1_A(\theta), \qquad A \subset S$$

corresponds to the Markov chain which stays for ever at the same state where it starts. Obviously, any probability distribution is an invariant distribution for this trivial chain. Staying put obviously preserves any target distribution, but at the same time, this is obviously useless for the purpose of exploring the target. Useful Markov chains are ergodic, and then the invariant density can be shown to be unique.

One simple way to ensuring that a Markov chain has a specified invariant density π is to construct the transition kernel K so that it is **reversible** with respect to π. This means that the condition

$$P_\pi(\Theta^{(0)} \in A, \Theta^{(1)} \in B) = P_\pi(\Theta^{(0)} \in B, \Theta^{(1)} \in A) \tag{11.6}$$

holds for every $A, B \subset S$. This means that

$$(\Theta^{(0)}, \Theta^{(1)}) \stackrel{\mathrm{d}}{=} (\Theta^{(1)}, \Theta^{(0)}), \qquad \text{when} \quad \Theta^{(0)} \sim \pi,$$

that is, the joint distribution of the pair $(\Theta^{(0)}, \Theta^{(1)})$ is the same as the joint distribution of the pair $(\Theta^{(1)}, \Theta^{(0)})$, when the chain is started from the invariant distribution. Of course, the same result then extends to all pairs $(\Theta^{(i)}, \Theta^{(i+1)})$, where $i \geq 0$.

Expressed in terms of the transition kernel, the condition (11.6) for reversibility becomes

$$\int_A \pi(\theta)\, K(\theta, B)\, \mathrm{d}\theta = \int_B \pi(\phi)\, K(\phi, A)\, \mathrm{d}\phi, \qquad \forall A, B \subset S. \tag{11.7}$$

These equations are also called the **detailed balance** equations.

Theorem 6. *If the transition kernel K is reversible for π, then π is an invariant density for the chain.*

Proof. For any $A \subset S$

$$P_\pi(\Theta^{(0)} \in A) = P_\pi(\Theta^{(0)} \in A, \Theta^{(1)} \in S) = P_\pi(\Theta^{(0)} \in S, \Theta^{(1)} \in A)$$
$$= P_\pi(\Theta^{(1)} \in A). \qquad \square$$

11.3 Finite state space

It is instructive to specialize the preceding concepts for the case of a finite state space, which may be familiar to the reader. Consider a Markov chain on the finite state space

$$S = \{1, \ldots, k\}.$$

Now we can identify the transition kernel with the transition matrix $P = (p_{ij})$ with entries

$$p_{ij} = P(\Theta^{(t+1)} = j \mid \Theta^{(t)} = i), \qquad i = 1, \ldots, k, \quad j = 1, \ldots, k.$$

It is customary to let the first index denote the present state, and the second index the possible values of the next state.

The entries of the transition matrix have obviously the following properties,

$$p_{ij} \geq 0 \quad \forall i, j; \qquad \sum_{j=1}^{k} p_{ij} = 1, \quad \forall i.$$

All the elements are non-negative and all the rows sum to one. Such a matrix is called a stochastic matrix. The transition kernel corresponding to the transition matrix is

$$K(i, A) = \sum_{j \in \{1,\ldots,k\} \cap A} p_{ij}.$$

If the pmf of the initial distribution is expressed as the row vector $\pi^T = [\pi_1, \ldots, \pi_k]$, then the pmf at time one is

$$\sum_i \pi_i p_{ij} = [\pi^T P]_j,$$

i.e., it is the j'th entry of the row vector $\pi^T P$.

The probability row vector $\pi^T = [\pi_1, \ldots, \pi_k]$ is stationary if and only if

$$\pi^T = \pi^T P,$$

which means that π^T has to be a left eigenvector of P corresponding to eigenvalue one, and π has to be a probability vector: its entries must be non-negative and sum to one. (A left eigenvector of P is simply the transpose of an ordinary eigenvector [or right eigenvector] of P^T).

In a finite state space the transition matrix P is reversible with respect to π, if

$$\pi_i \, p_{ij} = \pi_j \, p_{ji}, \quad \forall i, j.$$

Then π is an invariant pmf, since for any j

$$\sum_i \pi_i \, p_{ij} = \sum_i \pi_j \, p_{ji} = \pi_j \sum_i p_{ji} = \pi_j.$$

11.4 Combining kernels

A simulation algorithm, where one calculates the new state θ' based on the old state θ and some freshly generated random numbers corresponds to the kernel $K(\theta, A)$, where

$$K(\theta, A) = P(\Theta' \in A \mid \theta).$$

Now suppose that we have two simulation codes, which correspond to two different kernels $K_1(\theta, A)$ and $K_2(\theta, A)$. What is the transition kernel from θ to θ'', if we first calculate θ' by the code corresponding to $K_1(\theta, \cdot)$, and then calculate θ'' using the code corresponding to $K_2(\theta', \cdot)$? Notice that in the second step the initial value is the state where we ended up after the first step. The new piece of code corresponds to a transition kernel which we will denote by

$$K_1 K_2.$$

This is can be called the cycle of K_1 and K_2. In a finite state space $K_1 K_2$ corresponds to multiplying the transition matrices P_1 and P_2 to form the transition matrix $P_1 P_2$.

