Monte Carlo Methods, with an emphasis on Bayesian computation
Summer 2010

Petri Koistinen
Department of Mathematics and Statistics
University of Helsinki

Contents

Chapter 1

Introduction

This course gives an overview of computational methods which are useful in Bayesian statistics. Some of the methods (such as stochastic simulation or EM algorithm) are useful also for statisticians who follow the frequentist approach to inference.

1.1 Bayesian statistics: the basic components

Suppose we are going to observe **data** y in the form of a vector $y = (y_1, \ldots, y_n)$. Before the observation takes place, the values $y_1, \ldots, y_n$ are uncertain (due to measurement errors, the natural variation of the population or due to some other reason). To allow for this uncertainty, we consider y to be the observed value of a random vector $Y = (Y_1, \ldots, Y_n)$.

We consider a **parametric model** for the distribution of Y: the distribution of Y is governed by a parameter Θ which is unknown. Usually there are several (scalar) parameters, and then Θ is actually a vector. If $\Theta = \theta$, then the vector Y has the distribution with density

$$y \mapsto f_{Y|\Theta}(y \mid \theta). \tag{1.1}$$

This is called the **sampling distribution** (or data distribution). Having observed the data $Y = y$, the function

$$\theta \mapsto f_{Y|\Theta}(y \mid \theta)$$

(considered as a function of θ and with y equal to the observed value) is called the **likelihood function** (but often multiplicative constants are omitted from the likelihood).

In Bayesian statistics both observables and parameters are considered random. Bayesian inference requires that one sets up a a joint distribution for the data and the parameters (and perhaps other unknown quantities such as future observations). If the data and the parameter are jointly continuously distributed, then the density of the joint distribution can be written in the form

$$(y, \theta) \mapsto f_{Y,\Theta}(y, \theta) = f_{Y|\Theta}(y \mid \theta)\, f_{\Theta}(\theta),$$

where f_Θ is the density of the marginal distribution of Θ, which is called the **prior distribution**. The prior distribution reflects the statistician's uncertainty about plausible values of the parameter Θ before any data has been observed.

Having observed the data $Y = y$, the statistician constructs the conditional distribution of Θ given $Y = y$, which is called the **posterior distribution**. The posterior distribution summarizes the statistician's knowledge of the parameter after the data has been observed. The main goal of Bayesian inference is to gain an understanding of the posterior distribution.

Using **Bayes' rule** (Bayes' theorem) of elementary probability theory, the posterior distribution has the density

$$\theta \mapsto f_{\Theta|Y}(\theta \mid y) = \frac{f_{Y,\Theta}(y,\theta)}{f_Y(y)} = \frac{f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta)}{\int f_{Y|\Theta}(y \mid t)\, f_\Theta(t)\, \mathrm{d}t}. \tag{1.2}$$

Here f_Y, the density of the marginal distribution of Y, has been expressed by integrating the variable θ out from the density $f_{Y,\Theta}(y,\theta)$ of the joint distribution.

Notice that the posterior density is obtained, up to a constant of proportionality depending on the data, by multiplying the prior density by the likelihood,

$$f_{\Theta|Y}(\theta \mid y) \propto f_\Theta(\theta)\, f_{Y|\Theta}(y \mid \theta).$$

Once the full probability model has been set up, the formula of the posterior density is therefore available immediately, except for the nuisance that the normalizing constant $1/f_Y(y)$ is sometimes very hard to determine.

1.2 Remarks on notation

In Bayesian statistics one rarely uses as exact notation as we have been using up to now.

- It is customary to blur the distinction between a random variable and its observed (or possible) value by using the same symbol in both cases. This is especially handy, when the quantity is represented by such a lower-case Greek character which does not posess a useful upper-case version.

- It is customary to use the terms "distribution" and "density" interchangeably, and to use the same notation for density functions of continuous distributions and probability mass functions of discrete distributions.

- When the statistical model is complex, it very soon becomes cumbersome to differentiate all the different densities in question by subscripts. An alternative notation is to introduce a different symbol for each of the distributions of interest, e.g., in the style

  $$h(y,\theta) = g(\theta)\, f(y \mid \theta) = m(y)\, p(\theta \mid y),$$

  where h is what we previously denoted by $f_{Y,\Theta}$, g is f_Θ, f is $f_{Y|\Theta}$ and so on.

- However, many authors use a different system of notation, where one **abuses notation** to make the presentation more compact. For instance, one may use $p(\cdot)$ to stand generically for different densities, so that the argument of p shows both what random quantity is under consideration and the value it may assume. Further, it is customary to let an expression such as $g(\theta)$ denote the function g. Using such notation, e.g.,

  $$p(\theta) \quad \text{means the function } f_\Theta$$

  and

  $$p(y) \quad \text{means the function } f_Y$$

  even though f_Θ and f_Y may be quite different functions. Using such compact notation, Bayes' rule can be written as

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

- In the sequel, we will often use such compact notation, since it is important to become familiar with notational conventions typically used in the field. However, we will also use more explicit (and cumbersome) notation where one uses subscripts on the densities in order to avoid misunderstandings.

1.3 Frequentist statistics versus Bayesian statistics

The reader should be aware that the Bayesian approach is not the only approach to statistics. Since the 1930's, the dominant approach to statistical inference has been what we (nowadays) call **frequentist statistics** (or **classical statistics**). It is only since the 1990's that the Bayesian approach has gradually become widely spread largely due to the arrival of new computational techniques.

In frequentist statistics the parameter is considered a deterministic, unknown quantity, whose value, say θ_0, we seek to estimate. In frequentist statistics, one does not define any probability distributions on the parameter space, so concepts like prior or posterior distribution do not make any sense in that context. The typical way of estimation is by the principle of **maximum likelihood** although other methods are used, too. The maximum likelihood estimate is that point in the parameter space which maximizes the likelihood function. In some situations, the principle of maximum likelihood needs to be supplemented with various other principles in order to avoid nonsensical results.

Frequentist statistics assess the performance of a statistical procedure by considering its performance under a large number of **hypothetical repetitions** of the observations under identical conditions. Using the notation we have already introduced, this means that a frequentist statistician is interested in what happens, on the average, when data is repeatedly drawn from the sampling distribution with density $f_{Y|\Theta}(y \mid \theta_0)$. (A true frequentist would not use such notation but would use something like $f_Y(y; \theta_0)$ instead.) In contrast, Bayesian statisticians always condition on the observed data. Bayesians are not concerned with what would happen with data we might have observed but did not. A Bayesian makes probability statements about the parameter given the observed

data, rather than probability statements about hypothetical repetitions of the data conditional on the unknown value of the parameter.

There used to be a bitter controversy among followers of the two different schools of thought. The frequentists pointed out that the inferences made by Bayesians depend on the prior distribution chosen by the statistician. Therefore Bayesian inference is not objective but is based on the personal beliefs of the statistician. On the other hand, the Bayesians liked to poke fun at the many paradoxes one gets by adhering rigidly to the principles used in frequentist statistics and accused the field of frequentist statistics to be a hodgepodge of methods derived from questionable principles.

However, nowadays many statisticians use both Bayesian and frequentist inference. If the sample size is large, then the point estimates, confidence intervals and many other inferences using either approach are usually quite similar. However, the interpretations of these results are different. A Bayesian statistician might consider results he or she obtains using frequentist methods to be approximations to results one would obtain using proper Bayesian methodology, and vice versa.

One area where the two approaches differ clearly is hypothesis testing. In frequentist statistics it is very common to conduct a test of a sharp null hypothesis (or a point null hypothesis or a simple hypothesis) such as

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

Many Bayesians have objections to the whole idea of testing a sharp null hypothesis. What is more, in this setting one arrives at quite different results using Bayesian or frequentist methods.

1.4 A simple example of Bayesian inference

To illustrate the basic notions, consider the following example. Suppose that conditionally on $\Theta = \theta$, the random variables $Y_i, i = 1, \ldots, n$ are independently exponentially distributed with rate θ, i.e., that

$$p(y_i \mid \theta) = \theta \, \mathrm{e}^{-\theta y_i}, \qquad y_i > 0.$$

Then the likelihood is

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

Suppose that our prior is the gamma distribution $\mathrm{Gam}(a, b)$ with known hyperparameters $a, b > 0$, i.e.,

$$p(\theta) = \frac{b^a}{\Gamma(a)} \, \theta^{a-1} \mathrm{e}^{-b\theta}, \qquad \theta > 0.$$

Then, as a function of $\theta > 0$,

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \, p(\theta) \\ &\propto \theta^{a-1} \mathrm{e}^{-b\theta} \, \theta^n \, \exp(-\theta \sum_{i=1}^{n} y_i) \\ &= \theta^{a+n-1} \, \exp(-(b + \sum_{i=1}^{n} y_i)\theta). \end{aligned}$$

This shows that the posterior distribution is the gamma distribution

$$\mathrm{Gam}(a+n, b+\sum_{i=1}^{n} y_i).$$

Since the gamma distribution is a well-understood distribution, we can consider the inference problem solved.

In this case the prior distribution and posterior distribution belong to the same parametric family of distributions. In such a case we speak of a conjugate family (under the likelihood under consideration). In such a case Bayesian inference amounts to finding formulas for updating the so called hyperparameters of the conjugate family.

We might also want to consider a future observable Y^* whose distribution conditionally on $\Theta = \theta$ is also exponential with rate θ but which is conditionally indpendent of the already available observations $y_1, \ldots, y_n$. Then $p(y^* \mid y)$ is called the (posterior) **predictive distribution** of the future observable. Thanks to conditional independence, the joint posterior of Θ and Y^* can be shown to factorize as follows

$$p(y^*, \theta \mid y) = p(y^* \mid \theta)\, p(\theta \mid y)$$

and therefore, by marginalizing,

$$\begin{aligned} p(y^* \mid y) &= \int p(y^*, \theta \mid y)\, \mathrm{d}\theta = \int p(y^* \mid \theta)\, p(\theta \mid y)\, \mathrm{d}\theta \\ &= \int_0^\infty \theta\, \mathrm{e}^{-\theta y^*}\, \frac{(b+\sum_1^n y_i)^{a+n}}{\Gamma(a+n)}\, \theta^{a+n-1}\, \mathrm{e}^{-(b+\sum_1^n y_i)\theta}\, \mathrm{d}\theta \end{aligned}$$

where the integral can be expressed in terms of the gamma function. Hence also the predictive distribution can be obtained explicitely.

If we are not satistifid by any gamma distribution ias a representation of our prior knowledge, and we may pick our prior from another family of distributions. In this case the situation changes dramatically in that we must resort to numerical methods in order to understand the posterior distribution.

1.5 Introduction to Bayesian computations

Conceptually, Bayesian inference is simple. One simply combines the prior and the likelihood to derive the posterior. For a single parameter, this can be implemented quite simply by graphical methods or by numerical integration. However for more complex problems, Bayesian inference was traditionally extremely hard to implement except in some simple situations where it was possible to use conjugate priors and arrive at analytical solutions. In distinction, in classical statistics the conceptual underpinnings behind statistical inference are more complicated, but the calculations are simple, at least in the case of certain standard statistical models.

A breakthrough occurred in the 1980's, when people realized two things.

- Instead of an analytic expression, one can represent the posterior distribution on a computer by drawing a sequence of samples from it.

- In most situations it is easy to draw samples from the posterior using MCMC methods (Markov chain Monte Carlo methods). Such methods were introduced in the statistical physics literature already in the 1950's. Several computer programs, most notably BUGS (WinBUGS or OpenBUGS), are now available for constructing automatically MCMC algorithms for a wide variety of statistical models.

1.6 Literature

- See, e.g., Bernardo and Smith [2] for a clear exposition of the ideas of Bayesian statistics.
- Schervish [17] treats both Bayesian and frequentist statistics using a rigorous, measure theoretic formulation.
- See, e.g., Gelman et al. [8] and O'Hagan and Forster [14] for expositions of Bayesian analysis and its computational techniques.
- See, e.g., Bolstad [3], Robert and Casella [16] and Albert [1] for introductions to Bayesian computation and MCMC.
- More advanced books discussing Bayesian computation and MCMC include those by Tanner [18]; Robert and Casella [15]; Liu [12]; Chen, Shao and Ibrahim [4]; Gamerman and Lopes [7].
- Congdon [6, 5] and Ntzoufras [13] discuss a rich collection of Bayesian models using BUGS for implementing the computations.
- To gain a wider picture of computational statistics, consult Gentle [9, 10] or Givens and Hoeting [11].

Bibliography

[1] Jim Albert. *Bayesian Computation with R*. Springer, 2007.

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

[3] William M. Bolstad. *Understanding Computational Bayesian Statistics*. Wiley, 2010.

[4] Ming-Hui Chen, Qi-Man Shao, and Joseph G. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer, 2000.

[5] Peter Congdon. *Applied Bayesian Modelling*. Wiley, 2003.

[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] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, 2nd edition, 2004.

[9] James E. Gentle. *Elements of Computational Statistics*. Springer, 2002.

[10] James E. Gentle. *Computational Statistics*. Springer, 2009.

[11] Geof H. Givens and Jennifer A. Hoeting. *Computational Statistics*. Wiley-Interscience, 2005.

[12] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2001.

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

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

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

[16] Christian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Springer, 2010.

[17] Mark J. Schervish. *Theory of Statistics*. Springer series in statistics. Springer-Verlag, 1995.

[18] Martin A. Tanner. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer Series in Statistics. Springer, 3rd edition, 1996.

Chapter 2

Review of Probability

We are going to work with random vectors. Some of their components have discrete distributions and some continuous distributions, and a random vector may have both types of components. The reader is hopefully familiar with most of the concepts used in this chapter. We use uppercase letters such as X for random variables and random vectors, and lowercase letters such as x for their possible values. When there are several random variables under consideration, we may use subscripts to differentiate between functions (such as distribution functions, densities, ...) associated with the different variables.

2.1 Random variables and random vectors

While the student needs not know measure theoretic probability theory, it useful to at least recognize some concepts. The starting point of the theory is a **probability space** (or probability triple) $(\Omega, \mathcal{A}, P)$, where

- Ω is a set called a **sample space**,
- $\mathcal{A}$ is a collection of subsets of Ω. A set $E \in \mathcal{A}$ is called an **event**.
- P is a **probability measure**, which assigns a number

$$0 \leq P(E) \leq 1, \qquad E \in \mathcal{A}$$

  for each event E.

A **random variable** X is defined to be a function

$$X : \Omega \to \mathbb{R}\,.$$

Intuitively, a random variable is a number determined by chance. A **random vector** Y is a function

$$Y : \Omega \to \mathbb{R}^d$$

for some positive integer d. I.e., random vectors are vector-valued functions whose components are random variables. A random variable is a special case of a random vector (take $d = 1$). We will use the abbreviation RV to denote either a random variable or a random vector.

For technical reasons, which we will not discuss, the set of events $\mathcal{A}$ usually does not contain all subsets of Ω. Further, all RVs need to be Borel measurable. This is a technical condition, which ensures that everything is properly defined. Further, for technical reasons, all subsets of of $\mathbb{R}$ or $\mathbb{R}^d$ used in these notes are assumed to be Borel subsets, and this requirement is not going to be mentioned anymore.

If X is a random variable, then it is of interest to know how to calculate the probability that $X \in B$ for and arbitrary set $B \subset \mathbb{R}$. The function

$$B \mapsto P(X \in B), \qquad B \subset \mathbb{R}$$

is called the **distribution** of X. Here $P(X \in B)$ means the probability of the event

$$\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}.$$

In probability theory, it is customary to suppress the argument ω whenever possible, as was done here.

The distribution of a random vector Y is defined similarly as the set function

$$B \mapsto P(X \in B), \qquad B \subset \mathbb{R}^d\,.$$

The distribution of a RV defined as a set function is an abstract concept. In applications one usually deals with more concrete representations such as distribution functions, probability mass functions or probability densities.

2.2 Distribution function

The **cumulative distribution function** (cdf) of a random variable X is defined as

$$F_X(x) = P(X \le x), \qquad x \in \mathbb{R}\,. \tag{2.1}$$

(Probabilists usually use the shorter term **distribution function** (df).) If there is only one random variable under consideration, we may omit the symbol of that variable from the subscript. The distribution function is defined for any random variable no matter what type its distribution is (discrete, continuous, or something more complicated).

If F is the distribution function of any random variable, then it has the following properties.

- F is nondecreasing and right continuous.
- F has limits $F(-\infty) = 0$ and $F(+\infty) = 1$.

The distribution function determines the distribution. If two random variables X and Y have the same distribution functions, then they have the same distributions, i.e.,

$$F_X = F_Y \quad \Leftrightarrow \quad (P(X \in B) = P(Y \in B), \quad \forall B \subset \mathbb{R})\,.$$

The distribution function of a random vector $X = (X_1, \ldots, X_d)$ is defined analogously,

$$F_X(x) = P(X \le x) = P(X_1 \le x_1, \ldots, X_d \le x_d), \qquad x = (x_1, \ldots, x_d) \in \mathbb{R}^d\,.$$

The distribution function determines the distribution also for random vectors.

2.3 Discrete distributions

A discrete RV takes values in a finite or countable set. In this case also the distribution of that quantity is called discrete. The **probability (mass) function** (pmf) of a discrete RV is defined by

$$f_X(x) = P(X = x). \tag{2.2}$$

Usually the range of a discrete random variable is a subset of the integers.

A pmf f_X has the properties

$$0 \leq f_X(x) \leq 1, \quad \forall x,$$

and

$$\sum_x f_X(x) = 1,$$

which follow at once from the properties of the probability measure. Here the sum extends over all the possible values of X.

2.4 Continuous distributions

A RV X is called continuous and is said to have a continuous distribution, if its distribution has a **probability density function** (pdf) (or simply density), i.e., if there exists a function $f_X \geq 0$ such that for any set B,

$$P(X \in B) = \int_B f_X(x)\,\mathrm{d}x. \tag{2.3}$$

If X is a random variable, then $B \subset \mathbb{R}$, but if X is d-dimensional random vector, then $B \subset \mathbb{R}^d$, and the integral is actually a multiple integral.

The integral over the set B is defined as

$$\int_B f_X(x)\,\mathrm{d}x = \int 1_B(x) f_X(x)\,\mathrm{d}x,$$

where on the right the integral is taken over the whole space, and 1_B is the **indicator** function of the set B,

$$1_B(x) = \begin{cases} 1, & \text{if } x \in B \\ 0, & \text{otherwise.} \end{cases}$$

With integrals we follow the convention that if the range of integration is not indicated, then the range of integration is the whole space under consideration.

By definition, a probability density f_X satisfies

$$f_X(x) \geq 0, \qquad \forall x,$$

but a density need not be bounded from above. Also

$$\int f_X(x)\,\mathrm{d}x = 1,$$

(where the integral extends over the whole space). This follows since the probability that X takes on *some* value is 1.

The requirement (2.3) does not determine the density uniquely but only modulo sets of measure zero. In applications one works with continuous or piecewise-continuous versions of the densities, and does not worry about this non-uniqueness. We say that two densities f and g are equal, and write $f = g$, if f and g are densities of the same distribution, i.e., if f and g are equal almost everywhere.

The density can be obtained from the distribution function by differentiation. In one dimension,

$$f_X = F'_X$$

Here the derivative on the right is defined almost everywhere, and on the right we may extend the function arbitrarily to whole $\mathbb{R}$. After this we obtain a valid density function. In d dimensions one has an analogous result,

$$f_{X_1,\ldots,X_d}(x_1,\ldots,x_d) = \frac{\partial^d F_{X_1,\ldots,X_d}(x_1,\ldots,x_d)}{\partial x_1 \cdots \partial x_d},$$

almost everywhere, in the sense that the mixed derivative is defined almost everywhere and after an arbitrary extension one obtains a density for the joint distribution of $X_1, \ldots, X_d$.

The pmfs of discrete random variables and the pdfs of continuous random variables behave in many contexts in exactly the same way. That is why we use the same notation in both cases. Sometimes we use the word 'density' to refer to the pmf of a discrete random variable or even to the analogous concept for more complicated distributions. (The key mathematical concept is the Radon-Nikodym derivative with respect to some dominating sigma-finite measure.) If it is necessary to make a distinction, we will speak of the density of a continuous distribution or the density of a continuous RV.

2.5 Quantile function

A quantile function is the inverse function of the distribution function of a random variable whenever the distribution function is invertible. Otherwise the quantile function is defined as a generalized inverse function of the distribution function. Notice that quantile functions are defined only for univariate distributions.

Let us first consider the important case, where the quantile function can be obtained by inverting the distribution function. Consider a random variable X whose df F_X is continuous and strictly increasing on an interval (a, b) such that $F_X(a) = 0$ and $F_X(b) = 1$. In other words, we assume that $X \in (a, b)$ with probability one. The values $a = -\infty$ or $b = +\infty$ are permitted, in which case $F_X(a)$ or $F_X(b)$ has to be interpreted as the corresponding limit.

In this case, the equation

$$F_X(x) = u, \qquad 0 < u < 1,$$

has a unique solution $F_X^{-1}(u) \in (a, b)$ and we call the resulting function

$$q_X(u) = F_X^{-1}(u), \qquad 0 < u < 1 \tag{2.4}$$

the quantile function of (the distribution of) X. (We are abusing notation: we are actually using the inverse function of the df F_X restricted to the interval (a, b).) If a or b is finite, we could extend the domain of definition of q_X in a natural way to cover the points 0 or 1, respectively. However, we will not do this since this would lead to difficulties when $a = -\infty$ or $b = \infty$.

Since

$$P(X \leq q_X(u)) = F_X(q_X(u)) = F_X(F_X^{-1}(u)) = u, \qquad 0 < u < 1,$$

a proportion of u of the distribution of X lies to the left of the point $q_X(u)$. Similarly,

$$P(X > q_X(1-u)) = 1 - F_X(q_X(1-u)) = u, \qquad 0 < u < 1,$$

which shows that a proportion u of the distribution of X lies to the right of the point $q_X(1-u)$. If we assume that the distribution of X is continuous, then we can characterize its quantiles with the equations

$$\int_{-\infty}^{q_X(u)} f_X(x)\,\mathrm{d}x = u, \qquad \forall 0 < u < 1 \tag{2.5}$$

$$\int_{q_X(1-u)}^{\infty} f_X(x)\,\mathrm{d}x = u, \qquad \forall 0 < u < 1, \tag{2.6}$$

i.e., for any $0 < u < 1$, the area under the pdf in the left-hand tail $(-\infty, q_X(u))$ is u, and the area under the pdf in right-hand tail $(q_X(1-u), \infty)$ is also u.

Example 2.1. The unit exponential distribution Exp(1) has the density

$$f_X(x) = \mathrm{e}^{-x}\, 1_{[0,\infty)}(x)$$

and distribution function

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \begin{cases} 1 - \mathrm{e}^{-x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Hence the quantile function of this distribution is

$$q_X(u) = F_X^{-1}(u) = -\ln(1-u), \qquad 0 < u < 1.$$

△

The quantile function has important uses in simulation. Let $U \sim \mathrm{Uni}(0, 1)$, which means that U has the uniform distribution on $(0, 1)$. Recall that most programming environments have a random number generator for the $\mathrm{Uni}(0, 1)$ distribution. Let q_X be the quantile function of a random variable X. Then

$$q_X(U) \stackrel{\mathrm{d}}{=} X, \tag{2.7}$$

which means that $q_X(U)$ has the same distribution as X. We will check this claim shortly. Equation (2.7) shows how a uniformly distributed random variable U can be transformed to have a given distribution. We will refer to this method by the name **inverse transform** or **inversion**. This method has several other names in the literature: the **probability integral transform** the **inverse transformation method**, the **quantile transformation method** and others. The inverse transform is an excellent simulation method for certain distributions, whose quantile functions are easy to calculate.

Example 2.2. By the previous example, we can simulate a random draw from Exp(1) by generating $U \sim \text{Uni}(0,1)$ and then calculating

$$-\ln(1-U).$$

This procedure can be simplified a bit by noticing that when $U \sim \text{Uni}(0,1)$, then also $1-U \sim \text{Uni}(0,1)$ distribution. Therefore we may as well simulate Exp(1) by calculating

$$-\ln(U).$$

△

We now check the claim (2.7) in the case introduced before, where F_X is continuous and strictly increasing on (a,b) and $F_X(a) = 0$ and $F_X(b) = 1$.

Recall that the inverse function of a strictly increasing function is strictly increasing. Therefore

$$\{(u,x) \in (0,1) \times (a,b) : q_X(u) \le x)\} = \{(u,x) \in (0,1) \times (a,b) : u \le F_X(x)\}.$$

(Apply F_X to both sides of the first inequality, or $q_X = F_X^{-1}$ to the second.) Hence, for any $a < x < b$,

$$P(q_X(U) \le x) = P(U \le F_X(x)) = F_X(x).$$

This proves eq. (2.7).

A more general df F does not admit an inverse function defined on $(0,1)$. However, one can define a generalized inverse function by using the formula

$$F^{-1}(u) = \inf\{x : F(x) \ge u\}, \qquad 0 < u < 1. \tag{2.8}$$

Here $\inf B$ is the greatest lower bound of the set $B \subset \mathbb{R}$. Since a df is increasing and right continuous, the set $\{x : F(x) \ge u\}$ is of the form $[t, \infty)$ for some $t \in \mathbb{R}$, and then its infimum is t.

The inverse transform principle (2.7) holds for all univariate distributions, when we define the quantile function to be the generalized inverse of the distribution function.

2.6 Joint, marginal and conditional distributions

If we are considering two RVs X and Y, then we may form a vector V by concatenating the components of X and Y,

$$V = (X, Y).$$

Then the **joint distribution** of X and Y is simply the distribution of V. If the distribution of V is discrete or continuous, then we use the following notation for the pmf or density of the joint distribution

$$f_{X,Y}(x,y),$$

which means the same thing as $f_V(v)$, when $v = (x,y)$. The distribution of X or Y alone is often called its **marginal distribution**.

Recall the elementary definition of conditional probability. Suppose that A and B are events and that $P(A) > 0$. Then the conditional probability $P(B \mid A)$ of B given A (the probability that B occurs given that A occurs) is defined by

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}. \tag{2.9}$$

If the joint distribution of RVs X and Y is discrete, then the conditional distribution of Y given $X = x$ is defined by using (2.9). Given $X = x$, Y has the pmf

$$f_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f_{X,Y}(x, y)}{f_X(x)}. \tag{2.10}$$

Here f_X, the pmf of the marginal distribution of X is obtained by summing y out from the joint pmf,

$$f_X(x) = \sum_y f_{X,Y}(x, y),$$

Naturally, definition (2.10) makes sense only for those x for which $f_X(x) > 0$. If need be, we may extend the domain of definition of the conditional pmf $f_{Y|X}(y \mid x)$ by agreeing that

$$f_{Y|X}(y \mid x) = 0, \quad \text{if } f_X(x) = 0.$$

It is useful to have in mind some such extension in order to make sense of certain formulas. However, the exact manner in which we do this extensions does not really matter.

By rearranging the definition of the conditional pmf we see that for all x and y

$$f_{X,Y}(x, y) = f_X(x)\, f_{Y|X}(y \mid x).$$

By reversing the roles of X and Y, wee see that also the following holds,

$$f_{X,Y}(x, y) = f_Y(y)\, f_{X|Y}(x \mid y).$$

Hence, the pmf of the joint distribution can be obtained by multiplying the marginal pmf with the pmf of the conditional distribution. This result is called the **multiplication rule** or the **chain rule** (or the product rule).

When RVs X and Y have a continuous joint distribution, we define the conditional density $f_{Y|X}$ of Y given X as

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)}, \quad \text{when } f_X(x) > 0. \tag{2.11}$$

Here f_X is the density of the marginal distribution of X, which can be calculated by integrating y out from the joint distribution,

$$f_X(x) = \int f_{X,Y}(x, y)\, \mathrm{d}y,$$

Again, if need be, we may extend the definition by agreeing that $f_{Y|X}(y \mid x) = 0$ whenever $f_X(x) = 0$.

The multiplication rule holds also for jointly continuously distributed RVs. Considered as a function of x and y

$$f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y \mid x) = f_Y(y)\, f_{X|Y}(x \mid y).$$

(Equality is here interpreted as equality of density functions, i.e., it holds almost everywhere.)

If we have a discrete RV X and a continuous RV Y, then their joint distribution can be manipulated by making use of a function $f_{X,Y}(x,y)$ which yields probabilities when its summed over x and integrated over y, i.e.,

$$P(X \in A, Y \in B) = \sum_{x \in A} \int_B f_{X,Y}(x,y)\, \mathrm{d}y$$

for arbitrary sets A and B. For convenience, we call such a representation a density (of the joint distribution). We obtain the pmf of X by integrating y out from the joint density,

$$f_X(x) = \int f_{X,Y}(x,y)\, \mathrm{d}y,$$

and the density of Y by summing x out from the joint density,

$$f_Y(y) = \sum_x f_{X,Y}(x,y).$$

The multiplication rule holds,

$$f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y \mid x) = f_Y(y)\, f_{X|Y}(x \mid y).$$

Often a joint distribution like this is specified by giving the marginal distribution of one variable and the conditional distribution of the other variable.

Often we consider the joint distribution of more than two variables. E.g., consider three RVs X, Y and Z which have (say) continuous joint distribution. By conditioning on (X,Y) and by using the multiplication rule twice, we see that

$$f_{X,Y,Z}(x,y,z) = f_{X,Y}(x,y)\, f_{Z|X,Y}(z \mid x,y) = f_X(x)\, f_{Y|X}(y \mid x)\, f_{Z|X,Y}(z \mid x,y).$$

Of course, other factorizations are possible, too. We obtain the density of the marginal distribution of any set of variables, by integrating out the other variables from the joint density. E.g., the joint (marginal) density of X and Y is

$$f_{X,Y}(x,y) = \int f_{X,Y,Z}(x,y,z)\, \mathrm{d}z,$$

and the (marginal) density of X is

$$f_X(x) = \iint f_{X,Y,Z}(x,y,z)\, \mathrm{d}y\, \mathrm{d}z$$

The multiplication rule holds also for a random vector which has an arbitrary number of components some of which have discrete distributions and some of which continuous distributions as long as the joint distribution of the continuous

components is of the continuous type. In this case the joint density of any subset of the components can be obtained by marginalizing out the rest of the components from the joint density: the discrete variables have to be summed out and the continuous ones integrated out.

The multiplication rule holds also for conditional distributions. E.g., consider three variables X, Y and Z. As functions of x and y we have

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z)\, f_{Y|X,Z}(y \mid x, z) = f_{Y|Z}(y \mid z)\, f_{X|Y,Z}(x \mid y, z). \tag{2.12}$$

Notice that we use one vertical bar to indicate conditioning: on the right hand side of the bar appear the variables on which we condition, in some order, and on the left hand side those variables whose conditional distribution we are discussing, in some order. We can calculate the densities of marginals of conditional distributions using the same kind of rules as for unconditional distributions: we sum over discrete and integrate over continuous variables. E.g., if the distribution of Y is continuous, then

$$f_{X|Z}(x \mid z) = \int f_{X,Y|Z}(x, y \mid z)\, \mathrm{d}y, \tag{2.13}$$

and if Y is discrete, then

$$f_{X|Z}(x \mid z) = \sum_{y} f_{X,Y|Z}(x, y \mid z). \tag{2.14}$$

Once we have more than two RVs, it becomes tedious to write the RVs as subscripts and their potential values as arguments. We let p be the generic symbol of a density. The argument of $p(\cdot)$ indicates both the symbol of the RV and its potential value. Hence, e.g., $p(x, y)$ indicates, that there are two RVs X and Y under consideration, and that we are considering their joint density $f_{X,Y}(x, y)$. The multiplication rule for two variables can be written as

$$p(x, y) = p(x)\, p(y \mid x) = p(y)\, p(x \mid y).$$

However, in some other contexts this notation can be misleading. In those cases we will use subscripts to make the notation unambiguous.

2.7 Independence and conditional independence

If we have several RVs $X_1, X_2, \ldots, X_n$, then they are independent, if their joint distribution function factorizes as

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)\, F_{X_2}(x_2) \ldots F_{X_n}(x_n), \tag{2.15}$$

for all $x_1, x_2, \ldots, x_n$. If we have available some sort of a joint density, this is the case, if it factorizes as

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \ldots f_{X_n}(x_n),$$

for all $x_1, x_2, \ldots, x_n$.

If two random variables X and Y are independent, then their joint density has to satisfy

$$f_{X,Y}(x, y) = f_X(x)\, f_{Y|X}(y \mid x) = f_Y(y)\, f_{X|Y}(x \mid y) = f_X(x)\, f_Y(y)$$

by the multiplication rule and by independence. We conclude that X and Y are independent if and only if

$$f_{X|Y}(x \mid y) = f_X(x), \qquad f_{Y|X}(y \mid x) = f_Y(y)$$

for all x and y.

Sometimes we consider an infinite sequence of RVs $X_1, X_2, \ldots$. Then the sequence is independent, if for any n, the first n RVs $X_1, X_2, \ldots, X_n$ are independent. If all the RVs X_i in a finite or infinite sequence have the same distribution, then we say that $X_1, X_2, \ldots$ is an **i.i.d.** (independent, identically distributed) sequence.

Fact. If $X_1, X_2, \ldots$ are independent, and $f_1, f_2, \ldots$ are functions, then $f_1(X_1), f_2(X_2), \ldots$ are independent.

RVs $X_1, X_2, \ldots, X_n$ are **conditionally independent** given Y, if their conditional density factorizes as

$$f_{X_1,X_2,\ldots,X_n|Y}(x_1, x_2, \ldots, x_n \mid y) = f_{X_1|Y}(x_1 \mid y)\, f_{X_2|Y}(x_2 \mid y) \ldots f_{X_n|Y}(x_n \mid y),$$

for all $x_1, x_2, \ldots, x_n$ and y. Then the joint density of $X_1, X_2, \ldots, X_n$ and Y is

$$f_{X_1,\ldots,X_n,Y}(x_1, \ldots, x_n, y) = f_Y(y)\, f_{X_1|Y}(x_1 \mid y) \ldots f_{X_n|Y}(x_n \mid y).$$

We can obtain the marginal distribution of $X_1, X_2, \ldots, X_n$ from this by integrating (or summing) y out.

If conditionally on Y, the RVs $X_1, X_2, \ldots, X_n$ are not only independent but also have the same distribution, then we say that $X_1, X_2, \ldots, X_n$ are i.i.d. given Y (or conditionally on Y). It can be shown that in this case every permutation of $(X_1, \ldots, X_n)$ has the same (marginal) distribution as any other permutation. Such a collection of RVs is called **exchangeable**.

2.8 Expectations and variances

If X is a discrete RV and h is a function such that $h(X)$ is a scalar or a vector, then the **expected value** (or **expectation** or **mean**) of $h(X)$ is

$$Eh(X) = \sum_x h(x) f_X(x).$$

On the other hand, if X is a continuous RV, then

$$Eh(X) = \int h(x) f_X(x)\, \mathrm{d}x,$$

whenever that integral can be defined and the result is finite. In particular, EX is called the mean (or expectation or expected value) of X. If X is a random vector, then the mean is also a vector.

If X is a random variable, then its variance is

$$\operatorname{var} X = E((X - EX)^2).$$

The variance is always non-negative. By expanding the square, and by the linearity of expectation,

$$\operatorname{var} X = E(X^2) - (EX)^2.$$

If X is a random vector (a column vector), then we may consider its covariance matrix (variance matrix, dispersion matrix)

$$\operatorname{Cov} X = E[(X - EX)(X - EX)^T],$$

which has dimensions $d \times d$, when X has d scalar components.

Sometimes we consider the conditional expectation of a random variable Y given the value of another random variable X. Below, we write the formulas for the case when the joint distribution of X and Y is continuous. The conditional expectation of Y given $X = x$ is defined as the expectation of the conditional distribution $y \mapsto f_{Y|X}(y \mid x)$,

$$E(Y \mid X = x) = \int y\, f_{Y|X}(y \mid x)\, \mathrm{d}y.$$

The result is a function of x, say $m(x)$. When we plug the random variable X in that function, we get a random variable $m(X)$ which is called the conditional expectation of Y given the random variable X,

$$E(Y \mid X) = m(X), \quad \text{where } m(x) = E(Y \mid X = x).$$

$E(Y \mid X)$ is a random variable.