If we have d kernels $K_1, \ldots, K_d$, then we can define the **cycle** of the kernels $K_1, \ldots, K_d$ by

$$K_1 K_2 \cdots K_d,$$

which corresponds to executing the simulations corresponding to the kernels sequentially, always starting from the state where the previous step took us. If K_j is the transition kernel of the jth component Gibbs updating step, then the combined kernel $K_1 \cdots K_d$ is the kernel the deterministic scan Gibbs sampler, where the updates are carried out in the order $1, 2, \ldots, d$.

Now suppose that π is an invariant density for all kernels K_j. If the initial state Θ has the density π, then after drawing Θ' from the kernel $K_1(\theta, \cdot)$, the density of Θ' is π. When we then simulate Θ'' from the kernel $K_2(\theta', \cdot)$, its density is again π, and so on. Therefore the cycle kernel

$$K_1 K_2 \cdots K_d$$

also has π as its invariant density.

Now suppose that we have d transition kernels K_j. Suppose also that $\beta_1, \ldots, \beta_d$ is a probability vector. Then the kernel

$$K(\theta, A) = \sum_{j=1}^{d} \beta_j K_j(\theta, A)$$

is called a **mixture** of the kernels $K_1, \ldots, K_d$. It corresponds to the following simulation procedure. We draw j from the pmf $\beta_1, \ldots, \beta_d$ and then draw the new value θ' using the kernel $K_j(\theta, \cdot)$. If K_j is the jth updating step of a Gibbs sampler, then K is the transition kernel of the random scan Gibbs sampler corresponding to selecting the component to be updated using the probabilities $\beta_1, \ldots, \beta_d$.

Suppose that all the kernels K_j have π as an invariant density. Then also the mixture $K = \sum \beta_j K_j$ has the same invariant density, since

$$\int_A \pi(\theta)\, \mathrm{d}\theta = \int \pi(\theta)\, K_j(\theta, A)\, \mathrm{d}\theta, \qquad \forall j \quad \forall A \subset S,$$

and hence

$$\int_A \pi(\theta)\,\mathrm{d}\theta = \sum_{j=1}^{d} \beta_j \int_A \pi(\theta)\,\mathrm{d}\theta = \sum_{j=1}^{d} \beta_j \int \pi(\theta)\,K_j(\theta, A)\,\mathrm{d}\theta = \int \pi(\theta)\,K(\theta, A)\,\mathrm{d}\theta.$$

For this argument to work, it is critical that the mixing vector $\beta_1, \ldots, \beta_d$ does not depend on the present state θ.

We have proved the following theorem.

Theorem 7. *If π is an invariant density for each of the kernels $K_1, \ldots, K_d$, then it is also an invariant density for the cycle kernel $K_1 \cdots K_d$.*

If π is an invariant density for each of the kernels $K_1, \ldots, K_d$ and $\beta_1, \ldots, \beta_d$ is a probability vector, i.e., each $\beta_i \geq 0$ and $\beta_1 + \cdots + \beta_d = 1$, then π is also an invariant density for the mixture kernel $\sum_{j=1}^{d} \beta_j K_j$.

11.5 Invariance of the Gibbs sampler

Suppose that the target density is $\pi(\theta)$, where θ is divided into components

$$\theta = (\theta_1, \theta_2, \ldots, \theta_d).$$

Now consider the transition kernel K_j corresponding to the **jth component Gibbs sampler**. This sampler updates the jth component θ_j of θ only and keeps all the other components θ_{-j} at their original values. The sampler draws a new value θ_j' for θ_j from the corresponding full conditional density, which we denote by

$$\pi_j(\theta_j \mid \theta_{-j}).$$

A key observation is the identity

$$\pi(\theta) = \pi_j(\theta_j \mid \theta_{-j})\,\pi(\theta_{-j}),$$

where $\pi(\theta_{-j})$ is the marginal density of all the other components except θ_j.

Theorem 8. *The transition kernel corresponding to the jth component Gibbs sampler has π as its invariant density.*

Proof. Let the initial state Θ have density π, and let Θ_j' be drawn from the jth full conditional density. Then the joint distribution of Θ and Θ_j' has the density

$$\pi(\theta)\,\pi_j(\theta_j' \mid \theta_{-j}) = \pi_j(\theta_j \mid \theta_{-j})\,\pi(\theta_{-j})\,\pi_j(\theta_j' \mid \theta_{-j}).$$

After the update, the state is (Θ_j', Θ_{-j}). We obtain its density by integrating out the variable θ_j from the joint density of Θ and Θ_j', but

$$\begin{aligned}\int \pi_j(\theta_j \mid \theta_{-j})\,\pi(\theta_{-j})\,\pi_j(\theta_j' \mid \theta_{-j})\,\mathrm{d}\theta_j &= \pi_j(\theta_j' \mid \theta_{-j})\,\pi(\theta_{-j}) \int \pi(\theta_j \mid \theta_{-j})\,\mathrm{d}\theta_j \\ &= \pi_j(\theta_j' \mid \theta_{-j})\,\pi(\theta_{-j}) = \pi(\theta').\end{aligned}$$

Therefore the updated state has the density π. □

It now follows from theorem 7 that the systematic scan and the random scan Gibbs samplers have π as their invariant distribution.

It can also be shown that the transition kernel K_j of the jth Gibbs update is reversible with respect to π. From this it follows that the transition kernel $\sum_j \beta_j K_j$ of the random scan Gibbs sampler is also reversible with respect to π. However, the transition kernel of the systematic scan Gibbs sampler is not usually reversible. (The distinction between reversible and non-reversible kernels makes a difference when one discusses the regularity conditions needed for the Markov chain central limit theorems.)

11.6 Reversibility of the M–H algorithm

Proving that the Metropolis–Hastings update leaves the target density invariant requires more effort than proving the same property for the Gibbs sampler.

Let the initial state Θ be θ and let the next state be denoted by Φ. Recall that Φ is obtained from θ by the following steps.

- We generate the proposal Θ' from the proposal density $q(\theta' \mid \theta)$, and independently $U \sim \text{Uni}(0, 1)$.
- We set
$$\Phi = \begin{cases} \Theta', & \text{if } U < r(\theta, \Theta') \text{ (accept)} \\ \theta, & \text{otherwise (reject)}, \end{cases}$$
where the M–H ratio $r(\theta, \theta')$ is defined by
$$r(\theta, \theta') = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} \tag{11.8}$$

Notice that $r(\theta, \theta')$ can be greater than one, and hence the probability of acceptance, conditionally on $\Theta = \theta$ and $\Theta' = \theta'$ is given by

$$\alpha(\theta, \theta') = P(\text{accept} \mid \Theta = \theta, \Theta' = \theta') = \min(1, r(\theta, \theta')).$$

Theorem 9. *The Metropolis–Hastings sampler is reversible with respect to π, and hence has π as its invariant density.*

Proof. To prove reversibility, we must prove that

$$P_\pi(\Theta \in A, \Phi \in B) = P_\pi(\Theta \in B, \Phi \in A) \tag{11.9}$$

for all sets A and B in the state space. Here the subscript π means that the current state Θ is distributed according to the density π.

Now the left-hand side (LHS) of the claim (11.9) is

$$\begin{aligned} P_\pi(\Theta \in A, \Phi \in B) &= P_\pi(\Theta \in A, \Phi \in B, \text{accept}) + P_\pi(\Theta \in A, \Phi \in B, \text{reject}) \\ &= P_\pi(\Theta \in A, \Theta' \in B, \text{accept}) + P_\pi(\Theta \in A \cap B, \text{reject}) \end{aligned}$$

Similarly, the right-hand side (RHS) of the claim (11.9) is

$$P_\pi(\Theta \in B, \Phi \in A) = P_\pi(\Theta \in B, \Theta' \in A, \text{accept}) + P_\pi(\Theta \in B \cap A, \text{reject})$$

The contributions from rejection are equal on the LHS and on the RHS, and we need only show that the contributions from acceptance are also equal.

On the LHS, the contribution from acceptance is

$$P_\pi(\Theta \in A,\Theta' \in B, \text{accept}) = \int \mathrm{d}\theta\, 1_A(\theta)\, \pi(\theta) \int \mathrm{d}\theta'\, 1_B(\theta')\, q(\theta' \mid \theta)\, \alpha(\theta,\theta')$$
$$= \iint_{(\theta,\theta')\in A\times B} \pi(\theta)\, q(\theta' \mid \theta)\, \alpha(\theta,\theta')\, \mathrm{d}\theta\, \mathrm{d}\theta'.$$

Similarly, on the RHS, the contribution from acceptance is

$$P_\pi(\Theta \in B,\Theta' \in A, \text{accept}) = \iint_{(\theta,\theta')\in B\times A} \pi(\theta)\, q(\theta' \mid \theta)\, \alpha(\theta,\theta')\mathrm{d}\theta\, \mathrm{d}\theta'$$
$$= \iint_{(\theta,\theta')\in A\times B} \pi(\theta')\, q(\theta \mid \theta')\, \alpha(\theta',\theta)\, \mathrm{d}\theta\, \mathrm{d}\theta',$$

where in the last formula we just interchanged the names of the integration variables. Since the two integration sets are the same, and the equality has to hold for every integration set $A \times B$, the integrands must be proved to be the same, i.e., the claim (11.9) is true if and only if

$$\pi(\theta)\, q(\theta' \mid \theta)\, \alpha(\theta,\theta') = \pi(\theta')\, q(\theta \mid \theta')\, \alpha(\theta',\theta) \qquad \forall \theta,\theta', \tag{11.10}$$