An important property of conditional expectations is the following property (iterated expectation, tower rule),

$$EE(Y \mid X) = EY, \tag{2.16}$$

i.e., one can calculate the unconditional expectation by averaging the conditional expectation over the marginal distribution. This is valid whenever EY is a well-defined extended real number (possibly infinite). In the continuous case this follows from

$$EE(Y \mid X) = \int \left[\int y\, f_{Y|X}(y \mid x)\, \mathrm{d}y \right] f_X(x)\, \mathrm{d}x = \iint y\, f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

The conditional variance of Y given $X = x$,

$$\operatorname{var}(Y \mid X = x),$$

is defined as the variance of the conditional distribution of Y given $X = x$. The result is a function depending on x. When we substitute the random variable X for x, we get the conditional variance $\operatorname{var}(Y \mid X)$ of Y given the random variable X. We have the result

$$\operatorname{var} Y = E \operatorname{var}(Y \mid X) + \operatorname{var} E(Y \mid X). \tag{2.17}$$

This shows that conditioning decreases the variance: the variance of the conditional expectation, $\operatorname{var} E(Y \mid X)$, is less or equal to the unconditional variance $\operatorname{var} Y$.

2.9 Change of variable formula for densities

If X is a discrete RV and $Y = g(X)$ is some function X, then Y has the pmf

$$f_Y(y) = P(Y = y) = P(g(X) = y) = \sum_{x : g(x) = y} f_X(x).$$

However, for continuous distributions the situation is more complicated.

2.9.1 Univariate formula

Let us first consider the univariate situation. Suppose that X is a continuous random variable with density f_X and Y is defined by

$$Y = g(X),$$

where $g : A \to B$ is a continuously differentiable function such that

- The function $g : A \to B$ is a continuously differentiable bijection from an open interval $A \subset \mathbb{R}$ to an open interval $B \subset \mathbb{R}$.
- The inverse function $g^{-1} : B \to A$ is also continuously differentiable.
- $P(X \in A) = 1$.

Since g is a bijective function defined on an open interval, it has to be either increasing or decreasing. Suppose first that g is increasing. Suppose $a < b$ and $a, b \in B$. For convenience, let $h = g^{-1}$. Then h is increasing, and therefore

$$P(a < Y < b) = P(a < g(X) < b) = P(h(a) < X < h(b)) = \int_{h(a)}^{h(b)} f_X(x)\,\mathrm{d}x.$$

By making the change of variable

$$y = g(x) \quad \Leftrightarrow \quad x = h(y),$$

we get

$$P(a < Y < b) = \int_{h(a)}^{h(b)} f_X(x)\,\mathrm{d}x = \int_a^b f_X(h(y))\,h'(y)\,\mathrm{d}y.$$

Since this holds for all $a, b \in B$ such that $a < b$, and since $P(Y \in B) = 1$, we conclude that

$$f_Y(y) = f_X(h(y))\,h'(y), \qquad \text{when } y \in B,$$

and zero elsewhere.

On the other hand, if g is decreasing, then $h = g^{-1}$ is also decreasing, and the previous calculation holds except for a change of sign.

The end result of the calculations is that in either case Y has the density given by

$$f_Y(y) = f_X(h(y))\,|h'(y)|, \qquad \text{when } y \in B, \tag{2.18}$$

and zero elsewhere.

A useful heuristic, which helps to keep this in mind is to note that the formula

$$f_X(x)\,|\mathrm{d}x| = f_Y(y)\,|\mathrm{d}y| \tag{2.19}$$

holds under the bijective change of variable

$$y = g(x) \quad \Leftrightarrow \quad x = h(y).$$

Solving for $f_Y(y)$, we get

$$f_Y(y) = f_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = f_X(h(y))\,|h'(y)|.$$

Notice that the result holds on B, the image of A under the mapping g. Elsewhere $f_Y(y) = 0$.

The result can also be expressed by using the derivative of g instead of h, if one calculates as follows,

$$f_Y(y) = f_X(x)\frac{1}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = f_X(x)\frac{1}{|g'(x)|} = \frac{f_X(h(y))}{|g'(h(y))|}. \tag{2.20}$$

Also this formula holds on B and $f_Y(y) = 0$ elsewhere. Formula (2.20) is correct, since the formula

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{\frac{\mathrm{d}y}{\mathrm{d}x}}$$

expresses correctly the derivative of the inverse function.

This univariate case can usually be handled more easily by calculating first the cdf of $Y = g(X)$ and then by taking the derivative of the cdf. However, in higher-dimensional settings the change of variables formula becomes indispensable.

2.9.2 Multivariate formula

Consider a two-dimensional random vector $X = (X_1, X_2)$ with continuous distribution and pdf f_X, a function $g : A \to B$, where $A, B \subset \mathbb{R}^2$, and define the two-dimensional random vector Y by

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = g(X) = \begin{bmatrix} g_1(X) \\ g_2(X) \end{bmatrix}.$$

We assume that g is a **diffeomorphism**, i.e., that g is bijective, continuously differentiable, and that its inverse function is also continuously differentiable. We make the following assumptions.

- The set A is open and $P(X \in A) = 1$. The set B is the image of A under the function g. The function g is continuously differentiable.
- B is open and the inverse function $g^{-1} : B \to A$ is also continuously differentiable.

It can be shown that the random vector Y has the density

$$f_Y(y) = f_X(h(y))\,|J_h(y)|, \qquad y \in B \tag{2.21}$$

and zero elsewhere, where h is g^{-1}, the inverse function of g, and $J_h(y)$ is the **Jacobian determinant** (or Jacobian) of the function h evaluated at the point y,

$$J_h(y) = \det \begin{bmatrix} \dfrac{\partial h_1(y)}{\partial y_1} & \dfrac{\partial h_1(y)}{\partial y_2} \\ \dfrac{\partial h_2(y)}{\partial y_1} & \dfrac{\partial h_2(y)}{\partial y_2} \end{bmatrix} \tag{2.22}$$

The matrix, whose determinant the Jacobian is, is called the Jacobian matrix or the derivative matrix of the function h. This two-variate formula can be

derived in the same manner as the corresponding univariate formula by making a multivariate change of variable in a multivariate integral. Notice that we need the absolute value $|J_h(y)|$ of the Jacobian determinant in the change of variable formula (2.21).

A convenient standard notation for the Jacobian determinant is

$$J_h(y) = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}.$$

Notice that here J_h is a function of y. On the other hand, the Jacobian determinant of g,

$$J_g(x) = \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)}$$

is a function of x. When $y = g(x)$ which is the same as $x = h(y)$, then we have

$$\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \frac{\partial(y_1, y_2)}{\partial(x_1, x_2)} = 1,$$

since the two Jacobian matrices are inverses of each other, and $\det(A^{-1}) = 1/\det(A)$ for any invertible matrix A.

There is a useful heuristic also in the two-dimensional case. The formula

$$f_X(x)\, |\partial(x_1, x_2)| = f_Y(y)\, |\partial(y_1, y_2)| \tag{2.23}$$

has to hold under the bijective change of variable

$$y = g(x) \quad \Leftrightarrow \quad x = h(y).$$

Therefore

$$f_Y(y) = f_X(x) \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right| = f_X(h(y))\ |J_h(y)|$$

On the other hand, we may express $f_Y(y)$ as follows,

$$f_Y(y) = f_X(x) \frac{1}{\left| \dfrac{\partial(y_1, y_2)}{\partial(x_1, x_2)} \right|} = f_X(h(y)) \frac{1}{|J_g(h(y))|}, \tag{2.24}$$

where J_g is the Jacobian determinant of the function g (expressed as a function of x). These formulas for $f_Y(y)$ hold on the set B. Elsewhere $f_Y(y) = 0$.

The formulas (2.21) and (2.24) generalize also to higher dimensions, when one defines the Jacobians as

$$J_h(y) = \frac{\partial x}{\partial y} = \frac{\partial(x_1, \ldots, x_d)}{\partial(y_1, \ldots, y_d)} = \det \begin{bmatrix} \dfrac{\partial h_1(y)}{\partial y_1} & \cdots & \dfrac{\partial h_1(y)}{\partial y_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_d(y)}{\partial y_1} & \cdots & \dfrac{\partial h_d(y)}{\partial y_d} \end{bmatrix}$$

and

$$J_g(x) = \frac{\partial y}{\partial x} = \frac{\partial(y_1, \ldots, y_d)}{\partial(x_1, \ldots, x_d)} = \det \begin{bmatrix} \dfrac{\partial g_1(x)}{\partial x_1} & \cdots & \dfrac{\partial g_1(x)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial g_d(x)}{\partial x_1} & \cdots & \dfrac{\partial g_d(x)}{\partial x_d} \end{bmatrix}.$$

As an application of these formulas, consider a RV X, which has a d-dimensional continuous distribution, and define Y as an affine function of X,

$$Y = AX + b.$$

Here A is an invertible (i.e., nonsingular) $d \times d$ matrix and b is a d-vector, and A and b are constants (non-random quantities). Now

$$g(x) = Ax + b \quad \text{and} \quad h(y) = A^{-1}(y - b).$$

The Jacobian matrix of g is simply A and the Jacobian matrix of h is A^{-1}, so $J_g(x) = \det(A)$ and $J_h(y) = \det(A^{-1})$. By (2.21) or (2.24) we have

$$f_Y(y) = f_X(A^{-1}(y - b))|\det(A^{-1})| = \frac{f_X(A^{-1}(y - b))}{|\det(A)|}.$$

Chapter 3

Simulating Random Variables and Random Vectors

In this chapter we discuss methods for producing (on a computer) an endless supply of random values from a specified distribution, which we call the target distribution. Actually we should speak of **pseudo-random** values, since the calculated numbers are not random, but are calculated using deterministic, iterative algorithms. For practical purposes, however, the calculated values can be used as if they were the observed values of an i.i.d. sequence of RVs.

There are many terms in use for denoting this activity. Some authors speak of random variable/variate/deviate/number generation. Some say that they draw/generate/produce samples from a distribution. Some say that they simulate random variables/variates/deviates/numbers.

The aim of this chapter is not to present good (or the best) simulation methods for particular distributions. Rather, the emphasis is on explaining general principles on which such methods are based.

3.1 Simulating the uniform distribution

One speaks of **random numbers** especially when the target distribution is either the uniform distribution $\mathrm{Uni}(0,1)$ on the unit interval $(0,1)$ or the discrete uniform distribution on the set $\{0,\ldots,m-1\}$, where m is a large integer. Other distributions can be obtained from the uniform distribution by using a large variety of techniques.

Most programming languages and mathematical or statistical computing environments have available a generator for the uniform distribution $\mathrm{Uni}(0,1)$. The successive values $u_1, u_2, \ldots, u_n$ returned by a good uniform random number generator can be used as if they were the observed values of and i.i.d. sequence of random variables $U_1, U_2, \ldots, U_n$ having the uniform distribution $\mathrm{Uni}(0,1)$.

During the years, several tests have been devised for testing these key properties: uniformity and independence. (One famous test suite is the Diehard battery of tests assembled by G. Marsaglia.) Good uniform random number

generators are well documented and pass all the usual tests. Good quality mathematical and statistical computing environments have such good generators, but the reader is warned that some lower quality generators remain in use in some circles.

Mathematically, a uniform random number generator is of the form

$$s_i = g(s_{i-1}), \quad u_i = h(s_i), \qquad i = 1, 2, \ldots,$$

where s_i is the state of the generator at the ith step. (Typically, the state is either a scalar or a vector of a fixed dimension.) Notice that s_i is a deterministic function of the previous state s_{i-1}. The ith value returned by the generator is u_i, and it is obtained by applying a deterministic function to the state s_i. One needs an initial state s_0 to start the iteration. The initial state is usually called the seed state or the **seed**.

A random number generator usually provides means for

- querying and setting the seed (or state) of the generator,
- generating one or several random numbers.

If the random number generator is started on two different occasions from the same seed, one obtains exactly the same sequences of random numbers. Therefore it is important to be aware how one sets the seed and what happens if the seed is not explicitly set.

E.g., in the C programming language, there is available the uniform random number generator `random()` whose seed can be set with the functions `srandom()` or `initstate()`. If a program uses the function `random()` without setting the seed, then the seed is set to its default initial value with the consequence that different runs of the program make use of exactly the same "random" values.

From now on, it is assumed that the reader has available a uniform random number generator. Next we discuss how one can simulate i.i.d. random variables having some specified non-uniform target distribution. Basically, all methods are based on just two tricks, which are sometimes applied in a series,

- apply (one or several) deterministic transformations to uniform random numbers,
- apply a probabilistic transformation (such as random stopping in the accept–reject method) to an i.i.d. sequence of random numbers drawn from some distribution, the simulation of which is ultimately based on i.i.d. uniform random numbers.

3.2 The inverse transform

Let F be a univariate df, and let q be the corresponding quantile function. Recall from section 2.5 that if $U \sim \mathrm{Uni}(0, 1)$, then the random variable X defined by

$$X = q(U) \tag{3.1}$$

has the distribution function F. This is the *inverse transform* method or *inversion* (of the distribution function). Some other names for the method include the probability integral transform and the quantile transform(ation) method).

If $U_1, \ldots, U_n$ are i.i.d. and follow the Uni$(0,1)$ distribution, then also

$$X_1 = q(U_1), \ldots, X_n = q(U_n) \tag{3.2}$$

are i.i.d. with the distribution function F. Independence follows, since (deterministic) functions of independent random variables are themselves independent.

The inverse transform is a good choice if the quantile function of the target distribution is easy to calculate. This is the case, e.g., for

- the exponential distribution,
- the Weibull distribution,
- the Pareto distribution,
- the Cauchy distribution (which is same as the t_1 distribution); also the t_2 distribution.

Even though there may be available an iterative routine for calculating the quantile function of some given complicated target distribution, simulating it may be computationally more efficient with some other approach.

If one uses the inverse transform for simulating the general discrete distribution with pmf

$$f(i) = p_i, \qquad i = 1, 2, \ldots, k$$

with $\sum_{i=1}^{k} p_i = 1$, and remembers to use the generalized inverse function of the distribution function as the quantile function, then one obtains the following obvious algorithm.

Algorithm 1: The inverse transform method for the general discrete distribution.

Input: The pmf $p_1, p_2, \ldots, p_k$ of the target distribution.
Result: One sample I from the target distribution.
1 Generate $U \sim$ Uni$(0,1)$;
2 Return I, if

$$\sum_{j=1}^{I-1} p_j \leq U < \sum_{j=1}^{I} p_j.$$

This algorithm works by dividing the unit interval into n pieces whose lengths are $p_1, \ldots, p_k$ from left to right. Having generated U, the algorithm checks, into which of the intervals U falls, and returns the number of the interval. Notice that this algorithm requires a search, which may be time-consuming if k is large.

There are available more efficient algorithms such as the alias method for simulating the general discrete distribution. However, they require an initialization step. If one needs to generate just one value from a discrete distribution, then this simple method may well be the most efficient one.

3.3 Transformation methods

If we already know how to simulate a random vector $Y = (Y_1, \ldots, Y_k)$ with a known distribution, and we calculate (the scalar or vector) X as some function

of Y,

$$X = T(Y),$$

then X has *some* distribution. With careful choices for the distribution of Y and for the transformation T, we can obtain a wide variety of distributions for X. Of course, the inverse transform is an example of a transformation method.

Notice that if we apply the transformation T to an i.i.d. sequence $Y^{(1)}, Y^{(2)}, \ldots$ with the distribution of Y, then we obtain an i.i.d. sequence

$$X^{(1)} = T(Y^{(1)}),\ X^{(2)} = T(Y^{(2)}), \ldots$$

from the distribution of X.

In simulation settings one uses certain conventions, which are rarely explained in the literature. The main convention is the following. **If one generates several values in an algorithm, then they are generated independently.** This a natural convention, since the successive calls of the usual random number generators indeed do return values which can be considered independent (more pedantically: which can be considered to be observed values of independent random variables). E.g., a valid way of describing the preceding simulation of the i.i.d. sequence would be the following.

1. for $i = 1, 2, \ldots, n$ do
   - Generate Y from the appropriate distribution and set $X^{(i)} = T(Y)$.

   end
2. Return $(X^{(1)}, X^{(2)}, \ldots, X^{(n)})$.

In some contexts one may want to denote the generated values by lower case letters (since the actual numbers should be considered to be the observed values of random variables) and in some other contexts it is more convenient to use the corresponding upper case letters (especially when one is interested in the distributional aspects of the generated numbers). This should not cause serious confusion.

Sometimes we can use known connections between distributions to find the distribution of Y and the transformation T.

Example 3.1. The log-normal distribution. Random variable X has the log-normal distribution with parameters (μ, σ^2) if and only if its logarithm is normally distributed with mean μ and variance σ^2, i.e., if

$$\ln(X) \sim N(\mu, \sigma^2).$$

Therefore once we know how to simulate the normal distribution, we know how to simulate the log-normal distribution:

1. Generate $Y \sim N(\mu, \sigma^2)$.
2. Return $X = \exp(Y)$.

△

3.3.1 Scaling and shifting

If Y has a continuous distribution with the density g, and X is obtained from Y by scaling and shifting,

$$X = m + sY, \qquad m \in \mathbb{R}, s > 0, \tag{3.3}$$

then (by the change of variable formula for densities) X has the density

$$f(x \mid s, m) = g\left(\frac{x-m}{s}\right)\frac{1}{s}. \tag{3.4}$$

The density g is obtained with $s = 1$ and $m = 0$. If we know, how to simulate Y from the density g, then we can simulate from the density $f(\cdot \mid s, m)$ as follows.

1. Generate Y from the density g.
2. Return $X = m + sY$.

Many well-known families of continuous distributions have a scale parameter, i.e., their densities can be written in the form

$$x \mapsto g\left(\frac{x}{s}\right)\frac{1}{s}, \qquad s > 0. \tag{3.5}$$

In this case s is called a **scale parameter** of the family (and the family of distributions can be called a scale family). The density g is obtained, when $s = 1$. In this case we have the situation of (3.4) with $m = 0$, so simulation from the density with scale parameter s can be implemented as follows.

1. Generate Y from the density g.
2. Return $X = sY$.

Many families of distributions have a rate parameter, i.e., their densities can be represented as

$$x \mapsto \lambda\, g(\lambda x), \qquad \lambda > 0,$$

where g is a density. This means that the family is a scale family, with scale parameter $s = 1/\lambda$, i.e., the scale is the reciprocal of the rate.

As an example, consider the family of exponential distributions, which is usually parametrized using the rate parameter $\lambda > 0$. The density function of the Exp(λ) distribution (exponential with rate λ) is

$$\text{Exp}(x \mid \lambda) = \lambda \exp(-\lambda x) 1_{(0,\infty)}(x)$$

We see that $s = 1/\lambda$ is a scale parameter. Recall that we already know how to simulate the Exp(1) distribution (the unit exponential distribution) using the inverse transform. Therefore we can simulate the Exp(λ) distribution as follows.

1. Generate Y from the unit exponential distribution Exp(1).
2. Return $X = Y/\lambda$.

This simulation algorithm can also be derived directly using the inverse transform.

Some families of continuous distributions have both a scale and a location parameter, i.e., their densities can be written in the form (3.4). Such a family is called a location-scale family, and s is called the scale parameter and m the location parameter of the family. A familiar example is the family of normal distributions

$$\{N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}.$$

$N(\mu, \sigma^2)$, the normal distribution with mean μ and variance σ^2, has the density

$$N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) = \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Therefore μ is a location parameter, and the standard deviation (square root of variance) σ is a scale parameter of (univariate) normal distributions.

As a consequence, we can generate $X \sim N(\mu, \sigma^2)$ as follows.

1. Generate $Y \sim N(0, 1)$.
2. Return $X = \mu + \sigma Y$.

For another example of a location-scale family of distributions, consider $\mathrm{Uni}(a, b)$, the uniform distribution on the interval (a, b), where $a < b$. This distribution has the density

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

A moments reflection shows that one can simulate the $\mathrm{Uni}(a, b)$ distribution as follows.

1. Generate $U \sim \mathrm{Uni}(0, 1)$.
2. Return $X = a + (b - a)U$.

3.3.2 Polar coordinates

Consider the transformation from polar coordinates (r, ϕ) to the Cartesian coordinates (x, y),

$$x = r\cos(\phi), \qquad y = r\sin(\phi). \tag{3.6}$$

Here r is the radial coordinate and ϕ is the polar angle in radians. The mapping (3.6) is defined for all $r \geq 0$ and for all angles ϕ. However, if we want to use the change of variable formula with this mapping, we first have to restrict its domain so that the mapping becomes a bijection between its domain and its range. We obtain a bijective correspondence between (r, ϕ) and (x, y), if the domain of the mapping is selected so that $r > 0$ and ϕ is allowed to have values in any fixed open interval of length 2π.

We will use the following domain domain for the polar angle ϕ,

$$-\pi < \phi < \pi.$$

With this choice, the mapping (3.6) defines a bijective correspondence between the following open sets

$$(r,\phi) \in (0,\infty)\times(-\pi,\pi) \quad \to \quad (x,y) \in \mathbb{R}^2 \setminus \{(x,y) : x \le 0, y = 0\}. \tag{3.7}$$

Here the image of the domain $(0,\infty)\times(-\pi,\pi)$ is the coordinate plane cut along the negative x-axis. The Jacobian of the mapping $(r,\phi) \mapsto (x,y)$ is

$$\frac{\partial(x,y)}{\partial(r,\phi)} = \det \begin{bmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \phi} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \phi} \end{bmatrix} = \det \begin{bmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{bmatrix} = r.$$

The inverse function of the mapping (3.6) is a bit tricky to express. Many books state (not correctly) that we get r and ϕ form x and y by the formulas

$$r = \sqrt{x^2+y^2}, \qquad \phi = \arctan(y/x),$$

but if not an outright error, at least this is an instance of misuse of notation. If you have to program your own routines for the rectangular to polar conversion, do not use those formulas!

The formula for r is correct, and it is true that one has to select the value of ϕ so that $\tan(\phi) = y/x$. There is, however, a problem with the formula $\phi = \arctan(y/x)$, which stems from the fact, that the tangent function does not have a unique inverse function. Usually, the notation arctan means the principal branch of the (multivalued) inverse tangent function with the range

$$-\pi/2 < \arctan(u) < \pi/2, \qquad u \in \mathbb{R}\,.$$

If you use this convention and the formula $\phi = \arctan(y/x)$, then your polar coordinate point (r,ϕ) is guaranteed not to be in the second or third quadrant even if your original Cartesian coordinate point (x,y) is.

So, care is needed with the Cartesian to polar coordinate formula $(x,y) \mapsto (r,\phi)$. One expression, which is correct and easy to program, is given by

$$r = \sqrt{x^2+y^2}, \qquad \phi = \operatorname{atan2}(y,x), \tag{3.8}$$

where $\operatorname{atan2}(y,x)$ is the arc tangent function of two variables, which is defined for all $(x,y) \neq (0,0)$. It returns the counterclockwise (signed) angle in radians in the range $(-\pi,\pi]$ between the positive x axis and the vector (x,y). The function atan2 is available in most programming languages (but the order of the arguments is reversed in some programming environments). If (x,y) does not fall on the negative x-axis, then r and ϕ calculated by (3.8) satisfy $r > 0$ and $-\pi < \phi < \pi$.

The polar to Cartesian conversion formula (3.6) and the Cartesian to polar conversion formula (3.8) define a diffeomorphism between the sets in eq. (3.7).

After this preparation, suppose the two-dimensional random vector (X,Y) has a continuous density, and we want to express this distribution by means of polar coordinates (R,Φ) using the conversion formula (3.8). Now the probability that (X,Y) is exactly on the negative x-axis, $P(X \le 0, Y = 0) = 0$, since the joint distribution is continuous. Furthermore, we have a diffeomorphism

between the coordinates (r, ϕ) and (x, y) given by formulas (3.6) and (3.8). Hence, we can apply the change of variables formula with the result

$$f_{R,\Phi}(r,\phi) = f_{X,Y}(x,y)\left|\frac{\partial(x,y)}{\partial(r,\phi)}\right| = r f_{X,Y}(r\cos\phi, r\sin\phi), \qquad r>0, -\pi<\phi<\pi. \tag{3.9}$$

Actually, the same formula for $f_{R,\Phi}$ is valid, if we choose *any* open interval of length 2π as the domain of ϕ. This follows, since in that case one can define a diffeomorphism between rotated versions of the sets in eq. (3.7), and the Jacobian needed in the change of variables formula is still r.

Suppose in particular that the density $f_{X,Y}(x, y)$ is invariant under rotations about the origin, i.e., that

$$f_{X,Y}(x,y) = g(r), \quad \text{with } r = \sqrt{x^2+y^2}. \tag{3.10}$$

Then the polar coordinates of (X, Y) have the density

$$f_{R,\Phi}(r,\phi) = rg(r) = 2\pi r g(r)\,\frac{1}{2\pi}, \qquad r>0, -\pi<\phi<\pi.$$

This shows that R and Φ are independent, the polar angle Φ has the uniform distribution on its domain of length 2π (and this is obvious because of the rotational symmetry!), and the density of R can be read off from the previous formula. I.e., under the assumption (3.10), we have

$$R \perp\!\!\!\perp \Phi, \tag{3.11}$$

$$\Phi \sim \mathrm{Uni}(-\pi,\pi), \qquad f_R(r) = 2\pi r g(r), \quad r>0. \tag{3.12}$$

On the other hand, suppose we start with a density for the polar coordinates (R, Φ),

$$f_{R,\Phi}(r,\phi), \qquad r>0, -\pi<\phi<\pi$$

and let (X, Y) be (R, Φ) in Cartesian coordinates (formula (3.6)). By the change of variables formula,

$$f_{X,Y}(x,y) = \frac{f_{R,\Phi}(r,\phi)}{\left|\frac{\partial(x,y)}{\partial(r,\phi)}\right|} = \frac{f_{R,\Phi}(\sqrt{x^2+y^2}, \mathrm{atan2}(y,x))}{\sqrt{x^2+y^2}}, \tag{3.13}$$

where, initially, it is forbidden that (x, y) is on the negative x-axis. However, any continuous joint density for (X, Y) implies that

$$P(X \le 0, Y = 0) = 0,$$

and so we can let x and y to have any real values in (3.13). An exception is the origin $(x, y) = (0, 0)$, since the formula (3.13) is not defined at the origin, but one can use any value for $f_{X,Y}$ there, and the result remains correct.

As an application of the formulas in this section, consider the joint distribution of two independent variables, X and Y, having the standard normal distribution $N(0, 1)$. Their joint density is

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-y^2/2} = \frac{1}{2\pi}\exp(-r^2/2), \qquad \text{with } r^2 = x^2+y^2,$$

and so it is invariant under rotations about the origin. Let (R, Φ) be (X, Y) in polar coordinates. According to formulas (3.11) and (3.12), R and Φ are independent, $\Phi \sim \mathrm{Uni}(-\pi, \pi)$, and the density of R is

$$f_R(r) = r \exp(-r^2/2), \qquad r > 0.$$

The distribution of R belongs to the family of Rayleigh distributions. A statistician recognizes more easily the distribution of $Z = R^2$. A change of variables gives

$$f_Z(z) = f_R(\sqrt{z}) \frac{1}{2\sqrt{z}} = \frac{1}{2} \exp(-\frac{1}{2} z), \qquad z > 0,$$

so $Z \sim \mathrm{Exp}(1/2)$, the exponential distribution with rate $1/2$.

As a side product, we have obtained a way to simulate two independent samples X and Y from the standard normal distribution $N(0, 1)$. We have actually rediscovered the the famous method of Box and Muller, first published in 1958. (Notice: the name is Muller, not Müller.)

Algorithm 2: The method of Box and Muller, initial version.

Result: Two independent samples X and Y from $N(0, 1)$.

1 Generate independently $Z \sim \mathrm{Exp}(1/2)$ and $\Phi \sim \mathrm{Uni}(-\pi, \pi)$;

2 $X \leftarrow \sqrt{Z} \cos(\Phi), \quad Y \leftarrow \sqrt{Z} \sin(\Phi)$.

Of course, since we know how to simulate the $\mathrm{Exp}(1/2)$ and $\mathrm{Uni}(-\pi, \pi)$ distributions using the uniform distribution $\mathrm{Uni}(0, 1)$, we can implement the method of Box and Muller also as follows.

Algorithm 3: The method of Box and Muller, second version.

Result: Two independent samples X and Y from $N(0, 1)$.

1 Generate U and V independently from the $\mathrm{Uni}(0, 1)$ distribution;

2 $X \leftarrow \sqrt{-2 \ln U} \cos(\pi(2V - 1)), \quad Y \leftarrow \sqrt{-2 \ln U} \sin(\pi(2V - 1))$.

If you did not know about the explanation involving polar coordinates, these formulas would probably seem totally mysterious to you.

Actually, Box and Muller stated their method in the following form.

Algorithm 4: The method of Box and Muller, original version.

Result: Two independent samples X and Y from $N(0, 1)$.

1 Generate U and V independently from the $\mathrm{Uni}(0, 1)$ distribution;

2 $X \leftarrow \sqrt{-2 \ln U} \cos(2\pi V), \quad Y \leftarrow \sqrt{-2 \ln U} \sin(2\pi V)$.

This form uses the same idea, but corresponds to the convention that the polar angle belongs to the interval $(0, 2\pi)$.

There are also other methods for generating two independent draws from the standard normal, which are based on the use of polar coordinates (look up the Marsaglia polar method in Wikipedia). If one uses a bad uniform random number generator, then the method of Box and Muller leads to certain practical difficulties, although the method is exact if one uses uniform random variables.

3.3.3 The ratio of uniforms method

A nonnegative function $h \geq 0$ defined on some Euclidean space is called an **unnormalized density**, if its integral over the whole space is finite and nonzero. An unnormalized density can be converted to a density function f by

normalizing it,

$$f(x) = h(x) \Big/ \int h(t)\,\mathrm{d}t, \qquad x \in \mathbb{R}\,.$$

Unnormalized densities occur quite frequently in Bayesian statistics in the form

$$\text{prior} \times \text{likelihood}.$$

Truncated distributions (defined in the next section) provide other examples of unnormalized densities.

For still another example, consider the following definition for the uniform distribution on a set $A \subset \mathbb{R}^d$. Let $m(A)$ be the Lebesgue measure of $A \subset \mathbb{R}^d$, given by

$$m(A) = \int 1_A(x)\,\mathrm{d}x.$$

If $A \in \mathbb{R}$, then $m(A)$ is the length of set A; if $A \in \mathbb{R}^2$, then $m(A)$ is the area of A; if $A \in \mathbb{R}^3$, then $m(A)$ is the volume of A, and if $A \in \mathbb{R}^d$, we can call $m(A)$ the d-dimensional volume of A. Let $A \subset \mathbb{R}^d$. We assume that A has nonzero, finite d-dimensional volume, $0 < m(A) < \infty$. The **uniform distribution on the set** A, which we can denote by $\mathrm{Uni}(A)$, is the continuous distribution having the unnormalized density 1_A. The corresponding normalized density is, of course, $1_A/m(A)$.

Suppose that we want to generate samples from a distribution having a given unnormalized density h on the real line. Define the set $C \in \mathbb{R}^2$ by

$$C = \{(u, v) : 0 < u < \sqrt{h(v/u)}\}, \tag{3.14}$$

Kinderman and Monahan (1977) noticed that if we are able to generate the pair (U, V) from the uniform distribution on C, then V/U has the distribution corresponding to the unnormalized density h.

Algorithm 5: The ratio of uniforms method.

Assumption: We know how to simulate $\mathrm{Uni}(C)$, see eq. (3.14).
Result: One sample X from the distribution with unnormalized density h.

1 Generate $(U, V) \sim \mathrm{Uni}(C)$;
2 $X \leftarrow V/U$

The correctness of the algorithm can be proved by first completing the transformation by (e.g.) defining $Y = U$, after which we have a bijective correspondence between (U, V) and (X, Y), and then by calculating the density of X from the joint density of (X, Y). The joint density can be calculated easily by the change of variables formula. The details are left as an exercise for the reader. The uniform distribution on the set C can often be simulated in the manner described in the next section.

3.4 Naive simulation of a truncated distribution

Suppose that RV X has a continuous distribution with density f_X. Suppose A a set such that $P(X \in A) > 0$. Then we can consider the distribution of X

truncated (or restricted) to the set A, which has the unnormalized density given by

$$y \mapsto f_X(y)1_A(y). \tag{3.15}$$

This is also called the distribution of X conditionally on $X \in A$ (or given $X \in A$).

We can simulate this truncated distribution with the following, obvious method. Notice that we follow the usual convention: in the following algorithm, the successive draws within the repeat–unitl loop from the distribution with density f_X are supposed to be independent.

Algorithm 6: Naive method for simulating from a truncated distribution.

Input: Set A and simulation method for f_X.
Result: A sample Y from f_X truncated to the set A.

```
repeat
    Simulate X from the density f_X
until X ∈ A ;
Y ← X (i.e., accept X, if it is in A).
```

The correctness of this method follows from the following calculation,

$$P(Y \in B) = P(X \in B \mid X \in A) = \frac{\int_{A\cap B} f_X(x)\,\mathrm{d}x}{P(X \in A)} = \int_B f_Y(y)\,\mathrm{d}y,$$

where

$$f_Y(y) = \frac{1}{P(X \in A)}\, f_X(y)1_A(y).$$

The efficiency of this method depends on the acceptance probability

$$p = P(X \in A). \tag{3.16}$$

The number of simulations needed in order to get one acceptance has the geometric distribution on $1, 2, \ldots$ with success probability p. The mean of this distribution is $1/p$.

For example, suppose that we simulate the standard normal $N(0,1)$ truncated to the set $A = (5, \infty)$ using this naive method. Then the acceptance probability p turns out to be about $2.9 \cdot 10^{-7}$. With sample size of ten million from the $N(0,1)$ distribution, the expected number of accepted values would be 2.9. On the other hand, should we be interested in simulating $N(0,1)$ truncated to the complementary set $(-\infty, 5]$, then practically every point of the sample would be accepted by the naive method.

One important application for this naive simulation method is simulation of the uniform distribution on some complicated set A. Suppose that we are able to find a set B, such that $A \subset B$, and we already know how to simulate the uniform distribution on the set B. Then the uniform distribution on B truncated to the set A is the uniform distribution on A. This obvious fact can be proved by noting that the uniform distribution on B truncated to the set A has the unnormalized density

$$1_B 1_A = 1_{A\cap B} = 1_A,$$

where the last step follows from the inclusion $A \subset B$. As a consequence, we can simulate $Y \sim \mathrm{Uni}(A)$ as follows.

- Generate $X \sim \text{Uni}(B)$ until $X \in A$, and then return $Y = X$.

Often we are interested a set $A \subset \mathbb{R}^2$, which can be enclosed in a rectangle $B = (a, b) \times (c, d)$. The uniform distribution on the rectangle B can simulated by generating independently the first coordinate from $\text{Uni}(a, b)$ and the second coordinate from $\text{Uni}(c, d)$.

Sometimes it is costly to test whether $x \in A$. In such a case we can save some computational effort, if we can find a simpler set S such that $S \subset A$. So, now we have the inclusions

$$S \subset A \subset B, \tag{3.17}$$

and we know how to simulate $\text{Uni}(B)$. If now $X \in S$ with reasonable probability, and it is less costly to test, whether $x \in S$ than whether $x \in A$, then we can, on average, save some computational effort with the following algorithm.

Algorithm 7: Simulating from $\text{Uni}(A)$, with a pretest.

Assumption: The inclusions $S \subset A \subset B$ hold, and we know how to simulate $\text{Uni}(B)$

Result: One sample Y from $\text{Uni}(A)$.

```
repeat
    Generate X ~ Uni(B);
    if X ∈ S then accept ← true;
    else if X ∈ A then accept ← true;
    else accept ← false
until accept ;
Y ← X
```

The algorithm uses a Boolean variable `accept` to keep track of whether the proposed value X has been accepted or not.

If we use the naive method repeatedly (using an i.i.d. sequence of X's) to generate several values $Y_1, Y_2, \ldots, Y_n$, then they are i.i.d. On first thought this may seem obvious. After further thought this may, however, seem not so obvious anymore. The independence of the generated Y's can be proved either by elementary means or by appealing to the strong Markov property of i.i.d. sequences, but we skip the proof. The basic idea is that the sequence of X's starts afresh after each (random) time when a freshly generated Y is accepted.

3.5 Accept–reject method

In this section $f^* : \mathbb{R}^d \to [0, \infty)$ is an unnormalized density of some continuous target distribution. The corresponding normalized density function is

$$f(x) = f^*(x) \Big/ \int f^*(t)\, \mathrm{d}t.$$

In most of the applications of the method $d = 1$, but the method can be used in any dimension.

3.5.1 The fundamental theorem

Suppose $d = 1$ and consider **the set under the graph of** f^*, i.e., the set bounded by the x-axis and the graph of the function f^*,

$$A = \{(x, y) : 0 < y < f^*(x)\}. \tag{3.18}$$

The area of A is

$$m(A) = \int \left(\int_0^{f^*(x)} 1 \, \mathrm{d}y \right) \mathrm{d}x = \int f^*(x) \, \mathrm{d}x.$$

The same calculation for $m(A)$ holds for other values of d, too.

Suppose (X, Y) is uniformly distributed in the set A (3.18), and let us calculate (a) the marginal density of X and (b) the conditional density of Y given $X = x$. The joint density of (X, Y) is given by

$$f_{X,Y}(x, y) = \begin{cases} 1/m(A), & \text{if } (x, y) \in A, \\ 0, & \text{otherwise} \end{cases}$$

By the following calculation, the marginal density of X is simply f

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}y = \int_0^{f^*(x)} \frac{1}{m(A)} \, \mathrm{d}y = \frac{f^*(x)}{m(A)} = f(x).$$

If x is such that $f^*(x) > 0$ and y is such that $0 < y < f^*(x)$, we have

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{1}{f^*(x)},$$

while for other values of y, the conditional density is zero. In other words, given $X = x$, the random variable Y has the uniform distribution on the interval $(0, f^*(x))$.

We have incidentally proved the following theorem, which Robert and Casella call the fundamental theorem of simulation.

Theorem 1 (Fundamental theorem of simulation.)**.** *Suppose f^* is an unnormalized density on $\mathbb{R}^d$ and let f bet the corresponding normalized density. Let A be the set under the graph of f^*, i.e.,*

$$A = \{(x, y) : 0 < y < f^*(x)\}.$$

Then we have the following

1. *If $(X, Y) \sim \mathrm{Uni}(A)$, then $X \sim f$.*
2. *If $X \sim f$ and, conditionally on $X = x$, Y has the distribution $\mathrm{Uni}(0, f^*(x))$, then $(X, Y) \sim \mathrm{Uni}(A)$.*

3.5.2 Deriving the accept–reject method

Suppose that f^* is defined on the real line and that the set where $f^* > 0$ is a finite interval (a, b). Further, suppose f^* is bounded, $f^* \leq K$. Then we can enclose the set A in the rectangle $(a, b) \times (0, K)$, whose uniform distribution is simple to simulate. Hence we can simulate the uniform distribution on A by the naive method for truncated distributions. But not all pdfs of interest are supported on a finite interval. What to do in that case?

The solution is to apply the fundamental theorem twice. Suppose that we are able to find a (normalized) density function g such that

1. Mg majorizes (or envelopes) the unnormalized target density f^*, where $M > 0$ is a known (majorizing) constant, i.e.,

$$f^*(x) \leq Mg(x) \qquad \text{for all } x. \tag{3.19}$$

2. We know how simulate from g.

Then

$$A = \{(x, y) : 0 < y < f^*(x)\} \subset B = \{(x, y) : 0 < y < Mg(x)\}.$$

By the fundamental theorem, we can simulate (X, Y) from the uniform distribution on B as follows,

$$\text{Generate } X \sim g \text{ and } U \sim \text{Uni}(0, 1); \text{ set } Y = Mg(X)U.$$

Therefore we can use the naive method for a truncated distribution to simulate the uniform distribution on A: we simulate $(X, Y) \sim \text{Uni}(B)$ until (X, Y) falls under the graph of f^*. Combining these ideas, we get the following algorithm.

Algorithm 8: The accept–reject method.

Assumption: The unnormalized f^* is majorized by Mg
Result: One sample X from f.

```
repeat
    Generate Z ~ g and U ~ Uni(0, 1).
until Mg(Z)U < f*(Z) ;
X ← Z (i.e., accept the proposal Z).
```

Remarks

- Some people call the method acceptance sampling or the acceptance method; some others call it rejection sampling or the rejection method.
- The majorizing function $Mg(z)$ is also called the envelope of $f^*(z)$.
- The method can also described so that one accepts the proposal $Z \sim g$ with probability $f^*(Z)/(Mg(Z))$.
- The accept–reject method was originally published by John von Neumann in 1951.
- Although the method works in any dimension, finding useful envelopes in high-dimensional cases is very challenging.

The efficiency of the method depends crucially on the acceptance probability. Notice that the joint density Z and U before the acceptance test is

$$f_{Z,U}(z,u) = g(z)1_{(0,1)}(u).$$

Therefore the acceptance probability is

$$\begin{aligned} p &= P\left(U < \frac{f^*(Z)}{Mg(Z)}\right) \\ &= \int \mathrm{d}z \int_0^{f^*(z)/(Mg(z))} \mathrm{d}u \, g(z)1_{(0,1)}(u) \\ &= \int g(z)\frac{f^*(z)}{Mg(z)} \, \mathrm{d}z = \frac{\int f^*(z) \, \mathrm{d}z}{M}. \end{aligned} \tag{3.20}$$

If $d = 1$, this is the same as

$$\frac{\text{Area under } f^*}{\text{Area under the envelope } Mg.}$$

(Here, e.g., "area under f^*" actually means the area of the set bounded by the graph of f^* and the x-axis.) In order to get high efficiency, we need as high acceptance probability as possible. This is achieved by using a tightly fitting envelope Mg.

For a fixed g, if the majorizing condition

$$f^* \leq Mg$$

holds for $M = M_0$, then it holds for all $M \geq M_0$. In order to achieve the best efficiency, one should choose the least possible value for M such that the majorizing condition holds. However, the accept–reject method is valid for any choice of M such that the majorizing condition is true.

3.5.3 An example of accept–reject

Consider the unnormalized target density

$$f^*(x) = \exp(-x^2/2)(1 + 2\cos^2(x)\sin^2(4x)), \tag{3.21}$$

which is majorized by the function

$$Mg(x) = 3\exp(-x^2/2).$$

Here $Mg(x)$ is an unnormalized density of the $N(0,1)$ distribution, so g is the density of $N(0,1)$. Based on this fact, we could (but now need not) give an expression for M. The implied value of M is valid, but is not the best possible.

The following fragment coded in the R-language calculates $n = 1000$ independent values from the distribution corresponding to f^* using the accept–reject method and stores them in the vector `x`. The acceptance condition $Mg(Z)U < f^*(Z)$ has been converted to the equivalent condition

$$U < \frac{f^*(Z)}{Mg(Z)},$$

which now simplifies a bit.

```
n <- 1000;
x <- numeric(n)  # create a vector with n entries to store the results
for (i in 1:n) { # generate x[i]
  while (TRUE) {
    z <- rnorm(1); u <- runif(1)
    if (u < (1 + 2 * cos(z)^2 * sin(4 * z)^2) / 3) { # accept!
      x[i] <- z
      break
    }
  }
}
```

3.5.4 Further developments of the method

Sometimes the function f^* is costly to evaluate, but we can find a simpler function $s \geq 0$ which minorizes it,

$$s(x) \leq f^*(x) \leq Mg(x), \qquad (\text{all } x). \tag{3.22}$$

Then we can say that f^* has been squeezed between the lower envelope s and the upper envelope Mg. Sometimes such a function s is called a squeeze.

If s is less costly to evaluate than f^*, then we can save computation by using the following algorithm instead of the original version of accept–reject.

Algorithm 9: Accept–reject with squeezing.

Assumption: Inequality (3.22) holds
Result: One sample X from f.
repeat
  Generate $Z \sim g$ and $U \sim \text{Uni}(0, 1)$;
  $Y \leftarrow Mg(Z)U$;
  if $Y < s(Z)$ **then** accept $\leftarrow$ **true**;
  else if $Y < f^*(Z)$ **then** accept $\leftarrow$ **true**;
  else accept $\leftarrow$ **false**;
until *accept* ;
$X \leftarrow Z$

Here the test $Y < s(Z)$ is now the pretest. If it succeeds, then certainly $Y < f^*(Z)$ and there is no need to evaluate $f^*(Z)$.

Many familiar univariate continuous distributions have log-concave densities. A function is called log-concave, if its logarithm is a concave function. We are now interested in the case, where the density f is defined on an open interval (a, b), and f is strictly positive and twice differentiable on that interval. Then f is log-concave, if and only if

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log f(x) \leq 0, \qquad a < x < b.$$

The graph of a concave function lies below each of its tangents. Also, the graph of a concave function lies above each of its chords (secants). Therefore it is easy to find piecewise linear upper and lower envelopes for concave functions. If one constructs piecewise linear envelopes for $\log f$, then, by exponentiation, one gets piecewise exponential envelopes $s \leq f \leq g^*$. It turns out to be relatively easy

to generate values from the distribution, which has the piecewise exponential unnormalized density g^*. After this has been accomplished, we can immediately use the accept–reject method with squeezing to simulate from the log-concave density f.

It is even possible to construct iteratively better and better upper and lower envelopes for a log-concave density, so that the bounds get tighter every time a new value is generated from the density. This is called **adaptive rejection sampling (ARS)**, but there exist several different implementations of this basic idea.

3.6 Using the multiplication rule for multivariate distributions

Suppose we want to simulate the joint distribution of three variables X, Y and Z. The multiplication rule (i.e., the chain rule) gives us a decomposition of the joint distribution of the form

$$f_{X,Y,Z}(x, y, z) = f_X(x)\, f_{Y|X}(y \mid x)\, f_{Z|X,Y}(z \mid x, y).$$

If all the distributions on the right are available in the sense that we know how to simulate from them, then we can be interpret the multiplication rule as a recipe for simulating the joint distribution.

Algorithm 10: Using the multiplication rule for simulation, pedantic version

1 Generate the value x from f_X;
2 Generate the value y from $f_{Y|X}(\cdot \mid x)$;
3 Generate the value z from $f_{Z|X,Y}(\cdot \mid x, y)$.

If we repeat the process, we get i.i.d. samples

$$(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots$$

from the joint distribution of (X, Y, Z). Of course, one can generalize this to as many components as are needed. The components need not be scalars, but they may as well be vectors or even matrices.

Many people tend to describe the same algorithm more informally, e.g., as follows.

Algorithm 11: Using the multiplication rule for simulation, informal version

1 Generate $x \sim p(x)$;
2 Generate $y \sim p(y \mid x)$;
3 Generate $z \sim p(z \mid x, y)$.

This is acceptable, if both the writer and the reader understand what this is supposed to mean. However, the danger of misunderstanding (or rather, not understanding anything) is great.

3.7 Mixtures

It is instructive to consider the special case of the multiplication rule, when there are just two components. It is useful to the check what the marginal

distribution of the first component looks like. Simulating from the marginal distribution in this way is sometimes called the composition method.

Suppose X is continuous and J is discrete with values $1, 2, \ldots, k$. Then their joint distribution has the density

$$f_{X,J}(x, j) = f_{X|J}(x \mid j) f_J(j).$$

Let us denote

$$p_j = f_J(j), \quad \text{and} \quad f_j = f_{X|J}(\cdot \mid j), \qquad j = 1, 2, \ldots, k.$$

Then the marginal density of X is a convex combination of the densities f_j,

$$f_X(x) = \sum_{j=1}^{k} p_j f_j(x), \quad \text{where} \quad p_j \geq 0 \; \forall j, \quad \sum_{j=1}^{k} p_j = 1. \tag{3.23}$$

If we have a representation of the form (3.23), where the functions f_j are densities, then we say that the density of X is a (finite) mixture of the densities $f_1, \ldots, f_k$. The numbers $p_1, \ldots, p_k$ can be called mixing weights. We can simulate such a finite mixture distribution as follows.

Algorithm 12: Simulating from a finite mixture of distributions

1 Generate J from the pmf $(p_1, p_2, \ldots, p_k)$;
2 Generate X from density f_J;
3 Return X (and ignore J).

Similarly, if the distribution of (X, Y) is continuous, then the marginal distribution of X is

$$f_X(x) = \int f_{X|Y}(x \mid y) \, f_Y(y) \, \mathrm{d}y. \tag{3.24}$$

If we have a representation of the form (3.24), then we say that the distribution of X is a (continuous) mixture of the densities $f_{X|Y}$. In such a case, simulation can be implemented as follows.

Algorithm 13: Simulating from a continuous mixture of distributions

1 Generate y from density f_Y;
2 Generate $X \sim f_{X|Y}(\cdot \mid y)$;
3 Return X (and ignore y).

Some important distributions can be represented in the form (3.24) so that y is the scale parameter of the family of distributions

$$\{f_{X|Y}(\cdot \mid y) : y > 0\}.$$

In this case we can say that the distribution of X is a scale mixture of the distributions $f_{X|Y}$.

Example 3.2. [Simulating the multivariate t distribution] Let $\nu > 0$, $\mu \in \mathbb{R}^d$ and let Σ be a symmetric, positive definite $d \times d$ matrix. The multivariate t distribution $t_d(\nu, \mu, \Sigma)$ can be represented hierarchically as a scale mixture of multivariate normal distributions

$$X \mid Y \sim N_d(\mu, \Sigma/Y), \quad \text{where} \quad Y \sim \mathrm{Gam}(\nu/2, \nu/2).$$

Therefore it can be simulated as follows

1. Generate $Y \sim \text{Gam}(\nu/2, \nu/2)$.
2. Generate $Z \sim N_d(\mu, \Sigma)$.
3. Return $X = Z/\sqrt{Y}$.

We will discuss methods for simulating the multivariate normal distribution in the next Section.

The multivariate t distribution has become popular in Monte Carlo studies since its location and shape can be adjusted (by varying μ and Σ) and since it has heavier tails than the corresponding multivariate normal distribution. $\triangle$

3.8 Affine transformations

Affine transformations of random vectors are multivariate analogs of scaling and shifting of univariate random variables. If d-dimensional Z has density f_Z and X is defined by

$$X = b + AZ,$$

where $b \in \mathbb{R}^d$ is a constant vector, and A is an invertible, constant $d \times d$ matrix, then X has the density

$$f_X(x) = \frac{f_Z(A^{-1}(x-b))}{|\det(A)|}. \tag{3.25}$$

Example 3.3. [Another multivariate generalization of the t distribution] Let $Z_1, \ldots Z_d$ be independent random variables so that Z_i has the univariate t distribution with $\nu_i > 0$ degrees of freedom. Then the vector $Z = (Z_1, \ldots, Z_d)$ has the pdf

$$f_Z(z) = \prod_{i=1}^{d} t(z_i \mid \nu_i),$$

which is **not** the same as the elliptically contoured multivariate t density we discussed earlier (not even when $\nu_1 = \cdots = \nu_d$). The random vector $X = b + AZ$ has the density given by (3.25). Confusingly, many authors call the resulting distribution of X the multivariate t distribution. $\triangle$

We can apply affine transformations in order to simulate the multivariate normal distribution $N(\mu, \Sigma)$. Here $\mu \in \mathbb{R}^d$ is the mean (vector) of the distribution, and Σ, the covariance matrix of the distribution, is a $d \times d$ matrix. Σ is always symmetric and positive semidefinite. We now assume that Σ is positive definite, in which case it is also invertible. Then the $N(\mu, \Sigma)$ distribution has a density given by

$$f_X(x) = (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right). \tag{3.26}$$

For any symmetric, positive definite matrix Σ it is possible to find a matrix A such that

$$\Sigma = AA^T, \quad A \text{ is } d \times d \text{ and invertible} \tag{3.27}$$

One method for finding A is to use the Cholesky decomposition $\Sigma = LL^T$, where L is (the Cholesky factor of Σ) is a lower triangular matrix. Another possible

choice is to use the symmetric, positive definite square root of Σ, often denoted by $\Sigma^{1/2}$, as the matrix A.

Let us consider, what is the density of the vector $Z = (Z_1, \ldots, Z_d)$, when $Z_i \sim N(0,1)$ independently $i = 1, \ldots, d$. Then

$$f_Z(z) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z_i^2} = (2\pi)^{-d/2} \exp\left(-\frac{1}{2} z^T z\right).$$

This is the d-dimensional standard normal distribution $N(0, I_d)$.

Suppose we have available the decomposition (3.27) and calculate as follows.

1. Generate $Z \sim N_d(0, I)$.
2. Return $X = \mu + AZ$.

Then it can be proved that $X \sim N(\mu, \Sigma)$ either directly from eq. (3.25) or by using familiar properties of the multivariate normal distribution (i.e., an affine transform of a multivariate normal rv also has a multivariate normal distribution).

Sometimes one has to simulate a high-dimensional normal distribution $N(\mu, \Sigma)$ whose covariance matrix Σ is not explicitely available but whos precision matrix $Q = \Sigma^{-1}$ (inverse covariance matrix) is known. Suppose that one is able to obtain a decomposition

$$Q = BB^T$$

for the precision matrix. Then one can simulate the distribution as follows

1. Generate $Z \sim N(0, I)$.
2. Solve Y from the linear equation $B^T Y = Z$, and return $X = \mu + Y$.

This follows since Y now has the normal distribution $N(0, (B^T)^{-1}((B^T)^{-1})^T)$, where

$$(B^T)^{-1}((B^T)^{-1})^T = (B^T)^{-1}B^{-1} = (BB^T)^{-1} = Q^{-1}.$$

Another possibility is that one is able to generate efficiently from the normal distribution $N(0, Q)$ whose covariance matrix Q is the precision matrix of the target distribution. Then one can do as follows

1. Generate $Z \sim N(0, Q)$.
2. Solve Y from $QY = Z$, and return $X = \mu + Y$.

3.9 Literature

The following text books are good references for the topics of this chapter.

Bibliography

[1] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[2] Averll M. Law and W. David Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 2nd edition, 1991.

[3] Brian D. Ripley. *Stochastic Simulation*. John Wiley & Sons, 1987.

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

[5] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*. Wiley, 2nd edition, 2008.

Chapter 4

Monte Carlo Integration

In this chapter we discuss approximate integration methods, which use an i.i.d. sample $X_1, X_2, \ldots$ from some distribution. In a later chapter we will discuss MCMC methods, where the underlying random variables are not independent and where they do not have identical distributions.

Monte Carlo methods are computational methods, which depend on the use of random or pseudo random numbers. The name Monte Carlo refers to the famous casino located in Monaco. Like casino games, Monte Carlo methods are highly repetitive and depend on randomness.

4.1 Limit theorems

When the underlying sample is i.i.d., one can use the two most important limit theorems of probability theory to analyze the behavior of arithmetic means.

Theorem 2 (Strong law of large numbers, SLLN)**.** *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables such that $E|Y_i| < \infty$. Denote $\mu = EY_i$. Then*

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \to \mu,$$

almost surely, as $n \to \infty$.

Remark. The condition $E|Y_i| < \infty$ guarantees that the expectation EY_i is defined and finite. It is the best possible condition in the strong law of large numbers for i.i.d. random variables. If $E|Y_i| = \infty$, then it can be shown that

$$\limsup_{n \to \infty} \left| \frac{1}{n} \sum_{i=1}^{n} Y_i \right| \to \infty$$

almost surely, which means that the sample mean oscillates wildly and therefore diverges.

Theorem 3 (Central Limit Theorem, CLT)**.** *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables such that $EY_i^2 < \infty$. Denote*

$$\mu = EY_i, \qquad \sigma^2 = \operatorname{var} Y_i,$$

and assume that $\sigma^2 > 0$. *Then*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} N(0,1), \tag{4.1}$$

as $n \to \infty$.

In the CLT the arrow $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution. Random variables $Z_1, Z_2, \ldots$ converge in distribution to a limit distribution with df F, if

$$P(Z_n \leq z) \to F(z), \quad \text{as } n \to \infty$$

at all points of continuity z of F. Since in the CLT the df of the limit distribution $N(0,1)$ is continuous, in the CLT the convergence of the distribution functions holds at each point.

In CLT the quantity in (4.1) which has a limit distribution is the standardized mean of the n first random variables. I.e., if we denote

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

then

$$E\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

and

$$\begin{aligned}\operatorname{var} \bar{Y}_n &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)\right)^2\right] = \frac{1}{n^2} E\left[\sum_{i=1}^{n}(Y_i - \mu)\sum_{j=1}^{n}(Y_j - \mu)\right] \\ &= \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2.\end{aligned}$$

Therefore the numerator is $\bar{Y}_n$ minus its expectation, and the denominator is the standard deviation of $\bar{Y}_n$.

4.2 Confidence intervals for means and ratios

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with mean μ and finite variance $\sigma^2 > 0$. If the sample size n is large, then we can pretend that the standardized mean already follows its limit distribution, i.e., we can pretend that

$$\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \overset{\mathrm{d}}{=} N(0,1).$$

This an example of normal approximation.

Suppose we know σ but do not know μ. Then we can calculate a confidence limit for μ as follows. We seek a central $100(1-\alpha)\%$ confidence interval, for some $0 < \alpha < 1$. Let $z_{1-\alpha/2}$ be the value of the quantile function of the standard normal $N(0,1)$ at $1-\alpha/2$, i.e., a proportion $1-\alpha/2$ of the probability mass of $N(0,1)$ lies to the left of $z_{1-\alpha/2}$. E.g., a 95 % confidence interval corresponds to $\alpha = 0.05$ and $z_{0.975} \approx 1.96$. We start from the normal approximation

$$P\left(\left|\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}\right| \leq z_{1-\alpha/2}\right) \approx 1 - \alpha.$$

When we solve the inequality for μ, we see that approximately with probability $1-\alpha$ we have

$$\mu \in \bar{Y}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Usually not only μ but also σ would be unknown. However, we can still apply the preceding confidence interval, when we plug in a reasonable estimate $\hat{\sigma}$ of the standard deviation σ. Usually one uses the sample standard deviation of the Y_i values,

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\big(Y_i - \bar{Y}_n\big)^2},$$

which is a consistent estimate of σ. With this choice, we get the approximate $(1-\alpha)100\%$ confidence interval for the mean μ

$$\mu \in \bar{Y}_n \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}. \tag{4.2}$$

Here the quantity

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\big(Y_i - \bar{Y}_n\big)^2},$$

is called the standard error of the mean.

Instead of the critical values of the standard normal, one often uses the critical values of the t distribution with $(n-1)$ degrees of freedom in the previous construction. If the sample size is large, then the resulting confidence interval is in practice the same as (4.2).

There is nothing probabilistic about the coverage a single confidence interval: the interval either contains μ or does not. However, if one constructs a large number of $(1-\alpha)100\%$ confidence intervals (4.2), where n is large, then approximately proportion $(1-\alpha)$ of them covers μ and proportion α does not cover μ.

Our confidence interval (4.2) is valid only for a fixed sample size n. It is also possible to develop confidence bands for *the running mean plot*, which plots $\bar{Y}_m$ against m, see the books by Robert and Casella [5, 6].

Sometimes we need an estimate of the ratio

$$r = \frac{EX}{EY}, \tag{4.3}$$

when we have an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the joint distribution of X and Y. A natural estimator is the ratio of the averages,

$$\hat{r} = \frac{\bar{X}}{\bar{Y}}, \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{4.4}$$

This estimator is consistent but biased. Now we develop an approximate confidence interval for it.

Consider the random variable $Z = X - rY$, which has mean zero and variance

$$\sigma^2 = \operatorname{var}(Z) = \operatorname{var}(X) - 2r\operatorname{cov}(X,Y) + r^2\operatorname{var}(Y). \tag{4.5}$$

Let $Z_i = X_i - rY_i$. Then the Z_i are i.i.d. random variables and therefore the CLT ensures that

$$\frac{\bar{Z}}{\sigma/\sqrt{n}} = \frac{\bar{X} - r\bar{Y}}{\sigma/\sqrt{n}} = \frac{\hat{r} - r}{\sigma/(\sqrt{n}\bar{Y})}$$

converges in distribution to $N(0,1)$ as $n \to \infty$. Normal approximation gives now

$$P\left(\left|\frac{\hat{r} - r}{\sigma/(\sqrt{n}\,\bar{Y})}\right| \le z_{1-\alpha/2}\right) \approx 1 - \alpha$$

Therefore, an approximate $(1-\alpha)100\%$ confidence interval for the ratio EX/EY is

$$\hat{r} \pm z_{1-\alpha/2}\frac{S}{\sqrt{n}\,\bar{Y}} \tag{4.6}$$

where

$$S^2 = S_X^2 - 2\hat{r}\,S_{XY}^2 + 2\hat{r}^2\,S_Y^2$$

is the estimator of σ^2 where the unkwnown population parameters have been replaced by their consistent estimators, and

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$
$$S_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

We see from (4.6) that in large samples $\hat{r}$ is approximately normally distributed with mean r and with variance

$$\mathrm{var}(\hat{r}) \approx \frac{1}{n}\frac{S^2}{(\bar{Y})^2}. \tag{4.7}$$

This simple derivation of the results for the ratio estimator has been borrowed from [7, Sec. 4.3.2.2]. The same results can be derived based on a multivarate version of the CLT and the delta method.

4.3 Basic principles of Monte Carlo integration

Suppose f is a density, which we are able to to simulate from, and that we are interested in the expectation

$$I = \int h(x)f(x)\,\mathrm{d}x = Eh(X). \tag{4.8}$$

Suppose that we simulate $X_1, X_2, \ldots$ independently from the density f and set $Y_i = h(X_i)$. Then the sequence $Y_1, Y_2, \ldots$ is i.i.d. and

$$EY_i = Eh(X_i) = \int h(x)f(x)\,\mathrm{d}x = I.$$

If we calculate the mean of the N values $h(X_1), \ldots, h(X_N)$, then we obtain the estimate

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} h(X_i). \tag{4.9}$$

By the SLLN, $\hat{I}_N$ converges to I as N increases, provided that the condition $E|h(X)| < \infty$ holds. In Monte Carlo simulations we are free to select N as large as our budget (available computer time) allows.

We have

$$E\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} Eh(X_i) = I,$$

and therefore the estimate $\hat{I}_N$ is unbiased. It is also easy to express the variance and the standard error of the estimator. If the variance of a single term $h(X)$ is finite, then the variance of the average is

$$\operatorname{var} \hat{I}_N = \frac{1}{N} \operatorname{var} h(X). \tag{4.10}$$

This can be called the sampling variance, simulation variance or Monte Carlo variance of the estimator $\hat{I}_N$.

A more meaningful quantity for measuring the accuracy of $\hat{I}_N$ is the square root of the variance. Recall that the square root of the variance of an estimator (i.e., its standard deviation) is called its **standard error**. (This term is commonly used also for the estimate of the (theoretical) standard error.) The standard error of a Monte Carlo estimate can be called its sampling standard error, simulation standard error of Monte Carlo standard error. The Monte Carlo standard error is of the order $1/\sqrt{N}$, since

$$\sqrt{\operatorname{var} \hat{I}_N} = \frac{1}{\sqrt{N}} \sqrt{\operatorname{var} h(X)}. \tag{4.11}$$

The theoretical variance (population variance) $\operatorname{var} h(X)$, which is needed in both (4.10) and (4.11), is usually unknown. However, it it can be estimated by the sample variance of the $h(X_i)$ values,

$$s^2 = \widehat{\operatorname{var}}\, h(X) = \frac{1}{N-1} \sum_{i=1}^{N} \left(h(X_i) - \hat{I}_N \right)^2.$$

We get an approximate $100(1-\alpha)\%$ confidence interval for I from (4.2), namely

$$\hat{I}_N \pm z_{1-\alpha/2} \frac{s}{\sqrt{N}}. \tag{4.12}$$

Example 4.1. Calculating the 95 % confidence interval (4.12) with R. We assume that the sample from the density f is generated with the call `rname(N)`. We also assume that we have available a function `h`, which applies the function h element-by-element to its vector argument.

```
x <- rname(N)
# Calculate vector y so that y[i] = h(x[i]) for all i.
y <- h(x)
```

```
Ihat <- mean(y)
se <- sqrt(var(y) / N)
# or: se <- sd(y) / sqrt(N)
z <- qnorm(1 - 0.05/2)
ci <- c(Ihat - z * se, Ihat + z * se)
```

△

The accuracy of Monte Carlo integration goes to zero like $1/\sqrt{N}$ as N increases. To get an extra decimal place of accuracy it is necessary to increase N by a factor of 100. In practice, one usually achieves moderate accuracy with a moderate simulation sample size N. However, in order to achieve high accuracy, one usually needs an astronomical simulation sample size. Notice, however that Monte Carlo integration works equally well in a space of any dimensionality. In contrast, the classical quadrature rules of numerical analysis become prohibitively expensive in high dimensional spaces.

Notice, how versatile Monte Carlo integration is. If one wants to estimate several expectations $Eh_1(X), Eh_2(X), \ldots, Eh_k(X)$, then a single sample $X_1, \ldots, X_N$ from the density f suffices, since

$$Eh_j(X) \approx \frac{1}{N} \sum_{i=1}^{N} h_j(X_i), \qquad j = 1, \ldots, k.$$

In that case one uses *common random numbers* to estimate the different expectations.

4.4 Empirical quantiles

Often one wants to estimate the quantile function of a random variable X, when one has available a sample $X_1, \ldots, X_N$ (i.i.d. or not) from its distribution. Then one speaks of the **empirical quantile function**. This problem can be approached via Monte Carlo integration. One wants to solve x from the equation

$$E1_{(-\infty,x]}(X) = u, \qquad 0 < u < 1,$$

for various values of u. One can approximate the expectation by the Monte Carlo method. However, the resulting equation does not have a unique solution, as we will see in a moment.

Let $X_{(j)}$ be the j'th smallest observation, which is also called the j'th order statistic of the sample. I.e., the observations sorted from lowest to highest are

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}.$$

If

$$X_{(j)} < x < X_{(j+1)}$$

for some $j = 1, \ldots, N$, then by Monte Carlo

$$E1_{(-\infty,x]}(X) \approx \frac{1}{N} \sum_{i=1}^{N} 1_{(-\infty,x]}(X_i) = \frac{j}{N}.$$

Therefore a reasonable value for the empirical quantile function at $u = j/N$ is some value between $X_{(j)}$ and $X_{(j+1)}$, and one can use various interpolation methods to extend the definition to all values $0 < u < 1$.

Different statistical computer packages use slightly different formulas to define the empirical quantile function. There is latitude in selecting the exact point at which the empirical quantile function takes on the j'th order statistic and latitude in how one interpolates in between. E.g., in R the empirical quantile function is calculated by the function `quantile()`, and the user can choose between nine definitions of the empirical quantile function. For a large sample from a continuous distribution, all the definitions calculate approximately the same results.

4.5 Techniques for variance reduction

It is always possible to estimate the unknown integral by using different representations of the form

$$\int h(x)f(x)\,\mathrm{d}x.$$

A clever choice may imply a significantly lower variance for the Monte Carlo estimator. Then one speaks of **variance reduction** methods.

E.g., to reduce variance, it is always a good idea to try to carry out the computation analytically as far as possible, and then use Monte Carlo integration only as a last resort.

Suppose that we have two Monte Carlo methods for estimating the same integral. Let the variance in method i be

$$\frac{v_i}{N}, \qquad i = 1, 2,$$

where N is the sample size employed. Then, in order to achieve the same accuracy (e.g., the same variance or the same standard error), we should use in method two the sample size

$$\frac{v_2}{v_1}N,$$

where N is the sample size used in method one.

4.5.1 Conditioning

Conditioning decreases variance in the sense that

$$\operatorname{var} E(Z \mid Y) \leq \operatorname{var} Z$$

for any random variables Y and Z. In Monte Carlo integration it is therefore advantageous to use the conditional expectation of the integrand instead of the original integrand, whenever that is possible. Conditioning performs part of the original integration analytically, and the rest by Monte Carlo.

Conditioning is often called **Rao-Blackwellization**. (Explanation: in the celebrated Rao-Blackwell theorem one conditions on a sufficient statistic.)

To exemplify conditioning, suppose we want to estimate the integral

$$I = Eh(X, Y) = EE(h(X, Y) \mid Y),$$

and are able to compute the conditional expectation

$$m(y) = E[h(X,Y) \mid Y = y].$$

Then we can estimate I either by simulating $(X_i, Y_i), i = 1, \ldots, N$ from the joint distribution of (X, Y) and by calculating

$$\hat{I}_N^{(1)} = \frac{1}{N} \sum_{i=1}^{N} h(X_i, Y_i)$$

or by calculating

$$\hat{I}_N^{(2)} = \frac{1}{N} \sum_{i=1}^{N} m(Y_i).$$

Supposing that the computational effort required for evaluating $h(X_i, Y_i)$ or $m(Y_i)$ is about the same, the second method is better since its variance is lower.

One case where this idea can be used is in estimating posterior predictive expectations. We have often the situation, where in addition to the observed data we want to consider a future observation Y^*. The distribution of Y^* conditionally on the observed data $Y = y$ is its **(posterior) predictive distribution**. Typically, the data Y and future observation Y^* are modeled as conditionally independent given the parameter Θ. Then the joint posterior of Θ and Y^* factorizes as follows

$$p(y^*, \theta \mid y) = p(\theta \mid y)\, p(y^* \mid y, \theta) = p(\theta \mid y)\, p(y^* \mid \theta),$$

where the first identity follows by the multiplication rule for conditional distributions, and the second by conditional independence. Therefore we can simulate the joint posterior distribution of Y^* and Θ by first simulating θ_i from the posterior distribution $p(\theta \mid y)$ and then y_i^* from the sampling distribution of Y^* conditionally on the simulated value θ_i. We can estimate the mean $E[Y^* \mid Y = y]$ of the posterior predictive distribution by straightforward Monte Carlo as follows

$$\hat{I}_N^{(1)} = \frac{1}{N} \sum_{i=1}^{N} y_i^*.$$

However, in a typical situation we would know the mean of Y^* given the value of the parameter Θ, i.e., the mean of the sampling distribution of Y^*,

$$m(\theta) = E[Y^* \mid \Theta = \theta] = \int y^* p(y^* \mid \theta)\, \mathrm{d}y^*.$$

In this case we obtain a better estimator of $E[Y^* \mid Y]$ by conditioning,

$$\hat{I}_N^{(2)} = \frac{1}{N} \sum_{i=1}^{N} m(\theta_i).$$

The same approach applies also, when we want to estimate the expectation

$$E[h(Y^*) \mid Y = y],$$

where h is a function for which we know

$$\int h(y^*)\, p(y^* \mid \theta)\, \mathrm{d}y^*,$$

which is the expectation of $h(Y^*)$ given $\Theta = \theta$.

4.5.2 Control variates

Sometimes we want estimate the expectation $I = Eh(X)$ and know that

$$\mu = Em(X),$$

where m is a known function and μ is a known constant. By defining

$$W = h(X) - \beta(m(X) - \mu), \tag{4.13}$$

where β is a constant, we get a RV W, whose expectation is I. Since

$$\operatorname{var} W = \operatorname{var} h(X) - 2\beta \operatorname{cov}(h(X), m(X)) + \beta^2 \operatorname{var} m(X),$$

the lowest possible variance for W is obtained by selecting for β the value

$$\beta^* = \frac{\operatorname{cov}(h(X), m(X))}{\operatorname{var} m(X)}. \tag{4.14}$$

Here we must have $\operatorname{var}(m(X)) > 0$. If we use $\beta = \beta^*$ in (4.13), then

$$\operatorname{var} W = \operatorname{var} h(X) - \frac{\operatorname{cov}^2(h(X), m(X))}{\operatorname{var} m(X)}.$$

Notice that $\operatorname{var} W < \operatorname{var} h(X)$, if the RVs $h(X)$ and $m(X)$ are correlated, i.e., if $\operatorname{cov}(h(X), m(X)) \neq 0$. The stronger the correlation, the greater the variance reduction.

If we manage to select the value β so that $\operatorname{var} W < \operatorname{var} h(X)$, then we should estimate I as the mean of values W_i which are simulated from the distribution of W,

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} \left[h(X_i) - \beta(m(X_i) - \mu) \right]. \tag{4.15}$$

Here $X_1, \ldots, X_N$ is an i.i.d. sample with the distribution of X. Here $m(X)$ is the **control variate**, whose expectation we know. The variance of the control variate estimator (4.15) is less than the variance of the naive Monte Carlo estimator, which just averages the values $h(X_i)$.

To understand, why this happens, suppose that $\operatorname{cov}(h(X), m(X))$ is positive. Then also β should be selected positive. In this case an unusually high outcome for $\bar{h}$, the sample average of the $h(X_i)$ values, tends to be associated with an unusually high outcome for $\bar{m}$ the sample average of the $m(X_i)$ values In that case the control variate estimate adjusts the naive Monte Carlo estimate $\bar{h}$ of $Eh(X)$ downward, i.e.,

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} \left[h(X_i) - \beta(m(X_i) - \mu) \right] = \bar{h} - \beta(\bar{m} - \mu),$$

where

$$\bar{h} = \frac{1}{N} \sum_{i=1}^{N} h(X_i), \qquad \bar{m} = \frac{1}{N} \sum_{i=1}^{N} m(X_i).$$

Similar explanation works also when the correlation is negative.

The optimal β^* depends on the moments of RVs $h(X)$ and $m(X)$, and these are usually unknown. However, we can estimate the optimal β by using a pilot sample X_i', $i = 1, \ldots, n$. We then divide the sample covariance of $h(X_i')$ and $m(X_i')$ with the sample variance of $m(X_i')$. This is then our estimate of β^*, which is then used in eq. (4.15) with a fresh sample $X_1, \ldots, X_N$.