(almost everywhere). However, our choice (11.8) for $r(\theta',\theta)$ implies (11.10), since its LHS is

$$\pi(\theta)\, q(\theta' \mid \theta)\, \alpha(\theta,\theta') = \pi(\theta)\, q(\theta' \mid \theta)\, \min(1, r(\theta,\theta'))$$
$$= \min\left(\pi(\theta)\, q(\theta' \mid \theta), \pi(\theta)\, q(\theta' \mid \theta)\, \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)}\right)$$
$$= \min(\pi(\theta)\, q(\theta' \mid \theta), \pi(\theta')\, q(\theta \mid \theta')),$$

and its RHS is

$$\pi(\theta')\ q(\theta \mid \theta')\, \alpha(\theta',\theta) = \pi(\theta')\, q(\theta \mid \theta')\, \min(1, r(\theta',\theta))$$
$$= \min\left(\pi(\theta')\, q(\theta \mid \theta'), \pi(\theta')\, q(\theta \mid \theta')\frac{\pi(\theta)\, q(\theta' \mid \theta)}{\pi(\theta')\, q(\theta \mid \theta')}\right).$$

and therefore the two integrands are the same. □

Recall from the proof, that it is sufficient to show the reversibility of the acceptance part of the transition kernel by establishing (11.10), where $\alpha(\theta,\theta') = \min(1, r(\theta,\theta'))$. The formula (11.8) is not the only choice for r which works. E.g., Barker's formula

$$r(\theta,\theta') = \frac{\pi(\theta')q(\theta \mid \theta')}{\pi(\theta')q(\theta \mid \theta') + \pi(\theta)q(\theta' \mid \theta)}$$

(which was proposed by Barker in 1965) would also imply eq. (11.10). Indeed, Hastings considered Barker's formula and many other related formulas for $\alpha(\theta,\theta')$, which all guarantee (11.10). Later, Hasting's student Peskun showed that the acceptance probability $\alpha(\theta,\theta')$ implied by (11.8) is, in a certain sense,

the best possible [8]. Later, Tierney [12] extended Peskun's optimality argument from the discrete state space to the general state space.

If we use a Metropolis–Hastings update to update the jth component of θ only, then the corresponding kernel is reversible with respect to π and hence has π as its invariant density. This follows from our proof, when we treat the other components θ_{-j} as constants. We can then combine the jth component Metropolis–Hastings updates using a systematic scan or a random-scan strategy, and the resulting algorithm still has π as its invariant density. The random scan algorithm is still reversible with respect to π, but the systematic scan algorithm is usually not reversible.

11.7 State-dependent mixing of proposal distributions

As in Sec. 7.4.6 we calculate the proposal θ' as follows, when the current state is θ. We draw the proposal from a proposal density, which is selected randomly from a list of alternatives, and the selection probabilities are allowed depend on the current state.

- Draw j from the pmf $\beta(\cdot \mid \theta), j = 1, \ldots, K$.
- Draw θ' from the density $q(\theta' \mid \theta, j)$ which corresponds to the selected j.
- Accept the proposed value as the new state, if $U < r$, where $U \sim \mathrm{Uni}(0,1)$, and
$$r = \frac{\pi(\theta')\,\beta(j \mid \theta')\,q(\theta \mid \theta', j)}{\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)}. \tag{11.11}$$
Otherwise the chain stays at θ.

We now outline the proof why this yields a Markov chain which is reversible with respect to the target density $\pi(\theta)$.

As in ordinary Metropolis–Hastings, we only need to show reversibility when that the proposed value is accepted. That is, we need to show that

$$P_\pi(\Theta \in A, \Theta' \in B, \text{accept}) = P_\pi(\Theta \in B, \Theta' \in A, \text{accept}), \tag{11.12}$$

where the subscript indicates that the density of the current state Θ is assumed to be π.

Let

$$\begin{aligned}\alpha_j(\theta, \theta') &= P(\text{accept} \mid \Theta = \theta, \Theta' = \theta', \text{component } j \text{ was selected}) \\ &= \min\left(1, \frac{\pi(\theta')\,\beta(j \mid \theta')\,q(\theta \mid \theta', j)}{\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)}\right).\end{aligned}$$

The LHS of the condition (11.12) is

$$\begin{aligned}&\int \mathrm{d}\theta\, 1_A(\theta)\,\pi(\theta) \sum_{j=1}^{K} \beta(j \mid \theta) \int \mathrm{d}\theta'\, q(\theta' \mid \theta, j)\,\alpha_j(\theta, \theta')\,1_B(\theta') \\ &\qquad = \sum_j \iint 1_A(\theta)\,1_B(\theta')\,\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)\,\alpha_j(\theta, \theta')\,\mathrm{d}\theta\,\mathrm{d}\theta'\end{aligned}$$

Similarly, the RHS of the condition (11.12) is

$$\sum_j \iint 1_B(\theta)\, 1_A(\theta')\, \pi(\theta)\, \beta(j \mid \theta)\, q(\theta' \mid \theta, j)\, \alpha_j(\theta, \theta')\, \mathrm{d}\theta\, \mathrm{d}\theta'$$
$$= \sum_j \iint 1_A(\theta)\, 1_B(\theta')\, \pi(\theta')\, \beta(j \mid \theta')\, q(\theta \mid \theta', j)\, \alpha_j(\theta', \theta)\, \mathrm{d}\theta\, \mathrm{d}\theta'$$

The equality of LHS and RHS follows from the fact that the integration sets and the integrands are the same for each j, thanks to the formula (11.11) for the test ratio r.