Somewhat surprisingly, the same calculation can be done by fitting a linear model, as follows. We fit the linear model

$$h(X_i') = \alpha + \beta m(X_i') + \epsilon_i, \qquad i = 1, \ldots, n.$$

by least squares

$$\sum_{i=1}^{n} \left(h(X_i') - \alpha - \beta m(X_i')\right)^2 = \min!,$$

and this can be done by using any statistical package. Here the errors ϵ_i are definitely not normally distributed as would be required for linear models. We are just using the available software for linear models for our own purposes. This approach works, since the least squares estimate of β happens to be the same as calculated in the previous approach for estimating β^*. The estimated slope, $\hat{\beta}$, is then used in eq. (4.15) and the estimated intercept $\hat{\alpha}$ is ignored.

Example 4.2. Suppose that `rname(n)` simulates n values from the distribution of X and that `hfunc(x)` and `mfunc(x)` calculates the functions h and m for each value of its vector argument. Then the following code fragments demonstrates the two ways of estimating β^*.

```
x.pilot <- rname(n.pilot)
h <- hfunc(x.pilot); m <- mfunc(x.pilot)
beta <- cov(m, h) / var(m)
# Alternative; here the function lm() fits the linear model.
model <- lm(h ~ m)
# ... select for beta the estimated coefficient of m:
beta <- coef(model)[2]

# Then we estimate the integral and the simulation standard error
x <- rname(n)
h <- hfunc(x); m <- mfunc(x)
w <- h - beta * (m - mu)
Ihat <- mean(w)
se <- sd(w) / sqrt(n)
```

△

If one knows several expectations

$$\mu_j = Em_j(X), \qquad j = 1, \ldots, k,$$

then it is possible to use several control variates $m_1(X), \ldots, m_k(X)$. The values of the optimal coefficients can, again, be estimated using a pilot sample and by fitting a linear model.

4.5.3 Common random numbers

Often one wants to compare two expectations

$$I_1 = E_f h_1(X), \qquad \text{and} \qquad I_2 = E_f h_2(X),$$

where the functions h_1 and h_2 resemble one another. Suppose we estimate the expectations by the Monte Carlo estimators $\hat{I}_1$ and $\hat{I}_2$. We are interested in the sign of the difference $I_1 - I_2$. Since

$$\operatorname{var}\left(\hat{I}_1 - \hat{I}_2\right) = \operatorname{var}(\hat{I}_1) + \operatorname{var}(\hat{I}_2) - 2\operatorname{cov}(\hat{I}_1, \hat{I}_2),$$

it is worthwhile to use estimators, which have positive correlation. This is typically achieved by basing the estimators $\hat{I}_1$ and $\hat{I}_2$ on common random numbers, i.e., by using a single sample $X_1, \ldots, X_N$ instead of separate samples for the two estimators.

Using common random numbers is even more important in the case, where one tries to estimate a parametrized expectation

$$I(\alpha) = E_f h(X, \alpha)$$

for various values of the parameter α. Then the estimator using common random numbers produces a much smoother approximation

$$\alpha \mapsto \hat{I}(\alpha)$$

then what would be obtained by using separate samples at each α. Besides, by using common random numbers one saves a lot of computational effort.

4.6 Importance sampling

Suppose we want to estimate the integral

$$I = E_f[h(X)] = \int h(x)\, f(x)\, \mathrm{d}x, \tag{4.16}$$

where the density f might be difficult to sample from. We can rewrite the integral as

$$I = \int_{\{g>0\}} h(x) \frac{f(x)}{g(x)}\, g(x)\, \mathrm{d}x = E_g\left[h(X)\frac{f(X)}{g(X)}\right]. \tag{4.17}$$

Here the subscript of the expectation symbol shows, under what distribution the expectation is calculated. Robert and Casella [5] call this the importance sampling **fundamental identity**. In importance sampling, one draws a sample from g and uses the fundamental identity for developing a Monte Carlo estimate of the integral. This idea was used already in the early 1950's.

The new density g can be selected otherwise quite freely, but we must be certain that

$$g(x) = 0 \quad \Rightarrow \quad h(x)f(x) = 0,$$

since otherwise the integrals (4.16) and (4.17) are not guaranteed to be equal. In other words, the support of the function hf must be included in the support of the function g.

Of course, the fundamental identity can be formulated for other types of distributions, too. If X has a discrete distribution with pmf f, then we may estimate the sum

$$I = E_f[h(X)] = \sum_x h(x)\, f(x)$$

by drawing a sample from another pmf g with the same support, since

$$I = \sum_x h(x)\, \frac{f(x)}{g(x)}\, g(x).$$

Again, the support of the function hf must be included in the support of the function g.

In its most general form, the fundamental identity of importance sampling uses the concept of Radon–Nikodym derivative of measures. Let μ be the probability distribution of X. We assume that μ, when restricted to the set

$$\{h \neq 0\} = \{x : h(x) \neq 0\},$$

is absolutely continuous relative to another probability measure ν, which means that

$$\nu(B \cap \{h \neq 0\}) = 0 \quad \Longrightarrow \quad \mu(B \cap \{h \neq 0\}) = 0,$$

for all sets B. Then

$$\begin{aligned} Eh(X) &= \int h(x)\, \mu(\mathrm{d}x) = \int_{\{h\neq 0\}} h(x)\, \mu(\mathrm{d}x) \\ &= \int_{\{h\neq 0\}} h(x)\, \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\, \nu(\mathrm{d}x) \end{aligned}$$

where the function $\mathrm{d}\mu/\mathrm{d}\nu$ is the Radon–Nikodym derivative of (the restriction of) μ relative to (the restriction of) ν. When the distribution μ has pdf f and the probability measure ν has the pdf g, then the Radon–Nikodym derivative $\mathrm{d}\mu/\mathrm{d}\nu(x)$ is simply the ratio $f(x)/g(x)$.

4.6.1 Unbiased importance sampling

We assume the setting (4.17), where f is the pdf of a continuous distribution. We assume that we know f completely, including its normalizing constant.

We select a density g, which is easy to sample from. Then we generate a sample $X_1, \ldots, X_N$ from g and calculate

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} h(X_i)\, \frac{f(X_i)}{g(X_i)} \tag{4.18}$$

Let us call the ratio of the densities

$$w(x) = \frac{f(x)}{g(x)}$$

the importance ratio (or likelihood ratio), and the weights

$$w_i = w(X_i) = \frac{f(X_i)}{g(X_i)}, \qquad i = 1, \ldots, N \tag{4.19}$$

the **importance weights**. Then the importance sampling estimate (4.18) can be written as

$$\hat{I}_N = \frac{1}{N} \sum_{i=1}^{N} w_i h(X_i).$$

Importance sampling gives more weight for those sample points X_i for which $f(X_i) > g(X_i)$ and downweights the other sample points, in order to form an unbiased estimate of $I = E_f[h(X)]$, given a sample $X_1, \ldots, X_N$ from g.

Different authors use different names for g such as the importance sampling density, the approximation density, proposal density and so on. Following Robert and Casella [5], we call g the **instrumental** density.

We can interpret the procedure as producing a **weighted sample**

$$(w_1, X_1), \ldots, (w_N, X_N),$$

where the weights are needed in order to correct for the fact that the sample is produced from the wrong density. Since the estimator (4.18) is the arithmetic mean of terms $w_i\, h(X_i)$ each with mean I,

$$E_g[w_i\, h(X_i)] = E_g\left[\frac{f(X_i)}{g(X_i)}\, h(X_i)\right] = \int h(x) f(x)\, \mathrm{d}x = I,$$

the estimator is unbiased. Its variance can be estimated in the same way as the variance of the basic Monte Carlo estimator.

In importance sampling we should strive for low variance. In particular, the variance should be finite. This is the case, if and only if the expectation of the square of one term is finite, i.e., we should have

$$E_g\left[h^2(X)\frac{f^2(X)}{g^2(X)}\right] = \int h^2(x)\frac{f^2(x)}{g(x)}\, \mathrm{d}x < \infty.$$

If this condition is not satisfied, then the estimator behaves erratically.

One pair of conditions which guarantees finite variance is

$$\mathrm{var}_f[h(X)] < \infty, \quad \text{and} \quad \frac{f(x)}{g(x)} \leq M, \quad \forall x,$$

for some $M > 0$. The second of these conditions means that the tails of g should be at least as heavy as the tails of f.

In order to achieve minimal variance, one can show that it is optimal to choose the instrumental density g proportional to $|h|f$. Then the variance of the importance sampling estimator is smaller (or equal to) the variance of the naive Monte Carlo estimator, which uses samples from f. While the optimal choice

$$g \propto |h|f$$

can hardly ever be used in practice, it can still provide some guidance in choosing the form of g: the shape of the instrumental density should resemble the product $|h|f$ as closely as possible. One should focus sampling on the regions of interest where $|h|f$ is large in order to save computational resources.

On the other hand, if the integrand h is not fixed in advance (e.g., one wants to estimate expectations for many functions h) then the instrumental density g

should be selected so that $f(x)/g(x) = w(x)$ is nearly constant. If g is a good approximation to f, then all the importance weights will be roughly equal. If, on the other hand, g is a poor approximation to f, then most of the weights will be close to zero, and thus a few of the X_i's will dominate the sum, and the estimate will be inaccurate. Therefore it is a good idea to inspect the importance weights, e.g., by examining their variance or histogram.

Notice that the importance weights can be utilized to form a control variate. Denoting the importance weight w_i by $w(X_i)$, we have

$$E_g w(X_i) = \int \frac{f(x)}{g(x)}\, g(x)\, \mathrm{d}x = 1.$$

Therefore the average of the importance weights can be used as a control variate, whose expectation is known to be one.

4.6.2 Self-normalized importance sampling

It is possible to apply importance sampling also in the situation, where we want to estimate $I = E_f[h(X)]$, but only know an unnormalized version f^* of the density f. Here

$$f(x) = \frac{1}{c}\, f^*(x),$$

but the normalizing constant c is unknown. Of course, c can be expressed as the integral

$$c = \int f^*(x)\, \mathrm{d}x.$$

Such a situation is common in Bayesian statistics, but also when f^* corresponds to a truncated density. In these cases we cannot calculate (4.18) directly. However, we can express the integral as

$$I = \int h(x) f(x)\, \mathrm{d}x = \frac{\int h(x) f^*(x)\, \mathrm{d}x}{\int f^*(x)\, \mathrm{d}x},$$

and then estimate the numerator and denominator separately using importance sampling.

We sample $X_1, \ldots, X_N$ from an instrumental density g. We estimate the denominator by

$$\int f^*(x)\, \mathrm{d}x = \int \frac{f^*(x)}{g(x)}\, g(x)\, \mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} \frac{f^*(X_i)}{g(X_i)} = \frac{1}{N} \sum_{i=1}^{N} w_i,$$

where we use the importance weights w_i corresponding to the unnormalized density f^*, given by

$$w_i = \frac{f^*(X_i)}{g(X_i)}.$$

Our estimate of the numerator is

$$\int h(x) f^*(x)\, \mathrm{d}x \approx \frac{1}{N} \sum_{i=1}^{N} h(X_i) \frac{f^*(X_i)}{g(X_i)} = \frac{1}{N} \sum_{i=1}^{N} w_i h(X_i).$$

Canceling the common factor $1/N$, we obtain the following self-normalized importance sampling estimator (which is usually called just the importance sampling estimator without any further qualification).

1. Generate $X_1, X_2, \ldots, X_N$ from density g.
2. Calculate the importance weights
$$w_i = \frac{f^*(X_i)}{g(X_i)}$$
3. Estimate I by the weighted average
$$\hat{I} = \frac{\sum_{i=1}^N w_i h(X_i)}{\sum_{j=1}^N w_j}. \tag{4.20}$$

The same method can be described so that having calculated the (raw) importance weights w_i, one calculates the **normalized importance weights**,
$$\tilde{w}_i = \frac{w_i}{s}, \quad \text{where} \quad s = \sum_{j=1}^n w_j,$$
by dividing the raw weights by their sum, and then calculates the (self-normalized) importance sampling estimate as
$$\bar{I} = \sum_{i=1}^N \tilde{w}_i \, h(X_i).$$

Unlike the unbiased estimator (4.18), the self-normalized estimator (4.20) is not unbiased. Its bias is, however, negligible when N is large.

In both forms of importance sampling it is a good idea to inspect the importance weights w_i. If only few of the weights are large and others are negligible, then the estimate is likely not accurate. In self-normalized importance sampling one can examine the histogram or the coefficient of variation (which is the sample standard deviation divided by the sample mean) of the importance weights (standardized or not).

4.6.3 Variance estimator for self-normalized importance sampling

One should view the self-normalized estimator (4.20) as the ratio of two averages
$$\hat{I} = \frac{\frac{1}{N}\sum_{i=1}^N h(x_i) w_i}{\frac{1}{N}\sum_{j=1}^N w_j} = \frac{\bar{U}}{\bar{w}},$$
where $\bar{U}$ is the average of N RVs $U_i = h(X_i)w_i$ and $\bar{w}$ is the average of N raw importance weights w_i, and the pairs (U_i, w_i) are i.i.d. random vectors. Then we can apply the formula (4.7) for the approximate variance of the ratio,
$$\frac{1}{N}\frac{1}{\bar{w}^2} S^2,$$
where
$$\begin{aligned}(N-1)S^2 &= \sum (U_i - \bar{U})^2 - 2\frac{\bar{U}}{\bar{w}} \sum (U_i - \bar{U})(V_i - \bar{V}) + \frac{\bar{U}^2}{\bar{w}^2} \sum (w_i - \bar{w})^2 \\ &= \sum (h(X_i) - \hat{I})^2 w_i^2,\end{aligned}$$

after some straightforward calculations.

This implies the following formula for the Monte Carlo variance of the self-normalized importance sampling estimator,

$$\begin{aligned}\widehat{\operatorname{var}}\,\hat{I} &= \frac{1}{N(N-1)}\,\frac{1}{\bar{w}^2}\sum_{i=1}^{N}(h(X_i)-\hat{I})^2\,w_i^2 \\ &= \frac{N}{N-1}\,\frac{\sum_{i=1}^{N}(h(X_i)-\hat{I})^2\,w_i^2}{(\sum_{j=1}^{N} w_j)^2}.\end{aligned} \tag{4.21}$$

Geweke [2] omits the term $N/(N-1)$ which corresponds to using denominator N instead of $N-1$ in the formulas for the sample variances and covariances.

A $(1-\alpha)100\%$ confidence interval for I is then given by

$$\hat{I} \pm z_{1-\alpha/2}\,\sqrt{\widehat{\operatorname{var}}\,\hat{I}}.$$

4.6.4 SIR: Sampling importance resampling

Importance sampling can be interpreted so that it produces a weighted sample $(p_1, X_1), \ldots, (p_N, X_N)$, where now $(p_1, \ldots, p_N)$ is a probability vector (i.e., a probability mass function on $1, \ldots, N$). Then $I = E_f[h(X)]$ is approximated by

$$\sum_{i=1}^{N} p_i h(X_i).$$

The probability vector is here the vector of normalized importance weights.

However, for some purposes one needs a true sample; a weighted sample does not suffice. Such a sample can be produced approximately by sampling with replacement from the sequence

$$X_1, \ldots, X_N$$

with probabilities given by the vector $(p_1, \ldots, p_N)$. This is called SIR (sampling/importance resampling). Following Smith and Gelfand [9], this approach is sometimes called the weighted bootstrap.

If one samples without replacement, then one obtains an i.i.d. sample, which comes from an approximation to the target distribution. The approximation improves as the size of the initial sample N increases. (Sampling with replacement does not here result in an i.i.d. sample.)

4.7 Literature

Monte Carlo integration and variance reduction methods are discussed in the simulation literature, see e.g., Law and Kelton [3] (introductory level), Rubinstein and Kroese [7] and Asmussen and Glynn [1] (advanced level). Ripley [4] demonstrates how one can reduce the simulation variance by a factor of 10^8 by using variance reduction techniques cleverly.

Bibliography

[1] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation*. Springer, 2007.

[2] John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57:1317–1339, 1989.

[3] Averll M. Law and W. David Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 2nd edition, 1991.

[4] Brian D. Ripley. *Stochastic Simulation*. John Wiley & Sons, 1987.

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

[6] Christian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Springer, 2010.

[7] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*. Wiley, 2nd edition, 2008.

[8] Mark J. Schervish. *Theory of Statistics*. Springer series in statistics. Springer-Verlag, 1995.

[9] A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

Chapter 5

More Bayesian Inference

We use the generic $p(\cdot)$ notation for densities, if there is no danger of confusion.

5.1 Likelihoods and sufficient statistics

Let us consider n (conditionally) independent Bernoulli trials $Y_1, \ldots, Y_n$ with success probability θ. That is, the RVs Y_i are independent and Y_i takes on the value 1 with probability θ (success in the i'th Bernoulli experiment) and otherwise is zero (failure in the i'th Bernoulli experiment). Having observed the values $y_1, \ldots, y_n$, the likelihood corresponding to $y = (y_1, \ldots, y_n)$ is given by

$$\begin{aligned} p(y \mid \theta) &= \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^s (1-\theta)^{n-s}, \qquad 0 < \theta < 1, \end{aligned} \tag{5.1}$$

where

$$s = t(y) = \sum_{i=1}^{n} y_i$$

is the observed number of successes. Here the likelihood depends on the data y only through the value of $t(y)$, which is said to be a **sufficient statistic**. Since

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta) = \theta^{t(y)}(1-\theta)^{n-t(y)}\, p(\theta),$$

the posterior depends on the data only through the value of $t(y)$.

In a more general situation, a statistic $t(Y)$ is called sufficient, if the likelihood can be factored as

$$p(y \mid \theta) = g(t(y), \theta)\, h(y)$$

for some functions g and h. Then (as a function of θ)

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta) \propto g(t(y), \theta)\, p(\theta)$$

and therefore the posterior depends on the data only through the value $t(y)$ of the sufficient statistic.

In Bayesian inference, we might as well throw away the original data as soon as we have calculated the value of the sufficient statistic. (Do not try this at home. You might later want to consider other likelihoods for your data!) Sufficient statistics are very convenient, but not all likelihoods admit a sufficient statistic of a fixed dimension, when the sample size is allowed to vary. Such sufficient statistics exist only in what are known as exponential families, see, e.g., the text of Schervish [5, Ch. 2] for a discussion.

In the Bernoulli trial example, the random variable S corresponding to the sufficient statistic

$$S = t(Y) = \sum_{i=1}^{n} Y_i$$

has the binomial distribution $\mathrm{Bin}(n, \theta)$ with sample size n and success probability θ. I.e., if we observe only the number of success s (but not the order in which the successes and failures happened), then the likelihood is given by

$$p(s \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}, \qquad 0 < \theta < 1. \tag{5.2}$$

The two functions (5.1) and (5.2) describe the same experiment, and are proportional to each other (as functions of θ). The difference stems from the fact that there are exactly $\binom{n}{s}$ equally probable sequences $y_1, \ldots, y_n$, which sum to a given value of s, where s is one of the values $0, 1, \ldots, n$. Since the two functions are proportional to each other, we will get the same posterior with either of them if we use the same prior. Therefore it does not matter which of the expressions (5.1) and (5.2) we use as the likelihood for a binomial experiment.

Observations.

- When calculating the posterior, you can always leave out from the likelihood such factors, which depend only on the data but not on the parameter. Doing that does not affect the posterior.
- If your model admits a convenient sufficient statistic, you do not need to work out the distribution of the sufficient statistic in order to write down the likelihood. You can always use the likelihood of the underlying repeated experiment, even if the original data has been lost and only the sufficient statistic has been recorded.
- However, if you do know the density of the sufficient statistic (conditionally on the parameter), you can use that as the likelihood. (This is tricky; consult, e.g., Schervish [5, Ch. 2] for a proof.)

We can generalize the Bernoulli experiment (or binomial experiment) to the case, where there are $k \geq 2$ possible outcomes instead of two possible outcomes. Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from the discrete distribution with k different values $1, \ldots, k$ with respective probabilities $\theta_1, \ldots, \theta_k$, where $0 < \theta_j < 1$ and $\sum \theta_j = 1$. (Because of the sum constraint, there are actually only $k-1$ free parameters.) The likelihood corresponding to the data $y = (y_1, \ldots, y_n)$ is given by

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_j^{1(y_i = j)} = \theta_1^{n_1} \, \theta_2^{n_2} \cdots \theta_k^{n_k}, \tag{5.3}$$

where n_j is the number of y_is which take on the value j. This is the **multinomial likelihood**. Clearly the frequencies $n_1, \ldots, n_k$ form a sufficient statistic. Notice that $\sum_j n_j = n$.

In this case it is possible to work out the distribution of the sufficient statistic, i.e., the random frequency vector $N = (N_1, \ldots, N_k)$, where

$$N_j = \#\{i = 1, \ldots, n : Y_i = j\}, \qquad j = 1, \ldots, k.$$

Using combinatorial arguments it can be easily proven that

$$\begin{aligned} P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k \mid \theta_1, \theta_2, \ldots, \theta_k) \\ = \binom{n}{n_1, n_2, \cdots, n_k} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k}, \end{aligned} \tag{5.4}$$

when the integers $0 \leq n_1, \ldots, n_k \leq n$ and $\sum_j n_j = n$. Here

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!} \tag{5.5}$$

is called a **multinomial coefficient**. The multivariate discrete distribution with pmf (5.4) is called the **multinomial distribution** with sample size parameter n and probability vector parameter $(\theta_1, \ldots, \theta_k)$. The binomial distribution is a special case of the multinomial distribution: if $S \sim \text{Bin}(n, p)$, then the vector $(S, n - S)$ has the multinomial distribution with parameters n and $(p, 1 - p)$.

Notice that we can use the simple expression (5.3) for the likelihood of a multinomial observation even when we know very well that the pmf of the random vector $(N_1, \ldots, N_k)$ involves the multinomial coefficient.

5.2 Conjugate analysis

Some likelihoods have the property that if the prior is selected from a certain family of distributions $\mathcal{P}$, then the posterior also belongs to the same family $\mathcal{P}$. Such a family is called closed under sampling or a conjugate family (for the likelihood under consideration). A trivial and useless example of a conjugate family is provided by the set of all distributions. The useful conjugate families can be described by a finite number of hyperparameters, i.e., they are of the form

$$\mathcal{P} = \{\theta \mapsto f(\theta \mid \phi) : \phi \in S\}, \tag{5.6}$$

where S a set in an Euclidean space, and $\theta \mapsto f(\theta \mid \phi)$ is a density for each value of the hyperparameter vector $\phi \in S$. If the likelihood $p(y \mid \theta)$ admits this conjugate family, and if the prior $p(\theta)$ is $f(\theta \mid \phi_0)$ with a known value ϕ_0, then the posterior is of the form

$$\theta \mapsto p(\theta \mid y) = f(\theta \mid \phi_1),$$

where $\phi_1 \in S$. In order to find the posterior, we only need to find the value of the updated hyperparameter vector $\phi_1 = \phi_1(y)$.

If the densities $f(\theta \mid \phi)$ of the conjugate family have an easily understood form, then Bayesian inference is simple, provided we can approximate our prior

knowledge by some member $f(\theta \mid \phi_0)$ of the conjugate family and provided we know how to calculate the updated hyperparameters $\phi_1(y)$. However, nice conjugate families of the form (5.6) are possible only when the likelihood belongs to the exponential family, see, e.g., Schervish [5, Ch. 2].

The prior knowledge of the subject matter expert on θ is, unfortunately, usually rather vague. Transforming the subject matter expert's prior knowledge into a prior distribution is called **prior elicitation**. Supposing we are dealing with a scalar parameter, the expert might only have a feeling for the order of magnitude of the parameter, or might be able to say, which values would be surprisingly small or surprisingly large for the parameter. One approach for constructing the prior would then be to select from the family (5.6) some prior, which satisfies those kind of prior summaries.

As an example of conjugate analysis, consider the binomial likelihood (5.1) corresponding to sample size n and success probability θ. Recall that the beta density with (hyper)parameters $a, b > 0$ is given by

$$\mathrm{Be}(\theta \mid a, b) = \frac{1}{B(a,b)}\, \theta^{a-1}(1-\theta)^{b-1}, \qquad 0 < \theta < 1.$$

Suppose that the parameter θ has the beta prior $\mathrm{Be}(a, b)$ with known hyperparameters a and b. Then

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \\ &\propto \theta^s\, (1-\theta)^{n-s}\; \theta^{a-1}\, (1-\theta)^{b-1} \\ &\propto \mathrm{Be}(\theta \mid a+s, b+n-s), \qquad 0 < \theta < 1. \end{aligned}$$

Therefore we claim that the posterior is $\mathrm{Be}(a+s, b+n-s)$, where s is the number of successes (and $n-s$ is the number of failures). Notice the following points.

- We developed the posterior density, as a function of the parameter θ, dropping any constants (i.e., factors not involving θ).
- It is important to keep in mind, which is the variable we are interested in and what are the other variables, whose functions we treat as constants. The variable of interest is the one whose posterior distribution we want to calculate.
- We finished the calculation by recognizing that the posterior has a familiar functional form. In the present example we obtained a beta density except that it did not have the right normalizing constant. However, the only probability density on $0 < \theta < 1$ having the derived functional form is the beta density $\mathrm{Be}(\theta \mid a+s, b+n-s)$, and therefore the posterior distribution is this beta distribution.
- In more detail: from our calculations, we know that the posterior has the unnormalized density $\theta^{a+s-1}(1-\theta)^{b+n-s-1}$ on $0 < \theta < 1$. Since we know that the posterior density is a density on $(0, 1)$, we can find the normalizing constant by integration:

$$p(\theta \mid y) = \frac{1}{c(y)}\; \theta^{a+s-1}(1-\theta)^{b+n-s-1}, \qquad 0 < \theta < 1,$$

where

$$c(y) = \int_0^1 \theta^{a+s-1}(1-\theta)^{b+n-s-1}\,\mathrm{d}\theta = B(a+s, b+n-s),$$

where the last step is immediate, since the integral is the normalizing constant of the beta density $\mathrm{Be}(\theta \mid a_1, b_1)$, where $a_1 = a + s$ and $b_1 = b + n - s$. Therefore

$$p(\theta \mid y) = \mathrm{Be}(\theta \mid a+s, b+n-s).$$

- As soon as we have recognized the functional form of the posterior, we have recognized the posterior distribution.

5.3 More examples of conjugate analysis

5.3.1 Poisson likelihood and gamma prior

Suppose that

$$Y_i \mid \theta \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}(\theta), \qquad i = 1, \ldots, n,$$

which is shorthand notation for the statement that the RVs $Y_i, i = 1, \ldots, n$ are independently Poisson distributed with parameter θ. Then

$$p(y_i \mid \theta) = \frac{1}{y_i!}\,\theta^{y_i}\,\mathrm{e}^{-\theta}, \qquad y_i = 0, 1, 2, \ldots$$

and the likelihood is given by

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) \propto \theta^{\sum_1^n y_i}\,\mathrm{e}^{-n\theta}.$$

The likelihood has the functional form of a gamma density. If the prior for θ is the gamma distribution $\mathrm{Gam}(a, b)$ with known hyperparameters $a, b > 0$, i.e., if

$$p(\theta) = \frac{b^a}{\Gamma(a)}\,\theta^{a-1}\mathrm{e}^{-b\theta}, \qquad \theta > 0,$$

then

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \\ &\propto \theta^{\sum_1^n y_i}\mathrm{e}^{-n\theta}\theta^{a-1}\mathrm{e}^{-b\theta} \\ &\propto \theta^{a+\sum_1^n y_i - 1}\,\mathrm{e}^{-\theta(b+n)}, \qquad \theta > 0 \end{aligned}$$

and from this we recognize that the posterior is the gamma distribution

$$\mathrm{Gam}(a + \sum_1^n y_i, b + n).$$

5.3.2 Exponential likelihood and gamma prior

Suppose that

$$\begin{aligned} Y_i \mid \theta &\overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta), \qquad i = 1, \ldots, n \\ \Theta &\sim \text{Gam}(a, b), \end{aligned}$$

where $a, b > 0$ are known constants. Then

$$p(y_i \mid \theta) = \theta \, \mathrm{e}^{-\theta y_i}, \qquad y_i > 0,$$

and the likelihood is

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

We obtain $\text{Gam}(a + n, b + \sum y_i)$ as the posterior.

5.4 Conjugate analysis for normal observations

5.4.1 Normal likelihood when the variance is known

Suppose that we have one normally distributed observation $Y \sim N(\theta, \tau^2)$, where the mean θ is unknown but the variance τ^2 is a known value. Then

$$p(y \mid \theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\tau^2}\right).$$

Suppose that the prior is $N(\mu_0, \sigma_0^2)$ with known constants μ_0 and σ_0^2. Then the posterior is

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \\ &\propto \exp\left(-\frac{1}{2\tau^2}(y-\theta)^2 - \frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right) = \exp(L(\theta)), \end{aligned}$$

where $L(\theta)$ is a second degree polynomial in θ, and the coefficient of θ^2 in $L(\theta)$ is negative. Therefore the posterior is a normal distribution, but we need to calculate its mean μ_1 and variance σ_1^2.

Developing the density $N(\theta \mid \mu_1, \sigma_1^2)$ as a function of θ, we obtain

$$\begin{aligned} N(\theta \mid \mu_1, \sigma_1^2) &= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\sigma_1^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_1^2}\,\theta^2 + \frac{\mu_1}{\sigma_1^2}\,\theta\right) \end{aligned}$$

Next, we equate the coefficients of θ^2 and θ, firstly, in $L(\theta)$ and, secondly, in the previous formula to find out that we have

$$p(\theta \mid y) = N(\theta \mid \mu_1, \sigma_1^2),$$

where

$$\frac{1}{\sigma_1^2} = \frac{1}{\tau^2} + \frac{1}{\sigma_0^2}, \qquad \frac{\mu_1}{\sigma_1^2} = \frac{y}{\tau^2} + \frac{\mu_0}{\sigma_0^2}, \tag{5.7}$$

from which we can solve first σ_1^2 and then μ_1.

In Bayesian inference it is often convenient to parametrize the normal distribution by its mean and precision, where precision is defined as the reciprocal of the variance. We have just shown that the posterior precision equals the prior precision plus the datum precision.

If we have n independent observations $Y_i \sim N(\theta, \tau^2)$ with a known variance, then it is a simple matter to show that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is a sufficient statistic. In this case we know the distribution of the corresponding RV $\bar{Y}$ conditionally on θ,

$$[\bar{Y} \mid \theta] \sim N(\theta, \frac{\tau^2}{n}),$$

From these two facts we get immediately the posterior distribution from (5.7), when the prior is again $N(\mu_0, \sigma_0^2)$. (Alternatively, we may simply multiply the likelihood with the prior density, and examine the resulting expression.)

5.4.2 Normal likelihood when the mean is known

Suppose that the RVs Y_i are independently normally distributed,

$$Y_i \mid \theta \overset{\text{i.i.d.}}{\sim} N(\mu, \frac{1}{\theta}), \qquad i = 1, \ldots, n$$

where the mean μ is known but the variance $1/\theta$ is unknown. Notice that we parametrize the sampling distribution using the precision θ instead of the variance $1/\theta$. Then

$$p(y_i \mid \theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\theta(y_i - \mu)^2\right),$$

and the likelihood is

$$p(y \mid \theta) \propto \theta^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2\theta\right).$$

If the prior is $\mathrm{Gam}(a, b)$, then the posterior is evidently

$$\mathrm{Gam}(a + \frac{n}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2).$$

The previous result can be expressed also in terms of the variance $\phi = 1/\theta$. The variance has what is known as the inverse gamma distribution with density

$$\mathrm{Gam}(\frac{1}{\phi} \mid a_1, b_1)\,\frac{1}{\phi^2}, \qquad \phi > 0,$$

where a_1 and b_1 are the just obtained updated parameters, as can be established by the change of variable $\phi = 1/\theta$ in the posterior density. The inverse gamma distribution is also called the scaled inverse chi-square distribution, using a certain other convention for the parametrization.

5.4.3 Normal likelihood when the mean and the variance are unknown

Suppose that the RVs Y_i are independently normally distributed with unknown mean ϕ and unknown precision τ,

$$[Y_i \mid \phi, \tau] \overset{\text{i.i.d.}}{\sim} N(\phi, \frac{1}{\tau}), \qquad i = 1, \ldots, n.$$

In this case the likelihood for $\theta = (\phi, \tau)$ is conjugate for the prior of the form

$$p(\phi, \tau \mid a_0, b_0, \mu_0, n_0) = \text{Gam}(\tau \mid a_0, b_0)\, N(\phi \mid \mu_0, \frac{1}{n_0 \tau}).$$

Notice that the precision and the mean are dependent in this prior. This kind of a dependent prior may be natural in some problems but less natural in some other problems.

Often the interest centers on the mean ϕ while the precision τ is regarded as a nuisance parameter. The marginal posterior of ϕ (i.e., the marginal distribution of ϕ in the joint posterior) is obtained from the joint posterior by integrating out the nuisance parameter,

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

In the present case, this integral can be solved analytically, and the marginal posterior of ϕ can be shown to be a t-distribution.

5.4.4 Multivariate normal likelihood

When dealing with the multivariate instead of the univariate normal distribution, it is even more convenient to parametrize the normal distribution using the precision matrix, which is defined as the inverse of the covariance matrix, which we assume to be non-singular. Like the covariance matrix, also the precision matrix is a symmetric and and positive definite matrix.

The density of the multivariate normal $N_d(\mu, Q^{-1})$ with mean μ and precision matrix Q (i.e., of $N_d(\mu, \Sigma)$, where the covariance matrix $\Sigma = Q^{-1}$) is then given by

$$N_d(x \mid \mu, Q^{-1}) = (2\pi)^{-d/2} (\det Q)^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^T Q (x-\mu)\right)$$

where d is the dimensionality of x. Expanding the quadratic form, we get

$$(x-\mu)^T Q (x-\mu) = x^T Q x - x^T Q \mu - \mu^T Q x + \mu^T Q \mu$$

Now, the precision matrix is symmetric, and a scalar equal its transpose, so

$$\mu^T Q x = (\mu^T Q x)^T = x^T Q^T \mu = x^T Q \mu.$$

Therefore, as a function of x,

$$N_d(x \mid \mu, Q^{-1}) \propto \exp\left(-\frac{1}{2}(x^T Q x - 2 x^T Q \mu)\right). \tag{5.8}$$

Suppose that we have a single multivariate observation $Y \sim N(\theta, R^{-1})$, where the prior precision matrix R is known and suppose that the prior for the parameter vector θ is the normal distribution $N(\mu_0, Q_0^{-1})$ with known hyperparameters μ_0 and Q_0. Then

$$p(y \mid \theta) \propto \exp\left(-\frac{1}{2}(y-\theta)^T R(y-\theta)\right).$$

The prior is

$$p(\theta) \propto \exp\left(-\frac{1}{2}(\theta-\mu_0)^T Q_0(\theta-\mu_0)\right).$$

The posterior is proportional to their product,

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}(\theta-y)^T R(\theta-y) - \frac{1}{2}(\theta-\mu_0)^T Q_0(\theta-\mu_0).\right)$$

Here we have

$$\begin{aligned}
&(\theta-y)^T R(\theta-y) + (\theta-\mu_0)^T Q_0(\theta-\mu_0) \\
&= \theta^T R\theta - 2\theta^T Ry + y^T Ry + \theta^T Q_0\theta - 2\theta^T Q_0\mu_0 + \mu_0^T R\mu_0 \\
&= \theta^T(R+Q_0)\theta - 2\theta^T(Ry + Q_0\mu_0) + c,
\end{aligned}$$

where the scalar c does not depend on θ. Comparing this result with (5.8), we see that the posterior is the multivariate normal $N_d(\mu_1, Q_1^{-1})$, where

$$Q_1 = Q_0 + R, \qquad Q_1\,\mu_1 = Q_0\,\mu_0 + Ry. \tag{5.9}$$

Again, posterior precision equals the prior precision plus the datum precision.

As in the univariate case, this result can be extended to several (conditionally) independent observations, and also to the case where both the mean vector and the precision matrix are (partially) unknown, when we employ an appropriate conjugate prior.

5.5 Conditional conjugacy

In multiparameter problems it may be difficult or impossible to use conjugate priors. However, some benefits of conjugate families can be retained, if one has conditional conjugacy in the Bayesian statistical model.

Suppose we have parameter vector θ, which we partition as $\theta = (\phi, \psi)$, where the components ϕ and ψ are not necessarily scalars. The the **full conditional** (density) of ϕ in the prior distribution is defined as

$$p(\phi \mid \psi),$$

and the full conditional (density) of of ϕ in the posterior is defined as

$$p(\phi \mid \psi, y).$$

Then ϕ exhibits conditional conjugacy, if the full conditional of ϕ in the prior and and in the posterior belong to the same family of distributions.

In practice, one notices the conditional conjugacy of ϕ as follows. The prior full conditional of ϕ is

$$p(\phi \mid \psi) \propto p(\phi, \psi),$$

when we regard the joint prior as a function of ϕ. Similarly, the posterior full conditional of ϕ is

$$p(\phi \mid \psi, y) \propto p(\phi, \psi, y) = p(\phi, \psi)\, p(y \mid \phi, \psi),$$

when we regard the joint distribution $p(\phi, \psi, y)$ as a function of ϕ. If we recognize the functional forms of the prior full conditional and the posterior full conditional, then we have conditional conjugacy.

If we partition the parameter vector into k components, $\theta = (\theta_1, \ldots, \theta_k)$ (which are not necessarily scalars), then sometimes all the components are conditionally conjugate. In other cases, only some of the components turn out to be conditionally conjugate.

5.6 Reparametrization

Suppose that we have formulated a Bayesian statistical model in terms of a parameter vector θ with a continuous distribution, but then want to reformulate it in terms of a new parameter vector ϕ, where there is a diffeomorphic correspondence between θ and ϕ. I.e., the correspondence

$$\phi = g(\theta) \quad \Leftrightarrow \quad \theta = h(\phi)$$

is one-to-one and continuously differentiable in both directions. What happens to the prior, likelihood and the posterior under such a reparametrization?

We get the prior of ϕ using the change of variables formula for densities:

$$f_\Phi(\phi) = f_\Theta(\theta) \left| \frac{\partial \theta}{\partial \phi} \right| = f_\Theta(h(\phi))\, |J_h(\phi)|.$$

If we know ϕ then we also know $\theta = h(\phi)$. Therefore the likelihood stays the same in that

$$f_{Y|\Phi}(y \mid \phi) = f_{Y|\Theta}(y \mid h(\phi)).$$

Finally, the posterior density changes in the same way as the prior density (by the change of variables formula), i.e.,

$$f_{\Phi|Y}(\phi \mid y) = f_{\Theta|Y}(\theta \mid y) \left| \frac{\partial \theta}{\partial \phi} \right| = f_{\Theta|Y}(h(\phi) \mid y)\, |J_h(\phi)|.$$

5.7 Improper priors

Sometimes one specifies a prior by stating that

$$p(\theta) \propto h(\theta),$$

where $h(\theta)$ is a non-negative function, whose integral is infinite

$$\int h(\theta)\, \mathrm{d}\theta = \infty.$$

Then there does not exist a constant of proportionality that will allow $p(\theta)$ to be a proper density, i.e., to integrate to one. In that case we have an **improper prior**. Notice that this is different from expressing the prior by the means of an unnormalized density h, which can be normalized to be a proper density. Sometimes we get a proper posterior, if we multiply an improper prior with the likelihood and then normalize.

For example, consider one normally distributed observation $Y \sim N(\theta, \tau^2)$ with a known variance τ^2, and take

$$p(\theta) \propto 1, \qquad \theta \in \mathbb{R}\,.$$

This prior is intended to represent complete prior ignorance about the unknown mean: all possible values are deemed equally likely. Calculating formally,

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \propto \exp\left(-\frac{1}{2\tau^2}(y-\theta)^2\right) \\ &\propto N(\theta \mid y, \tau^2) \end{aligned}$$

We obtain the same result in the limit, if we take $N(\mu_0, \sigma_0^2)$ as the prior and then let the prior variance σ_0^2 go to infinity.

One often uses improper priors in a location-scale model, with a location parameter μ and a scale parameter σ. Then it is conventional to take the prior of the location parameter to be uniform and to let the logarithm of the scale parameter σ have a uniform distribution and to take them to be independent in their improper prior. This translates to an improper prior of the form

$$p(\mu, \sigma) \propto \frac{1}{\sigma}, \qquad \mu \in \mathbb{R}, \sigma > 0 \tag{5.10}$$

by using (formally) the change of variables formula,

$$p(\sigma) = p(\tau) \left| \frac{\mathrm{d}\tau}{\mathrm{d}\sigma} \right| \propto \frac{1}{\sigma},$$

when $\tau = \log \sigma$ and $p(\tau) \propto 1$.

Some people use the so called Jeffreys' prior, which is designed to have a form which is invariant with respect to one-to-one reparametrizations. Also this leads typically to an improper prior. There are also other processes which attempt produce non-informative priors, which often turn out to be improper. (A prior is called non-informative, vague, diffuse or flat, if it plays a minimal role in the posterior distribution.)

Whereas the posterior derived from a proper prior is automatically proper, the posterior derived from an improper prior can be either proper or improper. Notice, however, that **an improper posterior does not make any sense**. If you do use an improper prior, it is *your* duty to check that the posterior is proper.

5.8 Summarizing the posterior

The posterior distribution gives a complete description of the uncertainty concerning the parameter after the data has been observed. If we use conjugate

analysis inside a well-understood conjugate family, then we need only report the hyperparameters of the posterior. E.g., if the posterior is a multivariate normal (and the dimensionality is low) then the best summary is to give the mean and the covariance matrix of the posterior. However, in more complicated situations the functional form of the posterior may be so opaque that we need to summarize the posterior.

If we have a univariate parameter, then the best description of the posterior is the plot of its density function. Additionally, we might want to calculate such summaries as the posterior mean the posterior variance, the posterior mode, the posterior median, and other selected posterior quantiles. If we cannot plot the density, but are able to simulate from the posterior, we can plot the histogram and calculate summaries (mean, variance, quantiles) from the simulated sample.

If we have a two-dimensional parameter, then we can still make contour plots or perspective plots of the density, but in higher dimensions such plots are not possible. One practical approach in a multiparameter situation is to summarize the one-dimensional marginal posteriors of the scalar components of the parameter.

Suppose that (after a rearrangement of the components) $\theta = (\phi, \psi)$, where ϕ is the scalar component of interest. Then the marginal posterior of ϕ is

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

The indicated integration may be very difficult to perform analytically. However, if one has available a sample

$$(\phi_1, \psi_1), (\phi_2, \psi_2), ..., (\phi_N, \psi_N)$$

from the posterior of $\theta = (\phi, \psi)$, then $\phi_1, \phi_2, \ldots, \phi_N$ is a sample from the marginal posterior of ϕ. Hence we can summarize the marginal posterior of ϕ based on the sample $\phi_1, \ldots, \phi_N$.

5.9 Posterior intervals

We consider a univariate parameter θ which has a continuous distribution. One conventional summary of the posterior is a $100(1-\alpha)\%$ **posterior interval** of the parameter θ, which is any interval C in the parameter space such that

$$P(\Theta \in C \mid Y = y) = \int_C p(\theta \mid y)\, \mathrm{d}\theta = 1 - \alpha. \tag{5.11}$$

Some authors call such intervals **credible intervals** (or credibility intervals) and others may call them **Bayesian confidence intervals**.

The posterior intervals have the direct probabilistic interpretation (5.11). In contrast, the confidence intervals of frequentist statistics have probability interpretations only with reference to (hypothetical) sampling of the data under identical conditions.

Within the frequentist framework, the parameter is an unknown deterministic quantity. A frequentist confidence interval either covers or does not cover the true parameter value. A frequentist statistician constructs a frequentist confidence interval at significance level $\alpha 100\%$ in such a way that if it were possible

to sample repeatedly the data under identical conditions (i.e., using the same value for the parameter), then the relative frequency of coverage in a long run of repetitions would be about $1 - \alpha$. But for the data at hand, the calculated frequentist confidence interval still either covers or does not cover the true parameter value, and we do not have guarantees for anything more. Many naive users of statistics (and even some textbook authors) mistakenly believe that their frequentist confidence intervals have the simple probability interpretation belonging to posterior intervals.

The coverage requirement (5.11) does not by itself determine any interval in the parameter space but needs to be supplemented by other criteria. In practice it is easiest to use the **equal tail (area) interval** (or central interval), whose end points are selected so that $\alpha/2$ of the posterior probability lies to the left and $\alpha/2$ to the right of the intervals. By the definition of the quantile function, see (2.5) and (2.6), the equal tail posterior interval is given by

$$[q(\alpha/2), q(1 - \alpha/2)]. \tag{5.12}$$

If the quantile function is not available, but one has available a sample $\theta_1, \ldots, \theta_N$ from the posterior, then one can use the empirical quantiles calculated from the sample.

Many authors recommend the **highest posterior density (HPD)** region, which is defined as the set

$$C_t = \{\theta : f_{\Theta|Y}(\theta \mid y) \geq t\},$$

where the threshold t has to be selected so that

$$P(\Theta \in C_t) = 1 - \alpha.$$

Often (but not always) the HPD region turns out to be an interval. Then it can proven to be the shortest interval with the desired coverage $100(1 - \alpha)\%$. However, calculating a HPD interval is more difficult than calculating an equal tail interval.

In a multiparameter situation one usually examines one parameter at a time. Let ϕ be the scalar parameter of interest in $\theta = (\phi, \psi)$, and suppose that we have available a sample

$$(\phi_1, \psi_1), (\phi_2, \psi_2), ..., (\phi_N, \psi_N)$$

from the posterior. Then $\phi_1, \phi_2, \ldots, \phi_N$ is a sample from the marginal posterior of ϕ. Hence the central marginal posterior interval of ϕ can be calculated as in (5.12), when q is the empirical quantile function based on $\phi_1, \ldots, \phi_N$.

5.10 Literature

See, e.g., Bernardo and Smith [1] for further results on conjugate analysis. The books by Gelman *et al.* [3], Carlin and Louis [2] and O'Hagan and Forster [4] are rich sources of ideas on Bayesian modeling and analysis. Sufficiency is a central concept in parametric statistics. See, e.g., Schervish [5] for a discussion.

Bibliography

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

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

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, 2nd edition, 2004.

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

[5] Mark J. Schervish. *Theory of Statistics*. Springer series in statistics. Springer-Verlag, 1995.

Chapter 6

Approximations

6.1 The grid method

When one is confronted with a low-dimensional problem with a continuous parameter, then it is usually easy to approximate the posterior density at a dense grid of points that covers the relevant part of the parameter space. We discuss the method for a one-dimensional parameter θ.

We suppose that the posterior is available in the unnormalized form

$$f_{\Theta|Y}(\theta \mid y) = \frac{1}{c(y)}\, q(\theta \mid y),$$

where we know how to evaluate the unnormalized density $q(\theta \mid y)$, but do not necessarily know the value of the normalizing constant $c(y)$.

Instead of the original parameter space, we consider a finite interval $[a, b]$, which should cover most of the mass of the posterior distribution. We divide $[a, b]$ evenly into N subintervals

$$B_i = [a + (i-1)h, a + ih], \qquad i = 1, \ldots, N.$$

The width h of one subinterval is

$$h = \frac{b-a}{N}.$$

Let θ_i be the midpoint of the i'th subinterval,

$$\theta_i = a + (i - \frac{1}{2})h, \qquad i = 1, \ldots, N.$$

We use the midpoint rule for numerical integration. This means that we approximate the integral over the i'th subinterval of any function g by the rule

$$\int_{B_i} g(\theta)\, \mathrm{d}\theta \approx h g(\theta_i).$$

Using the midpoint rule on each of the subintervals, we get the following

approximation for the normalizing constant

$$\begin{aligned} c(y) &= \int q(\theta \mid y)\,\mathrm{d}\theta \approx \int_a^b q(\theta \mid y)\,\mathrm{d}\theta = \sum_{i=1}^{N} \int_{B_i} q(\theta \mid y)\,\mathrm{d}\theta \\ &\approx h \sum_{i=1}^{N} q(\theta_i \mid y) \end{aligned}$$

Using this approximation, we can approximate the value of the posterior density at the point θ_i,

$$f_{\Theta|Y}(\theta_i \mid y) = \frac{1}{c(y)}\, q(\theta_i \mid y) \approx \frac{1}{h} \frac{q(\theta_i \mid y)}{\sum_{j=1}^{N} q(\theta_j \mid y)}. \tag{6.1}$$

We also obtain approximations for the posterior probabilities of the subintervals,

$$\begin{aligned} P(\Theta \in B_i \mid Y = y) &= \int_{B_i} f_{\Theta|Y}(\theta \mid y)\,\mathrm{d}\theta \approx h f_{\Theta|Y}(\theta_i \mid y) \\ &\approx \frac{q(\theta_i \mid y)}{\sum_{j=1}^{N} q(\theta_j \mid y)}. \end{aligned} \tag{6.2}$$

These approximations can be surprisingly accurate even for moderate values of N provided we are able to identify an interval $[a, b]$, which covers the essential part of posterior distribution. The previous formulas give means for plotting the posterior density and simulating from it. This is the grid method for approximating or simulating the posterior distribution.

- First evaluate the unnormalized posterior density $q(\theta \mid y)$ at a regular grid of points $\theta_1, \ldots, \theta_N$ with spacing h. The grid should cover the main support of the posterior density.
- If you want to plot the posterior density, normalize these values by dividing by their sum and additionally by the bin width h as in eq. (6.1). This gives an approximation to the posterior ordinates $p(\theta_i \mid y)$ at the grid points θ_i.
- If you want a sample from the posterior, sample with replacement from the grid points θ_i with probabilities proportional to the numbers $q(\theta_i \mid y)$, cf. (6.2).

The midpoint rule is considered a rather crude method of numerical integration. In the numerical analysis literature, there are available much more sophisticated methods of numerical integration (or numerical quadrature) and they can be used in a similar manner. Besides dimension one, these kinds of approaches can be used in dimensions two or three. However, as the dimensionality of the parameter space grows, computing at every point in a dense multidimensional grid becomes more and more expensive.

6.2 Normal approximation to the posterior

We can try to approximate a multivariate posterior density by a multivariate normal density based on the behavior of the posterior density at its mode.

This approximation can be quite accurate, when the sample sizes is large, and when the posterior is unimodal. We will call the resulting approximation a normal approximation to the posterior, but the result is sometimes also called Laplace approximation or modal approximation. A normal approximation can be used directly as an approximate description of the posterior. However, such an approximation can be utilized also indirectly, e.g., to form a good proposal distribution for the Metropolis–Hastings method.

We first discuss normal approximation in the univariate situation. The statistical model has a single parameter θ, which has a continuous distribution. Let the unnormalized posterior density be given by $q(\theta \mid y)$. The normalizing constant of the posterior density can be unknown. We consider the case, where $\theta \mapsto q(\theta \mid y)$ is unimodal: i.e., it has only one local maximum. We suppose that we have located the mode $\hat{\theta}$ of $q(\theta \mid y)$. Actually, $\hat{\theta}$ depends on the data y, but we suppress this dependence in our notation. Usually we would have to run some numerical optimization algorithm in order to find the mode.

The basic idea of the method is to use the second degree Taylor polynomial of the logarithm of the posterior density centered on the mode $\hat{\theta}$,

$$\log f_{\Theta|Y}(\theta \mid y) \approx \log f_{\Theta|Y}(\hat{\theta} \mid y) + b(\theta - \hat{\theta}) - \frac{1}{2} A(\theta - \hat{\theta})^2, \tag{6.3}$$

where

$$b = \frac{\partial}{\partial \theta} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = \frac{\partial}{\partial \theta} \log q(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = 0,$$

and

$$A = -\frac{\partial^2}{\partial \theta^2} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = -\frac{\partial^2}{\partial \theta^2} \log q(\theta \mid y)\bigg|_{\theta=\hat{\theta}}.$$

Notice the following points.

- The first and higher order (partial) derivatives with respect to θ of $\log q(\theta \mid y)$ and $\log f_{\Theta|Y}(\theta \mid y)$ agree, since these function differ only by an additive constant (which depends on y but not on θ).
- The first order term of the Taylor expansion disappears, since $\hat{\theta}$ is also the mode of $\log f_{\Theta|Y}(\theta \mid y)$.
- $A \geq 0$, since $\hat{\theta}$ is a maximum of $q(\theta \mid y)$. For the following, we need to assume that $A > 0$.

Taking the exponential of the second degree Taylor approximation (6.3), we see that we may approximate the posterior by the function

$$\pi_{\text{approx}}(\theta) \propto \exp\left(-\frac{A}{2}(\theta - \hat{\theta})^2\right),$$

at least in the vicinity of the mode $\hat{\theta}$. Luckily, we recognize that $\pi_{\text{approx}}(\theta)$ is an unnormalized form of the density of the normal distribution with mean $\hat{\theta}$ and variance $1/A$. The end result is that the posterior distribution can be approximated with the normal distribution

$$N\left(\hat{\theta}, \frac{1}{-L''(\hat{\theta})}\right), \tag{6.4}$$

where

$$L(\theta) = \log q(\theta \mid y)$$

and $L''(\hat{\theta})$ is the second derivative of $L(\theta)$ evaluated at the mode $\hat{\theta}$.

The multivariate analog of the result starts with the second degree expansion of the log-posterior centered on the mode $\hat{\theta}$,

$$\log f_{\Theta|Y}(\theta \mid y) \approx \log f_{\Theta|Y}(\hat{\theta} \mid y) + 0 - \frac{1}{2}(\theta - \hat{\theta})^T A(\theta - \hat{\theta}),$$

where A is minus the Hessian matrix of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode,

$$A_{ij} = -\left.\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f_{\Theta|Y}(\theta \mid y)\right|_{\theta=\hat{\theta}} = -\left.\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(\theta)\right|_{\theta=\hat{\theta}} = -\left[\frac{\partial^2}{\partial\theta\, \partial\theta^T} L(\theta)\right]_{ij}$$

The first degree term of the expansion vanishes, since $\hat{\theta}$ is the mode of the log-posterior. Here A is at least positively semidefinite, since $\hat{\theta}$ is a maximum. If A is positively definite, we can proceed with the normal approximation.

Exponentiating, we find out that approximately (at least in the vicinity of the mode)

$$f_{\Theta|Y}(\theta \mid y) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T A(\theta - \hat{\theta})\right).$$

Therefore we can approximate the posterior with the corresponding multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix given by A^{-1}, i.e., the approximating normal distribution is

$$N\left(\hat{\theta}, \left(-L''(\hat{\theta})\right)^{-1}\right), \tag{6.5}$$

where $L''(\hat{\theta})$ is the Hessian matrix of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode $\hat{\theta}$. The precision matrix of the approximating normal distribution is minus the Hessian of the log-posterior (evaluated at the mode), and hence the covariance matrix is minus the inverse of the Hessian.

If the (unnormalized) posterior has K modes $\hat{\theta}_1, \ldots, \hat{\theta}_K$, which are well separated, then Gelman *et al.* [1, Ch. 12.2] propose that the posterior could be approximated by the normal mixture

$$\frac{1}{C}\sum_{k=1}^{K} q(\hat{\theta}_k \mid y) \exp\left(-\frac{1}{2}(\theta - \hat{\theta}_k)^T[-L''(\hat{\theta}_k)](\theta - \hat{\theta}_k)\right). \tag{6.6}$$

This approximation is reasonable, if

$$j \neq k \quad \Rightarrow \quad \exp\left(-\frac{1}{2}(\hat{\theta}_j - \hat{\theta}_k)^T[-L''(\hat{\theta}_k)](\hat{\theta}_j - \hat{\theta}_k)\right) \approx 0$$

The normalizing constant in the normal mixture approximation (6.6) is

$$C = \sum_{k=1}^{K} q(\hat{\theta}_k \mid y)\, (2\pi)^{d/2} (\det(-L''(\hat{\theta}_k))^{-1/2},$$

where d is the dimensionality of θ.

Before using the normal approximation, it is often advisable to reparameterize the model so that the transformed parameters are defined on the whole real line and have roughly symmetric distributions. E.g., one can use logarithms of positive parameters and apply the logit function to parameters which take values on the interval $(0, 1)$. The normal approximation is then constructed for the transformed parameters, and the approximation can then be translated back to the original parameter space. One must, however, remember to multiply by the appropriate Jacobians.

Example 6.1. We consider the unnormalized posterior

$$q(\theta \mid y) = \theta^{y_4} \, (1-\theta)^{y_2+y_3} \, (2+\theta)^{y_1}, \qquad 0 < \theta < 1,$$

where $y = (y_1, y_2, y_3, y_4) = (13, 1, 2, 3)$. The mode and the second derivative of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode are given by

$$\hat{\theta} \approx 0.677, \qquad L''(\hat{\theta}) \approx -37.113.$$

(The mode $\hat{\theta}$ can be found by solving a quadratic equation.) The resulting normal approximation in the original parameter space is $N(0.677, 1/37.113)$.

We next reparametrize by defining ϕ as the logit of θ,

$$\phi = \operatorname{logit}(\theta) = \ln \frac{\theta}{1-\theta} \quad \Leftrightarrow \quad \theta = \frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}}.$$

The given unnormalized posterior for θ transforms to the following unnormalized posterior for ϕ,

$$\begin{aligned} \tilde{q}(\phi \mid y) &= q(\theta \mid y) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right| \\ &= \left(\frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \right)^{y_4} \left(\frac{1}{1+\mathrm{e}^{\phi}} \right)^{y_2+y_3} \left(\frac{2+3\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \right)^{y_1} \frac{\mathrm{e}^{\phi}}{(1+\mathrm{e}^{\phi})^2}. \end{aligned}$$

The mode and the second derivative of $\tilde{L}(\phi) = \log \tilde{q}(\phi \mid y)$ evaluated at the mode are given by

$$\hat{\phi} \approx 0.582, \qquad \tilde{L}''(\hat{\phi}) \approx -2.259.$$

(Also $\hat{\phi}$ can be found by solving a quadratic.) This results in the normal approximation $N(0.582, 1/2.259)$ for the logit of θ.

When we translate that approximation back to the original parameter space, we get the approximation

$$f_{\Theta|Y}(\theta \mid y) \approx N(\phi \mid 0.582, 1/2.259) \left| \frac{\mathrm{d}\phi}{\mathrm{d}\theta} \right|,$$

i.e.,

$$f_{\Theta|Y}(\theta \mid y) \approx N(\operatorname{logit}(\theta) \mid 0.582, 1/2.259) \, \frac{1}{\theta(1-\theta)}.$$

Both of these approximations are plotted in Figure 6.1 together with the true posterior density (whose normalizing constant can be found exactly using computer algebra). △

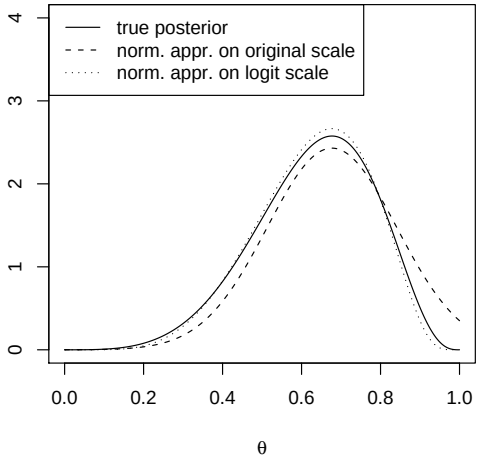


Figure 6.1: The exact posterior density (solid line) together with its normal approximation (dashed line) and the approximation based on the normal approximation for the logit of θ. The last approximation is markedly non-normal on the original scale, and it is able to capture the skewness of the true posterior density.

6.3 Posterior expectations using Laplace approximation

Laplace showed in the 1770's how one can form approximations to integrals of highly peaked positive functions by integrating analytically a suitable normal approximation. We will now apply this idea to build approximations to posterior expectations. We assume that the posterior density is highly peaked while the function h, whose posterior expectation we seek is relatively flat. To complicate matters, the posterior density is typically known only in the unnormalized form $q(\theta \mid y)$, and then

$$E[h(\Theta) \mid Y = y] = \frac{\int h(\theta)\, q(\theta \mid y)\, \mathrm{d}\theta}{\int q(\theta \mid y)\, \mathrm{d}\theta}. \tag{6.7}$$

Tierney and Kadane [5] approximated separately the numerator and the denominator of eq. (6.7) using Laplace's method, and analyzed the resulting error.

To introduce the idea of Laplace's approximation (or Laplace's method), consider a highly peaked function $L(\theta)$ of a scalar variable θ such that $L(\theta)$ has a unique mode (i.e., a maximum) at $\hat{\theta}$. Suppose that $g(\theta)$ is a function, which varies slowly. We seek an approximation to the integral

$$I = \int g(\theta)\, \mathrm{e}^{L(\theta)}\, \mathrm{d}\theta. \tag{6.8}$$

Heuristically, the integrand is negligible when we go far away from $\hat{\theta}$, and so we should be able to approximate the integral I by a simpler integral, where we

take into account only the local behavior of $L(\theta)$ around its mode. To this end, we first approximate $L(\theta)$ by its second degree Taylor polynomial centered at the mode $\hat{\theta}$,

$$L(\theta) \approx L(\hat{\theta}) + 0 \cdot (\theta - \hat{\theta}) + \frac{1}{2} L''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Since $g(\theta)$ is slowly varying, we may approximate the integrand as follows

$$g(\theta)\, \mathrm{e}^{L(\theta)} \approx g(\hat{\theta})\, \exp\left(L(\hat{\theta}) - \frac{1}{2}\tau^2(\theta - \hat{\theta})^2) \right),$$

where

$$\tau^2 = -L''(\hat{\theta}).$$

For the following, we must assume that $L''(\hat{\theta}) < 0$. Integrating the approximation, we obtain

$$\begin{aligned} I &\approx \int g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}\, \exp(-\frac{1}{2}\tau^2(\theta - \hat{\theta})^2)\, \mathrm{d}\theta \\ &= \frac{\sqrt{2\pi}}{\tau}\, g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})} \end{aligned}$$

This is Laplace's approximation. (Actually, it is only the leading term in a Laplace expansion, which is an asymptotic expansion for the integral.)

To handle the multivariate result, we use the normalizing constant of the $N_d(\mu, Q^{-1})$ distribution to evaluate the integral

$$\int \exp\left(-\frac{1}{2}(x - \mu)^T Q (x - \mu) \right)\, \mathrm{d}x = \frac{(2\pi)^{d/2}}{\sqrt{\det Q}}. \tag{6.9}$$

This result is valid for any symmetric and positive definite $d \times d$ matrix Q. Integrating the multivariate second degree approximation of $g(\theta)\, \exp(L(\theta))$, we obtain

$$I = \int g(\theta)\, \mathrm{e}^{L(\theta)}\, \mathrm{d}\theta \approx \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}, \tag{6.10}$$

where d is the dimensionality of θ, and Q is minus the Hessian of L evaluated at the mode,

$$Q = -L''(\hat{\theta}),$$

and we must assume that the $d \times d$ matrix Q is positively definite.

Using these tools, we can approximate the posterior expectation (6.7) in several different ways. One idea is to approximate the numerator by choosing

$$g(\theta) = h(\theta), \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y)$$

in eq. (6.10), and then to approximate the denominator by choosing

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y).$$

These choices yield the approximation

$$E[h(\Theta) \mid Y = y] \approx \frac{\frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, h(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}}{\frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, \mathrm{e}^{L(\hat{\theta})}} = h(\hat{\theta}), \tag{6.11}$$

where

$$\hat{\theta} = \arg\max L(\theta), \qquad Q = -L''(\hat{\theta}).$$

Here we need a single maximization, and do not need to evaluate the Hessian at all.

A less obvious approach is to choose

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = h(\theta)\, q(\theta \mid y)$$

to approximate the numerator, and

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y)$$

to approximate the denominator. Here we need to assume that h is a positive function, i.e., $h > 0$. The resulting approximation is

$$E[h(\Theta) \mid Y = y] \approx \left(\frac{\det(Q)}{\det(Q^*)}\right)^{1/2} \frac{h(\hat{\theta}^*)\, q(\hat{\theta}^* \mid y)}{q(\hat{\theta} \mid y)}, \tag{6.12}$$

where

$$\hat{\theta}^* = \arg\max[h(\theta)\, q(\theta \mid y)], \qquad \hat{\theta} = \arg\max q(\theta \mid y).$$

and Q^* and Q are the minus Hessians

$$Q^* = -L^{*''}(\hat{\theta}^*), \qquad Q = -L''(\hat{\theta}),$$

where

$$L^*(\theta) = \log(h(\theta)\, q(\theta \mid y)), \qquad L(\theta) = \log q(\theta \mid y).$$

We need two separate maximizations and need to evaluate two Hessians for this approximation.

Tierney and Kadane analyzed the errors committed in these approximations in the situation, where we have n (conditionally) i.i.d. observations, and the sample size n grows. The first approximation (6.11) has relative error of order $O(n^{-1})$, while the second approximation (6.12) has relative error of order $O(n^{-2})$. That is,

$$E[h(\Theta) \mid Y = y] = h(\hat{\theta})\, \left(1 + O(n^{-1})\right)$$

and

$$E[h(\Theta) \mid Y = y] = \left(\frac{\det(Q)}{\det(Q^*)}\right)^{1/2} \frac{h(\hat{\theta}^*)\, q(\hat{\theta}^* \mid y)}{q(\hat{\theta} \mid y)}\, \left(1 + O(n^{-2})\right).$$

Hence the second approximation is much more accurate (at least asymptotically).

6.4 Posterior marginals using Laplace approximation

Tierney and Kadane discuss also an approximation to the marginal posterior, when the parameter vector θ is composed of two vector components $\theta = (\phi, \psi)$.

The form of the approximation is easy to derive, and was earlier discussed by Leonard [2]. However, Tierney and Kadane [5, Sec. 4] were the first to analyze the error in this Laplace approximation. We first derive the form of the approximation, and then make some comments on the error terms based on the discussion of Tierney and Kadane.

Let $q(\phi, \psi \mid y)$ be an unnormalized form of the posterior density, based on which we try to approximate the normalized marginal posterior $p(\phi \mid y)$. Let the dimensions of ϕ and ψ be d_1 and d_2, respectively. We have

$$p(\phi \mid y) = \int p(\phi, \psi \mid y)\, \mathrm{d}\psi = \int \exp(\log p(\phi, \psi \mid y))\, \mathrm{d}\psi,$$

where $p(\phi, \psi \mid y)$ is the normalized posterior. The main difference with approximating a posterior expectation is the fact, that now we are integrating only over the component(s) ψ of $\theta = (\phi, \psi)$.

Fix the value of ϕ for the moment. Let $\psi^*(\phi)$ be the maximizer of the function

$$\psi \mapsto \log p(\phi, \psi \mid y),$$

and let $Q(\phi)$ be minus the Hessian matrix of this function evaluated at $\psi = \psi^*(\phi)$. Notice that we can equally well calculate $\psi^*(\phi)$ and $Q(\phi)$ as the maximizer and minus the $d_2 \times d_2$ Hessian matrix of $\psi \mapsto \log q(\phi, \psi \mid y)$, respectively,

$$\psi^*(\phi) = \arg\max_{\psi} (\log q(\phi, \psi \mid y)) = \arg\max_{\psi} q(\phi, \psi \mid y) \tag{6.13}$$

$$Q(\phi) = -\left[\frac{\partial^2}{\partial \psi\, \partial \psi^T} \log q(\phi, \psi \mid y) \right]_{|\psi = \psi^*(\phi)}. \tag{6.14}$$

For fixed ϕ, we have the second degree Taylor approximation in ψ,

$$\log p(\phi, \psi \mid y) \approx \log p(\phi, \psi^*(\phi) \mid y) - \frac{1}{2}(\psi - \psi^*(\phi))^T Q(\phi)(\psi - \psi^*(\phi)), \tag{6.15}$$

and we assume that matrix $Q(\phi)$ is positive definite.

Next we integrate the exponential function of the approximation (6.15) with respect to ψ, with the result

$$p(\phi \mid y) \approx p(\phi, \psi^*(\phi) \mid y)\, (2\pi)^{d_2/2}\, (\det Q(\phi))^{-1/2}.$$

To evaluate this approximation, we need the normalizing constant of the unnormalized posterior $q(\phi, \psi \mid y)$, which we obtain by another Laplace approximation, and the end result is

$$p(\phi \mid y) \approx (2\pi)^{-d_1/2}\, q(\phi, \psi^*(\phi) \mid y) \sqrt{\frac{\det Q}{\det Q(\phi)}}, \tag{6.16}$$

where Q is minus the $(d_1 + d_2) \times (d_1 + d_2)$ Hessian of the function

$$(\phi, \psi) \mapsto \log q(\phi, \psi \mid y)$$

evaluated at the MAP, the maximum point of the same function. However, it is often enough to approximate the functional form of the marginal posterior. When considered as a function of ϕ, we have, approximately,

$$p(\phi \mid y) \propto q(\phi, \psi^*(\phi) \mid y)\, (\det Q(\phi))^{-1/2}. \tag{6.17}$$

The unnormalized Laplace approximation (6.17) can be given another interpretation (see, e.g., [3, 4]). By the multiplication rule,

$$p(\phi \mid y) = \frac{p(\phi, \psi \mid y)}{p(\psi \mid \phi, y)} \propto \frac{q(\phi, \psi \mid y)}{p(\psi \mid \phi, y)}.$$

This result is valid for any choice of ψ. Let us now form a normal approximation for the denominator for a fixed value of ϕ, i.e.,

$$p(\psi \mid \phi, y) \approx N(\psi \mid \psi^*(\phi), Q(\phi)^{-1}).$$

However, this approximation is accurate only in the vicinity of the mode $\psi^*(\phi)$, so let us use it only at the mode. The end result is the following approximation,

$$\begin{aligned} p(\phi \mid y) &\propto \left[\frac{q(\phi, \psi \mid y)}{N(\psi \mid \psi^*(\phi), Q(\phi)^{-1})} \right]_{|\psi=\psi^*(\phi)} \\ &= (2\pi)^{d_2/2} \det(Q(\phi))^{-1/2} \, q(\phi, \psi^*(\phi) \mid y) \\ &\propto q(\phi, \psi^*(\phi) \mid y) \, (\det Q(\phi))^{-1/2}, \end{aligned}$$

which is the same as the unnormalized Laplace approximation (6.17) to the marginal posterior of ϕ.

Tierney and Kadane show that the relative error in the approximation (6.16) is of the order $O(n^{-1})$, when we have n (conditionally) i.i.d. observations, and that most of the error comes from approximating the normalizing constant. They argue that the approximation (6.17) captures the correct functional form of the marginal posterior with relative error $O(n^{-3/2})$ and recommend that one should therefore use the unnormalized approximation (6.17), which can then be normalized by numerical integration, if need be. For instance, if we want to simulate from the approximate marginal posterior, then we can use the unnormalized approximation (6.17) directly, together with accept–reject, SIR or the grid-based simulation method of Sec. 6.1. See the articles by H. Rue and coworkers [3, 4] for imaginative applications of these ideas.

Another possibility for approximating the marginal posterior would be to build a normal approximation to the joint posterior, and then marginalize. However, a normal approximation to the marginal posterior would only give the correct result with absolute error of order $O(n^{-1/2})$, so the accuracies of both of the Laplace approximations are much better. Since the Laplace approximations yield good relative instead of absolute error, the Laplace approximations maintain good accuracy also in the tails of the densities. In contrast, the normal approximation is accurate only in the vicinity of the mode.

Example 6.2. Consider normal observations

$$[Y_i \mid \mu, \tau] \overset{\text{i.i.d.}}{\sim} N(\mu, \frac{1}{\tau}), \qquad i = 1, \ldots, n,$$

together with the non-conjugated prior

$$p(\mu, \tau) = p(\mu) \, p(\tau) = N(\mu \mid \mu_0, \frac{1}{\psi_0}) \, \text{Gam}(\tau \mid a_0, b_0).$$

The full conditional of μ is readily available,

$$p(\mu \mid \tau, y) = N(\mu \mid \mu_1, \frac{1}{\psi_1})$$

where

$$\psi_1 = \psi_0 + n\tau \qquad \psi_1\,\mu_1 = \psi_0\,\mu_0 + \tau \sum_{i=1}^{n} y_i$$

The mode of the full conditional $p(\mu \mid \tau, y)$ is

$$\mu^*(\tau) = \mu_1 = \frac{\psi_0\,\mu_0 + \tau \sum_{i=1}^{n} y_i}{\psi_0 + n\tau}.$$

We now use this knowledge to build a Laplace approximation to the marginal posterior of τ.

Since, as a function of μ,

$$p(\mu, \tau \mid y) \propto p(\mu \mid \tau, y),$$

$\mu^*(\tau)$ is also the mode of $p(\mu, \tau \mid y)$ for any τ. We also need the second derivative

$$\frac{\partial^2}{\partial \mu^2} \left(\log p(\mu, \tau \mid y)\right) = \frac{\partial^2}{\partial \mu^2} \left(\log p(\mu \mid \tau, y)\right) = -\psi_1,$$

for $\mu = \mu^*(\tau)$, but the derivative does not in this case depend on the value of μ at all. An unnormalized form of the Laplace approximation to the marginal posterior of τ is therefore

$$p(\tau \mid y) \propto \frac{q(\mu^*(\tau), \tau \mid y)}{\sqrt{\psi_1}}, \quad \text{where} \quad q(\mu, \tau \mid y) = p(y \mid \mu, \tau)\, p(\mu)\, p(\tau).$$

In this toy example, the Laplace approximation (6.17) for the functional form of the marginal posterior $p(\tau \mid \mu)$ is exact, since by the multiplication rule,

$$p(\tau \mid y) = \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)}$$

for any choice of μ, in particular for $\mu = \mu^*(\tau)$. Here the numerator is known only in an unnormalized form.