11.8 Reversibility of RJMCMC

Recall that the reversible jump MCMC method (RJMCMC) allows transitions between parameter spaces of different dimensions. Green derived the RJMCMC algorithm starting from the requirement that the Markov chain should be reversible [3].

We consider reversibility proof for a simple case of the RJMCMC algorithm, where we have two alternative Bayesian models for the same data y. The setting is the same as in Sec. 10.7. The first model is indicated by $M = 1$ and the second model by $M = 2$. The two models have separate parameter vectors θ_1 and θ_2 which we assume to have different dimensionalities d_1 and d_2. Their values are in respective parameter spaces S_1 and S_2. The prior distributions within the two models are

$$p(\theta_1 \mid M = 1), \qquad p(\theta_2 \mid M = 2),$$

and the likelihoods are

$$p(y \mid M = 1, \theta_1), \qquad p(y \mid M = 2, \theta_2).$$

The RJMCMC algorithm constructs a Markov chain, whose state space is the sum space

$$S = (\{1\} \times S_1) \cup (\{2\} \times S_2).$$

Any point in S is of the form (m, θ_m), where m is either 1 or 2, and $\theta_m \in S_m$. The target distribution $\pi(m, \theta_m)$ of the chain is the posterior distribution

$$\pi(m, \theta_m) = p(M = m, \theta_m \mid y), \qquad m = 1, 2, \quad \theta_m \in S_m. \tag{11.13}$$

We suppose that the parameters θ_1 and θ_2 both have continuous distributions and that $d_1 < d_2$.

When the current state of the chain is (m, θ_m), then the algorithm chooses with probability $\beta(m \mid m)$ to attempt to move within the model m or with complementary probability $\beta(k \mid m)$ to attempt to move from the current model m to the other model $k \neq m$.

Recall that in RJMCMC, the moves $1 \to 2$ and $2 \to 1$ must be related in a certain way. Suppose that the move $1 \to 2$ is effected by the following steps, when the current state is $(1, \theta_1)$.

- Draw u_1 from density $g(\cdot \mid \theta_1)$.

- Calculate $\theta_2 = T_{1\to 2}(\theta_1, u_1)$.

We suppose that the function $T_{1\to 2}$ defines a diffeomorphic correspondence between θ_2 and (θ_1, u_1). The density of the noise $g(u_1 \mid \theta_1)$ is a density on the space of dimension $d_2 - d_1$. The test ratio is calculated as

$$r = \frac{\pi(2,\theta_2)}{\pi(1,\theta_1)} \frac{\beta(1 \mid 2)}{\beta(2 \mid 1)} \frac{1}{g(u_1 \mid \theta_1)} \left| \frac{\partial \theta_2}{\partial(\theta_1, u_1)} \right|, \qquad \text{(move } 1 \to 2\text{)}. \tag{11.14}$$

Our choice for the move $1 \to 2$ implies that the move $2 \to 1$ has to deterministic and has to be calculated by applying the inverse transformation $T_{1\to 2}^{-1} = T_{2\to 1}$ to θ_2, when the current state is $(2, \theta_2)$, i.e.,

$$(\theta_1, u_1) = T_{2\to 1}(\theta_2).$$

The value u_1 is also calculated from this requirement, and it is used when we evaluate the test ratio, which is given by

$$r = \frac{\pi(1,\theta_1)}{\pi(2,\theta_2)} \frac{\beta(2 \mid 1)}{\beta(1 \mid 2)} \frac{g(u_1 \mid \theta_1)}{1} \left| \frac{\partial(\theta_1, u_1)}{\partial \theta_2} \right|, \qquad \text{(move } 2 \to 1\text{)}. \tag{11.15}$$

The moves within the models are ordinary Metropolis–Hastings moves from some suitable proposal distributions and for them the test ratio is the ordinary M–H ratio.

To show that RJMCMC is reversible with respect to the target distribution, we should prove that

$$\begin{aligned} P_\pi(M^{(0)} = m, &\Theta^{(0)} \in A, M^{(1)} = k, \Theta^{(1)} \in B) \\ &= P_\pi(M^{(0)} = k, \Theta^{(0)} \in B, M^{(1)} = m, \Theta^{(2)} \in A) \end{aligned} \tag{11.16}$$

for all $m, k \in \{1, 2\}$ and all sets $A \in C_m$ and $B \in C_k$. Here $(M^{(i)}, \Theta^{(i)})$ is the state of the chain at iteration i, and the initial distribution is the target distribution π.