Figure 6.2 (a) illustrates the result using data $y = (-1.4, -1.6, -2.4, 0.7, 0.6)$ and hyperparameters $\mu_0 = 0$, $\psi_0 = 0.5$, $a_0 = 1$, $b_0 = 0.1$. The unnormalized (approximate) marginal posterior has been drawn using the grid method of Sec. 6.1. Figure 6.2 (b) shows an i.i.d. sample drawn from the approximate posterior

$$\tilde{p}(\tau \mid y)\, p(\mu \mid \tau, y),$$

where $\tilde{p}(\tau \mid y)$ is a histogram approximation to the true marginal posterior $p(\tau \mid y)$, which has been sampled using the grid method.

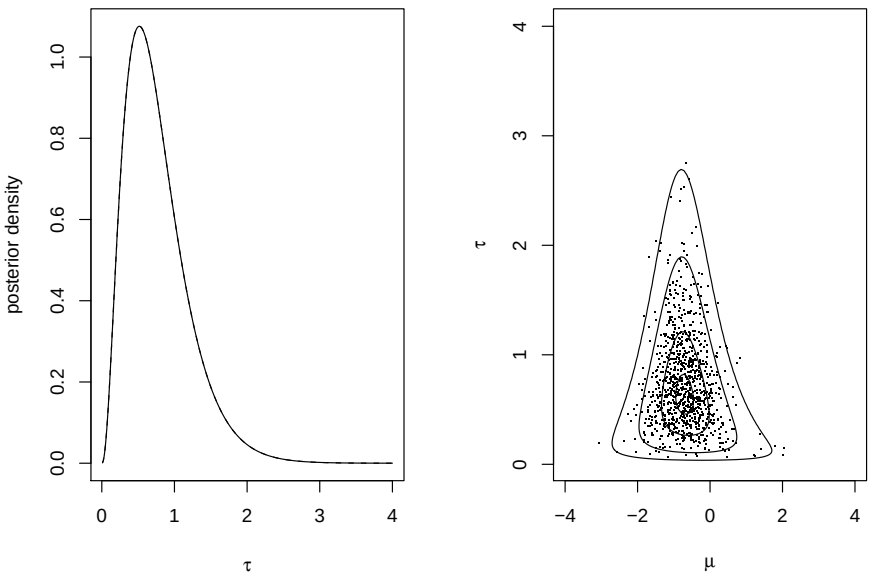
△

Bibliography

[1] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, 2nd edition, 2004.

[2] Tom Leonard. A simple predictive density function: Comment. *Journal of the American Statistical Association*, 77:657–658, 1982.

Figure 6.2: (a) Marginal posterior density of τ and (b) a sample drawn from the approximate joint posterior together with contours of the true joint posterior density.



(a) (b)

[3] H. Rue and S. Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192, 2007.

[4] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Lapalce approximations. *Journal of the Royal Statistical Society: Series B*, 2009. to appear.

[5] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

Chapter 7

MCMC algorithms

7.1 Introduction

In a complicated Bayesian statistical model it may be very difficult to analyze the mathematical form of the posterior and it may be very difficult to draw an i.i.d. sample from it. Fortunately, it is often easy to generate a correlated sample, which approximately comes from the posterior distribution. (In this context, the word *correlated* means *not independent*). However, we would very much prefer to have an i.i.d. sample from the posterior, instead. After one has available a sample, one can estimate posterior expectations and posterior quantiles using the same kind of techniques that are used with i.i.d. samples. This is the idea behind Markov chain Monte Carlo (MCMC) methods.

In this chapter we will introduce the basic MCMC sampling algorithms that are used in practical problems. The emphasis is on trying to understand what one needs to do in order to implement the algorithms. In Chapter 11 we will see why these algorithms work using certain concepts from the theory of Markov chains in a general state space.

There are available computer programs that can implement an MCMC simulation automatically. Perhaps the most famous such program is the BUGS system (Bayesian inference Using Gibbs Sampling), which has several concrete implementations, most notably WinBUGS and OpenBUGS. You can analyze most of the models of interest easily using BUGS. What the user of BUGS needs to do is to write the description of the model in a format that BUGS understands, read the data into the program, and then let the program do the simulation. Once the simulation has finished, one can let the program produce various summaries of the posterior. Using such a tool, it is simple to experiment with different priors and different likelihoods for the same data.

However, in this chapter the emphasis is on understanding how you can write your own MCMC programs. Why would this be of interest?

- If you have not used MCMC before, you get a better understanding of the methods if you try to implement (some of) them yourself.
- For some models, the automated tools fail. Sometimes you can, however, rather easily design and implement a MCMC sampler yourself, once you understand the basic principles. (In some cases, however, designing an efficient MCMC sampler can be an almost impossibly difficult task.)

- Sometimes you want to have more control over the sampling algorithm than is provided by the automated tools. In some cases implementation details can make a big difference to the efficiency of the method.

The most famous MCMC methods are the Metropolis–Hastings sampler and the Gibbs sampler. Where do these names come from?

- Nicholas (Nick) Metropolis (1915–1999) was an American mathematician, physicist and pioneer of computing, who was born in Greece. He published the Metropolis sampler in 1953 jointly with two husband-and-wife teams, namely A.W. and M.N. Rosenbluth and A.H. and E. Teller. At that time the theory of general state space Markov chains was largely unexplored. In spite of this, the authors managed to give a heuristic proof for the validity of the method.

- W. Keith Hastings (1930–) is a Canadian statistician, who published the Metropolis–Hastings sampler in 1970. It is a generalization of the Metropolis sampler. Hastings presented his algorithm using a discrete state space formalism, since the theory of general state space Markov chains was then known only to some specialists in probability theory. Hastings' article did not have a real impact on statisticians until much later.

- The name Gibbs sampler was introduced by the brothers S. and D. Geman in an article published in 1984. Related ideas were published also by other people at roughly the same time. The method is named after the American mathematician and physicist J. Willard Gibbs (1893–1903), who studied thermodynamics and statistical physics, but did not have anything to do with MCMC.

In the late 1980's and early 1990's there was an explosion in the number of studies, where people used MCMC methods in Bayesian inference. Now there was available enough computing power to apply the methods, and besides, the theory of general state space Markov chains had matured so that readable expositions of the theory were available.

Nowadays, many statisticians routinely use the concept of a Markov chain which evolves in a general state space. Unfortunately, their mathematical theory is still explained only in a handful of text books.

7.2 Basic ideas of MCMC

MCMC algorithms are based on the idea of a Markov chain which evolves in discrete time. A Markov chain is a stochastic process

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$$

Here $\theta^{(i)}$ (the state of the process at time i) is a RV whose values lie in a state space, which usually is a subset of some Euclidean space $\mathbb{R}^d$. The state space is the same for all times i. We write the time index as a superscript so that we can index the components $\theta^{(i)}$ using a subscript.

Markov chains have the following **Markov property**: the distribution of the next state $\theta^{(i+1)}$ depends on the history $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(i)}$ only through the

present state $\theta^{(i)}$. The Markov chains used in MCMC methods are **homogeneous**: the conditional distribution of $\theta^{(i+1)}$ given $\theta^{(i)}$ does not depend on the index i.

The following algorithm shows how one can simulate a Markov chain, in principle. Intuitively, a Markov chain is nothing else but the mathematical idealization of this simulation algorithm. (There are, howere, important Markov chains which are easiear to simulate using some other structure for the simulation program.)

Algorithm 14: Computer scientist's definition of a homogeneous Markov chain.

```
Generate θ^(0) from a given initial distribution;
for i = 0, 1, 2, ... do
    Generate a vector V^(i+1) of fresh random numbers from a suitable
    distribution;
    θ^(i+1) ← h(θ^(i), V^(i+1)) for a suitable function h(·, ·);
end
```

Some (but not all) Markov chains have an **invariant distribution** (or a stationary distribution or equilibrium distribution), which can be defined as follows. If the initial state of the chain $\theta^{(0)}$ follows the invariant distribution, then also all the subsequent states $\theta^{(i)}$ follow it.

If a Markov chain has an invariant distribution, then (under certain regularity conditions) the distribution of the state $\theta^{(i)}$ converges to that invariant distribution (in a certain sense). Under certain regularity conditions, such a chain is **ergodic**, which ensures that an arithmetic average (or an ergodic average) of the form

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)})$$

converges, almost surely, to the corresponding expectation calculated under the invariant distribution as $N \to \infty$. That is, the ergodic theorem for Markov chains then states that the strong law of large numbers holds, i.e.,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \to E_f h(\Theta) = \int h(\theta) f(\theta)\, \mathrm{d}\theta, \tag{7.1}$$

where f is the density of the invariant distribution. This will then hold for all functions h for which the expectation $E_f h(\Theta)$ exists, so the convergence is as strong as in the strong law of large numbers for i.i.d. sequences. There are also more advanced forms of ergodicity (geometric ergodicity and uniform ergodicity), which a Markov chain may either have or not have.

Under still more conditions, Markov chains also satisfy a central limit theorem, which characterizes the speed of convergence in the ergodic theorem. The central limit theorem for Markov chains is of the form

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) - E_f h(\Theta)\right) \xrightarrow{\mathrm{d}} N(0, \sigma_h^2).$$

The speed of convergence is of the same order of N as in the central limit theorem for i.i.d. sequences. However, estimating the variance σ_h^2 in the central limit theorem is lot trickier than with i.i.d. sequences.

After this preparation, it is possible to explain the basic idea of MCMC methods. The idea is to set up an ergodic Markov chain which has the posterior distribution as its invariant distribution. Doing this is often surprisingly easy. Then one simulates values

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$$

of the chain. When t is sufficiently large, then $\theta^{(t)}$ and all the subsequent states $\theta^{(t+i)}, i \geq 1$ follow approximately the posterior distribution. The time required for the chain to approximately achieve its invariant distribution is called the **burn-in**. After the initial burn-in period has been discarded, the subsequent values

$$\theta^{(t)}, \theta^{(t+1)}, \theta^{(t+2)}, \ldots$$

can be treated as a dependent sample from the posterior distribution, and we can calculate posterior expectations, quantiles and other summaries of the posterior distribution based on this sample.

After the burn-in period we need to store the simulated values of the chain for later use. So, for a scalar parameter we need a vector to store the results, for a vector parameter we need a matrix to store the results and so on. To save space, one often decides to **thin** the sequences by keeping only every kth value of each sequence and by discarding the rest.

Setting up *some* MCMC algorithm for a given posterior is usually easy. However, the challenge is to find an MCMC algorithm which converges rapidly and then explores efficiently the whole support of the posterior distribution. Then one can get a reliable picture of the posterior distribution after stopping the simulation after a reasonable number of iterations.

In practice one may want to try several approaches for approximate posterior inference in order to become convinced that the posterior inferences obtained with MCMC are reliable. One can, e.g., study simplified forms of the statistical model (where analytical developments or maximum likelihood estimation or other asymptotic approximations to Bayesian estimation may be possible), simulate several chains which are initialized from different starting points and are possibly computed with different algorithms, and compute approximations to the posterior.

7.3 The Metropolis–Hastings algorithm

Now we consider a target distribution with density $\pi(\theta)$, which may be available only in an unnormalized form $\tilde{\pi}(\theta)$. Usually the target density is the posterior density of a Bayesian statistical model,

$$\pi(\theta) = p(\theta \mid y).$$

Actually we only need to know an unnormalized form of the posterior, which is given, e.g., in the form of prior times likelihood,

$$\tilde{\pi}(\theta) = p(\theta)\, p(y \mid \theta).$$

The density $\pi(\theta)$ may be a density in the generalized sense, so we may have a discrete distribution for some components of θ and a continuous distribution for others.

For the Metropolis–Hastings algorithm we need a proposal density $q(\theta' \mid \theta)$, from which we are able to simulate. (Some authors call the proposal density the jumping density or candidate generating density.) As a function of θ', the proposal density $q(\theta' \mid \theta)$ is a density on the parameter space for each value of θ. When the current state of the chain is $\theta = \theta^{(i)}$, we propose a value for the next state from the distribution with density

$$\theta' \mapsto q(\theta' \mid \theta)$$

The proposed value θ' is then accepted or rejected in the algorithm. If the proposal is accepted, then the next state $\theta^{(i+1)}$ is taken to be θ', but otherwise the chain stays in the same state, i.e., $\theta^{(i+1)}$ is assigned the current state $\theta^{(i)}$.

The acceptance condition has to be selected carefully so that we get the target distribution as the invariant distribution of the chain. The usual procedure works as follows. We calculate the value of the Metropolis–Hastings ratio (M–H ratio)

$$r = r(\theta', \theta) = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta),} \tag{7.2}$$

where $\theta = \theta^{(i)}$ is the current state and θ' is the proposed state. Then we generate a value u from the standard uniform $\mathrm{Uni}(0, 1)$. If $u < r$, then we accept the proposal and otherwise reject it. For the analysis of the algorithm, it is essential to notice that the probability of accepting the proposed θ', when the current state is θ, is given by

$$\Pr(\text{proposed value is accepted} \mid \theta^{(i)} = \theta, \theta') = \min(1, r(\theta', \theta)). \tag{7.3}$$

We need here the minimum of one and the M–H ratio, since the M–H ratio may very well be greater than one.

Some explanations are in order.

- The denominator of the M–H ratio (7.2) is the joint density of the proposal θ' and the current state θ, when the current state already follows the posterior.
- The numerator is of the same form as the denominator, but θ and θ' have exchanged places.
- If $\pi(\theta^{(0)}) > 0$, then the denominator of the M–H ratio is always strictly positive during the algorithm. When $i = 0$ this follows from the observation that $q(\theta' \mid \theta^{(0)})$ has to be positive, since θ' is generated from that density. Also $\pi(\theta^{(1)})$ has to be positive, thanks to the form of the acceptance test. The rest follows by induction.
- We do not need to know the normalizing constant of the target distribution, since it cancels in the M–H ratio,

$$r = \frac{\pi(\theta')\, q(\theta \mid \theta')}{\pi(\theta)\, q(\theta' \mid \theta)} = \frac{\tilde{\pi}(\theta')\, q(\theta \mid \theta')}{\tilde{\pi}(\theta)\, q(\theta' \mid \theta)} \tag{7.4}$$

- If the target density is a posterior distribution, then the M–H ratio is given by

$$r = \frac{f_{Y|\Theta}(y \mid \theta')\, f_\Theta(\theta')\, q(\theta \mid \theta')}{f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta)\, q(\theta' \mid \theta)}. \tag{7.5}$$

- Once you know what the notation is supposed to mean, you can use an abbreviated notation for the M–H ratio, such as

$$r = \frac{p(\theta' \mid y)\, q(\theta \mid \theta')}{p(\theta \mid y)\, q(\theta' \mid \theta)}.$$

  Here, e.g., $p(\theta' \mid y)$ is the value of the posterior density evaluated at the proposal θ'.

An explanation of why the target distribution is the invariant distribution of the resulting Markov chain will be given in Chapter 11. Then it will become clear, that other formulas in place of eq. (7.2) would work, too. However, the formula (7.2) is known to be optimal (in a certain sense), and therefore it is the one that is used in practice.

In the Metropolis–Hastings algorithm the proposal density can be selected otherwise quite freely, but we must be sure that we can reach (with positive probability) any reasonably possible region in the parameter space starting from any initial state $\theta^{(0)}$ with a finite number of steps. This property is called **irreducibility** of the Markov chain.

Algorithm 15: The Metropolis–Hastings algorithm.

Input: An initial value $\theta^{(0)}$ such that $\tilde{\pi}(\theta^{(0)}) > 0$ and the number of iterations N.

Result: Values simulated from a Markov chain which has as its invariant distribution the distribution corresponding to the unnormalized density $\tilde{\pi}(\theta)$.

1 **for** $i = 0, 1, 2, \ldots, N$ **do**

2 $\quad \theta \leftarrow \theta^{(i)}$;

3 $\quad$ Generate θ' from $q(\cdot \mid \theta)$ and u from $\mathrm{Uni}(0, 1)$;

4 $\quad$ Calculate the M–H ratio

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

5 $\quad$ Set

$$\theta^{(i+1)} \leftarrow \begin{cases} \theta', & \text{if } u < r \\ \theta, & \text{otherwise.} \end{cases}$$

6 **end**

Algorithm 15 sums up the Metropolis–Hastings algorithm. When implementing the algorithm, one easily comes across problems, which arise because of underflow or overflow in the calculation of the M–H ratio r. Most of such problems can be cured by calculating with logarithms. E.g., when the target distribution is a posterior distribution, then one should first calculate $s = \log r$ by

$$\begin{aligned} s = {} & \log(f_{Y|\Theta}(y \mid \theta')) - \log(f_{Y|\Theta}(y \mid \theta)) \\ & + \log(f_\Theta(\theta')) - \log(f_\Theta(\theta)) + \log(q(\theta \mid \theta')) - \log(q(\theta' \mid \theta)) \end{aligned}$$

and only then calculate $r = \exp(s)$. Additionally, one might want cancel common factors from r before calculating its logarithm.

Implementing some Metropolis–Hastings algorithm for any given Bayesian statistical model is usually straightforward. However, finding a proposal distribution which allows the chain to converge quickly to the target distribution and allows it to explore the parameter space efficiently may be challenging.

7.4 Concrete Metropolis–Hastings algorithms

In the Metropolis–Hastings algorithm, the proposal θ' is in practice produced by running a piece of code, which can use the current state $\theta^{(i)}$, freshly generated random numbers from any distribution and arbitrary arithmetic operations. We must be able to calculate the density of the proposal θ', when the current state is equal to θ. This is then $q(\theta' \mid \theta)$, which we must be able to evaluate. Or at least we must be able to calculate the value of the ratio

$$q(\theta \mid \theta')/q(\theta' \mid \theta).$$

Different choices for the proposal density correspond to different choices for the needed piece of code. The resulting Metropolis–Hastings algorithms are named after the properties of the proposal distribution.

7.4.1 The independent Metropolis–Hastings algorithm

In the independent M–H algorithm (other common names: independence chain independence sampler), the proposal density is a fixed density, say $s(\theta')$, which does not depend on the value of the current state. In the corresponding piece of code, we only need to generate the value θ' from the proposal distribution.

If the proposal distribution happens to be the target distribution, then every proposal will be accepted, and as a result we will get an i.i.d. sample from the target distribution.

In order to to sample the target distribution properly with the independent M–H algorithm, the proposal density s must be positive everywhere, where the target density is positive. If there exist a majorizing constant M, such that

$$\pi(\theta) \leq M s(\theta) \qquad \forall \theta,$$

then the resulting chain can be shown to have good ergodic properties, but if this condition fails, then the convergence properties of the chain can be bad. (In the independent M–H algorithm one does not need to know the value of M.) This implies that the proposal density should be such that the accept–reject method or importance sampling using that proposal distribution would be possible, too.

In particular, the tails of the proposal density s should be at least as heavy as the tails of the target density. Finding such proposal densities may be difficult in high-dimensional problems. A natural choice would be a multivariate t distribution whose shape is chosen to match the shape of the posterior density. One should choose a low value (e.g. $\nu = 4$) for the degrees of freedom parameter in order to ensure heavy tails, and then one could choose the center μ of the multivariate t distribution $t(\nu, \mu, \Sigma)$ to be equal to the posterior mode and the

dispersion parameter Σ to be equal to the covariance matrix of an approximating normal distribution. Other choices for the center and dispersion matrix are possible, too. E.g., one could choose μ to be equal to the estimated posterior mean and Σ equal to the posterior covariance matrix.

7.4.2 Symmetric proposal distribution

If the proposal density is symmetric in that

$$q(\theta' \mid \theta) = q(\theta \mid \theta'), \qquad \forall \theta, \theta',$$

then the proposal density cancels from the M–H ratio,

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

This the sampling method that was originally proposed by Metropolis. Proposals leading to a higher value for the target density are automatically accepted, and other proposals may be accepted or rejected. Later Hastings generalized the method for non-symmetric proposal densities.

7.4.3 Random walk Metropolis–Hastings

Suppose that g is a density on the parameter space an that we calculate the proposal as follows,

$$\text{generate } w \text{ from density } g \text{ and set } \theta' \leftarrow \theta + w.$$

Then the proposal density is

$$q(\theta' \mid \theta) = g(\theta' - \theta).$$

This kind of a proposal is called a random walk proposal. If the density g is symmetric, i.e.,

$$g(-w) = g(w) \quad \forall w,$$

then the proposal density $q(\theta' \mid \theta)$ is also symmetric, and thus cancels from the M–H ratio. In the case of a symmetric random walk proposal, one often speaks of the random walk Metropolis (RWM) algorithm.

Actually, a random walk is a stochastic process of the form $X_{t+1} = X_t + w_t$, where the random variables w_t are i.i.d. Notice that the stochastic process produced by the random walk M–H algorithm is **not** a random walk, since the proposals can either be accepted or rejected.

The symmetric random walk Metropolis–Hastings algorithm (also known as the random walk Metropolis algorithm) is one of the most commonly used forms of the Metropolis–Hastings method. The most commonly used forms for g are the multivariate normal or multivariate Student's t density centered at the origin. This is, of course, appropriate only for continuous posterior distributions.

Often one selects the covariance matrix of the proposal distribution as

$$a\, C,$$

where C is an approximation to the covariance matrix of the target distribution (in Bayesian inference C is an approximation to the posterior covariance matrix) and the scalar a is a tuning constant which should be calibrated carefully. These kind of proposal distributions work well when the posterior distribution is approximately normal. One sometimes needs to reparametrize the model in order to make this approach work better.

The optimal value of a and the corresponding optimal acceptance rate has been derived theoretically, when the target density is a multivariate normal $N_d(\mu, C)$ and the random walk proposal is $N_d(0, aC)$, see [13]. The scaling constant a should be about $(2.38)^2/d$ when d is large. The corresponding acceptance rate (the number of accepted proposals divided by the total number of proposals) is from around 0.2 (for high-dimensional problems) to around 0.4 (in dimensions one or two). While these results have been derived using the very restrictive assumption that the target density is a multivariate normal, the results anyhow give rough guidelines for calibrating a in a practical problem.

How and why should one try to control the acceptance rate in the random walk M–H algorithm? If the acceptance rate is too low, then the chain is not able to move, and the proposed updating steps are likely to be too large. In this case one could try a smaller value for a. However, a high acceptance rate may also signal a problem, since then the updating steps may be too small. This may lead to the situation where the chain explores only a small portion of the parameter space. In this case one should try a larger value for a. From the convergence point of view, too high acceptance rate is a bigger problem. A low acceptance rate is a problem only from the computing time point of view.

In order to calibrate the random walk M–H algorithm, one needs an estimate of its acceptance rate. A simple-minded approach is just to keep track of the number of accepted proposals. A better approach is to calculate the average of the acceptance probabilities,

$$\frac{1}{N}\sum_{i=1}^{n} \min(1, r_i),$$

where r_i is the M–H ratio in the ith iteration.

In practice, one can try to tune a iteratively, until the acceptance rate is acceptable. The tuning iterations are discarded, and the MCMC sample on which the inference is based is calculated using the fixed proposal distribution, whose scale a is the selected value. Fixing the proposal distribution is necessary, since the theory of the Metropolis–Hastings algorithm requires a homogeneous Markov chain, i.e., a proposal density $q(\theta' \mid \theta)$ which does not depend on the iteration index.

Recently, several researchers have developed adaptive MCMC algorithms, where the proposal distribution is allowed to change all the time during the iterations, see [1] for a review. Be warned that the design of valid adpative MCMC algorithms is subtle and that their analysis requires tools which are more difficult than the general state space Markov chain theory briefly touched upon in Chapter 11.

Example 7.1. Let us try the random walk chain for the target distribution $N(0, 1)$ by generating the increment from the normal distribution $N(0, \sigma^2)$ using the following values for the variance: a) $\sigma^2 = 4$ b) $\sigma^2 = 0.1$ c) $\sigma^2 = 40$.

In situation a) the chain is initialized far away in the tails of the target distribution, but nevertheless it quickly finds its way to the main portion of the target distribution and then explores it efficiently. Such a chain is said to **mix** well. In situations b) and c) the chains are initialized at the center of the target distribution, but the chains mix less quickly. In situation b) the step length is too small, but almost all proposals get accepted. In situation c) the algorithm proposes too large steps, almost all of which get rejected. Figure 7.1 presents trace plots (or time series plots) of the chain in the three situations.

△

7.4.4 Langevin proposals

Unlike a random walk, the Langevin proposals introduce a drift which moves the chain towards the modes of the posterior distribution. When the current state is θ, the proposal θ' is generated with the rule

$$\theta' = \theta + \frac{\sigma^2}{2}\nabla(\log \pi(\theta)) + \sigma\,\epsilon, \qquad \epsilon \sim N_p(0, I).$$

Here $\sigma > 0$ is a tuning parameter and

$$\nabla(\log \pi(\theta)) = \nabla(\log \tilde{\pi}(\theta))$$

is the gradient of the logarithm of the (unnormalized) posterior density. The proposal distribution is motivated by a stochastic differential equation, which has π as its stationary distribution.

This proposal is then accepted or rejected using the ordinary Metropolis–Hastings rule, where the proposal density is

$$q(\theta' \mid \theta) = N_p(\theta' \mid \theta + \frac{\sigma^2}{2}\nabla(\log \pi(\theta)),\ \sigma^2 I).$$

7.4.5 Reparametrization

Suppose that the posterior distribution of interest is a continuous distribution and that we have implemented functions for calculating the log-prior and the log-likelihood in terms of the parameter θ. Now we want to consider a diffeomorphic reparametrization

$$\phi = g(\theta) \quad \Leftrightarrow \quad \theta = h(\phi).$$

Typical reparametrizations one might consider are taking the logarithm of a positive parameter or calculating the logit function of a parameter constrained to the interval $(0, 1)$. What needs to be done in order to implement the Metropolis–Hastings algorithm for the new parameter vector ϕ?

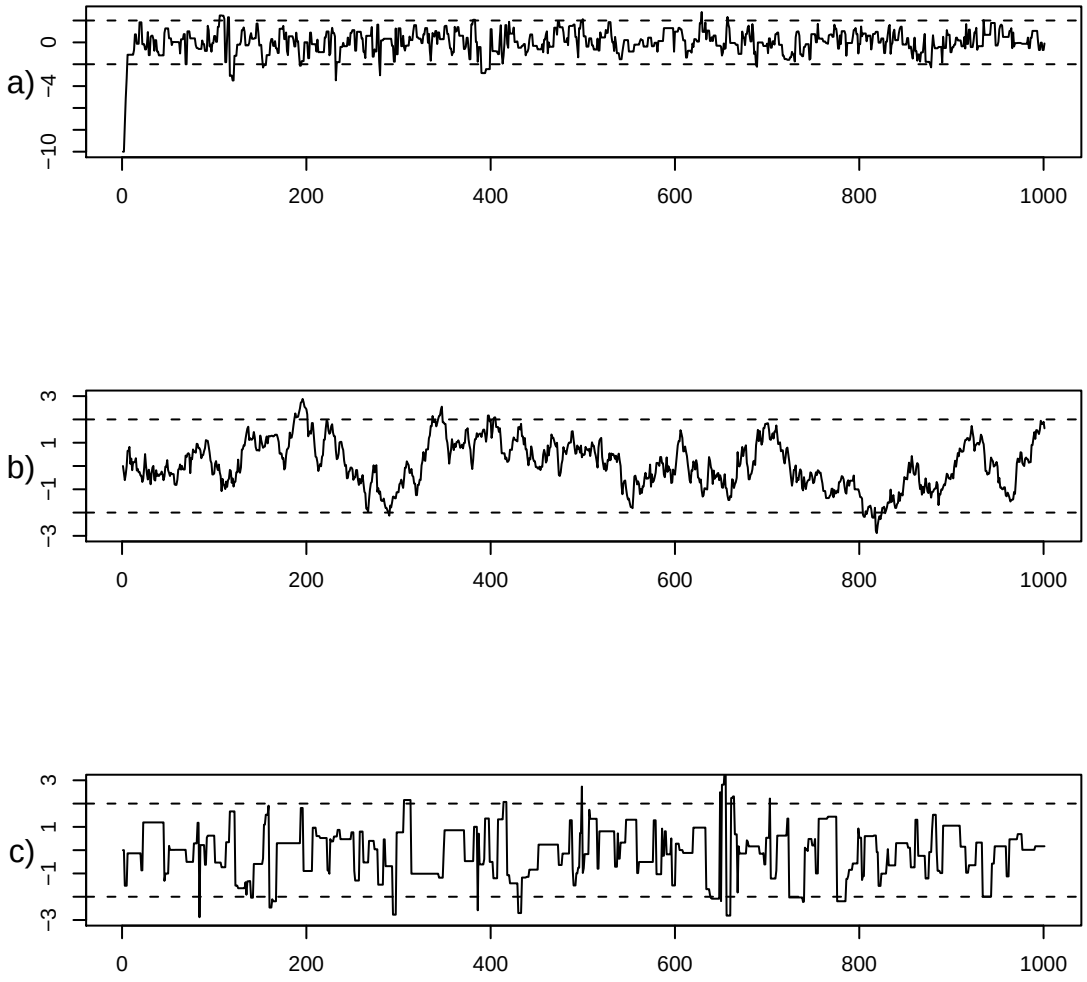
First of all, we need a proposal density $q(\phi' \mid \phi)$ and the corresponding code. We also need to work out how to compute one of the Jacobians

$$J_h(\phi) = \frac{\partial \theta}{\partial \phi} \quad \text{or} \quad J_g(\theta) = \frac{\partial \phi}{\partial \theta}.$$

In ϕ-space the target density is given by the change of variables formula

$$f_{\Phi|Y}(\phi \mid y) = f_{\Theta|Y}(\theta \mid y) \left|\frac{\partial \theta}{\partial \phi}\right| = f_{\Theta|Y}(\theta \mid y)\ |J_h(\phi)|\,,$$

Figure 7.1: Trace plots of the random walk chain using the three different proposal distributions.



where $\theta = h(\phi)$.

The M–H ratio, when we propose ϕ' and the current value is ϕ, is given by

$$\begin{aligned} r &= \frac{f_{\Phi|Y}(\phi' \mid y)\, q(\phi \mid \phi')}{f_{\Phi|Y}(\phi \mid y)\, q(\phi' \mid \phi)} \\ &= \frac{f_{\Theta|Y}(\theta' \mid y)\, |J_h(\phi')|\, q(\phi \mid \phi')}{f_{\Theta|Y}(\theta \mid y)\, |J_h(\phi)|\, q(\phi' \mid \phi)} \\ &= \frac{f_{Y|\Theta}(y \mid \theta')\, f_\Theta(\theta')\, q(\phi \mid \phi')}{f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta)\, q(\phi' \mid \phi)} \, \frac{|J_h(\phi')|}{|J_h(\phi)|} \end{aligned}$$

here $\theta' = h(\phi')$ and $\theta = h(\phi)$. Sometimes it is more convenient to work with the Jacobian J_g, but this is easy, since

$$J_g(\theta) = \frac{1}{J_h(\phi)}.$$

Above we viewed the Jacobians as arising from expressing the target density using the new ϕ parametrization instead of the old θ parametrization. An alternative interpretation is that we should express the proposal density in θ space instead of ϕ space and then use the ordinary formula for M–H ratio. Both viewpoints yield the same formulas.

In order to calculate the logarithm of the M–H ratio, we need to do the following.

- Calculate the θ and θ' values corresponding to the current ϕ and proposed ϕ' values.
- Calculate the log-likelihood and log-prior using the values θ and θ'.
- Calculate the logarithm s of the M–H ratio as

$$\begin{aligned} s = \log(f_{Y|\Theta}(y \mid \theta')) - \log(f_{Y|\Theta}(y \mid \theta)) \\ + \log(f_\Theta(\theta')) - \log(f_\Theta(\theta)) + \log(q(\phi \mid \phi')) - \log(q(\phi' \mid \phi)) \\ + \log(|J_h(\phi')|) - \log(|J_h(\phi)|). \end{aligned}$$

  Finally, calculate $r = \exp(s)$.

- The difference of the logarithms of the absolute Jacobians can be calculated either on the ϕ scale or on the θ scale by using the identity

$$\log(|J_h(\phi')|) - \log(|J_h(\phi)|) = \log(|J_g(\theta)|) - \log(|J_g(\theta')|).$$

7.4.6 State-dependent mixing of proposal distributions

Let θ be the current state of the chain. Suppose that the proposal θ' is drawn from a proposal density, which is selected randomly from a list of alternatives

$$q(\theta' \mid \theta, j), \qquad j = 1, \ldots K,$$

What is more, the selection probabilities may depend on the current state, as follows.

- 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 θ' as the new state, if $U < r$, where $U \sim \text{Uni}(0, 1)$, and
$$r = \frac{\pi(\theta')\,\beta(j \mid \theta')\,q(\theta \mid \theta', j)}{\pi(\theta)\,\beta(j \mid \theta)\,q(\theta' \mid \theta, j)}. \tag{7.6}$$
Otherwise the chain stays at θ.

This formula (7.6) for the M–H ratio r is contained in Green's article [6], which introduced the reversible jump MCMC method. The algorithm could be called the Metropolis–Hastings–Green algorithm.

The lecturer does know any trick for deriving formula (7.6) from the M–H ratio of the ordinary M–H algorithm. The beauty of formula (7.6) lies in the fact that one only needs to evaluate $q(\theta' \mid \theta, j)$ and $q(\theta \mid \theta', j)$ for the proposal density which was selected. A straightforward application of the M–H algorithm would require one to evaluate these densities for all of the K possibilities.

If the selection probabilities $\beta(j \mid \theta)$ do not actually depend on θ, then they cancel from the M–H ratio. In this case (7.6) is easily derived from the ordinary M–H algorithm.

7.5 Gibbs sampler

One of the best known ways of setting up an MCMC algorithm is Gibbs sampling, which is now discussed supposing that the target distribution is a posterior distribution. However, the method can be applied to any target distribution, when the full conditional distributions of the target distribution are available.

Suppose that the parameter vector has been divided into components

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

where θ_j need not be a scalar. Suppose also that the posterior full conditional distributions of each of the components are available in the sense that we know how to simulate them. This is the case when the statistical model exhibits conditional conjugacy with respect to all of the components θ_j. Then the basic idea behind Gibbs sampling is that we simulate successively each component θ_j from its (posterior) full conditional distribution. It is convenient to use the abbreviation θ_{-j} for the vector, which contains all the other components of θ but θ_j, i.e.

$$\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d). \tag{7.7}$$

Then the posterior full conditional of θ_j is

$$p(\theta_j \mid \theta_{-j}, y) = f_{\Theta_j \mid \Theta_{-j}, Y}(\theta_j \mid \theta_{-j}, y). \tag{7.8}$$

A convenient shorthand notation for the posterior full conditional is

$$p(\theta_j \mid \cdot),$$

where the dot denotes all the other random variables except θ_j.

The most common form of the Gibbs sampler is the systematic scan Gibbs sampler, where the components are updated in a fixed cyclic order. It is also possible to select at random which component to update next. In that case one has the random scan Gibbs sampler.

Algorithm 16 presents the systematic scan Gibbs sampler, when we update the components using the order $1, 2, \ldots, d$. In the algorithm i is the time index of the Markov chain. One needs d updates to get from $\theta^{(i)}$ to $\theta^{(i+1)}$. To generate the j'th component, $\theta_j^{(i+1)}$, one uses the most recent values for the other components, some of which have already been updated. I.e., when the value for $\theta_j^{(i+1)}$ is generated, it is generated from the corresponding full conditional using the following values for the other components,

$$\theta_{-j}^{\text{cur}} = (\theta_1^{(i+1)}, \ldots, \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, \ldots, \theta_d^{(i)}).$$

Algorithm 16: Systematic scan Gibbs sampler.

Input: An initial value $\theta^{(0)}$ such that $f_{\Theta|Y}(\theta^{(0)} \mid y) > 0$ and the number of iterations N.

Result: Values simulated from a Markov chain which has the posterior distribution as its invariant distribution.

```
θ^cur ← θ^(0)
for i = 0, 1, ..., N do
    for j = 1, ..., d do
        draw a new value for the jth component θ_j^cur of θ^cur from the
        posterior full conditional f_{Θ_j|Θ_{-j},Y}(θ_j | θ_{-j}^cur, y)
    end
    θ^(i+1) ← θ^cur
end
```

Usually the updating steps for the components of θ are so heterogeneous, that the inner loop is written out in full. E.g., in the case of three components, $\theta = (\phi, \psi, \tau)$, the actual implementation would probably look like the following algorithm 17. This algorithm also demonstrates, how one can write the algorithm using the abbreviated notation for conditional densities.