We consider the case $m = 1$ and $k = 2$, and leave the other cases for the reader to check. Let $A \in C_1$ and $B \in C_2$ be arbitrary sets. If the event on the LHS of (11.16) has taken place, then the move $1 \to 2$ has been selected and θ_2 has been proposed and accepted. Therefore the LHS is

$$\int \mathrm{d}\theta_1 \, 1_A(\theta_1) \, \pi(1,\theta_1) \, \beta(2 \mid 1) \int \mathrm{d}u_1 \, g(u_1 \mid \theta_1) \, \min(1, r_{1\to 2}(\theta_1, u_1, \theta_2)) \, 1_B(\theta_2),$$

where $r_{1\to 2}(\theta_1, u_1, \theta_2)$ is the expression (11.14), and θ_2 is short for $T(\theta_1, u_1)$. On the other hand, the RHS is given by

$$\int \mathrm{d}\theta_2 \, 1_B(\theta_2) \, \pi(2,\theta_2) \, \beta(1 \mid 2) \, \min(1, r_{2\to 1}(\theta_2, \theta_1, u_1)) \, 1_A(\theta_1)$$

where $r_{2\to 1}(\theta_2, \theta_1, u_1)$ is the expression (11.15), and the pair (θ_1, u_1) is short for $T_{2\to 1}(\theta_2) = T_{1\to 2}^{-1}(\theta_2)$. Make the change of variables from θ_2 to $(\theta_1, u_1) = T_{1\to 2}^{-1}(\theta_2)$. This changes the RHS to

$$\int \mathrm{d}\theta_1 \int \mathrm{d}u_1 \, 1_A(\theta_1) \, 1_B(\theta_2) \, \pi(2,\theta_2) \, \beta(1 \mid 2) \, \min(1, r_{2\to 1}(\theta_2, \theta_1, u_1)) \left| \frac{\partial \theta_2}{\partial(\theta_1, u_1)} \right|$$

where now θ_2 is short for $T(\theta_1, u_1)$. Taking into account the formulas for the test ratios and remembering that

$$\frac{\partial(\theta_1, u_1)}{\partial\theta_2} \frac{\partial\theta_2}{\partial(\theta_1, u_1)} = 1$$

(since the mappings are inverses of one another) it is routine matter to check that the integrands are the same, and therefore reversibility has been checked for the case $(m, k) = (1, 2)$.

11.9 Irreducibility

A Markov chain which has the target distribution as its invariant distribution may still be useless. For example, consider the trivial Markov chain which stays for ever at the same state where it starts. For this chain, any probability distribution on the state space is an invariant distribution. At the same time, this kernel is clearly useless for the purpose of generating samples from the target distribution. In order to be useful, a Markov chain should visit all parts of the state space. Irreducible chains have that desirable property. A Markov chain which is not irreducible is called reducible.

If the Markov chain has π as it invariant density, then it is called **irreducible**, if for any $\theta^{(0)} \in S$ and for any A such that $\int_A \pi(\theta)\,\mathrm{d}\theta > 0$ there exists an integer m such that

$$P(\Theta^{(m)} \in A \mid \Theta^{(0)} = \theta^{(0)}) > 0.$$

In other words, starting from any initial value, an irreducible chain can eventually reach any subset of the state space (which is relevant for π) with positive probability.

The Metropolis–Hastings sampler (which treats θ as a single block) is irreducible, e.g., if the proposal density is everywhere positive, i.e., if

$$q(\theta' \mid \theta) > 0 \qquad \forall \theta, \theta' \in S.$$

Then every set A which has positive probability under π can be reached with positive probability in one step starting from any θ. However, the positivity of the proposal density is not necessary for the irreducibility of the Metropolis–Hastings chain. It is sufficient that the proposal density allows the chain to visit any region of the space after a finite number of steps.

The jth component Gibbs sampler is, of course, reducible, since it can not change any other components than θ_j. By combining the component updates with a systematic or a random scan strategy, one usually obtains an irreducible chain. The same considerations apply to the Metropolis–Hastings sampler which uses componentwise transitions. However, irreducibility of the Gibbs sampler is not automatic, as the following example shows.

Example 11.1. Let $0 < p < 1$ and consider the density

$$\pi(\theta_1, \theta_2) = p\, 1_{[0,1]\times[0,1]}(\theta_1, \theta_2) + (1 - p)\, 1_{[2,3]\times[2,3]}(\theta_1, \theta_2).$$

The full conditional of θ_1 is the uniform distribution on $[0, 1]$, if $0 < \theta_2 < 1$ and the uniform distribution on $[2, 3]$, if $2 < \theta_2 < 3$. The full conditional of

θ_2 is similar. If we start the simulation using an initial value inside the square $[0,1] \times [0,1]$, then all the subsequent values of the Gibbs sampler will be inside the same square, and the square $[2,3] \times [2,3]$ will never be visited. On the other hand, if we start the simulation using an initial value inside the other square $[2,3] \times [2,3]$, then all the subsequent values of the Gibbs sampler will be inside the same square, and the square $[0,1] \times [0,1]$ will never be visited.