Algorithm 17: Systematic scan Gibbs sampler for three components $\theta = (\phi, \psi, \tau)$ given initial values for all the components except the one that gets updated the first.

```
ψ^cur ← ψ_0;   τ^cur ← τ_0;
for i = 0, 1, ..., N do
    draw φ^cur from p(φ | ψ = ψ^cur, τ = τ^cur, y);
    draw ψ^cur from p(ψ | φ = φ^cur, τ = τ^cur, y);
    draw τ^cur from p(τ | φ = φ^cur, ψ = ψ^cur, y);
    φ_{i+1} ← φ^cur;   ψ_{i+1} ← ψ^cur;   τ_{i+1} ← τ^cur;
end
```

Algorithm 18 presents the random scan Gibbs sampler. Now one time step of the Markov chain requires only one update of a randomly selected component.

In the random scan version, one can have different probabilities for updating the different components of θ, and this freedom can be useful for some statistical models.

Algorithm 18: Random scan Gibbs sampler.

Input: An initial value $\theta^{(0)}$ such that $f_{\Theta|Y}(\theta^{(0)} \mid y) > 0$, the number of iterations N and a probability vector $\beta_1, \ldots, \beta_d$: each $\beta_j > 0$ and $\beta_1 + \cdots + \beta_d = 1$.

Result: Values simulated from a Markov chain which has the posterior distribution as its invariant distribution.

```
θ^cur ← θ^(0);
for i = 0, 1, ..., N do
    select j from {1, ..., d} with probabilities (β_1, ..., β_d);
    draw a new value for the component θ_j^cur from the posterior full
    conditional f_{Θ_j|Θ_{-j},Y}(θ_j | θ_{-j}^cur, y);
    θ^(i+1) ← θ^cur;
end
```

If the statistical model exhibits conditional conjugacy with respect to all the components of θ, then the Gibbs sampler is easy to implement and is the method of choice for many statisticians. One only needs random number generators for all the posterior full conditionals, and these are easily available for the standard distributions. An appealing feature of the method is the fact that one does not need to choose the proposal distribution as in the Metropolis–Hastings sampler; the proposals of the Gibbs sampler are somehow automatically tuned to the target posterior. However, if some of the components of θ are strongly correlated in the posterior, then the convergence of the Gibbs sampler suffers. So one might want to reparametrize the model so that the transformed parameters are independent in their posterior. Unfortunately, most reparametrizations destroy the conditional conjugacy properties on which the attractiveness of the Gibbs sampler depends.

The name Gibbs sampling is actually not quite appropriate. Gibbs studied distributions arising in statistical physics (often called Gibbs distributions or Boltzmann distributions), which have densities of the form

$$f(x_1, \ldots, x_d) \propto \exp\left(-\frac{1}{kT} E(x_1, \ldots, x_d)\right),$$

where $(x_1, \ldots, x_d)$ is the state of physical system, k is a constant, T is the temperature of the system, and $E(x_1, \ldots, x_d) > 0$ is the energy of the system. The Geman brothers used a computational method (simulated annealing), where a computational parameter corresponding to the the temperature of a Gibbs distribution was gradually lowered towards zero. At each temperature the distribution of the system was simulated using the Gibbs sampler. This way they could obtain the configurations of minimal energy in the limit. The name Gibbs sampling was selected in order to emphasize the relationship with the Gibbs distributions. However, when the Gibbs sampler is applied to posterior inference, the temperature parameter is not needed, and therefore the reason for the name Gibbs has disappeared. Many authors have pointed this out this deficiency and

proposed alternative names for the sampling method, but none of them have stuck.

7.6 Componentwise updates in the Metropolis–Hastings algorithm

Already Metropolis *et al.* and Hastings pointed out that one can use componentwise updates in the Metropolis–Hastings algorithm. This is sometimes called single-site updating or blockwise updating.

The parameter vector is divided into d components (or blocks)

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

which need not be scalars. In addtion, we need d proposal densities

$$\theta'_j \mapsto q_j(\theta'_j \mid \theta^{\text{cur}}), \qquad j = 1, \ldots, d,$$

which may all be different.

When it is time to update the jth component, we do a single Metropolis–Hastings step. When the current value of the parameter vector is θ^{cur}, we propose the vector θ', where the jth component is drawn from the proposal density $q_j(\theta_j \mid \theta^{\text{cur}})$, and the rest of the components of θ' are equal to those of the current value θ^{cur}. Then the proposal is accepted or rejected using the M–H ratio

$$r = \frac{p(\theta' \mid y)\, q_j(\theta_j^{\text{cur}} \mid \theta')}{p(\theta^{\text{cur}} \mid y)\, q_j(\theta'_j \mid \theta^{\text{cur}})} \tag{7.9}$$

The vectors θ' and θ^{cur} differ only in the jth place, and therefore one can write the M–H ratio (for updating the jth component) also in the form

$$r = \frac{p(\theta'_j \mid \theta_{-j}^{\text{cur}}, y)\, q_j(\theta_j^{\text{cur}} \mid \theta')}{p(\theta_j^{\text{cur}} \mid \theta_{-j}^{\text{cur}}, y)\, q_j(\theta'_j \mid \theta^{\text{cur}})}, \tag{7.10}$$

where we used the multiplication rule to express the joint posterior as

$$p(\theta \mid y) = p(\theta_{-j} \mid y)\, p(\theta_j \mid \theta_{-j}, y)$$

both in the numerator and in the denominator, and then cancelled the common factor $p(\theta_{-j}^{\text{cur}} \mid y)$. Although eqs. (7.9) and (7.10) are equivalent, notice that in eq. (7.9) we have the M–H ratio when we regard the joint posterior as the target distribution, but in eq. (7.10) we have ostensibly the M–H ratio, when the target is the posterior full conditional of component j. If one then selects as q_j the posterior full conditional of the component θ_j for each j, then each proposal is accepted and the Gibbs sampler ensues.

One can use this procedure either a systematic or a random scan sampler, as is the case with the Gibbs sampler. The resulting algorithm is often called the Metropolis–within–Gibbs sampler. (The name is illogical: the Gibbs sampler is a special case of the Metropolis–Hastings algorithm with componentwise updates.) This is also a very popular MCMC algorithm, since then one does not have to design a single complicated multivariate proposal density but p simpler proposal densities, many of which may be full conditional densities of the posterior.

Small modifications in the implementation can sometimes make a big difference to the efficiency of the sampler. One important decision is how to divide the parameter vector into components. This is called **blocking** or **grouping**. As a general rule, the less dependent the components are in the posterior, the better the sampler. Therefore it may be a good idea to combine highly correlated components into a single block, with is then updated as a single entity.

It is sometimes useful to update the whole vector jointly using a single Metropolis–Hastings acceptance test, even if the proposed value is build up component by component taking advantage of conditional conjugacy properties. These and other ways of improving the performance of MCMC algorithms in the context of specific statistical models are topics of current research.

7.7 Analyzing MCMC output

After the MCMC algorithm has been programmed and tested, the user should investigate the properties of the algorithm for the particular problem he or she is trying to solve. There are available several tools, e.g., for

- diagnosing convergence
- estimating Monte Carlo standard errors.

We discuss some of the simpler tools.

A **trace plot** of a parameter ϕ is a plot of the iterates $\phi^{(t)}$ against the iteration number t. These are often examined for each of the components of the parameter vector, and sometimes also for selected scalar functions of the parameter vector. A trace plot is also called a *sample path*, a *history plot* or a *times series plot*. If the chain mixes well, then the trace plots move quickly away from their starting values and they wiggle vigorously in the region supported by the posterior. In that case one may select the length of the burn-in by examining trace plots. (This is not foolproof, since the chain may only have converged momentarily to some neighborhood of a local maximum of the posterior.) If the chain mixes poorly, then the traces will remain nearly constant for many iterations and the state may seem to wander systematically towards some direction. Then one may need a huge number of iterations before the traces show convergence.

An **autocorrelation plot** is a plot of the autocorrelation of the sequence $\phi^{(t)}$ at different iteration lags. These can be produced for all the interesting components of θ, but one should reject the burn-in before estimating the autocorrelation so that one analyzes only that part of the history where the chain is approximately stationary. The autocorrelation function (acf) of a stationary sequence of RVs (X_i) at lag k is defined by

$$R(k) = \frac{E[(X_i - \mu)(X_{i+k} - \mu)]}{\sigma^2}, k = 0, 1, 2, \ldots,$$

where $\mu = EX_i$, $\sigma^2 = \operatorname{var} X_i$, and the assumption of stationarity entails that μ, σ^2 and $R(k)$ do not depend on index i. For an i.i.d. sequence the autocorrelation function is one at lag zero and zero otherwise. A chain that mixes slowly exhibits slow decay of the autocorrelation as the lag increases. When there are more than one parameter, one may also examine cross-correlations between the parameters.

There exist tools for **convergence diagnostics**, which try to help in deciding whether the chain has already approximately reached its stationary distribution and in selecting the length of the burn-in period. E.g., in the approach of Gelman and Rubin, the chain is run many times starting from separate starting values dispersed over the support of the posterior. After the burn-in has been discarded, one calculates statistics which try to check whether all the chains have converged to the same distribution. In some other approches one needs to simulate only a single chain and one compares the behaviour of the chain in the beginning and in the end of the simulation run. Such convergence diagnostic are available in the `coda` R package and in the `boa` R package. However, convergence diagnostic tools can not prove that the chain has converged. They only help you to detect obvious cases of non-convergence.

If the chain seems to have converged, then it is of interest to estimate standard errors for the scalar parameters. The naive estimate (which is correct for i.i.d. sampling) would be to calculate the sample standard deviation of the last L iterations divided by $\sqrt{L}$ (after the burn-in has been discarded). However, MCMC iterates are typically positively correlated, and therefore this would underestimate severely the standard error.

A simple method for estimating the standard errors for posterior expectations

$$E[h(\Theta) \mid Y = y]$$

is the method of **batch means** [8], where the L last iterates are divided into a non-overlapping batches of length b. Then one computes the mean $\bar{h}_j$ of the values $h(\theta^{(t)})$ inside each of the batches $j = 1, \ldots, a$ and estimates the standard error of the grand mean $\bar{h}$ as the square roof of

$$\frac{1}{a} \frac{1}{a-1} \sum_{j=1}^{a} (\bar{h}_j - \bar{h})^2,$$

where $\bar{h}$ is the grand mean calculated from all the the L last iterates $h(\theta^{(t)})$. The idea here is to treat the batch means as i.i.d. random variables whose expected value is the posterior expectation. One should perhaps select the batch length as a function of the simulation length, e.g., with the rule $b = \lfloor \sqrt{L} \rfloor$.

7.8 Example

Consider the two dimensional normal distribution $N(0, \Sigma)$ as the target distribution, where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad -1 < \rho < 1, \quad \sigma_1, \sigma_2 > 0,$$

and ρ is nearly one. Of course, it is possible to sample this two-variate normal distribution directly. However, we next apply MCMC algorithms to this highly correlated toy problem in order to demonstrate properties of the Gibbs sampler and a certain Metropolis–Hastings sampler.

The full conditionals of the target distribution are given by

$$\begin{aligned} [\Theta_1 \mid \Theta_2 = \theta_2] &\sim N\left(\tfrac{\rho\sigma_1}{\sigma_2}\theta_2, (1-\rho^2)\sigma_1^2\right) \\ [\Theta_2 \mid \Theta_1 = \theta_1] &\sim N\left(\tfrac{\rho\sigma_2}{\sigma_1}\theta_1, (1-\rho^2)\sigma_2^2\right), \end{aligned}$$

Figure 7.2: The first ten iterations of the Gibbs sampler. The three contour lines enclose 50 %, 90 % and 99 % of the probability mass of the target distribution.

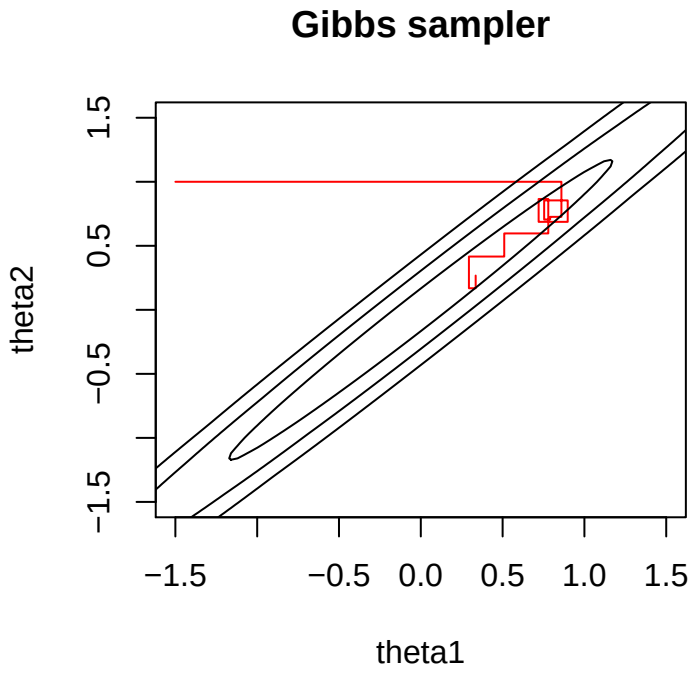
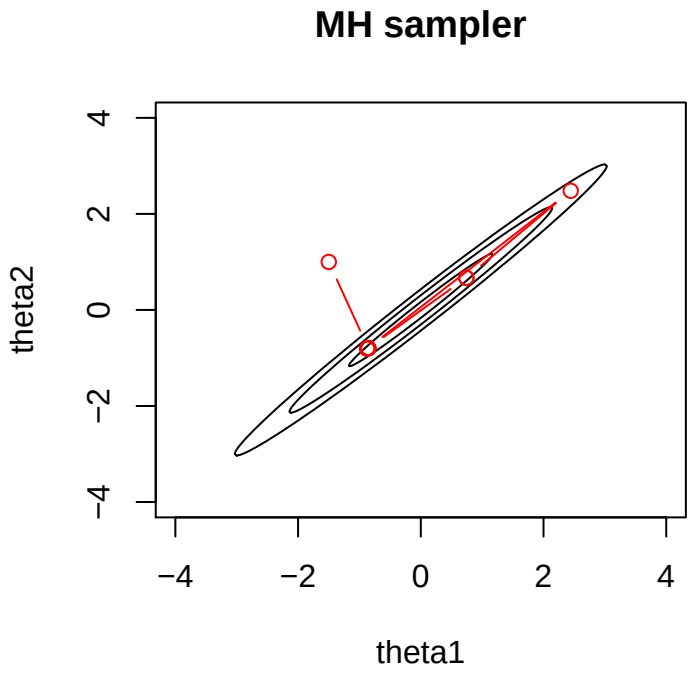


Figure 7.3: The first ten iterations of the Metropolis–Hastings sampler. Notice the sampler produced less than ten distinct θ values. The three contour lines enclose 50 %, 90 % and 99 % of the probability mass of the target distribution.



and these are easy to simulate. We now suppose that

$$\rho = 0.99, \quad \sigma_1 = \sigma_2 = 1.$$

Figure 7.2 shows the ten first steps of the Gibbs sampler, when all the component updates ("half-steps" of the sampler) are shown. Since ρ is almost one, the Gibbs sampler is forced to take small steps, and it takes a long time for it to explore the main support of the target distribution.

Another strategy would be to generate the proposal in two stages as follows. We first draw θ_1' from some convenient proposal distribution, e.g., by the random walk proposal

$$\theta_1' = \theta_1^{\text{cur}} + w,$$

where w is generated from (say) $N(0, 4)$. Then we draw θ_2' from the full conditional distribution of θ_2 conditioning on the proposed value θ_1'. Then the overall proposal density is given by

$$q((\theta_1', \theta_2') \mid (\theta_1^{\text{cur}}, \theta_2^{\text{cur}})) = N(\theta_1' - \theta_1^{\text{cur}} \mid 0, 4)\, N(\theta_2' \mid \frac{\rho\sigma_2}{\sigma_1}\theta_1', (1-\rho^2)\sigma_2^2)$$

We then either accept or reject the transition from θ^{cur} to θ' using the ordinary acceptance rule of the Metropolis–Hastings sampler. This algorithm explores

the target distribution much more efficiently, as can be guessed from Figure 7.3, which shows the first ten iterations of the sampler. The random walk proposal gives the component θ_1 freedom to explore the parameter space, and then the proposal from the full conditional for θ_2 draws the proposed pair into the main support of the target density.

Figure 7.4 shows the traces of the components using the two algorithms. The Metropolis–Hastings sampler seems to mix better than the Gibbs sampler, since there seems to be less dependence between the consecutive simulated values. Figure 7.5 shows the autocorrelation plots for the two components using the two different samplers. The autocorrelation functions produced by the Gibbs sampler decay more slowly than those produced by the Metropolis–Hastings sampler, and this demonstrates that we obtain better mixing with the Metropolis–Hastings sampler.

7.9 Literature

The original references on the Metropolis sampler, the Metropolis–Hastings sampler and the Gibbs sampler are [9, 7, 4]. The article by Gelfand and Smith [3] finally convinced the statistical community about the usefulness of these methods in Bayesian inference. The book [5] contains lots of information on MCMC methods and their applications.

The books by Nummelin [11] or Meyn and Tweedie [10] can be consulted for the theory of Markov chains in a general state space. The main features of the general state space theory are explained in several sources, including [2, Ch. 14] or [12, Ch. 6].

Bibliography

[1] Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373, 2008.

[2] Krishna B. Athreya and Soumendra N. Lahiri. *Measure Theory and Probability Theory*. Springer Texts in Statistics. Springer, 2006.

[3] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[4] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[5] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

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

[7] W. Hastings. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57:97–109, 1970.

Figure 7.4: Sampler traces for the two components θ_1 and θ_2 using the Gibbs sampler and the Metropolis–Hastings sampler.

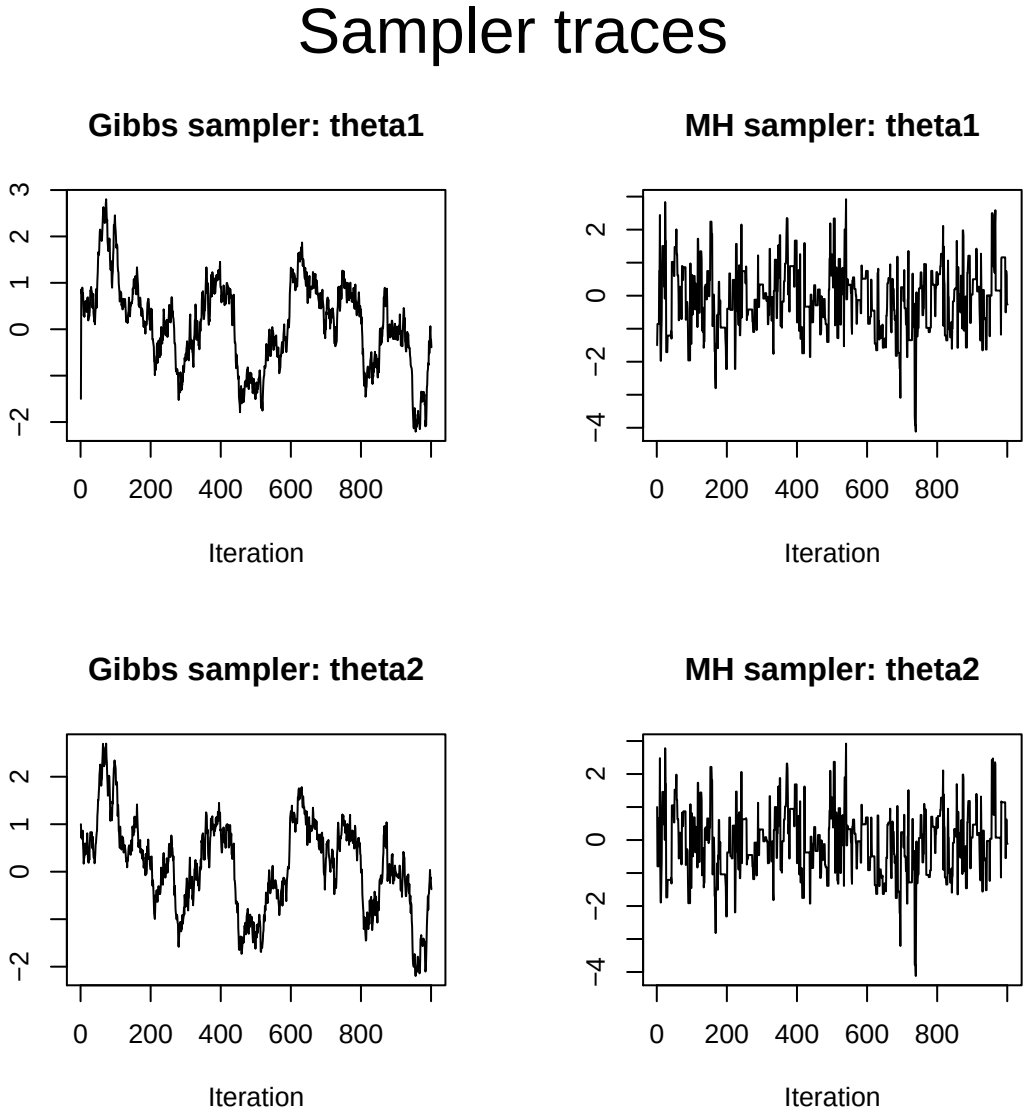
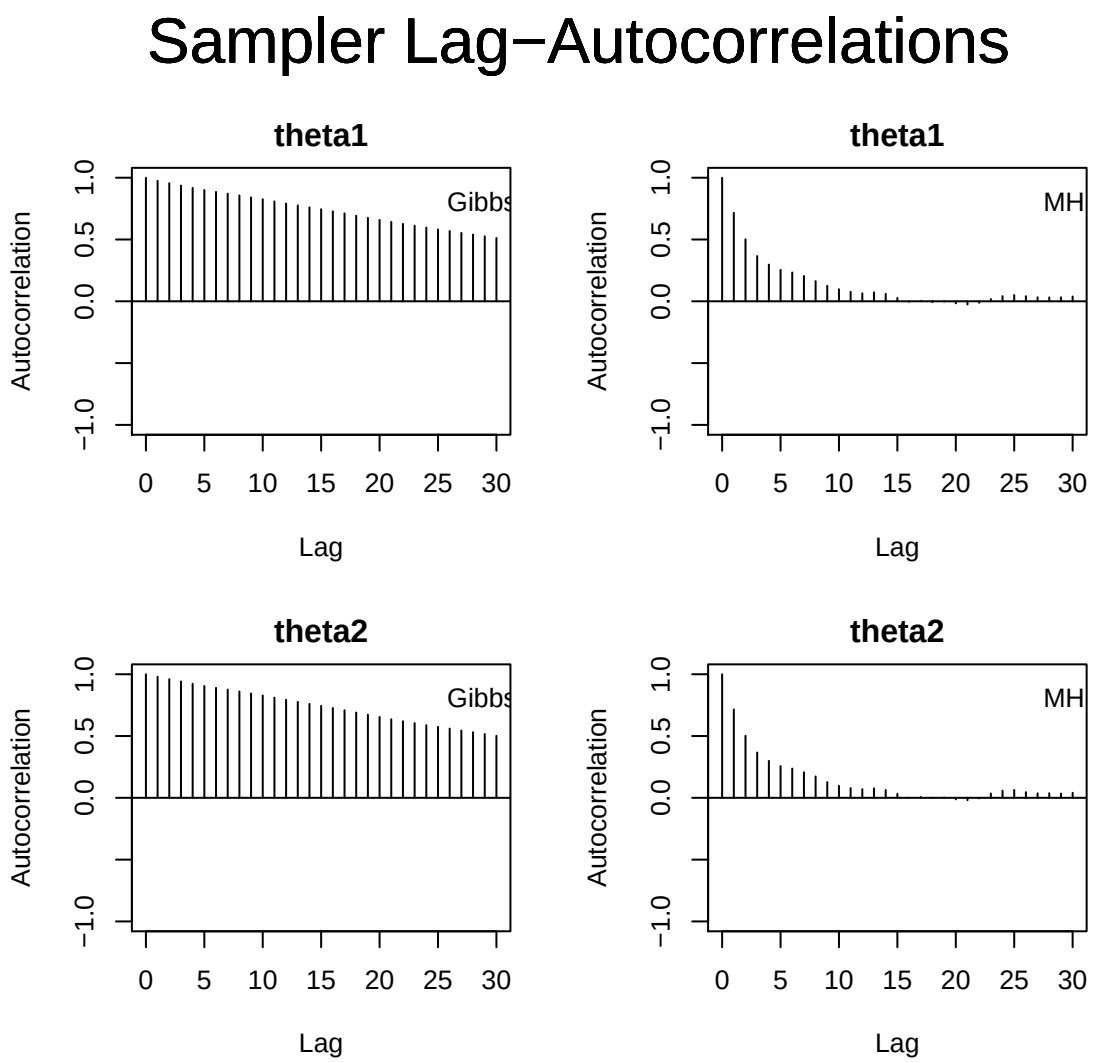


Figure 7.5: Sampler autocorrelation functions for the two components θ_1 and θ_2 using the Gibbs sampler and the Metropolis–Hastings sampler.



[8] Averll M. Law and W. David Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 2nd edition, 1991.

[9] N. Metropolis, A. Rosenbluth, , M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[10] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1993.

[11] Esa Nummelin. *General Irreducible Markov Chains and Nonnegative Operators*. Cambridge University Press, 1984.

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

[13] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

Chapter 8

Auxiliary Variable Models

8.1 Introduction

We are interested in an actual statistical model, with joint distribution

$$\mathrm{p}_{\mathrm{act}}(y, \theta) = \mathrm{p}_{\mathrm{act}}(y \mid \theta)\ \mathrm{p}_{\mathrm{act}}(\theta),$$

but where the posterior $\mathrm{p}_{\mathrm{act}}(\theta \mid y)$ is awkward to sample from. Suppose we are able to reformulate the original model by introducing a new random variable Z such that the marginal distribution of (y, θ) in the new model is the same as the joint distribution of (y, θ) in the original model, i.e., we assume that

$$\int \mathrm{p}_{\mathrm{aug}}(y, \theta, z)\, \mathrm{d}z = \mathrm{p}_{\mathrm{act}}(y, \theta). \tag{8.1}$$

When this is the case, we can forget the distinction between the actual model $\mathrm{p}_{\mathrm{act}}(\cdot)$ and the augmented model $\mathrm{p}_{\mathrm{aug}}(\cdot)$ and use the generic symbol $p(\cdot)$ to denote the densities calculated under either of the models. Here the *augmentation parameter*, the *auxiliary variable*, the *latent variable* or the *latent data* Z can be anything. However, it requires ingenuity and insight to come up with useful auxiliary variables.

Sometimes it is possible to sample much more efficiently from $p(\theta, z \mid y)$ than from $p(\theta \mid y)$. In such a case we can sample from the posterior $p(\theta, z \mid y)$, and we get a sample from the marginal posterior of θ by ignoring the z components of the (θ, z) sample. If both the full conditionals $p(\theta \mid z, y)$ and $p(z \mid \theta, y)$ are available in the sense that we know how to sample from these distribution, then implementing the Gibbs sampler is straightforward.

8.2 Slice sampler

Suppose we want to simulate from a distribution having the unnormalized density $q(\theta)$. By the fundamental theorem of simulation, this is equivalent to simulating (θ, z) from the uniform distribution under the graph of q, i.e., from $\mathrm{Uni}(A)$, the uniform distribution on the set

$$A = \{(\theta, z) : 0 < z < q(\theta)\}.$$

This distribution has the unnormalized density

$$p(\theta, z) \propto 1_A(\theta, z) = 1_{(0,q(\theta))}(z) = 1(0 < z < q(\theta))$$

The full conditional of Z is proportional to the joint density, considered as a function of z, i.e.,

$$p(z \mid \theta) \propto p(\theta, z) \propto 1(0 < z < q(\theta)),$$

and this an unnormalized density of the uniform distribution on the interval $(0, q(\theta))$.

Similarly, the full conditional of θ is the uniform distribution on the set (depending on z), where

$$1(0 < z < q(\theta)) = 1,$$

since the joint density is constant on this set. That is, the full conditional of θ is the uniform distribution on the set

$$B(z) = \{\theta : q(\theta) > z\}.$$

The resulting Gibbs sampler is called the slice sampler (for the distribution determined by q). The slice sampler is attractive, if the uniform distribution on the set $B(z)$ is easy to simulate.

Example 8.1. Let us consider the truncated standard normal distribution corresponding to the unnormalized density

$$q(\theta) = \exp\left(-\frac{1}{2}\theta^2\right) 1_{(\alpha,\infty)}(\theta),$$

where the truncation point $\alpha > 0$.

We can get a correlated sample $\theta_1, \theta_2, \ldots$ from this distribution as follows.

1. Pick an initial value $\theta_0 > \alpha$.
2. For $i = 1, 2 \ldots$
   - Draw z_i from $\mathrm{Uni}(0, q(\theta_{i-1}))$.
   - Draw θ_i from $\mathrm{Uni}(\alpha, \sqrt{-2 \ln z_i})$.

△

Simulating the uniform in the set $B(z)$ may turn out to be unwieldy. Usually, the target density can be decomposed into a product of functions,

$$p(\theta \mid y) \propto \prod_{i=1}^{n} q_i(\theta).$$

Then one may try the associated augmentation, where one introduces n auxiliary variables Z_i such that, conditionally on θ, the Z_i have independently the uniform distribution on $(0, q_i(\theta))$. In the augmented model, the full conditional of θ is the uniform distribution on the set

$$C(z) = \cap_{i=1}^{n} \{\theta : q_i(\theta) > z_i\},$$

and this may be easier to simulate. Typically, the more auxiliary variables one introduces, the slower is the mixing of the resulting chain.

8.3 Missing data problems

In many experiments the posterior distribution is easy to summarize if all the planned data are available. However, if some of the observations are missing, then the posterior is more complex. Let Z be the missing data and let y be the observed data. The full conditional

$$p(\theta \mid z, y)$$

is the posterior from the complete data, and it is of a simple form (by assumption). Often also the full conditional of the missing data

$$p(z \mid \theta, y)$$

is easy to sample from. Then it is straightforward to use the Gibbs sampler.

Here the joint distribution in the reformulated model is

$$\mathrm{p}_{\mathrm{aug}}(y, \theta, z) = \mathrm{p}_{\mathrm{act}}(\theta)\ \mathrm{p}_{\mathrm{aug}}(y, z \mid \theta).$$

In order to check the equivalence of the original and of the reformulated model, see (8.1), it is sufficient to check that

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

Example 8.2. Let us consider the famous genetic linkage example, where we have the multinomial likelihood

$$p(y \mid \theta) = \mathrm{Mult}\left((y_1, y_2, y_3, y_4) \mid n, \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right)\right).$$

Here $0 < \theta < 1$, and $y = (y_1, y_2, y_3, y_4)$, where the y_j:s are the observed frequencies of the four categories. We take the uniform prior $\mathrm{Uni}(0, 1)$ for θ. The posterior is proportional to

$$q(\theta) = \theta^{y_4}\,(1-\theta)^{y_2+y_3}\,(2+\theta)^{y_1}, \qquad 0 < \theta < 1$$

but thanks to the last factor, this is not of a standard form.

However, suppose that the first category with frequency y_1 is an amalgamation of two subclasses with probabilities $\theta/4$ and $1/2$, but the distinction between the subclasses has not been observed. Let Z be the frequency of the first subclass (with class probability $\theta/4$). Then the frequency of the second subclass (with class probability $1/2$) is $y_1 - Z$. Our reformulated model states that

$$p(z, y \mid \theta) = p(z, y_1, y_2, y_3, y_4 \mid \theta) = \\ \mathrm{Mult}\left((z, y_1 - z, y_2, y_3, y_4) \mid n, \left(\frac{1}{4}\theta, \frac{1}{2}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta\right)\right)$$

Let us check that the reformulated model and the original model are equivalent. If we combine the frequencies X_{11} and X_{12} in the the multinomial distribution

$$(X_{11}, X_{12}, X_2, X_3, X_4) \sim \mathrm{Mult}(n, (p_{11}, p_{12}, p_2, p_3, p_4)),$$

then we obtain the multinomial distribution

$$(X_{11} + X_{12}, X_2, X_3, X_4) \sim \text{Mult}(n, (p_{11} + p_{12}, p_2, p_3, p_4)),$$

and this is obvious when one thinks of the repeated sampling definition of the multinomial distribution. This shows that our original model and the reformulated model are equivalent.

The posterior of θ given the complete data consisting of y and z is given by

$$\begin{aligned} p(\theta \mid y, z) &\propto p(y, z \mid \theta)\, p(\theta) \\ &\propto \left(\frac{1}{4}\theta\right)^z \left(\frac{1}{2}\right)^{y_1 - z} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{1}{4}\theta\right)^{y_4} \\ &\propto \theta^{z+y_4}\,(1-\theta)^{y_2+y_3}. \end{aligned}$$

This is an unnormalized density of the beta distribution $\text{Be}(z+y_4+1, y_2+y_3+1)$, which can be sampled directly.

The full conditional of Z is trickier to recognize. Notice that Z is an integer such that $0 \le Z \le y_1$. It is critical to notice that the normalizing constant of the multinomial pmf $p(z, y \mid \theta)$ depends on z. While you can omit from the likelihood any terms which depend only on the *observed* data, you must keep those terms which depend on the unknowns: parameters or *missing* data.

As a function of z,

$$\begin{aligned} p(z \mid \theta, y) &\propto p(z, y \mid \theta)\, p(\theta) = p(z, y \mid \theta) \\ &= \frac{n!}{z!\,(y_1 - z)!\,y_2!\,y_3!\,y_4!} \left(\frac{1}{4}\theta\right)^z \left(\frac{1}{2}\right)^{y_1 - z} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{1}{4}\theta\right)^{y_4} \\ &\propto \frac{y_1!}{z!\,(y_1 - z)!} \left(\frac{\theta}{4}\right)^z \left(\frac{1}{2}\right)^{y_1 - z} \\ &= \binom{y_1}{z} \left(\frac{\frac{\theta}{4}}{\frac{\theta}{4} + \frac{1}{2}}\right)^z \left(\frac{\frac{1}{2}}{\frac{\theta}{4} + \frac{1}{2}}\right)^{y_1 - z} \left(\frac{\theta}{4} + \frac{1}{2}\right)^{z + y_1 - z} \\ &\propto \binom{y_1}{z} \left(\frac{\theta}{2+\theta}\right)^z \left(1 - \frac{\theta}{2+\theta}\right)^{y_1 - z}, \qquad z = 0, 1, \ldots, y_1. \end{aligned}$$

From this we see that the full conditional of Z is the binomial $\text{Bin}(y_1, \theta/(2+\theta))$, which we also are able to simulate directly. Gibbs sampling in the reformulated model is straightforward. △

8.4 Probit regression

We now consider a regression model, where each of the responses is binary: zero of one. In other words, each of the responses has the Bernoulli distribution (the binomial distribution with sample size one). Conditionally on the parameter vector θ, the responses Y_i are assumed to be independent, and Y_i is assumed to have success probability

$$q_i(\theta) = P(Y_i = 1 \mid \theta),$$

which is a function of the parameter vector θ. That is, the model assumes that

$$[Y_i \mid \theta] \overset{\text{ind}}{\sim} B(q_i(\theta)), \qquad i = 1, \ldots, n,$$

where $B(p)$ is the Bernoulli distribution with success probability $0 \leq p \leq 1$.

We assume that the success probability of the i'th response depends on θ and on the value of the covariate vector x_i for the i'th case. The covariate vector consists of observed characteristics which might influence the probability of success. We would like to model the success probability in terms of a linear predictor, which is the inner product $x_i^T\theta$ of the covariate vector and the parameter vector. For instance, if we have observed a single explanatory scalar variable t_i connected with the response y_i, then the linear predictor could be

$$x_i^T\theta = \alpha + \beta\, t_i, \qquad x_i = (1, t_i), \quad \theta = (\alpha, \beta).$$

Notice that we typically include the constant "1" in the covariate vector.

The linear predictor is not constrained to the range $[0, 1]$ of the probability parameter, and therefore we need to map the values of the linear predictor into that range. The standard solution is to posit that

$$q_i(\theta) = F(x_i^T\theta), \qquad i = 1, \ldots, n.$$

where F is the cumulative distribution function of some continuous distribution. Here F can be called a *link function*. Since $0 \leq F \leq 1$, here $q_i(\theta)$ is a valid probability parameter for the Bernoulli distribution for any value of θ.

In *probit regression* we take $F = \Phi$, where Φ is the cdf of the standard normal $N(0, 1)$, i.e., we assume that

$$q_i(\theta) = P(Y_i = 1 \mid \theta) = \Phi(x_i^T\theta), \qquad i = 1, \ldots, n. \tag{8.2}$$

We can complete the Bayesian model by taking as our prior, e.g., the normal distribution with mean μ_0 and precision matrix Q_0,

$$p(\theta) = N(\theta \mid \mu_0, Q_o^{-1}).$$

An even more popular choice for the link function in binary regression is the logit link, which corresponds to the choice

$$F(u) = \frac{\mathrm{e}^u}{1 + \mathrm{e}^u} = \mathrm{logit}^{-1}(u).$$

The probit and logit regression models belong to the class of generalized linear models (GLMs). The logit link has a special status in binary regression, since the logit link happens to be what is known as the canonical link function. The maximum likelihood estimate (MLE) for probit or logit regression can be calculated with standard software, e.g., using the function `glm` of R.

We can write the likelihood for probit or logit regression immediately, i.e.,

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

where

$$p(y_i \mid \theta) = F(x_i^T\theta)^{y_i} \left(1 - F(x_i^T\theta)\right)^{1-y_i}, \qquad i = 1, \ldots, n.$$

Posterior inference can be based directly on this expression. Gibbs sampling seems impossible, but a suitable MCMC algorithm could be, e.g., the independence sampler with a multivariate Student's t distribution, whose center and

covariance matrix are selected based on the MLE and its approximate covariance matrix, which can be calculated with standard software.

From now on, we will discuss the probit regression model, and its well-known auxiliary variable reformulation, due to Albert and Chib [1]. Let us introduce n latent variables (i.e., unobserved random variables)

$$[Z_i \mid \theta] \overset{\text{ind}}{\sim} N(x_i^T\theta, 1), \qquad i = 1, \ldots, n.$$

This notation signifies that the Z_i's are independent, conditionally on θ. We may represent the latent variables Z_i using n i.i.d. random variables $\epsilon_i \sim N(0,1)$ (which are independent of everything else),

$$Z_i = x_i^T\theta + \epsilon_i, \qquad i = 1, \ldots, n.$$

Consider n RVs Y_i which are defined by

$$Y_i = 1(Z_i > 0) = \begin{cases} 1, & \text{when } Z_i > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Conditionally on θ, the random variables Y_i are independent, Y_i takes on the value zero or one, and

$$P(Y_i = 1 \mid \theta) = P(Z_i > 0 \mid \theta) = P(x_i^T\theta + \epsilon_i > 0) = P(-\epsilon_i < x_i^T\theta) = \Phi(x_i^T\theta).$$

Here we used the fact that $-\epsilon_i \sim N(0,1)$ which follows from the symmetry of the standard normal. Therefore the marginal distribution of $Y = (Y_1, \ldots, Y_n)$ given θ is the same as in the original probit regression model (8.2). Our reformulated model has the structure

$$\mathrm{p}_{\mathrm{aug}}(y, \theta, z) = \mathrm{p}_{\mathrm{act}}(\theta)\ \mathrm{p}_{\mathrm{aug}}(y, z \mid \theta),$$

and we have just argued that

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

This shows that our reformulated model is equivalent with the original probit regression model.

The reformulated probit regression model has the following hierarchical structure,

$$\begin{aligned} \Theta &\sim N(\mu_0, Q_0^{-1}) && (8.3)\\ [Z \mid \Theta = \theta] &\sim N(X\theta, I) && (8.4)\\ Y &= 1_+(Z), && (8.5) \end{aligned}$$

where X is the known design matrix with ith row equal to x_i^T, Z is the column vector of latent variables, and $1_+(Z)$ means the vector

$$1_+(Z) = \begin{bmatrix} 1(Z_1 > 0) \\ \vdots \\ 1(Z_n > 0) \end{bmatrix},$$

where we write $1(Z_i > 0)$ for the indicator $1_{(0,\infty)}(Z_i)$. Therefore we can regard the original probit regression model as a missing data problem where we have a normal regression model on the latent data $Z = (Z_1, \ldots, Z_n)$ and the observed responses Y_i are incomplete in that we only observe whether $Z_i > 0$ or $Z_i \leq 0$.

The joint distribution of the reformulated model can be expressed as

$$p(\theta, y, z) = p(\theta)\, p(z \mid \theta)\, p(y \mid z),$$

where

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

and further

$$p(y_i \mid z_i) = 1(y_i = 1(z_i > 0)) = 1(z_i > 0)\, 1(y_i = 1) + 1(z_i \leq 0)\, 1(y_i = 0).$$

(Y_i is a deterministic function of Z_i. The preceding representation is possible, since Y_i has a discrete distribution.)

The full conditional of θ is easy, since

$$p(\theta \mid z, y) \propto p(\theta, y, z) \propto p(\theta)\, p(z \mid \theta),$$

but this is the same as the posterior for a linear regression model, which is given by a certain multivariate normal distribution $N(\mu_1, Q_1^{-1})$, whose parameters μ_1 and Q_1 depend on the conditioning variables z and y. It is easy to derive expressions for μ_1 and Q_1.

The other full conditional distribution is also easy to derive. As a function of z, we have

$$p(z \mid \theta, y) \propto p(z \mid \theta)\, p(y \mid z) = \prod_{i=1}^{n} N(z_i \mid x_i^T \theta, 1)\, p(y_i \mid z_i)$$

This is a distribution, where the components Z_i are independent, and follow truncated normal distributions, i.e.,

$$\begin{aligned} [Z_i \mid \theta, y] &\sim N(x_i^T \theta, 1)\, 1(Z_i > 0), \qquad \text{if } y_i = 1, \\ [Z_i \mid \theta, y] &\sim N(x_i^T \theta, 1)\, 1(Z_i \leq 0), \qquad \text{if } y_i = 0. \end{aligned}$$

Notice that the side of the truncation for Z_i depends on the value of the binary response y_i. Simulating the full conditional distribution $p(z \mid \theta, y)$ is also straightforward, since we only have to draw independently n values from truncated normal distributions with known parameters and known semi-infinite truncation intervals. Since all the needed full conditional distributions are easily simulated, implementing the Gibbs sampler is straightforward in the latent variable reformulation.

What is the practical benefit of the latent variable reformulation of the probit regression model? In the original formulation of the probit regression model, the components of θ are dependent in their posterior. MCMC sampling will be inefficient unless we manage to find a proposal distribution which is adapted to the form of the posterior distribution. After the reformulation, Gibbs sampling becomes straightforward. In the latent variable reformulation, most of the dependencies in the posterior are transferred to the multivariate normal distribution $p(\theta \mid z, y)$, where they are easy to handle. The components of Z are independent in the other needed full conditional distribution $p(z \mid \theta, y)$.

8.5 Scale mixtures of normals

Student's t distribution with $\nu > 0$ degrees of freedom can be expressed as a scale mixture of normal distributions as follows. If

$$\Lambda \sim \text{Gam}(\nu/2, \nu/2), \quad \text{and} \quad [W \mid \Lambda = \lambda] \sim N(0, \frac{1}{\lambda}),$$

then the marginal distribution of W is t_ν. We can use this property to eliminate Student's t distribution from any statistical model.

Albert and Chib considered approximating the logit link with the t_ν link in binary regression. The logit link is already well approximated by the probit link in the sense that

$$\text{logit}^{-1}(u) \approx \Phi\left(\sqrt{\frac{\pi}{8}}u\right),$$

when u is near zero. Here the scaling factor $\sqrt{\pi/8}$ has been selected so that the derivatives of the two curves are equal for $u = 0$. The approximation is not perfect away from zero. However, if one uses the distribution function F_ν of the t_ν distribution (e.g., with $\nu = 8$ degrees of freedom), then one can choose the value of the scaling factor s so that we have a much better approximation

$$\text{logit}^{-1}(u) \approx F_\nu(su)$$

for all real u. Making use of the scaling factor s, we can switch between a logit regression model and its t_ν regression approximation.

We now consider, how we can reformulate the binary regression model which has the t_ν link, i.e.,

$$[Y_i \mid \theta] \overset{\text{ind}}{\sim} B(F_\nu(x_i^T\theta)), \qquad i = 1, \ldots, n. \tag{8.6}$$

Here the degrees of freedom parameter ν is fixed. Also this reformulation is due to Albert and Chib [1].

The first step is to notice that we can represent the responses as

$$Y_i = 1(Z_i > 0), \quad \text{where} \quad Z_i = x_i^T\theta + W_i, \qquad i = 1, \ldots, n,$$

where $W_i \sim t_\nu$ are i.i.d. and independent of everything else. This holds since

$$P(Z_i > 0 \mid \theta) = P(x_i^T\theta + W_i > 0) = P(-W_i < x_i^T\theta) = F_\nu(x_i^T\theta).$$

Here we used the fact that $-W_i \sim t_\nu$ which follows from symmetry of the t distribution. Besides, the Z_i's are independent, conditionally on θ. Next we eliminate the t_ν distribution by introducing n i.i.d. latent variables Λ_i, each having the $\text{Gam}(\nu/2, \nu/2)$ distribution. If we choose $N(\mu_0, Q_0^{-1})$ as the prior for Θ, then we end up with the following hierarchical model

$$\begin{aligned}
\Theta &\sim N(\mu_0, Q_0^{-1}), && (8.7)\\
\Lambda_i &\overset{\text{i.i.d.}}{\sim} \text{Gam}(\nu/2, \nu/2), \qquad i = 1, \ldots, n && (8.8)\\
[Z \mid \Theta = \theta, \Lambda = \lambda] &\sim N\left(X\theta, [\text{diag}(\lambda_1, \ldots, \lambda_n)]^{-1}\right), && (8.9)\\
Y &= 1_+(Z). && (8.10)
\end{aligned}$$

This reformulation is equivalent with the original model (8.6).

The full conditionals in the reformulated model are easy to derive. The full conditional of θ is a multivariate normal. The full conditional of $\Lambda = (\Lambda_1, \ldots, \Lambda_n)$ is the distribution of n independent gamma distributed variables with certain parameters. The full conditional of Z is, once again, a distribution, where the components are independent and have truncated normal distributions.

Another well-known distribution, which can be expressed as a scale mixture of normal distributions is the Laplace distribution (the double exponential distribution), which has the density

$$\frac{1}{2}\mathrm{e}^{-|y|}, \qquad y \in \mathbb{R}\,.$$

If Y has the Laplace distribution, then it can be expressed as follows

$$V \sim \mathrm{Exp}(1/2) \quad \text{and} \quad [Y \mid V = v] \sim N(0, v).$$

This relationship can be used to eliminate the Laplace distribution from any statistical model.

Even the logistic distribution with distribution function $\mathrm{logit}^{-1}(z)$ can be expressed as a scale mixture of normals, but then one needs the Kolmogorov-Smirnov distribution, whose density and distribution function are, however, available only as series expansions. Using this device, one can reformulate the logistic regression model exactly using the Kolmogorov-Smirnov distribution, multivariate normal distribution and truncation, see Holmes and Held [3] for an implementation of the idea.

8.6 Analytical solutions

When the latent variable has a discrete distribution on a finite set, then the posterior distribution can usually be analyzed analytically.

Let us suppose that $p(z \mid \theta, y)$ is a discrete distribution on a finite set S_Z, and that the support S_Z is the same for all θ. We suppose that the parameter vector θ has a continuous distribution and that $p(\theta \mid z, y)$ is available analytically. Furthermore, we suppose that the support S_Θ of $p(\theta \mid z, y)$ does not depend on the value of z. Then the multiplication rule holds on the Cartesian product $S_\Theta \times S_Z$,

$$p(\theta, z \mid y) = p(z \mid y)\, p(\theta \mid z, y) = p(\theta \mid y)\, p(z \mid \theta, y), \qquad \forall \theta \in S_\Theta, \quad \forall z \in S_Z. \tag{8.11}$$

By the multiplication rule (8.11),

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

and by summing over S_Z we obtain

$$\frac{1}{p(\theta \mid y)} = \sum_{z \in S_Z} \frac{p(z \mid \theta, y)}{p(\theta \mid z, y)}. \tag{8.12}$$

Therefore the (marginal) posterior of θ can be expressed as a finite sum. For instance, this technique gives the exact normalizing constant in front of $q(\theta)$ in

the genetic linkage example, Example 8.2. However, in that example the normalizing constant could be derived more easily be first expanding the awkward term $(2+\theta)^{y_1}$ by the binomial formula

We can express the (marginal) posterior θ in another way by marginalizing z out from the joint posterior (8.11) as follows,

$$p(\theta \mid y) = \sum_{z \in S_Z} p(\theta, z \mid y) = \sum_{z \in S_Z} p(z \mid y)\, p(\theta \mid z, y). \tag{8.13}$$

In order to use this result, we need to be able to evaluate the marginal posterior $p(z \mid y)$.

An obvious attempt would now be to try to reverse the roles of z and θ in the argument which led to (8.12). Doing this we find

$$\frac{1}{p(z \mid y)} = \int_{S_\Theta} \frac{p(\theta \mid z, y)}{p(z \mid \theta, y)}\, \mathrm{d}\theta$$

Unfortunately, this integral is usually intractable. However, the multiplication rule (8.11) comes now to rescue, since

$$p(z \mid y) = \frac{p(\theta_0 \mid y)\, p(z \mid \theta_0, y)}{p(\theta_0 \mid z, y)}, \tag{8.14}$$

where θ_0 is an arbitrary point in S_Θ. Here the functions of θ_0 will cancel each other from the numerator and the denominator, since the left-hand side of the formula does not depend on θ_0. There are several ways one may use this result.

One way of using the representation (8.12) is to pick some $\theta_0 \in S_\Theta$ and then to evaluate the function

$$q(z, \theta_0) = \frac{p(z \mid \theta_0, y)}{p(\theta_0 \mid z, y)}, \qquad z \in S_Z.$$

By (8.12), this is an unnormalized version of $p(z \mid y)$, i.e.,

$$p(z \mid y) = \frac{q(z, \theta_0)}{\sum_{z' \in S_Z} q(z', \theta_0)}. \tag{8.15}$$

Another way of using (8.14) is to first evaluate $p(\theta_0 \mid y)$ by (8.12), since the other terms are readily available.

Since the marginal posterior $p(z \mid y)$ is a discrete distribution on the finite set S_Z and we now know how to calculate its pmf, we can readily simulate this distribution. Therefore we can simulate i.i.d. samples from the joint posterior $p(\theta, z \mid y)$ by the multiplication rule, and by marginalizing we obtain i.i.d. samples from the (marginal) posterior $p(\theta \mid y)$. This idea works as follows.

1. Evaluate $p(z \mid y), z \in S_Z$.
2. Simulate $z_1, \ldots, z_N$ from the pmf $p(z \mid y)$.
3. For $i = 1 : n$ draw θ_i from $p(\theta \mid z_i, y)$.
4. Now $\theta_1, \ldots, \theta_N$ is an i.i.d. sample from $p(\theta \mid y)$.

Tan, Tian and Ng [5] call this *exact IBF sampling*, where the acronym IBF stands for inverse Bayes formula.

It is also possible to reduce certain posterior expectations to finite sums. Suppose that $h(\theta)$ is such a function that the conditional expectation

$$E[h(\Theta) \mid Z = z, Y = y] = \int h(\theta)\, p(\theta \mid z, y)\, \mathrm{d}\theta$$

is available analytically. Then the posterior expectation of $h(\Theta)$ is given by

$$\begin{aligned} E[h(\Theta) \mid Y = y] &= \int h(\theta)\, p(\theta \mid y)\, \mathrm{d}\theta \\ &= \int h(\theta) \sum_{z \in S_Z} p(z \mid y)\, p(\theta \mid z, y)\, \mathrm{d}\theta \\ &= \sum_{z \in S_Z} p(z \mid y) \int h(\theta)\, p(\theta \mid z, y)\, \mathrm{d}\theta \\ &= \sum_{z \in S_z} p(z \mid y)\, E[h(\Theta) \mid Z = z, Y = y]. \end{aligned}$$

This can be evaluated once one has evaluated $p(z \mid y)$ on S_Z.

See [5, Ch. 3–4] for examples where theses approaches have been used successfully.

8.7 Literature

The slice sampler was proposed by Neal [4]. The data augmentation in the genetic linkage example is from the article by Tanner and Wong [6], who borrowed the idea from earlier work on the EM algorithm. The auxiliary variable formulation of probit regression was proposed by Albert and Chib [1]. Also the reformulation of the t link is from this article. Scale mixtures of normals were characterized by Andrews and Mallows [2].

Bibliography

[1] James H. Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

[2] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Series B*, 36:99–102, 1974.

[3] Chris C. Holmes and Leonhard Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168, 2006.

[4] Radford M. Neal. Slice sampling. *Annals of Statistics*, 23:705–767, 2003.

[5] Ming T. Tan, Guo-Liang Tian, and Kai Wang Ng. *Bayesian Missing Data Problems: EM, Data Augmentation and Noninterative Computation.* Chapman & Hall/CRC, 2010.

[6] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

Chapter 9

The EM Algorithm

The EM (Expectation–Maximization) algorithm is an iterative method for finding the mode of a marginal posterior density. It can also be used for finding the mode of a marginal likelihood function. The idea is to replace the original maximization problem by a sequence of simpler optimization problems. In many examples the maximizers of the simple problems can be obtained in closed form.

Often the EM algorithm is applied in an auxiliary variable (latent variable) formulation $p(y, \theta, z)$ of the original model $p(y, \theta)$, where θ is the parameter of interest, and Z is the auxiliary variable (or latent variable or missing data). Then the marginal posterior of θ, namely

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

is the posterior in the original model, and the marginal likelihood of θ, namely

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

is the likelihood in the original model. In such a case the EM algorithm can be used to find the posterior mode or the MLE (maximum likelihood estimate) of the original model.

9.1 Formulation of the EM algorithm

Let Z be the auxiliary variable and θ the parameter of interest. Often the auxiliary variable can be interpreted as missing data. The EM algorithm can be formulated either for the mode of the marginal posterior of θ or for the mode of the marginal likelihood of θ. In both cases one defines a function, usually called Q, which depends on two variables, θ and θ_0, where θ_0 stands for the current guess of the parameter vector θ_0. The function $Q(\theta \mid \theta_0)$ is defined as a certain expected value.

The EM algorithm alternates between two steps: first one calculates the Q function given the current guess θ_0 for the parameter vector (E-step), and then one maximizes $Q(\theta \mid \theta_0)$ with respect to θ in order to define the new guess for θ (M-step). This procedure is repeated until a fixed point of Q is obtained (or some other termination criterion is satisfied). This idea is formalized in

algorithm 19. There arg max denotes the maximizing argument (maximum point) of the function it operates on. If the maximizer is not unique, we may select any global maximizer.

Algorithm 19: The EM algorithm.

Input: An initial value $\theta^{(0)}$.

$k \leftarrow 0$;
repeat
  (E-step) Calculate the function $Q(\theta \mid \theta^{(k)})$;
  (M-step) Maximize $Q(\theta \mid \theta^{(k)})$ with respect to θ:

$$\theta^{(k+1)} \leftarrow \arg\max_{\theta} Q(\theta \mid \theta^{(k)})$$

  Set $k \leftarrow k + 1$
until *the termination criterion is satisfied* ;
Return the last calculated value $\theta^{(k)}$;

Next we define the function Q for the two different objectives. When we want to calculate the mode of the (marginal) posterior density, we define $Q(\theta \mid \theta_0)$ as the expected value of the log joint posterior density, conditioned on the data and on the current value θ_0,

$$\begin{aligned} Q(\theta \mid \theta_0) &= E\left[\log p(\theta, Z \mid y) \mid \theta_0, y\right] \\ &= E\left[\log f_{\Theta,Z|Y}(\theta, Z \mid y) \mid \Theta = \theta_0, Y = y\right] \\ &= \int \log f_{\Theta,Z|Y}(\theta, z \mid y)\, f_{Z|\Theta,Y}(z \mid \theta_0, y)\, \mathrm{d}z. \end{aligned} \tag{9.1}$$

The only random object in the above expected value is Z, and we use its distribution conditioned on the current value θ_0 and the data y.

When we want to calculate the mode of the (marginal) likelihood of θ, we define $Q(\theta \mid \theta_0)$ as the expected complete-data log-likelihood, conditioning on the data and on the current value θ_0,

$$\begin{aligned} Q(\theta \mid \theta_0) &= E\left[\log p(y, Z \mid \theta) \mid \theta_0, y\right] \\ &= E\left[\log f_{Y,Z|\Theta}(y, Z \mid \theta) \mid \Theta = \theta_0, Y = y\right] \\ &= \int \log f_{Y,Z|\Theta}(y, z \mid \theta)\, f_{Z|\Theta,Y}(z \mid \theta_0, y)\, \mathrm{d}z. \end{aligned} \tag{9.2}$$

The Q function is defined as an expectation of a sum of a number terms. Luckily, we can treat all of the terms which do not depend on θ as constants. Namely, in the M-step we select a maximum point of the function $\theta \mapsto Q(\theta \mid \theta_0)$, and the ignored constants only shift the object function but do not change the location of the maximum point. That is, the functions

$$Q(\theta \mid \theta_0) \quad \text{and} \quad Q(\theta \mid \theta_0) + c(\theta_0, y)$$

achieve their maxima at the same points, when the "constant" $c(\theta_0, y)$ does not depend on the variable θ. In particular, we can ignore any factors which depend solely on the observed data y.

The maximization problem (M-step) can be solved in closed form in many cases where the joint posterior (or complete data likelihood) belongs to the exponential family. Then the E- and M-steps boil down to the following steps: finding the expectations (given the current θ_0) of the sufficient statistics (which now depend on the missing data Z), and maximizing the resulting function with respect to the parameters θ.

If the maximizer cannot be solved analytically, then instead of the maximum point one can (in the M-step) select any value $\theta^{(k+1)}$ such that

$$Q(\theta^{(k+1)} \mid \theta^{(k)}) > Q(\theta^{(k)} \mid \theta^{(k)}).$$

The resulting algorithm is then called the generalized EM algorithm (GEM).

We will show later that the logarithm of the marginal posterior

$$\log f_{\Theta|Y}(\theta^{(k)} \mid y)$$

increases monotonically during the iterations of the EM or the GEM algorithms, if one defines Q by (9.1). On the other hand, if one defines Q by (9.2), then the log marginal likelihood

$$\log f_{Y|\Theta}(y \mid \theta^{(k)})$$

increases monotonically during the iterations. If these functions can be calculated easily, then a good check of the correctness of the implementation is to check that they indeed increase at each iteration.

Because of this monotonicity property, the EM algorithm converges to some local mode of the object function (except in some artificially constructed cases). If the object function has multiple modes, then one can try to find all of them by starting the EM iterations at many points scattered throughout the parameter space.

9.2 EM algorithm for probit regression

We return to the latent variable reformulation of the probit regression problem, i.e.,

$$\begin{aligned} \Theta &\sim N(\mu_0, R_0^{-1}) \\ [Z \mid \Theta = \theta] &\sim N(X\theta, I) \\ Y &= 1_+(Z), \end{aligned}$$

where X is the known design matrix, Z is the column vector of latent variables, and $1_+(Z)$ is the vector of indicators $1(Y_i > 0)$. We use the symbols ϕ and Φ for the density and df of the standard normal $N(0,1)$, and use R_0 to denote the precision matrix of the prior.

We have already obtained the distribution of the latent variables given θ and the data, $p(z \mid \theta, y)$. In it, the latent variables Z_i are independent and have the following truncated normal distributions

$$\begin{aligned} [Z_i \mid \theta, y] &\sim N(x_i^T\theta, 1)\, 1(Z_i > 0), \qquad \text{if } y_i = 1, \\ [Z_i \mid \theta, y] &\sim N(x_i^T\theta, 1)\, 1(Z_i \le 0), \qquad \text{if } y_i = 0. \end{aligned}$$

Now the joint posterior is

$$p(\theta, z \mid y) \propto p(y, \theta, z) = p(y \mid z)\, p(z \mid \theta)\, p(\theta).$$

Here $p(y \mid z)$ is simply the indicator of the constraints $y = 1_+(z)$. For any y and z values which satisfy the constraints $y = 1_+(z)$, the log joint posterior is given by

$$\begin{aligned}
\log p(\theta, z \mid y) &= \log p(z \mid \theta) + \log p(\theta) + c_1 \\
&= -\frac{1}{2}(\theta - \mu_0)^T R_0 (\theta - \mu_0) - \frac{1}{2}(z - X\theta)^T (z - X\theta) + c_2 \\
&= -\frac{1}{2}(\theta - \mu_0)^T R_0 (\theta - \mu_0) - \frac{1}{2} z^T z + \theta^T X^T z - \frac{1}{2}\theta^T X^T X \theta + c_2
\end{aligned}$$

where the constants c_i depends on the data y and the known hyperparameters, but not on z, θ or θ_0.

Since now

$$Q(\theta \mid \theta_0) = E[\log p(\theta, Z \mid y) \mid \Theta = \theta_0, Y = y],$$

at first sight it may appear that we need to calculate both the expectations

$$v(\theta_0) = E[Z^T Z \mid \Theta = \theta_0, Y = y], \quad \text{and} \quad m(\theta_0) = E[Z \mid \Theta = \theta_0, Y = y],$$

but on further thought we notice that we actually need only the expectation $m(\theta_0)$. This is so, since the term containing $z^T z$ in $\log p(\theta, z \mid y)$ does not depend on θ. In the maximization of $Q(\theta \mid \theta_0)$ its expectation therefore only shifts the object function but does not affect the location of the maximizer.

Let us next solve the maximizer of $\theta \mapsto Q(\theta \mid \theta_0)$ and then check which quantities need to be calculated. In the following, c_i is any quantity, which does not depend on the variable θ (but may depend on y, θ_0 or the known hyperparameters).

$$\begin{aligned}
Q(\theta \mid \theta_0) &= E[\log p(\theta, Z \mid y) \mid \Theta = \theta_0, Y = y] \\
&= -\frac{1}{2}(\theta - \mu_0)^T R_0 (\theta - \mu_0) - \frac{1}{2}\theta^T X^T X \theta + \theta^T X^T m(\theta_0) + c_3 \qquad (9.3) \\
&= -\frac{1}{2}\theta^T (R_0 + X^T X)\theta + \theta^T \left[R_0 \mu_0 + X^T m(\theta_0) \right] + c_4
\end{aligned}$$

We now make the following observations.

1. The matrix $R_0 + X^T X$ is symmetric and positive definite. Symmetry is obvious, and for any $v \neq 0$,

   $$v^T (R_0 + X^T X) v = v^T R_0 v + v^T X^T X v > 0,$$

   since $v^T R_0 v > 0$ and $v^T X^T X v = (Xv)^T (Xv) \geq 0$.

2. If the matrix K is symmetric and positive definite, then the maximizer of the quadratic form

   $$-\frac{1}{2}(\theta - a)^T K (\theta - a)$$

   is a, since the quadratic form vanishes if and only if $\theta = a$.

3. The preceding quadratic form can developed as

$$-\frac{1}{2}(\theta - a)^T K(\theta - a) = -\frac{1}{2}\theta^T K\theta + \theta^T Ka + \text{constant}.$$

Therefore, the maximum point of

$$-\frac{1}{2}\theta^T K\theta + \theta^T b + c,$$

where K is assumed to be symmetric and positive definite, is

$$\theta = K^{-1}b.$$

(An alternative way to derive the formula for the maximum point is to equate the gradient $-K\theta + b$ of the quadratic function to the zero vector, and to observe that the Hessian $-K$ is negative definite.)

Based on the preceding observations, the maximizer of $\theta \mapsto Q(\theta \mid \theta_0)$ given in eq. (9.3) is given by

$$\theta_1 = (R_0 + X^T X)^{-1}(R_0\mu_0 + X^T m(\theta_0)). \tag{9.4}$$

However, we still need to calculate a concrete formula for the vector

$$m(\theta_0) = E[Z \mid \Theta = \theta_0, Y = y].$$

We need a formula for the expected value of the truncated normal distribution $N(\mu, \sigma^2)1_{(\alpha,\beta)}$ corresponding to the unnormalized density

$$f(v) \propto N(v \mid \mu, \sigma^2)1_{(\alpha,\beta)}(v) \tag{9.5}$$

where we can have $\alpha = -\infty$ or $\beta = \infty$. The moment generating function of this distribution is easy to calculate. Then we obtain its expected value (and higher moments, if need be) by differentiating the result.

Let Φ be the distribution function and ϕ the density function of the standard normal $N(0, 1)$. If V has the truncated normal distribution (9.5), then a simple calculation shows that

$$\begin{aligned} M(t) &= E(\exp(tV)) \\ &= \exp(\mu t + \frac{1}{2}\sigma^2 t^2)\frac{\Phi\left(\frac{\beta-\mu}{\sigma} - \sigma t\right) - \Phi\left(\frac{\alpha-\mu}{\sigma} - \sigma t\right)}{\Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right)} \end{aligned} \tag{9.6}$$

The expected value of a distribution equals the first derivative of its moment generating function at $t = 0$, and hence

$$E[V] = M'(0) = \mu - \sigma\frac{\phi\left(\frac{\beta-\mu}{\sigma}\right) - \phi\left(\frac{\alpha-\mu}{\sigma}\right)}{\Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right)} \tag{9.7}$$

Using the preceding results, we see that the components $m(\theta_0)_i$ of the vector $m(\theta_0)$ are given by

$$m(\theta_0)_i = \begin{cases} x_i^T\theta_0 + \dfrac{\phi(-x_i^T\theta_0)}{1-\Phi(-x_i^T\theta_0)}, & \text{if } y_i = 1 \\ x_i^T\theta_0 - \dfrac{\phi(-x_i^T\theta_0)}{\Phi(-x_i^T\theta_0)}, & \text{if } y_i = 0. \end{cases} \tag{9.8}$$

Formulas (9.4) and (9.8) define one step of the EM algorithm for calculating the posterior mode in probit regression. The EM algorithm for the MLE of probit regression is obtained from formulas (9.4) and (9.8) by setting R_0 as the zero matrix. (Then we need to assume that X^TX is positive definite.)

The truncated normal distribution features in many other statistical models besides the latent variable formulation of probit regression. One famous example is the tobit regression model. This is a linear regression model, where the observations are censored. Since the truncated normal distribution pops up in many different contexts, it is useful to know that there is a simple formula (9.6) for its moment generating function.

9.3 Why the EM algorithm works

The proof of the monotonicity of the EM and GEM algorithms is based on the non-negativity of the Kullback-Leibler divergence. If f and g are two densities, then the K-L divergence (or relative entropy) of g from f is defined by

$$D(f \parallel g) = \int f \, \ln \frac{f}{g}, \tag{9.9}$$

where the integral is calculated over the whole space. If the supports of f and g are not the whole space, then we use the conventions

$$f(x) \, \ln \frac{f(x)}{g(x)} = \begin{cases} 0, & \text{if } f(x) = 0, \\ \infty, & \text{if } f(x) > 0 \text{ and } g(x) = 0. \end{cases}$$

We will show that the K-L divergence is always non-negative. Therefore we can use it to measure the distance of g from f. However, the K-L divergence is not a metric (on the space of densities), since it is even not symmetric.

The proof of the non-negativity can be based on the elementary inequality

$$\ln x \le x - 1 \qquad \forall x > 0, \tag{9.10}$$

where equality holds if and only if $x = 1$. This inequality follows from the concavity of the logarithm function. The graph of a concave function lies below each of its tangents, and right hand side of (9.10) is the tangent at $x_0 = 1$.

Theorem 4. *Let f and g be densities defined on the same space. Then*

$$D(f \parallel g) \ge 0,$$

and equality holds if and only if $f = g$ (almost everywhere).

Proof. We give the proof only in the case, when f and g have the same support, i.e., when the sets $\{x : f(x) > 0\}$ and $\{x : g(x) > 0\}$ are the same (except perhaps modulo a set of measure zero). Extending the proof to handle the general case is straightforward. In the following calculation, the integral extends only over the common support of f and g.

$$\begin{aligned}(-1)D(f \parallel g) &= \int -f \ln \frac{f}{g} = \int f \ln \frac{g}{f} \\ &\leq \int f(\frac{g}{f} - 1) \qquad \text{by (9.10)} \\ &= \int (g - f) = 1 - 1 = 0.\end{aligned}$$

We have equality if and only if

$$\ln \frac{g}{f} = \frac{g}{f} - 1,$$

almost everywhere, and this happens if and only if $f = g$ almost everywhere. □

The following theorem establishes the monotonicity of EM or GEM iterations.

Theorem 5. *Define the function Q by either the equation* (9.1) *or by* (9.2). *Let θ_0 and θ_1 be any values such that*

$$Q(\theta_1 \mid \theta_0) \geq Q(\theta_0 \mid \theta_0). \tag{9.11}$$

Then, with the definition (9.1) *we have*

$$f_{\Theta|Y}(\theta_1 \mid y) \geq f_{\Theta|Y}(\theta_0 \mid y),$$

and with the definition (9.2) *we have*

$$f_{Y|\Theta}(y \mid \theta_1) \geq f_{Y|\Theta}(y \mid \theta_0).$$

In either case, if we have strict inequality in the assumption (9.11), *then we have strict inequality also in the conclusion.*

Proof. We consider first the proof for the definition (9.1). We will use the abbreviated notations, and make use of the identity

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

For any θ, we have

$$\begin{aligned}\ln p(\theta \mid y) &= \int p(z \mid \theta_0, y) \ln p(\theta \mid y) \, \mathrm{d}z \\ &= \int p(z \mid \theta_0, y) \ln \frac{p(\theta, z \mid y)}{p(z \mid \theta, y)} \, \mathrm{d}z \\ &= Q(\theta \mid \theta_0) - \int p(z \mid \theta_0, y) \ln p(z \mid \theta, y) \, \mathrm{d}z\end{aligned}$$

Using this identity at the points θ_1 and θ_0, we obtain

$$\begin{aligned}
&\ln p(\theta_1 \mid y) - \ln p(\theta_0 \mid y) \\
&\qquad = Q(\theta_1 \mid \theta_0) - Q(\theta_0 \mid \theta_0) + \int p(z \mid \theta_0, y) \ln \frac{p(z \mid \theta_0, y)}{p(z \mid \theta_1, y)} \, \mathrm{d}z \\
&\qquad \geq Q(\theta_1 \mid \theta_0) - Q(\theta_0 \mid \theta_0),
\end{aligned}$$

since the K-L divergence is non-negative. This proves the claim for (9.1).

The proof for the definition (9.2) starts from the identity

$$\begin{aligned}
\ln p(y \mid \theta) &= \int p(z \mid \theta_0, y) \ln p(y \mid \theta) \, \mathrm{d}z \\
&= \int p(z \mid \theta_0, y) \ln \frac{p(y, z \mid \theta)}{p(z \mid \theta, y)} \, \mathrm{d}z \\
&= Q(\theta \mid \theta_0) - \int p(z \mid \theta_0, y) \ln p(z \mid \theta, y) \, \mathrm{d}z.
\end{aligned}$$

Rest of the proof is the same as before. □

9.4 Literature

The name EM algorithm was introduced by Dempster, Laird and Rubin in [1]. Many special cases of the method had appeared in the literature already in the 1950's, but this article gave a unified structure to the previous methods. The book [3] is dedicated to the EM algorithm and its variations. Many authors have extended the EM algorithm so that one obtains also the covariance matrix of the (marginal) posterior, or the approximate covariance matrix of the (marginal) MLE, see, e.g., [3] or [2].

Bibliography

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 45:1–38, 1977.

[2] Geof H. Givens and Jennifer A. Hoeting. *Computational Statistics*. Wiley-Interscience, 2005.

[3] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley series in probability and statistics. John Wiley & Sons, Inc., 1997.

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,

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

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. 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.,

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

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 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

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

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

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

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

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

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

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

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

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

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

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

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

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 $g_k(\theta_k \mid y)$, $k = 1, \ldots, K$ 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 (10.18) 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,

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

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 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

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

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.,

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

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, θ_m).
Assumption: The correspondences (10.20) are diffeomorphic.
Result: Proposed next value (k, θ_k) as well as the test ratio r.

1 Draw k from the pmf $\beta(k \mid m)$.
2 **if** $k = m$ **then**
3 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.
4 **else**
5 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.)
6 Calculate θ_k and u_k by the diffeomorphic correspondence specific for the move $m \to k$,

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

7 Calculate r by

$$r \leftarrow \frac{\pi(k, \theta_k)}{\pi(m, \theta_m)} \frac{\beta(m \mid k)}{\beta(k \mid m)} \frac{g(u_k \mid \theta_k, k \to m)}{g(u_m \mid \theta_m, m \to k)} \left| \frac{\partial(\theta_k, u_k)}{\partial(\theta_m, u_m)} \right|$$

8 **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.

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.