For this target distribution the Gibbs sample is reducible. This example has also the interesting feature that the two full conditional distributions do not determine the joint distribution, since all the joint distributions corresponding to the different $0 < p < 1$ have the same full conditional distributions. $\triangle$

The behavior of the previous example is be ruled out, if the target distribution satisfies what is known as the **positivity condition**. It requires that $\pi(\theta)$ is strictly positive for every θ for which each of the marginal densities of the target distribution $\pi(\theta_j)$ is positive. Thus the support of π has to be the Cartesian product of the supports of the marginal densities. The previous example clearly does not satisfy the positivity condition, since the Cartesian product of the supports of the marginal densities is

$$([0,1] \cup [2,3]) \times ([0,1] \cup [2,3]),$$

but $\pi(\theta) = 0$ for any $\theta \in [0,1] \times [2,3]$ or any $\theta \in [2,3] \times [0,1]$.

The positivity condition ensures irreducibility of the Gibbs sampler, since it allows transitions between any two values in a single cycle. The famous Hammersley–Clifford theorem shows that if the positivity condition is satisfied, then the full conditional distributions determine the joint distribution uniquely.

11.10 Ergodicity

A Markov chain which has an invariant density π is ergodic, if it is irreducible, aperiodic and Harris recurrent. Then the invariant density is unique. Of these conditions, π-irreducibility has already been discussed.

A Markov chain with a stationary density π is **periodic** if there exist $d \geq 2$ disjoint subsets $A_1, \ldots, A_d \subset S$ such that

$$\int_{A_1} \pi(\theta)\, \mathrm{d}\theta > 0,$$

and starting from A_1 the chain always cycles through the sets $A_1, A_2, \ldots A_d$. I.e., the chain with transition kernel K is periodic with period d, if for the sets A_i

$$K(\theta, A_{i+1}) = 1, \quad \forall \theta \in A_i, \qquad i = 1, \ldots, d-1$$

and

$$K(\theta, A_1) = 1, \quad \forall \theta \in A_d.$$

If the chain is not periodic then it is **aperiodic**. Aperiodicity holds virtually for any Metropolis–Hastings sampler or Gibbs sampler.

The chain is **Harris recurrent**, if for all A with $\int_A \pi(\theta)\, \mathrm{d}\theta > 0$, the chain will visit A infinitely often with probability one, when the chain starts from any initial state $\theta \in S$. For MCMC algorithms, π-irreducibility usually implies

Harris recurrence, so this property is usually satisfied, although generally π-irreducibility is a much weaker condition than Harris recurrence.

If the chain is ergodic in the above sense, then starting from any initial value $\Theta^{(0)} = \theta$, the distribution of $\Theta^{(n)}$ converges (in the sense of total variation distance) to the (unique) invariant distribution as n grows without limit.

Under ergodicity, the **strong law of large numbers** holds. Namely, for any real-valued function h, which is absolutely integrable in the sense that

$$\int |h(\theta)|\, \pi(\theta)\, \mathrm{d}\theta < \infty,$$

the empirical means of the RVs $h(\Theta^{(t)})$,

$$\hat{\pi}_n(h) = \frac{1}{n} \sum_{t=1}^{n} h(\Theta^{(t)}), \tag{11.17}$$

converge to the corresponding expectation

$$\pi(h) = \int h(\theta)\, \pi(\theta)\, \mathrm{d}\theta \tag{11.18}$$

with probability one, i.e.,

$$\lim_{n \to \infty} \hat{\pi}_n(h) = \pi(h), \tag{11.19}$$

and this holds for any initial distribution for $\Theta^{(0)}$.

11.11 Central limit theorem for Markov chains

We continue to use the notation (11.17) and (11.18). While the central limit theorem (CLT) does not hold for all Markov chains, it does hold for many chains generated by MCMC algorithms. Under regularity conditions on the Markov chain $\Theta^{(i)}$ and integrability conditions for the function h, the CLT then holds for the RVs $h(\Theta^{(i)})$ in the form

$$\sqrt{n}(\hat{\pi}_n(h) - \pi(h)) \xrightarrow{\mathrm{d}} N(0, \sigma_h^2), \qquad \text{as } n \to \infty. \tag{11.20}$$

As a function of the sample size n, the rate of convergence in the Markov chain CLT is the same as in the CLT for i.i.d. random variables. The required conditions on the Markov chain are easiest to state when the chain is reversible with respect to π, and this is why theoreticians recommend that one should favor reversible MCMC algorithms over non-reversible ones. However, these conditions require more advanced notions of ergodicity such as geometric ergodicity, which we bypass. See, e.g., Robert and Casella [9] or Roberts [10] for discussions of the regularity conditions for the CLT.

However, the variance σ_h^2 of the limit distribution is more difficult to estimate than in the i.i.d. setting, since in the Markov chain CLT it is given by the infinite sum

$$\sigma_h^2 = \mathrm{var}_\pi\, h(\Theta^{(0)}) + 2 \sum_{t=1}^{\infty} \mathrm{cov}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})). \tag{11.21}$$

Here the subscript π means that the covariances are calculated assuming that $\Theta^{(0)} \sim \pi$. Contrast this with the case of i.i.d. sampling from π, where the

variance of the limit distribution would be $\mathrm{var}_\pi\, h(\Theta^{(0)})$. If the chain is extended also for negative times, then this sum can be presented in the doubly-infinite form

$$\sigma_h^2 = \sum_{t=-\infty}^{\infty} \mathrm{cov}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})),$$

since the autocovariances at lags $-t$ and t are then equal.

One interpretation of the results (11.20) and (11.21) is that we can measure the loss in efficiency due to the use of the Markov chain instead of i.i.d. sampling by defining the parameter

$$\tau_h = \frac{\sigma_h^2}{\mathrm{var}_\pi\, h(\Theta^{(0)})} = 1 + 2\sum_{t=1}^{\infty} \mathrm{corr}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})),$$

which is called the **integrated autocorrelation time** for estimating $\pi(h)$ using the Markov chain under consideration (see e.g. [10]). Here $\mathrm{corr}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)}))$ is the autocorrelation at lag t for the sequence $\big(h(\Theta^{(t)})\big)$, when the chain is started from the invariant distribution π. We can also define the **effective sample size** (for estimating $\pi(h)$ using the Markov chain under consideration) as

$$n_{\mathrm{eff}}(h) = \frac{n}{\tau_h}$$

This is the sample size of an equivalent i.i.d. sample for estimating $\pi(h)$, when the Markov chain is run for n iterations.

Estimating the asymptotic variance can also be viewed as the problem of estimating the spectral density at frequency zero either for the autocovariance sequence or for the autocorrelation sequence. To simplify the notation, fix the function h and denote the autocovariance sequence of $(h(\Theta^{(t)})$ for the stationary chain by (R_t) and the autocorrelation sequence by (ρ_t),

$$R_t = \mathrm{cov}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})), \quad \rho_t = \mathrm{corr}_\pi(h(\Theta^{(0)}), h(\Theta^{(t)})), \qquad t = 0, 1, 2, \ldots$$

Further, let us extend these sequences to negative lags by agreeing that

$$R_{-t} = R_t, \quad \rho_{-t} = \rho_t, \qquad t = 1, 2, \ldots$$

Then the spectral density of the sequence (R_t) at angular frequency w is defined by the Fourier transform

$$g_R(w) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \mathrm{e}^{-itw} R_t, \qquad -\pi < w < \pi,$$

where $i = \sqrt{-1}$. (Warning: there are several related but slightly different definitions of the spectral density in the literature.) The spectral density $g_\rho(w)$ of the sequence (ρ_t) is defined similarly. Using these definitions,

$$\sigma_h^2 = 2\pi\, g_R(0), \qquad \tau_h = 2\pi\, g_\rho(0)$$

There are specialized methods available for the spectral density estimation problem, and these can be applied to estimating the asymptotic variance σ_h^2 or the integrated autocorrelation time τ_h.

All the usual methods for estimating Monte Carlo standard errors in MCMC are ultimately based on the CLT for Markov chains. The methods differ in how one estimates σ_h^2. Some of the methods are based on estimates for the integrated autocorrelation time or of the spectral density at zero. In the batch means method we have already implicitly formed an estimate for σ_h^2. See [2] for further discussion.

11.12 Literature

See the articles [5] or [11] and the book [1, Ch. 14] for surveys of the Markov chain theory needed in MCMC. See the books by Nummelin [6] or by Meyn and Tweedie [4] for comprehensive presentations of the general state space theory. See also the discussions in the books by Robert and Casella [9] and O'Hagan and Forster [7].

Bibliography

[1] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, 2005.

[2] J. M. Flegal, M. Haran, and G. L. Jones. Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science*, pages 250–260, 2008.

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

[4] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009. First ed. published by Springer in 1993.

[5] E. Nummelin. MC's for MCMC'ists. *Intenational Statistical Review*, 70(2):215–240, 2002.

[6] Esa Nummelin. *General Irreducible Markov Chains and Nonnegative Operators*. Cambridge University Press, first paperback edition, 2004. First published 1984.

[7] Anthony O'Hagan and Jonathan Forster. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Arnold, second edition, 2004.

[8] P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[9] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, second edition, 2004.

[10] Gareth O. Roberts. Linking theory and practice of MCMC. In Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, editors, *Highly Structured Stochastic Systems*. Oxford University Press, 2003.

[11] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[12] Luke Tierney. A note on Metropolis–Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8:1–9, 1998.